

Dataset para la Clasificación de Comportamientos Sedentarios en Entornos Cerrados



Proyecto de Trabajo de Grado

**William Stibent Possos Navia
Robinson Cruz Delgado**

Director:
PhD. (c) Jesús David Cerón Bravo

Co-Director:
PhD. Diego Mauricio López

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Departamento de Ingeniería Telemática
Línea de Investigación en eSalud
Popayán, febrero de 2017

William Stibent Possos Navia
Robinson Cruz Delgado

Dataset para la clasificación de comportamientos sedentarios
en entornos cerrados

Trabajo de grado presentado en la Facultad de Ingeniería Electrónica y
Telecomunicaciones de la Universidad del Cauca para la obtención del Título de

Ingeniero en
Electrónica y Telecomunicaciones

Director:
Jesús David Cerón Bravo
PhD(c) en Ingeniería Telemática

Co-Director:
Diego Mauricio López Gutiérrez
PhD. En Ciencias Biomédicas

Popayán
2017

*A todos aquellos que de una u otra manera fueron
parte de esta gran experiencia*

Agradecimientos

Fue un viaje largo, lleno de dificultades y que hayamos llegado a este destino solo es posible por el apoyo que hemos encontrado en familiares, amigos y aquellos que hacen parte de nuestra búsqueda diaria de conocimientos, alegrías, experiencias, distracciones, entre muchas más, que les dan sentido a nuestras vidas. Y aunque, en verdad son demasiadas las personas que deberían ser mencionadas en esta simple página, todos siempre estarán en nuestra mente y nuestros corazones.

Expresamos nuestro más sincero agradecimiento al actual candidato a doctor, ingeniero Jesús David Cerón Bravo cuya dirección fue en pocas palabras: “precisa y de alto nivel”, como era de esperarse de alguien con su personalidad y habilidades. Así como también al doctor ingeniero Diego Mauricio López, quién para nosotros es toda una autoridad mundial en ese interesante campo donde interactúan salud y tecnología. A pesar del formalismo que generalmente existe en la relación profesor-alumno, siempre sentimos a nuestros directores como amigos en los que confiamos uno de los más importantes procesos en la vida de un universitario.

También sentimos la necesidad de agradecer a nuestros padres: Abigail Delgado, Luz Argenis Navia, Pedro Cruz, William N. Possos. Ellos son los seres más admirables y los principales impulsores de las personas que hoy en día somos. Las contribuciones que ayer, hoy y mañana hagamos a la sociedad también serán sus logros.

A nuestros hermanos, tíos, sobrinos, primos y demás familiares que siempre nos brindaron su apoyo en todo momento para así lograr nuestras metas. Muchos de ellos siendo ejemplo para nuestras vidas y otros depositando su confianza en ser nosotros su ejemplo. A todos ellos gracias, ya que han sido más que una bendición para nuestras vidas.

A nuestros amigos, aquellos que estuvieron ahí para nosotros en cualquier situación. No cambiaríamos ni un solo instante de los que compartimos con ustedes y si fuese posible volver a elegir a las personas con quienes compartir esta experiencia, seguramente lo elegiríamos de nuevo.

Finalmente queremos agradecer a la Universidad del Cauca por las herramientas que nos brindó en este proceso de desarrollo profesional y personal.

Resumen estructurado

Antecedentes: El Comportamiento Sedentario (CS) se define frecuentemente como cualquier actividad caracterizada por bajos niveles de gasto energético (1.5 equivalentes metabólicos) mientras un individuo está sentado o reclinado. La evidencia epidemiológica muestra que los CS están asociados con el desarrollo de enfermedades no transmisibles (ENT), como las enfermedades cardiovasculares (ECV), la diabetes tipo 2 y el cáncer. Así que, el reconocimiento de este tipo de comportamientos se hace de vital importancia para que los profesionales de la salud puedan encontrar nuevas maneras de disminuir el sedentarismo, mediante estrategias mejor adaptadas al perfil de cada paciente.

Objetivo: Este trabajo pretende dar el primer paso para alcanzar un sistema de reconocimiento de comportamientos sedentarios. Esto mediante la construcción de un dataset que incluya datos de variables fisiológicas y contextuales que puedan ayudar al reconocimiento de ese tipo de comportamientos, así como el lugar donde se producen dentro de un entorno cerrado.

Metodología: El proceso de construcción del dataset involucró adaptar la metodología CRISP-DM, atender a las recomendaciones de algunas investigaciones relacionadas a la construcción de datasets, diseñar el escenario y sistema adecuados para la recolección de los datos y realizar la recolección de datos sobre una población definida. Después de realizada la recolección se describen los datos recolectados, así como los ajustes realizados sobre algunos de ellos. Finalmente se realiza una evaluación del dataset en tres etapas para verificar la calidad en la metodología de construcción del dataset, su estructura y las muestras incluidas en el mismo. Particularmente, una de las etapas consistió en realizar un experimento de clasificación para verificar la capacidad del dataset para brindar una buena precisión en futuros procesos de clasificación.

Conclusiones: La adaptación de CRISP-DM demostró un buen rendimiento para guiar el proceso de construcción del dataset, sobre todo la instanciación generada a partir de las recomendaciones dadas en investigaciones relacionadas a la construcción de datasets para reconocimiento de actividades de la vida diaria. Además, la evaluación realizada desde las tres perspectivas, logró demostrar la calidad del dataset en cuanto a la metodología de construcción, la estructura en que se encuentran organizados los datos, la documentación generada y pudo dar una idea sobre la precisión que puede alcanzarse en procesos de clasificación usando sus datos.

Structured abstract

Background: Sedentarism is associated with the development of noncommunicable diseases (NCD) such as cardiovascular diseases (CVD), type 2 diabetes, and cancer. Therefore, the identification of specific sedentary behaviors (TV viewing, sitting at work, driving, relaxing, etc.) is especially relevant for planning personalized prevention programs.

Objective: This work aims to take the first step to achieve a system of sedentary behaviors recognition. This through the construction of a public dataset that includes physiological and contextual variables that can help to recognize this type of behaviors, as well as the place where they occur indoors.

Methodology: The dataset's build process involved adapting the CRISP-DM methodology, attending to the recommendations of some research related to good practices in datasets construction, designing the appropriate scenario and system for collecting the data and performing data collection from a defined population. After the data collection, the data are described, as well as the adjustments made to some of them. Finally, an evaluation of the dataset is carried out in three phases to verify the quality in the methodology of construction of the resulting dataset, its structure and the samples included. In particular, one of the phases consisted of performing a classification experiment to verify the dataset's ability to provide good precision in future classification processes.

Conclusions: The adaptation of CRISP-DM demonstrated a good performance to guide the dataset's build process, especially the instantiation generated from the recommendations given in research related to good practices in the datasets construction for activities of daily living recognition. In addition, the evaluation carried out from the three perspectives was able to demonstrate the quality of the dataset in terms of the construction methodology, the structure in which the data are organized, the documentation generated and also, an evaluation phase gave an idea of the classification accuracy that can be achieved in future recognition process with data into result dataset.

Keywords: dataset, sedentary behavior, activities of daily living, data mining, symbolic location, wearables.

Tabla de contenido

Dedicatoria	v
Agradecimientos.....	vii
Resumen estructurado	ix
Structured abstract	xi
Tabla de contenido.....	xiii
Lista de tablas	xvii
Lista de figuras	xix
CAPÍTULO 1. INTRODUCCIÓN	1
1.1. Planteamiento del problema.....	1
1.1.1. Definición.....	1
1.1.2. Pregunta de investigación	3
1.1.3. Justificación	4
1.2. Objetivos	4
1.2.1. Objetivo General.....	4
1.2.2. Objetivos Específicos	4
1.3. Metodología	5
1.3.1. Fase 1: Comprensión del negocio.	5
1.3.2. Fase 2: Comprensión de los datos.	5
1.3.3. Fase 3: Evaluación del dataset propuesto.....	6
1.3.4. Fase 4: Entrega.	6
1.4. Estructura de la monografía	7
CAPÍTULO 2. ESTADO DEL ARTE	9
2.1. Metodología de la revisión	9
2.1.1. Estrategia de Búsqueda	9
2.1.2. Criterios de Inclusión y Exclusión	9
2.1.3. Identificación de Estudios Relevantes	9
2.1.4. Extracción y Síntesis de Datos	9
2.2. Resultados	10
2.2.1. Comportamientos sedentarios reconocidos por los sistemas.....	10
2.2.2. Participantes.....	12

2.2.3.	Variables.....	15
2.2.4.	Método de recolección de datos	17
2.2.5.	Precisión	18
2.2.6.	Datasets.....	19
2.2.7.	Dispositivos.....	20
2.3.	Proceso de obtención de los artículos incluidos en la revisión	23
2.4.	Discusión.....	23
2.4.1.	Principales hallazgos y conclusiones.....	24
2.5.	Conclusiones.....	26
CAPÍTULO 3. MARCO TEÓRICO, DISEÑO DEL SISTEMA Y ESCENARIO DE RECOLECCIÓN DE DATOS		27
3.1.	Metodología CRISP-DM.....	27
3.1.1.	Generalidades.....	28
3.1.2.	Fase 1: Entendimiento del negocio.....	29
3.2.	Metodología para la construcción del sistema de la recolección.....	56
3.2.1.	Pilares fundamentales	57
3.2.2.	Fases de referencia	58
3.3.	Conclusiones.....	58
CAPÍTULO 4. MODELADO Y ARQUITECTURA DEL SISTEMA DE RECOLECCIÓN DE DATOS		59
4.1.	Modelado del negocio	59
4.1.1.	Descripción de los actores	60
4.1.2.	Modelo de casos de uso del negocio.....	60
4.1.3.	Modelo de objetos de la organización.....	61
4.2.	Requisitos del sistema.....	62
4.2.1.	Características del Sistema	62
4.2.2.	Requisitos funcionales	64
4.2.3.	Requisitos no funcionales	64
4.3.	Descripción de la arquitectura del sistema	65
4.3.1.	Vista de Escenarios	66
4.3.2.	Vista Lógica	68
4.3.3.	Vista de Procesos	70

4.3.4.	Vista de Desarrollo	74
4.4.	Clean architecture	76
4.4.1.	Descripción de las capas	79
4.5.	Conclusiones.....	80
CAPÍTULO 5. IMPLEMENTACIÓN DEL SISTEMA		81
5.1.	Software y dispositivos.....	81
5.1.1.	Software	81
5.1.2.	Dispositivos	85
5.2.	Aplicación móvil	86
5.2.1.	Material Design.....	86
5.2.2.	Inicio (Información de la Investigación).....	87
5.2.3.	Formulario Personal	88
5.2.4.	Selección de Actividades.....	89
5.2.5.	Captura de Datos.....	90
5.2.6.	Validación de la Captura.....	91
5.2.7.	Recolección Terminada	92
5.3.	Base de datos	92
5.4.	Conclusiones.....	94
CAPÍTULO 6. RECOLECCIÓN Y LIMPIEZA DEL DATASET		95
6.1.	Recolección de los datos	95
6.1.1.	Posición de los dispositivos	97
6.2.	Descripción inicial de datos recolectados.....	99
6.2.1.	Datos iniciales y primer sondeo de variables.....	99
6.2.2.	Criterios de selección de variables dentro del dataset.....	100
6.2.3.	Selección de tablas o filas de interés.....	100
6.2.4.	Entradas con texto libre	101
6.2.5.	Atributos faltantes	101
6.2.6.	Mejoras en la recolección	101
6.2.7.	Análisis volumétrico de datos	101
6.2.8.	Métodos de captura	102
6.2.9.	Acceso a fuentes de datos	102
6.2.10.	Tipos de atributos.....	102

6.2.11.	Rangos de valores.....	103
6.2.12.	Archivos de datos	104
6.3.	Análisis de atributos	104
6.3.1.	Resultado de un análisis superficial.....	104
6.3.2.	Datos nulos.....	106
6.3.3.	Variables con significado similar	106
6.3.4.	Ortografía y el formato de los valores	107
6.3.6.	Factores no deseados	108
6.4.	Conclusiones.....	109
CAPÍTULO 7. EVALUACIÓN DEL DATASET CONSTRUIDO		111
7.1.	Reporte de datos.....	111
7.1.1.	Nulos por atributo.....	111
7.1.2.	Análisis de nulos para los beacons asociados a los dispositivos.....	112
7.1.3.	Análisis de nulos para los beacons asociados a un lugar.....	112
7.2.	Experimento de clasificación	113
7.2.1.	Resultados del experimento	114
7.3.	Verificación de la información acerca del dataset	116
7.4.	Conclusiones.....	118
CAPÍTULO 8. CONCLUSIONES Y TRABAJOS FUTUROS		119
8.1.	Conclusiones.....	119
8.2.	Aportes	120
8.3.	Trabajos futuros	121
REFERENCIAS.....		123

Lista de tablas

Tabla 2-1. Actividades de comportamiento sedentario generales.....	10
Tabla 2-2. Actividades de comportamiento sedentario específicas.	11
Tabla 2-3. Actividades que implican estar de pie.....	11
Tabla 2-4. Cantidad de participantes incluidos en la recolección de datos.	13
Tabla 2-5. Estado de salud de los participantes.	13
Tabla 2-6. Cantidad de hombres y mujeres.	14
Tabla 2-7. Intervalos de edad de los participantes.....	14
Tabla 2-8. Variables identificadas.	15
Tabla 2-9. Variables identificadas y artículos que las utilizan.	16
Tabla 2-10. Método de recolección de datos.	17
Tabla 2-11. Precisión en actividades de comportamiento sedentario generales.	18
Tabla 2-12. Estadísticas de datos en "Activity Prediction"	19
Tabla 2-13. Estadísticas de datos en "Actitracker"	19
Tabla 2-14. Dispositivos utilizados.....	22
Tabla 3-1. Metodologías de minería de datos.....	28
Tabla 3-2. Sensores de la Microsoft Band 2.	37
Tabla 3-3. Equipos requeridos para la recolección de datos y gastos que conllevan	48
Tabla 3-4. Propiedades a especificar del dataset	49
Tabla 3-5. Comportamientos sedentarios incluidos en el dataset.....	55
Tabla 4-1: Caso de uso, ingreso de datos personales.....	61
Tabla 4-2: Caso de uso, recolección de datos objetivos	61
Tabla 4-3: Caso de uso, validar recolección	61
Tabla 4-4. Características del sistema.....	64
Tabla 4-5: Caso de uso, Recolectar los datos objetivos.	67
Tabla 4-6: Caso de uso, Selección de actividad.	67
Tabla 4-7: Caso de uso, Ingresar.....	67
Tabla 4-8: Caso de uso, Registrar usuario.....	67
Tabla 4-9: Caso de uso, Validar datos capturados.	68
Tabla 6-1: Características de los participantes.	95
Tabla 6-2: Comportamientos sedentarios y su identificador en la Base de Datos. ...	96
Tabla 6-3: Tipos de datos de los sensores	103
Tabla 6-4: Rango de operación de los sensores.....	103
Tabla 6-5: Rangos aproximados de los sensores en la banda.	104
Tabla 6-6. Nulos de beacons en dispositivos.....	112
Tabla 6-7. Nulos en beacons de los lugares	113
Tabla 6-8: Porcentaje de precisión obtenida en la clasificación.....	115
Tabla 6-9. Verificación de información relacionada con el dataset	117

Lista de figuras

Figura 1. Proceso de obtención de artículos en la revisión	23
Figura 2. Ciclo de vida crisp-dm.....	29
Figura 3. Entendimiento del negocio.....	30
Figura 4. Arquitectura inicial del sistema.....	31
Figura 5: sensores disponibles en los wearables.....	33
Figura 6: sensores disponibles para desarrolladores.....	34
Figura 7. Arquitectura del sistema.....	41
Figura 8. Casos de uso del negocio.....	60
Figura 9. Modelo de objetos de la organización	62
Figura 10. Arquitectura 4+1 vistas.....	65
Figura 11. Casos de uso del sistema.....	66
Figura 12. Diagrama de clases del sistema.....	69
Figura 13. Diagrama de estados del sistema.....	70
Figura 14. Diagrama de actividad del sistema.....	71
Figura 15. Diagrama de secuencia, recolectar datos objetivos	72
Figura 16. Diagrama de secuencia, validar recolección de datos.....	73
Figura 17. Diagrama de secuencia, proceso interno en control.....	74
Figura 18. Diagrama de componentes	75
Figura 19. Diagrama de paquetes.....	76
Figura 20. Arquitectura clean.....	77
Figura 21. Arquitectura clean para desarrollo de aplicaciones móviles.....	79
Figura 22. Estructura básica de greendao.....	84
Figura 23. Características de rendimiento de greendao.....	84
Figura 24. Arquitectura soportada por event bus.....	85
Figura 25. Interfaz de inicio.....	87
Figura 26. Formulario personal	88
Figura 27. Selección de actividades.....	89
Figura 28. Captura de datos.....	90
Figura 29. Validación de la captura.....	91
Figura 30. Recolección terminada.....	92
Figura 31. Modelo de la base de datos.....	93
Figura 32. Posición de los dispositivos.....	97
Figura 33: activación del beacon.....	98
Figura 34: primer análisis de los datos del acelerómetro del teléfono principal	105
Figura 35: actividades y promedio de fc y vfc.....	106
Figura 36: ejemplo de la captura de eventos en la aplicación	108

CAPÍTULO 1. INTRODUCCIÓN

1.1. Planteamiento del problema

1.1.1. Definición

Las Enfermedades Cardiovasculares (ECV) son la causa número uno de muertes en el mundo. Esto ha sido evidenciado por la Organización Mundial de la Salud (OMS), la cual calcula que en 2012 fallecieron por esta causa 17,5 millones de personas, lo que representa un 31% de todas las muertes registradas en el mundo. Además, las proyecciones realizadas no son alentadoras, al concluir que las muertes derivadas por ECV ascenderán a 23.3 millones para el 2030 [1].

Uno de los factores de riesgo más importantes de las ECV es la diabetes mellitus tipo 2, de modo que el reconocimiento de sus síntomas previos ha tomado gran importancia. En consecuencia, la Federación Internacional de la Diabetes (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 creación de una nueva entidad clínica, denominada Síndrome Metabólico (SM). Este se define como un grupo de desórdenes médicos que incrementan el riesgo de desarrollar una ECV y/o diabetes mellitus tipo 2 [2], de manera que se convierte en una herramienta clínica simple para identificar a las personas en peligro de padecerlas [3]. Se estima que entre el 20 y 25 por ciento de la población adulta del mundo tiene SM. Esta población, en comparación con las personas que no poseen el síndrome, presentan el doble de probabilidades de morir por un ataque al corazón o un derrame cerebral, tres veces más de padecerlo y cinco veces más de desarrollar diabetes tipo 2 [4].

Según la FID, los factores de riesgo que constituyen el SM [4] son:

1. Presencia de obesidad abdominal (según clasificación de etnia).
2. Aumento de concentración de triglicéridos.
3. Disminución de las concentraciones del colesterol unido a las lipoproteínas de alta densidad (cHDL).
4. Aumento de presión arterial.
5. Alteración en la tolerancia a la glucosa.

Debido al interés reciente en el estudio del impacto del sedentarismo en la salud, se ha podido evidenciar que los estilos de vida sedentarios son una de las principales causas del SM [5] y de ECV en general [6]–[10]. Más aún, se ha encontrado que aquellas personas, con mayores niveles de comportamientos sedentarios, tienen un

mayor riesgo de morbilidad y mortalidad (asociados al incremento de factores de riesgo cardiovasculares como los del SM) independientemente del grado de intensidad de actividad física que practiquen, bien sea moderada o vigorosa [6]–[10]. Lo anterior sugiere que no realizar la cantidad de actividad física, de intensidad moderada o vigorosa (AFMV), recomendada (30 minutos al día por al menos 5 días a la semana según la OMS) e incurrir en comportamientos sedentarios por extensos periodos de tiempo a lo largo del día, constituyen factores de riesgo independientes para la salud.

La red para la investigación del comportamiento sedentario (CS) define el término comportamiento sedentario como: “cualquier actividad realizada por el individuo en posición sentada o inclinada con un gasto energético menor o igual 1.5 METs (equivalentes metabólicos), mientras está despierto”, además sugieren que el término inactivo se use para describir a aquellas personas que realicen una actividad física moderada o vigorosa insuficiente [11]. En relación a estos comportamientos, algunos estudios han evidenciado que durante las horas en que los adultos se encuentran despiertos (15,5 horas/día), aproximadamente el 60% de ese tiempo (9.3 horas/día) incurrir en algún comportamiento sedentario, el 35% (6.5 horas/día) practican una diversidad de actividades físicas de intensidad leve y tan solo el 5% (0.7 horas/día) realizan actividades físicas que representan una intensidad de moderada a vigorosa [12], [13]. Por otra parte, existe evidencia para señalar que el 90% de los comportamientos sedentarios suceden en lugares cerrados tales como la casa o el lugar de trabajo [14].

Desde esta perspectiva, existe un creciente interés en el campo de la investigación del comportamiento humano en evaluar dónde se producen las actividades sedentarias, como por ejemplo en la casa (en el sofá, frente al televisor, en el comedor, en la habitación, en el estudio) o el trabajo (en una sala de reuniones o en el escritorio personal). Debido a que estos factores darían a los investigadores un perfil mucho más completo y objetivo del contexto en el que sucede el comportamiento sedentario y de esa manera, podrían formular estrategias mejor adaptadas para cada individuo. Para que esto sea posible se requiere una medición precisa de la ubicación, especialmente en lugares cerrados[15]. El tiempo del comportamiento sedentario en estos sitios ha sido inferido tradicionalmente por la falta de una señal de GPS o también mediante el uso de un sensor de luz incorporado en wearables [16], [17]. Sin embargo, estos métodos sólo son capaces de diferenciar un entorno cerrado de uno externo, por lo tanto, no ofrecen información detallada de la ubicación en el interior del lugar. Es por eso que se deben explorar otras tecnologías que lo permitan, por ejemplo, RFID, WIFI, Bluetooth, ZigBee, entre otras que podrían proveer una alternativa adecuada. En consecuencia, uno de los ejes de este trabajo es encontrar un atributo de localización que mejore la clasificación de comportamientos sedentarios específicos.

Para obtener un registro personal del comportamiento sedentario, son ampliamente utilizadas algunas encuestas como el cuestionario internacional de actividad física

IPAQ [18] (por sus siglas en inglés, International Physical Activity Questionnaire), el cuestionario de comportamiento sedentario SIT-Q, etc. Estas brindan una gran facilidad y bajo costo de implementación. Sin embargo, presentan una naturaleza subjetiva, pues parten del hecho de que la persona es capaz de percibir por sí misma y recordar sus comportamientos sedentarios, así como el tiempo que permanece en ellos.

Entonces, con el fin de obtener una medida objetiva de los comportamientos sedentarios, se han utilizado una de las más prometedoras herramientas para el monitoreo de la actividad física. Tales herramientas son acelerómetros, giroscopios, podómetros y otros tipos de sensores portátiles, los cuales son moderadamente fiables, de bajo costo y se pueden utilizar en la vida cotidiana gracias a su creciente ubicuidad[19], [20] Adicionalmente, estos sensores están presentes en smartphones y dispositivos wearables, los cuales son fáciles de adquirir y manipular. De hecho, son muchos los estudios científicos que usan wearables para el reconocimiento de actividad física, y actividades de la vida diaria.

Hasta el momento, algunos investigadores han estudiado el impacto de diferentes variables fisiológicas (frecuencia cardíaca, temperatura y respuesta galvánica de la piel) y contextuales (clima, personas cercanas, ubicación y otras), disponibles en los wearables, para el monitoreo de actividad física y algunos pocos CS [21]–[23] Sin embargo, como se mostrará en el siguiente capítulo, estos estudios no han capturado de manera integrada las diferentes variables disponibles en dispositivos comúnmente usados hoy en día, ni tampoco han obtenido buenos resultados en la clasificación de algunos pocos CS específicos. Así pues, surge la necesidad de recolectar un dataset que integre las diferentes lecturas de sensores incluidos en dispositivos wearables comerciales o smartphones en la actualidad con el objetivo de explorar la clasificación de diversos CS por medio de técnicas de minería de datos.

1.1.2. Pregunta de investigación

Considerando los problemas descritos anteriormente, en este trabajo de grado se plantea la siguiente pregunta de investigación: ¿Cómo construir un dataset para la clasificación automática de comportamientos sedentarios que suceden en entornos cerrados?, la cual lleva a plantear la siguiente hipótesis: Es posible construir un dataset para la clasificación automática de comportamientos sedentarios empleando dispositivos ubicuos para la obtención de variables contextuales y fisiológicas.

1.1.3. Justificación

En el contexto del proyecto de investigación financiado por Colciencias y la Universidad Del Cauca, SimeTIC (VRI ID 3727), se encontró que el 36.7% de una muestra de 2100 personas laboralmente activas de la ciudad de Popayán poseen el síndrome metabólico [24] y la prevalencia de sedentarismo en una muestra de 589 personas laboralmente activas fue del 71,1% [25]. Adicionalmente, existen otros estudios que han estimado de manera subjetiva la prevalencia de sedentarismo en ciudades colombianas, los cuales han arrojado como resultado valores por encima del 70 por ciento [26], [27]. Por lo cual, con este trabajo de grado se pretende brindar una herramienta que permita la creación de sistemas capaces de medir el comportamiento sedentario objetivamente que contribuyan a la disminución de las tasas de SM encontradas. De hecho, este proyecto está enmarcado dentro de un trabajo realizado por el candidato a doctor Jesús David Cerón Bravo, quién pretende construir un sistema capaz de clasificar comportamientos sedentarios en un entorno real, apoyado en variables obtenidas de wearables y variables de localización. Sin embargo, para llegar a ese sistema primero es necesario entrenar algoritmos de clasificación con un universo de datos que permita reconocer precisamente los comportamientos sedentarios. Entonces, este trabajo gira alrededor de la creación de un dataset que incluya variables fisiológicas y de contexto, las cuales permitan una clasificación precisa de comportamientos sedentarios. Como resultado, este trabajo termina contribuyendo al proyecto SimeTIC, el cual está generando estrategias soportadas en las TIC para promover el autocuidado en personas con SM en la ciudad de Popayán.

1.2. Objetivos

1.2.1. Objetivo General

Construir un dataset para la clasificación de comportamientos sedentarios en entornos cerrados.

1.2.2. Objetivos Específicos

1. Diseñar el escenario y sistema apropiados para la recolección de datos referentes a la clasificación de comportamientos sedentarios.
2. Construir el sistema para la recolección de los datos en el escenario diseñado.
3. Evaluar la calidad del dataset propuesto.

1.3. Metodología

Para el desarrollo del proyecto se plantea una instanciación personalizada de la metodología para minería de datos CRISP-DM.

A continuación, se describen las fases del proyecto.

1.3.1. Fase 1: Comprensión del negocio.

Actividad 1: Revisión sistemática: revisión de la literatura científica respecto a la clasificación automática de comportamientos sedentarios, dirigiendo la atención a datasets desarrollados y aspectos de investigación como la ubicación o localización del sujeto mientras incurre en los mencionados comportamientos. Además, se dirige la atención a los dispositivos utilizados, las variables obtenidas y los comportamientos sedentarios realizados en cada estudio encontrado. Lo anterior con el objetivo de identificar las características que debería tener el dataset a construir.

Actividad 2: Selección de variables y métodos: selección de variables de entrada (atributos) que incluirá el dataset, así como también las variables de salida (etiquetas o clases), haciendo uso de la revisión literaria realizada y considerando prestaciones o limitaciones de los dispositivos hardware a utilizar.

Actividad 3: Plan del proyecto: descripción de cada uno de los procesos a llevar a cabo durante el proyecto. Además de la definición, desde un principio, de los criterios de evaluación, los objetivos y metas a lograr.

1.3.2. Fase 2: Comprensión de los datos.

Actividad 1: Implementación del sistema: se realizarán los desarrollos o adaptaciones hardware y software que conduzcan a la obtención de un prototipo funcional que permita la recolección de los datos de las variables seleccionadas en la fase previa.

Actividad 2: Recolección de datos: se recogen los datos de un grupo de personas definido, por medio del sistema desarrollado de acuerdo a un protocolo de recolección detallado.

Actividad 3: Descripción y exploración de datos: con el dataset completo se realiza una exploración superficial de su estructura y sus datos.

Actividad 4: Verificación de calidad de datos: revisión de integridad de los datos, errores y la frecuencia de los mismos.

Las fases posteriores de CRISP-DM (preparación de datos, modelado, evaluación y despliegue) se enfocan en la adecuación del dataset, exploración de la clasificación, evaluación de la clasificación y despliegue de los modelos obtenidos. Luego, tales fases están relacionadas a las técnicas de clasificación y los resultados de su implementación. En consecuencia y dado el contexto de este proyecto, estas fases no son adoptadas de forma rigurosa siguiendo la metodología CRISP-DM.

1.3.3. Fase 3: Evaluación del dataset propuesto.

La evaluación del dataset propuesto se realiza en tres actividades:

Actividad 1: Evaluación de los datos: en esta primera actividad se analizan los datos capturados, resultados deseados y resultados no deseados dentro de los registros.

Actividad 2: Clasificación: en esta actividad diferentes algoritmos de clasificación son aplicados al dataset construido con el objetivo de explorar la precisión de la clasificación, además de observar la importancia de los datos de localización en entornos cerrados, dados por los beacons, para realizar la clasificación de comportamientos sedentarios.

Actividad 3: Análisis comparativo: Las características del dataset son verificadas según las recomendaciones encontradas en la literatura sobre dataset bien documentados y construidos.

1.3.4. Fase 4: Entrega.

Actividad 1: Elaboración del documento final: Los conceptos y documentación en general surgida del desarrollo del proyecto son expuestas de forma ordenada y clara.

Actividad 2: Entrega de información del trabajo de grado: Esta entrega consta de la monografía, un disco compacto y un artículo.

1.4. Estructura de la monografía

La presente monografía se encuentra organizada de la siguiente manera:

Capítulo 1: Introducción

Capítulo 2: Estado del Arte

Capítulo 3: Marco teórico, diseño del sistema y escenario de recolección de datos

Capítulo 4: Modelado y arquitectura del sistema de recolección de datos.

Capítulo 5: Implementación del sistema.

Capítulo 6: Recolección y limpieza del dataset.

Capítulo 7: Evaluación del dataset construido.

Capítulo 8: Conclusiones y trabajos futuros.

CAPÍTULO 2.

ESTADO DEL ARTE

En este capítulo se realiza una revisión sistemática, la cual pretende identificar las tecnologías y métodos empleados para la clasificación de comportamientos sedentarios hasta la fecha. Además, se identifican los comportamientos sedentarios específicos que han sido clasificados, la precisión obtenida y las variables usadas en el proceso de su clasificación.

2.1. Metodología de la revisión

2.1.1. 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 cadena de búsqueda fue construida alrededor de cuatro grupos de palabras clave: comportamientos sedentarios (p.ej. screen time, body posture, sitting time, sedentary behavior), clasificación (p.ej. classification, recognition, dataset), método de monitoreo (p.ej. wearable, activity monitor, camera, RFID, bluetooth) y variables utilizadas (p.ej. accelerometer, heart rate, skin temperature, location).

2.1.2. Criterios de Inclusión y Exclusión

Para ser incluido, un estudio debe cumplir los siguientes requisitos: (1) estar en idioma inglés o español; (2) describir un sistema capaz de reconocer comportamientos sedentarios de manera objetiva. Ese sistema puede ser wearable o no; (3) haber sido publicado desde el año 2006.

2.1.3. Identificación de Estudios Relevantes

Para determinar la inclusión de los artículos, en primera instancia se tuvo en cuenta su título y abstract. Se excluyeron aquellos artículos que no cumplieron los criterios de inclusión. Los artículos potencialmente elegibles se obtuvieron para una lectura completa y luego se determinó su inclusión.

2.1.4. Extracción y Síntesis de Datos

Para la obtención de la información se desarrolló un banco de datos, dentro del cual se ajustó la información relevante extraída de cada uno de los artículos incluidos en la revisión. Las características de interés para esta revisión fueron: variables utilizadas, comportamientos sedentarios que reconoce el sistema, precisión alcanzada,

dispositivos usados, detalles del contexto de implementación y prueba, detalles sobre los participantes, duración de la recolección de datos y disponibilidad del dataset construido.

2.2. Resultados

Con la estrategia de búsqueda planteada se obtuvieron 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 sometidos a los criterios de inclusión y exclusión. 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, por cumplir los criterios de inclusión planteados. Por lo tanto, los artículos incluidos en esta revisión fueron 52.

2.2.1. Comportamientos sedentarios reconocidos por los sistemas.

En la Tabla 2-1 se consignan los comportamientos sedentarios considerados en los artículos incluidos. De ahí resulta relevante que los comportamientos sedentarios de mayor atención entre los estudiados son, permanecer acostado, sentado o de pie. Específicamente, de los artículos resultantes 51.92% clasifica la actividad de permanecer acostado, 94.23% estar sentado y 94.23% permanecer de pie. Otras actividades como trabajar en un computador, usar el ascensor y lavar platos son tenidas en cuenta en un porcentaje menor de estudios: 7.69%, 5.77% y 5.77% respectivamente. Las demás actividades mostradas en la Tabla 2-2 y la Tabla 2-3 son actividades específicas presentes en estudios únicos y en posturas diferentes. También hay que tener en cuenta que, con base en la definición de comportamiento sedentario, aquellos comportamientos que suceden en posición ‘de pie’, no son considerados comportamientos sedentarios como tal. Sin embargo, muchos de los artículos revisados no manejan esa misma definición de comportamiento sedentario (problema identificado por la red para la investigación del comportamiento sedentario [11] y por lo cual solicitan a los autores de estudios científicos manejar la definición propuesta por ellos), por lo tanto, para hacer una descripción general, en esta revisión también se consideran los CS que aborden la posición de estar de pie.

Actividades de comportamiento sedentario generales	Número de artículos que considera la actividad	Referencia(s)
Acostado	26	[28]–[53]
Sentado	48	[28]–[42], [44]–[51], [53]–[77]
Parado	49	[28]–[43], [45]–[58], [60]–[78]

Tabla 2-1. Actividades de comportamiento sedentario generales.

Actividades de comportamiento sedentario específicas	Número de artículos que considera la actividad	Referencia(s)
Trabajando en el computador	4	[43], [52], [53], [59]
Inclinado/encorvado en una silla	3	[29], [39], [44]
Jugar en el computador	1	[44]
Ver TV	1	[59]
Conducir un auto	1	[79]
Ir en bus	1	[79]
Sentado en el sofá	1	[50]
Sentado en la cama	1	[50]
Sentado en el baño	1	[50]
Leer	1	[59]
Jugar videojuegos	1	[59]
Escribir a mano	1	[44]
Organizar papeles	1	[52]
Acostado en la cama	1	[50]
Acostado en el sofá	1	[50]

Tabla 2-2. Actividades de comportamiento sedentario específicas.

Actividades de comportamiento sedentario que implican estar de pie	Número de artículos que considera la actividad	Referencia(s)
Usar ascensor	3	[28], [30], [78]
Lavar platos	3	[44], [52], [53]
Aspirar	2	[43], [52]
Barrer	2	[44], [52]
Cerca al sofá	1	[50]
Cerca de la cama	1	[50]
De pie en una silla	1	[50]
De pie en el baño	1	[50]
Lavar ventanas	1	[39]

Tabla 2-3. Actividades que implican estar de pie.

Con respecto a la actividad de barrer, los artículos [44], [76] la incluyen como comportamiento sedentario, dado que mientras se realiza la actividad, no se hacen movimientos importantes, sino pasos pequeños con frecuencia irregulares y pequeños. Sin embargo, en [28] la consideran actividad física de intensidad leve.

En cuanto al reconocimiento de comportamientos sedentarios utilizando variables del contexto en el que ocurren, el artículo [50] fue uno de los más completos en ese sentido. Entre las actividades consideradas se encuentran: estar sentado (en el sofá, el baño, la cama o una silla), estar acostado (en la cama o el sofá) y estar de pie (en

el baño, el sofá, la cama o una silla). Otro de los estudios que consigue obtener buenos resultados en la clasificación de distintos tipos de actividades es [59]. Este incluye, entre sus actividades, dos nuevas: escribir y lavar ropa.

En algunos estudios [44], [52], [59], se establecen rutinas en las que cada participante realiza un conjunto de actividades que al final del estudio son relacionadas con los datos recolectados, es decir, definen unos puntos de corte, con los cuales clasifican manualmente los datos adquiridos; pero no se hace una prueba de clasificación automática mediante el uso de alguna técnica como el aprendizaje de máquina supervisado.

2.2.2. Participantes

Los 52 estudios mencionan la información sobre los participantes involucrados en sus mediciones. A continuación, se describen las características registradas, entre las cuales resaltan: cantidad de personas, género, estado de salud y edad. Factores sumamente importantes para una buena descripción de un dataset y sobre todo para buscar la diversidad adecuada de una población.

2.2.2.1. Cantidad de Participantes

Para dar una descripción de la cantidad de personas que participaron en la recolección de datos en los 52 estudios, se emplearon 7 intervalos como se muestra en la Tabla 2-4. La mayor cantidad de artículos recolectaron datos de 1-10 personas (28.85% de los artículos). No obstante, el intervalo de 11-20 personas también es ampliamente considerado (cerca del 25% de las veces). Además, hay un 15.38% de artículos que emplean entre 21 y 30 personas. Así mismo, los intervalos 31-40, 41-50, 51-100, 101-500 son menos utilizados con 7.69%, 9.61%, 9.61% y 3.85% respectivamente. Eso era de esperarse por la cantidad de actividades consideradas y el número de sensores necesarios para clasificar actividades específicas en grandes poblaciones.

Intervalo de cantidad de participantes	Número de Artículos	Referencias
1-10	15	[32], [37], [42], [45], [46], [49], [51], [53], [60], [61], [70], [72], [77], [78], [80]
11-20	13	[28], [35], [36], [41], [47], [50], [54], [56], [57], [63], [65], [68], [79]
21-30	8	[31], [33], [38], [39], [43], [48], [62], [66]
31-40	4	[30], [34], [55], [58]

41-50	5	[40], [52], [71], [74], [75]
51-100	5	[29], [64], [67], [69], [73]
101-500	2	[44], [58]

Tabla 2-4. Cantidad de participantes incluidos en la recolección de datos.

2.2.2.2. Características de salud de los participantes

En total 19 artículos mencionan algún estado de salud de sus participantes (Tabla 2-5). Las enfermedades que describen los diferentes artículos son: Accidente cerebrovascular [28], [35], [68], [70], parálisis cerebral [36], [57], órtesis [37], distrofia muscular [46], Parkinson [51]. Esas enfermedades tienden a limitar la movilidad de las personas de alguna u otra manera. De ahí que en estos estudios se pretenda observar la vida de las personas que sufren estas enfermedades y cómo se ven afectadas durante el proceso de recuperación. Todo eso mediante la información de actividad física y midiendo el tiempo dedicado a comportamientos sedentarios.

Por otra parte, ningún artículo describe el nivel de actividad física o sedentarismo de sus participantes, los cuales son aspectos asociados con algunas variables como la frecuencia cardíaca y pueden contribuir en los procesos de clasificación.

Estado de salud	Cantidad de artículos	Referencia
Con alguna enfermedad	9	[28], [35]–[37], [46], [57], [68], [70], [80]
Normal	7	[41]–[43], [49], [53], [69], [77]
Sobrepeso u obesidad	2	[48], [69]

Tabla 2-5. Estado de salud de los participantes.

2.2.2.3. Género

De los 21 artículos que describen el género de los participantes, 18 emplean participantes de ambos géneros y solamente 3 miden a un solo grupo, específicamente solo mujeres [40], [66], [67]. A su vez, [66] aclara que el uso de solo las adolescentes en su estudio, se debe a que ellas tienden a ser más inactivas y sedentarias; por su parte [40] recolecta datos de solo un grupo de mujeres debido a que pretenden relacionar la actividad física y la mamografía.

En aquellos artículos que utilizan una población combinada de géneros, solamente en 3 estudios hay paridad en la cantidad de participantes hombres y participantes mujeres (véase Tabla 2-6). Esto se debe a la disponibilidad en cuanto a personas que cumplan con los criterios planteados y estén dispuestas a participar en la recolección de datos.

Cantidad Mujeres	Cantidad Hombres	Referencia
7	5	[28]
6	9	[54]
4	6	[60]
4	11	[36]
1	7	[37]
12	10	[62]
16	14	[39]
10	7	[65]
44	0	[40]
5	8	[41]
30	0	[66]
91	0	[67]
6	6	[68]
31	24	[69]
6	2	[70]
12	8	[79]
27	31	[73]
7	18	[44]
2	2	[49]
3	7	[51]
24	24	[52]

Tabla 2-6. Cantidad de hombres y mujeres.

2.2.2.4. Edad de los participantes

Para los intervalos de edad se usaron los datos y recomendaciones suministradas por la página [81].

Intervalos de Edad	Cantidad de Artículos	Referencias
Niño 1-11 años	8	[36], [44]–[46], [59], [64], [71], [73]
Adolescente 12-19	10	[29], [36], [38], [40], [44], [45], [59], [66], [71], [73]
Juventud 20-39	22	[29]–[31], [38], [39], [41], [43], [45], [47]–[49], [51]–[54], [60], [62], [65], [69], [72], [74], [79]
Adulthood 40-64	14	[28], [29], [38], [39], [47], [48], [60], [65], [67], [68], [70], [74], [79], [80]
Vejez >65	14	[28], [30], [32]–[34], [37]–[39], [48], [67], [68], [70], [74], [80]

Tabla 2-7. Intervalos de edad de los participantes.

De los 52 artículos, en 40 de ellos se logró obtener información acerca de la edad de sus participantes. Estos datos son plasmados en la Tabla 2-7. Se observa que 19 de

las investigaciones usan grupos de personas con edades pertenecientes a solo uno de los intervalos de edades definidos, en 15 son comprometidos dos de los intervalos de edad. De esos 15, 5 estudios involucran personas en adultez y vejez, 5 involucran niños y adolescentes, 4 miden jóvenes y adultos, y 1 toma datos de personas en juventud y vejez. Además, en 5 estudios participan personas en tres de los intervalos de edad estandarizados (en 3 estudios recolectan datos de jóvenes, adultos y viejos; 1 toma datos de adolescentes, jóvenes y adultos adolescente a adultez; y 1 mide a niños, adolescentes y jóvenes) y solamente en uno de los artículos se usa un grupo de personas pertenecientes a cuatro intervalos (todos los rangos entre adolescencia a vejez). De las observaciones anteriores, es claro que los jóvenes han sido los que más han participado en este tipo de proyectos y muy poco los niños.

Dos de los artículos no describen un intervalo de edad definido de sus participantes, sino que lo hacen cualitativamente, a saber, jóvenes en uno y adultos en el otro [42], [57].

2.2.3. Variables

En esta sección se obtuvo información de las variables asociadas en la clasificación de los comportamientos sedentarios. Entre las variables se consideran las fisiológicas y las variables de contexto. Las variables fisiológicas se refieren a información sobre respuestas o comportamientos producidos en el organismo biológico 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 [53]. La Tabla 2-8, muestra las variables utilizadas en los artículos revisados, mientras que la Tabla 2-9, muestra los estudios que las utilizan.

Variable Fisiológica	Variable de contexto
- Frecuencia cardiaca FC	Acelerómetro
- Saturación de oxígeno	Podómetro
- Frecuencia respiratoria	Sensor de presión (planta del zapato)
- Temperatura de la piel	Presión atmosférica
- Frecuencia Cardiaca	Giroscopio
	GPS
	Inclinómetro
	Localización con RFID
	Localización con sensor de presión
	Luz
	Temperatura ambiente
	Video/Imágenes
	Infrarrojo Pasivo

Tabla 2-8. Variables identificadas.

Variable	Artículos que emplean la variable	Referencias
Acelerómetro	49	[28]–[39], [41]–[55], [57]–[61], [63]–[75], [77]–[80]
Sensor de Presión	6	[50], [54], [56], [63], [68], [70]
Giroscopio	4	[28], [37], [46], [72]
Presión Barométrica	4	[28], [30], [41], [78]
Inclinómetro	3	[40], [55], [69]
Saturación de oxígeno	3	[44], [53], [66]
Podómetro	2	[40], [69]
Frecuencia Respiratoria	2	[53], [62]
Temperatura Ambiente	2	[53], [79]
Frecuencia Cardíaca	2	[53], [62]
Localización con RFID	1	[55]
Localización con sensor de presión	1	[50]
GPS	1	[60]
Nivel de luz	1	[79]
Video/Imágenes	1	[79]
Infrarrojo Pasivo	1	[79]
Temperatura de la piel	1	[53]
GSR	1	[53]

Tabla 2-9. Variables identificadas y artículos que las utilizan.

En general, los estudios utilizan variables de acelerometría para capturar patrones y características de movimiento. El sensor de presión es usado por una serie de artículos, que realizan el reconocimiento de actividades, mediante la obtención de datos de sensores de presión (puestos en la suela del zapato) y apoyando las mediciones con variables de acelerometría (también puestas en un zapato). El giroscopio es usado para percibir los cambios en la posición corporal de la persona y generalmente se usa junto con el acelerómetro. El podómetro es otra variable ampliamente utilizada para contar el número de pasos de las personas durante su rutina diaria. La presión atmosférica o barométrica es usada en los estudios principalmente para caracterizar los movimientos verticales como usar escaleras o el elevador.

En relación al número de variables utilizadas: [53] es el que recolecta la mayor cantidad, a saber, 7 (Acelerómetro, temperatura ambiente, frecuencia cardíaca, frecuencia respiratoria, temperatura de la piel, GSR y saturación de oxígeno), incorporados en 4 dispositivos (Shimmer, Zephyr sensor, BodyMedia y Cosmed). Seguido por [79] que utiliza 5, todos incorporados en la SenseCam (video/imágenes,

infrarrojo pasivo para detectar personas frente a la cámara, temperatura ambiente, nivel de luz y acelerómetro). Cuatro artículos usan 3 variables combinadas. De los cuatro, 1 usa acelerómetro, presión barométrica y giroscopio; otro utiliza acelerómetro, inclinómetro y localización con RFID; 1 usa acelerómetro, inclinómetro y podómetro; y 1 utiliza acelerómetro, sensor de presión y localización por presión. Además, 15 combinan 2 de las variables (4 usan acelerómetro y sensor de presión, 3 utilizan acelerómetro y presión barométrica, 3 mencionan el acelerómetro y el giroscopio, 2 trabajan con acelerómetro y saturación de oxígeno, 1 usa acelerómetro y GPS, otro utiliza inclinómetro y podómetro y 1 de los estudios usa frecuencia cardiaca y respiratoria. Finalmente, 31 de los artículos solamente hacen uso de los datos provenientes de un acelerómetro.

2.2.4. Método de recolección de datos

En la mayoría de los casos se especifica el método usado en la recolección de datos. Si no se especifica el método, este fue inferido indirectamente. Por ejemplo, cuando mencionan el uso de cuartos de calorimetría, se infiere que las condiciones están ligadas a un laboratorio. Así mismo, cuando se recolectan los datos en una casa o sitio del común, pero con alguna rutina dirigida, se toma como un método de recolección de datos semi-natural. La Tabla 2-10 muestra más información respecto a los métodos de recolección encontrados durante la revisión.

Método	Cantidad de artículos	Referencias
Libre o ambiente natural	14	[28], [30], [33], [34], [36], [40], [48], [64], [65], [67], [69], [71], [75], [79]
Laboratorio	21	[29], [33], [39], [43]–[46], [49], [50], [53], [54], [58], [59], [61], [62], [65], [66], [70], [72], [73], [77]
Semi-natural	11	[29], [38], [41], [47], [51], [52], [55], [56], [68], [77], [78]

Tabla 2-10. Método de recolección de datos.

Dos artículos [33] y [65] mencionan el uso de 2 métodos, natural y laboratorio. Según ellos, el de laboratorio es usado con el objetivo de entrenar los algoritmos clasificadores y obtener una buena precisión en el ambiente natural. En [29] y [77] usan los métodos de laboratorio y semi-natural y concluyen que los ambientes de laboratorio no permiten una apreciación tan buena de la vida cotidiana, como si puede permitirlo la recolección de datos de actividades realizadas con más libertad.

2.2.5. Precisión

La Tabla 2-11 muestra que, en general, la precisión en la clasificación de los comportamientos sedentarios (definidos como básicos: estar sentado y estar acostado), mediante el uso de acelerómetros utilizados singularmente o en combinación con giroscopios, inclinómetros y otros datos, es buena.

Posturas Estudias	N° de artículos que describen la precisión	Precisión mínima para la actividad (%)	Precisión máxima para la actividad (%)	media
Acostado	8	73 [35]	100 [35], [38], [59]	93,91
Parado	17	18 [38]	100 [35], [54]	87,1
Sentado	15	52,7 [29]	100 [58]	88,697

Tabla 2-11. Precisión en actividades de comportamiento sedentario generales.

2.2.5.1. Precisión en artículos que consideran actividades más específicas

Artículos como [28] recolectan datos y clasifican cinco actividades sedentarias (acostado, sentado, parado, bajando o subiendo en un ascensor), es decir dos adicionales a las comunes. Para esto se apoyan en datos provenientes de un acelerómetro, un giroscopio y un sensor de presión barométrica. Luego de desarrollar el procedimiento especificado identifican que la precisión en la identificación de esas actividades es, en general, 90.4%. Esta precisión es elevada en actividades como estar acostado (94.8%) o sentado (93.9%). Sin embargo, se presentan falencias en la clasificación de periodos mientras el participante se encuentra de pie (precisión para esta actividad 70.5%), y de pie usando un elevador (75.5% en el ascenso y 70% en el descenso).

los artículos [44], [52], [59] no escudriñan sobre la precisión en la clasificación de las actividades desarrolladas sino que usan datos de un acelerómetro u otras variables para establecer puntos de corte. Esto para que tales puntos de corte puedan ser usados en la distinción de 5 actividades sedentarias: estar sentado, leer un libro, ver televisión, usar un computador, jugar en la consola. Las actividades son realizadas en un ambiente controlado (laboratorio).

2.2.6. Datasets

De la revisión resulta que solo un artículo pone a disposición pública su dataset (relacionado a comportamientos sedentarios) [58]. Los investigadores, que conforma el WISDM (por las siglas en inglés “Wireless Sensor Data Mining”) Lab, usan el acelerómetro y el GPS disponibles en un smartphone, basado en la plataforma Android, para hacer la recolección de datos.

En la página web de WISDM, tienen a disposición 3 datasets, los cuales son:

- **Activity Prediction:** Este dataset data su última actualización hacia el 2 de diciembre del 2012. Los datos recolectados en este dataset, fueron realizados en un ambiente de laboratorio, y sus estadísticas se describen en la Tabla 2-12.

		Datos en crudo
Número de muestras		1.098.207
Número de atributos		6
Valores de atributo faltantes		Ninguno
Distribución de clases	Caminar	424.400 (38.6%)
	Trotar	342.177 (31.2%)
	Subir escaleras	122.869 (11.2%)
	Bajar escaleras	100.427 (9.1%)
	Sentado	59.939 (5.5%)
	Parado	48.395 (4.4%)

Tabla 2-12. Estadísticas de datos en "Activity Prediction"

- **Actitracker:** La última actualización de este dataset se realizó el 22 de octubre de 2013, y sus datos fueron recolectados en un ambiente natural. Para esto, ellos realizaron una aplicación presente en la Play Store, llamada actitracker. Las estadísticas de los datos recolectados se presentan en la Tabla 2-13. Además, datos demográficos en las cuales mencionan que el número de muestras fueron 563, el número de atributos 6 y ningún valor faltante.

		Datos en crudo
Número de muestras		2.980.765
Número de atributos		6
Valores de atributo faltantes		Ninguno
Distribución de clases	Caminar	1.255.923 (42.1%)
	Trotar	438.871 (14.7%)
	Escaleras	57.425 (1.9%)
	Sentado	663.706 (22.3%)
	Parado	288.873 (9.7%)
	Recostado	275.967 (9.3 %)

Tabla 2-13. Estadísticas de datos en "Actitracker"

También, proponen trabajos futuros para aprovechar otros sensores disponibles en el teléfono, como el micrófono, las cámaras, sensor de luz, sensor de proximidad, sensor de temperatura, sensor de presión, entre otros.

2.2.7. Dispositivos

2.2.7.1. Dispositivos utilizados en la recolección de los datos.

48 artículos mencionan el dispositivo usado. Entre estos, 2 fueron diseñados y construidos específicamente para el desarrollo del estudio, por tanto, no son descritos en la Tabla 2-14, ya que el análisis está dirigido a dispositivos de tipo comercial.

Debido a la cantidad de artículos revisados, 29 diferentes dispositivos fueron encontrados. Como se observa en la Tabla 2-14, los dispositivos más utilizados son el ActivPAL y el ActiGraph, los cuales son dispositivos que están diseñados para el reconocimiento de actividades principalmente físicas. Además, estos artefactos son considerados como fáciles de usar y portables, sin contar que, gracias a diversas validaciones en varios artículos, es comprobada su fiabilidad para medir características del movimiento y algunas posturas corporales. Luego se observan sistemas más elaborados y complejos que tienden a mejorar la precisión de las variables recolectadas. También, se observan sistemas con más sensores, capaces de recolectar un mayor número de variables y brindar más datos útiles en los procesos de clasificación.

Dispositivo	Variabes	Ubicación	N° de artículos que lo usan	Referencia
ActivPAL	Acelerómetro 1, 2 y 3 ejes, Inclinómetro (depende del modelo)	Muslo	13	[34], [36], [40], [51], [55], [57], [61], [64], [66], [67], [71], [75], [77]
ActiGraph	Acelerómetro 1, 2 y 3 ejes (depende del modelo)	Cintura	9	[44], [52], [59], [63], [65], [66], [71], [73], [74]
Smartphone	Acelerómetro, Giroscopio, Barómetro, entre otros, (depende del modelo)	Bolsillo	6	[30], [31], [42], [54], [58], [60], [78]
SmartShoe	Presión (fuerza), acelerómetro	Pies	4	[54], [56], [68], [70]
IDEEA	Acelerómetro	Muslos, pies y pecho	3	[39], [48], [63]

Movemonitor Dynaport	Acelerómetro	Cintura	2	[33], [38]
Actical	Acelerómetro	Cintura	2	[52], [63]
Cosmed K4B2	Intercambio pulmonar de gases, saturación de oxígeno	rostro	2	[53], [66]
Physilog 10D Silver GaitUP	Presión barométrica, acelerómetro, giroscopio	Muslo y pie	1	[28]
OpenBeacon RFID	Localización por RFID	No corporal	1	[55]
Fitbit Tracker	Acelerómetro	Muñeca	1	[63]
MotionLogs	Acelerómetro	Cadera	1	[29]
DirectLife	Acelerómetro	Muñeca	1	[63]
Oxycon mobile	Intercambio pulmonar de gases, saturación de oxígeno	Rostro	1	[44]
PAL2 (Gorman ProMed Pty. Ltd)	Acelerómetro	Arriba y abajo de la rodilla	1	[35]
Equival EQ-01, Hidalgo	Frecuencia cardiaca, frecuencia respiratoria	Pecho	1	[62]
ADXRS300	Acelerómetro, giroscopio	Pecho	1	[72]
SenseCam	Acelerómetro, temperatura ambiente, nivel de luz, video/imágenes, infrarrojo pasivo	Pecho	1	[79]
e-AR sensor	Acelerómetro	Oreja	1	[43]
Nokia wireless motion band	Acelerómetro	Tobillo	1	[45]
Tracmor	Acelerómetro	Espalda Baja	1	[47]
Freescale MMA7260Q	Acelerómetro	Pecho y cintura	1	[49]
M92962	Acelerómetro	Muslos	1	[80]

M92961	Acelerómetro	Pecho	1	[80]
Analog Devices ADXL202	Acelerómetro	Pecho	1	[51]
AMP-331	Acelerómetro	Tobillo	1	[52]
Shimmer	Acelerómetro	Pecho y muslo	1	[53]
Zephyr Sensor	Frecuencia cardiaca, frecuencia respiratoria, temperatura de la piel	Pecho	1	[53]
Body Media	Temperatura ambiente, temperatura de la piel, GSR	Brazo	1	[53]

Tabla 2-14. Dispositivos utilizados.

2.3. Proceso de obtención de los artículos incluidos en la revisión

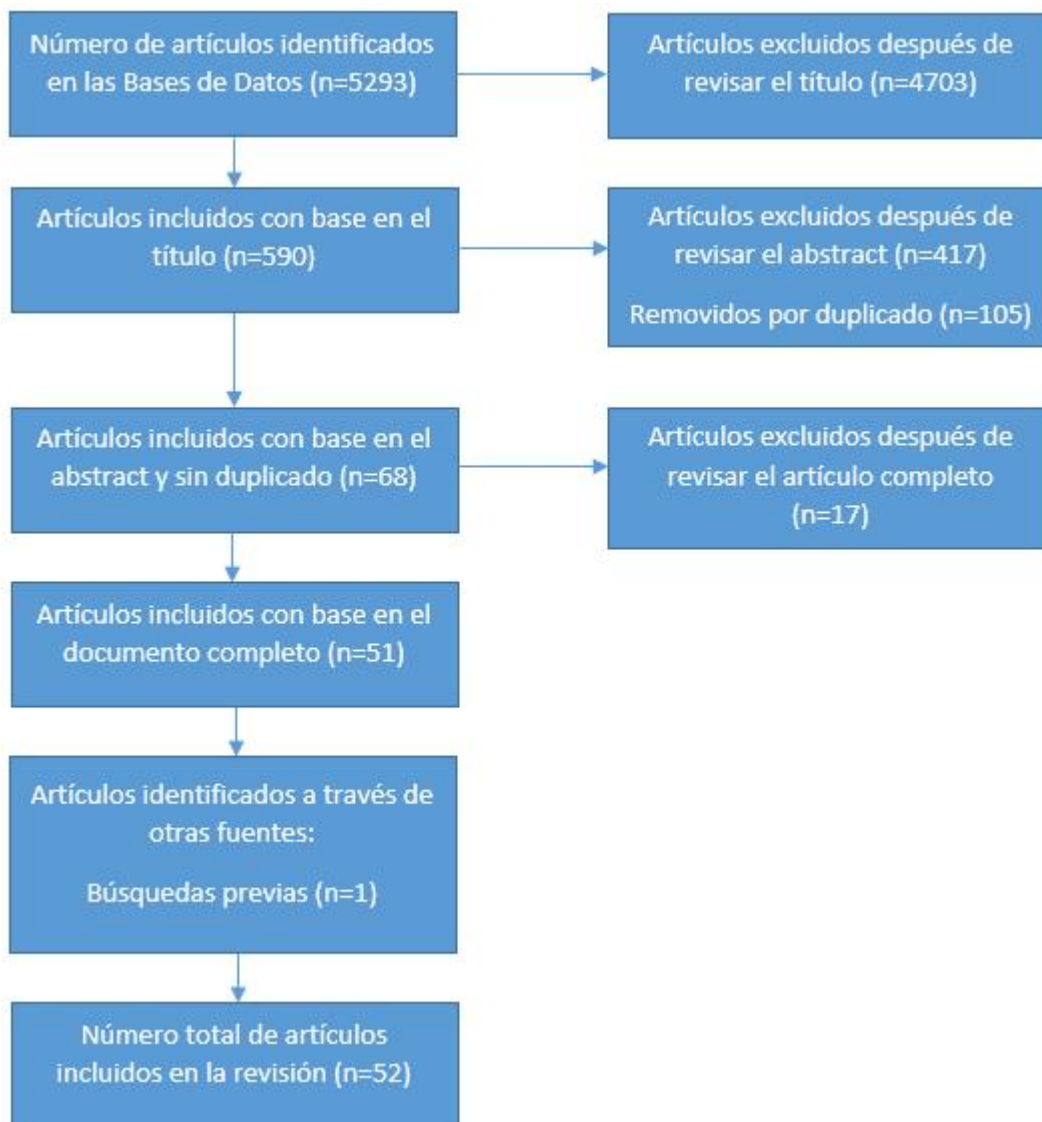


Figura 1. Proceso de obtención de artículos en la revisión

2.4. Discusión

El objetivo de esta revisión sistemática fue obtener información acerca de los sistemas usados para el reconocimiento de comportamientos sedentarios. Las características sobre las que se enfocó la búsqueda fueron: dispositivos, variables, sujetos y métodos considerados en la clasificación de CS. En los 52 artículos revisados, se identificaron 24 actividades que están ligadas a comportamiento sedentario y han sido clasificadas con los diferentes métodos, de las cuales los comportamientos sedentarios generales

(sentado, parado y acostado) han sido tratados en un buen número de estudios. En relación a los dispositivos se observó el uso de 29 tipos, los cuales miden 18 variables diferentes, principalmente acelerometría. En cuanto a datos de ubicación sólo 3 artículos consideran sensores de localización (incluyendo GPS).

2.4.1. Principales hallazgos y conclusiones

2.4.1.1. Sensores inerciales

La gran mayoría de los sistemas que realizan la clasificación de actividades en esta revisión, han utilizado sensores inerciales, en particular acelerómetros disponibles principalmente en el ActivPAL y ActiGraph. La combinación del acelerómetro con otros sensores es empleada constantemente. Aunque, algunos textos se limitan al uso exclusivo del acelerómetro [29], [33]–[36], [38], [42]–[45], [48], [49], [51], [52], [57]–[59], [61], [64], [65], [67], [71], [73]–[75], [77], [80]. Estos obtienen una buena precisión en la identificación de actividades, pero solo en las comunes. Un enfoque interesante es descrito por [31], [32]. Estos expresan que, optar por un enfoque de sensor único, lleva a proponer nuevas características con clasificaciones complejas para mejorar la precisión en el reconocimiento de actividades. Tales características avanzadas y las clasificaciones complejas exceden la capacidad de cálculo de la mayoría de las plataformas de sensores portátiles actuales. Por tal razón, terminan usando múltiples acelerómetros y la precisión para las actividades comunes está por encima del 96.4%.

2.4.1.2. Otras variables

Para realizar la clasificación de otros comportamientos sedentarios, diferentes a los comunes, suele usarse un sensor de presión barométrica [28], [30], [41], [78]. El barómetro es una herramienta utilizada para medir la presión atmosférica y con ello poder obtener la altura en la que se encuentra una persona respecto al nivel del mar, por tanto, esta medición ha sido usada con frecuencia para aportar en la clasificación de actividades que implican movimientos verticales (subir en un ascensor, usar las escaleras, etc.).

Los artículos [54], [56], [67], [70] 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 para ser estudiadas. Sin embargo, cada una de ellas es agrupada en una clase más general (sentado, parado, caminando/saltando y usando la bicicleta).

En cuanto a transmisión, los datos de los sensores en este estudio fueron enviados en tiempo real a través de Bluetooth a un smartphone. Una de las limitaciones en cuanto al tipo de dispositivos es que las personas no usan zapatos durante considerables períodos de tiempo mientras están en su hogar. El estudio es implementado en un ambiente de laboratorio. Así mismo, [60] 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).

En [55] se describe un sistema que se compone de dos partes: a) el monitor de actividad ActivPAL que detecta y registra los intervalos de tiempo en que se realizan actividades de estar sentado, de pie y 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 para localización, es integrado por tarjetas RFID localizadas de manera fija en una zona de oficinas, tarjetas 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 actividades desarrolladas por ellas.

De forma similar [50] 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 cuatro lugares (el sofá, la cama, el baño y una silla). La localización es conocida gracias a una serie de sensores de contacto, los cuales responden cuando los participantes se sientan o acuestan sobre ellos. Ligada a este estudio hay una limitación y es que la cantidad de sensores sería considerablemente elevada si se desea recolectar datos en una rutina diaria. Además, la estación receptora es un computador que está en constante comunicación con los sensores de presión y los acelerómetros. Esto implicaría que el equipo esté encendido todo el tiempo y que la persona lo llevase consigo siempre o que donde esté, haya un equipo listo para seguir con la captura de datos, cosa que en un ambiente real no sería viable.

En relación a la frecuencia cardíaca, en [17] se encuentra que esta aumenta cuando una persona pasa de estar sentada a pararse. Suceso que lleva a estimar este tipo de datos para hacer un sistema con más indicadores sobre una postura o una actividad.

En [34] y [66] se tienen en cuenta datos de consumo de oxígeno, que en combinación con acelerometría e información adicional de un Inclinómetro, identifican dos actividades (estar sentado o parado) con más del 98.1% de precisión. De manera similar [67] identifica, mediante sus datos (acelerometría y consumo de oxígeno), 4

actividades de tipo sedentarias: estar acostado, jugar en el computador, escribir a mano y lavar platos. Las estadísticas descriptivas muestran que los datos de consumo máximo de oxígeno (VO₂) están muy relacionadas a cada actividad. Incluso, para cada actividad la media de VO₂ es diferente. Esto implicaría altos beneficios cuando tales datos se tengan como entrada de un clasificador de actividades.

2.5. Conclusiones

Esta revisión sistemática identificó características relevantes en la clasificación de comportamientos sedentarios, dirigiendo la atención a datasets desarrollados y aspectos de investigación, como la ubicación del sujeto cuando ocurre tal comportamiento. Otros factores como los dispositivos utilizados y las variables consideradas, así como las actividades clasificadas en cada una de estas investigaciones también fueron tenidas en cuenta.

En general, los estudios examinan múltiples actividades de la vida diaria dentro de sus protocolos de experimentación. Sin embargo, estas se ven agrupadas en clases menos específicas, después de ser expuestas a los mecanismos de clasificación. En consecuencia, cuestiones como el contexto en que ocurren tales actividades es omitido. Además, se vislumbra la necesidad de mejorar la precisión de los sistemas de clasificación de comportamientos sedentarios, integrando en su dataset las diferentes variables que proporcionan los dispositivos wearables comerciales o smartphones en la actualidad. De ahí que diseñar un sistema que informe sobre el contexto en el que se presenta el comportamiento sedentario que posibilite la implementación adaptada en ambientes reales, se postula como uno de los retos cercanos en materia de reconocimiento de actividades de la vida diaria. Pero para lograr llevar a cabo un futuro proceso de clasificación automática preciso (haciendo uso del aprendizaje de máquina), es necesario tener datasets con la mayor cantidad de variables, que posibiliten el entrenamiento y evaluación de modelos de clasificación.

Adicionalmente, al crear estos datasets, son de vital importancia los participantes que hacen parte de ellos, por lo cual, como se observó, la variedad en género y edad puede ampliar las perspectivas del dataset, sobre todo si se mide la cantidad suficiente de individuos en cada uno de los grupos posibles.

CAPÍTULO 3.

MARCO TEÓRICO, DISEÑO DEL SISTEMA Y ESCENARIO DE RECOLECCIÓN DE DATOS

Este capítulo pretende describir algunas generalidades de la metodología seleccionada para desarrollar el proyecto, a saber, CRISP-DM. Además, el capítulo contiene el desarrollo de la primera fase propuesta en la metodología. Esta fase es la de entendimiento del negocio, y en ella se especifican: la evaluación de la situación, el inventario de recursos, requerimientos y aspectos fundamentales como el plan del proyecto. Adicionalmente, de este capítulo resulta una arquitectura inicial del sistema de recolección de datos. Esta arquitectura surge después de analizado el problema, los riesgos y criterios derivados de realizar la tareas y actividades de la fase de entendimiento del negocio propia de CRISP-DM.

3.1. Metodología CRISP-DM

Debido al crecimiento en el área de minería de datos 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 [82] 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. En este proyecto se opta por usar CRISP-DM por ser de uso libre y en la cual no se necesita más herramientas que las teóricas para poner en marcha un proyecto de minería de datos [83]. Además, es la metodología más utilizada actualmente según se puede observar en la Tabla 3-1; **Error! No se encuentra el origen de la referencia.**, posición que ha mantenido desde el año 2007 [84]. 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.

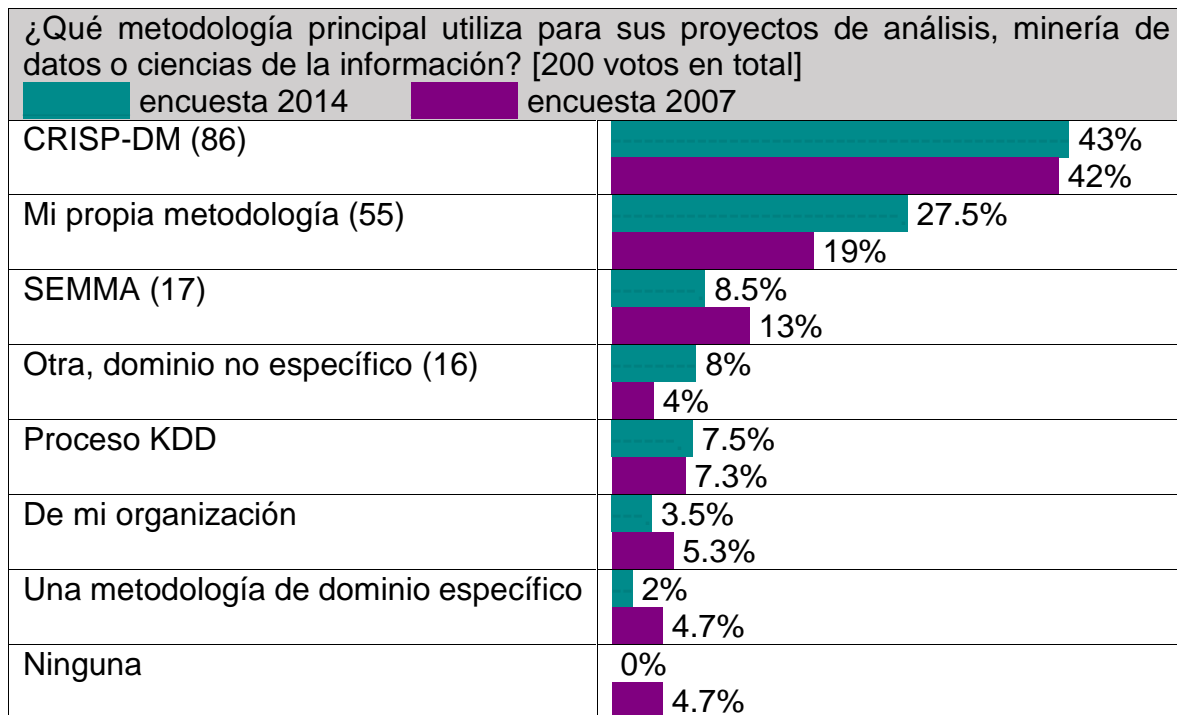


Tabla 3-1. Metodologías de minería de datos [84].

3.1.1. Generalidades

La metodología CRISP-DM está 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 [85].

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. *Completo* significa que cubre tanto al proceso entero de minería de datos y todas las aplicaciones de minería de datos posibles. *Estable* significa que el modelo debería ser válido para acontecimientos normales y aún para desarrollos imprevistos como 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.

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 hacer tareas anteriores repetidamente, así como repetir ciertas acciones.

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.

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 2. Estas no tienen una secuencia rígida y es comúnmente requerido el movimiento hacia adelante y hacia atrás entre fases diferentes.

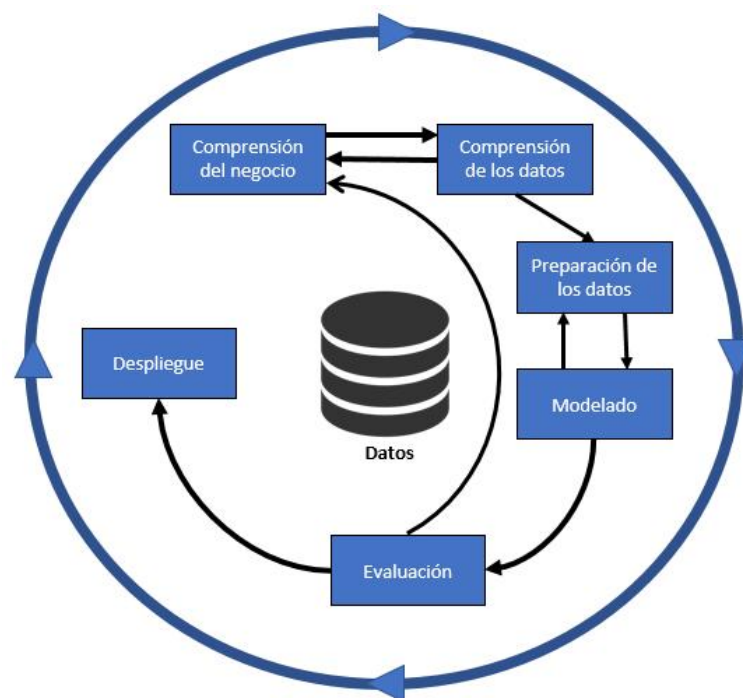


Figura 2. Ciclo de vida CRISP-DM [85].

Este proyecto es una implementación para fines de investigación y académicos, razón por la cual algunas actividades en cada fase no serán estrictamente desarrolladas. Este aspecto es posible por la flexibilidad de la metodología. Además, se aclara que no se hará uso de todas las fases de la metodología, llegando solamente hasta la fase de entendimiento de los datos, con la cual ya se tendrá el dataset, cumpliendo así el objetivo del proyecto.

3.1.2. Fase 1: Entendimiento del negocio

Como en la gran mayoría de metodologías de desarrollo de proyectos, la fase inicial es de vital importancia, ya que esta conlleva comprender qué es lo que necesita el

cliente y a qué se debe dirigir el desarrollo del proyecto. En la fase de entendimiento del negocio se pretende comprender los objetivos y requisitos del negocio, para posteriormente llegar a unos objetivos más técnicos (centrados en la minería de datos). De manera que finalmente surja un plan del proyecto, con el cual se tenga clara la dirección del proyecto. En consecuencia, después de hacer un análisis del contexto del trabajo de grado se obtiene el siguiente objetivo de negocio: Conocer características y patrones de los comportamientos sedentarios que permitan la creación de intervenciones en salud efectivas para la población. No obstante, este proyecto no cumplirá por sí solo con ese objetivo, sino que pretende ser la base de otros trabajos para alcanzarlo. La Figura 3 muestra más claramente las tareas y actividades generales de esta fase.

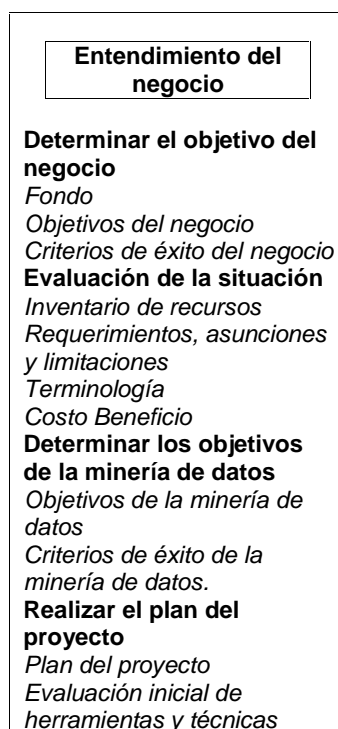


Figura 3. Entendimiento del negocio.

Dentro de esta fase se requiere elaborar un fondo (*background*), que consiste en registrar la información que se conoce sobre la situación de negocio al comienzo del proyecto. Esta actividad ha sido ejecutada en la introducción del capítulo 1 y en la revisión sistemática presentada en el capítulo 2.

Finalmente, en esta fase hay que definir los criterios de éxito del negocio. Estos deberían ser lo más específicos posible y capaces de ser medidos objetivamente. Luego, uno de los criterios de éxito fundamentales es: “el porcentaje en que se incurre en comportamientos sedentarios se ve disminuido en el grupo de personas que usen los sistemas e intervenciones que nazcan de usar los resultados obtenidos en este trabajo” y otro criterio de éxito es que “los investigadores encuentren en este trabajo soportes teóricos y técnicos para realizar sus proyectos de minería de datos en el

reconocimiento de comportamientos sedentarios mediante dispositivos móviles o wearables”.

3.1.2.1. 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 del análisis de los datos y en desarrollar el plan de proyecto.

3.1.2.1.1. Inventario de Recursos

A continuación, se presenta un **inventario de recursos** que lista aquellos recursos necesarios para desarrollar el proyecto y los criterios de selección de los mismos.

Arquitectura inicial

El sistema requiere módulos para recolectar datos sobre el contexto en que suceden los comportamientos sedentarios, así como capturar información de respuestas fisiológicas y datos sobre el movimiento de los participantes. En la Figura 4 se muestra una arquitectura inicial del sistema.

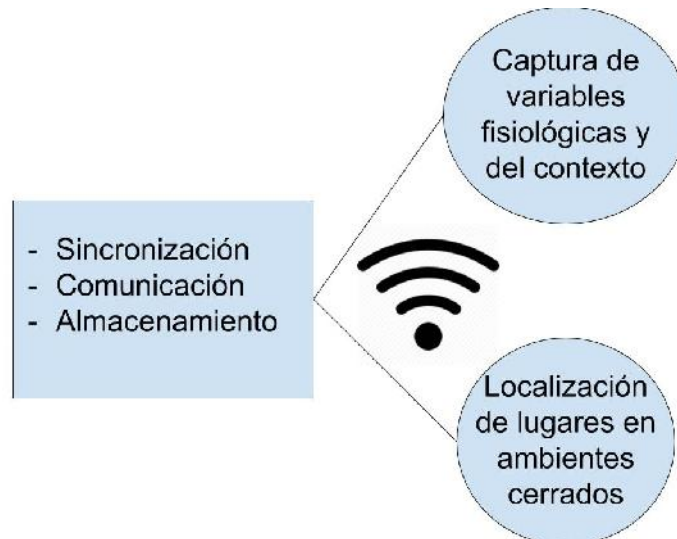


Figura 4. Arquitectura inicial del sistema.

Una de las características deseadas del sistema es que se componga de dispositivos no intrusivos y/o no invasivos, de manera que su implementación pueda llevarse a cabo en un ambiente real. De ahí que la miniaturización de sensores y dispositivos electrónicos se convierta en un enfoque importante dentro del proyecto, ya que hace posible incluir en la ropa, teléfonos móviles y accesorios de las personas una serie de características útiles para realizar seguimiento del comportamiento, actividad física y fisiológica de las mismas [86]. Esas propiedades son potencialmente explotables para construir el sistema esquematizado, todo mediante el uso de wearables y otros

dispositivos que con el tiempo serán de uso común en la población mundial. En la siguiente sección hay una serie de criterios que llevan a seleccionar los dispositivos y tecnologías más adecuados para este trabajo.

Recursos Hardware

- **Smartphone**

Este dispositivo, como se puede observar en la revisión sistemática, ha sido ampliamente usado en proyectos relacionados al reconocimiento de actividad física y actividades de la vida diaria en general. El resultado de su utilización en las diferentes arquitecturas ha sido bueno, a tal punto, que ha sido validado, aceptado y recomendado para trabajos. Algunas ventajas importantes son: el fácil acceso a internet (por wifi o datos), además de la alta capacidad de almacenamiento/procesamiento de datos (característica que lo diferencia de los wearables), y el uso de varios sensores que suelen tener incorporados. De manera que, el smartphone se convierte en una fuente de datos que, en combinación con los datos de otros sensores externos y mediante la aplicación de diferentes mecanismos, permitiría tener una visión más amplia del comportamiento de los sujetos y su contexto. Además, hoy en día el smartphone es un artículo de uso cotidiano, a tal punto que es casi una extensión de las personas. Incluso, es posible observar a individuos realizando todas sus tareas del día a día sin sentirse incómodos con este dispositivo electrónico. Gracias a las cualidades presentadas anteriormente y a que puede cumplir todos los requerimientos de conectividad necesarios entre los diferentes dispositivos del sistema a desarrollar, el smartphone es un componente esencial dentro de la arquitectura del proyecto.

Para la elección del dispositivo a utilizar se plantean los siguientes requerimientos, los cuales serán independientes de alguna marca en especial:

- a. Debe contener el sistema operativo Android, ya que este sistema operativo es el más usado alrededor del mundo [87], 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 en la etapa de recolección de datos. Además, entre los requerimientos definidos en algunas de las bandas inteligentes del mercado, se encuentra la compatibilidad solo con versiones del sistema operativo Android superiores al 4.4.
- b. Sus conexiones a una red Wi-Fi o de datos, o mediante cable usb deben estar funcionando óptimamente para permitir la transmisión de los datos.
- 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 y el giroscopio, dos sensores que por su naturaleza prometen ser adecuados para el reconocimiento de comportamientos sedentarios.

- **Dispositivos wearables comerciales (comparación y selección)**

Existe en el mercado varias alternativas de wearables con características prometedoras. La Figura 5 presenta los wearables comerciales más comunes y mejor equipados en cuanto al número de sensores disponibles, según las especificaciones técnicas presentadas en sus sitios web oficiales. La Figura 6 lista los mismos wearables, pero haciendo referencia a los sensores que están disponibles para desarrolladores.

CARACTERÍSTICAS DISPOSITIVO	Sensores disponibles según su página web													D. batería/GPS	
	Frecuencia cardíaca	Acelerómetro	Giroscopio	Temperatura de la piel	Respuesta galvánica de la piel	Radiación UV	Altimetro	Barómetro	Luz ambiente	GPS	Compás	Magnetómetro	Micrófono		Respiración
Microsoft Band	X	X	X	X		X							X		2D/ 5H
Microsoft Band 2	X	X	X	X	X	X	X	X					X		2D/ 5H
Fitbit Charge IIR	X	X					X								5D
Fitbit Surge	X	X	X				X		X	X		X			7D/ 10H
Apple Watch	X	X	X						X						18H
Pebble Time		X	X						X		X	X	X		7D
Garmin Vivoactive HR	X	X						X		X		X			8D/ 13H
Samsung Gear S	X	X	X			X		X	X	X		X			1-2 D
Jawbone UP3	X	X			X									X	7D

Figura 5: Sensores disponibles en los wearables.

CARACTERÍSTICAS DISPOSITIVO	Sensores disponibles para desarrolladores												
	Frecuencia cardíaca	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												

Figura 6: Sensores disponibles para desarrolladores.

De la Figura 5 y la Figura 6, se observa que algunos dispositivos tienen una gran variedad de sensores, pero sus datos no están disponibles por medio del SDK o API Web para quienes necesiten desarrollar sistemas que incluyan tal información.

El primer dispositivo en las figuras es la Microsoft Band [88]. Esta fue lanzada, en su primera versión, a finales del 2014 y en cantidades limitadas. Este wearable incluye GPS, pero esta característica no es alcanzable haciendo uso del SDK dispuesto en el centro de desarrolladores de Microsoft [89]. Un rasgo muy importante de este dispositivo es la compatibilidad, pues puede ser usada con móviles que cuenten con sistema operativo Android, Windows o iOS. La versión 2 de esta banda fue lanzada un año después y cuenta con sensores adicionales como GSR (Galvan Skin Response), barómetro y sensor de luz ambiental.

Fitbit para desarrolladores [90] propone una API Web, desde la cual puede accederse a los datos en crudo de la frecuencia cardíaca, los datos de calorías quemadas, distancia, pasos, minutos sedentarios, minutos de actividad leve, actividad moderada y alta, número de pisos subidos, velocidad y elevación. Por otra parte, los datos del acelerómetro, GPS, giroscopio, brújula digital, altímetro y sensor de luz ambiental (con los cuales cuenta la versión Surge) no son accesibles desde la API Web.

Garmin vivoactiveHR [91] a través de connect IQ SDK, proporciona datos en crudo del monitor de frecuencia cardíaca, altímetro y termómetro, pero no del acelerómetro. En lugar de eso, presenta resultados del procesamiento de dichos datos, es decir,

velocidad, el número de calorías quemadas, el número de pasos y la distancia recorrida.

Apple Watch [92], por su parte no tiene compatibilidad con dispositivos como Android o Windows. Lo cual se convierte en una limitación importante en regiones del mundo donde los sistemas más utilizados en dispositivos móviles son Android.

Misfit [93] provee una API Web mediante la cual se puede acceder a datos como las calorías, pasos y distancia recorrida. Sin embargo, datos crudos no son utilizables.

El siguiente de la lista es Samsung gear S. Este dispositivo presenta una amplia gama de características explotables. Mediante la documentación de su sistema operativo TIZEN es posible acceder a los datos en crudo de todos sus sensores (véase la Figura 6). Este dispositivo es compatible, inicialmente, con versiones de Android superiores a 4.3 y con móviles específicos de Samsung; aunque recientemente (no oficialmente) en Internet se han filtrado formas de sincronizarlo con sistemas iOS. Aun así, comercialmente es adquirido por aquellos que tienen un móvil Samsung compatible con el dispositivo, lo cual hace que se limite el alcance en cuanto a las personas que se benefician con el desarrollo de aplicaciones con este wearable.

Jawband [94] permite, a los desarrolladores, el acceso a datos en crudo de la frecuencia cardiaca a través de su Web API, pero al igual que en otros dispositivos es el único dato en crudo disponible. Mediante la API Web se puede acceder a distancia recorrida, número de pasos, tiempo activo, tiempo inactivo, calorías, BMR (por sus siglas en inglés Basal Metabolic Rate) estimado y velocidad.

Como conclusión de este repaso de tecnologías disponibles en el mercado, se ve conveniente utilizar la Microsoft Band 2 para este trabajo. Esta banda es apropiada por varias razones: Primero, el dispositivo pone a disposición de los desarrolladores una amplia gama de características útiles en el seguimiento de comportamientos sedentarios. Con respecto a esto, el número de sensores disponibles en la banda solo se compara con la cantidad que tiene el Samsung gear S. Segundo, uno de los fuertes de la banda es la compatibilidad con dispositivos Android, iOS y Windows Phone, característica que no aplica al Samsung gear S. Finalmente, al comparar el precio de la Microsoft Band 2 y el Samsung gear S es notable que: la primera tiene un valor mucho menor y tiene la capacidad de prestar un servicio adecuado para el proyecto.

- **Microsoft Band 2**

Para este proyecto es seleccionada la versión 2 de este wearable. Según la documentación oficial de este dispositivo, la banda cuenta con un número considerable de sensores que, mediante el SDK proporcionado a desarrolladores, pueden ser usados para desarrollar cualquier tipo de aplicaciones. La Tabla 3-2 muestra los sensores disponibles, algunos detalles y la frecuencia con la que recolectan los datos. Hay que anotar que algunos de los sensores incluidos en la Tabla 3-2 son sensores llamados subyacentes. Al igual como lo hace Android en sus especificaciones, en la Microsoft band un sensor subyacente es un sensor que obtiene los datos de un sensor físico y calcula directamente una variable específica. Por ejemplo, desde las lecturas del acelerómetro, que es un sensor físico, se obtiene el podómetro, el sensor de distancia y de calorías quemadas.

Sensor	Detalles	Frecuencia	Unidades
Acelerómetro	Provee la aceleración X, Y y Z en unidades g (9.81 m/s ²)	62/31/8 Hz	m/s ² (metros por segundos al cuadrado)
Giroscopio	Provee la velocidad angular en grados por segundo.	62/31/8 Hz	°/seg (grados por segundo)
Distancia	Provee la distancia total en centímetros y la velocidad en centímetros por segundo (cm/s).	1 Hz	cm (centímetros)
Frecuencia cardíaca	Proporciona el número de latidos por minuto. Además, indica si el sensor está completamente acoplado.	1 Hz	Latidos/min
Podómetro	Proporciona el número total de pasos que el usuario ha dado desde que la banda empieza a funcionar.	Cambio de valor	Número de pasos desde el último reset.
Temperatura de la piel	Provee la temperatura de quien la usa en grados centígrados.	1 Hz	° (grados en Celsius)
UV	Proporciona la intensidad de exposición a la radiación ultravioleta actual.	1 Hz	
Contacto de la banda	Proporciona el estado actual de la banda como siendo usado/no usado.	Cambio de valor	Usando / No Usando
Calorías	Proporciona el número total de calorías que el usuario ha quemado desde que la banda empieza a funcionar. Este es un contador de por vida y no un contador diario	Cambio de valor	kcal (kilocalorías)

	o un contador que vuelva a 0.		
Respuesta galvánica de la piel	Proporciona la resistencia de la piel del usuario en kohmios	0.2/5 Hz	Kohms (kilo ohms)
Intervalo RR	Proporciona el intervalo en segundos entre los dos últimos latidos del corazón continuos. Estos datos se deben usar sólo en el modo de reposo.	Cambio de valor	seg (segundos)
Luz ambiente	Provee la intensidad de luz en unidades Lux (Lumen/m ²)	2 Hz	Lux (lumes por metro al cuadrado)
Barómetro	Proporciona la presión del aire en hPa (hectopascales) y la temperatura en grados Celsius	1 Hz	hPa (hectopascales)
Altímetro	Proporciona datos de elevación como ganancia/pérdida total, escalones ascendidos/descendidos, vuelos ascenso/ descenso, y tasa de elevación	1 Hz	m (metros)

Tabla 3-2. Sensores de la Microsoft Band 2.

- **Comparación y selección de dispositivos para obtención de localización**

La localización es un factor muy importante dentro del sistema, ya que, conociendo el contexto en el cual se encuentra la persona, sería posible clasificar de una manera más precisa los comportamientos sedentarios. En la revisión sistemática se obtuvo dos sistemas que permiten obtener la localización de la persona y de forma simultánea detectan algunas actividades físicas o sedentarias. Uno de ellos lo hace mediante múltiples sensores de presión, por tanto, los participantes del estudio necesariamente debían entrar en contacto con cada sensor para que el sistema detecte la localización. Esto implicaría una cantidad elevada de sensores, uno por cada silla, sofá, cama o localización que se desee identificar. Por consiguiente, este proyecto busca superar esta limitación mediante un sistema que permita conocer la localización de la persona de la manera menos perceptible posible. Así pues, el medio de comunicación más conveniente a implementar es el inalámbrico. Luego, las siguientes tecnologías son exploradas:

a) Wi-Fi

Wi-Fi es una tecnología que utiliza el protocolo TCP/IP como protocolo de comunicación. Una red Wi-Fi cuenta con uno o más puntos de acceso a los que se conectan los terminales de la red. Esta tecnología está pensada para una cobertura local, es decir, algunos cientos de metros. Además, soporta redes con uno o muchos ordenadores [95].

Algunos investigadores proponen un sistema para el reconocimiento de actividades (AR) apoyado en datos provenientes de una topología Wi-Fi y acelerómetros. La topología se usa para identificar localización mientras los datos del acelerómetro se usan para el reconocimiento de posturas corporales y actividad motora [86]. En ese sistema son usados cuatro puntos de acceso, los cuales son puntos de referencia para inferir dónde se encuentra cada sujeto. Esta cantidad, vista desde un ambiente domiciliar, es demasiada para este tipo de sistemas, pues no se estarían utilizando todas las características con las que cuenta cada estación o punto de acceso. Además, en ambientes domiciliarios no es común tener más de un punto de acceso. Esto, sumado al gasto eléctrico que trae el uso de múltiples artefactos y la necesidad de conexiones cableadas entre los dispositivos y con las fuentes de electricidad, llevan a descartar el uso de esta tecnología en este trabajo (para cuestiones de localización).

b) Bluetooth

Esta tecnología fue creada en 1994 con el objetivo de concebir una tecnología alternativa de comunicación inalámbrica que permita la colaboración entre diferentes industrias y productos. Bluetooth fue creado como un estándar abierto con avances constantes y amplia aceptación. Gracias a esto, se desarrolló el nuevo Bluetooth Low Energy (BLE). BLE, como su nombre lo indica, produce un bajo consumo de energía, propiedad que le ha permitido ser implementado en gran cantidad de pequeños sensores capaces de funcionar por meses e incluso años [96]. Uno de los dispositivos que utiliza BLE y cuyo uso está en aumento vertiginoso es el beacon. Un beacon brinda reportes sobre localización con gran exactitud en lugares cerrados e incluso en exteriores gracias a su cobertura de hasta 50 metros. Cada beacon tiene un identificador (ID) único, así que, al implementarlos en un área común es posible diferenciarlos. El único requerimiento de este sistema es una aplicación alojada en un smartphone, tablet o pc con la que pueda interactuar [97].

c) RFID

Tal como el reconocimiento del código de barras, RFID se basa en objetos etiquetados con el fin de identificarlos [98]. A diferencia del código de barras, las etiquetas “tags” RFID no necesitan ser visibles para poder ser leídos. Un lector RFID envía una señal

de radio de corto alcance, la cual es recogida por el tag RFID. Luego, el tag retransmite una cadena corta de datos para su identificación en el sistema. Este enfoque es comúnmente usado para el monitoreo de localización y movimiento entre áreas referenciadas con estos tags. También se usa en combinación con monitores de actividad como el ActivPAL, el cual puede caracterizar periodos en los que se está sentado, parado o caminando. Esto con el objetivo de identificar la relación entre la localización y el tiempo en que permanece sentado o realizando actividad física leve [55].

Sin embargo, uno de los problemas de este enfoque es la proximidad análoga, es decir, el corto alcance del lector para identificar el tag. Esto, en un escenario de la vida cotidiana, supondría que para identificar el cambio de un área a otra; quien use el sistema tenga que acercarse al lector, de modo que se interrumpe la cotidianidad de las actividades del sujeto; o en su defecto, utilizar lectores RFID activos, los cuales para su mutua detección permiten mayores distancias, requieren de ubicar estos dispositivos en el sitio y que además cada persona cargue un tag RFID para que sea detectado.

d) Near field communication (NFC)

Es una forma de comunicación sin contacto entre dispositivos como Smartphone o tablets. Esta permite enviar información entre dispositivos, únicamente acercándolos, sin necesidad de que haya contacto entre ellos o realizar pasos previos como establecer una conexión, lo cual es necesario en otras tecnologías [99]. Pero al igual que los tags RFID, el corto alcance sería un limitante para el desarrollo del proyecto.

e) GPS

Originalmente desarrollado por el Departamento de Defensa de Estados Unidos, el sistema GPS consta de 24 satélites que orbitan la Tierra. Estos satélites transmiten señales a los receptores GPS y son capaces de determinar la localización, dirección y velocidad del receptor basado en trilateración entre tres o más satélites. Sin embargo, incluso en las mejores circunstancias, el GPS no provee la precisión suficiente para determinar la localización de personas en entornos cerrados tales como habitaciones, salas y demás lugares dentro de una casa u oficina [15]. De manera que, esta tecnología no se adapta a las necesidades de este trabajo.

f) Red celular

Esta se centra en la comunicación que tienen los dispositivos celulares o Smartphone con la central de comunicación a la cual se encuentra registrado. La telefonía celular utiliza la radiotransmisión mediante ondas, emitiendo y/o recibiendo información de la

antena más próxima o vía satélite. Esta recibe el nombre de celular debido a su arquitectura, ya que la cobertura que ofrece cada una de las antenas se relaciona a una célula y presta servicio a todos los usuarios presentes en esta área. Desde la tercera generación, este medio permite la convergencia de voz y datos teniendo acceso a internet inalámbrico, lo cual ha permitido el gran avance de los smartphome, sin contar la amplia cobertura alrededor del mundo que ofrecen este tipo de redes de comunicación [100]. Al igual que el GPS, esta tecnología no provee la precisión necesaria para el desarrollo de este trabajo.

Como resultado de la exploración de las anteriores tecnologías candidatas para conseguir una referencia de localización, se opta por el uso de los beacons. Estos dispositivos usan la tecnología bluetooth para comunicarse con los smartphome y así informar su localización. La selección se realiza después de considerar que los beacons son pensados especialmente para comunicaciones de corto y mediano alcance y sus mediciones en entornos cerrados es bastante precisa. De modo que solventa el problema relacionado a la medición de la localización en entornos cerrados.

- **Disponibilidad de las herramientas**

Para este proyecto se cuenta con dos bandas de Microsoft. Una es el prototipo o versión 1 (2014) que fue lanzada en un número limitado y la otra es la versión 2 (2015). La segunda versión presenta más sensores disponibles y su pantalla, al ser curva, es más cómoda para llevar. En caso de ser necesario puede ser usada las 24 horas al día (sin hacer uso del GPS). Incluso una carga completa puede durar hasta 48 horas, aunque puede variar significativamente dependiendo de la configuración, el uso y otros factores. Otro factor que puede ser útil en el diseño del experimento es el tiempo de carga completa, el cual es de menos de 1.5 horas.

Por otra parte, se cuenta con 6 beacons, los cuales tienen poco tiempo de uso y por tanto tienen la capacidad para funcionar por muchos meses con su batería original.

Además, para el funcionamiento del sistema, como ya se ha mencionado, se hace necesario el uso de un Smartphone que se usará para conectarse con la Microsoft band 2 y recolectar datos de sus sensores. El teléfono principal disponible para este proyecto es un LG G3. Adicionalmente, un Smartphone Huawei Ascend P7 es incorporado al estudio, en algunas secciones se refiere a él como secundario o adicional. La elección de estos modelos fue inherente a la marca y en comparación con los equipos disponibles se destacan cosas como: una elevada cantidad de sensores y mejores características de procesamiento y almacenamiento. Los demás dispositivos a los que se tenía acceso son: Motorola G1, Nexus S, Nexus 4, Huawei P6 y Avvio 795. Sin embargo, las versiones de Android no eran lo suficientemente actuales para sincronizarse con la Microsoft band 2 o para detectar los beacons BLE.

Dispositivos seleccionados

Con el inventario de recursos realizado, es posible mostrar de una forma más clara la arquitectura del sistema a implementar. En la Figura 7 se presenta la arquitectura resultante.

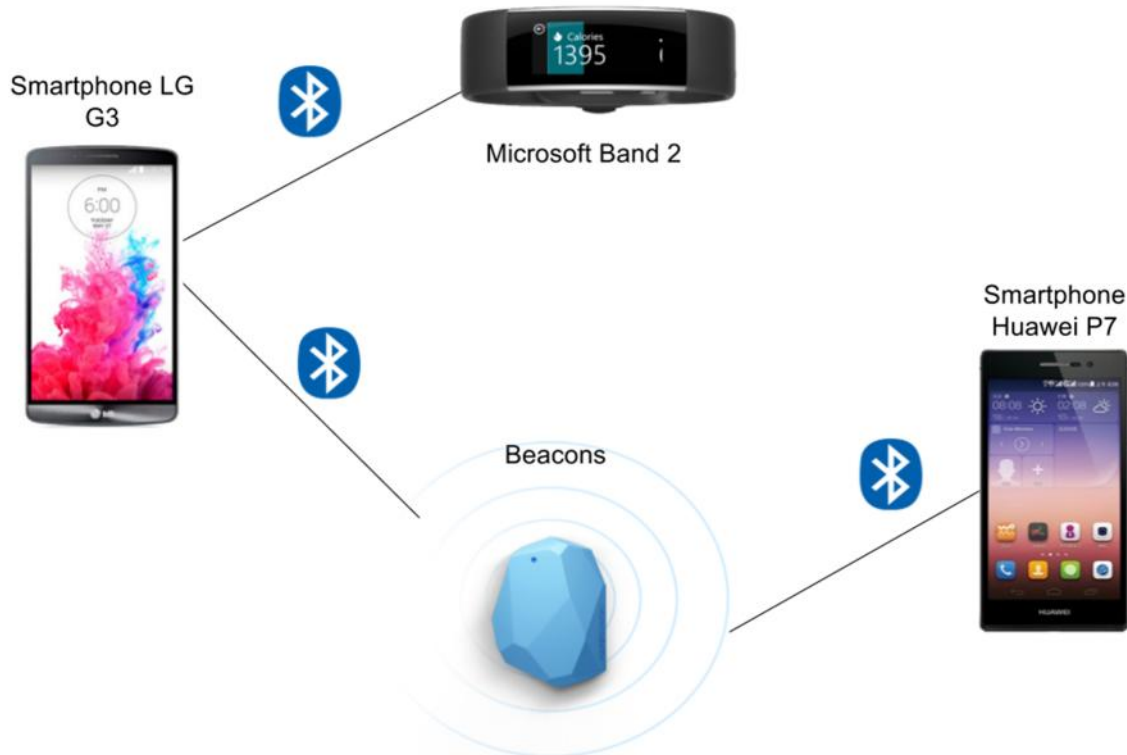


Figura 7. Arquitectura del sistema

A continuación, se describen las características de los dispositivos usados, obviando la Microsoft band 2 que fue descrita anteriormente.

- **Huawei Ascend P7**

El Huawei Ascend P7 hace parte de la gama alta de la línea de celulares de Huawei debido a su diseño y sus características [101]. Entre las características más importantes de este terminal se encuentran: cámara, procesamiento, memoria, conectividad, entre otros. En relación al proyecto, las características que sugieren su elección son las siguientes:

- Tamaño: Largo 139.8mm, Ancho 68.8mm y Grosor 6.5mm. Con lo cual se pretende que sea poco intrusivo para el usuario.
- CPU: Hisilicon Kirin 910T, Quad-Core 1.8 GHz.
- Sistema Operativo: Android 4.4.2 KitKat (Lo cual significa compatibilidad con la mayoría de wearables en el mercado, incluida la Microsoft Band 2).

- Memoria: RAM de 2GB y ROM de 16GB
- Conectividad: Wifi 802.11 b/g/n, Bluetooth 4.0 inteligente y Micro USB 2.0.
- Sensores: G-sensor, Acelerómetro, Sensor de proximidad, Sensor de luz ambiental y Compás.
- Batería: 2500mAh

- **LG G3**

Este celular fue galardonado en el 2015 como el mejor celular del mundo en el Global Mobile Awards debido a su gran avance y perfecto diseño [102]. Aunque haya pasado algún tiempo, LG G3 continúa siendo uno de los mejores celulares del mercado. Las características por las cuales fue seleccionado para esta investigación son:

- Tamaño: Largo 146.3mm, Ancho 74.6mm y Grosor 8.9mm
- CPU: Qualcomm Snapdragon Quad-Core 2.5Ghz
- Sistema Operativo: Android 5.0 Lollipop
- Memoria: RAM de 2GB y ROM de 16GB
- Conectividad: Wifi 802.11 a/b/g/n/ac, Bluetooth 4.0 inteligente y Micro USB 2.0
- Sensores: G-sensor, Acelerómetro, giroscopio, Sensor de proximidad y luz, Sensor de luz ambiental, Compás y Barómetro.
- Batería: 3000mAh.

- **Beacons Estimote de Proximidad**

Un Beacon Estimote es un pequeño computador [103]. Este elemento está conformado por un acelerómetro, sensor de temperatura y un módulo Bluetooth low energy. Estos dispositivos han sido ampliamente utilizados por desarrolladores, principalmente en el área de turismo. Eso se debe a que sirven como guías dentro de un edificio. Además, esta clase de beacons tienen la capacidad de almacenar poca información como una descripción del objeto al que se encuentran cerca o relacionados o una identificación regional.

Un aporte importante de Estimote es la aplicación que proveen para gestionar o configurar cada beacon. Pero sin lugar a duda, uno de los factores que más interesaron en la elección de estos dispositivos es el código que provee Estimote para desarrollos en Android y iOS. El código al que se tiene acceso es suficiente para crear nuevas y novedosas aplicaciones.

Las características con las que cuentan estos dispositivos son:

- Un CPU ARM Cortex de 32-bit
- Acelerómetro
- Sensor de temperatura
- Bluetooth 4.0 Inteligente
- Radio de hasta 70 metros en línea de vista.
- 3 años de batería en su configuración por defecto

Fuentes de datos y conocimiento

Las fuentes de datos usadas para este proyecto serán las disponibles en línea, charlas con expertos y documentación escrita que pueda encontrarse. Dichas fuentes serán las más formales y reconocidas. El motor de búsqueda de artículos científicos de Google (Google scholar) será usado con frecuencia para extraer información relevante como conceptos, técnicas o metodologías usadas en proyectos similares.

Fuentes personales

Patrocinador del proyecto: El proyecto es solventado por las personas encargadas del mismo, así como por la Universidad del Cauca. Esta institución financia la dirección y asesoramiento del proyecto por parte de docentes e investigadores competentes y con experiencia en el tema. Además, brinda los espacios necesarios dentro del campus universitario (aulas, laboratorios, etc.). En relación al desempeño de roles que se involucren en el desarrollo del proyecto (administrador del sistema, administrador de base de datos, soporte técnico y entre otros) serán desempeñados por los dos estudiantes encargados del trabajo de grado, contando siempre con la guía y asesoría del director y del asesor.

3.1.2.1.2. Requerimientos, supuestos y limitaciones

En esta sección se listan todos los requerimientos del proyecto, incluyendo el calendario de finalización, comprensibilidad y calidad de los resultados, la seguridad y cuestiones legales. Además de indicar las presunciones hechas para el proyecto. Estas pueden ser presunciones sobre los datos que pueden ser verificados durante el proceso de minería de datos, pero también puede incluir presunciones no-verificables sobre el negocio. Las limitaciones del proyecto también son abordadas en esta sección. Estas pueden ser limitaciones en cuanto a disponibilidad de los recursos, pero además pueden incluirse limitaciones tecnológicas.

Requerimientos

- **Grupo objetivo, factores y elementos de calidad.**

El dataset debe incluir datos de un buen número de sujetos con divergencia en género, edad, altura y peso. Esto debido a que es sumamente importante contar con un número considerable de sujetos de evaluación, con el fin de obtener resultados fiables. Además del número de sujetos, para la construcción de buenos modelos para el reconocimiento de actividad, la población objetivo debe ser lo más diversa posible [104][105].

El dataset debe documentar claramente los siguientes factores:

- Número de sujetos y diversidad
- Metodología de recolección de datos
- Sensores y características
- Funciones y generación de funciones

Además de contar con los siguientes elementos:

- **Linaje:** Definido como la información acerca del proceso de creación del dataset, por ejemplo, los instrumentos usados para hacer la recolección de datos, identidad de los responsables de la creación y el estándar definido para seleccionar los contenidos del dataset.
- **Integridad/completitud:** El grado en el cual el dataset captura todos los datos esperados.
- **Consistencia lógica:** Esto se refiere a la consistencia interna de los datos y la adherencia del dataset a sus propias reglas definidas.

Con estos elementos se asegura que el dataset construido pueda ser comprendido por quien lo desee.

- **Mantenibilidad**

La arquitectura realizada es altamente mantenible, gracias a que todos sus dispositivos son comerciales. Esto significa que cualquier persona puede adquirir tales elementos fácilmente. Además, no hay una restricción a los dispositivos descritos y utilizados en este proyecto.

Por otra parte, uno de los requerimientos de la aplicación móvil a desarrollar para el funcionamiento del sistema, es su modularidad, con lo cual la persona que la use, pueda hacer los cambios que necesite fácilmente.

Supuestos

- **Generales**

- Una sola Microsoft Band 2 es suficiente para desarrollar todo el proceso de medición.
- La capacidad de los dispositivos donde se aloje la información será suficiente.
- El tamaño de la Microsoft Band 2 será el adecuado para el tamaño de la muñeca de las personas involucradas en la recolección de datos.
- El smartphone a utilizar tiene las características software y hardware suficientes para realizar la sincronización y comunicación con los dispositivos, almacenar información necesaria y recolectar las variables requeridas.

- La cantidad de beacons será suficiente para obtener datos de los sitios y dispositivos en uso necesarios.
 - La aplicación que se desarrollará para realizar la recolección de los datos será capaz de diferenciar y categorizar las actividades y comportamientos sedentarios que se realizarán.
 - Habrán 30 o más personas que cumplan los requerimientos y estén dispuestas a hacer parte de la recolección de datos.
 - Los dispositivos a usar durante el proyecto no supondrán problemas de funcionamiento.
 - Entre más variables sean recolectadas e incorporadas al dataset, este será más útil y producirá mejores resultados en sus futuros usos.
- **Supuestos sobre calidad de datos**
 - Los dispositivos usados proveerán datos sin errores causados por mal funcionamiento de los sensores o por fallas de comunicación entre dispositivos.
 - El proceso usado en el desarrollo del proyecto permitirá obtener una buena calidad en el dataset.
 - El dataset contendrá todos los datos esperados en la etapa de planeación.
 - Los datos presentes en el dataset tendrán una lógica consistente, de tal manera que no se encuentren valores arbitrarios o fuera de sus intervalos.
- **Supuestos sobre factores externos**
 - La adquisición de los dispositivos (compra y envío) se llevará a cabo sin problemas y contratiempos.
- **Supuestos que conducen a alguno de los estimados**
 - El costo presentado de algunos de los equipos se asume dependiendo de algunas recomendaciones hechas para este tipo de proyectos (en general porcentajes del valor de compra del equipo en su momento). Esto debido a que varios de estos dispositivos no son adquiridos especialmente para este proyecto y mantienen un valor comercial variable.
- **Supuestos respecto a si es necesario entender y describir o explicar el modelo**
 - Aquellos que necesiten explorar el dataset construido en este proyecto deberán tener claro algunos conceptos de minería de datos como: atributos, clases y precisión. Además, de tener conocimientos sobre variables fisiológicas como la frecuencia cardíaca, el intervalo RR, GSR y temperatura de la piel, así como también variables de aceleración, presión y velocidad. En caso de que no tengan tales conocimientos se asume que la documentación

resultante de este trabajo será suficiente para entender y usar el dataset construido.

Limitaciones

- **Generales**

- El presupuesto para este proyecto es limitado, por esta razón el número de beacons es pequeño y solo se tiene a disposición una Microsoft band 2. Tales aspectos representan una limitante en cuanto al tiempo necesario para recolectar los datos en una población relativamente grande y adecuada.
- Encontrar la cantidad suficiente de personas con las características necesarias sugeridas en los requisitos y que estén dispuestas a participar en el proyecto.
- Solo podrán recolectarse datos de las personas que tengan una muñeca delgada debido al tamaño de la Microsoft Band.
- El tiempo mínimo de entrega del proyecto son 3 meses y el máximo son 9 meses, con posibilidades de prórroga de 3 meses más.
- La condición de que el sistema sea no intrusivo para los participantes, limita la recolección de los datos.

- **Derechos de acceso a las fuentes de datos o conocimientos relevantes**

- Son requeridos pagos por acceder a algunos de los artículos encontrados en la revisión sistemática y por información relacionada al proyecto.

- **Limitaciones en accesibilidad técnica de datos**

- La compatibilidad de la banda con los sistemas operativos del smartphone se limita a los 3 más importantes, Android, iOS y Windows Phone. Sin embargo, el proyecto se centra en llevar a cabo el desarrollo en un sistema operativo Android, ya que la gran mayoría de personas, cercanas al proyecto, usan este SO.
- La Microsoft Band 2 no tiene memoria para almacenar datos, por lo tanto, si hay alguna interrupción en la comunicación con el Smartphone, estos se perderían.

3.1.2.1.3. Riesgos y contingencias

En esta parte se enumeran los riesgos o acontecimientos que podrían retrasar el proyecto o causar que falle. Además, se listan los planes de contingencia correspondientes, donde se especifica qué medidas se tomarán si estos riesgos o eventos tienen lugar.

Según las eventualidades que puedan surgir en el desarrollo del sistema hay que considerar:

- La banda presente errores de funcionamiento.
 - En este caso habrá que recurrir al uso de la banda prototipo o versión 1 que se tiene disponible en el laboratorio, aunque con ella no se tenga acceso a todos los sensores que sí dispone su última versión.
- Pérdida de los datos.
 - Los datos capturados deben tener redundancia. Si los datos están alojados en un servidor externo, también deben estar en uno de los equipos portátiles. Si el servidor está alojado en uno de los equipos, la redundancia estará en el otro portátil disponible.
- Los participantes no sepan manejar los elementos del sistema o no realicen adecuadamente las actividades solicitadas.
 - Antes de cada sesión se va a instruir sobre el uso de cada dispositivo. Además, habrá una persona encargada de observar, registrar y corregir el protocolo de actividades sedentarias.
- La comunicación entre cada nodo del sistema sea interrumpida.
 - **Banda - móvil:** En caso de pérdida de comunicación, el sistema deberá detectar tal interrupción e informar, para así volver a recolectar los datos correspondientes.
 - **Beacon - móvil:** El procedimiento es similar al caso anterior.

3.1.2.1.4. Costos y beneficios

En este apartado se realiza un análisis de costo-beneficio para el proyecto. Entonces, se presenta una comparación de los gastos del proyecto con los beneficios potenciales, y así saber si este es viable.

- **Costo-Beneficio**

El desarrollo de la estimación de costos del proceso de recopilación de datos, implementación de la solución y costos operativos, es necesario para asegurarse que el proyecto cumple con las limitaciones presupuestarias y es factible su implementación.

- **Costo de la recopilación de datos**

En la Tabla 3-3 que se muestra a continuación hay una lista de las herramientas necesarias para cumplir con la recolección de datos, y así construir el dataset. Se aclara que, para el desarrollo de este proyecto, herramientas como los beacons, serán provistas por la universidad. Las personas que accedan a usar los dispositivos y

autoricen el tratamiento de sus datos, no tendrán ningún incentivo económico. Esto hace que no aumente el costo en términos monetarios del desarrollo de esta actividad.

Microsoft Band 2
Smartphone
Beacons
Equipos usados en la programación de los dispositivos
Consumo de energía de todos los dispositivos involucrados

Tabla 3-3. Equipos requeridos para la recolección de datos y gastos que conllevan

- **Costos de operación**

En el despliegue del sistema, el costo de operación estaría ligado a la adquisición de los dispositivos y al gasto energético de los mismos.

- **Beneficios**

En el campo de la minería de datos, un dataset hace parte fundamental para el descubrimiento de conocimiento. Cuando el dataset esté terminado, los investigadores podrán explorar técnicas que lleven a descubrir las características y patrones de la información recolectada. Esto lleva a que, en una futura implementación los sistemas que tomen como guía el presente dataset (así como los algoritmos que se hayan derivado del mismo) tengan resultados relevantes y satisfactorios.

La etapa de recolección de datos representa costos considerables dentro de un proyecto de minería de datos. Por esta razón, un dataset de calidad ahorraría costos adicionales en proyectos interesados en la evaluación de algoritmos, realizar clasificaciones o desarrollar alguna intervención para personas sedentarias. Por ende, en comparación, son muchos los beneficios que se consiguen y un bajo costo en su desarrollo, teniendo en cuenta que esta área del conocimiento es objeto constante de estudio.

3.1.2.2. Objetivos de la minería de datos

A diferencia del objetivo del negocio, los objetivos de la minería de datos se describen en términos técnicos. Vale la pena aclarar que el proyecto no alcanza a realizar clasificación, agrupamiento o predicciones con todos los datos recolectados. Sin embargo, como este proyecto pretende ser parte de proyectos de minería de datos, a continuación, se plantea un objetivo general que se llevaría a cabo (en un trabajo futuro) a partir del dataset generado en este trabajo.

Objetivos de la minería de datos

Proponer un clasificador para el monitoreo de comportamientos sedentarios en individuos, basado en su localización.

3.1.2.2.1. Criterios de éxito

Los criterios de éxito no van a estar estrictamente enfocados hacia la precisión alcanzada en los procesos de clasificación y predicción realizados al final de un proceso completo de minería de datos, sino en los puntos de referencia que se plantean en [105] relacionados a la construcción de un dataset bien estructurado. Entre los puntos de referencia se destaca que el dataset debe describir claramente las características de los datos y de la metodología de construcción del dataset. Esto para que los investigadores determinen si aquéllas propiedades se ajustan a sus necesidades. Para ello, es fundamental que el dataset especifique las propiedades listadas en la Tabla 3-4:

Tipo de modelo	Impersonal Híbrido Personal
Método de recolección	Totalmente natural Seminatural Laboratorio
Información del escenario <ul style="list-style-type: none"> • Número de participantes (sujetos) • Población • Rasgos • Actividades • Duración 	Ejemplo <ul style="list-style-type: none"> • 1,2,3 • estudiantes, ancianos, niños... • altura, peso, edad... • caminar, sentarse, acostarse...
Sensores <ul style="list-style-type: none"> • Tipo • Frecuencia de muestreo • Número • Localización 	<ul style="list-style-type: none"> • GPS, acelerómetro... • 20 Hz, 30 Hz, 50 Hz... • 1,2,3... • cintura, bolsillo, pecho...
Características/Presentación de <ul style="list-style-type: none"> • Unidades de los datos en bruto • Técnicas/Algoritmos • Ventanas 	<ul style="list-style-type: none"> • m/s, km/h... • estadística, transformadas... • 10 s, 30 s, 60 s...
Resultados <ul style="list-style-type: none"> • Exactitud • Consistencia 	<ul style="list-style-type: none"> • 94%, 95%... • media, varianza...

Tabla 3-4. Propiedades a especificar del dataset

Tipo de modelo

El tipo de modelo determina cómo los datos son recolectados, organizados y particionados en los datos de entrenamiento y de prueba. Los autores del artículo [105] encuentran tres tipos de modelos básicos:

- **Modelo impersonal:** Este modelo también se conoce como universal y se caracteriza porque usan los datos de entrenamiento de un panel de usuarios, quienes no usarán subsecuentemente el modelo, es decir, no estarán presentes en el conjunto de pruebas.
- **Modelo personal:** Solamente usa los datos de entrenamiento del usuario que usará el modelo. Es decir, los datos de entrenamiento y de prueba provendrán de la misma persona.
- **Modelo híbrido:** Es una combinación del modelo personal e impersonal. Los usuarios del modelo tendrán datos etiquetados en el conjunto de entrenamiento, pero este a su vez incluye datos de otros usuarios.

Los mismos autores subrayan la importancia de aplicar el modelo personal en las investigaciones relacionadas a reconocimiento de actividad física (estar de pie, sentado, caminando o corriendo). Esto implicaría obtener una precisión cercana al 98% en los resultados, mientras que con los modelos impersonal e híbrido se conseguirían valores de precisión cercanos al 76% y 95% respectivamente. **El modelo personal se caracteriza por arrojar buenos resultados con un muy pequeño número de datos de entrenamiento.** Mientras que los resultados obtenidos al aplicar un modelo impersonal son poco fiables si se analizan pocos usuarios.

Cuando el dataset esté construido será posible usarlo en los tres tipos de modelos. Si bien, cuando se use para entrenar un clasificador y se pruebe con un grupo de participantes diferentes a los de entrenamiento será usado en el **modelo impersonal**, este tipo de modelo será el que se use generalmente, ya que los datos serán anónimos por razones de seguridad de los participantes y por tanto no será posible usar los otros modelos. Sin embargo, si los mismos investigadores que realizaron la recolección, ejecutan una clasificación posterior (con los mismos sujetos que formaron la población de entrenamiento). Ellos podrían conseguir un **modelo personal** y/o el **modelo híbrido**.

Metodología de recolección de datos

El escenario en el cual los datos son recolectados es importante y por tanto debe ser claramente descrito. A un alto nivel se reconocen tres tipos:

- **Totalmente natural:** Los sujetos realizan sus actividades normales sin ser alterados por los mecanismos o técnicas utilizados en la lectura de datos.
- **seminatural:** Los sujetos se desenvuelven en su ambiente natural, pero modifican su comportamiento en forma leve para llevar a cabo las tareas incluidas en un protocolo propuesto.

- **Laboratorio:** Los sujetos realizan actividades estructuradas en un ambiente de laboratorio.

Para este proyecto se opta por desarrollar la recolección de datos mediante el método-escenario seminatural. Esto porque: un método de laboratorio interfiere demasiado en el comportamiento de las personas a medir y un ambiente totalmente natural es demasiado complejo en el sentido de necesitar demasiado tiempo para recolectar los datos a una cantidad suficiente de personas, además, con el fin de etiquetar las actividades realizadas, esto dependería en gran medida de la disposición de la persona para llenar constantemente información sobre aquello que esté realizando en un momento determinado o realizar el etiquetado basado en cámaras de video que involucraría un impacto enorme en la privacidad de los participantes. Por tanto, al usar el método seminatural, donde se estructura un protocolo en el cual la recolección de datos etiquetados no requiera tanto tiempo y se realice en un ambiente natural, se postula como la mejor opción para este trabajo.

Número de sujetos y diversidad

En el momento de seleccionar a las personas más convenientes para el estudio, surgen preocupaciones constantes por conseguir diversidad en la población; así como por procurar un número de sujetos válido. Estos factores tienen una relación directa con la calidad y la robustez de los modelos de clasificación que se induzcan a partir de este trabajo [105]. El número de sujetos no solo impacta en la calidad de los datos sino también en la capacidad de evaluar la consistencia de los resultados entre sujetos. Por esta razón el dataset a construir pretende recolectar datos de por lo menos 30 sujetos, con alta diversidad. Tal diversidad puede tomar muchas formas, entre las cuales sobresalen: el género, la edad, peso y estatura. Para el caso particular de este proyecto, la edad y el género son los factores de diversidad más sobresalientes.

- Al final de la recolección, los participantes deben pertenecer en años a los siguientes rangos: 20 a 29, 30 a 39, 40 a 49, 50 a 59 y mayores o iguales a 60. De manera que haya sujetos en todos los intervalos definidos. Así que hay 5 intervalos de edades, dentro de los cuales se deben encontrar datos de por lo menos 5 sujetos. Los 5 sujetos restantes pueden pertenecer a cualquiera de los rangos, siempre y cuando no estén todos dentro de uno solo.
- El sexo de los participantes también requiere que haya cierto balance entre el número de mujeres y hombres. De manera que se busca una relación lo más cercana a 50-50 en el porcentaje de sujetos masculinos contra femeninos.

Datos Personales: Datos personales usados para verificar la diversidad del datasets. A continuación, se describe cada uno de los datos personales recaudados.

- **Nombre de Usuario:** Este dato es solicitado con el objetivo de administrar la sesión de cada participante dentro de la aplicación móvil diseñada. La convención para este atributo es formar una cadena de texto donde se incluyan primer nombre y primer apellido del participante, en minúsculas. Si se presenta el caso en que haya dos personas con el primer nombre y primer apellido iguales, se agrega la primera letra del segundo nombre y por último la primera letra del segundo apellido.
- **Nombres y Apellidos:** Se toman los nombres y apellidos de la persona, con el objetivo de saber específicamente quién está registrado en el sistema. Además de evitar confusiones cuando se recauden los datos o se realice la respectiva revisión. Sin embargo, antes de ser publicado el dataset, este dato será eliminado por la seguridad y anonimato de los participantes.
- **Edad:** Mediante este dato se categoriza a los participantes en los diferentes intervalos de edad definidos. Este dato se ingresa en años.
- **Género:** Este dato solo puede ser masculino o femenino.
- **Peso:** Este dato debe ingresarse en kilogramos y debe medirse antes de iniciar la recolección de datos.
- **Altura:** Este dato debe ingresarse en centímetros y debe medirse antes de iniciar la recolección de datos.
- **Cintura:** Esta medida se toma debido a que la obesidad abdominal es un factor de riesgo del SM. Entonces, este dato puede ser importante en el momento de relacionar los diferentes datos con el SM. Este dato será tomado a la altura del ombligo y medido en centímetros.
- **Correo:** Este dato se toma con el objetivo de contactar con los participantes en caso de ser necesario. Sin embargo, en algunos casos las personas no poseen un correo electrónico, por lo cual se ingresará el correo del supervisor que se encuentre verificando la recolección de datos en ese caso. Al final este dato también será eliminado por el anonimato de la información.
- **Profesión:** Este dato, al igual que los siguientes tres, se toman para que se pueda relacionar el estilo de vida del participante con los datos capturados. Al no tener una profesión definida, este dato será reemplazado por la ocupación de la persona.

- **Frecuencia en el hábito de fumar:** Para el registro de este dato, se definen cinco opciones posibles: uno al día, varios al día, varios a la semana, rara vez y nunca.
- **Frecuencia en el consumo de alcohol:** Para este caso se presentan cuatro opciones: A diario, fines de semana, rara vez, nunca.
- **Medio de transporte habitual:** Para este caso se presentan cinco opciones entre las cuales el participante selecciona el medio de transporte que más usa en su día a día. Las opciones serán: Transporte público, carro, moto, bicicleta y caminando.

Sensores

Los sensores que se planean utilizar para recolectar los datos a incluir en el dataset son:

- **Microsoft Band 2:** Acelerómetro, luz ambiente, barómetro, giroscopio, respuesta galvánica de la piel, frecuencia cardíaca, intervalo de frecuencia cardíaca, podómetro, temperatura de la piel e índice UV.
- **Smartphones:** Acelerómetro, giroscopio, gravedad, campo magnético de la tierra, aceleración lineal, rotación y barómetro. Además, los smartphones reciben la potencia de la señal Bluetooth de los beacons, con lo cual podría deducirse la distancia a la que se encuentran.

3.1.2.3. Plan del proyecto

La idea de esta tarea es describir el plan previsto para alcanzar los objetivos de minería de datos y así lograr los objetivos del negocio.

3.1.2.3.1. Plan del proyecto

Para el desarrollo del proyecto se ha planteado el uso de determinadas fases y actividades propuestas en la metodología para minería de datos CRISP-DM.

Como primera acción se hace una contextualización sobre el comportamiento sedentario y el reconocimiento de actividades. De esta acción emergen necesidades y problemas en los procesos de reconocimiento de actividades. Luego de encontradas las necesidades en el campo de monitorización de comportamientos sedentarios se propone la construcción de una herramienta útil en futuros trabajos relacionados a minería de datos.

Antes de iniciar la recolección de los datos es necesario crear un sistema adecuado para ello. Entonces, diseñar las aplicaciones, diseñar los escenarios de recolección y seleccionar los dispositivos se convierten en tareas destacadas. La selección de

dispositivos se puede encontrar en la sección *inventario de recursos*. Donde como resultado de las comparaciones y disponibilidad de dispositivos, se decide por utilizar dos smartphones, una microsoft band 2 y 6 beacons. A continuación, se describe el proceso para la recolección de los datos, el cual detalla las tareas desde que se empieza a crear el sistema hasta que se publica el dataset elaborado.

- i. Desarrollo de prototipos y módulos del sistema con todas las conexiones entre los smartphone, la microsoft band 2 y los beacons. Se tiene en cuenta que el mismo sistema desarrollado se adaptará para el otro Smartphone. La aplicación móvil debe almacenar los datos en la memoria interna. La interfaz debe ser lo más simple e intuitiva posible, ya que posiblemente no sea manipulada solo por los desarrolladores, sino por colaboradores y participantes de la recolección.
- ii. Selección de las personas que van a participar en la recolección de los datos.
- iii. Descripción y firma de un consentimiento informado por parte del participante.
- iv. Entrenamiento de los participantes para el correcto uso del wearable y los smartphone. Esto con el fin de minimizar los errores de los datos debidos al mal uso de los dispositivos.
- v. Preparación del escenario de recolección de datos. El lugar de recolección es la casa de cada uno de los participantes y antes de empezar la recolección deben ubicarse los beacons en cada uno de los lugares y dispositivos de interés.
- vi. Luego de concluida la recolección de datos, hay que organizar y analizar los datos resultantes.
- vii. Finalmente hay que concluir sobre el proceso de desarrollo y los resultados obtenidos, así como los trabajos futuros y publicar el dataset.

3.1.2.3.2. Selección de comportamientos sedentarios a ser incluidos en el dataset.

La selección de los comportamientos sedentarios, se realiza basándose en la taxonomía de comportamientos sedentarios descrita en [106] y la definición de comportamiento sedentario presente en [11]. Con las cuales se definen tres posturas base: sentado, reclinado y acostado. Además, se enfocó la selección en el uso de dispositivos electrónicos, como el televisor, el computador y el celular, ya que la vida moderna, con el uso de estos dispositivos fomenta la ejecución de comportamientos sedentarios en la vida cotidiana [107].

Finalmente, basados en los lugares donde comúnmente se pueden usar estos dispositivos, se definieron la cama, el escritorio y la sala, como los lugares apropiados para realizar recolección de datos. En la Tabla 3-5, se presentan los comportamientos sedentarios seleccionados. Adicionalmente a estos 23 comportamientos, es adecuado recolectar datos de 2 actividades más: estar de pie y caminar, esto con el fin de que el dataset permita ampliar el tipo de experimentos y análisis hechos con el fin de obtener clasificadores capaces de ser implementados en un contexto real.

Sentado	Reclinado	Acostado
1. Usando PC en el escritorio.	10. Usando PC en el escritorio.	19. Usando celular en la cama.
2. Usando PC en la cama	11. Usando PC en la cama.	20. Usando celular en la sala.
3. Usando celular en el escritorio.	12. Usando celular en el escritorio.	21. Viendo TV en la sala.
4. Usando celular en la cama.	13. Usando celular en la cama.	22. En reposo en la sala.
5. Usando celular en la sala.	14. Usando celular en la sala.	23. En reposo en la cama.
6. Viendo TV en la sala.	15. Viendo TV en la sala.	
7. En reposo en el escritorio.	16. En reposo en el escritorio.	
8. En reposo en la sala.	17. En reposo en la sala	
9. En reposo en la cama.	18. En reposo en la cama.	

Tabla 3-5. Comportamientos sedentarios incluidos en el dataset

3.1.2.3.3. Evaluación inicial de herramientas y técnicas

Al final de la primera fase, el equipo del proyecto realiza una evaluación inicial de herramientas y técnicas más convenientes en el desarrollo del proyecto. Tales herramientas son descritas a continuación:

Android Studio

Para desarrollar la aplicación en un dispositivo móvil se hará uso de la herramienta Android Studio, la cual brinda un desarrollo ágil y estable en este tipo de sistemas. Además, se le incorporarán algunas librerías para permitir un mejor desarrollo.

Microsoft Band SDK

Desde el punto de vista del desarrollo Software, Microsoft proporciona un kit de desarrollo (SDK) que hace posible obtener datos de los sensores disponibles en la banda usada para este proyecto.

WISDM Lab

Hace recomendaciones concretas sobre recopilar datos de sensores, mediante smartphones y otros dispositivos móviles modernos, para vincularlos con ciertas actividades físicas [58]. Entre las recomendaciones presentan una serie de características relevantes en la construcción de un dataset. Es por este motivo que sus

contenidos se convierten en una herramienta metodológica base del desarrollo de este documento y del dataset esperado.

3.2. Metodología para la construcción del sistema de la recolección

CRISP-DM no tiene una fase, tarea o actividad que guie los procesos de desarrollo o implementación de los sistemas usados para realizar la recolección de datos. Es por ese motivo que se opta por el uso de elementos del Modelo de Construcción de Soluciones M.C.S para el desarrollo del sistema de recolección de datos. Este modelo se caracteriza por ser: dirigido por objetivos, iterativo e incremental, manejado por riesgos, guiado por casos de uso, centrado en arquitectura y basado en componentes [108] La elección de ese proceso de desarrollo se hace debido a los buenos resultados que previamente se han obtenido en prácticas de laboratorio y ciertas asignaturas. Además, el M.C.S es ampliamente usado en los Trabajos de Grado llevados a cabo en la Facultad de Ingeniería Electrónica y Telecomunicaciones de la Universidad del Cauca, especialmente en el Departamento de Telemática. A continuación, se describen los tres propósitos fundamentales del M.C.S.

- **Construir una solución de calidad y oportuna:** En el presente trabajo de grado se aplica al cumplimiento de los objetivos, dentro de los plazos de tiempo fijados en la reglamentación de la facultad para la presentación de este tipo de trabajos.
- **Que la solución planteada tenga costos competitivos:** Uno de los problemas planteados y encontrados en la revisión de la literatura es precisamente el elevado costo que involucran las enfermedades cardiovasculares en su prevención, detección y tratamiento.
- **Que la solución planteada pueda contribuir a la creación y enriquecimiento de la base de conocimiento/experiencia institucional:** El presente trabajo busca generar conocimiento y herramientas para la línea de investigación en e-Salud.

El M.C.S. está compuesto por tres macrocomponentes, cada uno de los cuales es una perspectiva o dimensión de cualquier proceso que se debe llevar a cabo para la construcción de la solución. Estos tres macrocomponentes son [108]:

- **Estructura para la descripción del sistema:** Este macrocomponente está formado por tres modelos, denominados Modelo para el Establecimiento de Responsabilidades (M.E.R.), Modelo para la Descripción del Sistema (M.D.S.) y Modelo para la Implementación del Sistema (M.I.S.); organizados

de mayor nivel de abstracción (M.E.R.) a un menor nivel de abstracción (M.I.S.).

- **Modelo del proceso de desarrollo:** Define las fases del proceso para la construcción de la solución, y determina las transiciones entre las mismas.
- **Modelo para la organización del talento:** Define una referencia para organizar al talento (recurso humano) que trabaja en el proyecto. En este caso el equipo de trabajo se encuentra definido por los estudiantes y el director del Trabajo de Grado.

3.2.1. Pilares fundamentales

Los pilares esenciales para el proceso de desarrollo de la solución, con el fin de lograr los propósitos descritos anteriormente, son cinco. Estos cinco pilares son [108]:

1. **Ingeniería del sistema:** Este pilar implica que el sistema debe tratar de cumplir y satisfacer las necesidades del ambiente en el que se va a implantar. Debido a esto, se debe tener una comprensión del entorno y los fenómenos que lo rodean.
2. **Orientación a objetos:** Hace referencia a la reusabilidad de componentes (utilizar al máximo los componentes ya existentes, y al generar nuevos componentes, que estos sean diseñados para que puedan ser usados en el futuro).
3. **Orientación a diseño:** Este pilar indica que el proceso de desarrollo debe enfocarse en el diseño del sistema o solución, no en la implementación del mismo. Si se obtiene un diseño claro y entendible por todo el equipo de trabajo, entonces la implementación es considerada una derivación con cierta automatización. Por lo tanto, los sistemas deben ser validados principalmente en su diseño.
4. **Técnicas de descripción formal:** Este pilar indica que se deben buscar lenguajes de descripción que favorezcan una clara comunicación entre los integrantes del equipo de trabajo, y de estos con los clientes. El autor del M.C.S. sugiere una posición intermedia, con notaciones semi-formales, pero con lenguajes de modelamiento formales.
5. **Abstracción como técnica fundamental:** Hace referencia a la abstracción como uno de los pilares más importantes, con lo cual se puede entender completamente la finalidad del sistema.

3.2.2. Fases de referencia

La metodología M.C.S. plantea un proceso de desarrollo con cuatro fases, con las respectivas etapas para pasar de una fase a la siguiente (estas etapas se denominan Revisión y Compromiso). Las cuatro fases son: estudio de prefactibilidad, formulación del proyecto, ejecución del proyecto, y validación de la solución [108].

Fase 1, estudio de prefactibilidad: Con esta fase se pretende analizar el problema y determinar si se continúa con el proyecto. De manera que pueda definirse claramente el alcance del sistema.

Fase 2, formulación del proyecto: Mediante esta fase se pretende analizar los aspectos esenciales de la construcción del sistema, y asegurar su viabilidad. De manera que pueda establecerse una arquitectura de referencia.

Fase 3, ejecución del proyecto: Con esta fase se pretende construir el sistema o solución. De manera que el sistema construido tenga una capacidad operacional adecuada.

Fase 4, validación de la solución: Mediante esta fase se pretende probar el sistema y corregir errores detectados. Como resultado se llega a un sistema en versión final (a satisfacción del cliente o las necesidades)

3.3. Conclusiones

Este y el primer capítulo contienen las recomendaciones, tareas y actividades propuestas en la fase 1 de CRISP-DM. El principal objetivo de este capítulo es entender, de la forma más clara, el proyecto. De modo que, este capítulo describe aspectos fundamentales, conceptos, requisitos y consideraciones sobre el proyecto en general. En consecuencia, a partir de esta fase surgen factores como la arquitectura inicial y final del sistema, la selección de los dispositivos, requerimientos, limitaciones, supuestos, criterios de éxito, metodología de recolección y el plan del proyecto.

CAPÍTULO 4.

MODELADO Y ARQUITECTURA DEL SISTEMA DE RECOLECCIÓN DE DATOS

El objetivo de este capítulo es modelar el sistema para la recolección de datos. El modelado se realiza mediante una serie de diagramas y vistas comúnmente utilizadas en UML. Además, el modelado debe tener en cuenta que el sistema pretende integrar los diferentes dispositivos seleccionados: smartphone, Microsoft band 2 y beacons, y que éste a su vez, debe interactuar con investigadores y participantes. Inicialmente y contando con los requerimientos descritos para el funcionamiento del sistema, se realiza una arquitectura y una conceptualización del mismo. Adicionalmente, se describe la arquitectura desde las múltiples vistas en el modelo 4+1. De esta manera, se procura clarificar cada uno de los pasos a seguir en el desarrollo del software deseado.

4.1. Modelado del negocio

El modelado del negocio se define en [109] como un conjunto de actividades, cuyo objetivo es ayudar a visualizar y comprender los procesos del negocio. En otras palabras, el modelado del negocio es una guía que actúa como plano o borrador para la construcción de un sistema software u otros tipos de sistemas [109].

El modelado consta de dos elementos claves: un modelo de casos de uso del negocio y uno de objetos del negocio. El primero describe las acciones que realiza un proceso particular del negocio con el fin de ofrecer un valor a los actores del negocio. El otro modelo describe cómo un proceso del negocio lograría las acciones descritas en el modelo de casos de uso del negocio. Este modelo ayuda a observar cómo las personas y las cosas están relacionadas y cómo ellos interactúan para realizar el proceso en cuestión [109]. En estas relaciones es necesario tener claro que un rol o un conjunto de roles en el negocio pueden ser representados como actores del negocio; quienes interactúan entre sí y manipulan entidades del negocio, las cuales representan una “cosa” o elemento manejado o usado por el negocio.

Para la realización de los modelos se hace uso de Unified Modeling Language (UML), el cual es un lenguaje creado para visualizar, especificar, construir, y documentar sistemas Software. En relación a la herramienta con la cual se realizan estos modelos, se usa StarUML, que es una herramienta dirigida a modelar sistemas software, teniendo a su disposición modelos tales como diagramas de casos de uso, de clases, componentes, entre otros. Para el desarrollo de este proyecto se usa la versión 1, ya que se puede usar de una forma gratuita [110].

4.1.1. Descripción de los actores

En este proyecto se plantean dos actores que interactúan con el sistema. Tal interacción va desde permitir la toma de los datos hasta la validación de los datos adquiridos en el proceso.

Sujeto de Pruebas: Es la persona que utilizará el sistema siguiendo un protocolo predeterminado para la recolección de datos.

Supervisor: Es la persona encargada de instruir a los sujetos de pruebas en el uso del sistema y del protocolo a realizar. Además, tendrá la tarea de vigilar y validar el correcto desarrollo de todo el procedimiento, de modo que se eviten, en la medida de lo posible, errores durante el proceso de recolección de datos.

4.1.2. Modelo de casos de uso del negocio

En la Figura 8 se presenta el modelo de casos de uso del negocio. En este se observan los actores del negocio identificados y el rol o roles que cada uno desempeña.

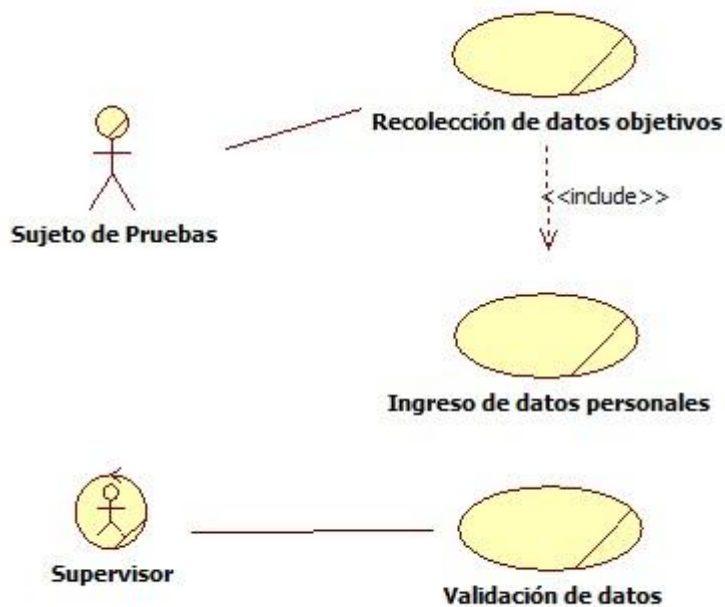


Figura 8. Casos de uso del negocio.

En las 3 tablas que se presentan a continuación se realiza una descripción más detallada de los casos de uso propuestos para el proyecto.

Caso de uso del negocio No. 1: Ingreso de datos personales
Iniciador: Sujeto de Pruebas
Propósito: Crear un registro con los datos personales del sujeto de pruebas.
Resumen: Este caso de uso solicita al sujeto de pruebas que ingrese una serie de datos personales, por ejemplo, nombre, fecha de nacimiento, altura, peso y otros. Esto con el fin de generar un registro de todas las personas e incorporar al dataset datos que permitan hallar patrones según esas características.

Tabla 4-1: Caso de uso, ingreso de datos personales

Caso de uso del negocio No. 2: Recolección de datos objetivos.
Iniciador: Sujeto de Pruebas
Propósito: Iniciar y realizar la recolección de los datos
Resumen: Es en este caso donde se ofrece la posibilidad de iniciar la medición de los datos, en la actividad seleccionada y cuando el sujeto de pruebas esté listo. También muestra el porcentaje de progreso en la recolección.

Tabla 4-2: Caso de uso, recolección de datos objetivos

Caso de uso del negocio No. 3: Validación de los datos.
Iniciador: Supervisor
Propósito: Validar el procedimiento de toma de datos.
Resumen: El supervisor acompañará al sujeto de pruebas durante todo el proceso de medición. De esta manera se garantiza el correcto desarrollo del proceso de recolección. Esto se realiza al final de cada actividad y los datos serán registrados o no, según el porcentaje de errores que despliega la interfaz de validación.

Tabla 4-3: Caso de uso, validar recolección

4.1.3. Modelo de objetos de la organización

La Figura 9 describe las interacciones entre los actores del negocio y los objetos del negocio. El eje del sistema es una aplicación móvil, que es usada por el sujeto de pruebas para proporcionar sus datos personales y realizar la recolección de variables fisiológicas y contextuales. Por otro lado, el supervisor solo tendrá contacto con la aplicación cuando sea necesario validar la recolección de los datos. Además, la aplicación móvil interactúa con los sensores de cada uno de los smartphone donde está alojada, la microsoft band 2 y los beacons. Estos datos son almacenados en la base de datos y según las acciones en la validación pueden eliminarse los registros o no.

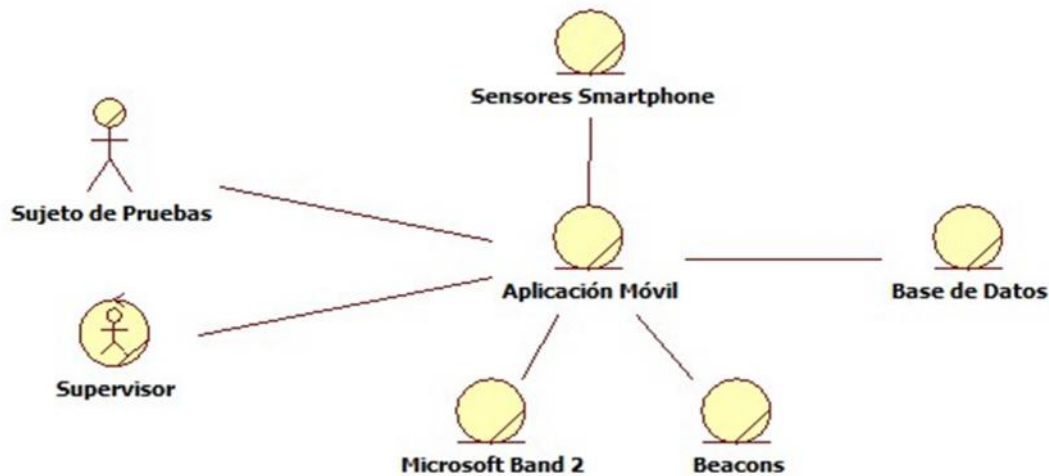


Figura 9. Modelo de objetos de la organización

4.2. Requisitos del sistema

En esta sección se presentan las características del sistema. A pesar de haber encontrado ciertas características del sistema deseado, en esta sección se detallan de forma más específica los requisitos funcionales y no funcionales que deberán ser implementados dentro del desarrollo.

4.2.1. Características del Sistema

La selección de las características del sistema se hizo considerando los requerimientos descritos (Capítulo 3, evaluación de la situación), el análisis del problema, la revisión sistemática y el modelado del negocio. Con todas las consideraciones hechas hasta esta sección, se obtienen las características puntuales del sistema deseado. Estas son mostradas en la Tabla . Adicionalmente, para cada una de las características se presenta una breve descripción, riesgo y prioridad. Además, el riesgo y la prioridad se dan en 4 categorías, crítico, alto, importante y ordinario, del más grave al menos relevante respectivamente.

Característica	Descripción	Riesgo	Prioridad
Captura datos generales	Mediante un formulario se capturan ciertos rasgos de la persona (nombre, peso, edad). Esos datos son enviados a la base de datos.	Alto: La persona debe ingresar sus datos con la mayor veracidad posible.	Importante: los datos generales son necesario para caracterizar a las personas y los datos.
Recolección datos de la Microsoft Band	Captura los datos de los sensores de la Microsoft Band 2	Crítico: Se debe garantizar el buen posicionamiento de la banda, además de las adecuadas frecuencias de muestreo de cada uno de los sensores.	Crítico: La característica provee buena parte de los datos fisiológicos y contextuales necesarios en el dataset.
Recolección datos del smartphone	Por medio de los sensores del Smartphone, son capturados algunos datos contextuales de la persona.	Crítico: Infieren las limitaciones de procesamiento del smartphone, ya que aparte de capturar datos de sensores internos, también recolecta y almacena datos de la banda, de los beacons.	Crítico: Los sensores del smartphone permiten tener más datos importantes para el dataset.
Recolección datos de los beacons	Con los datos de los beacons se puede saber la ubicación aproximada de una persona dentro de un entorno cerrado.	Crítico: Deben considerarse los tiempos de conexión entre el beacon y el Smartphone (tiempo que toma al teléfono reconocer al beacon).	Crítico: Esta es una de las características importantes y centrales del proyecto.
Visualización del proceso de recolección (inicio, reinicio, terminación)	Describe constantemente el tiempo restante para finalizar el comportamiento sedentario, como iniciar, reiniciar o finalizar una recolección.	Ordinario: La persona debe permanecer realizando el comportamiento sedentario durante el tiempo establecido en el protocolo.	Ordinario. Es importante la interacción de la aplicación con los actores.

Almacenamiento de los datos recolectados	Todas las muestras recolectadas se almacenan en una base de datos.	Alto: Debe considerarse la calidad en el almacenamiento de los datos y la correcta ubicación y estructura de los mismos para facilitar futuros procesos de limpieza o tratamiento.	Crítico: Hace parte de los objetivos del proyecto.
--	--	---	---

Tabla 4-4. Características del sistema.

4.2.2. Requisitos funcionales

A continuación, se presentan los requisitos funcionales del sistema:

- Cada dispositivo del sistema se conectará vía bluetooth low energy (BLE) con el smartphone donde se encuentra alojada la aplicación.
- La Microsoft band 2 deberá capturar todos los datos de sus sensores con una frecuencia mayor a 20 Hz para los sensores de movimiento (acelerómetro y giroscopio) y la mayor frecuencia disponible para el resto de sensores.
- Los beacons deberán transmitir su señal con una potencia de transmisión que permita cubrir la zona en la cual se haga la captura de datos.
- El sistema deberá tener una cuenta regresiva y una forma para notificar a la persona que ya ha terminado de recolectar los datos del comportamiento sedentario seleccionado.
- El sistema deberá almacenar la información registrada de todos los sensores y eliminar los registros en caso de presentarse fallos durante la medición.
- La aplicación permitirá a la persona seleccionar qué comportamiento sedentario desea realizar y en qué momento quiere empezar a recolectar los datos.
- El sistema deberá permitir que se inicie o termine la recolección de datos de la actividad de comportamiento sedentario que se encuentre en medición.

4.2.3. Requisitos no funcionales

Basados en las características del sistema descritas en la Tabla 4-4, a continuación, se presentan los requisitos no funcionales del sistema:

- La interfaz de la aplicación móvil deberá ser intuitiva y deberá contener la información necesaria para identificar el usuario registrado, los comportamientos sedentarios pendientes y los finalizados.
- La capacidad de almacenamiento del smartphone deberá ser suficiente para la recolección de los datos.

- Los datos capturados por los sensores deberán tener la fiabilidad más alta posible, en relación a su disponibilidad y el número de datos nulos.
- La aplicación desarrollada deberá ser fácilmente escalable y estable para futuros cambios.

4.3. Descripción de la arquitectura del sistema

La selección de la arquitectura base del sistema y de los elementos que la compondrán, están descritos en el capítulo 3. En esta sección se hace una descripción más a fondo mediante el modelo de vistas de arquitectura 4+1 [111]. Este modelo permite presentar la descripción completa de un sistema desde múltiples perspectivas, cuatro principales (vista lógica, vista desarrollo, vista de procesos y vista física) y una vista (escenarios) transversal a las otras 4 donde son descritas las necesidades funcionales del sistema [112]. En la Figura 10, se presenta el modelo de vistas 4+1.

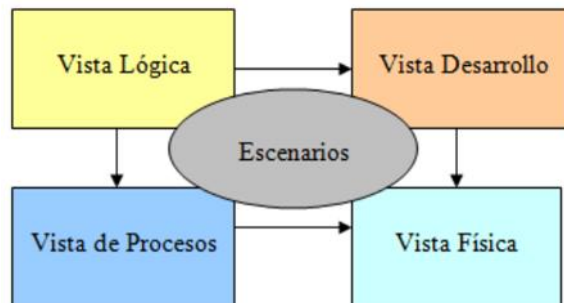


Figura 10. Arquitectura 4+1 Vistas [112].

Recomendaciones de qué es lo que se debería documentar en cada una de las vistas, se pueden encontrar en [112], más no define de qué manera se debe hacer. Gracias al desarrollo de UML se puede hacer una relación de las 5 vistas del 4+1 con diagramas y símbolos. Así se obtiene una manera clara de desarrollar cada una de las vistas [112]. A continuación, se presenta una breve descripción de las 5 vistas y los diagramas UML relacionados a estas.

Vista Lógica: Se encarga de describir mediante diagramas de clases, de estados o colaboración la estructura y las funciones del sistema. Todo eso siempre apoyada en los requisitos funcionales del sistema identificados [112].

Vista de Procesos: Mediante diagramas de actividad, de estados o de secuencia se representan algunos requisitos no funcionales, con los cuales se busca especificar el hilo del proceso que desarrolla cada una de las clases de la vista lógica desde la perspectiva de un integrador de sistemas [112].

Vista de Desarrollo: Usando diagramas de componentes o paquetes organiza los módulos software, en el entorno de desarrollo, en forma de capas jerárquicas. De esa

manera, esta vista describe las dependencias entre los componentes, lo cual hace parte de la perspectiva del programador y como este ve el sistema [112].

Vista Física: Con el uso de diagramas de despliegue se describen los componentes físicos del sistema y las conexiones entre ellos [112].

Vista de Escenarios: Usa diagramas de casos de usos para representar las interacciones entre los actores y los casos de uso, con el objetivo de unir y relacionar las 4 vistas principales del sistema. Esto significa que con cada caso de uso se van a ir ligando las diferentes vistas [112].

A continuación, se presentan las vistas y diagramas más importantes, que permiten comprender de la mejor manera la aplicación a desarrollar, que será el eje central del sistema de recolección de datos. En el Anexo A, se encontrarán la vista de despliegue, los diagramas de componentes y algunos diagramas de secuencia que complementan los diagramas que sugiere la arquitectura de 4+1 vistas.

4.3.1. Vista de Escenarios

Casos de Uso del Sistema

En la Figura 11, se presenta el diagrama de casos de usos del sistema,

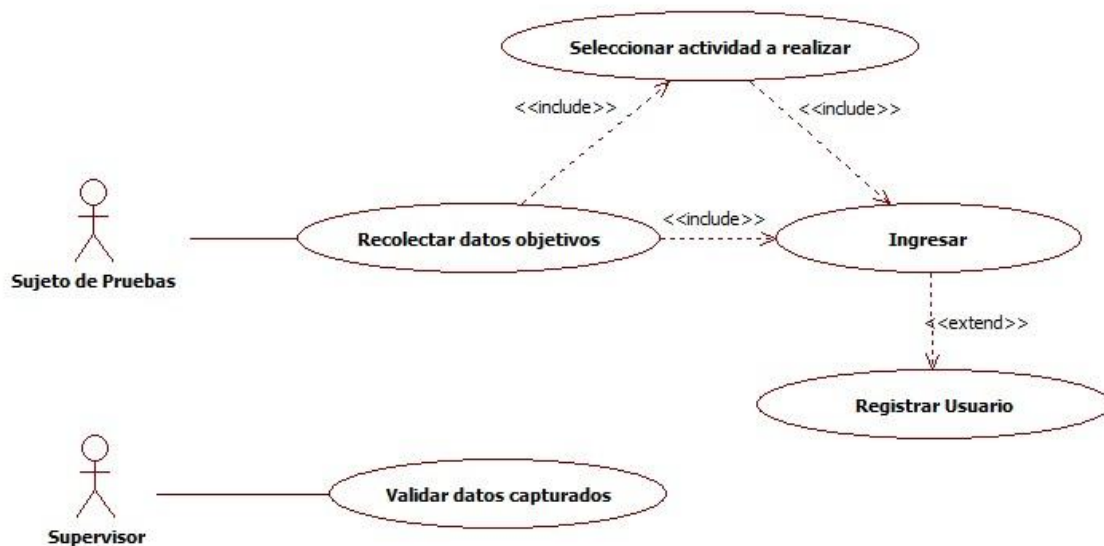


Figura 11. Casos de uso del sistema.

Las cinco tablas que se presentan a continuación presentan con más detalle los casos de usos encontrados.

Caso de uso del sistema No. 1: Recolectar datos objetivos.
Iniciador: Sujeto de Pruebas
Propósito: Iniciar y realizar la recolección de los datos sobre el sujeto de pruebas.
Resumen: Brinda la posibilidad de iniciar la medición de los datos cuando el sujeto de pruebas esté listo. Los datos empezarán a ser almacenados 40 segundos después del momento en que se inicie la recolección, para dar un intervalo de tiempo para que se ubiquen los dispositivos en su lugar y la persona adopte la posición deseada.

Tabla 4-5: Caso de uso, Recolectar los datos objetivos.

Caso de uso del sistema No. 2: Seleccionar actividad
Iniciador: Sujeto de Pruebas
Propósito: Seleccionar el comportamiento sedentario que se va a realizar.
Resumen: Seleccionar uno de los 23 comportamientos sedentarios que se va a realizar. (o alguna de las dos actividades adicionales: estar de pie y caminando)

Tabla 4-6: Caso de uso, Selección de actividad.

Caso de uso del sistema No. 3: Ingresar
Iniciador: Sujeto de Pruebas
Propósito: Ingresar al sistema cuando ya tiene un registro válido.
Resumen: El objetivo de este caso de uso es permitir el ingreso de un usuario, que ya ha realizado el proceso de registro en el sistema.

Tabla 4-7: Caso de uso, Ingresar.

Caso de uso del sistema No. 4: Registrar Usuario
Iniciador: Sujeto de Pruebas
Propósito: Registrar los datos personales del usuario.
Resumen: Se le solicita al usuario que ingrese sus datos personales una sola vez. Entre los datos solicitados se encuentran: nombre de usuario, nombre, apellido, edad, género, peso, altura, circunferencia de la cintura, profesión, correo, frecuencia con la que fuma, frecuencia en el consumo de alcohol y medio de transporte habitual. Esto con el fin de generar un registro de todas las personas e incorporar al dataset datos que permitan hallar patrones según esas características. Adicionalmente a los datos personales, la persona debe dar un previo consentimiento para el uso de estos.

Tabla 4-8: Caso de uso, Registrar usuario.

Caso de uso del sistema No. 5: Validar datos capturados
Iniciador: Supervisor
Propósito: Validar el procedimiento de toma de datos.
Resumen: El supervisor acompañará al sujeto de pruebas durante todo el proceso de medición. De esta manera se garantiza el correcto desarrollo del proceso de recolección. Al final de la toma de datos en cada una de las actividades, el supervisor validará el correcto desarrollo del protocolo y el comportamiento sedentario será marcado como realizado o no, quedando los datos almacenados en la base de datos o borrados según sea el caso.

Tabla 4-9: Caso de uso, Validar datos capturados.

4.3.2. Vista Lógica

La vista lógica, en este trabajo, se obtiene creando dos tipos de diagramas: el diagrama de clases del sistema y su diagrama de estados.

4.3.2.1. Diagrama de Clases

El diagrama de clases, en la Figura 12, describe las diferentes clases e interfaces dentro de la aplicación móvil. Este desarrollo es basado en la idea de Robert C. Martin sobre una arquitectura de desarrollo que él mismo denomina *clean architecture*, la cual se describe en la sección “4.4. Clean Architecture”. En el siguiente diagrama se observan las clases distribuidas en diferentes paquetes. El paquete “Vista” contiene la estructura visual para la interfaz con la que el usuario puede interactuar. Esto en el desarrollo Android es representado por *layouts* y *activities*. En el paquete de control se pueden observar los siguientes tipos de clases: Presentador, Interactuador y Repositorio. Además de eventos utilizados para comunicar repositorio y presentador. Estas clases hacen parte de las recomendaciones descritas en *clean architecture*. Además, se muestra una clase encargada de gestionar los datos para el componente “Base de Datos”.

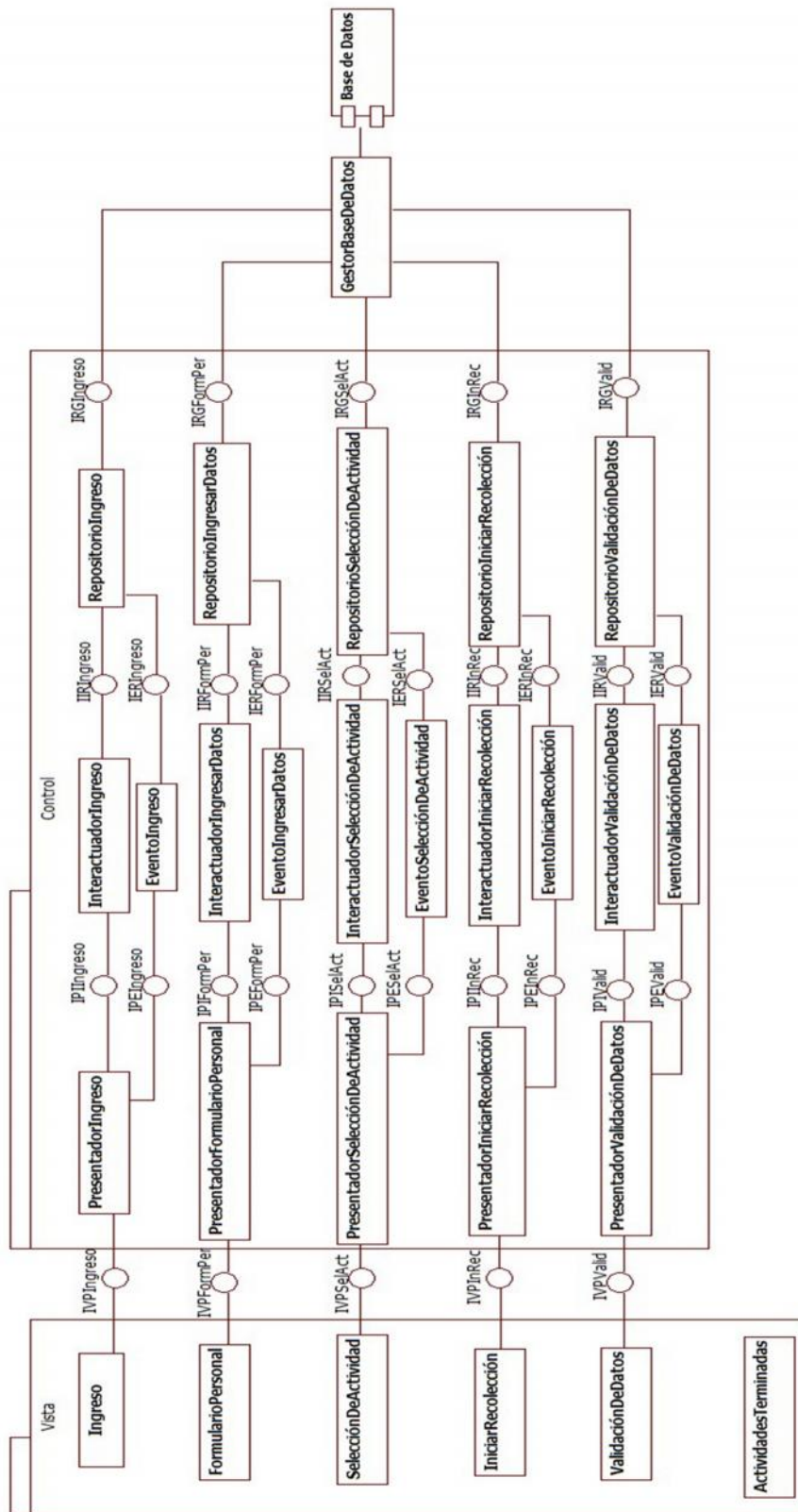


Figura 12. Diagrama de clases del sistema.

4.3.2.2. Diagrama de Estados

En el diagrama de la Figura 13 se pueden observar los estados en los que se mantendrá la aplicación mientras espera que el usuario realice una acción o que los procesos internos de la aplicación conduzcan a otros estados. Con este diagrama se infieren las diferentes interfaces que serán necesarias en el desarrollo; así como también los procesos automáticos requeridos.

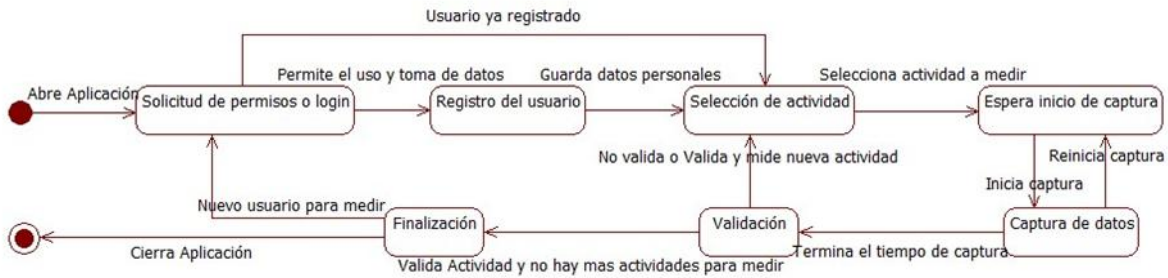


Figura 13. Diagrama de estados del sistema.

4.3.3. Vista de Procesos

4.3.3.1. Diagrama de Actividad

En el diagrama, de la Figura 14, se muestran las diferentes interacciones entre los actores y la aplicación. Como se observa en el diagrama, el usuario de la aplicación puede iniciar y reiniciar la recolección de datos. También es preciso señalar que sin el consentimiento del usuario no se puede ingresar a un nuevo sujeto de pruebas. En cuanto al ingreso de usuarios, hay una verificación para validar que el nombre de usuario no exista. Otra de las actividades presentadas, y que es realizada por el sujeto de pruebas, es la selección de los comportamientos definidos en este estudio.

Por otra parte, en el siguiente diagrama también se muestran las actividades que lleva a cabo el supervisor: validar que se hayan recogido todos los datos implicados correctamente y en caso de terminar todo el procedimiento con un usuario, decidir por un registro nuevo.

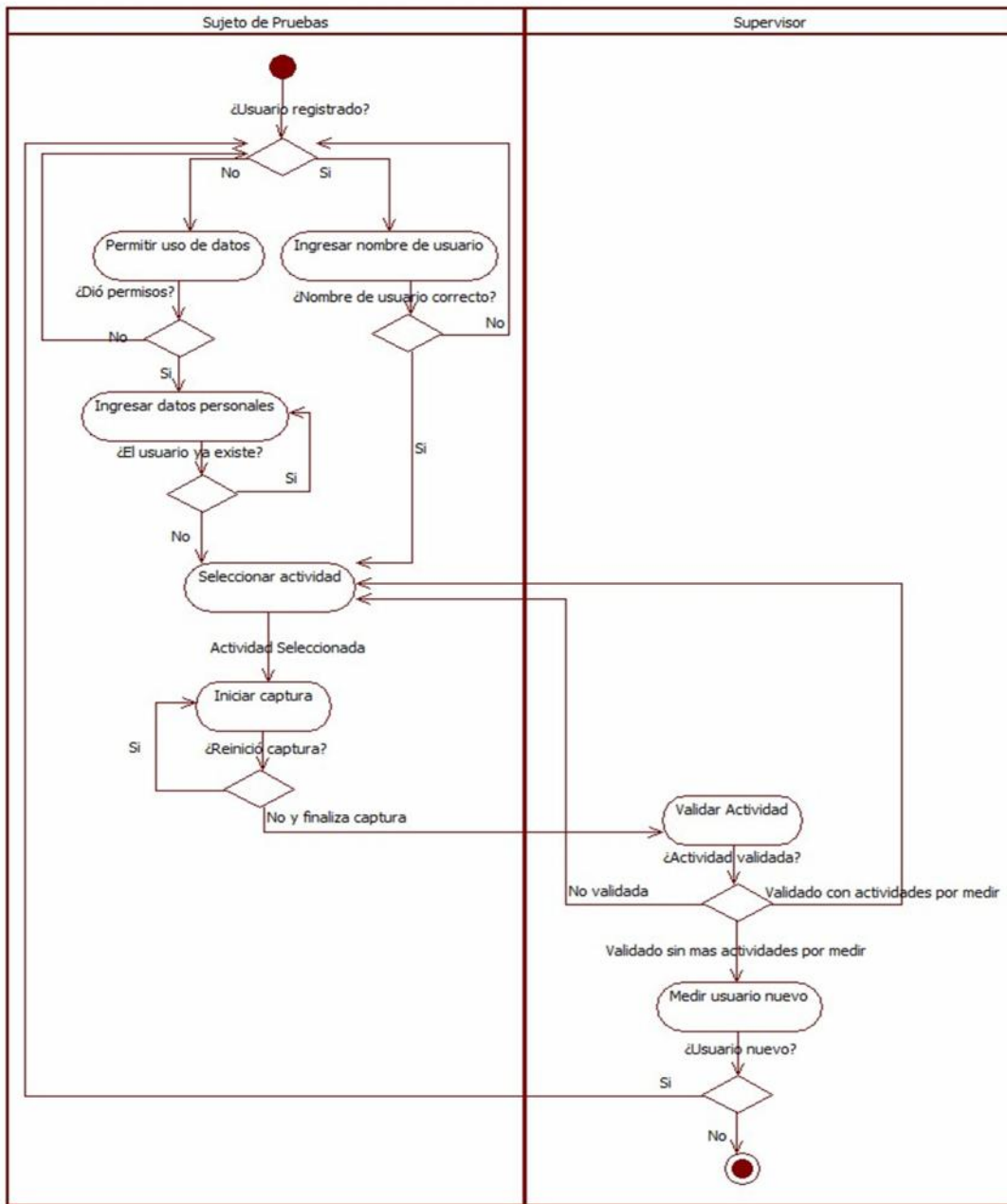


Figura 14. Diagrama de actividad del sistema.

4.3.3.2. Diagramas de Secuencia

El tres diagrama presentados a continuación son usados para mostrar las interacciones entre objetos, en el orden secuencial que tales interacciones ocurren [113]. Además de presentar la secuencia esperada en el desarrollo de cada proceso, en estos diagramas también aparecen las secuencias derivadas de errores.

Diagrama de Secuencia Recolectar Datos Objetivos

En la Figura 15 se presenta uno de los procesos más importantes, ya que es donde se inicia y termina la recolección. Es decir, el sistema captura y guarda los datos en la base de datos.

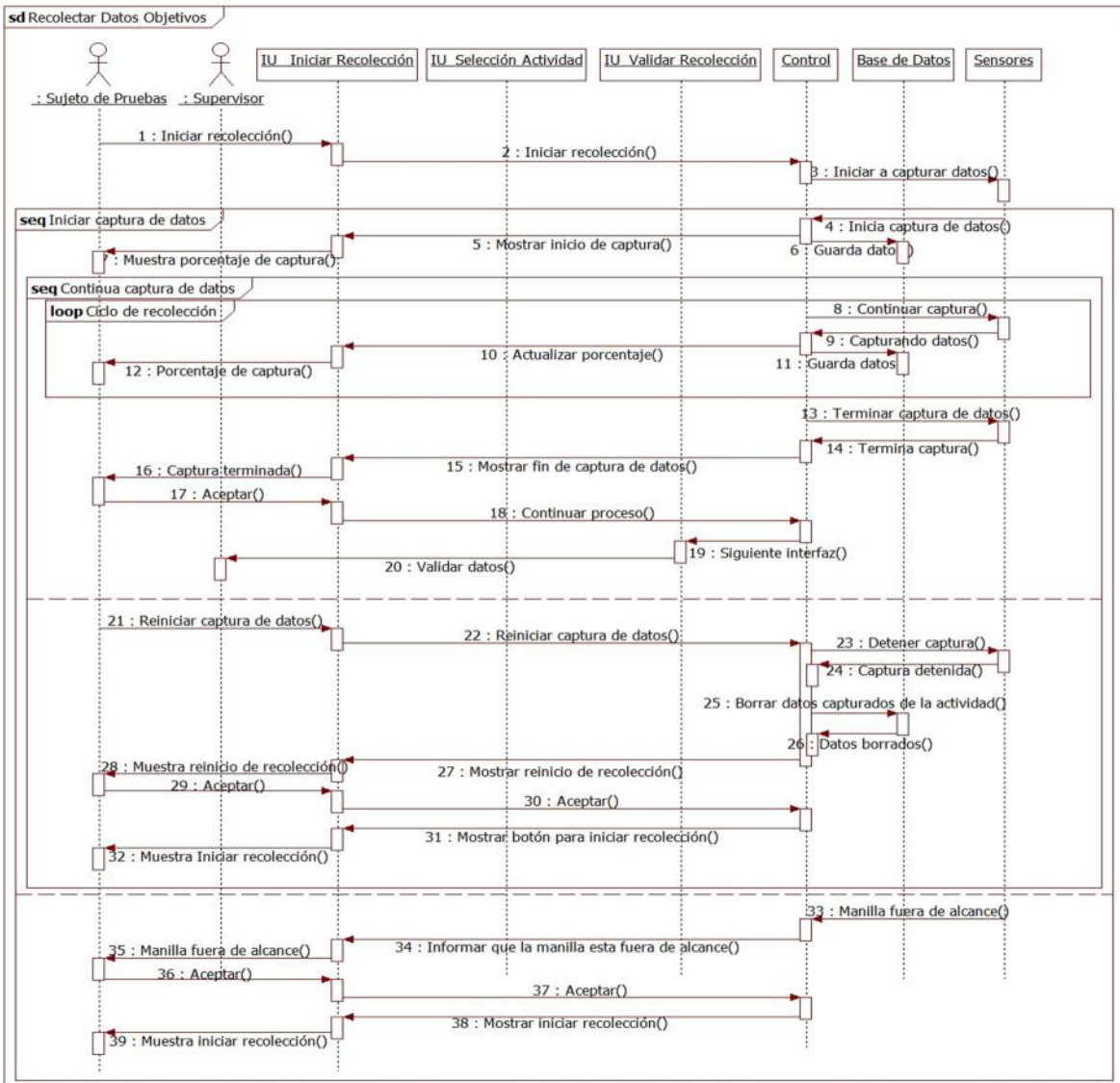


Figura 15. Diagrama de secuencia, Recolectar datos objetivos

Diagrama de Secuencia para Validar Recolección de Datos

La Figura 16 muestra el proceso seguido por el supervisor para verificar si la captura fue exitosa. Esto mediante el porcentaje de nulos registrados en el tiempo que duda la captura de una actividad o comportamiento. Si el porcentaje está por debajo del nivel determinado, el supervisor validará la captura y la actividad se marcará como

realizada. Entonces tal actividad no se podrá realizar nuevamente. Por el contrario, si no se valida la actividad, el estado de la actividad seguirá siendo sin realizar y el usuario de la aplicación podrá seleccionar nuevamente una actividad cualquiera.

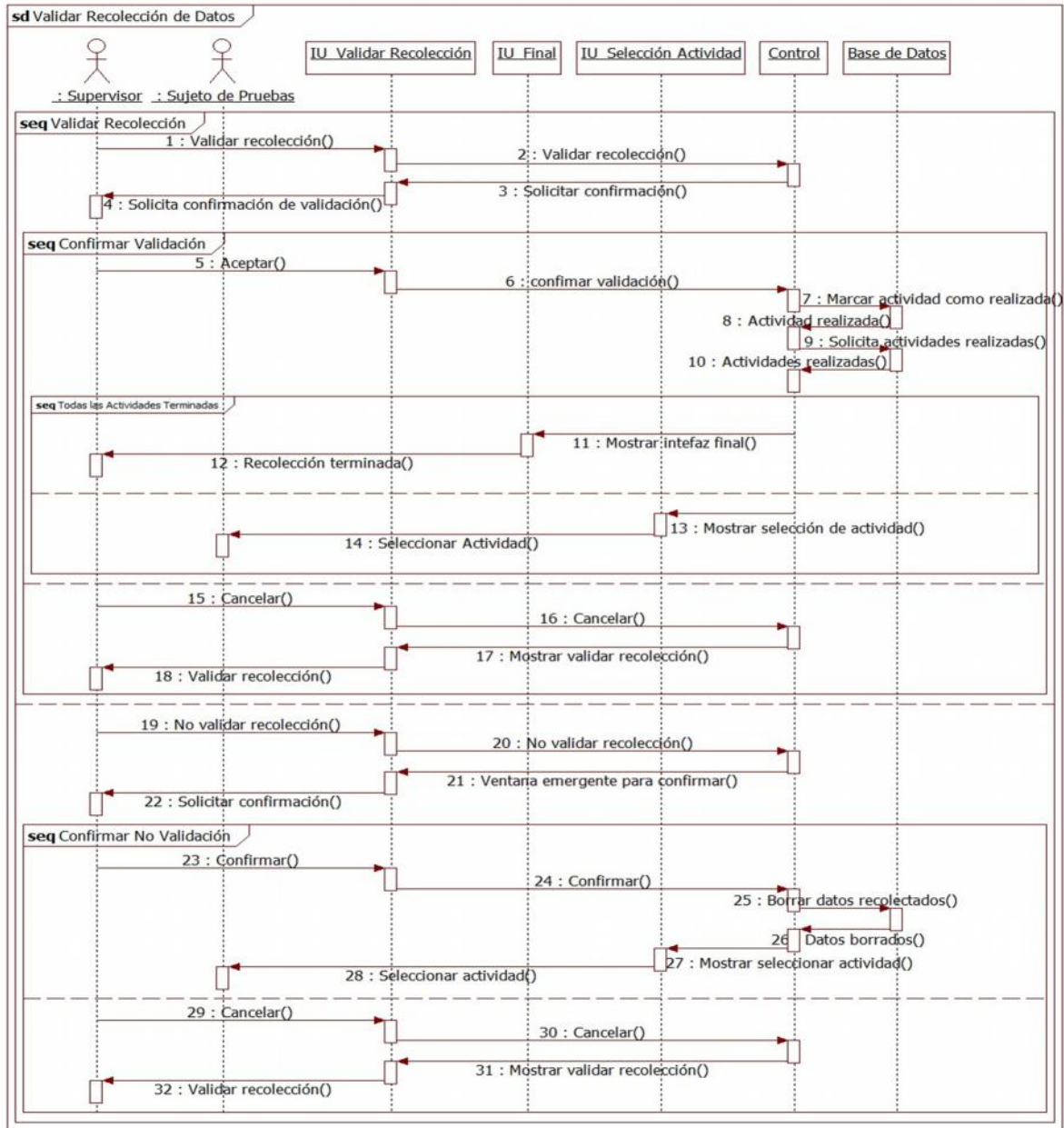


Figura 16. Diagrama de secuencia, Validar recolección de datos.

Diagrama de Secuencia para Proceso Interno en Control

El objeto llamado “control”, en los anteriores diagramas de secuencia, es donde se realizan todas las operaciones de control del sistema. Sin embargo, este objeto internamente tiene otros objetos que lo hacen efectivo, pero debido a que en todos los

casos se realiza la misma secuencia, solo se presenta una vez. En la Figura 17 puede verse la secuencia interna que tiene el proceso “control”.

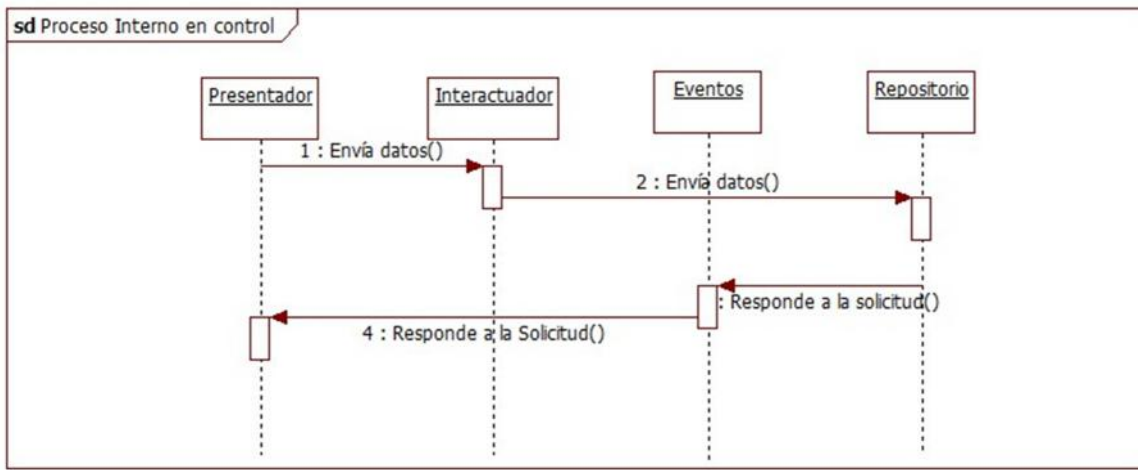


Figura 17. Diagrama de Secuencia, Proceso interno en control.

4.3.4. Vista de Desarrollo

4.3.4.1. Diagrama de Componentes

En la Figura 18 se puede observar el diagrama de componentes que describe cómo está dividida la aplicación móvil.

Los componentes están relacionados entre sí, ya sea por asociaciones, interfaces, dependencias, etc. Estos componentes pueden representar archivos, cabeceras, librerías, módulos, ejecutables o paquetes que posibilitan la construcción de una buena vista del sistema software. Los componentes presentes son:

- **Base de Datos:** Este componente representa el archivo que se crea, modifica y actualiza a lo largo de la aplicación. Entre los registros del archivo se encuentran los datos personales, así como los datos provenientes de cada uno de los dispositivos que integran el sistema y las actividades que quieran medirse.
- **Control Aplicación:** En este componente se encuentran todas las clases desarrolladas para gestionar el funcionamiento de la aplicación. Las clases para gestionar la base de datos (Database helper), las encargadas de comunicar las capas o los componentes de la aplicación (EventBus) y las que permiten la interacción del usuario (Acivities).
- **SDK Microsoft Band 2:** Este componente es suministrado por Microsoft y permite la integración de la banda con la aplicación, es decir, provee los mecanismos necesarios para capturar la información que la banda envía a través de Bluetooth. Además de la información de los sensores, también llegan datos de sincronización.
- **SDK Beacons:** Este componente tiene como objetivo permitir el tratamiento de los datos emitidos por cada beacon dentro del rango del dispositivo. Tiene funciones

para identificar y ordenar los beacons de un entorno o área. Así como para determinar la potencia percibida de los beacons, el nivel de batería o la descripción imprimible sobre ellos.

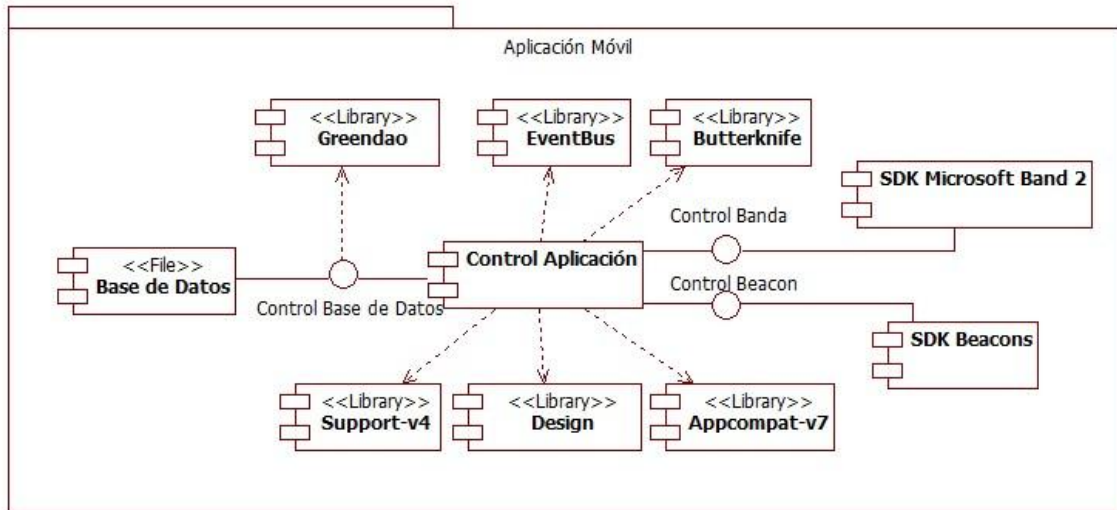


Figura 18. Diagrama de componentes

4.3.4.2. Diagrama de Paquetes

Para el diagrama de paquetes, de la Figura 19, se exhiben dos grandes paquetes que se relacionan para permitir el funcionamiento deseado de la aplicación. Uno de los paquetes es el paquete de recursos. En este se encuentran los layouts, imágenes y valores (cadenas, colores, dimensiones, etc.) de la aplicación. El otro paquete es el de archivos de clases java, el cual se subdivide en múltiples paquetes con la intención de organizar el desarrollo y procurar que el código sea entendible. Entre estos paquetes se encuentran:

- **control:** paquete que contienen todas las clases java que controlan el funcionamiento de la aplicación y donde se encuentran cuatro tipos de clases, a saber, presentador, interactuador, repositorio y eventos
- **vista:** paquete que aloja una actividad, relacionada con un layout en el paquete de recursos.
- **librerías:** software usado por los diferentes paquetes para realizar sus tareas
- **gestor de base de datos:** donde ocurre la administración de conexiones a la base de datos (greendao).

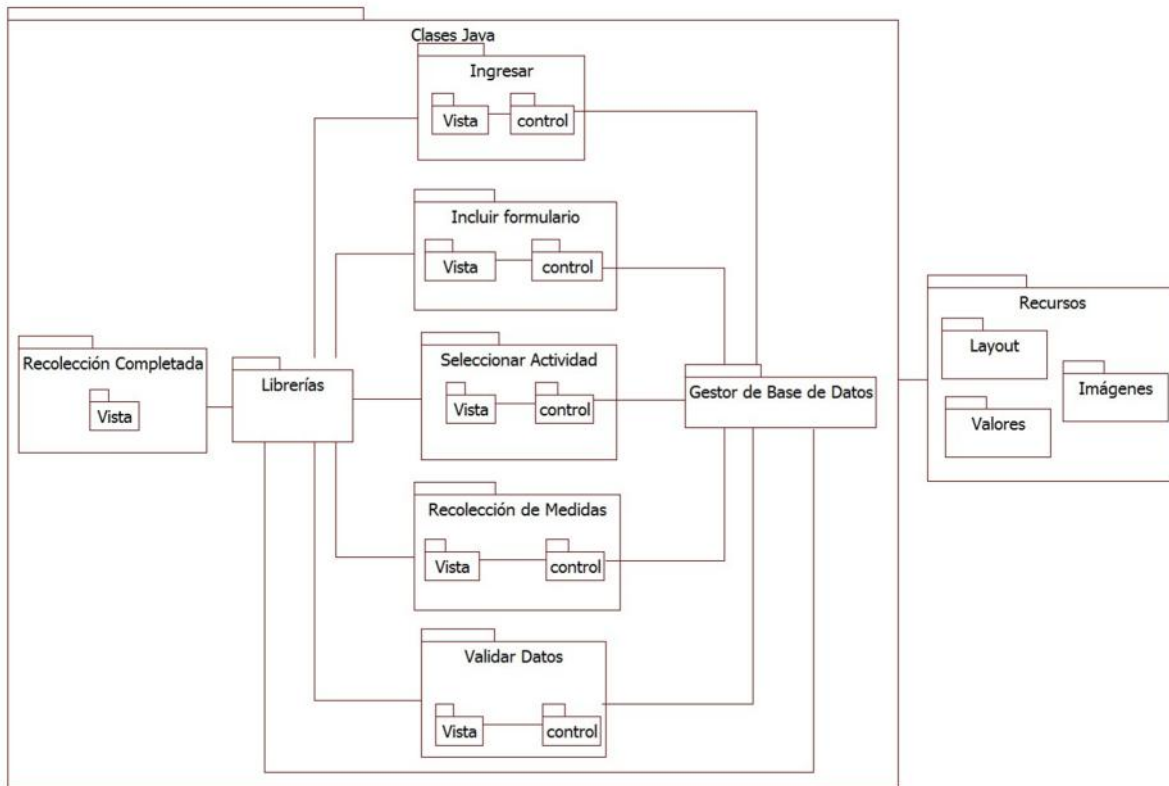


Figura 19. Diagrama de paquetes.

4.4. Clean architecture

Relacionado a la construcción del sistema y específicamente en lo concerniente a la arquitectura para el desarrollo software de la aplicación, hay una tendencia marcada a utilizar las especificaciones que Robert C. Martin describe en “The Clean Architecture” [114]. Las razones para esta elección son tratadas a lo largo de esta sección.

Clean Architecture es una integración de otras arquitecturas para el desarrollo software. Tal integración es realizada por Robert C. Martin y las arquitecturas que toma como referencia son:

- Arquitectura Hexagonal (Hexagonal Architecture) realizada por Alistair Cockburn y adoptada por Steve Freeman y Nat Pryce en [115].
- Arquitectura de cebolla (Onion Architecture) propuesta por Jeffrey Palermo
- DCI de James Coplien y Trygve Reenskaug.
- BCE propuesta por Ivar Jacobson en [116].

Estas perspectivas son similares y buscan la separación de intereses o preocupaciones. El término *separation of concerns* está enfocado en distinguir todas las preocupaciones importantes de una aplicación en unidades modulares y en permitir que tales unidades sean manejadas independientemente [117]. La forma en que se

llega a esa separación es dividiendo el software en capas. Entre las arquitecturas listadas, por lo menos existe una capa para las reglas del negocio y otra para interfaces. Como consecuencia de tal idea de desarrollo, se producen sistemas que son:

- Independientes de Frameworks: La arquitectura no depende de la existencia de librerías. Esto permite que los Frameworks se conviertan en herramientas, en lugar de atestar el sistema con sus limitaciones.
- Testeables: Las reglas del negocio pueden ser probadas sin una interfaz de usuario, base de datos, servidor o cualquier otro elemento externo.
- Independiente de la interfaz de usuario UI: La interfaz de usuario puede cambiar fácilmente, sin que se perciban cambios notables en el resto del sistema.
- Independiente de la base de datos: Posibilita intercambiar desde una base de datos relacional (Oracle, SQL Server, etc.) a una no relacional (p. ej. Mongo, Crunch, etc.) y viceversa. Esto se debe a que la lógica del negocio no está ligada a la base de datos.
- Independiente de factores externos: De hecho, las reglas del negocio simplemente no saben nada en absoluto sobre el mundo exterior.

Después de aclaradas las generalidades de la arquitectura, a continuación, se presenta el diagrama que resume la idea de Rober C. Martin, Figura 20 .

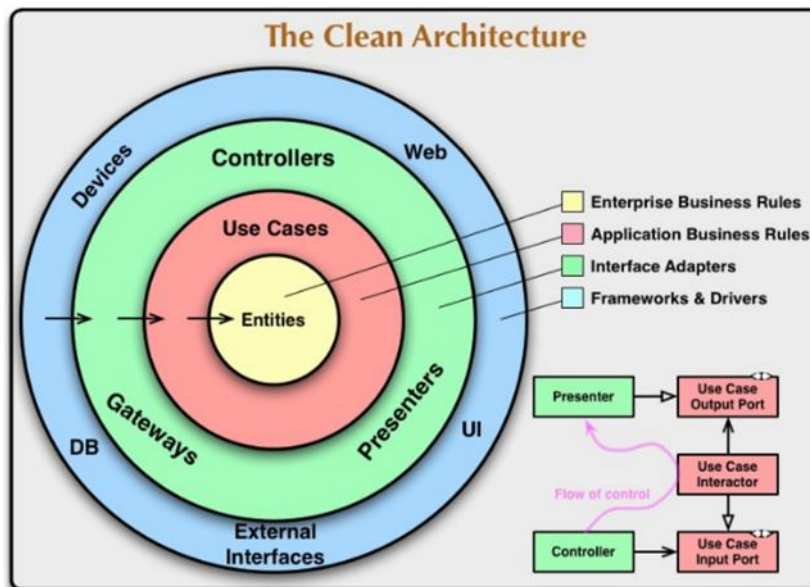


Figura 20. Arquitectura Clean [114].

Antes de hablar de cada una de las capas es importante mencionar la **regla de dependencia**.

Regla de dependencia: En la Figura 20, cada círculo del diagrama representa un área de software diferente. En general, el nivel de software aumenta entre más se aleje del centro. Los círculos exteriores son mecanismos mientras que los interiores son políticas.

La regla primordial que hace que esta arquitectura funcione es la regla de dependencia (o en inglés *The dependency Ruler*). Esta regla determina que las dependencias de código fuente sólo pueden apuntar hacia adentro. Es decir, nada en un círculo interior puede conocer algo sobre un círculo exterior. En particular, el nombre de algo declarado en un círculo exterior no puede ser mencionado por el código en un círculo interno (p. ej. funciones, clases, variables o cualquier otra entidad de software nombrada).

Entidades: Las entidades son objetos de negocio de la aplicación y encapsulan las reglas o lógica de negocio más general y de alto nivel. Además, son las menos propensas a cambiar por influencias externas.

Casos de uso: El software en esta capa contiene las reglas del negocio específicas de la aplicación. Además, encapsula e implementa, precisamente, los casos de uso del sistema. Estos casos de uso dirigen el flujo de datos desde y hacia las entidades que usan su lógica del negocio para alcanzar los objetivos del caso. En el contexto de este proyecto también se conocen como interactores.

Adaptadores de interfaz: La esencia de esta capa es convertir los datos del formato utilizado para los casos de uso y las entidades, al formato más conveniente para la base de datos. Para el caso especial de este proyecto, esta capa contiene totalmente la arquitectura MVP de una interfaz gráfica para el usuario. En consecuencia, presentadores, repositorios y vistas pertenecen a esta sección.

Controladores y frameworks: Es la capa más exterior y generalmente la conforman frameworks y herramientas como la base de datos.

En esta capa es donde van todos los detalles. La interfaz es un detalle, la base de datos es un detalle, los dispositivos son un detalle, etc.

Para terminar con el recorrido general de la arquitectura propuesta por Robert C. Martin se enumeran algunos aspectos relevantes:

- Los círculos son esquemáticos y no hay una regla que limite su número a únicamente 4. Es decir, según los requerimientos del proyecto puede surgir la necesidad de más capas.
- Los datos que cruzan los límites entre cada capa son estructuras de datos simples (p. ej. argumentos en una función, datos empaquetados en un hashmap, o construir los datos dentro en un objeto) u objetos de transferencia de datos.
- Cuando cualquiera de las partes del sistema se vuelve obsoleta, como la base de datos, o el framework, es posible reemplazarlos sin mucho esfuerzo.

Adaptación de la arquitectura para el desarrollo de aplicaciones Android

Gracias a los esfuerzos de la comunidad de desarrolladores de aplicaciones Android hay una adaptación bien aceptada de Clean Architecture. De hecho, el diagrama en la Figura 21 describe bien el resultado de los esfuerzos de la comunidad [118].

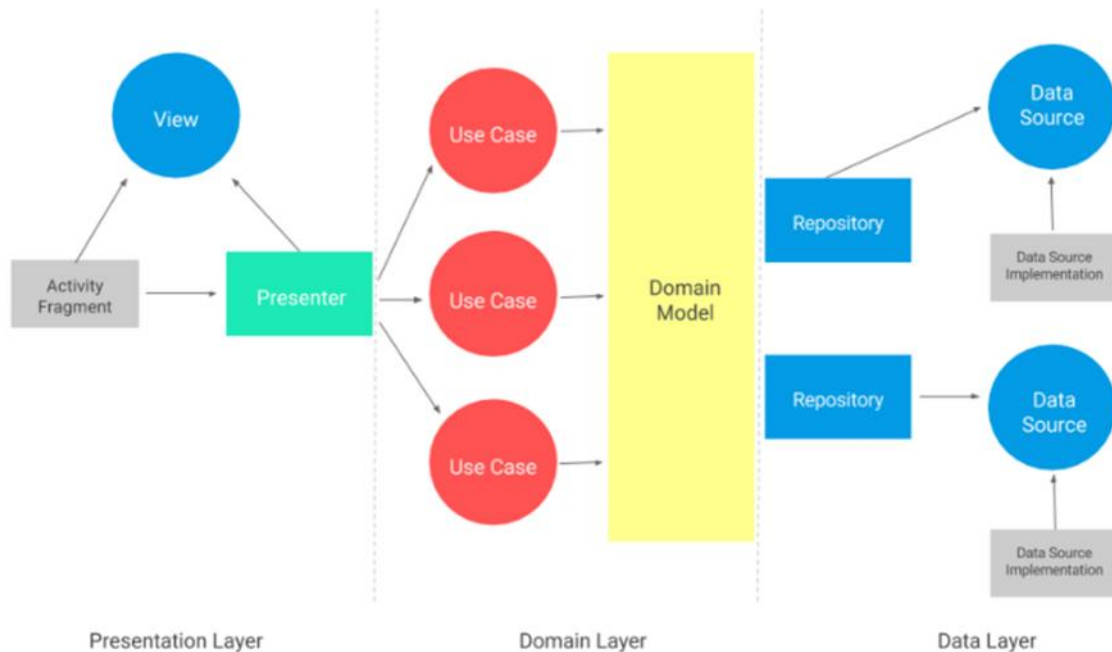


Figura 21. Arquitectura Clean para desarrollo de aplicaciones móviles [118].

4.4.1. Descripción de las capas

Capa de presentación

Esta capa se relaciona con las actividades (*activities*) y los fragmentos (*fragments*), los cuales están enfocados a cómo y cuándo mostrar las vistas al usuario. En relación a esto para el buen manejo de las vistas suele usarse patrones como MVC, MVP, etc. En este proyecto el patrón utilizado es el MVP (Modelo-Vista-Presentador).

Capa del dominio

La capa del dominio es donde ocurre la lógica del negocio y se enfoca en lo que debe hacer la aplicación.

Capa de datos

Esta capa provee los datos necesarios para el funcionamiento de la aplicación. Tales datos pueden llegar desde una base de datos, red o memoria. Con esto, otro aporte que hace la comunidad Android es el patrón repositorio. Este patrón permite abstraer

el origen de datos, con lo cual no importa de dónde lleguen los datos, siempre estarán definidas las acciones para su tratamiento.

4.5. Conclusiones

En este capítulo se presentó el modelado del negocio, con el objetivo de describir el entorno de aplicación del sistema. Además, se mostraron los requisitos funcionales, no funcionales y características que el sistema debe afrontar para conseguir un óptimo funcionamiento. Adicionalmente, se presentó el modelado de la aplicación móvil mediante la arquitectura de vistas 4+1. Para ello, se realizan los diagramas de casos de uso, clases, estados, actividad, secuencia, colaboración, componentes, paquetes y despliegue. Así, se obtiene una clara visión desde diferentes enfoques tanto de la arquitectura como de la aplicación y por tanto del sistema. Finalmente se describe la arquitectura "Clean Architecture", la cual sirve de guía para el desarrollo de la aplicación móvil, permitiendo la independencia entre diferentes capas, generando así un sistema de mejor calidad.

CAPÍTULO 5.

IMPLEMENTACIÓN DEL SISTEMA

En este capítulo se describe el proceso llevado a cabo para implementar el sistema usado para la recolección. Primero se presentan las herramientas y dispositivos utilizados para el desarrollo, (p. ej. ambiente de desarrollo, librerías, celulares, banda, beacons, etc.) y las respectivas configuraciones necesarias para integrarlas en pro del funcionamiento deseado. Luego se presentan las diferentes interfaces existentes en la aplicación móvil. Aplicación creada con el objetivo de recolectar datos de las diferentes actividades de forma cómoda y cuidadosa. Para cada una de estas interfaces se hace una descripción de su objetivo y funcionamiento. Por último, se presenta el esquema de la base de datos donde se almacenan las muestras recolectadas y la información general de cada participante.

5.1. Software y dispositivos

En esta sección se presentan las herramientas software que fueron necesarias y facilitaron el desarrollo del sistema de recolección. Además, se mencionan los dispositivos hardware que hicieron parte de este sistema.

5.1.1. Software

5.1.1.1. Genymotion

Genymotion es un emulador del sistema operativo Android [119]. Este emulador permite agilizar las pruebas sobre las aplicaciones. Unas de las características más relevantes de Genymotion son:

- Emula más de 40 dispositivos Android.
- Usa la cámara del computador como si fuese la del celular emulado
- Es compatible con Android Studio, Eclipse y Android SDK tools
- Puede variarse el nivel de batería para probar el funcionamiento de la aplicación
- Trabaja en Windows, Linux y Mac OS X.
- Puede emular el GPS y el Acelerómetro.

5.1.1.2. DB Browser For SQLite

DB Browser es una herramienta con un entorno gráfico , para crear, diseñar y editar bases de datos compatibles con SQLite [120]. Su interfaz de usuario es simple y fácil de manejar y evita muchos de los comandos complicados SQL; claro que también presenta una consola para ingresar consultas SQL personalizada. Unas de las características que se pueden encontrar en este entorno son:

- Crea y compacta bases de datos.
- Crea, define, modifica y elimina tablas.
- Crea, define y elimina índices.
- Examina, edita, agrega y elimina registros.
- Busca registros.
- Importa y Exporta registros como textos.
- Importa y Exporta tablas desde y hacia archivos CSV.
- Importa y Exporta bases de datos desde y hacia archivos de volcado SQL.
- Ejecuta comandos SQL e inspecciona los resultados.
- Mantiene un registro de todos los comandos SQL ejecutados por la aplicación.
- Grafica resultados, relacionando dos o más de sus columnas.

5.1.1.3. Ambiente de Desarrollo

Un ambiente de desarrollo en el campo de la ingeniería se describe como un conjunto de procesos y herramientas que permiten la creación, prueba y depuración de programas o productos software. Estos entornos se conocen normalmente como IDEs (Integrated Development Environment). [121], [122]

Android Studio es el IDE utilizado para el desarrollo de la aplicación en este trabajo. Esta herramienta es el entorno oficial de desarrollo para Android en todas sus plataformas, proporcionando gran estabilidad y cuenta con herramientas que permiten un desarrollo ágil, como *Instant Run* [123].

Para el manejo de versiones del proyecto se elige Bitbucket, que entre otras cosas permite: mantener el desarrollo seguro en la nube, tener control de versiones y trabajar de forma colaborativa. Todo lo anterior de manera privada, es decir, solo tienen acceso las personas que tengan permisos brindados por el administrador [124].

5.1.1.4. Librerías

Las librerías son un conjunto de implementaciones funcionales o un conjunto generalizado de algoritmos que se enfocan en el reúso de código [125] y mediante las cuales se facilitan las tareas de programación. [126]. En el desarrollo de este proyecto se usan las siguientes librerías:

Butterknife

Butterknife, que en español significa cuchillo para mantequilla, es una librería creada para Android y tiene el objetivo de facilitar la inyección de vistas y recursos dentro de las clases java de una aplicación (*Activities*). Esto se hace con el propósito de tener un código más organizado y evitar instanciaciones largas o complejas. [127].

Design

Con el lanzamiento de Android 5.0 (Lollipop), llegaron nuevos criterios de diseño, características, funciones y componentes. En consecuencia, para que las versiones anteriores de Android puedan acceder a tales innovaciones, nace una librería de soporte “design”. Algunos de los componentes que se pueden encontrar en esta librería son: botones de acción flotantes, pestañas, paneles laterales de navegación y snack bars. [128].

Support-v4

Las librerías de support v4 y appcompat v7 son altamente recomendadas por el sitio oficial de desarrolladores. Support v4 incluye soporte para componentes de aplicación, funciones de interfaz de usuario, accesibilidad, manejo de datos, conectividad de redes y utilidades de programación. [129].

Appcompat-v7

Appcompat v7, al igual que desing, ofrece compatibilidad para el uso de elementos de material design. En particular, está directamente enfocada en barras de acciones de la aplicación y en general “proporcionan una estructura visual y elementos interactivos conocidos por los usuarios”. [130].

GreenDAO

GreenDAO es un ORM (por sus siglas en inglés Object/Relation Mapping) de código abierto para Android [131]. Este ORM tiene como objetivo simplificar muchas de las tareas tediosas dentro del manejo de una base de datos SQL y se caracteriza por:

- Obtener alto rendimiento
- Ser fácil de usar
- Poco consumo de memoria
- Ser una librería liviana
- Ofrecer encriptación de la base de datos
- Tener una comunidad fuerte de desarrolladores y aportantes
- Ofrecer generación de código

En la Figura 22, se muestra la estructura básica de este componente.



Figura 22. Estructura básica de greenDAO [131].

En cuanto a rendimiento, la Figura 23 muestra diferencias significativas con otras opciones de gestión de base de datos. Las características, el rendimiento de greenDAO, los importantes aportes de greenrobot para la comunidad de desarrollo de aplicaciones Android y las pruebas de eficiencia en cuanto al uso de la batería mostradas en [132] son los aspectos más relevantes para escoger greenDAO dentro de este trabajo de grado.

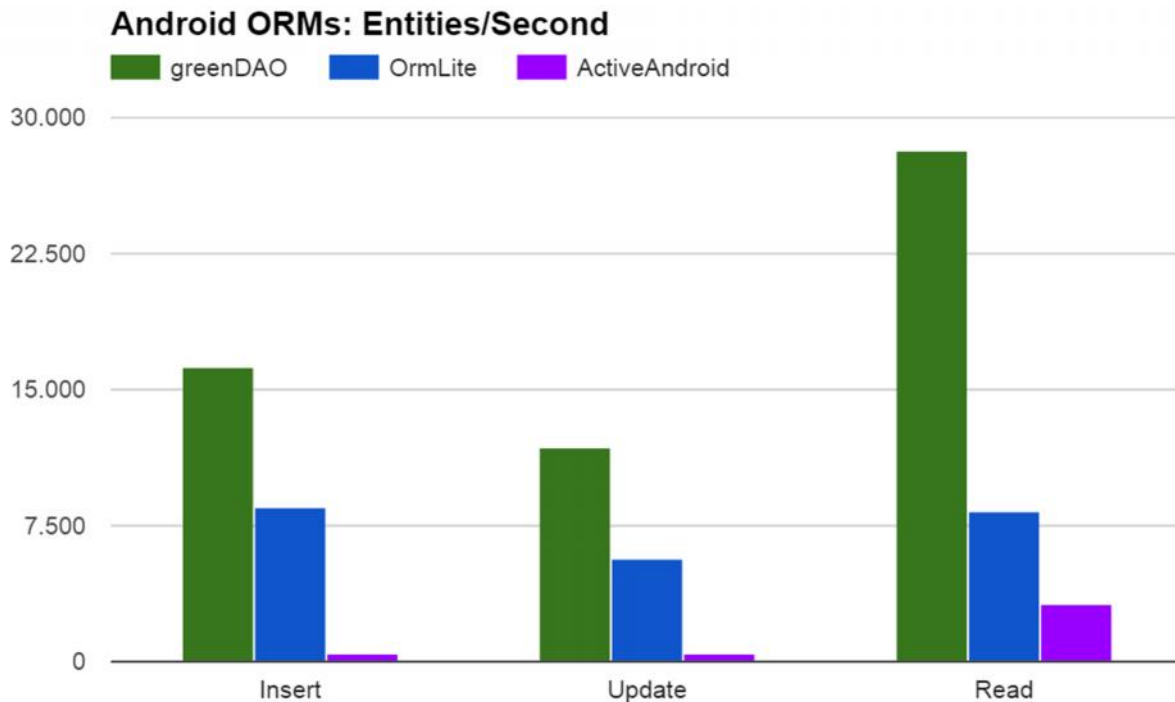


Figura 23. Características de rendimiento de greenDAO [132].

La imagen muestra algo muy importante y es el elevado número de lecturas por segundo que permite realizar este framework. Lo anterior es realmente importante si se considera que una de las funcionalidades del sistema es validar los datos recolectados. Para lograr tal verificación el sistema debe ser capaz de leer todas las muestras, encontrar las muestras de una actividad específica y luego buscar nulos. Esto requiere analizar más de 170000 muestras en caso de estar realizando la recolección de datos de la actividad final.

EventBus

EventBus es una librería basada en el patrón *publisher/subscriber* (editor/suscriptor) [133]. EventBus hace posible comunicar clases que están completamente desacopladas sin ningún inconveniente gracias a la arquitectura presentada en la Figura 24:

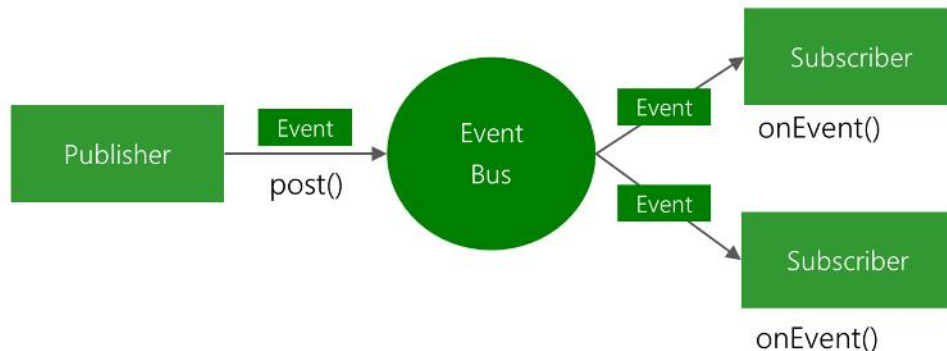


Figura 24. Arquitectura soportada por Event Bus [133].

En forma simple, EventBus es un estilo de comunicación entre componentes de la forma editor/suscriptor. El editor envía un evento a una entidad determinada que no es directamente el suscriptor. Mientras uno o varios suscriptores están siempre al pendiente de recibir los eventos que se hayan recogido del editor [133].

Algunas de las características más importantes de EventBus son:

- Simplifica la comunicación entre componentes
- Desacopla al emisor y receptor.
- Trabaja bien con actividades, fragmentos e hilos
- Evita dependencias complejas y propensas a errores, además de problemas del ciclo de vida.
- Es rápido y está específicamente optimizado para alto rendimiento.
- Es liviano, ya que pesa menos de 50kB
- tiene características avanzadas como hilos de entrega, prioridades de suscriptor, etc.

Para este trabajo y como se describe en los diagramas de clases y de secuencia, EventBus es utilizado para realizar la comunicación de los repositorios con cada presentador.

5.1.2. Dispositivos

Los dispositivos utilizados son los que se describen en el Capítulo 3, en la subsección “3.1.2.”.

5.2. Aplicación móvil

Una aplicación móvil es un desarrollo software dirigido a dispositivos móviles, como teléfonos inteligentes, bandas inteligentes, entre otros. De acuerdo al estándar 729 de la IEEE: “Software es el conjunto de los programas de cómputo, procedimientos, reglas, documentación y datos asociados, que forman parte de las operaciones de un sistema de computación” [134]. En este estándar se basa el concepto de software, sin discriminar el dispositivo de destino, sin embargo, para los desarrolladores es claro que el software para PC y dispositivos móviles difieren en algunos conceptos.

Las características básicas de una aplicación móvil son:

- Fáciles de usar, con interfaces intuitivas adaptadas a pantallas pequeñas.
- Diseñadas para periodos cortos de tiempo.
- Evitar el gasto de batería innecesario.
- Respuesta rápida.
- Implementación de las características necesarias y no las posibles.
- Reducción del número de elementos en pantalla
- Reducción del número de pasos para llevar a cabo una tarea

Para este proyecto se define a un dispositivo móvil como el centro de la arquitectura y es aquí donde se procesan todos los datos que llegan de los beacons y la banda. De esta manera apareció la necesidad de una aplicación móvil. En el capítulo 4 se encuentra el modelado y la arquitectura definida para el desarrollo de la aplicación. A continuación, se describe cada una de las secciones o interfaces desarrolladas en la aplicación.

5.2.1. Material Design

El sitio oficial para desarrolladores de Android define Material Design como: “una guía integral para el diseño visual, de movimientos y de interacción en distintas plataformas y dispositivos.” El sitio oficial de Material define que es una especificación, un “documento vivo” o un documento abierto que se actualiza a medida que continúa el desarrollo de los principios y especificaciones del diseño del material. En cierto modo, Material Design es un lenguaje diseñado para el desarrollo de interfaces de usuario y su objetivo es sintetizar los principios clásicos de buen diseño con la innovación y las posibilidades que brindan la ciencia y la tecnología. Por lo tanto, al usar Material Design se consigue que las vistas sean agradables y fáciles de usar para el usuario [135]. Inclusive cuenta con recomendaciones para la combinación de los colores.

La aplicación va a entrar en contacto tanto con los participantes como con el grupo de investigadores. Así que, se requiere enfatizar en aspectos como la usabilidad de la aplicación, siendo este un factor de alta relevancia dentro del sistema requerido. De hecho, sin este enfoque sería difícil obtener una recolección de datos rápida y sin inconvenientes, pues los participantes (tanto quienes monitorean el desarrollo de la recolección de datos, como los participantes) necesitan interactuar fluidamente con la

aplicación para agilizar cada proceso. Es así que se considera utilizar la especificación material design para el desarrollo de la aplicación móvil.

La aplicación es planteada con el propósito de servir en futuros desarrollos que integren nuevas variables y contextos. Es por ese motivo que las interfaces deben caracterizarse por ser intuitivas y agradables a la vista.

5.2.2. Inicio (Información de la Investigación)

La Figura 25 presenta la interfaz con lo que se encuentra el participante del estudio cuando inicia la aplicación por primera vez.

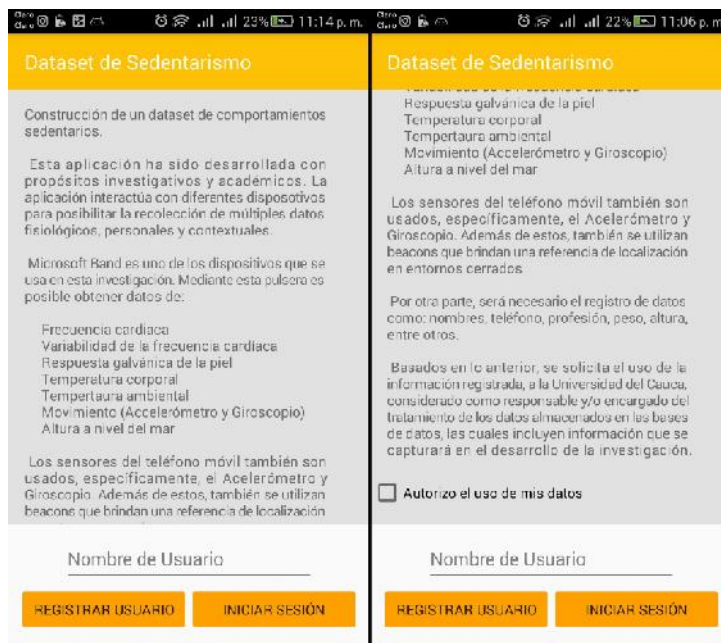


Figura 25. Interfaz de inicio.

En esta interfaz se pretende instruir al participante acerca de la investigación, los dispositivos usados y los datos recolectados. Todo esto tiene el objetivo de informar al usuario para que tenga el conocimiento necesario acerca del estudio y apruebe o no su participación dentro del mismo. La información presentada en esta interfaz se muestra en el anexo B:

Adicionalmente, dentro de esta interfaz se presentan dos posibilidades para continuar con el proceso. La primera y en caso de existir un usuario registrado, es un inicio de sesión simple. La segunda es un registro que se posibilita si y sólo si la casilla de verificación, que concede los permisos para el tratamiento de la información, está marcada.

Finalmente, la aplicación cuenta con control de sesiones. Entonces, si un usuario ingresa a la aplicación, se presentan dos situaciones: La primera, cuando no existe

una sesión y el participante debe gestionar el inicio de sesión mediante su nombre de usuario. La segunda cuando el participante haya salido de la aplicación (sin cerrar o eliminar la sesión) y accede directamente a la interfaz de actividades.

5.2.3. Formulario Personal

La Figura 26 muestra los campos que debe llenar el participante antes de iniciar la recolección de datos por actividad o comportamiento.

Datos requeridos	
Nombre de Usuario	
Nombre	
Apellido	
Edad	
Sexo	Masculino
Peso (Kg)	
Altura (cm)	
cintura (cm)	

Datos requeridos	
Sexo	Masculino
Peso (Kg)	
Altura (cm)	
cintura (cm)	
Profesión	
example@mail.com	
Fuma	Varios al día
Bebe Licor	A diario
Medio de Transporte Habitual	Transporte Público
CONTINUAR	

Figura 26. Formulario personal

Entre los datos requeridos se encuentran: nombre, edad, peso, altura, diámetro de la cintura, profesión y correo electrónico. El objetivo de capturar estos rasgos es caracterizar la población y aumentar la información que pueda utilizarse en estudios futuros. Además, se solicitan otros comportamientos habituales como la frecuencia de consumo de alcohol, la frecuencia con la que fuma o el medio de transporte que generalmente usa. Estos estilos de vida pueden brindar características de una población, que algunos estudios mencionan como influyentes en factores como la frecuencia cardíaca, frecuencia respiratoria y estado físico y de salud en general. Es necesario aclarar que los nombres y apellidos de las personas no van a estar presentes en el dataset final por seguridad y protección de identidad.

5.2.4. Selección de Actividades

En la Figura 27 se pueden observar los distintos *tabs* o pestañas definidas en la aplicación.

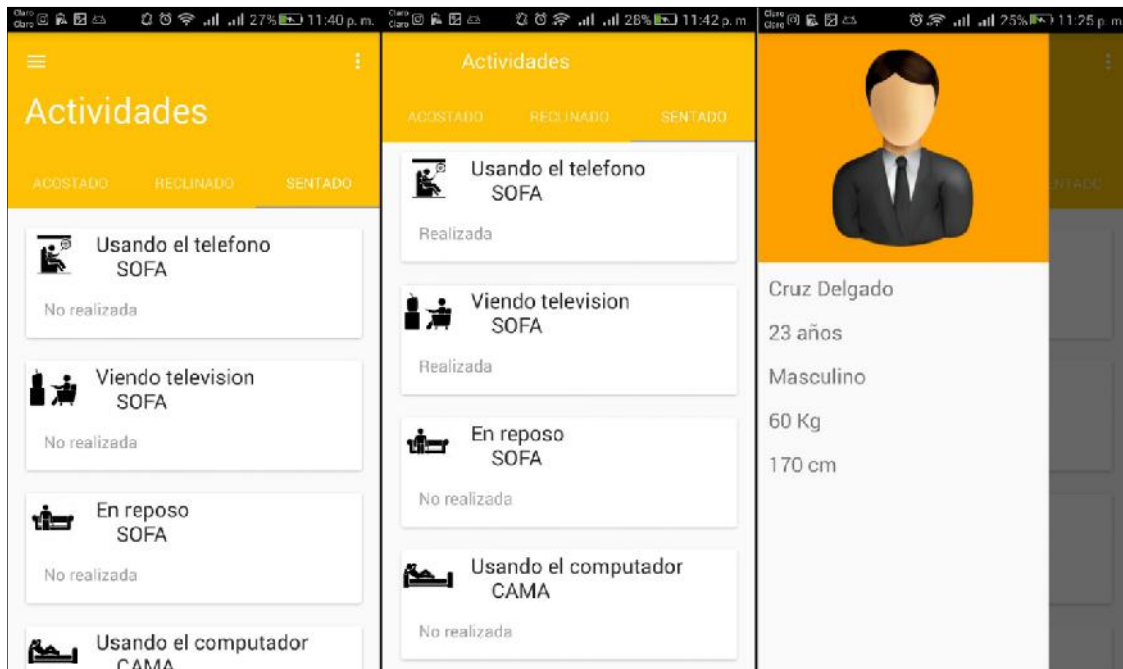


Figura 27. Selección de actividades

Esta interfaz implementa la mayor cantidad de componentes descritos por material design. Dentro de esta interfaz es posible interactuar con: pestañas, un panel de navegación lateral, *cardviews*, diálogos y notificaciones. En el panel lateral se presentan los datos básicos de la persona, así como una imagen que representa el género de la persona.

En la vista principal se tiene una barra de app (*AppBar*) con el nombre de la interfaz y 3 pestañas (*Tabs*). Cada pestaña lleva el nombre de una postura relacionada a comportamiento sedentario. Para mostrar el contenido de cada pestaña se usan fragmentos (*fragments*) y en cada uno se muestra los diferentes *cardviews* que contienen comportamientos sedentarios específicos, su respectiva localización y una descripción para saber si la actividad ya fue realizada o no. Además, la barra de app tiene la cualidad de reducirse cuando hay un desplazamiento dirigido a la parte inferior de la pantalla y volver a su estado inicial cuando hay un desplazamiento contrario (*collapsing*, utilizado muy a menudo por permitir organizar y aprovechar el espacio de la pantalla).

La interacción del usuario y la aplicación consiste en:

- El usuario puede moverse entre pestañas (posturas)
- El usuario de la aplicación puede seleccionar cualquiera de las actividades que no hayan sido realizadas.
- Con tocar alguno de los *cardview* (actividades) el sistema identifica la actividad seleccionada y salta a la interfaz para capturar los datos de esa actividad específica.

5.2.5. Captura de Datos

La Figura 28 muestra la interfaz encargada de iniciar, reiniciar o terminar la recolección para pasar a la interfaz de validación.

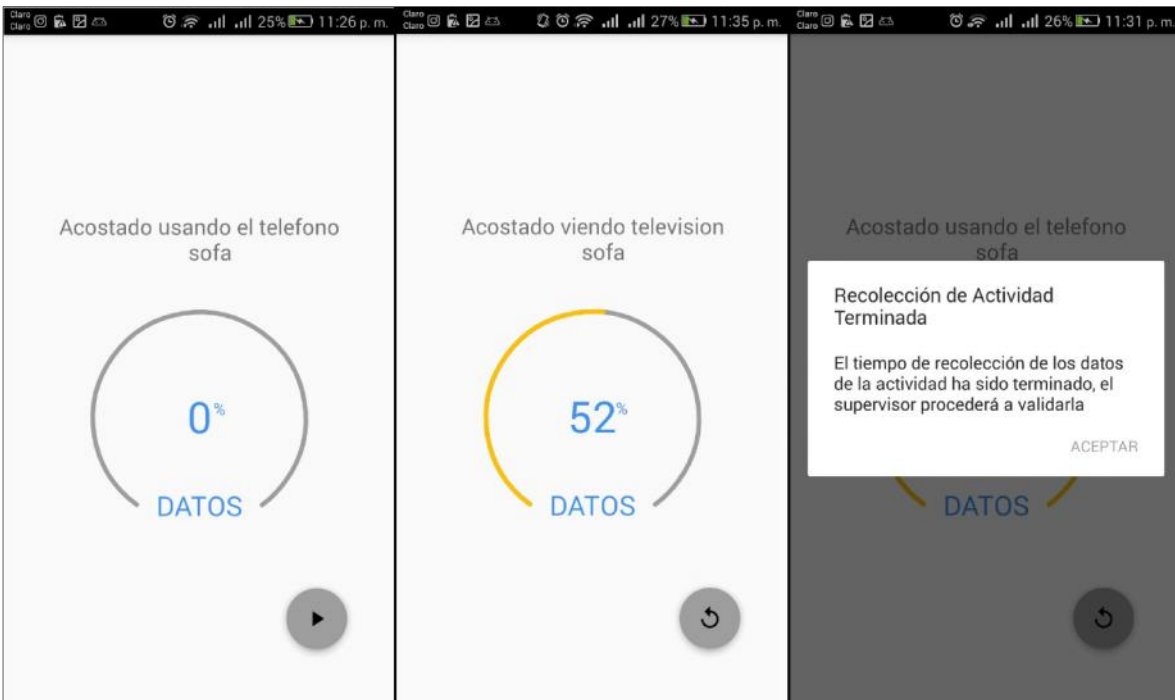


Figura 28. Captura de datos.

Esta interfaz es una de las más importantes en la recolección de datos. Aquí inicia y termina la comunicación con los demás dispositivos del sistema. Además, se implementan los métodos necesarios para la captura y almacenamiento de la información que llega de todos los sensores dispuestos a la actividad o comportamiento seleccionado. La interfaz que se muestra al usuario es básicamente una barra de progreso circular, que depende del tiempo elegido para la toma de datos (3' 10'') y un botón que permite iniciar o reiniciar el proceso. En cuanto al tiempo de captura de datos es preciso mencionar que los primeros 40 segundos no son almacenados, por tanto, el tiempo de datos recolectados finalmente son 2 minutos y medio. Esto ocurre para evitar guardar información que no es relevante para el caso

de este proyecto (p. ej, cuando la persona está adoptando la posición o cuando está ubicando los dispositivos en su lugar). También ocurrió que la primera lectura de los dispositivos no es inmediata, entonces este tiempo es ideal para esperar a que todas las variables tengan un valor diferente de nulo cuando inicia la recolección.

5.2.6. Validación de la Captura

En la Figura 29 se presenta la interfaz que ve el supervisor para tomar la decisión sobre continuar con la recolección de datos de otra actividad o repetir la recién hecha.

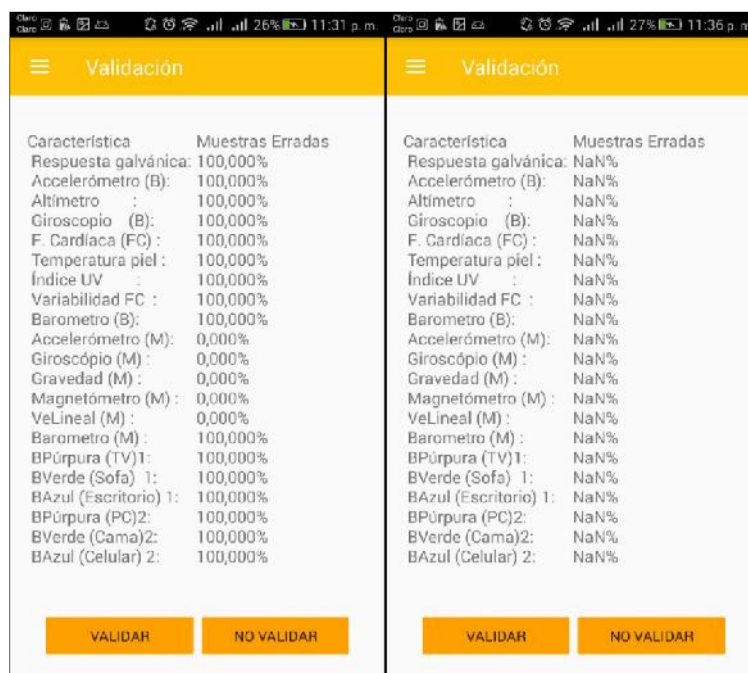


Figura 29. Validación de la captura.

El objetivo de esta interfaz es presentar el porcentaje de datos nulos en la captura de los datos. Con esta información, el supervisor puede validar la recolección de datos de la actividad correspondiente o descartar los datos recolectados y volver a iniciar la captura. La validación se lleva a cabo mediante un botón y enseguida con la confirmación en un diálogo. Esta interfaz brinda una ayuda considerable, pues con los valores dispuestos se puede inferir si uno de los dispositivos está funcionando correctamente o no. La validación es particularmente útil, pues la banda deja de enviar los datos mediante bluetooth cuando tiene la carga baja; el bluetooth de los celulares es apagado constantemente para ahorrar energía y en ocasiones no se vuelve a activar.

En el porcentaje de datos errados es posible encontrar valores de 0% a 100%. Donde 0% representa que no se han presentado inconvenientes y todos los datos se han capturado con éxito; mientras 100% sugiere que todos los datos son nulos y posiblemente el sensor o el bluetooth se encuentren apagados. En la captura de

cualquiera de las variables, se definió como porcentaje máximo de valores nulos un 19%, lo cual asegura conseguir un mínimo 2 minutos de datos por actividad recomendados en [105].

5.2.7. Recolección Terminada

En la Figura 30 aparece la interfaz que indica el final de la recolección para cada sujeto de pruebas.

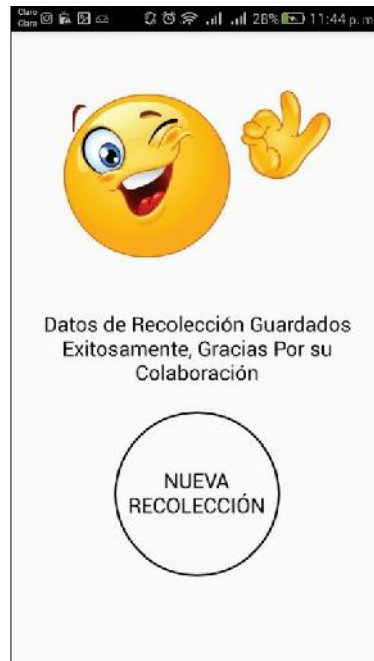


Figura 30. Recolección terminada

Esta es la interfaz que se encuentra después de la última actividad y se termina la recolección. En esta interfaz se agradece, al sujeto de pruebas, por la colaboración prestada y se posibilita un nuevo registro. La sesión del usuario es destruida o finalizada cuando se crea esta interfaz.

5.3. Base de datos

La base de datos, usada para almacenar los datos recolectados de los múltiples sensores disponibles, se modela de la siguiente forma:

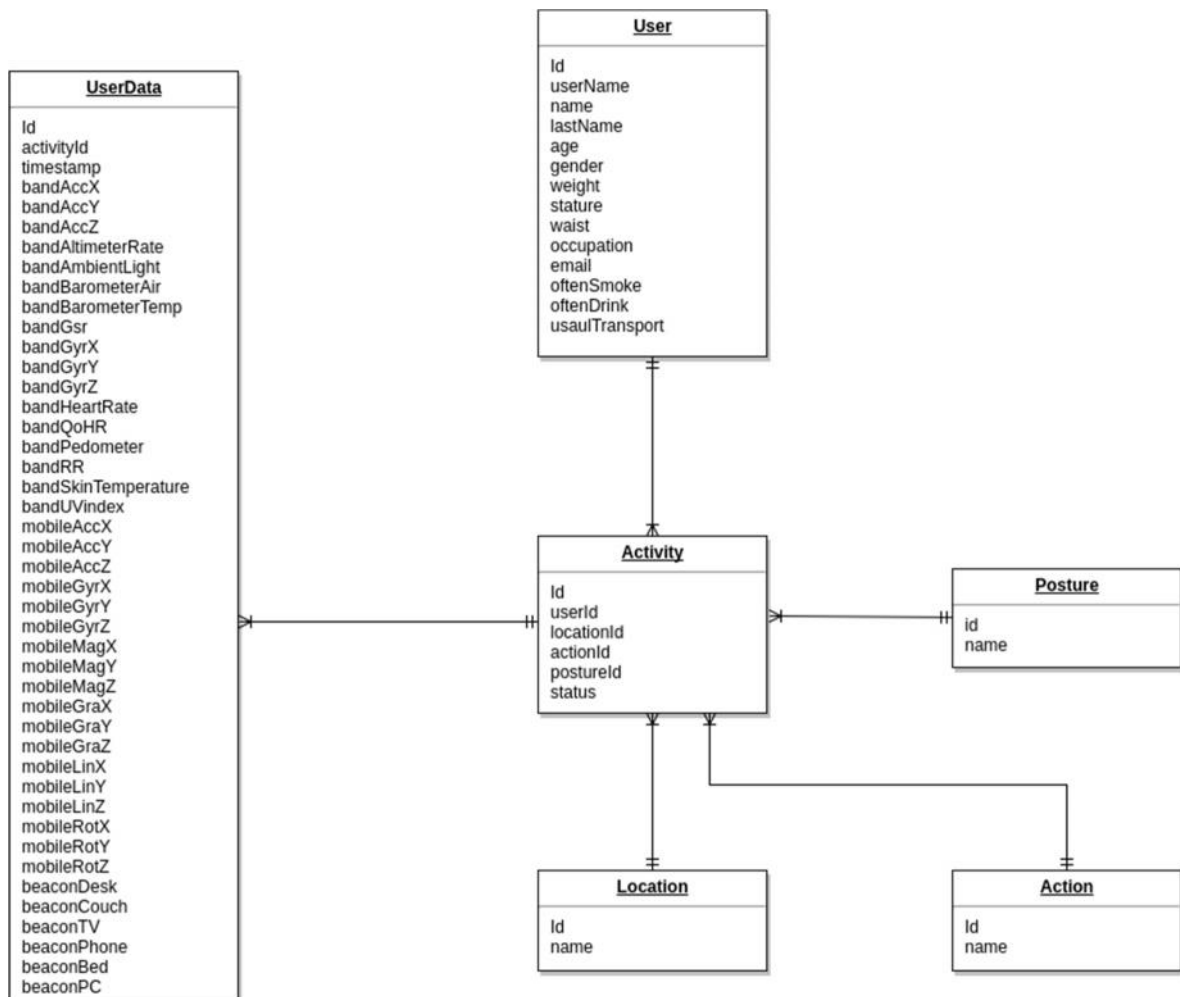


Figura 31. Modelo de la base de datos.

La notación para las relaciones en este diagrama es Crow's Foot. En la tabla "UserData" se encuentran los datos que llegan de los sensores, tanto del teléfono como de la banda y los beacons. Además, esta tabla presenta una marca de tiempo *timestamp* que informa sobre el instante en que se realiza la inserción del objeto en la base de datos. Es válido hablar de insertar un objeto puesto que es lo que en realidad pasa, gracias al componente greenDAO utilizado para gestionar la base de datos SQLite. Finalmente, la tabla "Activity" se relaciona con "UserData" mediante una relación *one-to-many* (uno a muchos) y se lee así: una actividad (comportamiento sedentario) **contiene** de uno a muchos datos fisiológicos y contextuales.

En la tabla "User" son almacenados los datos que se requieren en el formulario. Esta tabla presenta una relación *one-to-many* con la tabla Activity. Esto se leería de la siguiente forma: Un usuario **realiza** de una a muchas actividades (en este caso comportamientos sedentarios).

Finalmente, las tablas “Location”, “Action” y “Posture” también presentan una relación de uno a muchos (*one-to-many*) con la tabla “Activity”. Esto es particularmente útil en caso de presentarse la necesidad de conocer o listar los comportamientos sedentarios que se realizan en una determinada localización o postura.

5.4. Conclusiones

En el presente capítulo fue presentado el desarrollo de la aplicación móvil. Entonces, el software y los dispositivos utilizados fueron los primeros aspectos en describirse. Luego se especificaron ciertos detalles de la aplicación creada, presentando cada una de las interfaces y las tareas que cumple el sistema, tanto visual como internamente. Finalmente fue presentada la arquitectura de la base de datos donde se almacenan los datos recolectados, así como las relaciones entre las diferentes tablas.

CAPÍTULO 6.

RECOLECCIÓN Y LIMPIEZA DEL DATASET

En el presente capítulo se presenta todo lo referente a la recolección llevada a cabo. Entre otras cosas, son indicados los aspectos tenidos en cuenta para la toma de datos (p.ej. la posición de cada dispositivo durante el proceso o el tiempo gastado en cada actividad). Además, se describe el tratamiento realizado a los datos después de efectuada su captura. Esto con el propósito de limpiar y preparar el dataset creado para ponerlo a disposición del público. Esta sección del trabajo pretende abordar las actividades y sugerencias de la metodología CRISP-DM descritas en la fase 2: entendimiento de los datos.

6.1. Recolección de los 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 6-1 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 6-1: Características de los participantes.

Cada actividad fue realizada durante 3 minutos y 10 segundos, mientras cada participante usaba la Microsoft band 2 en la muñeca, un smartphone fijo en la cadera y otro en el bolsillo como se muestra en la Figura 32. Además, el teléfono en el bolsillo fue manipulado en las actividades que requerían el uso o interacción con el dispositivo móvil. Cada sujeto realizó los 23 comportamientos sedentarios y las dos actividades adicionales listadas en la Tabla 6-2, la tabla también muestra cómo están nombrados cada uno de los comportamientos sedentarios dentro de las bases datos.

Comportamiento sedentario	ID del comportamiento dentro de la base de datos.
Sentado usando el computador en el escritorio	1
Sentado usando el computador en la cama	2
Reclinado usando el computador en el escritorio	3
Reclinado usando el computador en la cama	4
Sentado usando el teléfono en el escritorio	5
Sentado usando el teléfono en la cama	6
Sentado usando el teléfono en el sofá	7
Reclinado usando el teléfono en el escritorio	8
Reclinado usando el teléfono en la cama	9
Reclinado usando el teléfono en el sofá	10
Acostado usando el teléfono en la cama	11
Acostado usando el teléfono en el sofá	12
Sentado viendo televisión en el sofá	13
Reclinado viendo televisión en el sofá	14
Acostado viendo televisión en el sofá	15
Sentado en reposo en el escritorio	16
Sentado en reposo en el sofá	17
Sentado en reposo en la cama	18
Reclinado en reposo escritorio	19
Reclinado en reposo en el sofá	20
Reclinado en reposo en la cama	21
Acostado en reposo en el sofá	22
Acostado en reposo en la cama	23
Estar de pie	24
Caminando	25

Tabla 6-2: Comportamientos sedentarios y su identificador en la Base de Datos.

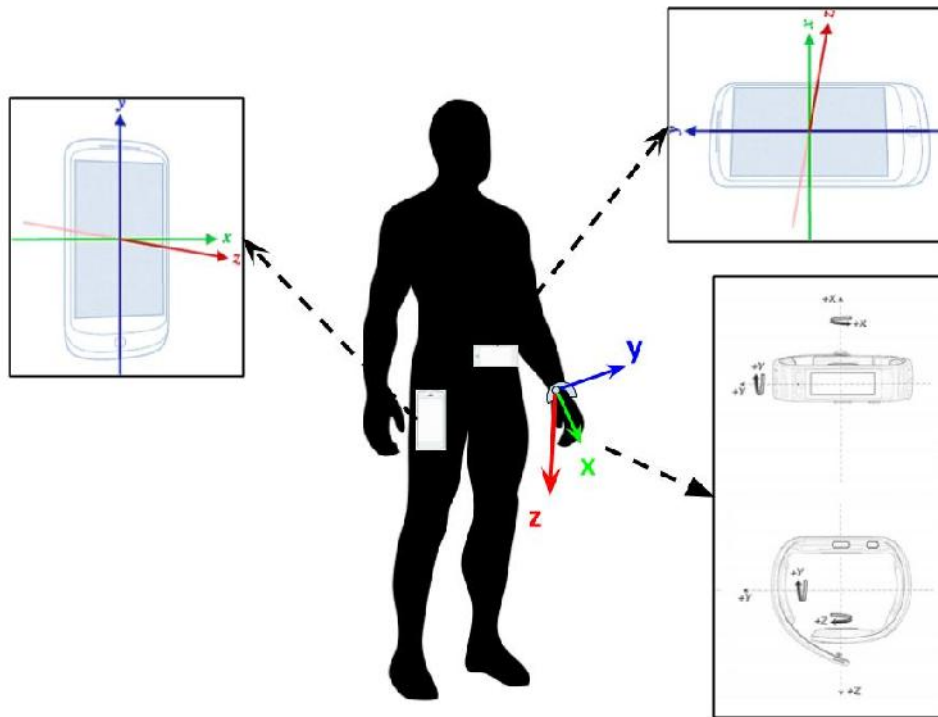


Figura 32. Posición de los dispositivos

El método usado en la recolección de los datos fue el seminatural. Ya que, la recolección de datos se hizo en un ambiente natural al participante (su casa). Pero el sujeto de pruebas estuvo realizando actividades ya definidas y hubo una persona supervisando su desarrollo, lo cual puede influenciar el comportamiento normal de las personas comparado con un ambiente natural. Aunque debido a que la recolección se hizo con familiares cercanos a los supervisores, puede pensarse que la influencia en la conducta de los sujetos, debido a la presencia del supervisor, no es muy relevante.

Por otra parte, puesto que la recolección es realizada en lugares diferentes, hay una influencia notoria en los datos de los beacons. Esto es algo razonable, ya que la distancia entre las habitaciones, salas y estudios cambia de una casa a otra.

6.1.1. Posición de los dispositivos

Para cada uno de los dispositivos utilizados se definieron determinadas posiciones en el cuerpo del participante. Las posiciones definidas se mantuvieron en todo el proceso de recolección. Esta consideración tuvo el objetivo de mantener las mismas condiciones entre los participantes durante la etapa de recolección y además procurar que los datos no sean negativamente afectados por tales factores.

Beacons: antes de iniciar la recolección de datos se pidió a cada participante ubicar 3 beacons de la siguiente manera: uno en el sofá donde ve televisión, otro en su cama

y otro en su escritorio. Los beacons restantes, a diferencia de los anteriores que permanecieron activos durante toda la recolección de datos, solo debían activarse cuando alguno de los tres dispositivos (TV, computador o smartphone) estuviera siendo usado. Debido a que los Estimote beacons usados no tienen la capacidad de conectarse al dispositivo, se utilizó la funcionalidad de voltear el beacon para ser desactivado. Esta funcionalidad es aclarada en la Figura 33. De esa forma, cuando el participante realizó un CS que involucrara el uso de un dispositivo, como por ejemplo el televisor, el beacon asociado se voltea para que sea detectable por los smartphones.

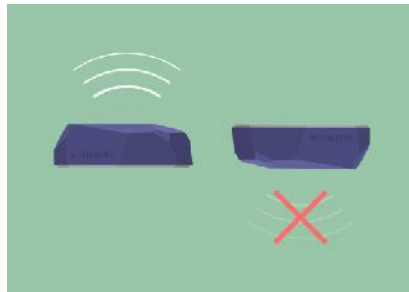


Figura 33: Activación del beacon

Microsoft Band 2: La banda se ubicó en la muñeca del brazo no dominante de cada participante y se dispone como un reloj común. Se elige el brazo no dominante por las recomendaciones encontradas en la revisión sistemática. Además, la banda fue ubicada de tal manera que los botones de encendido y selección queden apuntando hacia el antebrazo del participante.

Celular LG G3: Este equipo simuló el celular que usan los participantes en su día a día. Entonces, puede ubicarse en el bolsillo o en las manos del participante. El sujeto usó el celular como quiso mientras lo tuvo en sus manos. En los comportamientos que no impliquen el uso del celular, el dispositivo se acomoda en el bolsillo de forma vertical con la pantalla hacia el frente y la cámara hacia arriba. Es decir, la parte trasera del dispositivo roza la pierna.

Celular Huawei Ascend P7: Este celular estuvo fijo en la cintura de la persona con la ayuda de un estuche como se muestra en la Figura 32. Se ubicó horizontalmente con la pantalla apuntando hacia afuera.

Además, antes de realizar la recolección de los datos de cada persona, se le informaba de la existencia de un consentimiento informado (Anexo C), el cual era leído en su compañía y se respondía cualquier duda al respecto. Posteriormente era firmado como constancia de su consentimiento para realizar la recolección y la libertad para el uso de datos con fines científicos.

6.2. Descripción inicial de datos recolectados

Para la visualización de estos datos se usó la herramienta DB Browser for SQLite, con la cual es posible observar el estado de las bases de datos y brindar una idea de las posibles acciones para limpiar su contenido.

Durante la recolección, se creó por cada celular una base de datos diferente para cada persona. Esto sucedió por la necesidad de verificar que los dispositivos estaban trabajando con normalidad. Para lo cual, el método usado es un recorrido por los registros en busca de nulos, que tardaría más si se almacenan todos los registros en una sola base de datos.

A continuación, se presentan las primeras observaciones realizadas sobre las bases de datos.

6.2.1. Datos iniciales y primer sondeo de variables

Con la primera observación realizada de las bases de datos fue posible discernir sobre los datos útiles dentro del dataset y aquellos que no son significativos. Entre los atributos no relevantes se encuentran:

- Calidad de frecuencia cardiaca en la banda: Este dato permite saber si la banda se encuentra midiendo la frecuencia cardiaca. Sin embargo, después de observar la columna que almacena tal información, se determinó que este dato es irrelevante. Ya que, en todas las filas se almacena “LOCKED”, lo cual significa que la banda está capturando adecuadamente la frecuencia cardiaca del usuario.
- Índice UV: Esta variable ofrece la cantidad de radiación UV alrededor de la persona. No obstante, los datos se recolectaron siempre en ambientes cerrados y el sensor de radiación siempre se mantuvo en “NONE”.

Vale la pena aclarar que, para este trabajo, los atributos de calidad de frecuencia cardiaca, e índice UV no son relevantes. Pero, en trabajos donde se desee tratar de manera simultánea ambientes cerrados y exteriores, la medida de UV posiblemente sería sumamente útil. En conclusión, los datos descritos no son sobresalientes para este trabajo, pero eso no significa que sea así para otros estudios. Es así que en la aplicación móvil desarrollada no se eliminan las funciones que adquieren la información de los sensores en cuestión.

De los datos personales recolectados también hay información como el nombre de usuario, el nombre, el apellido y el correo que pueden considerarse irrelevantes, debido a que no se almacenan en el dataset final por cuestiones de anonimato.

En esta sección es preciso mencionar que ambos dispositivos (celular principal y adicional) usan la misma aplicación para recolectar y almacenar los datos de los sensores. No obstante, el celular ubicado en la cintura tiene varios atributos o columnas con todos sus datos nulos (correspondientes a las columnas reservadas para la banda de Microsoft, la cual no puede sincronizarse con este teléfono adicional). Además, el teléfono de la cintura no cuenta con barómetro, es decir, otra columna llena de nulos para la base de datos secundaria. Entonces, es claro que esas columnas, correspondientes a datos de la banda y barómetro, debieron eliminarse de esta base de datos.

6.2.2. Criterios de selección de variables dentro del dataset

El objetivo del dataset es brindar la mayor cantidad de atributos que puedan usarse para clasificar CS específicos. Por ende, los criterios de selección están centrados en descartar los atributos que no brindan alguna información significativa. De hecho, columnas con valores que solo contienen un solo valor durante todo el proceso (p.ej la columna donde se almacenan los datos UV) se descartaron por no aportar en futuros procesos de clasificación alrededor de comportamientos sedentarios. Aquellas variables que por el contrario si presentan cambios durante la recolección, todas se mantienen en el dataset.

6.2.3. Selección de tablas o filas de interés

La revisión sistemática muestra que las variables más usadas en los estudios relacionados son aquellas obtenidas del acelerómetro. Respecto a esto, este trabajo comprueba el cambio notorio de los datos del acelerómetro entre posturas. Además, al observar la base de datos, ciertamente aquellos datos relacionados a cuando se está manipulando el teléfono pueden ser bien diferenciadas desde un análisis superficial de las columnas. Por otra parte, los datos de los beacons son realmente interesantes, por su notable impacto en cuanto a diferenciar los comportamientos sedentarios definidos. El impacto de las demás variables, y la combinación entre ellas, pueden ser considerables en la precisión de los modelos y clasificadores que se generen. Pero tales conclusiones deben hacerse en trabajos posteriores y en esta etapa sería equivocado descartar las columnas específicas (diferentes a las que se mantienen constantes).

El uso de la información presente en el dataset depende del objetivo de las investigaciones que exploten los datos reunidos en este trabajo. Entonces, según sea el caso, puede que los investigadores usen las columnas o atributos que hayan sido más relevantes en otros estudios, como también puede suceder que experimenten con las demás variables para definir su contribución en los modelos de clasificación. Esos experimentos pueden recurrir a usar las variables en solitario o en forma conjunta con

otras y mejorar los procesos de clasificación, además de explorar las bondades de las tres ubicaciones en el cuerpo en las cuales se toma los datos (muñeca, muslo o cintura)

6.2.4. Entradas con texto libre

En los datos capturados solamente se tiene un dato que es de texto libre, a saber, la ocupación de la persona. Esto se dejó de esa manera porque las descripciones hechas en este campo entregan aclaraciones valiosas sobre la ocupación de la persona. Podría pensarse en dejar un menú de opciones con carreras profesionales, pero muchas de las personas, aunque tengan el mismo título, pueden cumplir actividades que impliquen más comportamientos sedentarios que otras. De manera que se decidió no limitar esta entrada.

6.2.5. Atributos faltantes

Durante la recolección de datos, si se presenta una situación de datos faltantes, la aplicación desarrollada estuvo en capacidad de detectarlos e informar al investigador encargado de supervisar la recolección de datos en ese momento con el fin de iniciar nuevamente la recolección.

6.2.6. Mejoras en la recolección

Después de recolectados los datos y debido a que están bien estructurados y almacenados en un base de datos, se pueden crear scripts o aplicaciones para modificarlos y ejecutar tareas de limpieza sobre los mismos. Una de las maneras en que se mejoraron los datos consistió en recortar las muestras a 7500, para cada una de las actividades recolectadas. Circunstancias causadas por la frecuencia inconstante de los dispositivos y la forma manual de iniciar la recolección.

Otra operación realizada sobre los datos fue eliminar aquellas columnas que tenían datos de los beacons, cuando deberían estar nulas. Ejemplo, si un beacon no se voltea con la anticipación necesaria seguirá emitiendo su señal y el celular captará y almacenará tales registros.

6.2.7. Análisis volumétrico de datos

En total resultaron 60 carpetas, 30 carpetas (una por cada participante) con los datos del celular principal y 30 carpetas contenedoras de los datos del celular adicional. El nombre de las carpetas es el código del participante.

Los datos fueron recolectados con una frecuencia de 50Hz, por lo tanto, 7500 líneas de datos fueron recolectadas por actividad (2 minutos 30 segundos), 187500 por

participante (62 minutos 30 segundos), 5'625.000 en total (31 horas 15 minutos). Cada participante realizó los 23 comportamientos sedentarios y las 2 actividades adicionales listadas en la Tabla 6-2.

6.2.8. Métodos de captura

Los dispositivos que hicieron parte del estudio trabajan a frecuencias diferentes. Entonces, uno de los criterios para almacenar los registros fue que se insertaran los datos según el sensor con la frecuencia de muestreo más alta y para evitar valores nulos en el resto de la fila, los datos de sensores más lentos corresponderían al último valor conocido. Además, la frecuencia de muestreo del sensor que define la inserción de datos no es constante. Así que, si se considera el uso de un temporizador para notificar y terminar la recolección, es de esperar que la cantidad de datos obtenidos siempre variaba de una actividad a otra y de un teléfono a otro. Razón por lo cual se hace necesario definir un límite de muestras por actividad. Por lo tanto, al observar las diferentes bases de datos, se concluye que es posible obtener en cada actividad 7500 muestras (2 minutos con 30 segundos). Esta cantidad de datos cumple el requerimiento de obtener como mínimo 2 minutos de datos recolectados con un adicional de 30 segundos.

6.2.9. Acceso a fuentes de datos

Las fuentes de datos son las personas que hicieron parte de la recolección en función de participantes. Por lo tanto, el acceso a las fuentes de datos siempre estuvo ligado a la disponibilidad que cada participante tenía. También son fuentes de datos los sensores y dispositivos utilizados.

6.2.10. Tipos de atributos

Una base de datos relacional está formada por múltiples tablas, con relaciones entre tablas, columnas y registro o filas. Es preciso que en la creación de una columna haya que asignar el tipo de datos que contendrá en cada registro.

En la tabla de datos personales se tienen 5 atributos numéricos (Id, edad, peso, estatura y cintura) y 9 atributos de tipo cadena (nombre, apellido, nombre de usuario, género, profesión u ocupación, correo, frecuencia en el hábito de fumar, frecuencia en el consumo de alcohol y medio de transporte). Mientras que, en la tabla contenedora de la información de los sensores, los tipos de datos asignados se definen según la documentación encontrada de Android [136] y la documentación de la Microsoft band [137].

Datos del sensor	Dispositivo	Tipo de dato
Acelerómetro ejes X, Y y Z	Banda	double
Altímetro	Banda	float
Luz ambiente	Banda	int
Presión de aire	Banda	double
Temperatura ambiental	Banda	double
Respuesta galvánica de la piel	Banda	int
Giroscopio ejes X, Y y Z	Banda	double
Frecuencia cardíaca	Banda	int
Variabilidad de frecuencia cardíaca	Banda	double
Temperatura de la piel	Banda	double
Acelerómetro ejes X, Y y Z	Teléfonos	float
Giroscopio ejes X, Y y Z	Teléfonos	float
Campo magnético ejes X, Y y Z	Teléfonos	float
Gravedad ejes X, Y y Z	Teléfonos	float
Aceleración lineal ejes X, Y y Z	Teléfonos	float
Rotación ejes X, Y y Z	Teléfonos	float
Barómetro ejes X, Y y Z	Teléfono principal	float
Potencia	Beacon	int

Tabla 6-3: Tipos de datos de los sensores

6.2.11. Rangos de valores

Los acelerómetros, en los dispositivos móviles actuales, generalmente funcionan en los rangos: $\pm 2g$, $\pm 4g$, $\pm 8g$. Los dispositivos que hicieron parte del estudio contaban con acelerómetros que trabajan en el rango $\pm 4g$. Así mismo el giroscopio trabaja en el rango de ± 35 rad/seg.

Sensores móviles	Rango de operación	Valor Mínimo	Valor Máximo
Acelerómetro	$\pm 4g$	-22,9652 m/s ²	37,3728 m/s ²
Giroscopio	$\pm 34,9066$ rad/s	-11,6037 rad/s	11,3833 rad/s
Magnético	$\pm 4915,2002$ uT	-1161,2244 μ T	684,1827 μ T
Gravedad	$\pm 9,8067$ m/s ²	-9,7327 m/s ²	12,0730 m/s ²
Aceleración Lineal	$\pm 39,2266$ 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 6-4: Rango de operación de los sensores

La documentación de la banda no especifica los rangos entre los que funcionan sus sensores. De modo que, para tener una idea de los valores, se obtienen los máximos y mínimos de los datos capturados y se presentan en la siguiente tabla.

Sensor Microsoft Band	Mínimo Encontrado	Máximo Encontrado	Posible rango de operación
Acelerómetro	-6.89 m/s ²	+7.99 m/s ²	±1 g
Giroscopio	-17.44 rad/s	17.44 rad/s	±17.44 rad/s

Tabla 6-5: Rangos aproximados de los sensores en la banda.

6.2.12. Archivos de datos

Los datos inicialmente no se almacenan en texto plano sino de forma estructurada mediante el gestor de base de datos SQLite para dispositivos móviles. De manera que los delimitadores no aparecen sino hasta que es necesario exportar la información registrada al formato txt. El delimitador usado en los archivos txt es el punto y coma. Mientras que el carácter de nueva línea es “\n”.

6.3. Análisis de atributos

La cantidad de variables y actividades es demasiado grande como para pensar en mostrar medidas estadísticas para cada variable, de forma ordenada en una tabla o en un diagrama. De manera que en esta sección se presentan las medidas estadísticas más significativas. Pero, con el dataset estará adjunta una hoja de cálculo con el análisis estadístico de las variables: mediana, promedio, desviación estándar, varianza, máximos y mínimos.

6.3.1. Resultado de un análisis superficial

Los datos del acelerómetro propio del teléfono principal cambian drásticamente entre las actividades que implican la manipulación del dispositivo y aquellas cuando el teléfono está en el bolsillo como se muestra en la Figura 34.

	A	P	Q	R
1		MOBILE_ACC_X	MOBILE_ACC_Y	MOBILE_ACC_Z
2	SENTADO-USANDOCOMPUTADOR-ESCRITORIO	3,153021473	3,28340996	7,212509791
3	SENTADO-USANDOCOMPUTADOR-CAMA	3,556270726	2,122101827	7,407765425
4	RECLINADO-USANDOCOMPUTADOR-ESCRITORIO	3,935187502	3,253402277	7,069201253
5	RECLINADO-USANDOCOMPUTADOR-CAMA	4,707536099	1,146568595	6,991428298
6	SENTADO-USANDOTELEFONO-ESCRITORIO	-0,674766574	5,860520125	7,501642739
7	SENTADO-USANDOTELEFONO-CAMA	-0,561784331	4,450396786	8,387160244
8	SENTADO-USANDOTELEFONO-SOFA	-0,33482282	4,996193815	8,211544458
9	RECLINADO-USANDOTELEFONO-ESCRITORIO	-0,65307607	6,515071787	6,815470456
10	RECLINADO-USANDOTELEFONO-CAMA	-0,003344843	6,947448828	5,976116305
11	RECLINADO-USANDOTELEFONO-SOFA	-0,396843175	7,024344845	6,355960068
12	ACOSTADO-USANDOTELEFONO-CAMA	0,370303654	9,169249854	0,285190699
13	ACOSTADO-USANDOTELEFONO-SOFA	0,621218022	8,998423018	0,038093146
14	SENTADO-VIENDOTELEVISION-SOFA	3,866156138	2,262990934	7,590211901
15	RECLINADO-VIENDOTELEVISION-SOFA	4,434679907	1,879808364	7,521214599
16	ACOSTADO-VIENDOTELEVISION-SOFA	3,777137717	-0,17281731	7,557688474
17	SENTADO-ENREPOSO-ESCRITORIO	3,42467415	3,37422077	7,502547661
18	SENTADO-ENREPOSO-SOFA	4,041352688	2,796943621	7,49181383
19	SENTADO-ENREPOSO-CAMA	4,13318842	2,285734123	7,128698022
20	RECLINADO-ENREPOSO-ESCRITORIO	4,155552816	3,544035912	7,08021095
21	RECLINADO-ENREPOSO-SOFA	4,305265801	1,881981937	7,489488628
22	RECLINADO-ENREPOSO-CAMA	4,69525573	1,081494396	7,383101499
23	ACOSTADO-ENREPOSO-SOFA	4,490591376	-0,182892473	7,642471137
24	ACOSTADO-ENREPOSO-CAMA	4,807668575	-0,02140808	7,221064136

Figura 34: Primer análisis de los datos del acelerómetro del teléfono principal

Además, desde este primer análisis se hace evidente la marcada influencia de los beacons en el reconocimiento de lugares donde se realizan los comportamientos sedentarios. Lo mismo pasa con los beacons que marcan el uso de los dispositivos. La función de estos beacons es facilitar la captura de datos que en otros estudios ya se han realizado. Por ejemplo, la información de uso de celulares, computadoras y los nuevos televisores, ya ha sido obtenida con aplicaciones que se interconectan mediante Internet [138].

En general, para obtener buenos resultados, se necesita que las muestras de un determinado atributo, estén bien delimitadas o reunidas en una zona. Además, las zonas de diferentes clases no deben tener una superposición demasiado grande entre ellas. Precisamente eso es lo que pasa con los datos de los beacons, especialmente cuando se detecta el beacon que se relaciona con la ubicación más cercana.

En otros estudios ya se había notado la influencia de las posturas en la frecuencia cardíaca. En [139] ya se evidenció que la frecuencia cardíaca y la variabilidad de frecuencia cardíaca cambia cuando una persona está sentada, de pie o caminando. En los datos recolectados se evidenció que mientras los participantes estaban

acostados, la frecuencia bajó respecto a las demás actividades realizadas como se nota en la Figura 35.

	A	M	N
1		BAND_HEART_RATE	BAND_RR
2	SENTADO-USANDOCOMPUTADOR-ESCRITORIO	70,88416	0,839875979
3	SENTADO-USANDOCOMPUTADOR-CAMA	71,46287111	0,843954513
4	RECLINADO-USANDOCOMPUTADOR-ESCRITORIO	71,49076	0,836342251
5	RECLINADO-USANDOCOMPUTADOR-CAMA	70,25631556	0,8521531
6	SENTADO-USANDOTELEFONO-ESCRITORIO	71,86541333	0,837267182
7	SENTADO-USANDOTELEFONO-CAMA	73,01396889	0,822431519
8	SENTADO-USANDOTELEFONO-SOFA	73,27247556	0,817685912
9	RECLINADO-USANDOTELEFONO-ESCRITORIO	73,16388	0,81933287
10	RECLINADO-USANDOTELEFONO-CAMA	70,30254222	0,857531194
11	RECLINADO-USANDOTELEFONO-SOFA	74,11176	0,805523607
12	ACOSTADO-USANDOTELEFONO-CAMA	68,34430667	0,886164709
13	ACOSTADO-USANDOTELEFONO-SOFA	69,88698222	0,866615056
14	SENTADO-VIENDOTELEVISION-SOFA	72,73544444	0,822529743
15	RECLINADO-VIENDOTELEVISION-SOFA	72,12475556	0,826322969
16	ACOSTADO-VIENDOTELEVISION-SOFA	68,00723111	0,88338625
17	SENTADO-ENREPOSO-ESCRITORIO	72,33861609	0,829782807
18	SENTADO-ENREPOSO-SOFA	73,11294667	0,817392049
19	SENTADO-ENREPOSO-CAMA	71,73718667	0,829491451
20	RECLINADO-ENREPOSO-ESCRITORIO	72,34961778	0,829279664
21	RECLINADO-ENREPOSO-SOFA	72,41010667	0,828773497
22	RECLINADO-ENREPOSO-CAMA	70,27017778	0,852232668
23	ACOSTADO-ENREPOSO-SOFA	69,42948	0,864626302
24	ACOSTADO-ENREPOSO-CAMA	67,85877778	0,885446312

Figura 35: Actividades y promedio de FC y VFC.

6.3.2. Datos nulos

Situaciones de datos nulos se presentaron en las columnas que almacenan la potencia de los beacons. Aquellas situaciones son causadas generalmente por la inactividad del bluetooth en el dispositivo receptor y ciertamente fueron errores humanos no percibidos en la interfaz de validación.

6.3.3. Variables con significado similar

En cuanto a datos similares, puede mencionarse los datos obtenidos del acelerómetro y los del sensor de gravedad. El acelerómetro obtiene la fuerza de la aceleración aplicado sobre el dispositivo, incluyendo la fuerza de la gravedad. De manera que, en

situaciones donde el movimiento de los dispositivos es mínimo, los datos de la aceleración recolectados siempre van a ser similares a los producidos por la gravedad. Luego, en algún momento para agilizar el procesamiento de datos obtenidos en estas condiciones podría optarse por descartar tales columnas.

6.3.4. Ortografía y el formato de los valores

El formulario está diseñado para capturar el tipo de dato correctamente. Es decir, no se permite texto en campos como el peso, edad o altura. Además, se implementan menús que facilitan el almacenamiento común de los ítems en los listados. En otras palabras, cuando un participante selecciona algún elemento del menú, el sistema guarda una cadena que será la misma siempre que se realice esa selección.

Hay un campo dentro del formulario que puede prestarse a diferencias ortográficas, a saber, el espacio para insertar la ocupación del participante. Sin embargo, de igual manera se mantuvo un formato en el cual, la primera letra es mayúscula y el resto minúscula.

6.3.5. Plausibilidad de los valores

La construcción del dataset implicó interactuar con múltiples dispositivos, los cuales trabajan a frecuencias diferentes. De modo que, el procedimiento que se emplea para almacenar los datos es usar la frecuencia de muestreo más alta entre los sensores a utilizar como llave de la conexión a la base de datos e inserción de los valores en la misma. Luego, se va a presentar de forma constante que mientras el sensor más rápido (acelerómetro) capture un evento (proceso mediante el cual se reciben datos producidos por cambios en el sensor) en un instante de tiempo, en ese instante posiblemente no exista un valor conocido de otros sensores, por ejemplo, de la variabilidad de frecuencia cardíaca que se presenta una vez cada segundo. La Figura 36 aclara la situación. Suponga que el sensor A (SA) tiene una frecuencia de 5 Hz y el sensor B una de 1 Hz. De modo que, en ocasiones se presentan registros del sensor A y no del sensor B.

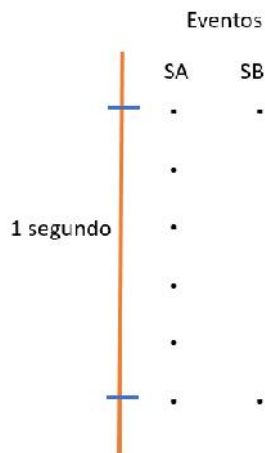


Figura 36: Ejemplo de la captura de eventos en la aplicación

La base de datos se diseñó de manera que la información de los sensores y los beacons se guarden en una tabla única. La razón de ese criterio es simplificar la exportación de los valores capturados a formatos comunes de minería y análisis. Una de las opciones fue implementar una tabla por sensor, pero al final implicaría más operaciones para relacionar los registros entre tablas. Entonces, atendiendo la situación descrita y relacionada a la frecuencia de muestreo de cada sensor, aplica un procedimiento simple. Tal procedimiento implica evitar nulos entre registros, mediante variables que mantienen el último valor conocido del sensor. Así se asegura que en cada registro haya un valor diferente de nulo para cada columna.

Es preciso e importante recalcar que, gracias a la marca temporal de cada registro, obtenida del sensor más rápido, y con el pre-conocimiento de la frecuencia de todos los sensores, pueden aplicarse operaciones software simples para generar tablas por cada sensor en caso de ser necesario. Con las aclaraciones en este párrafo es de imaginar que datos como el del sensor de frecuencia cardíaca se van a repetir cerca de 50 veces.

6.3.6. Factores no deseados

Cada dispositivo interactúa a su manera con las señales sinusoidales que producen los datos que finalmente se disponen para los desarrolladores. Es decir, que cada dispositivo detecta el ruido e implementa funciones para tratarlo como mejor se adapte a su hardware. Por tal razón, el sistema se desarrolló con el objetivo de obtener los datos tal y como son capturados en los sensores. Es por eso que no se aplican filtros u otras operaciones adicionales que además implicarían mayor procesamiento y gasto energético de los smartphone.

Un tipo de ruido que afectó de una forma más directa al sistema y por consiguiente a los datos fue la irregularidad de la frecuencia del acelerómetro, en ambos teléfonos. Se observó que la mayoría de las veces, el teléfono tomó una muestra cada 20 mili segundos. Sin embargo, existieron situaciones en que el tiempo entre muestras era mayor o menor a ese valor. Guiados por situaciones como estas, al inicio del proyecto se decidió en dejar una frecuencia mayor a la recomendada en estudios previos (20 Hz) acerca de la influencia de los acelerómetros en la clasificación de actividades de la vida diaria.

6.4. Conclusiones

En este capítulo fue realizada la descripción sobre la recolección de los datos. Para ello, fueron mencionados aspectos como los participantes de la recolección y sus características. Además, fueron descritas las posiciones en que se ubicó cada uno de los dispositivos. Luego, la atención estuvo centrada en los datos recolectados y a su descripción inicial. Entre las descripciones realizadas están: la cantidad de información obtenida, los atributos de menos relevancia, los tipos de atributos y otros. Por último, se realizó un análisis de los atributos superficialmente debido a la cantidad de información, para ello, basados en valores calculados de media, varianza, desviación estándar, máximos, mínimos y mediana; se percibe algunas tendencias en los datos presentes en las diferentes bases de datos

CAPÍTULO 7.

EVALUACIÓN DEL DATASET CONSTRUIDO

La evaluación del dataset es realizada a través de tres etapas. La primera etapa consiste en elaborar un reporte sobre los resultados obtenidos, la segunda pone a prueba la validez de los datos recolectados aplicando un algoritmo de clasificación y finalmente se hace una verificación para determinar si la documentación del dataset sigue las recomendaciones de [105].

Un aspecto importante, a partir de este capítulo, es que se deja de seguir formalmente la metodología CRISP-DM. Esto se hace porque las siguientes actividades y fases de la metodología van enfocadas a seleccionar y preparar los datos para su análisis.

7.1. Reporte de datos

El reporte pretende analizar aspectos como:

- Cantidad de nulos por atributo en la base de datos.
- Cantidad de valores que deberían ser nulos, pero no lo son. Esto ocurre generalmente con los beacons y es causado por el tiempo que requieren para desactivarse o activarse.

Además, el reporte contiene información sobre los datos, luego de someterlos a una serie de métricas de calidad tradicionales como la simple ratio, completitud, consistencia, etc. Finalmente se describen las técnicas que se usaron para corregir los problemas encontrados con el reporte.

7.1.1. Nulos por atributo.

Los valores nulos aparecieron especialmente con los datos de los beacons. Esto sucede por la potencia de radiación con la que se configuraron. Entonces, es claro que en las ocasiones donde el beacon esté demasiado lejos del celular receptor, la columna donde se almacena cada registro será nula. Sin embargo, los beacons lejanos no deberían incidir en los propósitos de clasificación de los comportamientos sedentarios. Ya que se supone que son más valiosas las señales de los beacons más cercanos.

7.1.2. Análisis de nulos para los beacons asociados a los dispositivos.

En el capítulo 6 ya se mencionó que hubo ocasiones en que por limitaciones de tiempo decide validarse la recolección, aun cuando se conoce la falta de datos, pero pueden recuperarse con los datos previamente conocidos.

La cantidad de nulos obtenidos en las columnas que contienen la información de los beacons, asociados a los dispositivos electrónicos son resumidos en la tabla. Una de las métricas usadas para verificar la calidad en un conjunto de datos recolectados es la de *simple ratio* o relación simple. Esta métrica considera el número de actividades en que se presentaron los nulos, sobre la cantidad de actividades realizadas durante todo el experimento. Por ejemplo, si la actividad “sentado viendo televisión en el sofá” no tiene valores en la columna que representa al beacon asociado a la televisión, tal situación implica un resultado no deseado. Para el caso de la tabla, el total de resultados es el número de actividades realizadas en todo el experimento y que implican el uso del dispositivo, para el caso del ejemplo, el televisor (4 actividades por persona, 240 actividades que implican usar el televisor en todo el estudio).

	resultados no deseados	resultados no deseados/total resultados	Simple ratio
Computador	5	5/180	0.027777777777
Teléfono	3	3/480	0.00625
Televisor	0	0/240	0

Tabla 6-6. Nulos de beacons en dispositivos

El análisis hasta el momento se hizo para las bases de datos obtenida del teléfono principal. Sin embargo, como las condiciones siempre fueron las mismas, el celular adicional presentó las mismas columnas nulas que el principal, en cuanto a beacons se refiere. Entonces, si se considera todo el dataset, el número de resultados no deseados es 16. Pero en todos los casos fue posible recuperar los valores perdidos mediante los registros conocidos. De modo que, en todos los casos la relación simple pasa a ser cero, es decir, se corrigen todas las columnas con datos faltantes.

7.1.3. Análisis de nulos para los beacons asociados a un lugar.

Una de las descripciones hechas sobre los beacons es que para el estudio se activó la característica de voltear para activarlos o desactivarlos. Entonces, en ocasiones accidentalmente los beacons se voltearon. De ahí se generan columnas con registros nulos.

	resultados no deseados	resultados no deseados/total resultados	Simple ratio
Escritorio	0	0/360	0
Sofá	1	1/540	0.00185185185
Cama	6	6/480	0.0125

Tabla 6-7. Nulos en beacons de los lugares

La Tabla 6-7 muestra los resultados no deseados para los archivos del celular principal, que tienen que ver con las columnas de los beacons asociados a alguno de los lugares definidos. En la tabla se evidencia que en 7 actividades hubo problemas para detectar los beacons ligados a la localización. Es decir que, para los dos celulares corresponde a 14 de 2760 resultados. Sin embargo, al igual que los nulos encontrados para los beacons de los dispositivos, los registros pueden ser recuperados mediante los valores conocidos de actividades previa y posteriormente realizadas por el mismo participante.

Finalmente, la aplicación desarrollada para la recolección de datos y el dataset con toda su información relacionada se hace pública en el siguiente repositorio: <https://github.com/WilliamPossos/sedentary-behaviors-dataset>

7.2. Experimento de clasificación

Luego de haber realizado el reporte de los datos y las correcciones a que hubo lugar, el dataset puede ser usado con los propósitos que los investigadores determinen. Entonces, en el caso de este proyecto se determinó por realizar un experimento de clasificación, aprovechando que fruto de un trabajo realizado en la línea de eSalud de la Universidad del Cauca, se tiene acceso a un programa en java que facilita la transformación de datos y la inducción de modelos de clasificación [140].

Ya que para esta etapa de la evaluación se propone un experimento, es necesario plantear una hipótesis que se enfoque en verificar la calidad del dataset.

Hipotesis: Mediante los datos presentes en el dataset puede alcanzarse una precisión elevada (>90%) en la clasificación de comportamientos sedentarios específicos.

En el experimento se compara la precisión alcanzada con diferentes técnicas de clasificación, sobre los 23 comportamientos sedentarios incluidos en el dataset. La clasificación se realiza respecto al lugar del cuerpo en la que se adquirieron los datos (la muñeca, la cintura y el bolsillo). Además, con el fin de visualizar la contribución del uso de los beacons, se analiza la precisión incluyendo o excluyendo la variable de localización. Los algoritmos de clasificación estándar no pueden ser aplicados directamente a datos en series de tiempo como los incluidos en el dataset construido [141], entonces es necesario convertirlos en ejemplos (resumen de características de

un subconjunto de registros). Para realizar esto, se utilizó una aplicación en java que divide el dataset en segmentos de 5 segundos (250 lecturas) y luego genera las características (features) que pueden procesarse con WEKA para obtener resultados más estables. Fue escogida la duración de example (DE) de 5 segundos con base en los resultados obtenidos en [140].

Debido a que las fases de preparación de datos, modelado y evaluación, toman un determinado tiempo por la cantidad de datos que deban procesar, se decide por hacer el análisis con los atributos más usados para la clasificación de comportamientos sedentarios según la revisión sistemática, es decir, el acelerómetro y giroscopio. De manera que pueda generarse un punto de referencia o una idea de la precisión que pueda alcanzarse en futuros procesos de clasificación. De estos dos atributos las siguientes características fueron generadas para cada uno de sus 3 ejes: promedio, desviación estándar y desviación absoluta promedio. También fue calculado el promedio resultante (promedio de las raíces cuadradas de la suma de los valores de cada eje al cuadrado sobre el DE) para los datos de los dos sensores. Adicionalmente, se desarrolló un pequeño algoritmo dentro de la aplicación java para obtener los beacons ordenados del más lejano al más cercano según la potencia BLE recibida por los smartphones, sumando así 6 características más, cada una de las cuales fue representando por medio de un id, uno de los seis beacons usados. Cuando un beacon no era detectado, el id fue asignado a cero.

En total, 41400 muestras fueron obtenidas, 1380 muestras por cada participante, 60 por cada CS. Cada ejemplar contiene en total 26 características más la etiqueta del CS realizado. Estos datos fueron incluidos en un mismo archivo, sobre el cual, seis técnicas de clasificación de WEKA *data mining suite* [142] fueron ejecutadas. Tales técnicas son: árboles de decisión (J48, random forest), redes neuronales (multilayer perceptron), vecinos cercanos (*nearest neighbor* (IB1)), Bayes (Naive Bayes) y Máquina de vectores de soporte (*support vector machine* (SVM)). Para cada técnica se usan las configuraciones por defecto para la clasificación y validación cruzada de diez pliegues (*ten-fold cross validation*) para su evaluación.

7.2.1. Resultados del experimento

La Tabla 6-8 presenta el porcentaje de precisión promedio en la clasificación de los 23 comportamientos sedentarios para cada una de las seis técnicas de clasificación. Tales técnicas son seleccionadas dado que son, según la literatura, bien conocidas *well-known* y han sido ampliamente probadas y verificadas para contextos similares al de este trabajo.

	Cintura		Bolsillo		Muñeca	
	a	b	a	b	a	b
Random Forest	93.6	99.1	86.2	96.4	74	93.8
Ib1	93.7	99.1	85.5	96.3	74.3	93.9
J48	89.8	98.6	80.8	94.9	63.9	90.9
Multi. Perceptron	33.2	78.6	32.3	71	28.7	67.9
SVM	27.3	53.2	21.1	49.2	11.5	42.3
Naive Bayes	19.7	10.5	27.9	16	21.6	9.7

Tabla 6-8: Porcentaje de precisión obtenida en la clasificación. a) sin incluir las seis características de los beacons. b) Incluyendo las seis características de los beacons.

Como se puede observar en la Tabla 6-8, incluyendo las características de localización para inducir el modelo de clasificación (análisis b) se mejora la precisión para la mayoría de los casos, respecto a no incluirlas (análisis a). Esto indica que las características extraídas de los beacons son importantes en la clasificación. Respecto a las técnicas de clasificación empleadas, se evidencia que Random Forest, IB1 y J48 son las que mejor precisión ofrecen respectivamente.

Random Forest consiste en un método de ensamblaje, el cual utiliza como clasificador base al algoritmo J48. Se comprueba entonces que utilizando un método de ensamblaje se obtiene una mejor clasificación de los CS, al igual que ha sucedido en algunos trabajos cuyo objetivo fue la clasificación de actividad física [143], [144]. Muy a la par de RF se encuentra el nivel de precisión del algoritmo IBk. Para IBk se indujeron modelos extra con diferentes valores de k (1,2,3,4,5,10,30,50,100) con el objetivo de encontrar el número de vecinos más cercanos óptimo para mejorar el porcentaje de la clasificación, siendo k=1 el más adecuado. De acuerdo con los resultados, la mejor estrategia para realizar la clasificación de los CS, es recolectar los datos desde la cintura, por lo que un dispositivo diseñado específicamente para ser usado en ese lugar del cuerpo sería óptimo. El alto porcentaje en la precisión de clasificación, alcanzado al utilizar los datos recolectados desde la cintura y los algoritmos RF o IB1, se obtiene por dos razones: la primera es que el smartphone ubicado en la cintura estuvo siempre fijo, mientras se recolectaban los datos de todas las actividades, incluyendo las que requerían utilizar o manipular un smartphone (Para esos casos fue utilizado el smartphone que se llevaba en el bolsillo). La otra razón es que la clasificación de las posturas sentado, inclinado y acostado, se hace de una mejor manera desde esta parte del cuerpo. Es importante notar que la clasificación realizada utilizando el smartphone puesto en el bolsillo es bastante buena. Además, al incluir las características de localización, la clasificación usando los datos tomados desde la banda se incrementa a un nivel más que aceptable.

Un porcentaje de precisión de casi 94% y 86% es logrado sin usar las características de localización en el caso de usar los datos del acelerómetro y giroscopio desde la cintura y el muslo de los participantes respectivamente. Este es un resultado inesperado, e indica que los algoritmos RF e IBk son capaces de discriminar cambios muy sutiles en la posición de las personas, tendiendo a sobre estimar el nivel de clasificación, lo que podría llevar a cometer muchos errores en la clasificación de CS en un sistema desplegado en la vida real. Debido a lo anterior, además de llevar a cabo un proceso de selección de atributos formal, se hace necesario evaluar diferentes tipos de modelos tales como el personal, híbrido y universal teniendo en cuenta por ejemplo un enfoque de validación cruzada dejando uno fuera *leave-one-user-out* para este último.

En comparación con la precisión documentada en los estudios de la revisión sistemática (donde en promedio, para las posiciones “acostado”, “sentado” y “parado” alcanzaron el 93,91%, 87,1% y 88,697% respectivamente), es evidente que las precisiones obtenida en este experimento de clasificación (incluyendo o excluyendo la información de localización y usando diferentes algoritmos) son similares e incluso mejores.

7.3. Verificación de la información acerca del dataset

Basados en las recomendaciones realizadas en [105], sobre la información que se debe suministrar acerca de un dataset. En el dataset publicado se brindan junto a este, 3 archivos; 2 en formato pdf y uno en formato excel, en los cuales se presenta la siguiente información:

descripcionDataset.pdf: Este documento presenta la información general del dataset, como su tipo de modelo, el método de recolección, los comportamientos sedentarios medidos, un resumen de los participantes, entre otros. Con lo cual, la persona que desee usar el dataset, tenga la información suficiente para entender la información que contiene y su estructura. En el Anexo D se presenta una copia de este documento.

descripcionParticipantes.pdf: En este archivo se presentan todos los datos capturados de los participantes. En el Anexo E se presenta una copia de este documento.

estadisticas.xlsx: En estas hojas de cálculo se presentan varios cálculos estadísticos (máximos, mínimos, promedio, desviación estándar, varianza y mediana), con los cuales se puede formar una imagen inicial del dataset y sus datos, además de observar la importancia de unos atributos sobre otros en la clasificación de comportamientos

sedentarios. Sin embargo, esta labor pertenece a los investigadores que utilicen el dataset y el objetivo con el q lo hagan.

En la Tabla 6-9, se tiene un checkList, en el cual se presentan las características que sugieren en [105] como información base de un dataset y en frente de cada una de ellas, una casilla en que se encuentra marcada si la característica se cumple (✓) o no (⊙) y en que archivo se encuentra este dato.

Característica	Verificación
Tipo de modelo	✓ En descripcionDataset.pdf
Método de recolección	✓ En descripcionDataset.pdf
Información del escenario <ul style="list-style-type: none"> • Número de participantes (sujetos) • Población • Rasgos • Actividades • Duración 	<ul style="list-style-type: none"> • ✓ En descripcionDataset.pdf • ✓ En descripcionDataset.pdf • ✓ En descripcionParticipantes.pdf • ✓ En descripcionDataset.pdf • ✓ En descripcionDataset.pdf
Sensores <ul style="list-style-type: none"> • Tipo • Frecuencia de muestreo • Número • Localización 	<ul style="list-style-type: none"> • ✓ En descripcionDataset.pdf • ✓ En descripcionDataset.pdf • ✓ En descripcionDataset.pdf • ✓ En descripcionDataset.pdf
Características/Presentación de <ul style="list-style-type: none"> • Unidades de los datos en bruto • Técnicas/Algoritmos • Ventanas (duración de example) 	<ul style="list-style-type: none"> • ✓ En descripcionDataset.pdf • ✓ En descripcionDataset.pdf • ✓ En descripcionDataset.pdf
Resultados <ul style="list-style-type: none"> • Exactitud • Consistencia 	<ul style="list-style-type: none"> • ✓ En descripcionDataset.pdf • ✓ En estadisticas.xlsx

Tabla 6-9. Verificación de información relacionada con el dataset

Como resultado de este chequeo es válido decir que la información suministrada junto con el dataset construido cumple con todas las especificaciones dadas por investigadores con experiencia en el campo del reconocimiento de actividades a partir de dispositivos celulares. De hecho, la metodología aplicada para realizar y describir el dataset hace que las limitaciones en trabajos futuros sean menores, según los mismos autores.

7.4. Conclusiones

En este capítulo se presentó la calidad del dataset construido. La primera etapa consistió en un reporte sobre los resultados obtenidos, la cantidad de nulos encontrados y los aspectos corregidos. La segunda etapa puso a prueba la validez de los datos recolectados mediante la aplicación de algoritmos de clasificación estándar. Finalmente se hizo una verificación para determinar si la documentación del dataset siguió las recomendaciones de [105]. Como resultado, todas las perspectivas evidenciaron la calidad tanto en la metodología de construcción como de la estructura y los registros obtenidos en el dataset obtenido en este trabajo.

CAPÍTULO 8.

CONCLUSIONES Y TRABAJOS FUTUROS

8.1. Conclusiones

El uso de CRISP-DM permitió una amplia comprensión en el desarrollo del proyecto, dando paso a que cada una de las actividades que se propusieron fueran claras y ordenadas en el momento de su desarrollo en general. Es recomendable para la solución de proyectos de minería de datos y construcción de DataSets.

Las recomendaciones propuestas por Lochart y Weiss, para la construcción de DataSets que están relacionadas con actividad física y comportamientos sedentarios, son una base importante en esta clase de proyectos, ya que; gracias a ellos, se tienen en cuenta factores tales como: las características de los participantes, el método de recolección, los sensores, la documentación, entre otros, que permiten que la calidad del DataSet mejore.

El sistema diseñado y los dispositivos seleccionados, como la Microsoft band 2, los teléfonos inteligentes (LG G3 y Huawei P7) y los estimote beacons, lograron cumplir con los objetivos planteados, ya que permitieron recolectar una gran cantidad de atributos y un funcionamiento óptimo.

La aplicación desarrollada con base a “Clean architecture”, utilizando herramientas como EventBus, greenDao y atendiendo los criterios de diseño de material design, fueron útiles para lograr tener la modularidad deseada con un funcionamiento correcto durante toda la fase de recolección. A partir de ello se concluye que éstas tienen la capacidad de ser un buen soporte para el desarrollo de aplicaciones móviles.

Las treinta personas presentaron diversidad en sus características, basados en la información recolectada en el formulario, como lo fue su edad, género, peso, altura, cintura, ocupación, etcétera.

En el escenario de recolección diseñado, se obtuvieron datos de 23 comportamientos sedentarios, más dos actividades adicionales: estar de pie y caminando. Tales datos representan 5'625.000 muestras que equivalen a 31:15 horas de recolección.

La evaluación del DataSet realizada desde las tres perspectivas, logró demostrar su calidad en la metodología de construcción, la estructura en que se encuentran organizados los datos, la precisión de las muestras obtenidas y los documentos generados.

8.2. Aportes

Con el desarrollo del presente trabajo de grado fueron generados los siguientes aportes:

- Un dataset disponible al público para la clasificación de comportamientos sedentarios. Disponible en:

<https://drive.google.com/drive/folders/0B8oVIqp033LdMXBFM2VVRzRLckE?usp=sharing>

- Una aplicación móvil Android para la recolección de datos centrada en comportamientos sedentarios y siendo el eje de un sistema conformado por un smartphone, un wearable y beacons. Pero gracias a su construcción de forma modular, permite ser adecuada para enfocarla en otro tipo de recolección y con cualquier otro tipo de dispositivo. Disponible en:

<https://github.com/WilliamPossos/sedentary-behaviors-dataset>

- Artículo titulado: “Open Dataset for the Automatic Recognition of Sedentary Behaviors Recognition”, el cual fue aceptado como “full paper”, para ser presentado en el evento “pHealth 2017 14th International Conference on Wearable, micro & Nano Technologies for Personalized Health”, a llevarse a cabo del 14 al 16 de mayo del 2017 en Eindhoven, Holanda.
- Artículo titulado: “Datasets for Sedentary Behaviors Recognition: Systematic Review” el cual se encuentra en preparación para ser sometido a una revista internacional.

8.3. Trabajos futuros

Desarrollar las fases faltantes de la metodología CRISP-DM a partir de los desarrollos adelantados en el presente trabajo. De manera que se llegue a un descubrimiento efectivo de conocimiento sobre los datos y se encuentren los mejores algoritmos para el reconocimiento de comportamientos sedentarios en escenarios similares al presentado. Un aporte importante dentro de este trabajo futuro sería:

- Aplicar técnicas de clasificación sobre todas las variables en el dataset construido para determinar su influencia en la clasificación de comportamientos sedentarios.

Relacionar las características de los participantes (edad, peso, altura, diámetro de la cintura, frecuencia en el consumo de cigarrillos y demás información capturada de la población) con los datos recolectados para generar caracterización de la población medida.

Incluir más comportamientos sedentarios o actividades físicas dentro del dataset. Así como también, incluir diferentes wearables y tecnologías que brinden información relevante en el seguimiento de actividades de la vida diaria.

Usar el dataset construido para generar modelos que permitan implementar un sistema para el reconocimiento de comportamientos sedentarios en tiempo real y en la vida cotidiana. De manera que se les brinde a los profesionales de la salud una herramienta para generar estrategias o intervenciones mejor adaptadas según el perfil sedentario de los pacientes.

REFERENCIAS

- [1] C. D. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030," *PLoS Med.*, vol. 3, no. 11, pp. 2011–2030, Nov. 2006.
- [2] 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. 3066–72, Nov. 2005.
- [3] S. G. Wannamethee *et al.*, "Metabolic Syndrome vs Framingham Risk Score for Prediction of Coronary Heart Disease, Stroke, and Type 2 Diabetes Mellitus," *Arch. Intern. Med.*, vol. 165, no. 22, p. 2644, Dec. 2005.
- [4] G. Alberti, P. Zimmet, J. Shaw, and S. Grundy, "The IDF consensus worldwide definition of the metabolic syndrome." International Diabetes Federation, p. 24, 2006.
- [5] P. E. M. Soca, "El síndrome metabólico: un alto riesgo para individuos sedentarios," *ACIMED*, vol. 20, no. 2, pp. 0–0, 2009.
- [6] D. W. Dunstan *et al.*, "Television viewing time and mortality: the Australian Diabetes, Obesity and Lifestyle Study (AusDiab).," *Circulation*, vol. 121, no. 3, pp. 384–91, Jan. 2010.
- [7] A. Grøntved *et al.*, "Television Viewing and Risk of Type 2 Diabetes, Cardiovascular Disease, and All-Cause Mortality," *JAMA*, vol. 305, no. 23, p. 2448, Jun. 2011.
- [8] 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. Sport. Exerc*, vol. 41, no. 5, pp. 998–1005, 2009.
- [9] 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, 1996–2011," *Am. J. Prev. Med.*, vol. 41, no. 2, pp. 207–215, 2011.
- [10] 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. 150–9, Feb. 2011.
- [11] Sedentary Behaviour Research Network, "Letter to the editor: standardized use of the terms 'sedentary' and 'sedentary behaviours'.,," *Appl. Physiol. Nutr. Metab. = Physiol. Appl. Nutr. metabolisme*, vol. 37, no. 3, pp. 540–542, Jun. 2012.
- [12] 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. 1384–9, Jun. 2007.
- [13] 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.
- [14] I. A. D. Office of Air and Radiation, Office of Atmospheric and Indoor Air Programs, "Report to Congress on Indoor Air Quality, Volume II: Assessment and Control of Indoor Air Pollution," Washington D.C., 1989.

- [15] A. Loveday, L. B. Sherar, J. P. Sanders, P. W. Sanderson, and D. W. Esliger, "Technologies That Assess the Location of Physical Activity and Sedentary Behavior: A Systematic Review," *J. Med. Internet Res.*, vol. 17, no. 8, p. e192, 2015.
- [16] 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.
- [17] 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," *Am. J. Prev. Med.*, vol. 41, no. 5, pp. 508–515, 2011.
- [18] D. E. Rosenberg, F. C. Bull, A. L. Marshall, J. F. Sallis, and A. E. Bauman, "Assessment of Sedentary Behavior With the International Physical Activity Questionnaire," *J. Phys. Act. Health*, vol. 5, no. 1, pp. 30–44, 2008.
- [19] M. Lustrek, B. Cvetkovic, and S. Kozina, "Energy expenditure estimation with wearable accelerometers," in *2012 IEEE International Symposium on Circuits and Systems*, 2012, pp. 5–8.
- [20] S. Kozey-Keadle, A. Libertine, K. Lyden, J. Staudenmayer, and P. S. Freedson, "Validation of Wearable Monitors for Assessing Sedentary Behavior," *Med. Sci. Sport. Exerc.*, vol. 43, no. 8, pp. 1561–1567, 2011.
- [21] N. F. Butte, U. Ekelund, and K. R. Westerterp, "Assessing Physical Activity Using Wearable Monitors: Measures of Physical Activity," *Med. Sci. Sport. Exerc.*, vol. 44, no. 1S, pp. 5–12, 2012.
- [22] S. E. O'Connell, P. L. Griffiths, and S. A. Clemes, "Seasonal variation in physical activity, sedentary behaviour and sleep in a sample of UK adults," <http://dx.doi.org/10.3109/03014460.2013.827737>, vol. 41, no. 1, pp. 1–8, 2013.
- [23] H. Gjoreski, B. Kaluža, M. Gams, R. Mili, and M. Luštrek, "Context-based ensemble method for human energy expenditure estimation," *Appl. Soft Comput.*, vol. 37, pp. 960–970, 2015.
- [24] R. Alvarez, L. Urbano, C. Vélez, J. Vidarte, and H. Sierra, "Prevalencia de síndrome metabólico en una población laboralmente activa del sur occidente Colombiano," *Datos sin publicar, está siendo revisado por el Profr. Hernán*.
- [25] J. Vidarte Claros, C. Vélez Álvarez, and R. Álvarez, "Actividad física en población diagnosticada con síndrome metabólico del Departamento del Cauca, Niveles de sedentarismo en una población laboralmente activa," *Aceptado para publicación. Rev. salud pública México*.
- [26] M. L. 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. Fac. Med.* 61, vol. 1, pp. 3–8, 2013.
- [27] J. A. Vidarte Claros, C. Vélez Álvarez, A. Angel, and J. Ignacio, "Niveles de sedentarismo en población entre 18 y 60 años: Sincelejo (Colombia)," *Rev. Salud Uninorte*, vol. 31, no. 1, pp. 70–77, 2015.
- [28] 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.
- [29] 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.
- [30] 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.
- [31] 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.
- [32] L. Gao, A. K. Bourke, and J. Nelson, "Evaluation of accelerometer based multi-sensor versus single-sensor activity recognition systems," 2014.
- [33] 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.
- [34] 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.
- [35] 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.
- [36] 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.
- [37] 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.
- [38] 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.
- [39] 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.
- [40] K. P. Dowd *et al.*, "The measurement of sedentary patterns and behaviors using the activPAL™ Professional physical activity monitor," *Physiol. Meas.*, vol. 33, no. 11, pp. 1887–1899, Nov. 2012.
- [41] 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.
- [42] 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.
- [43] 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.
- [44] S. G. Trost, "Comparison of Accelerometer Cut Points for Predicting Activity Intensity in Youth," *Comparison of Accelerometer Cut Points for Predicting Activity Intensity in Youth. Med*

- .*Sci.SportsExerc*, vol. 43, no. 7, pp. 1360–1368, 2011.
- [45] J. Parkka, 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.
- [46] 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.
- [47] 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.
- [48] 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.
- [49] 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.
- [50] 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.
- [51] A. Godfrey, K. M. Culhane, and G. M. Lyons, “Comparison of the performance of the activPAL™ Professional physical activity logger to a discrete accelerometer-based activity monitor,” 2007.
- [52] 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.
- [53] H. Gjoreski, B. Kaluža, M. Gams, R. Mili, and M. Luštrek, “Context-based ensemble method for human energy expenditure estimation,” *Appl. Soft Comput.*, vol. 37, pp. 960–970, 2015.
- [54] 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.
- [55] 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.
- [56] 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.
- [57] 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.
- [58] 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.
- [59] 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.
- [60] 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.
- [61] 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.
- [62] 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.
- [63] 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.
- [64] 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.
- [65] 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.
- [66] K. P. Dowd *et al.*, “Criterion and Concurrent Validity of the activPAL™ Professional Physical Activity Monitor in Adolescent Females,” *PLoS One*, vol. 7, no. 10, p. e47633, Oct. 2012.
- [67] 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.
- [68] 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.
- [69] 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.
- [70] 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.
- [71] 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.
- [72] 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.
- [73] 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.
- [74] 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.
- [75] 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.
- [76] 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.
- [77] 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.
- [78] 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.
- [79] 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.
- [80] 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.
- [81] Psicólogos Argentinos, “Ciclo vital - Infancia - Adolescencia - Juventud - Adulterez - Vejez.” .
- [82] J. M. Moine, “Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo,” Facultad de Informática, 2013.
- [83] 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.
- [84] Gregory Piatetsky-Shapiro, “KDnuggets Data Mining, Analytics, Big Data, and Data Science.” [Online]. Available: <http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html>. [Accessed: 01-Aug-2016].
- [85] P. Chapman *et al.*, “CRISP-DM 1.0 Step-by-step data mining guide,” 2000.
- [86] A. Alvarez-Alvarez, J. M. Alonso, and G. Trivino, “Human activity recognition in indoor environments by means of fusing information extracted from intensity of WiFi signal and accelerations,” *Inf. Sci. (Ny)*, vol. 233, pp. 162–182, 2013.
- [87] “IDC: Smartphone OS Market Share 2016, 2015.” [Online]. Available: <http://www.idc.com/promo/smartphone-market-share/os;jsessionid=DDA511E125A3697FC15669F399C9EA0B>. [Accessed: 19-Jan-2017].
- [88] “Microsoft Band | Official Site.” [Online]. Available: <https://www.microsoft.com/microsoft-band/en-us>. [Accessed: 18-Jan-2017].

- [89] Microsoft, “Microsoft Band - Developers.” [Online]. Available: <https://developer.microsoftband.com/>. [Accessed: 18-Jan-2017].
- [90] I. Fitbit, “Fitbit Docs.” [Online]. Available: <https://dev.fitbit.com/co>. [Accessed: 18-Jan-2017].
- [91] L. Garmin, “Programmer’s Guide | Connect IQ | Garmin Developers.” [Online]. Available: <https://developer.garmin.com/connect-iq/programmers-guide/>. [Accessed: 18-Jan-2017].
- [92] I. Apple, “watchOS - Apple Developer.” [Online]. Available: <https://developer.apple.com/watchos/>. [Accessed: 18-Jan-2017].
- [93] “Build Misfit.” [Online]. Available: https://build.misfit.com/docs/cloudapi/get_started. [Accessed: 18-Jan-2017].
- [94] “UP for Developers: Home.” [Online]. Available: <https://jawbone.com/up/developer>. [Accessed: 18-Jan-2017].
- [95] J. A. Carballar and J. A. C. Falcón, *Wi-Fi: lo que se necesita conocer*. RC Libros, 2010.
- [96] Bluetooth SIG Inc, “The Story Behind Bluetooth Technology | Bluetooth Technology Website.” [Online]. Available: <https://www.bluetooth.com/what-is-bluetooth-technology/bluetooth>. [Accessed: 02-Aug-2016].
- [97] I. Dragojevic, “Qué es un beacon y cuáles son sus beneficios - FayerWayer,” 2014. [Online]. Available: <https://www.fayerwayer.com/2014/10/que-es-un-beacon/>. [Accessed: 02-Aug-2016].
- [98] L. M. Ni, Y. Liu, Y. C. Lau, and A. P. Patil, “Landmarc: Indoor location sensing using active RFID Texas,” *USA March*, 2003.
- [99] NearFieldCommunication.org, “Near Field Communication: What is Near Field Communication?” [Online]. Available: <http://nearfieldcommunication.org/>. [Accessed: 02-Aug-2016].
- [100] datateca.unad.edu.co, “Lección 13: Fundamentos de redes de telefonía celular.” [Online]. Available: http://datateca.unad.edu.co/contenidos/2150509/Contenido_en_linea/leccin_13_fundamentos_de_redes_de_telefona_celular.html. [Accessed: 02-Aug-2016].
- [101] L. Huawei Technologies Co., “Huawei - HUAWEI P7 - Teléfonos Móviles - Especificaciones.” [Online]. Available: <http://consumer.huawei.com/mx/mobile-phones/tech-specs/ascend-p7-mx.htm>. [Accessed: 18-Jan-2017].
- [102] L. Electronics, “LG G3 - El mejor smartphone | Mobile World Congress 2015.” [Online]. Available: <http://www.lg.com/es/telefonos-moviles/lg-G3-D855>. [Accessed: 18-Jan-2017].
- [103] Estimote, “Beacon Tech Overview - Estimote Developer.” [Online]. Available: <http://developer.estimote.com/>. [Accessed: 05-Jan-2017].
- [104] M. ZHANG and A. SAWCHUK, “A Daily Activity Dataset for Ubiquitous Activity Recognition Using Wearable Sensors,” in *Proc. UbiComp*, vol. 12.
- [105] J. W. Lockhart and G. M. Weiss, “Limitations with activity recognition methodology & data sets,” in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, 2014, pp. 747–756.
- [106] S. F. M. Chastin *et al.*, “Development of a Consensus Taxonomy of Sedentary Behaviors (SIT): Report of Delphi Round 1,” *PLoS One*, vol. 8, no. 12, p. e82313,

Dec. 2013.

- [107] D. M. Camargo Lemos, L. C. Orozco, and L. C. Orozco, "Factores asociados a la disponibilidad y uso de medios electrónicos en niños desde pre-escolar hasta 4° grado," *Biomédica*, vol. 33, no. 2, pp. 175–85, 2013.
- [108] C. E. S. Castaño, *Modelo integral para el profesional en ingeniería*. Universidad del Cauca, 2005.
- [109] B. Bryon, "Business Modeling with UML: The light at the end of the Tunnel," *Proc. Ration. Edge*, 2001.
- [110] "StarUML - La Plataforma Open Source UML / MDA." [Online]. Available: <http://staruml.sourceforge.net/v1/about.php>. [Accessed: 25-Feb-2017].
- [111] P. Krüchten, "The 4+ 1 view model of software architecture," *IEEE Softw.*, pp. 42–50, 1995.
- [112] "Foro trimestre II (Modelo Arquitectural)." [Online]. Available: <http://proy-pnfi.forosactivos.net/t3-foro-trimestre-ii-modelo-arquitectural>. [Accessed: 19-Jan-2017].
- [113] "UML basics: The sequence diagram." [Online]. Available: <http://www.ibm.com/developerworks/rational/library/3101.html>. [Accessed: 19-Jan-2017].
- [114] "The Clean Architecture | 8th Light." [Online]. Available: <https://8thlight.com/blog/uncle-bob/2012/08/13/the-clean-architecture.html>. [Accessed: 19-Jan-2017].
- [115] B. Meyer, *Object-oriented software construction*, vol. 2. Prentice hall New York, 1988.
- [116] I. Jacobson, *Object-oriented software engineering: a use case driven approach*. Pearson Education India, 1993.
- [117] D. L. Parnas, "On the Criteria To Be Used in Decomposing Systems into Modules."
- [118] E. Caffrey, "Aplicando Clean Architecture en Android." [Online]. Available: <https://erikcaffrey.github.io/ANDROID-clean-architecture/#scroll>. [Accessed: 20-Jan-2017].
- [119] "Features – Genymotion Android Emulator." [Online]. Available: <https://www.genymotion.com/features/>. [Accessed: 19-Jan-2017].
- [120] "DB Browser for SQLite." [Online]. Available: <http://sqlitebrowser.org/>. [Accessed: 19-Jan-2017].
- [121] "What is development environment? - Definition from WhatIs.com." [Online]. Available: <http://searchsoftwarequality.techtarget.com/definition/development-environment>. [Accessed: 19-Jan-2017].
- [122] "What is Development Environment? - Definition from Techopedia." [Online]. Available: <https://www.techopedia.com/definition/16376/development-environment>. [Accessed: 19-Jan-2017].
- [123] "Funciones de Android Studio | Android Studio." [Online]. Available: <https://developer.android.com/studio/features.html?hl=es-419>. [Accessed: 19-Jan-2017].
- [124] "Features | Bitbucket." [Online]. Available: <https://bitbucket.org/product/features>. [Accessed: 19-Jan-2017].
- [125] J. I. Hong and J. A. Landay, "An Infrastructure Approach to Context-Aware

- Computing,” 2001.
- [126] “What is a Software Library? - Definition from Techopedia.” [Online]. Available: <https://www.techopedia.com/definition/3828/software-library>. [Accessed: 19-Jan-2017].
 - [127] “Butterknife. Sorprendentemente sencillo, absolutamente imprescindible.” [Online]. Available: <https://geekytheory.com/butterknife-sorprendentemente-sencillo-absolutamente-imprescindible/>. [Accessed: 19-Jan-2017].
 - [128] “Biblioteca de compatibilidad | Android Developers.” [Online]. Available: <https://developer.android.com/topic/libraries/support-library/index.html>. [Accessed: 19-Jan-2017].
 - [129] “Funciones de la biblioteca de compatibilidad | Android Developers.” [Online]. Available: <https://developer.android.com/topic/libraries/support-library/features.html?hl=es>. [Accessed: 19-Jan-2017].
 - [130] “Adición de la barra de app | Android Developers.” [Online]. Available: <https://developer.android.com/training/appbar/index.html?hl=es>. [Accessed: 19-Jan-2017].
 - [131] “greenDAO: Android ORM for your SQLite database - Open Source by greenrobot.” [Online]. Available: <http://greenrobot.org/greendao/>. [Accessed: 25-Feb-2017].
 - [132] “Android Best Practices to Improve Battery Efficiency.”
 - [133] “EventBus: Events for Android - Open Source by greenrobot.” [Online]. Available: <http://greenrobot.org/eventbus/>. [Accessed: 19-Jan-2017].
 - [134] “Lección 2 ¿Que es una Aplicación Móvil?” [Online]. Available: http://datateca.unad.edu.co/contenidos/233016/EXE_SAM/leccin_2_que_es_un_a_aplicacin_mvil.html. [Accessed: 19-Jan-2017].
 - [135] “Introduction - Material design - Material design guidelines.” [Online]. Available: <https://material.io/guidelines/#>. [Accessed: 19-Jan-2017].
 - [136] “Sensors Overview | Android Developers.” [Online]. Available: https://developer.android.com/guide/topics/sensors/sensors_overview.html. [Accessed: 21-Jan-2017].
 - [137] Microsoft, “Microsoft Band, SDK Documentation.” [Online]. Available: https://developer.microsoftband.com/content/docs/microsoft_band_sdk.pdf. [Accessed: 21-Jan-2017].
 - [138] A. Sano *et al.*, “Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones,” in *Wearable and Implantable Body Sensor Networks (BSN), 2015 IEEE 12th International Conference on*, 2015, pp. 1–6.
 - [139] F.-T. Sun, C. Kuo, H.-T. Cheng, S. Buthpitiya, P. Collins, and M. Griss, “Activity-aware mental stress detection using physiological sensors,” in *International Conference on Mobile Computing, Applications, and Services*, 2010, pp. 211–230.
 - [140] 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. 804–6, 2016.
 - [141] G. M. Weiss and H. Hirsh, “Learning to predict rare events in event sequences,” *Proceedings of the Fourth International Conference on Knowledge Discovery*

- and Data Mining*. AAAI Press, pp. 359–363, 1998.
- [142] I. H. (Ian H. . Witten, E. Frank, and M. A. (Mark A. Hall, *Data mining : practical machine learning tools and techniques*. Morgan Kaufmann, 2011.
 - [143] A. Reiss, G. Hendeby, and D. Stricker, “A novel confidence-based multiclass boosting algorithm for mobile physical activity monitoring,” *Pers. Ubiquitous Comput.*, vol. 19, no. 1, pp. 105–121, Jan. 2015.
 - [144] Z. Zengtao Feng, L. Lingfei Mo, and M. Meng Li, “A Random Forest-based ensemble method for activity recognition,” in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, vol. 2015, pp. 5074–5077.