



LA FORMACIÓN DE INGENIEROS:
UN COMPROMISO PARA EL
DESARROLLO Y LA SOSTENIBILIDAD

15 al 18
DE SEPTIEMBRE

20
20

www.acofi.edu.co/eiei2020

MODELO DE MACHINE LEARNING PARA LA DETECCIÓN DE EVENTOS DE MARCHA HUMANA

Dayana Muñoz Muñoz, Sandra Patricia Castillo Landínez

**Corporación Universitaria Autónoma del Cauca
Popayán, Colombia**

Resumen

La marcha humana es un proceso en el cual una persona traslada su centro de masa hacia adelante alternando las piernas. Diferentes investigaciones han determinado que la marcha es una de las acciones más complejas efectuadas por el ser humano; se puede describir mediante los movimientos y estudiar las fuerzas ejercidas por los músculos (cinemática y cinética). Para su análisis existen diferentes métodos, entre ellos los sistemas de captura de movimiento 2D y 3D los cuales utilizan herramientas tecnológicas para la recolección de datos (cámaras, infrarrojos, marcadores, etc.), recreando estas acciones en diversas formas, dimensiones y formatos.

La marcha es cíclica y puede ser descompuesta en ciclos o zancadas, los cuales presentan dos fases: la primera, apoyo, donde el pie está en contacto con el suelo y equivale al 60% del ciclo, la segunda, oscilación o balanceo, cuando el pie no está en contacto con el suelo y representa el 40% restante. Las dos fases están limitadas por eventos de marcha humana. Cada marcha posee diferentes variables como longitud de paso, ancho de paso, longitud de zancada, ancho de zancada, velocidad, aceleración, cadencia.

Es importante estudiar el patrón de marcha de un individuo, ya que es un indicador de diferentes patologías, trastornos de movimiento, entre otros, y sirve como base para que los expertos en la salud propongan planes de rehabilitación y métodos preventivos para problemas motores. El análisis biomecánico de la marcha genera un gran volumen de datos variados, que difícilmente pueden ser analizados e integrados a través de métodos tradicionales, lo que constituye una barrera para su máximo aprovechamiento.

Este trabajo propone el uso de algoritmos predictivos de Machine Learning para la detección de eventos de marcha humana, los cuales constituyen la base del estudio biomecánico de la caminata en personas.

Palabras clave: marcha humana; machine learning; algoritmos predictivos

Abstract

Human walking is a process in which a person moves his center of mass forward by alternating his legs. Different investigations have determined that walking is one of the most complex actions carried out by humans; It can be described through movements and study of the forces exerted by the muscles (kinematic and kinetic). For its analysis, there are different methods, including 2D and 3D motion capture systems, which use technological tools for data collection (cameras, infrared, markers, inertial units), recreating these actions in various forms, dimensions, and formats.

The gait is cyclical and can be divided into cycles or strides, which have two phases: the first, support, where the foot is in contact with the ground and is equivalent to 60% of the cycle, the second, oscillation or rocking when the foot is not in touch with the ground and represents the remaining 40%. Human walking events limit the two phases. Each gait has different variables such as step length, step width, stride length, stride width, speed, acceleration, and cadence.

It is important to study the gait pattern of an individual, since it is an indicator of different pathologies, movement disorders, among others, and serves as a basis for health experts to propose rehabilitation plans and preventive methods for motor problems. Biomechanical gait analysis generates a large volume of varied data that can hardly be analyzed and integrated through traditional methods, which constitutes a barrier to its maximum use.

This work proposes the use of predictive algorithms of Machine Learning for the detection of human gait events, which constitute the basis of the biomechanical study of walking in people.

Keywords: human gait; machine learning; predictive algorithms

1. Introducción

La marcha humana es considerada un elemento esencial de la salud y el bienestar de las personas, es descrita como un movimiento corporal humano constituido por patrones y factores motrices alternantes, rítmicos de las extremidades y el tronco que determinan un desplazamiento hacia adelante (Agudelo, et al., 2013), siendo esta un proceso de locomoción que presenta una eficiencia y funcionalidad única, es decir que cada individuo posee un patrón determinado en su ejecución, dicho patrón es adquirido en la infancia mediante la práctica (Marco Sanz, 2016).

El análisis de la marcha humana tiene una gran variedad de aplicaciones como el diagnóstico y tratamiento de pacientes con diferentes alteraciones patológicas y lesiones traumáticas que les generan discapacidad y dependencia. Para la realización del análisis es necesario obtener los parámetros cuantitativos característicos de cada individuo haciendo uso de equipos que capturan estos datos (cámaras, infrarrojos, marcadores, etc.) y de esta manera poder recrear las acciones de cada sujeto.

Esta marcha se caracteriza por tener dos momentos o fases durante su ciclo, la primera nombrada como apoyo con un porcentaje del 60% y la segunda conocida como balanceo con un porcentaje del 40% (Martín Nogueras, et al., 1999). Es posible observar cada ciclo desde diferentes planos (sagital, horizontal, frontal) lo que permite tener un visón en dos o tres dimensiones. Estas fases están delimitadas por eventos de marcha humana; los cuales son claves para la descripción del movimiento (Figuereido, et al., 2018)

Se han realizados diferentes estudios en el análisis de la marcha, ente los cuales podemos encontrar a (Kone, et al., 2020) quien utiliza tres algoritmos: bosque aleatorio (RF), refuerzo de gradiente basado en histograma (HGB) y red de memoria a largo plazo (LSTM), con el objetivo de encontrar detectores de velocidad cero (ZVD) basados en aprendizaje automático. Del mismo modo el (Ana, et al., 2017) hace uso de los registros cinéticos y técnicas de minería de datos para obtener un modelo basado en metaclasificadores con regresión logística con el objetivo de obtener un modelo de clasificación diagnostica para diferentes patologías. Finalmente (Subero, et al., 2011) quien implementa técnicas de árboles de regresión logística para sugerir modelos de aproximación diagnóstica en pacientes patológicos con hemiplejia espástica.

Una de las poblaciones más vulnerables a sufrir cambios en su marcha e incluso desarrollar patologías del movimiento son los adultos mayores, como consecuencia de su envejecimiento presentan alteraciones en los distintos componentes de su marcha, haciendo que ésta se torne lenta e inestable aumentando la probabilidad de sufrir caídas, fracturas e incluso la muerte. Cartier, ha encontrado que anualmente en los EE. UU. 250.000 fracturas de cadera se dan en adultos mayores (Cartier,2002), 90% son resultado de una caída, la mayoría de estas fracturas ocurren con relación a la marcha, es decir que la marcha se ve afectada por muchos factores adquiridos durante el desarrollo y el paso de los años de cada individuo; por este motivo es difícil de analizar a través de métodos cotidianos debido a la gran cantidad de información que se obtiene de ella; por lo tanto, es necesario hacer uso de algoritmos y técnicas de Machine Learning para medir y analizar los diferentes eventos que son generados durante el estudio de esta, con el objetivo de prevenir y detectar diferentes alteraciones presentes en la marcha que puedan ser de utilidad en diferentes áreas del conocimiento.

2. Materiales y métodos

En este trabajo se utilizó un conjunto de datos biomecánicos obtenidos mediante un sistema óptico de captura de movimiento de ocho cámaras Vicon Vantage con una tasa de muestreo de 100Hz; participaron doce adultos jóvenes con edad de 23.67 ± 2.19 años, una altura de 1.68 ± 0.04 m, una masa corporal de 66.88 ± 4.04 Kg y un índice de masa corporal de 23.46 ± 1.77 Kg/m²; cada uno de ellos realizó siete marchas en una pasarela de 8.5m. Todos los participantes tienen género biológico masculino.

Las 245 variables que conforman este dataset se refieren a las coordenadas cartesianas de las trayectorias correspondientes a los marcadores reflectivos utilizados por los participantes siguiendo el protocolo propuesto por Vargas y colaboradores (Vargas et al, 2020). Además, incluye las variables de identificación del participante, del número de marcha, el tiempo y la clase de evento.

Para el análisis de los datos se utilizó el software R versión 4.0. con el propósito de implementar algoritmos de clasificación que permitieran identificar entre cinco eventos posibles: no evento (0), golpe de talón izquierdo (1), levantamiento de dedos izquierdo (2), golpe de talón derecho (3) y levantamiento de dedos derecho (4).

Random Forest

Se probó la técnica de machine learning Random Forest, también conocido como bosques aleatorios formados por un conjunto de árboles de decisión (árboles de clasificación o regresión). Es uno de los modelos de clasificación más utilizados en estadística y análisis de datos, por ser un método rápido, eficiente y fácil de interpretar (Cánovas García, Alonso Sarría, and Gomariz Castillo, 2016).

Los bosques aleatorios son entrenados por subconjuntos aleatorios con la finalidad de disminuir la correlación entre los árboles, estos son obtenidos del bagging también denominados in-bag correspondiente al rango entre 63% y 80% de la muestra original, el resto de los casos forman el out-of-bag con un porcentaje entre el 37% y 20% restante. Cada nodo se considera como un solo subconjunto aleatorio de predictores. En el momento de la presencia de un nuevo caso, este debe recorrer cada árbol y finalmente se realiza la predicción por medio de votación, se determina la probabilidad de que dicho dato pertenezca a cierta clase (Cánovas García, Alonso Sarría, and Gomariz Castillo, 2016). El out-of-bag es una medida de error aplicada cuando se hace uso del bootstrapping, también designado como OOB, encargado del error de la predicción cuando el bosque tiene en cuenta las variables que han quedado fuera de la bolsa (García Ruiz de León 2018).

Para la construcción del modelo se generaron dos muestras aleatorias teniendo en cuenta el número de instancias que contiene el dataset; la primera representa al 70% y es considerada como la muestra de entrenamiento (in-bag) y la segunda equivale al 30% y es nombrada como la muestra de prueba (out-of-bag). El modelo es entrenado con el 70% de los datos, haciendo uso de parámetros fundamentales como Ntree (número de árboles que conforman el bosque) con un valor de 1500, Mtree (número de variables o predictores que se presentan en cada nodo) equivalente a 6, y replace (admite que las muestras tengan reemplazo) igual a true. Este bosque aleatorio es puesto a prueba con la muestra out-of-bag con el objetivo de predecir las clases perteneciente dichos registros. La ejecución de este algoritmo permite obtener la importancia de las variables según el índice de gini, matrices de confusión, validación y estimación del modelo.

Índice gini

Uno de los indicadores sintéticos más utilizados para el análisis estadístico de la desigualdad en una distribución es el coeficiente de Gini, el cual fue ideado por el estadístico italiano Corrado Gini en 1912 (Corrado, 1992); El valor del índice se define como una relación con valores entre 0 y 1, siendo cero (0) la máxima igualdad y uno (1) la máxima desigualdad (Chaves, 2003). En la construcción de los árboles de clasificación, el índice de Gini es una medida de la varianza total en el conjunto de las K clases del nodo m.

$$G_m = \sum_{k=1}^K \hat{p}_{mk}(1 - \hat{p}_{mk})$$

Figura 1. Expresión matemática índice de Gini. Fuente (Amat Rodrigo, et al., 2017).

Cuando \hat{p}_{mk} es cercano a 0 o a 1 (el nodo contiene mayoritariamente observaciones de una clase), el término $\hat{p}_{mk}(1 - \hat{p}_{mk})$ es muy pequeño. Como consecuencia, cuanto mayor sea el criterio de partición de los nodos, menor el valor del índice Gini G (Amat Rodrigo, et al., 2017).

3. Resultados

Luego de la ejecución del modelo con los respectivos parámetros (Ntree y Mtree) que gobiernan la función; se obtiene un resumen donde se puede observar la estimación de la tasa de error (OOB), la cual tiene un porcentaje del 5.14%, esta se calcula usando las observaciones que se encuentran fuera de la bolsa; como también una matriz de confusión (Barrios Arce, 2019) formada con la muestra de entrenamiento donde las columnas hacen referencia al número de veces que se predice dicha clase, mientras que las filas representan la cantidad de instancias pertenecientes a la clase real; En la Figura 1 se muestra un modelo que intenta predecir las variables (clase0, clase1, clase2, clase3 y clase4) las cuales son obtenidas de los eventos generados por la marcha. La imagen puede ser interpretada de la siguiente forma: la OOB se simboliza con la línea de color negro, la clase 0 (línea color rojo) no genera evento y posee un error del 2,75%, La clase 1 (línea color magenta) tiene un error de 13,7%, la clase 2 (línea color azul rey) obtiene un error del 14,71%, la clase 4 (línea color verde) tiene un error del 13,71% y la clase 3 (línea color azul agua marina) posee un mayor error del 15,31% respecto a las demás clases.

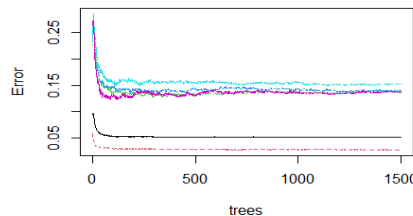


Figura 2. Random forest: Error vs número de árboles. Fuente propia.

Importancia de variables

De las 245 variables que presentaron al modelo random forest, fueron seleccionadas cinco (5) de las primeras más importantes según el criterio Gini (figura 2), donde el predictor que manifiesta mayor cantidad de respuestas es RTIA_3 con un valor de importancia de 155.88546, seguida de las variables LTIA_3 (152.52839), RTOL3 (151.57223), LTOO_3 (148.85893), RTOO_3 (148.80065), Estas presentan una reducción significativa de impuridad, por ende son las que mejor pueden explicar el bosque aleatorio ya que presentan mayor variabilidad en sus salidas.

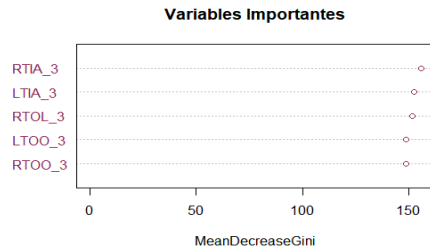


Figura 3. Variables más relevantes según el criterio de Gini. Fuente propia.

Evaluación del modelo

Para la estimación del modelo se genera una matriz de confusión (Barrios Arce, 2019) formada con la muestra de prueba (out of bag), donde se presenta la alineación entre las predicciones y los datos reales. Esta evaluación suministra unas estadísticas generales, de las cuales por ejemplo se obtiene la clasificación de los verdaderos positivos, verdaderos negativos, falsos positivos y fasos negativos. Como resultado de este proceso, el presente modelo tiene un acierto del 95.18%.

Tabla 1 Matriz de confusión con sus respectivas estadísticas

| | Class:0 | Class:1 | Class:2 | Class:3 | Class:4 |
|-----------------------|---------|---------|---------|---------|---------|
| Sensitivity | 0.9713 | 0.87602 | 0.89041 | 0.84936 | 0.89107 |
| Specificity | 0.8861 | 0.99259 | 0.99494 | 0.99110 | 0.99493 |
| Pos pred Value | 0.9709 | 0.86546 | 0.89450 | 0.85401 | 0.89890 |
| Neg pred Value | 0.8875 | 0.99325 | 0.99472 | 0.99076 | 0.99449 |

4. Discusión

El modelo alcanzo una estimación de error (OOB) del 5.14%, en la figura 2 se puede apreciar la diferencia entre los errores de clasificación que presentan las clases 1,2,3 y 4, siendo la clase 3 (golpe de pie derecho) la que presenta un mayor error con un 15.31% y la clase 0 (no genera evento) exhibe el error de clasificación más bajo con un 2.75% encontrándose por debajo de la OOB, del mismo modo se observa que a partir de los 1500 árboles el error de generalización se encuentra estabilizado, lo que indica que si se toman valores anteriores para la construcción del modelo es posible que exista un aumento en el error de estimación ya que el error no se encuentra equilibrado.

La figura 3 hace referencia a las primeas cinco (5) variables más relevantes (RTOO_3, LTOO_3, RTOL_3, LTIA_3 y RTIA_3) que se obtuvieron del modelo de Random Forest, haciendo uso del MeanDecreaseGini, cuyo valor de importancia se encuentra en entre 148,859 y 155,885; mientras que las cuatro (4) variables menos relevantes (LHJC_2, LKJC_2, SACRD_2 y LFEO_2) están ubicadas entre 11,330 y 11,972; con base en esto se obtuvo una diferencia de 144,155 entre la variable más relevante RTIA_3 con un valor de 155,885 y la variable menos relevante LHJC_2 con un valor de 11,330, lo que significa que se maneja una diferencia considerable, permitiendo que las variables más influyentes sean las que mejor puedan explicar el modelo ya que presentan un alto grado de clasificación. Esto es congruente con la teoría biomecánica ya que RTOO y LTOO



son los marcadores ubicados en los metatarsianos derechos e izquierdo respectivamente; mientras que las variables menos significativas están asociadas a las articulaciones de rodilla y pelvis.

En la tabla 1 se obtiene una matriz de confusión donde se encuentran unas de las estadísticas más importantes como lo es la sensibilidad, la cual describe la tasa de valores que se clasifican correctamente, la clase 0 presenta mayor sensibilidad (97%) en comparación con la clase 3 (84%), pues entre mayor sea la sensibilidad que presente la prueba, más zonas de eventos se podrán clasificar correctamente; como segunda estadística se tiene la especificidad o tasa de valores negativos donde la clase cero muestra un menor valor (88%) respecto a las demás clases, las cuales reflejan un porcentaje superior al 99%, además de valores predictivos que permitan determinar si el evento pertenece o no a la zona.

El trabajo de (Kidziński, Delp), presenta algunas similitudes con el actual estudio, en cuanto a la detección de eventos de marcha; además de realizar el análisis y seguimiento de la cinética y cinemática que contienen las articulaciones (cadera, tobillo, pies, pelvis, etc) que intervienen en la ejecución de la misma. Dicha investigación, realiza una comparación entre los modelos basados en coordenadas, velocidad y red de memoria a largo plazo. Donde el método más sobresaliente fue la red neuronal, la cual maneja dos eventos de marcha: el pie de contacto, el cual clasifico correctamente un 99% las instancias, mientras que el pie de despegue clasifico correctamente el 95% debido a que los participantes con patologías suelen no despegar el pie completamente, no obstante, los autores no tomaron en cuenta eventos relevantes que si fueron incluidos en la presente investigación; por ejemplo: pie izquierdo, golpe de pie izquierdo, pie derecho, golpe de pie derecho y no evento, al tener en cuenta estas variables podemos afirmar que el modelo presentado en este estudio tiene mayor exactitud aunque su porcentaje de acierto es del 95.18%.

5. Conclusión

- En el presente trabajo el modelo random forest presenta una estimación de error del 5,14%, lo que significa que, si este es sometido al ingreso de nuevos casos que se encuentren asociados con los eventos de marcha, se realizara la predicción con un margen de error máximo a 5,14%.
- El índice de gini en el bosque aleatorio permitió la identificación de las variables más influyentes; es decir, las que obtienen un menor error en cuanto a predicción del objetivo y presentan una mayor contribución en la construcción del modelo. Estas variables son las que pueden explicar de una mejor manera el modelo clasificador y son de gran utilidad para estudios posteriores como ajuste y depuración, con la finalidad de obtener un modelo más preciso.
- En trabajos futuros es necesario probar la eficiencia de otros modelos de clasificación que permitan comparar parámetros como carga computacional, precisión, especificidad, región de convergencia, entre otros.

6. Referencias

Artículos de revista

- Agudelo, A. I., Briñez, T. J., Guarín, V., & Ruiz, J. P. (2013). Marcha: descripción, métodos, herramientas de evaluación y parámetros de normalidad reportados en la literatura. *CES Movimiento y Salud*, 1(1), 29–43. <http://revistas.ces.edu.co/index.php/movimientoysalud/article/view/2481> Marco Sanz, Carmen. 2016. Cinesiología de La Marcha Humana Normal.
- Ana I. Aguilera, L. D. (2017). Modelo basado en metaclassificadores para diagnóstico en marcha patológica mediante análisis cinético. *Revista Ingeniería UC*. <http://servicio.bc.uc.edu.ve/ingenieria/revista/v17n2/art01.pdf>
- Andrade Saltos, V. A., & Flores M., P. (2018). Comparativa Entre Classification Trees, Random Forest Y Gradient Boosting; En La Predicción De La Satisfacción Laboral En Ecuador. *Ciencia Digital*, 2(4.1.), 43–56. <https://doi.org/10.33262/cienciadigital.v2i4.1..189>
- Campos, V. P. G. (2017). Modelado mediante Random Forest de las emisiones de autobuses urbanos en función de los ciclos cinemáticos. http://oa.upm.es/45914/1/TFG_VICTOR_PITA_GONZALEZ_CAMPOS.pdf
- Cánovas García, F., Alonso Sarría, F., & Gomariz Castillo, F. (2016). Modificación Del Algoritmo Random Forest Para Su Empleo En Clasificación De Imágenes De Teledetección. In *Aplicaciones de las Tecnologías de la Información Geográfica (TIG) para el desarrollo económico sostenible XVII Congreso Nacional de Tecnologías de Información Geográfica*, Málaga, 29, 30 de junio y 1 de julio 2016 (pp. 359–368). http://tig.age-geografia.es/2016_Malaga/Cánovas-García.pdf
- Cartier R., L. (2002). Caídas y alteraciones de la marcha en los adultos mayores. *Revista de Chile*, 332-337. https://scielo.conicyt.cl/scielo.php?script=sci_arttext&pid=S0034-98872002000300014.
- Chaves, E. (2003). Distribución y coeficiente de gini, curva paramétrica de Lorenz sugerida y calculos. *Tendencias*, 4(2), 5. <https://dialnet.unirioja.es/descarga/articulo/5029680.pdf>
- Figueiredo, J., Felix, P., Costa, L. L., Moreno, J. C., Santos, C. P., Félix, P., Costa, L. L., Moreno, J. C., & Santos, C. P. (2018). Gait Event Detection in Controlled and Real-Life Situations: Repeated Measures from Healthy Subjects. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(10), 1945-1956. <https://doi.org/10.1109/TNSRE.2018.2868094>
- Kidziński, Ł., Delp, S., & Schwartz, M. (2019). Automatic real-time gait event detection in children using deep neural networks. *PLoS ONE*, 14(1), 1–11. <https://doi.org/10.1371/journal.pone.0211466>
- Kone, Y., Zhu, N., Renaudin, V., & Ortiz, M. (2020). Machine Learning based Zero-Velocity Detection for Inertial Pedestrian Navigation. *IEEE Sensors Journal*, 1–1. <https://doi.org/10.1109/jsen.2020.2999863>
- Martín Nogueras, A., Calvo Arenillas, J. L., Orejuela Rodríguez, J., Barbero Iglesias, F. J., & Sánchez Sánchez, C. (1999). Fases de la marcha humana. *Revista Iberoamericana de Fisioterapia y Kinesiología*, 2(1), 44–49. <http://www.elsevier.es/es-revista-revista-iberoamericana-fisioterapia-kinesiologia-176-articulo-fases-marcha-humana-13012714>

- Medina-Merino, R. F., & Ñique-Chacón, C. I. (2017). Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python. *Interfases*, 0(010), 165. <https://doi.org/10.26439/interfases2017.n10.1775>
- Ruiz-Maya, L. (1978). Sobre la metodología del Índice de Gini. *Cuadernos de Economía*, 6(16), 327–346. <http://hdl.handle.net/10486/5861>
- Subero, A. A. (2011). Modelos de clasificación en marcha patológica usando árboles de regresión logística. *Multiciencias*. <https://www.redalyc.org/pdf/904/90421736012.pdf>

Libros

- Gini, Corrado (1912). *Variabilità e mutabilità. Contributo allo studio delle distruzione e delle relazione statistiche*. Bologna. Tipografia di Paolo Cuppini.
- Paradis, E., & Ahumada, J. A. (2003). *R para Principiantes*. Evolution. https://cran.r-project.org/doc/contrib/rdebuts_es.pdf.

Fuentes electrónicas

- Barrios Arce, J. I. (2019, July 26). La matriz de confusión y sus métricas. *Ciencia de Datos*. <https://www.juanbarrios.com/matriz-de-confusion-y-sus-metricas/>

Sobre los autores

- **Dayana Muñoz Muñoz:** Estudiante de 10 semestre de Ingeniería de Sistemas Informáticos, miembro del Semillero de Investigación en Minería de Datos (SIMD). dayana.munoz.m@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, Magister(c) en Ciencia de los Datos y Procesamiento de Datos Masivos (Big-Data). 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

Los puntos de vista expresados en este artículo no reflejan necesariamente la opinión de la Asociación Colombiana de Facultades de Ingeniería.

Copyright © 2020 Asociación Colombiana de Facultades de Ingeniería (ACOFI)