MODELO PREDICTIVO PARA EL RECONOCIMIENTO DE ACTIVIDADES HUMANAS BASADO EN TÉCNICAS DE MACHINE LEARNING Y DE SELECCIÓN DE CARACTERÍSTICAS

JANNS ALVARO PATIÑO SAUCEDO

MAESTRIA EN INGENIERIA CON ENFASIS EN SISTEMAS

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN Y ELECTRÓNICA

UNIVERSIDAD DE LA COSTA

BARRANQUILLA - COLOMBIA

2020

MODELO PREDICTIVO PARA EL RECONOCIMIENTO DE ACTIVIDADES HUMANAS BASADO EN TÉCNICAS DE MACHINE LEARNING Y DE SELECCIÓN DE CARACTERÍSTICAS

JANNS ALVARO PATIÑO SAUCEDO

Trabajo de Investigación presentado para optar por el título de Magister en Ingeniería

Directores

PhD. Emiro de La Hoz Franco

MSc. Jorge Diaz Martínez

MAESTRIA EN INGENIERIA CON ENFASIS EN SISTEMAS

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN Y ELECTRÓNICA

UNIVERSIDAD DE LA COSTA

BARRANQUILLA - COLOMBIA

2019

El Maestrante **Janns Alvaro Patiño Saucedo** y sus tutores de proyecto de fin de master **PhD. Emiro De la Hoz Franco y MSc. Jorge Diaz Martínez** garantizamos, al firmar este documento, que el trabajo ha sido realizado por el maestrante bajo la dirección de los tutores y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

Barranquilla, ¿??? - Agosto de 2020

PhD. Emiro De la Hoz Franco

Tutor

Firma

MSc. Jorge Diaz Martínez

Cotutor

Firma

Janns Alvaro Patiño Saucedo

Maestrante

Firma

Agradecimientos

A Dios por darme la fuerza para llevar a cabo este proyecto y por poner ángeles a mi alrededor que me han dado el impulso para lograr este tan anhelado objetivo, a mi familia que han realizado un esfuerzo enorme para que yo esté hoy en este punto y en especial a mis tutores PhD. Emiro De la Hoz Franco y MSc. Jorge Diaz Martínez les voy a estar muy agradecido siempre por su invaluable aporte a mi formación profesional.

A Dios, a mis padres, mi esposa, mis hijos y familia en general, esto es por ustedes.

Resumen

Los ambientes asistidos para la vida - AAL por sus siglas en inglés (*Ambient Assisted Living*), se enfocan en generar productos y servicios innovadores en aras de proporcionar asistencia y atención médica a personas de avanzada edad que padezcan enfermedades neurodegenerativas o alguna discapacidad. Esta área de investigación se encarga del desarrollo de sistemas para el reconocimiento de actividad - ARS (*Activity Recognition Systems*) los cuales están basados en el reconocimiento de actividades humanas - HAR (*Human Activity Recognition*), específicamente en actividades de la vida diaria - ADL (*Activities of Daily Living*) en ambientes interiores (*indoor*). Estos sistemas permiten identificar el tipo de actividad que realizan las personas, ofreciendo una posibilidad de asistencia efectiva que les permita llevar a cabo actividades cotidianas con total normalidad.

El desempeño de los ARS en el proceso de HAR, debe ser evaluado a través del planteamiento de escenarios experimentales con conjuntos de datos dispuestos por la comunidad científica en repositorios en linea, este trabajo plantea una variedad de combinaciones de técnicas de machine learning con técnicas de selección de características, obteniendo como resultado un modelo funcional para el HAR, que combina la técnica de clasificación árboles para el modelamiento logístico - LMT por sus siglas en inglés (Logistic Model Trees) y la técnica de selección de características One R.

Palabras clave: *Reconocimiento de actividades humanas (HAR), Machine learning, Clasificación, Selección de características.*

Abstract

Ambient assisted living (AAL), focus on generating innovative products and services in order to aid and medical attention to elderly people who suffer from neurodegenerative diseases or a disability. This research area is responsible for the development of activity recognition systems (ARS) which are based on Human Activity Recognition (HAR), specifically in activities of daily life (ADL) in indoor environments. These systems make it possible to identify the type of activity that people carry out, offering a possibility of effective assistance that allows them to carry out daily activities with total normality.

The performance of the ARS in the HAR process must be evaluated through the approach of experimental scenarios with data sets available by the scientific community in online repositories, this work proposes a variety of combinations of machine learning algorithms with feature selection algorithms, obtaining as a result a functional model for the HAR, which combines the classification algorithm Logistic model trees (LMT) and the feature selection algorithm One R.

Keywords: *Human Activity Recognition (HAR), Machine Learning, Classification, Feature Selection.*

TABLA DE CONTENIDO

Pág

LISTA DE TABLAS

Pág

LISTA DE FIGURAS

Pág

PRESENTACIÓN

El área de investigación de ambientes asistidos para la vida - AAL por sus siglas en inglés (*Ambient Assisted Living*), se enfoca en generar tecnología, productos y servicios innovadores en aras de proporcionar asistencia, atención médica y rehabilitación a personas de avanzada edad, con el propósito de incrementar el tiempo en que estas personas puedan vivir de manera autónoma, ya sea que padezcan o no enfermedades neurodegenerativas o alguna discapacidad. Esta importante área, se encarga del desarrollo de sistemas para el reconocimiento de actividad - ARS (*Activity Recognition Systems*) los cuales son una valiosa herramienta al momento de identificar el tipo de actividad que realizan las personas mayores, para así brindarles una asistencia efectiva que les permita ejecutar las actividades cotidianas con total normalidad.

Los ARS se basan en el reconocimiento de actividades humanas - HAR (*Human Activity Recognition*), que para efectos de este trabajo de grado, está enfocado específicamente en actividades de la vida diaria - ADL (*Activities of Daily Living*) en ambientes interiores (*indoor*), dado que existen otros tipos de actividades. Para evaluar el desempeño de los ARS en el reconocimiento de actividades de la vida diaria, se hace necesario el uso de conjuntos de datos de prueba, en escenarios experimentales acordes con esta temática, los cuales han sido dispuestos por la comunidad científica para el reconocimiento de actividades humanas HAR.

Actualmente, gran parte de la población mundial de avanzada edad padece enfermedades neurodegenerativas, este tipo de enfermedades afectan ostensiblemente a las personas que las padecen, dado que inciden en la pérdida del equilibrio, reducción de la movilidad, deficiencias en

el habla, alteración de la respiración y otras alteraciones propias de la función cardiovascular, lo cual directamente repercute en una disminución de las capacidades cognitivas de los individuos y en gran medida dificultan la realización de actividades de la vida diaria ADL (U.S. National Library of Medicine, 2019). El Alzheimer, la Demencia, la esclerosis lateral amiotrófica - ALS por sus siglas en inglés (*Amyotrophic Lateral Sclerosis*) y el Parkinson, son algunos de los tipos de enfermedades neurodegenerativas más frecuentes.

Según la organización mundial de la salud (World Health Organization, 2019), se estima que en el mundo alrededor de 50 millones de personas padecen demencia y entre el 60% y 70% de los casos son particularmente pacientes con Alzheimer, la cual es la forma más común de demencia. La demencia es una de las principales causas de discapacidad y dependencia entre las personas mayores en todo el mundo impactando física, psicológica, social y económicamente no solo a quien presenta la enfermedad, sino también a las personas en su entorno: sus cuidadores, familias y la sociedad en general, de acuerdo a lo indicado en (World Health Organization, 2019). En Colombia, entre los años 2009 al 2015, se atendieron un total de 252.577 personas que padecen esta enfermedad. Según la evidencia científica, el factor de riesgo más importante es la avanzada edad de algunos focos poblacionales (Ministerio de Salud y Protección Social, 2017).

Estas enfermedades dificultan la ejecución de actividades tan triviales como lavar los platos, preparar la comida o tomar un baño, generando en quienes las padecen, dependencia a terceros para poder llevar a cabo las diferentes acciones y en algunos casos causan un aislamiento en las personas con estos padecimientos, afectando considerablemente su calidad de vida.

Los AAL ofrecen a las personas que padecen de enfermedades neurodegenerativas una alternativa de solución, procurando mejorar significativamente su calidad de vida, gracias al

desarrollo de ARS. Dichos sistemas, permiten identificar el tipo de actividades que las personas realizan, con el fin de recrear los mecanismos de ayuda que favorezcan la ejecución de acciones cotidianas, a quienes padecen de este tipo de enfermedades, y así poder llevar a cabo una vida independiente, el mayor tiempo posible.

Sin embargo, antes de implementar estos sistemas se hace necesario evaluar su desempeño en el proceso de HAR. para efectos de optimización del proceso de clasificación de las actividades en ambientes indoor. En este proyecto se construyó un modelo funcional para el HAR, que combina la técnica de clasificación árboles para el modelamiento logístico - LMT por sus siglas en inglés (*Logistic Model Trees*) y la técnica de selección de características *One R*, identificando a partir de esta última, las 33 características que mayormente inciden en la mejora de las tasas de acierto del modelo. Las métricas empleadas para determinar el nivel de calidad del modelo fueron el *Recall* y la *Precision*, ambas con un 95,90%. Este documento se ha organizado en cinco capítulos con sus respectivas secciones, al final se presentan un listado de las referencias consultadas.

En el capítulo 1, denominado “Fundamentos relativos al reconocimiento de actividades humanas”, se recopila la base documental que fundamenta la presente investigación, por ello en primera instancia, en la sección 1.1 se aborda una introducción sobre el HAR (ubicándolo dentro del contexto los AAL), se define el tipo de actividades que fueron objeto de estudio en esta investigación y se describen los ARS. En segunda instancia, en la sección 1.2 se mencionan las diferentes colecciones de datos (que en lo sucesivo denominaremos *dataset*) desarrolladas y promovidas por la comunidad científica para el HAR, seguidamente se identifica y se hace una

descripción del *dataset* que fue escogido para las diferentes experimentaciones, en esta investigación.

En el capítulo 2, denominado "Proceso de construcción de modelos predictivos para el HAR", en primera instancia, se abordan las bases conceptuales y metodologías usadas en procesos de minería de datos, en la sección 2.1, luego se detalla el tema de preprocesamiento del *dataset*, en la sección 2.2, en el cual se describen las técnicas usadas para agregarle calidad a los datos. Seguidamente, en la sección 2.3, se define el proceso de clasificación, se mencionan las diferentes técnicas valoradas (de acuerdo con la categoría a la que pertenecen) y se describen las técnicas con las que se obtuvieron los mejores resultados en el proceso de evaluación. En la sección 2.4 se fundamenta el proceso de selección de características y se describen las técnicas con las que se obtuvieron los mejores resultados. El capítulo finaliza, en la sección 2.5, con la descripción del proceso de evaluación de modelos, donde se detallan cada una de las métricas de calidad que fueron utilizadas para medir la efectividad del proceso.

En el capítulo 3, denominado "Proceso experimental", en primera instancia se describe la propuesta en la sección 3.1, luego se detalla el preprocesamiento de los diferentes conjuntos de datos en la sección 3.2. Posteriormente, en la sección 3.3 se define el proceso de construcción del modelo funcional. El capítulo finaliza describiendo en la sección 3.4 de manera sucinta, los diferentes escenarios de experimentación efectuados en esta investigación.

En el capítulo 4, denominado "Escenarios de experimentación", se exponen detalladamente los resultados obtenidos en los diferentes escenarios experimentales que fueron planteados en esta investigación. Inicialmente se presenta un análisis comparativo de las técnicas de clasificación que arrojaron los mejores resultados, en cuanto a métricas de desempeño. En una segunda instancia,

se combinan las mejores técnicas de clasificación con las técnicas de selección de características. El capítulo finaliza con una evaluación exhaustiva para las combinaciones de técnicas de clasificación y selección que arrojaron los mejores resultados en cuanto a métricas de calidad.

En el capítulo 5, denominado “Conclusiones y recomendaciones”, se exponen las conclusiones a las cuales se ha llegado producto del análisis de los resultados de los experimentos, que se desarrollaron en esta investigación. El capítulo finaliza con las recomendaciones y a su vez se plantean algunos trabajos futuros, para darle continuidad a esta investigación.

1. FUNDAMENTOS RELATIVOS AL RECONOCIMIENTO DE ACTIVIDADES HUMANAS

Actualmente, gran porcentaje de la población mundial de avanzada edad padece enfermedades neurodegenerativas que afectan no solo la memoria, el pensamiento y el comportamiento, sino que además afectan la movilidad, impidiendo la realización de algunas actividades cotidianas (World Health Organization, 2019). Estas enfermedades deterioran significativamente la calidad de vida de quien la padece dado que al no poder realizar actividades del diario vivir, se ven obligados a depender de terceros para tratar de llevar una vida con normalidad y en algunos casos sufren de aislamiento social. El área de investigación denominada ambientes asistidos para la vida – AAL, por sus siglas en inglés (*Ambient Assisted Living*), ofrece a los adultos mayores diversas soluciones para que puedan ejecutar estas acciones diarias y vivir de manera autónoma el mayor tiempo posible.

Esta importante área se encarga de investigar técnicas innovadoras basadas en Tecnologías de la Información y la Comunicación – TIC, que proporcionen asistencia, atención médica y rehabilitación a personas mayores, con el fin de mejorar su calidad de vida (R. Li, Lu, & McDonald-Maier, 2015). Para ello, los AAL proporcionan un ecosistema de sensores, computadoras, redes inalámbricas y aplicaciones de software, que permiten monitorear la atención médica y su objetivo principal es facilitar la vida y generar un mayor nivel de independencia en las personas mayores (Memon, Wagner, Pedersen, Aysha Beevi, & Hansen, 2014). Dentro de los productos y servicios que ofrecen los AAL, se encuentran el desarrollo de sistemas de reconocimientos de actividades – ARS, por sus siglas en inglés (*Activity Recognition Systems*).

Con tales sistemas, se procura efectuar procesos de reconocimiento de actividades humanas - HAR (*Human Activity Recognition*), la cual es una de las funcionalidades más importantes a implementar en los AAL.

1.1 RECONOCIMIENTO DE ACTIVIDADES HUMANAS - HAR

El HAR tiene como objetivo identificar las acciones llevadas a cabo por una persona a través de un conjunto de observaciones de sí mismo y del entorno en que se desenvuelve (Anguita, Ghio, Oneto, Parra, & Reyes-Ortiz, 2013). Esta es un área de investigación aplicada a la visión por computadora (Aggarwal & Ryoo, 2011), la video vigilancia implementada en bancos o aeropuertos, el análisis de la técnica deportiva, los sistemas que permiten interactuar con videojuegos a través de los gestos, las tácticas militares (Kwon, Kim, & Lee, 2017), además de los ambientes de vida asistida para el cuidado de la salud de personas de avanzada edad o con enfermedades mentales. Este amplio campo de aplicación, hace del HAR un tema de actualidad y gran relevancia.

El HAR trabaja en el reconocimiento de patrones de actividad humana a partir de diferentes tipos de datos, los cuales son recolectados mediante diferentes tipos de dispositivos que contienen variedad de sensores, por ejemplo: 1) dispositivos usables o vestibles (*wearables*) que están integrados por acelerómetros, giroscopios, GPS y para la recolección de datos del ritmo cardiaco, entre otros, 2) sensores ambientales que recolectan datos numéricos o categóricos directamente de los sensores y de las cámaras, que registran datos de imagen o video, en ese orden de ideas, el reconocimiento de las actividades humanas se ha abordado de dos maneras diferentes en cuanto a la fuente que da origen a los datos, la primera pude ser a través de sensores portátiles (wearable) los cuales están conectados al usuario. La segunda fuente son sensores externos, los cuales se fijan

predeterminadamente en objetos ubicados un área de interés, con los cuales las personas van a interactuar (Lara & Labrador, 2013).

Con respecto a las actividades humanas, en (Aggarwal & Ryoo, 2011) se han categorizado o clasificado en diferentes niveles de acuerdo con su complejidad: los gestos, acciones, interacciones y actividades grupales. En donde se consideran los gestos como movimientos elementales de una parte del cuerpo de la persona tales como estirar un brazo o levantar una pierna, Las acciones como actividades que pueden estar compuestas por múltiples gestos organizados en un espacio de tiempo realizadas por una persona, tales como caminar o saltar. Las interacciones como actividades humanas que involucran dos o más personas y/u objetos, por ejemplo dos personas peleando o una persona lavando los platos y las actividades grupales como aquellas que involucran múltiples personas y/u objetos, tales como una marcha de grupo o una lucha de dos grupos.

Aunque el HAR es muy amplio, este trabajo de investigación se enmarca en el reconocimiento de actividades de la vida diaria - ADL (*Activities of Daily Living*), las cuales fueron definidas en (Reed & Sanderson, 1999), como el conjunto de actividades que una persona realiza independientemente, para su cuidado personal, desplazamiento y comunicación, tales como la movilidad personal, la alimentación, el aseo y descanso, entre otras.

Si bien es cierto que el desarrollo de ARS basados en HAR para el reconocimiento de actividades de la vida diaria, constituye una mejora significativa en la calidad de vida de las personas que padecen enfermedades neurodegenerativas, el desempeño de estos sistemas debe ser medido previamente, a través de la evaluación de diversos conjuntos de datos de prueba en escenarios experimentales, para lograr este fin, la comunidad científica ha desarrollado y promovido una variedad de colecciones de datos disponibles en la web, que contienen información

referente a actividades de la vida diaria, realizadas tanto en ambientes interiores (*indoor*), como en exteriores (*outdoor*), dichos conjuntos de datos son también conocidos con el nombre de *dataset*.

1.2 DATASET

El término *dataset* es un anglicismo que se ha adoptado en la lengua española para referirse a una colección de datos. Estos conjuntos de datos usualmente están estructurados de manera tabular, donde cada columna corresponde a un atributo o variable y cada fila corresponde a una instancia del conjunto de datos. Tales atributos son también conocidos como características y pueden ser de diversos tipos: 1) cuantitativos, que poseen un valor numérico; 2) ordinales, que tienen una relación de orden y 3) nominales, que son etiquetas para clasificar los datos en categorías (PRADENA, 2013).

En (De-La-Hoz-Franco, Ariza-Colpas, Quero, & Espinilla, 2018), se compararon siete de los más referenciados *dataset* en la literatura científica, en el ámbito del reconocimiento de ADL. Los conjuntos de datos fueron capturados a partir de diversos dispositivos, tales como: giroscopios, acelerómetros, sensores para detección de movimiento, balizas (*beacons*), sensores de temperatura y sensores binarios de contacto, entre otros. Los datos capturados a partir de dichos sensores son recolectados a partir de infraestructuras de comunicaciones basadas en redes inalámbricas de sensores WSN - *Wireless Sensor Network* y Wifi, entre otras tecnologías. Algunos de estos *dataset* registran eventos correspondientes a tres tipos de ocupación: única (*single activity*), intercalada (interleaved activity) o de ocupación múltiple (multi-occupancy).

Las ocupación *single activity*, indica que solo un individuo interactúa con el ambiente, la ocupación *interleaved activity*, indica que varios individuos interactúan con el ambiente, pero no lo hacen de forma simulatánea y multi-occupancy, indica que varios individuos pueden interactuar en el ambiente en el mismo instante de tiempo.

Los conjuntos de datos más relevantes son: 1) el *dataset* Van Kasteren (Van Kasteren, Englebienne, & Kröse, 2010), que es una colección de valores binarios a partir de la implementación de una red de sensores inalámbricos -WSN, recolectado en un recinto ocupado por dos hombres; 2) los *dataset* Kyoto (Cook, Crandall, Thomas, & Krishnan, 2013), Aruba (Cook, 2012) y Multiresident (Singla, Cook, & Schmitter-Edgecombe, 2010), todos ellos hacen parte del proyecto CASAS (Cook et al., 2013) llevado a cabo por WSU - *Washington State University*, la cual desplegó una variedad de sensores ambientales en un apartamento, que constó de tres dormitorios, un baño, una cocina y una sala de estar.

Para este estudio en particular, se ha tomado la decisión de evaluar el dataset Aruba CASAS, dado que es un dataset bastante completo cuyos archivos *raw* (en bruto) se encuentran disponibles en línea en el sitio oficial del proyecto. Si bien se ha evidenciado que las métricas de evaluación en el estudio (Shahi, Woodford, & Lin, 2017) son de 100% en cuanto a acuraccy, en este estudio se obtuvo el que hasta ahora es el segundo mejor resultado presentando mejora en cuanto a los tiempos computacionales producto de la evaluación de otras técnicas de clasificación y de la reducción de la dimensionalidad de los datos al aplicar diversas técnicas de selección de características.

El *dataset* de ocupación única y múltiple, llamado Aruba CASAS (Cook, 2012), del proyecto de casas inteligentes de *Washington State University* - WSU, fue recolectado a partir de

diferentes fuentes de datos, en la casa de un adulto voluntario. La residente de la casa fue una mujer que recibió visitas de sus hijos y nietos regularmente, durante el periodo comprendido entre el 2010-11-04 y el 2011-06-11. Dos fuentes de datos dieron origen a la información, la primera fuente es binaria y está compuesta por sensores de movimiento y de contacto, la segunda fuente está conformada por sensores de temperatura.

La fuente binaria consta de 35 sensores de los cuales 31 son de movimiento, identificados por la letra M. Estos sensores están instalados en el piso y detectan la presión ejercida por el individuo al pisar el suelo, representado los estados de activación y desactivación (ON/OFF). Los cuatro (4) sensores restantes son de contacto, instalados en las puertas e identificados por la letra D, este tipo de sensores detectan los estados de apertura y cierre de las puertas (OPEN/CLOSE). Por otra parte, la segunda fuente está conformada por 5 sensores de temperatura, ubicados en diferentes lugares de la casa, e identificados por la letra T, este tipo de sensores detectan la temperatura del ambiente en valores continuos representados en grados *Celsius*.

La información contenida en este *dataset* está conformada por los eventos registrados, producto de las interacciones del individuo con cada uno de los sensores anteriormente mencionados. Por cada evento se registra la fecha y hora, tanto de inicio como de finalización, por cada actividad que realiza el individuo. En total se etiquetaron once actividades, pero en este estudio solo se tuvieron en cuenta nueve (9), debido a que las otras dos actividades tienen un número muy bajo de muestras. Para efectos de la evaluación se consideraron las siguientes actividades: preparación de comidas (*Meal_Preparation*), descansar (*Relax*), comer (*Eating*), trabajar (*Work*), dormir (*Sleeping*), ir de la cama al baño (*Bed_to_Toilet*), entrar a casa (*Enter_Home*), salir de casa (*Leave_Home*) y limpieza (*Housekeeping*).

2. PROCESO DE CONSTRUCCIÓN DE MODELOS PREDICTIVOS PARA EL HAR

El término minería de datos ha sido ampliamente usado en la literatura científica en diversos ámbitos de aplicación, a continuación, se mencionan algunas definiciones de diversos autores, entendiendo que no existe una definición única de este proceso de extracción de conocimiento. En (Fayyad, Piatetsky-Shapiro, & Smyth, 1996) se definió a la minería de datos como un paso particular en el proceso de descubrimiento de conocimiento en bases de datos, el cual consiste en la aplicación de algoritmos específicos para extraer patrones o modelos desde los datos.

Luego, en el año 2003 se definió como una forma de descubrimiento de conocimiento esencial, para resolver problemas en dominios que involucran grandes volúmenes de datos. Los conjuntos de datos individuales pueden recopilarse y estudiarse colectivamente para fines distintos de aquellos para los que fueron creados originalmente. También se puede obtener nuevo conocimiento en el proceso, al tiempo que se elimina el costo de la recopilación de datos adicionales (Mitra & Acharya, 2003).

Por otra parte, en (Witten, Frank, & Hall, 2011) se define la minería de datos como el proceso de descubrir patrones en los datos. Tal descubrimiento debe ser automático o semiautomático. Los patrones descubiertos, deben ser lo suficientemente relevantes para obtener un provecho por lo general económico.

En el año 1996 la comunidad científica aceptó el primer modelo de explotación de información, conocido como descubrimiento de conocimiento en bases de datos (KDD -

Knowledge Discovery in Databases), definido en (Fayyad et al., 1996), el cual establece a la minería de datos como uno de los pasos principales que se encarga de la extracción de patrones a partir de los datos.

Gracias al crecimiento que tuvo el área de la minería de datos a partir del año 2000, surgieron nuevas metodologías aceptadas y validadas ampliamente por la comunidad científica para la implementación de este tipo de procesos (Moine, Haedo, & Gordillo, 2011). En la figura 1 se presenta un estudio comparativo de las metodologías más utilizadas según la comunidad KDnuggets (Data Mining Community's Top Resource), en el año 2014.

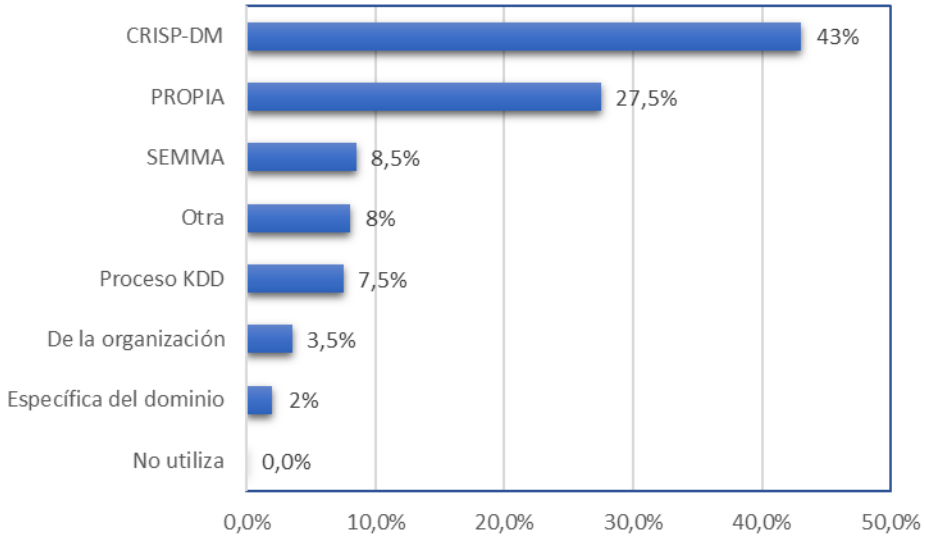

Figura 1. Metodologías utilizadas en procesos basados en minería de datos

Fuente: (KDnuggets, 2014)

A continuación, se detallan las metodologías más usadas en el proceso de minería de datos, además, se describen cada uno de los pasos de la metodología escogida para el desarrollo de esta investigación y finalmente, se conceptualizan las técnicas de preprocesamiento y clasificación usadas para la construcción del modelo.

2.1 METODOLOGIAS UTILIZADAS EN PROCESOS DE MINERIA DE DATOS

En los últimos años, el área de la minería de datos ha experimentado un importante crecimiento, y como consecuencia se han propuesto diversas metodologías ampliamente avaladas por la comunidad científica. A continuación, se conceptualizan algunas de las más utilizadas: 1) proceso estándar de la industria cruzada para la minería de datos - CRISP-DM por sus siglas en inglés (*Cross Industry Standard Process for Data Mining*), 2) muestreo, exploración, modificación, modelado y evaluación – SEMMA por sus siglas (*Sample, Explore, Modify, Model and Assess*) y 3) descubrimiento de conocimiento en bases de datos – KDD (*Knowledge Discovery in Databases*).

2.1.1. Metodología CRISP-DM

El proceso estándar de la industria cruzada para la minería de datos es actualmente, es la metodología más usada para el desarrollo de proyectos de minería de datos y fue creada por un grupo de empresas (SPSS, NCR y Daimer Chrysler) en el año 2000, según (Pete et al., 2000).
Esta metodología estructura el ciclo de vida de un proyecto de minería de datos en seis (6) fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado y Evaluación e Implantación, ver la figura 2.

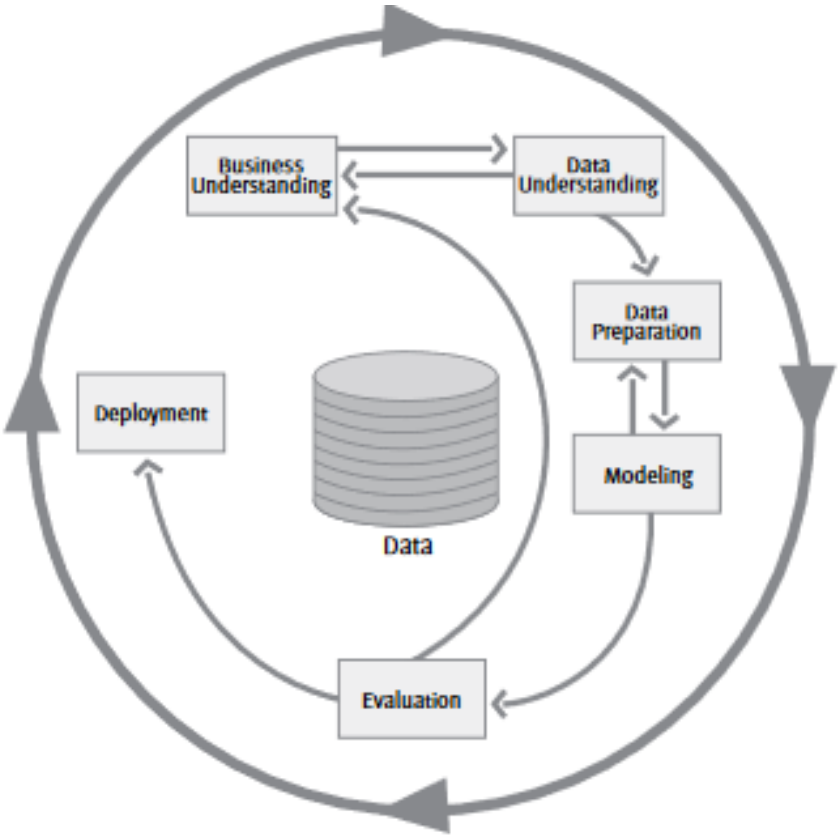

Figura 2. Fases de CRISP – DM

Fuente: (Pete et al., 2000)

A continuación, se detallan cada una de las fases de esta metodología:

- **Comprensión del negocio**: la fase de compresión del negocio *(Business understanding)* hace referencia al entendimiento de los objetivos y requerimientos del proyecto.
- **Comprensión de los datos:** la fase de comprensión de los datos (*Data understanding*) se refiere a la obtención y exploración del conjunto de datos, también a la identificación de los problemas de calidad de los datos.

- **Preparación de los datos:** en la fase de preparación de los datos (*Data preparation*) se realizan tareas de selección de atributos, limpieza y transformación de datos, estas tareas se pueden realizar de manera repetitiva y en cualquier orden.
- **Modelamiento:** en la fase de modelamiento (*Modeling*) se seleccionan y aplican varias técnicas de minería de datos, algunas de ellas tienen requisitos en cuanto a la forma de los datos, por lo que se hace necesario volver a la fase de preparación de los datos.
- **Evaluación:** en la fase de evaluación (*Evaluation*) se debe probar la calidad del modelo construido antes de su implementación, es decir, se realiza una evaluación a fondo del modelo para asegurar de que los resultados coincidan con los objetivos del negocio.
- **Despliegue:** la fase de despliegue (*Deployment*), hace referencia a la implementación del modelo construido y evaluado. Aun cuando el propósito del modelo sea aumentar el conocimiento de los datos, este se le debe presentar al cliente de manera organizada para que pueda ser utilizado. Esta fase puede incluir tareas tan sencillas como la generación de un informe, o tan complejas como la implementación de un proceso de minería de datos de forma repetida o continua, dentro de las empresas. CRISP – DM establece tareas y actividades para cada una de las fases, la secuencia de ellas no es necesariamente rígida (Pete et al., 2000).

2.1.2. Metodología SEMMA

La metodología SEMMA es la segunda más usada, para desarrollo de proyectos de minería de datos, fue creada por el SAS Institute el cual la define como el proceso de selección, exploración y modelado de grandes volúmenes de datos para el descubrimiento de patrones de negocio desconocidos (Milley, Seabolt, & Williams, 1998). Su nombre es un acrónimo que corresponde a

las cinco fases del proceso: muestreo (*Sample),* exploración *(Explore),* modificación (*Modify*), modelado (*Model) y* Evaluación (*Assess).*

- **Muestreo:** la fase de muestreo (*Sample),* hace referencia a la recolección de una muestra representativa de los datos, la cual debe ser lo suficientemente grande para que contenga información significativa y lo suficientemente pequeña para procesarlas de manera rápida.
- **Exploración:** la fase de exploración *(Explore),* se refiere a la búsqueda y detección de relaciones anticipadas, tendencias imprevistas y anomalías.
- **Modificación:** la fase de modificación (*Modify*), hace referencia a la selección y transformación de variables con el objetivo de preparar los datos para el proceso de selección del modelo.
- **Modelado:** en la fase de modelado (*Model),* se lleva a cabo la construcción del modelo a través de la aplicación de técnicas de minería de datos para problemas de clasificación, agrupamiento (*clustering*) y asociación.
- **Evaluación:** en la fase de evaluación (*Assess)* se determina la calidad del modelo, evaluando la utilidad y confiabilidad de los patrones a través de diferentes métricas del proceso de minería de datos (Milley et al., 1998).

2.1.3. Metodología KDD

Esta metodología consiste en extraer conocimiento a partir de la identificación de patrones de comportamientos de los datos. En 1996 se definió como: *"proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y comprensibles a partir de los datos"*.

Este proceso consta de diversos pasos para la generación de conocimiento, tales como: selección, preprocesamiento, transformación, minería de datos y evaluación e implantación, según (Fayyad et al., 1996), para una mayor descripción analizar la figura 3.

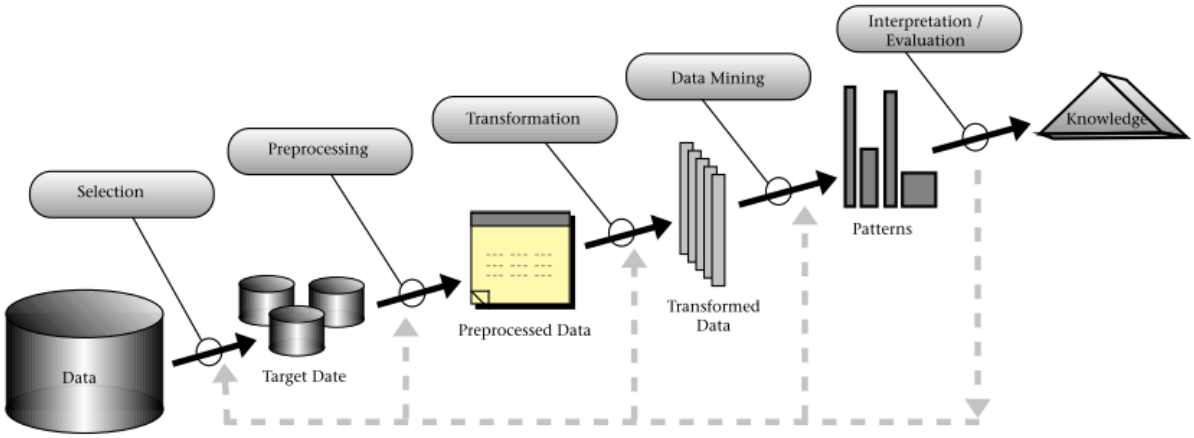

Figura 3. Pasos que componen el proceso KDD

Fuente: (Fayyad et al., 1996).

- **Selección:** en el paso de selección *(Selection)*, se desarrolla una comprensión y entendimiento del dominio o contexto de la aplicación y conocimientos previos, además, se identifica el objetivo del proceso de KDD desde el punto de vista del cliente. Seguidamente se selecciona el conjunto de datos objetivo o subconjunto de variables o muestra de datos, sobre el que se va a realizar el proceso.
- **Preprocesamiento:** en el preprocesamiento (*Preprocesing*), se realiza limpieza y preprocesamiento de los datos a través de operaciones básicas como la eliminación de ruido. Además, se llevan a cabo estrategias para manejar los campos de datos faltantes, también se realiza reducción de los datos a través de la búsqueda de características útiles que representen los datos en función del objetivo de la tarea.

- **Transformación:** en el paso de transformación (*Transformation*) o reducción de la dimensionalidad, se pueden disminuir el número de variables que se van a tener en cuenta en el proceso.
- **Minería de datos:** en este paso (*Data Mining*), primeramente se selecciona un algoritmo de minería de datos (clasificación, clustering, asociación, etc.) alineado a los objetivos del proceso de KDD definidos en el paso de selección, posteriormente se realiza una búsqueda de patrones de interés y se construye el modelo.
- **Evaluación e implantación:** en este último paso (*Interpretation / Evaluation*), se interpretan los patrones extraídos, se evalúa el modelo construido a través de diferentes métricas de calidad, y es posible que se vuelva a cualquiera de los pasos anteriores de forma iterativa para ajustes del modelo. Además, la implementación del conocimiento puede ser a través de documentación y reportes a los interesados o incorporándolo en sistemas para acciones posteriores.

Esta investigación se ha desarrollado bajo la metodología KDD. A continuación, se detallan cada uno de sus pasos. Es decir: de selección de datos, preprocesamiento y proceso de minería de datos. En cuanto al paso de selección de datos, se debe precisar que se utilizó un *dataset* colectado con diferentes infraestructuras tecnológicas, las cuales constituyen diferentes fuentes de datos. A este *dataset* se accedió a través de un repositorio dispuesto públicamente en línea en el proyecto de casas inteligentes llevado a cabo por *Washington State University* - WSU, el cual dispone de varios conjuntos de datos en el ámbito de reconocimiento de actividades de la vida diaria (HAR).

En el siguiente paso, denominado preprocesamiento, se afinó el *dataset* a través de las estrategias de reducción y transformación de los datos, se aplicaron técnicas de selección de características tanto al *dataset* de entrenamiento como al de prueba, con el objetivo de reducir la dimensionalidad de los datos y entregar un conjunto de datos de calidad para la fase de construcción del modelo.

En el paso denominado minería de datos, se lleva a cabo el proceso de construcción del modelo que según (Chen et al., 2012) puede ser de dos enfoques: 1) Enfoque orientado a objetos - DDA (*Data-Driven Approaches*) o 2) Enfoque orientado al conocimiento- KDA (*knowledge-Driven Approaches*). El enfoque DDA, está basado en técnicas de aprendizaje automático (*Machine learning*), las cuales requieren un conjunto de datos preexistentes sobre los comportamientos del usuario. Generalmente, se realiza un proceso de entrenamiento (*training*) para construir un modelo de actividad, seguido de un proceso de prueba (*testing*) para evaluar la generalización del modelo en la clasificación de actividades (C. Li, Lin, Yang, & Ding, 2014).

En cuanto al enfoque KDA, un modelo de actividad se construye a través de la incorporación de un rico conocimiento previo obtenido del dominio de la aplicación, a través de técnicas de ingeniería y gestión del conocimiento (Chen & Nugent, 2009).

La construcción del modelo de minería de datos se basó en el enfoque DDA, del cual hacen parte técnicas de análisis de umbral (*Threshold analysis*), métodos de regresión (*Regression methods*), métodos de aprendizaje automático (*Machine learning methods*), entre otros. Luego de haber analizado exhaustivamente la literatura científica se evidenció que los métodos de aprendizaje automático son los más utilizados para el proceso de clasificación. Algunos de estos métodos son: árboles de decisión (*Decision trees*), redes neuronales artificiales (*Artificial Neural*

Networks), modelos de Markov (*Markov Models*) y clasificadores bayesianos (*Bayesian Classifiers*), entre otros.

En el paso de evaluación e implementación, se procedió a evaluar el modelo construido, con el objetivo de medir su calidad y confiabilidad. Este paso se llevó a cabo a través de la obtención de la matriz de confusión y de diferentes métricas de calidad, las cuales serán detalladas al final de este capítulo.

En este estudio se profundizó en algunos de los pasos del KDD, tales como: preprocesamiento, minería de datos y evaluación e implementación. A continuación, se detalla el preprocesamiento específicamente la implementación de técnicas de selección de características. También se describen algunas las categorías de técnicas de clasificación utilizadas para la construcción del modelo, tales como: árboles de decisiones, métodos de regresión, redes neuronales artificiales y clasificadores bayesianos. Finalmente se describen las métricas de calidad utilizadas en el proceso de evaluación del modelo.

2.2 PREPROCESAMIENTO

El preprocesamiento es la preparación, limpieza y afinamiento de la información que conforma el *dataset*, con lo cual se busca incrementar la calidad del conjunto de datos según (Pyle, 1999). Normalmente los datos sin procesar (*Raw* - en bruto) presentan inconsistencias, ruido o están incompletos, a través del preprocesamiento se pueden obtener datos de mejor calidad e incluso reducir el tamaño del *dataset*, con el fin de lograr mejores resultados en referencia a la extracción de conocimiento a través de técnicas de minería de datos como las de clasificación, entre otras (Zhang, Zhang, & Yang, 2003). Este procedimiento se puede llevar a cabo aplicando

cualquiera de las siguientes estrategias: limpieza, reducción, integración y transformación de datos (Herrera & Cano, 2006).

- **Limpieza de datos:** es un proceso que se lleva a cabo para aumentar la calidad de los datos a través de la eliminación de datos inconsistentes o erróneos. Además, este proceso resuelve problemas de ruido y valores perdidos (Kim, Won and Choi, Byoung-Ju and Hong, Eui and Kim, Soo-Kyung and Lee, 2003).
- **Reducción de datos:** consiste en la reducción del volumen de datos, manteniendo los más relevantes, para su posterior uso a través de técnicas como la discretización, el muestreo (Liu & Motoda, 2013) o la selección de características (Liu & Motoda, 2012).
- **Integración de datos:** consiste en la combinación de información proveniente de diferentes fuentes de datos. Es decir, la unión de dos o más tablas que presentan información (registros) hetereogenea sobre los mismos objetos (Detours, Dumont, Bersini, & Maenhaut, 2003).
- **Transformación de datos:** consiste en modificaciones sintácticas sobre los datos, sin que se ejecute un cambio en el significado de estos (Lin, 2002), por ejemplo la normalización de los datos.

La clasificación, predicción, el agrupamiento (*clustering*) y el análisis de correlación son tareas mediante las cuales se aborda la fase de construcción de un modelo en un proceso de minería de datos. El desarrollo de investigación se basó en el proceso de clasificación, a continuación se detallan cada una de las tareas efectuadas en este proceso.

2.3 CLASIFICACIÓN

La clasificación es uno de los procesos que hace parte de la minería de datos. Gracias al conocimiento de registros históricos etiquetados de un tema en particular, pretende predecir el comportamiento de un conjunto de datos con los mismos atributos. Estas técnicas pretenden clasificar elementos de datos en una de varias clases predefinidas. Es decir, en alguno de los valores posibles que puede tomar el atributo de clase (Weiss & Kulikowski, 1991). Según (Mitra & Acharya, 2003), las técnicas de clasificación se pueden dividir en las siguientes categorías: árboles de decisión, clasificadores bayesianos, técnicas de regresión y redes neuronales artificiales (Mitra & Acharya, 2003):

- **Árboles de decisión:** son estructuras organizadas en forma jerárquica, a modo de árbol que permiten obtener de forma visual las reglas de decisión bajo las cuales operan las variables y parámetros, a partir de datos históricos almacenados. Sirven de apoyo en la toma de decisiones, con el objetivo de seleccionar la mejor opción basada en un punto de vista probabilístico (Aluja, 2001).
- **Clasificadores bayesianos:** calculan las probabilidades de hipótesis basadas en el teorema de Bayes, estas técnicas funcionan bajo el supuesto de que el mecanismo generador está representado por un modelo estadístico, estos modelos representan un conjunto de variables y sus dependencias probabilísticas. Es decir, que pueden calcular la distribución de probabilidad de cualquier subconjunto de variables (Mitra & Acharya, 2003).
- **Regresión:** son un conjunto de técnicas empleadas para formar relaciones entre los datos, la regresión lineal es de las más usadas para tal fin, a pesar de su rapidez y eficacia es insuficiente en espacios multidimensionales. Es decir, donde se relacionan más de dos variables. La

naturaleza de los datos muestra comportamientos no lineales en mayor medida, por lo que se hace necesario la implementación de técnicas de regresión no lineal para obtener resultados que se ajusten mejor a la realidad (García, 2016).

- **Redes neuronales:** son algoritmos del área de la inteligencia artificial, que tratan de emular el comportamiento del cerebro humano, en cuanto a aprender de datos históricos y aplicar lo aprendido a la solución de nuevos problemas, a través, de la formación conceptual. Se trata de un sistema de interconexión de neuronas que colaboran entre sí para producir un estímulo de salida (Aluja, 2001).

Es importante resaltar que en este estudio se evaluaron 31 técnicas de clasificación, de las subcategorías: algoritmos de árboles de decisión (*Decision trees*), algoritmos basados en reglas (*Rules*), algunas funciones (*functions*) tales cómo los algoritmos de regresión logística y redes neuronales artificiales (*Artificial Neural Networks*), algoritmos multiclasificadores (*Meta*) y algoritmos perezosos (*Lazy*) .

De la subcategoría de algoritmos de árboles de decisión se evaluaron las técnicas: Logistic Model Trees - LMT (Landwehr, Hall, & Frank, 2003), J48 (C4.5 decision tree) (Quinlan, 1994), Reduced-Error Pruning Tree - REPTree (Frank & Witten, 1998), RandomForest (Breiman, 2001), Random Tree (Frank & Witten, 1998), y DecisionStump (Marks Hall, 1994).

De la subcategoría de algoritmos basados en reglas se evaluaron las técnicas: JRip (Cohen, 1995), Partial Decision Trees – PART (Frank & Witten, 1998), Decision Table (Ron Kohavi, 1995), ZeroR Stacking(Wolpert, 1992) y OneR(Holte, 1993).

De la subcategoría de funciones se evaluaron la técnica de regresión Logistic (Cessie & Houwelingen, 1992) y la técnica de redes neuronales artificiales MultilayerPerceptron (Van Der

Malsburg, 1986). Por otra parte, se evaluaron de la subcategoría de algoritmos multiclasificadores (metaclasificadores) las técnicas: Random Committee (Frank & Witten, 1998),LogitBoost (Friedman, Hastie, & Tibshirani, 2000), Classification Via Regression (Frank, Wang, Inglis, Holmes, & Witten, 1998), MultiClass Classifier (Read, Puurula, & Bifet, 2015), MultiScheme (Eibe, Holmes, & Witten, 2007), Bagging (Breiman, 1996), AdaBoostM1(Freund & Schapire, 1996), AttributeSelectedClassifier (Frank, Wang, Inglis, Holmes, & Witten, 1998), Vote(Kittler, Hatef, Duin, & Matas, 1998), CVParameterSelection (R Kohavi, 1995), RandomSubSpace (T.K. Ho, 1998), Stacking (Wolpert, 1992) y Filtered Classifier (Eibe et al., 2007).

Finalmente, de la subcategoría de algoritmos perezosos se evaluaron las técnicas: IB1,IB2, IB3 Instance-based Learning Algorithms (Aha, Kibler, & Albert, 1991), KStar(Cleary & Trigg, 1995) y LWL(Frank, Hall, & Pfahringer, 2003).

Los mejores resultados a nivel de métricas de calidad, se obtuvieron en la subcategoría de algoritmos de árboles de decisión y algoritmos basados en reglas, específicamente las técnicas LMT, JRip y J48, las cuales son detalladas a continuación.

2.3.1. Logistic model trees – LMT

Es una técnica usada para problemas de clasificación que combina dos de los algoritmos más populares para este tipo de tareas, la regresión logística y los árboles de decisión. Este método consiste en una estructura de árbol de decisión estándar, con funciones de regresión logística en las hojas, al igual que en los árboles de decisión ordinarios, una prueba en uno de los atributos está asociada con cada nodo interno del árbol. Para cada atributo nominal con k valores, el nodo tiene

k nodos secundarios y cada una de las instancias se clasifican en una de las k ramas, de acuerdo con el valor que tome el atributo. Por otro lado, para los atributos numéricos, el nodo tiene dos nodos secundarios y la prueba consiste en la comparación del valor del atributo con un umbral, entonces, una instancia se clasificará en la rama izquierda si el valor para ese atributo es menor que el umbral. Por lo contrario, se clasificará en la rama derecha. Entonces, un modelo logístico de árbol consta de una estructura de árbol compuesto por un conjunto de nodos internos o no terminales N y un conjunto de hojas o nodos terminales T. Por otra parte, S denota todo el espacio de instancia, abarcado por todos los atributos que están presentes en los datos. Luego, la estructura de árbol da una subdivisión disjunta de S en regiones St, y cada región está representada por una hoja en el árbol. En (1) se define la ecuación que lo sustenta.

$$S = \bigcup_{t \in T} S_t\,, \quad S_t \cap S_{t'} = \emptyset \; para \; t \neq t' \tag{1}$$

Las hojas t ∈ T tienen una función de regresión logística asociada ft, en lugar de solo una etiqueta de clase (a diferencia de los árboles de decisión ordinarios). La función de regresión ft tiene en cuenta un subconjunto Vt ⊆ V de todos los atributos presentes en los datos (donde se supone que los atributos nominales se han binarizado con el fin de la regresión), y modela las probabilidades de pertenencia a la clase como:

$$\Pr(G = j \mid X = x) = \frac{e^{F_j(x)}}{\sum_{k=1}^{j} e^{F_k(x)}} \tag{2}$$

Donde

$$F_j(x) = \propto_0^j + \sum_{v \in V_t} \propto_v^j . v \quad (3)$$

o su equivalente,

$$F_j(x) = \propto_0^j + \sum_{k=1}^{m} \propto_{vk}^j . v_k \quad (4)$$

Si $\propto_{vk}^j = 0$ para $v_k \notin V_t$. El modelo representado por el LMT modelo logístico de árbol esta entonces dado por:

$$F(x) = \sum_{t \in T} F_t(x) . I(x \in S_t) \quad (5)$$

Donde $I(x \in S_t)$ es 1 si $x \in S_t$ y 0 de lo contrario (Landwehr, Hall, & Frank, 2003).

2.3.2. J48

El algoritmo J48 también conocido como C4.5 es una mejora al algoritmo ID3, el cual fue construido a partir del trabajo de Hovelant and Hunt, desarrollado a final de la década de los 50. Este método de clasificación está apoyado en la teoría de la información de Shanon y para su construcción se parte de un conjunto de datos de entrenamiento T.

Sean las clases {C1, C2, . ., Ck}. Existen tres posibilidades:

- En la primera, cuando el conjunto de datos T contiene uno o más casos, todos pertenecen a una sola clase Cj. En este caso, el árbol de decisión para T es una hoja identificando la clase Cj .
- En la segunda, el conjunto de datos T no contiene casos. El árbol de decisión es una hoja de nuevo, pero la clase asociada debe ser determinada por información que no pertenece a T. Es decir, una hoja puede escogerse de acuerdo con conocimientos de base del dominio, como la clase mayoritaria general.
- La tercera posibilidad es cuando el conjunto de datos T contiene casos que pertenecen a una mezcla de clases. Entonces se refina T en subconjuntos de casos que tiendan, o parezcan tender hacia una colección de casos pertenecientes a una única clase. Para esto, se elige una prueba basada en un único atributo, que tiene uno o más resultados, mutuamente excluyentes {O1, O2,. . ., On}. T se particiona en los subconjuntos T1, T2,. . ., Tn donde Ti contiene todos los casos de T que tienen el resultado Oi para la prueba elegida.

El árbol de decisión para T consiste en un nodo de decisión de acuerdo con la identificación de una prueba, con una rama para cada salida posible. la misma maquinaria de construcción de árboles se aplica de manera recursiva a cada subconjunto de datos de entrenamientos de modo que la rama i-ésima conduce al árbol de decisión construido a partir del subconjunto Ti de datos de entrenamiento.

Para la construcción del árbol en los casos en donde el conjunto de datos T contiene instancias que pertenecen a distintas clases, se realiza una prueba a cada uno de los atributos para determinar cuál es el mejor. Es decir, el atributo que logra separar mayormente bien las clases. Para este proceso se hace uso de la teoría de la información, que sostiene que la información se maximiza cuando la entropía se hace menor.

Primero se calcula la entropía de la clase, la fórmula es:

$$H(S_i) = \sum_{i=1}^{n} - P_i \log P_i \quad (6)$$

Donde P corresponde a la probabilidad o proporción de casos que pertenecen a cada valor de la clase. El atributo at divide al conjunto S en subconjuntos Si, i = 1,2, ... , n, entonces, la entropía total del sistema de subconjuntos será:

$$H(S, at) = \sum_{i=1}^{n} P(S_i).H(S_i) \quad (7)$$

Donde $H(S_i)$ es la entropía del conjunto S_i definida anteriormente y es la probabilidad de que una instancia pertenezca a S_i. Se calcula a partir de los tamaños relativos de cada subconjunto:

$$P(S_i) = \frac{|S_i|}{S_i} \quad (8)$$

Entonces, la ganancia de información está dada por la disminución en entropía.

$$I(S, at) = H(S) - H(S, at) \quad (9)$$

$H(S)$ es el valor de la entropía de todo el conjunto, antes de realizar la subdivisión.

$H(S, at)$ es el valor de la entropía del sistema de subconjuntos generados por la partición (Quinlan, 1994).

A continuación, se detalla el pseudocódigo del algoritmo J48:

```
Función C4.5
R: conjunto de atributos no clasificadores,
C: atributo clasificador,
S: conjunto de entrenamiento, devuelve un árbol de decisión
Comienzo
  Si S está vacío,
  Devolver un único nodo con Valor Falla; para formar el nodo raiz
  Si todos los registros de S tienen el mismo valor para el
  atributo clasificador,Devolver un único nodo con dicho valor; un
  unico nodo para todos
  Si R está vacío,
  Devolver un único nodo con el valor más frecuente del atributo
  Clasificador en los registros de S [Nota: habrá errores, es
  decir, Registros que no estarán bien clasificados en este caso];
  Si R no está vacío,
  D  < - atributo con mayor Proporción de Ganancia (D,S) entre los
  atributos de R;
  Sean {dj | j=1,2,...., m} los valores del atributo D;
  Sean {dj | j=1,2,...., m} los subconjuntos de S correspondientes
  a los valores de dj respectivamente;
  Devolver un árbol con la raíz nombrada como D y con los arcos
  nombrados d1, d2,....,dm, que van respectivamente a los árboles
  C4.5(R-{D}, C, Sl), C4.5(R-{D}, C, S2), C4.5(R-{D}, C, Sm);
Fin
```

2.3.3. Jrip (RIPPER)

La técnica JRIP o RIPPER por sus siglas en inglés (*Repeated Incremental Pruning to Produce Error Reduction*), según (Cohen, 1995), puede resolver problemas multiclase y es una mejora al algoritmo IREP de (Fürnkranz & Widmer, 1996). Este clasificador construye un conjunto de reglas usando covering. Es decir, establece reglas o condiciones de atributo–valor que cubran la mayor cantidad de instancias de una clase y la menor del resto de las clases. luego se añaden pruebas a cada regla para tratar de maximizar la cobertura y minimizar los errores.

JRIP está basado en diferentes conceptos y medidas al mismo tiempo, por un lado hace uso de la ganancia de información para crecer las reglas y por otro lado utiliza la medida del algoritmo

IREP para podar las reglas. Finalmente, establece una medida como criterio de paro para el conjunto global de reglas. Se pueden resumir en tres puntos las mejoras que introdujo JRIP:

1. Métrica alternativa para la fase de poda, RIPPER basa su poda en el criterio siguiente:

$$\upsilon\,(Rule, D_{prune-pos}, D_{prune-neg,}) = \frac{Pos - Neg}{Pos \mp Neg} \qquad (10)$$

2. Incorporación de Heurística para determinar cuando parar el proceso de añadir reglas
3. RIPPER ejecuta una búsqueda local para optimizar el conjunto de reglas (ruleset) de dos formas diferentes:

   a) Reemplazando una regla Ri que forma parte del ruleset$\{R_1, \dots, R_{i-1}, R_i, R_{i+1}, \dots R_k\}$ por R_i', cuando el ruleset correspondiente tenga un menor error en la clasificación en:

$$D_{prune-pos,} \text{ U } D_{prune-neg,} \qquad (11)$$

   b) Revisar una determinada Ri añadiendo literales para que a si se consiga un menor error en:

$$D_{prune-pos,} \text{ U } D_{prune-neg,} \qquad (12)$$

Este algoritmo usa el enfoque divide y vencerás de manera iterativa, para construir reglas que cubren inicialmente ejemplos de entrenamiento anómalos (ejemplos positivos), y a estas reglas iniciales se le van agregando condiciones para cubrir ejemplos comunes (ejemplos negativos) (Camaré, 2008).

Una vez descritas las técnicas de clasificación que mejor resultado generaron, en relación a las métricas de calidad, se detalla el proceso de selección de características en cada una de sus subcategorías. A continuación, se describen las técnicas utilizadas para el desarrollo de esta investigación, con el objetivo de reducir la carga computacional, además de reducir el tiempo de procesamiento de los *dataset*.

2.4 SELECCIÓN DE CARACTERÍSTICAS

La selección de características es un proceso de optimización en el que se intenta seleccionar el mejor subconjunto del conjunto total de características originales, de acuerdo con un criterio y un objetivo de procesamiento especifico según (Hota & Shrivas, 2014). Estas técnicas son muy importantes y de uso frecuente para la reducción de la dimensionalidad, en el proceso de extracción de conocimiento a partir de grandes volúmenes de datos. Con la aplicación de estas técnicas se eliminan atributos irrelevantes, redundantes o ruidosos, lo cual genera grandes efectos para las aplicaciones, tales como la aceleración de un algoritmo de minería de datos, el mejoramiento de la precisión del aprendizaje y la comprensión del modelo (Liu, Motoda, Setiono, & Zhao, 2010).

En la figura 4, se presenta una vista unificada para un proceso de selección de características, el cual consta de dos fases: 1) fase de selección de características y 2) fase de ajuste del modelo y evaluación del rendimiento. En la primera fase, se genera un subconjunto de las características originales el cual es evaluado para estimar la utilidad de las características en dicho conjunto, de acuerdo con la evaluación algunas características pueden descartarse o agregarse al conjunto final de características seleccionadas según su relevancia, luego se determina si ese

conjunto de características es lo suficientemente bueno, usando cierto criterio de detención. Si es así, un algoritmo de selección de características devolverá el conjunto de características seleccionadas, de lo contrario, el proceso se repetirá hasta que se cumpla el criterio de detención.

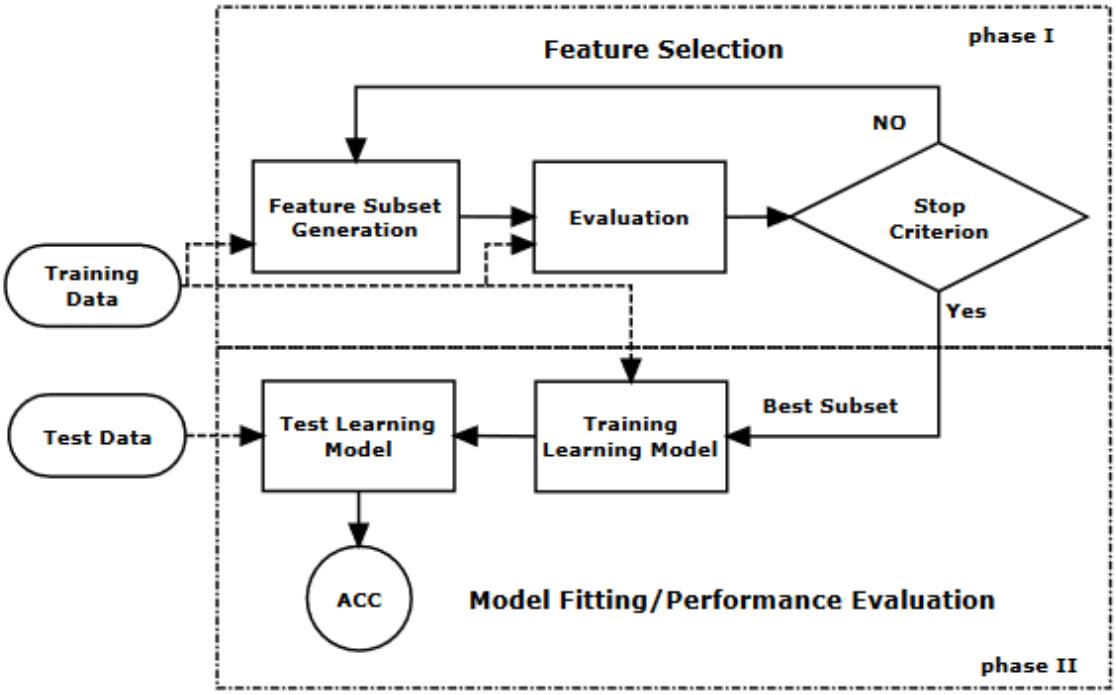

Figura 4. Vista unificada para un proceso de selección de características.

Fuente: (Liu et al., 2010)

En la segunda fase, el conjunto de características seleccionado se utiliza para filtrar los datos de entrenamiento (*train*) y prueba (*test*) para el ajuste y predicción del modelo (Liu et al., 2010).

Para el proceso de selección de características se pueden adoptar diferentes estrategias clasificadas en tres categorías, dependiendo de cómo y cuándo se evalúa la utilidad de los atributos seleccionados, ellas son: filtros (*filter*), envoltorios (*Wrapper*) e integrados (*Embedded*), según (Guyon & Elisseeff, 2003; Liu et al., 2010).

- **Filtros:** este tipo de técnicas de selección de características se basan en el análisis de las características generales de los datos y la evaluación de las características sin involucrar ningún algoritmo de aprendizaje (Liu et al., 2010).
- **Envoltorios:** este tipo de técnicas de selección de características requieren un algoritmo de aprendizaje predeterminado y utilizan su rendimiento en las características proporcionadas en el paso de evaluación para identificar la característica relevante. Estos métodos evalúan un subconjunto de características según la precisión de un predictor dado (Guyon & Elisseeff, 2003; Ron Kohavi & John, 1997).
- **Integrados:** Este tipo de técnicas incorporan la selección de características como parte del proceso de ajuste del modelo / proceso de entrenamiento, estos métodos realizan la función de selección durante el proceso de entrenamiento y, en general, son específicos de determinadas máquinas de aprendizaje (Guyon & Elisseeff, 2003).

En este estudio se aplicaron un total de cinco (5) técnicas de selección de características, las cuales están ubicadas en la categoría de filtros. Dichas técnicas son: Info Gain (Shaltout, Elhefnawi, Rafea, & Moustafa, 2014), Gain Ratio [36], Symmetrical Uncert (Hall & Holmes, 2003), OneR [48] y Relieff (Kira & Rendell, 1992). A continuación, se detallan las cuatro que mejor inciden en el proceso de clasificación con miras a la predicción.

2.4.1 Information Gain

Esta técnica comúnmente se utiliza en la construcción de árboles de decisión a partir de un conjunto de datos de entrenamiento, evaluando la ganancia de información para cada variable y seleccionando la variable que maximiza la ganancia de información, que a su vez minimiza la entropía y divide mejor el conjunto de datos en grupos para una clasificación efectiva.

La ganancia de información (*Information gain*) también se puede utilizar para la selección de características, evaluando la ganancia de cada variable en el contexto de la variable objetivo. En este uso ligeramente diferente, el cálculo se denomina información mutua entre las dos variables aleatorias (Shaltout et al., 2014).

$$IG(D, X_3) = entropy(D) - \sum_v \frac{|D_v| entropy(D_v)}{|D|} \quad (13)$$

2.4.2. Gain Ratio

La medida anterior de ganancia de información, favorece a las características con muchos valores. Puede ocurrir que esta sobreestimación no sea un comportamiento deseable y para evitarlo se puede usar como medida el ratio entre la ganancia de información y la entropía de la característica. Esta medida fue usada por Quinlan en el algoritmo C4.5 (Quinlan, 1994).

$$Gain\ Ratio = \frac{I(F;C)}{H(F)} \quad (14)$$

Este método es muy utilizado y cuenta con popularidad en tener buenos resultados al clasificar, además se caracteriza por tener una particularidad donde una modificación de la ganancia de información que reduce su sesgo, la relación de ganancia toma el número y el tamaño de las ramas en cuenta a elegir un atributo su función se define de la siguiente forma como se expresa en esta fórmula:

$$Gain\ Ratio(x) = gain(x)/sp(x) \quad (15)$$

Esta ecuación expresa la proporción de información generada por la división de la variable candidata que es útil para normalizar la ganancia.

2.4.3. Relieff

El algoritmo de Relieff fue desarrollado por Larry A Rendell (Kira & Rendell, 1992), el cual se limita a los problemas de clasificación con dos clases. Así mismo, la extensión Relief puede hacer frente a los problemas multiclase. El algoritmo es capaz de tratar con datos incompletos y ruidosos.

Por lo tanto, una idea clave del algoritmo Relief, se puede visualizar en la Tabla 1, cuya función es estimar la calidad de los atributos de acuerdo con lo apropiado de sus valores, y estos se distinguen entre instancias que están cerca el uno al otro. Para ese propósito, dada una instancia seleccionada al azar Ri (línea 3), Relief busca por sus dos vecinos más cercanos. uno de la misma clase llamada (Nearesthit H), y otra clase diferente llamada (Nearest miss M) (línea 4).

Tabla 1. Pseudocódigo del algoritmo básico relieff

Algoritmo Relief
Input: para cada instancia de entrenamiento, un vector de valores de atributo y el valor de clase Output: el vector W de estimaciones de las cualidades de los atributos. 1. ajuste todos los pesos W[A] := 0.0; 2. for i := 1 to m do begin 3. seleccionar aleatoriamente una instanciaRi ; 4. encontrar el resultado H más cercano y el error M más cercano; 5. for A := 1 to a do 6. W[A] := W[A]−diff(A,Ri ,H)/m+diff(A,Ri ,M)/m; 7. end;

Fuente: (Kira & Rendell, 1992)

2.4.4. One R

Este método es caracterizado por basarse en la tasa de error de las reglas generadas a partir de un conjunto de atributos, a diferencia de otros algoritmos donde la función solo contiene un atributo que se induce probando sobre el conjunto de entrenamiento en todas las combinaciones de atributos y los valores. Resultante de este proceso el cual permite quedarse con la regla de menos errores, este método funciona como un clasificador de forma muy rápida; adicionalmente, sus resultados son muy buenos en comparación con algoritmos muchos más complejos.

De igual manera este algoritmo también conocido en muchas revisiones documentales que se hicieron como 1R, propuesto por holte en 1993, es un clasificador es un clasificador muy sencillo, que únicamente utiliza un atributo para la clasificación. A pesar de que el autor lo cataloga como "Program 1R is ordinary in most respects." sus resultados pueden ser muy buenos en comparación con algoritmos mucho más complejos y su rendimiento promedio está por debajo de los de C4.5 en solo 5,7 puntos porcentuales de aciertos de clasificación según los estudios realizados por el autor del algoritmo (Holte, 1993), ver tabla 2.

Tabla 2. Pseudocódigo del algoritmo One R

Algoritmo One R
1. Para cada atributo (A) 2. Para cada tributo del (Ai) 3. Contar el número de apariciones de cada clase con Ai 4. Obtener la clase más frecuente (Cj) 5. Crear una regla de tipo Ai->Cj 6. Calcular el error de las reglas del atributo A 7. Escoger las reglas con menor error}

Fuente: (Robles Aranda & Sotolongo, 2013)

Una vez aplicadas las diferentes técnicas de selección de características para la reducción de los datos, se procedió a comparar el rendimiento de las varias técnicas de clasificación, el criterio utilizado para evaluar dicho rendimiento está determinado por las métricas de calidad que son definidas a continuación.

2.5 EVALUACIÓN DE MODELOS

Para efectos de la medición de la calidad de los clasificadores en un conjunto de datos multiclase se detallan a continuación las métricas que han sido utilizadas para tal fin. Primeramente, se detalla la matriz de confusión para analizar las métricas básicas. Luego, se detallan las diferentes métricas de calidad a partir de las métricas básicas anteriormente definidas.

2.5.1 Matriz de confusión

La matriz de confusión es una tabla de frecuencia donde las columnas muestran la clase predicha y las filas a la clase actual, una matriz de confusión es de tamaño l x l , donde l es el número de valores de la clase (Ron Kohavi & Provost, 1998), vea la figura 5.

↓ actual \ predicted→	negative	positive
negative	a	b
positive	c	d

Figura 5: Matriz de confusión para l=2.

Fuente: (Ron Kohavi & Provost, 1998)

Donde:

- **TP - verdaderos positivos**: representados por la letra “d” en la matriz, corresponde al número de predicciones correctas de una instancia que es positiva, donde el algoritmo clasifica algo como verdadero y la salida real es verdadera.
- **TN - verdaderos negativos:** representados por la letra “a” en la matriz, corresponde al número de predicciones correctas de una instancia que es negativa, donde se clasifica algo como falso y la salida real es falsa.
- **FN - falsos negativos:** representados por la letra “c” en la matriz, corresponde al número de predicciones incorrectas de una instancia que es negativa, donde se clasifica algo como falso y la salida real es verdadera.
- **FP - Falsos positivos:** representados por la letra “b” en la matriz, corresponde al número de predicciones incorrectas de una instancia que es positiva, donde se clasifica algo como verdadero y la salida real es falsa (Berges Gonzalez, 2010).

Las métricas básicas se utilizan para calcular métricas de calidad, tales como: TP-Rate (tasa de verdaderos positivos), FP-Rate (tasa de verdaderos negativos), Precision, Recall, F-Measure y Roc área. Cada una de estas métricas se define en detalle a continuación.

2.5.2. Métricas de calidad

Recall: es el número de elementos identificados correctamente como positivos del total de positivos verdaderos. Se le conoce también como tasa de verdaderos positivos TP-Rate, es la proporción de instancias que se informaron como relevantes contra todas las instancias verdaderamente relevantes (Anderson, Bergés, Ocneanu, Benitez, & Moura, 2012; Berges Gonzalez, 2010).

$$R = \frac{TP}{TP+FN} \quad (16)$$

FP-Rate: la tasa de falsos positivos es la proporción de falsos positivos frente a los resultados negativos reales (Anderson et al., 2012; Ron Kohavi & Provost, 1998).

$$FP - Rate = \frac{FP}{FP + TN} \quad (16)$$

Precision: también llamado valor predictivo positivo, VPP), es la proporción de instancias relevantes que fueron reportadas como relevantes contra todas las instancias que fueron reportadas como relevantes (Berges Gonzalez, 2010; Ron Kohavi & Provost, 1998)**.**

$$R = \frac{TP}{TP + FP} \quad (17)$$

F-Measure: es la media armónica entre dos métricas: Recall y Precision (Anderson et al., 2012; Berges Gonzalez, 2010).

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (18)$$

Roc area: *Receive Operating Characteristics*, también denominada la curva de operación del receptor, representa el valor de razón de **TP** contra la razón de **FP.** El AUC (*area under the curve*) área bajo la curva ROC, se usa como indicador de la calidad del clasificador mientras más cerca esté a la unidad mejor su desempeño (Berges Gonzalez, 2010; Provost & Fawcett, 2001).

Una vez detalladas las métricas de calidad utilizadas en este estudio para valorar la confiabilidad de los modelos construidos a partir de técnicas de clasificación, se procede a describir el proceso de experimentación, a continuación, se presenta el proceso de construcción del modelo propuesto en esta investigación.

3. PROCESO EXPERIMENTAL

Esta investigación se fundamenta en el reconocimiento de actividades humanas HAR y las diferentes técnicas tanto desde el punto de vista de la clasificación, como de la selección de características, para la construcción de modelos funcionales que permitan identificar las actividades de la vida diaria ADL. Teniendo en cuenta que dicho modelo funcional es entrenado con unas colecciones de datos o *dataset* y que una vez entrenado debe ser evaluado a partir del análisis de métricas de calidad, profundamente documentadas en la literatura científica y anteriormente mencionadas en la sección 2.5.2. A continuación se presenta el proceso de construcción del modelo propuesto, detallando las diferentes fases de la experimentación.

3.1 DESCRIPCIÓN DE LA PROPUESTA

La propuesta aquí descrita parte del preprocesamiento de los datos originales, provistos por el *dataset* Aruba CASAS. Gracias a este procedimiento se obtuvo como producto el *dataset* procesado, a partir del cual se generaron tres nuevos subconjuntos de datos: Aruba CASAS - raw, Aruba CASAS - duration y Aruba CASAS - sensor based. Para cada uno de los tres dataset se realizó un proceso de construcción de un modelo funcional, posteriormente se hizo una comparativa de métricas de calidad para cada modelo, la cual arrojó como resultado la elección del mejor modelo validado además de la correcta configuración del dataset en cuanto a las categorías de características, ver figura 6.

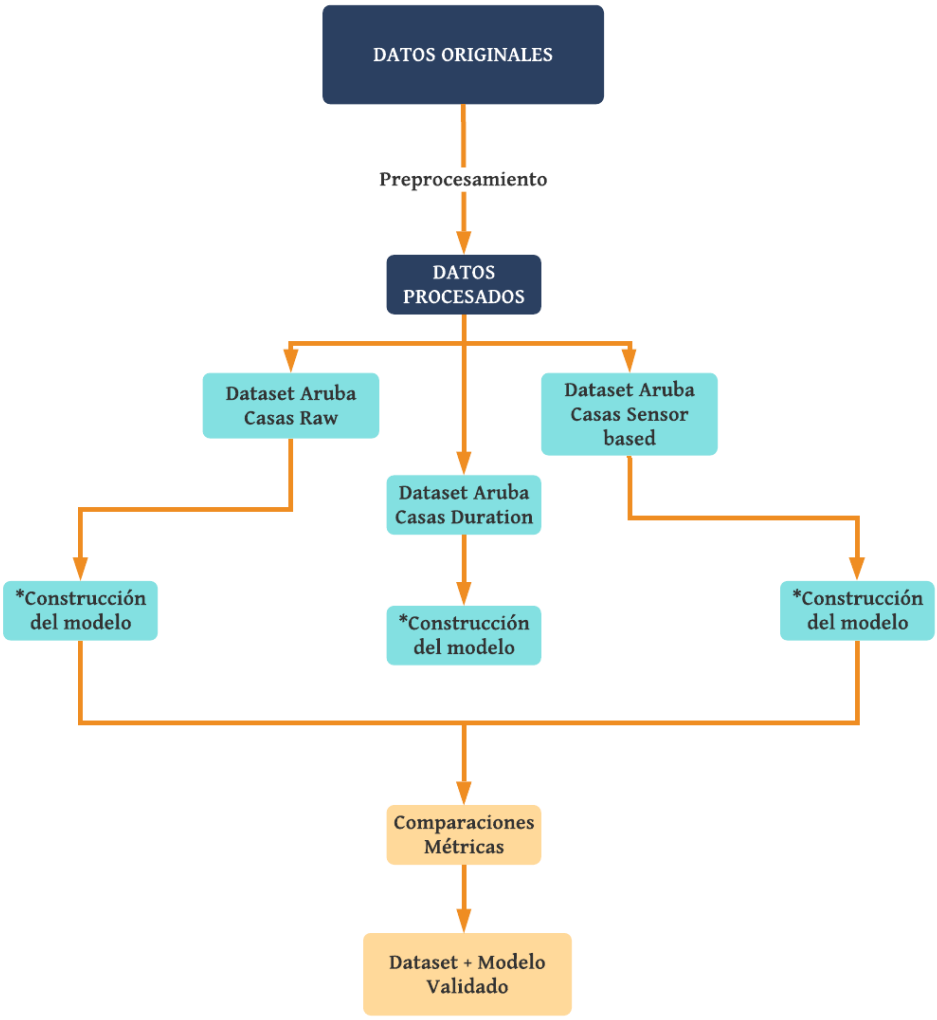

Figura No 6. Preparación de los datos y construcción del modelo propuesto

Fuente: Elaboración propia

3.2 PREPROCESAMIENTO DE LOS DATASET

Esta investigación tuvo como punto de partida los datos originales provistos por el *dataset* Aruba CASAS detallados en la sección 1.2, los cuales están conformados por el número de eventos registrados tanto por los sensores binarios (de movimiento y de contacto), como por los sensores de temperatura. Además, incluye fecha y hora de inicio y de finalización de cada actividad.

Inicialmente se realizó una fase preprocesamiento la cual consistió en la generación de características a partir de la representación de los marcos de tiempo de duración de las actividades, extraídos de las instancias de datos originales. Este procedimiento dio origen al dataset procesado, cuya estructura se detalla a continuación.

El dataset procesado está constituido en total por 69 características, divididas en cuatro (4) categorías: características representadas mediante conteo, características representadas mediante promedios, características representadas mediante funciones de agregación y características originales.

Las características representadas mediante conteo, se construyen a partir de los sensores de contacto en las puertas, en total existen cuatro (4) sensores de contacto y se efectuó conteo tanto para la apertura como para el cierre (OPEN/CLOSE), en los marcos de tiempo de duración de las actividades. Por lo tanto, se generaron ocho (8) características de sensores de contacto de puertas.

Por otra parte, a partir de los sensores de movimiento también se generaron características, representadas a partir del conteo. Sin embargo, dado que los sensores de movimiento tienen los estados de activación y desactivación (ON/OFF) casi simultáneos (es decir, un estado OFF se ejecuta inmediatamente posterior al estado ON), el conteo de un evento se realizó a partir del par de estados (ON y OFF) para cada sensor. Por ello, se generan 31 características de sensores de movimiento. Otras características representadas mediante conteo, son el número de eventos que en el marco de tiempo se realizó una determinada actividad y la característica de duración que corresponde al cálculo en segundos a partir de la diferencia entre la fecha y hora de inicio y fecha y hora de finalización de la actividad.

En cuanto a las características representadas mediante promedios, éstas han sido calculadas a partir de los sensores de temperatura, debido a que los valores que toman son datos continuos,

adicionando un total de cinco (5) características al dataset. Adicionalmente, a partir de estos sensores se generaron otras características representadas mediante funciones de agregación, y dado que se usaron cuatro (4) fórmulas estadísticas (rango, desviación estándar, sesgo y curtosis) por cada uno de los sensores, en total se generaron 20 características representadas para esta categoría. Las tres (3) características restantes, hacen parte de la categoría de características originales y corresponden a la etiqueta de clase, la fecha y hora de inicio y la fecha y hora de finalización de la actividad realizada. Para mayores precisiones la tabla 3 contiene la estructura del *dataset* procesado.

A partir del *dataset* procesado se generaron tres subconjuntos de datos que se denominarán: Aruba CASAS - *raw*, Aruba CASAS - *duration* y Aruba CASAS - *sensor based*, los cuales difieren en el número de características y tienen la siguiente configuración:

El *dataset* Aruba CASAS - *raw* tiene un total de 47 características, de las cuales 39 corresponden a la categoría de características representadas por conteo, cinco (5) a la categoría de características representadas por promedio y los tres (3) restantes a la categoría de características originales. El *dataset* Aruba CASAS - *duration* tiene un total de 49 características de las cuales 41 corresponden a la categoría de características representadas por conteo, cinco (5) a la categoría de características representadas por promedio y los tres (3) restantes a la categoría de características originales. En cuanto al *dataset* Aruba CASAS - *sensor based*, éste está conformado por un total de 67 características de las cuales 39 corresponden a la categoría de características representadas por conteo, cinco (5) a la categoría de características representadas por promedio, 20 a la categoría de características representadas mediante funciones de agregación (detalladas en la sección 3.2.1) y los tres (3) restantes a la categoría de características originales.

Tabla 3.Estructura dataset procesado

Características	**Características representadas por conteo**				**Características representadas por promedio**	**Características representadas por funciones de agregación**	**Características originales**			**Total**
	Sensores de contacto en puertas	Sensores de movimiento	Numero de eventos	Duración de la actividad	Sensores de temperatura	Sensores de temperatura	Inicio actividad	Fin actividad	Clase	
	D001-open, D001-close, D002-open, D002-close, D003-open, D003-close, D004-open y D004-close (Total: 8)	M001, M002, M003, M004, M005, M006, M007, M008, M009, M010, M011, M012, M013, M014, M015, M016, M017, M018, M019, M020, M021, M022, M023, M024, M025, M026, M027, M028, M029, M030 y M031 (Total: 31)	Eventos (Total: 1)	Duración (Total: 1)	T001, T002, T003, T004 Y T005 (Total: 5)	T001-RANGO, T001-DESV, T001-SESGO, T001-KURT, T002-RANGO, T002-DESV, T002-SESGO, T002-KURT, T003-RANGO, T003-DESV, T003-SESGO, T003-KURT, T004-RANGO, T004-DESV, T004-SESGO, T004-KURT, T005-RANGO, T005-DESV, T005-SESGO y T005-KURT (Total: 20)	Fecha y hora de inicio (Total: 1)	Fecha y hora de fin (Total: 1)	Actividad (Total: 1)	**69 características**

Tabla 4. Configuración dataset Aruba CASAS - raw, Aruba CASAS - duration y Aruba CASAS - sensor based

Dataset	**Características representadas por conteo**				**Características representadas por promedio**	**Características representadas por funciones de agregación**	**Características originales**			**Número de Características**
	Sensores de contacto en puertas	Sensores de movimiento	Numero de eventos	Duración de la actividad	Sensores de temperatura	Sensores de temperatura	Inicio actividad	Fin actividad	Clase	
Aruba CASAS – raw	8	31	0	0	5	0	1	1	1	**47**
Aruba CASAS - duration	8	31	1	1	5	0	1	1	1	**49**
Aruba CASAS - sensor based	8	31	0	0	5	20	1	1	1	**67**

Estos dataset se generaron para efectuar pruebas posteriores e identificar cual conjunto de datos produce mejores resultados en cuanto a la capacidad de clasificación de las técnicas de *machine learning*. Dichas técnicas fueron propuestas con el objeto de evaluar la incidencia de una u otra categoría de características en la capacidad de clasificación de la técnica. En la tabla 4 se identifican el número de características de estos tres *dataset* a partir de las categorías de características que los integran.

Por otra parte, los subconjuntos de datos usados para el entrenamiento (*train*) y prueba (*test*), en el proceso de construcción del modelo, siguen la distribución de instancias de datos presentadas en la tabla 5. Las proporciones corresponden al 69,90% para el subconjunto *train* y 30,10% para el subconjunto *test*, en cada uno de los tres dataset (Aruba CASAS - *raw*, Aruba CASAS - *duration* y Aruba CASAS - *sensor based*).

Tabla 5. Distribución de instancias para subconjuntos de datos train y test de los dataset Aruba CASAS - raw, Aruba CASAS - duration y Aruba CASAS - sensor based

Dataset	**Subconjunto train**		**Subconjunto test**		**Total instancias**
	Instancias	Porcentaje	Instancias	Porcentaje	
Aruba CASAS - raw	4460	69,9%	1916	30,1%	6376
Aruba CASAS - duration	4460	69,9%	1916	30,1%	6376
Aruba CASAS - sensor based	4460	69,9%	1916	30,1%	6376

Para la construcción de cada subconjunto (*train* y *test*), se seleccionaron las instancias de manera aleatoria, aproximadamente la misma proporción de instancias por cada etiqueta de clase. Es decir, aproximadamenre el 70% para el entrenamiento y el 30% para la prueba, ver la tabla 6.

Tabla 6. Distribución de instancias de datos por clase para subconjuntos de datos train y test para los dataset Aruba CASAS - raw, Aruba CASAS - duration y Aruba CASAS - sensor based

Dataset / Class		Sleeping	Bed to Toilet	Meal Preparation	Relax	House keeping	Eating	Leave Home	Enter Home	Work
Aruba CASAS - raw	**Train**	265	112	1105	2036	23	181	298	319	121
	Test	136	45	482	878	9	71	133	112	50
Aruba CASAS - duration	**Train**	265	112	1105	2036	23	181	298	319	121
	Test	136	45	482	878	9	71	133	112	50
Aruba CASAS - sensor based	**Train**	265	112	1105	2036	23	181	298	319	121
	Test	136	45	482	878	9	71	133	112	50

3.2.1 Funciones de agregación

Para la conformación del *dataset* Aruba CASAS - *sensor based* fue necesario el cálculo de varias características a partir de funciones de agregación. En dicho proceso, las instancias se agruparon por el criterio de clase. Es decir, por actividad, específicamente a partir de las características de temperatura, las funciones usadas fueron: rango (*Range*), desviación estándar (*Standard Deviation*), sesgo (*Skewness*) y curtosis (*Kurtosis*), utilizando las funciones definidas en (Rice, 2006). Cada una de las cuales se detalla a continuación:

- **Rango:** es la diferencia entre el valor más grande y el valor más pequeño de un conjunto de datos.

$$o_{range} = o_{max} - o_{min} \quad (16)$$

- **Desviación estándar:** es la raíz cuadrada de la varianza. La varianza es la suma de todas las diferencias al cuadrado de cada valor de ocurrencia a la media, dividida por el número de sensores S menos 1.

$$\sigma = \sqrt{\frac{1}{s-1}\sum_{i=1}^{s}(o_{i-}\mu^2)} \quad (17)$$

- **Sesgo:** se define como el cociente del tercer momento central m_3 de un conjunto de datos y la desviación estándar al cubo.

$$\gamma = \frac{m_3}{\sigma^3} = \frac{\frac{1}{s}\sum_{i=1}^{s}(o_{i-}\mu)^3}{\sqrt{\frac{1}{s}\sum_{i=1}^{s}(o_{i-}\mu^2)}^3} \quad (18)$$

- **Curtosis:** se define como el cociente del cuarto momento central de un conjunto de datos m_4, y la desviación estándar σ a la cuarta potencia.

$$\kappa = \frac{m_4}{\sigma^4} = \frac{\frac{1}{s}\sum_{i=1}^{s}(o_{i-}\mu)^4}{\sqrt{\frac{1}{s}\sum_{i=1}^{s}\left(o_{i-}\mu^2\right)}^4} \quad (19)$$

3.3 CONSTRUCCIÓN DEL MODELO

Se construyeron diferentes modelos a partir de los tres dataset, implementando técnicas de clasificación, integradas con técnicas de selección de características. Producto de la evaluación de las métricas de calidad, se identificaron los mejores resultados por cada *dataset*. Es decir, cada evaluación permitió identificar cuáles eran las mejores combinaciones de técnicas de clasificación con técnicas de selección de características, que generaron las más altas métricas de calidad por

cada *dataset* evaluado (Aruba CASAS - *raw*, Aruba CASAS - *duration* y Aruba CASAS - *sensor based*). Un análisis comparativo integral, de los resultados arrojados por estas evaluaciones permitió identificar el dataset que generaba mejores resultados a nivel de clasificación y las respectivas técnicas de clasificación y selección de características, que conllevaron a generar esos mejores resultados, ver figura 7.

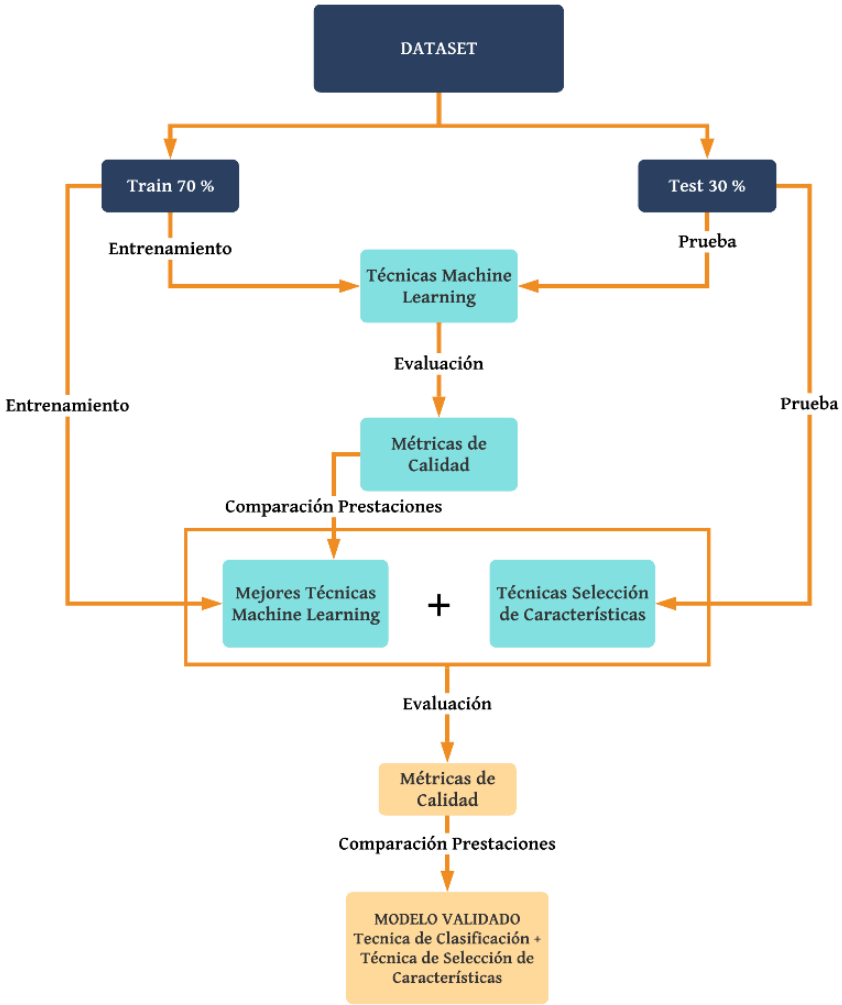

Figura 7. Construcción del modelo

Fuente: Elaboración propia

3.4 EXPERIMENTACIÓN

Con el objetivo de construir un modelo que arroje los mejores resultados a nivel de métricas de calidad. Además de evaluar diferentes configuraciones de categorías de características para el *dataset*, que incidan mayormente en el proceso de clasificación, se plantearon tres escenarios de experimentación.

En un primer escenario experimental, fueron aplicadas diferentes técnicas de clasificación a cada uno de los tres subconjuntos de datos (Aruba CASAS - *raw*, Aruba CASAS - *duration* y Aruba CASAS - *sensor based*), mencionados en la sección 3.2. Con el objetivo de identificar las técnicas que generan mejores métricas de calidad en cada uno de los experimentos. Para dicha evaluación, se realizó un muestreo aleatorio de instancias de cada dataset con el propósito de dividirlos en entrenamiento y prueba, de los cuales cada conjunto de datos de entrenamiento (*train*) es el 70% de las muestras, y cada conjunto de datos de prueba (*test*) es el 30% aproximadamente.

En un segundo escenario experimental, se aplicaron diferentes técnicas de selección de características a los dataset train y test, de cada subconjunto de datos, y se identificó el óptimo número de características con la técnica de clasificación que mejor afecta el proceso de evaluación para cada uno de los subconjuntos de datos. En un tercer escenario experimental, para cada dataset, se evaluó exhaustivamente el rendimiento de la mejor hibridación, de técnica de clasificación con técnica de selección de características, usando validación cruzada con 10 fold (pliegues).

En la sección 4, se detallan cada uno de los escenarios de experimentación planteados, incluyendo los experimentos realizados para cada subconjunto de datos (Aruba CASAS - *raw*, Aruba CASAS - *duration* y Aruba CASAS - *sensor based*).

4. ESCENARIOS DE EXPERIMENTACIÓN

En este capítulo, se describen diferentes escenarios de experimentación para la creación de un modelo predictivo de HAR (Reconocimiento de Actividades Humanas), aplicando variación de técnicas de clasificación y selección de características a los subconjuntos de datos Aruba CASAS - *raw*, Aruba CASAS - *duration* y Aruba CASAS - *sensor based* (descritos en la sección 3.2), generados a partir del dataset original, Aruba CASAS. Posteriormente, con el fin de confrontar el rendimiento de diferentes enfoques de aprendizaje automático, se aplicó un análisis comparativo de las métricas de calidad detalladas en la sección 2.5.2, para cada uno de los tres escenarios de experimentación recreados: 1) con técnicas de clasificación, 2) mediante la hibridación de técnicas de clasificación y selección, y 3) evaluando los mejores resultados mediante la aplicación de validación cruzada. En los tres escenarios, se utilizaron los tres subconjuntos de datos preprocesados, para identificar cual es el subconjunto de datos, al ser procesado mediante las respectivas técnicas, genera mejores métricas de calidad en el proceso predictivo.

4.1 ESCENARIO EXPERIMENTAL No 1: ANÁLISIS COMPARATIVO DE TÉCNICAS DE CLASIFICACIÓN A SUBCONJUNTOS DE DATOS

En éste primer escenario, se realizaron tres experimentos, evaluando en cada uno de ellos 31 técnicas de clasificación, en los tres dataset (Aruba CASAS - *raw*, Aruba CASAS - *duration* y Aruba CASAS - *sensor based*). Por cada experimento se utilizaron subconjuntos de datos, a partir del respectivo *dataset*, para el proceso de entrenamiento (*train*) y el proceso de prueba (*test*). Las técnicas de clasificación evaluadas en los diferentes experimentos de este escenario se presentan en la tabla 7, indicando la subcategoría a la cual corresponden.

Tabla 7. Técnicas evaluadas en los diferentes experimentos por subcategorías

Subcategorías	Técnica	Función
Árboles de decisión	Logistic Model Trees - LMT (Landwehr et al., 2003)	Construye árboles de modelo logístico
	J48 (C4.5 decision tree) (Quinlan, 1994)	árbol de decisión basado en algoritmo C4.5
	Reduced-Error Pruning Tree - REPTree (Frank & Witten, 1998)	Aprendizaje de árbol rápido que usa la poda en la reducción de errores
	RandomForest (Breiman, 2001)	Construcción de árboles aleatorios.
	Random Tree (Frank & Witten, 1998)	Construir un árbol que considera un número aleatorio de características dadas en cada nodo.
	DecisionStump (Marks Hall, 1994)	Construye árboles de decisión de un nivel.
Reglas	JRip (Cohen, 1995)	Algoritmo RIPPER (poda incremental reducida para producir reducción de error) para rapidez, regla de inducción eficaz
	Partial Decision Trees – PART (Frank & Witten, 1998)	Obtiene reglas a partir de árboles de decisión construidos usando J4.8.
	Decision Table (Ron Kohavi & Provost, 1998)	Construye una tabla de decisión simple del clasificador mayoritario.
	ZeroR, Stacking (Wolpert, 1992)	Predice la clase mayoritaria (si es nominal) o el valor promedio (si es numérico). Funciones
	OneR (Holte, 1993)	Clasificador de una regla
Funciones	Logistic (Cessie & Houwelingen, 1992)	Construye modelos de regresión logística lineal.
	MultilayerPerceptron (Van Der Malsburg, 1986)	Red neuronal de propagación hacia atrás
Multiclasificadores (Meta)	Random Committee (Frank & Witten, 1998)	Construye un conjunto de clasificadores base aleatorios
	Stacking (Wolpert, 1992)	Combina varios clasificadores usando el método apilado (stacking).
	LogitBoost (Friedman, Hastie, & Tibshirani, 2000)	Realiza regresión logística aditiva
	Classification Via Regression (Frank et al., 1998)	Realiza clasificación mediante un método de regresión
	MultiClass Classifier (Read, Puurula, & Bifet, 2015)	Usa un clasificador de dos clases para conjuntos de datos multiclases.
	Bagging (Breiman, 1996)	Un clasificador bolsa (bag), trabaja por regresión también.
	AdaBoostM1 (Freund & Schapire, 1996)	Usa el método AdaBoostM1
	Vote (Kittler, Hatef, Duin, & Matas, 1998)	Combina clasificadores usando promedio de estimados de probabilidad o predicciones numéricas.
	CVParameterSelection (R Kohavi, 1995)	Realiza la selección de parámetros mediante validación cruzada
	MultiScheme (Eibe et al., 2007)	Utiliza la validación cruzada para seleccionar un clasificador de varios candidatos
	AttributeSelectedClassifier (Frank et al., 1998)	Reduce la dimensionalidad de los datos mediante la selección de atributos
	RandomSubSpace (T.K. Ho, 1998)	construye un clasificador basado en árbol de decisión que mantiene la mayor precisión en los datos de entrenamiento
	Filtered Classifier (Eibe et al., 2007)	Ejecuta un clasificador en datos filtrados
Algoritmos perezosos (Lazy)	IB1 Instance-based Learning Algorithms (Aha, Kibler, & Albert, 1991)	Aprendizaje basado en instancia un vecino más cercano básico.
	IB2 Instance-based Learning Algorithms (Aha et al., 1991)	Clasificador k vecino más cercano.
	IB3 Instance-based Learning Algorithms (Aha et al., 1991)	Clasificador k vecino más cercano.
	KStar (Cleary & Trigg, 1995)	Vecino más cercano con función de distancia generalizado.
	LWL (Frank, Hall, & Pfahringer, 2003)	Algoritmo general para aprendizaje localmente pesado.

Fuente: (Witten et al., 2011)

Para los experimentos con los dataset Aruba CASAS - *raw* y Aruba CASAS - *sensor based*, ver tabla 8, los clasificadores con mejores resultados en cuanto a la métrica *Recall* fueron LMT con 94,50% y *LogitBoost* con 94,20% cuando se evaluaron ambos dataset. Complementariamente, se puedo identificar que en estos casos la métrica ROC area fue de 99,60% y 99,70% respectivamente. En cuanto a la prueba con el dataset Aruba CASAS - *duration*, las técnicas de clasificación con más alto *Recall* fueron J48 y JRIP con 95,60% para ambos clasificadores, presentando JRIP la más alta métrica ROC Area, con un 99,30%.

Se evidencia que LMT es la técnica de clasificación que presenta mejores resultados en cuanto a la métrica *Recall* tanto con el dataset Aruba CASAS - *raw* como con Aruba CASAS - *sensor based*. En cuanto al dataset Aruba CASAS – *duration*, a pesar de que LMT no fue la técnica con los mejores resultados a nivel de clasificación, presentó un representativo *recall*, del 95,40%, tal como se muestra en la tabla 8.

Tabla 8. Comparativa de las mejores técnicas de clasificación para los dataset (Train y Test)

Aruba CASAS

Dataset	Métricas de calidad					Técnica de clasificación
	FP Rate	Precision	Recall	F-Measure	ROC Area	
Aruba CASAS – raw	0,50%	94,80%	**94,50%**	94,50%	**99,60%**	**LMT**
	0,60%	94,60%	**94,20%**	94,00%	**99,70%**	LogitBoost
	0,60%	94,30%	94,10%	94,10%	99,70%	ClassificationViaRegression
	0,50%	94,30%	94,00%	94,00%	99,00%	J48
Aruba CASAS – duration	0,50%	95,70%	**95,60%**	95,60%	**99,00%**	J48
	0,60%	95,70%	**95,60%**	95,50%	**99,30%**	JRIP
	0,60%	95,40%	**95,40%**	95,40%	**99,60%**	**LMT**
	0,70%	95,20%	95,30%	95,10%	99,80%	RandomSubSpace
Aruba CASAS - sensor based	0,70%	94,80%	**94,50%**	94,40%	**99,70%**	**LMT**
	0,60%	94,60%	**94,20%**	94,00%	**99,70%**	LogitBoost
	0,60%	94,30%	94,10%	94,10%	99,70%	ClassificationViaRegression
	0,60%	94,20%	93,90%	93,90%	99,00%	J48

4.2 ESCENARIO EXPERIMENTAL No 2: ANÁLISIS COMPARATIVO DE LA HIBRIDACIÓN DE TÉCNICAS DE SELECCIÓN Y CLASIFICACIÓN A SUBCONJUNTOS DE DATOS

En este escenario se efectuaron tres experimentos, cada uno con los respectivos dataset antes mencionados (Aruba CASAS - *raw*, Aruba CASAS - *duration* y Aruba CASAS - *sensor based*), con el objetivo de minimizar los tiempos computacionales, se buscó reducir la dimensionalidad de los tres dataset, identificando el conjunto de características que mejor afectan la clasificación. Por esta razón, las técnicas de selección de características *Info Gain* (Hall & Holmes, 2003), *Gain Ratio* (Hall & Holmes, 2003), *Symmetrical Uncert* (Hall & Holmes, 2003),

OneR (Holte, 1993) y *Relieff* (Robnik-Šikonja & Kononenko, 1997), fueron combinadas con cada una de las cuatro técnicas de clasificación, que generaron mejores resultados, a partir del análisis del primer escenario, para cada *dataset*.

Una vez evaluadas las métricas de calidad, fue posible determinar que la hibridación de las técnicas de clasificación con las técnicas de selección de características, que generaron los más altos resultados fueron: 1) LMT con *Gain Ratio* utilizando tanto 27 como 24 características, para el dataset Aruba CASAS - *raw*, ver tabla 9; 2) JRIP con *One R* utilizando 47 características y LMT con *One R* utilizando 33 características, para el dataset Aruba CASAS - *duration*, ver tabla 10; y 3) LMT con *Info Gain* utilizando 47 características y LMT con *Gain Ratio* utilizando 31 características, para el dataset Aruba CASAS - *sensor based*, ver tabla 11.

Tabla 9. Comparativa entre la hibridación de técnicas LMT + Gain Ratio con diferente número de características para los dataset Train y Test de Aruba CASAS - raw

Class	LMT + Gain Ratio (27 características)					LMT + Gain Ratio (24 características)				
	FP Rate	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**
Sleeping	0,00%	100,00%	100,00%	100,00%	100,00%	0,00%	100,00%	100,00%	100,00%	100,00%
Bed_to_Toilet	0,00%	100,00%	100,00%	100,00%	100,00%	0,00%	100,00%	100,00%	100,00%	100,00%
Meal_Preparation	0,10%	99,80%	98,50%	99,20%	99,90%	0,10%	99,60%	98,50%	99,10%	99,80%
Relax	0,60%	99,30%	99,80%	99,50%	100,00%	0,40%	99,50%	99,40%	99,50%	99,90%
Housekeeping	0,00%	100,00%	88,90%	94,10%	100,00%	0,10%	90,00%	100,00%	94,70%	100,00%
Eating	0,20%	94,70%	100,00%	97,30%	100,00%	0,30%	92,20%	100,00%	95,90%	100,00%
Leave_Home	1,50%	73,00%	54,90%	62,70%	98,10%	1,50%	73,30%	55,60%	63,20%	98,10%
Enter_Home	3,30%	59,00%	75,90%	66,40%	97,90%	3,20%	59,40%	75,90%	66,70%	97,90%
Work	0,00%	100,00%	100,00%	100,00%	100,00%	0,00%	100,00%	100,00%	100,00%	100,00%
Average	0,60%	95,20%	**94,90%**	94,90%	**99,70%**	0,50%	95,10%	**94,90%**	94,90%	**99,70%**

Para el experimento con el dataset Aruba CASAS - *raw*, ambas propuestas (27 y 24 características) lograron **aumentar la métrica *recall* a 94,90% y *ROC area* a 99,70%** (respecto al escenario inicial en el cual se utilizaron los datasets con todas las 47 características, obteniendo un recal del 94,50% y un ROC area de 99,60%), siendo **LMT con *Gain Ratio* (24 características)** la combinación que logra **una mayor reducción en cuanto al número de atributos**, a ser usados en el proceso de clasificación, ver tabla 9.

Tabla 10. Comparativa entre la hibridación de técnicas JRIP + One R y LMT + One R con los dataset Train y Test de Aruba CASAS – duration

Class	JRIP + One R (47 características)					LMT + One R (33 características)				
	FP Rate	Precision	Recall	F-Measure	ROC Area	FP Rate	Precision	Recall	F-Measure	ROC Area
Sleeping	0,00%	100,00%	98,50%	99,30%	99,60%	0,00%	100,00%	100,00%	100,00%	100,00%
Bed_to_Toilet	0,00%	100,00%	97,80%	98,90%	98,90%	0,00%	100,00%	97,80%	98,90%	100,00%
Meal_Preparation	0,30%	99,20%	98,60%	98,90%	99,20%	0,10%	99,80%	98,60%	99,20%	99,90%
Relax	0,50%	99,40%	99,50%	99,50%	99,60%	0,70%	99,20%	99,70%	99,40%	99,80%
Housekeeping	0,10%	87,50%	77,80%	82,40%	88,90%	0,00%	100,00%	77,80%	87,50%	100,00%
Eating	0,40%	90,90%	98,60%	94,60%	98,80%	0,30%	92,10%	98,60%	95,20%	98,80%
Leave_Home	2,20%	73,50%	81,20%	77,10%	98,30%	2,30%	72,50%	83,50%	77,60%	98,50%
Enter_Home	1,30%	75,30%	65,20%	69,90%	97,00%	1,30%	75,30%	62,50%	68,30%	98,20%
Work	0,10%	98,00%	100,00%	99,00%	100,00%	0,10%	98,00%	96,00%	97,00%	100,00%
Average	0,50%	95,80%	**95,80%**	95,80%	**99,20%**	0,60%	95,90%	**95,90%**	95,80%	**99,70%**

En el experimento con el dataset Aruba CASAS – *duration*, a pesar de que el clasificador J48 (en el primer escenario de experimentación), había generado muy buenos resultados obteniendo un 95,60% en la métrica *recall* (con 49 características, como se puede ver en la tabla 3), las propuestas de hibridación JRIP con One R utilizando 47 características y LMT con One R utilizando 33 características (y ejecutadas en este segundo escenario), incrementaron el *recall*, alcanzando un 95,80% y 95,90% respectivamente. Además, se logró una reducción significativa

del número de características para el proceso de clasificación. Para una mejor apreciación vea la tabla 10. Es evidentre que de estas dos combinaciones de técnicas **es mejor utilizar LMT con One R porque genera un mayor recall (95,90%) y porque solo requiere 33 características para el proceso de clasificación.**

Tabla 11. Comparativa entre la hibridación de técnicas LMT + Info Gain y LMT + Gain Ratio con los dataset Train y Test de Aruba CASAS - sensor based

Class	LMT + Info Gain (47 características)					LMT + Gain Ratio (31 características)				
	FP Rate	Precision	Recall	F-Measure	ROC Area	FP Rate	Precision	Recall	F-Measure	ROC Area
Sleeping	0,10%	99,30%	100,00%	99,60%	100,00%	0,00%	100,00%	100,00%	100,00%	100,00%
Bed_to_Toilet	0,00%	100,00%	100,00%	100,00%	100,00%	0,00%	100,00%	100,00%	100,00%	100,00%
Meal_Preparation	0,10%	99,80%	99,20%	99,50%	100,00%	0,20%	99,40%	98,30%	98,90%	100,00%
Relax	0,50%	99,40%	99,80%	99,60%	99,90%	0,60%	99,30%	99,80%	99,50%	100,00%
Housekeeping	0,00%	100,00%	44,40%	61,50%	91,00%	0,00%	100,00%	88,90%	94,10%	100,00%
Eating	0,20%	94,70%	100,00%	97,30%	100,00%	0,20%	94,60%	98,60%	96,60%	100,00%
Leave_Home	1,50%	73,00%	54,90%	62,70%	98,00%	1,50%	73,30%	55,60%	63,20%	98,10%
Enter_Home	3,30%	59,00%	75,90%	66,40%	97,80%	3,20%	59,40%	75,90%	66,70%	97,90%
Work	0,10%	98,00%	100,00%	99,00%	100,00%	0,00%	100,00%	100,00%	100,00%	100,00%
Average	0,50%	95,10%	**94,90%**	94,80%	**99,70%**	0,60%	95,10%	**94,90%**	94,80%	**99,70%**

Si bien en el primer escenario, utilizando el dataset Aruba CASAS - *sensor based* con 67 características, se obtuvo un recall de 94,50% y un ROC area de 99,70% con la técnica LMT; en este escenario, en con el mismo dataset, ambas propuestas (LMT + Info Gain con 47 características y LMT + Gain ratio con 31 características) presentaron un incremento en el ***recall***, el cual fue de **94,90%, siendo LMT con Gain Ratio la combinación que logró una mayor disminución en cuanto al número de características** (lo cual incide en el tiempo computacional utilizado por el modelo predictivo), como se puede apreciar en la tabla 11.

Tabla 12. Comparativa entre las mejores hibridaciones de técnicas de clasificación y selección de características con los dataset Train y Test de cada subconjunto de datos

Dataset	Métricas de calidad					Hibridación Técnica de clasificación + Selección de características
	FP Rate	Precision	Recall	F-Measure	ROC Area	
Aruba CASAS - raw	0,60%	95,20%	94,90%	94,90%	99,70%	LMT + Gain Ratio (27 Características)
	0,50%	**95,10%**	**94,90%**	**94,90%**	**99,70%**	**LMT + Gain Ratio (24 características)**
Aruba - duration	0,50%	95,80%	95,80%	95,80%	99,20%	JRIP + One R (47 Características)
	0,60%	**95,90%**	**95,90%**	**95,80%**	**99,70%**	**LMT + One R (33 Características)**
Aruba - sensor based	0,50%	95,10%	94,90%	94,80%	99,70%	LMT + Info Gain (47 Características)
	0,60%	**95,10%**	**94,90%**	**94,80%**	**99,70%**	**LMT + Gain Ratio (31 Características)**

Al hacer la comparativa entre las dos mejores hibridaciones, para cada *dataset*, se pudo evidenciar que para el dataset Aruba CASAS - *raw* las dos combinaciones presentaron los mismos resultados en cuanto a las métricas *recall*, *F-Measure* y *ROC Area*, siendo LMT con *Gain Ratio* utilizando 24 características, la que presentó la más baja métrica *FP-Rate* con un valor de 0.5%.

En la evaluación del dataset dataset Aruba CASAS - *duration*, la combinación que presentó el mejor *recall* de 95.90% y un *ROC area* de 99.70%, fue la de LMT con *One R*, utilizando 33 características. En cuanto a la evaluación del dataset Aruba CASAS - *sensor based*, los resultados para las dos hibridaciones de técnicas de clasificación y selección utilizadas, coincidieron en los respectivos resultados de las métricas *precision*, *recall*, *F-Measure* y *ROC Area*. Pese a que LMT con *Gain Ratio* para 31 características, es la combinación que presentó el más alto FP Rate del 0.6%, es importante resaltar que la otra combinación (LMT con Info Gain), utiliza 16 características más, ver tabla 12. Hasta este punto se puede deducir que **el dataset que permite generar el mejor modelo predictivo es Aruba CASAS duration**, luego de aplicar la hibridación de técnicas LMT con One R, utilizando solo 33 características de las 49 características originales.

En este orden de ideas, estas 33 características son las que mejor inciden en el proceso clasificatorio con miras a la predicción de las diferentes actividades humanas, en la tabla 13 se indica la prioridad de incidencia en la predicción identificada a partir de la técnica de selección One R.

Tabla 13. Atributos de mayor incidencia en la clasificación de la técnica LMT identificados con la técnica de selección de características One R para el dataset Aruba CASAS – duration

ID	Atributo	Prioridad	ID	Atributo	Prioridad	ID	Atributo	Prioridad
1	M018	70,12	12	D004-close	51,97	23	M026	47,49
2	M019	69,94	13	D004-open	51,80	24	M004	47,43
3	M015	69,08	14	M030	51,49	25	T001	47,36
4	M009	68,30	15	M007	51,18	26	M022	47,34
5	M017	68,12	16	M003	50,89	27	M005	47,25
6	duracion	66,02	17	M020	50,78	28	M027	47,03
7	M016	65,04	18	M002	50,58	29	T005	46,96
8	M013	64,17	19	T003	48,58	30	M028	46,85
9	M014	59,61	20	M029	47,98	31	M008	46,52
10	M021	55,99	21	T004	47,65	32	M006	45,94
11	eventos	52,15	22	T002	47,63	33	M023	45,87

4.3 ESCENARIO EXPERIMENTAL No 3: ANÁLISIS COMPARATIVO DE LOS MEJORES RESULTADOS OBTENIDOS, APLICANDO VALIDACIÓN CRUZADA

En este escenario, se llevó a cabo una evaluación más exhaustiva, para valorar si existe una mejor combinación de técnicas de clasificación y selección de características, respecto al escenario anterior, para cada dataset (Aruba CASAS - *raw*, Aruba CASAS - *duration* y Aruba CASAS - *sensor based*). Cada conjunto de datos fue entrenado y probado utilizando validación cruzada con 10 *fold* (pliegues), generando tres experimentos, cuyos resultados se detallan en las tablas 14, 15 y 16.

En el proceso de validación cruzada, cada *dataset* completo fue dividido en 10 *fold* (pliegues) de igual tamaño. Luego se realizaron pruebas iterativas en las que el modelo se entrenó con 9 pliegues y se probó con el pliegue restante. Finalmente se promediaron las métricas de calidad obtenidas en cada una de las 10 iteraciones para calcular el resultado final.

Para la prueba con el dataset Aruba CASAS – *raw*, con técnica de clasificación LMT y selección de característica Gain Ratio (24 características), el r*ecall* fue del 94,10% (ver la tabla 14), lo cual no representa una mejora con respecto a la evaluación efectuada para este *dataset* con la misma combinación de técnicas, en el segundo escenario, donde el *recall* fue del 94,90% (ver tabla 12).

Tabla 14. Resultados clasificación LMT + Gain Ratio con validación cruzada con 10 folds para dataset Aruba CASAS - raw

Class	LMT + Gain Ratio (24 Características)				
	FP Rate	**Precision**	**Recall**	**F-Measure**	**ROC Area**
Sleeping	0,00%	99,20%	99,60%	99,40%	100,00%
Bed_to_Toilet	0,00%	98,20%	100,00%	99,10%	100,00%
Meal_Preparation	0,30%	99,20%	99,00%	99,10%	99,50%
Relax	0,80%	99,10%	99,40%	99,20%	99,70%
Housekeeping	0,10%	86,40%	82,60%	84,40%	96,40%
Eating	0,10%	98,30%	96,10%	97,20%	99,80%
Leave_Home	2,00%	65,20%	53,40%	58,70%	97,20%
Enter_Home	3,30%	63,20%	73,40%	67,90%	97,50%
Work	0,10%	97,50%	97,50%	97,50%	100,00%
Average	0,80%	94,10%	**94,10%**	94,10%	**99,30%**

Tabla 15. Resultados clasificación LMT + One R con validación cruzada con 10 folds para dataset Aruba CASAS - duration

Class	LMT + One R (33 Características)				
	FP Rate	**Precision**	**Recall**	**F-Measure**	**ROC Area**
Sleeping	0,00%	99,60%	99,60%	99,60%	99,80%
Bed_to_Toilet	0,00%	99,10%	100,00%	99,60%	100,00%
Meal_Preparation	0,40%	98,80%	98,60%	98,70%	99,60%
Relax	0,80%	99,10%	99,30%	99,20%	99,70%
Housekeeping	0,20%	66,70%	69,60%	68,10%	94,90%
Eating	0,10%	97,10%	93,40%	95,20%	97,70%
Leave_Home	2,80%	62,80%	66,80%	64,70%	97,10%
Enter_Home	2,30%	68,00%	63,30%	65,60%	97,20%
Work	0,20%	94,50%	99,20%	96,80%	100,00%
Average	0,80%	94,10%	**94,10%**	94,00%	**99,20%**

En la prueba con el dataset Aruba CASAS – *duration*, con técnica de clasificación LMT y selección de característica One R (33 características), el *recall* fue del 94,10% (ver tabla 15), lo cual no representan una mejora con respecto a la evaluación efectuada para este *dataset* con la misma combinación de técnicas, en el segundo escenario, donde el *recall* fue del 95,90% (ver tabla 12).

Tabla 16. Resultados clasificación LMT + Gain Ratio con validación cruzada con 10 folds para dataset Aruba CASAS - sensor based

Class	LMT + Gain Ratio (31 Características)				
	FP Rate	**Precision**	**Recall**	**F-Measure**	**ROC Area**
Sleeping	0,10%	98,90%	100,00%	99,40%	100,00%
Bed_to_Toilet	0,10%	97,40%	100,00%	98,70%	100,00%
Meal_Preparation	0,40%	98,80%	99,00%	98,90%	99,70%
Relax	0,70%	99,20%	99,20%	99,20%	99,40%
Housekeeping	0,10%	86,40%	82,60%	84,40%	90,70%
Eating	0,00%	99,40%	94,50%	96,90%	99,80%
Leave_Home	2,10%	64,90%	53,40%	58,60%	97,30%
Enter_Home	3,30%	63,10%	73,40%	67,80%	97,70%
Work	0,10%	95,90%	97,50%	96,70%	99,90%
Average	0,80%	94,00%	**94,00%**	93,90%	**99,30%**

En cuanto a la prueba con el *dataset* Aruba CASAS - *sensor based*, con técnica de clasificación LMT y selección de características *One R* (31 características), el *recall* fue del 94,00% (ver tabla 16), lo cual no representa una mejora con respecto a la evaluación efectuada para este *dataset* con la misma combinación de técnicas, en el segundo escenario, donde el *recall* fue del 94,90% (ver tabla 12).

Tabla 17. Comparativa entre las mejores hibridaciones de técnicas de clasificación y selección de características con validación cruzada para cada dataset

Dataset	Métricas de calidad					Hibridación Técnica de clasificación + Selección de características (Validación cruzada 10 fold)
	FP Rate	**Precision**	**Recall**	**F-Measure**	**ROC Area**	
Aruba CASAS – *raw*	0,80%	94,10%	**94,10%**	94,10%	99,30%	LMT + Gain Ratio (24 Características)
Aruba CASAS – *duration*	0,80%	94,10%	**94,10%**	94,00%	99,20%	LMT + One R (33 Características)
Aruba CASAS - *sensor based*	0,80%	94,00%	**94,00%**	93,90%	99,30%	LMT + Gain Ratio (31 Características)

Los resultados obtenidos en este tercer escenario de experimentación, en cuanto a las métricas *recall* y *ROC area* luego de aplicar validación cruzada, no representaron mejoras frente a los obtenidos en el segundo escenario. Este comportamiento se presentó en cada uno de los experimentos llevados a cabo con los *dataset* (Aruba CASAS - *raw*, Aruba CASAS - *duration* y Aruba CASAS - *sensor based*) debido al sobreajuste (*Overfitting*), Ver la tabla 17. El sobreajuste es el resultado de sobre-entrenar un modelo con datos ajustados a unas características específicas del *dataset*. Es decir, se genera un aprendizaje excesivo de algunos comportamientos singulares de la clase, a su vez, se imposibilita la comprensión de comportamientos diferentes de la etiqueta de clase, producto del desbalanceo del conjunto de datos de entrenamiento, según (Camaré, 2008).

Para comprobar si hay diferencias significativas entre los modelos propuestos, se llevó acabo un análisis estadístico a través del estudio de la varianza de cada uno de ellos por lo cual se plantea la hipótesis nula *Ho* que plantea la igualdad entre las medias de los modelos con un nivel de significancia alfa del 5% y una hipótesis alternativa *H1* que rechaza dicha igualdad.

Tabla 18. Análisis estadístico – ANOVA

Modelos	F	Probabilidad	Valor crítico para F
M1 (LMT + Gain Ratio 24 características) vs M2(LMT + One R 33 características)	0,058300716	0,812269355	4,493998478
M1 (LMT + Gain Ratio 24 características) vs M3(LMT + Info Gain 47 características)	0,034326866	0,855341542	4,493998478
M2(LMT + One R 33 características) vs M3(LMT + Info Gain 47 características)	0,00182054	0,966494302	4,493998478

En la tabla 18, se puede apreciar que los valores para la probabilidad en las tres comparaciones: M1 (LMT + Gain Ratio 24 características) vs M2(LMT + One R 33 características), M1 (LMT + Gain Ratio 24 características) vs M3(LMT + Info Gain 47 características) y M2(LMT + One R 33 características) vs M3(LMT + Info Gain 47 características) son mucho mayores que el nivel de significancia alfa del 5% por lo que se decide aceptar la hipótesis nula *Ho* que plantea la igualdad entre las medias de los modelos, lo cual indica que no hay diferencia significativa entre los tre modelos planteados, además, de la consistencia de los datos tenidos en cuenta para la experimentación.

5. CONCLUSIONES Y RECOMENDACIONES

En este capítulo se presentan las conclusiones a las cuales se ha llegado con el desarrollo de este trabajo de investigación, luego de evaluar cada escenario de experimentación planteado anteriormente, además se plasman los resultados obtenidos para cada experimento llevado a cabo en el transcurso de dicha investigación.

En el primer escenario, la métrica de calidad *recall* con un 95,60%, representa el mas alto resultado, cuando se evalúa el *dataset* Aruba CASAS – *duration*, utilizando 49 características, con las técnicas de clasificación J48 y JRIP. Superando a los resultados obtenidos luego de evaluar los *dataset* Aruba CASAS – *raw* y Aruba CASAS - *sensor based*, en los que se obtuvo un *recall* de 94,50%, en ambas evaluaciones. Esto indica, que haber agregado las dos características de conteo del número de eventos y duración de la actividad, mejoró porcentualmente en 1,1% la métrica del *recall*. Por otro lado, el dataset Aruba CASAS - *sensor based* (que tiene 20 características adicionales, calculadas mediante las funciones de agregación aplicadas a las carácterísticas, generadas a partir de los sensores de tremperatura, y que han sido calculadas mediante el agrupamiento de las instancias del *dataset* original, segmentado por clases - actividades), no generó ninguna mejora respecto al *dataset* Aruba CASAS – *duration*, por el contrario, los resultados evidencian un incremento en los tiempos computacionales para el proceso de clasificación.

En el segundo escenario, nuevamente el experimento que presenta el mejor resultado en cuanto a la métrica de calidad *recall*, es cuando se utiliza el *dataset* Aruba CASAS – *duration*. La hibridación de la técnica de clasificación LMT con la técnica de selección de características One R, utilizando 33 características, generó un *recall* del 95,90% frente al 95,80% logrado por JRIP y

One R utilizando 47 características. Adicionalmente, la hibridación de LMT y One R lograron una significativa reducción del 32,65% en el número de características (una reducción de 16 características) frente al apenas 4.08% (una reducción de dos características) logrado por la hibridación de JRIP y One R, lo cual va ha incidir directamente en una disminiución de los tiempos computacionales necesarios para la construcción y evaluación del modelo predictivo si se usa la combinación LMT con One R.

Por otra parte, en el segundo escenario, en cuanto a los experimentos con los *dataset* Aruba CASAS – raw y Aruba CASAS – sensor based, también hubo una reducción representativa del número de características. Específicamente, con la hibridación de técnicas LMT con *Gain Ratio*, utilizando 24 Características, para Aruba CASAS – *raw* y la hibridación de técnicas LMT con *Gain Ratio*, utilizando 31 Características, para Aruba CASAS – *sensor based*. Aunque la disminución, en cuanto al número de características, es del 48.94% y el 53,73%, respectivamente, la métrica del *recall* esta por debajo un 1,00% de la obtenida en el experimento para es el *dataset* Aruba CASAS – *duration*. Es importante resaltar que la técnica de clasificación que arrojó los mejores resultados, en cuanto a métricas de calidad, para cada uno de los experimentos con los tres *dataset*, fue la técnica LMT. Por otra parte, en la tabla 12, se presenta el ranking de las 33 características que mas inciden en el proceso de clasificación, según lo determinado por la técnica de selección de características One R.

En el tercer escenario, si bien se realizó una evaluación exhaustiva para la mejor hibridación de cada *dataset*, utilizando validación cruzada con 10 fold (pliegues), se evidenció un decremento en cuanto a la métrica del *recall* en cada uno de los experimentos con los tres dataset (Aruba CASAS - *raw*, Aruba CASAS - *duration* y Aruba CASAS - *sensor based*) a causa del sobre-entrenamiento (*overfitting*).

En cuanto a la métrica de calidad *Recall*, correspondiente a cada etiqueta de clase (actividad), en cada uno de los experimentos, se debe indicar, que la hibridación ganadora para el *dataset* Aruba CASAS – *duration*, a pesar de haber arrojado un bajo 83,50% para la actividad "salir de casa (*leave_home*)", ha logrado generar un 28,24% por encima del 55,60% alcanzado en ambos casos por las hibridaciones ganadoras de los de los dataset Aruba CASAS – raw y Aruba CASAS – *sensor based*, es posible que esto se deba a la inclusión de las 2 características adicionales de número de eventos y duración de la actividad, incluidas para el *dataset* Aruba CASAS – *duration*, dado que particularmente para dicha actividad el número de eventos (lecturas de los sensores) es muy bajo, ver tabla 18.

Por otra parte, el *Recall* para la actividad limpieza (*Housekeeping*), ha arrojado valores diferentes en el experimento con cada dataset. A pesar de haber logrado un 100,00% con el *dataset* Aruba CASAS – *raw*, su resultado con los otros dos dataset no fueron los mejores, con Aruba CASAS – *duration* fue del 77,80% y con Aruba CASAS – sensor based fue del 88,90%, la diferencia en los resultados obtenidos para el *Recall* en cada *dataset* se debió al bajo número de instancias de esta actividad frente a las demás, apenas 32 instancias de datos, ver la tabla 6.

Las mayores tasas de éxito en cuanto a métricas de calidad se obtuvieron al entrenar el modelo con el *dataset* Aruba CASAS – *duration*, con un 95,90% en el *recall*, indicando una alta proporción de casos positivos. Es decir, una alta detección de actividades que se identificaron correctamente y un 99,70% en *ROC área*, lo cual indica que es un modelo con muy alta calidad predictiva, ver tabla 18. Además, una muy baja tasa promedio de detección de falsos positivos *FP Rate* del 0,60%. También se alcanzó un promedio de precisión del 95,90% lo cual indica que hay una alta proporcionalidad entre el número de predicciones correctas tanto positivas como negativas y el total de predicciones. y un F-Measure del 95,80%, ver Tabla 18.

Tabla 19. Comparativa entre la mejor hibridación de técnicas de clasificación y selección de características con train y test para cada dataset

Class	**Aruba CASAS – raw (LMT + Gain Ratio 24 características)**		**Aruba CASAS – duration (LMT + One R 33 características)**		**Aruba CASAS - Sensor based (LMT + Gain Ratio 31 características)**	
	Recall	ROC Area	Recall	ROC Area	Recall	ROC Area
Sleeping	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Bed_to_Toilet	100,00%	100,00%	97,80%	100,00%	100,00%	100,00%
Meal_Preparation	98,50%	99,80%	98,60%	99,90%	98,30%	100,00%
Relax	99,40%	99,90%	99,70%	99,80%	99,80%	100,00%
Housekeeping	100,00%	100,00%	77,80%	100,00%	88,90%	100,00%
Eating	100,00%	100,00%	98,60%	98,80%	98,60%	100,00%
Leave_Home	55,60%	98,10%	83,50%	98,50%	55,60%	98,10%
Enter_Home	75,90%	97,90%	62,50%	98,20%	75,90%	97,90%
Work	100,00%	100,00%	96,00%	100,00%	100,00%	100,00%
Average	94,90%	99,70%	**95,90%**	**99,70%**	94,90%	99,70%

En consecuencia, el modelo propuesto en esta investigación, integra la técnica de clasificación LMT con la técnica de selección de características One R, usando solo 33 de las 49 características disponibles en el dataset Aruba CASAS - *duration* para el reconocimiento de las actividades humanas: preparación de comidas (*Meal_Preparation*), descansar (*Relax*), comer (*Eating*), trabajar (*Work*), dormir (*Sleeping),* ir de la cama al baño (*Bed_to_Toilet*), entrar a casa (*Enter_Home*), salir de casa (*Leave_Home*) y limpieza (*Housekeeping*), dichos datos fueron recolectados de un ambiente interior (*Indoor*) por el proyecto de casas inteligentes de *WSU - Washington State University.*

Finalmente, esta investigación hace dos importantes aportes al área de reconocimiento de las actividades humanas HAR, primeramente, el preprocesamiento del *dataset* a partir del conjunto de datos original Aruba CASAS provisto por el proyecto de casas inteligentes de *WSU* -

Washington State University. El cual se encuentra disponible en repositorio en línea con todos sus registros en crudo (*raw*). Por último, la identificación de la técnica de clasificación y selección de características que mejores métricas arrojan por criterio de clase a partir de la construcción de un modelo que evalúa dicho *dataset.*

REFERENCIAS

Aggarwal, J. K., & Ryoo, M. S. (2011). Human activity analysis: A review. *ACM Computing Surveys*, *43*(3). https://doi.org/10.1145/1922649.1922653

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-Based Learning Algorithms. *Machine Learning*, *6*(1), 37–66. https://doi.org/10.1023/A:1022689900470

Aluja, T. (2001). La minería de datos, entre la estadística y la inteligencia artificial. *Questiio*, *25*(3), 479–498. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-0035573389&partnerID=40&md5=2d59288d728bde451b4bf19d5855e4ba

Anderson, K. D., Bergés, M. E., Ocneanu, A., Benitez, D., & Moura, J. M. F. (2012). Event detection for Non Intrusive load monitoring. *IECON 2012 - 38th Annual Conference on IEEE Industrial Electronics Society*, 3312–3317. https://doi.org/10.1109/IECON.2012.6389367

Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-Ortiz, J. L. (2013). A public domain dataset for human activity recognition using smartphones. *ESANN 2013 Proceedings, 21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, (April), 437–442.

Aprende Machine Learning - Qué es overfitting y underfitting y cómo solucionarlo. (2017). Retrieved from https://www.aprendemachinelearning.com/que-es-overfitting-y-underfitting-y-como-solucionarlo/

Berges Gonzalez, M. E. (2010). *A Framework for Enabling Energy-Aware Facilities through Minimally-Intrusive Approaches*. Carnegie Mellon University, USA.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/bf00058655

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Camaré, L. J. M. (2008). *Aprendizaje Automático a partir de Conjuntos de Datos No Balanceados y su Aplicación en el Diagnóstico y Pronóstico Médico*.

Cessie, S. Le, & Houwelingen, J. C. Van. (1992). Ridge Estimators in Logistic Regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *41*(1), 191–201. Retrieved from http://www.jstor.org/stable/2347628

Chen, L., Hoey, J., Nugent, C. D., Cook, D. J., Yu, Z., & Member, S. (2012). *Sensor-Based Activity Recognition*. *42*(6), 790–808.

Chen, L., & Nugent, C. (2009). Ontology-based activity recognition in intelligent pervasive environments. *International Journal of Web Information Systems*.

Cleary, J. G., & Trigg, L. E. (1995). An Instance-based Learner Using an Entropic Distance Measure. *Elsevier*, *5*, 1–14. https://doi.org/10.1016/B978-1-55860-377-6.50022-0

Cohen, W. W. (1995). Fast Effective Rule Induction. *Differences*.

Cook, D. J. (2012). Learning setting-generalized activity models for smart spaces. *IEEE Intelligent Systems*, *27*(1), 32–38. https://doi.org/10.1109/MIS.2010.112

Cook, D. J., Crandall, A. S., Thomas, B. L., & Krishnan, N. C. (2013). CASAS: A smart home in a box. *Computer*, *46*(7), 62–69. https://doi.org/10.1109/MC.2012.328

De-La-Hoz-Franco, E., Ariza-Colpas, P., Quero, J. M., & Espinilla, M. (2018). Sensor-based datasets for human activity recognition - A systematic review of literature. *IEEE Access*, *6*, 59192–59210. https://doi.org/10.1109/ACCESS.2018.2873502

Detours, V., Dumont, J. E., Bersini, H., & Maenhaut, C. (2003). Integration and cross-validation of high-throughput gene expression data: Comparing heterogeneous data sets. *FEBS Letters*,

546(1), 98–102. https://doi.org/10.1016/S0014-5793(03)00522-2

Eibe, F., Holmes, G., & Witten, I. H. (2007). *Weka 3 - Data Mining with Open Source Machine Learning Software in Java*. Retrieved from https://www.cs.waikato.ac.nz/ml/weka/

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37–53. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-0002283033&partnerID=40&md5=266faf7bded790e22bc3754ab7e2caa1

Frank, E., Hall, M., & Pfahringer, B. (2003). *Locally Weighted Naive Bayes*. 249–256. Retrieved from http://arxiv.org/abs/1212.2487

Frank, E., Wang, Y., Inglis, S., Holmes, G., & Witten, I. H. (1998). Using model trees for classification. *Machine Learning*, *32*(1), 63–76. https://doi.org/10.1023/A:1007421302149

Frank, E., & Witten, I. H. (1998). Generating accurate rule sets without global optimization. *Proceedings of the Fifteenth International Conference on Machine Learning*, 144–151. https://doi.org/1-55860-556-8

Freund, Y., & Schapire, R. E. (1996). Experiments with a New Boosting Algorithm. *Proceedings of the 13th International Conference on Machine Learning*, 148–156. https://doi.org/10.1.1.133.1040

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *The Annals of Statistics*, *28*(2), 337–407. https://doi.org/10.1214/aos/1016218223

Fürnkranz, J., & Widmer, G. (1996). Incremental Reduced Error Pruning. *Machine Learning Proceedings 1994*, (January), 70–77. https://doi.org/10.1016/b978-1-55860-335-6.50017-9

García, J. A. (2016). Líneas de investigación en minería de datos en aplicaciones en ciencia e

ingeniería: Estado del arte y perspectivas. *Arxiv, Artificial Intelligence (Cs.AI)*, *1*(1609.05401), 1–17. https://doi.org/10.1007/s003350010211

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*(Mar), 1157–1182.

Hall, M. A., & Holmes, G. (2003). Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 15, pp. 1437–1447. https://doi.org/10.1109/TKDE.2003.1245283

Herrera, F., & Cano, J. R. (2006). Técnicas de reducción de datos en KDD. El uso de Algoritmos Evolutivos para la Selección de Instancias. *Actas Del I Seminario Sobre Sistemas Inteligentes (SSI'06), Universidad Rey Juan Carlos, Madrid (Spain).*, 165–181.

Holte, R. C. (1993). Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. *Machine Learning*, *11*, 63–91. https://doi.org/10.1023/A:1022631118932

Hota, H. S., & Shrivas, A. K. (2014). Decision tree techniques applied on NSL-KDD data and its comparison with various feature selection techniques. In *Advanced Computing, Networking and Informatics-Volume 1* (pp. 205–211). https://doi.org/http://doi.org/10.1007/978-3-319-07353-8

KDnuggets. (2014). *What main methodology are you using for your analytics, data mining, or data science projects? Poll*. Retrieved from https://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html

Kim, Won and Choi, Byoung-Ju and Hong, Eui and Kim, Soo-Kyung and Lee, D. (2003). A Taxonomy of Dirty Data. *Data Min. Knowl. Discov.*, *7*, 81–99. https://doi.org/10.1023/A:1021564703268

Kira, K., & Rendell, L. A. (1992). The Feature Selection Problem: Traditional Methods and a New Algorithm. *Proceedings of the Tenth National Conference on Artificial Intelligence*, 129–134. AAAI Press.

Kittler, J., Hatef, M., Duin, R. P. W., & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(3), 226–239. https://doi.org/10.1109/34.667881

Kohavi, R. (1995). *Wrappers for performance enhancement and obvious decision graphs*. (November). Retrieved from https://dl.acm.org/citation.cfm?id=241090

Kohavi, Ron, & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1), 273–324. https://doi.org/https://doi.org/10.1016/S0004-3702(97)00043-X

Kohavi, Ron, & Provost, F. (1998). Glossary of Terms. *Machine Learning*, *2*, 271–274. https://doi.org/10.1023/A:1017181826899

Kwon, B., Kim, J., & Lee, S. (2017). An enhanced multi-view human action recognition system for virtual training simulator. *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2016*, 1–4. https://doi.org/10.1109/APSIPA.2016.7820895

Landwehr, N., Hall, M., & Frank, E. (2003). Logistic model trees. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, *2837*, 241–252. https://doi.org/10.1007/s10994-005-0466-3

Lara, Ó. D., & Labrador, M. A. (2013). A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys and Tutorials*, *15*(3), 1192–1209. https://doi.org/10.1109/SURV.2012.110112.00192

Li, C., Lin, M., Yang, L. T., & Ding, C. (2014). Integrating the enriched feature with machine

learning algorithms for human movement and fall detection. *The Journal of Supercomputing*, *67*(3), 854–865. https://doi.org/https://doi.org/10.1007/s11227-013-1056-y

Li, R., Lu, B., & McDonald-Maier, K. D. (2015). Cognitive assisted living ambient system: a survey. *Digital Communications and Networks*, *1*(4), 229–252. https://doi.org/10.1016/j.dcan.2015.10.003

Lin, T. Y. (2002). Attribute transformations for data mining I: Theoretical explorations. *International Journal of Intelligent Systems*, *17*(2), 213–222.

Liu, H., & Motoda, H. (2012). *Feature selection for knowledge discovery and data mining* (Vol. 454). Springer Science & Business Media.

Liu, H., & Motoda, H. (2013). *Instance selection and construction for data mining* (Vol. 608). Springer Science & Business Media.

Liu, H., Motoda, H., Setiono, R., & Zhao, Z. (2010). Feature Selection : An Ever Evolving Frontier in Data Mining. *Journal of Machine Learning Research: Workshop and Conference Proceedings 10: The Fourth Workshop on Feature Selection in Data Mining*, 4–13.

Marks Hall, G. H. (1994). *WEKA: Practical Machine Learning Tools and Techniques with JAva Implementations*. Retrieved from https://researchcommons.waikato.ac.nz/bitstream/handle/10289/1040/uow-cs-wp-1999-11.pdf?sequence=1&isAllowed=y

Memon, M., Wagner, S. R., Pedersen, C. F., Aysha Beevi, F. H., & Hansen, F. O. (2014). Ambient Assisted Living healthcare frameworks, platforms, standards, and quality attributes. *Sensors (Switzerland)*, *14*(3), 4312–4341. https://doi.org/10.3390/s140304312

Milley, A. H., Seabolt, J. D., & Williams, J. S. (1998). *Data Mining and the Case for Sampling. A SAS Institute Best Practices*. 1–36. Retrieved from

http://sceweb.uhcl.edu/boetticher/ML_DataMining/SAS-SEMMA.pdf

Ministerio de Salud y Protección Social. (2017). Boletín de salud mental - Demencia. Retrieved from Ministerio de Salud website: https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/PP/ENT/boletin-depresion-marzo-2017.pdf

Mitra, S., & Acharya, T. (2003). Data Mining: Multimedia, Soft Computing, and Bioinformatics. In *Technometrics* (Vol. 46). https://doi.org/10.1198/tech.2004.s207

Moine, J. Mi., Haedo, A., & Gordillo, S. (2011). Estudio comparativo de metodologías para minería de datos. *XIII Workshop de Investigadores En Ciencias de La Computación*, 278–281. Retrieved from http://sedici.unlp.edu.ar/handle/10915/20034

Pete, C., Julian, C., Randy, K., Thomas, K., Thomas, R., Colin, S., & Wirth, R. (2000). Crisp-Dm 1.0. *CRISP-DM Consortium*, 76.

PRADENA, P. C. P. A. (2013). *VISUALIZACIÓN ESPACIO/TEMPORAL DE EVENTOS NOTICIOSOS* (UNIVERSIDAD DE CHILE). https://doi.org/10.1787/9789264197565-3-es

Provost, F., & Fawcett, T. (2001). Robust Classification for Imprecise Environments. *Machine Learning*, *42*, 203–231. https://doi.org/10.1023/A:1007601015854

Pyle, D. (1999). *Data preparation for data mining*. Morgan Kaufmann.

Quinlan, J. R. (1994). C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. In *Machine Learning* (Vol. 16). https://doi.org/10.1007/BF00993309

Read, J., Puurula, A., & Bifet, A. (2015). Multi-label Classification with Meta-Labels. *Proceedings - IEEE International Conference on Data Mining, ICDM*, *2015-Janua*(January), 941–946. https://doi.org/10.1109/ICDM.2014.38

Reed, K. L., & Sanderson, S. N. (1999). *Concepts of occupational therapy*. Retrieved from https://books.google.com.co/books?hl=es&lr=&id=1ZE47g_IRTwC&oi=fnd&pg=PR7&dq=Concepts+of+Occupational+Therapy.&ots=sMksfVhmYK&sig=wlabmL9W01HtUuzpARaj6BUDtHI#v=onepage&q=Concepts of Occupational Therapy.&f=false

Rice, J. A. (2006). *Mathematical statistics and data analysis*. Cengage Learning.

Robnik-Šikonja, M., & Kononenko, I. (1997). An adaptation of {R}elief for attribute estimation in regression. *Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97)*, *5*, 296–304. Retrieved from http://dl.acm.org/citation.cfm?id=645526.657141

Shahi, A., Woodford, B. J., & Lin, H. (2017). Dynamic real-time segmentation and recognition of activities using a multi-feature windowing approach. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 26–38. https://doi.org/https://doi.org/10.1007/978-3-319-67274-8_3

Shaltout, N., Elhefnawi, M., Rafea, A., & Moustafa, A. (2014). Information Gain as a Feature Selection Method for the Efficient Classification of Influenza Based on Viral Hosts. *Lecture Notes in Engineering and Computer Science*, *1*, 625–631.

Singla, G., Cook, D. J., & Schmitter-Edgecombe, M. (2010). Recognizing independent and joint activities among multiple residents in smart environments. *Journal of Ambient Intelligence and Humanized Computing*, *1*(1), 57–63. https://doi.org/10.1007/s12652-009-0007-1

T.K. Ho. (1998). The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(8), 832–844. https://doi.org/10.1109/34.709601

U.S. National Library of Medicine. (2019). Enfermedades neurodegenerativas: MedlinePlus en español. Retrieved January 13, 2020, from Medlineplus website:

https://medlineplus.gov/spanish/degenerativenervediseases.html

Van Der Malsburg, C. (1986). Frank Rosenblatt: Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. *Brain Theory*, (February), 245–248. https://doi.org/10.1007/978-3-642-70911-1_20

Van Kasteren, T. L. M., Englebienne, G., & Kröse, B. J. A. (2010). Activity recognition using semi-Markov models on real world smart home datasets. *Journal of Ambient Intelligence and Smart Environments*, *2*(3), 311–325. https://doi.org/10.3233/AIS-2010-0070

Weiss, S. M., & Kulikowski, C. A. (1991). *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems*. Morgan Kaufmann Publishers Inc.

Witten, I. H., Frank, E., & Hall, M. a. (2011). Data Mining: Practical Machine Learning Tools and Techniques. In *Complementary literature None*. Retrieved from http://books.google.com/books?id=bDtLM8CODsQC&pgis=1

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

World Health Organization. (2019). Dementia. Retrieved January 13, 2020, from https://www.who.int/news-room/fact-sheets/detail/dementia

Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, *17*(5–6), 375–381. https://doi.org/10.1080/713827180