Encuentro Internacional de Educación en Ingeniería ACOFI
GESTIÓN, CALIDAD Y DESARROLLO EN LAS FACULTADES DE INGENIERÍA
CARTAGENA, COLOMBIA
18 al 21 de septiembre de 2018

IDENTIFICACIÓN DE PATRONES EN ACCIDENTES DE TRÁNSITO EN COLOMBIA DURANTE EL PERIODO 2010-2016 MEDIANTE EL USO DE TÉCNICAS DE MINERÍA DE DATOS

Juan Pablo Henao Pereira, Andrea Esperanza Tovar León, Fabián Andrés Urrea Ceballos, Sandra Patricia Castillo Landínez

Corporación Universitaria Autónoma del Cauca
Popayán, Colombia

Resumen

Según informes de la Organización Mundial de la Salud – OMS, los accidentes de tránsito se han convertido en un problema de salud pública, siendo uno de los mayores generadores de pérdidas de vidas que se presentan en las carreteras. En este proyecto se utilizaron los datos reportados por el observatorio de delitos de la Policía Nacional en el periodo comprendido entre 2010 y 2016, que involucra lesiones y homicidios en accidentes de tránsito; después de usar técnicas de preprocesamiento para mejorar la calidad del dataset, se emplearon algoritmos de minería de datos con el fin de identificar patrones que permitieron caracterizar la accidentalidad en Colombia. Adicionalmente se obtuvieron representaciones gráficas, resultado de un análisis visual que exhiben la situación de la Ciudad de Popayán (Cauca) durante el mismo periodo.

El trabajo de investigación inició en septiembre de 2017 y está enmarcado en el proyecto "Identificación de patrones en datasets gubernamentales: caso de estudio hurtos y accidentes de tránsito en Colombia"; actualmente se estudian otros modelos y técnicas de machine learning. A futuro se busca explorar otras fuentes de datos e incluir nuevas variables (sociales, económicas, demográficas) a fin de generar patrones más completos.

Los resultados buscan llamar la atención de los entes gubernamentales y la sociedad en general para tomar medidas efectivas que reduzcan los daños económicos, físicos, psicológicos y emocionales que genera un accidente en las vías.

Palabras clave: minería de datos; accidentes de tránsito; análisis visual

Abstract

According to reports from the World Health Organization - WHO, traffic accidents have become a public health problem, being one of the largest generators of loss of life that occur on the roads. In this project the data reported by the observatory of crimes of the National Police in the period between 2010 and 2016, involving injuries and homicides in traffic accidents, were used; After using preprocessing techniques to improve the quality of the dataset, data mining algorithms were used in order to identify patterns that allowed us to characterize the accident rate in Colombia. Additionally, graphic representations were obtained, as a result of a visual analysis that shows the situation of the City of Popayán (Cauca) during the same period.

The research work began in September 2017 and is part of the project "Identification of patterns in government datasets: theft case study and traffic accidents in Colombia"; other models and techniques of machine learning are currently being studied. In the future, we seek to explore other data sources and include new variables (social, economic, demographic) in order to generate more complete patterns.

The results seek to draw the attention of government entities and society in general to take effective measures that reduce the economic, physical, psychological and emotional damages caused by an accident on the roads.

Keywords*: data mining; traffic accidents; visual analysis*

1. Introducción

Según la Organización Mundial de la Salud – OMS, un accidente de tránsito es un evento en las vías en el cual se ve involucrado al menos un vehículo automotor en movimiento, y en donde ciclistas, peatones y motociclistas llevan la peor parte; este fenómeno ha sido catalogado como un problema de salud pública ya que su impacto no solo se limita a pérdidas materiales (OMS, 2017), sino lo más importante, vidas humanas y las consecuencias que lo preceden para un núcleo familiar con afectaciones psicológicas, físicas y económicas, que también se manifiestan en los altos costos para los servicios de salud y movilidad.

Los accidentes de tránsito se han convertido en un inconveniente progresivo, que cada día cobra más vidas en Colombia, durante el 2017 en promedio 18 personas perdieron la vida diariamente, 162 fallecieron durante enero y diciembre del mismo año en el departamento del Cauca y en total a nivel nacional 6.479 fueron reportados por el Instituto Nacional de Medicina Legal y Ciencias Forenses. Se suma a esto 38.073 lesionados durante el mismo periodo. A nivel general se percibe un leve decremento en las cifras reportadas de fallecidos y lesionados comparado con las cifras en un mismo periodo de 2016; sin embargo, contrasta con el aumento de peatones víctimas en accidentes de tránsito (Martínez et al., 2018).

Las características generales de los incidentes y demás los datos generados durante el proceso de valoración de la escena (Ministerio de Transporte, 2006), incluyendo móviles del accidente y

detalles de las personas involucradas reconocidas como fallecidas o lesionadas durante el evento están a disposición de la ciudadanía en general gracias a la Ley 1712 de 2014 de Transparencia y acceso a la información, cuyo principal objetivo es involucrar activamente a personas del común con el aprovechamiento de datos abiertos para fines de mejoramiento de procesos que generen impacto social.

Analizar y procesar una cantidad tan considerable de información, ya sea de forma manual o mediante técnicas tradicionales resulta improcedente, por lo que se requiere el uso de herramientas robustas que proporcionen algoritmos novedosos y eficientes para generar resultados relevantes que soporten la adopción de políticas y medidas efectivas por parte de las autoridades competentes. En consideración a lo expuesto anteriormente, se decidió emplear técnicas de minería de datos para estudiar los datasets disponibles.

2. Minería de datos

Hace parte del proceso de extracción y análisis de grandes cantidades de datos (Knowledge Discovery from Data Base - KDD) en busca de conocimiento valioso no explícito a simple vista debido sus complejas relaciones dentro de bases de datos extensas (Clinic Cloud, 2016) (Owen Duncan, 2017). Actualmente la gran influencia de las Tecnologías de la Información y la Comunicación (TIC) ha hecho de la minería de datos un gran diferenciador en la toma de decisiones, desde la manera como se hace comercio hasta la obtención de patrones predictivos para el diagnóstico de enfermedades. Por sus características, no solo es posible encontrar relaciones entre la gravedad de las lesiones y las características del conductor de un vehículo involucrado o entre variables como estado de las vías y condiciones ambientales (Chang & Wang, 2006), sino también patrones dentro de los datos reportados en accidentes de tránsito que permitan orientar medidas preventivas.

La construcción de modelos para descubrir patrones o relaciones mediante algoritmos de minería de datos tiene dos orientaciones: la predictiva (estimar valores desconocidos a partir de variables independientes) y la descriptiva (establecer patrones que permitan explicar los dataset) (Rodríguez, J. 2010). Dentro de esta última categoría se encuentran las técnicas de agrupación o clustering.

Clustering, es una técnica que permite analizar y examinar datos que no se encuentran etiquetados, formando conjuntos de grupos a partir de su similitud (Figura 1). Su principal objetivo es dividir un conjunto de objetos en dos o más grupos, basándose en la similitud de un conjunto de variables que los caracterizan. La similitud puede medirse a través de funciones de distancia, y los objetos se agrupan de acuerdo a todas las variables y por ello, una variable irrelevante puede generar ruido en los resultados obtenidos (Cáceres, 2016).

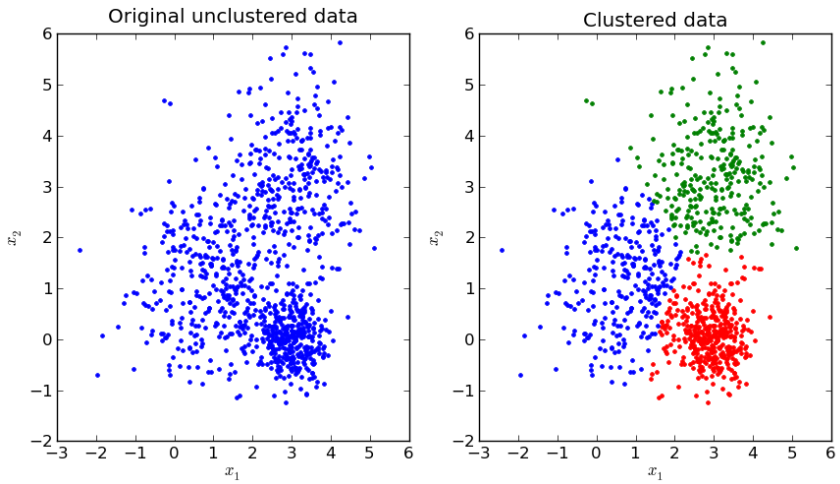

Figura 1. Creación de clúster. Fuente: https://towardsdatascience.com

El trabajo realizado buscó establecer relaciones y encontrar características similares entre los datos de accidentes de tránsito mediante la generación de Cluster (Jaramillo & Paz, 2015). Se definieron dos conjuntos de datos objetivo para el análisis, el primero un archivo compuesto de lesionados y fallecidos en accidentes de tránsito en la ciudad de Popayán durante el 2016 y el segundo, el consolidado nacional de similares características en el periodo 2010 a 2016.

3. Materiales y métodos

En este proyecto se usaron datos de libre acceso para hacer un análisis de accidentes de tránsito en Colombia. Los datos usados provienen de los reportes de la Policía Nacional disponibles en la página de Datos Abiertos www.datos.gov.co; después de unir los informes anuales comprendidos en el periodo 2010 a 2016 de lesionados y fallecidos, el dataset final estaba compuesto de 267.317 registros y 20 variables tales como departamento, ciudad, barrio, sexo, edad de la víctima, entre otras.

El análisis se realizó por medio de dos métodos diferentes:

- Análisis visual: orientado a crear una representación gráfica de los datos con el fin de simplificar su comprensión y facilitar la identificación de patrones. Se usó la herramienta ArcGIS para georreferenciar los puntos de accidentalidad en la ciudad de Popayán y producir cartografía temática con visualizaciones de las zonas con mayor ocurrencia de accidentes de tránsito y algunas de las condiciones bajo las que se presentan.
- Análisis descriptivo usando clústers para encontrar relaciones significativas entre los atributos y de esta manera identificar patrones interesantes y antes desconocidos que aporten conocimiento a los entes gubernamentales encargados de establecer de medidas preventivas en el control y manejo de los accidentes de tránsito.

El análisis y procesamiento de los datos se realizó por etapas, tomando como referencia la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) (Urrego, 2010), la cual

plantea que el ciclo de vida de un proyecto de minería de datos está basado en fases, cambiantes entre sí, y ésta a su vez se componen de actividades o secuencia de pasos ordenados (Jaramillo & Paz, 2015). A continuación, se describe el proceso:

- Fase 1: comprensión del problema. Es comprender y definir el problema, ya que es importante para entender los objetivos del proyecto. Por lo que en esta fase es necesario la realización de tareas específicas como:

  - Evaluación de la situación
  - Entender las necesidades y las condiciones en las que sucede
  - Valoración inicial de las posibles técnicas y herramientas a utilizar.

- Fase 2: comprensión de los datos. Involucra la búsqueda de información y de las variables que se utilizaran durante el proceso, contiene las tareas:

  - Recolección inicial de datos
  - Exploración de los datos
  - Verificación de calidad de los datos

- Fase 3: preparación de los datos. La preparación de los datos se realiza para adaptarlos de tal manera que sean óptimos para aplicar la técnica elegida. Se realizó un preprocesamiento sobre los atributos que involucró la identificación de valores anómalos, faltantes, no reportados o no identificados durante el reporte de un accidente en particular, finalmente se eligieron los atributos considerados como relevantes: fecha, franja horaria(madrugada, mañana, tarde, noche), día, barrio, municipio, departamento, zona, móvil agresor, como datos de la víctima: móvil víctima, edad, sexo, estado civil, país de nacimiento, escolaridad y tipo(lesionado, homicidio). Las actividades desarrolladas fueron:

  - Seleccionar los datos.
  - Depuración de los datos.
  - Estructuración de los datos
  - Integración de los datos.
  - Formateo de los datos

- Fase 4: análisis visual. Con el fin de mejorar el entendimiento respecto a la ubicación geográfica de los registros y producir mejores resultados al analizar sus características e impacto se requirió:

  - Definir el tipo de visualizaciones a crear
  - Construir visualizaciones.
  - Evaluar visualizaciones.

- Fase 5: aplicación de técnicas de minería de datos. Para la generación de los clústers se trabajó específicamente con el algoritmo k-modes propuesto por Zhexue Huang en 1998 (Huang, 1998), el cual permite simplificar la agrupación de datos categóricos, se realizaron las tareas:

- Construcción de los modelos.
- Evaluación de resultados.

4. Resultados

- Producto del análisis visual de accidentes de tránsito (AT) que involucra homicidios y lesionados en la ciudad de Popayán se obtuvo:

  - El 69.5 % de las personas involucradas en accidentes de tránsito en la ciudad de Popayán son hombres y el 30.5% restante son mujeres (Figura 2).

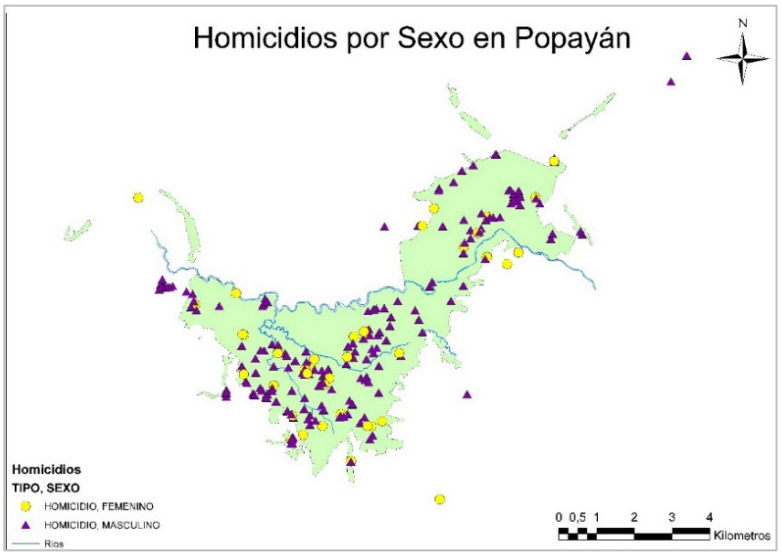

Figura 2. Análisis de homicidios en accidentes de tránsito por sexo durante el 2016 en Popayán. Fuente propia

Teniendo en cuenta la franja horaria en que se presentan los eventos, se tiene que el 4.5% de los accidentes ocurre en la madrugada (12:00 am-4:59 am), 33.8% en horas de la mañana (5:00 am-11:59 am), 18.6% ocurre en la tarde (12:00 pm-6:59 pm) y el 43.1% restante sucede en la noche (7:00 pm-11:59 am).

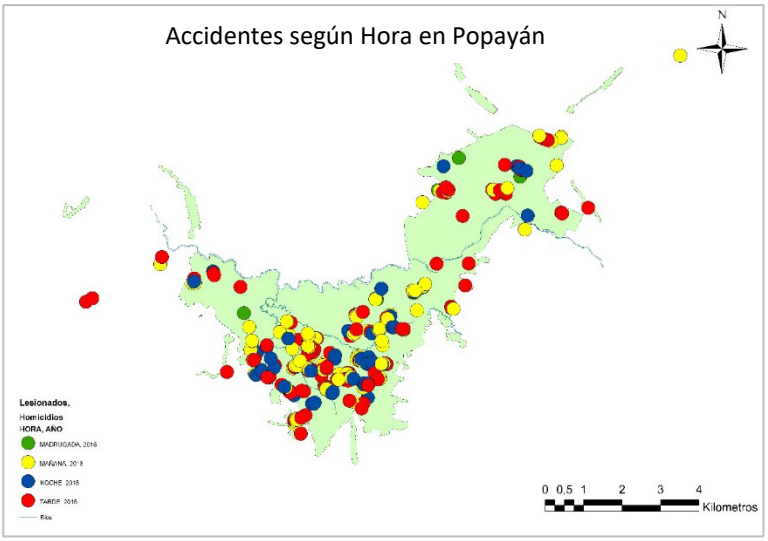

Figura 3. Análisis de accidentes según la hora durante el 2016 en Popayán. Fuente propia

- La zona donde más se presentan accidentes de tránsito es el centro de la ciudad (7.2%), seguido del barrio La Esmeralda con 6.3%

- Como resultado de usar la técnica de minería de datos denominada clustering o agrupación, y específicamente el algoritmo K-Modes, se generaron dos modelos: uno con tres grupos y otro cinco. A continuación, se describen las características de cada uno.

3 clúster (k=3)

- **Clúster 1:** es el más grande de todos y está conformado por 96.908 registros, la mayoría de accidentes de tránsito se presentaron en el mes de julio del año 2015, principalmente ocurrieron en la zona urbana los departamentos de Antioquia (Candelaria) y Cundinamarca (Bogotá) dos de los departamentos más grandes del país y con gran número de población, también se observa que los días lunes y viernes de este mes en horas de la tarde fue cuando más sucedieron los eventos; se vieron involucrados conductores de motocicletas con una edad promedio de la victima de entre 26 y 27 años de sexo masculino, un nivel de escolaridad de básica secundaria y estado civil soltero. La mayoría de los hechos agrupados en este clúster corresponden a accidentes de tránsito con lesionados.

- **Clúster 2:** compuesto por 83.981 registros, la mayoría de los accidentes se presentaron en el mes de mayo del año 2016 en el departamento del Valle los días sábado y domingo en horas de la tarde, principalmente en la zona urbana. Se vieron involucrados conductores de motocicletas con una edad promedio de las víctimas de 28 a 30 años de sexo masculino, solteros y con un nivel de escolaridad de secundaria. La mayoría de los hechos agrupados en este clúster corresponden a accidentes de tránsito con lesionados.

- **Clúster 3:** comprende 86.428 registros; la mayor parte de los accidentes se presentaron en el mes de enero de 2013 en el departamento de Santander específicamente en Bucaramanga, los días viernes y sábado en horas de la mañana; en la mayoría de los hechos se asocian

vehículos y conductores de motocicletas de sexo masculino, con una edad promedio de entre 49 y 50 años, casados y nivel de escolaridad secundaria. , La mayoría de los hechos agrupados en este clúster corresponden a accidentes de tránsito con lesionados.

5 clúster (k=5)

- **Clúster 1:** compuesto por 36.306 registros de accidentes de tránsito, la mayoría ocurrieron los días martes en horas de la tarde durante el mes de mayo de 2011 en el departamento del Valle (en la zona del centro de Cali). En los hechos se vieron involucrados vehículos y transeúntes lesionados de sexo masculino, con una edad promedio de 44 años, estado civil unión libre, y nivel de escolaridad de secundaria.

- **Clúster 2:** posee 56.779 registros. Los accidentes se concentran en el centro de la ciudad de Bogotá los días viernes y domingos de septiembre del año 2010, en horas de la tarde. Se involucraron vehículos y conductores de motocicletas que resultaron lesionados con un promedio de edad de la victima de 37 años, de sexo masculino, casados y un nivel de escolaridad de secundaria.

- **Clúster 3:** contiene 50.224 registros; la mayoría de los sucesos se registraron en el mes de abril del año 2016, en el departamento de Santander (en el centro de Bucaramanga) los días lunes durante la tarde; los afectados fueron conductores de motocicletas[1] con una edad promedio de 30 años, de sexo masculino, de estado civil unión libre y nivel de escolaridad secundaria.

- **Clúster 4:** comprende 97.247 registros. Los accidentes se presentaron en el Valle y la zona centro de Medellín durante el mes de febrero del año 2011, los sábados en horas de la tarde. Los involucrados fueron vehículos y conductores de sexo masculino de motocicletas, con promedio de 20 años de edad, la mayoría eran solteros y nivel de escolaridad secundaria.

- **Clúster 5:** formado por 26.761 registros. La mayor parte de los siniestros se presentaron en Bogotá durante los viernes en la tarde en el mes de julio de 2010; los implicados fueron vehículos y transeúntes femeninas y casadas con una edad promedio de 65 años y nivel de escolaridad primaria.

5. Conclusiones

- A partir del análisis visual se pudo determinar que los homicidios en accidentes de tránsito (AT) a nivel nacional han aumentado respecto a años anteriores.
- Aunque en la actualidad hay libre acceso a los reportes de accidentalidad en Colombia, los datos carecen de detalles en cuanto a la ubicación exacta de la ocurrencia de los hechos, donde a pesar de coincidir los atributos reportados tanto para homicidios como lesionados falta estandarizar en el método de captura de los datos y el mecanismo usado (planilla, formulario). Lo anterior permitiría evitar errores de digitación, ausencia de datos y de esta

[1] Según la Dirección de Tránsito de Bucaramanga, en el año 2016 habían registradas 368.206 motos.

manera se puede garantizar la fiabilidad de la información

- El uso de minería de datos permitió extraer información que se encontraba oculta en los reportes de accidentalidad en Colombia; se identificaron patrones que describen las características bajo las cuales más se presentan homicidios y lesionados en (AT).
- El conocimiento generado a partir de la aplicación de técnicas de minería puede ayudar a los organismos gubernamentales y de seguridad a tomar decisiones eficaces relacionadas a la implementación de planes de prevención de accidentalidad en el país.
- Como trabajo futuro, serán incluidas y estudiadas otras variables de tipo socio-económico para identificar nuevos patrones y la generación de modelos de tipo predictivo.

6. Referencias

Artículos de revistas

- Jaramillo, A., & Paz, H. (2015). Aplicación de Técnicas de Minería de Datos para Determinar las Interacciones de los Estudiantes en un Entorno Virtual de Aprendizaje. *Revista Tecnológica ESPOL – RTE*, Vol. *28,* No.1, pp. 64–90.
- Urrego, C. (2010). Metodología crisp para la implementación Data Warehouse. *Tecnura*, Vol. *14, No.* 26, pp. 35–48.
- Huang, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, Vol. *2,* No. 3, pp. 283–304.
- Chang, L. Y., & Wang, H. W. (2006). Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis and Prevention*, Vol. 38, No. 5, pp. 1019-1027.
- Cáceres, J. H. (2016). Clustering technique based on k- means algorithm for the identification of clusters of surgical patients. *Universidad Santo Tomás, Seccional Bucaramanga*, pp. 1–8.

Libros

- Rodríguez, J. E. (2010). *Fundamentos de minería de datos*. (Universidad Distrital Fco José de Caldas., Ed.) (1st ed.), pp. 205.

Fuentes electrónicas

- Clinic Cloud. (2016). ¿Qué es el data mining? La definición de la minería de datos. Consultado el 01 de junio de 2018 en https://clinic-cloud.com/blog/data-mining-que-es-definicion-mineria-de-datos/
- Martínez, A. M., Ansv, D., López, M., Director, B., Observatorio, T., Equipo, A., … Peinado, V. (2018). *Cifras para Colombia Fallecidos y Lesionados en hechos de tránsito* (Vol. marzo). Consultado el 13 de junio de 2018 en http://ansv.gov.co/observatorio/public/documentos/boletin_nacional.pdf
- Ministerio de Transporte. (2006). Manual para el diligenciamiento del formato del informe policial de accidentes de tránsito adoptado según resolución 004040 del 28 de diciembre de

2004 modificada por la resolución 1814 del 13 de julio de 2005., 109. Consultado el 01 de junio de 2018 en https://www.mintransporte.gov.co/descargar.php?idFile=6412
- OMS. (2017). OMS | 10 datos sobre la seguridad vial en el mundo. *WHO*. Consultado el 20 de abril de 2018 en http://www.who.int/features/factfiles/roadsafety/es/#.Wtn4xoyVvLg.
- Owen Duncan. (2017). Conceptos de minería de datos | Microsoft Docs. Consultado el 20 de marzo de 2018 en https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-concepts?view=sql-server-2017.

Sobre los autores

- **Juan Pablo Henao Pereira**: Estudiante de 10 semestre de Ingeniería de Sistemas Informáticos, miembro del Semillero de Investigación en Minería de Datos (SIMD). juan.henao.p@uniautonoma.edu.co
- **Andrea Esperanza Tovar León**: Estudiante de 10 semestre de Ingeniería de Sistemas Informáticos, miembro del Semillero de Investigación en Minería de Datos (SIMD). andrea.tovar.l@uniautonoma.edu.co
- **Fabian Andrés Urrea Ceballos:** Estudiante de 9 semestre de Ingeniería de Sistemas Informáticos, miembro del Semillero de Investigación en Minería de Datos (SIMD). fabian.urrea.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

---

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

Copyright © 2018 Asociación Colombiana de Facultades de Ingeniería (ACOFI)