Sistema para la Clasificación Automática de Comportamientos Sedentarios en Entornos Cerrados

Jesús David Cerón Bravo

Tesis de Maestría en Ingeniería Telemática

Director:
Diego Mauricio López Gutiérrez
PhD en Ciencias Biomédicas

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Departamento de Telemática
Línea de Investigación e-Salud
Popayán, abril de 2017

Jesús David Cerón Bravo

Sistema para la Clasificación Automática de Comportamientos Sedentarios en Entornos Cerrados

Tesis presentada a la Facultad de Ingeniería Electrónica y Telecomunicaciones de la Universidad del Cauca para la obtención del Título de

Magister en:
Ingeniería Telemática

Director:
Diego Mauricio López Gutiérrez
PhD en Ciencias Biomédicas

Popayán
2017

Universidad del Cauca

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Programas de Posgrado

FORMATO I
ACTA DE SUSTENTACIÓN DE
TRABAJO DE GRADO DE MAESTRÍA/TESIS DOCTORAL

Los jurados de:

TRABAJO DE GRADO DE MAESTRÍA (X) TESIS DOCTORAL ( )

Titulado: Sistema para la clasificación automática de comportamientos sedentarios en entornos cerrados.

Bajo la dirección de: _____

HACEN CONSTAR:

Que siendo las 7:30 del día 12 del mes de mayo de 2017, el(la) estudiante: Jesús David Cerón Bravo

realizó la Sustentación del Trabajo de Grado de Maestría/Tesis Doctoral, obteniendo la calificación de:

APROBADO (X) NO APROBADO ( ).

Para constancia, se firma en Popayán, a los 12 días del mes de mayo de 2017.

JURADO 1:

Nombre: José Armando Ordóñez

JURADO 2 (Doctorado):

Nombre:

JURADO COORDINADOR:

Nombre: Gustavo Adolfo Ramírez G

COORDINADOR DEL PROGRAMA (*):

Nombre:

(*) Firma el Coordinador del Programa con autorización del(los) Jurado(s):

Agradecimientos

Mis más sinceros agradecimientos a todos aquellos que hicieron posible la realización de este trabajo. Debo resaltar, por supuesto, al doctor Diego Mauricio López Gutiérrez, quien ha sido más que mi director, mi guía y ejemplo desde mis estudios de pregrado. Agradezco al Departamento Administrativo de Ciencia, Tecnología e Innovación (Colciencias) ya que este trabajo se gestó y se llevó a feliz término en el contexto del proyecto de investigación SIMETIC (convocatoria 596-2012) el cual fue financiado por esta entidad. Mi gratitud para el Dipl.-Ing Christian Hofmann, director del grupo Medical Sensor Systems del Fraunhofer Institute for Integrated Circuits ISS de Alemania, quien me recibió en su grupo durante un valioso mes en el que adquirí valioso conocimiento para el desarrollo de este trabajo.

Resumen estructurado

Antecedentes

Diversos estudios han demostrado la relación entre el sedentarismo y/o inactividad física con el desarrollo del SM, diabetes mellitus tipo 2 y ECV]. Además, se ha encontrado que las personas con mayores niveles de comportamientos sedentarios sufren de un mayor riesgo de morbilidad y mortalidad, independientemente de que practiquen Actividad Física Moderada o Vigorosa (AFMV) . Esto significa que pasar mucho tiempo realizando algún comportamiento sedentario genera un incremento del riesgo de padecer enfermedades cardiovasculares, diabetes o síndrome metabólico independiente que el generado por realizar poca o no realizar AFMV. Según los resultados de una encuesta realizada en el proyecto de investigación SIMETIC, se encontró que el 36.7 por ciento de una muestra de 2100 personas laboralmente activas de la ciudad de Popayán poseen el síndrome metabólico y la prevalencia de sedentarismo en una muestra de 589 personas laboralmente activas fue del 71,1%. Adicionalmente, existen otros estudios que han estimado de manera subjetiva la prevalencia de sedentarismo en otras ciudades Colombianas, los cuales han arrojado como resultado valores por encima del 70%. Según Owen y colaboradores [14], con el fin de desarrollar y probar intervenciones para influir en un estilo de vida sedentario, es necesario medir con precisión los comportamientos sedentarios (clasificarlos correctamente) y conocer sus determinantes contextuales: esto es, identificar el lugar y/o momento en que estos ocurren, por ejemplo en el hogar, trabajo, transporte y recreación; e identificar los determinantes de los comportamientos sedentarios. Considerando los problemas descritos anteriormente y teniendo en cuenta el estado del arte presentado en el capítulo 2, en este trabajo de grado de maestría se plantea la siguiente pregunta de investigación: ¿Cómo soportar la clasificación automática de comportamientos sedentarios en entornos cerrados?

La hipótesis planteada establece que es factible desarrollar un sistema para la clasificación automática de comportamientos sedentarios en entornos cerrados implementando un sistema de localización en entornos cerrados usando BLE beacons y un monitor de actividad física.

Objetivos

El objetivo general planteado para responder a la pregunta de investigación establecida es: Proponer un sistema para la clasificación automática de comportamientos sedentarios en individuos basado en su localización en entornos cerrados. Para lograr el cumplimiento de ese gran objetivo, fueron planteados los siguientes tres objetivos específicos:

1. Construir un conjunto de datos (dataset) de comportamientos sedentarios el cual incluya datos de acelerometría y localización.
2. Seleccionar un algoritmo clasificador de comportamientos sedentarios basado en aprendizaje supervisado empleando el dataset obtenido.
3. Evaluar experimentalmente la precisión del sistema desarrollado.

Métodos

El planteamiento de los anteriores objetivos lleva a deducir que el presente proyecto debe ser abordado como un proyecto de minería de datos. Por tal razón la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), la más usada para proyectos de minería de datos al momento de realizar esta tesis es seguida rigurosamente, y sus seis fases dan la estructura base a este documento. Se inicia con la fase de Entendimiento del negocio donde es evaluado el estado actual del conocimiento y por medio de este se plantea un plan de proyecto. Acto seguido, en la fase de entendimiento de los datos se recolectan los datos requeridos, ya que según el estado del arte realizado en la fase 1, no existen datasets que permitan continuar con la ejecución del proyecto. La fase 3 es la fase de preparación de los datos, donde se da formato a los datos recolectados para que puedan ser utilizados en la fase 4, la fase de modelado. En esa fase se construyen los modelos para la clasificación de los comportamientos sedentarios seleccionados para posteriormente ser evaluados en la fase 6, la fase de evaluación. Finalmente la fase de despliegue es la encargada de poner en uso el modelo o modelos aprobados para realizar la clasificación en un entorno real.

Resultados

Un completo análisis acerca de la exactitud en la clasificación de 23 comportamientos sedentarios teniendo en cuenta el lugar del cuerpo en el cual fueron obtenidos los datos y el tipo de modelo generado a partir de ellos arroja como resultado que recolectando los datos desde la cintura y usando modelos personales se obtiene una exactitud de 99%.

Conclusiones

1. La necesidad de realizar la clasificación de comportamientos sedentarios está soportada en el reciente interés de la comunidad científica acerca de su impacto en la salud de las personas. Ese interés ha llevado a la propuesta de una taxonomía de comportamientos sedentarios, que fue la base para la selección de los comportamientos sedentarios clasificados en el presente proyecto. Sin duda alguna, las contribuciones aquí aportadas servirán de línea de base para futuras investigaciones acerca de la clasificación automática de ese tipo de comportamientos.

2. Los análisis realizados acerca de los tipos de modelos (personal, híbrido e impersonal), evidencian que la generación de un modelo universal, el cual pueda ser utilizado en frío por cualquier persona, es una tarea que llevaría bastante trabajo, ya que según los muestran las curvas de aprendizaje obtenidas, se necesita una gran cantidad de personas para lograr un modelo universal que provea una exactitud de al menos un 80%.

3. La retribución de recolectar datos personales para la clasificación de los CS y con ellos generar modelos personales, es el buen nivel de exactitud obtenido en la clasificación.

4. A saber, este es el primer sistema dirigido hacia la clasificación de comportamientos sedentarios. El sistema está compuesto por una aplicación móvil Android que se ejecuta en background, recibiendo continuamente la señal bluetooth de los beacons ubicados en el contexto del entorno cerrado, en este caso la casa.

5. El seguimiento de la metodología CRISP-DM guio la ejecución de este proyecto, las tareas incluidas en cada una de las fases son consistentes con lo realmente necesario para proyectos de minería de datos, por tal razón es la metodología para minería de datos más utilizada a nivel mundial.

Palabras Clave

Clasificación de comportamientos sedentarios, minería de datos, conjunto de datos

Structured abstract

Background

Several studies have demonstrated the relationship between sedentary lifestyle and / or physical inactivity with the development of MS, type 2 diabetes mellitus and CVD. In addition, it has been found that people with higher levels of sedentary behaviors suffer from an increased risk of morbidity and mortality, regardless of whether they practice Moderate or Vigorous Physical Activity (MVPA). This means that spending a lot of time performing some sedentary behaviors generates an increased risk of cardiovascular disease, diabetes or metabolic syndrome independent of that generated by performing little or not performing AFMV. According to the results of a survey carried out in the SIMETIC research project, it was found that 36.7 percent of a sample of 2100 workers in the city of Popayán has the metabolic syndrome and the prevalence of sedentary lifestyle in a sample of 589 people was 71.1%. In addition, there are other studies that have subjectively estimated the prevalence of sedentarism in other Colombian cities, which have resulted in values above 70%. According to Owen et al. [14], in order to develop and test interventions to influence a sedentary lifestyle, it is necessary to accurately measure sedentary behaviors (classify them correctly) and to know their contextual determinants: that is, identify the place and / or the moment they occur, for example at home, work, transportation and recreation; And identify the determinants of sedentary behavior. Considering the problems described above and taking into account the state of the art presented in Chapter 2, in this master's degree the following research question is posed: How to support the automatic classification of sedentary behaviors in indoor environments?
The hypothesis stated that it is feasible to develop a system for the automatic classification of sedentary behaviors in indoors environments by implementing a localization system in indoors environments using BLE beacons and a physical activity monitor.

Aims

The general objective to answer the established research question is: To propose a system for the automatic classification of sedentary behaviors in individuals based on

their location in indoors environments. To achieve the fulfillment of this general objective, the following three specific objectives were proposed:

1. To build a dataset of sedentary behaviors that includes accelerometry and location data.
2. To select a sedentary behavior classifier algorithm based on supervised learning using the obtained dataset.
3. To evaluate experimentally the accuracy of the developed system.

Methods

The approach of the previous objectives leads to deduce that the present project must be approached as a project of data mining. For this reason, the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology, the most used for data mining projects at the time of this thesis was carried out is strictly followed, and its six phases give the base structure to this document. It starts with the understanding phase of the business where the current state of knowledge is evaluated and through this a project plan is proposed. Then, in the understanding of the data phase, the required data are collected, since according to the state of the art made in phase 1, there are no datasets that allow to continue the execution of the project. Phase 3 is the data preparation phase, where the data collected are formatted so that they can be used in phase 4, the modeling phase. In this phase, the models for the classification of the selected sedentary behaviors are constructed and later evaluated in phase 6, the evaluation phase. Finally the deployment phase is in charge of putting into use the model or models approved to perform the classification in a real environment.

Results

A complete analysis of the classification accuracy of 23 sedentary behaviors taking into account the location of the body in which the data were obtained and the type of model generated from them resulted in that collecting data from the waist and using personal models an accuracy of 99% is obtained.

Conclusions

1. The need to perform the classification of sedentary behaviors is supported by the recent interest of the scientific community about its impact on the health. This interest has led to the proposal of a taxonomy of sedentary behaviors, which was the basis for the selection of the 23 sedentary behaviors classified in this project. Undoubtedly, the contributions provided here will serve as a baseline for future research into the automatic classification of such behaviors.

2. Analysis of the types of models (personal, hybrid and impersonal) shows that the generation of a universal model, which can be used by anyone (without collecting personal data), is a task that would take a lot of work, since according to the learning curves obtained, data taken from a lot of people is needed to achieve a universal model that provides an accuracy of at least 80%.

3. The retribution of collecting personal data and with them generate personal models for the classification of sedentary behaviors, is the good level of accuracy obtained in the classification.

4. Namely, this is the first system aimed at the classification of sedentary behaviors. The system consists of an Android mobile application that runs in background, continuously receiving the Bluetooth signal from the beacons located in the context of the indoor environment.

5. Following the CRISP-DM methodology guiding the execution of this project, the tasks included in each of the phases are consistent with what is really necessary for data mining projects, for this reason is the methodology for data mining used worldwide.

Keywords

Sedentary behaviors classification, data mining, dataset.

Contenido

Lista de figuras

Lista de tablas

Capítulo 1

1 Introducción.

La Organización Mundial de la Salud (OMS) ha evidenciado que las enfermedades cardiovasculares (ECV) han sido la mayor causa de muerte en la última década [1]. Según esta organización, se calcula que en 2012 murieron por esta razón 17,5 millones de personas, lo cual representa un 31% de todas las muertes registradas en el mundo. Las proyecciones realizadas en el informe de Mathers y colaboradores [2] concluyen que las muertes derivadas de enfermedades cardiovasculares ascenderán a 23.3 millones en el 2030. La diabetes mellitus tipo 2 es un factor de riesgo importante para ECV. La Federación Internacional de la Diabetes (FID) estipula que cada año 3.2 millones de personas mueren a causa de complicaciones relacionadas con diabetes, siendo la diabetes tipo 2 la causante del 90% de estas muertes [3].

Por lo anterior, el reconocimiento de síntomas previos a la diabetes tipo 2 se ha convertido en un fin de gran importancia, de esta manera, la FID y la comunidad cardiovascular en general están fuertemente unidas en el estudio de un grupo de factores de riesgo comunes agrupados a través de la definición de una nueva entidad clínica, denominada síndrome metabólico (SM) [4], [5]. El síndrome metabólico se define como un grupo de desórdenes médicos que incrementan el riesgo de desarrollar una ECV y diabetes [6]. Se estima que alrededor del 20-25 por ciento de la población adulta del mundo tienen el SM [7]. Esta población, en comparación con las personas que no poseen el síndrome, presenta el doble de probabilidades de morir por un ataque al corazón o un derrame cerebral y tres veces más probabilidades de padecerlo. Además, las personas con SM tienen un riesgo cinco veces mayor de desarrollar diabetes tipo 2 [4].

Diversos estudios han demostrado la relación entre el sedentarismo y/o inactividad física[1] con el desarrollo del SM, diabetes mellitus tipo 2 y ECV [8]–[13]. Además, se ha encontrado que las personas con mayores niveles de comportamientos sedentarios[2] sufren de un mayor riesgo de morbilidad y mortalidad, independientemente de que practiquen Actividad Física Moderada o Vigorosa (AFMV) [14], [15]. Esto significa que pasar mucho tiempo realizando algún comportamiento sedentario genera un incremento del riesgo de padecer enfermedades cardiovasculares, diabetes o síndrome metabólico independiente que el generado por realizar poca o no realizar AFMV.

1.1 Definición del problema.

Según Owen y colaboradores [14], con el fin de desarrollar y probar intervenciones para influir en un estilo de vida sedentario, es necesario medir con precisión los comportamientos sedentarios (clasificarlos correctamente) y conocer sus determinantes contextuales: esto es, identificar el lugar y/o momento en que estos ocurren, por ejemplo en el hogar, trabajo, transporte y recreación; e identificar los determinantes de los comportamientos sedentarios. Según la revisión hecha por Sanders et al. [16], a mayo del 2016 existían 73 dispositivos wearables comerciales para registrar el seguimiento de la actividad física y solamente 9 para realizar el registro del tiempo sedentario. En el caso del registro del tiempo sedentario, este se realiza comúnmente de dos formas: con sensores de presión, por ejemplo, puestos en los zapatos o con algoritmos propietarios que reconocen posturas como estar sentado, inclinado o acostado usando otros tipos de wearables. Es importante notar que ninguno de estos dispositivos hace una clasificación explícita de los comportamientos sedentarios, solo registran el tiempo del comportamiento, es decir, no son capaces de determinar, por ejemplo, si la persona está viendo televisión, leyendo un libro o trabajando en el computador en su casa o sitio de trabajo. Mediante el uso de dispositivos wearables o de teléfonos inteligentes, varios estudios han evidenciado que durante las horas que los adultos se encuentran despiertos en un día, aproximadamente el 60% de ese tiempo (9.3 horas/día) incurre en algún comportamiento sedentario, el 35% (6.5 horas/día) a una diversidad de actividades físicas de intensidad leve y tan solo el 5% (0.7 horas/día) los dedican a actividades físicas que representan una intensidad de moderada a vigorosa [17]–[20]. Además,

[1] Se considera a una persona ‘inactiva’ cuando no cumple con las recomendaciones de cantidad de actividad física moderada o vigorosa (AFMV) para su salud [33], que para el caso de la OMS equivale a 150 minutos semanales.

[2] El concepto ‘hábito sedentario’ (en adelante ‘comportamientos sedentarios’) se define como: “cualquier actividad realizada por el individuo en posición sentada o inclinada con un gasto energético ≤1.5 METs, mientras se está despierto” [33].

se ha encontrado que el 90% de los comportamientos sedentarios ocurre en entornos cerrados tales como la casa o el lugar de trabajo [21], [22].

Al igual que al realizar la clasificación objetiva de actividad física, realizar una clasificación objetiva de comportamientos sedentarios requiere emplear métodos de clasificación, los cuales son de naturaleza probabilística, razón por la cual, la inclusión de datos basados en la localización en la cual estos ocurren en un entorno cerrado dentro de los diferentes algoritmos de clasificación, podría elevar su nivel de precisión. Además, lo anterior ayudaría a dar cumplimiento a los requerimientos propuestos por Owen y colaboradores descritos anteriormente [14]. El tiempo que pasa una persona en entornos cerrados se ha inferido tradicionalmente a través de la ausencia de una señal de GPS [23] o mediante el uso de un sensor de luminosidad incorporado en wearables [24]. Sin embargo, estos métodos sólo son capaces de diferenciar entre un entorno cerrado y uno externo, por lo tanto no ofrecen información detallada de la localización en el entorno cerrado. El GPS no es el dispositivo adecuado para obtener la localización en el entorno cerrado debido a su poca precisión en esta clase de lugares [25], es por eso que se deben explorar otras tecnologías que lo permitan. En ese sentido, tecnologías como RFID, WIFI, Bluetooth, ZigBee, entre otras, podrían proveer una alternativa adecuada. Comparado con las demás tecnologías, Bluetooth Low Energy se encuentra actualmente como una solución adecuada por las siguientes características:

Es poco invasiva gracias a las etiquetas (beacons) de tamaño reducido.

1. Las etiquetas tienen un bajo costo (su costo por cantidad está alrededor de 7 USD según las tiendas virtuales).
2. La autonomía energética de una etiqueta puede ser mayor a un año.
3. El alcance de la señal de una etiqueta puede variarse según la potencia que se le suministre.
4. Una gran cantidad de Smartphones incluye la tecnología Bluetooth Low Energy.

Con base en la revisión realizada en [16] y según la revisión sistemática realizada en este trabajo y presentada en el próximo capítulo, se encuentra que no existe un sistema que permita obtener una clasificación automática de comportamientos sedentarios relacionado con su localización en el entorno cerrado empleando la tecnología Bluetooth.

1.2 Motivación

Este trabajo de maestría es relevante para el proyecto de investigación SimeTIC (VRI ID 3127) financiado por Colciencias y la Universidad Del Cauca. El objetivo del proyecto es proponer y evaluar estrategias soportadas en las TIC para promover el autocuidado en personas con SM en la ciudad de Popayán.

Como uno de los resultados de este proyecto de investigación, se encontró que el 36.7 por ciento de una muestra de 2100 personas laboralmente activas de la ciudad de Popayán poseen el síndrome metabólico y la prevalencia de sedentarismo en una

muestra de 589 personas laboralmente activas fue del 71,1%. Adicionalmente, existen otros estudios que han estimado de manera subjetiva la prevalencia de sedentarismo en otras ciudades Colombianas, los cuales han arrojado como resultado valores por encima del 70% [26]–[28].

1.3 Pregunta de investigación e hipótesis

Considerando los problemas descritos anteriormente y teniendo en cuenta el estado del arte presentado en el capítulo 2, en este trabajo de grado de maestría se plantea la siguiente pregunta de investigación: ¿Cómo soportar la clasificación automática de comportamientos sedentarios en entornos cerrados?
La hipótesis planteada establece que es factible desarrollar un sistema para la clasificación automática de comportamientos sedentarios en entornos cerrados implementando un sistema de localización en entornos cerrados usando BLE beacons y un monitor de actividad física.

1.4 Objetivos

Para dar respuesta a la pregunta de investigación planteada se describen a continuación los objetivos de este trabajo.

1.4.1 Objetivo General

Proponer un sistema para la clasificación automática de comportamientos sedentarios en individuos basado en su localización en entornos cerrados.

1.4.2 Objetivos Específicos

4. Construir un conjunto de datos (dataset) de comportamientos sedentarios el cual incluya datos de acelerometría y localización.
5. Seleccionar un algoritmo clasificador de comportamientos sedentarios basado en aprendizaje supervisado empleando el dataset obtenido.
6. Evaluar experimentalmente la precisión del sistema desarrollado. Contribuciones indirectas.

1.5 Metodología

La búsqueda de una solución a la pregunta de investigación planteada se hará bajo los parámetros de un proyecto de minería de datos. Debido al número creciente de proyectos de minería de datos en la última década, han surgido diversos enfoques que plantean una perspectiva sistemática para llevar a cabo el proceso de extracción de patrones a partir de los datos. Entre estos enfoques sobresalen KDD, SEMMA, CRISP-DM y Catalyst. Entre los cuales según [29] solo deberían ser considerados metodologías de minería de datos las últimas dos, ya que además de describir las actividades específicas de cada fase, proveen una guía de cómo llevar a cabo el trabajo. Para la consecución de los objetivos planteados se decide usar CRISP-DM ya que además de ser la metodología más utilizada en proyectos de minería de datos desde al año 2007 [30] como se evidencia en la Tabla 1, es una metodología de libre uso en la cual no se necesita más herramientas que las teóricas para poner en marcha un proyecto de ese tipo [31]. Esto hace posible encontrar una gran cantidad de información y proyectos en que se detalla su uso y se evidencia los resultados positivos de su implementación.

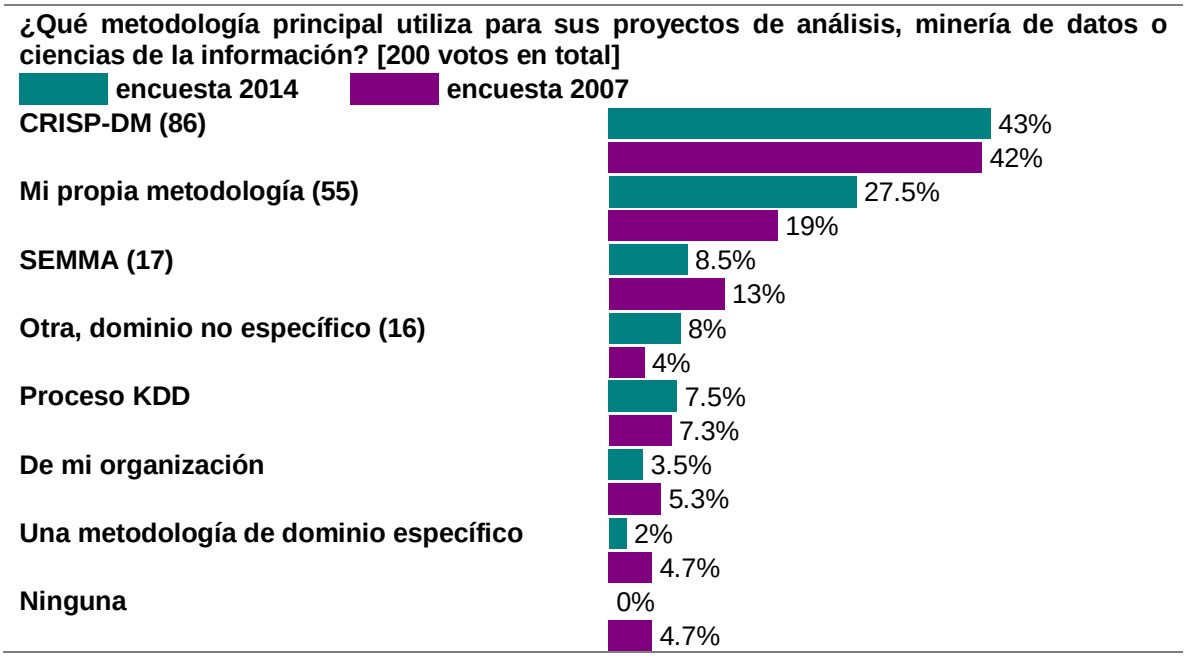

Tabla 1. Resultados de encuesta sobre metodologías para proyectos de minería de datos más utilizadas.

CRISP-DM es una metodología descrita en términos de un modelo de proceso jerárquico, consistente en un conjunto de tareas descritas en cuatro niveles de abstracción (de lo general a lo específico): fase, tarea genérica, tarea especializada, e instancia de procesos [32].

En el nivel superior, el proceso de minería de datos es organizado en seis fases. Cada fase consiste de varias tareas genéricas de segundo nivel. Este segundo nivel

es llamado genérico porque está destinado a ser bastante general para cubrir todas las situaciones posibles de minería de datos. Las tareas genéricas están destinadas a ser tan completas y estables como sea posible. Por *Completo* se entiende que cubre el proceso entero de minería de datos para cualquier proyecto de minería de datos posible. *Estable* significa que el modelo debería ser válido para acontecimientos normales y aún para desarrollos imprevistos como por ejemplo técnicas de modelado nuevas. El tercer nivel, el nivel de tarea especializado, es el lugar para describir cómo deberían ser realizadas las acciones en las tareas genéricas dadas ciertas situaciones específicas. Por ejemplo, en el segundo nivel podría haber una tarea genérica llamada limpieza de datos. El tercer nivel describe cómo esta tarea se diferencia en situaciones diferentes, como la limpieza de valores numéricos contra la limpieza de valores categóricos, o si el tipo de problema es de agrupamiento o de modelado predictivo. El cuarto nivel, la instancia de proceso, es un registro de las acciones, decisiones, y de los resultados de un trabajo real de minería de datos.

La descripción de fases y tareas como pasos discretos realizados en un orden específico representa una secuencia idealizada de eventos. En la práctica, muchas de las tareas pueden ser realizadas en un orden diferente, y a menudo será necesario volver a ejecutar tareas anteriores repetidamente, así como repetir ciertas acciones. Según el modelo de referencia CRISP-DM, el ciclo de vida de un proyecto de minería de datos consiste en seis fases, como se muestra en la Figura 1.

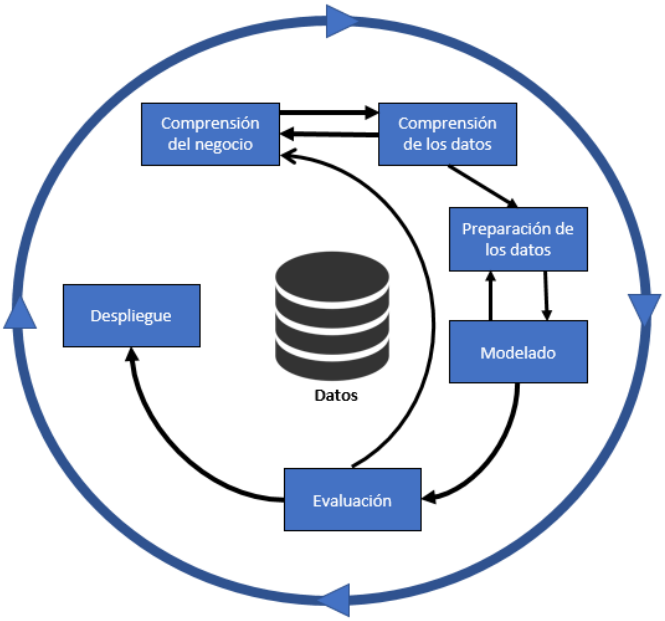

Figura 1. Ciclo de vida de la metodología CRISP-DM

1.6 Contenido de la monografía.

La monografía contiene siete capítulos, siendo el primero de ellos este capítulo introductorio al tema de investigación y los seis restantes corresponden a cada una de las seis fases de la metodología CRISP-DM.

Capítulo 2

2 Fase 1: Entendimiento del negocio.

La fase inicial de todo proyecto de minería de datos usando CRISP-DM es de vital importancia porque en ella se tiene que comprender con claridad el objetivo de negocio que se deberá dar por cumplido al finalizar el proyecto. En la fase de entendimiento del negocio se pretende comprender el o los objetivos y requisitos del negocio, para posteriormente llegar a unos objetivos más técnicos (centrados en la minería de datos), de manera que al finalizar la fase surja un plan del proyecto. La Tabla 2 muestra las tareas de la presente fase.

Entendimiento del negocio
Determinar el objetivo del negocio
Fondo
Objetivos del negocio
Criterios de éxito del negocio
Evaluación de la situación
Inventario de recursos
Requerimientos, supuestos y limitaciones
Terminología
Costo Beneficio
Determinar los objetivos de la minería de datos
Objetivos de la minería de datos
Criterios de éxito de la minería de datos.
Realizar el plan del proyecto
Plan del proyecto
Evaluación inicial de herramientas y técnicas

Tabla 2. Tareas de la fase de entendimiento de negocio.

2.1 Determinar los objetivos del negocio.

El fin último de esta tarea es obtener el o los objetivos del negocio, así como sus criterios de éxito. Para ello, se debe empezar con la elaboración de un (background) con el cual se ponga en evidencia el estado actual del tema central del proyecto, en este caso, los sistemas para la clasificación de comportamientos sedentarios. Este es el momento oportuno para realizar lo que en el contexto de investigación se conoce como estado actual de conocimiento o estado del arte.

2.1.1 Fondo (background).

Entendiendo esta actividad como un requerimiento indispensable para la realización del proyecto de minería de datos que se está abordando, no se han escatimado esfuerzos para obtener un óptimo estado actual del conocimiento que vislumbrará los objetivos del proyecto de minería de datos por medio del reconocimiento de las brechas existentes. Por esa razón, la estrategia para obtener el estado del arte, o estado actual del conocimiento fue una revisión sistemática.

2.1.1.1 Estrategia de búsqueda.

La búsqueda se realizó en abril de 2016 en tres bases de datos electrónicas: PubMed, Science Direct e IEEE Xplore Digital Library. La estrategia de búsqueda fue construida alrededor de cuatro grupos de palabras clave: comportamientos sedentarios (screen time, body posture, sitting time), clasificación (classification, recognition, dataset), método de monitoreo (wearable, activity monitor, camera, RFID, bluetooth) y variables utilizadas (acelerómetro, heart rate, skin temperature, location).

2.1.1.2 Criterios de inclusión.

Los artículos de interés debieron cumplir los siguientes criterios: (1) estar en idioma inglés o español. (2) haber sido publicado a partir del año 2006, incluyendo los publicados en ese año. (3) describir un sistema capaz de reconocer comportamientos sedentarios de manera objetiva.

2.1.1.3 Proceso de identificación de artículos relevantes.

Antes de ejecutar la búsqueda en las bases de datos electrónicas, en cada una de estas se configuraron las opciones de encontrar solamente artículos en idioma inglés o español y que hayan sido publicados desde el año 2006. Con lo anterior se aseguró que los artículos resultantes al ejecutar la búsqueda ya cumplieran con los criterios de inclusión 1 y 2. Luego de ejecutar la búsqueda en cada una de las tres bases de datos electrónicas, el primer paso para la identificación de artículos relevantes fue leer uno a uno su título y abstract y verificar si se cumplía el criterio de inclusión número 3. Los artículos potencialmente relevantes se obtuvieron para una lectura completa y así determinar el cumplimiento del tercer criterio de inclusión.

2.1.1.4 Resultados.

Con la estrategia de búsqueda planteada se obtuvo 5293 artículos, de los cuales fueron seleccionados 590 después de considerar su título y posteriormente 68 luego de revisar el abstract. Estos fueron descargados y leídos para verificar el cumplimiento del tercer criterio de inclusión. En consecuencia, 17 fueron descartados, es decir, 51 aprobaron los criterios de inclusión para esta revisión. Además, se agregó 1 artículo proveniente de fuentes externas, que cumple los criterios de inclusión planteados. Por lo tanto, los artículos incluidos en la revisión sistemática fueron 52. El diagrama de flujo de lo anteriormente descrito se puede observar en la Figura 2.

2.1.1.4.1 Comportamientos sedentarios reconocidos.

Como fue descrito en el capítulo introductorio, la definición de comportamiento sedentario (en adelante CS) tomada para la realización de esta tesis es la propuesta por La Red de Investigación del Comportamiento Sedentario (RICS): “cualquier

actividad realizada por el individuo en posición sentada o inclinada con un gasto energético menor o igual a 1.5 METs, mientras se está despierto" [33]. Esta red ha solicitado a los editores de diferentes revistas científicas que sea esa la definición estándar que se acepte debido a las diferencias conceptuales que se pueden encontrar fácilmente en numerosos artículos hasta la fecha. Un ejemplo de ello son algunos de los artículos incluidos en la presente revisión sistemática, donde toman como CS actividades de la vida diaria que se realizan estando de pie.

Partiendo de la definición de CS, existen dos tipos de CS generales: permanecer sentado o inclinado. La postura que implica estar sentado es clara, pero permanecer inclinado involucra dos posibles posturas corporales: estar acostado (postura corporal totalmente horizontal) o estar reclinado (postura corporal separada de una postura vertical u horizontal). La Tabla 3 muestra los CS generales y la Tabla 4 muestran los CS específicos que han sido clasificados en los estudios incluidos en la revisión.

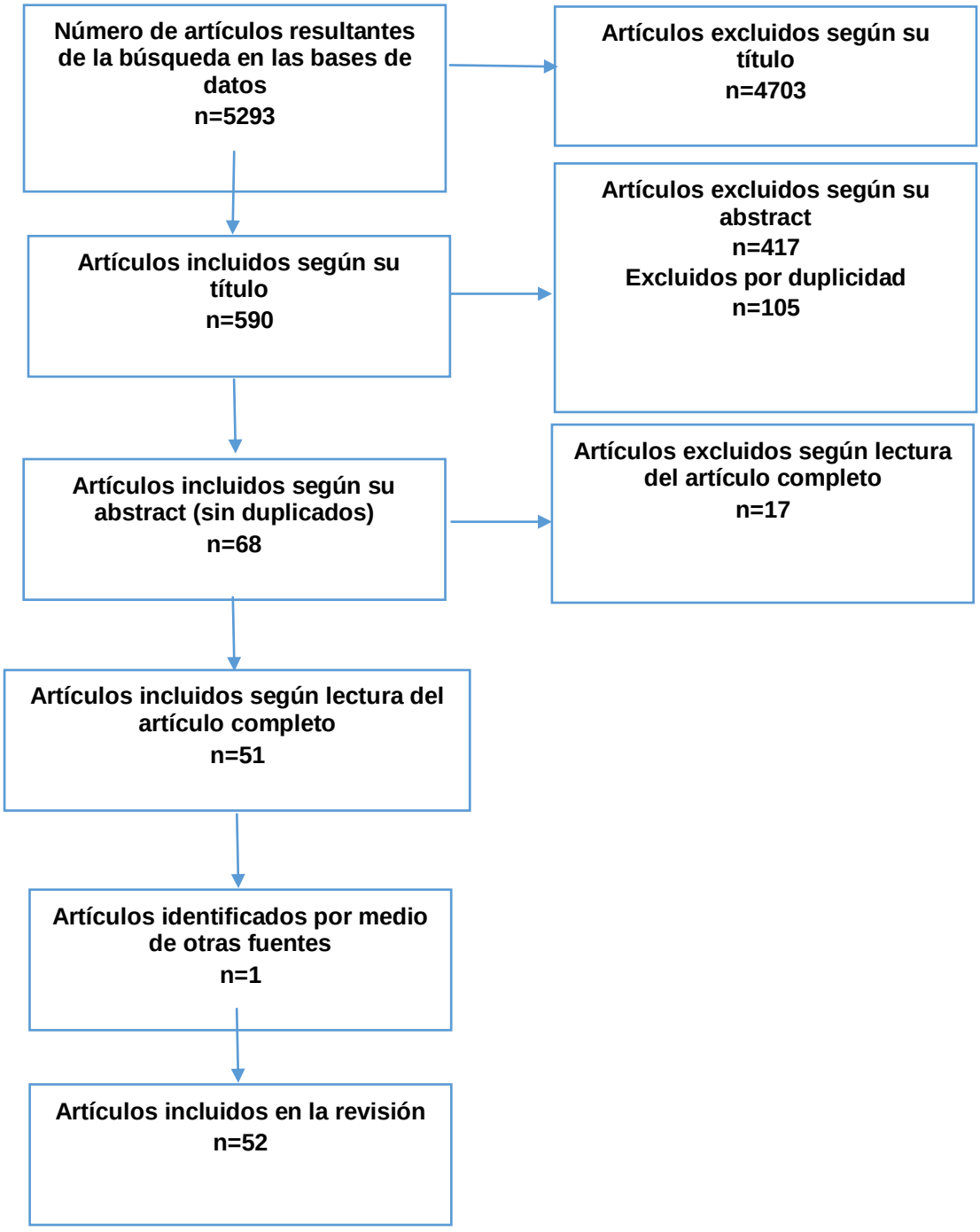

Figura 2. Diagrama de flujo del proceso de selección de artículos

CS comunes de inactividad estudiados	Número de artículos que incluyen clasificación de cada CS	Referencia
Acostado	26	[34]–[59]
Sentado	48	[34]–[48], [50]–[57], [59]–[83]
Inclinado	3	[25][35][40]
(Parado)	49	[34]–[49], [51]–[64], [66]–[84]

Tabla 3. Comportamientos sedentarios generales reconocidos por los sistemas descritos en los artículos seleccionados

CS Específicos	Referencia	CS Específicos	Referencia
Trabajando en el computador (sentado)	[49], [58], [59], [65]	Viajar en bus	[50], [85]
Organizar papeles (sentado)	[58]	Sentado en el sofá	[56]
Inclinado en una silla	[35], [45], [50]	Sentado en la cama	[56]
Jugar en el computador (sentado)	[50]	Sentado en el baño	[56]
Escribir a mano (sentado)	[50]	Acostado en la cama	[56]
Ver TV (sentado)	[65]	Acostado en el sofá	[56]
Jugar videojuegos (sentado)	[65]	Sentado en oficina	[50]
Leer (sentado)	[65]	Sentado en la cafetería	[50]
Conducir automóvil (sentado)	[50],[85]	Sentado en el baño	[50]

Tabla 4. Comportamientos sedentarios específicos reconocidos por los sistemas descritos en los artículos seleccionados

2.1.1.4.2 Participantes.

2.1.1.4.3 La cantidad y edad de los participantes involucrados en los artículos incluidos en la revisión sistemática se detallan en la Tabla 5 y

2.1.1.4.4

Tabla *6* respectivamente.

Intervalo de cantidad de participantes	N° de Artículos	Referencias
1-10	15	[38], [43], [48], [51], [52], [55], [57], [59], [66], [67], [76], [78], [82]–[84]
11-20	13	[34], [41], [42], [47], [53], [56], [60], [62], [63], [69], [71], [74], [85]
21-30	8	[37], [39], [44], [45], [49], [54], [68], [72]
31-40	4	[36], [40], [61], [64]
41-50	5	[46], [58], [77], [80], [81]
51-100	5	[35], [70], [73], [75], [79]

101-500	2	[50], [64]

Tabla 5. Cantidad de participantes involucrados en los artículos incluidos en la revisión sistemática

Intervalos de Edad	N° de Artículos	Referencias
Niño 1-11 años	8	[42], [50]–[52], [65], [70], [77], [79]
Adolescente 12-19	10	[35], [42], [44], [46], [50], [51], [65], [72], [77], [79]
Juventud 20-39	22	[35]–[37], [44], [45], [47], [49], [51], [53]–[55], [57]–[60], [66], [68], [71], [75], [78], [80], [85]
Adultez 40-64	14	[34], [35], [44], [45], [53], [54], [66], [71], [73], [74], [76], [80], [82], [85]
Vejez >65	14	[34], [36], [38]–[40], [43]–[45], [54], [73], [74], [76], [80], [82]

Tabla 6. Edad de los participantes involucrados en los artículos incluidos en la revisión sistemática.

2.1.1.4.5 Variables-Dispositivos utilizados.

Esta sección brinda información acerca de las variables, sensores y dispositivos asociados en los artículos incluidos en la revisión al realizar la tarea de clasificación de CS. La Tabla 7 muestra los dispositivos utilizados y las variables que estos capturan para realizar la tarea de clasificación de comportamientos sedentarios. Entre las variables se consideran las fisiológicas y las variables contextuales. Las variables fisiológicas se refieren a información sobre respuestas o comportamientos producidos en el organismo biológico estudiado ante situaciones específicas. Estas pueden revelar las condiciones de salud de una persona, capturar acciones corporales e incluso reaccionar ante emociones. Mientras que una variable de contexto contiene información característica de la circunstancia en la que ocurre una situación o evento [59].

Dispositivos	Variables	Ubicación	Referencias
ActivPAL	Acelerómetro 1, 2 y 3 ejes, Inclinómetro (depende del modelo)	Muslo	[40], [42], [46], [57], [61], [63], [67], [70], [72], [73], [77], [81], [83]
ActiGraph	Acelerómetro 1, 2 y 3 ejes (depende del modelo)	Cintura	[50], [58], [65], [69], [71], [72], [77], [79], [80]
Smartphone	Acelerómetro, Giroscopio, Barómetro, entre otros,	Bolsillo	[36], [37], [48], [60], [64], [66],

	(depende del modelo)		[84]
SmartShoe	Presión (fuerza), acelerómetro	Pie	[55],[60], [62], [74], [76]
IDEEA	Acelerómetro	Muslos, pie y pecho	[45], [54], [69]
Movemonitor Dynaport	Acelerómetro	Cintura	[39], [44]
Actical	Acelerómetro	Cintura	[58], [69]
Cosmed K4B2	Intercambio pulmonar de gases, saturación de oxígeno	Rostro	[59], [72]
Physilog 10D Silver GaitUP	Presión barométrica, acelerómetro, giroscopio	Muslo y pie	[34]
OpenBeacon RFID	Localización con RFID	No corporal	[61]
Fitbit Tracker	Acelerómetro	Muñeca	[69]
MotionLogs	Acelerómetro	Cadera	[35]
DirectLife	Acelerómetro	Muñeca	[69]
Oxycon mobile	Intercambio pulmonar de gases, saturación de oxigeno	Rostro	[50]
PAL2 (Gorman ProMed Pty. Ltd)	Acelerómetro	Arriba y abajo de la rodilla	[41]
Equivital EQ-01, Hidal- go	Frecuencia cardiaca y respiratoria	Pecho	[68]
ADXRS300	Acelerómetro, giroscopio	Pecho	[78]
SenseCam	Acelerómetro, temperatura ambiente, nivel de luz, video/imágenes, infrarrojo pasivo	Pecho	[85]
e-AR sensor	Acelerómetro	Oreja	[49]
Nokia wireless motion band	Acelerómetro	Tobillo	[51]
Tracmor	Acelerómetro	Espalda baja	[53]
Freescale MMA7260Q	Acelerómetro	Pecho y cintura	[55]
M92962	Acelerómetro	Muslos	[82]
M92961	Acelerómetro	Pecho	[82]
Analog Devices ADXL202	Acelerómetro	Pecho	[57]
AMP-331	Acelerómetro	Tobillo	[58]
Shimmer	Acelerómetro	Pecho y muslo	[59]
Zephyr Sensor	Frecuencia cardiaca, respiratoria y temperatura de la piel	Pecho	[59]
Body Media	Temperatura ambiente,	Brazo	[59]

	temperatura de la piel, GSR

Tabla 7. Dispositivos utilizados para realizar la clasificación de comportamientos sedentarios

2.1.1.4.6 Datasets

Fruto de la revisión resulta que solamente un artículo pone a disposición pública su dataset [64], aunque solamente integra un CS general: estar sentado. El "Wireless Sensor Data Mining" recolecta datos del acelerómetro y el GPS disponibles en un smartphone. Incluso proponen trabajos futuros para aprovechar otros sensores disponibles en el teléfono.

2.1.1.5 Discusión.

La Tabla 3 muestra que de los sistemas descritos en los 52 artículos incluidos en la revisión sistemática, el 94.23% (n=48) reconoce estar sentado, el 51.92% (n=26) permanecer acostado y tan solo 5.77% (n=3) reconoce estar inclinado. Como se puede observar en la Tabla 4, 10 artículos describen sistemas que son capaces de reconocer CS específicos, es decir, no tan solo reconocen la postura corporal de la persona sino que describen con más detalle el CS que está siendo realizado. Aunque son muchos los dispositivos que se han utilizado para realizar la clasificación de algunos CS, el sensor más común en ellos es el acelerómetro, como se puede comprobar en la Tabla 7.

Los artículos [44], [59] están enfocados en clasificar algunos CS por medio del cálculo del gasto energético (energy expenditure). Este cálculo se realiza con un sistema portable como por ejemplo el Cosmed K4B2, el cual calcula los METs gastados al realizar cada actividad de interés. En contraste con los anteriores tres estudios, en [50], [58], [65] son buscados puntos de corte utilizando solamente las lecturas de un acelerómetro para clasificar algunos CS. Los artículos [60], [62], [73], [80] son un esfuerzo paulatino para construir un calzado especial con múltiples sensores. El sistema descrito proporciona monitoreo de presión en puntos clave de soporte del peso corporal y es capaz de diferenciar posturas estáticas (sentado, parado). Además, intenta reducir la carga computacional y el uso de memoria para poder implementar los algoritmos en un Smartphone en tiempo real. En realidad, dentro de la metodología describen varias actividades físicas y CS para ser clasificados. Sin embargo finalmente cada una de ellos es agrupado en un solo CS general: permanecer sentado. En cuanto a transmisión, los datos de los sensores en este estudio fueron enviados en tiempo real a través de Bluetooth a un Smartphone. Así mismo, [66] diseña y prueba calzado con sensores de fuerza en el pie y el GPS. La novedad de este sistema es que gracias al GPS pueden reconocerse actividades como (sentado, sentado en un autobús o automóvil, así como también de pie en ambos contextos).

El sistema presentado en [61] es el único que ha utilizado un sistema de localización para entornos cerrados para la clasificación de algunos CS. La finalidad de ese

estudio es conocer las interacciones entre los empleados y el espacio físico de una empresa. Este sistema está compuesto de dos módulos: a) el monitor de actividad ActivPAL que detecta y registra los intervalos de tiempo en que se la persona está sentada, de pie o caminando, así como también registra el número de pasos y el número de transiciones de pie a sentado). b) el OpenBeaconSystem, descrito como un sistema de localización para entornos cerrados, que es integrado por tarjetas RFID ubicadas de manera fija en una zona de oficinas, otras que cargan los empleados y un lector RFID. Las variables utilizadas en este trabajo son las de un acelerómetro/inclinómetro dispuestos en el activPAL, además de variables de ubicación y proximidad del OpenBeaconSystem. Finalmente concluyen que el sistema de localización es muy eficiente y logra identificar con gran validez, en un conjunto de oficinas, la ubicación de las personas y algunas actividades desarrolladas por ellas. De forma similar [56] presenta un método para la clasificación de la actividad basado en el conocimiento de la localización y la identificación del usuario. El sistema está compuesto por tres módulos (detección de presión, detección de actividad y estación receptora) y reconocen en general 3 actividades que ellos llaman atómicas (acostado, sentado y parado) en tres lugares (el sofá, la cama, el baño y una silla). La localización es conocida gracias a una serie de sensores de contacto, los cuales detectan cuando los participantes se sientan o acuestan sobre ellos. La estación receptora es un computador que está en constante comunicación con los sensores de presión y los acelerómetros.

Respecto a la cantidad de participantes con los que los estudios han realizado sus experimentos, se ve en la Tabla 5 que en su mayoría han incluido menos de 30 y respecto a su edad (

Tabla *6*) se evidencia que muchos de ellos han estado enfocados solamente a un rango de edad específico, lo que daría lugar a que los modelos de clasificación obtenidos por ellos solo fueran aplicables a esa población.

2.1.1.6 Brechas detectadas

Teniendo en cuenta los resultados de la revisión sistemática, se obtienen las siguientes brechas que soportan el objetivo principal de este trabajo de investigación, ya que ninguno de los estudios encontrados lo ha abordado:

1. No han sido encontrados estudios en los cuales el tema de la clasificación de CS haya sido su principal objetivo.
2. Solamente un estudio ha usado datos de localización dados por un sistema de localización para entornos cerrados en conjunto con datos de aceleración para reconocer algunas actividades de los empleados de una empresa.
3. No ha sido posible encontrar un dataset de dominio público el cual posea datos de variables fisiológicas y/o contextuales que permitan realizar un proceso de minería de datos con el objetivo de lograr una clasificación automática de CS.

2.1.2 Objetivos del negocio

Habiendo obtenido el estado actual del conocimiento y detectado las brechas con base en él, muy acorde al objetivo general de este trabajo investigativo, el objetivo de negocio planteado es:
Desarrollar un sistema para la clasificación automática de comportamientos sedentarios en individuos basado en su localización en entornos cerrados.

2.1.3 Criterios de éxito del negocio

El único criterio de éxito del objetivo del negocio planteado es:
Presentación del Sistema para la Clasificación automática de Comportamientos Sedentarios (SCaCS).

2.2 Evaluación de la situación

Esta tarea implica la descripción más detallada sobre todos los recursos, restricciones, presunciones, y otros factores que deberían ser considerados en la determinación del objetivo de la minería de datos.

2.2.1 Inventario de Recursos.

2.2.1.1 Personal

El recurso humano fundamental para la realización del proyecto es el estudiante de maestría Jesús David Cerón Bravo. El estudiante está soportado en su director de tesis, el Doctor Diego López y en general por el grupo de ingeniería telemática (GIT) de la Universidad del Cauca, el cual está conformado por profesionales de alto nivel dispuestos a asesorar al estudiante. Además de esto, el programa de maestría en ingeniería telemática de la Universidad del Cauca tiene alianzas de cooperación con universidades y centros de investigación de orden nacional e internacional. En este caso particular, el estudiante de maestría realizó una estancia de un mes en el Instituto de Circuitos Integrados Fraunhofer, estancia en la cual tuvo experiencia con el grupo de sistemas de sensores biomédicos.

2.2.1.2 Datos

Una de las brechas detectadas fue la inexistencia de un dataset con el que se puedan realizar las siguientes fases del CRISP-DM. Se evidencia entonces la necesidad de construir uno propio.

2.2.1.3 Recursos computacionales

Al iniciar el proyecto se dispone del conjunto de wearables listados en la Tabla 8, además se cuenta con dos smartphones (LG G3, Huawei P7), un computador personal (Asus K555U con procesador core i7 de sexta generación y 8 GB de ram) y un servidor (Dell Power Edge T320).

Wearables	Tipo
Microsoft Band	Manilla
Microsoft Band 2	Manilla
Pebble Classic	Manilla
Misfit Flash	Manilla
MiFit	Manilla
Fitbit Zip	Manilla
e-Health Sensor Platform	Banda para usar en el pecho
Samsung Gear	Manilla

Tabla 8. Wearables disponibles.

2.2.2 Requisitos, supuestos y restricciones.

2.2.2.1 Requisitos

A continuación se listan los requisitos enfocados en el proceso de minería de datos:

1. Es fundamental tener un dataset para lograr el objetivo del negocio planteado, así que la recolección de este es un requisito esencial. Esto implica los siguientes sub-requisitos:
    a. Plantear una arquitectura de recolección de datos que sea compatible con la arquitectura final del SCaCS.
    b. Obtener un grupo determinado de personas quienes voluntariamente participen en la recolección de datos. Para ello deben firmar un consentimiento informado.
    c. Seleccionar un escenario de recolección de datos. Como se verá en el Capítulo 3, existe tres opciones: natural, semi-natural y laboratorio.
    d. Reservar tiempo adecuado para la recolección de los datos y llevar a cabo la recolección.
2. El SCaCS en entornos cerrados propuesto debe ser un sistema apto para ser usado en un contexto real, como por ejemplo en el proyecto de investigación SIMETIC. Por ello debe ser lo menos intrusivo posible y de fácil uso.
3. El tiempo máximo planteado para la culminación del presente proyecto es de 9 meses contados a partir de la fecha de aprobación del anteproyecto de grado.

2.2.2.2 Supuestos

Acorde a los requisitos previamente descritos, al inicio del proyecto es asumido lo siguiente:

1. El dataset elaborado será un dataset de calidad.
2. Es posible encontrar dispositivos que permitan obtener los datos en crudo de sus sensores (al menos de los datos de aceleración).
3. La capacidad de procesamiento del computador o servidor disponibles será suficiente para llevar a buen término la fase de modelado.
4. El SCaCS podrá ser usado por cualquier tipo de persona en un entorno real.

2.2.2.3 Limitaciones

La principal limitante radica en la no existencia de un dataset que contenga datos de aceleración y localización de la persona cuando realiza diferentes CS, por lo que se recolectará uno. Por otra parte, la cantidad de recursos económicos es limitada, pero como se describió en la sección de inventario de recursos, para este proyecto se tiene disponibilidad de varios dispositivos, así que un criterio para la elección de los dispositivos a usar en adelante será su disponibilidad previa.

2.2.3 Riesgos y contingencias.

A continuación se presentan los eventos que podrían retardar o hacer fallar el proyecto y sus respectivas acciones de contingencia:

1. El número de dispositivos disponibles impiden realizar la recolección de datos de manera simultánea con varias personas: se debe asignar un cronograma de recolección de datos riguroso con el fin de recolectar datos de la cantidad de personas requerido y en un tiempo previamente establecido.
2. La calidad del dataset recolectado no es buena: no debe suceder este riesgo debido a que el proceso de recolección de datos debe integrar una evaluación de la calidad de los datos in situ.
3. Pérdida de los datos recolectados: se debe asegurar un almacenamiento redundante de los datos recolectados.
4. La capacidad de procesamiento del computador no es suficiente para el proceso de modelado: en este caso se dispone con un servidor el cual posee mucha más capacidad de procesamiento en caso de que no sea suficiente la del computador portátil.
5. Ningún modelo obtenido es capaz de reconocer los CS recolectados en el dataset: se cree que clasificar este tipo de comportamientos es una tarea compleja, así que se deben tener en cuenta las variables-sensores recolectadas, la frecuencia de muestreo, el lugar del cuerpo en el que son recolectadas y la precisión del sistema de localización para entornos cerrados empleado. En ese sentido, es importante recolectar datos desde diferentes partes del cuerpo y de la mayor cantidad de sensores que se alojen en los dispositivos usados para tal fin. Con esto se podría experimentar haciendo la

selección de las variables que aporten más a una buena clasificación de los CS.

2.2.3.1.1 Terminología

En esta tarea se deben explicar los términos empleados en el proyecto con los que el cliente se tiene que familiarizar. Para efectos de este trabajo investigativo, los términos utilizados corresponden a dos grandes componentes: términos relevantes acerca del campo de investigación del comportamiento sedentario y acerca del proceso de minería de datos. En ambos casos, cada término de importancia es presentado en su momento oportuno.

2.3 Determinar los objetivos de la minería de datos

A diferencia del objetivo del negocio, los objetivos de la minería de datos se describen en términos técnicos enfocados específicamente en la minería de datos.

2.3.1 Objetivos de la minería de datos

Reconociendo que el tipo de problema de minería de datos abordado es de clasificación, el objetivo de la minería de datos establecido es:
Clasificar un conjunto de CS con base en datos de variables fisiológicas y de localización en entornos cerrados.

2.3.2 Criterios de éxito de la minería de datos

Para proyectos de minería de datos que involucran clasificación, como se verá en el capítulo 6, existen diversas métricas que indican la bondad de la clasificación realizada. En este proyecto se tomará la exactitud como métrica principal, por lo que el criterio de éxito de la minería de datos será:
Lograr un nivel de exactitud igual o mayor a 80% en la clasificación de los CS elegidos.

2.4 Realizar el plan del proyecto.

Al final de la fase 1, un plan de proyecto previsto debe ser elaborado. Este plan debe contener los pasos a ser ejecutados en las siguientes fases del proceso, y

corresponde a un plan de proyecto dinámico, pues al final de cada fase se deberá revisar a modo checklist y evaluar si es necesario modificarlo.

2.4.1 Plan de proyecto

La ejecución de todas las tareas correspondientes a cada fase siguiente es detalladamente descrita en su momento, por lo tanto, un plan general del proyecto se presenta en la Tabla 9.

Fase	Pasos	Duración	Recursos	Entradas	Salidas
2	Recolección de un dataset	3 meses	Humano: Participantes para recolectar datos. Computacionales: Smartphones, wearables, Bluetooth beacons, entornos de programación	Aplicación para la recolección de datos	Dataset (datos en crudo)
3	Preparar el dataset para aplicar técnicas de clasificación	1 mes	Un computador o servidor. Entornos de programación	Dataset con datos en crudo. Aplicación para la transformación de los datos	Dataset transformado
4	Generación y evaluación de los modelos de clasificación	2 meses	Un computador o servidor. Una herramienta para la generación de modelos de clasificación. Entornos de programación	Dataset transformado. Aplicación que agilice la generación	Modelos de clasificación y los resultados de su evaluación (métricas seleccionadas).
5	Análisis de los resultados de la evaluación de los modelos	1 mes	Un computador.	Resultados de evaluación de los modelos. (Matthews correlation coefficient y exactitud)	Elección del o los modelos adecuados para dar cumplimiento al objetivo de minería de

					datos
6	Implementación del o los modelos seleccionados en el SCaCS	2 mes	Computador, servidor smartphones, wearables, Bluetooth beacons, entornos de programación.	Modelos seleccionados	SCaCS final.

Tabla 9. Plan general del proyecto.

2.4.2 Evaluación inicial de herramientas y técnicas

De la herramienta para minería de datos seleccionada (herramientas software usadas para implementar algoritmos de aprendizaje automático) dependen las acciones que se deben ejecutar en las tareas propias de las siguientes fases, por esa razón, al finalizar la fase de entendimiento del negocio, es fundamental su elección.

Existen diferentes herramientas que pueden usarse para implementar algoritmos de aprendizaje automático. Prácticamente para cada lenguaje de programación existe al menos una. Algunas de ellas son de libre uso y otras de pago. Entre las de pago se encuentran SPSS, MATLAB, Microsoft Azure entre muchas otras. De libre uso se destacan actualmente R, Scikit-learn y WEKA. Hay disponibles gran variedad de componentes para implementar algoritmos de aprendizaje automático en el lenguaje de programación R. El Scikit-learn es una biblioteca escrita en Python. WEKA está escrita en Java. La elección de WEKA como la herramienta para minería de datos en este proyecto responde al conocimiento previo del lenguaje de programación Java, además que WEKA provee una librería que puede ser importada en cualquier tipo de aplicación Java, incluyendo una aplicación móvil Android y un servicio web que como se verá posteriormente son partes esenciales del SCaCS propuesto.

Capítulo 3

3 Fase 2: Entendimiento de los datos

La Tabla 10 muestra las tareas de la fase de entendimiento de los datos, las cuales abarcan desde la recolección de los datos hasta la verificación de su calidad. Esta fase, junto a la siguiente son las que más demandan tiempo y esfuerzo en el proceso de la minería de datos.

Entendimiento de los datos
Recolectar los datos iniciales
Reporte de recolección de los datos
Descripción de los datos
Reporte de descripción de datos
Exploración de los datos
Reporte de exploración de los datos
Verificar la calidad de los datos
Reporte de calidad de los datos

Tabla 10. Tareas de la fase de entendimiento de los datos

3.1 Recolectar los datos iniciales.

Como fue encontrado en la sección 2.1.1 (Fondo) son numerosos los dispositivos que han sido utilizados para realizar la clasificación de los CS generales. De hecho ninguno de los estudios revisados tiene como foco la clasificación de CS sino de actividad física o actividades de la vida diaria (ADL, por sus siglas en inglés: Activities of Daily Living). Es posible categorizar estos dispositivos en dos grandes grupos; wearables y Smartphones. Los wearables son dispositivos que como su nombre lo indica, son vestidos por las personas y que deben ser tan poco intrusivos como sea posible. Los wearables comerciales más comunes son tipo manilla, tales como la FitBit y la ActiGraph las cuales poseen sensores como acelerómetro y giroscopio entre otros. Otro tipo de wearable son los que se colocan en el muslo o la cadera como el ActivPAL o el Shimmer, las cámaras de video como la SenseCam la cual se coloca en el pecho de la persona con una banda, y soluciones no tan comerciales como las plantillas para los zapatos las cuales integran sensores de presión. Por otra parte se tiene al Smartphone, que debido a su gran presencia en la cultura actual, donde cada persona lleva consigo uno y el gran despliegue de sensores integrados en él, se ha convertido en un dispositivo que ha llamado la atención de muchos investigadores en el tema de clasificación de actividad física y actividades de la vida diaria. En ese sentido, y más aun teniendo en cuenta que no existe un dataset que incluya datos que permita la clasificación de CS, es indispensable obtener un dataset propio. Como fue comentado en la sección 2.2.3 (riesgos y contingencias) existe un riesgo de no obtener niveles de precisión aceptables al clasificar los CS debido a la elección de las partes del cuerpo en las cuales son recolectados los datos. Piense por ejemplo en estos dos CS: estar sentado en el escritorio trabajando en el computador y estar sentado en el escritorio leyendo un libro. Es fácil de suponer que clasificar de manera correcta estos CS con datos recolectados con un dispositivo puesto en la cintura será bastante complejo. Quizá con datos tomados desde la muñeca de la persona la tarea de clasificación sea algo más sencilla. Entonces, para minimizar ese riesgo, se plantea que el dataset a construir obtenga datos recolectados desde diferentes lugares del cuerpo.

A continuación se presenta el reporte de recolección de los datos, empezando con la selección de los CS que serán clasificados por el sistema, detallando luego la selección de los dispositivos para conformar la arquitectura para la recolección de datos y finalizando con un resumen de los datos recolectados.

3.1.1 Reporte de recolección de los datos.

3.1.1.1 Selección de los comportamientos sedentarios a ser reconocidos por el sistema.

Antes de describir la arquitectura para la recolección de los datos, es necesario elegir los CS que se desea que el sistema reconozca de manera automática. Existe una taxonomía desarrollada en consenso por 69 expertos en el tema de CS [86], en la

cual proponen 9 facetas con las que se puede caracterizar de manera detallada cada posible CS (ver Figura 3).

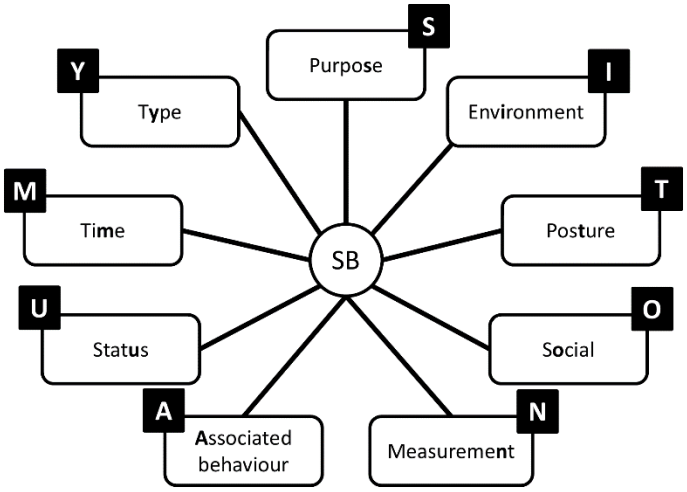

Figura 3. Taxonomía de comportamientos sedentarios, sus facetas y etiquetas de codificación

Las facetas más importantes según los 69 expertos son, en su orden: El propósito por el cual sucede el CS , el entorno, el tipo, la postura, el contexto social, la hora del día o estación climática en la que ocurre, los comportamientos asociados, el estado de la persona y el método de medición. A continuación se pueden observar las nueve facetas, sus correspondientes sub-dominios, categorías y las etiquetas de codificación que representan cada uno.

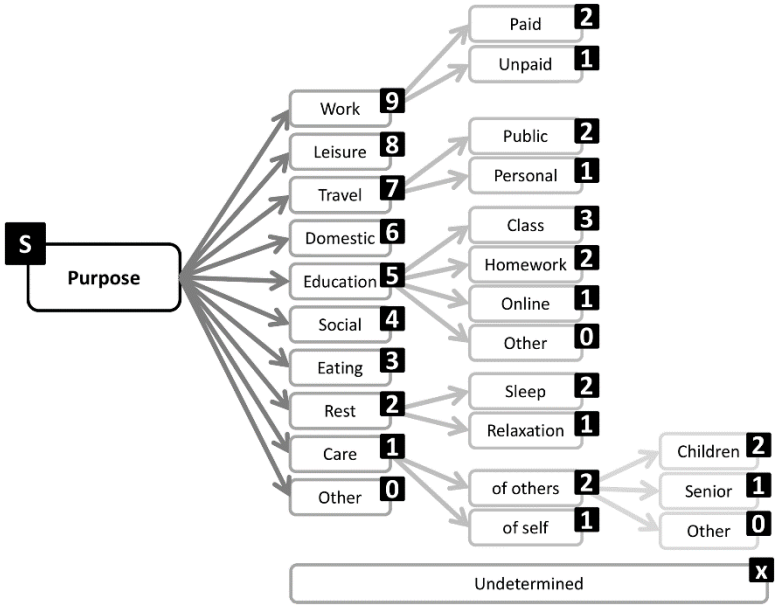

Figura 4. Faceta 1: Propósito

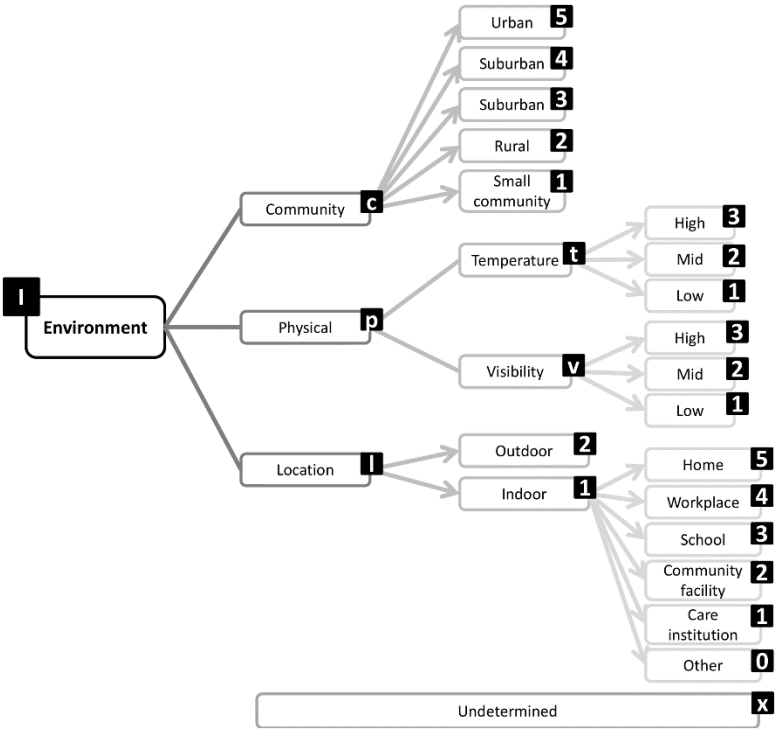

Figura 5. Faceta 2: Entorno

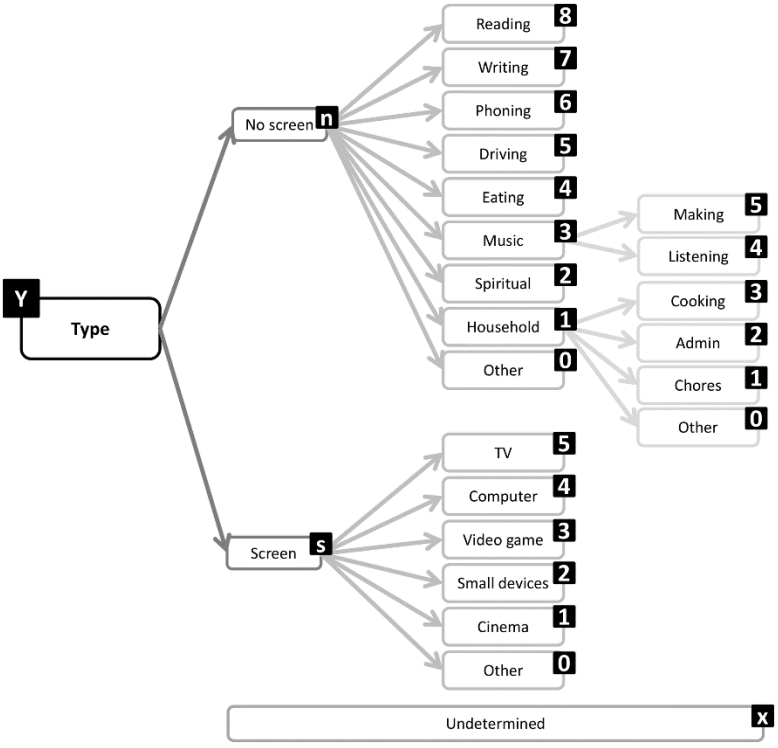

Figura 6. Faceta 9: Tipo

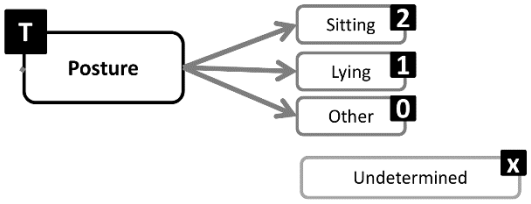

Figura 7. Faceta 3: Postura

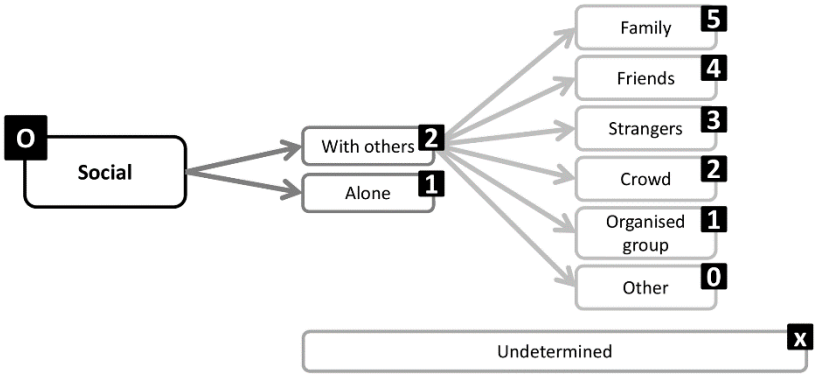

Figura 8. Faceta 4: Social

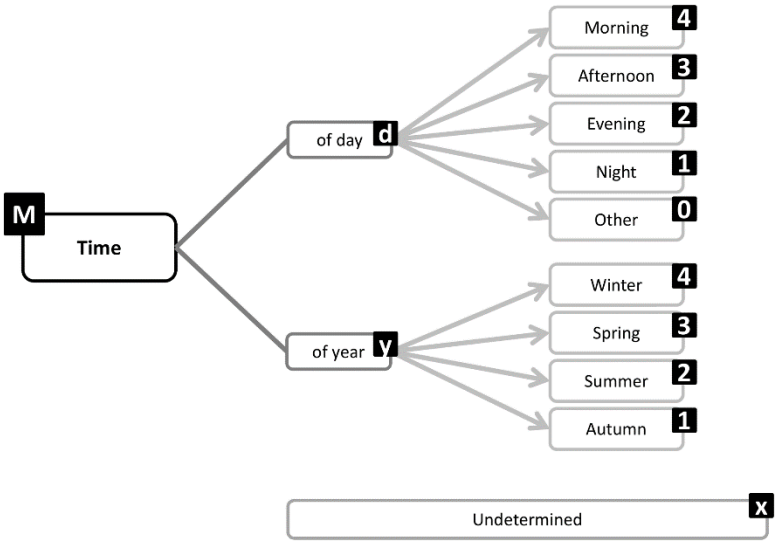

Figura 9. Faceta 8: Tiempo

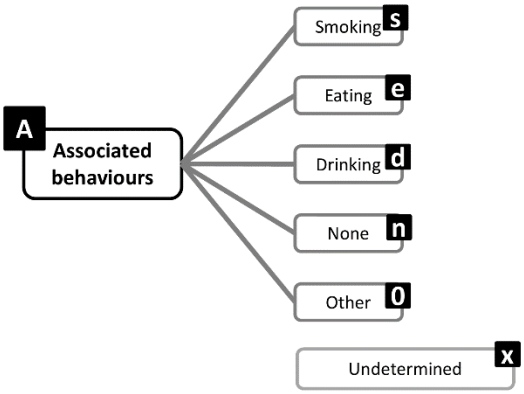

Figura 10. Faceta 6: Comportamientos asociados

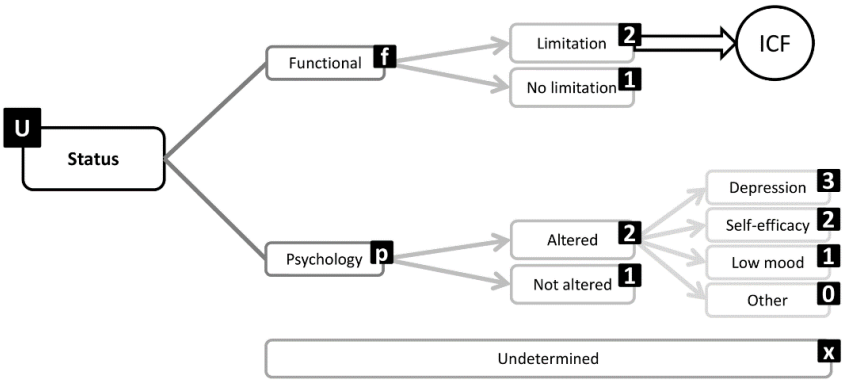

Figura 11. Faceta 7: Estado

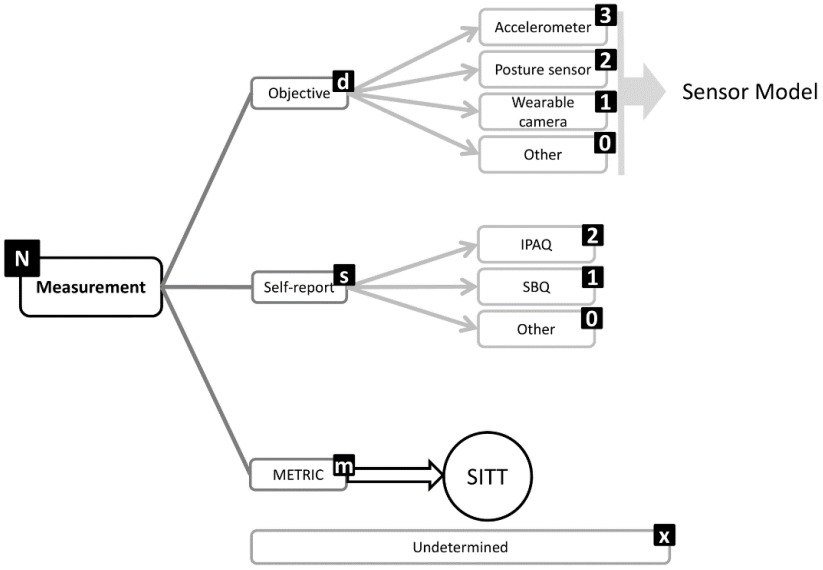

Figura 12. Faceta 5: Medición

Luego de analizar en detalle la taxonomía y considerando que en la actualidad realizar un CS está fuertemente influenciado por el uso de dispositivos electrónicos como los smartphones, computadores y televisores, son seleccionados los 23 CS incluidos en la Tabla 11. El SCaCs deberá incluir información acerca de la hora en la que sucede cada CS y el método de medición (el sistema dará una medición objetiva) para así describir cada CS de manera más específica al incluir información acerca de esas dos facetas restantes según la taxonomía.

El SCaCS debe poderse usar en un entorno real, y sabemos que realizar una clasificación objetiva de CS requiere emplear métodos de clasificación, los cuales son de naturaleza probabilística, por ese motivo es necesario asegurar la completitud del dataset a construir. Por ello es necesario incluir dos actividades físicas básicas que pueden suceder en un entorno cerrado: permanecer de pie y caminar. Si no se incluyeran estas dos actividades, cuando una persona realice alguna de ellas, el SCaCS caería en un error, pues los modelos de clasificación habrían sido entrenados solo para clasificar CS.

Comportamiento sedentario	**Codificación del CS según la taxonomía de CS**
Sentado usando el computador en el escritorio	S8IL15T2Nd30Ys4
Sentado usando el computador en la cama	S8 IL15 T2 Nd30 Ys4
Reclinado usando el computador en el escritorio	S8 IL15 T0 Nd30 Ys4
Reclinado usando el computador en la cama	S8 IL15 T0 Nd30 Ys4
Sentado usando el teléfono en el escritorio	S8 IL15 T2 Nd30 Ys2
Sentado usando el teléfono en la cama	S8 IL15 T2 Nd30 Ys2
Sentado usando el teléfono en el sofá	S8 IL15 T2 Nd30 Ys2
Reclinado usando el teléfono en el escritorio	S8 IL15 T0 Nd30 Ys2
Reclinado usando el teléfono en la cama	S8 IL15 T0 Nd30 Ys2
Reclinado usando el teléfono en el sofá	S8 IL15 T0 Nd30 Ys2
Acostado usando el teléfono en la cama	S8 IL15 T1 Nd30 Ys2
Acostado usando el teléfono en el sofá	S8 IL15 T1 Nd30 Ys2
Sentado viendo televisión en el sofá	S8 IL15 T2 Nd30 Ys5
Reclinado viendo televisión en el sofá	S8 IL15 T0 Nd30 Ys5
Acostado viendo televisión en el sofá	S8 IL15 T1 Nd30 Ys5
Sentado en reposo en el escritorio	S8 IL15 T2 Nd30
Sentado en reposo en el sofá	S8 IL15 T2 Nd30
Sentado en reposo en la cama	S8 IL15 T2 Nd30
Reclinado en reposo escritorio	S8 IL15 T0 Nd30
Reclinado en reposo en el sofá	S8 IL15 T0 Nd30
Reclinado en reposo en la cama	S8 IL15 T0 Nd30
Acostado en reposo en el sofá	S8 IL15 T1 Nd30
Acostado en reposo en la cama	S8 IL15 T1 Nd30

Tabla 11. Comportamientos sedentarios que serán clasificados por el sistema.

3.1.1.2 Sistema de recolección de datos.

La obtención del dataset fue soportado por una tesis a nivel de pregrado para optar por el título de ingeniero en electrónica y telecomunicaciones en la Universidad del Cauca. Estuvo a cargo de los estudiantes Stibent Possos y Robinson Cruz, bajo la

dirección del autor del presente proyecto de maestría. Su trabajo fue satisfactorio, obteniendo la aprobación de sus tesis y la aceptación del artículo "*Open Dataset for the Automatic Recognition of Sedentary Behaviors."* En la conferencia PHealth 2017 a llevarse a cabo en el mes de mayo de 2017. El artículo será publicado en los proceedings "*Studies in health technology and informatics*." Esta revista es reconocida en Colombia como una revista de calidad, siendo homologada como categoría B según publindex de Colciencias. A continuación se presenta el proceso seguido para la obtención del dataset.

Para obtener la arquitectura del sistema de recolección de datos fue utilizado el modelo de vistas de arquitectura 4+1 [87]. A continuación se presentan las consideraciones y selección de los dispositivos del sistema de recolección de datos para luego presentar la arquitectura final.

La arquitectura del sistema de recolección de datos será la base de la arquitectura final del sistema a proponer en este trabajo, ya que los modelos obtenidos en la fase de modelado corresponden a los datos con los cuales hayan sido entrenados, y estos a su vez dependen de cómo fueron recolectados (lugar del cuerpo en el que fueron recolectados, frecuencia de recolección y localización de la persona en el entorno cerrado). En ese sentido, son planteadas las siguientes consideraciones:

a) Es importante recolectar datos desde un dispositivo puesto de manera fija en la cintura de las personas. Esto debido a que en diversos estudios enfocados en la clasificación de actividad física han encontrado que usando los datos de aceleración obtenidos desde la cintura, la clasificación de posturas es más exacta que usando datos de aceleración obtenidos desde la muñeca, o el muslo [88]–[90].
b) Siguiendo la tendencia de muchos trabajos actuales acerca de la clasificación de actividad física en los cuales solamente usan el smartphone personal, se deben recolectar datos desde este dispositivo para realizar la clasificación de los CS. Sería una solución viable, ya que reduciría costos por el hecho de no necesitar un dispositivo 'extra' como un wearable.
c) Recolectar datos desde la muñeca con un wearable puede ser útil para clasificar entre CS que ocurren en un mismo lugar del entorno cerrado y que requieren la detección del movimiento de las manos.
d) La frecuencia de recolección de los datos debe permitir la detección de patrones que describan los CS.
e) El sistema de localización para entornos cerrados deberá evitar el uso de elementos o dispositivos extra que tengan que ser cargados por los usuarios.

Teniendo en cuenta estas consideraciones iniciales a continuación se describe la elección de dispositivos y su integración en un sistema para la recolección del dataset.

3.1.1.2.1 Smartphones

Para la elección del Smartphone a utilizar se plantean los siguientes requerimientos:

a. Debe contener el sistema operativo Android, ya que este sistema operativo es el más usado alrededor del mundo [91], con un porcentaje de participación en el mercado de 86.8% en el tercer cuarto del 2016. Esto permite tener un mayor alcance en cuanto a la cantidad de personas que podrán usar el sistema final.
b. Como se verá en la siguiente sección, las manillas inteligentes del mercado utilizan Bluetooth Low Energy (BLE) para enlazarse con un Smartphone, por lo que es necesario un Smartphone con sistema operativo Android superior al 4.4.
c. Debe contener una capacidad de memoria interna suficiente para soportar las necesidades de almacenamiento requeridas en la recolección de datos o tener la posibilidad para insertar una memoria externa que logre cumplir los requerimientos.
d. Debe contener la mayor cantidad de sensores posible, entre los cuales debe estar incluido el acelerómetro.

Se dispone de dos Smartphones que cumplen con los anteriores requisitos: un LG G3 y un Huawei Ascend P7.

3.1.1.2.2 Dispositivos wearables comerciales

Los requisitos que debe cumplir el wearable elegido son:

a. Ser de tipo manilla.
b. Brindar acceso a los datos que capture.
c. Tener la mayor cantidad de sensores posible.

Para la elección de la manilla, fue realizada una búsqueda de manillas comerciales en la web. La Tabla 12 muestra los sensores que incluye cada una, así como cuáles de los sensores incluidos están disponibles o no para los desarrolladores con el fin de obtener sus datos. Todos los wearables listados en la Tabla 12 tienen en común la tecnología de comunicación con la cual se conectan al Smartphone para obtener los datos recolectados: BLE.

CARACTERÍSTICAS / DISPOSITIVO		Sensores disponibles para desarrolladores											
	Frecuencia cardiaca	Acelerómetro	Giroscopio	Temperatura de la piel	Respuesta galvánica de la	Radiación UV	Altímetro	Barómetro	Luz ambiente	GPS	Inclinómetro	Magnetómetro	Temperatura/Humedad ambiente
Microsoft Band	X	X	X	X		X							
Microsoft Band 2	X	X	X	X	X	X	X	X					
Fitbit Charge HR	X												
Fitbit Surge	X								X				
Apple Watch	X	X	X				X					X	
Pebble		X							X			X	
Garmin	X						X	X		X			
Samsung Gear S	X	X	X			X		X	X		X	X	X
Jawbone UP3	X												

Tabla 12. Wearables comerciales.

Se puede observar que los wearables con mayor número de sensores son el Samsung Gear S y la Microsoft Band 2. Debido a que uno de los dispositivos disponibles en nuestro laboratorio es la Microsoft Band 2, esta fue la manilla elegida.

3.1.1.2.3 Sistema de localización para entornos cerrados.

Hay diversas tecnologías con las cuales se puede implementar un sistema de este tipo. Entre las más conocidas están: WIFI, BLE y RFID. La elección entre estas tecnologías depende básicamente en el contexto en el que se usará y su costo total de implementación. Para efectos de este trabajo, se parte del hecho de que no es necesario obtener una localización física precisa de la persona en el entorno cerrado, ya que es suficiente obtener una localización simbólica. Este tipo de localización, a diferencia de la localización física, la cual está dada por coordenadas, solamente da una referencia del lugar en el cual la persona está, por ejemplo: en el sofá, en la cama o en la cocina [92]. Por otro lado, al analizar el costo de implementación del sistema de localización para entornos cerrados y teniendo en mente la consideración (e) presentada al inicio de la sección 3.1.1.2, se debe considerar que la localización de la persona se puede realizar mediante la localización de su smartphone o wearable seleccionados, de esa manera se reducen costos de implementación y se evita intrusividad, ya que por ejemplo, al usar RFID se necesitan lectores y etiquetas activas o pasivas (la persona debería cargar un lector RFID o una etiqueta activa, según el enfoque preferido), mientras usando BLE es el propio smartphone que trabaja como lector y los dispositivos tipo beacons como puntos de referencia que trabajan como si fueran una etiqueta RFID activa. Por estas razones se decide utilizar beacons BLE para implementar el sistema de localización para entornos

cerrados. Un beacon BLE es un dispositivo que utiliza BLE para emitir una señal de radio de tipo broadcast, por lo que puede ser detectada por cualquier dispositivo que soporte BLE. Los lectores de estas señales usualmente son los smartphones, los cuales obtienen el indicador de fuerza de señal de recepción (en inglés RSSI; received signal strength indication) con el que es posible deducir la distancia o proximidad entre el smartphone y un beacon. Cada beacon tiene un identificador único, así que, al funcionar varios de ellos en un área común, es posible diferenciarlos. El único requerimiento al usar beacons es una aplicación móvil que reciba su señal y accione los eventos deseados.

3.1.1.2.4 Sensores disponibles en los dispositivos seleccionados.

La Tabla 13 muestra las variables que pueden ser recolectadas desde los sensores de cada dispositivo. En este punto es valioso mencionar que existen dos clases de sensores a los cuales los dispositivos dan acceso para obtener sus datos: los basados en hardware y los basados en software. Las lecturas de los sensores basados en software son calculadas tomando las lecturas de uno o varios de los sensores basados en hardware. Por ejemplo, la rotación vectorial es calculada a partir de los datos del acelerómetro, giroscopio y magnetómetro (este proceso es conocido como fusión de datos).

El sensor ‘Movimiento corporal significativo’, es usualmente usado para disparar eventos en una aplicación Android cuando es detectado un movimiento corporal fuerte. Básicamente este sensor evalúa si los valores del sensor de aceleración sobrepasan un umbral establecido. Una funcionalidad soportada en ese sensor es la detección de caídas. Pruebas preliminares con este sensor dan como resultado que detecta movimientos no muy bruscos del smartphone, por ejemplo cuando se éste está siendo manipulado o cuando está siendo cargado en un automóvil. Por lo tanto, no es un sensor de utilidad para ser incluido en el dataset.

Los sensores ‘Detección de pasos’ y ‘Contador de pasos’, los cuales indican si la persona está caminando y cuántos pasos ha dado respectivamente, a diferencia de los demás sensores los cuales realizan una fusión de los datos en crudo tomados por diferentes sensores basados en hardware, son el resultado de la implementación de algoritmos para la detección de caminata. La inclusión de los datos provenientes de estos sensores implicaría obligatoriamente considerar la precisión de sus algoritmos, lo que influiría directamente en la precisión final del SCaCS ya que, como se verá más adelante en el desarrollo de esta monografía, la precisión final del SCaCS es obtenida de la concatenación de las precisiones de los algoritmo que hayan sido empleados. Por esta razón, los datos de estos sensores no son incluidos en el dataset.

Sensor	Dispositivo	Tipo de sensor	Tipo de dato
Acelerómetro 3D (ejes X, Y ,Z)	Manilla	Hardware	double
Altímetro	Manilla	Software	float
Luz ambiente	Manilla	Hardware	int
Índice ultravioleta	Manilla	Hardware	float
Presión de aire	Manilla	Hardware	double
Temperatura ambiental	Manilla	Hardware	double
Respuesta galvánica de la piel	Manilla	Hardware	int
Giroscopio 3D (ejes X, Y ,Z)	Manilla	Hardware	double
Frecuencia cardíaca	Manilla	Hardware	int
Calidad de frecuencia cardiaca	Manilla	Software	Dicotómico
Variabilidad de frecuencia cardíaca	Manilla	Software	double
Temperatura de la piel	Manilla	Hardware	double
Acelerómetro 3D (ejes X, Y ,Z)	Smartphones	Hardware	float
Giroscopio 3D (ejes X, Y ,Z)	Smartphones	Hardware	float
Campo magnético 3D (ejes X, Y ,Z)	Smartphones	Hardware	float
Gravedad ejes 3D (ejes X, Y ,Z)	Smartphones	Software	float
Aceleración lineal ejes 3D (ejes X, Y ,Z)	Smartphones	Software	float
Rotación vectorial ejes 3D (ejes X, Y ,Z)	Smartphones	Software	float
Movimiento significativo	Smartphones	Software	Dicotómico
Detección de pasos	Smartphones	Software	Dicotómico
Contador de pasos	Smartphones	Software	int
Barómetro ejes 3D (ejes X, Y ,Z)	Smartphone principal	Hardware	float
RSSI	Beacons	NA	int

Tabla 13. Sensores disponibles para ser recolectar sus datos

Los sensores ‘Detección de pasos’ y ‘Contador de pasos’, los cuales indican si la persona está caminando y cuántos pasos ha dado respectivamente, a diferencia de los demás sensores los cuales realizan una fusión de los datos en crudo tomados por diferentes sensores basados en hardware, son el resultado de la implementación de algoritmos para la detección de caminata. La inclusión de los datos provenientes de estos sensores implicaría obligatoriamente considerar la precisión de sus algoritmos, lo que influiría directamente en la precisión final del SCaCS ya que, como se verá más adelante en el desarrollo de esta monografía, la precisión final del SCaCS es obtenida de la concatenación de las precisiones de los algoritmo que hayan sido empleados. Por esta razón, los datos de estos sensores no son incluidos en el dataset.

3.1.1.2.5 Arquitectura del sistema de recolección de datos.

La arquitectura empleada para la recolección del dataset es la mostrada en la *Figura 13. Arquitectura del sistema para la recolección de datos.*Figura 13. El smartphone LG G3 (en adelante smartphone principal), además de recolectar los datos de sus sensores, recibe los datos de la Microsoft Band 2 (en adelante la manilla) y obtiene el

RSSI respecto a la potencia de transmisión de los beacons que se encuentren en el rango de detección. Por su parte, el smartphone Huawei P7 (en adelante smartphone secundario) recolecta datos de sus sensores y también obtiene el RSSI de los beacons cercanos.

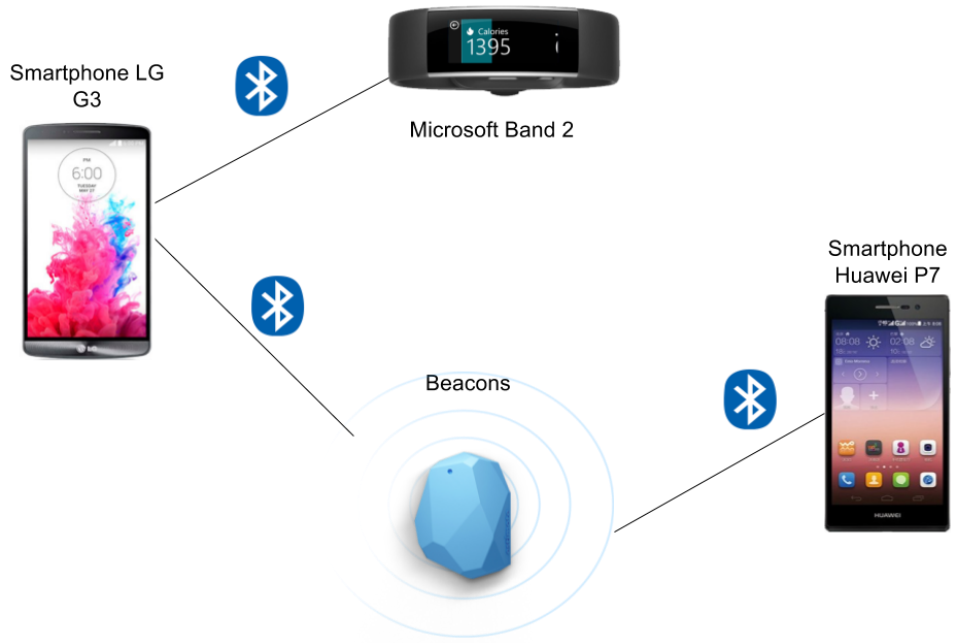

Figura 13. Arquitectura del sistema para la recolección de datos.

La *Figura 14. Posición de los dispositivos*Figura 14 muestra la forma cómo las personas deben ubicar los dispositivos en su cuerpo para la recolección de datos. La manilla se usa en la mano no dominante, el smartphone secundario en la cintura, en el mismo lado que su mano no dominante, con su cámara frontal apuntando hacia el cuerpo y la pantalla hacia afuera. El smartphone principal es cargado en el bolsillo del lado contrario al que se encuentra el smartphone secundario. Cuando un CS involucra el uso del smartphone (por ejemplo usar el smartphone en el sofá), las personas deben usar el smartphone principal.

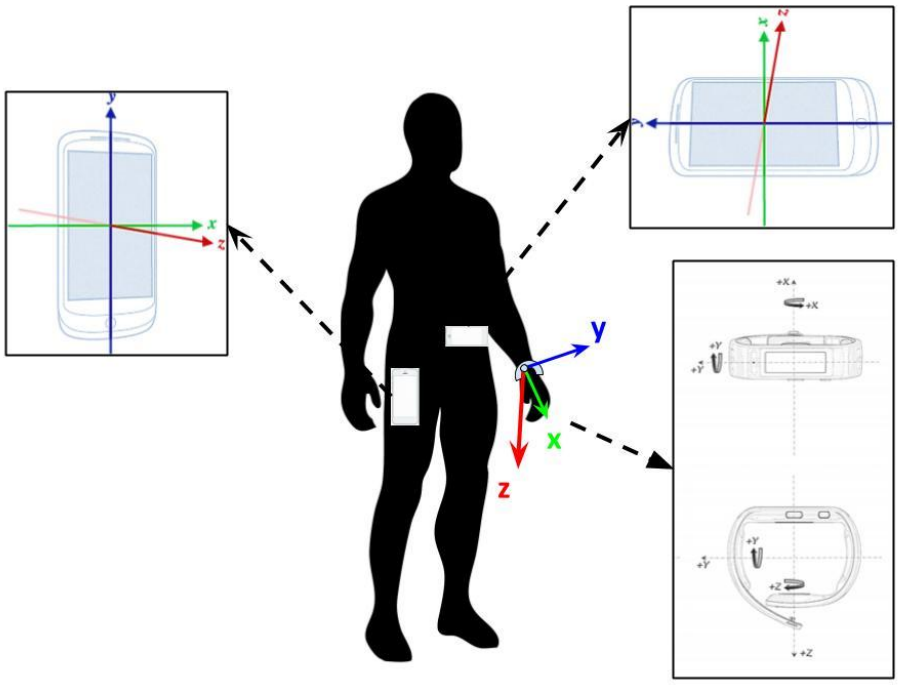

Figura 14. Posición de los dispositivos

Se disponen de seis BLE beacons marca Estimote, así que teniendo en cuenta los CS elegidos según la taxonomía de CS, estos deberán ser ubicados así: uno en la cama, otro en el escritorio y otro en el sofá de la sala de estar. Estos beacons deberán estar encendidos de manera continua y los llamaremos de ahora en adelante beacons de ubicación. Los tres beacons restantes indicarán el uso de dispositivos, por tanto, deberán estar encendidos únicamente cuando la persona esté utilizando el televisor, el computador o el smartphone. En adelante estos tres beacons serán nombrados como beacons de uso de dispositivos. Los beacons Estimote disponibles son pequeñas balizas que funcionan con una pila de botón, tal como se puede observar en la Figura 15. La potencia con la que emiten la señal fue configurada a -12dBm, logrando así un alcance máximo de 15 metros. Este tipo de beacons no posee un conector USB que permita su conexión a los dispositivos y así obtener la funcionalidad deseada, es decir, que se enciendan solamente cuando el dispositivo al que estén conectados esté encendido. Afortunadamente Estimote provee la funcionalidad de girar para apagar, que consiste en colocar el beacon hacia abajo para que se apague, de esa manera se obtiene el requerimiento deseado.

Figura 15. Beacon Estimote y funcionalidad flip to sleep

Finalmente, fue necesario desarrollar una aplicación móvil para dispositivos con sistema operativo Android la cual recolecta los datos de las variables descritas en la Tabla 13. Esta aplicación utiliza las clases que brindan los métodos para la recolección directa de los datos de los sensores necesarios y fue programada siguiendo las directrices de Clean Arquitecture, utilizando herramientas como Genymotion para realizar pruebas de desarrollo, DB Browser for SQLite para la elaboración de su base de datos, material design para la creación de vistas agradables y librerías como GreenDao, EventBus, entre otras. La Figura 16 muestra la interfaz de inicio, en la cual cada persona que participa en la recolección de datos debe leer un consentimiento informado donde acepta de manera voluntaria su participación.

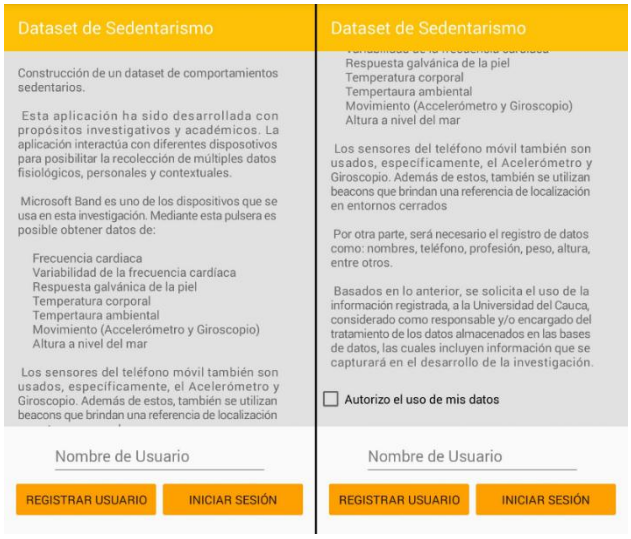

Figura 16.Interfaz de inicio

La Figura 17 muestra los datos requeridos para crear una nueva cuenta luego de aceptar participar.

Figura 17. Formulario de registro

La Figura 18 muestra la interfaz de selección de actividades, donde se encuentran los 23 CS y las dos actividades físicas a realizar (25 tareas en total).

Figura 18. Selección de actividades

Al seleccionar uno de los CS desplegados en la interfaz anterior, se muestra la interfaz de la Figura 19, que permite empezar la recolección de los datos. La aplicación móvil fue configurada para recolectar los datos con una frecuencia de 50 Hz, es decir, recolecta 50 muestras por segundo de cada variable capturada. Debido a que algunos sensores tienen una frecuencia menor, como por ejemplo el sensor de frecuencia cardiaca y temperatura de la piel (1Hz), los datos capturados de estos fueron replicados en los registros faltantes.

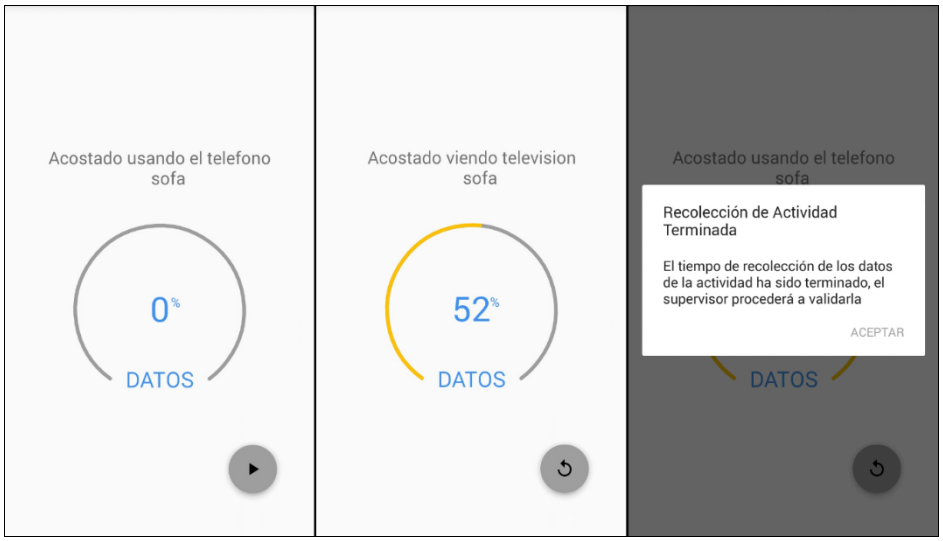

Figura 19. Captura de datos

Como se verá en la sección 3.4.1 (Reporte de calidad de datos), asegurar que exista la menor cantidad de datos perdidos en cada tarea de recolección de datos constituye un atributo de calidad del dataset. La aplicación muestra la interfaz de la Figura 20 al finalizar cada tarea de recolección, donde se puede observar el porcentaje de datos perdidos por cada sensor y decidir así si se repite o no la tarea previamente realizada.

Figura 20. Validación de datos

Finalmente, cuando la persona termina de realizar los 23 CS y las 2 actividades establecidas, se muestra la interfaz de la Figura 21.

Figura 21. Recolección de datos finalizada

3.1.1.3 Protocolo y recolección de datos.

En total, 30 participantes saludables (13 hombres y 17 mujeres) formaron parte de la recolección de datos. La mayoría de ellos radicados en la ciudad de Popayán, Colombia. La Tabla 14 provee más información acerca de los participantes.

	Edad (años)	**Peso (kg)**	**Altura (cm)**	**Cintura (cm)**
Promedio	44.167	67.433	162	88.8
Desviación Estándar	14.556	11.511	8.15	9.73
Máximo	73	91	174	108
Mínimo	20	47	146	65

Tabla 14. Características de los participantes

El método de recolección de datos fue semi-natural, ya que un asistente de investigación se dirigió a la casa de cada participante (algunos vivían en la misma casa) y les solicitó que colocaran ellos mismos los beacons de ubicación en los lugares establecidos: la cama, el escritorio y el sofá de la sala de estar. En el caso de los beacons de uso de dispositivos, se le indicó a cada participante girarlos como se indica en la anterior sección para que empezaran a emitir en los casos que fuera necesario. Cada participante realizó las 25 tareas. Cada tarea fue ejecutada por un lapso de tiempo de 2 minutos y medio, dando 40 segundos extra al inicio con el fin de que el participante ubique en su cuerpo los dispositivos y se acomode en la postura y lugar correctos.

3.2 Descripción de los datos

Finalizada la recolección del dataset, en esta sección son presentadas sus características en detalle.

3.2.1 Reporte de descripción de los datos

El tamaño del dataset adquirido es de 3.54 GB y está dividido en 30 carpetas, una para cada participante. En cada carpeta se alojan dos archivos en formato de texto, uno correspondiente al smartphone principal y otro al secundario. Ya que la frecuencia de recolección de los datos es de 50Hz y cada tarea es realizada por 2 minutos y medio, 7500 muestras (líneas) son obtenidas por cada tarea. Por lo tanto, cada archivo contiene 187500 muestras en total correspondientes a las 25 tareas.
Cada muestra de los archivos correspondientes al smartphone principal tiene 42 atributos divididos por comas:
- 1 Id (Id de cada persona)
- 2 Timestamp (ms)
- 3-16 Datos de la manilla
- 17-35 Datos Smartphone
- 36-41 Datos Beacons
- 42 Label (Id de la tarea realizada)

Los atributos recolectados desde la manilla (columnas 3 a 16) están organizados de la siguiente manera:

- 1-3 Acelerómetro 3D (m/s^2)
- 4 Altímetro (cm)
- 5 Luz ambiente (lux)
- 6 Presión del aire con barómetro (hPa)
- 7 Temperatura ambiente con barómetro (°C)
- 8 Respuesta galvánica de la piel (Kohms)
- 9-11 Giroscopio 3D (rad/s)
- 12 Frecuencia Cardiaca (latidos/min)
- 13 Intervalo RR (seg)
- 14 Temperatura de la piel (°C)

Los datos recolectados desde los smartphones están organizados de la siguiente manera:

- 1-3 Acelerómetro 3D (m/s^2) rango ±4g
- 4-5 Giroscopio 3D (rad/s) rango ±34.91
- 6-8 Magnético 3D (µT) rango ±4915.20

- 9-11 Gravedad 3D (m/s^2) rango ±1g
- 12-14 Aceleración lineal 3D (m/s^2) rango ±39.23
- 15-17 Rotación Vectorial 3D rango ±1
- 18 Barómetro (hPa) rango ±1100

Por último, ya que cada beacon representa un lugar u dispositivo diferente, sus datos se encuentran organizados de la siguiente manera:

- 1 Beacon Escritorio (RSSI)
- 2 Beacon Sofá (RSSI)
- 3 Beacon Televisor (RSSI)
- 4 Beacon Smartphone (RSSI)
- 5 Beacon Cama (RSSI)
- 6 Beacon Computador (RSSI)

Cada muestra incluida en los archivos correspondientes al smartphone secundario tiene 27 atributos divididos por comas: (A diferencia del smartphone principal, el secundario no recibe los datos recolectados desde la manilla y no posee barómetro)
- 1 Id
- 2 Timestamp (ms)
- 3-20 Datos Smartphone
- 21-26 Datos Beacons
- 27 Label (Id de la tarea realizada)

3.3 Exploración de los datos

Realizar una exploración de datos es necesario como antesala a la eventual preparación de los datos que se efectuará en la siguiente fase del CRISP-DM. En esta actividad son analizadas una a una las características de cada variable incluida en el dataset.

3.3.1 Reporte de exploración de los datos

La exploración de los datos fue realizada con la suite para data mining de WEKA [93]. Para ello, se siguieron los siguientes pasos:

1. Cada archivo del dataset fue configurado según el formato '.arff'.
2. Fue cargado archivo por archivo en WEKA para así poder visualizar las tablas de frecuencia, el valor máximo y mínimo, la media, la desviación estándar y la cantidad de valores perdidos de cada atributo (WEKA llama 'atributo' a cada variable de entrada).

Se encuentra que existen valores perdidos exclusivamente en los atributos relacionados a los beacons. Además se encuentra lo siguiente:

1. El valor del índice de luz ultravioleta es constante e igual a cero. Esto no se debe al mal funcionamiento del sensor sino al hecho de que todos los datos fueron recolectados en entornos cerrados.
2. El sensor 'Calidad de frecuencia cardiaca' indica cuándo las lecturas del sensor de frecuencia cardiaca están siendo tomadas correctamente. Este sensor indicó una correcta toma de la frecuencia cardiaca en todo el proceso de recolección de datos.

3.4 Verificar la calidad de los datos

Siendo esta la última tarea de la fase 2, luego de su cumplimiento se espera que el dataset se encuentre listo para ejecutar sobre él los futuros procesos propios de la fase 3 (Preparación de datos). En esta tarea se da respuesta a preguntas como: ¿Los datos cubren todos los casos requeridos?, ¿Los datos contienen errores?, ¿Hay valores perdidos? De ser necesario, se implementan los cambios o ajustes necesarios.

3.4.1 Reporte de calidad de los datos

Gracias a que el dataset usado en este proyecto investigativo fue recolectado en el contexto mismo del proyecto, se logran anticipar algunas estrategias para asegurar su calidad. A continuación se detallan los atributos de calidad del dataset.

3.4.1.1 Completitud

El dataset debería integrar datos de los CS que se desean clasificar, pero eso no sería suficiente para que los modelos de clasificación inducidos posteriormente puedan ser aplicables en un sistema que eventualmente será usado en un contexto real. Por esa razón, se incluyeron dos actividades físicas genéricas que suceden en cualquier entorno cerrado: estar de pie y caminar. La inclusión de estas dos tareas garantiza la completitud del dataset.

3.4.1.2 Existencia de errores en los datos

Se considera un dato erróneo cuando un dato perteneciente a un determinado atributo marca error o 'NONE' o si está fuera del rango de operación del sensor por el cual fue capturado o cuando su tipo es diferente, por ejemplo cuando un dato debe ser de tipo entero y es representado como una cadena de texto. En la tarea anterior, la de exploración de los datos, se usó WEKA para visualizar la tabla de frecuencia de cada atributo, comprobar sus máximos y mínimos, media, desviación estándar y

cantidad de valores perdidos. Fue requerido bastante tiempo para revisar atributo por atributo para cada uno de los archivos que integran el dataset, pero siendo ese un proceso netamente manual, pudo ocurrir algún error humano, así que teniendo como base los rangos de operación de cada uno de los sensores y el tipo de datos que manejan como muestra la Tabla 15, fue escrito un script en Python el cual realizó la búsqueda de errores de manera automática. Como resultado se obtiene que efectivamente no existe ningún error en los datos.

Sensores	Rango de operación	Valor Mínimo	Valor Máximo
Acelerómetro	±4g	-22,9652 m/s^2	37,3728 m/s^2
Giroscopio	±34,9066 rad/s	-11,6037 rad/s	11,3833 rad/s
Magnético	±4915,2002 uT	-1161,2244 µT	684,1827 µT
Gravedad	±9,8067 m/s^2	-9,7327 m/s^2	12,0730 m/s^2
Aceleración Lineal	±39,2266 m/s^2	-30,6367 m/s^2	30,9765 m/s^2
Rotación Vectorial	±1	-0,9996	0,9995
Barómetro	0-1100 hPa	822,2275 hPa	873,2682 hPa

Tabla 15. Rango de operación de sensores

3.4.1.3 Valores perdidos

Situaciones de datos perdidos se presentaron en las columnas que almacenan el RSSI de los beacons. Un valor perdido en este caso significa que no fue detectado el RSSI de un beacon que debería haber sido detectado. Por ejemplo cuando un participante está viendo televisión recostado en el sofá, los beacons correspondientes al TV y al sofá deben haber sido detectados durante toda la ejecución de dicho CS. Aquellos valores perdidos son causa ciertamente de errores humanos cometidos durante la recolección, por ejemplo al dejar los beacons volteados y no percatarse en la interfaz de validación.

3.4.1.3.1 Análisis de valores perdidos por los beacons de uso de dispositivos.

La cantidad de valores perdidos obtenidos en los atributos que contienen la información de los beacons asociados al uso de dispositivos son resumidos en la Tabla 16. Una de las métricas usadas para verificar la calidad en un conjunto de datos recolectados es la de relación simple (*simple ratio)*. Esta métrica considera el número de tareas (CS o actividades físicas) en las cuales se presentaron los valores perdidos, sobre la cantidad total de tareas realizadas durante la recolección de datos. Para realizar un análisis de los valores perdidos derivados de los beacons de uso de dispositivos, en la Tabla 16, el número total de tareas realizadas es el número de tareas realizadas que implican el uso del correspondiente dispositivo, por ejemplo en el caso del televisor, 3 tareas implican utilizarlo por participante, como son 30 los

participantes, 90 es el número de veces que la tarea fue realizada en total, pero como son 2 smartphones, se obtiene un total de 180 veces.

	Número de tareas con valores perdidos	Tareas con valores perdidos / número total de tareas	Simple ratio
Televisor	10	10/180	0.02777777777
Smartphone	6	6/480	0.00625
Computador	0	0/240	0

Tabla 16. Valores perdidos para los atributos de los beacons de uso de dispositivos

3.4.1.3.2 Análisis de valores perdidos por los beacons de ubicación.

Probablemente por la característica de girar para apagar que tienen los Estimote beacons, en 14 ocasiones pudo ser la causa de la generación de valores perdidos en los atributos relacionados al uso de los beacons de ubicación.

	Resultados no deseados	Resultados no deseados/ #total de tareas	Simple ratio
Escritorio	0	0/360	0
Sofá	2	2/540	0.00185185185
Cama	12	12/480	0.0125

Tabla 17. Valores perdidos para los atributos de los beacons de ubicación

En la Tabla 17 se evidencia que en 14 tareas hubo problemas para detectar los beacons de ubicación, doce de ellas tuvieron que ver con el beacon que se ubicaba en la cama.

En el próximo capítulo se describe en detalle la fase 3 de CRISP-DM de preparación de los datos, que como su nombre lo indica, es la fase encargada de preparar los datos para luego obtener modelos de aprendizaje con base en ellos.

Capítulo 4

4 Fase 3: Preparación de los datos

Hay diversos tipos de problemas de minería de datos: descripción y resumen de datos, segmentación, predicción, clasificación y descripción de concepto. No todas las técnicas de modelado son aplicables a un problema en particular, por lo tanto, esta fase se encuentra estrechamente relacionada con la fase siguiente, la fase de modelado, ya que dependiendo de las técnicas de modelado aplicables y elegidas, los datos deben ser preparados de diferentes maneras. De eso se ocupa esta fase de preparación de los datos.

Es claro que el problema de minería de datos que compete a este proyecto de investigación es de clasificación, donde lo que se pretende es obtener un sistema capaz de clasificar diferentes CS. Como insumo fundamental para abordar el problema de clasificación, se obtuvo el dataset presentado en la fase anterior. En la sección 3.2 fue descrito en detalle el dataset, donde se indica que cada muestra (línea del dataset) contiene varios atributos relacionados con una tarea específica (alguno de los 23 CS o 2 actividades físicas) por medio de un atributo de clase (en inglés: Class attribute). Algunas técnicas apropiadas para abordar un problema de clasificación son los análisis discriminantes, métodos de inducción de reglas, árboles de decisión, redes neuronales y vecinos más cercanos. Con el fin de aplicar esas técnicas con el dataset construido, este debe ser preparado, para ello, son realizadas las tareas que se muestran en la Tabla 18. Como resultado final del seguimiento de esta fase se presentará una descripción detallada del dataset completamente listo para ser utilizado en la fase de modelado.

Preparación de los datos
Seleccionar los datos
Inclusión/exclusión de datos
Limpiar los datos
Reporte de limpieza de los datos
Estructurar los datos
Derivación de atributos
Generación de registros
Integrar los datos
Unificación de datos
Formateo de los datos
Dar formato a los datos

Tabla 18. Tareas de la fase de preparación de los datos

4.1 Seleccionar los datos

La selección de datos implica dos pasos:

- La selección de un subconjunto representativo de datos en caso que el dataset sea demasiado grande para ser procesado posteriormente.
- Seleccionar los atributos de interés.

Se considera que el tamaño del dataset construido no es un elemento crítico para los posteriores procesos, así que se trabajará con la totalidad del dataset, el cual asegura completitud y fue depurado de errores y valores perdidos en la fase 2. Es entonces de particular atención la selección de atributos.

Todos los atributos que componen el dataset recolectado son de interés ya que podrían aportar, cada uno a su medida, en la correcta clasificación de CS. Por consiguiente, si se llegare a excluir algún atributo debería ser por algún tipo de inconsistencia que afecte el proceso para la obtención de un clasificador. En ese sentido, en la sección 3.3 (Exploración de los datos) se encontró que dos atributos describían un comportamiento constante: el valor del índice de luz ultravioleta y la calidad de frecuencia cardiaca. El índice de luz ultravioleta es constante e igual a cero en todo el dataset. La razón de esto es que todos los datos fueron recolectados en entornos cerrados, donde el nivel de luz ultravioleta es muy débil para ser detectado por el sensor de la manilla. Por otro lado, el atributo de calidad de frecuencia cardiaca indica cuándo las lecturas del sensor de frecuencia cardiaca están siendo tomadas correctamente. Este atributo indicó una correcta toma de la frecuencia cardiaca en todo el proceso de recolección de datos, por lo tanto es constante. Estos atributos, al ser constantes en todo el dataset no aportan ningún valor en un eventual proceso de modelado, por lo tanto es necesario excluirlos del dataset.

En conclusión, solamente dos atributos tuvieron que ser excluidos del dataset.

4.2 Limpiar los datos

El objetivo de esta actividad es idear decisiones y tomar las acciones que den lugar para abarcar los problemas reportados en la actividad de verificación de la calidad de los datos de la fase de entendimiento de los datos (sección 513.4). Las acciones tomadas para la corrección de los datos se describen a continuación, primero para el caso de los valores perdidos que están relacionados con los beacons de uso de dispositivos (a), seguido de los valores perdidos relacionados con los beacons de ubicación (b):

a) El número tareas en los que hubo valores perdidos fue 16. Cada caso fue revisado individualmente, y fue posible dar solución a cada uno de ellos de la siguiente manera: Si en una tarea existían valores perdidos aislados, es decir, solamente faltaban algunas muestras puntuales (faltaba el RSSI en una o varias líneas del dataset), se asignó el valor promedio del RSSI del resto del atributo. En caso que faltaran todos los datos del RSSI para un atributo, se asignaron los mismos datos del atributo faltante que fueron recolectados para una tarea similar, por ejemplo, si faltaban los datos del atributo correspondientes al beacon del TV para la tarea “Sentado viendo TV en el sofá”, se tomaron los datos del mismo atributo que fueron recolectados para la tarea “Recostado viendo TV en el sofá”. De modo que, en todos los casos la relación simple mostrada en la Tabla 16 pasa a ser cero, es decir, se corrigen todos los atributos con datos faltantes.
b) Empleando las mismas estrategias descritas para el manejo de los valores perdidos de la sección anterior, estos fueron completamente solucionados.

4.3 Estructurar, integrar y formatear los datos

¿Cómo facilitar el proceso de modelado usando transformación de atributos? Esta es una pregunta crítica que se debe realizar al empezar la actividad de estructuración de los datos. De hecho es de gran relevancia para proyectos de minería de datos con un enfoque de clasificación como este. Al respecto, en la literatura científica se pueden encontrar trabajos que pertenecen al campo de clasificación de actividad física o de actividades de la vida diaria en los cuales se vislumbran dos técnicas bien definidas. La primera y más común de ellas realizando un proceso de extracción de características (Feature extraction en inglés) con el que se hace una transformación de los datos en crudo incluidos en el dataset recolectado a un nuevo ‘dataset transformado’ que está compuesto de características que describen a los datos en crudo usualmente en términos de propiedades estadísticas tales como la media, la mediana y la desviación estándar y/o en términos de conceptos de energía usando por ejemplo la transformada de Fourier. La otra técnica empleada en algunos escasos trabajos es el uso de los datos en crudo directamente para inducir los modelos de clasificación. Esta última técnica tiene dos grandes desventajas: la

primera es que son necesarias técnicas de clasificación más complejas que puedan trabajar con base en los datos en crudo de los sensores, lo que implica un inminente procesamiento elevado de datos que eventualmente es un limitante en el contexto de la computación ubicua de vanguardia, en el cual los dispositivos wearables, incluidos los smartphones poseen limitaciones en cuanto a capacidad de procesamiento y la segunda desventaja es que hasta el momento no se ha demostrado que empleando los datos en crudo se obtengas mejores resultados de clasificación comparado con la técnica de transformación de los datos.

Como se expuso en la Figura 1, el ciclo de vida de la metodología CRISP-DM es iterativo, lo cual permite regresar desde la fase de modelado hacia la presente fase si es necesario. De hecho, en el transcurso de este proyecto se realizaron diversas iteraciones entre estas dos fases con el fin, por ejemplo, de generar diferentes características y probar sus aportes en la construcción de los modelos de clasificación. A continuación se detalla cómo fue el proceso completo de extracción de características, incluyendo el formateo de los datos.

En la sección 3.2 se describió el contenido del dataset, donde se menciona que para cada participante se tienen dos archivos de texto diferentes: Uno con los datos recolectados en el smartphone principal y otro con los datos recolectados en el smartphone secundario. Por razones expuestas en la siguiente sección, se decidió no integrar los datos de estos dos archivos en uno solo, por lo tanto, la transformación de los datos se realizó de manera separada para cada archivo.

El proceso de transformación de los datos consistió en dos pasos:

1. Segmentación: Los datos de cada archivo son divididos en segmentos de 5 segundos (250 muestras). Esta duración, conocida como ‘duración del ejemplo’ (en inglés: example duration) se determinó a partir de un experimento previo realizado en el contexto de este trabajo, en el cual se utilizó una duración de ejemplo de 10 y 5 segundos para la clasificación de algunos CS, obteniendo mejores resultados con esta última [94].
2. Extracción de características: cada segmento es convertido en un ejemplo (example en inglés); Las características extraídas, para todos los atributos, a excepción de los atributos relacionados a los beacons fueron: Promedio simple del valor de las 250 muestras, desviación estándar del valor de las 250 muestras, promedio de la diferencia absoluta entre el valor de cada una de las 250 lecturas y el valor promedio sobre esos 250 valores (Ecuación 1). Para los sensores que ofrecen lecturas en los 3 ejes del plano cartesiano (acelerómetro, giroscopio y magnetómetro) se calculó el promedio resultante de las raíces cuadradas de la suma de los valores de cada eje al cuadrado (Ecuación 2).

$$Promedio\ de\ diferencia\ absoluta = \frac{\sum_{i=0}^{n} |x_i - promedio(x)|}{n}$$

Ecuación 1. Diferencia absoluta media.

$$Promedio\ resultante = \frac{\sum_{i=0}^{n} \sqrt{(x_i^2 + y_i^2 + z_i^2)}}{n}$$

Ecuación 2. Promedio resultante

3. Para obtener la localización simbólica de la persona, a cada beacon le fue asignado un identificador numérico, como se muestra en la Tabla 19. Se calculó el promedio del RSSI proveniente de los beacons para cada segmento y según este las características extraídas fueron seis: ID de los beacons de ubicación del más cercano al más lejano e ID de los beacons de uso de dispositivos del más cercano al más lejano.

Beacon	Beacon ID
TV	1
PC	2
Smartphone	3
Cama	4
Sofá	5
Escritorio	6

Tabla 19. Identificadores numéricos de los beacons

Con la anterior estrategia se obtiene una localización simbólica de la persona dentro del entorno cerrado. Un ejemplo práctico para comprender mejor la estrategia se muestra en la Tabla 20. Allí se muestra que el beacon 1, es decir, el beacon de ubicación más cercano tiene ID 5, lo que indica que el beacon más cercano es el que está en el sofá. Por otra parte, el Beacon 4 corresponde al beacon de uso de dispositivo más cercano, en el caso del ejemplo tiene un ID igual a 1, lo que indica que está siendo usado el televisor. Con esto se podría inferir que la persona está en el sofá viendo televisión.

Beacon 1	Beacon 2	Beacon 3	Beacon 4	Beacon 5	Beacon 6
5	4	6	1	0	0

Tabla 20. Ejemplo Sistema de Localización Simbólica

Para realizar todo el proceso de extracción de características descrito se desarrolló una aplicación de escritorio en Java en el entorno de desarrollo NetBeans. Al finalizar el proceso de extracción de características de cada archivo del dataset, la aplicación genera un archivo en formato '.arff' listo para ser interpretado por WEKA en la fase de modelado. Para efectos de futuros proyectos, esta aplicación permite la configuración, por medio de interfaz gráfica, del tamaño del ejemplo.

4.4 Descripción del dataset transformado

En resumen, los archivos transformados correspondientes al smartphone principal contienen 113 características, 44 que pertenecen a los sensores de la manilla, 6 a los beacons y 63 a los sensores del smartphone como tal. Por su parte, los archivos correspondientes al smartphone secundario, tienen 66 características, 60 que pertenecen a los sensores del smartphone y 6 a los beacons. La estructura de los archivos arff se ve en la Tabla 21. La primera línea corresponde al nombre del dataset, las siguientes líneas corresponden a la declaración de atributos con su nombre y tipo, (en adelante las características serán llamadas como atributos para concordar con la terminología usada en WEKA). Luego se deben colocar los atributos de clase o etiquetas que se desean clasificar. Para el caso que nos compete, los atributos de clase son los 23 CS y las dos actividades físicas (estar de pie y caminar), las cuales en este paso son fusionadas en una sola clase llamada "no_sedentary_behavior"

```
@relation nombre_del_dataset

@attribute nombre_del_atributo_1 tipo
@attribute nombre_del_atributo_2 tipo
.
.
.
@attribute nombre_del_atributo_n tipo

@attribute class {clase_1, clase_2,…clase_n}
@data
atributo_1, atributo_2,…, atributo_n, clase_x;
```

Tabla 21. Estructura archivo ARFF

Debido a que cada archivo del dataset sin transformar tenía 7500 muestras por cada CS o actividad realizada, se generaron 30 ejemplos por cada actividad, es decir, cada archivo transformado contiene 750 ejemplos.

Realizada la preparación de los datos es obtenido un dataset transformado. Ese dataset es el finalmente empleado para la siguiente fase, la fase de modelado, descrita en el siguiente capítulo.

Capítulo 5

5 Fase 4: Modelado

En esta fase de CRISP-DM se seleccionan las técnicas de modelado adecuadas para el proyecto, se obtienen los modelos y finalmente se evalúan. Ya desde la fase 2 del proyecto se eligió la suite para minería de datos de WEKA para realizar el modelado y el dataset transformado que se obtuvo en la fase anterior está configurado específicamente para cumplir con ese propósito. Las tareas propias de la fase de modelado están descritas en la Tabla 22.

Modelado
Seleccionar la técnica de modelado
Técnica(s) seleccionadas
Supuestos de los modelos
Generar el plan de prueba
Plan de prueba
Construir los modelo
Configuración de parámetros
Modelos
Descripción de los modelos
Evaluar los modelo
Evaluar los modelos
Revisión de los parámetros

Tabla 22. Tareas de la fase de modelado

5.1 Seleccionar la técnica de modelado

Esta tarea ha sido prevista desde fases anteriores, de tal manera que se ha reconocido previamente que el tipo de problema de minería de datos que se está abordando es de clasificación, más aún, la solución a este problema estará enfocada en aprendizaje de máquina supervisado, para lo cual se cuenta con el atributo de clase que permite a las diferentes técnicas de clasificación entrenarse según los demás atributos para luego clasificar de manera autónoma, sin necesidad de atributo de clase, el CS que se esté realizando. Con esta introducción se entiende entonces que en este punto se tiene ya un dataset listo para realizar la búsqueda de modelos eficientes para la clasificación de los CS.

Ahora bien, existen diferentes técnicas de clasificación incluidas en WEKA, las cuales podrían ser utilizadas para inducir modelos. La pregunta ahora es ¿Cómo seleccionar la técnica adecuada? Pues bien, es posible indagar en la literatura científica acerca de las técnicas que han sido más exitosas en la clasificación de actividad física y/o actividades de la vida diaria, pero esto no aseguraría que esas técnicas funcionen para este nuevo campo de estudio llamado clasificación de CS. Por ese motivo, es interesante aplicar la mayor cantidad posible de técnicas de clasificación que provee WEKA y así obtener un perfil de las más adecuadas para este tema en particular. Las 7 técnicas de clasificación que brinda WEKA por defecto para afrontar problemas de aprendizaje de máquina supervisado están descritas en la Tabla 23. Para cada técnica de clasificación WEKA provee diferentes algoritmos. En total, en este trabajo de maestría se utilizarán los 23 algoritmos correspondientes a las 7 técnicas de clasificación. No es el objetivo de este trabajo dedicarse al ajuste de parámetros en cada algoritmo, por lo que inicialmente se emplearon los parámetros por defecto dados por WEKA. Con fines de reproducción de resultados, la configuración de parámetros por defecto para casa técnica está consignada en la Tabla 23.

Técnicas de clasificación		Configuración de parámetros por defecto
Rule-induction	Decisión table	-X 1 -S \"weka.attributeSelection.BestFirst -D 1 -N 5\"
	JRip	-F 3 -N 2.0 -O 2 -S 1
	OneR	-B 6
	PART	-M 2 -C 0.25 -Q 1
	ZeroR	
Decision trees	Decision Stump	
	Hoeffding Tree	-L 2 -S 1 -E 1.0E-7 -H 0.05 -M 0.01 -G 200.0 -N 0.0
	J48	-C 0.25 -M 2
	LMT	-I -1 -M 15 -W 0.0
	Random Forest	-P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
	Random Tree	-K 0 -M 1.0 -V 0.001 -S 1
	REPTree	-M 2 -V 0.001 -N 3 -S 1 -L -1 -I 0.0
Instance-based	IBk	-K 1 -W 0 -A \"weka.core.neighboursearch.LinearNNSearch -A \\\"weka.core.EuclideanDistance -R first-last\\\"\"
	KStar	-B 20 -M a
	LWL	-U 0 -K -1 -A \"weka.core.neighboursearch.LinearNNSearch -A \\\"weka.core.EuclideanDistance -R first-last\\\"\" -W weka.classifiers.trees.DecisionStump
Functions	Logistic Regression	-R 1.0E-8 -M -1 -num-decimal-places 4
	Simple logistic	-I 0 -M 500 -H 50 -W 0.0
Support vector machine	SMO	-C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K \"weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007\" -calibrator \"weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4\"
Neural Networks	Bayes Net	-D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5
	Naive Bayes	
	Multilayer perceptron	-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
Meta-algorithms	AdaBoostM1	-P 100 -S 1 -I 10 -W weka.classifiers.trees.DecisionStump
	AdaBoostM1 (J48 como clasificador base)	-P 100 -S 1 -I 10 -W weka.classifiers.trees.J48 -- -C 0.25 -M 2
	Bagging	-P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.trees.REPTree -- -M 2 -V 0.001 -N 3 -S 1 -L -1 -I 0.0
		-P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.trees.J48 -- -C 0.25 -M 2
	Bagging (J48 como clasificador base)	-P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.trees.J48 -- -C 0.25 -M 2

Tabla 23. Técnicas de clasificación utilizadas

5.2 Generación del plan de prueba

El objetivo de esta tarea es generar un procedimiento que permita evaluar la calidad y validez de los modelos generados.

5.2.1 Preparación para el plan de prueba

Antes de plantear el plan de prueba es necesario describir:

1. Los tres tipos de modelos que se pueden inducir según la partición de los datos en datos de entrenamiento y evaluación.
2. El enfoque de clasificación que se adoptará para realizar el SCaCS.

Adicionalmente, hay que tener en cuenta que debe ser construido un modelo de cada tipo para las tres diferentes fuentes de datos (smartphone principal llevado en el bolsillo y usado según el CS, manilla puesta en la muñeca y smartphone secundario cargado en la cintura).

5.2.1.1 Tipos de modelos

Tal como sucede en el campo de clasificación de actividad física, al abordar la tarea de clasificar CS es posible obtener tres tipos de modelos dependiendo la manera en que se haga la partición de los datos en datos de entrenamiento y evaluación: modelo personal, híbrido e universal. Enseguida se da una descripción de cada uno de ellos.

5.2.1.1.1 Modelo personal

Descripción: Un modelo personal es aquel que ha sido construido a partir de ejemplos únicamente de la persona que lo va a utilizar. En consecuencia, este modelo es entrenado y evaluado con ejemplos de una misma persona. Claro está que los ejemplos de entrenamiento y evaluación tienen que ser diferentes, de no ser así, los resultados de evaluación del modelo estarían completamente sesgados.

Forma de particionado de los datos: Para cada uno de los 30 participantes se debe utilizar un archivo que contenga ejemplos únicamente de la persona en particular (tal como están preparados en el dataset transformado) y como método de evaluación se debe usar validación cruzada, en este caso se usará validación cruzada de 10 pliegues (10-fold cross validation), el cual asegura que los ejemplos con los cuales se entrena cada modelo sean diferentes a los que este es evaluado.

Posibles ventajas y desventajas: Una ventaja de este modelo es precisamente la personalización del modelo, lo que se traduce en que el modelo podría ser muy eficiente en la clasificación debido a que fue entrenado justamente con datos de la persona que lo utilizará. Desafortunadamente esta posible bondad se ve opacada al tener obligatoriamente que recolectar datos de la persona para entrenar su modelo.

Lo anterior implicaría que si se desea que el SCaCS reconozca los 23 CS abordados en este proyecto, cada persona que desee utilizarlo deberá recolectar previamente datos de cada CS por su propia cuenta.

5.2.1.1.2 Modelo universal

Descripción: Este tipo de modelo es entrenado con ejemplos de un conjunto de personas y es evaluado con ejemplos de una persona que no perteneció al conjunto de personas que lo entrenó.
Forma de particionado de datos: Se utilizará el método de dejar un participante fuera (Leave-One-User-Out) para evaluar el rendimiento de los modelos. Para esto es necesario tener dos archivos por participante, un archivo en el cual se incluyan los ejemplos de los demás 29 participantes, dejando fuera los datos del participante correspondiente y otro archivo que contiene los ejemplos del participante. Con el primer archivo se hace el entrenamiento del modelo y con el segundo se evalúa.
Posibles ventajas y desventajas: La gran bondad de este modelo es que puede ser utilizado por cualquier persona directamente sin necesidad de recolectar datos personales previamente como sucede con el modelo personal. Como desventaja se tiene que es posible que se necesiten datos de un número elevado de personas para obtener una exactitud adecuada en la clasificación de los CS.

5.2.1.1.3 Modelo híbrido

Descripción: Este tipo de modelo es una combinación entre un modelo personal y uno universal. Este modelo es entrenado con ejemplos de un conjunto de personas que incluye la persona que lo usará.
Forma de particionado de datos: Los ejemplos de los 30 participantes deben estar en un mismo archivo y se usará validación cruzada de 10 pliegues para la evaluación. Para obtener unos resultados confiables, se utilizará 10 diferentes semillas (seeds) para entrenar y generar 10 modelos híbridos diferentes y promediar su resultado.
Posibles ventajas y desventajas: Puede que el rendimiento de los modelos sea mayor en comparación con los anteriores tipos de modelos, ya que el modelo se construye con base en los ejemplos del participante que lo usará y adicionalmente tendrá ejemplos de los demás 29 participantes. Al igual que el modelo personal, este modelo necesita ejemplos de la persona que lo va a utilizar, así que se debe hacer una recolección de datos inicial para luego obtener el modelo, además se necesita la colaboración de posiblemente muchas más personas que previamente hayan contribuido a la construcción de un dataset previo como sucede con el modelo universal.

5.2.1.2 Enfoque de clasificación adoptado para la construcción del SCaCS

El enfoque de clasificación por capas será usado para la construcción del SCaCS. La idea principal de este enfoque es obtener un modelo de clasificación diferente en

cada capa, de tal manera que en cada capa se puedan utilizar diferentes técnicas de clasificación e incluso diferentes atributos con el objetivo de mejorar la precisión de la clasificación final. Además, este enfoque facilita la integración de nuevas capas si se llegare a necesitar, sin la necesidad de generar nuevos modelos. Un ejemplo del uso de este enfoque es el estudio realizado en [95], donde proponen un algoritmo en el que su primera capa clasifica tres actividades usando datos de un acelerómetro: estar de pie, sentado o caminando. Cuando la primera capa clasifica que la persona está caminando, la segunda capa entra en acción. Allí se clasifican tres actividades: estar caminando normalmente y subir o bajar escaleras. La clasificación en las dos capas se realiza con dos algoritmos de clasificación diferentes, y en la segunda capa, adicional a los datos del acelerómetro utilizados en la capa 1, son utilizadas las lecturas de un barómetro. Adoptando entonces un enfoque por capas para el SCaCS, queda abierta la posibilidad de escalar el sistema a uno que abarque la clasificación de diferentes tipos de actividad física y demás actividades de la vida diaria.

El algoritmo de dos capas propuesto es el mostrado en la Figura 22. La primer capa es la encargada de clasificar si la persona está realizando o no un CS con base en datos de los sensores. En la segunda capa, en caso de que en la primera se obtenga como resultado que la persona está realizando un CS, se clasifica cuál CS en específico está realizando. Una ventaja significativa que se puede deducir del uso de este enfoque es que el uso de los datos provenientes de los beacons es tenido en cuenta solamente por la capa 2. De esa forma, el modelo utilizado en la primera capa no estaría afectado por la falta de información de ubicación posiblemente en algún lugar que no haya beacons, es más, daría una clasificación básica del comportamiento de la persona: en CS o no. Por otra parte, este enfoque provee la cualidad de escalar el algoritmo de clasificación, ya que por ejemplo, si la capa 1 detecta que la persona no está realizando algún CS, significa que la persona está realizando algún tipo de actividad física, por lo cual podría incluirse una capa paralela a la capa 2 en la cual se clasifiquen de manera específica un conjunto de actividades físicas.

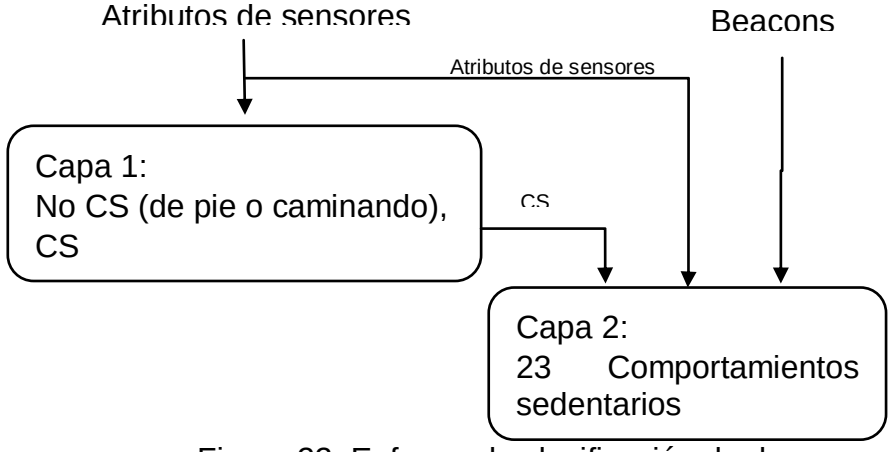

Figura 22. Enfoque de clasificación de dos capas

5.2.2 Plan de prueba

El plan de prueba contempla los siguientes pasos:

5.2.2.1 Partición de datasets de entrenamiento y evaluación respectivos

Los datos del dataset transformado deben ser ordenados en archivos según se requiera para la generación de los tres tipos de modelos para cada capa tal como fue descrito en la sección 5.2.1.1.

5.2.2.2 Generar los modelos

Con el ordenamiento de los datos realizado en el paso anterior, se debe realizar la construcción de los modelos de clasificación. Es de interés comparar el la exactitud obtenida por cada uno de los tres tipos de modelos (personal híbrido e universal), así como también comparar la exactitud de las técnicas de clasificación y de las tres fuentes de datos (smartphone principal, secundario y manilla).

5.2.2.3 Evaluar los modelos obtenidos.

Existe un diverso número de criterios de evaluación para evaluar un modelo. La elección de ellos debe ser acorde al tipo problema de minería de datos que se esté abordando y a las características del dataset empleado para ello. La base de los criterios de evaluación para proyectos de minería de datos referentes a tareas de clasificación es la descripción del resultado de clasificación de cada una de los ejemplos, teniendo cuatro categorías: verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN). Un ejemplo de la aplicación de estas categorías se puede realizar fácilmente en el caso de nuestra capa 1: si un ejemplo que originalmente está etiquetado como ‘CS’ y un algoritmo clasificador los clasifica como tal, eso corresponde a un TP, en caso contrario, si ese ejemplo es clasificado como ‘No CS’ se considerará como un FN. Si un ejemplo que originalmente está etiquetado como ‘No CS’ el algoritmo clasificador lo clasifica como tal, esto será un TN, en caso contrario se considerará un FP. Como se puede observar en las Ecuaciones 3 a 8, la taza de verdaderos positivos, verdaderos negativos, la exactitud (accuracy), precisión, recall, F-Measure, Mathews Correlation Coefficient (MCC), entre otros, son métricas de evaluación que utilizan los cuatro conceptos descritos.

$$TPR = \frac{\sum TP}{\sum TP + \sum FN}$$

Ecuación 3. Taza de verdaderos positivos

$$TNR = \frac{\sum TN}{\sum TN + \sum FP}$$

Ecuación 4. Taza de verdaderos negativos

$$Accuracy = \frac{\sum TP + \sum TN}{\sum TP + \sum TN + \sum FP + \sum FN}$$

Ecuación 5. Accuracy (exactitud)

$$Precision = \frac{\sum TP}{\sum TP + \sum FP}$$

Ecuación 6. Precisión

$$F\ measure = \frac{2 * \sum TP}{2 * \sum TP + \sum FP + \sum FN}$$

Ecuación 7. F- measure

$$MCC = \frac{(\sum TP * \sum TN) - (\sum FP * \sum FN)}{\sqrt{(\sum TP + \sum FP) * (\sum TP + \sum FN) * (\sum TN + \sum FN) * (\sum TN + \sum FN)}}$$

Ecuación 8. Matthews Correlation Coefficient

Debido a que en la primera capa hay un evidente desbalance de clases ya que la cantidad de ejemplos de la clase etiquetada como 'CS' es bastante superior a la clase etiquetada como 'No CS' esa capa será evaluada según los resultados del Mathews Correlation Coefficient (MCC). Esta métrica de evaluación ha probado ser muy robusta cuando suceden casos de desbalance de clases [96]. El rango de resultados del MCC va desde -1 hasta 1, siendo 1 el mejor valor, 0 indica una clasificación aleatoria y -1 una clasificación inversa. La segunda capa posee balance entre la cantidad de ejemplos para cada CS (30 ejemplos para cada CS), así que los modelos obtenidos para la capa 2 serán evaluados según su exactitud (accuracy en inglés). La exactitud será descrita en este trabajo como un porcentaje, donde el rango va de 0% a 100%, siendo 100% una exactitud perfecta.

La evaluación de los modelos se describe inicialmente en la sección 5.4 y se aborda ampliamente en la fase 5 del proceso de minería.

5.2.2.4 Analizar los modelos obtenidos

Ya con los resultados obtenidos en el paso anterior, es posible realizar un análisis profundo. Es de interés comparar el rendimiento de los tres tipos de modelos (personal híbrido e universal), así como también comparar el rendimiento de las técnicas de clasificación y de las tres fuentes de datos (smartphone principal, secundario y manilla). Estos análisis se realizan en la fase 5.

5.3 Construir los modelos

El primer paso del plan de prueba fue realizado de manera manual. Para dar cumplimiento al segundo paso, hay que considerar que para cada participante incluido en el dataset deben generarse 414 modelos, ya que son 3 los tipos de modelos, por 3 fuentes de datos, por 23 algoritmos de clasificación y por 2 capas. Como son 30 los participantes incluidos en el dataset, 12420 es el número total de modelos que se deben generar. Realizar esa tarea usando la interfaz gráfica de WEKA requeriría una cantidad de tiempo considerable debido en principio a la configuración manual y posteriormente al tiempo que tarda la obtención de un modelo para cada uno de los 23 algoritmos. Por esa razón se programó una aplicación de escritorio en Java utilizando el entorno NetBeans con la cual se agilizó dicho proceso. Para el desarrollo de esta aplicación fue necesaria la importación de la librería de WEKA al proyecto para poder utilizar las técnicas de modelado y demás recursos por medio de las clases y métodos que esta ofrece.

La estancia de investigación realizada durante un mes en el Instituto de Circuitos Integrados Fraunhofer en Erlangen, Alemania, permitió profundizar el conocimiento acerca del tema abordado en este trabajo. En dicha estancia se experimentó con wearables, específicamente con la camiseta AmbioTex, que incorpora una unidad central llamada TechUnit donde son alojados un acelerómetro, giroscopio y barómetro. Lo anterior permitió experimentar posteriormente, sobre un dataset inicial recolectado con 15 personas y utilizando algunas técnicas de selección de características incluidas en WEKA. Fruto de ello se obtuvo el artículo “Two-Layer Method for Sedentary Behaviors Classification Using Smartphone and Bluetooth Beacons” el cual a la fecha de abril de 2017 está en proceso de publicación. Es claro que es posible obtener infinidad de modelos para la clasificación de los CS debido a que se posee un dataset que contiene diversos atributos relacionados a varios sensores de los dispositivos con los que fue recolectado, pero con el fin de poder realizar un análisis comparativo de igual a igual entre los tipos de modelos y las fuente de datos respecto a la clasificación de CS, para la obtención de los modelos que serán inducidos se utilizaron únicamente los 10 atributos correspondientes al acelerómetro (promedio, desviación estándar, y promedio de diferencia absoluta para cada eje y promedio resultante.) y los 6 correspondientes a los beacons.

Finalmente, empleando la aplicación de escritorio descrita para la generación de modelos, el proceso completo de modelado y evaluación tomó 96 horas y 47 minutos. Los resultados consolidados se encuentran en el Anexo A. Los criterios de evaluación allí consignados para cada caso son: exactitud, taza de verdaderos positivos, taza de falsos positivos, precisión, recall, F-measure, Matthews Correlation Coefcient (MCC), Area Under Receiver Operating Characteristic (AUROC) y Area Under Precisioin-Recall Curve.

5.4 Evaluar los modelos

La evaluación inicial requerida en esta fase debe hacer una comprobación del criterio de éxito de la minería de datos con base en los resultados de los modelos obtenidos. El criterio de éxito planteado es: “Lograr un nivel de exactitud igual o mayor a 80% en la clasificación de los CS elegidos”. El primer paso para comprobar ese objetivo es la selección del algoritmo que provea mayor exactitud para cada tipo de modelo y fuente de datos. La Tabla 24 y la Tabla 25 muestran el resultado de dicha selección para cada capa. Cabe aclarar que los valores presentados en esas tablas para el caso del modelo personal y universal corresponden al promedio de los resultados obtenidos para cada uno de los 30 participantes incluidos en el dataset, y para el caso del modelo híbrido, el promedio de los 10 modelos obtenidos con 10 diferentes semillas de aleatorización. Se incluye el algoritmo que presentó mayor exactitud.

	Personal			**Hibrido**			**Universal**		
	MCC	Accuracy	Algoritmo	MCC	Accuracy	Algoritmo	MCC	Accuracy	Algoritmo
Smartphone secundario	0.9916	99.8755	MP	0.9536	99.3111	Ib1	0.7581	95.7155	RF
Smartphone principal	0.9849	99.7777	Ib1	0.9495	99.2613	RF	0.886	98.2444	RF
Manilla	0.9193	98.8577	RF	0.8688	98.1511	RF	0.7716	97.1778	RF

Tabla 24. Resultados de la evaluación de la capa 1.

	Personal		**Hibrido**		**Universal**	
	Accuracy	Algoritmo	Accuracy	Algoritmo	Accuracy	Algoritmo
Smartphone secundario	99.3671	RF	99.0126	RF	63.8261	ABJ48
Smartphone principal	98.7536	RF	96.7121	ABJ48	51.1642	RF
Manilla	97.0386	RF	92.9111	RF	52.087	BJ48

Tabla 25. Resultados de la evaluación de la capa 2.

Al ser los resultados de cada capa independientes, la exactitud total del algoritmo con enfoque de dos capas propuesto está dado por la Ecuación 9.

$$Exactitud_{total} = Exactitud_{capa\,1} * Exactitud_{capa\,2}$$

Ecuación 9. Ecuación para el cálculo de exactitud total

De esa forma, son calculadas las exactitudes totales incluidas en la Tabla 26.

	Personal	Hibrido	Universal
Smartphone secundario	99.2434	98.3305	61.0915
Smartphone principal	98.534	95.9976	50.266
Manilla	95.9301	91.1932	50.617

Tabla 26. Exactitud total algoritmo de dos capas

Una comprobación inicial del objetivo de minería de datos desde los resultados mostrados en las Tablas 24, 25 y 26 permite vislumbrar que es posible alcanzar el objetivo de minería de datos usando modelos personales o híbridos, porque la exactitud obtenida por los modelos universales es inferior al 80% planteado. En la siguiente fase se realiza una evaluación formal de los modelos obtenidos.

Capítulo 6

6 Fase 5: Evaluación

En esta fase se evalúan los modelos obtenidos. La evaluación consiste en determinar si se cumple el criterio de éxito del problema de negocio. Las tareas propias de esta fase son las tres descritas en la Tabla 27.

Evaluación
Evaluar los resultados *Valoración de los resultados* *Modelos aprobados* **Revisión del proceso** *Revisión del proceso* **Determinar próximos pasos** *Técnicas modeladas* *Listado de las posibles acciones*

Tabla 27. Tareas de la fase de evaluación

6.1 Evaluar los resultados

En un proyecto de minería de datos usual, en esta tarea se esperaría tener una respuesta definitiva a si el o los modelos obtenidos cumplen con los objetivos de negocio planteados, pero en el presente proyecto investigativo esto no será posible hasta la siguiente fase, donde se implemente el SCaCS final. Por lo tanto, la

siguiente evaluación está dada exclusivamente en torno a la utilidad que los modelos obtenidos tienen en el SCaCS.

6.1.1 Valoración de los resultados

El enfoque de dos capas propuesto para la clasificación de los CS implica realizar un análisis para cada capa. A continuación se consigna dicha evaluación.

6.1.1.1 Evaluación de los modelos para la capa 1

Para la capa 1, debido al desbalance de clases que existe, se tiene una línea base de exactitud del 92%. Este resultado se obtiene con el algoritmo ZeroR, que solamente basa su clasificación respecto a la clase con mayor número de ejemplos, en este caso, la clase 'No CS'. Por lo tanto, lo esperado es que con otros algoritmos se obtengan valores muy superiores a ese valor. Como podrá comprobar en el Anexo B, así sucedió. Solamente el algoritmo KStar estuvo por debajo del 92% para los tres tipos de modelos, lo que indica un pésimo rendimiento.
Como se ve en la Tabla 28, las exactitudes más altas en la capa 1 se obtienen al usar los modelos personales, seguidos de cerca por los modelos híbridos y por último los modelos universales.

	Personal		**Hibrido**		**Universal**	
	Exactitud	Algoritmo	Exactitud	Algoritmo	Exactitud	Algoritmo
Smartphone secundario	99.876	MP	99.3111	Ib1	95.7155	RF
Smartphone principal	99.778	Ib1	99.2613	RF	98.2444	RF
Manilla	98.858	RF	98.1511	RF	97.1778	RF

Tabla 28. Resultados de exactitud capa 1

En el caso de los modelos personales, para el caso del smartphone secundario es obtenida la mejor exactitud y esta se logra con el algoritmo Multilayer Perceptron (MP). Muy cerca le sigue el smartphone principal, pero esta vez utilizando el algoritmo Ib1. Finalmente, con un punto porcentual por debajo de la exactitud alcanzada por el smartphone secundario, utilizando la manilla se obtiene una exactitud del 98.858% empleando el algoritmo RandomForest (RF). La Figura 23 muestra la distribución de los resultados de exactitud para los modelos personales. Como se ve, la mayoría de participantes obtuvo exactitudes en el rango del 99% y 100% utilizando los datos tomados con el smartphone principal y secundario, mientras que en el caso de la manilla, 12 participantes estuvieron por fuera de ese rango.

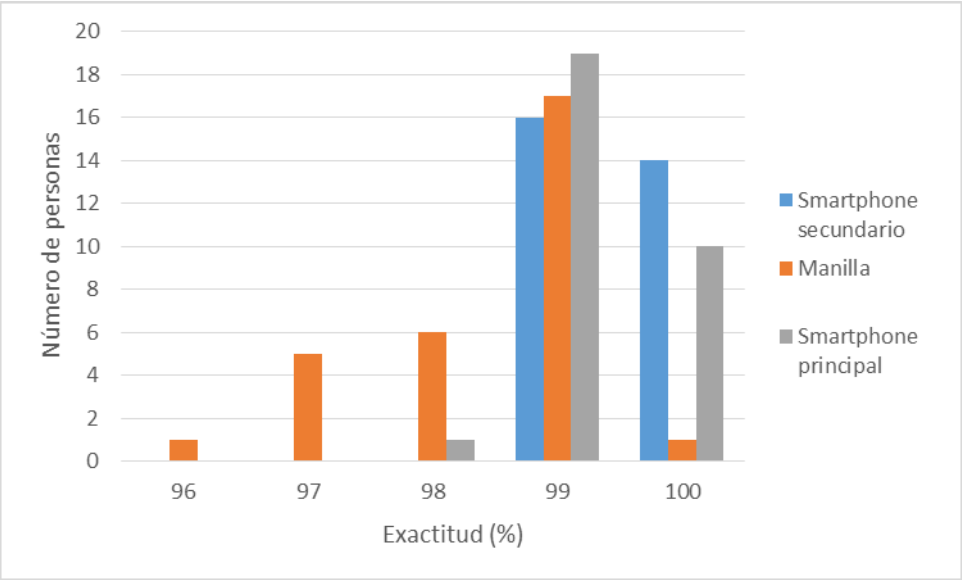

Figura 23. Histograma de exactitud de los modelos personales para todos los participantes en la capa 1

El mismo orden en cuanto a la exactitud de los modelos personales respecto a las fuentes de datos se obtiene para los modelos híbridos, donde la mejor exactitud es obtenida con el algoritmo Ib1 utilizando los datos tomados desde el smartphone secundario, seguido muy de cerca por la exactitud alcanzada al usar el smartphone principal. La Figura 24 muestra los rangos de exactitud alcanzados por cada uno de los 10 modelos híbridos generados cada uno con una semilla de aleatorización diferente. En todos los casos, el nivel de exactitud se mantiene en el rango comprendido del 98 al 100%.

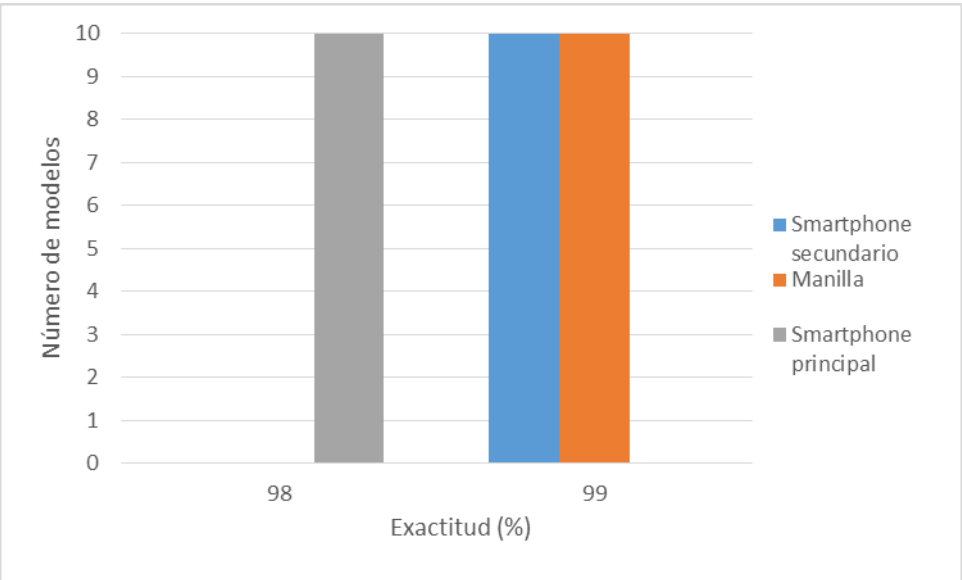

Figura 24. Histograma de exactitud de los modelos híbridos en la capa 1

Para el caso de los modelos universales, el algoritmo RF fue con el que se obtuvo mejor exactitud para las tres fuentes de datos, y a diferencia de los modelos personal e híbrido, la exactitud más alta se consigue con los datos tomados desde el smartphone principal. Teniendo en cuenta que el nivel de exactitud base es del 92%, con la Figura 25 se evidencia que existe una amplia diferencia en la distribución de los resultados de la exactitud para este tipo de modelos, incluso hay 6 participantes que obtienen resultados de exactitud menores o iguales a 92%.

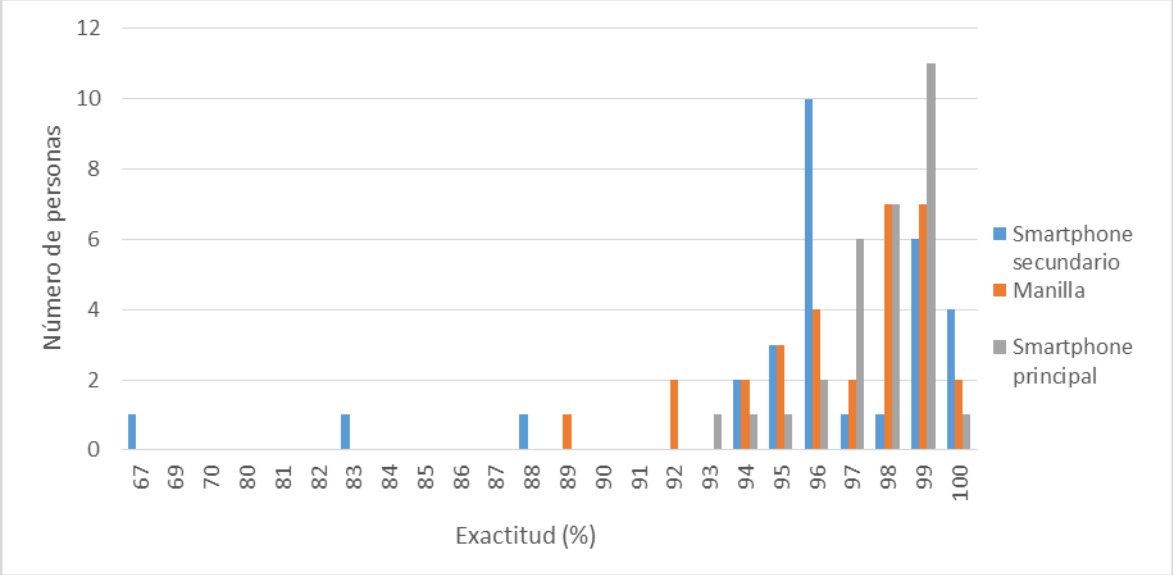

Figura 25. Histograma de exactitud de los modelos universales para todos los participantes en la capa 1

En conclusión, en la capa 1 se obtienen niveles de exactitud elevados para las tres fuentes de datos, y se obtiene una clara ventaja al utilizar los modelos personales e híbridos respecto a los universales. Para el caso de los modelos personales, para los 30 participantes generaron resultados de exactitud por encima del 96% para las tres fuentes de datos. Los resultados obtenidos para los modelos híbridos son muy buenos, estuvieron muy cercanos pero siempre por debajo de los resultados obtenidos por los modelos personales. Por último, los modelos universales no obtuvieron malos resultados, pero es de notar la existencia de 6 resultados que se encuentran por debajo del umbral de exactitud esperado.

6.1.1.2 Evaluación de los modelos para la capa 2

La capa 2 es la encargada de clasificar los 23 CS, por lo que en un principio se cree que los resultados de exactitud no resultarán muy elevados considerando el número de CS y la complejidad de la clasificación de los mismos.

	Personal		Hibrido		Universal	
	Accuracy	Algoritmo	Accuracy	Algoritmo	Accuracy	Algoritmo
Smartphone secundario	99.3671	RF	99.0126	RF	63.8261	ABJ48
Smartphone principal	98.7536	RF	96.7121	ABJ48	51.1642	RF
Manilla	97.0386	RF	92.9111	RF	52.087	BJ48

Tabla 29. Resultados de exactitud capa 2.

Los resultados de exactitud para cada tipo de modelo y fuente de datos se ven en la Tabla 29, donde de manera simple se evidencia el buen rendimiento en cuanto a clasificación de los modelos personal e híbrido. El modelo personal es el tipo de modelo con el que se obtienen mejores resultados de exactitud al clasificar los 23 CS, seguido por el modelo híbrido y muy de lejos por el modelo universal. Se puede observar también que la mejor fuente de datos, para la clasificación de los CS, utilizando cualquiera de los tres tipos de modelos es el smartphone secundario, es decir, tomar los datos desde la cintura de las personas. Se puede deducir que este resultado se debe a que los datos recolectados desde la cintura con un acelerómetro facilitan la clasificación de las tres posturas base: sentado, recostado y acostado.

La exactitud resultante de la clasificación de los 23 CS empleando modelos personales con cualquiera de las tres fuentes de datos es excelente. Si se observa la Figura 26 se encuentra que solo un participante obtiene una exactitud por debajo del 99% al 100% en el caso de utilizar los datos tomados por el smartphone secundario. El rango de exactitud al usar los datos tomados por el smartphone principal es de 96% a 100%, mientras que usando la manilla ese rango se amplía desde 93% a 99%

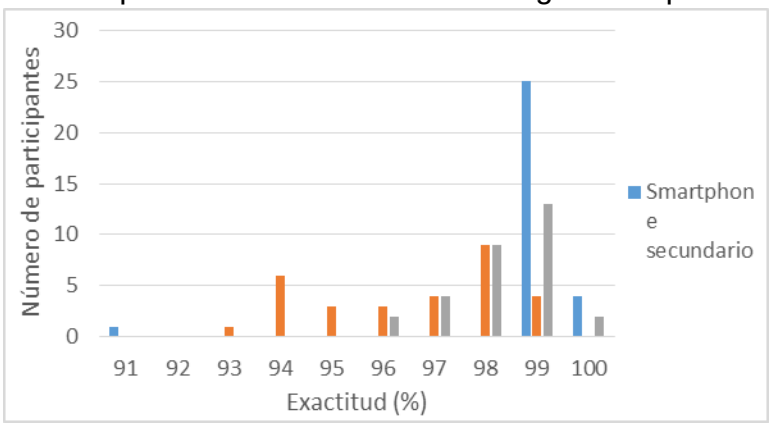

Figura 26. Histograma de exactitud de los modelos personales para todos los participantes en la capa 2

Para el caso de los modelos híbridos, se visualiza desde la Figura 27 que la exactitud al utilizar la manilla como fuente de datos es la peor comparada con las otras dos fuentes, pero aun así, la exactitud no sale del rango del 92% al 93%.

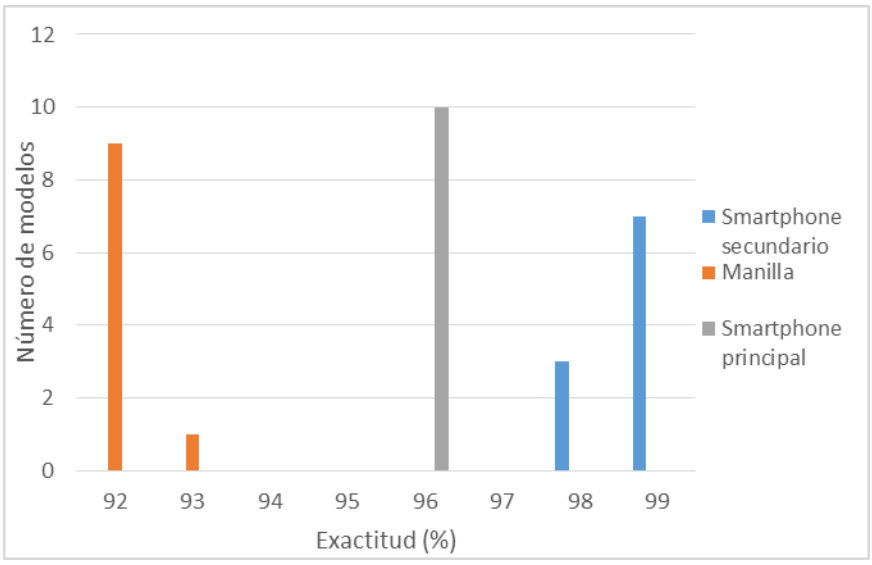

Figura 27. Histograma de exactitud de los modelos híbridos en la capa 2

El rendimiento obtenido con los modelos universales es bastante pobre, la Tabla 27 muestra que la exactitud más alta alcanzada por este tipo de modelo es 63.8261% al utilizar los datos tomados por el smartphone secundario y emplear el algoritmo AdaBoostM1 configurado con un clasificador base J48. La Figura 28 muestra que para el caso del smartphone secundario, 8 participantes obtienen exactitudes comprendidas en el rango de 95% a 100%, y 6 en el rango entre 70% a 80%, lo que aumenta el nivel de exactitud en comparación a las otras dos fuentes de datos, para las cuales la distribución de los resultados de exactitud comprende valores entre el 25% y el 65%.

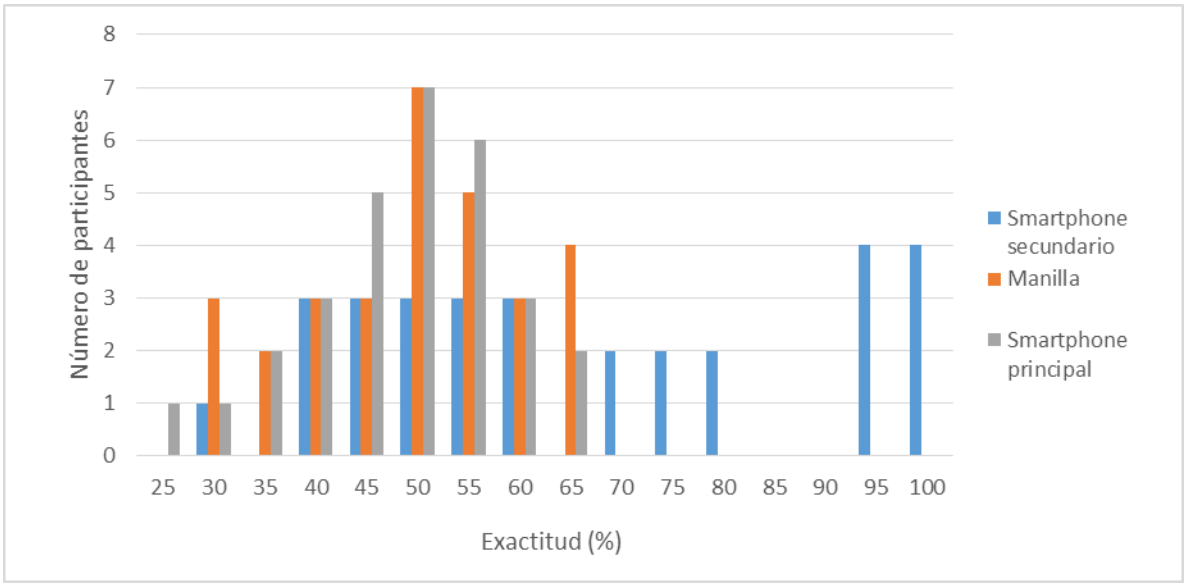

Figura 28. Histograma de exactitud de los modelos universales para todos los participantes en la capa 2

Como conclusión, la clasificación de los 23 CS es exitosa empleando modelos personales o híbridos, pero no lo es para los modelos universales. Teniendo en mente que el rendimiento de los modelos universales se ve afectado por el número de personas con las cuales han sido entrenados, de aquí se desprende la siguiente pregunta: ¿Con datos de cuántas personas se debe entrenar un modelo universal para obtener una exactitud que dé cumplimiento al objetivo de minería de datos planteado? Para responder a esta pregunta fue generada la curva de aprendizaje de cada fuente de datos como se puede observar en las Figuras 29, 30 y 31. Para la obtención de estas curvas se toman los datos de un participante base como datos de evaluación. Al inicio, el dataset de entrenamiento es compuesto solamente por los datos de otro participante, luego por dos, luego por tres y así sucesivamente hasta llegar a ser compuesto por datos de los 29 participantes restantes. Este proceso se sigue para cada uno de los 30 participantes. La inclusión aditiva de los participantes se realizó de manera aleatoria y para hacer reproducible la obtención de las curvas de aprendizaje, el Anexo C contiene el template y los datos para su generación.

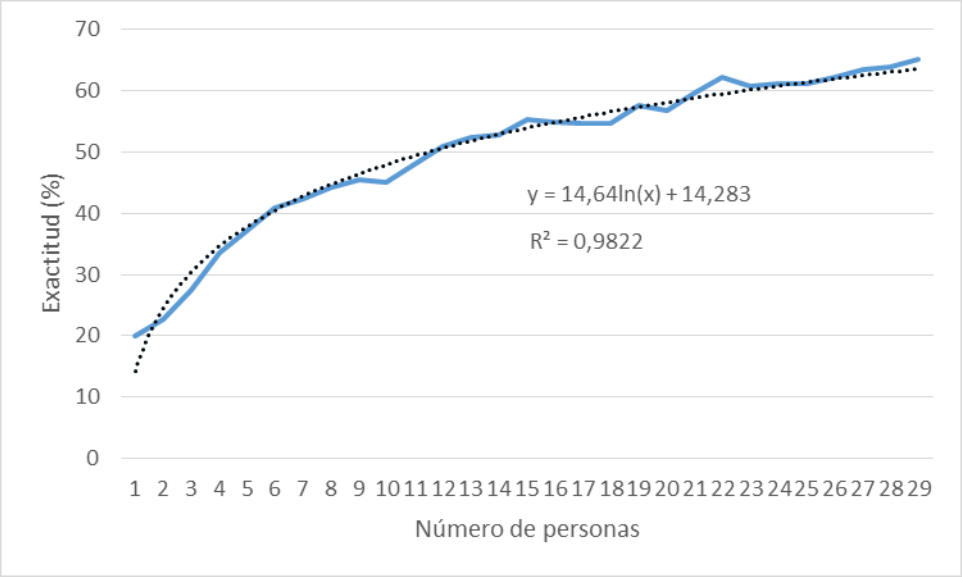

Figura 29. Curva de aprendizaje modelo universal. Smartphone secundario

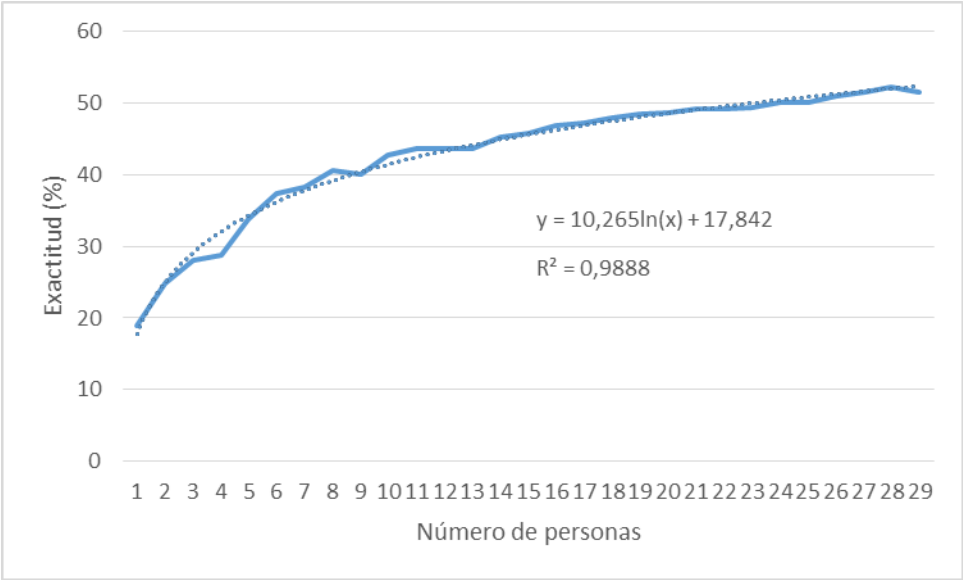

Figura 30. Curva de aprendizaje modelo universal. Smartphone principal

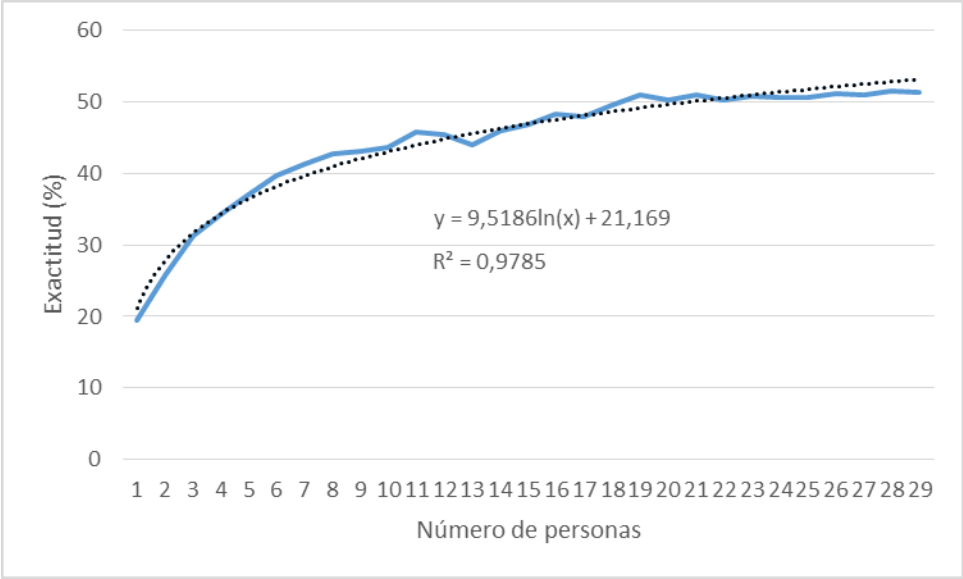

Figura 31. Curva de aprendizaje modelo universal. Manilla

Para los tres casos, una línea de tendencia logarítmica ha podido ser obtenida, con un R^2 muy cercano a 1, valor que indica un excelente ajuste. Para un mejor entendimiento de los resultados obtenidos, se decide graficar las líneas de tendencia obtenidas.

Para el caso del smartphone secundario, la Figura 32 muestra que son necesarias alrededor de 100 personas para alcanzar una exactitud de al menos 80%, 75 personas más para alcanzar una exactitud de 90% y en teoría, al menos 350 personas en total para tener una exactitud cercana al 100%.

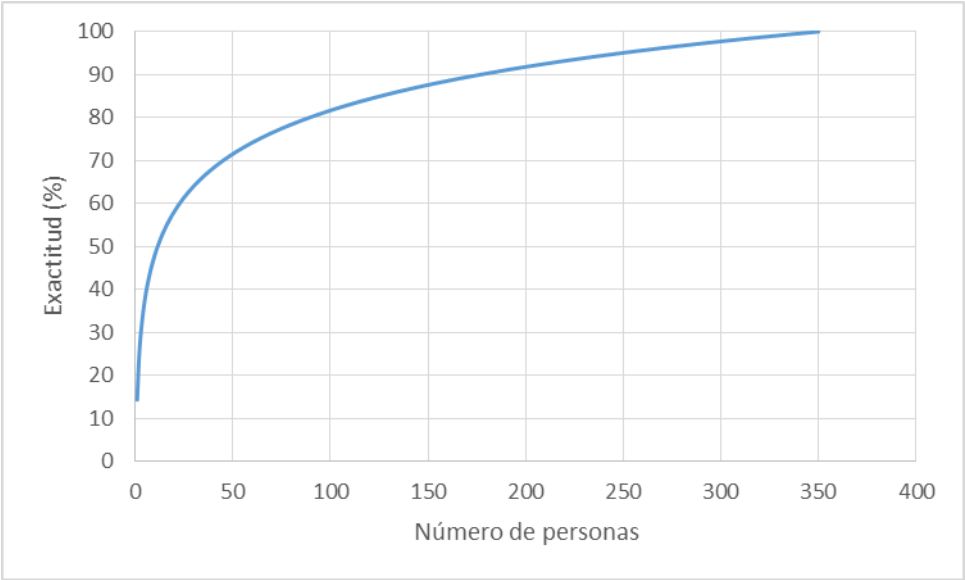

Figura 32.Función de tendencia y = 14,64ln(x) + 14,283. Smartphone secundario

La Figura 33 nos dice que son necesarias alrededor de 500 personas para obtener una exactitud del 80% y que hacen falta unas 2500 personas para obtener un nivel de exactitud parecido al obtenido por los modelos personales al usar los datos obtenidos desde el smartphone principal.

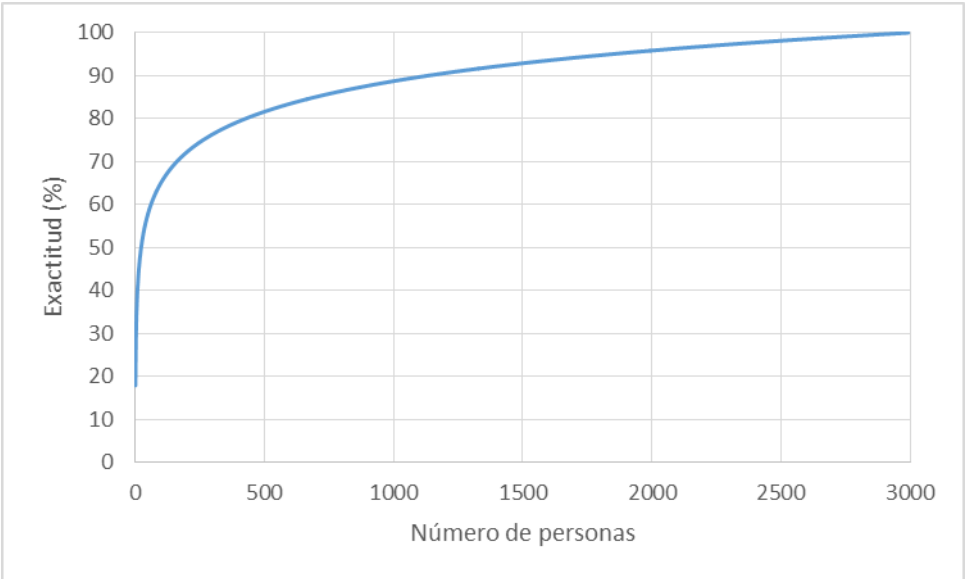

Figura 33. Función de tendencia y = 10,265ln(x) + 17,842. Smartphone principal

Para el caso de la manilla, los resultados son más críticos. Son necesarias alrededor de 500 personas para obtener una exactitud del 80%, pero cerca de 4000 personas para obtener una exactitud cercana al 100%.

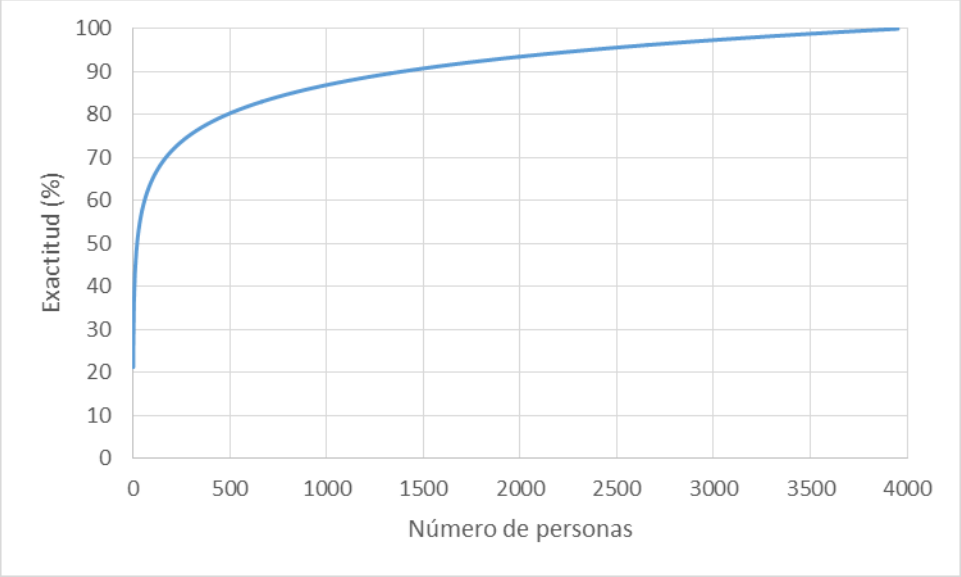

Figura 34. Función de tendencia y = 9,5186ln(x) + 21,169. Manilla

6.1.1.3 Evaluación de la contribución hecha por los datos de localización en entornos cerrados a la exactitud del SCACS

Es de interés saber si los datos de localización en entornos cerrados contribuyen de manera significativa o no en la clasificación de los CS. Para averiguarlo se tuvo que realizar el mismo proceso descrito en la fase de modelado para la generación de los modelos de clasificación, pero en esta ocasión excluyendo los datos de localización y generando y evaluando los modelos únicamente para los algoritmos con mejor exactitud obtenida, mostrados en la Tabla 29.

La hipótesis nula o H_0 plantea que: No existe diferencia significativa en las medias de exactitud al incluir los datos de localización dados por los beacons para la clasificación de los 23 CS en la capa 2.

Por su parte, la hipótesis alternativa establece que: Existe diferencia significativa en las medias de exactitud al incluir los datos de localización dados por los beacons para la clasificación de los 23 CS en la capa 2.

El intervalo de confianza seleccionado es de 0.05, es decir, 5%.

Los resultados de la ejecución de la prueba para cada tipo de modelo se consignan en las Tablas 30, 21 y 32. Cada tabla muestra la exactitud obtenida incluyendo o no los datos de localización en entornos cerrados que se obtienen utilizando los beacons, la diferencia entre esos resultados, el algoritmo utilizado y el nivel de significancia p.

Los resultados de exactitud que se obtienen en la capa 2 utilizando modelos personales y sin utilizar los beacons son bastante sorpresivos. Como se observa en la Tabla 30, cuando se usa el smartphone secundario o principal, se logra una exactitud por encima del 95%. Esto sugiere que el algoritmo RF es capaz de clasificar cambios muy sutiles en la posición de cada persona, distinguiendo de esa

manera los 23 CS sin necesidad de datos de localización con una exactitud muy buena. En este punto es importante recordar que cada participante realizó una única vez cada CS en el proceso de recolección de datos, lo que posiblemente permita al algoritmo RF clasificar los CS según la postura particular en el que estos fueron realizados sin importar los datos de localización en el entorno cerrado. En ese sentido, es necesario experimentar en un trabajo futuro la exactitud de los modelos personales incluyendo o no los datos de localización en entornos cerrados en un dataset en el que cada persona haya recolectado datos de cada CS varias veces, preferiblemente a diferentes horas del día. Luego del análisis anterior, se puede observar que aunque sin utilizar los datos de localización la exactitud es muy buena, al incluirlos la exactitud mejora y según el valor p, el cual es menor a 0.05 que equivale al intervalo de confianza tomado para realizar la prueba, se rechaza la hipótesis nula y se acepta la hipótesis alternativa, es decir, existe diferencia significativa al incluir los datos de localización en entornos cerrados para la clasificación de los 23 CS.

	Exactitud			**Algoritmo**	**p**
	Sin beacons	Con beacons	Diferencia		
Smartphone secundario	96.9324	99.3671	2.4347	RF	0.000
Smartphone principal	95.1594	98.7536	3.5942	RF	0.000
Manilla	87.285	97.0386	9.7536	RF	0.000

Tabla 30. Resultado prueba T-Student modelos personales

Como se describió anteriormente, los resultados de exactitud para los modelos personales sin utilizar los datos de localización fueron sorpresivos y se indicó el escenario de un posible experimento para comprobar la generalización de los resultados allí obtenidos. Pues bien, al revisar los resultados obtenidos por los modelos híbridos sin utilizar los datos de localización en la Tabla 31 y teniendo en cuenta que el rendimiento de los modelos personales ha estado muy cercano, pero siempre por encima de estos en los experimentos realizados, se tiene un indicio de que posiblemente los resultados del experimento planteado como trabajo futuro pueda desembocar en resultados satisfactorios. No está de más aclarar aquí que lo anteriormente comentado aplicaría única y exclusivamente para la clasificación de los 23 CS aquí discutidos. Se presume que la inclusión de un CS diferente a los aquí utilizados y aún más, realizado en una localización diferente a las aquí tratadas (por ejemplo los CS realizados en la oficina) reduciría significativamente la exactitud alcanzada por los algoritmos si solamente utilizaran los datos de aceleración, dando como resultado una necesidad obligatoria de utilizar datos de localización. Ahora bien, al igual que para los modelos personales, para los modelos híbridos el valor de significancia es menor a 0.05, lo que indica que existe una diferencia significativa al incluir los datos de localización datos por los beacons para la clasificación de los 23 CS. Esta diferencia es mayormente notable en el caso de usar la manilla como fuente de datos, ya que pasa de ser del 69% a casi el 93% y de esa forma se garantiza el cumplimiento del objetivo de minería de datos planteado.

	Exactitud			Algoritmo	p
	Sin beacons	Con beacons	Diferencia		
Smartphone secundario	93.2212	99.0125	5.7913	RF	0.000
Smartphone principal	85.7401	96.7121	10.972	ABJ48	0.000
Manilla	69.8584	92.9111	23.0526	RF	0.000

Tabla 31. Resultado prueba T-Student modelos híbridos

Finalmente, para el caso de los modelos universales se puede notar claramente en la Tabla 32 la diferencia entre incluir o no los datos de localización para la clasificación de los CS. Desafortunadamente esta diferencia, aunque es significativa para las tres fuentes de datos, no es suficiente para lograr el cumplimiento del objetivo de minería de datos.

	Exactitud			Algoritmo	p
	Sin beacons	Con beacons	Diferencia		
Smartphone secundario	23.2802	63.8261	40.5459	ABJ48	0.000
Smartphone principal	15.285	51.1642	35.8792	RF	0.000
Manilla	11.913	52.087	40.174	BJ48	0.000

Tabla 32. Resultado prueba T-Student modelos universales

6.1.2 Modelos aprobados

En esta sección se deben aprobar los modelos que contribuirán a la solución del problema del negocio declarado al inicio del proceso de minería de datos. Es clave tener en cuenta que este objetivo va dirigido hacia la creación de un sistema para la clasificación de CS en entornos cerrados, por lo tanto, considerando los resultados de la exactitud total del clasificador de dos capas presentados en la Tabla 26, se llega a que es posible utilizar cualquiera de las tres fuentes de datos ya sea utilizando modelos personales o híbridos. La implementación del clasificador con un modelo universal no es posible a menos que se recolecten datos de una cantidad considerable de personas tal como se analizó en la sección 6.1.1.2. Para la selección entre usar modelos personales o híbridos se debe considerar que para ambos casos se deben recolectar datos de la persona para generar un modelo que posteriormente puede ser usado para la clasificación y según lo analizado hasta el momento, son los modelos personales los que ofrecen un nivel de exactitud mayor, esto gracias a que solamente son entrenados con datos de la misma persona por la que serán usados, logrando de esa manera reconocer más fácilmente los CS. Solamente con 2 minutos de datos personales recolectados por cada CS es suficiente para alcanzar las exactitudes presentadas. Por las anteriores razones, el SCaCS usará modelos personales. Luego de la selección del tipo de modelo, corresponde seleccionar la

fuente de datos que se usará para el SCaCS. Al respecto se ha elaborado la Tabla 33, donde se evalúan criterios como la exactitud obtenida, la penetración en el mercado y la accesibilidad de los datos para cada fuente de datos.

	Exactitud total del clasificador	Penetración del el mercado	Accesibilidad de datos
Cintura (smartphone secundario)	99.2434	Baja	Baja
Smartphone de uso diario (smartphone principal)	98.534	Alta	Alta
Muñeca (manilla)	95.9301	Media	Media

Tabla 33. Características de las fuentes de datos

Empezando con los wearables que se llevan vestidos en la cintura, la mayoría de dispositivos diseñados con ese propósito están enfocados en temas de investigación, son pocos los dispositivos comerciales de ese tipo, un ejemplo es el Fitbit Zip. Adicionalmente, en el caso del Fitbit Zip, no se tiene acceso a los datos en crudo de sus sensores. Son muy pocas las personas que cargan su smartphone de uso diario en la cintura en un estuche, por lo tanto, aunque la exactitud lograda recolectando datos de aceleración desde este lugar del cuerpo es la mejor, a la hora de implementar el SCACS, usar esta opción restringe dramáticamente el alcance de uso del sistema. Siguiendo con el tema de wearables, los que se visten en la muñeca ha tenido una penetración en el mercado superior a la de los que se visten en la cintura. Como se presentó en la Tabla 12, existen una gran variedad de este tipo de dispositivos y algunos de ellos dan acceso a los datos en crudo que toman sus sensores. Una desventaja de los wearables en la actualidad es que no incorporan la funcionalidad de detectar la señal Bluetooth de los beacons y por esa razón se hace necesario, al igual que en el sistema de recolección de datos utilizado en este trabajo, utilizar el smartphone para que realice esa tarea.

Sin lugar a dudas, ha sido el smartphone el dispositivo que ha penetrado el mercado con números exorbitantes en ventas y que además permite obtener los datos en crudo de todos sus sensores. De acuerdo con el estudio “Smartphones users and penetration worldwide, 2013-2018”, a principios de 2017, 2 billones de personas alrededor del mundo tienen un smartphone. Para el caso de Colombia, para esa fecha se tenía un estimado de 18,2 millones de usuarios de smartphones. Ahora bien, a diferencia de los wearables, los cuales las personas visten de manera fija en una parte del cuerpo, es posible que las personas no carguen su smartphone en el bolsillo cuando estén en un entorno cerrado. No se ha encontrado algún estudio en el cual se indague al respecto, por lo que es una desventaja por considerar.

Habiendo considerado las ventajas y desventajas, se decide que el SCaCS emplee el smartphone como su dispositivo principal. La Figura 35 muestra la configuración final del clasificador. Es importante anotar que para el caso del algoritmo Ibk, se experimentó cambiando el número k (número de vecinos más cercanos) a 2, 5, 10, 25, 50 y 100 para evaluar si se mejoraba la exactitud. El mejor resultado obtenido fue con k igual a 1.

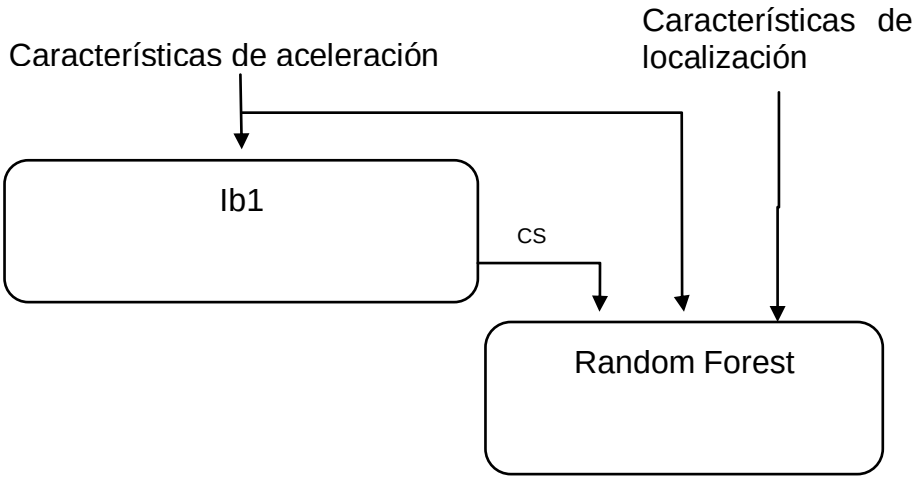

Figura 35. Enfoque de clasificación de dos capas / configuración final

De esta manera, se da por cumplido finalmente el criterio de éxito de minería de datos.

6.2 Revisión del proceso y determinación de próximos pasos.

Cabe recordar que el ciclo de vida de la metodología CRISP-DM no es lineal, así que, aunque no se haya mencionado a lo largo de la monografía todas las iteraciones que se hayan realizado, se da por hecho que el proceso descrito en cada fase de la metodología es al que se ha llegado luego de realizar las iteraciones que hayan sido necesarias. De esa manera, la revisión del proceso involucró reiteradas tareas de modelado y preparación de datos para finalmente describir los resultados obtenidos en la presente monografía. Las revisiones hechas en esta tarea dieron lugar a la ejecución de un paso vital como lo fue la inclusión de la clasificación con enfoque de dos capas, así como la inclusión de un análisis acerca del número de personas necesario para cumplir el objetivo de la minería de datos planteado utilizando modelos universales en la capa 2 del algoritmo clasificador propuesto.

Capítulo 7

7 Fase 6: Despliegue

Ahora es tiempo de transformar el conocimiento y los resultados obtenidos en las anteriores fases para desplegar finalmente el SCaCS. Las tareas de esta fase son las consignadas en la Tabla 34.

Despliegue
Plan de despliegue *Plan de despliegue* **Plan de monitoreo y mantenimiento** *Plan de monitoreo y mantenimiento* **Informe final** *Informe final* **Revisión del proyecto** *Documentación de experiencias*

Tabla 34. Tareas de la fase de despliegue

7.1 Plan de despliegue, monitoreo y mantenimiento

El uso de modelos personales para la clasificación de los CS implica que el SCaCS tenga capacidad para recolectar datos, procesarlos y realizar la clasificación final. El componente central del SCaCS será una aplicación móvil que deberá realizar esas

tareas. Para esto se reutiliza y se modifica la aplicación previamente desarrollada para la recolección de datos. A manera de resumen, el SCaCS tiene las siguientes tareas:

1. Darle a sus usuarios la facilidad para recolectar los datos de aceleración y localización de cada CS.

La recolección de los datos de aceleración y localización se realiza igual como se realizó en la Fase 2 del proceso. Para ello es necesario que los beacons de localización y uso de dispositivos sean ubicados en los lugares y dispositivos correspondientes, a excepción del smartphone, para el cual se implementó, por medio de código, la detección de cuándo el usuario lo está utilizando. El usuario debe por supuesto tener encendido el bluetooth de su smartphone para que así detecte la señal de los beacons. No se almacenarán los datos en crudo recolectados, en su lugar, la aplicación desarrollada obtendrá los ejemplos directamente al realizar el proceso de transformación de datos. De esa forma, las 16 características son calculadas cada 5 segundos, 10 correspondientes a la aceleración y 6 a la localización.

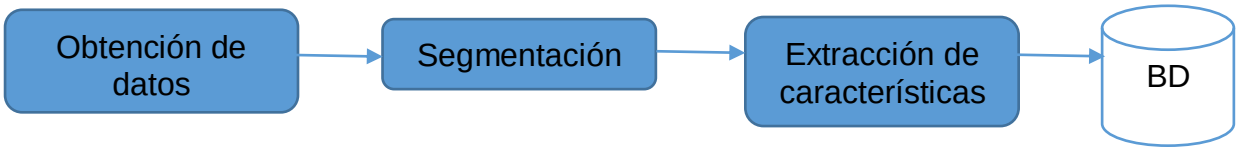

Figura 36. Proceso local de transformación de los datos

2. Al terminar la recolección de los 23 CS, debe generar el modelo personal correspondiente.

Los 750 ejemplos correspondientes a los 23 CS y las dos actividades físicas (de pie y caminando) son enviados a un servidor por medio de internet. El servidor recibe los ejemplos y utilizando los algoritmos IB1 y RF, genera los dos modelos personales que integran el algoritmo clasificador de dos capas. Estos modelos son enviados hacia el smartphone del usuario. Es de notar que el servidor almacena los ejemplos como estrategia para recolectar un dataset con los datos de las personas que usen la aplicación y así poder realizar investigaciones posteriores.

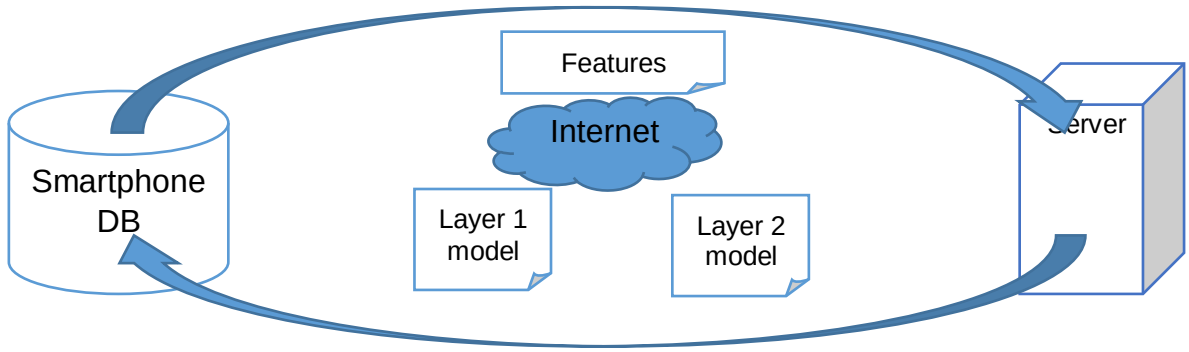

Figura 37. Proceso de generación de modelos.

3. Ya con el modelo de clasificación generado, deberá ejecutar el seguimiento de los CS en segundo plano.

El smartphone, habiendo recibido los modelos personales que integran el clasificador de dos capas, está en capacidad de empezar la clasificación continua de los CS. El proceso de la clasificación es el mostrado en la Figura 38.

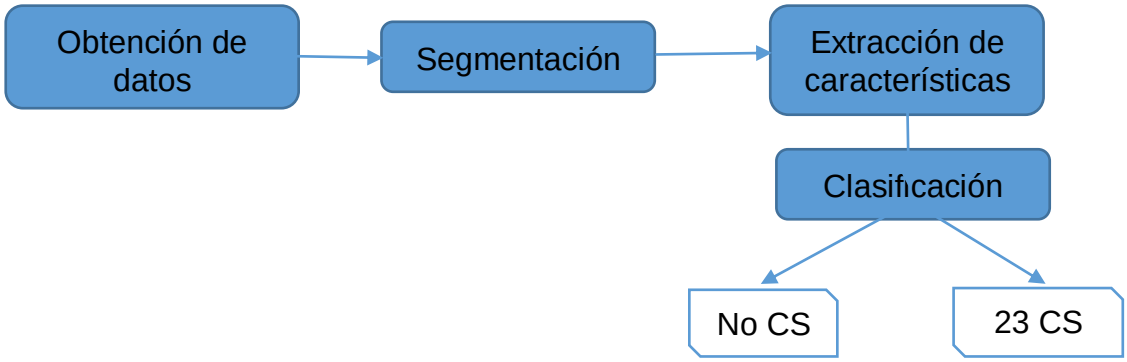

Figura 38. Proceso de clasificación del SCACS

4. Deberá mostrar estadísticas del seguimiento de los CS.

A continuación se presentan las interfaces de la aplicación móvil desarrollada.

Figura 39. App móvil interfaz principal

La aplicación permite visualizar de manera sencilla los minutos transcurridos en el día en los cuales la persona ha realizado actividad física y en los que ha realizado un CS.

En ambas opciones, la persona puede mirar más detalles seleccionando la opción de interés

Figura 40. App móvil interfaz 2

Si la persona selecciona algún CS de la lista, se muestra una nueva lista en la cual se hace explicita la duración y la hora en la cual estuvo realizando el CS seleccionado.

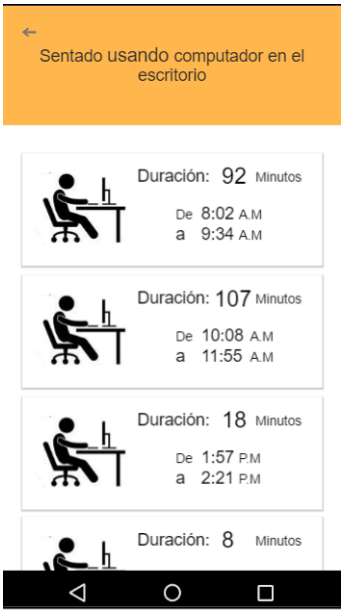

Figura 41. App móvil interfaz 3

Finalmente, la persona puede recolectar sus datos para entrenar su propio modelo personal. Para ello debe ir al menú de la aplicación y seleccionar la opción indicada. Luego de eso, se empezará la recoelcción de datos tal como se realizó en la fase 2 para la recolección del dataset utilizado a lo largo del proceso.

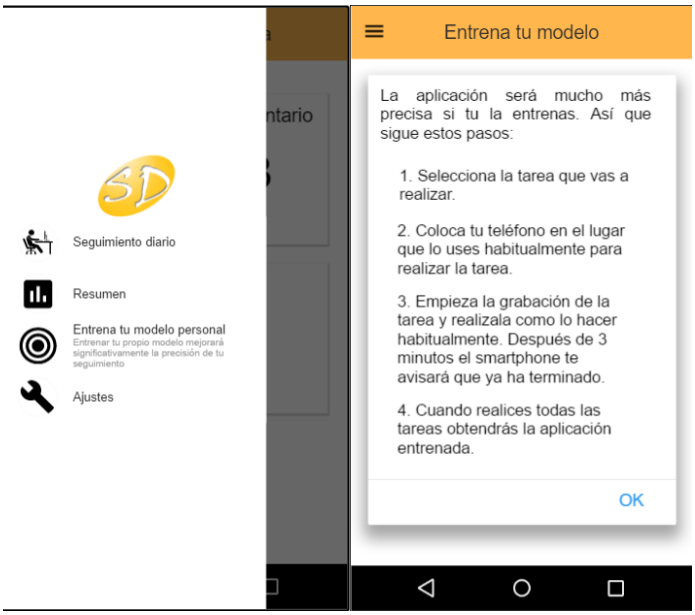

Figura 42. App móvil. Menú

Con la presentación del SCaCS final se da por cumplido finalmente el objetivo del negocio.

7.2 Informe Final

Al final del proyecto, CRISP-DM deja a libre decisión realizar un reporte, una presentación, un resumen, un artículo, o cualquier método para presentar los resultados del proyecto de minería de datos. Esa decisión debe tomarse respecto a la audiencia del informe final. En esta monografía, este espacio es aprovechado para presentar las contribuciones, conclusiones y los trabajos futuros.

7.2.1 Contribuciones

Las contribuciones directas del trabajo de grado presentado son:

1. La mayor contribución de este trabajo de grado de maestría consiste en el cumplimiento de su objetivo general, el cual es la obtención de un sistema para la clasificación de comportamientos sedentarios en entornos cerrados.

2. Un dataset que contiene datos recolectados de diversos sensores de tres dispositivos puestos en diferentes partes del cuerpo (cintura, muñeca y muslo) mientras 30 personas realizaban 23 CS.

3. Un clasificador con enfoque de dos capas para la clasificación de 23 CS el cual brinda una exactitud del 98.534% y además provee flexibilidad respecto

al uso de diferentes técnicas de clasificación y sensores entre sus capas y escalabilidad al permitir implementar capas adicionales para la clasificación de otras actividades o ser integrado dentro de otros clasificadores.

4. Un análisis de la exactitud alcanzada en la clasificación de los 23 CS respecto a la fuente de datos y el tipo de modelo utilizado.

5. Una evaluación de la contribución obtenida en cuanto a exactitud, utilizando los datos de localización en entornos cerrados.

6. Un análisis de las curvas de aprendizaje de los modelos universales para la clasificación de los 23 CS.

7. Publicación de 4 artículos:
    a. J. D. Ceron and D. M. Lopez, “Towards a Personal Health Record System for the Assesment and Monitoring of Sedentary Behavior in Indoor Locations,” *Stud. Health Technol. Inform.*, vol. 228, pp. 804–806, 2016.
    b. J. D. Ceron and D. M. Lopez, G. A. Hoffman, (2017) Two-Layer Method for Sedentary Behaviors Classification Using Smartphone and Bluetooth Beacons. *Studies in health technology and informatics*.
    c. *W. Possos, R. Cruz, J. D. Ceron, D. M. Lopez, H. Sierra,* (2017) *Open Dataset for the Automatic Recognition of Sedentary Behaviors. Studies in health technology and informatics*.
    d. J. D. Ceron and D. M. Lopez, G. A. Ramirez, (2017) A mobile system for sedentary behavior classification based on accelerometer and location data. *Computers in industry (en revisión)*

8. Un framework para realizar el proceso de minería de datos en el ámbito de la clasificación de CS y actividad física en general. Este framework está compuesto por los desarrollos software realizados en respuesta al seguimiento de las fases de la metodología CRISP-DM. Como se observa en la Figura 43, el primer componente consiste en una aplicación móvil Android que es utilizada en la fase 2 con el fin de hacer la recolección de los datos, el segundo componente es empleado en las fases 3 y 4, el cual consiste en una aplicación de escritorio que permite hacer la segmentación y extracción de características del dataset recolectado eligiendo la duración de los ejemplos. El tercer componente agiliza los procesos de generación y evaluación de modelos de diferentes técnicas de clasificación. Consiste en una aplicación escrita en Java y que integra la librería de Weka para tomar los archivos generados por el anterior componente y así entrenar y evaluar los modelos de clasificación. Finalmente, el cuarto componente es implementado en la última fase del proceso, donde los modelos aprobados son utilizados en una aplicación móvil Android que realiza la clasificación continua de los 23 CS.

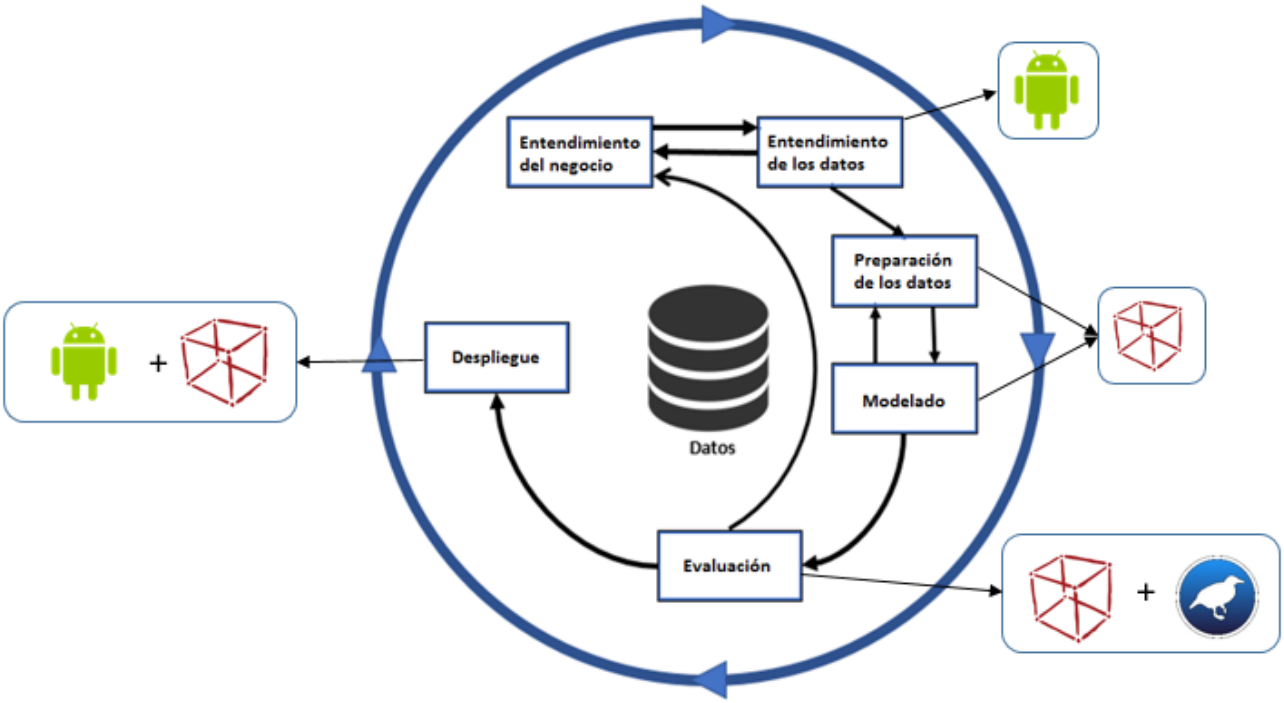

Figura 43. Framework para proceso de minería de datos en clasificación de CS o actividad física

Como contribuciones indirectas se tienen:

1. Ampliación del conocimiento en el tema de clasificación de comportamientos sedentarios, lo que promoverá el surgimiento de nuevas investigaciones dentro y fuera de la Universidad del Cauca acerca de clasificación de actividad física, comportamientos sedentarios y/o actividades complejas o de la vida diaria.

2. Dirección del trabajo de grado titulado “Dataset para la clasificación de comportamientos sedentarios en entornos cerrados” a cargo de los estudiantes del programa de Ingeniería en Electrónica y Telecomunicaciones de la Universidad del Cauca, Stibent Possos y Robinson Cruz.

7.2.2 Conclusiones

6. La necesidad de realizar la clasificación de comportamientos sedentarios está soportada en el reciente interés de la comunidad científica acerca de su impacto en la salud de las personas. Ese interés ha llevado a la propuesta de una taxonomía de comportamientos sedentarios, que fue la base para la selección de los comportamientos sedentarios clasificados en el presente proyecto. Sin duda alguna, las contribuciones aquí aportadas servirán de línea de base para futuras investigaciones acerca de la clasificación automática de ese tipo de comportamientos.

7. Los análisis realizados acerca de los tipos de modelos (personal, híbrido e impersonal), evidencian que la generación de un modelo universal, el cual pueda ser utilizado en frío por cualquier persona, es una tarea que llevaría bastante trabajo, ya que según los muestran las curvas de aprendizaje obtenidas, se necesita una gran cantidad de personas para lograr un modelo universal que provea una exactitud de al menos un 80%.

8. La retribución de recolectar datos personales para la clasificación de los CS y con ellos generar modelos personales, es el buen nivel de exactitud obtenido en la clasificación.

9. A saber, este es el primer sistema dirigido hacia la clasificación de comportamientos sedentarios. El sistema está compuesto por una aplicación móvil Android que se ejecuta en background, recibiendo continuamente la señal bluetooth de los beacons ubicados en el contexto del entorno cerrado, en este caso la casa.

10. El seguimiento de la metodología CRISP-DM guio la ejecución de este proyecto, las tareas incluidas en cada una de las fases son consistentes con lo realmente necesario para proyectos de minería de datos, por tal razón es la metodología para minería de datos más utilizada a nivel mundial.

7.2.3 Trabajos Futuros

1. El dataset recolectado abre muchas posibilidades para trabajos futuros. Algunos de ellos son la experimentación de la clasificación de los CS con los atributos no empleados generando nuevas características a partir de ellos, utilizar selección de características con el dataset transformado, probar otras técnicas de clasificación, ajustar parámetros de algunos algoritmos de clasificación, entre otros.

2. Es necesario experimentar en un trabajo futuro la exactitud de los modelos personales incluyendo o no los datos de localización en entornos cerrados en un dataset en el que cada persona haya recolectado datos de cada CS varias veces, preferiblemente a diferentes horas del día. Lo anterior para comprobar el nivel de exactitud obtenido en la clasificación de los 23 CS sin hacer uso de los datos de localización en entornos cerrados.

3. Efectuar la evaluación del SCACS aquí propuesto en un entorno real.

4. Evaluar el impacto energético que tiene el uso de del SCaCS para el smartphone.

8 Bibliografía

[1] World Health Organization, “Global status report on noncommunicable diseases 2010,” *World Health*, 2010. [Online]. Available: http://whqlibdoc.who.int/publications/2011/9789240686458_eng.pdf.

[2] C. D. Mathers and D. Loncar, “Updated projections of global mortality and burden of disease, 2002-2030: data sources, methods and results.,” *World Health*, no. October, pp. 2002–2030, 2005.

[3] D. R. Whiting, L. Guariguata, C. Weil, and J. Shaw, “IDF Diabetes Atlas: Global estimates of the prevalence of diabetes for 2011 and 2030,” *Diabetes Res. Clin. Pract.*, vol. 94, no. 3, pp. 311–321, 2011.

[4] K. G. M. M. Alberti, P. Zimmet, and J. Shaw, “Metabolic syndrome--a new world-wide definition. A Consensus Statement from the International Diabetes Federation.,” *Diabet. Med.*, vol. 23, no. 5, pp. 469–80, 2006.

[5] K. G. M. M. A. y M. S. R. Zimmeta, Paul, “Una nueva definición mundial del síndrome metabólico propuesta por la Federación Internacional de Diabetes: fundamento y resultados,” *Rev Esp Cardiol*, vol. 58, no. 12, pp. 1371–6, 2005.

[6] P. W. F. Wilson, R. B. D’Agostino, H. Parise, L. Sullivan, and J. B. Meigs, “Metabolic syndrome as a precursor of cardiovascular disease and type 2 diabetes mellitus,” *Circulation*, vol. 112, no. 20, pp. 3066–3072, 2005.

[7] F. Hadaegh, A. Zabetian, M. Tohidi, A. Ghasemi, F. Sheikholeslami, and F. Azizi, “Prevalence of metabolic syndrome by the Adult Treatment Panel III, International Diabetes Federation, and World Health Organization definitions and their association with coronary heart disease in an elderly Iranian population.,” *Ann. Acad. Med. Singapore*, vol. 38, no. 2, pp. 142–149, Feb. 2009.

[8] P. Enrique and M. Soca, “El síndrome metabólico : un alto riesgo para individuos sedentarios The metabolic syndrome : a high risk for sedentary persons,” *Acimed*, vol. 20, no. 1, pp. 1–8, 2009.

[9] D. W. Dunstan *et al.*, “Television viewing time and mortality: The australian diabetes, obesity and lifestyle study (ausdiab),” *Circulation*, vol. 121, no. 3, pp. 384–391, 2010.

[10] A. Grøntved and F. B. Hu, “Television viewing and risk of type 2 diabetes, cardiovascular disease, and all-cause mortality: a meta-analysis.,” *JAMA*, vol. 305, no. 23, pp. 2448–55, 2011.

[11] P. T. Katzmarzyk, T. S. Church, C. L. Craig, and C. Bouchard, “Sitting time and mortality from all causes, cardiovascular disease, and cancer,” *Med. Sci. Sports Exerc.*, vol. 41, no. 5, pp. 998–1005, 2009.

[12] A. A. Thorp, N. Owen, M. Neuhaus, and D. W. Dunstan, “Sedentary behaviors and subsequent health outcomes in adults: A systematic review of longitudinal studies, 19962011,” *Am. J. Prev. Med.*, vol. 41, no. 2, pp. 207–215, 2011.

[13] K. Wijndaele *et al.*, “Television viewing time independently predicts all-cause and cardiovascular mortality: The EPIC Norfolk study,” *Int. J. Epidemiol.*, vol. 40, no. 1, pp. 150–159, 2011.

[14] N. Owen, G. N. Healy, C. E. Matthews, and D. W. Dunstan, “Too much sitting: the population health science of sedentary behavior.,” *Exerc. Sport Sci. Rev.*, vol. 38, no. 3, pp. 105–13, 2010.

[15] M. S. Tremblay, R. C. Colley, T. J. Saunders, G. N. Healy, and N. Owen, “Physiological and health implications of a sedentary lifestyle.,” *Appl. Physiol. Nutr. Metab. = Physiol. Appl. Nutr. Metab.*, vol. 35, no. 6, pp. 725–740, Dec. 2010.

[16] J. P. Sanders *et al.*, “Devices for Self-Monitoring Sedentary Time or Physical Activity: A Scoping Review.,” *J. Med. Internet Res.*, vol. 18, no. 5, p. e90, 2016.

[17] G. N. Healy *et al.*, “Objectively measured light-intensity physical activity is independently associated with 2-h plasma glucose,” *Diabetes Care*, vol. 30, no. 6, pp. 1384–1389, 2007.

[18] G. N. Healy *et al.*, “Objectively measured sedentary time, physical activity, and metabolic risk the Australian Diabetes, Obesity and Lifestyle Study (AusDiab),” *Diabetes Care*, vol. 31, no. 2, pp. 369–371, 2008.

[19] N. Owen, A. Bauman, and W. Brown, “Too much sitting: a novel and important predictor of chronic disease risk?,” *Br. J. Sports Med.*, vol. 43, no. 2, pp. 80–81, 2009.

[20] D. W. Dunstan, G. N. Healy, T. Sugiyama, and N. Owen, “‘Too much sitting’ and metabolic risk - Has modern technology caught up with us?,” *US Endocrinol.*, vol. 5, pp. 29–33, 2009.

[21] U. . E. P. Agency, “Report to Congress on Indoor Air Quality. Volume 2,” 1990.

[22] Eurostat, *How Europeans spend their time: everyday life of women and men : data 1998-2002*. Office for Official Publications of the European Communities, 2004.

[23] A. R. Cooper *et al.*, “Mapping the Walk to School Using Accelerometry Combined with a Global Positioning System,” *Am. J. Prev. Med.*, vol. 38, no. 2, pp. 178–183, 2010.

[24] P. S. Tandon, B. E. Saelens, C. Zhou, J. Kerr, and D. A. Christakis, “Indoor versus outdoor time in preschoolers at child care,” *Am. J. Prev. Med.*, vol. 44, no. 1, pp. 85–88, 2013.

[25] P. J. Krenn, S. Titze, P. Oja, A. Jones, and D. Ogilvie, “Use of global positioning systems to study physical activity and the environment: A systematic review,” *American Journal of Preventive Medicine*, vol. 41, no. 5. pp. 508–515, 2011.

[26] M. Alfonso-Mora, J. A. Vidarte-Claros, C. Vélez-Álvarez, and C. Sandoval-Cuéllar, “Prevalencia de sedentarismo y factores asociados, en personas de 18 a 60 años en Tunja, Colombia,” *Rev. la Fac. Med.*, vol. 61, no. 1, pp. 3–8, 2013.

[27] J. A. V. Claros, C. V. Álvarez, and J. H. P. Sánchez, “El nivel de sedentarismo en nueve ciudades colombianas: an{á}lisis de cl{ú}ster,” *Arch. Med. del Deport. Rev. la Fed. Espa{ñ}ola Med. del Deport. y la Confed. Iberoam. Med. del Deport.*, vol. 33, no. 174, pp. 253–258, 2016.

[28] J. a. Vidarte-Claros, C. Vélez-Álvarez, and J. H. Parra-Sánchez, “Niveles de sedentarismo en población de 18 a 60 años. Manizales, Colombia,” *Rev. salud pública*, vol. 14, no. 3, pp. 417–428, 2012.

[29] J. M. Moine, “Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo,” Facultad de Informática, 2013.

[30] Gregory Piatetsky-Shapiro, “KDnuggets Data Mining, Analytics, Big Data, and Data Science.” .

[31] M. R. T. Perales, O. N. R. Montalvo, and C. A. C. Mundaca, “MODELO DE CLASIFICACIÓN DE OPINIONES SUBJETIVAS EN REDES SOCIALES,” *Rev. Científica Ing. Ciencia, Tecnol. e Innovación*, vol. 1, no. 1, p. 77, 2015.

[32] P. Chapman *et al.*, “CRISP-DM 1.0 Step-by-step data mining guide,” 2000.

[33] “Letter to the editor: standardized use of the terms ‘sedentary’ and ‘sedentary behaviours’.,” *Applied physiology, nutrition, and metabolism = Physiologie appliquee, nutrition et metabolisme*, vol. 37, no. 3. Canada, pp. 540–542, Jun-2012.

[34] F. Massé *et al.*, “Improving activity recognition using a wearable barometric pressure sensor in mobility-impaired stroke patients,” *J. Neuroeng. Rehabil.*, vol. 12, no. 1, p. 72, Dec. 2015.

[35] T. Bastian *et al.*, “Automatic identification of physical activity types and sedentary behaviors from triaxial accelerometer: laboratory-based calibrations are not enough.,” *J. Appl. Physiol.*, vol. 118, no. 6, pp. 716–722, Mar. 2015.

[36] M. B. Del Rosario *et al.*, “A comparison of activity classification in younger and older cohorts using a smartphone,” *Physiol. Meas.*, vol. 35, no. 11, pp. 2269–2286, Nov. 2014.

[37] J. J. Guiry, P. van de Ven, J. Nelson, L. Warmerdam, and H. Riper, “Activity recognition with smartphone support,” *Med. Eng. Phys.*, vol. 36, no. 6, pp. 670–675, 2014.

[38] L. Gao, A. K. Bourke, and J. Nelson, “Evaluation of accelerometer based multi-sensor versus single-sensor activity recognition systems,” 2014.

[39] L. M. Taylor, J. Klenk, A. J. Maney, N. Kerse, B. M. MacDonald, and R. Maddison, “Validation of a Body-Worn Accelerometer to Measure Activity Patterns in Octogenarians,” *Arch. Phys. Med. Rehabil.*, vol. 95, no. 5, pp. 930–934, 2014.

[40] N. Reid *et al.*, “Objectively Measured Activity Patterns among Adults in Residential Aged Care,” *Int. J. Environ. Res. Public Health*, vol. 10, no. 12, pp. 6783–6798, Dec. 2013.

[41] S. F. Kramer, T. Cumming, L. Churilov, and J. Bernhardt, “Measuring Activity Levels at an Acute Stroke Ward: Comparing Observations to a Device,” *Biomed Res. Int.*, vol. 2013, pp. 1–8, 2013.

[42] K. T. Tang, A. M. Richardson, D. Maxwell, W. D. Spence, and B. W. Stansfield, “Evaluation of an Activity Monitor for the Objective Measurement of Free-Living Physical Activity in Children With Cerebral Palsy,” *Arch. Phys. Med. Rehabil.*, vol. 94, no. 12, pp. 2549–2558, 2013.

[43] C. M. Archer, J. Lach, S. Chen, M. F. Abel, and B. C. Bennett, “Activity classification in users of ankle foot orthoses,” *Gait Posture*, vol. 39, no. 1, pp. 111–117, 2014.

[44] S. de Groot and M. G. Nieuwenhuizen, “Validity and reliability of measuring activities, movement intensity and energy expenditure with the DynaPort MoveMonitor,” *Med. Eng. Phys.*, vol. 35, no. 10, pp. 1499–1505, 2013.

[45] Y. Jiang and J. L. Larson, “IDEEA activity monitor: validity of activity recognition for lying, reclining, sitting and standing,” *Front. Med.*, vol. 7, no. 1, pp. 126–131, Mar. 2013.

[46] K. P. Dowd *et al.*, “The measurement of sedentary patterns and behaviors using the activPALTM Professional physical activity monitor,” *Physiol. Meas.*, vol. 33, no. 11, pp. 1887–1899, Nov. 2012.

[47] J. Wang *et al.*, “Energy expenditure estimation during normal ambulation using triaxial accelerometry and barometric pressure,” *Physiol. Meas.*, vol. 33, no. 11, pp. 1811–1830, Nov. 2012.

[48] Y. Xia, V. Cheung, E. Garcia, H. Ding, and M. Karunaithi, “Development of an automated physical activity classification application for mobile phones.,” *Stud. Health Technol. Inform.*, vol. 168, pp. 188–194, 2011.

[49] L. Atallah, J. J. H. Leong, B. Lo, and G.-Z. Yang, “Energy expenditure prediction using a miniaturized ear-worn sensor.,” *Med. Sci. Sports Exerc.*, vol. 43, no. 7, pp. 1369–1377, Jul. 2011.

[50] S. G. Trost, “Comparison of Accelerometer Cut Points for Predicting Activity Intensity in Youth,” *ComparisonofAccelerometerCutPointsforPredictingActivityIntensityinYouth.Med.Sci.SportsExerc*, vol. 43, no. 7, pp. 1360–1368, 2011.

[51] J. Pärkkä, L. Cluitmans, and M. Ermes, “Personalization Algorithm for Real-Time Activity Recognition Using PDA, Wireless Motion Bands, and Binary Decision Tree,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 5, pp. 1211–1215, Sep. 2010.

[52] P.-Y. Jeannet, K. Aminian, C. Bloetzer, B. Najafi, and A. Paraschiv-Ionescu, “Continuous monitoring and quantification of multiple parameters of daily physical activity in ambulatory Duchenne muscular dystrophy patients,” *Eur. J. Paediatr. Neurol.*, vol. 15, no. 1, pp. 40–47, 2011.

[53] A. G. Bonomi, G. Plasqui, A. H. C. Goris, and K. R. Westerterp, “Aspects of activity behavior as a determinant of the physical activity level,” *Scand. J. Med. Sci. Sports*, vol. 22, no. 1, pp. 139–145, Feb. 2012.

[54] M. Benedetti *et al.*, “Physical activity monitoring in obese people in the real life environment,” *J. Neuroeng. Rehabil.*, vol. 6, no. 1, p. 47, 2009.

[55] S. Sa-kwang Song, J. Jaewon Jang, and S.-J. Soo-Jun Park, “Dynamic activity classification based on automatic adaptation of postural orientation,” in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2009, pp. 6175–6178.

[56] J. Jaewon Jang, S. Sa-kwang Song, and S. Park, “An effective method for component activity classification supporting location awareness and user identification,” in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2008, pp. 5258–5261.

[57] A. Godfrey, K. M. Culhane, and G. M. Lyons, “Comparison of the performance of the activPALTM Professional physical activity logger to a discrete accelerometer-based activity monitor,” 2007.

[58] S. E. Crouter, J. R. Churilla, and D. R. Bassett, “Estimating energy expenditure using accelerometers,” *Eur. J. Appl. Physiol.*, vol. 98, no. 6, pp. 601–612, Nov. 2006.

[59] H. Gjoreski, B. Kaluža, M. Gams, R. Milić, and M. Luštrek, “Context-based ensemble method for human energy expenditure estimation,” *Appl. Soft Comput.*, vol. 37, pp. 960–970, 2015.

[60] E. Sazonov, N. Hegde, R. C. Browning, E. L. Melanson, and N. A. Sazonova, “Posture and Activity Recognition and Energy Expenditure Estimation in a Wearable Platform,” *IEEE J. Biomed. Heal. Informatics*, vol. 19, no. 4, pp. 1339–1346, Jul. 2015.

[61] R. Spinney *et al.*, “Indoor Tracking to Understand Physical Activity and Sedentary Behaviour: Exploratory Study in UK Office Buildings,” *PLoS One*, vol. 10, no. 5, p. e0127688, May 2015.

[62] N. Sazonova, R. Browning, E. Melanson, and E. Sazonov, “Posture and activity recognition and energy expenditure prediction in a wearable platform,” in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2014, pp. 4163–4167.

[63] D. O’ Donoghue *et al.*, “Validity of an activity monitor in young people with cerebral palsy gross motor function classification system level I,” *Physiol. Meas.*, vol. 35, no. 11, pp. 2307–2318, Nov. 2014.

[64] M. Arif, M. Bilal, A. Kattan, and S. I. Ahamed, “Better Physical Activity Classification using Smartphone Acceleration Sensor,” *J. Med. Syst.*, vol. 38, no. 9, p. 95, Sep. 2014.

[65] Y. Kim *et al.*, “Examination of Different Accelerometer Cut-Points for Assessing Sedentary Behaviors in Children,” *PLoS One*, vol. 9, no. 4, p. e90630, Apr. 2014.

[66] Z. Zhang and S. Poslad, “Design and Test of a Hybrid Foot Force Sensing and GPS System for Richer User Mobility Activity Recognition,” *Sensors*, vol. 13, no. 11, pp. 14918–14953, Nov. 2013.

[67] M. T. McAloon, S. Hutchins, M. Twiste, R. Jones, and S. Forchtner, “Validation of the activPAL activity monitor in children with hemiplegic gait patterns resultant from cerebral palsy.,” *Prosthet. Orthot. Int.*, vol. 38, no. 5, pp. 393–399, Oct. 2014.

[68] J. L. Miles-Chan *et al.*, “Heterogeneity in the Energy Cost of Posture Maintenance during Standing Relative to Sitting: Phenotyping According to Magnitude and Time-Course,” *PLoS One*, vol. 8, no. 5, p. e65827, May 2013.

[69] K. L. Dannecker, N. A. Sazonova, E. L. Melanson, E. S. Sazonov, and R. C. Browning, “A comparison of energy expenditure estimation of several physical activity monitors.,” *Med. Sci. Sports Exerc.*, vol. 45, no. 11, pp. 2105–2112, Nov. 2013.

[70] E. A. Hinckson, W. G. Hopkins, S. Aminian, and K. Ross, “Week-to-week differences of children’s habitual activity and postural allocation as measured by the ActivPAL monitor,” *Gait Posture*, vol. 38, no. 4, pp. 663–667, 2013.

[71] J. Skotte, M. Korshøj, J. Kristiansen, C. Hanisch, and A. Holtermann, “Detection of Physical Activity Types Using Triaxial Accelerometers,” *J. Phys. Act. Heal.*, vol. 11, pp. 76–84, 2014.

[72] K. P. Dowd *et al.*, “Criterion and Concurrent Validity of the activPALTM Professional Physical Activity Monitor in Adolescent Females,” *PLoS One*, vol. 7, no. 10, p. e47633, Oct. 2012.

[73] L. L. Craft *et al.*, “Evidence that women meeting physical activity guidelines do not sit less: An observational inclinometry study,” *Int. J. Behav. Nutr. Phys. Act.*, vol. 9, no. 1, p. 122, 2012.

[74] G. D. Fulk, S. R. Edgar, R. Bierwirth, P. Hart, P. Lopez-Meyer, and E. Sazonov, “Identifying activity levels and steps of people with stroke using a novel shoe-based sensor.,” *J. Neurol. Phys. Ther.*, vol. 36, no. 2, pp. 100–107, Jun. 2012.

[75] S. L. Schmidt, K. A. Harmon, T. A. Sharp, E. H. Kealey, and D. H. Bessesen, “The Effects of Overfeeding on Spontaneous Physical Activity in Obesity Prone and Obesity Resistant Humans,” *Obesity*, vol. 20, no. 11, pp. 2186–2193, Nov. 2012.

[76] G. D. Fulk and E. Sazonov, “Using Sensors to Measure Activity in People with Stroke,” *Top. Stroke Rehabil.*, vol. 18, no. 6, pp. 746–757, 2015.

[77] N. D. Ridgers *et al.*, “Agreement between activPAL and ActiGraph for assessing children’s sedentary time,” *Int. J. Behav. Nutr. Phys. Act.*, vol. 9, no. 1, p. 15, 2012.

[78] A. Godfrey, A. K. Bourke, G. M. Ólaighin, P. van de Ven, and J. Nelson, “Activity classification using a single chest mounted tri-axial accelerometer,” *Med. Eng. Phys.*, vol. 33, no. 9, pp. 1127–1135, 2011.

[79] S. I. de Vries, M. Engels, and F. G. Garre, “Identification of children’s activity type with accelerometer-based neural networks.,” *Med. Sci. Sports Exerc.*, vol. 43, no. 10, pp. 1994–1999, Oct. 2011.

[80] S. I. De Vries, F. G. Garre, L. H. Engbers, V. H. Hildebrandt, and S. Van Buuren, “Evaluation of Neural Networks to Identify Types of Activity Using Accelerometers,” *Med. Sci. Sport. Exerc*, vol. 43, no. 1, pp. 101–107, 2011.

[81] J. A. Cuthill, K. Fitzpatrick, and J. Glen, “Anaesthesia - a sedentary specialty? Accelerometer assessment of the activity level of anaesthetists while at work,” *Anaesthesia*, vol. 63, no. 3, pp. 279–283, Feb. 2008.

[82] D. K. White, R. C. Wagenaar, M. E. Del Olmo, and T. D. Ellis, “Test-retest reliability of 24 hours of activity monitoring in individuals with Parkinson’s disease in home and community.,” *Neurorehabil. Neural Repair*, vol. 21, no. 4, pp. 327–340, 2007.

[83] P. M. Grant, C. G. Ryan, W. W. Tigbe, and M. H. Granat, “The validation of a novel activity monitor in the measurement of posture and motion during everyday activities.,” *Br. J. Sports Med.*, vol. 40, no. 12, pp. 992–997, Dec. 2006.

[84] S. Vanini, F. Faraci, A. Ferrari, and S. Giordano, “Using barometric pressure data to recognize vertical displacement activities on smartphones,” *Comput. Commun.*, vol. 87, pp. 37–48, 2016.

[85] P. Kelly *et al.*, “Can we use digital life-log images to investigate active and sedentary travel behaviour? Results from a pilot study,” *Int. J. Behav. Nutr. Phys. Act.*, vol. 8, no. 1, p. 44, 2011.

[86] S. F. M. Chastin, U. Schwarz, and D. A. Skelton, “Development of a consensus taxonomy of sedentary behaviors (SIT): Report of Delphi round 1,” *PLoS One*, vol. 8, no. 12, 2013.

[87] P. B. Kruchten, “4+1 view model of architecture,” *IEEE Softw.*, vol. 12, no. 6, pp. 42–50, 1995.

[88] K. Ellis *et al.*, “A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers,” *Physiol. Meas.*, vol. 35, no. 11, pp. 2191–2203, 2014.

[89] S. G. Trost, Y. Zheng, and W.-K. Wong, “Machine learning for activity recognition: hip versus wrist data.,” *Physiol. Meas.*, vol. 35, no. 11, pp. 2183–9, 2014.

[90] M. E. Rosenberger, W. L. Haskell, F. Albinali, S. Mota, J. Nawyn, and S. Intille, “Estimating activity and sedentary behavior from an accelerometer on the hip or wrist,” *Medicine and Science in Sports and Exercise*, vol. 45, no. 5. pp. 964–975, 2013.

[91] “IDC: Smartphone OS Market Share 2016, 2015.” .

[92] J. Hightower and G. Borriello, “Location systems for ubiquitous computing,” *Computer (Long. Beach. Calif).*, vol. 34, no. 8, pp. 57–66, 2001.

[93] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, “The WEKA data mining software: An update,” *SIGKDD Explor.*, vol. 11, no. 1, pp. 10–18, 2009.

[94] J. D. Ceron and D. M. Lopez, “Towards a Personal Health Record System for the Assesment and Monitoring of Sedentary Behavior in Indoor Locations.,” *Stud. Health Technol. Inform.*, vol. 228, pp. 804–806, 2016.

[95] M. Rulsch, J. Busse, M. Struck, and C. Weigand, “Method for daily-life movement classification of elderly people.,” *Biomed. Tech. (Berl).*, vol. 57 Suppl 1, Sep. 2012.

[96] P. Baldi, S. Brunak, Y. Chauvin, C. a Andersen, and H. Nielsen, “Assessing the accuracy of prediction algorithms for classification: an overview.,” *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000.