Encuentro Internacional de Educación en Ingeniería ACOFI
2º Congreso Latinoamericano de Ingeniería
2019 10 al 13 de septiembre - Cartagena de Indias, Colombia
RETOS EN LA FORMACIÓN DE INGENIEROS EN LA ERA DIGITAL

APLICACIÓN DE MÉTODOS DE SELECCIÓN DE ATRIBUTOS PARA DETERMINAR FACTORES RELEVANTES EN LA CLASIFICACIÓN DE ACTIVIDADES FÍSICAS COTIDIANAS

Yesica Lorena Zúñiga Mamián, Kevin Felipe Meneses Palta, Néstor Iván Martínez Cobo, Sandra Patricia Castillo Landínez

Corporación Universitaria Autónoma del Cauca
Popayán, Colombia

Resumen

En la vida moderna el sedentarismo o la falta de actividad física se considera una de las principales causas asociadas a enfermedades tales como obesidad, hipertensión arterial, accidentes cerebrovasculares y diabetes. Por lo anterior, en el diario vivir los diferentes tipos de movimientos se consideran un indicador de salud física y mental.

Existen múltiples técnicas para la medición de este indicador, como encuestas u otras basadas en dispositivos electrónicos; las primeras pueden tener tendencias generadas por la inexperiencia del encuestador, los segundos, pueden requerir de ambientes controlados como los sistemas optoelectrónicos de captura de movimiento o el entendimiento de complejos procesos matemáticos. Ante este panorama, los sistemas de medición inercial constituyen una buena alternativa para recopilar una cantidad de datos suficiente para desarrollar estudios, sin generar costos excesivos y midiendo la actividad física del participante en su ambiente cotidiano. Típicamente las unidades inerciales están conformadas por acelerómetros, giroscopios y magnetómetros. Otro factor importante que incide en la medición de la actividad física es su intensidad, la cual debe ser suficientemente alta de forma tal que permita un buen estado físico, pero sin llegar al extremo de generar daños en huesos y músculos.

Este trabajo presenta los resultados del análisis exploratorio del conjunto de datos publicados por Kerem Altun y Billur Barshan, los cuales son base para el estudio de diferentes actividades físicas cotidianas simples como sentarse, pararse, subir y bajar escaleras, y otras no tan frecuentes como

el canotaje o caminar en un parqueadero. Además, se comparan estos resultados con los métodos de selección de atributos disponibles en la herramienta WEKA.

La identificación de atributos relevantes es útil al momento de monitorear diferentes actividades físicas ya que puede llevar a un conocimiento más profundo de factores que afectan la salud de personas que tienen problemas o dificultades relacionadas con la motricidad. Los datos generados a partir de mediciones con unidades inerciales son útiles en la medida en que se pueden obtener patrones de movimiento que permitan el diagnóstico preventivo en problemas relacionados con la salud, haciendo uso de técnicas de minería de datos y algoritmos de clasificación, entre otros.

Palabras clave: actividad física; análisis exploratorio; atributos relevantes

Abstract

In modern life sedentarism or lack of physical activity is cosidered one of the main causes asociated to deseases like obesity, arterial hypertension, cerebrovascular accidents and diabetes. Therefore, in everyday life diferent type of movement are considerated an indicator of mental and physical health.

There are multiples techniques for measuring this indicator, as surveys or others based in electronic devices; the first ones may have tendencies generated for the lack of experience of the pollster and the second ones may require controlled environments like optoelectronic motion capture systems or the understanding of complex mathematical processes. In this context, inertial measurement systems are a good alternative to collect enough amount of data to develop studies, without generating excessive costs and measuring the physical activity of the participant in its everyday environment. Typically, inertial units are conformed by accelerometers, gyroscopes and magnetometers. Another important factor that affects the measuring of physical activity is its intensity, which has to be enough high to have a good physical condition without reaching the extreme of generating damage to bones and muscles.

This paper presents the results of the exploratory analysis of the dataset published by Kerem Altun and Billur Barshan, which are the basis for the study of different daily physical activities such as sitting, standing, going up and down stairs, and others not so frequent as boating or walking in a parking lot. Also, this results are compared with attribute selection methods available in software WEKA.

The identification of relevant attributes is useful when monitoring different physical activities since it can lead to a deeper knowledge of factors that affect the health of people that have problems or deseases related to motor skills. The data generated from inertial units measurements are useful insofar as movement patterns that allow the preventive diagnostic of issues related to health can be obtained, using data mining techniques and classification algorithms, among others.

Keywords*: physical activity; exploratory analysis; relevant attributes*

1. Introducción

Algunas patologías como son los ataques cardíacos, enfermedades cerebrovasculares, el aumento de obesidad, la diabetes e incluso la muerte temprana están cada vez más relacionados con el estilo de vida poco saludable que lleva cada individuo, básicamente continúan siendo causados en gran medida por la ausencia de actividad física (Cardiaco, Atlas, & Northwell, 2019).

Según la Organización Mundial de la Salud (OMS), a nivel mundial en el año 2010, más del 80% de los adolescentes de 11 a 17 años de edad escolar no eran suficientemente activos físicamente y en el 2016, el 28% de los adultos mayores de 18 años tampoco lo eran (OMS, 2019); también destaca que las enfermedades cardiovasculares se encuentran en el top número 1 de muertes por esta causa en todo el mundo y establecieron que en 2015 murieron 17,7 millones de personas, lo cual representa un 31% de todas las muertes registradas en el mundo por esta causa (OMS, 2017).

De cara a este flagelo, la OMS informa que un cuarto de la población mundial (27,5%), equivalente a 1.400 millones de personas, tienen su salud en riesgo por no realizar un mínimo de 150 minutos de actividad física por semana (Salas, 2018), e inclusive mencionan que el sedentarismo es la principal causa de otros padecimientos como los cánceres de mama y colon (entre el 21% y 25%), el 27% de los casos de diabetes; y el 30% de las cardiopatías isquémicas (OMS, 2013).

En este contexto, se hace necesario medir y analizar los movimientos involucrados en las diferentes actividades físicas cotidianas, acciones importantes para el monitoreo de personas con dificultades o enfermedades relacionadas con la motricidad, deportistas y adultos de edad avanzada. Estas mediciones resultan fundamentales para establecer acciones preventivas y correctivas en áreas como la medicina, la fisiología y el deporte (Altun, Barshan, & Tunçel, 2010).

El volumen de datos obtenidos con un sistema de medición inercial (Mannini & Sabatini, 2011) está fuera del alcance de los métodos de procesamiento tradicionales, por lo que se hace necesario utilizar técnicas y algoritmos de minería de datos, que apoyen la construcción de modelos como clasificadores y predictores.

2. Minería de datos y selección de atributos

La minería de datos surgió como una opción para aprovechar la gran cantidad de información que tienen almacenada las organizaciones, resultado del desarrollo de sus actividades; el procesamiento de una gran cantidad de datos se asemeja a una mina de la cual se extraen minerales o metales valiosos, para este caso se busca obtener modelos que apoyen la toma de decisiones y permitan la generación de nuevo conocimiento mediante la identificación de patrones (Han & Kamber, 2006). Es necesario mencionar que la minería de datos hace parte del proceso KDD (Knowledge Discovery in Databases), concebido como una metodología que consta de varias fases (Figura 1), cuyo objetivo es generar conocimiento que resulta potencialmente útil para una problemática específica. Este proceso resulta complejo ya que los volúmenes de datos que se procesan tienden a ser bastantes robustos (Landa, 2018).

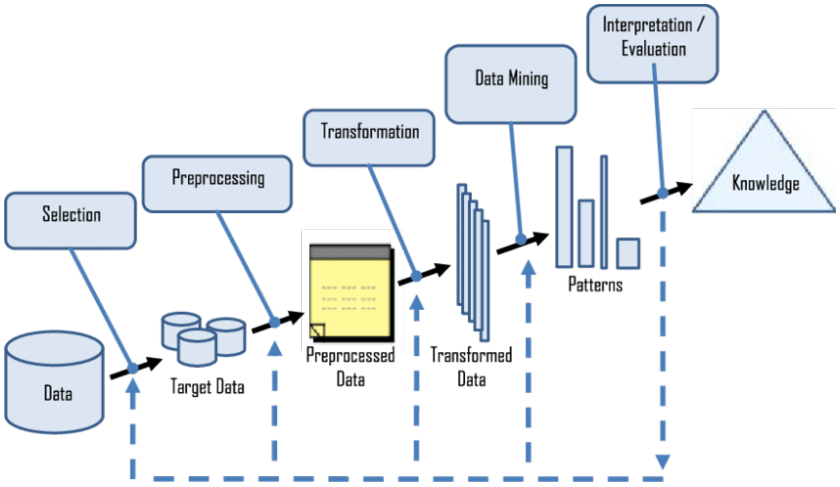

Figura 1. Etapas del proceso KDD. Fuente (Fayyad, Piatetsky-shapiro & Smyth,1996)

Como parte de la etapa de preprocesamiento del dataset, es necesario realizar la selección de atributos, un método usado para eliminar datos no relevantes o redundantes y de esta manera reducir la dimensionalidad del dataset (Qinbao Song, Jingjie Ni, & Guangtao Wang, 2011), esto se realiza con el fin de reducir la carga computacional de los algoritmos de procesamiento.

La herramienta Weka implementa varios algoritmos de selección de atributos que se clasifican en dos tipos: métodos contenedores y filtros. Los primeros usan clasificadores para determinar la importancia de los atributos y por lo general son mejores que los filtros, debido a la optimización que los algoritmos de aprendizaje hacen al proceso de selección, pero con la desventaja que requieren más tiempo de procesamiento (Karegowda, Manjunath, & Jayaram, 2010). Por su parte, los filtros usan las características generales de los datos (su relación con la clase o su grado de entropía en un modelo) para evaluar los atributos, haciéndolos más rápidos y escalables (Hall & Holmes, 2002).

Para este trabajo se utilizaron tres métodos de selección de atributos: dos filtros (Selección de Basada en Correlación y Ganancia de Información) y un contenedor (OneR junto con el algoritmo J48 de árboles de decisión); cada uno de ellos fue aplicado junto a un método de búsqueda como BestFirst o Ranker, que ayudan a realizar las combinaciones en cada grupo de atributos evaluados por el método de selección.

La Selección de Atributos Basada en Correlación (CFS) evalúa la capacidad de predicción de varios conjuntos de atributos eligiendo aquellos que se encuentran más relacionados con la actividad física y menos relacionados entre sí para evitar o minimizar la redundancia (Pandey, Pandey, Jaiswal, & Sen, 2013).

La Ganancia de Información (InformationGain) mide la importancia de un atributo a partir de la entropía de la información que presenta un modelo, cuando dicho atributo está presente o ausente, lo anterior significa, que entre menos entropía mejor es el atributo (Paul, 2005).

Por último, el evaluador de atributos OneR utiliza una metodología para generar árboles de decisión expresados en forma de un conjunto de reglas que ponen a prueba un atributo en particular respecto a la clase para predecir su calidad (Liu, 2009).

3. Materiales y métodos

Para realizar este trabajo se utilizaron los datos publicados por Kerem Altum y Billur Barshan, y obtenidos del repositorio de la página web de la IEEE(https://ieee-dataport.org/documents/daily-and-sports-activities-data-set), contempla 19 actividades físicas habituales tales como sentarse, pararse, subir y bajar escaleras y otras no tan cotidianas como canotaje o caminar en un parqueadero. Cada actividad fue ejecutada por 8 personas, cada una de ellas fue medida (censada) durante 5 minutos (éstos fueron divididos en segmentos de 5 segundos para un total: 60 segmentos). La tasa de muestreo fue de 25 Hz y se usaron 5 sensores (cada uno captura 9 variables) ubicados así: torso (T), brazo derecho (RA) e izquierdo (LA) y pierna derecha (RL) e izquierda (LL). De acuerdo a lo anterior, se tenía un dataset de 1.140.000 registros y 45 variables, las cuales, representan datos como aceleración, velocidad angular y campo magnético que fueron obtenidos a través del acelerómetro, magnetómetro y giroscopio de forma tridimensional.

Se utilizaron herramientas como Weka, MatLab y R para realizar el preprocesamiento de los datos, que incluyó actividades como integración de los datos, y generación de nuevas variables utilizando métodos estadísticos. Se optó por reducir el número de registros y ampliar la cantidad de variables: el dataset final tenía las siguientes características:

- 9120 filas: valores de las medidas de 60 registros por actividad por participante.
- 228 columnas: el original tenía 45 variables y para cada una se calculó: media, mediana, IQR, rango y desviación estándar.

La metodología usada para el desarrollo del proyecto consta de tres fases:

- Fase 1: alcance del problema. Se indagó sobre la problemática relacionada con patologías causadas por el sedentarismo, importancia y formas de medir el movimiento en actividades físicas humanas; se desarrollan las siguientes actividades:
  - Revisión bibliográfica.
  - Valoración de la situación.

- Fase 2: preparación de los datos. Después de obtener los datos se realizaron las actividades orientadas a conseguir un dataset con características adecuadas para generar nuevo conocimiento:
  - Recolección de datos iniciales.
  - Integración los datos.
  - Limpieza de los datos.
  - Generación de nuevas variables.

- Fase 3: análisis exploratorio de los datos. Se emplearon técnicas analíticas con el propósito de entender la naturaleza de los datos y establecer las relaciones existentes, las actividades desarrolladas fueron:
  - Identificación de relaciones entre variables por medio gráficas comparativas de caja y bigotes (boxplot).
  - Determinación de atributos relevantes por medio evaluadores de atributos y métodos de búsqueda.
  - Análisis de los resultados.

4. Resultados

- **Identificación de variables relevantes a partir de diagramas Boxplot**

Del total de 225 variables se establecieron 20 como relevantes; el criterio considerado fue la posibilidad de diferenciar una actividad de las demás. En las Figuras 2 y 3 se presentan los diagramas de caja y bigotes de la variable mediana de la aceleración en el eje z de los sensores ubicados en la pierna derecha e izquierda; se observa que la actividad 4 (girarse hacia la derecha estando acostado) se distingue del resto, el resultado sugiere que los sensores ubicados en ambas piernas captan información destacada de esta postura; por tanto, estas dos variables se constituyen como relevantes.

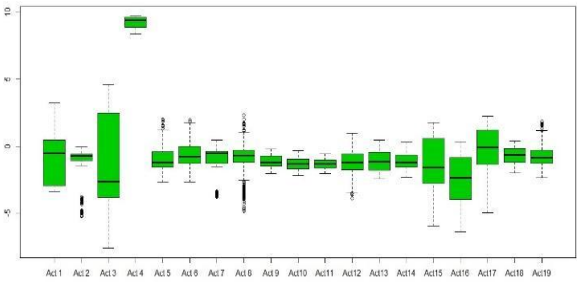

Figura 2. Mediana de la aceleración en el eje z del sensor ubicado en la pierna izquierda. Fuente propia.

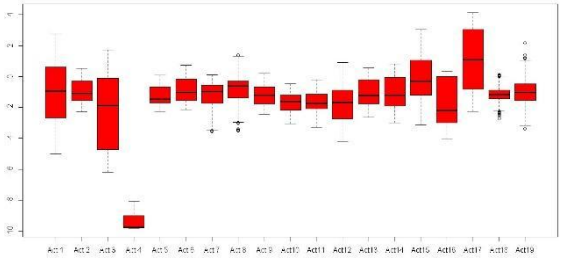

Figura 3. Mediana de la aceleración en el eje z del sensor ubicado en la pierna derecha. Fuente propia

En la Figura 4, se observa que las cajas que representan la actividad 4 mencionada anteriormente y la actividad 3 (acostarse sobre la espalda) se encuentran aisladas del resto; lo anterior indica que el sensor ubicado en el torso es contiene información relevante sobre la actividad.

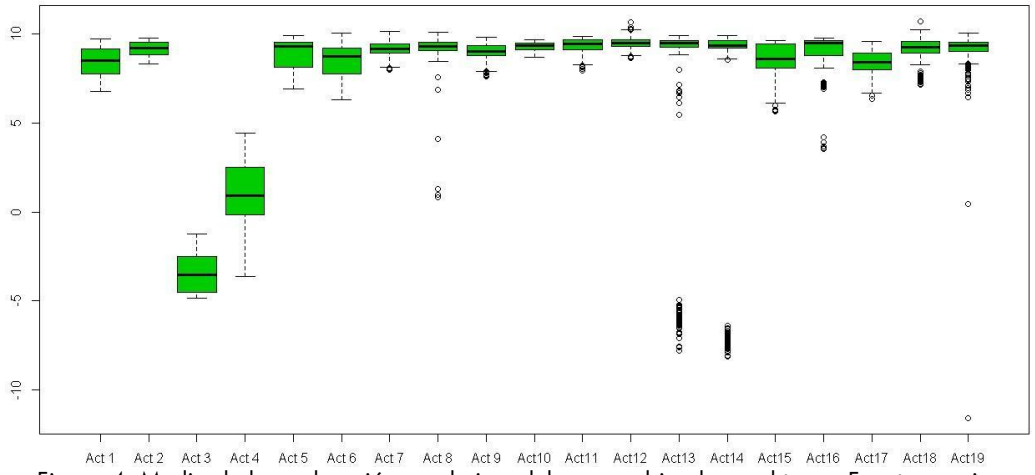

Figura 4. Media de la aceleración en el eje x del sensor ubicado en el torso. Fuente propia.

Bajo las mismas consideraciones se generaron otras gráficas que permitieron diferenciar actividades tales como ciclismo en posición horizontal, vertical y canotaje (actividades 15, 16 y 17), en las cuales es fundamental el uso de las dos piernas en su ejecución y esto es corroborados por la información de los sensores.

- **Identificación de variables relevantes a partir de algoritmos de selección**

Se realizaron diferentes pruebas con evaluadores y métodos de búsqueda, que permitieran validar los resultados obtenidos previamente.

En la tabla 1 se comparan las variables relevantes identificadas con cada uno de los métodos usados. En el caso de OneR y Ganancia de Información se empleó el método Ranker, que posiciona cada atributo de acuerdo a la calificación que le da el evaluador, y en la Selección de Atributos Basada en Correlación se utilizó BestFirst que muestra la cantidad de veces que aparece el atributo para un número determinado de iteraciones.

Método Gráfico Boxplot	Selección de Atributos Basada en Correlación	OneR	Ganancia de Información
T_zacc_median	RL_xacc_mean	LL_xacc_mean	LL_xacc_mean
T_xacc_mean	RL_yacc_mean	LL_xmag_mean	LL_xmag_mean
RL_yacc_mean	LL_xacc_mean	RL_xacc_mean	RL_xacc_mean
RL_yacc_iqr	RL_yacc_median	LL_xmag_median	LL_xmag_median
RL_xmag_median	RL_xmag_median	RL_yacc_median	LL_xacc_median
RL_xacc_mean	LL_xacc_median	LL_xacc_median	T_xacc_std
LL_zacc_median	LL_yacc_median	RL_xacc_median	RL_zmag_median
LL_yacc_iqr	RL_yacc_iqr	RL_xmag_mean	LA_xacc_std
LL_xmag_median	LL_yacc_iqr	RL_yacc_mean	RL_xmag_mean
LL_xacc_mean	RL_zmag_median	RL_zmag_median	RL_zmag_mean

Tabla 1. Variables relevantes identificadas por diferentes métodos. Fuente propia

Se observa coincidencia en algunos resultados, por ejemplo, la *"mediana del campo magnético en el eje z del sensor ubicado en la pierna derecha (RL_zmag_median)"* fue identificada por los tres evaluadores; la variable *"media de la aceleración en el eje x del sensor ubicado en la pierna derecha (RL_xacc_mean)"* fue seleccionada por todos los métodos.

Cabe mencionar que algunas variables fueron registradas únicamente por un método:

- Variable *"desviación estándar de la aceleración en el eje x del sensor ubicado en el brazo izquierdo (LA_xacc_std)"* seleccionada por el evaluador de Ganancia de Información
- Variable *"media del campo magnético en el eje x del sensor ubicado en la pierna derecha (RL_xmag_mean)"* escogida por el algoritmo OneR.

5. Conclusiones

- Este trabajo permitió identificar la ubicación de los sensores que son determinantes en la ejecución de las diferentes actividades físicas cotidianas.
- Se logró determinar que el movimiento en los brazos y la velocidad angular presentan una baja participación en las actividades físicas diarias y no son tan relevantes para su identificación.
- Los métodos de selección de atributos en Weka son útiles para validar y establecer atributos relevantes para la construcción de clasificadores y otros modelos de minería de datos, lo cual hace parte del trabajo futuro.
- En la próxima fase de este proyecto se busca determinar las relaciones estadísticas entre las diferentes variables y con todos estos resultados establecer si existe la necesidad de generar nuevas variables que permitan la clasificación de actividades a partir de información inercial.

6. Referencias

Artículos de revistas

- Altun, K., Barshan, B., & Tunçel, O. (2010). Comparative study on classifying human activities with miniature inertial and magnetic sensors. Pattern Recognition, 43(10), 3605–3620. https://doi.org/10.1016/j.patcog.2010.04.019
- Fayyad, U., Piatetsky-shapiro, G. & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Mag, 3, 37–54
- Hall, M., & Holmes, G. (2002). Benchmarking Attribute Selection Techniques for Discrete Class Data Mining, (April).
- Karegowda, A. G., Manjunath, A. S., & Jayaram, M. A. (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection. International Journal of Information Technology and Knowledge and Knowledge Management, 2(2), 271–27
- Mannini, A., & Sabatini, A. M. (2011). Classification of human physical activities from on-body accelerometers - A Markov Modeling Approach, 201–208. https://doi.org/10.5220/0003151102010208
- Pandey, A. K., Pandey, P., Jaiswal, K. L., & Sen, A. K. (2013). DataMining Clustering Techniques in the Prediction of Heart Disease using Attribute Selection Method, 2(10), 2003–2008.

- Paul, J. P. (2005). Feature selection methods and algorithms. British Journal of Plastic Surgery, 31(2), 181. https://doi.org/10.1016/s0007-1226(78)90078-4
- Qinbao Song, Jingjie Ni, & Guangtao Wang. (2011). A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data. IEEE Transactions on Knowledge and Data Engineering, 25(1), 1–14. https://doi.org/10.1109/tkde.2011.1817.

Libros

- Han, J., & Kamber, M. (2006). Data Mining: Concepts and Techniques (2nd edition). Soft Computing (Vol. 54). https://doi.org/10.1007/978-3-642-19721-5
- Liu, X. (2009). Linking Competence to Opportunities to Learn. Linking Competence to Opportunities to Learn. Springer. https://doi.org/10.1007/978-1-4020-9911-3

Fuentes electrónicas

- Cardiaco, H., Atlas, S., & Northwell, B. De. (2019). ¿Desea vivir más? Simplemente pase un poco menos de tiempo sentado cada día, 15-16. Recuperado de http://web.b.ebscohost.com.bdigital.sena.edu.co/nrc/delivery?vid=16&sid=8aa19d10-7aaa-4a34-a016-eea313717d11%40pdc-v-sessmgr03
- Landa, j. (2018). http://fcojlanda.me. Obtenido de http://fcojlanda.me/es/sin-categoria-es/kdd-y-mineria-de-datos-espanol/
- Salas, J., (2018). La OMS alerta de la caída de la actividad física en el siglo XXI. Recuperado de https://elpais.com/elpais/2018/09/04/ciencia/1536054340_198371.html
- OMS. (2013). Actividad física. Recuperado de https://www.who.int/dietphysicalactivity/pa/es/
- OMS. (2017). Enfermedades cardiovasculares. Recuperado de https://www.who.int/es/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
- OMS. (2019). Prevalencia de actividad física insuficiente. Recuperado de https://www.who.int/gho/ncd/risk_factors/physical_activity/en/

Sobre los autores

- **Yesica Lorena Zúñiga Mamián:** Estudiante de 10 semestre de Ingeniería de Sistemas Informáticos, miembro del Semillero de Investigación en Minería de Datos (SIMD). yesica.zuniga.m@uniautonoma.edu.co
- **Kevin Felipe Meneses Palta:** Estudiante de 10 semestre de Ingeniería de Sistemas Informáticos, miembro del Semillero de Investigación en Minería de Datos (SIMD). kevin.meneses.p@uniautonoma.edu.co
- **Néstor Iván Martínez Cobo:** Estudiante de 10 semestre de Ingeniería de Sistemas Informáticos, miembro del Semillero de Investigación en Minería de Datos (SIMD). nestor.martinez.c@uniautonoma.edu.co
- **Sandra Patricia Castillo Landínez:** Ingeniera de Sistemas (Universidad Nacional de Colombia), Especialista en Administración de la Información y Bases de Datos (Colegio Mayor del Cauca), Certified Big Data Professional, Certified Big Data Scientist. Docente de la Facultad

de Ingeniería, investigadora adscrita al Grupo de Investigación en Tecnología y Ambiente (GITA), coordinadora de la línea de Investigación en Ingeniería de Software y líder del Semillero de Investigación en Minería de Datos (SIMD). sandra.castillo.l@uniautonoma.edu.co

---