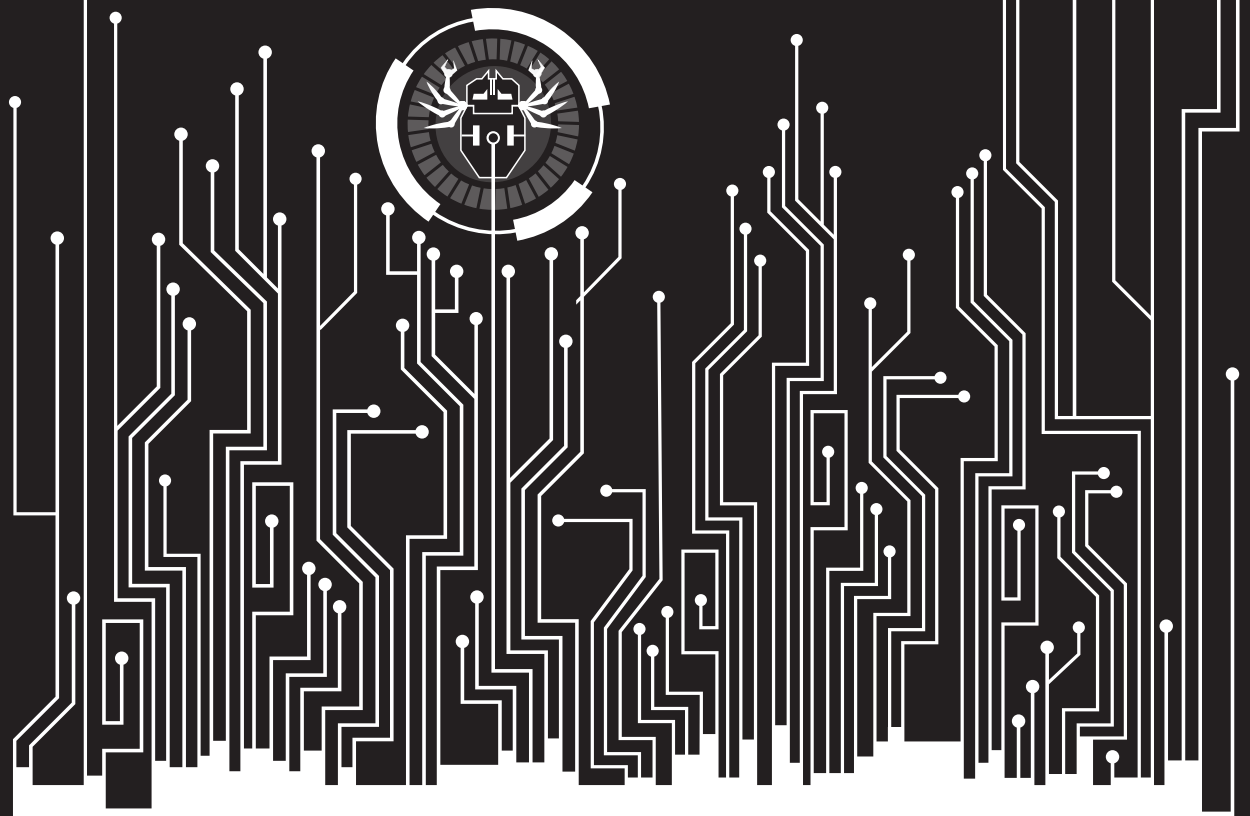
CHRISTIAN CAMILO URCUQUI LÓPEZ
MELISA GARCÍA PEÑA
JOSÉ LUIS OSORIO QUINTERO
ANDRÉS NAVARRO CADAVID



CIBERSEGURIDAD
UN ENFOQUE DESDE LA CIENCIA DE DATOS

UNIVERSIDAD ICESI
Editorial Universidad Icesi

Ciberseguridad: un enfoque desde la ciencia de datos

Ciberseguridad: un enfoque desde la ciencia de datos

Christian Camilo Urcuqui
Melisa García Peña
José Luis Osorio Quintero
Andrés Navarro Cadavid

Editorial Universidad Icesi, 2018

Ciberseguridad: un enfoque desde la ciencia de datos

1 ed. Cali, Colombia. Universidad Icesi, 2018
86 p., 19x24 cm
Incluye referencias bibliográficas
ISBN: 978-958-8936-55-0
https://doi.org/10.18046/EUI/ee.4.2018

1. Machine learning 2. Intelligent systems 3. Cyber security I.Tit
006 – dc22

Facultad de Ingeniería

Rector: Francisco Piedrahita Plata
Decano Facultad de Ingeniería: Gonzalo Ulloa Villegas
Coordinador editorial: Adolfo A. Abadía

Producción y diseño: Claros Editores SAS.
Editor: José Ignacio Claros V.
Diseño de portada: Jhoan Sebastian Pérez.
Impresión: Carvajal Soluciones de Comunicación.

Impreso en Colombia / *Printed in Colombia.*

Editorial Universidad Icesi

Calle 18 #122-135 (Pance), Cali-Colombia
editorial@icesi.edu.co
www.icesi.edu.co/editorial
Teléfono: +57(2) 555 2334

Christian Camilo Urcuqui López

Máster en Informática y Telecomunicaciones e Ingeniero de Sistemas con énfasis en Administración e Informática de la Universidad Icesi (Cali, Colombia), y Especilista en *Deep Learning*. Docente investigador, miembro del Grupo de Investigación en Informática y Telecomunicaciones (i2t) y Coordinador del Club de Hacking de la Universidad Icesi. Sus áreas de interés profesional son: ciberseguridad, ciencia de datos y *machine learning*. ulcamilo@gmail.com

Melisa García Peña

Ingeniera de Sistemas de la Universidad Icesi (Cali, Colombia) actualmente vinculada al Banco de Occidente (Cali). Durante la preparación de su trabajo de grado "Sistemas Open Source para la detección de ataques a páginas web", fue monitora en el Grupo de Investigación en Informática y Telecomunicaciones (i2t), donde participó en el proyecto Sniff, primordialmente en la elaboración de estado del arte sobre *antidefacement*. megape82@hotmail.com

José Luis Osorio Quintero

Ingeniero de Sistemas de la Universidad Icesi (Cali, Colombia), actualmente vinculado al Banco de Occidente (Cali). Durante la preparación de su trabajo de grado "Sistemas Open Source para la detección de ataques a páginas web", fue monitor en el Grupo de Investigación en Informática y Telecomunicaciones (i2t), donde participó en el proyecto Sniff, primordialmente en la elaboración del estado del arte sobre *antidefacement*. josquin@outlook.com

Andrés Navarro Cadavid

Doctor Ingeniero en Telecomunicaciones de la Universidad Politecnica de Valencia (España), Máster en Gestión Tecnológica e Ingeniero Electrónico de la Universidad Pontificia Bolivariana (Medellín, Colombia). Es profesor de tiempo completo y Director del Grupo de Investigación en Informática y Telecomunicaciones (i2t) de la Universidad Icesi (Cali, Colombia). Es Investigador Senior (Colciencias); miembro senior del IEEE y presidente del capítulo Comunicaciones del IEEE Colombia; consultor internacional; y miembro de Grupo de Estudio 1 de la Unión Internacional de Telecomunicaciones. anavarro@icesi.edu.co

Tabla de contenido

Índice de Tablas

Índice de Figuras

Acrónimos

APK	Android Application Package
ASEF	Android Security Evaluation Framework
ASUM-DM	Analytics Solutions Unified Method for Data Mining
AUC	Area Under Curve
AVD	Android Virtual Devices
CDA	Confirmatory Data Analysis
CFG	Control Flow Graph
CRISP-DM	Cross-Industry Standard Process for Data Mining
DoS	Denegation of Services
DT	Decision Tree
EDA	Exploratory Data Analysis
FN	Falsos Negativos
FP	Falsos Positivos
GUI	Graphical User Interface
HTTP	Hypertext Transfer Protocol
IDS	Intrusion Detection System
IoT	Internet of Things
JAR	JAva Archiver

JVM	Java Virtual Machine
KNN	K-Nearest Neighbor
KPI	Key Performance Index
LIBSVM	Library for Support Vector Machines
MLP	Multi-Layer Perceptron
NB	Naïve Bayes
NN	Neural Networks
OWASP	Open Web Application Security Project
PCA	Principal Components Analysis
RF	Random Forest
RFI	Remote File Inclusion
ROC	Receiver Operating Characteristics
SAAF	Static Android Analysis Framework
SEMMA	Sample, Explore, Modify, Model, and Assess
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine
TIC	Tecnologías de la Información y las Comunicaciones
URL	Uniform Resource Locator
VN	Verdaderos Negativos
VP	Verdaderos Positivos
XSS	Cross-Site Scripting
XXE	Entidades Externas XML

Los proyectos base de este documento: "Safe Candy: plataforma de análisis, validación y configuración de seguridad en aplicaciones Android" y "Ciberseguridad 2.0: Mejoramiento de la seguridad de la información a los sitios web de las entidades públicas, por medio de una arquitectura de control informático", fueron implementados gracias al financiamiento del Departamento Administrativo de Ciencia, Tecnología e Innovación (Colciencias), Password Consulting Services y la Univerisidad Icesi.

Resumen

El desarrollo de las tecnologías de la información y las comunicaciones requiere de mecanismos de ciberseguridad que garanticen la confidencialidad, integridad y disponibilidad de la información, y del desarrollo de habilidades para detectar y controlar oportunamente las nuevas amenazas. A pesar del gran desarrollo de las líneas de defensa, la creatividad y la creciente capacidad tecnológica de los *hakers* hacen más compleja la tarea, por lo que metodologías tradicionales (e.g., los sistemas determinísticos basados en perfiles y firmas, y los análisis descriptivos y diagnósticos), no son suficientes.

Con base en los resultados de dos proyectos de investigación en seguridad informática, uno focalizado en la detección de *malware* en dispositivos móviles con sistema operativo Android, el otro en el control de *defacement* en sitios web, se exploró la viabilidad de aplicar la ciencia de datos en el desarrollo de soluciones a problemas de ciberseguridad. En el primer tema, se evaluaron seis clasificadores de *machine learning* para la detección de malware a partir de permisos accedidos, sus resultados mostraron la factibilidad de preparar algoritmos de clasificación capaces de generalizar a otro conjunto de datos, a partir de un conjunto de entrenamiento. En el segundo, luego de generar *datasets* apropiados (con URL benignos y malignos), se seleccionaron algoritmos de clasificación y medidas de evaluación, se realizó un análisis exploratorio de los *datasets* (capas de aplicación y de red), y se revisaron las bondades y limitaciones de los clasificadores propuestos. A partir de sus resultados, se propone un procedimiento –base para la construcción de un *framework*–, con las actividades necesarias para: el entrenamiento y la evaluación de modelos de *machine learning* para la detección de *malware* en dispositivos con sistema operativo Android, y la identificación *a priori* de aplicaciones web maliciosas.

El desarrollo de las tecnologías de la información y las comunicaciones requiere de mecanismos de ciberseguridad que garanticen la confidencialidad, integridad y disponibilidad de la información, y del desarrollo de habilidades para detectar y controlar oportunamente nuevas amenazas. A pesar del gran desarrollo de las líneas de defensa, la creatividad y la creciente capacidad tecnológica de los *hakers* hacen más compleja la tarea, por lo que metodologías tradicionales (e.g., los sistemas determinísticos basados en perfiles y firmas, y los análisis descriptivos y diagnósticos), no son suficientes.

Con base en los resultados de dos proyectos de investigación en seguridad informática, uno focalizado en la detección de *malware* en dispositivos móviles con sistema operativo Android, el otro en el control de *defacement* en sitios web, se exploró la viabilidad de aplicar la ciencia de datos en el desarrollo de soluciones a problemas de ciberseguridad. En el primer tema, se evaluaron seis clasificadores de *machine learning* para la detección de malware a partir de permisos accedidos, sus resultados mostraron la factibilidad de preparar algoritmos de clasificación capaces de generalizar a otro conjunto de datos, a partir de un conjunto de entrenamiento. En el segundo, luego de generar *datasets* apropiados (con URL benignos y malignos), se seleccionaron algoritmos de clasificación y medidas de evaluación, se realizó un análisis exploratorio de los *datasets* (capas de aplicación y de red), y se revisaron las bondades y limitaciones de los clasificadores propuestos. A partir de sus resultados, se propone un procedimiento –base para la construcción de un *framework*–, con las actividades necesarias para: el entrenamiento y la evaluación de modelos de *machine learning* para la detección de *malware* en dispositivos con sistema operativo Android, y la identificación *a priori* de aplicaciones web maliciosas.

Presentación

La ciencia de datos se ha venido aplicando en distintos campos gracias a su capacidad de proveer soluciones aproximadas a problemas no resolubles por sistemas convencionales. Ha tomado fuerza en la medida en que responde adecuadamente a los retos que representan los grandes volúmenes de información que se manejan hoy en día y a las robustas capacidades computacionales requeridas para su uso.

La ciberseguridad requiere de un esfuerzo constante para el desarrollo de soluciones que garanticen, no solo la integridad, disponibilidad y confidencialidad de la información, sino también su capacidad de detección de nuevas amenazas informáticas. Esto representa un reto, tanto para los sistemas, como para los investigadores, por la complejidad de sus variables, especialmente por el desarrollo acelerado de las tecnologías y la notable y creciente astucia y capacidad técnica de los cibercriminales.

La comunidad de desarrolladores e investigadores ha estado trabajando en técnicas, marcos de trabajo (*frameworks*) y soluciones para mejorar los niveles de seguridad de las distintas tecnologías. Entre las propuestas se encuentran: los análisis estático, dinámico e hibrido; la aplicación de la inteligencia artificial; y las metodologías de detección de amenazas cibernéticas, tales como: la identificación por firmas y el descubrimiento a través de anomalías. Un camino prometedor es la aplicación de la ciencia de datos para el desarrollo de soluciones de software, por ejemplo, el uso de modelos predictivos especializados para la detección de *malware* y la predicción de ciberataques web.

En este libro se aborda la aplicación de la ciencia de datos para el desarrollo de soluciones aproximadas a problemas en ciberseguridad a partir de un

recorrido que va desde los fundamentos teóricos de la ciencia de datos y la ciberseguridad, hasta su aplicación en proyectos de investigación. De estos últimos, se tratan dos investigaciones con enfoque de soluciones para defensa: la primera, orientada al análisis de software malicioso para dispositivos con sistema operativo Android; la segunda, a la detección de sitios web con contenido perjudicial.

Estas dos investigaciones surgieron de las necesidades, experiencias y resultados de dos proyectos de ciberseguridad desarrollados por el Grupo de Investigación en Informática y Telecomunicaciones de la Universidad Icesi (i2t) en asocio Password Consulting Services, con financiamiento del Departamento Administrativo de Ciencia, Tecnología e Innovación (Colciencias): "Safe Candy: plataforma de análisis, validación y configuración de seguridad de aplicaciones Android" y "Ciberseguridad 2.0: Mejoramiento de la seguridad de la información a los sitios web de las entidades públicas, por medio de una arquitectura de control informático". Las experiencias de estos dos proyectos han sido incluidas con el fin de resaltar por qué la ciencia de datos es un área importante para el análisis de amenazas cibernéticas.

En el texto se presentan además los estudios más recientes de la aplicación de ciencia de datos en ciberseguridad, las ventajas y desventajas de las técnicas de análisis para detección de software malicioso y los caminos pendientes de explorar en la investigación y el desarrollo de soluciones de seguridad informática.

Finalmente, a partir de la revisión del estado del arte y de las lecciones aprendidas en los proyectos realizados, se propone un marco de trabajo de ciencia de datos para la detección de amenazas digitales en un contexto de investigación, el cual le permite al lector conocer cuáles son las actividades necesarias para el entrenamiento y la evaluación de modelos de *machine learning* para la detección de *malware* en dispositivos con sistema operativo Android y la identificación *a priori* de páginas web maliciosas.

Ciberseguridad y ciencia de datos

Introducción

La seguridad informática o ciberseguridad no es tema reciente, ha presentado retos desde el momento en que los primeros computadores se conectaron a redes de comunicaciones, sin embargo, el crecimiento de la Internet; la popularización de los dispositivos móviles inteligentes; y la aparición y el crecimiento de tecnologías como la Internet de las Cosas (IoT, *Internet of Things*), las ciudades inteligentes y sostenibles, y las redes eléctricas inteligentes, han aumentado su importancia y complejidad. Es ahí donde la ciencia de datos surge como una opción para mejorar los mecanismos de análisis que requieren los sistemas cibernéticos para hacerle frente a los diferentes tipos de riesgos de seguridad que existen en la actualidad. La ciencia de datos, si bien puede ayudar a mejorar la seguridad, también puede servir para erosionarla, por lo que es importante mantener la dinámica de análisis y desarrollo de nuevas estrategias que garanticen el mejoramiento continuo de la seguridad informática. A continuación se presenta la revisión de algunos conceptos relacionados con la ciberseguridad y algunas técnicas de la ciencia de datos y de los sistemas de aprendizaje de máquina (*machine learning*).

Ciberseguridad

La ciberseguridad es el área de las ciencias de la computación encargada del desarrollo y la implementación de los mecanismos de protección de la información y de la infraestructura tecnológica. Debido a la importancia de tener sistemas seguros, que garanticen la confidencialidad, integridad y

disponibilidad de los datos, se han propuesto: métodos para la evaluación de amenazas cibernéticas (técnicas de análisis y marcos de trabajo); prácticas y herramientas para el desarrollo de software y hardware seguros; y sistemas de seguridad, entre otros. Estas propuestas han permitido tener líneas de defensa contra los cibercriminales y el software malicioso, sin embargo, dado el desarrollo de las Tecnologías de la Información y las Comunicaciones (TIC) –con sus nuevas propuestas, como la IoT y el *Big Data*–, la presencia de nuevas vulnerabilidades [1] y los ataques de día cero, la ciberseguridad requiere de un trabajo constante para poder mitigar los riesgos. Existen distintas herramientas y técnicas para el análisis de amenazas cibernéticas, algunas de las más representativas se presentan en la Tabla 1.

Tabla 1. Métodos para el análisis de amenazas cibernéticas

Método	Descripción
Análisis estático	Técnica que evalúa los comportamientos maliciosos en el código fuente, los datos o los archivos binarios, sin ejecutar directamente la aplicación [2]. Su complejidad ha aumentado debido a la experiencia que han adquirido los cibercriminales en el desarrollo de aplicaciones. Se ha demostrado que es posible evadir este análisis a partir de técnicas de ofuscación, como las descritas por Sharif et al. [3].
Análisis dinámico	Métodos automatizados que estudian el comportamiento del *malware* en ejecución mediante un análisis de la interactividad del atacante y permiten evaluar características que solo pueden ser obtenidas mientras el software está en funcionamiento, como por ejemplo: la inyección de código en ejecución [4], los procesos en ejecución, la interfaz de usuario, las conexiones de red y la apertura de sockets [5]. Existen técnicas que permiten evadir este análisis, como las descritas por Petsas et al., [6] donde el *malware* tiene la capacidad de detectar ambientes sandbox y detener su comportamiento malicioso.
Análisis híbrido	Método que combina las ventajas de la aplicación de los análisis dinámico y estático.
Inteligencia artificial	Área que provee de una serie de técnicas para dar soluciones aproximadas a problemas complejos. Una de ellas, el *machine learning* tiene como propósito proveer a los sistemas de la capacidad de aprender cómo identificar a un *malware* sin ser programado de forma explícita. Gran cantidad de propuestas [7-12] usan algoritmos de clasificación, tales como: *Support Vector Machines* (SVM), *Bagging*, *Neural Network* (NN), *Decision Tree* (DT) y *Naïve Bayes* (NB) . Existen otras aplicaciones de la inteligencia artificial, como por ejemplo, las técnicas de programación genética para la detección de anomalías en peticiones HTML.

Para la detección de ciberataques se han propuesto dos marcos de trabajo [13]: el método basado en firmas, que tiene como finalidad detectar amenazas a partir de una base de datos que contiene distintas características (firmas) de peticiones

o archivos maliciosos; y el método de detección por anomalías, que tiene dos actividades, una dedicada a la construcción de un perfil del sitio de análisis a partir de ciertas variables, y otra enfocada en el monitoreo y la detección de anomalías (cambios no registrados) en un perfil previamente creado.

¿Es posible desarrollar un sistema determinista que garantice el cien por ciento de seguridad? Probablemente no, debido a que las soluciones de seguridad informática deben enfrentar variables que hacen de sus retos tareas crecientes en complejidad, entre ellas: el continuo aumento de las habilidades y capacidades de desarrollo de los cibercriminales, los ataques de día cero, las malas prácticas de desarrollo de software y hardware, las nuevas tecnologías, los *insiders* y los requerimientos no funcionales de los sistemas informáticos.

El software convencional de seguridad para la identificación de amenazas cibernéticas requiere de un esfuerzo humano que implica tiempo y recursos, la aplicación de la ciencia de datos es actualmente uno de los caminos más prometedores, porque a través de sus técnicas permite conseguir soluciones aproximadas a problemas complejos [14].

Ciencia de datos

La ciencia de datos tiene como objetivo obtener elementos de valor de distintas fuentes de información –incluso datos que pueden ser incompatibles y estar en distintos formatos–, a través de técnicas y herramientas que incluyen métodos de estadística, minería de datos, *machine learning* y visualización. Esta ciencia busca encontrar y entender patrones en los datos con el fin de generar modelos que representan al contexto de la información.

Un modelo es una representación general de las relaciones entre los atributos de los datos y una función matemática, estadística o una serie de reglas que asocia la información. Existen dos tipos de modelos: descriptivos, que tienen como objetivo proveer mayor información acerca del contexto de los datos –relaciones causa efecto–; y predictivos, encargados de estimar un objetivo a partir de una serie de valores.

En la ciencia de datos existen dos enfoques de análisis que dependen del nivel de conocimiento del científico acerca de la información: el análisis exploratorio de los datos (EDA, *Exploratory Data Analysis*), que tiene como finalidad conocer las relaciones o patrones que existen en los datos, aplicable cuando de antemano

no se tiene una hipótesis o un entendimiento de ellos; y análisis de datos confirmatorio (CDA, *Confirmatory Data Analysis*), que asume que el científico tiene una hipótesis acerca de la información y tiene como objetivo probarla o descartarla a partir de los modelos creados.

El EDA busca entender las relaciones y la calidad de los datos a partir de la extracción de atributos cuantitativos y de la producción de indicadores y resúmenes visuales basados en la aplicación de métodos estadísticos.

Los análisis numéricos hacen parte de la estadística descriptiva, la cual se subdivide en tres categorías: medidas de tendencia central, medidas de dispersión y medidas de asociación: las medidas de tendencia central permiten entender cómo los datos están organizados alrededor del centro de la distribución y cuáles son los valores de mayor ocurrencia (e.g., la media, la mediana y la moda); las medidas de variación o dispersión calculan las distancias entre los datos, es decir, proveen los mecanismos para determinar qué tan esparcidos del centro están (e.g., el rango, el rango intercuartil, la varianza y la desviación estándar); las medidas de asociación permiten conocer la existencia y la fuerza de la relación entre las variables (e.g., la correlación y la covarianza).

Los modelos en la ciencia de datos son de dos tipos: modelos basados en la estadística, desarrollados a partir del análisis de la distribución de los datos y de la probabilidad de la predicción de posibles resultados a partir de una ecuación matemática –que debe ser descubierta– y de los parámetros que mejor estén relacionados a partir de los datos analizados; y modelos basados en algoritmos de *machine learning*, cuyo objetivo es encontrar los datos mejor relacionados (patrones y reglas) y evaluar los resultados de los algoritmos que mejor se ajusten a la solución del problema.

Actualmente se pueden encontrar varias propuestas metodológicas para la aplicación de la ciencia de datos, entre ellas: CRISP-DM (*Cross-Industry Standard Process for Data Mining*), ASUM-DM (*Analytics Solutions Unified Method for Data Mining*) y SEMMA (*Sample, Explore, Modify, Model, and Assess*). Aunque cada una tiene un punto de vista distinto, su objetivo es el mismo: el aprovechamiento de la información.

Para un contexto de grandes volúmenes de información (*Big Data*), se plantea el ciclo de vida de la analítica de datos [14], que se presenta en la Tabla 2. Las etapas son secuenciales y se describen en su orden, salvo la etapa de análisis de datos, la que por su naturaleza es iterativa a sí misma.

Tabla 2. Ciclo de vida de la analítica de datos en *Big Data* [14]

Etapa	Descripción
Evaluación del *Business Case*.	Se enuncian y definen las necesidades, justificaciones y motivaciones del negocio para la definición de los objetivos del análisis. Esta iteración es importante ya que permite obtener una primera idea acerca de los distintos recursos y retos del proyecto, por ejemplo: la identificación de indicadores clave de rendimiento (KPI, *Key Performance Index*), para la futura evaluación de los modelos de datos; las bases de datos internas y externas; los procesos, las tecnologías y la inversión.
Identificación de datos.	Una vez evaluadas las distintas variables del negocio, se procede a la identificación de las fuentes de datos externas e internas. Para datos externos (especialmente de texto, e.g., blogs) puede ser necesario contar con herramientas automatizadas que permitan recolectar la información.
Adquisición y filtrado de datos.	Se recogen los datos desde todas las fuentes que se identificaron durante la etapa anterior y se someten a un filtrado automático para eliminar los datos corruptos y los que no tienen valor para los objetivos de análisis.
Extracción de datos.	Se extraen los datos que estén en un formato no compatible con la solución *Big Data* y se llevan a un formato que ella pueda usar para su análisis.
Validación y limpieza de datos.	Se limpian los datos erróneos a través de reglas o mecanismos preestablecidos y se valida que los datos estén completos y no presenten incoherencias que vayan a dificultar los futuros análisis.
Adición y representación de los datos.	Se llevan a un mismo contexto las distintas fuentes de información, de tal manera que los registros obtenidos, internos y externos, tengan el mismo significado y estén en sus respectivas representaciones (e.g., las tablas en las bases de datos).
Análisis de los datos.	Esta es una actividad iterativa a sí misma, busca aplicar análisis exploratorio o confirmatorio (EDA o CDA) a través del análisis, el entrenamiento y la validación de los resultados de distintos modelos.
Visualización de los datos.	Para que la capacidad de analizar grandes cantidades de datos y encontrar información útil no esté limitada a los analistas, se usan técnicas y herramientas para comunicar gráficamente los resultados del análisis y facilitar su interpretación efectiva por parte de usuarios no expertos.
Utilización de los resultados.	Para obtener el mayor provecho de los datos en términos de su aporte a la toma de decisiones, se define cómo y dónde lo hallado puede ser utillizado.

El análisis con visualización incluye algunas herramientas que facilitan la abstracción de la información, las más usuales son: los mapas de calor, el análisis de series de tiempo y el análisis de redes. El mapa de calor permite apreciar la cantidad de datos cualitativos presentes en unas áreas específicas representadas con colores en matrices o mapas geográficos; el análisis de series de tiempo facilita la comprensión y abstracción de patrones de datos que han sido registrados en intervalos de tiempo; y el análisis de redes permite comprender las relaciones que existen entre los nodos de una red.

Los algoritmos de *machine learning* se utilizan para encontrar soluciones aproximadas a distintos problemas y analizar la información. Existen algoritmos de clasificación, algoritmos de agrupamiento, algoritmos de detección de datos atípicos y algoritmos de filtrado.

Los algoritmos de clasificación a través del aprendizaje supervisado permiten predecir las categorías de una variable nominal u ordinal de un conjunto de datos; los algoritmos de agrupamiento están basados en el aprendizaje no supervisado y buscan encontrar patrones en los datos a través de su segmentación en distintos grupos, es decir, buscan crear agrupaciones de datos que se encuentran relacionados a través de sus propiedades; los algoritmos de detección de datos atípicos tienen como finalidad proveer los mecanismos para la identificación de anomalías o de irregularidades en la información, sean ellas favorables o no al contexto en que se aplica el análisis; y los algoritmos de filtrado buscan y ofrecen pequeñas cantidades de registros provenientes de un gran conjunto de datos por medio de la identificación del comportamiento (patrón) de un usuario (filtrado basado en contenido) o de varios (enfoque colaborativo).

El análisis semántico busca obtener elementos de valor por medio de la extracción de los datos que se encuentran en formato de texto y del reconocimiento de voz, incluye el procesamiento de lenguaje natural, la analítica de texto y el análisis de sentimiento. En la aplicación del procesamiento de lenguaje natural se automatiza el reconocimiento de voz y se ejecutan acciones computacionales dependiendo de los resultados obtenidos; en la analítica de texto se buscan elementos de valor en archivos textuales (datos no estructurados), como correos electrónicos, registros y mensajes en redes sociales; y en el análisis de sentimiento, se busca información acerca de los sentimientos de las personas y su intensidad a través del análisis de texto, sus resultados usualmente son utilizados para la toma de decisiones de una empresa o para la sistematización respuestas dado un texto de entrada.

Machine Learning

La inteligencia artificial presenta distintas definiciones,las cuales varían en dos dimensiones [15]: la primera, orientada a los procesos de pensamiento y razonamiento; y la segunda, encaminada al comportamiento. Estas dimensiones se enfocan al estudio de la fidelidad del rendimiento humano o al concepto de racionalidad.

Machine learning es la parte de la inteligencia artificial que busca que un sistema tenga la capacidad de aprender en entornos variables, sin que sea programado de forma explícita. Su uso ha venido creciendo debido a los volúmenes de información que se presentan en los medios electrónicos (*Big Data*) y a las mayores capacidades computacionales.

Dependiendo del problema a abarcar, se pueden emplear tres técnicas para el entrenamiento de los modelos: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo [16]. En el aprendizaje supervisado se conoce el conjunto de datos, principalmente se comprende la salida correcta del algoritmo y su relación con la entrada, abarca problemas de regresión y clasificación; en el aprendizaje no supervisado no se conoce el conjunto de datos, por lo que su objetivo es encontrar la estructura y los patrones presentes en la información; en el aprendizaje por refuerzo se busca que el algoritmo esté en capacidad de encontrar el conjunto de operaciones más adecuadas para cumplir un objetivo, a través del aprendizaje de reglas y acciones.

Los algoritmos de *machine learning* pueden aplicarse en distintos contextos, dependiendo de su tipo de aprendizaje: clasificación, para predecir la categoría de un ítem; regresión, para predecir un valor continuo; detección de anomalías, para identificar datos atípicos; agrupamiento, para encontrar grupos de elementos similares; asociación, para encontrar reglas de concurrencia; secuencia, para encontrar sucesiones de eventos; resumen, para simplificar la representación de una información; y visualización, para facilitar la comprensión y el descubrimiento.

En los problemas de clasificación existen distintas medidas que permiten evaluar la eficacia de los algoritmos (Tabla 3), entre las que podemos encontrar cuatro muy usuales que hacen parte de la tabla de confusión (Tabla 4) y otros indicadores que facilitan la elección del modelo que mejor cumple con el objetivo del proyecto (Tabla 5).

Existen otros mecanismos de valoración de los algoritmos, como por ejemplo: las Características Operativas del Receptor (ROC, *Receiver Operating Characteristics*), que permite comparar el conjunto de medidas encontradas; el Área Bajo la Curva (AUC, *Area Under Curve*), que permite conocer visualmente la eficacia del clasificador [16]; y el coeficiente de kappa de Cohen, de gran ayuda para aislar el efecto del azar de los datos observados y conocer así el desempeño de los modelos que fueron entrenados a través de un desbalance en

Tabla 3. Medidas de evaluación de la eficacia de los algoritmos de *machine learning*

Tipo	Medida	Descripción
Confusión	Verdaderos Positivos (VP).	Tasa de instancias identificadas correctamente y que hacen parte de su respectiva clase.
	Falsos Negativos (FN).	Tasa de instancias que se identificaron incorrectamente, pero que no hacen parte de una clase especifica.
	Falsos Positivos (FP).	Tasa de datos que fueron identificadas incorrectamente y que pertenecen a una clase especifica.
	Verdaderos Negativos (VN).	Tasa de instancias que fueron identificadas correctamente, pero que no pertenecen a una clase especifica.
Desempeño	Sensitividad.	Probabilidad de obtener un verdadero positivo.
	Error.	Tasa de instancias incorrectamente identificadas de todos los datos estudiados.
	Exactitud (*Accuracy*).	Proporción de datos que fueron correctamente identificados a través de todas las instancias utilizadas.
	Especificidad.	Probabilidad de obtener un verdadero negativo.
	Recuperación (*recall*).	Proporción de datos correctamente clasificados contra el total de datos de su clase.
	Precisión.	Tasa de datos identificados que son realmente relevantes.

Tabla 4. Medidas de confusión para problemas de dos clases [16]

Clase verdadera	Clase predicta		Total
	Positiva	Negativa	
Positiva	Verdadero Positivo (VP)	Falso Negativo (FN)	P
Negativa	Falso Positivo (FP)	Verdadero Negativo (VN)	N
	p’	n’	N

Tabla 5. Medidas de desempeño para problemas de dos clases [16]

Medida	Fórmula
Sensitividad	tn / n =1-fp rate
Error	(fp+fn) / N
Exactitud (*Accuracy*)	(tp+tn) / N=1-error
Especificidad	tp / p= tp rate
Recuperación (*recall*)	tp / p =p
Precisión	tp / p’

la variable objetivo. El tiempo de respuesta, el número de niveles de un árbol de decisión y el número de neuronas o de niveles en una red neuronal entre otras, son también técnicas útiles para evaluar modelos basados en *machine learning*.

Durante el proceso de entrenamiento y evaluación de *machine learning* se debe evitar incurrir en: el sesgo (bias), una medida que indica qué tan lejos está el modelo de la verdad o de la solución del problema, e ilustra la variación de las predicciones respecto de una instancia; y el sobreajuste (*overfitting*), un fenómeno que se presenta cuando los algoritmos son entrenados y ajustados a una proporción de los datos, es decir, cuando los modelos no cuentan con la capacidad de generalizar a otra información que pertenece al contexto del problema y que no fue parte de los procesos de entrenamiento y validación. En la Tabla 6 se describen algunos algoritmos de uso común como clasificadores.

Tabla 6. Algoritmos de clasificación

Algoritmo	Descripción
Naïve Bayes	Clasificador probabilístico para aprendizaje supervisado basado en la aplicación del teorema de Bayes, que asume la independencia entre cada par de variables utilizadas. Muy útil para conjuntos de datos grandes, simple, rápido y muy buen método clasificador. Tiene buen rendimiento con variables categóricas. Para las variables numéricas usa una distribución normal. Durante el estudio McCallum y Nigam [17] se utilizaron dos tipos de Naïve Bayes: Gaussian y de Bernoulli, cuya diferencia radica en el tipo de distribución de los datos (Gaussiana o de Bernoulli, respectivamente).
Bagging	Su finalidad es combinar el resultado de la predicción de múltiples clasificadores aplicados a un remuestreo del conjunto inicial. Hace parte de los métodos de ensamblado que permiten reducir la varianza y el sobre aprendizaje.
K Nearest Neighbours	Algoritmo supervisado que clasifica a cada dato del conjunto en la clase más frecuente de sus k vecinos más cercanos.
Support Vector Machines (SVM)	Construye un modelo con un plano de puntos dividido en dos subconjuntos a partir de un conjunto de puntos, una función (lineal, polinómica, de base radial o sigmoide) y de otros parámetros, para con base en él predecir si un nuevo punto desconocido pertenece a alguna de esas dos categorías [18].
Stochastic Gradient Descent	Algoritmo de optimización que implementa una rutina de aprendizaje de primer orden. El método reduce el costo de las funciones lineales por cada iteración que se produce sobre el conjunto de entrenamiento.
Decision Tree	Algoritmo para problemas de clasificación que crea un modelo con una estructura lógica (árbol) a partir de las características de los datos de entrenamiento.
Regresión logística	Algoritmo que intenta predecir una característica categórica a partir de otras variables independientes.
C.45	Algoritmo que genera reglas para la predicción de variables objetivas con la ayuda de un algoritmo de árbol de clasificación [19]. Su aplicación en el lenguaje R o la plataforma Weka es posible a través de una librería de J48.

Ciencia de datos y ciberseguridad

El software convencional de seguridad requiere de un esfuerzo humano para identificar vulnerabilidades a través de un proceso que permita encontrar sus características y desarrollar la solución sobre la herramienta. Esta es una labor que puede ser más eficiente si se aplica un proceso de análisis a través de la ciencia de datos y de los algoritmos de *machine learning* [20]. La importancia de la aplicación de la inteligencia artificial en la seguridad informática ha sido destacada en varios trabajos, algunos de ellos se relacionan a continuación.

Gandotra, Bansal y Sofat [21] proponen un marco de trabajo para clasificar los *malware* en MS-Windows a través de las características extraídas de los análisis estático y dinámico; en su estudio emplearon como algoritmos de clasificación: MultiLayer Perceptron (MLP), IB1, Decision Tree (DT) y Random Forest (RF) y concluyeron que es posible conseguir buenos resultados empleando en conjunto la información de los análisis estático y dinámico.

Rieck [22] evaluó la aplicación de *machine learning* en la seguridad informática y enunció los problemas, retos y perspectivas de trabajar con ciencia de datos y ciberseguridad en conjunto. Indica además que para aplicar ambas áreas en un sistema, este debe ser efectivo, eficiente, transparente, controlable y robusto. Como parte de las ventajas de contar con un sistema de seguridad con *machine learning* menciona: la detección proactiva de los ataques, el análisis automático de amenazas y el descubrimiento asistido de vulnerabilidades. Concluye que hay buenas esperanzas de este enfoque de trabajo con ambas áreas y propone seguir fomentando investigaciones que hagan uso de ellas.