WEKA, ÁREAS DE APLICACIÓN Y SUS ALGORITMOS: UNA REVISIÓN SISTEMÁTICA DE LITERATURA

WEKA, AREAS OF APPLICATION AND THEIR ALGORITHMS: A SYSTEMATIC REVIEW OF LITERATURE

Marcos Antonio Espinoza Mina, Mgs.
Magíster en Negocios Internacionales y Gestión en Comercio Exterior (Ecuador).
Doctorando en Administración de Empresas
Pontificia Universidad Católica Argentina Santa María de los Buenos Aires.
Docente de Universidad Tecnológica ECOTEC, Ecuador.
Docente de la Universidad Agraria del Ecuador.
mespinoza@ecotec.edu.ec
mespinoza@uagraria.edu.ec

ARTÍCULO DE REFLEXIÓN
Recibido: 5 de septiembre de 2018.
Aceptado: 20 de noviembre de 2018.

RESUMEN

Actualmente se generan grandes cantidades de datos almacenados en dispositivos digitales que cada día, debido a los avances tecnológicos, crecen también en su capacidad de almacenamiento. Muchos de estos datos no se encuentran adecuadamente estructurados, resultando una tarea difícil su explotación. Ante estas realidades y dificultades es necesario hacer uso de técnicas automatizadas que permitan reducir, analizar y utilizar de forma eficiente los datos. Se han desarrollado variadas metodologías, programas y complementos aplicados específicamente al conjunto de datos con el que se trabaje; se destacan las herramientas informáticas que implementan técnicas para el aprendizaje automático y la minería de datos, una de ellas es Weka. El presente trabajo hace una revisión sistemática de la literatura cuyo objetivo fue buscar los campos o áreas de aplicación de Weka y los algoritmos más utilizados de este programa. Los resultados exponen que Weka está siendo empleada en campos como: informática, medicina, educación y agricultura. Los algoritmos más utilizados dependiendo del tipo de datos y propósito de la evaluación son: Naïve Bayes, J48, Decision Tree, Random Forest, Support Vector Machine (SMV) y Sequential Minimal Optimization (SMO).

Palabras clave: Minería de datos, aprendizaje automático, algoritmos, predicción, clasificación.

ABSTRACT

Currently, large amounts of data stored in digital devices are generated every day, due to technological advances, also grow in their storage capacity. Many of these data are not properly structured, resulting in a difficult task exploitation. Faced with these realities and difficulties, it is necessary to use automated techniques to reduce, analyze and use data efficiently. Several methodologies, programs and complements have been developed specifically applied to the data set with which he works; highlights are the computer tools that implement techniques for automatic learning and data mining, one of which is Weka. The present work makes a systematic review of the literature whose objective was to search the fields or application areas of Weka and the algorithms most used in this program. The results show that Weka is being used in fields such as computers, medicine, education and agriculture. The most used algorithms depending on the type of data and purpose of the evaluation are: Naïve Bayes, J48, Decision Tree, Random Forest, Support Vector Machine (SMV) and Sequential Minimal Optimization (SMO).

Keywords: Data mining, machine learning, algorithms, prediction, classification.

INTRODUCCIÓN

La evolución tecnológica supone un crecimiento exponencial y con ese desarrollo las oportunidades en el mundo empresarial también se ven acrecentadas, siempre y cuando las organizaciones puedan hacer un uso eficiente de esa tecnología. El crecimiento explosivo de las bases de datos junto con las conexiones en internet, aceleran la búsqueda de técnicas y herramientas que, de manera automática y eficiente, generen información a partir de los datos almacenados. El poseer abundantes datos no es lo mismo que tener conocimiento. El exceso de datos no permite la comprensión de una determinada situación. Existe una gran diversidad de datos, en conjuntos cada día más voluminosos, y que abarcan periodos de tiempo más largos. En cada uno de ellos hay innumerables preguntas a responder y casos extraños a encontrar. Estas preguntas pueden ser desde medidas específicas hasta modelos para el comportamiento de millones de personas (Baeza, 2009).

Cada vez que: se paga un bien o servicio con una tarjeta de crédito en un comercio, se envía un correo electrónico, se escribe un mensaje en un celular, se utilizan las redes sociales o se activan determinados servicios como geolocalización; se están generando datos y con ellos se puede fácilmente, por ejemplo, crear el perfil digital de una persona y así, las empresas pueden conocer mejor a sus clientes, mejorar su fidelidad, aumentar los volúmenes de ventas, atraer al cliente por un producto específico.

No es difícil descubrir que es de vital importancia para las organizaciones incorporar herramientas que automaticen el análisis de datos desde múltiples fuentes, los transforme, los mida y permita mostrarlos en efectivos informes gráficos que se actualicen en tiempo real. A partir de estas realidades, han surgido conceptos que agrupan variadas herramientas como el de big data, minería de datos (Data Mining), ciencia de datos (Data Scientist), descubrimiento del conocimiento en bases de datos (Knowledge Discovery in Databases o KDD) e inteligencia de negocios (Business Intelligence), que inquieren desde la infinidad de información y aprovechan las nuevas formas de trabajo digital que son fundamentales para la competitividad de las empresas.

Muchas organizaciones poseen montañas de datos, pero no saben qué hacer con ellos, sin ser capaces de accederlos y consolidarlos. Se puede sumar a esta realidad el hecho de que muchos datos están incompletos o con errores. Partiendo de los datos generados luego del tratamiento de un conjunto de metodologías, aplicaciones y tecnologías de big data se pueden elaborar estrategias que contemplen la totalidad de la información disponible y la simplifiquen. El concepto de big data está asociado a esos conjuntos de datos que crecen rápidamente y que por su cantidad se dificulta su procesamiento, análisis y gestión.

Los datos siguen generándose cada día y es obligatorio tener alternativas para almacenar, clasificar, analizar y compartir esa información para conseguir resultados en beneficio de la organización. El proceso de extracción no trivial, de información implícita, previamente desconocida y potencialmente útil no es nuevo, para ello una muestra se tiene en el análisis supervisado de las opiniones políticas en tiempo real con técnicas de aprendizaje automático o de evaluación tendencial, vía el análisis de datos masivos, que se dio en las elecciones presidenciales 2012 en EUA (Estados Unidos de América), del finalmente presidente reelecto Obama, a través de una solución big data de Oracle (con un coste medio de 3 a 4 millones de dólares EUA) y la implicación de equipos interdisciplinares de análisis y redacción de informes de campaña (Arcila, Ortega, Jiménez y Trullenque, 2017).

Para optimizar el proceso de toma de decisiones y reducir los tiempos de respuesta se requiere que los empleados en las empresas puedan: entender los datos, los procesos y la información generada; realizar conexiones con diferentes fuentes de datos; utilizar herramientas para exportar, importar y transformar datos; manejar editores de consultas; seleccionar filtros adecuados y hacer segmentaciones; incluso diseñar y modelar relaciones en los datos. Destrezas que hasta hace pocos años eran exigidas solo a los profesionales en sistemas e informática. Así mismo son responsables de transformar los datos (internos y externos) en información y a su vez la información en conocimiento, con el fin de dar insumos para la toma de decisiones de la organización de forma eficaz y exitosa.

Hay que mencionar, además, que las organizaciones necesitan contar con colaboradores diestros en herramientas para la generación de modelos que faciliten el procesamiento de datos y permitan el reconocimiento de patrones; también, con habilidades para la definición de hipótesis de trabajo para su verificación o negación en la búsqueda de nuevos conocimientos. Se debe agregar que deben poseer experiencias en la transformación de datos desorganizados en información estructurada. Las empresas en todo el mundo están comenzando a valorar los conocimientos que tengan los profesionales en las múltiples herramientas para tratamiento de datos, por lo que aprender a manejarlas es de gran aporte para los distintos perfiles profesionales.

Los profesionales especialistas en informática que incursionan en estos nuevos campos, no solo deben manejar muy bien los conceptos y nuevas herramientas ya expuestas, también deben comprender y analizar el contexto del negocio y los procesos de las organizaciones, de forma tal que le permita fusionar todos estos conocimientos, y con los resultados de la aplicación de los mismos, generen diseños e implementación de mejoras que sirvan para aumentar la productividad, mejorar la eficacia en la toma de decisiones alto impacto y ganar mayor competitividad.

La aplicación de la minería de datos, además de permitir el descubrimiento de conocimientos para el sector comercial, soporta las investigaciones en muy variadas ramas, como por ejemplo la biológica; encuentran en ella una herramienta insustituible para enfrentar la avalancha de datos que producen las investigaciones genómicas y proteómicas (Febles y González, 2002).

El análisis ambiental del entorno urbano para evaluación de la afectación externa en actividades de las empresas o de los ciudadanos; la determinación de una producción agrícola de ciclos cortos o largos en cantidades justas para satisfacer la demanda; el tener una visión y control completo de las tierras fértiles, pronosticando las áreas de enfermedades y detección de las mismas en base a condiciones previamente evaluadas; la personalización de los productos turísticos según el perfil de los clientes; son ejemplos de los resultados que se puede obtener hoy en día, ya que la tecnología permite procesar grandes volúmenes de datos e información y descubrir comportamientos, tendencias e incluso determinar información que no es obvia por el mismo hecho de tener esa amplitud de datos.

Weka es una colección de algoritmos de aprendizaje automático para tareas de minería de datos; contiene herramientas para la preparación de datos, clasificación, regresión, clustering, minería de reglas de asociación y visualización. Weka es un software de código abierto emitido bajo la Licencia Pública General de GNU («Weka - University of Waikato», 2018).

Las técnicas de aprendizaje automático están basadas sobre un modelo explícito o implícito establecido que posibilita categorizar los patrones analizados (Rivero Pérez, 2014); mientras tanto, la minería de datos abarca todo un conjunto de técnicas enfocadas en la extracción de conocimiento implícito en las bases de datos (Corso, 2009).

El objetivo principal de esta revisión sistemática de literatura es compilar una lista de los campos o áreas de actividad en los que se utiliza Weka como una herramienta de análisis de datos y los algoritmos que se emplean. A los investigadores les resulta útil este trabajo, ya que adicionalmente proporciona de forma resumida una descripción de los usos que se le está dando a esta herramienta.

El artículo le permite abrir una nueva visión a otros actores, diferente a los profesionales de informática, como los empresarios o científicos de otras ramas, que, al emplearlo como un documento informativo preliminar, les permitirá conocer sobre qué algoritmos existen y son mejor evaluados según su formulación y tipo de datos a explotar. Además, también pueden consultar el documento como referencia para descubrir el estado de Weka frente a otras herramientas similares.

Este trabajo está estructurado de la siguiente forma: en los puntos 2.1, 2.2, 2.3 y 2.4 se presenta la metodología utilizada para realizar la revisión sistemática, la formulación de

preguntas, la selección de fuentes y estudios, extracción de información y el proceso de mapeo. En el punto 2.5, se muestran los resultados obtenidos después de realizar el mapeo sistemático de la revisión. En el 2.6 se expone una discusión de los resultados, para finalmente, en el punto 3 presentar las conclusiones de la investigación.

1. REVISIÓN TEÓRICA

1.1. Revisión sistemática

El objetivo de la investigación fue desarrollar un mapeo sistemático de revisión de la literatura relacionada con la herramienta informática Weka, sus áreas de aplicación y los algoritmos más utilizados a través de este programa.

El método de revisión sistemática planteado presentado por Kitchenham (2004) ha sido adaptado para llevar a cabo esta investigación. El objetivo básico de una revisión sistemática es compilar y evaluar todas las investigaciones disponibles relacionadas con una cuestión de interés, logrando así resultados imparciales, auditables y repetibles.

1.2. Definición de la pregunta de investigación

Los campos en los que se utiliza Weka y los algoritmos más utilizados fueron obtenido formulando la siguiente pregunta: ¿En qué campos se emplea la herramienta Weka y cuáles son los algoritmos más utilizados?

La lista de palabras claves utilizadas para llevar a cabo la investigación fueron:

- Weka
- Minería de datos / Data mining
- Aprendizaje automático / Machine learning

La pregunta de investigación formulada, una vez que se completó la revisión sistemática del mapeo, proporcionó en diferente medida, los siguientes resultados:

- Reconocimiento de los campos en los cuales se utiliza el software Weka.
- Listado de los principales algoritmos utilizados en el software Weka

1.3. Selección de fuentes

El objetivo de esta sección es exponer las fuentes donde se realizaron búsquedas de las investigaciones realizadas referentes al tema. Se definieron los siguientes criterios para la selección de las fuentes:

- Realización de la búsqueda de versiones digitales de artículos, revistas y conferencias documentadas sobre Weka utilizando las palabras claves establecidas.
- Las fuentes bibliográficas utilizadas deben tener implementado un motor de búsqueda que permita la ejecución de búsquedas avanzadas
- Los estudios tenían que estar escritos en idioma inglés.
- Las bibliotecas digitales seleccionadas para la búsqueda deben combinar como tipos de publicación los especializados en informática y otras ciencias.

Con los criterios establecidos, se realizó la búsqueda de estudios en las bibliotecas de investigaciones digitales siguientes:

- IEEE Xplore (IEEEX)
- Springer

1.3.1. Estrategias de búsqueda

- Se seleccionó una serie de términos y palabras claves para dar respuesta a la pregunta: ¿En qué campos se emplea la herramienta Weka y cuáles son los algoritmos más utilizados? y obtener los resultados requeridos en el presente trabajo.
- La estrategia de búsqueda se basó en las palabras "Weka", "data mining" y “machine learning”.
- La cadena de búsqueda estructurada fue: “Weka” AND (“data mining” OR “machine learning”).
- Se aplicó las cadenas al título y al resumen.
- Cuando el resumen encajaba con el objeto de la investigación, se obtenía y revisaba el artículo en su totalidad.
- El idioma o lenguaje de búsqueda de las publicaciones fue inglés.
- La temporalidad de las publicaciones fue desde el año 2015.

1.3.2. Criterios de inclusión y exclusión

Los criterios de inclusión que se definieron fueron:

- Artículos publicados desde el año 2015. Las aplicaciones informáticas tienen un tiempo de soporte determinado y algunas de ellas dejan de recibir actualizaciones; para aquellas que continúan evolucionando se crean nuevas funcionalidades; por lo tanto, se considera que cualquier documento que estudie una herramienta antes de este año podría no ser útil.
- Artículos de conferencias, revistas y talleres internacionales.
- Cuando un artículo se repite en las bibliotecas digitales, se selecciona a uno de ellos.

Los criterios de exclusión fueron:

- Artículos cuyo contenido no esté relacionado con el uso de Weka.
- Trabajos en diapositivas y libros.
- Trabajos publicados fuera del rango especificado.
- Literatura gris.

1.4. Extracción de información y revisión de trabajos

En esta sección se expone los documentos más relevantes identificados de acuerdo a la relación que tienen con los objetivos de la revisión sistemática y el alcance de la pregunta de investigación.

Una dificultad encontrada al inicio del proceso de extracción de información fue que los términos utilizados en la pregunta condujeron a resultados muy amplios. Por ello, para el proceso de extracción y revisión de los trabajos de investigación se asumió que la calidad de los de los mismos estaría garantizada por la evaluación que realizan las propias fuentes bibliográficas de donde se los tomó, ya que las plataformas generan y presentan sus resultados por orden de relevancia. Para la recopilación se tomaron solo aquellos que estuvieron relacionados a la pregunta de investigación y ubicados en los primeros resultados.

Los documentos más abundantes se obtuvieron mediante la realización basada en cadena de búsqueda en “cualquier campo”. En esa primera fase, se obtuvo una gran cantidad de documentos de investigación, pero esos resultados no fueron los más pertinentes, ya que

consistían en comentarios, cartas o trabajos repetidos que no solo se limitaban a presentar los diversos usos de Weka, sino que se exponían información alejados al objetivo.

En una posterior fase se realizó la búsqueda por los campos “título” y “resumen”, permitió eliminar algunos resultados no útiles, pero todavía no resultaba suficiente para satisfacer la investigación. Finalmente, para obtener la lista definitiva de estudios primarios, se realizó una revisión basada en la cadena de búsqueda definida aplicada con campos seleccionados por “título” y “resumen”, además de seleccionar solo los artículos académicos, llegando a un total de 27 artículos.

2. ANÁLISIS DE RESULTADOS

Los estudios seleccionados y principalmente revisados se muestran en la Tabla 1, con datos que permiten hacer una rápida comparación de información entre ellos. Su orden no determina el grado de importancia con respecto a los objetivos de este trabajo.

Tabla 1. Resumen de artículos relacionados con Weka, detalle del área de aplicación y algoritmos revisados, sus características y el área de aplicación.

TRABAJO	ESTUDIO	ÁREA DE APLICACIÓN	ALGORITMOS REVISADOS
Decision Tree for Manual Material Handling Tasks Using WEKA (Rajesh, Maiti, & Reena, 2018)	Manejo de Materiales	Ergonomía	J48, Random forest, REP, LMT
WekaBioSimilarity—Extending Weka with Resemblance Measures (Domínguez, Heras, Mata, & Pascual, 2016)	Presenta WekaBioSimilarity, una extensión de Weka implementando varias medidas de semejanza para comparar diferentes	Biología	k-means, k-nearest neighbour

	tipos de descriptores.		
Performance improvement of data mining in Weka through multi-core and GPU acceleration: opportunities and pitfalls. (Engel, Charão, Kirsch-Pinheiro, & Steffenel, 2015)	Usando el perfil de rendimiento de Weka, identifica las operaciones que podrían mejorar el rendimiento de la minería de datos cuando se paralelizan.	Informática	M5P, QuickSort
A Classification on Brain Wave Patterns for Parkinson's Patients Using WEKA. (Mahfuz et al., 2015)	Presenta la clasificación de la onda cerebral utilizando datos del mundo real de pacientes con Parkinson en la producción de un modelo emocional.	Medicina	Bayes, multilayer perceptron, K-Means
A comparative study of classification techniques by utilizing WEKA. (Pandey & Rajpoot, 2016)	Análisis comparativo de varios algoritmos de clasificación	Educación	J48, Random Forest, DecisionStump, NaiveBayes, BayesNet.
Classification and prediction based data mining algorithms to predict students' introductory programming performance. (Sivasakthi, 2017)	Predice el rendimiento en el tema introducción a la programación de los estudiantes de primer año de licenciatura en el curso de Aplicación Informática.	Educación	Multilayer Perception, Naive Bayes, SMO, J48, REPTree

Comparison of applications for educational data mining in Engineering Education. (Fernández & Luján-Mora, 2017)	Compara las características técnicas de tres herramientas de código abierto.	Educación	K-means
Student academic performance and social behavior predictor using data mining techniques. (Athani, Kodli, Banavasi, & Hiremath, 2017)	Predice el desempeño y el comportamiento académico de los estudiantes	Educación	Naive Bayes
Identification and Evaluation of Discriminative Lexical Features of Malware URL for Real-Time Classification. (Olalere, Abdullah, Mahmod, & Abdullah, 2016).	Estudio identifica y evalúa las características léxicas discriminatorias de las URL de malware para construir una clasificación de URL de malware en tiempo real.	Informática	Support vector machine (SVM)
Evaluation of the performance of a machine learning algorithms in Swahili-English emails filtering system relative to Gmail classifier. (Omar & Tjahyanto, 2018)	Compara los algoritmos de aprendizaje automático con el clasificador de Gmail.	Informática	Naïve Bayes, Sequential Minimal Optimization (SMO), J48
Preprocessing compressed 3D kinect skeletal joints in enhancing human motion classification. (Lee, Loh, & Chin, 2016)	Se hacen comparaciones de análisis de preprocesamiento de datos de movimiento humanos en	Informática	Bayes Function, Lazy, Tree

	articulaciones esqueléticas 2D y 3D		
Extracting sentiment from healthcare survey data: An evaluation of sentiment analysis tools. (Georgiou, MacFarlane, & Russell-Rose, 2015)	Este proyecto tuvo como objetivo examinar una serie de herramientas analíticas con respecto a su idoneidad para los datos de salud.	Medicina	Naïve Bayes
Rice crop yield prediction in India using support vector machines. (Gandhi, Armstrong, Petkar, & Tripathy, 2016)	Busca mejorar la predicción del rendimiento de los cultivos en diferentes escenarios climáticos.	Agricultura	Sequential Minimal Optimization (SMO)
Vegetation indices based segmentation for automatic classification of brown spot and blast diseases of rice. (Phadikar & Goswami, 2016)	Clasificación de las enfermedades del arroz.	Agricultura	Naïve Bayes updateable, Naïve Bayes, Bayes Net, Part, J48, Decisionstump, LMT, randomforest, Jrip, OneR, FilteredClassifier, Multiclassalassifier, IBK, Logistic LibSVM
Prediction of breast cancer using classification rule mining techniques in blood test datasets. (Muthuselvan & Soma, 2016)	Se evalúa y examina examina los datos recopilados del Instituto de Cáncer Arignar Anna para	Medicina	Naïve Bayes, Zero R, One R, J48, Random Tree

	implementar las Técnicas de Minería de Datos		
Beatmap generator for Osu Game using machine learning approach. (Perkasa & Maulidevi, 2015)	Se plantea una alternativa de un mapa de compás que se considera jugable para el juego Osu utilizando la detección de tiempo y melodía mediante el enfoque de aprendizaje automático	Entretenimiento	Support Vector Machine (SVM)
Decision support systems for predicting diabetes mellitus - A Review. (Vijayan & Anjali, 2015)	Se hace una revisión de los beneficios de las diferentes técnicas de preprocesamiento para los sistemas de apoyo a la toma de decisiones para predecir diabetes	Medicina	Support Vector Machine (SVM), Naive Bayes classifier, Decision Tree.
Towards an Application for Real-Time Travel Mode Detection in Urban Centers. (Soares, de MS Quintella, & Campos, 2017)	Propuesta de una aplicación de detección de modo de viaje en tiempo real basada en rastros de GPS utilizando una técnica de minería de datos	Informática	Support Vector Machine (SVM)
Fault location in radial distribution systems based on decision trees and	Estudio que estima la ubicación de fallas monofásicas en el	Eléctrico	J48

optimized allocation of power quality meters. (da Silva Pessoa & Oleskovicz, 2017)	sistema de distribución IEEE 34-bus		
A cross-platform evaluation of various decision tree algorithms for prognostic analysis of breast cancer data. (Jhajharia, Verma, & Kumar, 2016)	Evalúa el rendimiento relativo de diferentes variantes de un algoritmo de aprendizaje supervisado para implementar un modelo para la evaluación pronóstica de los datos de cáncer de mama.	Medicina	J48
Smart-walk: An intelligent physiological monitoring system for smart families. (Sundaravadivel, Mohanty, Kougianos, Yanambaka, & Ganapathiraju, 2018)	Propone un diseño de sensor acelerómetro que ayuda a rastrear las actividades físicas de familiares y amigos.	Informática	SMO, Gaussian Process, M5 Rules, Decision Table, Linear Regression, Multilayer Perceptron, Additive Regression
Priority based decision tree classifier for breast cancer detection. (Hamsagayathri & Sampath, 2017)	Se analiza el rendimiento de algoritmos basados en prioridad, para la clasificación de cáncer de mama.	Medicina	J48, decision tree classifier
Use of data mining in crop yield prediction. (Mishra,	Implementación del sistema de predicción del	Agricultura	J48, LAD Tree, LWL, IBK.

Paygude, Chaudhary, & Idate, 2018)	rendimiento de los cultivos mediante el uso de técnicas de minería de datos.		
Evaluation of deceptive mails using filtering & WEKA. (More & Kalkundri, 2015)	Evaluación de diferentes métodos supervisados; estudia el impacto de diferentes algoritmos en mensajes engañosos.	Informática	Naïve Bayes, Neural Network, Random Forest, Decision Trees, SVM, IB1, Decision Trees.
New classification system for protein sequences. (Kabli, Hamou, & Amine, 2017)	Presenta un marco global basado en el proceso de extracción de conocimiento a partir de datos biológicos basados en las reglas de asociación.	Informática	PART, One-R, JRip, Decision table
Comparative analysis of breast cancer and hypothyroid dataset using data mining classification techniques. (Verma & Mishara, 2017)	Se discutió la clasificación de las técnicas de minería de datos, a través de conjuntos de datos de cáncer de mama y de hipotiroidismo.	Medicina	Naïve Bayes, MLP
Comparative analysis of classification algorithms on three different datasets using WEKA. (Duriqi, Raca, & Cico, 2016)	Analiza los algoritmos de clasificación más útiles y populares utilizados por los sistemas Machine Learning,	Informática	Naive Bayes, Random Forest, K *

	implementados en Weka.		

Fuente: Elaboración propia.

3. DISCUSIÓN

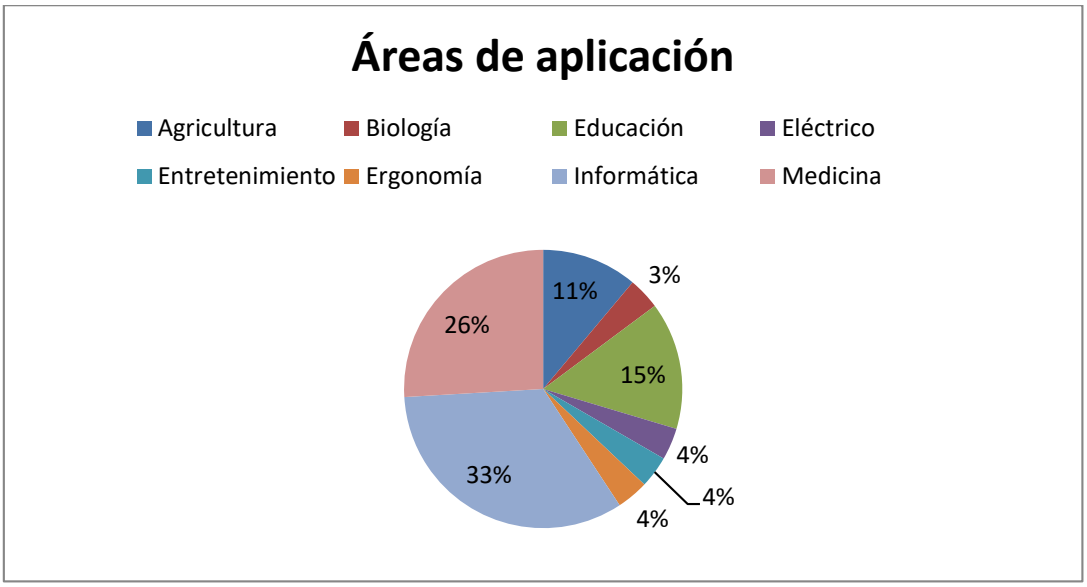
Después de haber determinado, mediante la búsqueda en diferentes bases de datos científicas y académicas, la no existencia de trabajos de investigación que hicieran una revisión sistemática que determine los usos de la herramienta Weka, se procedió a realizar un mapeo a la información recabada. De acuerdo a las preguntas determinadas en el proceso de revisión, a continuación, se presenta una discusión de sus resultados.

El objetivo principal, al realizar la revisión sistemática de literatura y su mapeo, fue identificar en qué campos o áreas de aplicación se está empleando Weka como herramienta de extracción de información, así como los algoritmos más utilizados; por lo que a continuación se resaltan las diferencias y coincidencias de los resultados.

Teniendo en cuenta la complejidad computacional de algunos algoritmos de minería de datos, los análisis de datos pueden tardar horas en completarse, convirtiendo la actividad en improductiva. Además, normalmente los algoritmos proporcionan parámetros que cuando se ajustan pueden mejorar los resultados (por ejemplo, precisión). Sin embargo, a menudo la precisión y el tiempo de cálculo están correlacionados y cuando el primero se requiere alto, el segundo también se ve afectado (Engel, Charão, Kirsch-Pinheiro, & Steffenel, 2015).

Considerando lo expuesto, tal como lo muestra la figura 1, al revisar los artículos científicos primarios se encontró que nueve de ellos están aplicados a la ciencia de la informática, por cuanto demuestran la efectividad de Weka como herramienta computacional de aprendizaje automático y minería de datos. Son muchos los profesionales que están evaluando y evidenciando la aplicación en diversos campos, el comportamiento, rendimiento y la eficacia de los diferentes algoritmos con los que trabaja la herramienta. Se demuestra cómo en el proceso de filtrado de datos, propio de los modelos de minería de datos, se utilizan técnicas de transformación para reducir las instancias insignificantes de los modelos de capacitación; evaluando diferentes algoritmos de clasificación y eligiendo clasificadores basados en el rendimiento de simulación (More & Kalkundri, 2015).

Figura 1. Áreas de aplicación de estudios relacionados con Weka



Fuente: Elaboración propia.

Hoy en día, la minería de datos aplicado al cuidado de la salud es un campo popular, de gran importancia; permite proporcionar predicciones y una comprensión más profunda de los datos médicos. Los autores están utilizando técnicas de minería de datos en el diagnóstico de varias enfermedades, como diabetes, derrames cerebrales, cáncer, enfermedades renales y cardíacas, etc. (Verma & Mishara, 2017). Siguiendo con la revisión del campo de aplicación de la herramienta Weka, se evidencia que en las ciencias médicas se está aprovechando este software estudiado, debido a que se cuenta con amplios datos relacionados a diferentes patologías y existe el interés de poder encontrar modelos de clasificación y pronóstico que tengan un menor error en sus resultados. Las ciencias médicas ocupan un segundo lugar entre las áreas en las que se aplica Weka.

Las nuevas tecnologías han revolucionado el concepto de enseñanza-aprendizaje, por ejemplo, en la educación virtual, en donde la gestión de la interacción entre el profesor/instructor y el estudiante, en la que se basa fundamentalmente, ayuda a la monitorización el progreso de los estudiantes durante el desarrollo de un curso; esto ha dado lugar a una abundante cantidad de información de cada estudiante conocida como big data. Este conjunto de datos que puede superar, en ocasiones, la capacidad del software habitual (Fernández y Ivantchev, 2013).

El programa Weka también está siendo utilizado en el campo educativo, son varios los estudios relacionados a esta herramienta, que ayudan a darle una perspectiva diferente a los múltiples procesos regulares de control y operación, como es la evaluación del desempeño y rendimiento académico de estudiantes en el proceso de enseñanza-aprendizaje. Asimismo, Weka es usado como un instrumento para la prevención de la aparición de factores de riesgo psicosocial en docentes de colegios (Mosquera, Parra y Castrillón, 2016).

Revisando las necesidades de otro sector de aplicación para análisis de datos, como es el agrícola, se reconoce que, debido a factores como el tipo de suelo, la precipitación, la calidad de la semilla, la falta de instalaciones técnicas, etc., el rendimiento de los cultivos se ven directamente afectados. Por lo tanto, las nuevas tecnologías son necesarias para satisfacer la creciente necesidad de mejores prácticas de cultivo y los agricultores deben trabajar inteligentemente optando por nuevas tecnologías en lugar de ir por métodos triviales (Mishra, Paygude, Chaudhary, & Idate, 2018).

La variabilidad en las condiciones climáticas estacionales puede tener un efecto perjudicial, con incidentes de producción en las áreas de cultivo. Desarrollar mejores técnicas para predecir la productividad de los cultivos en diferentes condiciones climáticas puede ayudar al agricultor y a otras partes interesadas a tomar mejores decisiones en términos de agronomía y elección de cultivos. Las técnicas de aprendizaje automático pueden utilizarse para mejorar la predicción del rendimiento de los cultivos en diferentes escenarios climáticos (Gandhi, Armstrong, Petkar, & Tripathy, 2016). Herramientas como Weka están siendo utilizadas en el campo agrícola, por ejemplo, para la determinación y predicción del rendimiento en cultivos de diferentes tipos, bajo variantes en las condiciones climáticas o también en las evaluaciones de las enfermedades a las que pueden verse expuestos.

A través WEKA es posible aplicar algoritmos de minería de datos en un sin número de actividades o necesidades, sea en el campo científico como en el comercial o productivo; por ejemplo, en la localización de las situaciones de cortocircuito monofásico, en busca de una mejora en la continuidad de la fuente de alimentación de los sistemas de distribución eléctricos (da Silva Pessoa & Oleskovicz, 2017); también en la identificación de las principales causas que generan los tiempos muertos en las líneas de producción (Garcés & Castrillón, 2017). La utilización de las herramientas de análisis de datos es tan variada que incluso se las emplea en el área de entretenimiento, una muestra se puede revisar en el estudio de Perkasa & Maulidevi (2015) en el cual buscan la creación de un mapa de compás para un video juego

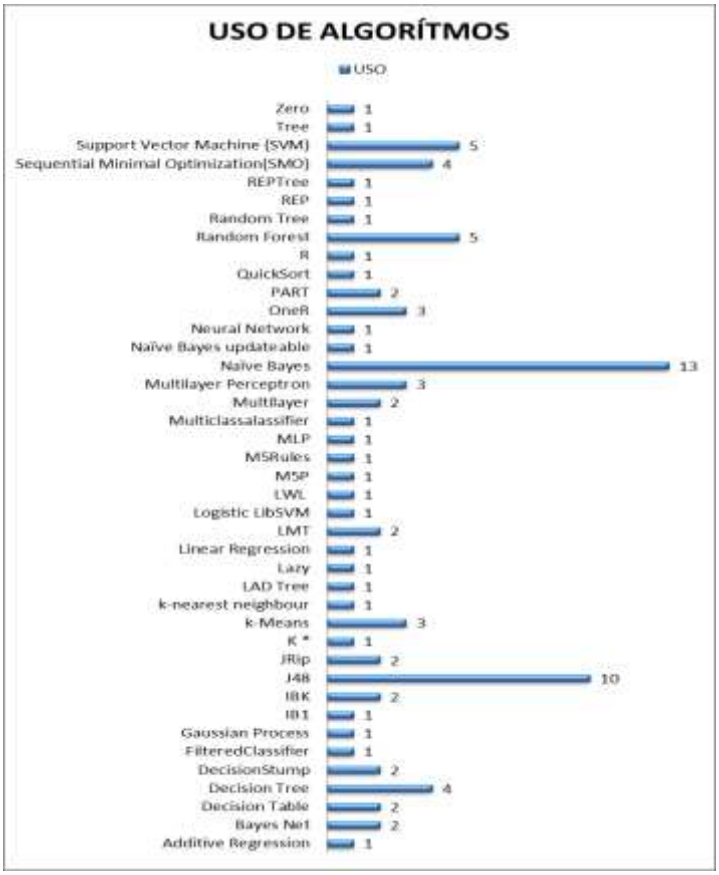
de ritmo utilizando la detección de tiempo y melodía mediante el enfoque de aprendizaje automático.

Weka está disponible para su uso libremente bajo la licencia pública general de GNU, posee una colección de algoritmos de minería de datos con una mayor fortaleza en aquellos para realizar filtros y clasificación, así mismo en algoritmos para el aprendizaje de asociación y el agrupamiento de datos.

Un clasificador permite encontrar la pertenencia de un vector de datos a una clase. La revisión sistemática realizada presenta al algoritmo Naïve Bayes como el más utilizado por los autores de los estudios valorados, ver Figura 2.

El clasificador Naïve Bayes es un clasificador probabilístico fundamentado en el teorema de Bayes y algunas hipótesis simplificadoras adicionales. Naïve Bayes es uno de los algoritmos de aprendizaje inductivo más eficientes y efectivos. Simplifica considerablemente el aprendizaje mediante el supuesto de independencia de los atributos (Pereira, López, & Quintero, 2017).

Figura 2. Algoritmos de Weka evaluados en los estudios



Fuente: Elaboración propia.

Los árboles de decisión proveen de una herramienta de clasificación muy potente. Su uso en el manejo de datos la hace ganar en popularidad dadas las posibilidades que brinda y la facilidad con que son comprendidos sus resultados por cualquier usuario. El árbol en sí mismo, al ser obtenido, determina una regla de decisión (Bouza y Santiago, 2012). Un árbol de decisión como J48 aplicado a un conjunto de datos con una lista de predictores o variables independientes y una lista de objetivos o variables dependientes, permitirá predecir la variable de destino de un nuevo registro de conjunto de datos. El algoritmo J48 de Weka es la segunda herramienta de mayor uso en los estudios revisados.

Los árboles de decisión soportan problemas de clasificación y regresión. El algoritmo trabaja creando un árbol para evaluar una instancia de datos, comienzan en la raíz del árbol y mueven todo a las hojas (raíces) hasta que se pueda realizar una predicción. La creación de un árbol de decisión funciona seleccionando con avidez el mejor punto de división para hacer predicciones y repetir el proceso hasta que el árbol tenga una profundidad fija. El algoritmo árbol de decisión (decision tree) consta de tres elementos fundamentales, nodo raíz, nodo interno y nodo hoja. El más fundamental es el nodo raíz. El nodo hoja es el terminal

fundamental de la estructura y los nodos intermedios se llaman nodo interno. Cada nodo interno denota prueba en un atributo, cada rama representa un resultado de la prueba, y cada nodo hoja tiene una etiqueta de clase (Sewaiwar & Verma, 2015). En cambio, la máquina de aprendizaje radom forest es un meta-aprendiz; significa que consiste en muchos aprendices individuales (árboles); el random forest combina múltiples árboles aleatorios que votan sobre un resultado particular; en el algoritmo random forest cada voto tiene un peso. El bosque clasifica el que contiene más votos (Livingston, 2005). Los algoritmos decision tree y random forest también son empleados en varios de los documentos evaluados en la revisión sistemática, ver figura 2.

El algoritmo Support Vector Machine (SMV) es un método de clasificación supervisada lineal que determina la frontera óptima entre dos grupos que pueden ser linealmente separables o no. Sequential Minimal Optimization (SMO) es otro algoritmo que resuelve un problema que surge en SVM de optimización de una función cuadrática de varias variables, pero sujetas a una restricción lineal de esas variables (Barbona, 2015). De acuerdo a la figura 2, los dos métodos de clasificación SMV y SMO, aunque en menor puntuación que Naïve Bayes y J48, son también de preferencia para realizar estudios con Weka.

CONCLUSIONES

Este trabajo presenta una revisión sistemática de mapeo para la herramienta informática Weka, sus áreas de aplicación y los algoritmos más utilizados; se realizó siguiendo las pautas de (Kitchenham, 2004) y produjo las conclusiones a continuación expuestas.

La tercera parte de los artículos revisados son aplicados a la ciencia de la informática, con este porcentaje, encabeza el listado de las áreas de aplicación de Weka; son muchos los investigadores que hacen referencia a la efectividad, eficacia, comportamiento y rendimiento de Weka como herramienta computacional de aprendizaje automático y minería de datos. Como segunda área de aplicación puntuada en este estudio, se encuentra medicina y salud, Weka es una herramienta muy utilizada para estudiar y mejorar los diagnósticos de varias enfermedades.

En tercer lugar, de campo de aplicación de Weka detectado en la revisión, está el educativo; interviniendo en los procesos regulares de control y operación, con resultados clasificadores y predictivos.

El área Agrícola busca en Weka una alternativa en minería de datos principalmente para mejorar los métodos y estrategias de cultivo. Otras áreas como biología, electricidad, ergonomía y entretenimiento también son exploradas con Weka, dejando claramente establecido que esta herramienta puede aplicarse a cualquier tipo de área en la cual se generen datos para explotar y generar información válida.

Naïve Bayes es el algoritmo más utilizado por los autores de los estudios valorados de la herramienta Weka, le sigue J48, en menor proporción, pero con buenos resultados en su uso también se destacan los algoritmos decision tree, random forest, Support Vector Machine (SMV) y Sequential Minimal Optimization (SMO).

Se debe diferenciar el uso de las herramientas como Weka y la programación de algoritmos específicos para determinados problemas; la programación tiene que ver principalmente con el desarrollo de una solución a un problema específico, que herramientas ya desarrolladas no pueden resolverlos de forma óptima y esto lo que buscan los algoritmos encontrados en los artículos revisados. En el futuro se espera poder realizar una investigación que permita ampliar el conocimiento de diferentes lenguajes de programación que permitan ampliar o complementar este masivo grupo de algoritmos. Adicionalmente se propone llevar esta investigación a una revisión sistemática de la literatura de otras herramientas similares a Weka y que se incluyan más bibliotecas digitales.

REFERENCIAS BIBLIOGRÁFICAS

Abdulla, R., & Tjahyanto, I. A. (2018). Evaluation of the Performance of a Machine Learning Algorithms in Swahili-English Emails Filtering System Relative to Gmail Classifier. *2018 International Conference on Information and Communications Technology (ICOIACT)*. https://doi.org/10.1109/ICOIACT.2018.8350713

Arcila-Calderón, C., Ortega-Mohedano, F., Jiménez-Amores, J., & Trullenque, S. (2017). Análisis supervisado de sentimientos políticos en español: clasificación en tiempo real de tweets basada en aprendizaje automático. *El Profesional de la Información*, *26*(5), 973. https://doi.org/10.3145/epi.2017.sep.18

Athani, S. S., Kodli, S. A., Banavasi, M. N., & Hiremath, M. (2017). Student Academic Performance and Social Behavior Predictor using Data Mining Techniques. Presentado en IEEE International Conference on Computing, Communication and

Automation (ICCCA-2017) on 5th-6th May, 2017: proceeding. https://doi.org/10.1109/CCAA.2017.8229794

Baeza-Yates, R. (2009). Tendencias en minería de datos de la Web. *El Profesional de la Informacion*, *18*(1), 5-10. https://doi.org/10.3145/epi.2009.ene.01

Barbona, I. (2015). Comparación de métodos de clasificación aplicados a textos Científicos y No Científicos. *Revista INFOSUR. Grupo INFOSUR. Rosario*.

Bouza, C. N., & Santiago, A. (2012). La minería de datos: árboles de decisión y su aplicación en estudios médicos. *Modelación Matemática de Fenómenos del Medio Ambiente y la Salud*, *2*.

Corso, C. L. (2009). Aplicación de algoritmos de clasificación supervisada usando Weka. *Córdoba: Universidad Tecnológica Nacional, Facultad Regional Córdoba*.

da Silva Pessoa, A. L., & Oleskovicz, M. (2017). Fault location in radial distribution systems based on decision trees and optimized allocation of power quality meters. En *PowerTech, 2017 IEEE Manchester* (pp. 1–6). IEEE. https://doi.org/10.1109/PTC.2017.7980907

Domínguez, C., Heras, J., Mata, E., & Pascual, V. (2016). WekaBioSimilarity—Extending Wek with Resemblance Measures. En O. Luaces, J. A. Gámez, E. Barrenechea, A. Troncoso, M. Galar, H. Quintián, & E. Corchado (Eds.), *Advances in Artificial Intelligence* (Vol. 9868). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-44636-3

Duriqi, R., Raca, V., & Cico, B. (2016). Comparative analysis of classification algorithms on three different datasets using WEKA. En *2016 5th Mediterranean Conference on Embedded Computing (MECO)*. IEEE. https://doi.org/10.1109/MECO.2016.7525775

Engel, T. A., Charão, A. S., Kirsch-Pinheiro, M., & Steffenel, L.-A. (2015). Performance improvement of data mining in Weka through multi-core and GPU acceleration: opportunities and pitfalls. *Journal of Ambient Intelligence and Humanized Computing*, *6*(4), 377-390. https://doi.org/10.1007/s12652-015-0292-9

Febles Rodríguez, J. P., & González Pérez, A. (2002). Aplicación de la minería de datos en la bioinformática. *Acimed*, *10*(2), 69–76.

Fernández, D. B., & Luján-Mora, S. (2017). Comparison of applications for educational data mining in Engineering Education. En *World Engineering Education Conference (EDUNINE), IEEE* (pp. 81–85). IEEE. https://doi.org/10.1109/EDUNINE.2017.7918187

Fernández-Sainz, A., & Ivantchev, S. (2013). Cómo el aprendizaje automático y el big data pueden ayudarnos con el aprendizaje humano: un experimento con Introducción a la

Econometría. *Revista electrónica sobre la enseñanza de la Economía Pública Págs*, *24*, 35.

Gandhi, N., Armstrong, L. J., Petkar, O., & Tripathy, A. K. (2016). Rice crop yield prediction in India using support vector machines. En *Computer Science and Software Engineering (JCSSE), 2016 13th International Joint Conference on* (pp. 1–5). IEEE. https://doi.org/10.1109/JCSSE.2016.7748856

Garcés, D. A., & Castrillón, O. D. (2017). Diseño de una Técnica Inteligente para Identificar y Reducir los Tiempos Muertos en un Sistema de Producción. *Información Tecnológica*, *28*(3), 157-170. https://doi.org/10.4067/S0718-07642017000300017

Georgiou, D., MacFarlane, A., & Russell-Rose, T. (2015). Extracting sentiment from healthcare survey data: An evaluation of sentiment analysis tools. En *Science and Information Conference (SAI), 2015* (pp. 352–361). IEEE. https://doi.org/10.1109/SAI.2015.7237168

Hamsagayathri, P., & Sampath, P. (2017). Priority based decision tree classifier for breast cancer detection. En *Advanced Computing and Communication Systems (ICACCS), 2017 4th International Conference on* (pp. 1–6). IEEE. https://doi.org/10.1109/ICACCS.2017.8014598

Jhajharia, S., Verma, S., & Kumar, R. (2016). A cross-platform evaluation of various decision tree algorithms for prognostic analysis of breast cancer data. En *Inventive Computation Technologies (ICICT), International Conference on* (Vol. 3, pp. 1–7). IEEE. https://doi.org/10.1109/INVENTIVE.2016.7830107

Kabli, F., Hamou, R. M., & Amine, A. (2017). New classification system for protein sequences. En *Embedded & Distributed Systems (EDiS), 2017 First International Conference on* (pp. 1–6). IEEE. https://doi.org/10.1109/EDIS.2017.8284029

Lee, P. Y., Loh, W. P., & Chin, J. F. (2016). Preprocessing compressed 3D kinect skeletal joints in enhancing human motion classification. En *Computer and Information Sciences (ICCOINS), 2016 3rd International Conference on* (pp. 357–362). IEEE. https://doi.org/10.1109/ICCOINS.2016.7783241

Livingston, F. (2005). Implementing Breiman's random forest algorithm into Weka. En *ECE591Q machine learning conference papers* (Vol. 27). Citeseer.

Mahfuz, N., Ismail, W., Noh, N. A., Jali, M. Z., Abdullah, D., & Nordin, M. J. bin. (2015). A Classification on Brain Wave Patterns for Parkinson's Patients Using WEKA. En A. Abraham, A. K. Muda, & Y.-H. Choo (Eds.), *Pattern Analysis, Intelligent Security and the Internet of Things* (Vol. 355, pp. 21-33). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-17398-6_3

Mishra, S., Paygude, P., Chaudhary, S., & Idate, S. (2018). Use of Data Mining in Crop Yield Prediction. En *Proceedings of the Second International Conference on Inventive Systems and Control (ICISC 2018)*. https://doi.org/10.1109/ICISC.2018.8398908

More, S., & Kalkundri, R. (2015). Evaluation of deceptive mails using filtering & WEKA. En *Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015 International Conference on* (pp. 1–4). IEEE. https://doi.org/10.1109/ICIIECS.2015.7193262

Mosquera, R., Parra-Osorio, L., & Castrillón, O. D. (2016). Metodología para la Predicción del Grado de Riesgo Psicosocial en Docentes de Colegios Colombianos utilizando Técnicas de Minería de Datos. *Información Tecnológica*, *27*(6), 259-272. https://doi.org/10.4067/S0718-07642016000600026

Muthuselvan, S., & Soma, K. (2016). Prediction of Breast Cancer UsingClassification Rule Mining Techniques in Blood Test Datasets. En *International Conference On Information Communication And Embedded System (ICICES 2016)*. IEEE. https://doi.org/10.1109/ICICES.2016.7518932

Olalere, M., Abdullah, M. T., Mahmod, R., & Abdullah, A. (2016). Identification and Evaluation of Discriminative Lexical Features of Malware URL for Real-Time Classification (pp. 90-95). IEEE. https://doi.org/10.1109/ICCCE.2016.31

Pandey, A. K., & Rajpoot, D. S. (2016). A comparative study of classification techniques by utilizing WEKA. En *Signal Processing and Communication (ICSC), 2016 International Conference on* (pp. 219–224). IEEE. https://doi.org/10.1109/ICSPCom.2016.7980579

Pereira, A., López, J., & Quintero, L. (2017). Estudio experimental para la comparación del desempeño de Naïve Bayes con otros clasificadores bayesianos. *Revista Cubana de Ciencias Informáticas*, *11*, 67-84.

Perkasa, D. B., & Maulidevi, N. U. (2015). Beatmap generator for Osu Game using machine learning approach. En *Electrical Engineering and Informatics (ICEEI), 2015 International Conference on* (pp. 77–81). IEEE. https://doi.org/10.1109/ICEEI.2015.7352473

Phadikar, S., & Goswami, J. (2016). Vegetation Indices Based Segmentation for Automatic Classification of Brown Spot and Blast Diseases of Rice. En *3rd Int'l Conf. on Recent Advances in Information Technology*. IEEE. https://doi.org/10.1109/RAIT.2016.7507917

Rajesh, R., Maiti, J., & Reena, M. (2018). Decision Tree for Manual Material Handling Tasks Using WEKA. En P. K. Ray & J. Maiti (Eds.), *Ergonomic Design of Products and Worksystems - 21st Century Perspectives of Asia* (pp. 13-24). Singapore: Springer

Singapore. https://doi.org/10.1007/978-981-10-5457-0_2

Rivero Pérez, J. L. (2014). Técnicas de aprendizaje automático para la detección de intrusos en redes de computadoras. *Revista Cubana de Ciencias Informáticas*, *8*(4), 52–73.

Sewaiwar, P., & Verma, K. K. (2015). Comparative study of various decision tree classification algorithm using WEKA. *International Journal of Emerging Research in Management &Technology*, *4*, 2278–9359.

Sivasakthi, M. (2017). Classification and Prediction based Data Mining Algorithms to Predict Students' Introductory programming Performance. *2017 International Conference on Inventive Computing and Informatics (ICICI)*, 346-350. https://doi.org/10.1109/ICICI.2017.8365371

Smith, T. C., & Frank, E. (2016). Introducing Machine Learning Concepts with WEKA. En E. Mathé & S. Davis (Eds.), *Statistical Genomics* (Vol. 1418, pp. 353-378). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4939-3578-9_17

Soares, E. F. de S., de MS Quintella, C. A., & Campos, C. A. V. (2017). Towards an Application for Real-Time Travel Mode Detection in Urban Centers. En *Vehicular Technology Conference (VTC-Fall), 2017 IEEE 86th* (pp. 1–5). IEEE. https://doi.org/10.1109/VTCFall.2017.8288311

Sundaravadivel, P., Mohanty, S., Kougianos, E., Yanambaka, V., & Ganapathiraju, M. (2018). Smart-Walk: An Intelligent Physiological Monitoring System for Smart Families. En *2018 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE. https://doi.org/10.1109/ICCE.2018.8326065

Valero, C. S. (s. f.). Aplicación de métodos de aprendizaje automático en el análisis y la predicción de resultados deportivos (Application of automated learning methods for analyzing and predicting sports outcomes). *Retos*, (34), 377–382.

Verma, D., & Mishara, N. (2017). Comparative analysis of breast cancer and hypothyroid dataset using data mining classification techniques. En *IEEE International Conference on Power, Control Signals and Instrumentation Engineering (ICPCSI-2017)*. IEEE. https://doi.org/10.1109/ICPCSI.2017.8391987

Vijayan, V., & Anjali. (2015). Decision Support Systems for Predicting Diabetes Mellitus –A Review. En *Proceedings of 2015 Global Conference on Communication Technologies(GCCT 2015)*. IEEE. https://doi.org/10.1109/GCCT.2015.7342631

Weka - University of Waikato. (2018). Recuperado de https://www.cs.waikato.ac.nz/ml/weka/