Modelado de señales electrofisiológicas para detectar el desempeño cognitivo de niños con TEA

Katherin Gómez Guzmán

Proyecto de Trabajo de Pregrado

Director: PhD (C) Carolina Rico Olarte
Co-director: PhD Diego Mauricio López Gutiérrez

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Departamento de Telemática
Línea de investigación en eSalud
Popayán, marzo de 2022

Katherin Gómez Guzmán

Modelado de señales electrofisiológicas para detectar el desempeño cognitivo de niños con TEA

Trabajo presentado a la
Facultad de Ingeniería Electrónica y Telecomunicaciones
De la Universidad del Cauca, Colombia
Para la adquisición del título académico

Pregrado en:
Ingeniería Electrónica y Telecomunicaciones

Director: PhD (C) Carolina Rico Olarte
Co-director: PhD Diego Mauricio López Gutiérrez

Popayán
2022

Agradecimientos

Structured abstract

Background: Cognitive skills are essential to perform daily activities self-sufficiently. Cognitive problems usually appear in childhood, manifesting in Specific Learning Disorders (Specific Learning Disorders - SLD). SLD are a condition in which academic and cognitive skills are significantly lower than expected for age, considerably affecting school success and daily activities to which the child is exposed. The University of Cauca in collaboration with the Fraunhofer IDMT Institute in Germany developed the HapHop-Fisio system to support rehabilitation therapies for children with SLD based on the collection of physiological signals and cognitive performance data.

Objectives: The main objective of this thesis is to determine the cognitive performance of children with SLD from the physiological signals collected during therapies using the HapHop-Physio system. To achieve this objective, in the first instance, it is necessary to analyze the collected physiological signals to identify incomplete signals. Subsequently, complete signals must be processed with supervised learning techniques for the recognition of cognitive performance. Finally, the model generated must be evaluated from the creation of a confidence index in the classification**.**

Methods: To achieve this objective, two methodologies were used: one for documentary research and another for the analysis and experimentation stage. For the documentary research, systematic mapping and systematic review were carried out. Regarding the analysis and experimentation, the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology was applied, with a special emphasis on the data preparation stage.

Results: This undergraduate thesis produced three important results. Firstly, for the imputation of the missing data, a comparison was made through a statistical analysis of two imputation techniques: a simple one (K nearest neigbors - KNN) and a multiple one (MICE - Multiple Imputation by Chained Equations), in which found that the best way to impute for this case is the simple nearest neighbor method with three neighbors. Second, during the understanding of the data, two main shortcomings were identified in the dataset: unbalanced classes and high dimensionality. To balance the classes, the SMOTE (synthetic minority oversampling technique) algorithm was used, which allowed training more accurate models. To reduce the dimensionality, two approaches were used: with a Wrapper and filter, the first being the one that presented the best results, significantly reducing the number of features to be analyzed without reducing the performance of the classification models. Finally, three algorithms were tested and refined for classification: Random Forest, Support Vector Machines (SVM), and Multilayer Perceptron (MLP). For the tuning of these models, a 10-fold cross-validation was used, observing their accuracy. Once the models were tuned, four metrics were evaluated in the predictions with the test set: accuracy, precision, sensitivity and measure F1. The classifier that presented the best result was SVM.

Conclusions: Although multiple imputation techniques more robust and efficient, the technique that best fitted the data in this project was a simple method. However, the imputed data did not have a positive impact on the classification models. Regarding the cleaning process, even though the class imbalance was not very significant, balancing the dataset had a significant positive impact on the results. Furthermore, although the dimensionality reduction process did not improve the results of the classifiers, it did substantially reduce the number of features to be analyzed and therefore the training and prediction times. Finally, the results of this work contribute significantly to the doctoral thesis of Carolina Rico, from the generation of new workflows and experimentation, to decision-making on the inclusion of other types of physiological signals obtained with the E4 wristband and other wearable devices.

Keywords: Physiological Signal, Specific Learning Disorders, Imputation, Cognitive Performance, Interbeat Interval, Supervised Learning.

Resumen estructurado

Antecedentes: Las habilidades cognitivas son fundamentales para realizar actividades diarias de manera autosuficiente. Normalmente los problemas cognitivos aparecen en la infancia, manifestándose en Trastornos Específicos del Aprendizaje (TEA, en inglés: *Specific Learning Disorders* - SLD). Los TEA son una condición en la cual las habilidades académicas y cognitivas son significativamente más bajas de lo esperado según la edad, afectando considerablemente el éxito escolar y las actividades diarias a las que el niño está expuesto. La Universidad del Cauca en colaboración con el Instituto Fraunhofer IDMT de Alemania desarrolló el sistema HapHop-Fisio para apoyar las terapias de rehabilitación de niños con TEA a partir de la recopilación de señales fisiológicas y datos de desempeño cognitivo.

Objetivos: El objetivo principal de esta tesis es determinar el desempeño cognitivo en niños con TEA a partir de las señales fisiológicas recolectadas durante las terapias apoyadas por el sistema HapHop-Fisio. Para cumplir con dicho objetivo, en primera instancia, es necesario analizar el conjunto de señales fisiológicas para determinar las señales incompletas. Posteriormente, se deben procesar las señales completas con técnicas de aprendizaje supervisado para el reconocimiento del desempeño cognitivo. Finalmente se debe evaluar el modelo generado a partir de la creación de un índice de confianza en la clasificación.

Métodos: Para alcanzar este objetivo se utilizaron dos metodologías: una para la investigación documental y otra para la etapa de análisis y experimentación. Para la investigación documental se realizó mapeo sistemático y revisión sistemática. En cuanto al análisis y experimentación, se aplicó la metodología CRISP-DM (*Cross Industry Standard Process for Data Mining*) haciendo un énfasis especial en la etapa de preparación de los datos.

Resultados: Este trabajo de grado produjo tres resultados importantes. En primer lugar, para la imputación de los datos perdidos se realizó una comparación a través de un análisis estadístico de dos técnicas de imputación: uno simple (vecinos cercanos) y uno múltiple (MICE, *multiple imputation by chained equations*), en el cual se obtuvo que la mejor manera de imputar para este caso es el método simple de vecinos cercanos con tres vecinos. En segundo lugar, durante el entendimiento de los datos se identificaron dos falencias en el conjunto de datos: clases desbalanceadas y alta dimensionalidad. Para balancear las clases se utilizó el algoritmo SMOTE (*synthetic minority oversampling technique*), el cual permitió entrenar modelos más precisos. Para reducir la dimensionalidad se utilizaron dos aproximaciones: con *Wrapper* y con filtro, siendo la primera la que mejores resultados presentó, reduciendo significativamente la cantidad de características a analizar sin reducir el desempeño de los modelos de clasificación. Finalmente, para la clasificación se probaron y afinaron tres algoritmos: Bosques aleatorios (*Random Forest*), Maquinas de vector soporte (SVM, *support-vector machines*) y perceptrón multicapa (MLP, *multilayer perceptron*). Para la afinación de estos modelos se utilizó validación cruzada de 10 pliegues observando su exactitud. Una vez afinados los modelos, se evaluaron cuatro metricas en las predicciones con el conjunto de prueba: exactitud, precisión, sensibilidad y medida F1. El clasificador que presento el mejor resultado fue SVM.

Conclusiones: Aunque en la literatura las técnicas de imputación multiple son más robustas y eficientes, en el caso de este trabajo de grado, la técnica que mejor se ajustó a los datos fue un método simple. No obstante, los datos imputados no tuvieron un impacto positivo en los modelos de clasificación. Con respecto al proceso de limpieza, balancear el conjunto de datos tuvo un impacto positivo en los resultados. Además, aunque el proceso de reducción de dimensionalidad no mejoró los resultados de los clasificadores, sí redujo sustancialmente el tiempo de entrenamiento y predicción. Finalmente, los resultados de este trabajo aportan de manera significativa al trabajo de doctorado de la magíster Carolina Rico, desde la generación de nuevos flujos de trabajo y experimentación, hasta la toma de decisiones sobre la inclusión de otro tipo de señales fisiológicas obtenidas con la pulsera E4 y otros dispositivos wearables.

Palabras Clave: Señal Fisiológica, Trastornos Específicos del Aprendizaje, Imputación, Desempeño cognitivo, *Interbeat Interval*, Aprendizaje Supervisado.

Contenido

Lista de figuras

Lista de tablas

Capítulo 1

Introducción

1.1. Planteamiento del problema

Las habilidades cognitivas como la memoria y la atención son fundamentales para realizar actividades diarias de manera autosuficiente. Los problemas cognitivos normalmente aparecen en la infancia, manifestándose en trastornos escolares y de aprendizaje. Según la encuesta nacional de salud mental realizada por el Ministerio de Salud en 2015 [1], el 21,6% de los niños en Colombia entre 7 y 11 años sufren dificultades de aprendizaje, de los cuales, cerca de la tercera parte sufren de Trastornos Específicos del Aprendizaje (TEA, en inglés: *Specific Learning Disorders - SLD*). Usualmente estos niños son tan inteligentes como sus compañeros, pero se malinterpretan como perezosos o poco inteligentes. Con el seguimiento, tratamiento y apoyo adecuados, estos niños pueden superar estos trastornos y mejorar su rendimiento escolar y social [2].

Los TEA son una condición en la cual las habilidades académicas y cognitivas son significativamente más bajas de lo esperado según la edad [3], afectando considerablemente el éxito escolar y las actividades diarias a las que el niño este expuesto [4]. Las causas de los TEA son muchas y variadas, pero se sabe

relativamente poco acerca de ellas. Sin embargo, se acepta que tienen una base biológica que interactúa con factores genéticos y sociales, tales como la falta de oportunidades para aprender y la calidad de la enseñanza [3]–[5].

Para el diagnóstico y tratamiento de los TEA, el especialista cuenta con varios tipos de pruebas psicométricas o baterías estandarizadas, las cuales analizan las habilidades cognitivas como: razonamiento, aprendizaje, memoria, procesamiento auditivo/visual, comprensión auditiva, expresión verbal y función ejecutiva. De igual forma, la evaluación analiza las habilidades académicas como: lectura, escritura, matemáticas y ortografía [6]. Convencionalmente, la intervención o tratamiento en el niño se planifica a partir del resultado obtenido en la prueba de diagnóstico: el especialista desarrolla una estrategia de aprendizaje para mejorar las habilidades del niño a partir de sus fortalezas cognitivas [7]. Usualmente este tratamiento se centra en mejorar las habilidades cognitivas básicas como la memoria y la atención, por medio de terapias estandarizadas en papel. Estas sesiones normalmente no motivan al paciente debido a la similitud con los métodos utilizados en la escuela, llevando a una temprana finalización de la terapia [8].

La Universidad del Cauca en colaboración con el Instituto Fraunhofer IDMT de Alemania desarrolló el sistema HapHop-Fisio, un proyecto de transferencia tecnológica. Este sistema apoya la rehabilitación de niños con discapacidades intelectuales y cognitivas combinando actividad física con entrenamientos cognitivos. Este programa de rehabilitación altamente motivador se centra en terapias de memoria y atención, destacando su componente auditivo, junto con la práctica de procesos de lectura y escritura [9]. En las pruebas de efectividad terapéutica del sistema, se demostró que todos los niños mostraron un aumento significativo en el rendimiento de la memoria después de 8 semanas de entrenamiento con el sistema; igualmente, se logró la recolección de un conjunto de señales fisiológicas con la pulsera E4 (*wearable*) durante la interacción del niño con el juego con el fin de realizar la inferencia psicofisiológica desde las señales fisiológicas, especialmente de la actividad electrodermal (En inglés: Electrodermal Activity - EDA) hacia algunos aspectos cognitivos medidos a partir de los resultados del sistema, como el desempeño y la carga cognitiva [10].

HapHop-Fisio tiene como propósito apoyar al especialista en el reconocimiento de los cambios cognitivos (evolución) de los niños con TEA. El reconocimiento se da a partir

de técnicas de aprendizaje supervisado aplicadas a las señales fisiológicas procesadas, especialmente de la EDA, para llegar al reconocimiento del aspecto psicológico denominado “desempeño cognitivo” en los dominios de memoria y atención. El resultado preliminar en la validación de la clasificación tuvo una precisión del 80% [11]. Este porcentaje de clasificación puede estar asociado a la incorrecta colocación de la pulsera en la muñeca del paciente, a los movimientos de gran impacto durante su utilización [12] y principalmente, a la falta de análisis de las otras señales que se recolectaron durante el proceso, como la temperatura, la acelerometría y la frecuencia cardíaca, la cual se logra obtener a partir del volumen sanguíneo gracias a un algoritmo propio del software de la pulsera E4.

El reconocimiento de cambios cognitivos a partir del análisis de la señal EDA sirve como soporte a los especialistas en neuropsicopedagogía para evaluar el rendimiento cognitivo de los niños durante el juego, pero es necesario el análisis de más señales para que el sistema sea capaz de identificar si el paciente presenta un avance positivo durante el tratamiento. Posiblemente las primeras causas que no permitieron llegar a un porcentaje óptimo de clasificación infirieron en la obtención completa del ritmo cardiaco en la mayoría de las sesiones. En el 57,14% de los casos, la señal de frecuencia cardíaca está completa sólo entre un 75% y un 99% y entre el 50% y 75% en el 18,18% de las sesiones; 7,79% de las sesiones de grabación están completas en un 25% a 50%, pero el 16,23% de las sesiones no presentan señal de frecuencia cardíaca. Por lo tanto, es necesario corregir las señales que están incompletas para proporcionar un conjunto de datos más amplio, lo que permitirá que un análisis de una mayor cantidad de información brinde un porcentaje de clasificación mayor al obtenido en la anterior investigación. El aumento de este porcentaje de clasificación debe garantizar al especialista que el modelo tiene un mejor desempeño al momento de clasificar los procesos cognitivos de los niños.

Con base en lo anterior, se identifica la necesidad de generar un índice de confianza respecto a los cambios cognitivos del niño durante su interacción con el juego, con el fin de garantizar progresos favorables del paciente al especialista o profesional de la salud y que eventualmente, los niños con TEA tengan una adecuada intervención. Por esta razón, se plantea la siguiente pregunta de investigación:

¿Cómo regenerar las señales incompletas de frecuencia cardíaca para realizar la clasificación por aprendizaje supervisado de las señales fisiológicas

completas y recolectadas de HapHop-Fisio para detectar el desempeño cognitivo de niños con TEA?

1.2. Motivación

El sistema HapHop-Fisio apoya a niños en edad escolar con TEA en su proceso de rehabilitación por medio de entrenamientos cognitivos. Igualmente, el proyecto ayuda al especialista en el reconocimiento de los cambios cognitivos del niño, por medio de un conjunto de señales fisiológicas recolectadas durante la interacción del paciente con el juego; estas señales permiten medir aspectos cognitivos como el desempeño y la carga cognitiva durante el juego, pero a causa de varios factores externos los datos recolectados no están completos lo cual no permite realizar un análisis eficaz de estas señales para obtener resultados más reales y certeros del avance en el tratamiento del paciente. Por esta razón es necesario completar estas señales.

De acuerdo a diferentes metodologías de análisis de datos y algoritmos de procesamiento, los datos incompletos representan una problema para llevar a cabo tareas de clasificación con alta precisión [13]. Igualmente, la pérdida de datos en proyectos e investigaciones es una realidad que trae consigo otros inconvenientes cómo la perdida de validez o confiabilidad del análisis de estos datos, colocando en riesgo el resultado de la investigación [14]. Así mismo, no es viable realizar la eliminación de los datos de aquellos pacientes que presentan señales incompletas ya que esto puede conducir a resultados sesgados y poco fiables. De hecho, un estudio mostró que la eliminación por lista conduce a una disminución en el poder estadístico entre el 35% (con el 10% de los datos perdidos) y el 98% (con el 30% de los datos perdidos) [15]. Completar los datos faltantes del sistema HapHop-Fisio permitirá aumentar el porcentaje de precisión en el análisis de las señales, lo cual sirve como soporte a los especialistas en neuropsicopedagogía para identificar si el paciente presenta un avance positivo durante el tratamiento.

1.3. Objetivos

1.3.1. Objetivo general

Determinar el desempeño cognitivo en niños con TEA a partir de un análisis multimodal [1] de las señales fisiológicas periféricas.

1.3.2. Objetivos específicos [2]

1. Analizar el conjunto de señales fisiológicas que reportan los niños con TEA durante su interacción con el juego HapHop-Fisio para determinar las señales incompletas.
2. Completar por medio de un proceso de imputación las señales incompletas detectadas.
3. Procesar las señales completas con técnicas de aprendizaje supervisado para el reconocimiento del desempeño cognitivo en niños con TEA.
4. Evaluar el modelo generado a partir de la creación de un índice de confianza en la clasificación.

[1] Analisis multimodal comprende todas las señales de la manilla E4 tanto completas como impcompletas.
[2] Despues de evaluar las actividades para lograr el objetivo general fue necesario agregar otro objetivo especifico que permitiera dimensionar las contribuciones de esta tesis.

1.4. Contribuciones

Este proyecto de investigación implementó un modelo de datos completos capaz de obtener un alto índice de confianza en los resultados medidos por el sistema HapHop-Fisio durante las sesiones de rehabilitación de los niños que sufren de TEA, con el fin de identificar los cambios cognitivos generados en el niño durante la interacción humano-computador.

Este proyecto es importante para (1) el área de dominio (TICs para la salud), ya que se realizó un aporte significativo al completar y proveer un dataset de señales fisiológicas segmentadas y clasificadas para el reconocimiento del desempeño cognitivo a partir de la inserción de valores inexistentes en las señales medidas; (2) el área de aplicación (salud) para tratamientos y empresas dedicadas a la rehabilitación neuropsicológica infantil, ya que el proyecto dará soporte al profesional de la salud realizando un seguimiento objetivo de los avances del paciente con TEA durante las sesiones de rehabilitación y de esta forma, tratar de garantizar una mejor calidad de vida a los pacientes.

Por último, este trabajo de pregrado realizó una contribución a la línea eSalud del Grupo de Ingeniería Telemática al construir un conjunto de datos robusto y de calidad, basado en las señales fisiológicas obtenidas por medio de la pulsera E4, durante la interacción con el producto HapHop-Fisio, producto también asociado al GIT.

1.5. Metodología general del trabajo

La metodología CRISP-DM (*Cross-Industry Standard Process for Data Mining*) [16] describe el ciclo de vida de un proyecto de ciencia de datos que comprende seis fases (Figura 1). Es un conjunto de pautas y guías para ayudar en la planificación, organización e implementación de proyecto de aprendizaje automático como el de este trabajo de grado.

La fase de entendimiento del negocio se centra en la comprensión de los objetivos y requisitos del proyecto. Es importante establecer un sólido entendimiento ya que puede considerarse como construir los cimientos de una casa: absolutamente esencial. Dentro de este proyecto de investigación, el planteamiento del problema junto con la descripción de los conceptos base y el levantamiento del estado del arte es parte de los resultados de esta primera fase.

La fase del entendimiento de los datos comprende el enfoque para identificar, recopilar y analizar los conjuntos de datos que pueden ayudar a lograr los objetivos del proyecto. Este entendimiento se ve reflejado en los procesos y resultados descritos en la sección 3.1 de este documento.

La preparación de datos, a la que a menudo se hace referencia como "recopilación de datos", prepara los conjuntos de datos finales para el modelado. Una regla general común es que el 80% del proyecto es la preparación de datos [17]. En este trabajo de grado se realizaron dos procesos de preparación de datos que fueron complementarios para la obtención final de los conjuntos de datos. Los datos fuente del primer proceso fueron las señales fisiológicas y la preparación de este tipo de datos está descrito en la sección 3.2. El segundo proceso prepara las señales y determina sus características para ser convertidas en los conjuntos de datos a utilizar en el modelado; estos procesos son detallados en la sección 4.1.

En la fase de modelado se construyen y validan varios modelos basados en diferentes técnicas de modelado. Esta fase se considera como la más emocionante de la ciencia de datos, aunque también suele ser la fase más corta del proyecto. Para este proyecto, los resultados de modelado se presentan en la sección 4.2.

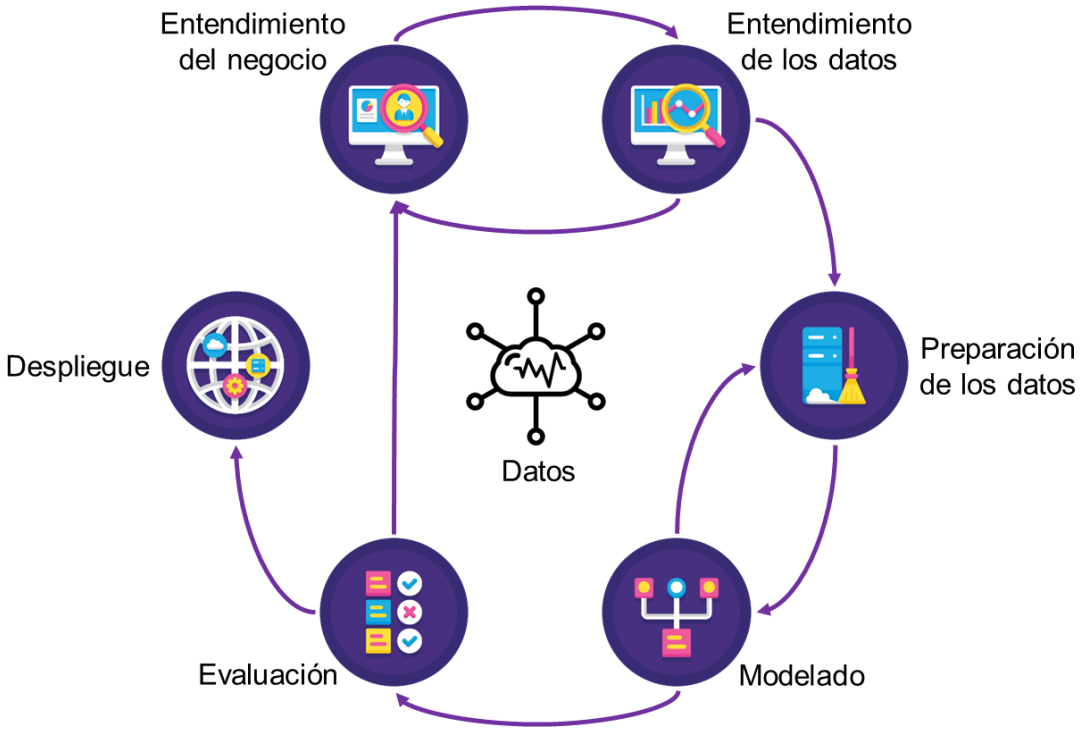

Figura 1: Metodología CRISP-DM [16]

Mientras que la tarea de validación del modelo de la fase de modelado se centra en la evaluación del modelo técnico, la fase de evaluación analiza de manera más amplia qué modelo se adapta mejor al negocio y qué hacer a continuación [18]. Respecto a esta fase, en la sección 4.3 se encuentran las conclusiones de esta fase, especialmente, en relación con la generación del índice de confianza en el reconocimiento del desempeño cognitivo.

Finalmente, la fase de despliegue es necesaria ya que un modelo no es particularmente útil a menos que el cliente pueda acceder a sus resultados. Sin embargo, esta fase no se encuentra dentro del alcance de los objetivos de este proyecto. No obstante, en los trabajos futuros (sección 5.2) si se consideran unas pautas para la culminación de un proyecto como el que se presenta aquí.

1.6. Contenido del documento

Este documento monográfico se divide en 5 capítulos, los cuales se describen a continuación:

1.6.1 Capitulo 1. Introducción

En este capítulo se presenta el problema planteado en la investigación, la razón por la cual se decidió abordar este problema, los objetivos propuestos para hallar una solución al mismo, las contribuciones que dejaran este proyecto y el contenido de este documento.

1.6.2. Capítulo 2. Estado del arte

En este capítulo se presenta una breve descripción de conceptos relevantes para esta tesis. Además, se analizan diferentes proyectos de investigación que trabajan en aplicaciones utilizadas para generar un diagnostico o detectar niños con TEA para establecer nuestras brechas de investigación.

1.6.3. Capítulo 3. Caracterización e imputación de señales

En este capítulo se describen las señales fisiológicas que se van a utilizar en la investigación, se identifican sus características por medio del análisis descriptivo estadístico y el ajuste de la distribución de probabilidad para finalmente imputar los datos faltantes de las señales identificadas.

1.6.4. Capítulo 4. Preparación y modelado de datos

En este capítulo son presentados dos procesos complementarios en la preparación de los datos para conformar los conjuntos de datos que fueron utilizados para generar los modelos de clasificación que permiten realizar el reconocimiento del desempeño cognitivo de los niños con TEA a partir del análisis multimodal de las señales fisiológicas obtenidas con la pulsera E4.

1.6.5. Capítulo 5. Conclusiones y trabajos futuros

Finalmente, desde el proceso de imputación de las señales identificadas como incompletas hasta la generación y evaluación de los modelos de clasificación, se obtuvieron conclusiones importantes e ideas para trabajos futuros para continuar con el desarrollo de la investigación en la detección de patrones cognitivos a partir de fuentes objetivas como las señales fisiológicas.

Capítulo 2

Estado del arte

En este capítulo, se presenta la descripción de los conceptos más importantes para tener en cuenta en esta tesis. Se describe el análisis de los trabajos relacionados y las brechas de investigación encontradas a partir de los estudios que relacionan aplicaciones utilizadas para generar un diagnostico o detectar niños con TEA con técnicas de aprendizaje supervisado.

2.1. Antecedentes

Con el objetivo de ofrecer un trasfondo general para este trabajo de grado, esta sección incluye la explicación de un concepto clave en esta investigación y cuatro conceptos compuestos (construidos a partir de otros conceptos) que rodean el contexto del proyecto.

2.1.1. Tecnologías de la Información y la Comunicación (TIC) en salud

Las TIC comprenden desarrollos relacionados con computadores, internet, telefonía, aplicaciones multimedia, realidad virtual y básicamente con tecnologías que proporcionan información y canales de comunicación, brindando un beneficio a todos los campos del saber [19].

Particularmente, el uso de las TIC en el área de la salud tiene como objetivo principal el mejoramiento de la calidad de vida de la población urbana y rural, esto gracias a la cantidad de herramientas que han permitido mayor acceso a este servicio, tales como consultas médicas en tiempo real asignadas o realizadas desde casa con profesionales de la salud de manera remota, apoyo a la toma de decisiones para algunas enfermedades, acceso e intercambio de información de fuentes confiables y control de información del paciente para brindarle un servicio más eficiente y cómodo, es decir de mejor calidad [20].

2.1.2. Cognición

La cognición es la capacidad de obtener información del entorno, procesarla e interpretarla. Los procesos cognitivos dependen tanto de las capacidades sensoriales como del sistema nervioso central e involucra el uso de habilidades mentales como la percepción (captación de los estímulos externos e internos a través de los sentidos), la atención (enfocar habilidades mentales en la información que está recibiendo), el aprendizaje (adquisición de conocimientos nuevos) y la memoria (capacidad de almacenar, codificar y recuperar esa información), el lenguaje (oral, escrito o gestual), la emoción (proceso similar al de la cognición), el razonamiento y la solución de problemas (el razonamiento permite evaluar la información obtenida y facilita la identificación de soluciones), hasta la metacognición (conciencia que el sujeto desarrolla sobre su propio proceso cognitivo) [21].

2.1.3. Señales fisiológicas

Las señales fisiológicas son un medio de transmisión de información de los sistemas fisiológicos del organismo, es decir, de todos los órganos y sistemas que forman nuestro cuerpo. Estas señales permiten extraer información para describir el funcionamiento del cuerpo y emitir un diagnóstico [22].

El principal propósito de las señales fisiológicas ha sido ayudar en el diagnóstico clínico o detección de patologías en los pacientes; de igual forma la ciencia ha empezado a utilizar estas señales para tratar de entender el comportamiento del ser humano en diferentes situaciones, gracias al análisis de procesos cognitivos que son realizados por el ser humano en el momento de tomar decisiones o experimentar emociones tanto positivas como negativas. De esta forma es posible evaluar el impacto que puede producir en una persona la interacción con distintos dispositivos computacionales y así establecer las características y propiedades que debería tener un producto o servicio para que sea beneficioso o atractivo al cliente [23], [24].

2.1.4. Reconocimiento cognitivo a partir de señales fisiológicas

La psicofisiología busca hallar relación entre las señales fisiológicas y los eventos psicológicos, por esta razón desde esta área se ha tratado de hacer el reconocimiento cognitivo a partir de las señales fisiológicas por medio de una relación de inferencia, es decir que a partir de señales se pueda deducir un evento psicológico ya sea emocional/afectivo, cognitivo o comportamental [22]; para identificar estas emociones y comportamientos que permiten adquirir información del usuario se debe recurrir a la computación fisiológica, es decir al procesamiento, análisis y clasificación de las señales fisiológicas de una persona, dando lugar a la representación del estado psicofisiológico del usuario para que el software pueda responder de forma dinámica y específica [25].

2.1.5. Aprendizaje automático

El aprendizaje automático o mejor conocido como Machine Learning (ML) es un tipo de inteligencia artificial (En inglés: *Artifical Intelligence - AI*) encargada de desarrollar técnicas por las cuales los computadores tengan la capacidad de aprender. El proceso de ML consiste en crear algoritmos que generalicen comportamientos y reconozcan patrones para ajustar las acciones del programa en consecuencia a estos, sin haber sido programados específicamente para ello [26], [27].

Los algoritmos del aprendizaje automático se clasifican comúnmente como supervisados, no supervisados o por refuerzo. En el aprendizaje supervisado, los algoritmos pueden aplicar lo que se ha aprendido en el pasado a nuevos datos; mientras que los algoritmos no supervisados pueden reconocer patrones para dividir los datos en grupos que posean características similares entre sí; por último, en el aprendizaje por refuerzo, los algoritmos aprenden observando el mundo que los rodea y refuerzan aquellas acciones de las cuales reciben una respuesta positiva [27], [28].

2.2. Trabajos relacionados

Para proporcionar información acerca de las aplicaciones utilizadas para generar un diagnostico o detectar niños con TEA, se realizó una búsqueda en la plataforma SCOPUS, de la cual se seleccionaron los artículos más relevantes que contenían elementos cercanos a una posible solución del problema planteado en este proyecto. Estos trabajos sirvieron como referencia para definir las brechas de investigación y las contribuciones de este trabajo de grado. Los resultados finales se clasificaron en las temáticas que se presentan a continuación:

2.2.1. TIC para diagnóstico de los TEA

Los trabajos de investigación encontrados tratan acerca de herramientas software [29] y aplicaciones web [30] para el diagnóstico, las cuales evalúan una habilidad determinada en los niños por medio de pruebas en forma de juegos [29] y tests [30]. Estas herramientas ayudan en la detección de problemas en el área específica en la cual son evaluados como la memoria [29] o el lenguaje [30]. Así mismo, existen proyectos basados en aprendizaje automático capaces de identificar trastornos relacionados con la cognición como el espectro autista [31], déficit de atención, trastorno de hiperactividad, trastorno de conducta [32] y trastorno del habla y del lenguaje [33]; estos proyectos utilizan redes neuronales [31], [33] y el algoritmo Support Vector Machine (SVM) de aprendizaje automático supervisado [32] para analizar las respuestas de los tests [32], [33] y de las señales fisiológicas obtenidas del movimiento ocular del paciente mientras se trata de comprender un video de concentración [31].

Por otra parte, es importante resaltar la existencia de un sistema de evaluación psicométrico computarizado que identifica las fortalezas y dificultades cognitivas de los niños de jardín y primaria por medio de pruebas en forma de juego, las cuales evalúan memoria secuencial y asociativa, discriminación auditiva y de color, y conciencia fonológica. El programa es capaz de mostrar el nivel de desarrollo del niño en las áreas relevantes y reconocer debilidades en varios aspectos cognitivos a partir de las respuestas del paciente. Sin embargo como en la mayoría de las investigaciones, el diagnóstico se basa en un puntaje que el paciente ha obtenido en un test o prueba, sin realizar evaluaciones más objetivas como las inferencias psicofisiológicas [29].

2.2.2. Reconocimiento cognitivo a partir de señales fisiológicas

Una pequeña parte de los trabajos de investigación encontrados hablan acerca de la creación de dispositivos para la lectura del EDA [34]–[36] o de la inducción de reacciones corporales para inferir estados emocionales [37]. Por el contrario, fue común encontrar el uso de dispositivos disponibles comercialmente para la recolección de una o varias señales fisiológicas tales como electrocardiograma (en inglés:

Electrocardiogram - ECG), fotopletismografia (en inglés: *Photoplethysmography - PPG*), electromiografía (en inglés: *electromyography - EMG*), respiración (en inglés: *respiration - RESP*), EDA y SKT (*Skin Temperature*). Las señales fueron recolectadas mientras el paciente era sometido a pruebas cognitivas y/o estímulos emocionales por medio de juegos, videos o imágenes afectivas [34], [36], [38]–[46]. De igual forma, existen investigaciones que recolectaron los datos a partir de la interacción humano–computadora [43], humano-robot [47], [48], niño-adulto [49], o simplemente se monitorizó al paciente durante un día de actividad ordinaria [35]. Los resultados más interesantes de estas investigaciones se exponen a continuación:

- El nivel de conductancia de la piel se mantenía alta durante las tareas cognitivas y momentos de estrés, porque la conductividad eléctrica de la piel permite evaluar la actividad del Sistema Nervioso Central y Periférico, y así, reflejar el estado emocional del paciente. De esta forma, la conductancia de la piel se mantiene alta cuando el nivel de estrés aumenta y disminuye cuando el paciente entra en un estado de relajación [34]. Se ha evidenciado que un niño con un grado de estrés alto no continúa la terapia, por esta razón es importante identificar estos aspectos durante las sesiones para ayudar al especialista a encontrar la forma correcta de realizar el tratamiento en cada paciente.

- La amplitud de la señal de flujo sanguíneo (*Blood Volume Pulse – BVP*) disminuye en el momento que se genera ira y disgusto en el usuario por medio de videos, igualmente los datos de EDA presentan un aumento en su amplitud del nivel de conductancia de la piel (en inglés: *Skin Conductance Level- SCL*) con la misma emoción, pero en el caso de tristeza se encontró que la amplitud de la señal EDA disminuye gradualmente para el SCL y la respuesta de conductancia de la piel (en inglés: *Skin Conductance Responses - SCR*) [44]. Con este precedente es posible plantear una relación entre la señal BVP y EDA, teniendo en cuenta que estas señales tienen una relación inversa en cuanto a la ira y disgusto, y directa en cuanto a la tristeza.

2.2.3. Análisis de datos

Procesamiento de señales fisiológicas

A partir de los artículos encontrados fue posible identificar el uso frecuente de las siguientes técnicas para procesar la señal:

- Como primera medida se elimina todo tipo de contaminantes que pueden afectar la señal, entre estos encontramos el ruido eléctrico, y los artefactos de movimiento [44], [48], las metodologías comúnmente utilizadas son: segmentación [50] y filtrado; los filtros usualmente implementados son: filtro de paso bajo, filtro paso alto, filtro pasa banda, filtro Hanning, algoritmo de Hamilton y el algoritmo de gamboa [34], [35], [39], [41], [44], [49], [50].

- A continuación, se extraen características de la señal; su cantidad puede variar por el número y tipos de señales obtenidas [48] y estas pueden ser estadísticas o geométricas [47]. Existen diferentes técnicas para la extracción, en la bibliografía se destacan: ventanas temporales parametrizadas para calcular los ángulos de la variación electrodermal local [38], la transformada discreta de Wavelet para detectar bordes y cambios bruscos de la señal [39], [41] y un diferenciador de primer orden para generar una señal que varía alrededor del origen [35].

Aprendizaje automático a partir de señales fisiológicas

Inicialmente, se realiza un procedimiento para seleccionar algunas de las características extraídas, esto con el propósito de descartar datos que no sean relevantes para el parámetro que se busca relacionar, y de esta forma, mejorar el tiempo y la eficiencia de la memoria computacional gracias a la reducción de la dimensionalidad [39], [41], [48]. En algunos de los trabajos de investigación se utilizó el método de selección por Wrapper, el cual evalúa repetidamente los conjuntos de características o atributos utilizando un clasificador para seleccionar un conjunto independiente de características que produzcan el mejor rendimiento [39], [41].

Posteriormente, se selecciona un algoritmo de aprendizaje de clasificación. Un 4,5% de las investigaciones encontradas utilizan el aprendizaje no supervisado [48], la mayoría de las investigaciones que están relacionadas con la implementación del ML sobre señales fisiológicas utilizan métodos de aprendizaje supervisado. A continuación, se describen las técnicas más relevantes implementadas en cada uno de los tipos de aprendizaje:

- En los métodos de aprendizaje supervisado se halló una amplia variedad de clasificadores, tales como: LDF (*Linear Discriminant Function*), CART (*Classification and Regression Tree*), SOM (*Self Organizing Map*), Naive Bayes, SVM, k-NN (*Nearest Neighbour*), redes neuronales (*Neural Networks*), árbol de decisión como Random Forest y regresión lógica [39], [41], [43], [44], [46], [49], [50]. Estos algoritmos de clasificación se implementaron para el reconocimiento de emociones (felicidad, tristeza, miedo, sorpresa, disgusto, ira, dolor, aburrimiento y estrés) [43], [44], [46].
  Por medio de pruebas se identificó que el porcentaje más alto de precisión se obtuvo con SVM, el cual registró porcentajes entre 99.04% y 100%, a diferencia de CART con un valor entre 74,93% y 93,3%, Naive Bayes con un valor entre 66,44% y 83,4%, SOM con 37,67%, k-NN con 31% y LDF con 30,14% [43], [44], [46].

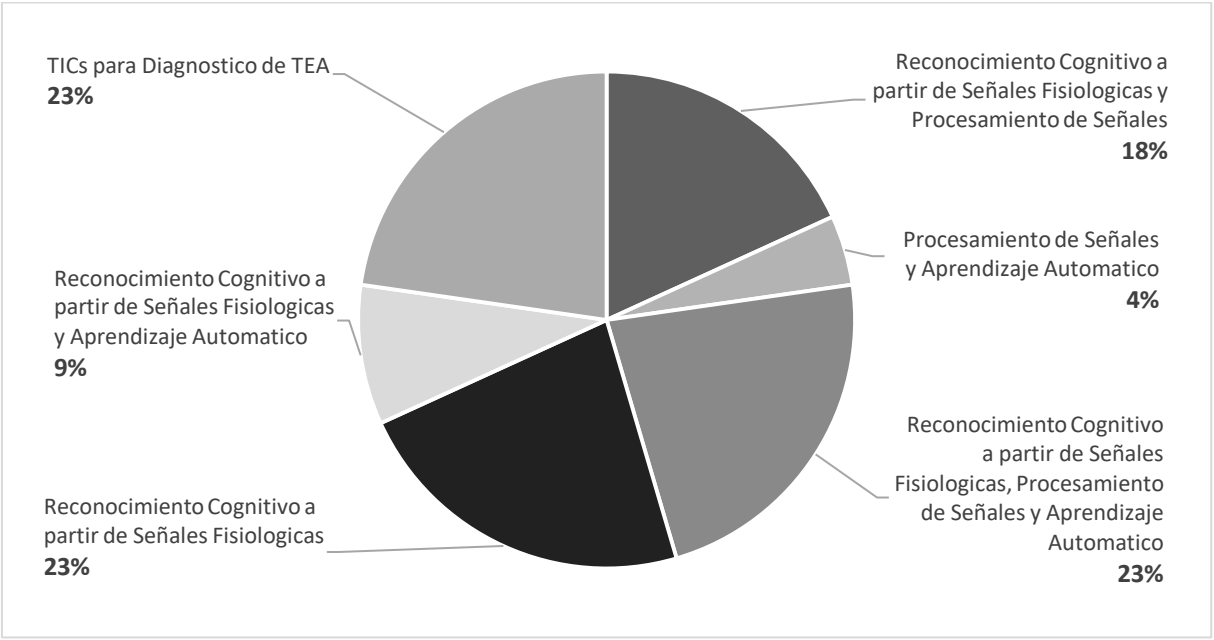

Figura 2. Artículos de trabajos relacionados

2.2.4. Brechas

Con base en los resultados obtenidos de las investigaciones encontradas y los artículos revisados (Figura 2), se identificaron las siguientes brechas:

- Varios de los proyectos recolectan y analizan la señal fisiológica EDA [34], [35], [39], [48] porque consideran que está estrechamente ligada a la parte cognitiva del ser humano; igualmente, un gran porcentaje de los trabajos realizan la recolección y análisis de otras señales como HR (*Heart Rate*), EMG, ECG, BVP, RESP, PPG, SKT [38], [40], [43], [44], pero estas investigaciones tienen como objetivo principal el reconocimiento de emociones, no se habla de reconocimiento de la cognición y/o procesos cognitivos.

- La mayoría de los trabajos se centran en la fase de reconocimiento de emociones [35], [37], [38], [40], [42]–[46], [48], y otros pocos en el reconocimiento de la cognición y/o procesos cognitivos a partir de señales fisiológicas [9], pero no se discuten estrategias de inserción de datos a las señales incompletas para obtener un mejor conjunto de datos en el inicio del proyecto y que finalmente permita aumentar el porcentaje de precisión.

- La mayoría de las investigaciones son evaluadas en pacientes sanos o sin un diagnóstico de tipo cognitivo [34], [35], [37]–[47], [49]; algunas incluyen pacientes con trastornos cognitivos como niños con autismo [50]. Sin embargo, no se encontró registro de estudios de tipo comparativo en niños con TEA.

- Los trabajos no implementan proyectos TIC para tratar pacientes con TEA, como el sistema HapHop-Fisio. Igualmente, no se realiza un análisis de señales fisiológicas para evaluar el progreso durante la intervención. La mayoría de los proyectos que utilizan TIC solo se centran en detectar los trastornos como el espectro autista [31], déficit de atención, trastorno de conducta y trastorno de hiperactividad [32]; o bien sea solo en un problema específico del aprendizaje como la memoria de trabajo [29] y el lenguaje [30], [33].

2.3. Resumen

En este capítulo, se describió un concepto simple de cognición y conceptos compuestos de Tecnologías de la información y la comunicación (TIC) en salud, señales fisiológicas, reconocimiento cognitivo a partir de señales fisiológicas y aprendizaje automático; estos conceptos básicos se definen para que el lector comprenda el objetivo final del trabajo de grado, con el que se buscó llegar a un reconocimiento cognitivo, por medio del análisis con técnicas de aprendizaje automático de las señales fisiológicas obtenidas durante el tratamiento del paciente.

Además, a través del estudio de los resultados obtenidos de la búsqueda en la plataforma SCOPUS, se halló proyectos en los que se crean herramientas software y aplicaciones web para el diagnóstico de trastornos y problemas cognitivos. Igualmente se encontraron algunos trabajos en los que había un acercamiento al reconocimiento cognitivo a partir de señales fisiológicas. Particularmente, se deseaba identificar métodos de análisis de datos como el aprendizaje automático en proyectos que estuvieran orientados al reconocimiento cognitivo o específicamente al diagnóstico de niños con TEA. Finalmente, se especificaron las brechas identificadas en los artículos hallados en la investigación.

Teniendo en cuenta que el principal problema que se afronta en este proyecto son los datos incompletos de la señal fisiológica recolectada HR, se tomó la decisión de investigar inicialmente las características de las señales fisiológicas y el posible método para completarlas, con el fin de obtener una base integral de conceptos en el primer aporte de esta tesis de pregrado, los resultados serán expuestos en el próximo capítulo.

Capítulo 3

Caracterización e imputación de señales

En este capítulo se hará una breve introducción a las señales fisiológicas recolectadas por la pulsera E4 y que serán utilizadas en el desarrollo de este trabajo de grado. Además, se hará un reconocimiento estadístico de estas señales y un ajuste en su distribución de probabilidad para finalmente, encontrar la forma más acertada de organizarlas y completar los datos de la señal IBI, la señal con más datos perdidos.

3.1. Caracterización de las señales fisiológicas

Para realizar las pruebas del juego HapHop-Fisio, se contó con un grupo de 14 pacientes monitorizados con la pulsera E4 durante cada sesión. En total, se recolectaron los datos de 154 sesiones dónde cada paciente realizó entre 9 y 17 sesiones grabadas con una duración que osciló entre 13 y 33 minutos cada una, dependiendo del criterio del profesional de salud que acompañó las sesiones de juego. El tiempo total de grabación fue de 61 horas, 42 minutos y 25 segundos, en los que se logró recolectar las señales fisiológicas EDA, BVP, ACC y TEMP en su totalidad.

3.1.1. Señales fisiológicas medibles con la pulsera E4

La pulsera Empatica E4 es un dispositivo inalámbrico, cómodo y portátil, diseñado para la adquisición de señales fisiológicas en tiempo real, las cuales son almacenadas en una aplicación web que permite ver y administrar los datos en formato CSV [10]. El dispositivo es tan fácil de llevar como un reloj (Figura 3).

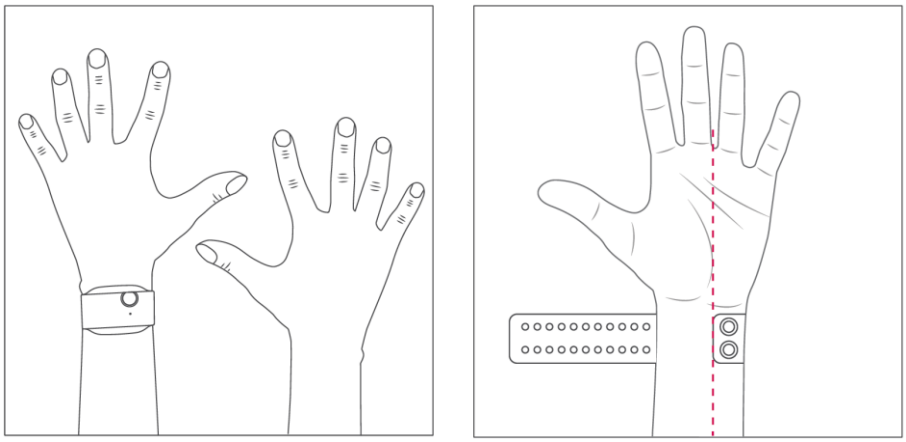

Figura 3. Colocación del dispositivo E4

La pulsera E4 utiliza la metodología exosomática (el uso de corriente externa) para adquirir las señales fisiológicas. A continuación, se presentan las especificaciones técnicas del dispositivo E4 para la recolección de señales en la Tabla 1.

Tabla 1. Especificaciones técnicas E4

Característica	Especificación EDA	Especificación PPG	Especificación ACC	Especificación termopila infrarroja
Frecuencia de muestreo	4 Hz	64 Hz	32 Hz	4 Hz
Resolución	1 digito - 900 pSiemens	0.9 nW/Dígito.	8 bits del rango seleccionado	0.02 °C.
Rango	0.01 µSiemens – 100 µSiemens		± 4g o ± 8g	-40 - 85 °C temperatura ambiente. -40 - 115 °C temperatura de la piel.

El modo de grabación de la E4 permite registrar datos y analizarlos posteriormente con su sistema de gestión, el cuál sincroniza los datos de la sesión de grabación, configura el reloj y administra el *firmware*[3] del dispositivo. En el panel web de la E4, es posible ver, administrar y descargar todas las sesiones grabadas (Figura 4), además de

[3] Software integrado dentro del hardware

observar los detalles de la sesión por series de tiempo para cada tipo de señal con etiquetas de marca de evento superpuestas.

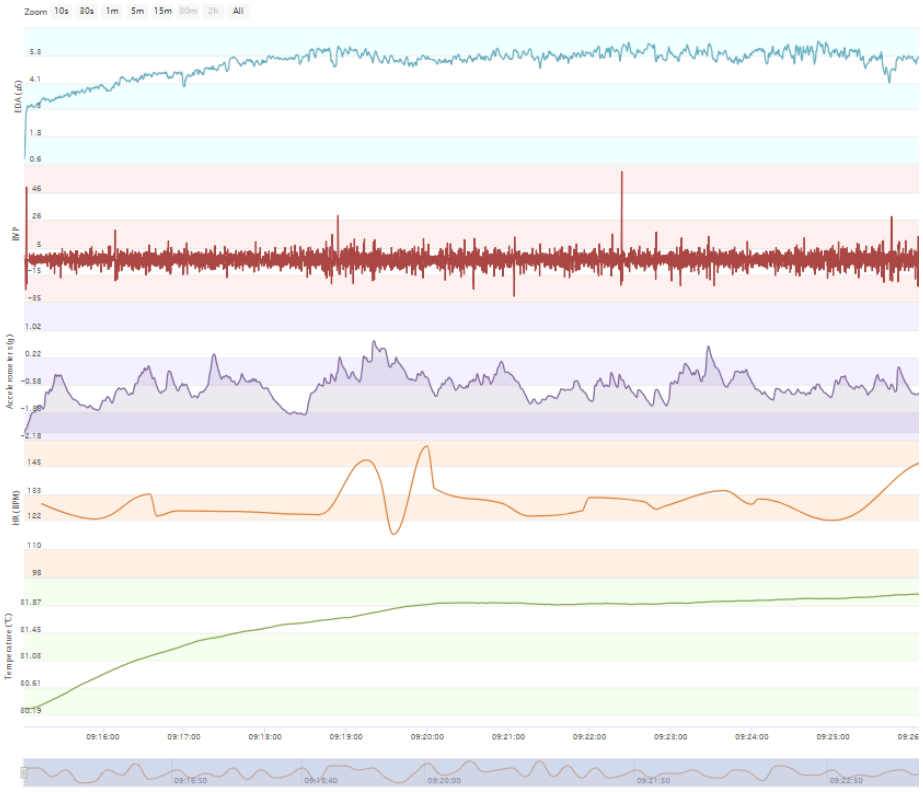

Figura 4. Tablero para la gestión de las sesiones E4.

La pulsera E4 contiene cuatro sensores. A continuación, se detalla cada uno de ellos.

Sensor EDA

La EDA es un proceso neuropsicológico, exactamente del Sistema Nervioso Simpático. Este proceso es indicador de la actividad y variación de las glándulas sudoríparas y de los tejidos dérmicos y epidérmicos asociados a la piel, es decir, es un proceso que hace referencia a las propiedades eléctricas de la piel como respuesta a la sudoración del cuerpo humano. Generalmente la activación de la EDA va asociada a un aumento de los niveles de estrés, atención, cansancio, compromiso, procesamiento de información y estado emocional, por lo que se ha empleado como predictor de la conducta humana, ya que el Sistema Nervioso Autónomo es el

responsable del aumento de dichos niveles, activando la sudoración y, en consecuencia, la conductancia en la piel. Es importante tener en cuenta que cuando estos aumentos se producen, lo hace también el ritmo cardiaco y la presión arterial [51], [52].

Existen dos componentes principales en la actividad electrodermal. (i) Componente tónico general: se considera el nivel de conductancia de la piel en ausencia de cualquier evento ambiental discreto particular o estímulos externos. El nivel de conductancia tónica de la piel puede variar lentamente con el tiempo en un individuo dependiendo de su estado psicológico, hidratación, sequedad de la piel y regulación autónoma. La medida más común de este componente es el Nivel de Conductancia de la Piel (*Skin Conductance Level* - *SCL*) [53]. (ii) Componente fásico: generalmente, se asocia con eventos a corto plazo y ocurre en presencia de estímulos ambientales como la vista, el sonido, el olfato, procesos cognitivos que preceden a un evento, toma de decisiones, entre otros. Estos estímulos aumentan bruscamente la conductancia de la piel generando "picos". Estos picos se denominan Respuestas de Conductancia de la Piel (*Skin Conductance Response* - *SCR*) [53].

La pulsera Empatica E4 pasa una cantidad minúscula de corriente entre dos electrodos en contacto con la piel, de esta forma captura la conductancia eléctrica. Los datos de este sensor están en µSiemens ya que es la unidad de medida para la conductancia y están muestreados a 4 Hz [53].

Sensor PPG (Señal BVP)

Los datos del sensor PPG (fotopletismografía) se conocen en la literatura científica como pulso del volumen sanguíneo (En inglés: *Blood Volume Pulse* - BVP) [54]. La señal BVP a menudo se analiza para evaluar los estados fisiológicos, psicológicos y emocionales de los individuos y derivar variables informativas como la frecuencia cardíaca (En inglés: Heart Rate - HR), intervalos de tiempo entre latidos (En inglés: *Inter-Beat Interval* - *IBI*), variabilidad de la frecuencia cardíaca (En inglés: Heart Rate Variability - HRV) y presión arterial (En inglés: Blood Pressure - BP) [53], [55]. Estas variables ayudan a detectar anomalías del ritmo cardíaco y niveles de estrés. Por lo tanto, su medición precisa es importante en muchas aplicaciones, como el control de la salud de los recién nacidos y la predicción de enfermedades cardíacas [56].

La medida de la fotopletismografía está basada en los cambios de absorción de la luz incidente en la piel provocados por el bombeo de sangre en cada pulsación. Normalmente, en este tipo de mediciones se emplea luz estructurada o infrarroja por contacto [57]. La pulsera E4 utiliza luces producidas por los LED verde y rojo orientados hacia la piel, que son absorbidas por la sangre, la luz se refleja y el receptor de luz la mide. La luz medida durante la exposición verde contiene la mayor parte de la información sobre la onda del pulso (latidos del corazón) y se caracteriza típicamente por una secuencia de valles, cuyas ocurrencias temporales se utilizan para estimar los latidos del corazón. Es importante tener en cuenta que cuanto más se oxigena la sangre, más se absorbe la luz. Por lo tanto, durante un latido cardíaco, hay una alta absorción de luz, que se observa como un valle en la señal de salida de luz. La luz medida durante la exposición roja contiene un nivel de luz de referencia que se utiliza para cancelar artefactos de movimiento.

Los datos del sensor no proporcionan una unidad de medida para BVP ya que esta se deriva de la combinación de las dos medidas diferentes: la cantidad de luz que se refleja y la que vuelve al sensor. Estos datos se encuentran muestreados a 64 Hz. El PPG/BVP es la señal de entrada a los algoritmos propios de la pulsera que calculan los intervalos de tiempo entre latidos (IBI) y la frecuencia cardíaca (HR) como salidas; el algoritmo también se encarga de eliminar los picos incorrectos debido al ruido en la señal BVP [53].

1. Señal IBI
   La señal IBI representa un intervalo de tiempo entre dos latidos cardíacos sucesivos y, por lo tanto, indica un estado cardíaco instantáneo en una escala más fina que la frecuencia cardíaca [56]. La pulsera Empatica E4 calcula el tiempo entre latidos detectando picos (latidos) del BVP y calculando las longitudes de los intervalos entre latidos adyacentes, como se explica en la Figura 5 [53].

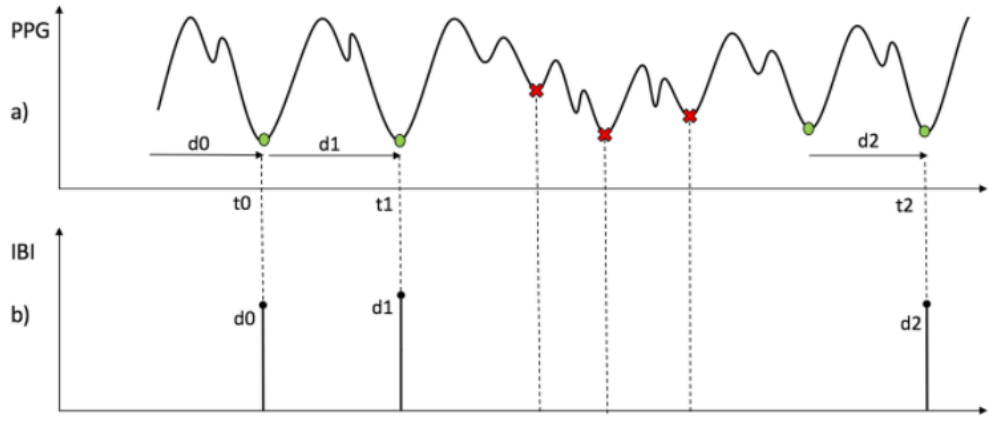

Figura 5. Señal IBI a partir de la señal BVP [58]

Los datos IBI están representados por las distancias, en este caso d0, d1 y d2, los tiempos de los latidos incorrectos no se incluyen en el IBI.csv y, por lo tanto, puede suceder que dos filas consecutivas en el IBI.csv no sean consistentes con un tacograma estándar; igualmente, no se encuentran filas vacías por los datos incorrectos que no se toman en cuenta. Cada dato IBI está asociado a un dato *t* que indica el momento exacto en el que se registra. A continuación, en la Figura 6 se expone un ejemplo de lo que se encontrará en el archivo .csv, con respecto a la situación que se muestra en la figura anterior [58].

A	B
UNIX start t	IBI
t0	d0
t1	d1
t2	d2
...	...
...	...

Figura 6. Señal IBI en archivo CSV [58]

Los datos de la señal IBI son los tiempos entre cada pico, los latidos incorrectos no se incluyen; de acuerdo con la gráfica los datos utilizados para la señal IBI serán d0, d1, d2 [59].

2. Señal HR

   La frecuencia cardíaca representa el número de latidos del corazón en una determinada ventana de tiempo y normalmente se describe en forma de HR media para ese tiempo específico [56]. El archivo final que facilita la página web de la pulsera E4 contiene los valores promedio de frecuencia cardíaca calculados en tramos de 10 segundos. No se derivan de una lectura en tiempo real, sino que se crean solo después de cargar la sesión en la página web y realizar el análisis de BVP de la sesión. En el archivo, la velocidad de muestreo utilizada es de 1 Hz, por lo que los valores proporcionados en el archivo están separados por 1 segundo. Existe una señal de HR que se muestra en tiempo real; este HR instantáneo es derivado del IBI en durante la sesión, puede visualizarse en la aplicación web de la pulsera E4 y no debe compararse con los otros valores de HR que se descargan de cada sesión de grabación [53].

Acelerómetro de 3 Ejes

El análisis de la orientación tiempo-espacial de una persona durante un tiempo determinado permite evaluar los movimientos motrices y por medio de estos detectar emociones del paciente, puesto que la motricidad está ligada estrechamente con la parte intelectual y cognitiva del ser humano. Así como lo plantea la psicomotricidad descrita como la psicología del movimiento en la que entran en contacto: cuerpo, mente y emociones, es decir cuando un niño realiza una acción, ésta se encuentra directamente relacionada con un pensamiento y con una emoción [60].

Los dispositivos Empatica tienen un acelerómetro integrado de 3 ejes que mide la fuerza gravitacional continua (g) aplicada a cada una de las tres dimensiones espaciales (x, y, z). La escala está limitada a ± 2g (por defecto) en la E4; el eje “y” puede extenderse a ± 8g con un *firmware* personalizado. En este caso, el acelerómetro está configurado para medir la aceleración en el rango [-2g, 2g] y sus datos están muestreados a 32 Hz [53].

Termopila infrarroja (señal TEMP)

Este sensor de la pulsera Empatica E4 lee la temperatura periférica de la piel. Esta temperatura es una de las bioseñales más utilizadas para proporcionar retroalimentación biológica tanto en investigaciones como en el ámbito clínico y puede ser registrada en los miembros superiores e inferiores, en las narinas o en las orejas. En estado de tensión o estrés, la temperatura periférica de la piel de los miembros disminuye, debido a que el flujo sanguíneo se dirige a los grupos de músculos largos, en los que aumenta la temperatura para prepararlos ante una eventual situación de lucha o huida; a medida que nos relajamos, el flujo que va hacía las extremidades aumenta, elevando la temperatura nuevamente [61], [62].

Los datos del sensor de temperatura están expresados en grados en la escala Celsius (°C) y están muestreados a 4 Hz [10].

3.1.2. Análisis estadístico descriptivo de las señales fisiológicas

La descripcion de los datos consiste en estudiar y observar su comportamiento para encontrar las tendencias y relaciones existentes entre ellos; el objetivo es hacer síntesis de la información para brindar información sencilla y precisa, además de dar orden a los datos lo que permite encaminar la investigación hacia resultados mejorados [63].

Para obtener una base de conocimiento de la información y características de los datos más representativos obtenidos por medio de la pulsera Empatica E4 fue indispensable realizar un análisis estadístico de las señales fisiológicas, especialmente de la señal con datos faltantes; a partir de los resultados obtenidos, se tomaron las decisiones que permitieron una mejor organización de los datos y la selección del método más acertado para complementarlos.

Realizar un análisis básico descriptivo implica calcular las medidas simples de composición y distribución de variables, tales como: medidas de tendencia central,

medidas de dispersión, percentiles o medidas de posición y medidas de forma [64], [65].

A continuación, en la Tabla 2 se da a conocer un ejemplo de las medidas extraídas con el análisis descriptivo realizado sobre los datos recolectados en una sola sesión de uno de los pacientes. La información sobre el resto de las sesiones analizadas se encuentra en el anexo A.

Tabla 2. Análisis descriptivo de las señales fisiológicas

	x	y	z	BVP	EDA	HR	IBI	TEMP
MEDIA	18,247	-21,399	17,121	0,004	12,761	100,550	0,629	31,326
MEDIANA	20,120	-40,120	16,750	0,400	11,317	94,220	0,625	31,350
MODA	['19,880']	['-60,620']	['14,250']	['6,110']	['2,182', '2,225']	['90,670']	['0,625']	['31,530']
MINIMO	-73,500	-128,000	-83,500	-317,240	0,827	55,670	0,375	30,670
MAXIMO	125,380	79,500	115,620	262,330	23,849	129,880	0,859	31,810
RANGO	198,880	207,500	199,120	579,570	23,022	74,210	0,484	1,140
DESVIACION STD	26,872	47,875	21,748	44,168	7,307	15,359	0,067	0,300
VARIANZA	722,124	2292,013	472,967	1950,840	53,399	235,886	0,004	0,090
PRIMER CUARTIL	4,380	-59,380	5,380	-25,310	5,668	88,900	0,594	31,130
SEGUNDO CUARTIL	20,120	-40,120	16,750	0,400	11,317	94,220	0,625	31,350
TERCER CUARTIL	35,120	30,880	31,250	25,570	20,282	116,500	0,672	31,590
ASIMETRIA	-0,314	0,508	-0,385	-0,043	0,045	0,368	-0,177	-0,397
CURTOSIS	0,360	-1,156	0,972	2,324	-1,611	-1,159	1,185	-0,818
NOMBRE DISTRIBUCION DE PROBABILIDAD							gamma	

Medidas de tendencia central

Medidas estadísticas que resumen en un solo valor todo un conjunto de valores, esto debido a que, en la mayoría de los casos, el conjunto de datos tiende a reunirse alrededor de un valor central que es el valor típico de todo el conjunto de datos, este se denomina medida de tendencia central. Las medidas de tendencia central más representativas son: media, mediana y moda [64]–[66].

1. Media

   Valor promedio de un conjunto de datos numéricos, es una de las medidas de posición central más conocida y utilizada, se calcula con la suma de todos los valores de los datos y se divide entre el número total de sumandos. La media emplea en su cálculo toda la información disponible, esto la hace muy sensible a cualquier cambio en los datos, especialmente cuando el grupo de muestras tiene valores extremos, por esta razón esta medida también puede ser utilizada como un detector de variaciones en los valores [64], [65], [67], [68].

   En el caso de las señales fisiológicas, la media da una idea cercana del valor que maneja el paciente en cada señal recolectada durante la sesión, este valor puede ser útil en bioseñales que no tienen gran variación en el niño, como la temperatura. Si observamos el ejemplo en la tabla 2, la temperatura tiene un rango de variación de 1,14 y una media de 31,32; en cualquiera de los casos la temperatura seguiría siendo baja y no se haría una diferencia importante, pero las señales x, y, z del ACC, BVP, EDA, HR y IBI tienen rangos grandes de variación de acuerdo a su escala, incluso varias de estas señales tienen valores negativos y positivos, lo que hace que la media no refleje valores importantes para detectar cambios de comportamiento en el paciente en el momento de analizar las variables, lo cual es el objetivo más importante de este ejercicio investigativo.

2. Mediana

   Variable que divide los datos en dos partes iguales, de tal manera que deja por debajo el 50% de las observaciones y por encima el otro 50%. Para calcular la mediana los datos deben estar ordenados en función de su magnitud de mayor a menor, o en sentido contrario. Si el número de observaciones es impar, la mediana es el valor que ocupa la posición central, en el caso de que el número de datos del grupo sea par, la mediana se calcula como la media aritmética de los dos valores situados en el centro. A diferencia de la media, la mediana no se ve influenciada por valores extremos, ya que está basada en la posición que ocupan los datos y no en su magnitud [64], [65], [68], [69].

   Se puede considerar que esta medida no es muy útil en el análisis de las bioseñales, principalmente teniendo en cuenta que muchos de los valores de

las señales se repiten y al ser organizados de acuerdo con su magnitud el valor central en cuanto a posición no representa el valor medio entre la magnitud de todos los valores, ni tampoco un valor en la posición del tiempo durante la terapia.

3. Moda
   La moda es el valor que se presenta más frecuentemente en un conjunto de muestra, no es necesariamente 'central', pero es útil cuando el valor más común es de interés en la investigación. Se pueden encontrar varias modas en un grupo de datos, cuando existe una moda la distribución de los datos es unimodal, cuando existen dos modas su distribucion es bimodal, y así sucesivamente. Si los datos tienen una distribución unimodal y es aproximadamente simétrica (se distribuyen de forma similar a ambos lados de la media), la media, la mediana y la moda coinciden o tienen valores muy próximos [64], [65].

   Durante la sesión, el paciente experimenta diferentes emociones que permiten entender la efectividad del juego sobre las necesidades del niño, esto se puede manifestar en la repetición de varias señales fisiológicas, por esta razón, la moda toma un papel importante en el análisis de estas señales. En el caso de ACC la moda indica geográficamente un movimiento motriz repetitivo en el niño, o los valores de EDA, HR, IBI y TEMP más frecuentes permiten determinar cuál fue el comportamiento del niño la mayor cantidad de tiempo que interactuó con el juego.

Medidas de dispersión

Las medidas de dispersión entregan un resumen de la información sobre la variación o agrupación de los datos alrededor de una medida de centralización. En los análisis estadísticos, las medidas de dispersión más representativas son: rango, varianza y desviación estándar [63]–[65].

1. Rango
   El rango se considera la medida de dispersión más simple, es un valor numérico que indica la diferencia entre el valor máximo y el mínimo de un conjunto de

datos. Si bien el rango es una variable fácil de calcular, generalmente tiene un uso limitado por el hecho de estar basada sólo en los valores extremos, lo que la hace una medida ineficiente y sensible a la presencia de datos atípicos, aunque suele ser muy útil cuando se desea conocer qué tan extremos son los límites máximos y mínimos de un grupo de datos [64], [65], [70], [71].

En el análisis de las señales es importante tener en cuenta los picos que presentan, puesto que estos datos pueden asociarse a un comportamiento especial del paciente, así que conocer los valores máximos, mínimos y su rango de dispersión es de gran ayuda en el momento de verificar el funcionamiento correcto del dispositivo e identificar el momento que se manifestó este comportamiento y cuál fue su posible causa.

En la Tabla 2 se puede observar que el rango de HR es de 74,21 con un máximo de 129,88 y un mínimo de 55,67. De acuerdo con la Tabla 3 y teniendo en cuenta que el paciente tiene 9 años, se puede asegurar que la frecuencia cardíaca que presenta durante la sesión es normal de acuerdo con su edad, pero se debe evaluar en que momentos subió y bajo de su valor medio y que pudo causar esto.

Tabla 3. Valores de frecuencia cardíaca [72]

Edad	Rango (media) [lpm]
Neonato	95 - 150 (123)
1-2 meses	121 - 179 (149)
3-5 meses	106 - 186 (141)
6-11 meses	109 - 169 (134)
1-2 años	89 - 151 (119)
3-4 años	73 - 137 (108)
5-7 años	65 - 133 (100)
8-11 años	62 - 130 (91)
12-15 años	60 - 119 (85)

2. Varianza

La varianza permite hacerse una idea del grado de dispersión de una variable respecto a su media. Para calcularla se halla la diferencia de cada dato con la media aritmética, cada diferencia se reconoce como desviación respecto al promedio o media. Al sumar el total de las desviaciones respecto al promedio, éste tiende a cero por la compensación de las desviaciones positivas con las desviaciones negativas si en el conjunto de datos existen valores positivos y

negativos. De esta manera, no es posible obtener efectivamente la desviación de los datos respecto del promedio, por lo cual se hace necesario elevar cada desviación al cuadrado, garantizando así que todas las desviaciones obtenidas presenten cantidades positivas; el resultado quedará en unidades cuadradas y finalmente se podrá dividir entre el valor de la media del conjunto de datos.

Al calcular la varianza, los datos se elevan al cuadrado, por tanto, el resultado final se ve distorsionado. En consecuencia, en la mayoría de los análisis estadísticos se emplea la varianza como medida para comparar la dispersión entre dos o más variables, identificando el valor mayor como aquel que posee mayor dispersión o variabilidad. La importancia de la varianza está en que es una medida transitoria para el cálculo de la desviación típica o estándar de un conjunto de datos [64], [65], [73].

3. Desviación Estándar

   La desviación estándar o desviación típica es considerada la medida de dispersión con mayor representatividad para un conjunto de datos; matemáticamente, se calcula como la raíz cuadrada positiva de la varianza, por lo que está expresada en las mismas unidades y escala de medida que la media. La desviación estándar ofrece información sobre la dispersión de la variable, ya que indica qué tan dispersos están los datos con respecto a la media. Mientras mayor sea la desviación estándar, mayor será la dispersión de los datos, es decir, una desviación estándar baja indica que la mayor parte de los datos de una muestra tienden a estar agrupados cerca de su media, mientras que una desviación estándar alta indica que los datos se extienden sobre un rango de valores más amplio; esta medida es siempre mayor o igual que cero [64], [65], [74].

   La desviación estándar y la varianza son medidas muy similares, la diferencia radica en que la desviación estándar al estar en unidades de la media es una medida precisa y más fácil de interpretar. Aplicar varianza o desviación típica en las bioseñales nos va a permitir ver la variación de su comportamiento entre sesiones o juegos, por ejemplo, si el rango de movimiento motriz es más limitado o si BVP, HR e IBI tienen una dispersión más pequeña o si la temperatura varía menos.

Medidas de Posición

Las medidas de posición o cuantiles permiten calcular valores en la distribución de los datos y dividen un conjunto de datos en partes iguales, de tal forma que los intervalos generados por los cuantiles contienen el mismo número de datos. Para calcular las medidas de posición es necesario que los datos estén ordenados de menor a mayor, en caso de que la posición del cuantil involucre dos valores, el valor del cuantil será el promedio de ambos datos. Para dividir el conjunto de datos en N grupos es necesario definir N-1 cuantiles, que reciben distintos nombres en función del valor de N: percentiles (N=100), deciles (N=10), quintiles (N=5), cuartiles (N=4) o mediana (N=2), los cuantiles más utilizados son los cuartiles, deciles y percentiles [64], [65].

1. Cuartiles
   Los cuartiles (Qk) son tres valores que dividen el grupo de datos ordenados en cuatro partes iguales, cada una de las partes representa un 25% de los datos. El primer cuartil (Q1) deja por debajo el 25% de los datos y el 75% restante por encima de él. El segundo cuartil (Q2) tiene el 50% de los datos por debajo y el otro 50% restante por encima de él (por esta razón es igual a la mediana) y el tercer cuartil (Q3) deja por debajo el 75% de los datos y por encima el 25% restante [64], [65], [75].

2. Deciles
   Los deciles (Dk) son nueve valores que fraccionan el conjunto de observaciones en diez partes iguales, cada una representa el 10% del grupo de datos. El primer decil (D1) deja por debajo el 10% de los datos y por encima el 90% restante, el segundo decil (D2) deja por debajo el 20% de los datos y por encima el 80% restante y así sucesivamente hasta el noveno decil (D9), valor que deja por debajo el 90% de los datos y por encima el 10% restante [65].

3. Percentiles:
   Los percentiles (Pk) son noventa y nueve valores que fraccionan la distribución de los datos en cien partes iguales. El primer percentil (P1) deja por debajo el 1% de los datos y por encima el 99% restante, el segundo percentil (P2) deja por debajo el 2% de los datos y por encima el 98% restante, se aplica de forma

similar hasta llegar al percentil noventa y nueve (el k percentil deja por debajo el k% de la distribución) [65].

Al igual que la mediana, para hallar los cuantiles se deben reorganizar los datos de menor a mayor, y teniendo en cuenta que gran cantidad de señales fisiológicas pueden repetirse varias veces, se pierde el objetivo de hallar el valor de ciertas posiciones para identificar distribución de los datos según su magnitud, por lo que calcular estos valores no es muy relevante durante el análisis.

Medidas de forma

Medidas particulares de un conjunto de datos que permiten conocer la forma de la curva que los representa e identificar si la distribución de los datos presenta uniformidad, para ello toman como referencia la media aritmética y la desviación estándar de la población o la muestra. Para caracterizar el perfil de una distribución de valores existen dos coeficientes útiles para describir la forma de una distribución: coeficientes de asimetría y de curtosis [64], [65], [75].

1. Asimetría o Sesgo
   La asimetría permite identificar si las observaciones de un conjunto de datos se distribuyen simétricamente alrededor de la media, este valor también es importante para conocer la zona en la que se concentran los datos. Para su interpretación se debe tener en cuenta que es posible que exista una tendencia de los datos hacia uno de los extremos (derecho o izquierdo), esta tendencia se denomina sesgo y para describirlo se comparan la media aritmética, la mediana y la moda, pueden existir los siguientes casos [64], [65], [75]:
   - Las medidas de la media, mediana y moda son exactamente iguales, la variable es simétrica y el coeficiente de asimetría toma el valor cero.
   - La media aritmética es inferior a la mediana y a la moda, la distribución de los datos presenta un sesgo o cola hacia la izquierda y el coeficiente toma un valor negativo (asimetría negativa)
   - La media aritmética es superior a la mediana y a la moda, la distribución de los datos esta sesgada hacia la derecha y el valor del coeficiente es positivo (asimetría positiva).

Los anteriores casos se representan en la Figura 7:

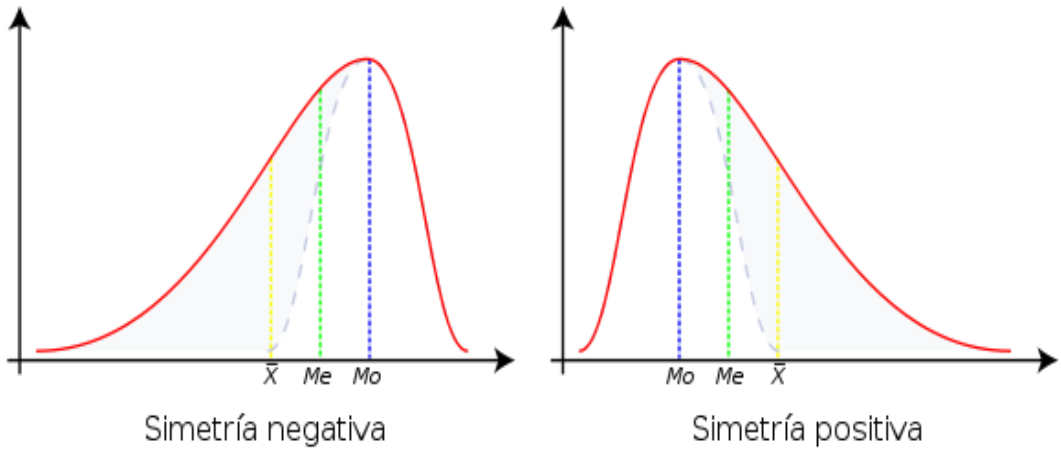

Figura 7. Asimetría de la distribución de los datos [76].

En la Tabla 2 se puede detallar los valores de asimetría de cada variable, en esta sesión que se ha tomado como ejemplo, el eje y de ACC, EDA y HR tienen asimetría positiva, mientras que las variables restantes tienen asimetría negativa. Este valor nos permite identificar el comportamiento que tiene la variable y tener una idea del intervalo en el que se ubican la mayoría de los datos.

2. Curtosis

   La curtosis mide el grado de similitud de la distribución de los datos respecto a la distribución normal, diferencia tres clases de distribuciones [64], [65], [75]:

   - La distribución presenta el mismo perfil que la distribución normal con la misma media y varianza, entonces el coeficiente de curtosis toma el valor cero (distribución mesocúrtica);
   - La distribución es más puntiaguda que la distribución normal, dado que las frecuencias altas están alrededor de la media, en este caso el valor del coeficiente es positivo (distribución leptocúrtica).
   - La distribución es más “aplastada” que la distribución normal, dado que las frecuencias bajas están alrededor de la media, en este caso el valor del coeficiente es negativo (distribución platicúrtica).

   Los anteriores casos se representan en la Figura 8.

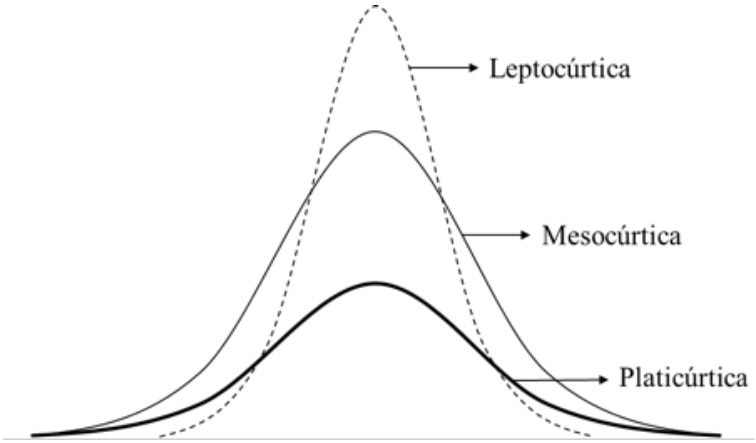

Figura 8. Tipos de distribución según la curtosis [77].

Según la Tabla 2, las variables que presenta una distribución leptocúrtica son el eje x, y y z de ACC, EDA, HR y TEMP. La variable x de ACC se encuentra muy cercana a la distribución mesocúrtica o normal. Las demás variables presentan una distribución platicúrtica.

Determinación de las señales incompletas

Después de realizar el análisis descriptivo estadístico de cada una de las señales obtenidas por medio de los sensores y de los algoritmos de la pulsera E4, se logró determinar cuáles señales estaban incompletas dentro del conjunto de datos recolectados.

Cómo hipótesis inicial, se había considerado que la señal con datos perdidos era el HR, dado que visualmente, por medio de la aplicación web de la pulsera E4, se puede apreciar que es la señal que no esta completa (Figura 9). Sin embargo, al descargar los datos de las sesiones, se evidenció que la señal de HR se encontraba completa, puesto que la señal de HR fue computada a partir del algoritmo de BVP.

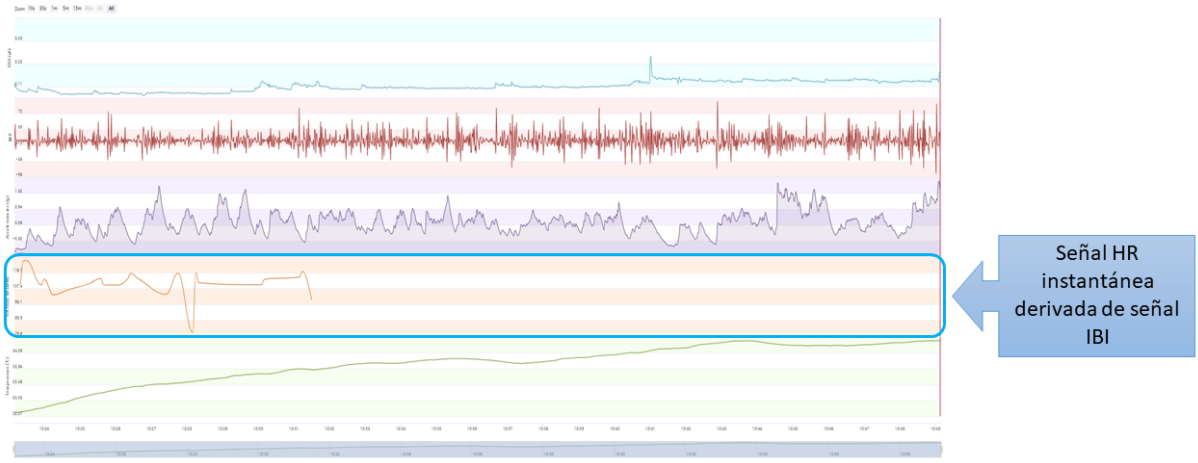

Figura 9. Sesión de grabación en la que se evidencia HR incompleta

Al explorar el resto de las señales, fue evidenciado que la señal de IBI es la que se encontraba incompleta, dado el formato que presenta esta señal, explicada en la anterior subsección. Por lo tanto, a partir de esta conclusión, la señal IBI será objeto de más análisis para resolver el problema de pérdida de datos.

3.1.3. Ajuste de distribución de probabilidad para señal IBI

Estadísticamente, la distribución de probabilidad hace referencia a la función que asigna la probabilidad de que ocurra un suceso sobre una variable [78]; la distribución de probabilidad sobre los datos EDA, BVP, HR, IBI y TEMP describe el porcentaje de la cantidad de veces que se encuentra presente cada dato a lo largo de la sesión, esto sería la probabilidad de que cada valor de la señal fisiológica se repita durante el tiempo que el niño interactúa con el juego.

Identificar el tipo de distribución que tiene una variable es un paso esencial para la creación de modelos por aprendizaje estadístico y de máquina, puesto que con una base de conocimiento acerca de los datos es posible tomar decisiones más eficaces. Recordando que el problema identificado consiste en que la señal IBI presenta una gran cantidad de datos perdidos, la imputación de datos a la señal IBI es una solución viable. Para realizar este proceso, es importante contar con variables que permitan comparar el comportamiento de los datos antes y después de la imputación con el fin

de evaluar la eficiencia del método utilizado y la conveniencia de utilizar los datos resultantes.

Para identificar la distribución de probabilidad que tiene la variable IBI en cada sesión, es necesario analizar cada parámetro de probabilidad de la señal y seleccionar la distribución estadística que tenga métricas similares a las de la señal analizada. Este proceso se conoce como ajuste de distribución [79], en el cual se pueden aplicar múltiples métricas que permiten cuantificar que tan bien se ajusta una distribución a los datos observados. En este proceso fueron empleadas las métricas AIC (por sus siglas en inglés, *Akaike information criterion* - criterio de información de Akaike) y BIC (por sus siglas en inglés, *Bayesian information criterion* - criterio de información Bayesiano). Estos criterios de información están asociados con el método de máxima verosimilitud, el cual está basado en el supuesto de que las variables observadas siguen una distribución normal multivariante[4].

El ajuste de distribución se realizó a la señal fisiológica IBI de 107 sesiones de las 154 en total; ya que las sesiones restantes (30,5%) carecían completamente de datos IBI. El resultado de este análisis arrojó como principales distribuciones de probabilidad: Gamma y Weibull para el 45,8% y 20,5% respectivamente del total de las sesiones, otras de las distribuciones que se presentaron en cantidades entre el 1% y 4% de las sesiones fueron: Burr, Burr12, Gennorm, Gengamma, Genlogistic, Laplace, Loglaplace y Triang. El total de distribuciones obtenidas y sus porcentajes sobre el total de sesiones se describen en la Tabla 4.

Tabla 4. Distribuciones de probabilidad de la señal IBI

Distribución de Probabilidad	Porcentaje de Sesiones
Gamma	45,8 %
Weibull	20,56%
Gennorm	4,67%
Burr	3,73 %
Burr12	2,8%
Gengamma	1,87%
Laplace	1,87%
Genlogistic	1,87%

[4] Generalización de la distribución normal unidimensional a dimensiones superiores.

Loglaplace	1,87%
Triang	1,87%
Loguniform	0.93%
T	0.93%
Skewnorm	0.93%
Invgauss	0.93%
Tukeylambda	0.93%
Beta	0.93%
Recipinvgauss	0.93%
Kstwobign	0.93%
Gausshyper	0.93%
Pearson3	0.93%
Foldcauchy	0.93%
Mielke	0.93%
Johsonsu	0.93%
Hypsecant	0.93%

Teniendo en cuenta estos resultados, se realiza una breve descripción de las distribuciones que presentan un comportamiento más cercano al de la señal IBI.

- Distribución Gamma:
  Esta distribución representa el comportamiento de variables aleatorias continuas con asimetría positiva. Es decir, variables que tienen mayor cantidad de datos a la izquierda de la media que a la derecha [80]. Teniendo en cuenta que la mayoría de los datos se ajustan en esta distribución, se puede concluir que en la mayoría de las sesiones gran parte de los datos de la señal IBI se concentran en un intervalo corto por debajo de la media, es decir, una frecuencia cardíaca en gran parte estable y dentro de un rango que no excede los límites que se consideran normales. Esto se podría interpretar como que el paciente no experimentó angustia o estrés en la mayoría del tiempo que duro la sesión.

- Distribución Weibull:
  Es una distribución de probabilidad continua y triparamétrica, es decir, está definida por tres parámetros que definen su forma. La distribución Weibull es la

más empleada en el campo de la confiabilidad, se utiliza para modelar situaciones del tipo tiempo-fallo, modelar tiempos de vida o en el análisis de supervivencia. Los parámetros que la describen son [78]:
β = Parámetro de forma.
η= Parámetro de escala
γ= Parámetro de posición.

3.1.4. Organización de las variables

Para realizar la imputación de la señal IBI, primero se organizó por sesión todas las variables en un único conjunto de datos, en el cual cada variable representa una columna. Para realizar el análisis de los datos de cada variable, estos debían coincidir con su ubicación respecto al tiempo con las demás variables, es decir, su frecuencia de muestreo debía ser la misma para que los datos a su lado coincidieran con el momento en que se registraron. La frecuencia de muestreo que utilizó la pulsera empática E4 para cada una de las señales fisiológica es la siguiente:

- ACC = 32 Hz (32 muestras por segundo)
- BVP = 64 Hz
- EDA = 4 Hz
- HR = 1 Hz
- TEMP = 4Hz
- IBI = No tiene frecuencia de muestreo, cada dato IBI está asociado a la variable tiempo que indica exactamente el segundo de la sesión en el que se registró, como se describió anteriormente en la definición de esta señal.

Inicialmente se consideró dejar una frecuencia de 4 Hz para todas las señales, es decir cuatro muestras por cada segundo, con el fin de aprovechar al máximo los datos de las variables completas. Sin embargo, teniendo en cuenta la definición de la señal IBI, es posible afirmar que la cantidad de datos en 60 segundos va a ser igual o cercana a la frecuencia cardíaca promedio en esos mismos 60 segundos.

Basándose en este análisis, se concluyó que era un error imputar los datos con una frecuencia de 4 Hz, ya que la imputación se haría para 4 muestras IBI por segundo, es decir 240 muestras en total por minuto, algo similar a afirmar que el paciente presentaba 240 lpm (latidos por minuto), lo cual era contrario a la realidad. Por esta razón y con el objetivo de identificar la cantidad de datos que deberían presentarse en la señal IBI, se decide extraer el valor mínimo y máximo de la señal HR, es decir el mínimo y máximo de datos que debería tener la señal IBI. Posteriormente, se extrajeron estos valores por cada minuto de la sesión, con el fin de estudiar las variaciones de la frecuencia cardíaca ya que normalmente esta debería ser variante durante la sesión. De hecho, esta variación es la que permite identificar cambios en el comportamiento del niño y si la interacción que está desarrollando con el juego HapHop-Fisio puede ser medida y caracterizada fisiológicamente.

Finalmente, los valores máximos y mínimos de HR por minuto se comparan con la cantidad de datos IBI equivalentes al mismo rango de tiempo, con el fin de dar una aproximación de los valores que deberían estar presentes en cada minuto de esta señal y los que se tienen en realidad. De esta forma, se obtiene un porcentaje aproximado de los datos presentes y faltantes en cada archivo IBI. A continuación, en la Tabla 5, se exponen los resultados del análisis mencionado anteriormente en un ejemplo de una sesión de un paciente con 11 años.

Tabla 5. Valores máximos y mínimos de señal HR, porcentaje de datos IBI presentes y faltantes

Minuto	MaxHR	MinHR	DatosIBI	AusentesIBIMaximoHR	AusentesIBIMinimoHR
1	116.0	75.25	23	93	52
2	93.65	90.85	22	71	68
3	90.8	83.58	24	66	59
4	97.5	83.75	27	70	56
5	97.7	92.77	40	57	52
6	94.28	92.23	65	29	27
7	92.12	88.32	48	44	40
8	90.63	85.73	34	56	51
9	91.3	85.33	33	58	52
10	90.88	86.67	34	56	52
11	91.22	87.32	80	11	7
12	103.92	91.48	53	50	38
13	101.02	88.28	29	72	59
14	94.17	88.27	46	48	42

15	94.22	86.75	26	68	60
16	92.72	85.32	25	67	60
17	92.9	82.67	42	50	40
18	99.45	93.25	42	57	51
19	97.52	90.87	41	56	49
20	93.23	89.2	35	58	54
21	89.05	85.5	51	38	34
22	88.02	84.27	4	84	80
23	83.92	78.0	0	83	78
24	90.92	82.5	0	90	82
Total	2256	2067	824	1432	1243
Porcentaje (%)	100	100	['36.52', '39.86']	63.48	60.14

Los pacientes que interactuaron con el juego HapHop-Fisio tienen edades entre los 6 y los 15 años. De acuerdo con la Tabla 3, el rango de frecuencia cardíaca que deben manejar estos niños para considerarse entre los parámetros normales es de: niños entre 5 y 7 años deben presentan una frecuencia cardíaca mínima de 65 latidos por minuto (lpm) y máxima de 133 lpm, niños de 8 a 11 años de 62 a 130 lpm y niños de 12 a 15 años entre 60 y 119 lpm.

Se conoce que el paciente que realizo la sesión del anterior ejemplo tiene 11 años, por lo que su rango normal de frecuencia cardíaca debería variar entre 62 a 130 lpm. Al realizar la comparación con la Tabla 4, se puede observar que el valor mínimo de frecuencia cardíaca durante la sesión fue de 75,25 lpm y el máximo 116 lpm, lo que se considera un rango adecuado para su edad. Tomando en cuenta estos datos, se concluyó que tampoco era viable tomar una frecuencia de muestreo de 3, 2 o 1 Hz, ya que un muestreo de 3 Hz nos indicaba 180 muestras por minuto, representados en 180 lpm, que es un dato muy superior a los rangos establecidos. De igual manera, en el caso de aplicar un muestreo de 2 Hz, se afirmaba que el niño manejó valores cercanos a su frecuencia cardíaca máxima durante toda la sesión, o, por lo contrario, si se tomaba una sola muestra por segundo, es decir 60 muestras por minuto se asumía que el niño mantuvo una frecuencia cardíaca por debajo del valor mínimo normal durante su interacción con el juego.

Por lo tanto, con el propósito de organizar los datos acordes a una variación “normal” en la señal IBI durante la sesión, se buscó identificar un valor de tendencia central en

cada minuto, que representara el rango de frecuencia cardíaca característico del paciente en cada intervalo de tiempo. Como se mencionó anteriormente, las variables de tendencia central más representativas son: media, media y moda. Hallar un promedio de la señal HR por minuto podía ser una buena aproximación a la cantidad de datos que debería tener IBI, pero se debía tener en cuenta que la media podía verse afectada por los valores extremos, en este caso, se podía tomar en cuenta la mediana que representa el valor central entre los datos organizados de acuerdo a su magnitud. Sin embargo, se debía recordar que en la frecuencia cardíaca se pueden repetir varios datos, lo que distorsionaba la posición del dato buscado. Finalmente, tomar el valor de la moda por minuto fue la solución más viable, ya que este valor representa la frecuencia cardíaca más común en cada minuto de la sesión. Por lo tanto, se decidió que el valor de la moda representaría la cantidad de datos IBI y que sería variable en cada minuto de la sesión.

3.2. Imputación de señales

A pesar del esfuerzo al realizar la planificación de una investigación en distintos ámbitos, es común la aparición de pérdida de datos, este es un fenómeno que siempre está presente, es una realidad y un problema por considerar. Está demostrado que una pérdida de datos trae consigo inconvenientes como complicaciones en el análisis de estos, la pérdida de eficiencia y resultados sesgados que disminuyen la validez de la investigación. Por esta razón, los datos perdidos son un punto muy importante que se debe de tener en cuenta en cualquier proyecto [14], [81].

Durante muchos años el tratamiento de datos perdidos consistió en la eliminación de los sujetos con información incompleta. La formalización y estudio sistemático de este problema se inicia hasta mediados de los años setenta e inicios de los años ochenta, destacando principalmente el trabajo de Rubin [82], quien propuso un marco conceptual para el análisis de datos faltantes sustentado en métodos de inferencia estadística [81], [83], [84]. En la década de los noventa, se produjo un cambio en el procedimiento de los datos incompletos, buscando modelar la incertidumbre alrededor de él y no solo buscando un valor para completar [83].

En los últimos tiempos, se han desarrollado formas más eficientes y fáciles para el manejo de datos perdidos, obteniéndose una variedad de técnicas basadas en diferentes enfoques según las características de los datos [85], [86]. Esta práctica de reemplazar los datos faltantes con nuevos valores se denomina imputación de datos. Las ventajas de utilizar las técnicas de imputación se centran en que el poder estadístico de los datos no disminuye, porque la imputación conserva el tamaño completo de la muestra, lo que permite utilizar métodos de análisis estándar y software, sin preocuparse por los datos que faltaban anteriormente [87].

3.2.1. Mecanismos y patrones de imputación de datos

Los datos faltantes son aquellos valores no disponibles, pero que son necesarios en el análisis de datos si fuera posible utilizarlos; es común que se desconozca la causa de la perdida de datos, por esta razón existen mecanismos para identificar la causa de la falta de datos y patrones para clasificar los datos perdidos.

La estrategia más conveniente para el tratamiento de datos incompletos está principalmente en reconocer los diferentes tipos de mecanismos y patrones para la imputación de datos. Los mecanismos describen el proceso mediante el cual se produce la ausencia de los valores y su implicación en la inferencia estadística, de esta forma se puede decidir cómo manejar la información incompleta y obtener una idea de los efectos que pueden causar los ajustes a partir de los procedimientos que se utilicen; por su parte, los patrones describen tanto los valores presentes en el conjunto de datos como los valores perdidos [83], [87], [88].

Mecanismos de imputación

Existen tres tipos de clasificación para los mecanismos de imputación: proceso completamente aleatorio (*missing completely at random* - MCAR), proceso aleatorio (*missing at random* - MAR) y proceso no aleatorio (*missing not at random* - MNAR) [83], [86].

- Datos perdidos completamente al azar (MCAR)
  En este caso la perdida de datos no está relacionada con ninguna variable presente en el conjunto de datos, incluyendo la misma variable [83], [86], [87], [89]. Por ejemplo, el grupo de los datos incompletos pertenecen a la variable X y el conjunto de datos cuenta con otros dos grupos de variables representadas por Y y Z. Los valores podrían considerarse perdidos completamente al azar si la probabilidad de valores perdidos en X no depende de Y, Z o X en sí misma [86]. MCAR se considera el escenario ideal de datos incompletos ya que es un proceso totalmente aleatorio, en el cual los valores perdidos son totalmente arbitrarios y no habría sesgo alguno. Es debido a esto que la mayoría de los métodos tradicionales utilizados para el manejo de datos perdidos requieren de su cumplimiento [86], [89].

- Datos perdidos al azar (MAR)
  En este caso la ausencia de datos no está relacionada con la variable que está incompleta, pero si lo está con otras variables completamente observadas [83], [87]. Por ejemplo, se tiene una variable X con valores perdidos y el conjunto de datos tiene otra variable Y con valores completos. En este caso la perdida de datos en X está relacionada con la variable Y pero no con los valores de X en sí [86], [89].

- Datos perdidos no aleatorios (MNAR)
  En este caso la probabilidad de datos perdidos no se puede relacionar con la información contenida en otras variables. Este mecanismo es contrario al mecanismo MAR, por lo que la perdida de datos está relacionada con su misma variable. MNAR frecuentemente conduce a estimaciones sesgadas de los parámetros en sus análisis estadísticos. Por esto, se considera como los peores tipos de datos que faltan. Con MCAR y MAR, esto es un problema menor, aunque los datos incompletos pueden llevar a una pérdida de poder estadístico [83], [85], [90], [91].

Patrones de imputación

Existen dos tipos principales de patrones de datos perdidos [88], [91]:

- Patrón univariado
  Describe el grupo de datos en el cual solo una variable contiene valores faltantes.

- Patrón multivariado
  Hace referencia al conjunto de datos donde más de una variable contiene datos faltantes. Este patrón se puede dividir en dos grupos: patrón monótono y patrón arbitrario. El patrón monótono describe el conjunto de observaciones en el cual faltan los datos de varias variables de una parte localizada de la estructura de datos. En el patrón arbitrario, los datos faltantes se encuentran en cualquier lugar y no aparece ninguna estructura especial, independientemente de cómo se organicen las variables.

Para el conjunto de señales fisiológicas utilizadas en este trabajo de grado, se observa que los datos de las señales están perdidos al azar (MAR), además de presentar un patrón univariado.

3.2.2. Técnicas para el tratamiento de datos perdidos

A continuación, se explican los diferentes procedimientos para sustituir la falta de datos:

Técnicas de eliminación

Consisten en eliminar la información de los casos en los que haya pérdida de datos. Son los más sencillos de implementar y los que menos recursos computacionales necesitan, por lo que se usan ampliamente, pero no conduce a la utilización más eficiente de los datos y reducen el tamaño de la muestra, lo que disminuye la eficiencia

y aumenta el porcentaje de error en los parámetros de interés. Solo deben utilizarse en situaciones donde la cantidad de valores faltantes es mínima. Esta técnica tiene dos formas: eliminación por lista (en inglés, *listwise* o *case deletion* - LD) y eliminación por pares (en inglés: *pairwise deletion* - PD o *available case* - AC) [81], [88], [91]:

- Eliminación por lista o eliminación de casos
  Este método se utiliza con mayor frecuencia, se caracteriza por excluir los casos que tienen datos faltantes, es decir, elimina la fila en la que se presenta un dato ausente, por lo que solo hace uso de aquellos casos que no contienen valores perdidos. Su aplicación conduce a una gran pérdida de observaciones, lo que puede resultar en conjuntos de datos demasiado pequeños si la cantidad de datos perdidos es alta.

- Eliminación por pares o método de caso disponible
  Este método considera cada variable por separado. Para cada variable, se consideran todos los valores registrados en cada caso y se ignoran los datos faltantes. Esto significa que diferentes cálculos utilizarán diferentes casos y tendrán diferentes tamaños de muestra. Esto no es lo más conveniente, aunque proporciona mejores estimaciones que la eliminación por lista. Es viable utilizarlo cuando el tamaño total de la muestra es pequeño o el número de casos con datos faltantes es grande.

Técnicas de imputación simple

Estas técnicas reemplazan los datos faltantes por un único valor basándose en los valores de la propia variable o de otras variables y tienen una implementación sencilla, lo que evita que sufran una importante pérdida de eficiencia en comparación con técnicas más robustas. Existen diferentes tipos de técnicas de imputación simple, algunas de las más conocidas son [27], [84], [92]:

1. Imputación de la media o moda: el dato perdido de la variable es reemplazada con la media de las muestras existentes para las variables cuantitativas; para el caso de las variables cualitativas se hace con la moda. Una variación de esta técnica es consiste en agrupar las respuestas de cada variable en clases

distintas con diferentes medias, y a cada valor faltante se le imputará con la media respectiva del grupo al que se asignó.

2. Imputación por regresión: se estima un modelo de regresión lineal sobre la base de los valores observados en la variable objetivo y en algunas variables explicativas; el modelo se utiliza para predecir valores para los casos que faltan en la variable objetivo. Existen dos clases de imputación por regresión. (1) Aleatoria: se emplea cuando los datos de la variable a imputar son números; la ventaja de este procedimiento está en el hecho de que no hay pérdida de información, puesto que se trabaja con todas las unidades que fueron estudiadas. (2) Logística: esta regresión se emplea cuando los datos de la variable incompleta son categóricos.

3. Imputación por el vecino más cercano: identifica la distancia entre la variable a imputar y cada una de las variables auxiliares; la medida de distancia más cercana es utilizada para imputar el valor faltante.

4. Algoritmo EM (*Expectation Maximization*): esta técnica está basada en la función de máxima verosimilitud (MV). Se selecciona como valor estimado del parámetro, aquél que tiene mayor probabilidad de ocurrir según lo observado.

5. Redes neuronales: en esta técnica, los patrones de los datos completos son caracterizados para obtener los valores a imputar sobre el conjunto de datos incompleto. Estas redes son más utilizadas en variables cualitativas que cuantitativas, y son más adecuadas cuando la distribución es no lineal.

6. Modelos de series de tiempo: estos modelos asumen que los datos perdidos ocurren de tal forma, que el problema se reduce a una situación en la cual, hay una serie de tiempo. La técnica hace un óptimo uso de las interrelaciones entre cada serie de tiempo [84], [85].

Técnicas de imputación múltiple

Esta técnica sustituye cada valor perdido por un conjunto de *m* valores, lo que permite minimizar el sesgo estadístico. Los objetivos principales de esta técnica de imputación

es utilizar de la forma más adecuada y eficiente los datos que se han recogido, así obtener estimadores no sesgados y mostrar adecuadamente la incertidumbre en la estimación de parámetros causada por los valores perdidos [93]. La imputación múltiple puede presentar una aproximación muy eficaz a los datos reales, pero también tiene algunas limitaciones como el gasto computacional que supone el mantenimiento de conjuntos de datos con un tamaño muy elevado [92]. La imputación múltiple se realiza en tres pasos [83], [88], [94]

1. Imputación: a cada dato perdido se asignan *m* valores extraídos de una distribución predictiva, lo que produce *M* bases de datos.

2. Análisis: para cada base de datos se realiza un análisis estadístico definido de acuerdo con el propósito del estudio; estos pueden variar desde obtener estimaciones puntuales y sus intervalos de confianza hasta modelos de regresión.

3. Agrupación: finalmente, se combinan estos análisis mediante la regla de Rubin [95], para llegar a la estimación definitiva, que es la que se interpreta como resultado final.

Algunas de las imputaciones múltiples más reconocidas son [88]:

- Imputación multivariante por ecuación encadenada (en inglés: *Multiple Imputation with Chained Equations (MICE)*): es un método de imputación múltiple basado en ecuaciones encadenadas, las variables faltantes se reemplazan una a una con la concatenación de procedimientos de imputaciones simples o univariantes.

- *Missing Forest*: es una técnica de imputación no paramétrica basada en el método de predicción Random Forest. Asume una distribución normal multivariante y los datos faltantes como MAR. Es un método iterativo, el cual crea múltiples árboles de decisión sobre muestras de un conjunto de datos. El objetivo es realizar una gran cantidad de predicciones con pocas variables y obtener un promedio de estas.

3.2.3. Plan de imputación de señales

Técnica seleccionada

El manejo de los datos perdidos es una tarea importante y compleja ya que la exactitud del análisis final de los datos depende en gran parte de este procedimiento. Al examinar la teoría, se identificó que la imputación múltiple era más robusta y adecuada para el manejo de datos, ya que toma en cuenta una gran cantidad de factores de las variables que componen el conjunto de datos. De esta manera y considerando las fortalezas y limitaciones del grupo de señales fisiológicas, se decidió realizar un proceso de imputación múltiple. Sin embargo, teniendo en cuenta lo expuesto en el trabajo presentado en [81], se pueden presentar situaciones en que las técnicas de imputación simple pueden entregar resultados satisfactorios, además de ser sencillas de implementar y requerir un gasto computacional bajo. La elección de la técnica de imputación depende enteramente del comportamiento y tipo de variables que se están tratando, por lo que no es posible establecer reglas para decidir cuándo es favorable la aplicación de un método simple o múltiple; por esta razón, también se decidió implementar un método de imputación simple.

Las técnicas seleccionadas para imputación de datos fueron: (1) Imputación múltiple: **MICE**, esta técnica se escogió con base en la literatura que describe este método como uno de los más conocidos y utilizados en análisis de datos tanto clínicos como de diferentes áreas científicas; adicionalmente, brinda mejores resultados que otros métodos de imputación como *Missing Forest* [88], [96], [97]. (2) Imputación simple: técnica **KNN**, esta técnica se ha cotejado con otras imputaciones simples en estudios ya realizados, en los cuales presentó mejores resultados sobre métodos como imputación media, imputación mediana, coincidencia media predictiva, regresión lineal bayesiana, regresión lineal no bayesiana y muestra aleatoria [20], [98].

La utilización de estas técnicas se detalla a continuación:

MICE

Esta técnica asume que la probabilidad de que falte una variable depende de los datos observados, es decir una ausencia de datos al azar (MAR). MICE entrega múltiples

valores para un valor faltante por medio de la creación de una serie de modelos de regresión u otro modelo que se considere más conveniente para el manejo de los datos. Cada variable perdida se maneja como una variable dependiente y los datos auxiliares se consideran variables independientes. El proceso se describe brevemente con un ejemplo a continuación [88], [98].

La Tabla 6 representa el conjunto de datos con valores perdidos que se desea imputar.

Tabla 6. Ejemplo de imputación múltiple - Parte 1

Edad	Años de experiencia	Salario
25		50.000
27	3	
29	5	80.000
31	7	90.000
33	9	100.000
	11	130.000

Paso 1: se reemplaza cada valor faltante implementando una imputación simple, como la imputación de la media, en la cual para cada valor ausente se halla la media de su respectiva columna.

Tabla 7. Ejemplo de imputación múltiple - Parte 2

Edad	Años de experiencia	Salario
25	7	50.000
27	3	90.000
29	5	80.000
31	7	90.000
33	9	100.000
29	11	130.000

Paso 2: se eliminan los valores imputados de una de las columnas que pasaran a ser la variable objetivo, en este caso se inicia con la columna “Edad”.

Tabla 8. Ejemplo de imputación múltiple - Parte 3

Edad	Años de experiencia	Salario
25	7	50.000
27	3	90.000

29	5	80.000
31	7	90.000
33	9	100.000
	11	130.000

Paso 3: las columnas que acompañan la variable objetivo se convierten en un nuevo conjunto de datos. Al nuevo conjunto se le aplica un modelo de regresión lineal tomando como datos de entrenamiento las filas que no tienen datos perdidos de la variable objetivo y como datos de prueba, las filas de los valores ausentes. De esta forma se obtiene una primera predicción de los datos perdidos de una de las variables o columnas.

Tabla 9. Ejemplo de imputación múltiple - Parte 4

Edad	Años de experiencia	Salario
25	7	50.000
27	3	90.000
29	5	80.000
31	7	90.000
33	9	100.000
34,99	11	130.000

Paso 4: se repiten los pasos 2 y 3 para cada una de las columnas que presenten valores perdidos. La Tabla 10 muestra el cambio para la columna “Años de Experiencia” y la Tabla 11 muestra el cambio para la columna “Salario”.

Tabla 10. Ejemplo de imputación múltiple - Parte 5

Edad	Años de experiencia	Salario
25		50.000
27	3	90.000
29	5	80.000
31	7	90.000
33	9	100.000
34,99	11	130.000

Edad	Años de experiencia	Salario
25	0,98	50.000
27	3	90.000
29	5	80.000
31	7	90.000
33	9	100.000
34,99	11	130.000

Tabla 11. Ejemplo de imputación múltiple - Parte 6

Edad	Años de experiencia	Salario
25	0,98	50.000
27	3	
29	5	80.000
31	7	90.000
33	9	100.000
34,99	11	130.000

Edad	Años de experiencia	Salario
25	0,98	50.000
27	3	70.000
29	5	80.000
31	7	90.000
33	9	100.000
34,99	11	130.000

El dataset resultante es el siguiente (Tabla 12):

Tabla 12. Resultados parciales del ejemplo de imputación múltiple

Edad	Años de experiencia	Salario
25	0,98	50.000
27	3	70.000
29	5	80.000
31	7	90.000
33	9	100.000
34,99	11	130.000

Paso 5: se restan los datos obtenidos en el conjunto de datos del paso 1 con los datos resultantes del paso 4. El conjunto de datos final quedaría de la siguiente forma:

Tabla 13. Resultados finales del ejemplo de imputación múltiple

Edad	Años de experiencia	Salario
25	6,02	50.000
27	3	20.000
29	5	80.000
31	7	90.000
33	9	100.000
-5,99	11	130.000

El objetivo de la imputación con MICE es reducir estas diferencias lo más cercano a 0 que sea posible. Para lograr este objetivo se debe repetir el procedimiento desde el paso 1 hasta el paso 4, partiendo de los valores obtenidos en la última repetición. Cada repetición se define como una iteración y se realizaran iteraciones la cantidad de veces

que se considere necesario de acuerdo con los resultados presentados al final de cada una.

Vecinos más cercanos (En inglés: K-Nearest Neighbours (KNN)):
El algoritmo de KNN para imputación sustituye cada valor faltante de una columna en específico por la media de sus k valores más cercanos del conjunto de entrenamiento. Un aspecto clave para la aplicación de este procedimiento es la función de distancia entre la variable a imputar y cada una de las variables auxiliares, la función es definida por el lenguaje de programación y las librerías que se utilicen para implementar el método [84], [85], [99]

Herramientas

a. **Lenguaje de programación: Python**
   Python es un lenguaje de programación dinámico, multiparadigma ya que permite varios estilos de programación como: programación orientada a objetos, programación imperativa y programación funcional. Una de las razones de su éxito es que cuenta con una licencia de código abierto que permite su utilización en cualquier escenario. Esto hace que sea uno de los lenguajes de iniciación de muchos programadores siendo impartido en escuelas y universidades de todo el mundo. Adicionalmente se usa como lenguaje de scripting y es un lenguaje interpretado, es decir, no necesita ser preprocesado mediante un compilador.

   Python es ideal para trabajar con grandes volúmenes de datos ya que, el ser multiplataforma, favorece su extracción y procesamiento, por esta razón lo eligen las empresas de Big Data. A nivel científico, tiene una gran biblioteca de recursos con especial énfasis en las matemáticas para aspirantes a programadores en áreas especializadas.

   Librerías de Python

   NumPy: es el paquete más usado para computación científica con Python. NumPy es una extensión de Python, que le agrega mayor soporte para vectores

y matrices, constituyendo una biblioteca de funciones matemáticas de alto nivel para operar con esos vectores o matrices. Contiene, entre otras cosas:

- **Potentes matrices n-dimensionales:** Rápidos y versátiles, los conceptos de vectorización, indexación y transmisión de NumPy son los estándares de facto de la computación de arreglo en la actualidad.

- **Interoperable:** NumPy es compatible con una amplia gama de plataformas informáticas y de hardware, y funciona bien con bibliotecas distribuidas, GPU y de arreglos dispersos.

- **Herramientas de cómputo numérico:** NumPy ofrece funciones matemáticas completas, generadores de números aleatorios, rutinas de álgebra lineal, transformadas de Fourier y más.

- **Rendimiento:** El núcleo de NumPy es un código C bien optimizado. Es posible disfrutar de la flexibilidad de Python con la velocidad del código compilado.
- **Fácil de usar:** La sintaxis de alto nivel de NumPy lo hace accesible y productivo para programadores de cualquier origen o nivel de experiencia.

- **Fuente abierta:** Distribuido bajo una licencia BSD liberal, NumPy es desarrollado y mantenido públicamente en GitHub por una comunidad vibrante, receptiva y diversa.

NumPy forma la base de potentes bibliotecas de aprendizaje automático como scikit-learn y SciPy. A medida que crece el aprendizaje automático, también lo hace la lista de bibliotecas creadas en NumPy.

Pandas: es un paquete de Python que proporciona estructuras de datos rápidas, flexibles y expresivas diseñadas para que el trabajo con datos "relacionales" o "etiquetados" sea fácil e intuitivo. Sus objetivos son:

- Ser el bloque de construcción fundamental de alto nivel para realizar análisis de datos prácticos del mundo real en Python.

- Convertirse en la herramienta de análisis / manipulación de datos de código abierto más potente y flexible disponible en cualquier idioma.

Pandas es adecuado para muchos tipos diferentes de datos:

- Datos tabulares con columnas de tipos heterogéneos, como en una tabla SQL o una hoja de cálculo de Excel

- Datos de series de tiempo ordenados y desordenados (no necesariamente de frecuencia fija).

- Datos matriciales arbitrarios (homogéneos o heterogéneos) con etiquetas de fila y columna.

Cualquier otra forma de conjuntos de datos observacionales/estadísticos. Los datos en realidad no necesitan etiquetarse en absoluto para ser colocados en una estructura de datos de pandas.
Para los científicos de datos, el trabajo con datos generalmente se divide en múltiples etapas: recopilar y limpiar datos, analizarlos/modelarlos y luego organizar los resultados del análisis en una forma adecuada para trazarlos o mostrarlos en forma de tabla. Pandas es la herramienta ideal para todas estas tareas.

Matplotlib: es una librería de trazado 2D que produce gráficas de buena calidad en una variedad de formatos y entornos interactivos. Se puede generar gráficas, histogramas, espectros de potencia, gráficas de barras, gráficas de errores, diagramas de dispersión, etc., con unas pocas líneas de código.

Scikit-learn: es una biblioteca en Python que proporciona muchos algoritmos de aprendizaje supervisados y no supervisados. Proporciona una selección de herramientas eficientes para el aprendizaje automático y el modelado estadístico. A través de una interfaz de consistencia en Python, provee funciones para la clasificación, la regresión, el agrupamiento y la reducción de la dimensionalidad.

b. Entorno de programación: Colab

Colaboratory, o "Colab" para abreviar, es un servicio de Google Research en la nube basado en Jupyter Notebook que no requiere configuración para usarlo, se puede compartir y brinda acceso gratuito a recursos computacionales, incluidas GPU.

Colab permite que todos puedan escribir y ejecutar código arbitrario de Python 3 en el navegador, el código se ejecuta en una máquina virtual exclusiva para la cuenta desde la que se está utilizando, es importante tener en cuenta que las máquinas virtuales se borran cuando están inactivas durante un tiempo prolongado y tienen una vida útil máxima determinada por el sistema de Colab. Google Colaboratory es un proyecto que tiene como objetivo difundir la educación, la investigación del aprendizaje automático y el análisis de datos.

Pruebas

Para realizar la imputación de las señales incompletas se tomaron las sesiones que presentaban el 5% o más en cantidad de datos IBI existentes. A partir de este límite, fueron imputadas un total de 58 sesiones de las 142 totales.

La imputación MICE se implementó con 1, 5, 10, 20 y 30 iteraciones, mientras que la imputación por KNN se aplicó para 4, 5, 6, 7 y 8 vecinos. Estos valores fueron seleccionados teniendo en cuenta los valores de iteraciones y vecinos más cercanos que se utilizan comúnmente en este tipo de practica y que se recomiendan utilizar en la literatura revisada [85], [92], [96], [97].

3.2.4. Resultados de la imputación de señales

A continuación, en la Tabla 14 se presenta el análisis estadístico de la señal IBI sin imputar comparada con el análisis de la misma señal imputada con los métodos mencionados anteriormente. Tal como en las tablas anteriores, se presenta un ejemplo de una sesión de datos de un solo paciente. En el Anexo A, se encuentran las demás tablas obtenidas.

Tabla 14. Comparación análisis estadístico con técnicas de imputación utilizadas

	IBI	MICE1	MICE 5	MICE 10	MICE2 0	MICE 30	KNN3	KNN4	KNN5	KNN6	KNN7	KNN8
Media	0.825	0.824	0.824	0.824	0.824	0.824	0.816	0.816	0.816	0.817	0.818	0.818
Mediana	0.830	0.830	0.830	0.830	0.830	0.830	0.813	0.818	0.814	0.817	0.817	0.818
Moda	0.860	0.860	0.860	0.860	0.860	0.860	0.840	0.825	0.800	0.810	0.860	0.860
Mínimo	0.560	0.560	0.560	0.560	0.560	0.560	0.560	0.560	0.560	0.560	0.560	0.560
Máximo	1.110	1.110	1.110	1.110	1.110	1.110	1.110	1.110	1.110	1.110	1.110	1.110
Rango	0.550	0.550	0.550	0.550	0.550	0.550	0.550	0.550	0.550	0.550	0.550	0.550
Desviación estándar	0.071	0.050	0.050	0.050	0.050	0.050	0.060	0.058	0.056	0.055	0.055	0.054
Varianza	0.005	0.002	0.002	0.002	0.002	0.002	0.004	0.003	0.003	0.003	0.003	0.003
Primer cuartil	0.780	0.801	0.801	0.801	0.801	0.801	0.780	0.780	0.780	0.783	0.786	0.789
Segundo cuartil	0.830	0.830	0.830	0.830	0.830	0.830	0.813	0.818	0.814	0.817	0.817	0.818
Tercer cuartil	0.880	0.847	0.847	0.847	0.847	0.847	0.853	0.853	0.850	0.852	0.850	0.850
Asimetría	-0.085	-0.310	-0.310	-0.310	-0.310	-0.310	-0.041	-0.019	0.023	0.026	0.041	0.019
Curtosis	0.373	2.351	2.351	2.351	2.351	2.351	0.427	0.576	0.734	0.847	0.902	0.998

La señal IBI de la sesión 1 del paciente 4 tiene un total de 73% de datos perdidos. Debido a que los datos de imputación se acercan bastante a los datos estadísticos de la señal original según la Tabla 4, el método final de imputación seleccionado fue la imputación KNN con 3 vecinos. Para esta selección, se tomó como prioridad las medidas registradas en la media, el rango, la desviación estándar, la asimetría y la curtosis. Esta decisión se debe a que estas medidas estadísticas describen con mayor exactitud la forma como se comportan los datos. Se dio poca prioridad a las medidas que representan una posición como la mediana, primer, segundo y tercer cuartil, ya que las posiciones de los datos existentes pueden variar durante el proceso de imputación, lo cual no permite que sean una referencia de comparación.

En el ejemplo expuesto en la Tabla 14, se puede observar que los valores estadísticos de KNN3 son los más cercanos a los valores originales, presentando una diferencia de 0,009 en la media; de 0,011 en la desviación estándar; de 0,044 en la asimetría y de 0,054 en la curtosis.

Para el caso de este trabajo de grado, la técnica de imputación simple se ajustó mucho mejor a los datos que la imputación múltiple. Esto puede presentarse por la cantidad de datos perdidos y el tamaño del conjunto de datos. Existen registros de más

investigaciones en las que aplicar una imputación simple puede llegar a ser más benéfico y mostrar mejores resultados en los datos que una imputación múltiple [94].

3.3. Resumen

En este capítulo se abordaron las etapas de CRISP-DM (sección 1.5) acerca del entendimiento de los datos y la primera parte de la preparación de los datos teniendo como fuente las señales fisiológicas coleccionadas con la pulsera E4. Para entender los datos, se analizaron las variables estadísticas descriptivas de cada una de las señales. Para la preparación de los datos, se realizó una aproximación desde la imputación de datos para completar las partes perdidas de la señal IBI.

Con los resultados obtenidos y descritos en este capítulo, se cumple el primer, segundo objetivo específico de este trabajo de investigación, y parte del tercer objetivo específico.

Capítulo 4

Preparación y modelado de datos

Este capítulo narra el proceso de preparación de los datos que complementa el proceso de imputación de los datos. La preparación de los datos consta de los pasos: procesamiento de la señal, segmentación y corroboración de la calidad de los segmentos, etiquetado de los segmentos y la extracción de las características. Este capítulo también comprende el proceso de modelado de los datos por medio del aprendizaje automático supervisado. Una vez se ha generado el conjunto de datos con base en las señales, se revisa la calidad del conjunto y se determina que deben llevarse a cabo dos estrategias antes de la clasificación. Finalmente, el proceso de clasificación se lleva a cabo dividiendo el conjunto para realizar el entrenamiento, la validación, el ajuste de hiperparámetros y obtener las métricas finales que determinan el desempeño del modelo generado.

4.1. Preparación de los datos

El proceso de preparación de los datos obtenidos durante la recolección de datos consta de una serie de pasos que entregan como resultado final un conjunto de datos.

Un conjunto de datos se encuentra formado por atributos, instancias y una clase objetivo; contar con esta última parte dependerá de la naturaleza del problema que se quiera resolver (para el caso que se tiene en este trabajo de grado se cuenta con la clase objetivo ya que el abordaje es sobre un problema de clasificación). Un conjunto de datos se puede observar cómo una tabla de datos; al hacer la comparación, los atributos son las columnas de la tabla, las instancias son las filas de la tabla. Aquello que hace a un conjunto de datos diferente a la información en una tabla es la clase objetivo, la cuál es la observación que se hace a los datos que están en una misma fila.

Con el fin de construir los conjuntos de datos necesarios para continuar con la generación de los modelos (modelo multimodal sin IBI y modelo multimodal con IBI imputado), la Figura 10 ilustra los pasos a seguir. (1) Procesamiento de la señal: los datos se preprocesan para reducir el ruido y eliminar artefactos. (2) Segmentación: los registros de datos iniciales se segmentan de acuerdo con las observaciones en el tiempo en que sucedieron. (3) Extracción de características: a menudo, los clasificadores deben entrenarse de acuerdo con características en lugar de datos sin procesar; esto sucede sobre todo por el tipo de dato fuente que se maneja, tales como las señales fisiológicas. Estas características intentan delinear una representación informativa de los datos originales [100]. (4) Etiquetado: una vez definidas las características, deben ser identificables por medio de la observación realizada sobre cada dato o segmento de datos.

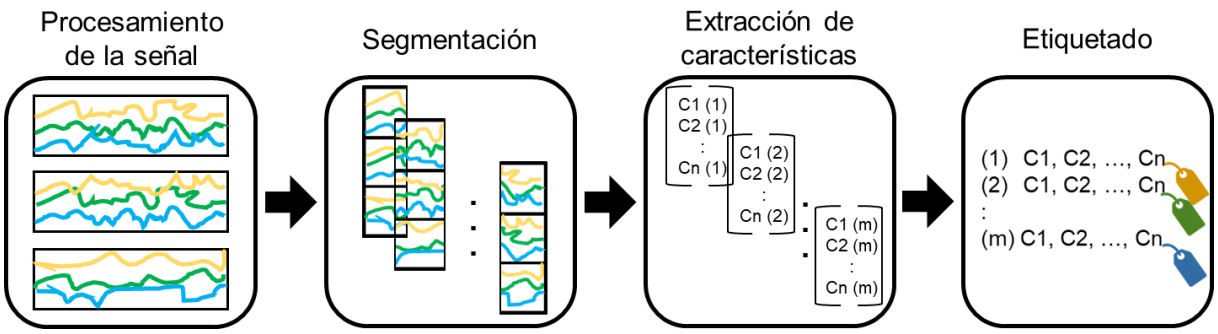

Figura 10. Proceso de preparación de los datos

4.1.1. Procesamiento de la señal

Las señales fisiológicas obtenidas por medio de los sensores de la pulsera E4 se descargan desde el gestor web que guarda la información. Estas señales en crudo presentan comúnmente algún tipo de contaminación cómo ruido, interferencia externa y artefactos de movimiento [101].

Generalmente, la remoción de la contaminación identificada se realiza a través de diferentes métodos, siendo los más aceptados y utilizados: descartar las bandas de señal inicial y final; aplicar filtros de suavizado; aplicar filtros pasa bajo, restar la línea de base, normalizar y discretizar [36], [49], [102].

Dado que en el capítulo anterior se realizó un proceso previo de análisis estadístico descriptivo de las señales, que dio como resultado que los datos se encontraban en rangos normales de operación, las señales completas de los dos conjuntos de sesiones (con IBI imputado y sin valores de IBI) no fueron procesadas por ruido o artefactos antes de la segmentación, ya que algunos otros artefactos pueden detectarse mejor con una mayor resolución de las señales en el tiempo.

4.1.2. Segmentación

La segmentación de las señales fisiológicas es un proceso necesario teniendo en cuenta el objetivo final de la recolección de los datos: reconocer el desempeño cognitivo de niños con TEA durante su interacción con el juego serio HapHop-Fisio. Esta segmentación se hizo según el contenido del juego, el cual fue abordado durante las sesiones de terapia. Cabe aclarar que la segmentación para este caso no es uniforme, ya que depende de la duración de cada juego y de los niveles alcanzados por cada niño. Es necesario aclarar que, además de las señales, también se obtuvo información sobre el desempeño del niño en cada juego, por medio de grabaciones en video de las sesiones con cada niño; esto se constituye como el insumo principal para la segmentación y posterior etiquetado de los datos.

Para segmentar las señales se tuvieron en cuenta los videos y la información de la base de datos del juego, para extraer los siguientes datos: modulo, categoría, modalidad, nivel, puntaje, tiempo inicial, tiempo final y duración de cada uno de los mini-juegos jugados por los niños en las sesiones de tratamiento. El tiempo de inicio, final y duración de cada mini-juego fue cotejado con la frecuencia de muestreo de cada una de las señales de los archivos tipo CSV (datos crudos de la pulsera E4) de cada conjunto de sesiones. Aquellas señales recolectadas durante los intervalos de tiempo en los que el niño no presentó ningún tipo de influencia cognitiva de parte del juego fueron descartadas, mientras que las señales recolectadas en los intervalos de tiempo que el niño jugo un mini-juego formaron los segmentos finales de esta etapa.

Calidad de los segmentos de las señales

Para garantizar la calidad de los datos de los segmentos de las señales, el proceso de limpieza de los segmentos siguió tres reglas [103]:

1. Valores del segmento fuera del rango: para evitar artefactos de "suelo" cuando el sensor pierde contacto la superficie (en este caso, la piel). Se eligieron valores mínimos aceptados para cada señal.

2. Cambios rápidos en los valores: esta regla se usa para prevenir altas frecuencias en intervalos cortos de tiempo.

3. Valores alrededor de artefactos: hay que tener en cuenta los efectos de transición para los valores cercanos en el tiempo a los artefactos. Dentro de los cinco segundos alrededor del artefacto, es mejor realizar la remoción de datos.

En la Tabla 15, se muestra cómo fue aplicada cada regla a las señales procesadas en este trabajo de grado.

Tabla 15. Reglas de calidad de los datos de las señales

Regla	Aplicación para cada señal
1. Valores fuera del rango	EDA: valor mínimo de 0.05 µS ACC: valor mínimo de -2g y máximo de 2g BVP: valor mínimo de -5 y máximo de 5 HR: valor mínimo de 60 lpm y máximo de 135 lpm IBI: valor mínimo de 0.025 TEMP: valor mínimo de 25 °C

2. Cambios rápidos	EDA: cambios mayores a ±10 µS/seg ACC: no hay valores restrictivos BVP: cambios mayores a ±150/seg HR: cambios mayores a ±30 lpm/seg IBI: cambios mayores a ±0.3/seg TEMP: cambios mayores a ±10 °C/seg

Las tres reglas descritas fueron aplicadas para cada segmento de señal. La Tabla 16 resume el número inicial de cada segmento de señal por niño y el número final de segmentos de señal que fueron utilizados para crear los conjuntos de datos. Al final, se tiene un conjunto de muestras con IBI imputado con 399 segmentos de señales y un conjunto de muestras sin valores IBI con 457 segmentos de señales.

Tabla 16. Número de segmentos de señal por paciente

Identificador del paciente	Sesiones	Segmentos	Segmentos válidos	Segmentos imputados
Paciente_1	13	77	71	58
Paciente_2	9	68	61	39
Paciente_3	17	127	89	9
Paciente_4	11	90	79	26
Paciente_5	10	59	49	0
Paciente_6	10	59	47	13
Paciente_7	10	70	54	27
Paciente_8	10	84	72	42
Paciente_9	9	66	56	12
Paciente_10	12	99	81	34
Paciente_11	10	67	47	9
Paciente_12	10	100	76	76
Paciente_13	11	89	74	54
Total			856	399

4.1.3. Extraction de características

Tras la segmentación, se identificaron cada uno de los fragmentos de las señales como el estímulo generado a partir de un cambio en la actividad cognitiva del niño durante el juego con HapHop-Fisio. Las ocho señales (X de ACC, Y de ACC, Z de ACC, BVP, EDA, HR, IBI y TEMP) se procesaron para obtener las características de cada una.

Dado que las señales fisiológicas se clasifican como señales no estacionarias, las técnicas de análisis multiresolución son apropiadas para representar la información contenida en estas señales. La Transformada Wavelet (TW) es una técnica ampliamente utilizada para el análisis multirresolución de datos variantes en el tiempo. La TW consiste en comparar la señal de interés con todas las traslaciones y escalas posibles de una misma función wavelet, conocida como la Wavelet madre. Así, se lleva a cabo el análisis de las señales para obtener las características de en el dominio wavelet. Como resultado, este dominio brinda una representación más detallada de la señal en comparación con la señal sin procesar en el dominio del tiempo.

Específicamente, para obtener el equivalente en el dominio de wavelet se consideraron 10 coeficientes de detalle wavelet, para cada coeficiente se calcularon 36 características resultantes de las combinaciones entre tres variaciones de amplitud (amplitud total de la señal, la amplitud normalizada y la amplitud absoluta), por tres razones matemáticas de cambio (señal básica, la primera derivada y la segunda derivada), por cuatro variables estadísticas (la media, la varianza, la curtosis y la desviación estándar). La Tabla 17 muestra, de forma visual, la combinación de estas características.

Tabla 17. Características extraídas en el dominio Wavelet

Variaciones de amplitud	Razones matemáticas de cambio	Variables estadísticas
Amplitud total Amplitud normalizada Amplitud absoluta	Señal básica Primera derivada Segunda derivada	Media Varianza Curtosis Desviación estándar

En total se extrajeron 360 características para cada señal fisiológica de los segmentos en el dominio Wavelet. Con la extracción de las características, finalmente se generaron dos conjuntos de muestras: el conjunto con IBI imputada que cuenta con 399 muestras (instancias) y 2880 características (atributos); el conjunto de datos sin valores IBI cuenta con 457 muestras (instancias) y 2520 características (atributos).

4.1.4. Etiquetado

De acuerdo con lo establecido hasta ahora, el interés que enmarca este trabajo de grado es conocer los aspectos reconocibles de la cognición a través de las señales obtenidas de la pulsera E4. El etiquetado final de los conjuntos de muestras se realizó de acuerdo con la variable de la cognición que indica el desempeño del niño durante cada uno de los mini-juegos. Este tipo de información es variable ya que la dificultad de los mini-juegos aumenta con el nivel, lo que consume más tiempo en el estímulo y en las respuestas del niño, tanto fisiológicas como cognitivas.

Para la tarea de etiquetar, se estableció la variable de desempeño en función de la puntuación del juego según el factor económico creado para HapHop-Fisio. La recompensa del juego está representada en estrellas: el jugador obtiene una, dos o tres estrellas por cada mini-juego completado. El jugador nunca pierde estrellas por responder incorrectamente a los desafíos de los mini-juegos y acumula mínimamente una estrella por cada mini-juego.

La clase objetivo que debe agregarse a los conjuntos de muestras para generar los conjuntos de datos finales que se esperan para la tarea de clasificación tiene tres clases: desempeño alto, desempeño medio y desempeño bajo. Los segmentos de las señales que corresponden a cada uno de los mini-juegos en los cuáles los niños obtuvieron la calificación de tres estrellas, se etiquetan con desempeño alto; mini-juegos calificados con dos estrellas se etiquetan con desempeño medio; y mini-juegos calificados con una sola estrella se etiquetan con desempeño bajo.

Finalmente, los conjuntos de datos tienen el siguiente arreglo: el conjunto de datos con IBI imputado es de la forma (399, 2881) con clase múltiple y el conjunto de datos sin valores IBI es de la forma (457, 2551) con clase múltiple. Por lo tanto, el problema de clasificación de los conjuntos de datos creados es multiclase.

4.2. Modelado basado en aprendizaje supervisado

El aprendizaje automático proporciona a los sistemas la capacidad de aprender y mejorar automáticamente a partir de la experiencia sin ser programados explícitamente. Para este trabajo de grado, algunos algoritmos de aprendizaje automático supervisados fueron utilizados; estos aplican lo que aprenden (entrenamiento) a nuevos datos usando muestras etiquetadas para predecir eventos futuros. Dado que las variables de etiquetado son cualitativas, fue necesario trabajar con algoritmos de clasificación.

4.2.1. Calidad del conjunto de datos

Una vez obtenidos los conjuntos de datos a partir de las señales fisiológicas para el reconocimiento del desempeño cognitivo de niños con TEA, se realizó una inspección de la calidad del conjunto de datos formado, dando como resultado la identificación de dos problemas: (1) el desbalance de clases en la clase objetivo y (2) la alta dimensionalidad del conjunto de datos.

Desbalance de clases

En un problema de clasificación con clases desbalanceadas se encuentra que la distribución de muestras entre las clases está sesgada o inclinada hacía alguna clase en particular. El desbalance podría presentarse desde un ligero sesgo hasta uno grave donde habría una muestra en la clase minoritaria y miles o millones de muestras en la clase o clases mayoritarias. La mayoría de los algoritmos de clasificación fueron diseñaron partiendo de la suposición de un número igual de muestras para cada clase, por ello, es necesario tratar de superar este desbalance.

Cuando se presenta este escenario en el conjunto de datos, los modelos tienen un rendimiento predictivo deficiente, específicamente para la clase minoritaria, por lo que el modelo es más sensible a cometer errores de clasificación para esta clase. En la Tabla 18 se presenta la información que da cuenta del desbalance de las clases en los conjuntos de datos obtenidos de las señales fuente. Cabe resaltar que la clase 0 corresponde a la etiqueta “desempeño bajo”, la clase 1 corresponde a la etiqueta “desempeño medio”, y la clase 2 corresponde a la etiqueta “desempeño alto”.

Tabla 18. Conjuntos de datos con desbalanceo de clases

Conjuntos	Con IBI	Sin IBI
Instancias totales	457 instancias	399 instancias
Muestras en clase 0	92	100
Muestras en clase 1	107	101
Muestras en clase 2	258	198

La relación de desbalance entre la clase mayoritaria y las clases minoritarias del conjunto de datos con IBI es de 2,8:1 (clase 0) y de 2,4:1 (clase 1), es decir, por cada 14 muestras en la clase 2, hay 5 muestras en la clase 0. Para el caso del conjunto de datos sin IBI, las relaciones son 1,98:1 (clase 0) y 1,96:1 (clase 1).

Se pueden identificar dos grupos principales de causas del desbalance: el proceso de muestreo de los datos y las propiedades del dominio. Para el caso de los dos conjuntos de datos que fueron obtenidos, el desbalance se presentaba por propiedad del dominio del problema. La clasificación de “alto desempeño” del conjunto de datos era el que más muestras presentaba, esto debido a que el objetivo de las terapias con los niños era la completitud de cada uno de los mini-juegos con el mejor puntaje posible. Dado que, se trataba de un juego, los pacientes siempre hicieron un esfuerzo por ganar, lo que ayudaba en el cumplimiento del propósito de HapHop-Fisio, el cuál es brindar un soporte para las terapias de aprendizaje con los niños que sufren de TEA.

Sin embargo, este dominio de la clase “alto desempeño” debe manejare con cuidado, pues para un posterior desarrollo e implementación del juego, la clase minoritaria sería de mayor interés, es decir, la clase “bajo desempeño”. Desde el dominio del problema, la detección de las áreas en que el niño presenta más dificultad es de mayor interés para el terapeuta.

Existen muchas soluciones para el desbalance de clases, una de ellas es cambiar el conjunto de datos para que existan datos balanceados. Este cambio se llama muestreo del conjunto de datos y hay dos métodos principales que igualan las clases: (1) agregar copias de instancias de la clase subrepresentada, esto es sobremuestreo, o (2) eliminar instancias de la clase sobrerrepresentada, esto es submuestreo. Ya que la cantidad de datos es baja para el estándar del aprendizaje automático, se consideró utilizar el sobremuestreo.

Uno de los métodos más utilizados es el SMOTE (*Synthetic Minority Over-sampling Technique*). Funciona creando muestras sintéticas de la clase menor en lugar de crear copias. El algoritmo selecciona dos o más instancias similares (usando una medida de distancia) y perturba una instancia un atributo a la vez por una cantidad aleatoria dentro de la diferencia con las instancias vecinas [104].

Después de utilizar esta técnica de sobremuestreo, el conjunto de datos sin valores IBI tiene un arreglo final con 774 instancias y cada una de sus clases con 258 muestras por igual; así mismo, el conjunto de datos con IBI imputado tiene un arreglo final con 594 instancias y cada una de sus clases con 198 muestras.

Alta dimensionalidad

El rendimiento de los algoritmos de aprendizaje automático puede verse altamente afectado si el conjunto de datos presenta demasiadas variables de entrada. En un conjunto de datos tabulares que contiene filas y columnas, las columnas representan las dimensiones del espacio de características. Por ello, la alta dimensionalidad se refiere a un conjunto de datos en el que el número de características es mucho mayor que el número de muestras, es decir, el número de atributos supera ampliamente al número de instancias [105].

La alta dimensionalidad impide que el algoritmo de aprendizaje pueda encontrar un modelo que describa la relación entre las variables predictoras (atributos) y la clase objetivo porque no existen suficientes instancias para entrenar el modelo. Los conjuntos de datos con alta dimensionalidad son comunes en los conjuntos de datos de salud (variables fisiológicas) donde la cantidad de características para un individuo determinado puede ser enorme [106].

Por lo tanto, a menudo es deseable reducir el número de variables de entrada. Este proceso es conocido como “reducción de dimensionalidad”. Menos dimensiones de entrada pueden significar menos parámetros o una estructura más simple en el modelo de aprendizaje automático. Al mismo tiempo que la reducción de dimensionalidad reduce la cantidad de atributos, debe velarse porque mantenga la mayor variación posible en el conjunto de datos original. Igualmente, un menor número de dimensiones en los datos significa menos tiempo de entrenamiento y menos recursos

computacionales y podría aumentar el rendimiento de los algoritmos que generan los modelos.

Algunas ventajas de la reducción de dimensionalidad:

- Evita el problema del sobreajuste del modelo
- Es útil para la visualización de datos.
- Es útil para el análisis factorial.
- Elimina el ruido en los datos [107]

Existen tres enfoques de reducción de la dimensionalidad. (i) Filtro: calcula los coeficientes de correlación entre los atributos y la clase objetivo, luego selecciona las características con la correlación más alta. (ii) *Wrapper*: construye modelos con todas las combinaciones de características, donde el subconjunto de funciones es seleccionado con base en el rendimiento del modelo. (iii) Embebido: durante el proceso de entrenamiento, agrega la selección de características por métodos de *Wrapper* reduciendo el tiempo de cálculo necesario para reclasificar los diferentes subconjuntos [108].

En este trabajo de grado, se realizaron experimentos para dos de los tres enfoques de reducción de dimensionalidad: filtro y *wrapper*. En la aproximación por filtro, se definió que los límites en el valor de la correlación entre las características y la variable objetivo serían 0,1 para la correlación positiva y -0,1 para la correlación negativa. La aproximación por Wrapper utilizó como algoritmo base para la creación de los subconjuntos de características el algoritmo de RandomForest. En la Tabla 19 se encuentra el resumen de los resultados de los experimentos de reducción de dimensionalidad.

Tabla 19: Número de atributos después de la reducción de dimensionalidad

Enfoque	Conjunto de datos	
	Con IBI	Sin IBI
Filtro	245 atributos	153 atributos
Wrapper	20 atributos	14 atributos

Selección de características por filtro

La correlación entre los atributos y la clase objetivo se obtuvo con el método de Pearson[5].

[5] Muestra la relación lineal entre dos conjuntos de datos.

Selección de características por *Wrapper*

La reducción de dimensionalidad por el enfoque de Wrapper arrojó cómo resultados finales un conjunto de características seleccionadas a partir de la evaluación de 47.711 subconjuntos para el conjunto de datos sin valores IBI y de 86.021 subconjuntos para seleccionar los atributos del conjunto de datos con IBI imputado.

Se obtuvieron un total de 14 atributos seleccionados para el conjunto de datos sin valores IBI. La tabla 20 describe el número de atributo y el nombre.

Tabla 20. Características seleccionadas por Wrapper para conjunto de datos sin IBI

Número ordinal del atributo	Nombre del atributo
6	X_v_d2d2n
141	X_m_d4d2a
193	X_m_d6d1r
289	X_m_d9r
335	X_k_d10a
351	X_k_d10d2r
553	Y_m_d6d1r
693	Y_m_d10a
1322	BVP_v_d7d2r
1324	BVP_std_d7d2r
1361	BVP_m_d8d2n
1533	EDA_m_d3d1a
2303	TEMP_k_d4d2a
2377	TEMP_m_d7r

Para el conjunto de datos con IBI imputados, se obtuvieron un total de 20 atributos seleccionados. La tabla 21 describe el número de atributo y el nombre.

Tabla 21. Características seleccionadas por Wrapper para conjunto de datos con IBI imputado

Número ordinal del atributo	Nombre del atributo
97	X_m_d3d2r
121	X_m_d4d1r
270	X_v_d8d1n
401	Y_m_d2n
414	Y_v_d2d1n
417	Y_m_d2d1a
485	Y_m_d4d1n
608	Y_std_d7d2n
646	Y_v_d8d2a
937	Z_m_d7r
1021	Z_m_d9d1r
1097	BVP_m_d1d1n

1191	BVP_k_d4r
1229	BVP_m_d5n
1737	EDA_m_d9a
1763	EDA_k_d9d2a
1802	HR_v_d1r
2449	TEMP_m_d9r
2609	IBI_m_d3d1n
2618	IBI_v_d3d2r

4.2.2. Clasificación y generación del modelo

Después de realizar la preparación de los datos (incluyendo la evaluación de la calidad de los conjuntos de datos) se procede a generar los modelos de aprendizaje automático por medio de la clasificación. En el aprendizaje automático, la clasificación se refiere a un problema de modelado predictivo en el que se "predice" una etiqueta de clase para un ejemplo dado de datos de entrada (instancias).

La clasificación requiere de un conjunto de datos de **entrenamiento** con muchos ejemplos de entradas y salidas de las cuales aprender. Un modelo utiliza dicho conjunto de datos de entrenamiento y calcula la mejor manera de asignar ejemplos de datos de entrada a etiquetas de clase específicas.

Cómo se había mencionado anteriormente, la tarea de clasificación para los conjuntos de datos que se tienen es una clasificación multiclase. Con el fin de crear un modelo de aprendizaje automático confiable, se dividieron los conjuntos de datos en tres tipos de conjuntos: de entrenamiento, de validación y de prueba.

División del conjunto de datos

Conjunto de entrenamiento

Es el conjunto de datos que se utiliza para entrenar y hacer que el modelo aprenda las características y/o patrones que se encuentran ocultos en los datos. El conjunto de entrenamiento debería presentar diversidad en sus entradas para que el modelo se entrene en todos los escenarios y pueda predecir cualquier muestra de datos futura.

Conjunto de validación

Este conjunto se utiliza para validar el rendimiento del modelo durante el entrenamiento. La validación ayuda en la toma de decisiones sobre el ajuste de los hiperparámetros y las configuraciones del algoritmo que está generando el modelo. El modelo se entrena en el conjunto de entrenamiento y, simultáneamente, la evaluación del modelo se realiza en el conjunto de validación después de iteraciones. La idea principal de tener un conjunto de validación es evitar un sobreajuste en el modelo.

Conjunto de prueba

El conjunto de prueba es un conjunto separado de datos que se utiliza para probar el modelo después de completar el entrenamiento. Proporciona las métricas de rendimiento del modelo final de manera imparcial. Con este conjunto se conoce que tan bien funciona el modelo [109].

Existe un flujo de trabajo estándar para utilizar los conjuntos de datos en la generación de modelos de clasificación:

1. El conjunto de datos de entrenamiento se usa para entrenar algunos modelos candidatos
2. El conjunto de datos de validación se utiliza para evaluar los modelos candidatos
3. Uno de los candidatos es elegido
4. El modelo elegido y entrenado se evalúa con el conjunto de datos de prueba

En los pasos 1 y 2, no se pretende evaluar los modelos candidatos una sola vez. Es preferible evaluar cada modelo varias veces con diferentes conjuntos de datos y tomar la puntuación promedio para el paso 3. Al no contar con una alta cantidad de datos para poder generar los tres conjuntos de datos, es necesario utilizar la técnica de validación cruzada [110]. En la validación cruzada, se determina un número fijo de pliegues (o particiones) de los datos, se ejecuta el análisis en cada pliegue y luego, se promedia la estimación general del error [111].

Para el caso de este trabajo de grado, el método de validación cruzada se utilizó con 10 pliegues. De acuerdo con la literatura, el conjunto de datos de entrenamiento debe representar una gran parte de las muestras de todo el conjunto de datos total. Se definió que el 70% de datos fuera para el entrenamiento y la validación (proceso de

refinamiento de los hiperparámetros del modelo) y el 30% de datos restantes fuera para las pruebas finales del modelo. De esta manera, el detalle en la partición final de los conjuntos de datos obtenidos se ilustra en la Figura 11.

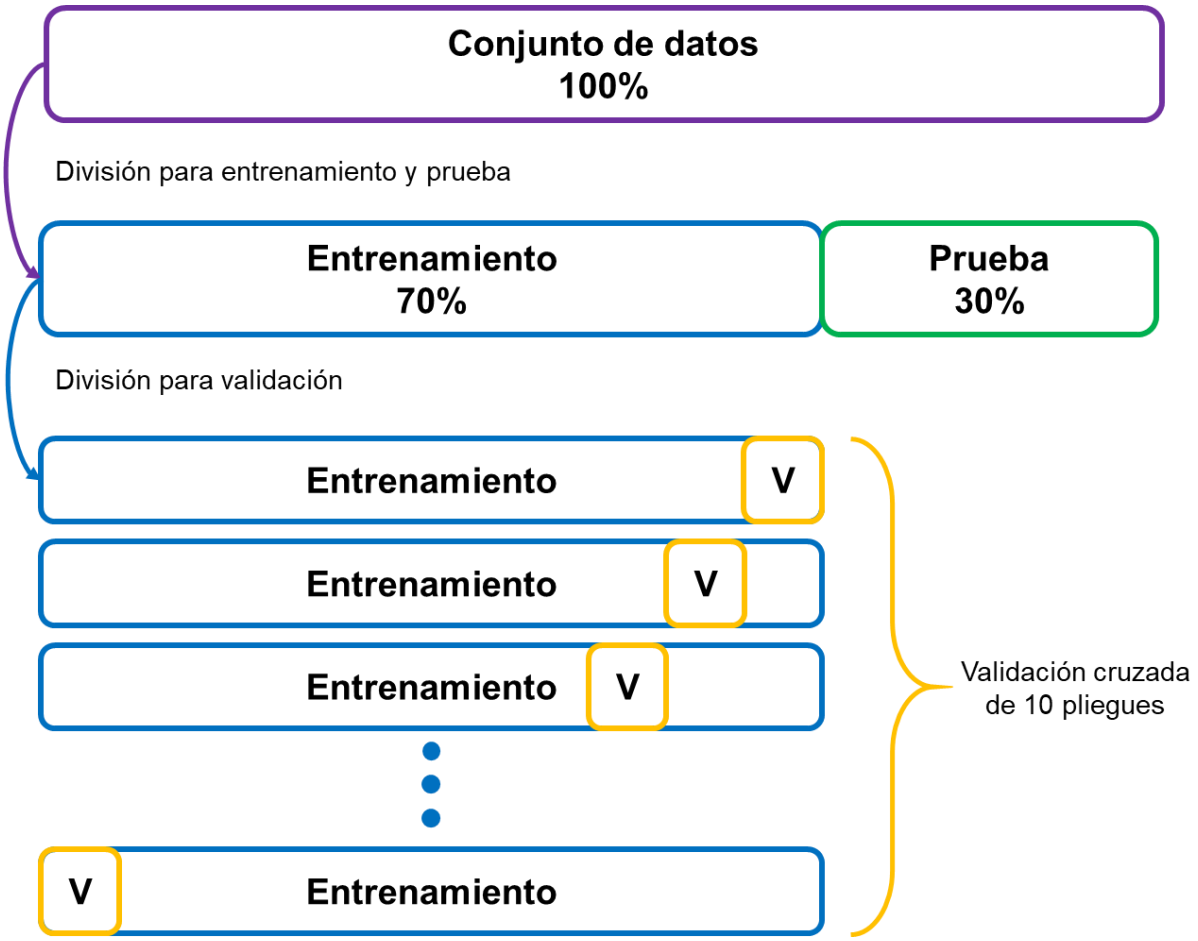

Figura 11. División para entrenamiento/prueba con validación cruzada

Resultados de clasificación

La figura 12 muestra un resumen del flujo para la ejecución de todos los experimentos de clasificación de los datos obtenidos. Se tiene un total de seis experimentos dadas las variaciones que los conjuntos de datos podían presentar.

1. Resultado 1: No se evalúa previamente la calidad del conjunto de datos, se aplicó directamente métodos de clasificación.

2. Resultado 2: Se evalúa calidad del conjunto de datos, inicialmente se corrige desbalance de clases con algoritmo SMOTE, posteriormente se aplicó métodos de clasificación al conjunto de datos balanceado.
3. Resultado 3: Se evalúa calidad del conjunto de datos, inicialmente se corrige la alta dimensionalidad de los datos con el método de filtro, posteriormente se aplicó métodos de clasificación al conjunto de datos con dimensionalidad reducida.
4. Resultado 4: Se evalúa calidad del conjunto de datos, inicialmente se corrige la alta dimensionalidad de los datos con el método Wrapper, posteriormente se aplicó métodos de clasificación al conjunto de datos con dimensionalidad reducida.
5. Resultado 5: Se evalúa calidad del conjunto de datos, se corrige el desbalance de clases con algoritmo SMOTE y la alta dimensionalidad de datos con el método de filtro, posteriormente se aplicó métodos de clasificación al conjunto de datos balanceado y con dimensionalidad reducida por filtro.
6. Resultado 6: Se evalúa calidad del conjunto de datos, se corrige el desbalance de clases con algoritmo SMOTE y la alta dimensionalidad de datos con el método Wrapper, posteriormente se aplicó métodos de clasificación al conjunto de datos balanceado y con dimensionalidad reducida por Wrapper.

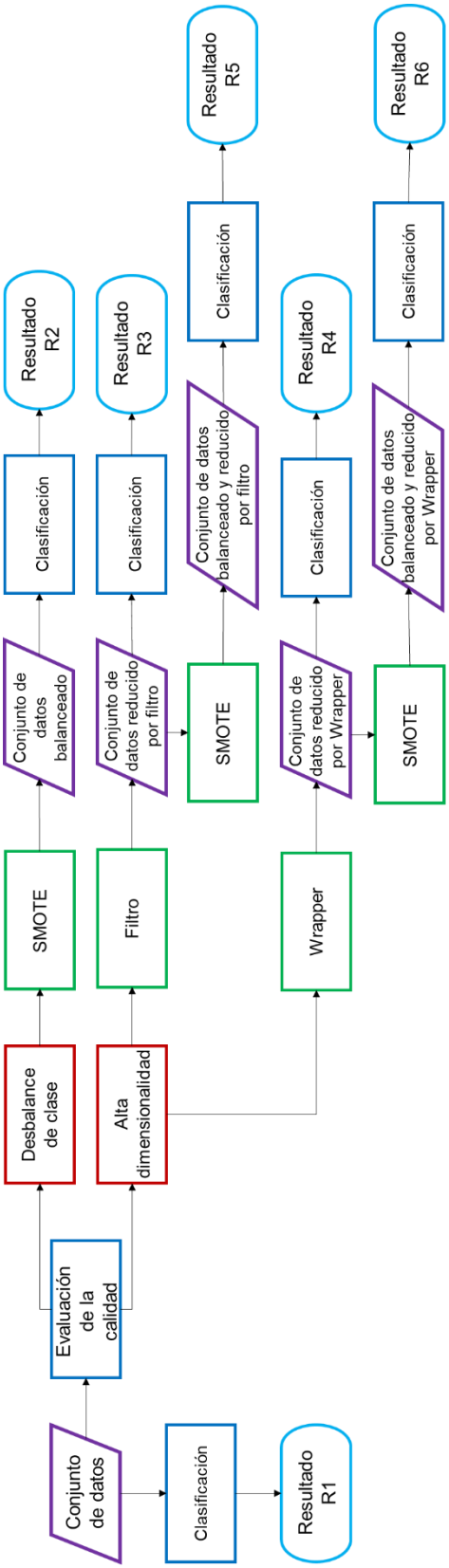

Figura 12. Flujo de experimentos de clasificación

En este punto del desarrollo del trabajo de grado investigativo, se han llevado a cabo todas las comprobaciones necesarias (calidad del conjunto de datos y división de los datos) para iniciar con el modelado por medio de diferentes algoritmos de aprendizaje supervisado. Se recuerda que este trabajo fue concebido desde un problema claro de clasificación: una exactitud del 79,80% con el algoritmo RandomForest para el reconocimiento unimodal del desempeño cognitivo a partir de las características de la señal EDA.

Por ello, a continuación, se presentan los resultados de clasificación de acuerdo con los siguientes algoritmos:

- RandomForest
- SupportVectorMachine (SVM)
- NeuralNetwork (MultiLayer Perceptron - MLP)

Experimento 1: Resultado R1

En este experimento se toma el conjunto de datos cómo se obtuvo en la sección 4.1.

Tabla 22. R1 (Random Forest) – Conjunto de datos sin IBI

Métricas	Numero de árboles										
	100	150	200	250	300	350	400	450	500	550	600
Validación cruzada	0,548	0,551	0,551	0,555	0,558	0,554	0,554	0,558	0,558	0,558	0,558
Resultados conjunto de pruebas											
Exactitud	0,572	0,565	0,565	0,565	0,550	0,550	0,557	0,550	0,557	0,557	0,557
Precisión	0,634	0,522	0,523	0,523	0,521	0,521	0,520	0,186	0,520	0,522	0,520
Sensibilidad	0,359	0,340	0,340	0,340	0,332	0,332	0,336	0,324	0,336	0,336	0,336
Medida F1	0,303	0,262	0,263	0,263	0,259	0,259	0,260	0,236	0,260	0,261	0,260

Tabla 23. R1 (Random Forest) – Conjunto de datos con IBI

Métricas	Número de árboles										
	100	150	200	250	300	350	400	450	500	550	600
Validación cruzada	0,440	0,451	0,458	0,455	0,465	0,476	0,473	0,451	0,458	0,469	0,465
Resultados conjunto de pruebas											
Exactitud	0,45	0,458	0,466	0,483	0,491	0,483	0,483	0,483	0,491	0,508	0,491
Precisión	0,340	0,286	0,269	0,342	0,344	0,342	0,309	0,307	0,377	0,380	0,377
Sensibilidad	0,289	0,289	0,294	0,304	0,309	0,304	0,304	0,304	0,315	0,325	0,315
Medida F1	0,261	0,247	0,248	0,255	0,257	0,255	0,254	0,252	0,273	0,279	0,273

Tabla 24. R1 (SVM) – Conjunto de datos sin IBI

Métricas	Tipo de kernel	
	lineal	rbf
Validación cruzada	0,413407	0,56744
Resultados conjunto de pruebas		
Exactitud	0,347826	0,565217
Precisión	0,292492	0,189781
Sensibilidad	0,289263	0,333333
Medida F1	0,28721	0,24186

Tabla 25. R1 (SVM) – Conjunto de datos con IBI

Métricas	Tipo de kernel	
	lineal	rbf
Validación cruzada	0,362169	0,47328
Resultados conjunto de pruebas		
Exactitud	0,366667	0,55
Precisión	0,331895	0,183333
Sensibilidad	0,335898	0,333333
Medida F1	0,330518	0,236559

Experimento 2: Resultado R2

Tabla 26. R2 (Random Forest) – Conjunto de datos sin IBI

Métricas	Numero de árboles										
	100	150	200	250	300	350	400	450	500	550	600
Validación cruzada	0,746	0,755	0,754	0,761	0,759	0,752	0,754	0,752	0,752	0,752	0,750
Resultados conjunto de pruebas											
Exactitud	0,712	0,725	0,708	0,712	0,729	0,721	0,716	0,716	0,716	0,716	0,725
Precisión	0,723	0,739	0,718	0,727	0,741	0,736	0,727	0,731	0,725	0,730	0,740
Sensibilidad	0,721	0,735	0,718	0,722	0,737	0,731	0,726	0,726	0,726	0,727	0,735
Medida F1	0,709	0,720	0,704	0,709	0,727	0,716	0,712	0,713	0,712	0,712	0,721

Tabla 27. R2 (Random Forest) – Conjunto de datos con IBI

Métricas	Numero de árboles										
	100	150	200	250	300	350	400	450	500	550	600
Validación cruzada	0,609	0,614	0,624	0,629	0,638	0,631	0,643	0,641	0,638	0,634	0,638
Resultados conjunto de pruebas											
Exactitud	0,698	0,703	0,692	0,698	0,687	0,675	0,664	0,681	0,670	0,664	0,664
Precisión	0,708	0,710	0,698	0,705	0,692	0,679	0,667	0,689	0,676	0,670	0,671
Sensibilidad	0,703	0,707	0,697	0,703	0,692	0,681	0,670	0,686	0,675	0,670	0,670
Medida F1	0,696	0,702	0,690	0,695	0,684	0,672	0,661	0,678	0,666	0,660	0,660

Tabla 28. R2 (SVM) – Conjunto de datos sin IBI

Métricas	Tipo de kernel	
	lineal	rbf
Validación cruzada	0,707845	0,380808
Resultados conjunto de pruebas		
Exactitud	0,742489	0,360515
Precisión	0,761547	0,287037
Sensibilidad	0,757543	0,381248
Medida F1	0,730561	0,275806

Tabla 29. R2 (SVM) – Conjunto de datos con IBI

Métricas	Tipo de kernel	
	lineal	rbf
Validación cruzada	0,636295	0,397619
Resultados conjunto de pruebas		
Exactitud	0,636872	0,318436
Precisión	0,642986	0,326517
Sensibilidad	0,643501	0,331785
Medida F1	0,629303	0,211444

Tabla 30. R2 (MLP) – Conjunto de datos sin IBI

Métricas	Número de capas		
	1	2	3
Validación cruzada	0,632458	0,698586	0,664141
Resultados conjunto de pruebas			
Exactitud	0,686695	0,532189	0,72103
Precisión	0,685408	0,587636	0,722864
Sensibilidad	0,698791	0,555488	0,734956
Medida F1	0,680564	0,468776	0,711657

Tabla 31. R2 (MLP- 1 capa) – Conjunto de datos con IBI

Métricas	Número de capas		
	1	2	3
Validación cruzada	0,646225	0,622184	0,636585
Resultados conjunto de pruebas			
Exactitud	0,558659	0,648045	0,519553
Precisión	0,568388	0,647392	0,538272
Sensibilidad	0,559912	0,652177	0,526415
Medida F1	0,55943	0,645002	0,50896

Experimento 3: Resultado R3

Tabla 32. R3 (Random Forest) – Conjunto de datos sin IBI

Métricas	Numero de árboles										
	100	150	200	250	300	350	400	450	500	550	600
Validación cruzada	0,532	0,551	0,542	0,533	0,536	0,545	0,545	0,539	0,536	0,539	0,533
Resultados conjunto de pruebas											
Exactitud	0,586	0,601	0,586	0,579	0,572	0,572	0,572	0,579	0,586	0,579	0,572
Precisión	0,492	0,482	0,461	0,425	0,416	0,416	0,448	0,451	0,411	0,394	0,371
Sensibilidad	0,372	0,388	0,379	0,375	0,365	0,365	0,365	0,369	0,373	0,369	0,365
Medida F1	0,321	0,346	0,339	0,336	0,320	0,320	0,320	0,323	0,326	0,323	0,320

Tabla 33. R3 (Random Forest) – Conjunto de datos con IBI

Métricas	Numero de árboles										
	100	150	200	250	300	350	400	450	500	550	600
Validación cruzada	0,426	0,437	0,426	0,419	0,433	0,430	0,422	0,419	0,426	0,437	0,433
Resultados conjunto de pruebas											
Exactitud	0,491	0,483	0,516	0,491	0,5	0,508	0,508	0,508	0,516	0,508	0,5
Precisión	0,387	0,353	0,406	0,387	0,403	0,431	0,431	0,431	0,434	0,431	0,428
Sensibilidad	0,347	0,338	0,368	0,347	0,352	0,363	0,363	0,363	0,368	0,363	0,358
Medida F1	0,340	0,328	0,360	0,340	0,345	0,360	0,360	0,360	0,363	0,360	0,357

Tabla 34. R3 (SVM) – Conjunto de datos sin IBI

Métricas	Tipo de kernel	
	lineal	rbf
Validación cruzada	0,548488	0,570565
Resultados conjunto de pruebas		
Exactitud	0,536232	0,572464
Precisión	0,384291	0,303704
Sensibilidad	0,356074	0,345238
Medida F1	0,330502	0,265637

Tabla 35. R3 (SVM) – Conjunto de datos con IBI

Métricas	Tipo de kernel	
	lineal	rbf
Validación cruzada	0,45172	0,47328
Resultados conjunto de pruebas		
Exactitud	0,391667	0,55
Precisión	0,312624	0,183333
Sensibilidad	0,296174	0,333333

Medida F1	0,301346	0,236559

Experimento 4: Resultado R4

Tabla 36. R4 (Random Forest) – Conjunto de datos sin IBI

	100	150	200	250	300	350	400	450	500	550	600
Validación cruzada	0,579	0,567	0,576	0,561	0,567	0,567	0,567	0,567	0,567	0,567	0,570
Resultados conjunto de pruebas											
Exactitud	0,528	0,543	0,528	0,543	0,543	0,536	0,543	0,543	0,543	0,550	0,543
Precisión	0,348	0,368	0,312	0,368	0,357	0,346	0,374	0,374	0,374	0,386	0,374
Sensibilidad	0,344	0,346	0,331	0,346	0,340	0,336	0,346	0,346	0,346	0,350	0,346
Medida F1	0,308	0,303	0,284	0,303	0,289	0,286	0,302	0,302	0,302	0,306	0,302

Tabla 37. R4 (Random Forest) – Conjunto de datos con IBI

Métricas	Numero de árboles										
	100	150	200	250	300	350	400	450	500	550	600
Validación cruzada	0,469	0,480	0,476	0,473	0,455	0,466	0,480	0,483	0,483	0,483	0,491
Resultados conjunto de pruebas											
Exactitud	0,508	0,491	0,5	0,516	0,516	0,533	0,525	0,533	0,533	0,533	0,533
Precisión	0,378	0,335	0,353	0,370	0,418	0,447	0,445	0,447	0,447	0,447	0,447
Sensibilidad	0,347	0,332	0,348	0,353	0,353	0,363	0,358	0,363	0,363	0,363	0,363
Medida F1	0,328	0,307	0,331	0,332	0,336	0,344	0,341	0,344	0,344	0,344	0,344

Tabla 38. R4 (SVM) – Conjunto de datos sin IBI

Métricas	Tipo de kernel	
	lineal	rbf
Validación cruzada	0,564214	0,557964
Resultados conjunto de pruebas		
Exactitud	0,550725	0,557971
Precisión	0,297678	0,187348
Sensibilidad	0,330929	0,32906
Medida F1	0,254897	0,23876

Tabla 39. R4 (SVM) – Conjunto de datos con IBI

Métricas	Tipo de kernel	
	lineal	rbf
Validación cruzada	0,469709	0,47328
Resultados conjunto de pruebas		
Exactitud	0,55	0,541667
Precisión	0,183333	0,182073

Sensibilidad	0,333333	0,328283
Medida F1	0,236559	0,234234

Experimento 5: Resultado R5

Tabla 40. R5 (Random Forest) – Conjunto de datos sin IBI

Métricas	Numero de árboles										
	100	150	200	250	300	350	400	450	500	550	600
Validación cruzada	0,685	0,682	0,709	0,704	0,708	0,715	0,719	0,708	0,709	0,709	0,715
Resultados conjunto de pruebas											
Exactitud	0,639	0,678	0,682	0,673	0,669	0,669	0,669	0,665	0,678	0,686	0,682
Precisión	0,648	0,685	0,687	0,679	0,672	0,678	0,675	0,669	0,684	0,693	0,690
Sensibilidad	0,649	0,687	0,692	0,683	0,678	0,680	0,679	0,675	0,687	0,696	0,693
Medida F1	0,634	0,674	0,677	0,670	0,666	0,663	0,664	0,660	0,674	0,682	0,676

Tabla 41. R5 (Random Forest) – Conjunto de datos con IBI

Métricas	Numero de árboles										
	100	150	200	250	300	350	400	450	500	550	600
Validación cruzada	0,631	0,619	0,638	0,633	0,628	0,641	0,643	0,638	0,638	0,636	0,633
Resultados conjunto de pruebas											
Exactitud	0,603	0,592	0,592	0,592	0,592	0,592	0,586	0,597	0,592	0,586	0,603
Precisión	0,602	0,591	0,592	0,591	0,594	0,593	0,588	0,600	0,592	0,588	0,605
Sensibilidad	0,608	0,597	0,597	0,596	0,597	0,597	0,592	0,603	0,597	0,592	0,609
Medida F1	0,600	0,588	0,589	0,589	0,589	0,588	0,582	0,592	0,587	0,582	0,598

Tabla 42. R5 (SVM) – Conjunto de datos sin IBI

Métricas	Tipo de kernel	
	lineal	rbf
Validación cruzada	0,554377	0,354815
Resultados conjunto de pruebas		
Exactitud	0,540773	0,32618
Precisión	0,542547	0,320261
Sensibilidad	0,544443	0,337706
Medida F1	0,540857	0,230711

Tabla 43. R5 (SVM) – Conjunto de datos con IBI

Métricas	Tipo de kernel	
	lineal	rbf
Validación cruzada	0,527584	0,368699
Resultados conjunto de pruebas		

Exactitud	0,536313	0,402235
Precisión	0,53521	0,415245
Sensibilidad	0,54225	0,405468
Medida F1	0,531361	0,389769

Experimento 6: Resultado R6

Tabla 44. R6 (Random Forest) – Conjunto de datos sin IBI

Métricas	Numero de árboles										
	100	150	200	250	300	350	400	450	500	550	600
Validación cruzada	0,678	0,665	0,682	0,676	0,682	0,685	0,685	0,680	0,680	0,685	0,683
Resultados conjunto de pruebas											
Exactitud	0,652	0,660	0,673	0,690	0,682	0,678	0,686	0,690	0,682	0,682	0,678
Precisión	0,653	0,663	0,675	0,693	0,683	0,678	0,687	0,690	0,682	0,682	0,677
Sensibilidad	0,660	0,668	0,681	0,697	0,689	0,684	0,693	0,697	0,688	0,689	0,684
Medida F1	0,649	0,659	0,672	0,691	0,681	0,678	0,686	0,690	0,682	0,681	0,678

Tabla 45. R6 (Random Forest) – Conjunto de datos con IBI

Métricas	Numero de árboles										
	100	150	200	250	300	350	400	450	500	550	600
Validación Cruzada	0,628	0,647	0,662	0,657	0,660	0,665	0,667	0,665	0,672	0,667	0,672
Resultados conjunto de pruebas											
Exactitud	0,620	0,608	0,620	0,631	0,625	0,608	0,631	0,620	0,620	0,620	0,625
Precisión	0,623	0,616	0,626	0,636	0,634	0,616	0,642	0,628	0,631	0,628	0,632
Sensibilidad	0,624	0,612	0,625	0,635	0,630	0,614	0,636	0,625	0,625	0,625	0,630
Medida F1	0,618	0,608	0,617	0,630	0,624	0,606	0,628	0,617	0,617	0,617	0,624

Tabla 46. R6 (SVM) – Conjunto de datos sin IBI

Métricas	Tipo de kernel	
	lineal	rbf
Validación cruzada	0,707845	0,380808
Resultados conjunto de pruebas		
Exactitud	0,742489	0,360515
Precisión	0,761547	0,287037
Sensibilidad	0,757543	0,381248
Medida F1	0,730561	0,275806

Tabla 47. R6 (SVM) – Conjunto de datos con IBI

Métricas	Tipo de kernel	
	lineal	rbf

Validación cruzada	0,438908	0,380778
Resultados conjunto de pruebas		
Exactitud	0,363128	0,340782
Precisión	0,367273	0,293839
Sensibilidad	0,364258	0,351713
Medida F1	0,363309	0,233525

Discusión de los resultados

A partir de los resultados obtenidos con el flujo de experimentos, el mejor modelo presentó una exactitud del 74,2% sobre el conjunto de datos de prueba. Este resultado corresponde al experimento 6 con el algoritmo SVM sobre el conjunto de datos sin valores IBI. El mismo resultado se presenta en el experimento 2 con el mismo algoritmo y sobre el mismo conjunto de datos.

De estas observaciones, es importante notar que la diferencia entre los dos resultados radica en las diferencias sobre los conjuntos: mientras que en el experimento 2 se tiene alta dimensionalidad, este problema no existe en el experimento 6 al reducir la dimensionalidad por el método Wrapper. Por ello, se considera que el experimento 6 es mejor, ya que los recursos computacionales que puedan invertirse en la implementación de este modelo son mucho menores.

4.3. Índice de confianza

Aunque para la tarea de clasificación que se tenía en este proyecto de investigación se necesitaba la identificación de los problemas de calidad de los conjuntos de datos finales obtenidos de la preparación exhaustiva de los datos fuente, la realidad es que los conjuntos de datos obtenidos en un ambiente “real” van a presentar el problema del desbalance de clases. Por lo tanto, el despliegue de una posible solución a nivel de los usuarios potenciales del juego HapHop-Fisio y su herramienta de detección del desempeño cognitivo de niño con TEA durante las terapias debe contar con métricas

especiales que permitan dar confianza acerca de la predicción que se realiza con el modelo obtenido.

Las métricas de umbral son aquellas que cuantifican los errores de predicción de la clasificación. Están diseñadas para resumir la proporción o tasa de cuando una clase pronosticada no coincide con la clase esperada en un conjunto de datos. Hay dos grupos de métricas que pueden ser útiles para la clasificación con desbalance de clases, porque se centran en una clase.

El primer grupo son las métricas de sensibilidad-especificidad. La sensibilidad se refiere a la tasa positiva verdadera y resume qué tan bien se predijo la clase positiva; esta métrica es útil para el caso de este trabajo de grado. El segundo grupo son las métricas de precisión-recuperación. La precisión resume la fracción de ejemplos asignados a la clase positiva que pertenecen a la clase positiva. La recuperación resume qué tan bien se predijo la clase positiva y es el mismo cálculo que se hace era la sensibilidad. La precisión y la recuperación se pueden combinar en una sola métrica que busca equilibrarlas, llamada “medida F1”.

$$Medida\ F1 = (2 * Precisión * Recuperación) / (Precisión + Recuperación)$$

Ecuación 2. Cálculo de la medida F1

La medida F1 es una métrica popular para la clasificación con desbalance de clases. Una limitación de estas métricas es que asumen que la distribución de clases observada en el conjunto de datos de entrenamiento coincidirá con la distribución en el conjunto de prueba y en los datos reales cuando el modelo se usa para hacer predicciones. Aunque este suele ser el caso, cuando no es así, el rendimiento puede ser bastante engañoso [112].

Por lo tanto, cómo índices de confianza, se tienen la sensibilidad y la medida F1. Estas métricas fueron obtenidas del desempeño del modelo sobre el conjunto de pruebas. Los mejores resultados obtenidos fueron del 75,7% para la sensibilidad del modelo y del 73,1% para la medida F1

4.4. Resumen

En este capítulo, se continuó con la implementación de la metodología de CRISP-DM en las fases de preparación de los datos, modelado y evaluación del modelo. La parte de preparación de los datos, que fue iniciada con la imputación de la señal IBI, se continuó con el procesamiento de las señales en sus etapas de filtrado, segmentación, etiquetado y posterior extracción de las características de las señales fisiológicas para conformar los conjuntos de datos.

El modelado de los datos se realizó en tres etapas, dónde en la primera, se realizó la validación de la calidad del conjunto de datos y se implementaron las estrategias para superar dichos problemas; en una segunda etapa, se realizó el análisis y posterior división del conjunto de datos para las fases de entrenamiento, validación y prueba del modelo. Finalmente, se realizaron los procesos de clasificación, dónde también se detallaron las métricas de evaluación necesarias para comprobar la calidad de los modelos generados.

Con la culminación de estas actividades, se da por cumplido el segundo objetivo específico. Para el tercer objetivo específico, se recurre a la teoría sobre diferentes métodos de medición y evaluación para generar el índice de confianza a partir de las métricas obtenidas en la clasificación, por lo que este objetivo también se da por cumplido.

Capítulo 5

Conclusiones y trabajo futuro

En este trabajo de investigación, se presentó la implementación de la metodología CRISP-DM aplicada a un problema de clasificación de un tipo de dato especial: las señales fisiológicas. Con esto, se obtuvieron diferentes resultados que permitieron abordar cada una de las etapas de la metodología para continuar con la siguiente, asegurando, por medio de diferentes técnicas, la calidad de los resultados obtenidos. Finalmente, se obtuvieron los resultados del modelado para la clasificación de las señales con el fin de reconocer el desempeño cognitivo de niños con TEA.

En este capítulo, se presentan las conclusiones obtenidas del desarrollo de este trabajo de grado, especialmente considerando los resultados del proceso en términos de la aplicación del conocimiento previo, la obtención de nuevo conocimiento aplicado al dominio del problema y de la solución, además del refinamiento de habilidades duras y blandas para la culminación de este proceso por parte de la estudiante aspirante al título de Ingeniera en Electrónica y Telecomunicaciones de la Universidad del Cauca.

5.1. Conclusiones

Con base en los resultados obtenidos y descritos en este trabajo investigativo y de desarrollo, se puede concluir que:

- Llevar a cabo un análisis estadístico descriptivo de las señales fisiológicas, fuente de datos de este trabajo, permitió obtener un conocimiento amplio acerca de la forma de capturar estas señales (además de comprender el proceso de obtención desde los sensores), de los aciertos y las posibles fallas a la hora de realizar la captura, de los procesamientos internos de los dispositivos captadores de estas señales y del tratamiento de los datos entregados de acuerdo al concepto práctico y real de cada señal (como con la señal de HR).

- Realizar el proceso de imputación de datos sobre las señales incompletas implicó el entendimiento del tipo de datos que fueron utilizados como insumo principal en esta tesis. Bajo esta premisa, se comprendieron las dinámicas sobre los mecanismos y patrones de los datos perdidos además de las técnicas más adecuadas para completarlos en cada caso, lo que derivó en un proceso efectivo de imputación para el caso particular de la señal IBI.

- El procesamiento de las seis señales fisiológicas implicó una alta complejidad organizativa y computacional para realizar la extracción exhaustiva de la información más relevante de cada señal. El seguimiento del proceso definido y detallado en la Figura 10 brindó las garantías para que el procesamiento fuera exitoso y al final se pudiera obtener un conjunto completo de datos y con la mayor cantidad de información posible. No obstante, este proceso generó una alta dimensionalidad en los datos que, finalmente, tuvo que ser manejada.

- Al definir un flujo de experimentos para realizar la tarea de clasificación sobre los conjuntos de datos, llevó a que la evaluación y la comparación de los resultados fuera más sencilla. A partir de esto, se notó que el conjunto de datos con la señal IBI imputada presentó un desempeño menor que su contraparte sin valores IBI, este resultado implica que hay necesidad de imputar con otros

métodos y que la imputación se debería hacer con conjuntos de datos que preseneten un porcentaje mayor al 5% de datos existentes en la variable a imputar. Igualmente, se identificó que para el clasificador que mejor resultado dio, existe un mejor comportamiento del modelo al sólo tener que utilizar 14 características obtenidas por el método de *Wrapper*.

- El hecho de utilizar IBI para la clasificación del desempeño de los niños no mejoró la evaluación de los modelos. Esto puede deberse a varias razones relacionadas con el IBI o con el proceso de imputación realizado:
    - El IBI podría en realidad no estar correlacionado con el desempeño que presentan los niños en cada juego de HapHop-Fisio, por lo que esta variable no agregaría ningún valor adicional a la clasificación.
    - El IBI puede estar muy correlacionado con otras covariables que ya están utilizando los modelos de clasificación (como el BVP y el HR), por lo que no estaría aportando información adicional. De hecho, esto podría explicar por qué el agregar IBI empeoró los resultados de los clasificadores, ya que tener covariables con una alta correlación puede confundir el modelo y reducir su capacidad para estimar correctamente la clase.
    - La señal IBI presenta un porcentaje de valores perdidos muy alto (95%), por lo que al imputarlos no es posible generar un nivel de confianza en el conjunto de datos, ya que no se cuenta con suficiente evidencia de datos para que la imputación sea capaz de generar valores cercanos a la realidad. Además, el hecho de que un porcentaje tan alto del IBI fuera generado a partir de sus covariables podría confundir el modelo de clasificación, ya que potencialmente estaría brindando la misma información.
    - El conjunto de datos está conformado por la información recolectada durante las sesiones de varios pacientes, esta característica puede haber afectado la capacidad de la técnica de imputación para estimar correctamente el IBI en cada caso.

  No obstante, al tener tan pocos datos no es posible esclarecer la verdadera razón por la cual el IBI no mejoró el índice de exactitud; para esto se requeriría una mayor cantidad de datos y pruebas adicionales.

- El entendimiento profundo de la teoría detrás de las métricas utilizadas en la evaluación de modelos de clasificación permitió generar, lo que, desde el principio, se concibió como un “índice de confianza”, una métrica que permita que la evaluación realizada durante un potencial y futuro despliegue del modelo genere mayor confianza para la ayuda en la toma de decisiones respecto a los resultados de desempeño cognitivo de los niños con TEA.

- Aunque la medida de comparación que dio lugar al problema que se abordó en este trabajo de grado fue la exactitud en la clasificación (79,80% en [113]) es importante aclarar que dicho porcentaje se logró con la validación cruzada sin tener en cuenta un conjunto de prueba. En cambio, en este trabajo se mejoró el proceso en general al tener en cuenta buenas prácticas de ciencia de datos, separando al principio el conjunto en entrenamiento y prueba, utilizando validación cruzada en el subconjunto de entrenamiento para afinar los clasificadores, y haciendo una evaluación final con los subconjuntos de prueba. Esta evaluación es más rígida que la validación cruzada, pues se simula un escenario real en el cual los datos de prueba son totalmente desconocidos hasta que el modelo se pasa a producción. Por lo tanto, el resultado de 74,20% obtenido en este trabajo puede considerarse bueno e incluso más prometedor que el de [113].

- Aunque los resultados de este trabajo no estuvieron dentro de los cánones regulares esperados (obtener una mejor clasificación), la aproximación de este trabajo aportará de manera significativa al trabajo de doctorado de la magíster Carolina Rico, desde la generación de nuevos flujos de trabajo y experimentación hasta la toma de decisiones sobre la inclusión de otro tipo de señales fisiológicas obtenidas con la pulsera E4 y otros dispositivos *wearables*.

- Teniendo en cuenta el auge que actualmente tiene la ciencia de datos en ámbitos académicos e industriales es importante resaltar las habilidades adquiridas por la estudiante durante el desarrollo del trabajo de grado para el desarrollo de proyectos de analítica de datos utilizando distintas librerías del lenguaje python, en particular en procesos de imputación y clasificación de datos. Estas habilidades serán de gran importancia para su futuro profesional.

5.2. Trabajos futuros

Desde los procesos utilizados e implementados durante este trabajo de grado, surgieron diferentes opciones e ideas de mejoramiento. Considerando los resultados obtenidos, se propone como trabajo futuro:

- Aumentar el tamaño del conjunto de datos para mejorar la precisión del modelo de clasificación obtenido. Esto debería hacerse por medio de la realización de más sesiones por paciente y aumentando el número de pacientes.

- A partir de este aumento de datos, realizar comparaciones de modelos personales y generales para el reconocimiento de patrones individuales y colectivos en las señales fisiológicas respecto al desempeño cognitivo.

- Aprovechando una mayor cantidad de datos, es necesario realizar un estudio de la influencia de la señal IBI y del proceso de imputación de la señal para comprobar y entender la verdadera razón de porque no aporta de manera positiva a los modelos de clasificación generados.

- Completar la implementación de la metodología de CRISP-DM con el despliegue final del modelo obtenido dentro del marco de uso del juego HapHop-Fisio en ambientes de rehabilitación de niños con TEA.
- Desarrollar un algoritmo que permita la generación de la señal IBI, no en tiempo real, a partir de la señal BVP que se colecciona durante la utilización de la pulsera E4.

- Probar obtener los datos del paciente estando en reposo mientras se utiliza la versión móvil de HapHop-Fisio, lo cual evitaría los artefactos de movimiento en las señales y podría mejorar sensiblemente la clasificación, o por lo menos, reduciría la cantidad de datos perdidos.

- Complementar el índice de confianza con métricas de ranking y métricas probabilísticas.

Referencias

[1] A. O. A. VIEIRA *et al.*, *Encuesta Nacional Salud Mental 2015*, vol. 117, n.º 2. 2014. doi: 10.1016/j.oooo.2013.12.297.

[2] S. H. Horowitz y M. C. Whittaker, «The State of Learning Disabilities: Understanding the 1 in 5.», *Natl. Cent. Learn. Disabil.*, p. xi, 2017, doi: 10.1016/S0264-410X(12)01439-9.

[3] M. Araz Altay y I. Görker, «Assessment of psychiatric comorbidity and WISC-R profiles in cases diagnosed with specific learning disorder according to DSM-5 criteria», *Noropsikiyatri Arsivi*, vol. 55, n.º 2, pp. 127-134, 2018, doi: 10.5152/npa.2017.18123.

[4] M. Magaña y P. Ruiz-lázaro, «Trastornos específicos del aprendizaje», *Faros*, pp. 21-28, 2015.

[5] N. Gordon, «The “ medical ” investigation of specific learning disorders», vol. 2, n.º 1, pp. 3-8, 2004.

[6] M. A. Betancourt, L. S. Dethorne, K. Karahalios, y J. G. Kim, «Skin Conductance as an *In Situ* Marker for Emotional Arousal in Children with Neurodevelopmental Communication Impairments», *ACM Trans. Access. Comput.*, vol. 9, n.º 3, pp. 1-29, 2017, doi: 10.1145/3035536.

[7] M. Rosselli, E. Matute, y A. Ardila, «Neuropsychological Assessment of Children : A test battery for children Evaluación Neuropsicológica Infantil (ENI): una batería para la evaluación de niños entre 5 y 16 años de edad . Estudio normativo colombiano», *Neurol*, vol. 38, n.º April, pp. 720-731, 2004.

[8] K. Moll, S. M. Göbel, D. Gooch, K. Landerl, y M. J. Snowling, «Cognitive Risk Factors for Specific Learning Disorder: Processing Speed, Temporal Processing, and Working Memory», *J. Learn. Disabil.*, vol. 49, n.º 3, pp. 272-281, 2014, doi: 10.1177/0022219414547221.

[9] C. Rico-Olarte, D. Lopez, S. Narváez, C. D. Farinango, y P. S. Pharow, «HapHop-Physio: a computer game to support cognitive therapies in children», *Psychol. Res. Behav. Manag.*, vol. Volume 10, pp. 209-217, jul. 2017, doi: 10.2147/PRBM.S130998.

[10] Empatica, «Empatica E4 User Manual», pp. 1-32, 2015.

[11] C. R. Olarte, «Objective Method for User Experience Evaluation of Children with SLD on HapHop- Physio», n.º September, 2018.
[12] Empatica Support, «What should I know to use the PPG/IBI data in my experiment? – Empatica Support». https://support.empatica.com/hc/en-us/articles/203621335-What-should-I-know-to-use-the-PPG-IBI-data-in-my-experiment- (accedido 17 de febrero de 2019).
[13] «Fundamentos teóricos», pp. 4-18, 2002.
[14] P. J. Mallol Roselló, «Importancia del tratamiento de datos perdidos. Aplicación en estudios longitudinales pequeños», may 2017.
[15] Q. A. W. Raaijmakers, «Effectiveness of different missing data treatments in surveys with likert-type data: Introducing the relative mean substitution approach», *Educ. Psychol. Meas.*, vol. 59, n.º 5, pp. 725-748, 1999, doi: 10.1177/0013164499595001.
[16] P. Chapman, J. Clinton, R. Kerber, y T. Khabaza, *CRISP-DM 1.0 Step-by-step data mining guide*. SPSS, 2000.
[17] Data Science Process Alliance, «CRISP-DM», *Data Science Process Alliance*, 2022. https://www.datascience-pm.com/crisp-dm-2/ (accedido 29 de marzo de 2022).
[18] IBM Documentation, «Conceptos básicos de ayuda de CRISP-DM», 17 de agosto de 2021. https://prod.ibmdocs-production-dal-6099123ce774e592a519d7c33db8265e-0000.us-south.containers.appdomain.cloud/docs/es/spss-modeler/SaaS?topic=dm-crisp-help-overview (accedido 29 de marzo de 2022).
[19] Dr. Pere Marquès Graells, «Documento utilizado con fines exclusivamente educativos por la Universidad Autónoma Metropolitana Unidad Iztapalapa, Oficina de Educación Virtual, para el Curso Gestión de Páginas Web Educativas, el cual no persigue un fin lucrativo, directo o indirecto».
[20] Ramos Gonzalez Victoria, «Las TIC en el sector de la salud», *Bit 163 JUN-JUL*, vol. 163, pp. 41-45, 2007.
[21] M. R. Navarro, *Procesos Cognitivos Y Aprendizaje Significativo*.
[22] J. T. Cacioppo, U. of Chicago, L. G. Tassinary, T. A. & M. University, G. G. Berntson, y O. S. University, *HANDBOOK OF PSYCHOPHYSIOLOGY*. 1385.
[23] A. C. Granero, «Analisis de Señales Fisiológicas. Aplicacion a la Neuroeconomia», 2013.
[24] H. F. L. LÓPEZ, «MEDICIÓN DE CARGAS COGNITIVAS DURANTE ACTIVIDADES DE INTERACCIÓN HUMANO COMPUTADOR EN AMBIENTE MÓVIL USANDO SENSORES PSICO-FISIOLÓGICOS», 2015.
[25] G. Jacucci, S. Fairclough, y E. T. Solovey, «Physiological Computing», *Computer*, vol. 48, n.º 10, pp. 12-16, 2015, doi: 10.1109/MC.2015.291.
[26] L. E. I. Éticas, «M Achine Learning in Health : Applications, Limitations and Ethical», vol. 4, n.º 4, pp. 1-3, 2019.
[27] A. Moreno *et al.*, *Aprendizaje automático*. 1994.
[28] C. S. Gómez, «Desarrollo de soluciones software mediante Aprendizaje Automático en el ámbito de la Salud»:, pp. 0-209, 2019.

[29] A. Drigas, G. Kokkalia, y M. D. Lytras, «ICT and collaborative co-learning in preschool children who face memory difficulties», *Comput. Hum. Behav.*, vol. 51, pp. 645-651, 2015, doi: 10.1016/j.chb.2015.01.019.
[30] M. L. M. Ruiz, M. Á. V. Duboy, C. T. Loriente, y I. P. De La Cruz, «Evaluating a web-based clinical decision support system for language disorders screening in a nursery school», *J. Med. Internet Res.*, vol. 16, n.º 5, 2014, doi: 10.2196/jmir.3263.
[31] J. B. Romuald Carette, Federica Cilia, Gilles Dequen y and L. V. Jean-Luc Guerin, «Automatic Autism Spectrum Disorder Detection Thanks to Eye-Tracking and Neural Network-Based Approach», *Postscapes*, vol. 1, pp. 75-81, 2018, doi: 10.1007/978-3-319-76213-5.
[32] A. S. Drigas, R. E. Ioannidou, G. Kokkalia, y M. D. Lytras, «ICTs, mobile learning and social media to enhance learning for attention difficulties», *J. Univers. Comput. Sci.*, vol. 20, n.º 10, pp. 1499-1510, 2014.
[33] E. I. Toki, J. Pange, y T. A. Mikropoulos, «An online expert system for diagnostic assessment procedures on young children's oral speech and language», *Procedia Comput. Sci.*, vol. 14, n.º Dsai, pp. 428-437, 2012, doi: 10.1016/j.procs.2012.10.049.
[34] M. Poh, S. Member, N. C. Swenson, y R. W. Picard, «A Wearable Sensor for Unobtrusive , Long-Term Assessment of Electrodermal Activity», vol. 57, n.º 5, pp. 1243-1252, 2010.
[35] S. Madhuri, J. Dorathi Jayasheeli, D. Malathi, y K. Senthilkumar, «Electrodermal activity (eda) based wearable device for qunatifying normal and abnormal emotions in humans», *ARPN J. Eng. Appl. Sci.*, vol. 12, n.º 12, pp. 3730-3735, 2017.
[36] R. Zangróniz, A. Martínez-Rodrigo, J. M. Pastor, M. T. López, y A. Fernández-Caballero, «Electrodermal activity sensor for classification of calm/distress condition», *Sens. Switz.*, vol. 17, n.º 10, pp. 1-14, 2017, doi: 10.3390/s17102324.
[37] M. N. Saadatzi, F. Tafazzoli, K. C. Welch, y J. H. Graham, «EmotiGO: Bluetooth-enabled eyewear for unobtrusive physiology-based emotion recognition», *IEEE Int. Conf. Autom. Sci. Eng.*, vol. 2016-Novem, n.º c, pp. 903-909, 2016, doi: 10.1109/COASE.2016.7743498.
[38] T. K. L. Hui y R. S. Sherratt, «Coverage of emotion recognition for common wearable biosensors», *Biosensors*, vol. 8, n.º 2, 2018, doi: 10.3390/bios8020030.
[39] V. Xia, N. Jaques, S. Taylor, S. Fedor, y R. Picard, «Active Learning for Electrodermal Activity Classification», doi: 10.1109/SPMB.2015.7405467.
[40] S. V. Wass, K. de Barbaro, y K. Clackson, «Tonic and phasic co-variation of peripheral arousal indices in infants», *Biol. Psychol.*, vol. 111, pp. 26-39, 2015, doi: 10.1016/j.biopsycho.2015.08.006.
[41] S. Taylor, N. Jaques, W. Chen, S. Fedor, A. Sano, y R. Picard, «Automatic Identification of Artifacts in Electrodermal Activity Data», *IEEE Eng. Med. Biol.*, pp. 1-14, 2017, doi: 10.1109/EMBC.2015.7318762.Automatic.
[42] K. de Barbaro, K. Clackson, y S. V. Wass, «Infant attention is dynamically modulated with changing arousal levels», *Child Dev.*, vol. 88, n.º 2, pp. 629-639, 2017, doi: 10.1111/cdev.12689.

[43] E.-H. Jang, B.-J. Park, S.-H. Kim, Y. Eum, y J. Sohn, «Identification of the optimal emotion recognition algorithm using physiological signals», *Eng. Ind. ICEI 2011 Int. Conf. On*, pp. 1-6, 2011, doi: 10.1111/1460-6984.12102.
[44] F. Canento, A. Fred, H. Silva, H. Gamboa, y A. Lourenço, «Multimodal biosignal sensor data handling for emotion recognition», *Proc. IEEE Sens.*, pp. 647-650, 2011, doi: 10.1109/ICSENS.2011.6127029.
[45] M. Callejas-Cuervo, L. A. Martínez-Tejada, y A. C. Alarcón-Aldana, «Emotion recognition techniques using physiological signals and video games –Systematic review–», *Rev. Fac. Ing.*, vol. 26, n.º 46, pp. 19-28, 2017, doi: 10.19053/01211129.v26.n46.2017.7310.
[46] E. Jang, B. Park, S. Kim, y J. Sohn, «THREE DIFFERENTIAL EMOTION CLASSIFICATION BY MACHINE LEARNING ALGORITHMS USING PHYSIOLOGICAL SIGNALS - Discriminantion of Emotions by Machine Learning Algorithms», *Proc. 4th Int. Conf. Agents Artif. Intell.*, pp. 528-531, 2012, doi: 10.5220/0003880605280531.
[47] I. Leite, R. Henriques, C. Martinho, y A. Paiva, «Sensors in the wild: Exploring electrodermal activity in child-robot interaction», *ACMIEEE Int. Conf. Hum.-Robot Interact.*, pp. 41-48, 2013, doi: 10.1109/HRI.2013.6483500.
[48] R. Henriques, A. Paiva, y C. Antunes, «On the Need of New Methods to Mine Electrodermal Activity in Emotion-Centered Studies», pp. 203-215, 2013, doi: 10.1007/978-3-642-36288-0_18.
[49] J. Hernandez, I. Riobo, A. Rozga, G. D. Abowd, y R. W. Picard, «Using electrodermal activity to recognize ease of engagement in children during social interactions», *Proc. 2014 ACM Int. Jt. Conf. Pervasive Ubiquitous Comput. - UbiComp 14 Adjun.*, pp. 307-317, 2014, doi: 10.1145/2632048.2636065.
[50] A. Youn, T. Chaspari, R. Gupta, L. I. S. Duker, S. A. Cermak, y S. S. Narayanan, «Capturing the Structure of Electrodermal Activity with Deep Neural Networks», vol. 45, n.º 9, p. 2888, 2016.
[51] «Actividad Electrodermal en niños TEA dentro de un entorno virtual inmersivo». https://www.redcenit.com/electrodermal-ninos-tea-dentro-de-un-entorno-virtual-inmersivo/ (accedido 5 de junio de 2020).
[52] «Respuesta galvánica de la piel (GSR) - Brainsigns». https://www.brainsigns.com/es/science/s2/technologies/gsr (accedido 5 de junio de 2020).
[53] «Pulsera E4 | Señales fisiológicas en tiempo real | PPG, EDA, temperatura, sensores de movimiento portátiles». https://www.empatica.com/en-eu/research/e4/ (accedido 22 de febrero de 2021).
[54] «Utilizando la señal PPG/BVP – Soporte de Empatica». https://support.empatica.com/hc/en-us/articles/204954639-Utilizing-the-PPG-BVP-signal (accedido 19 de enero de 2022).
[55] S. Samadi *et al.*, «Modeling Blood Volume Pulse Signal Using Exercise Intensity», *2019 IEEE EMBS Int. Conf. Biomed. Health Inform. BHI 2019 - Proc.*, pp. 1-4, 2019, doi: 10.1109/BHI.2019.8834662.
[56] Y. Maki, Y. Monno, K. Yoshizaki, M. Tanaka, y M. Okutomi, «Inter-Beat Interval Estimation from Facial Video Based on Reliability of BVP Signals», *Proc. Annu.*

Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS, pp. 6525-6528, 2019, doi: 10.1109/EMBC.2019.8857081.
[57] D. A. Ferrer y R. C. Laguna, «Comparativa de algoritmos para Fotopletismografía Remota e implementación de una arquitectura en tiempo real para cámaras web de bajo coste», p. 13, 2020.
[58] «Datos E4 - Señal esperada IBI – Soporte Empatica». https://support.empatica.com/hc/en-us/articles/360030058011-E4-data-IBI-expected-signal (accedido 26 de enero de 2022).
[59] «E4 data - IBI expected signal – Empatica Support». https://support.empatica.com/hc/en-us/articles/360030058011-E4-data-IBI-expected-signal (accedido 2 de marzo de 2021).
[60] P. P. Berruezo, *El contenido de la Piscomotricidad*, n.º 2000. 2000.
[61] «Los marcadores psicofisiológicos. Dando certeza al fenómeno psicológico». https://www.uaeh.edu.mx/scige/boletin/atotonilco/n8/e6.html (accedido 20 de enero de 2022).
[62] «Cómo el estrés puede provocar fiebre | Actualidad | Investigación y Ciencia». https://www.investigacionyciencia.es/noticias/cmo-el-estrs-puede-provocar-fiebre-18502 (accedido 20 de enero de 2022).
[63] M. del P. B. Lucio, C. F. Collado, y R. H. Sampieri, *Metodologia de la invetigacion*, n.º 1. 2003. doi: 10.16309/j.cnki.issn.1007-1776.2003.03.004.
[64] Junta General de la Organización Panamericana de Salud, «Epidat 4: Ayuda de Análisis descriptivo», 2014.
[65] G. Posada, *Elementos básicos de estadística descriptiva para el análisis de datos*. 2016.
[66] «Mean, Mode and Median - Measures of Central Tendency - When to use with Different Types of Variable and Skewed Distributions | Laerd Statistics». https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median.php (accedido 21 de enero de 2022).
[67] «Media - Qué es, definición y significado | Economipedia». https://economipedia.com/definiciones/media.html (accedido 14 de diciembre de 2020).
[68] «Ventajas y desventajas, – “Estadística Descriptíva”». https://estadisticassandmary.wordpress.com/ventajas-y-desventajas/ (accedido 7 de diciembre de 2020).
[69] «Mediana - Qué es, definición y concepto | Economipedia». https://economipedia.com/definiciones/mediana.html (accedido 7 de diciembre de 2020).
[70] «Rango (estadística) - Qué es, definición y concepto | Economipedia». https://economipedia.com/definiciones/rango-estadistica.html (accedido 14 de diciembre de 2020).
[71] «Ventajas y desventajas del rango La principal ventaja del rango radica en que | Course Hero». https://www.coursehero.com/file/p6vdgos/Ventajas-y-desventajas-del-rango-La-principal-ventaja-del-rango-radica-en-que/ (accedido 14 de diciembre de 2020).
[72] «FAPap - El-electrocardiograma». https://fapap.es/articulo/134/el-electrocardiograma (accedido 24 de enero de 2022).

[73] «Varianza - Qué es, definición y significado | Economipedia». https://economipedia.com/definiciones/varianza.html (accedido 14 de diciembre de 2020).
[74] «¿Qué es la desviación estándar? - Minitab». https://support.minitab.com/es-mx/minitab/18/help-and-how-to/statistics/basic-statistics/supporting-topics/data-concepts/what-is-the-standard-deviation/ (accedido 14 de diciembre de 2020).
[75] Spss, «10. Frecuencias y descriptivo», *Spss*, 2009.
[76] «Medidas de sesgo o asimetría - Probabilidad y Estadística». https://sites.google.com/site/probyestacecytechig/parcial-ii/medidas-de-forma/medidas-de-sesgo-o-asimetria (accedido 24 de enero de 2022).
[77] «Volatilidad II: Volatilidad, curtosis y asimetría - Rankia». https://www.rankia.com/blog/tomas-garcia/2226562-volatilidad-ii-curtosis-asimetria (accedido 24 de enero de 2022).
[78] W. Padilla Pardo, «Distribuciones de Probabildiad», *Ayuda Distrib. Probab.*, p. 72, 2014.
[79] «What Is Distribution Fitting ?», pp. 1-26.
[80] F. Muñoz, «Distribuciones Poisson y Gamma: Una Discreta y Continua Relación.», *Prospectiva*, vol. 12, n.º 1, p. 99, 2014, doi: 10.15665/rp.v12i1.156.
[81] F. Medina y M. Galván, *Imputación de datos: teoría y práctica*, vol. 4. 2007. doi: 978-92-1-323101-2.
[82] D. B. Rubin, «Inference and missing data», *Biometrika*, vol. 63, n.º 3, pp. 581-592, 1976, doi: 10.1093/biomet/63.3.581.
[83] M. Cañizares, I. Barroso, y K. Alfonso, «Datos incompletos: una mirada crítica para su manejo en estudios sanitarios», *Gac. Sanit.*, vol. 18, n.º 1, pp. 58-63, 2004, doi: 10.1016/s0213-9111(04)72000-2.
[84] C. Viada, C. N. Bouza, y M. Fors, «REVISION SISTEMATICA DE LOS METODOS DE IMPUTACION DE DATOS FALTANTES», n.º January, 2016.
[85] R. E. Quiroz, «UNA INTRODUCCION A LA IMPUTACION DE VALORES PERDIDOS», vol. 11, n.º 17, pp. 339-361, 2009.
[86] E. Montenegro-Montenegro, Y. Oh, y S. Chesnut, «No le tema a los datos perdidos: enfoques modernos para el manejo de datos perdidos», *Actual. En Psicol.*, vol. 29, n.º 119, p. 29, 2015, doi: 10.15517/ap.v29i119.18812.
[87] I. Schmidt y B. Mosima, «To Impute or Not Impute : That Is the Question ?», pp. 2008-2009, 2008.
[88] Á. P. Gómez, J. A. Pacheco, y M. J. López Herrero, «Comparativa de análisis de imputación de datos faltantes con análisis de casos completos en pruebas diagnósticas», 2017.
[89] T. D. Little, T. D. Jorgensen, K. M. Lang, y E. W. G. Moore, «On the joys of missing data», *J. Pediatr. Psychol.*, vol. 39, n.º 2, pp. 151-162, 2014, doi: 10.1093/jpepsy/jst048.
[90] R. T. Sataloff, M. M. Johns, y K. M. Kost, «Report of the task force on imputation».
[91] M. Shepperd, «Missing Data Imputation Techniques Qinbao Song *», vol. 2, n.º 3, 2007.
[92] F. J. Muñoz Rosas y E. Álvarez Verdejo, «Métodos de imputación para el tratamiento de datos faltantes: Aplicación mediante R/Splus», *Rev. Metodos Cuantitativos Para Econ. Empresa*, vol. 7, n.º 7, pp. 3-30, 2009.

[93] A. Puerta Goicoechea, «Imputación Basada En Árboles De Clasificación», pp. 1-76, 2002.
[94] F. M. Shrive, H. Stuart, H. Quan, y W. A. Ghali, «Dealing with missing data in a multi-question depression scale: A comparison of imputation methods», *BMC Med. Res. Methodol.*, vol. 6, pp. 1-10, 2006, doi: 10.1186/1471-2288-6-57.
[95] W. M. Campion y D. B. Rubin, «Multiple Imputation for Nonresponse in Surveys», *J. Mark. Res.*, vol. 26, n.º 4, p. 485, 1989, doi: 10.2307/3172772.
[96] C. Velasco-Gallego y I. Lazakis, «Real-time data-driven missing data imputation for short-term sensor data of marine systems. A comparative study», *Ocean Eng.*, vol. 218, n.º October, p. 108261, 2020, doi: 10.1016/j.oceaneng.2020.108261.
[97] H. Wang, X. Tan, Z. Huang, B. Pan, y J. Tian, «Mining incomplete clinical data for the early assessment of Kawasaki disease based on feature clustering and convolutional neural networks», *Artif. Intell. Med.*, vol. 105, n.º April, p. 101859, 2020, doi: 10.1016/j.artmed.2020.101859.
[98] S. I. Khan y A. S. M. L. Hoque, «SICE: an improved missing data imputation technique», *J. Big Data*, vol. 7, n.º 1, 2020, doi: 10.1186/s40537-020-00313-w.
[99] X. Zhang, C. Yan, C. Gao, B. A. Malin, y Y. Chen, «Predicting Missing Values in Medical Data Via XGBoost Regression», *J. Healthc. Inform. Res.*, vol. 4, n.º 4, pp. 383-394, 2020, doi: 10.1007/s41666-020-00077-1.
[100] P. Gouverneur, F. Li, W. M. Adamczyk, T. M. Szikszay, K. Luedtke, y M. Grzegorzek, «Comparison of Feature Extraction Methods for Physiological Signals for Heat-Based Pain Recognition», *Sensors*, vol. 21, n.º 14, Art. n.º 14, ene. 2021, doi: 10.3390/s21144838.
[101] R. Henriques, A. Paiva, y C. Antunes, «On the need of new methods to mine electrodermal activity in emotion-centered studies», en *International Workshop on Agents and Data Mining Interaction*, 2012, pp. 203-215.
[102] R. Henriques y A. Paiva, «Descriptive Models of Emotion-Learning Useful Abstractions from Physiological Responses during Affective Interactions.», en *PhyCS*, 2014, pp. 393-400.
[103] I. R. Kleckner *et al.*, «Simple, transparent, and flexible automated quality assessment procedures for ambulatory electrodermal activity data», *IEEE Trans. Biomed. Eng.*, vol. 65, n.º 7, pp. 1460-1467, 2018.
[104] J. Brownlee, «A Gentle Introduction to Imbalanced Classification», *Machine Learning Mastery*, 22 de diciembre de 2019. https://machinelearningmastery.com/what-is-imbalanced-classification/ (accedido 2 de marzo de 2022).
[105] J. Brownlee, «Introduction to Dimensionality Reduction for Machine Learning», *Machine Learning Mastery*, 5 de mayo de 2020. https://machinelearningmastery.com/dimensionality-reduction-for-machine-learning/ (accedido 3 de marzo de 2022).
[106] ZACH, «What is High Dimensional Data? (Definition & Examples)», *Statology*, 10 de febrero de 2021. https://www.statology.org/high-dimensional-data/ (accedido 3 de marzo de 2022).
[107] R. Pramoditha, «11 Dimensionality reduction techniques you should know in 2021», *Medium*, 28 de septiembre de 2021. https://towardsdatascience.com/11-

dimensionality-reduction-techniques-you-should-know-in-2021-dcb9500d388b (accedido 3 de marzo de 2022).

[108] D. Corrales, A. Ledezma, y J. Corrales, «From Theory to Practice: A Data Quality Framework for Classification Tasks», *Symmetry*, vol. 10, n.º 7, p. 248, 2018.

[109] P. Baheti, «Train, Validation, and Test Set: How to Split Your Machine Learning Data», 2022. https://www.v7labs.com/blog/train-validation-test-set, https://www.v7labs.com/blog/train-validation-test-set (accedido 3 de marzo de 2022).

[110] A. Tam, «Training-validation-test split and cross-validation done right», *Machine Learning Mastery*, 22 de septiembre de 2021. https://machinelearningmastery.com/training-validation-test-split-and-cross-validation-done-right/ (accedido 3 de marzo de 2022).

[111] P. Upretee y M. E. Yüksel, «Accurate classification of heart sounds for disease diagnosis by using spectral analysis and deep learning methods», en *Data Analytics in Biomedical Engineering and Healthcare*, K. C. Lee, S. S. Roy, P. Samui, y V. Kumar, Eds. Academic Press, 2021, pp. 215-232. doi: 10.1016/B978-0-12-819314-3.00014-8.

[112] J. Brownlee, «Tour of Evaluation Metrics for Imbalanced Classification», *Machine Learning Mastery*, 7 de enero de 2020. https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/ (accedido 3 de marzo de 2022).

[113] C. Rico-Olarte, D. M. López, L. Becker, y B. Eskofier, «Towards Classifying Cognitive Performance by Sensing Electrodermal Activity in Children With Specific Learning Disorders», *IEEE Access*, vol. 8, pp. 196187-196196, 2020, doi: 10.1109/ACCESS.2020.3033769.

Anexo A

Análisis de las señales fisiológicas

En este anexo se encuentran los análisis básicos descriptivos de las señales fuente de este trabajo. Las tablas que se presentan son una muestra de una de las sesiones de cada uno de los pacientes (por motivos de espacio, no es posible mostrar cada una de las 142 sesiones)

A.1. Tablas del análisis estadístico

Las tablas A1 - A13 muestran la información del cálculo de las medidas simples de composición y distribución de variables, tales como: medidas de tendencia central, medidas de dispersión, percentiles o medidas de posición y medidas de forma.

Tabla A1. Análisis estadístico de las señales fisiológicas – Paciente 1

	x	y	z	BVP	EDA	HR	IBI	TEMP
MEDIA	-31,279	-19,632	6,528	-0,003	0,107	89,649	0,671	32,474
MEDIANA	-31,120	-43,000	9,000	0,710	0,108	89,520	0,672	32,530
MODA	['-27,750']	['-55,000']	['14,000']	['3,180']	['0,118']	['88,780']	['0,672']	['32,530']
MINIMO	-97,750	-121,750	-89,620	-317,940	0,076	57,500	0,469	31,870
MAXIMO	74,120	76,880	93,380	199,830	0,155	101,000	0,891	32,770
RANGO	171,870	198,630	183,000	517,770	0,079	43,500	0,422	0,900
DESVIACION STD	19,880	43,638	20,915	25,697	0,011	5,048	0,069	0,187
VARIANZA	395,220	1904,257	437,437	660,317	0,000	25,481	0,005	0,035
PRIMER CUARTIL	-44,500	-55,000	-3,000	-9,070	0,099	88,000	0,625	32,390
SEGUNDO CUARTIL	-31,120	-43,000	9,000	0,710	0,108	89,520	0,672	32,530
TERCER CUARTIL	-22,880	21,620	18,250	9,998	0,117	92,430	0,719	32,610
ASIMETRIA	0,772	0,626	-0,681	-0,904	-0,005	-0,976	0,166	-0,951
CURTOSIS	1,583	-1,164	1,344	13,853	-0,407	3,003	0,542	0,313
NOMBRE DISTRIBUCION DE PROBABILIDAD							gamma	

Tabla A2. Análisis estadístico de señales fisiológicas – Paciente 2

	x	y	z	BVP	EDA	HR	IBI	TEMP
MEDIA	26,945	-43,063	8,144	-0,012	2,282	99,214	0,583	34,279
MEDIANA	22,500	-58,380	1,750	0,915	2,419	101,580	0,563	34,290
MODA	['21,000']	['-59,000']	['2,000']	['-7,240', '-5,230']	['2,441']	['108,200']	['0,547']	['34,290']
MINIMO	-26,750	-128,000	-66,000	-336,670	0,827	48,000	0,438	33,590
MAXIMO	94,250	62,500	65,500	276,710	3,655	110,430	1,656	34,610
RANGO	121,000	190,500	131,500	613,380	2,829	62,430	1,219	1,020
DESVIACION STD	13,747	31,615	20,706	37,993	0,599	9,060	0,107	0,185
VARIANZA	188,974	999,485	428,729	1443,502	0,358	82,081	0,011	0,034
PRIMER CUARTIL	18,750	-60,620	-4,000	-16,325	1,964	95,730	0,531	34,230
SEGUNDO CUARTIL	22,500	-58,380	1,750	0,915	2,419	101,580	0,563	34,290
TERCER CUARTIL	35,500	-21,315	15,940	17,768	2,693	105,930	0,609	34,410
ASIMETRIA	0,849	0,963	0,529	-0,338	-0,535	-1,151	5,531	-1,003
CURTOSIS	0,639	-0,146	0,357	8,998	-0,576	0,972	42,073	1,247
NOMBRE DISTRIBUCION DE PROBABILIDAD							burr	

Tabla A3. Análisis estadístico de señales fisiológicas – Paciente 3

	x	y	z	BVP	EDA	HR	IBI	TEMP
MEDIA	39,681	-35,303	9,040	0,002	0,280	84,054	0,718	34,254
MEDIANA	47,380	-35,380	8,000	0,215	0,154	84,020	0,719	34,230
MODA	['51,250']	['-33,000']	['3,380', '5,000']	['0,120', '0,520']	['0,136']	['81,020']	['0,719']	['34,130']
MINIMO	-47,620	-117,380	-63,750	-344,500	0,062	61,000	0,500	33,930
MAXIMO	91,620	69,500	75,620	249,850	2,292	97,650	0,922	34,710
RANGO	139,240	186,880	139,370	594,350	2,231	36,650	0,422	0,780
DESVIACION STD	17,649	23,185	16,163	12,751	0,355	6,289	0,081	0,171
VARIANZA	311,498	537,564	261,249	162,594	0,126	39,556	0,007	0,029
PRIMER CUARTIL	27,590	-45,750	1,500	-3,450	0,135	79,330	0,656	34,110
SEGUNDO CUARTIL	47,380	-35,380	8,000	0,215	0,154	84,020	0,719	34,230
TERCER CUARTIL	51,880	-30,000	17,120	3,552	0,183	88,490	0,770	34,390
ASIMETRIA	-1,086	1,733	-0,322	-2,953	2,832	0,078	0,229	0,439
CURTOSIS	0,542	5,241	2,660	152,121	7,339	-0,645	-0,194	-0,583
NOMBRE DISTRIBUCION DE PROBABILIDAD							invgauss	

Tabla A4. Análisis estadístico de señales fisiológicas – Paciente 4

	x	y	z	BVP	EDA	HR	IBI	TEMP
MEDIA	-28,688	17,877	4,995	0,001	2,193	77,228	0,824	29,081
MEDIANA	-35,000	49,620	6,120	-0,675	2,115	76,030	0,828	29,070
MODA	['-37,000']	['-62,000']	['6,000']	['-17,680', '-13,660']	['2,134']	['69,900', '72,200']	['0,859']	['28,770']
MINIMO	-92,500	-72,750	-56,000	-317,240	0,832	60,380	0,563	28,630
MAXIMO	15,250	99,750	67,250	439,490	3,984	97,700	1,109	29,590
RANGO	107,750	172,500	123,250	756,730	3,153	37,320	0,547	0,960
DESVIACION STD	17,921	50,100	11,608	38,536	0,578	6,921	0,071	0,288
VARIANZA	321,172	2509,961	134,739	1485,040	0,334	47,898	0,005	0,083
PRIMER CUARTIL	-39,120	-54,030	1,500	-13,713	1,785	72,300	0,781	28,790
SEGUNDO CUARTIL	-35,000	49,620	6,120	-0,675	2,115	76,030	0,828	29,070
TERCER CUARTIL	-16,470	52,750	10,000	12,725	2,594	81,670	0,875	29,350
ASIMETRIA	0,486	-0,817	-1,093	1,208	0,434	0,674	-0,063	0,206
CURTOSIS	-0,441	-1,189	5,026	16,760	-0,377	0,425	0,381	-1,477
NOMBRE DISTRIBUCION DE PROBABILIDAD							weibull	

Tabla A5. Análisis estadístico de señales fisiológicas – Paciente 5

	x	y	z	BVP	EDA	HR	IBI	TEMP
MEDIA	29,497	-35,447	11,689	-0,022	3,711	92,697		29,386
MEDIANA	27,880	-50,250	11,620	-1,215	3,926	91,165		29,190
MODA	['16,500']	['-60,880']	['7,620']	['-17,900', '-9,270']	['4,174', '4,324']	['108,870']		['29,130']
MINIMO	-86,750	-128,000	-90,000	-317,510	0,838	71,680		28,730
MAXIMO	118,500	95,120	118,120	355,440	7,368	118,280		30,150
RANGO	205,250	223,120	208,120	672,950	6,530	46,600		1,420
DESVIACION STD	21,463	40,811	22,128	35,711	1,299	11,112		0,401
VARIANZA	460,675	1665,558	489,642	1275,296	1,688	123,481		0,161
PRIMER CUARTIL	16,620	-61,000	0,120	-16,585	2,641	82,850		29,110
SEGUNDO CUARTIL	27,880	-50,250	11,620	-1,215	3,926	91,165		29,190
TERCER CUARTIL	43,380	-11,250	25,500	13,973	4,666	102,100		29,730
ASIMETRIA	0,037	0,802	-0,311	0,402	-0,253	0,238		0,570
CURTOSIS	1,557	-0,034	1,299	6,512	-0,789	-1,070		-1,022
NOMBRE DISTRIBUCION DE PROBABILIDAD								

Tabla A6. Análisis estadístico de señales fisiológicas – Paciente 6

	x	y	z	BVP	EDA	HR	IBI	TEMP
MEDIA	18,247	-21,399	17,121	0,004	12,761	100,550	0,629	31,326
MEDIANA	20,120	-40,120	16,750	0,400	11,317	94,220	0,625	31,350
MODA	['19,880']	['-60,620']	['14,250']	['6,110']	['2,182', '2,225']	['90,670']	['0,625']	['31,530']
MINIMO	-73,500	-128,000	-83,500	-317,240	0,827	55,670	0,375	30,670
MAXIMO	125,380	79,500	115,620	262,330	23,849	129,880	0,859	31,810
RANGO	198,880	207,500	199,120	579,570	23,022	74,210	0,484	1,140
DESVIACION STD	26,872	47,875	21,748	44,168	7,307	15,359	0,067	0,300
VARIANZA	722,124	2292,013	472,967	1950,840	53,399	235,886	0,004	0,090
PRIMER CUARTIL	4,380	-59,380	5,380	-25,310	5,668	88,900	0,594	31,130
SEGUNDO CUARTIL	20,120	-40,120	16,750	0,400	11,317	94,220	0,625	31,350
TERCER CUARTIL	35,120	30,880	31,250	25,570	20,282	116,500	0,672	31,590
ASIMETRIA	-0,314	0,508	-0,385	-0,043	0,045	0,368	-0,177	-0,397
CURTOSIS	0,360	-1,156	0,972	2,324	-1,611	-1,159	1,185	-0,818
NOMBRE DISTRIBUCION DE PROBABILIDAD							gamma	

Tabla A7. Análisis estadístico de señales fisiológicas – Paciente 7

	x	y	z	BVP	EDA	HR	IBI	TEMP
MEDIA	35,961	-14,037	30,448	-0,000	0,103	79,004	0,802	32,757
MEDIANA	37,000	-0,250	33,620	0,205	0,109	76,470	0,797	32,610
MODA	['54,000']	['1,000']	['52,000']	['2,240']	['0,132']	['75,670']	['0,797']	['32,590']
MINIMO	-37,380	-124,380	-86,620	-345,570	0,060	66,000	0,563	31,650
MAXIMO	81,880	66,880	67,120	181,710	0,143	108,270	1,219	33,630
RANGO	119,260	191,260	153,740	527,280	0,083	42,270	0,656	1,980
DESVIACION STD	15,801	29,938	20,789	10,035	0,024	8,050	0,092	0,531
VARIANZA	249,686	896,287	432,174	100,699	0,001	64,795	0,008	0,282
PRIMER CUARTIL	24,250	-40,880	11,000	-3,670	0,081	74,193	0,734	32,390
SEGUNDO CUARTIL	37,000	-0,250	33,620	0,205	0,109	76,470	0,797	32,610
TERCER CUARTIL	49,250	1,750	50,000	3,640	0,126	81,170	0,859	33,290
ASIMETRIA	-0,863	-0,488	-0,605	-4,703	-0,231	1,816	0,217	0,020
CURTOSIS	1,157	-0,160	-0,120	248,181	-1,352	3,039	0,123	-1,136
NOMBRE DISTRIBUCION DE PROBABILIDAD							weibull	

Tabla A8. Análisis estadístico de señales fisiológicas – Paciente 8

	x	y	z	BVP	EDA	HR	IBI	TEMP
MEDIA	24,828	-39,502	-7,881	0,001	4,890	93,613	0,608	32,878
MEDIANA	25,620	-50,500	-5,380	1,760	5,544	95,380	0,609	32,890
MODA	['17,500']	['-60,620']	['1,000']	['1,760']	['0,268']	['94,620']	['0,594']	['32,930']
MINIMO	-69,250	-126,380	-74,880	-332,070	0,195	55,170	0,406	32,310
MAXIMO	98,250	66,500	72,250	384,250	9,874	114,130	1,078	33,180
RANGO	167,500	192,880	147,130	716,320	9,679	58,960	0,672	0,870
DESVIACION STD	19,583	33,986	18,088	44,390	2,666	10,378	0,075	0,162
VARIANZA	383,504	1155,043	327,168	1970,488	7,107	107,694	0,006	0,026
PRIMER CUARTIL	16,500	-59,750	-17,000	-21,960	3,073	87,602	0,563	32,810
SEGUNDO CUARTIL	25,620	-50,500	-5,380	1,760	5,544	95,380	0,609	32,890
TERCER CUARTIL	38,250	-39,380	3,250	23,180	6,665	100,450	0,656	32,950
ASIMETRIA	-0,779	1,832	-0,330	-0,505	-0,315	-0,836	1,329	-0,574
CURTOSIS	1,188	2,435	0,795	5,853	-0,895	0,494	6,491	0,731
NOMBRE DISTRIBUCION DE PROBABILIDAD							gamma	

Tabla A9. Análisis estadístico de señales fisiológicas – Paciente 9

	x	y	z	BVP	EDA	HR	IBI	TEMP
MEDIA	14,059	-56,919	-3,388	-0,006	13,156	101,827	0,526	33,585
MEDIANA	14,620	-61,250	-2,380	0,965	13,744	110,375	0,516	33,570
MODA	['12,880']	['-61,380']	['-1,750']	['-7,030', '-0,930']	['14,424']	['111,050', '112,500']	['0,500']	['33,870']
MINIMO	-59,750	-118,380	-77,750	-317,240	0,825	60,000	0,391	32,930
MAXIMO	73,120	70,880	74,750	278,780	17,957	125,400	0,688	33,930
RANGO	132,870	189,260	152,500	596,020	17,132	65,400	0,297	1,000
DESVIACION STD	13,563	20,333	14,275	45,743	2,203	16,193	0,058	0,240
VARIANZA	183,948	413,416	203,783	2092,423	4,851	262,221	0,003	0,058
PRIMER CUARTIL	9,250	-63,620	-8,000	-22,170	12,115	87,337	0,484	33,430
SEGUNDO CUARTIL	14,620	-61,250	-2,380	0,965	13,744	110,375	0,516	33,570
TERCER CUARTIL	19,590	-58,380	3,880	23,045	14,670	113,750	0,563	33,810
ASIMETRIA	-0,221	3,360	-0,772	-0,313	-1,197	-0,532	0,304	-0,600
CURTOSIS	3,305	13,060	2,797	5,385	1,611	-1,118	0,015	-0,270
NOMBRE DISTRIBUCION DE PROBABILIDAD							gamma	

Tabla A10. Análisis estadístico de señales fisiológicas - Paciente 10

	x	y	z	BVP	EDA	HR	IBI	TEMP
MEDIA	-33,785	-27,167	2,922	0,013	1,479	106,826	0,595	33,052
MEDIANA	-34,250	-35,880	9,500	1,390	0,961	106,350	0,594	33,370
MODA	['-23,120']	['-58,620']	['12,120', '16,000']	['-7,660']	['0,356']	['105,450']	['0,594']	['33,390']
MINIMO	-115,380	-126,750	-122,880	-442,230	0,311	80,380	0,391	31,010
MAXIMO	75,880	81,880	87,380	297,700	4,445	130,970	1,141	33,550
RANGO	191,260	208,630	210,260	739,930	4,134	50,590	0,750	2,540
DESVIACION STD	19,315	34,587	29,305	63,362	0,999	11,592	0,083	0,623
VARIANZA	373,081	1196,256	858,792	4014,801	0,998	134,371	0,007	0,388
PRIMER CUARTIL	-47,620	-54,848	-16,500	-27,130	0,692	100,770	0,547	32,970
SEGUNDO CUARTIL	-34,250	-35,880	9,500	1,390	0,961	106,350	0,594	33,370
TERCER CUARTIL	-21,880	-3,880	22,750	30,660	2,421	117,080	0,625	33,430
ASIMETRIA	0,700	0,786	-0,584	-0,691	0,699	-0,346	1,265	-1,715
CURTOSIS	2,478	-0,052	-0,129	5,110	-0,963	-0,371	5,231	1,872
NOMBRE DISTRIBUCION DE PROBABILIDAD							weibull	

Tabla A11. Análisis estadístico de señales fisiológicas - Paciente 11

	x	y	z	BVP	EDA	HR	IBI	TEMP
MEDIA	-25.690	24.991	12.277	-0.006	10.794	100.596	0.568	33.560
MEDIANA	-25.250	35.620	9.880	0.540	11.291	103.040	0.563	33.570
MODA	['-20.000']	['63.000']	['9.120']	['-36.010', '-24.000']	['12.057']	['106.080']	['0.547']	['33.660']
MINIMO	-125.000	-72.120	-107.750	-567.340	0.834	62.170	0.406	33.310
MAXIMO	79.750	125.120	111.000	444.040	14.943	134.430	0.813	33.790
RANGO	204.750	197.240	218.750	1.011.380	14.109	72.260	0.406	0.480
DESVIACION STD	25.474	41.786	24.526	70.592	2.104	13.604	0.058	0.113
VARIANZA	648.922	1746.031	601.528	4983.273	4.426	185.070	0.003	0.013
PRIMER CUARTIL	-44.380	-9.000	-3.250	-36.770	9.626	89.810	0.531	33.470
SEGUNDO CUARTIL	-25.250	35.620	9.880	0.540	11.291	103.040	0.563	33.570
TERCER CUARTIL	-11.380	61.120	30.750	36.370	12.385	109.290	0.594	33.650
ASIMETRIA	0.382	-0.352	-0.107	-0.101	-0.910	-0.196	0.510	-0.112
CURTOSIS	0.281	-0.990	0.100	3.137	0.716	-0.285	1.482	-0.876
NOMBRE DISTRIBUCION DE PROBABILIDAD							weibull	

Tabla A12. Análisis estadístico de señales fisiológicas - Paciente 12

	x	y	z	BVP	EDA	HR	IBI	TEMP
MEDIA	20.291	-30.094	20.171	0.020	0.125	90.070	0.663	27.993
MEDIANA	11.000	-53.880	16.380	0.720	0.128	89.820	0.656	27.990
MODA	['7.000']	['-61.000']	['14.000']	['-8.100', '-3.180']	['0.135']	['87.250', '88.380']	['0.688']	['27.970']
MINIMO	-68.750	-103.000	-46.750	-317.240	0.051	75.250	0.469	27.270
MAXIMO	90.250	68.620	82.380	298.260	0.628	116.000	0.953	28.510
RANGO	159.000	171.620	129.130	615.500	0.577	40.750	0.484	1.240
DESVIACION STD	23.095	39.090	15.180	31.829	0.015	4.904	0.072	0.210
VARIANZA	533.391	1527.995	230.445	1013.075	0.000	24.050	0.005	0.044
PRIMER CUARTIL	6.120	-61.120	12.500	-11.190	0.117	87.170	0.609	27.930
SEGUNDO CUARTIL	11.000	-53.880	16.380	0.720	0.128	89.820	0.656	27.990
TERCER CUARTIL	38.120	-6.120	27.500	11.355	0.133	92.780	0.703	28.130
ASIMETRIA	-0.044	0.925	0.619	-0.427	8.528	0.310	0.429	-0.757
CURTOSIS	0.678	-0.331	1.106	12.891	272.333	0.947	0.759	1.602
NOMBRE DISTRIBUCION DE PROBABILIDAD							weibull	

A.2. Valores máximos y mínimos perdidos de IBI

En las tablas A14 – A25, los valores máximos y mínimos de HR por minuto se comparan con la cantidad de datos IBI equivalentes al mismo rango de tiempo para obtener un porcentaje aproximado de los datos presentes y faltantes de IBI.

Tabla A14. Organización y porcentaje de datos IBI perdidos – Paciente 1

Minuto	MaxHR	MinHR	DatosIBI	AusentesIBIMaximoHR	AusentesIBIMinimoHR
1	101.0	57.5	26	75	31
2	89.48	85.35	28	61	57
3	89.57	87.28	23	66	64
4	93.13	89.05	5	88	84
5	94.17	91.92	16	78	75
6	92.42	88.3	9	83	79
7	88.9	88.08	14	74	74
8	89.88	87.98	4	85	83
9	90.4	88.58	27	63	61
10	90.4	87.35	30	60	57
11	90.5	87.37	9	81	78
12	97.08	88.65	7	90	81
13	98.45	90.8	12	86	78
14	93.47	90.08	9	84	81
15	93.48	86.15	22	71	64
16	89.42	86.15	12	77	74
17	90.42	88.0	20	70	68
18	89.8	82.83	23	66	59
19	92.17	81.67	7	85	74
20	93.02	87.13	25	68	62
21	97.08	89.73	14	83	75
22	91.27	80.42	12	79	68
23	80.17	73.4	9	71	64
24	94.47	74.92	10	84	64
25	97.93	94.78	16	81	78
26	97.08	91.78	4	93	87
27	100.43	93.0	0	100	93
Total	2495	2306	393	2102	1913
Porcentaje (%)	100	100	['15.75', '17.04']	84.25	82.96

Tabla A15. Organización y porcentaje de datos IBI perdidos – Paciente 2

Minuto	MaxHR	MinHR	DatosIBI	AusentesIBIMaximoHR	AusentesIBIMinimoHR
1	90.75	48.0	9	81	39
2	88.5	83.57	3	85	80
3	96.75	83.6	26	70	57
4	83.13	77.18	21	62	56
5	99.23	81.43	28	71	53
6	107.57	99.77	43	64	56
7	107.42	104.13	37	70	67
8	108.82	107.57	14	94	93
9	110.43	103.65	55	55	48
10	103.38	99.62	29	74	70
11	108.98	100.27	9	99	91
12	110.32	98.55	20	90	78
13	98.32	88.85	11	87	77
14	107.75	97.35	12	95	85
15	104.63	97.8	40	64	57
16	106.1	99.73	65	41	34
17	105.95	102.6	7	98	95
18	109.23	105.72	32	77	73
19	104.7	96.07	38	66	58
20	102.83	98.85	12	90	86
21	104.98	98.32	14	90	84
Total	2148	1962	525	1623	1437
Porcentaje (%)	100	100	['24.44', '26.76']	75.56	73.24

Tabla A16. Organización y porcentaje de datos IBI perdidos – Paciente 3

Minuto	MaxHR	MinHR	DatosIBI	AusentesIBIMaximoHR	AusentesIBIMinimoHR
1	88.72	61.0	1	87	60
2	91.3	80.3	4	87	76
3	79.73	73.55	9	70	64
4	93.1	80.02	3	90	77
5	92.03	84.08	23	69	61
6	84.57	74.4	19	65	55
7	75.5	72.47	9	66	63
8	93.13	75.57	0	93	75
9	97.65	90.55	3	94	87
10	96.42	86.33	5	91	81
11	96.4	78.28	12	84	66

12	85.72	78.52	6	79	72
13	84.28	78.9	12	72	66
14	86.85	79.83	0	86	79
15	86.2	78.85	2	84	76
16	84.8	72.87	0	84	72
17	88.4	73.38	2	86	71
18	89.5	82.03	0	89	82
19	91.65	82.68	0	91	82
20	96.5	88.85	4	92	84
21	88.33	75.4	7	81	68
Total	1861	1638	121	1740	1517
Porcentaje (%)	100	100	['6.50', '7.39']	93.50	92.61

Tabla A17. Organización y porcentaje de datos IBI perdidos – Paciente 4

Minuto	MaxHR	MinHR	DatosIBI	AusentesIBIMaximoHR	AusentesIBIMinimoHR
1	88.4	60.38	27	61	33
2	72.72	65.45	5	67	60
3	88.6	72.62	2	86	70
4	97.7	88.98	9	88	79
5	95.27	71.38	25	70	46
6	71.57	68.92	6	65	62
7	75.52	71.8	9	66	62
8	75.57	73.45	27	48	46
9	88.62	74.97	20	68	54
10	87.15	77.83	19	68	58
11	78.4	75.9	2	76	73
12	80.03	78.17	36	44	42
13	78.83	73.17	5	73	68
14	86.03	76.18	33	53	43
15	85.48	72.8	28	57	44
16	72.75	69.38	18	54	51
17	73.07	67.92	38	35	29
18	74.85	72.6	61	13	11
19	74.08	71.17	60	14	11
20	72.67	70.33	64	8	6
21	76.17	70.27	18	58	52
22	78.92	76.45	9	69	67
23	83.08	77.98	0	83	77
24	88.17	82.63	0	88	82

Total	1933	1747	521	1412	1226
Porcentaje (%)	100	100	['26.95', '29.82']	73.05	70.18

Tabla A18. Organización y porcentaje de datos IBI perdidos – Paciente 5

Minuto	MaxHR	MinHR	DatosIBI	AusentesIBIMaximoHR	AusentesIBIMinimoHR
1	93.0	55.67	0	93	55
2	93.45	81.02	0	93	81
3	87.7	78.08	14	73	64
4	90.1	87.63	18	72	69
5	91.65	87.35	39	52	48
6	90.4	86.58	16	74	70
7	105.48	89.58	0	105	89
8	105.9	90.63	20	85	70
9	93.95	90.65	46	47	44
10	91.93	79.75	3	88	76
11	108.63	79.97	3	105	76
12	120.1	109.08	0	120	109
13	124.58	115.35	6	118	109
14	128.63	119.48	0	128	119
15	119.2	106.93	55	64	51
16	106.58	94.32	55	51	39
17	94.9	92.85	0	94	92
18	121.83	95.1	0	121	95
19	129.88	122.2	0	129	122
20	125.58	114.53	0	125	114
21	126.45	112.83	0	126	112
22	126.08	90.25	0	126	90
23	89.58	81.47	0	89	81
Total	2453	2150	275	2178	1875
Porcentaje (%)	100	100	['11.21', '12.79']	88.79	87.21

Tabla A19. Organización y porcentaje de datos IBI perdidos – Paciente 6

Minuto	MaximoHR	MinimoHR	DatosIBI	AusentesIBIMaximoHR	AusentesIBIMinimoHR
1	76.0	66.0	13	63	53
2	94.28	74.73	7	87	67
3	108.27	95.17	0	108	95
4	100.25	84.68	13	87	71
5	100.13	88.78	11	89	77

6	99.37	87.63	7	92	80
7	87.58	78.55	51	36	27
8	79.55	74.77	23	56	51
9	77.4	73.48	16	61	57
10	86.33	73.15	8	78	65
11	85.87	74.92	20	65	54
12	76.95	71.68	66	10	5
13	72.62	71.0	49	23	22
14	72.13	71.53	34	38	37
15	73.53	70.02	38	35	32
16	77.1	70.87	24	53	46
17	76.95	70.8	23	53	47
18	77.97	71.2	14	63	57
19	82.18	75.27	29	53	46
20	81.48	74.1	26	55	48
21	76.87	73.03	45	31	28
22	77.68	74.48	40	37	34
23	75.73	73.03	61	14	12
24	77.23	73.1	44	33	29
25	76.72	72.27	15	61	57
26	81.95	74.67	7	74	67
27	83.73	77.35	34	49	43
28	77.77	76.53	31	46	45
Total	2299	2101	749	1550	1352
Porcentaje (%)	100	100	['32.58', '35.65']	67.42	64.35

Tabla A20. Organización y porcentaje de datos IBI perdidos – Paciente 7

Minuto	MaxHR	MinHR	DatosIBI	AusentesIBIMaximoHR	AusentesIBIMinimoHR
1	74.88	55.17	27	47	28
2	94.52	75.29	28	66	47
3	94.78	89.7	30	64	59
4	100.0	95.08	32	68	63
5	95.6	93.55	15	80	78
6	96.38	91.75	13	83	78
7	96.22	79.42	11	85	68
8	94.02	79.5	3	91	76
9	104.67	94.37	6	98	88
10	104.92	101.57	2	102	99
11	101.38	98.33	16	85	82

12	108.7	101.12	11	97	90
13	106.5	93.42	27	79	66
14	96.35	81.28	5	91	76
15	80.58	71.85	36	44	35
16	103.57	79.18	43	60	36
17	114.13	104.05	36	78	68
18	106.45	93.62	43	63	50
19	99.6	93.42	0	99	93
20	106.67	100.0	27	79	73
21	105.58	98.18	6	99	92
22	100.7	96.37	69	31	27
23	99.83	97.07	24	75	73
24	102.18	94.98	9	93	85
25	94.62	77.62	0	94	77
26	83.68	76.17	0	83	76
Total	2553	2302	519	2034	1783
Porcentaje (%)	100	100	['20.33', '22.55']	79.67	77.45

Tabla A21. Organización y porcentaje de datos IBI perdidos – Paciente 8

Minuto	MaxHR	MinHR	DatosIBI	AusentesIBIMaximoHR	AusentesIBIMinimoHR
1	76.67	60.0	12	64	48
2	83.03	72.51	0	83	72
3	89.4	83.23	0	89	83
4	89.22	85.68	3	86	82
5	85.73	74.1	12	73	62
6	75.0	72.13	1	74	71
7	87.47	75.37	0	87	75
8	97.03	87.63	25	72	62
9	98.83	92.58	7	91	85
10	99.25	92.37	6	93	86
11	125.4	94.63	10	115	84
12	124.83	110.63	31	93	79
13	124.28	110.27	9	115	101
14	125.08	110.18	17	108	93
15	111.9	109.68	5	106	104
16	114.67	111.78	6	108	105
17	114.67	111.27	7	107	104
18	119.75	114.88	0	119	114
19	121.63	114.85	20	101	94

20	114.6	110.07	0	114	110
21	113.75	112.03	11	102	101
22	112.85	107.4	0	112	107
23	114.32	106.4	0	114	106
Total	2408	2210	182	2226	2028
Porcentaje (%)	100	100	['7.56', '8.24']	92.44	91.76

Tabla A22. Organización y porcentaje de datos IBI perdidos – Paciente 9

Minuto	MaxHR	MinHR	DatosIBI	AusentesIBIMaximoHR	AusentesIBIMinimoHR
1	122.0	95.4	22	100	73
2	103.73	100.12	23	80	77
3	103.43	93.12	33	70	60
4	93.53	90.12	3	90	87
5	90.0	82.33	7	83	75
6	83.82	80.38	20	63	60
7	89.8	81.55	15	74	66
8	108.62	90.25	38	70	52
9	108.58	101.78	9	99	92
10	102.78	101.65	31	71	70
11	101.52	99.97	18	83	81
12	106.0	101.27	13	93	88
13	104.58	98.4	15	89	83
14	105.12	100.33	3	102	97
15	107.9	105.1	15	92	90
16	116.43	107.57	17	99	90
17	119.27	116.63	11	108	105
18	118.18	115.6	5	113	110
19	122.32	115.18	15	107	100
20	115.02	110.47	5	110	105
21	118.92	113.42	16	102	97
22	127.65	118.87	8	119	110
23	130.97	119.97	7	123	112
24	119.57	106.35	16	103	90
25	107.67	104.13	1	106	103
26	115.2	105.13	0	115	105
27	122.83	115.37	1	121	114
28	122.87	116.58	0	122	116
Total	3074	2875	367	2707	2508
Porcentaje (%)	100	100	['11.94', '12.77']	88.06	87.23

Tabla A23. Organización y porcentaje de datos IBI perdidos – Paciente 10

Minuto	MaxHR	MinHR	DatosIBI	AusentesIBIMaximoHR	AusentesIBIMinimoHR
1	83.0	62.17	0	83	62
2	97.02	72.37	55	42	17
3	108.87	97.35	32	76	65
4	113.07	105.82	52	61	53
5	105.47	101.85	31	74	70
6	105.47	101.52	34	71	67
7	116.82	102.12	9	107	93
8	120.43	114.68	2	118	112
9	114.32	106.67	30	84	76
10	107.28	103.55	13	94	90
11	134.43	107.77	0	134	107
12	134.25	109.62	1	133	108
13	109.38	104.22	1	108	103
14	108.63	105.25	1	107	104
15	112.88	105.28	2	110	103
16	117.48	112.25	3	114	109
17	112.02	79.77	1	111	78
18	101.33	77.7	9	92	68
19	105.78	86.53	0	105	86
20	86.73	80.88	0	86	80
21	117.23	86.77	2	115	84
22	122.27	114.62	7	115	107
23	114.5	103.67	6	108	97
24	103.33	89.67	7	96	82
25	90.28	82.35	0	90	82
26	86.48	81.8	0	86	81
27	98.93	86.83	6	92	80
28	96.58	75.13	0	96	75
29	97.4	77.12	0	97	77
30	103.7	97.48	0	103	97
31	101.28	95.42	0	101	95
Total	3313	2912	304	3009	2608
Porcentaje (%)	100	100	['9.18', '10.44']	90.82	89.56

Tabla A24. Organización y porcentaje de datos IBI perdidos – Paciente 11

Minuto	MaxHR	MinHR	DatosIBI	AusentesIBIMaximoHR	AusentesIBIMinimoHR
1	116.0	75.25	23	93	52
2	93.65	90.85	22	71	68
3	90.8	83.58	24	66	59
4	97.5	83.75	27	70	56
5	97.7	92.77	40	57	52
6	94.28	92.23	65	29	27
7	92.12	88.32	48	44	40
8	90.63	85.73	34	56	51
9	91.3	85.33	33	58	52
10	90.88	86.67	34	56	52
11	91.22	87.32	80	11	7
12	103.92	91.48	53	50	38
13	101.02	88.28	29	72	59
14	94.17	88.27	46	48	42
15	94.22	86.75	26	68	60
16	92.72	85.32	25	67	60
17	92.9	82.67	42	50	40
18	99.45	93.25	42	57	51
19	97.52	90.87	41	56	49
20	93.23	89.2	35	58	54
21	89.05	85.5	51	38	34
22	88.02	84.27	4	84	80
23	83.92	78.0	0	83	78
24	90.92	82.5	0	90	82
Total	2256	2067	824	1432	1243
Porcentaje (%)	100	100	['36.52', '39.86']	63.48	60.14

Tabla A25. Organización y porcentaje de datos IBI perdidos – Paciente 12

Minuto	MaxHR	MinHR	DatosIBI	AusentesIBIMaximoHR	AusentesIBIMinimoHR
1	69.88	59.0	4	65	55
2	95.88	70.24	20	75	50
3	109.33	96.25	21	88	75
4	116.77	109.47	16	100	93
5	112.18	97.57	66	46	31
6	97.28	91.17	43	54	48
7	94.52	89.92	72	22	17
8	89.85	88.42	65	24	23

CURTOSIS	0.519	10.147	10.147	10.147	10.147	10.147	1.271	2.180	2.868	3.547	4.101	4.735
9	94.38		89.47		42		52			47		
10	94.05		91.47		23		71			68		
11	97.4		91.85		4		93			87		
12	109.28		90.93		4		105			86		
13	112.65		109.37		16		96			93		
14	109.62		108.35		6		103			102		
15	111.88		107.03		3		108			104		
16	112.7		106.6		1		111			105		
17	119.88		112.1		18		101			94		
18	122.55		118.63		5		117			113		
19	118.23		90.02		4		114			86		
20	89.58		77.82		7		82			70		
21	111.78		85.87		11		100			74		
22	114.35		111.53		0		114			111		
23	112.3		103.98		0		112			103		
24	114.98		103.95		5		109			98		
25	115.28		112.72		2		113			110		
26	112.93		95.95		0		112			95		
27	95.45		80.25		0		95			80		
28	84.37		77.88		0		84			77		
29	88.83		84.73		0		88			84		
Total	3012		2737		458		2554			2279		
Porcentaje (%)	100		100		['15.21', '16.73']		84.79			83.27		

A.3. Imputación de datos simple y múltiple

En las tablas A27 – A40, se presenta el analisis estadístico de la señal IBI sin imputar comparada con el análisis de la misma señal imputada con el método MICE con 1, 5, 10, 20, y 30 iteraciones y con la técnica KNN con 3, 4, 5, 6, 7 y 8 vecinos más cercanos.

Tabla A27. Comparación análisis estadístico con las técnicas de imputación utilizadas – Paciente 1

	IBI	MICE1	MICE5	MICE10	MICE20	MICE30	KNN3	KNN4	KNN5	KNN6	KNN7	KNN8
MEDIA	0.672	0.670	0.670	0.670	0.670	0.670	0.674	0.674	0.674	0.674	0.673	0.673
MEDIANA	0.670	0.670	0.670	0.670	0.670	0.670	0.670	0.670	0.672	0.672	0.671	0.671
MODA	['0.670']	['0.670']	['0.670']	['0.670']	['0.670']	['0.670']	['0.660']	['0.660']	['0.670']	['0.640']	['0.660']	['0.670']
MINIMO	0.470	0.470	0.470	0.470	0.470	0.470	0.470	0.470	0.470	0.470	0.470	0.470
MAXIMO	0.890	0.890	0.890	0.890	0.890	0.890	0.890	0.890	0.890	0.890	0.890	0.890
RANGO	0.420	0.420	0.420	0.420	0.420	0.420	0.420	0.420	0.420	0.420	0.420	0.420
DESVIACION STD	0.068	0.031	0.031	0.031	0.031	0.031	0.046	0.041	0.039	0.038	0.037	0.036
VARIANZA	0.005	0.001	0.001	0.001	0.001	0.001	0.002	0.002	0.002	0.001	0.001	0.001
PRIMER CUARTIL	0.630	0.657	0.657	0.657	0.657	0.657	0.643	0.647	0.650	0.652	0.653	0.654
SEGUNDO CUARTIL	0.670	0.670	0.670	0.670	0.670	0.670	0.670	0.670	0.672	0.672	0.671	0.671
TERCER CUARTIL	0.720	0.681	0.681	0.681	0.681	0.681	0.700	0.698	0.698	0.695	0.693	0.693
ASIMETRIA	0.099	0.338	0.338	0.338	0.338	0.338	0.102	0.089	0.064	0.070	0.100	0.112
CURTOSIS	0.519	10.147	10.147	10.147	10.147	10.147	1.271	2.180	2.868	3.547	4.101	4.735

Tabla A28. Comparación análisis estadístico con las técnicas de imputación utilizadas – Paciente 2

	IBI	MICE1	MICE5	MICE10	MICE20	MICE30	KNN3	KNN4	KNN5	KNN6	KNN7	KNN8
MEDIA	0.585	0.580	0.580	0.580	0.580	0.580	0.582	0.583	0.582	0.581	0.581	0.580
MEDIANA	0.580	0.575	0.575	0.575	0.575	0.575	0.570	0.573	0.572	0.570	0.570	0.570
MODA	['0.550']	['0.550']	['0.550']	['0.550']	['0.550']	['0.550']	['0.590']	['0.550']	['0.550']	['0.560']	['0.550']	['0.560']
MINIMO	0.440	0.425	0.424	0.424	0.424	0.424	0.440	0.440	0.440	0.440	0.440	0.440
MAXIMO	1.660	1.660	1.660	1.660	1.660	1.660	1.660	1.660	1.660	1.660	1.660	1.660
RANGO	1.220	1.235	1.236	1.236	1.236	1.236	1.220	1.220	1.220	1.220	1.220	1.220
DESVIACION STD	0.110	0.065	0.065	0.065	0.065	0.065	0.080	0.081	0.077	0.074	0.071	0.069
VARIANZA	0.012	0.004	0.004	0.004	0.004	0.004	0.006	0.007	0.006	0.006	0.005	0.005
PRIMER CUARTIL	0.530	0.550	0.550	0.550	0.550	0.550	0.547	0.550	0.550	0.550	0.550	0.550
SEGUNDO CUARTIL	0.580	0.575	0.575	0.575	0.575	0.575	0.570	0.573	0.572	0.570	0.570	0.570
TERCER CUARTIL	0.610	0.599	0.599	0.599	0.599	0.599	0.600	0.600	0.594	0.593	0.591	0.591
ASIMETRIA	5.449	6.446	6.437	6.437	6.437	6.437	5.137	5.325	5.497	5.705	5.972	6.333
CURTOSIS	40.390	82.186	82.019	82.019	82.019	82.019	43.404	43.157	48.375	52.965	59.830	68.487

Tabla A29. Comparación análisis estadístico con las técnicas de imputación utilizadas – Paciente 3

	IBI	MICE 1	MICE5	MICE1 0	MICE2 0	MICE3 0	KNN3	KNN4	KNN5	KNN6	KNN7	KNN8
MEDIA	0.718	0.712	0.712	0.712	0.712	0.712	0.719	0.715	0.714	0.714	0.714	0.714
MEDIANA	0.720	0.718	0.718	0.718	0.718	0.718	0.720	0.720	0.718	0.713	0.714	0.718
MODA	['0.720']	['0.720']	['0.720']	['0.720']	['0.720']	['0.720']	['0.773']	['0.752']	['0.728']	['0.713']	['0.714']	['0.719']
MINIMO	0.500	0.363	0.363	0.363	0.363	0.363	0.500	0.500	0.500	0.500	0.500	0.500
MAXIMO	0.920	0.920	0.920	0.920	0.920	0.920	0.920	0.920	0.920	0.920	0.920	0.920
RANGO	0.420	0.557	0.557	0.557	0.557	0.557	0.420	0.420	0.420	0.420	0.420	0.420
DESVIACION STD	0.084	0.085	0.085	0.085	0.085	0.085	0.054	0.048	0.044	0.041	0.040	0.038
VARIANZA	0.007	0.007	0.007	0.007	0.007	0.007	0.003	0.002	0.002	0.002	0.002	0.001
PRIMER CUARTIL	0.660	0.678	0.678	0.678	0.678	0.678	0.683	0.685	0.688	0.693	0.696	0.695
SEGUNDO CUARTIL	0.720	0.718	0.718	0.718	0.718	0.718	0.720	0.720	0.718	0.713	0.714	0.718
TERCER CUARTIL	0.780	0.752	0.752	0.752	0.752	0.752	0.770	0.752	0.732	0.730	0.734	0.731
ASIMETRIA	0.241	-1.311	-1.311	-1.311	-1.311	-1.311	-0.128	-0.026	0.056	0.175	0.135	0.153
CURTOSIS	-0.264	3.567	3.567	3.567	3.567	3.567	-0.135	0.654	1.438	2.107	2.340	2.865

Tabla A30. Comparación análisis estadístico con las técnicas de imputación utilizadas – Paciente 4

	IBI	MICE 1	MICE5	MICE1 0	MICE2 0	MICE3 0	KNN3	KNN4	KNN5	KNN6	KNN7	KNN8
MEDIA	0.825	0.824	0.824	0.824	0.824	0.824	0.816	0.816	0.816	0.817	0.818	0.818
MEDIANA	0.830	0.830	0.830	0.830	0.830	0.830	0.813	0.818	0.814	0.817	0.817	0.818
MODA	['0.860']	['0.860']	['0.860']	['0.860']	['0.860']	['0.860']	['0.840']	['0.825']	['0.800']	['0.810']	['0.860']	['0.860']
MINIMO	0.560	0.560	0.560	0.560	0.560	0.560	0.560	0.560	0.560	0.560	0.560	0.560
MAXIMO	1.110	1.110	1.110	1.110	1.110	1.110	1.110	1.110	1.110	1.110	1.110	1.110
RANGO	0.550	0.550	0.550	0.550	0.550	0.550	0.550	0.550	0.550	0.550	0.550	0.550
DESVIACION STD	0.071	0.050	0.050	0.050	0.050	0.050	0.060	0.058	0.056	0.055	0.055	0.054
VARIANZA	0.005	0.002	0.002	0.002	0.002	0.002	0.004	0.003	0.003	0.003	0.003	0.003
PRIMER CUARTIL	0.780	0.801	0.801	0.801	0.801	0.801	0.780	0.780	0.780	0.783	0.786	0.789
SEGUNDO CUARTIL	0.830	0.830	0.830	0.830	0.830	0.830	0.813	0.818	0.814	0.817	0.817	0.818
TERCER CUARTIL	0.880	0.847	0.847	0.847	0.847	0.847	0.853	0.853	0.850	0.852	0.850	0.850
ASIMETRIA	-0.085	-0.310	-0.310	-0.310	-0.310	-0.310	-0.041	-0.019	0.023	0.026	0.041	0.019
CURTOSIS	0.373	2.351	2.351	2.351	2.351	2.351	0.427	0.576	0.734	0.847	0.902	0.998

Tabla A31. Comparación análisis estadístico con las técnicas de imputación utilizadas – Paciente 5

	IBI	MICE 1	MICE5	MICE10	MICE20	MICE30	KNN3	KNN4	KNN5	KNN6	KNN7	KNN8
MEDIA	0.630	0.591	0.591	0.591	0.591	0.591	0.593	0.593	0.594	0.594	0.595	0.595
MEDIANA	0.630	0.589	0.589	0.589	0.589	0.589	0.590	0.588	0.584	0.583	0.583	0.585
MODA	['0.630']	['0.630']	['0.630']	['0.630']	['0.630']	['0.630']	['0.563']	['0.560']	['0.564']	['0.575']	['0.571']	['0.573']
MINIMO	0.380	0.362	0.362	0.362	0.362	0.362	0.380	0.380	0.380	0.380	0.380	0.380
MAXIMO	0.860	1.006	1.006	1.006	1.006	1.006	0.860	0.860	0.860	0.860	0.860	0.860
RANGO	0.480	0.644	0.644	0.644	0.644	0.644	0.480	0.480	0.480	0.480	0.480	0.480
DESVIACION STD	0.065	0.109	0.109	0.109	0.109	0.109	0.058	0.056	0.054	0.052	0.051	0.049
VARIANZA	0.004	0.012	0.012	0.012	0.012	0.012	0.003	0.003	0.003	0.003	0.003	0.002
PRIMER CUARTIL	0.590	0.510	0.510	0.510	0.510	0.510	0.563	0.560	0.564	0.565	0.567	0.560
SEGUNDO CUARTIL	0.630	0.589	0.589	0.589	0.589	0.589	0.590	0.588	0.584	0.583	0.583	0.585
TERCER CUARTIL	0.670	0.660	0.660	0.660	0.660	0.660	0.627	0.623	0.628	0.627	0.628	0.630
ASIMETRIA	-0.265	0.545	0.545	0.545	0.546	0.546	0.033	0.165	0.306	0.397	0.429	0.447
CURTOSIS	1.252	0.257	0.257	0.258	0.258	0.258	0.885	0.812	0.612	0.647	0.593	0.540

Tabla A32. Comparación análisis estadístico con las técnicas de imputación utilizadas – Paciente 6

	IBI	MICE 1	MICE5	MICE10	MICE20	MICE30	KNN3	KNN4	KNN5	KNN6	KNN7	KNN8
MEDIA	0.804	0.786	0.786	0.786	0.786	0.786	0.789	0.790	0.790	0.790	0.790	0.789
MEDIANA	0.800	0.800	0.800	0.800	0.800	0.800	0.790	0.795	0.792	0.793	0.794	0.796
MODA	['0.800']	['0.800']	['0.800']	['0.800']	['0.800']	['0.800']	['0.810']	['0.800']	['0.800']	['0.750']	['0.833']	['0.800']
MINIMO	0.560	0.560	0.560	0.560	0.560	0.560	0.560	0.560	0.560	0.560	0.560	0.560
MAXIMO	1.220	1.220	1.220	1.220	1.220	1.220	1.220	1.220	1.220	1.220	1.220	1.220
RANGO	0.660	0.660	0.660	0.660	0.660	0.660	0.660	0.660	0.660	0.660	0.660	0.660
DESVIACION STD	0.092	0.070	0.070	0.070	0.070	0.070	0.079	0.078	0.075	0.074	0.073	0.073
VARIANZA	0.008	0.005	0.005	0.005	0.005	0.005	0.006	0.006	0.006	0.005	0.005	0.005
PRIMER CUARTIL	0.730	0.720	0.720	0.720	0.720	0.720	0.730	0.730	0.730	0.738	0.733	0.730
SEGUNDO CUARTIL	0.800	0.800	0.800	0.800	0.800	0.800	0.790	0.795	0.792	0.793	0.794	0.796
TERCER CUARTIL	0.880	0.830	0.830	0.830	0.830	0.830	0.840	0.840	0.838	0.837	0.833	0.838
ASIMETRIA	0.212	0.302	0.302	0.302	0.302	0.302	0.288	0.266	0.278	0.300	0.301	0.289
CURTOSIS	0.107	0.978	0.978	0.978	0.978	0.978	0.289	0.327	0.513	0.629	0.668	0.626

Tabla A33. Comparación análisis estadístico con las técnicas de imputación utilizadas – Paciente 7

	IBI	MICE 1	MICE5	MICE10	MICE20	MICE30	KNN3	KNN4	KNN5	KNN6	KNN7	KNN8
MEDIA	0.613	0.612	0.612	0.612	0.612	0.612	0.615	0.615	0.614	0.614	0.614	0.614
MEDIANA	0.610	0.613	0.613	0.613	0.613	0.613	0.610	0.612	0.612	0.613	0.613	0.614
MODA	['0.590']	['0.590']	['0.590']	['0.590']	['0.590']	['0.590']	['0.610']	['0.610']	['0.610']	['0.590']	['0.590']	['0.640']
MINIMO	0.410	0.410	0.410	0.410	0.410	0.410	0.410	0.410	0.410	0.410	0.410	0.410
MAXIMO	1.080	1.080	1.080	1.080	1.080	1.080	1.080	1.080	1.080	1.080	1.080	1.080
RANGO	0.670	0.670	0.670	0.670	0.670	0.670	0.670	0.670	0.670	0.670	0.670	0.670
DESVIACION STD	0.076	0.039	0.039	0.039	0.039	0.039	0.056	0.051	0.048	0.046	0.045	0.044
VARIANZA	0.006	0.002	0.002	0.002	0.002	0.002	0.003	0.003	0.002	0.002	0.002	0.002
PRIMER CUARTIL	0.560	0.594	0.594	0.594	0.594	0.594	0.580	0.583	0.586	0.590	0.590	0.590
SEGUNDO CUARTIL	0.610	0.613	0.613	0.613	0.613	0.613	0.610	0.612	0.612	0.613	0.613	0.614
TERCER CUARTIL	0.660	0.629	0.629	0.629	0.629	0.629	0.643	0.642	0.640	0.638	0.639	0.637
ASIMETRIA	1.332	1.943	1.943	1.943	1.943	1.943	1.129	1.078	1.073	1.152	1.240	1.291
CURTOSIS	6.571	24.714	24.712	24.712	24.712	24.712	5.835	7.514	9.437	11.665	13.417	14.799

Tabla A34. Comparación análisis estadístico con las técnicas de imputación utilizadas – Paciente 8

	IBI	MICE 1	MICE5	MICE10	MICE 20	MICE30	KNN3	KNN4	KNN5	KNN6	KNN 7	KNN 8
MEDIA	0.528	0.525	0.525	0.525	0.525	0.525	0.529	0.529	0.529	0.529	0.529	0.528
MEDIANA	0.525	0.526	0.526	0.526	0.526	0.526	0.530	0.530	0.528	0.528	0.527	0.526
MODA	['0.500']	['0.500']	['0.500']	['0.500']	['0.500']	['0.500']	['0.520']	['0.530']	['0.530']	['0.520', '0.525']	['0.530']	['0.530']
MINIMO	0.390	0.390	0.390	0.390	0.390	0.390	0.390	0.390	0.390	0.390	0.390	0.390
MAXIMO	0.690	0.690	0.690	0.690	0.690	0.690	0.690	0.690	0.690	0.690	0.690	0.690
RANGO	0.300	0.300	0.300	0.300	0.300	0.300	0.300	0.300	0.300	0.300	0.300	0.300
DESVIACION STD	0.059	0.023	0.023	0.023	0.023	0.023	0.040	0.036	0.034	0.032	0.031	0.030
VARIANZA	0.003	0.001	0.001	0.001	0.001	0.001	0.002	0.001	0.001	0.001	0.001	0.001
PRIMER CUARTIL	0.480	0.513	0.513	0.513	0.513	0.513	0.503	0.502	0.506	0.508	0.510	0.510
SEGUNDO CUARTIL	0.525	0.526	0.526	0.526	0.526	0.526	0.530	0.530	0.528	0.528	0.527	0.526
TERCER CUARTIL	0.560	0.540	0.540	0.540	0.540	0.540	0.553	0.555	0.552	0.552	0.550	0.548
ASIMETRIA	0.297	0.381	0.381	0.381	0.381	0.381	-0.037	-0.055	0.020	0.093	0.148	0.190
CURTOSIS	-0.029	6.681	6.681	6.681	6.681	6.681	0.281	0.507	0.728	1.044	1.348	1.645

Tabla A35. Comparación análisis estadístico con las técnicas de imputación utilizadas – Paciente 9

	IBI	MICE 1	MICE 5	MICE 10	MICE2 0	MICE3 0	KNN3	KNN4	KNN5	KNN6	KNN7	KNN8
MEDIA	0.595	0.579	0.579	0.579	0.579	0.579	0.586	0.586	0.586	0.586	0.586	0.586
MEDIANA	0.590	0.575	0.575	0.575	0.575	0.575	0.580	0.583	0.584	0.583	0.584	0.585
MODA	['0.590']	['0.590']	['0.590']	['0.590']	['0.590']	['0.590']	['0.590']	['0.578']	['0.550', '0.580']	['0.590']	['0.590']	['0.550', '0.590']
MINIMO	0.390	0.390	0.390	0.390	0.390	0.390	0.390	0.390	0.390	0.390	0.390	0.390
MAXIMO	1.140	1.140	1.140	1.140	1.140	1.140	1.140	1.140	1.140	1.140	1.140	1.140
RANGO	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750
DESVIACION STD	0.084	0.053	0.053	0.053	0.053	0.053	0.061	0.056	0.053	0.051	0.049	0.048
VARIANZA	0.007	0.003	0.003	0.003	0.003	0.003	0.004	0.003	0.003	0.003	0.002	0.002
PRIMER CUARTIL	0.550	0.539	0.539	0.539	0.539	0.539	0.543	0.550	0.552	0.553	0.556	0.557
SEGUNDO CUARTIL	0.590	0.575	0.575	0.575	0.575	0.575	0.580	0.583	0.584	0.583	0.584	0.585
TERCER CUARTIL	0.630	0.611	0.611	0.611	0.611	0.611	0.623	0.620	0.616	0.617	0.616	0.614
ASIMETRIA	1.308	0.974	0.974	0.974	0.974	0.974	0.659	0.715	0.807	0.875	0.922	0.971
CURTOSIS	5.431	5.413	5.410	5.410	5.410	5.410	2.622	3.856	5.081	6.176	7.123	7.916

Tabla A36. Comparación análisis estadístico con las técnicas de imputación utilizadas – Paciente 10

	IBI	MICE 1	MICE5	MICE1 0	MICE2 0	MICE3 0	KNN3	KNN4	KNN5	KNN6	KNN7	KNN8
MEDIA	0.568	0.561	0.561	0.561	0.561	0.561	0.557	0.558	0.558	0.558	0.559	0.559
MEDIANA	0.560	0.562	0.562	0.562	0.562	0.562	0.553	0.552	0.552	0.553	0.554	0.555
MODA	['0.550']	['0.550']	['0.550']	['0.550']	['0.550']	['0.550']	['0.563']	['0.550']	['0.550']	['0.542']	['0.550']	['0.530']
MINIMO	0.410	0.410	0.410	0.410	0.410	0.410	0.410	0.410	0.410	0.410	0.410	0.410
MAXIMO	0.810	0.810	0.810	0.810	0.810	0.810	0.810	0.810	0.810	0.810	0.810	0.810
RANGO	0.400	0.400	0.400	0.400	0.400	0.400	0.400	0.400	0.400	0.400	0.400	0.400
DESVIACION STD	0.059	0.034	0.034	0.034	0.034	0.034	0.043	0.039	0.036	0.035	0.033	0.033
VARIANZA	0.003	0.001	0.001	0.001	0.001	0.001	0.002	0.002	0.001	0.001	0.001	0.001
PRIMER CUARTIL	0.530	0.539	0.539	0.539	0.539	0.539	0.530	0.532	0.534	0.533	0.536	0.535
SEGUNDO CUARTIL	0.560	0.562	0.562	0.562	0.562	0.562	0.553	0.552	0.552	0.553	0.554	0.555
TERCER CUARTIL	0.590	0.584	0.584	0.584	0.584	0.584	0.580	0.580	0.580	0.580	0.579	0.579
ASIMETRIA	0.584	0.122	0.122	0.122	0.122	0.122	0.552	0.637	0.724	0.811	0.871	0.904
CURTOSIS	1.558	2.633	2.633	2.633	2.633	2.633	1.024	1.406	2.061	2.681	3.206	3.542

Tabla A37. Comparación análisis estadístico con las técnicas de imputación utilizadas – Paciente 11

	IBI	MICE 1	MICE5	MICE1 0	MICE2 0	MICE3 0	KNN3	KNN4	KNN5	KNN6	KNN7	KNN8
MEDIA	0.667	0.666	0.666	0.666	0.666	0.666	0.668	0.668	0.668	0.668	0.668	0.668
MEDIANA	0.670	0.664	0.664	0.664	0.664	0.664	0.667	0.670	0.670	0.670	0.670	0.670
MODA	['0.67 0']	['0.67 0']	['0.670 ']	['0.670 ']	['0.670 ']	['0.670 ']	['0.640 ']	['0.690 ']	['0.670 ']	['0.640 ']	['0.660 ']	['0.690 ']
MINIMO	0.470	0.470	0.470	0.470	0.470	0.470	0.470	0.470	0.470	0.470	0.470	0.470
MAXIMO	0.950	0.950	0.950	0.950	0.950	0.950	0.950	0.950	0.950	0.950	0.950	0.950
RANGO	0.480	0.480	0.480	0.480	0.480	0.480	0.480	0.480	0.480	0.480	0.480	0.480
DESVIACION STD	0.073	0.046	0.046	0.046	0.046	0.046	0.058	0.056	0.054	0.053	0.052	0.051
VARIANZA	0.005	0.002	0.002	0.002	0.002	0.002	0.003	0.003	0.003	0.003	0.003	0.003
PRIMER CUARTIL	0.610	0.649	0.649	0.649	0.649	0.649	0.630	0.635	0.638	0.640	0.640	0.640
SEGUNDO CUARTIL	0.670	0.664	0.664	0.664	0.664	0.664	0.667	0.670	0.670	0.670	0.670	0.670
TERCER CUARTIL	0.700	0.680	0.680	0.680	0.680	0.680	0.700	0.700	0.700	0.698	0.696	0.695
ASIMETRIA	0.398	0.673	0.673	0.673	0.673	0.673	0.329	0.281	0.281	0.289	0.283	0.291
CURTOSIS	0.836	6.021	6.021	6.021	6.021	6.021	1.326	1.713	2.099	2.503	2.781	3.050

Tabla A38. Comparación análisis estadístico con las técnicas de imputación utilizadas – Paciente 12

	IBI	MICE 1	MICE5	MICE1 0	MICE2 0	MICE3 0	KNN3	KNN4	KNN5	KNN6	KNN7	KNN8
MEDIA	0.616	0.546	0.546	0.546	0.546	0.546	0.574	0.575	0.576	0.576	0.577	0.578
MEDIANA	0.630	0.554	0.554	0.554	0.554	0.554	0.573	0.562	0.566	0.565	0.567	0.569
MODA	['0.61 0']	['0.61 0']	['0.610 ']	['0.610 ']	['0.610 ']	['0.610 ']	['0.573 ']	['0.560 ']	['0.550 ']	['0.563 ']	['0.537 ']	['0.569 ']
MINIMO	0.380	0.279	0.279	0.279	0.279	0.279	0.380	0.380	0.380	0.380	0.380	0.380
MAXIMO	0.840	0.840	0.840	0.840	0.840	0.840	0.840	0.840	0.840	0.840	0.840	0.840
RANGO	0.460	0.561	0.561	0.561	0.561	0.561	0.460	0.460	0.460	0.460	0.460	0.460
DESVIACION STD	0.082	0.086	0.086	0.086	0.086	0.086	0.062	0.059	0.058	0.057	0.056	0.055
VARIANZA	0.007	0.007	0.007	0.007	0.007	0.007	0.004	0.004	0.003	0.003	0.003	0.003
PRIMER CUARTIL	0.560	0.479	0.479	0.479	0.479	0.479	0.530	0.535	0.538	0.540	0.540	0.544
SEGUNDO CUARTIL	0.630	0.554	0.554	0.554	0.554	0.554	0.573	0.562	0.566	0.565	0.567	0.569
TERCER CUARTIL	0.670	0.613	0.613	0.613	0.613	0.613	0.610	0.610	0.608	0.607	0.609	0.610
ASIMETRIA	-0.178	-0.044	-0.044	-0.044	-0.045	-0.045	0.538	0.613	0.679	0.721	0.735	0.729
CURTOSIS	-0.074	-0.620	-0.620	-0.620	-0.620	-0.620	0.434	0.566	0.725	0.843	0.882	0.926