
# **Detectar accesos no autorizados tempranos a través del análisis de pistas de auditoria generadas por la plataforma SIIPNE3W de la Policía Nacional usando aprendizaje automático**.

Silva Yumi, Medardo Angel

Vicerrectorado de Investigación, Innovación y Transferencia de Tecnología

Centro de Postgrados

Maestría en Gestión de Sistemas de Información e Inteligencia de Negocios

Trabajo de titulación, previo a la obtención del título de Magister en Gestión de Sistemas de Información e Inteligencia de Negocios

Mgs. Zaldumbide Proaño, Juan Pablo

16 de Diciembre del 2021

COPYLEAKS

## Tesis Medardo_obsCorregidas.docx

Scanned on: 14:13 January 27, 2022 UTC


Overall Similarity Score


Results Found


Total Words in Text

| | |
|---|---|
| Identical Words | 822 |
| Words with Minor Changes | 185 |
| Paraphrased Words | 163 |
| Ommited Words | 0 |

Firmado electrónicamente por:
JUAN PABLO ZALDUMBIDE PROANO

**VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN**

**Y TRANSFERENCIA DE TECNOLOGÍA**

**CENTRO DE POSGRADOS**

## CERTIFICACIÓN

Certifico que el trabajo de titulación, "**Detectar accesos no autorizados tempranos a través del análisis de pistas de auditoria generadas por la plataforma SIIPNE3w de la Policía Nacional usando aprendizaje automático**" fue realizado por el señor **Silva Yumi, Medardo Angel,** el mismo que ha sido revisado y analizado en su totalidad, por la herramienta de verificación de similitud de contenido; por lo tanto cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí, 16 de diciembre de 2021

Firma:

Firmado electrónicamente por:
JUAN PABLO ZALDUMBIDE PROANO

MSC. Zaldumbide Proaño, Juan Pablo

Director

**C.C.:** 1715467948

**VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN**

**Y TRANSFERENCIA DE TECNOLOGÍA**

**CENTRO DE POSGRADOS**

## RESPONSABILIDAD DE AUTORÍA

Yo **Silva Yumi, Medardo Angel**, con cédula de ciudadanía n° 0603176348, declaro que el contenido, ideas y criterios del trabajo de titulación: **Detectar accesos no autorizados tempranos a través del análisis de pistas de auditoria generadas por la plataforma SIIPNE3W de la Policía Nacional usando aprendizaje automático** es de mí autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

**Sangolquí, 16 de diciembre del 2021**

Firma

MEDARDO
ANGEL SILVA

.....

Silva Yumi, Medardo Angel

**C.C.:** 0603176348

**VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN**

**Y TRANSFERENCIA DE TECNOLOGÍA**

**CENTRO DE POSGRADOS**

## AUTORIZACIÓN DE PUBLICACIÓN

Yo **Silva Yumi, Medardo Angel,** con cédula de ciudadanía n° 0603176348, autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: **Detectar accesos no autorizados tempranos a través del análisis de pistas de auditoria generadas por la plataforma SIIPNE3W de la Policía Nacional usando aprendizaje automático** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi responsabilidad.

**Sangolquí, 16 de diciembre del 2021**

Firma

Firmado electrónicamente por:
MEDARDO ANGEL SILVA

.....

Silva Yumi, Medardo Angel

**C.C.**: 0603176348

## Dedicatoria

El presente trabajo de titulación se lo dedico a mis padres y aquellas personas que día a día laboran para mejorar y mantener los sistemas informáticos funcionales, a pesar de ser una labor invisibilizada, su vocación y profesionalismo los impulsan a generar mejores ideas y soluciones para enfrentar los retos tecnológicos.

**Agradecimiento**

A todos los profesionales que impulsan el acceso a la información y herramientas de manera libre y su predisposición a compartir su conocimiento de modo sencillo a través de diversos medios como blogs, publicaciones, traducciones, repositorios académicos, etc.

A la institución policial por facilitar la labor de la presente investigación orientada en aunar esfuerzos para brindar un mejor servicio día con día a través de su plataforma tecnológica en beneficio de sus uniformados y ciudadanía en general.

**Índice de Contenido**

**Índice de Tablas**

**Índice de Figuras**

**Resumen**

Los delitos informáticos cada vez son más frecuentes en la actualidad, uno de dichos delitos constituye el acceso no consentido a un sistema informático el cual ha tenido un crecimiento notable de denuncias en los últimos años.

El objetivo del presente proyecto es predecir éste tipo de accesos inusuales a través del análisis de información de las pistas de auditoria de la plataforma SIIPNE3w de la Policía Nacional del Ecuador haciendo uso de herramientas de Aprendizaje Automático.

El propósito del estudio es comparar algoritmos de clasificación supervisado y no supervisado, su precisión/sensibilidad en la detección de anomalías (accesos no autorizados o inusuales) para determinar cuál otorga mejor resultado. Los algoritmos elegidos incluyen: árbol de decisiones y bosque de aislamiento. El etiquetado de datos necesario para el algoritmo supervisado se lo alcanzó utilizando clúster k-means.

Para una adecuada gestión se utilizan el estudio de caso como metodología de investigación y CRISP-DM para el desarrollo de los modelos de aprendizaje.

Como resultado se obtiene que la etiquetación de datos juega un papel importante al momento de la clasificación para los algoritmos supervisados, además que el algoritmo de bosque de aislamiento nos brinda mejor resultado de clasificación para el caso que se estudia.

Palabras clave:

- **DETECCIÓN DE INTRUSOS**
- **PISTAS DE AUDITORIA**
- **APRENDIZAJE AUTOMATICO**
- **APRENDIZAJE SUPERVISADO**

**Abstract**

Nowadays cybercrime are more frequently every time, one of them is the informatics system unauthorized access who has been having a significant growing of complaints in the last years. The target of the project is to predict those unauthorized access through information's analysis from audit tracks produced by the SIIPNE3w platform of the Ecuador National Police using machine learning tools.

The study aim is to compare the supervised and not supervised classification algorithms, its accuracy/sensibility to detect outliers (unauthorized access) and determine which of them offer a better result. The used algorithms include: Decision Tree and Isolation Forest. Data annotation for supervised algorithm is performed with k-means clustering.

For an appropriate management of the project were used Study Case as a research methodology and CRISP-DM in the development of the machine learnings models.

As a result was obtained that annotation is an important factor for classification in supervised algorithms, also that the Isolation Forest algorithm give us a better classification result for the study case.

Key words:

- **INTRUSION DETECTION**
- **AUDIT TRAILS**
- **MACHINE LEARNING**
- **SUPERVISED LEARNING**

# Capítulo I

## Problema de la Investigación

### Antecedentes

Varios casos de ataques cibernéticos desarrollados en los últimos tiempos han puesto en evidencia la vulnerabilidad de los sistemas de información de algunas entidades privadas como públicas.

Entre los casos más conocidos podemos destacar los intentos de ataques cibernéticos registrados a raíz de la detención de Julian Assange, los cuales alcanzaron la cifra de los 8000 intentos según registros del MINTEL. (DATTA, 2019). Adicional a dicho caso tenemos la exposición de información sensible de casi toda la población ecuatoriana de un servidor administrado por la empresa Novaestrat dedicada a las actividades de marketing, desarrollo de software y análisis de datos. (VpnMentor, 2019)

Los sistemas de información (SIIPNE3w) de la Policía Nacional del Ecuador también han sido víctimas de uno de estos tipos de ataques, siendo el acceso no autorizado uno de los más frecuentes, en su mayoría perpetrado por parte de usuarios internos.

Por esta clase de riesgos producidos más que todo por vulnerabilidades debidas a la complejidad, deficiencia o falta de controles que los sistemas como tal presentan, es que los mencionados suelen utilizar como elemento de control las denominadas bitácoras o pistas de auditoria, cuya importancia radica en que permiten detectar comportamientos inusuales, recuperar información, obtener datos para la atención de problemas, generar evidencia de auditoría legal y forense, entre otras.

Existen variadas herramientas en el mercado para automatizar el proceso de análisis de la información almacenada en dichas bitácoras, pero muchas no son útiles en ciertos casos, por

lo que el análisis de este tipo de mecanismos suele realizarse a pie o de manera particular elaborando herramientas para cada tipo de bitácora. (Guido, 2012)

Los avances en la tecnología en materia de inteligencia artificial, machine learning, minería de datos, técnicas de clusterización, regresión, podrían facilitar la extracción de valor de la información de manera temprana en el caso de estudio que se pretende desarrollar.

**Contexto del Problema**

La plataforma SIIPNE3w de la Policía Nacional del Ecuador constituye un sistema de información integral en la cual se manejan diversos procesos concernientes a la gestión tanto interna como externa. La mayoría de sus usuarios deben ser miembros de dicha institución o laborar dentro de ella para poder acceder a los recursos o servicios digitales que brinda dicho sistema de información.

La Institución en su afán de mejora continua se encuentra migrando en ciertos casos hacia esta plataforma orientada a la web todos aquellos sistemas de información que han quedado descontinuados y sin soporte tecnológico, y en otros desarrollando nuevos módulos de tal manera de contar con una sola fuente de información fiable y válida.

Sin embargo esta integración de información en un repositorio central conlleva el estar al pendiente de la seguridad de dichos datos con el propósito de prevenir y evitar ataques como el acceso no autorizado a su sistema de información utilizando diversas técnicas entre las cuales tenemos aquellas que incluyen el análisis de datos de sus bitácoras o pistas de auditoria.

**Planteamiento del problema**

Es común que la gran mayoría de las personas dedicadas al análisis, control u obtención de información de las bitácoras de la plataforma SIIPNE3w en la actualidad efectúen a pie el análisis de datos o en el mejor de los casos desarrollando un analizador particular para cada tipo de bitácora.

La falta de recursos, obsolescencia de equipos, capacitaciones deficientes, aplicación de métodos no adecuados, etc., suelen ser algunas de las causas que motivan a mantener dicho habito de trabajo, lo cual se ve reflejado en la obtención tardía de valor de la información recopilada en sus pistas de auditoria.

Uno de los casos constituye los accesos no autorizados o inusuales de sus usuarios en su mayoría internos a su sistema de información denominado SIIPNE3w con el fin de obtener provecho de la información a la cual tienen acceso, los cuales han sido detectados posterior a que se ha efectuado el hecho inusual o cuando los procesos investigativos así lo han requerido.

## Objetivos

### *Objetivo general*

Desarrollo de un modelo de aprendizaje automático para detectar accesos no autorizados de manera temprana a través del análisis de pistas de auditoria generadas por la plataforma SIIPNE3w de la Policía Nacional.

### *Objetivos Específicos*

**OE1:** Realizar una revisión de la literatura para determinar posibles soluciones y recomendaciones existentes para detectar accesos no autorizados de manera temprana a través del análisis de pistas de auditoria utilizando herramientas de aprendizaje automático.

**OE2:** Aplicación de algoritmos supervisados para la detección de accesos no autorizados de manera temprana a través del análisis de pistas de auditoria.

**OE3:** Evaluar los resultados obtenidos mediante métricas comunes de aprendizaje automático, para verificar la eficiencia del modelo usado para la detección de accesos no autorizados.

**Hipótesis de investigación**

El uso de aprendizaje automático como herramienta de Business Intelligence permitirá detectar accesos no autorizados de manera temprana en la plataforma SIIPNE3w de la Policía Nacional a través del análisis de sus pistas de auditoria.

Señalamiento de Variables:

Variable Dependiente: Accesos no autorizados en SIIPNE3w.

Variable Independiente: Aplicación de un modelo supervisado para detectar accesos no autorizados.

La demostración de la hipótesis planteada se realizará mediante el análisis y evaluación de los datos obtenidos del modelo de aprendizaje automático que se aplique.

**Justificación, importancia y alcance del proyecto**

Según (Jackson et al., 1992), aproximadamente un 80% de todos los fraudes, robos, sabotajes o accidentes relacionados con los sistemas informáticos son responsabilidad de empleados o ex empleados de la compañía a la que pertenecen dichos sistemas.

Según datos de la Fiscalía General del Estado en los años 2017, 2018 y 2019 existe un mayor número de denuncias relacionadas a vulneración de sistemas informáticos entre las cuales se encuentra el acceso no consentido a un sistema informático, el ataque a la integridad de sistemas informáticos, la interceptación ilegal de datos y la revelación ilegal de bases de datos, los cuales se encuentran tipificados en los artículos 234, 232, 230 y 229 del Código Integral Penal de la legislación ecuatoriana. (Primicias, 2019)

En dichos datos entre los años 2018 al 2019 es notable el incremento del número de accesos no consentidos a los sistemas informáticos.

El presente proyecto pretende optimizar la detección temprana de accesos no autorizados a través del análisis de pistas de auditoria por medio de herramientas de Inteligencia Artificial.

Esta investigación realizará la recolección y análisis de la información de la bitácora de accesos de la plataforma SIIPNE3w, luego aplicará algoritmos supervisados para establecer un modelo o línea base de normalidad con los datos históricos, para esto de la base de datos obtenidos se seleccionará el 70% de los datos para las pruebas y aprendizaje y el 30% restante se utilizará para verificar la eficacia de la aplicación.

***Preguntas de Investigación***

Para la consecución del objetivo general del proyecto: Desarrollo de un modelo de aprendizaje automático para detectar accesos no autorizados de manera temprana a través del análisis de pistas de auditoria generadas por la plataforma SIIPNE3w de la Policía Nacional, se diseñan las siguientes preguntas:

OE1-RQ1.1: ¿Cuáles son los estudios existentes sobre detección de accesos no autorizados a sistemas de información usando datos de pistas de auditoría?

OE1-RQ1.2: ¿Cuáles son las herramientas de ML open source que mejor se ajustan para el cumplimiento de los objetivos?

OE2-RQ2.1: ¿Cuáles son los algoritmos supervisados que mejor se ajustan al caso de estudio?

OE2-RQ2.2: ¿Cuáles son las fuentes de datos que mejor se ajustan al caso de estudio?

OE3-RQ3.1: ¿Es posible la generación de una línea base para detectar accesos no autorizados mediante las herramientas de Business Intelligence que se va a utilizar?

OE3-RQ3.2: ¿Cuáles son las métricas usadas para evaluar la aplicación de algoritmos de aprendizaje automático?

# Capítulo II

## Marco Teórico

### Fundamentación de la Variable Independiente

#### *Machine Learning*

El aprendizaje automático es el estudio de algoritmos que mejoran su rendimiento con experiencia y están destinados a informatizar ejercicios; la máquina toma todos los pasos necesarios de manera consumada además de una manera mantenida. Esto permite a las computadoras aprender de un modo similar a como lo realiza un humano sin la necesidad de tener que recodificar algoritmos para cada caso. Incluye varias técnicas de aprendizaje clasificadas como aprendizaje supervisado, no supervisado y de refuerzo, dependiendo de la presencia o ausencia de datos etiquetados. El aprendizaje supervisado entrena el programa con muestras etiquetadas; por lo tanto, el programa capacitado puede predecir muestras similares sin marcar. Incluye tareas de predicción, extracción de conocimiento y compresión. El aprendizaje no supervisado no tiene muestras de entrenamiento; Utiliza el enfoque estadístico de la estimación de densidad. El aprendizaje no supervisado funciona según el principio de encontrar el diseño oculto de los datos agrupando o agrupando datos de un tipo similar. Incluye trabajos como reconocimiento de patrones y detección de valores atípicos. El aprendizaje de refuerzo se centra en los agentes de software que necesitan actuar en un entorno para maximizar la recompensa acumulativa. Cada paso del agente no se considera individualmente para el éxito o el fracaso, pero en una secuencia de acciones tomadas en conjunto debe tener una dirección hacia una buena política. Este aprendizaje se usa mucho en la teoría de juegos y la navegación de robots. (Hamid et al., 2016)

### *Algoritmos No supervisados*

El aprendizaje no supervisado trabaja directamente con datos no etiquetados, en ausencia de etiquetas que permitan orientar el proceso de aprendizaje el algoritmo aplicado debe descubrirlas. (Palacio-Niño & Berzal, 2019).

Algunos algoritmos no supervisados son:

#### *Kmeans-Clúster*

El algoritmo de clusterización kmeans permite agrupar elementos que coinciden o poseen similitud entre sí, para lo cual utiliza como medida la distancia y como método el particionamiento. Kmeans requiere inicialmente establecer un número de centroides, luego calcula la distancia de cada elemento con los centroides y va agrupando cada punto con el centroide más cercano para posterior volver a recalcular los centroides; el proceso se repite iterativamente hasta alcanzar la convergencia. Una vez tenemos cada punto asociado a un clúster, podemos etiquetarlo en el dataframe original asociándolo a dicho grupo y "catalogando" por tanto nuestros datos. (Gliese710, 2019).

#### *Isolation Forest*

Es un eficiente método de detección de outlier en aprendizaje automático el cual aísla las anomalías en lugar de perfiles de instancias normales, para alcanzar dicho propósito el algoritmo toma ventaja de dos propiedades de las anomalías: la primera que las anomalías son minoría consistiendo de pocas instancias y segundo que tienen atributos-valores que son muy diferentes de aquellos de las instancias normales. En otras palabras las anomalías son pocas y diferentes, lo cual las hace más susceptibles de aislar que los puntos normales. (Liu et al., 2008)

***Algoritmos Supervisados***

El aprendizaje supervisado es una técnica de la inteligencia artificial que permite entrenar modelos mediante la utilización de datos etiquetados previamente. Mediante dichos datos el modelo va aprendiendo como clasificarlos y una vez que se logra obtener un modelo preciso, se lo alimenta con datos de entrada nuevos para predecir su etiqueta o resultado. Algunos algoritmos de aprendizaje supervisado son los siguientes

***Support vector machine***

SVM es un algoritmo que trabaja en primera instancia con un conjunto de datos de entrenamiento para determinar límites de decisión o conocidos como hiperplano que permitan dividir el espacio en dos mitades y posterior separar las muestras en una u otra clase. (Jiawei Han et al., 2012)

***Naive Bayes classifier***

Es un clasificador que utiliza la teoría de la probabilidad fundamentada en el teorema de Bayes para efectuar la clasificación de muestras. El algoritmo utiliza el concepto de independencia incondicional de clase, el cual asume que la presencia o ausencia de una característica particular es independiente de los valores de las otras características. Una de sus ventajas radica en que es considerado como uno de los algoritmos con alta precisión y velocidad cuando se trabaja con grandes cantidades de datos. (Eric-Joel Blanco-Hermida Sanz, 2016, p. 17)

***Decision Tree Learning***

El árbol de decisión constituye uno de los algoritmos de aprendizaje automático que permite clasificar elementos siguiendo una serie de reglas de clasificación por las cuales cada elemento va siendo evaluado con el propósito de ser ubicado dentro de una categoría. El árbol comprende una estructura compuesta por un nodo raíz y otros nodos internos que representa

una prueba a un atributo, las ramas los resultados que conducen y las hojas o nodos terminales que representan etiquetas de clase. (Eric-Joel Blanco-Hermida Sanz, 2016)

***Pistas de Auditoría***

Según (Guido, 2012) son datos e informaciones históricas que están disponibles para su examen, con objeto de probar la corrección e integridad con la cual los procedimientos convenidos se seguridad, relativos a una clave o transacción(es), han sido seguidos. Pueden usarse en la investigación de incidentes relacionados con la seguridad o para reconstruir datos dañados o destruidos.

En todos los sistemas, sean automatizados o no, las pistas de auditoría deben estar siempre presentes, y deben de cumplir con tres requisitos básicos:

1. Que cualquier transacción pueda ser seguida, desde el documento fuente que la originó, por medio del proceso a que es sometida, hasta las salidas (archivos, consultas por pantalla o informes) y los totales a los cuales se agrega.

2. Que cada salida o resumen de datos, se pueda seguir hacia atrás, hasta la transacción o cálculos que los produjeron.

3. Que cualquier transacción generada automáticamente, se pueda rastrear hasta el evento o condición que la generó.

Lo que se persigue es poder determinar a ciencia cierta que toda transacción sea procesada, almacenada y presentada sin ninguna transgresión, que no haya cambios o modificaciones en ella, y los resultados que produce sean autorizados y sean producto fiel y cabal de la transacción que le dio origen; especialmente porque existen procesos que son complejos y que los resultados no son exactamente iguales que el origen y la entrada; se aclara con algunos ejemplos.

Un salario mensual, ya está almacenado en la base de datos y no cambia a menudo, lo más una vez cada seis meses; la salida, en forma de pago, es muy similar mes a mes, con pequeñas variaciones por las deducciones aplicadas, por lo tanto tiene un comportamiento igual, y habría que revisar aquellas variaciones que se vea que se puedan salir de lo normal. (p. 4).

**Fundamentación de la Variable Dependiente**

***Anomalías***

Las anomalías según (Alcalde, 2018) se definen de la siguiente manera:

Una anomalía es un dato muy distinto del resto. Esto puede deberse a fallos en mediciones, o a la propia naturaleza del dato. Por ejemplo, una intrusión a un sistema informático puede considerarse una anomalía, ya que por norma general el resto de actividades en dicho sistema serán legítimas. Por lo general, un dato se considera anómalo si escapa a los rangos de normalidad del resto de los datos. El tratamiento de los datos anómalos debe hacerse con cuidado, ya que en ocasiones se podrá descartar (cuando son errores de medición) y en otras será importante (Intrusiones/ataques a un sistema).

***Accesos no autorizados a sistemas de Información***

Constituye un tipo de delito establecido en el Código Integral Penal de la legislación ecuatoriana y es definido como:

Acceso no consentido a un sistema informático, telemático o de telecomunicaciones.- La persona que sin autorización acceda en todo o en parte a un sistema informático o sistema telemático o de telecomunicaciones o se mantenga dentro del mismo en contra de la voluntad de quien tenga el legítimo derecho, para explotar ilegítimamente el acceso logrado, modificar un portal web, desviar o re direccionar el tráfico de datos o voz u ofrecer servicios que estos

sistemas proveen a terceros, sin pagarlos a los proveedores de servicios legítimos, será sancionada con la pena privativa de la libertad de tres a cinco años. (CODIGO ORGANICO INTEGRAL PENAL, COIP, 2018, p. 81)

**Trabajos relacionados (estado del arte)**

Con el propósito de determinar artículos científicos que nos sirvan para el caso de estudio planteado se eligió como estrategia usar el proceso planteado para llevar a cabo un SMS (Systematic Mapping Study) y como fuentes confiables de información a consultar se eligió tres repositorios académicos, los mismos que fueron: IEEExplore, ACM Digital, Springer. El proceso adoptado es el siguiente:

***Definición del objetivo.***

El objetivo es identificar el estado del arte de los temas planteados en las preguntas de los objetivos específicos.

***Definición de los criterios de inclusión y exclusión:***

Se tomaron en cuenta los siguientes criterios:

**Criterios de Inclusión**. Artículos científicos publicados en el idioma inglés.

Artículos que contengan información referente a detección de intrusiones, pistas de auditoria, técnicas de aprendizaje automático, herramientas de ML.

**Criterios de Exclusión**. Se excluyen los artículos cuyos artículos se relacionan a la aplicación de ML en disciplinas diferentes a la del caso de estudio.

Artículos que hacen alusión a la utilización de herramientas comerciales.

Artículos en los que no se incluya datos cualitativos en el análisis.

***Definición de la estrategia de búsqueda***

**Revisión Inicial.** Para encontrar información específica que permita responder las preguntas de investigación planteadas se ejecutó una revisión inicial buscando en las bases de datos académicas elegidas.

**Validación cruzada de estudios.** Se utiliza validación cruzada para la elección de los estudios a utilizarse, verificando que los mismos cumplan los criterios de inclusión y exclusión definidos.

**Integración del Grupo de Control.** Se analizan las características tales como título, introducción, conclusiones y palabras claves de los estudios que conforman el grupo de control y que son los siguientes:

**Tabla 1.**

*Estudios por grupo de control*

| GRUPO DE CONTROL | TÍTULO | PALABRAS CLAVE |
|---|---|---|
| EC1 | Intrusion Detection<br>Applying Machine Learning to Solaris Audit Data | Intrusion detection, Machine learning,<br>Data security, Information security, Detectors, Humans, Operating systems, Sun, Neural networks, Computer errors |
| EC2 | Open Source Platforms and Frameworks for Artificial<br>Intelligence and Machine Learning | diversity, artificial intelligence, machine learning,<br>capstone course, project-based learning |
| EC3 | Machine Learning Techniques for Intrusion Detection: A Comparative Analysis | False Positive, IDS, Machine Learning, Precision, ROC, True Positive |
| EC4 | Innovative Genetic Approach for Intrusion Detection by Using Decision Tree | Genetic algorithm, layered approach, PCA |
| EC5 | Intrusion Detection Based on Behavior Mining and<br>Machine Learning Techniques | Support Vector Machine Intrusion Detection Frequent |

| GRUPO DE CONTROL | TÍTULO | PALABRAS CLAVE |
|---|---|---|
| EC6 | Research of Recognition System of Web Intrusion Detection Based on Storm | Pattern Terminal Node Intrusion Detection System Strom; Big Data; Web Intrusion Detection System; TF-IDF; cosine similarity |
| EC7 | Authentication Anomaly Detection: A Case Study On A Virtual Private Network. | Security, Data Mining, Authentication, Anomaly Detection, Virtual Private Network |

*Nota*. Listado de estudios por grupo de control.

***Construcción de la cadena de búsqueda***

En la construcción de la cadena de búsqueda se contabilizan las palabras claves más usadas en cada uno de los estudios como se indica en la Tabla 2.

**Tabla 2.**

*Cadena de Búsqueda*

| PALABRA CLAVE | EC1 | EC2 | EC3 | EC4 | EC5 | EC6 | EC7 | NUMERO DE REPETICIONES |
|---|---|---|---|---|---|---|---|---|
| Intrusion Detection | X | | | | X | X | | 3 |
| Anomaly Detection | | | | | | | X | 1 |
| Audit Trails | X | | | | | | | 1 |
| Machine Learning | X | X | X | | X | | | 4 |
| Supervised Learning | | | | | | | | 0 |

*Nota*. Conformación de la Cadena de Búsqueda

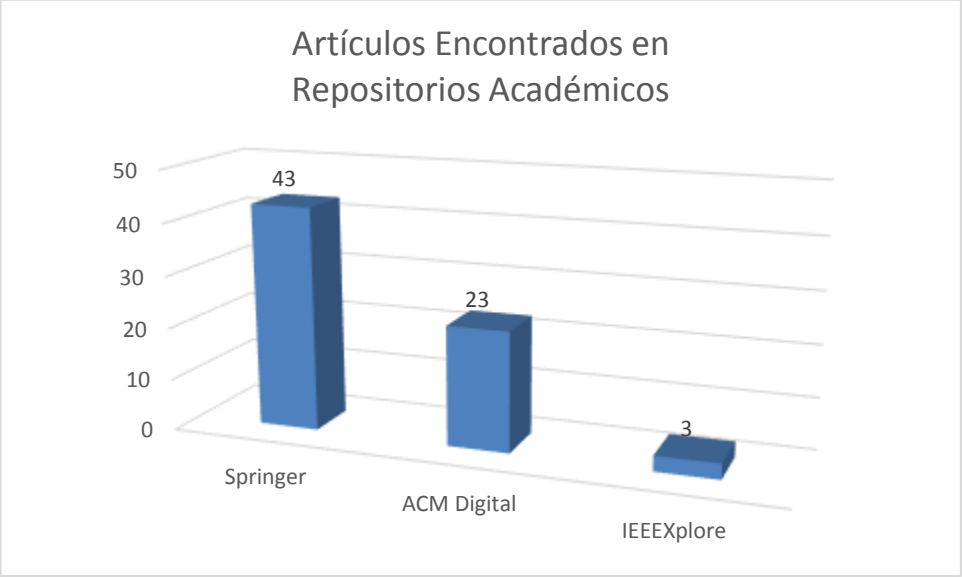
Se eligen los términos que presentan mayor número de repeticiones para formar la cadena de búsqueda concatenando las palabras con los conectores AND para añadir nuevos términos y OR para añadir sinónimos, estableciéndose la siguiente cadena:

**(Intrusion Detection) AND (Audit Trails) AND ((Machine Learning) OR (Supervised Learning))**

Con la cadena definida se efectuó la búsqueda en cada uno de los repositorios académicos seleccionados para el estudio, obteniéndose un total de 69 artículos, 43 en Springer, 23, en ACM Digital y 3 en IEEEXplore como se muestra en la Figura 1.

**Figura 1.**

*Resultados - Cadena de Búsqueda*




*Nota*. Diagrama de los artículos encontrados al aplicar la cadena de búsqueda a las bases de datos académicas.

De los resultados obtenidos, para el análisis se escogió aquellos documentos que mencionan comparativas o uso de herramientas de aprendizaje automático de libre uso. Y que guardan relación con análisis de pistas de auditoria o que detectan intrusiones, anomalías, comportamientos inusuales.

Finalmente se realizó la revisión de los documentos encontrados, algunos de los cuales se listan a continuación:

**Intrusion Detection Applying Machine Learning to Solaris Audit Data** (Endler, 1998)

El documento en cuestión realiza un análisis de detección de intrusos usando datos de auditoria presentes en el sistema operativo Solaris, para dicho propósito usa algoritmos de clasificación de histograma y de redes neuronales combinándolos con el propósito de mejorar los resultados, haciendo hincapié en su conclusión de los beneficios de combinar ambos aspectos de detección en futuros IDS para disminuir errores de falso positivo y falso negativo.

**Machine Learning Techniques for Intrusion Detection: A Comparative Analysis.**(Hamid et al., 2016)

En el documento el autor revisa todos los algoritmos de aprendizaje automático presentes en la herramienta WEKA y hace un análisis comparativo para la detección de intrusos en la red en un conjunto de datos de referencia y reporta los datos obtenidos con diferentes métricas de evaluación.

**Authentication Anomaly Detection: A Case Study on a Virtual Private Network.** (Chapple et al., 2007)

El documento describe un enfoque novedoso para identificar anomalías en los registros de autenticación mediante el uso de Agrupación de maximización de expectativas. Específicamente, usa registros de conexión de una red privada virtual de la universidad para desarrollar modelos de actividad típica basados en tipo de usuario, rol, fecha y hora de conexión, y distancia geográfica desde la fuente de conexión al punto final VPN.

**Research of Recognition System of Web Intrusion Detection Based on Storm.** (Bo et al., 2016)

El documento usa el algoritmo TF-IDF para detectar comportamientos maliciosos en tiempo real de peticiones http por medio de la plataforma de Big Data Storm. El experimento muestra que dicho sistema puede efectivamente identificar ataques maliciosos y detectar intrusos en tiempo real así como también muestra las ventajas de utilizar la plataforma Storm.

La revisión de literatura permite observar lo siguiente:

Como herramientas de aprendizaje automático se ha encontrado el uso de WEKA, Jupyter Notebook, Python, ELK, Apache Storm.

Existen muchas aplicaciones de aprendizaje automático para la detección de intrusos enfocadas en accesos a la red que se podrían aplicar a nuestro caso de estudio.

La validación de los modelos de aprendizaje automático se las realiza mediante algunas métricas como: curva de análisis ROC, precisión, Accuracy, kappa, matriz de confusión.

En cuanto a los algoritmos de aprendizaje supervisado, para la detección de anomalías los recomendados son los de clasificación y en adición se debe contar con datos etiquetados para poder entrenar el modelo de aprendizaje.

En la literatura no se ha identificado un estudio sobre la detección de accesos no autorizados a sistemas de información web, por lo que se considera será un aporte de nuestra investigación.

## Capítulo III

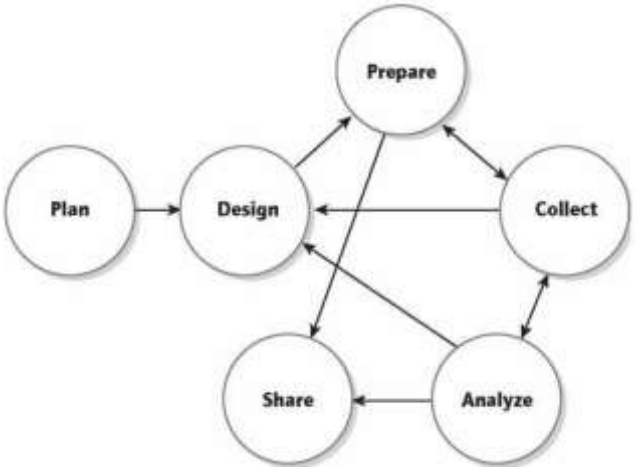
### Metodología de la Investigación

La metodología elegida para llevar la investigación es el estudio de caso, Robert Yin es un profesional en psicología experimental con muchas publicaciones basada en casos, su definición de caso de estudio es la siguiente:

El estudio de caso es una de las maneras de hacer investigación científica y constituye una estrategia preferida cuando las preguntas “como” y “porque” son realizadas, cuando el investigador tiene poco control sobre los eventos y cuando el foco está en un fenómeno contemporáneo dentro de un contexto de la vida real. Los estudios de caso pueden ser basados en cualquier mezcla de evidencia cuantitativa y cualitativa. Además, los estudios de caso no siempre necesitan incluir las observaciones directas y detalladas como una fuente de evidencia. (Robert K. Yin, 2009)

**Figura 2.**

*Metodología Caso de Estudio*




*Nota*. Diagrama de la metodología para la investigación de Casos de Estudio. Adaptado de (Robert K. Yin, 2009)

La metodología consta de 5 componentes importantes como se muestra en la Tabla 3.

**Tabla 3.**

*Metodología Yin*

| METODOLOGÍA YIN | OBJETIVO ESPECÍFICO | PREGUNTAS DE INVESTIGACIÓN |
|---|---|---|
| DISEÑO DEL CASO DE ESTUDIO | Realizar una revisión de la literatura para determinar posibles soluciones y recomendaciones existentes para detectar accesos no autorizados de manera temprana a través del análisis de pistas de auditoria utilizando herramientas de aprendizaje automático | OE1-RQ1.1. ¿Cuáles son los estudios existentes sobre detección de accesos no autorizados a sistemas de información usando datos de pistas de auditoría?<br>OE1-RQ1.2. ¿Cuáles son las herramientas de ML open source que mejor se ajustan para el cumplimiento de los objetivos? |
| CONDUCCIÓN DEL CASO DE ESTUDIO: PREPARACIÓN DE LA RECOLECCIÓN DE DATOS. | Aplicación de algoritmos supervisados para la detección de accesos no autorizados de manera temprana a través del análisis de pistas de auditoria | OE2-RQ2.1. ¿Cuáles son los algoritmos supervisados que mejor se ajustan al caso de estudio? |
| CONDUCCIÓN DEL CASO DE ESTUDIO: RECOLECCIÓN DE DATOS. | | OE2-RQ2.2. ¿Cuáles son las fuentes de datos que mejor se ajustan al caso de estudio? |
| ANÁLISIS DEL CASO DE ESTUDIO. | Evaluar los resultados obtenidos mediante métricas comunes de aprendizaje automático, para verificar la eficiencia del modelo usado para la detección de accesos no autorizados | OE3-RQ3.1. ¿Es posible la generación de una línea base para detectar accesos no autorizados mediante las herramientas de Business Intelligence que se va a utilizar? |
| ELABORACIÓN DEL REPORTE DE CASO DE ESTUDIO. | | OE3-RQ3.2. ¿Cuáles son las métricas usadas para evaluar la aplicación de algoritmos de aprendizaje automático? |

*Nota*. Componentes según metodología YIN para el caso de estudio planteado.

## Diseño de Caso de Estudio

Los componentes que consta el diseño de la investigación son:

### *Preguntas de Estudio*

Las preguntas de estudio han sido planteadas en el numeral 0

### *Proposiciones*

El propósito es determinar si al analizar las pistas de auditoria con algoritmos supervisados de aprendizaje automático se puede detectar accesos no autorizados al Sistema Integrado de la Policía Nacional del Ecuador.

### *Unidades de Análisis*

Accesos no autorizados en el sistema Integrado de la Policía Nacional del Ecuador.

### *Relación lógica entre los datos y proposiciones*

Se usa la técnica de modelo emparejando, para lo cual se aplicará un modelo no supervisado y otro supervisado para clasificar los datos y posterior determinar cuál de los dos empareja mejor.

### *Criterios para interpretar los resultados*

Como criterios de interpretación de resultados se elegirá la comparación de proposiciones rivales. Para lo manifestado se compararán los valores obtenidos en las métricas de validación de los modelos de aprendizaje automático utilizados para el análisis de los datos, aquel modelo con valores más altos será el que mejor resultados de clasificación ofrece o dicho en otras palabras empareja mejor.

## Preparación de la recolección de Datos

### *Visión general del proyecto de caso de estudio*

La metodología indica que la visión general debe incluir información acerca del proyecto y los objetivos.

La visión general la abordamos con detalle en el acápite 0 0, y los objetivos, se los plantea en el punto 0 del presente documento.

### *Procedimiento de campo*

Esta etapa que incluye el grado de accesibilidad al lugar de estudio y a la información necesaria para la investigación la desglosamos en profundidad en el ítem 0,

El procedimiento utilizado en el campo para la recolección de datos está basado en las actividades propuestas por la metodología CRISPDM en su fase 2 al cual se le han agregado algunas consideraciones, dicho protocolo incluye:

Recopilar información en sitio que incluya datos de logs de auditoria, usuarios policiales, georreferenciación relacionadas con los accesos al sistema, características del usuario policial, coordenadas de lugar desde el cual accede al sistema, necesarias para el caso de estudio.

Manejar nombres diferentes para las variables recopiladas y sus fuentes para proteger la privacidad y confidencialidad.

Evitar la recopilación de tipos de datos como claves, cuentas de usuario, ubicaciones de las fuentes, nombres propios, configuraciones.

Describir los datos incluyendo nombre, tipo y descripción.

Efectuar un análisis exploratorio de datos usando métodos estadísticos y herramientas de uso libre.

Verificar la calidad de los datos.

Almacenar el set de datos resultante en un archivo CSV.

**Recolección de Datos**

La metodología YIN para esta etapa establece seguir el protocolo del caso de estudio, usar múltiples fuentes de evidencia, crear una base de datos y mantener una cadena de evidencia.

Dicha recolección de datos se la realizó siguiendo el protocolo de campo definido anteriormente usando diferentes fuentes de datos que incluyen log de auditoria, tablas de datos de información policial, base de datos de geolocalización.

A la información obtenida se aplicó un conjunto de actividades que nos recomienda la fase III de la metodología CRISP-DM, las mismas que tienen como objetivo disponer del juego de datos final sobre el cual se aplicará los modelos de aprendizaje.

Mencionadas actividades incluyen efectuar una selección de las variables finales, limpieza de datos, ejecución de las transformaciones necesarias al juego de datos, integración de la información de las distintas fuentes y formateo del set de datos final.

Todo el detalle y evidencia de lo mencionado para el caso de estudio planteado en el presente trabajo se lo desarrolla a profundidad en el ítem 0

**Análisis del Caso de Estudio. (Modelo de Aprendizaje Automático)**

Para ésta etapa de la investigación, el planteamiento es usar una rama de la Inteligencia artificial que es el aprendizaje automático, el cual permite a las maquinas detectar patrones en los datos con la finalidad de efectuar predicciones. Sin embargo éste proceso suele ser complejo y requerir de un sinnúmero de tareas para alcanzar el objetivo, por lo que es indispensable seguir una metodología para llevar una adecuada gestión.

La metodología elegida para llevar la gestión del desarrollo del modelo de aprendizaje automático para la investigación planteada es Crisp-DM.

Crisp-DM comprende una guía para la ejecución de proyectos de minería de datos desarrollada por dos empresas cuya actividad se basa en brindar servicios de éste tipo.

Esta metodología consiste de un número de fases que permiten de manera organizada la ejecución adecuada de un proyecto de minería de datos, empezando desde la definición de los objetivos y necesidades del cliente hasta su implementación. (Jordi Gironés Roig, s. f.)

Para alcanzar el objetivo de la investigación propuesta nos apoyaremos en las diferentes fases que ofrece esta metodología exceptuando la fase de implementación. Por lo que nuestro trabajo llegará hasta la evaluación del modelo que diseñaremos, de tal manera de obtener una perspectiva de los beneficios que pueden ofrecernos las herramientas de minería y aprendizaje automático y ofrecer una alternativa a la institución policial para análisis de implementación o no en su plataforma tecnológica.

***Fase I. Comprensión del Negocio. Definición de necesidades del cliente***

**Identificar los Objetivos del Negocio**

***Entorno***. En el acápite 0 y 0 del presente documento se ha detallado la problemática planteada en la presente investigación, sin embargo en esta parte se amplía algunas causas para mejor comprensión de la misma y del entorno donde se genera ésta, entre las cuales podemos citar que:

Los usuarios de la plataforma son personal que pertenecen a la institución, los cuales poseen ciertos privilegios según su función y unidad en la que prestan su servicio.

En adición a los permisos otorgados en la plataforma para la ejecución de sus funciones, cada usuario posee acceso para consultar información inherente a sí mismo, propia de cada uno y su gestión personal.

La naturaleza de su trabajo involucra constantes movimientos y traslados entre unidades de servicio y con dichos procesos se generan otros, uno de los cuales incluye la inactivación de permisos de su antigua función y activación conforme a su nueva función y unidad.

Los periodos de tiempo que toma la legalización de ciertos traslados en ciertas ocasiones generan que los usuarios aun posean permisos que ya no le corresponden, lo cual incrementa la probabilidad de verse tentados hacer mal uso de la información de la que tienen acceso.

La informatización de trámites y procesos puede generar complicaciones en algunas personas que se encuentran en ciertos rangos de edad, haciendo que pidan ayuda a terceros para el manejo de la plataforma, a los cuales debido a la falta de cultura informática hace que compartan la información de sus credenciales pudiendo generarse un mal uso de la misma.

Debido a lo mencionado se requiere predecir de alguna manera cuales accesos pueden ser catalogados como sospechosos o inusuales (no autorizados) y cuáles no.

***Objetivo***. Las necesidades del cliente son identificar aquellas conexiones inusuales o sospechosas (no autorizadas), para en base a dicha identificación poder indagar las transacciones de dicha conexión en mayor profundidad con el fin de prever malos usos de la información, con esto se pretende reducir la lentitud del trabajo manual, así como el uso de métodos no adecuados. De ésta manera poseer un sistema a medida que se pueda ir puliendo, afinando y escalando conforme los resultados obtenidos para mejorar el proceso de detección de anomalías.

***Criterio de Éxito***. Para medir el éxito del objetivo se entrenará dos modelos de aprendizaje automático, uno de aprendizaje supervisado utilizando el algoritmo de Árbol de Decisión y otro de aprendizaje no supervisado utilizando el algoritmo Isolation Forest.

La selección de los algoritmos se basa en lo siguiente:

Los dos constituyen algoritmos de clasificación con un enfoque contrario tanto en su técnica como en su género, lo cual permite enfrentarlos como proposiciones rivales. El algoritmo de Árbol de Decisión es un algoritmo de aprendizaje supervisado que para clasificar los datos intentan definir las instancias normales para luego catalogar las anomalías. Por el contrario el algoritmo Isolation Forest es un algoritmo de aprendizaje no supervisado que trabaja de modo opuesto, centrándose en aislar las anomalías para clasificar los datos.

La estructura de los dos algoritmos está basada en árboles lo cual facilita su entendimiento y representación.

Ambos algoritmos se ajustan mejor para el caso de estudio debido a su eficiencia computacional teniendo en consideración el hardware disponible para el análisis y procesamiento de los datos.

Los dos algoritmos al estar basados en árboles son muy buenos clasificando datos categóricos.

Para establecer el criterio de éxito cada modelo será evaluado según las métricas comunes de validación externa e interna de las que dispone la librería scikit-learn para los proyectos de aprendizaje automático.

Una vez validados ambos modelos se realizará una comparativa de los valores de las métricas de validación obtenidos, con el propósito de determinar cuál modelo es el que mejor resultados brinda.

**Evaluar la situación actual**. La plataforma tecnológica SIIPNE3w cuenta con diversas bases de datos con información que data del año 2005, fechas desde las cuales siempre han mantenido pistas de auditoria para el control y monitoreo de las acciones de sus usuarios. Sin

embargo para nuestro estudio de caso nos centraremos en la utilización de las pistas de auditoria de su versión web, de cierto año en específico.

***Recursos***. Los recursos en cuanto a fuentes de datos que usaremos se detallan a continuación:

- Logs de auditorías producidos por el sistema SIIPNE3w.
- Base de Datos Personal Policial
- Bases de Datos de Geolocalización de IP.

Los logs de auditoria producidos por el sistema SIIPNE3w seleccionados son aquellos que almacenan información relacionada al momento en que un usuario se conecta y desconecta al sistema. Dichos logs de auditoria únicamente almacenan los intentos exitosos de conexión al sistema.

Existen otros tipos de logs que recaban información de los intentos no exitosos, sin embargo para nuestro caso de estudio se los deja de lado ya que el ámbito de investigación está orientado a determinar los accesos sospechosos no autorizados.

La población elegida es los registros producidos durante el año 2016, con un total de 4 899 877 de registros generados por 50215 usuarios.

Dichos logs por cada vez que un usuario se conecta y desconecta del sistema almacena un registro con ciertos datos importantes como: la IP de conexión, fecha y hora, e id de usuario que se conecta.

Con esta información se calculará los tiempos de conexión en segundos de cada usuario. Además se complementará con datos adicionales de otras fuentes (Bases de Datos Personal Policial, Bases de Datos de Geolocalización de IP), como son: la edad, sexo, estado civil, situación en la institución, ciudad, país, latitud y longitud de donde se conecta.

Como recurso humano que se utilizará de apoyo para la comprensión de los datos, reglas del negocio y criterios de expertos se cuenta con aquellas personas que han laborado el mayor tiempo en la unidad encargada de la administración de ésta plataforma.

Estas personas especialistas incluyen:

Administradores de Base de Datos.

Desarrolladores con mayor experiencia y mayor tiempo de permanencia.

Administradores de Infraestructura.

Especialistas de Seguridad de la Información.

En cuanto a recursos tecnológicos que se usaran para la ejecución del análisis del caso de estudio serán los descritos a continuación:

Hardware

- Computador portátil.
  - Dell
  - 16Gb Ram
  - Procesador Intel Core i7 2.9 Ghz
  - 500Gb Disco Duro SSD

Software

Los recursos tecnológicos que se van a utilizar son:

- Python, Jupyter Notebook
- Pandas, SckiLearn, SQL

***Costos y Beneficios***. La utilización de herramientas de libre acceso no genera costos en cuanto a licencias y adquisición.

La adquisición de datos de igual manera no representa costos adicionales.

Se trabajó con hardware disponible por lo que no existen costos para su adquisición.

Los costos adicionales tales como: Limpieza, formateo, depuración, integración, modelado y evaluación de datos, descargas de herramientas son asumidos por el autor de la presente investigación.

La institución policial al ser una entidad pública está reglamentada de tal manera que no puede generar beneficios económicos y más bien todo tipo de inversión debe traducirse en mejorar la calidad de servicio a la sociedad, Sin embargo la implementación a un ambiente de producción no ha sido contemplada en la presente investigación por lo que no se plantea un análisis de beneficios de dicho tipo.

**Objetivos Minería de Datos**. Predecir accesos inusuales o sospechosos (no autorizados) de manera temprana a través del análisis de pistas de auditoria.

Verificar la eficiencia del modelo usado para la detección de accesos no autorizados, mediante métricas comunes del aprendizaje automático.

**Plan del Proyecto**.

***Fases y Actividades a realizar***. Realizar la revisión bibliográfica de los estudios existentes sobre detección de accesos no autorizados a sistemas de información usando datos de pistas de auditoría.

Definir las herramientas de ML open source a utilizar.

Definir los algoritmos supervisados a usar en el caso de estudio.

Obtener los datos de pruebas, entrenamiento y validación

Desarrollar el modelo de aprendizaje automático.

Realizar la revisión bibliográfica de métricas de evaluación para aprendizaje automático.

Validar el modelo de aprendizaje automático.

Conclusiones y Recomendaciones.

***Evaluación inicial de Herramientas y técnicas***. Conforme a los dispositivos tecnológicos de hardware que se dispone y la restricción en recursos económicos, se elige el uso de herramientas libres que permitan un manejo adecuado y procesamiento aceptable de la cantidad de información que se dispone del sistema, estas herramientas son:

Librería SckiLearn de Python. Para el modelado ML

PySpark, Para el procesamiento de datos a través de Python.

Pandas. Para la ejecución del Análisis exploratorio de Datos (EDA)

Jupyter Notebook. Como Ambiente de desarrollo (IDE)

Apache Spark. Para el procesamiento de datos en gran escala en caso de ser necesario.

Python. Como lenguaje de Programación.

Docker. Como contenedor de software para instalar las aplicaciones a usarse.

En cuanto a técnicas, métodos y metodologías; se han encontrado una gran variedad para ser utilizadas en cada una de las etapas del proyecto de Aprendizaje Automático, se detalla a continuación algunas en las cuáles nos apoyaremos para la resolución del problema planteado:

CRISP-DM. Para la organización del proyecto de Aprendizaje Automático

Sobre-muestreo, sub-muestreo, ajuste de parámetro. Para el tratamiento de datos no balanceados.

Método del Codo. Para calcular la elección del número de clústeres adecuado.

Ganancia de Información (entropía). Para calcular la profundidad del Árbol de Decisiones.

Diagrama de Caja y Bigotes, Límites superior e inferior. Para el cálculo de valores atípicos

Analítica descriptiva y cuartiles. Para categorizar datos.

Estadística descriptiva. Para la comprensión de datos

Matriz de confusión. Para la evaluación y validación del modelo.

Clustering, Criterio de Expertos, técnicas de asociación y técnicas para codificar variables categóricas. Para el etiquetado de datos

Técnicas de validación Interna y Externa. Para la validación de los modelos de aprendizaje automático.

Los criterios de selección para la elección de una u otra técnica en cada fase se lo realizó mediante prueba y error hasta elegir la que mejor resultados nos proporciona.

De igual manera en la selección de herramientas y librerías se usó el mismo método, añadiendo que dichas herramientas sean de libre acceso (open source), y sean las más comúnmente usadas. (Nguyen et al., 2019)

***Fase II. Estudio y comprensión de los datos***

**Captura de Datos**. La captura de datos de la bitácora se la ejecutó mediante un script en lenguaje SQL. El script permite extraer la información que requerimos hacia un archivo CSV combinando el registro de ingreso con el de salida en uno solo, con el propósito de poder efectuar cálculos en ciertos datos que estimamos importantes para la selección y análisis.

A través de otro script desarrollado en el lenguaje de programación Python los datos obtenidos de la bitácora son complementados con información adicional del usuario, obtenida del repositorio de datos de funcionarios policiales a través del código de usuario.

Finalmente el script desarrollado usa la información de la dirección IP para adicionar información del lugar como país, ciudad, latitud, longitud utilizando una base de datos de georreferencia por IP de acceso libre.

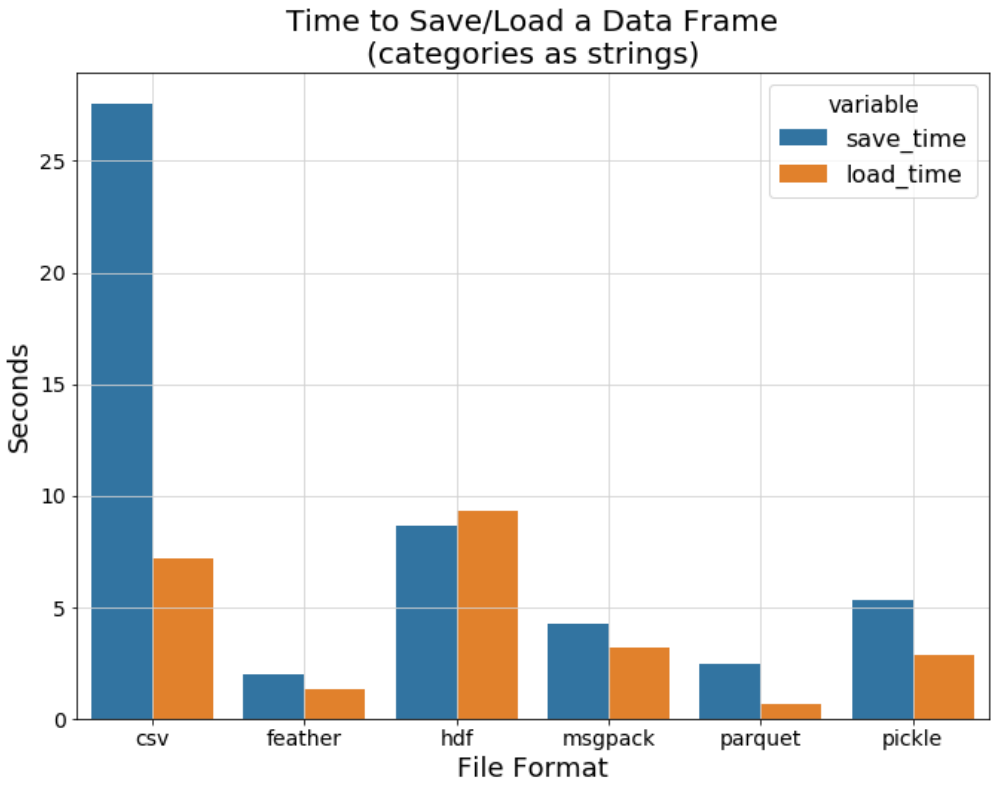
Para dicha integración de información el script desarrollado en Python ejecutaba un recorrido secuencial para complementar la información y escritura línea a línea para almacenar dicha información a un archivo CSV.  Dicha estrategia en sí resultó funcional pero no óptima ya

que resultó ser demasiado costosa en cuanto a tiempo y rendimiento tanto en la integración como en la carga y almacenamiento de la información.

Para solventar lo manifestado se modificó el script para utilizar la librería pandas y otro tipo de formato de almacenamiento, obteniendo una notable mejora, en la captura e integración de datos con dicha librería y una mejor eficiencia y reducción de tiempo al momento del almacenamiento y carga de datos con el formato parquet a relación de otros formatos tales como CSV, picklet, etc. La elección de lo manifestado se lo realizó mediante prueba error y según lo determinado por (Zaitsev, 2019) quien realiza una pequeña comparación de varias formas de serializar un marco de datos de pandas en un almacén persistente.

**Figura 3.**

*Comparación Formatos de Almacenamiento*




*Nota*. Cuadro comparativo según (Zaitsev, 2019)

**Descripción de los Datos**.

***Logs de Auditoría***. Constituyen estructuras de tipo TABLE LOG que almacena pistas de auditoría de la plataforma web en estudio.

Existen varias estructuras que almacenan pistas de auditoria, en nuestro caso de estudio usaremos aquella que almacena las conexiones login y logout de los usuarios.

***Tabla Auditoria Log-in-out***. Esta tabla almacena los accesos de los usuarios. Los campos que almacena son:

Id_Serial. Tipo numérico. Código de identificación única del registro

Id_ serial_padre. Tipo numérico. Código de identificación única del registro que mantiene relación de dependencia con el registro en mención.

Operación. Tipo Alfabético. Tipo de operación que realiza el usuario (IN, OUT)

**Tabla 4.**

*Valores para Operación*

| Descripción | Código | Abreviatura |
|---|---|---|
| Ingreso | I | IN |
| Salida | O | OUT |

*Nota*. Esta tabla muestra los posibles valores que almacena el campo Operación.

Id_usuario. Tipo numérico. El código único del usuario que genera la pista de auditoria al ingresar o salir del sistema.

Ip. Tipo Alfanumérico. Almacena información de la IP del lugar desde donde se conecta el usuario al sistema, además un texto descriptivo del tipo de conexión.

Fecha. Tipo Timestamp. Almacena la fecha y hora del momento exacto cuando el usuario genera la pista de auditoria.

***Tabla de Información Policial***. En dicha tabla se almacena la información del personal que pertenece a la institución policial, todos los datos de los miembros de ésta institución especial que posee ciertas características para el cumplimiento de su misión.

Situacion_policial. Tipo Alfabético. Almacena la situación del funcionario policial, entendiéndose como situación policial al estado en que se encuentra el funcionario los cuales puede ser Alta, Baja, Cesación, Transitoria, A disposición, Se almacena un carácter como código para cada tipo de situación.

**Tabla 5.**

*Valores para Situacion_policial*

| **Descripción** | **Código** |
|---|---|
| Activo | A |
| Baja | B |
| Cesación | C |
| Transitoria | T |
| Disposición | D |

*Nota*. Esta tabla muestra los posibles valores que almacena el campo Situación Policial.

Estado_civil. Tipo Alfabético. Guarda información relacionada al estado civil del funcionario, el cual puede encontrarse como soltero, casado, viudo, divorciado, unión libre. La codificación almacena dos caracteres para identificar cada tipo de estado.

**Tabla 6.**

*Valores para Estado_civil*

| **Descripción** | **Código** |
|---|---|
| Soltero | SO |
| Casado | CA |
| Divorciado | DI |
| Unión Libre | UL |
| Viudo | VI |

*Nota*. Esta tabla muestra los posibles valores que almacena el campo Estado Civil.

Fecha_nacimiento. Tipo fecha. Almacena la fecha de nacimiento del funcionario policial

Sexo. Tipo Alfabético. Almacena un carácter para identificar las características fisiológicas y sexuales con las que nace el funcionario policial.

**Tabla 7.**

*Valores para Sexo*

| Descripción | Código |
|---|---|
| M | Masculino |
| F | Femenino |

*Nota*. Esta tabla muestra los posibles valores que almacena el campo Sexo.

***Tabla Geolocalización de IP. (GeoIPCity.dat)***. La base de datos de geolocalización por IP, permite obtener información específica de una dirección IP. Los siguientes campos se pueden extraer.

**Tabla 8.**

*Campos Base de Datos Geolocalización por IP.*

| Campo | Tipo Dato | Descripción |
|---|---|---|
| country_code. | Tipo Alfabético. | Código del país. |
| country_code3. | Tipo Alfabético. | Código del país. |
| country_name. | Tipo Alfanumérico. | El nombre del país |
| región. | Tipo Alfabético. | Nombre de la Región |
| city. | Tipo Alfanumérico. | El nombre de la ciudad |
| latitude. | Coordenada | La coordenada de latitud aproximada WGS84 del lugar asociado con la red. |
| longitude. | Coordenada | La coordenada de longitud aproximada WGS84 del lugar asociado con la red. |
| dma_code. | Tipo Numérico. | DMA código de región en US. |
| area_code. | Tipo Numérico. | Código de área teléfono (Solo US) |

*Nota*. Describe los campos que se pueden extraer de una Base de Datos de Geolocalización por IP.

***Set de Datos***. Como hemos podido notar por cada registro de ingreso en los log de auditoría se almacena un registro de salida, por lo que fue necesario unir ambos en uno solo, de tal manera de poseer información completa del acceso de un usuario. Al ejecutar dicho paso se ha podido generar mediante sentencias SQL nuevos campos calculados a partir de los disponibles con información más precisa y exacta. Y finalmente adicionar información relacionada al funcionario policial y de georreferenciación por IP, quedando nuestro set de datos como se describe a continuación:

**Tabla 9.**

*Set de Datos*

| Campo | Tipo de Dato | Descripción |
|---|---|---|
| usercode | int64 | Código único del Usuario |
| fini | Object | Fecha Inicio de conexión |
| fFin | Object | Fecha Fin de conexión |
| diaSemana | int64 | Día de la semana (0-7) conectado |
| hini | Object | Hora de inicio de conexión (hh:mm:ss) |
| hfin | Object | Hora fin de conexión (hh:mm:ss) |
| hini1 | int64 | Hora inicio de conexión. (0-24) |
| ipini | Object | IP inicial de conexión |
| ipfin | Object | IP final de conexión |
| iptipo | Object | Tipo de IP |
| segundos | int64 | Tiempo en segundos conectado |
| estadoCivil | Object | Estado Civil del usuario |
| sexo | Object | Sexo del usuario |
| situacion | Object | Situación en la institución del usuario. |
| edad | float64 | Edad del usuario |
| long | float64 | Coordenada longitud |
| lat | float64 | Coordenada latitud |
| pais | Object | País desde donde se conecta |
| ciudad | Object | Ciudad desde donde se conecta |

*Nota*. Listado de campos elegidos como set de Datos.

Con el set de datos procedemos a calcular algunas medidas para poder obtener una visión global de la información recopilada haciendo uso de las librerías, herramientas y técnicas detalladas en los anteriores ítems.

Mediante el calculó de algunas medidas de tendencia central, tenemos una apreciación general de los datos.

**Figura 4.**

*Medidas de tendencia Central.*

```
In [13]: df.describe()
```

Out[13]:

| | auldGenUsuario | diaSemana | hini1 | segundos | edad | long | lat |
|---|---|---|---|---|---|---|---|
| **count** | 1.900314e+06 | 1.900314e+06 | 1.900314e+06 | 1.900314e+06 | 1.900145e+06 | 1.900314e+06 | 1.900314e+06 |
| **mean** | 2.496764e+04 | 3.519566e+00 | 1.332815e+01 | 6.196718e+02 | 3.769760e+01 | -7.911580e+01 | -3.382950e-01 |
| **std** | 1.602152e+04 | 1.851944e+00 | 4.748695e+00 | 1.038719e+03 | 6.366905e+00 | 6.482059e+00 | 6.802110e+00 |
| **min** | 1.000000e+00 | 1.000000e+00 | 0.000000e+00 | 0.000000e+00 | 1.500000e+01 | -1.231552e+02 | -5.314710e+01 |
| **25%** | 1.077200e+04 | 2.000000e+00 | 1.000000e+01 | 1.350000e+02 | 3.300000e+01 | -7.990110e+01 | -2.166400e+00 |
| **50%** | 2.361600e+04 | 3.000000e+00 | 1.300000e+01 | 3.090000e+02 | 3.700000e+01 | -7.861670e+01 | -1.999800e+00 |
| **75%** | 3.870200e+04 | 5.000000e+00 | 1.700000e+01 | 6.950000e+02 | 4.200000e+01 | -7.850170e+01 | -2.143000e-01 |
| **max** | 5.871500e+04 | 7.000000e+00 | 2.300000e+01 | 4.054300e+04 | 7.800000e+01 | 1.530215e+02 | 5.994520e+01 |

*Nota*. El grafico representa las medidas de tendencia central de las variables del set de datos.

Posterior para determinar si existe algún tipo de asociación entre las variables elegidas del set de datos de tal manera de conocer si una variable puede explicarnos el comportamiento de otra, se efectuó un cálculo de correlación.

**Figura 5.**

*Cálculo de correlación.*

```
df.corr()
```

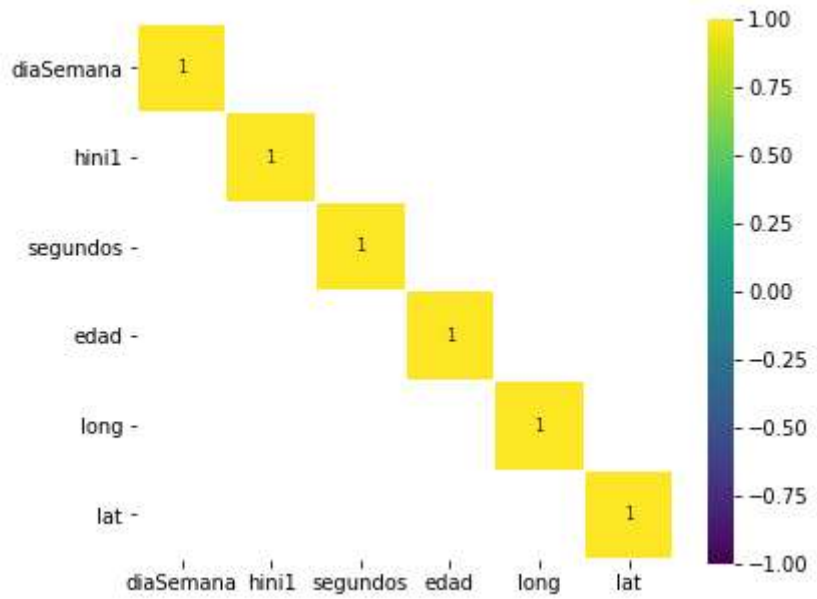
| | auldGenUsuario | diaSemana | hini1 | segundos | edad | long | lat |
|---|---|---|---|---|---|---|---|
| **auldGenUsuario** | 1.000000 | 0.015638 | -0.003014 | -0.089943 | -0.548344 | -0.015964 | 0.060996 |
| **diaSemana** | 0.015638 | 1.000000 | -0.010452 | 0.003843 | -0.015517 | -0.011605 | 0.012547 |
| **hini1** | -0.003014 | -0.010452 | 1.000000 | -0.006570 | 0.001106 | 0.001835 | -0.008007 |
| **segundos** | -0.089943 | 0.003843 | -0.006570 | 1.000000 | 0.072627 | 0.005487 | -0.018262 |
| **edad** | -0.548344 | -0.015517 | 0.001106 | 0.072627 | 1.000000 | 0.027058 | -0.049536 |
| **long** | -0.015964 | -0.011605 | 0.001835 | 0.005487 | 0.027058 | 1.000000 | -0.102045 |
| **lat** | 0.060996 | 0.012547 | -0.008007 | -0.018262 | -0.049536 | -0.102045 | 1.000000 |

*Nota*. Valores de correlación entre las variables de análisis

Para una mejor visualización y comprensión, se elaboró un gráfico mapa de calor de correlación, en el cual se elimina la variable usercod y se resaltan aquellos valores mayores a 0,5 y menores a -0,4.

**Figura 6.**

*Diagrama de Correlación.*




*Nota*. Mapa de calor de los valores de correlación.

Como se puede apreciar en la Figura 6, no existe una correlación importante entre las variables de análisis.

Otro punto importante de recalcar es que no se ha encontrado alguna variable para establecerla como clase o etiqueta, por lo que fue necesario buscar una estrategia para etiquetar los datos a fin de generar un modelo que permita alcanzar los objetivos de la investigación, la cual se expone en los ítems posteriores.

De igual manera se puede observar que existen variables categóricas que requieren ser tratadas para poder ser analizadas, así como variables que deben ser discriminadas ya que no serán de utilidad para el estudio, el tipo de tratamiento se lo definirá en el siguiente paso según lo que se pueda advertir con una mayor exploración de los datos.

**Exploración de los Datos**. Agrupamos los datos por países, para conocer sus totales. Esto nos da como resultado que existe una notable cantidad de conexiones desde países externos.

**Figura 7.**

*Conexiones por Países.*

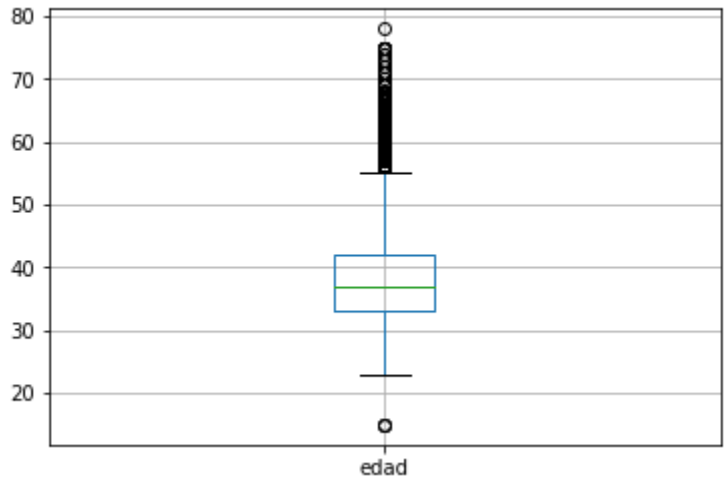
| pais | |
|---|---|
| | 14 |
| Argentina | 449 |
| Australia | 2 |
| Belgium | 1 |
| Bolivia | 2 |
| Bonaire, Sint Eustatius and Saba | 1 |
| Brazil | 257 |
| Canada | 421 |
| Chile | 161 |
| China | 2 |
| Colombia | 2592 |
| Costa Rica | 2 |
| Czech Republic | 7 |
| Dominican Republic | 2 |
| Ecuador | 1842657 |
| El Salvador | 26 |
| Europe | 8 |
| France | 65 |
| Germany | 224 |
| Hong Kong | 638 |

*Nota*. Totales de conexiones agrupadas por países.

Revisamos datos atípicos (Outliers).

**Figura 8.**

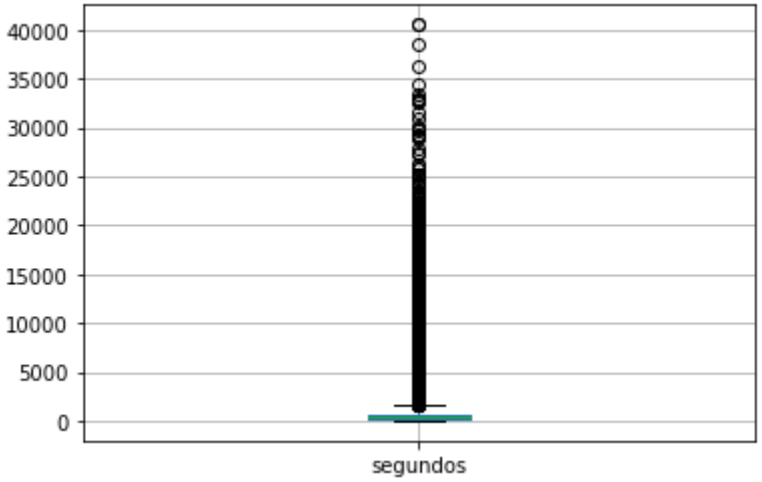
*Valores atípicos de variable edad*




*Nota*. Diagrama de Caja y Bigotes para apreciar los valores atípicos de la variable edad.

**Figura 9.**

*Valores atípicos de variable segundos.*




*Nota*. Diagrama de Caja y Bigotes para apreciar los valores atípicos de la variables segundos

Revisamos si los datos poseen valores nulos

**Figura 10.**

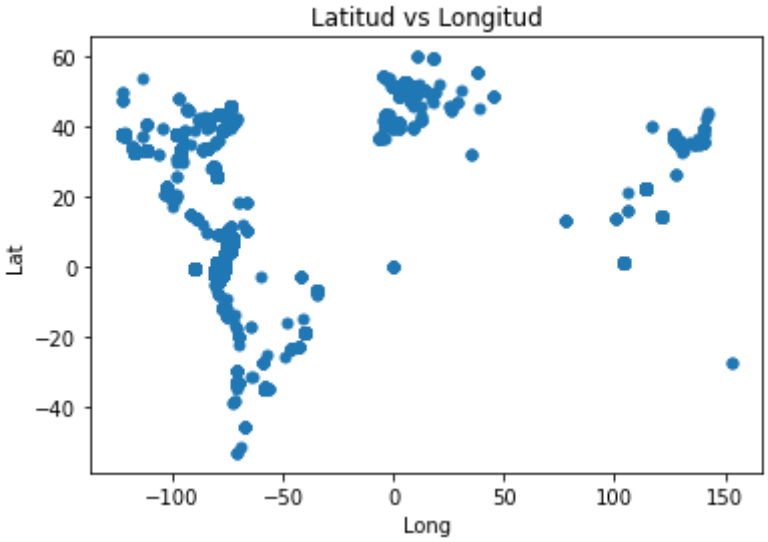
*Revisión de valores nulos*

```
auIdGenUsuario         0
fini                   0
fFin                   0
diaSemana              0
hini                   0
hfin                   0
hini1                  0
ipini                  0
ipfin                  0
iptipo                 0
segundos               0
estadoCivil        15583
sexo                 169
situacion           3388
edad                 169
long                   0
lat                    0
pais                   0
ciudad            341993
dtype: int64
```

*Nota*. Análisis de las variables con valores nulos.

**Figura 11.**

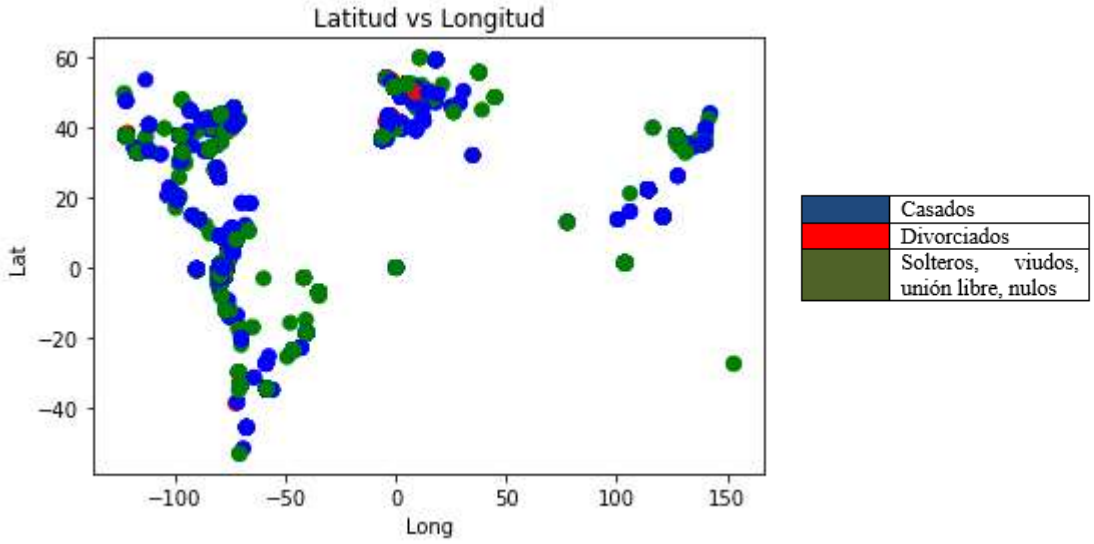
*Diagrama Latitud vs Longitud*




*Nota*. Análisis de las variables Latitud y Longitud.

**Figura 12.**

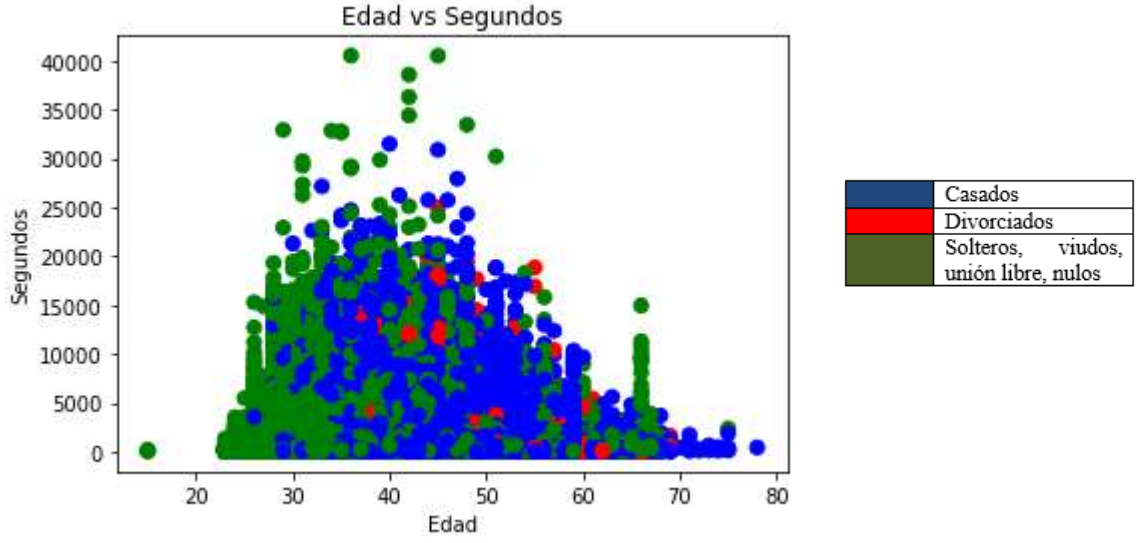
*Diagrama Latitud versus Longitud por Estado civil*




*Nota*. Análisis de las variables latitud y longitud por estado civil.

**Figura 13.**

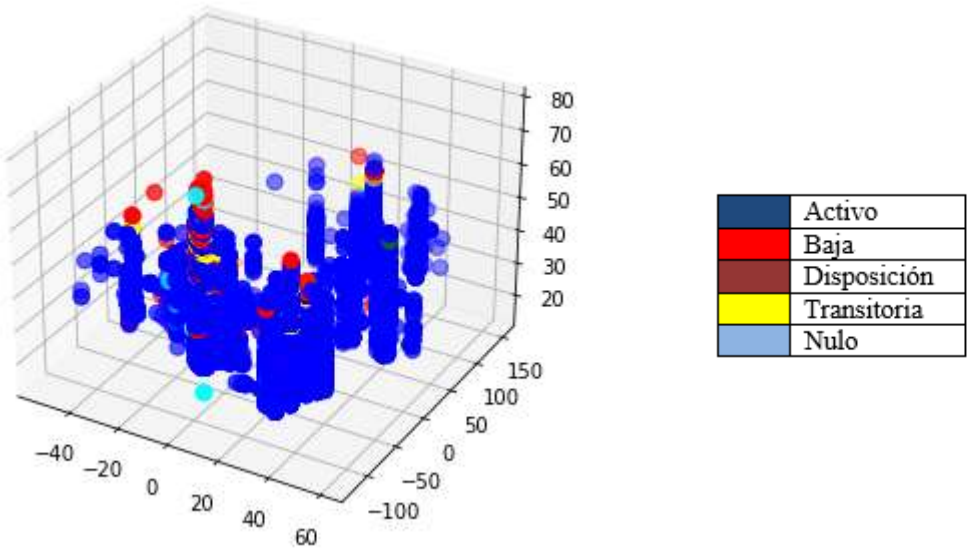
*Diagrama Edad vs segundos por estado civil*




*Nota*. Análisis de las variables edad y segundos por estado civil.

**Figura 14.**

*Diagrama 3D variables Latitud, longitud, edad por situación policial.*




*Nota*. Análisis en 3 dimensiones de las variables Latitud, longitud y edad por situación policial.

El análisis exploratorio nos proporciona información relevante respecto a las variables. De esta manera encontramos variables categóricas como: sexo, estado civil y situación policial que pueden ser consideradas en la clasificación de los datos.

Igual manera tenemos variables discretas como: latitud, longitud, edad, segundos que podríamos considerar.

La exploración de datos además nos brinda información importante de las variables que poseen valores atípicos y nulos, los cuales deben ser tomados en cuenta para la aplicación de los algoritmos de clasificación supervisados, en los cuales se requiere contar con clases etiquetas para poder entrenar el modelo de predicción.

**Verificaciones y gestión de la calidad**. Los diagramas de bigotes de las variables edad y segundos, nos muestran una cantidad considerable de datos atípicos. Dichos datos atípicos se los pondrá a consideración de expertos para establecer el tipo de tratamiento.

De igual manera se ha encontrado campos con valores nulos así como valores en blanco que pueden requerir limpieza, o quizá tener un análisis bajo el juicio de expertos para determinar su tratamiento.

No se ha encontrado valores truncados durante la exploración del set de datos.

***Fase III. Preparación de los Datos. Análisis de los datos y selección de características***

**Selección de Datos**. Para la selección de datos el análisis de la problemática nos brinda información importante, es así que la variable situacion_policial se ha elegido debido a que el análisis de la problemática nos indica que el 80 % de ataques a sistemas informáticos son producidos por personas internas y ex trabajadores de una empresa.

El significado de dicha variable en el caso de estudio que se investiga posee una característica especial por cuánto difiere con el de un trabajador normal, esto se debe a que los funcionarios policiales no solo manejan estados como Alta y Baja, sino que pueden encontrarse en estados especiales denominados Transitoria y A Disposición. Estos estados para el primer caso permite al funcionario mantenerse aun siendo parte de la institución mientras se establece la tramitología para su proceso de Baja o Cesación y en el segundo caso por motivos de mantener algún proceso judicial debido a las funciones específicas de su profesión.

Para la selección de la variable edad se ha considerado otro antecedente de la problemática, mismo que indica que muchas personas que se encuentran en ciertos rangos de edad suelen solicitar la ayuda de terceros para la ejecución de sus trámites personales en la plataforma, a los cuales suelen compartir sus credenciales de usuarios.

Latitud y Longitud. De la exploración se ha podido notar que los lugares de acceso son variados, al fijarnos en la variable país existe un gran número de conexiones realizadas desde países externos, este resultado lo hemos puesto en consideración del juicio de expertos, los cuales en esta parte nos indican que al ser un sistema que se usa a nivel nacional por parte de

funcionarios policiales, la mayoría de conexiones deberían provenir de lugares pertenecientes al país, por lo que llama la atención aquellas conexiones desde países externos debiendo considerar lo encontrado ya que nos genera sospechas.

Segundos. Es importante analizar esta variable que almacena el tiempo de conexión de una sesión, el análisis exploratorio nos da una idea de los tiempos promedio de la mayoría de sesiones. Aquellas conexiones que sobrepasan o se encuentran por debajo de dichos tiempos nos generan sospechas. Ha lo expuesto debemos añadir que los tiempo de caducidad de sesiones establecidos para el sistema rondan lo determinado en la media encontrada.

Se incluyen las variables Sexo y estadoCivil con el propósito de obtener una mejor perspectiva en la clasificación para conocer si hay algún comportamiento inusual o influencia de parte de dichas variables en la detección de anomalías.

Se excluye las variables usercode, fini, ffin, hini, hfin, ipini, ipfin, iptipo, país, ciudad ya que dichas variables serán tratadas y depuradas para generar nuevas variables con información que pueda ser incluida en el modelamiento de los algoritmos que se pretende generalizar para solventar el problema planteado dentro de la investigación que se estudia. Las variables generadas a partir de éstas y que incluyen información relevante para el caso de estudio son long, lat, hini1, segundos, diaSemana.

**Tabla 10.**

*Set de Datos Final*

| Campo | Tipo de Dato | Descripción |
|---|---|---|
| diaSemana | int64 | Día de la semana (0-7) conectado |
| hini1 | int64 | Hora inicio de conexión. (0-24) |
| segundos | int64 | Tiempo en segundos conectado |
| estadoCivil | Object | Estado Civil del usuario |
| sexo | Object | Sexo del usuario |

| Campo | Tipo de Dato | Descripción |
|---|---|---|
| situacion | Object | Situación en la institución del usuario. |
| edad | float64 | Edad del usuario |
| long | float64 | Coordenada longitud |
| lat | float64 | Coordenada latitud |

*Nota*. Listado de campos elegidos como set de Datos final.

**Limpieza de Datos**. Las variables que poseen valores nulos y en blanco deben incluirse en el análisis por cuanto constituyen un insumo de importancia para el modelo a aplicar, en vista que lo que se requiere es determinar aquellas instancias atípicas.

De acuerdo al juicio de expertos el algoritmo debe estar en la capacidad de analizarlos para determinar su clasificación o no dentro del grupo de accesos no autorizados.

Los valores atípicos encontrados en la etapa de exploración de la información no serán eliminados, ya que son necesarios para el estudio.

**Construcción del Juego de Datos**. Para el cálculo de nuevos campos necesarios para el análisis se aplicaron funciones SQL al momento de la descarga de información de las pistas de auditoria, las cuales se detallan a continuación.

```
CASE when DATE_FORMAT(a.fechaLogeo,'%w') = 0 then 7 else DATE_FORMAT
(a.fechaLogeo,'%w')  end diaSemana
```

Calculo del día de la Semana en función de la fecha de conexión

```
DATE_FORMAT(a.fechaLogeo,'%Y-%m-%d') fini,
DATE_FORMAT(b.fechaLogeo,'%Y-%m-%d') fFin
```

Extracción de las fechas.

```
DATE_FORMAT(a.fechaLogeo,'%H:%i:%s') hini,
DATE_FORMAT(b.fechaLogeo,'%H:%i:%s') hfin,
DATE_FORMAT(a.fechaLogeo,'%H') hini1,
```

Extracción de la hora de conexión en función de la fecha

```
TIMESTAMPDIFF(SECOND,a.fechaLogeo,b.fechaLogeo) as segundos
```

Calculo del tiempo de conexión en segundos en función de las fechas.

```
REGEXP_substr(a.obs,'[0-9,.]+') ipini,
REGEXP_substr(b.obs,'[0-9,.]+') ipfin,
```

Extracción de la cadena IP a partir del campo donde se almacena la IP.

```
REGEXP_substr(a.obs,'[^0-9,:]+') iptipo
```

Extracción del tipo de IP a partir del campo donde se almacena la IP.

**Integración de Datos**. Como se detalló en la captura de datos, la recopilación inicial de la información almacenada en las pistas de auditoria se la ejecutó mediante un script SQL.

Sin embargo mencionada información no posee detalles adicionales, ventajosamente es posible extraer información adicional de algunos de los valores almacenados en sus campos si se integra con información de otras fuentes, es así que utilizando la variable ipini se adiciona datos de geolocalización empleando la base de datos GeoIP de libre acceso. (*Free updated GeoIP legacy databases*, 2021). Y de igual manera de la variable usercode empleando la base de datos de servidores policiales.

Para la extracción de la información citada se desarrollaron scripts en lenguaje de programación Python.

**Formateo de Datos**. Para las variables categóricas estado civil, sexo y situación policial se aplicará la técnica de codificación One Hot Encoding (Shipra Saxena, 2020), para cambiar su contenido por valores numéricos por cuanto los algoritmos de aprendizaje automático trabajan en su mayoría con este tipo de datos. La elección de dicha técnica se debe a que los valores no poseen un orden determinado por lo tanto es la que mejor se ajusta para el objetivo.

Para la aplicación del algoritmo Clustering KMeans no se aplicó formateo de datos adicional en las variables que se utilizó.

Para la aplicación del algoritmo de Árbol de Decisiones, se aplicó el siguiente formateo de datos para las variables edad y segundos.

Se dividió los valores en 5 categorías según lo siguiente:

Valores menores al límite inferior

Valores comprendidos entre límite inferior y primer cuartil

Valores comprendidos entre el primer y tercer cuartil

Valores comprendidos entre el tercer cuartil y límite superior

Valores comprendidos mayores al límite superior

Para el cálculo de los límites superior e inferior se usó la formula estadística:

$$\text{Límite inferior} = Q1 - 1.5*(Q3-Q1)$$

$$\text{Límite superior} = Q3 + 1.5*(Q3-Q1)$$

Para la aplicación del algoritmo Isolation Forest, no se aplica formateo de datos adicional.

***Fase IV. Modelado***

Dado que la investigación se orienta a la identificación de accesos no autorizados, lo que se pretende encontrar es aquellos patrones inusuales que son diferentes de la normalidad, que podrían indicarnos que se trata de accesos no autorizados.

Ante lo analizado lo que se requiere es poder diferenciar la información en patrones usuales y patrones inusuales, por lo que los algoritmos de clasificación son los que mejor se ajustan al problema planteado.

Una de las estrategias a considerar para resolver el problema mencionado es determinar los patrones de normalidad para posterior identificar los anómalos, lo cual se lo puede alcanzar con la aplicación de algoritmos mayormente conocidos como clustering y árboles de decisiones. Sin embargo el objetivo también se lo puede lograr usando una estrategia diferente propuesta por (Liu et al., 2008), que consiste en explícitamente aislar las anomalías en lugar de los perfiles normales a través de la aplicación de un algoritmo denominado Isolation Forest.

Un punto importante a considerar es que los datos recolectados no poseen una clase etiqueta, la cual es necesaria para poder aplicar un algoritmo de clasificación supervisado y

realizar predicciones. En virtud de aquello el estudio del arte nos sugiere algunas técnicas de las cuales se trabajó con la combinación de dos de ellas: etiquetación a través de clustering y a través de criterio de expertos.

El objetivo consiste en aplicar clustering y etiquetar las instancias del set de datos como “usual” e “inusual” usando las variables de mayor importancia, posterior localizar los valores atípicos del resto de variables y etiquetarlos como anormales. Con éste procedimiento alcanzaríamos un set de datos etiquetado al cual le aplicaremos el algoritmo de Árbol de Decisiones para entrenar un modelo de clasificación.(Pius Owoh et al., 2018)

Para la validación del modelo usaremos una matriz de confusión (técnica de validación externa) comparando los resultados arrojados por el modelo entrenado con el algoritmo de Árbol de Decisiones y las etiquetas iniciales, para finalmente calcular las métricas de validación necesarias tales como: F1-measure, accuracy, precisión, recall.

Lo siguiente a realizar es obtener un nuevo modelo entrenado con el set de datos que carece de una clase etiqueta aplicando el algoritmo Isolation Forest, con el fin de comparar algoritmos y determinar cuál de ellos se ajusta mejor al caso de estudio

Para validar éste modelo nos apoyaremos en la técnica de Twin-sample validation (Priyanshu Jain, 2020), para generar lo que el autor denomina como muestras gemelas y poder obtener dos sets de resultados que puedan ser comparables con las mismas técnicas de validación externa del primer modelo.

Finalmente obtendremos los resultados de las métricas y los compararemos para establecer las conclusiones correspondientes y responder a las preguntas de investigación y objetivos planteados en el caso de estudio.

**Etiquetado con Clustering**. Como algoritmo de clúster se eligió K-means.

Se elige las variables de nuestro dataset con las cuales vamos a trabajar. Para este caso latitud y longitud.

**Figura 15.**

*Selección de variables en Clustering.*

```
df.head()
```

| | long | lat |
|---|---|---|
| **0** | -79.9011 | -2.1664 |
| **1** | -79.5077 | -1.8250 |
| **2** | -77.4966 | -1.9998 |
| **3** | -79.9011 | -2.1664 |
| **4** | -79.9011 | -2.1664 |

*Nota*. Selección de variables para aplicar Clustering

Verificamos si nuestras dimensiones tienen valores nulos o NAN

**Figura 16.**

*Determinar valores nulos en Clustering*

```
df.isnull().sum()
```

```
long    14
lat     14
dtype: int64
```

*Nota*. Análisis de nulos

Rellenamos los valores nulos y NAN.

Utilizando métodos ffill, Existen diferentes métodos.

**Figura 17.**

*Rellenado de nulos para Clustering.*

```
df.fillna(method='ffill',inplace=True)
```

*Nota*. Uso de método ffill para llevar valores nulos.

Escalamos o Normalizamos las variables.

Normalizamos los datos con dos variables: latitud, longitud. El método de normalización usado es Z-score, su fórmula es la siguiente:

$$Z = (x-\mu) / \sigma$$

$z = x - \mu \sigma$.

La normalización permite manejar los datos en una sola escala [0,1].

Varios métodos existen: 0-1 estandarización, Z-score y MaxAbs

**Figura 18.**

*Estandarización en Clustering*

```
#Escalamos los datos con estandarizacion Z
from sklearn.preprocessing import StandardScaler
ss = StandardScaler()
Xs = ss.fit_transform(X)
Xs
```

```
array([[-0.1211492 , -0.26875565],
       [-0.06045859, -0.21856532],
       [ 0.24979784, -0.24426324],
       ...,
       [ 0.02031823, -0.37395832],
       [-0.1211492 , -0.26875565],
       [ 0.09473905,  0.01822891]])
```

*Nota*. Normalización de variables.

Calculamos el número de cluster, usamos método Elbow. Ver técnicas.

**Figura 19.**

*Cálculo de número de clusters.*

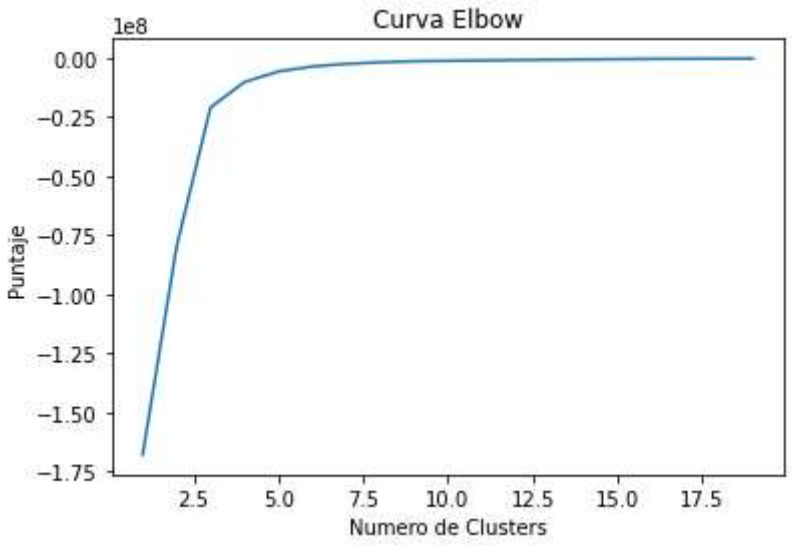
```
X = np.array(df[["long","lat"]])
```

```
#Calculamos el numero de clusters con metodo codo
Nc = range(1, 20)
kmeans = [KMeans(n_clusters=i) for i in Nc]
kmeans
score = [kmeans[i].fit(X).score(X) for i in range(len(kmeans))]
score
plt.plot(Nc,score)
plt.xlabel('Numero de Clusters')
plt.ylabel('Puntaje')
plt.title('Curva Elbow')
plt.show()
```

*Nota*. Uso del método del codo para cálculo de número de clúster.

**Figura 20.**

*Diagrama Curva Elbow*




*Nota*. Análisis del número de clusters.

Numero de clusters para latitud y longitud, Numero encontrado igual a 3

Calculo de centroides para las dos variables sin escalar.

**Figura 21.**

*Centroides sin escalar*

```
: kmeans = KMeans(n_clusters=3).fit(X)
centroids = kmeans.cluster_centers_
print(centroids)
```

```
[[-78.9847268   -1.4783693 ]
 [-93.42628651  38.1346471 ]
 [ 31.51411727  44.20269592]]
```

*Nota*. Cálculo de número de centroides usando kmeans para Clustering con datos normalizados.

Calculo de centroides para las dos variables con datos escalados

**Figura 22.**

*Centroides con escalado*

```
kmeans = KMeans(n_clusters=3).fit(Xs)
centroids = kmeans.cluster_centers_
print(centroids)
```

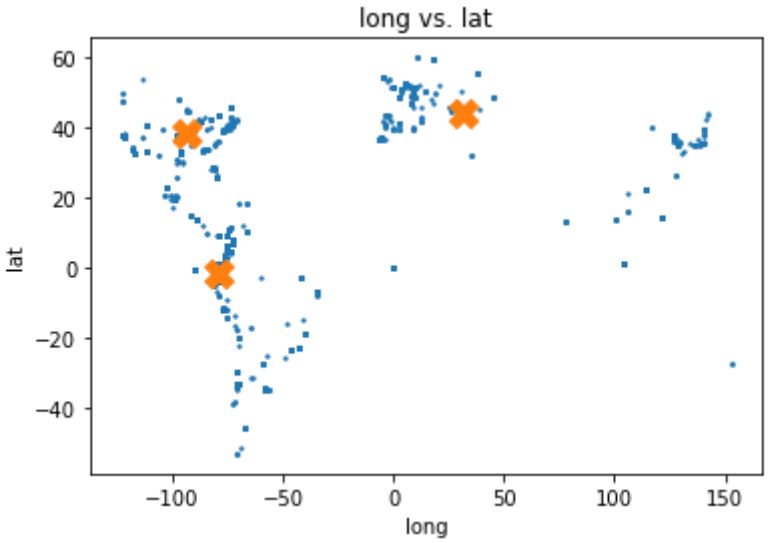
```
[[ 0.02022118 -0.16760613]
 [-2.20770684  5.65603204]
 [17.05111467  6.52674937]]
```

Nota. Cálculo de número de centroides usando kmeans para Clustering con datos sin normalizar.

Visualizamos las variables y sus centroides.

**Figura 23.**

*Diagrama de Centroides.*




*Nota*. Diagrama de variables y sus centroides.

Predecimos con el modelo y obtenemos las etiquetas que las añadimos al set de datos

**Figura 24.**

*Etiquetación con clustering*

```
#Etiquetamos nuestros datos
labels = kmeans.predict(X)
df['label'] = labels
```

*Nota*. Etiquetado de datos.

Verificamos la cantidad de instancias que pertenecen a cada grupo.

**Figura 25.**

*Agrupación de datos por etiqueta*

```
df.groupby('label').size()
```

```
label
0    1846273
1      49797
2       4244
dtype: int64
```

*Nota*. Determinación de valores normales y anomalías.

Para completar la etiquetación de datos, agrupamos las instancias con menor cantidad en una sola siendo las etiquetas con valor cero las normales y las etiquetas con valor 1 anomalías.

Además también etiquetamos como anomalías aquellas instancias cuyos valores están en el rango de atípicos en una o más de sus dimensiones. Para la consecución de este objetivo se aplicó técnicas de asociación para ciertos campos en función del juicio de expertos.

**Modelado con Árbol de Decisión**. Continuando las actividades para alcanzar el objetivo del estudio, una vez etiquetado el set de datos ya podemos entrenar un modelo de predicción. Para lo indicado se usa el algoritmo de árbol de decisión con el propósito de clasificar los datos en usuales e inusuales.(Juan Ignacio Bagnato, 2018)

Empezamos cargando el set de datos y escogemos las variables con las cuales vamos a trabajar.

No es necesario escalar el set de datos, ya que el algoritmo de árbol de decisiones puede trabajar sin requerir una transformación en una representación común. (Pang-Ning Tan, 2019)

Dividimos set de datos en prueba, entrenamiento y validación.

**Figura 26.**

*Muestras de pruebas, entrenamiento y validación. Árbol de Decisión*

```
from sklearn.model_selection import train_test_split
y = df['label']
x = df.drop(['label'], axis=1)
X_train, X_rem, Y_train, Y_rem = train_test_split(x,y, train_size=0.7)
X_valid, X_test, y_valid, y_test = train_test_split(X_rem,Y_rem, test_size=0.5)
```

```
X_train.head()
```

| | situacion_cod | ecivil_cod | sex_cod | seg_cod | edad_cod |
|---|---|---|---|---|---|
| **1458225** | 0 | 2 | 1 | 0.0 | 2.0 |
| **826590** | 0 | 2 | 1 | 1.0 | 3.0 |
| **1121627** | 0 | 0 | 1 | 1.0 | 3.0 |
| **1355674** | 0 | 3 | 1 | 0.0 | 2.0 |
| **87918** | 1 | 0 | 1 | 3.0 | 4.0 |

*Nota*. División de set de datos para aplicar algoritmo Árbol de Decisión.

Calculamos la profundidad que mejor se ajusta. En este punto debemos considerar si los datos se encuentran balanceados o no ya que de esto dependerá que el algoritmo generalice mejor la información.

Una muestra desbalanceada puede perjudicar a las clases minoritarias generando que el modelo no las prediga adecuadamente, algunas técnicas para el manejo de estos datos son:

Ajustar los parámetros del modelo, que consiste en ajustar el parámetro class_weight=”balanced” que poseen algunos algoritmos de aprendizaje automático.

Técnicas de muestreo como sub-sampling que consiste en quitar muestras de la clase mayoritaria y over-sampling que añade copias de la clase minoritaria. (Juan Ignacio Bagnato, 2019)

Para el caso de estudio se utilizó el ajuste de parámetro para entrenar el modelo con el algoritmo árbol de decisión.

**Figura 27.**

*Profundidad del árbol de decisión.*

```
from sklearn.model_selection import KFold
from sklearn import tree
cv = KFold(n_splits=10) # Numero deseado de "folds" que haremos
accuracies = list()
max_attributes = len(list(df))
depth_range = range(1, max_attributes + 1)

# Testearemos la profundidad de 1 a cantidad de atributos +1
for depth in depth_range:
    fold_accuracy = []
    tree_model = tree.DecisionTreeClassifier(criterion='entropy',
                                             min_samples_split=20,
                                             min_samples_leaf=5,
                                             max_depth = depth,
                                             class_weight='balanced')
    for train_fold, valid_fold in cv.split(df):
        f_train = df.loc[train_fold]
        f_valid = df.loc[valid_fold]

        model = tree_model.fit(X = f_train.drop(['label'], axis=1),
                               y = f_train["label"])
        valid_acc = model.score(X = f_valid.drop(['label'], axis=1),
                                y = f_valid["label"]) # calculamos la precision con el segmento de validacion
        fold_accuracy.append(valid_acc)

    avg = sum(fold_accuracy)/len(fold_accuracy)
    accuracies.append(avg)

# Mostramos los resultados obtenidos
dfr = pd.DataFrame({"Max Depth": depth_range, "Average Accuracy": accuracies})
dfr = dfr[["Max Depth", "Average Accuracy"]]
print(dfr.to_string(index=False))
```

```
 Max Depth  Average Accuracy
         1          0.965348
         2          0.973218
         3          0.973218
         4          0.973260
         5          0.973321
         6          0.973356
```

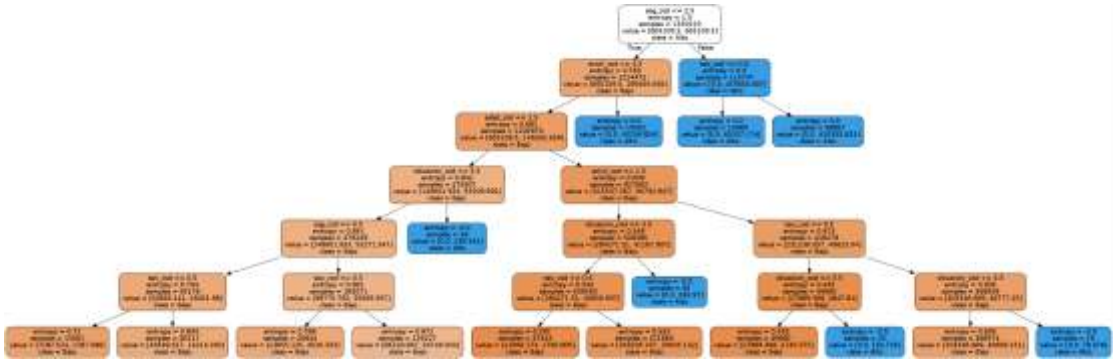
*Nota*. Cálculo de la profundidad del Árbol de Decisión.

Entrenamos el modelo con la profundidad encontrada que mejor accuracy nos proporciona, en este caso max Depth = 6 con una precisión de 0.97335.

Representación gráfica del modelo entrenado.

**Figura 28.**

*Diagrama del Árbol de Decisión entrenado.*



*Nota*. Diagrama de árbol del modelo entrenado con árbol de decisión.

Validación con matriz de Confusión

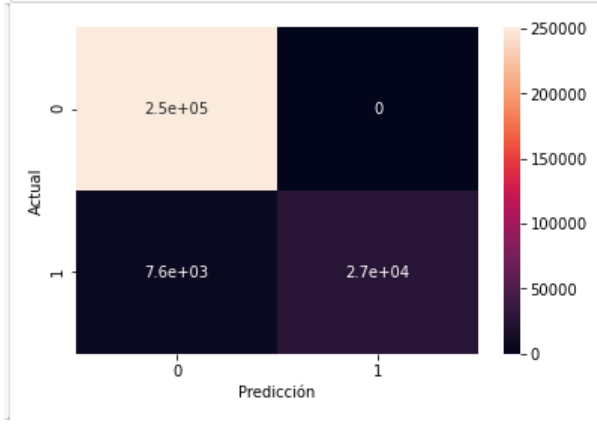
**Figura 29.**

*Matriz de Confusión Árbol de Decisión*

```
import seaborn as sn
import matplotlib.pyplot as plt

confusion_matrix = pd.crosstab(y_valid, labelp, rownames=['Actual'], colnames=['Predicción'])

sn.heatmap(confusion_matrix, annot=True)
plt.show()
```




*Nota*. Mapa de Calor de la matriz de confusión del árbol de decisión.

Métricas de validación

**Figura 30.**

*Medidas de validación Árbol de Decisión.*

```
from sklearn.metrics import classification_report
print (classification_report(y_valid, labelp))
```

```
              precision    recall  f1-score   support

           0       0.97      1.00      0.99    250566
           1       1.00      0.78      0.88     34481

    accuracy                           0.97    285047
   macro avg       0.99      0.89      0.93    285047
weighted avg       0.97      0.97      0.97    285047
```

*Nota*. Análisis de métricas de validación del árbol de decisión.

**Modelado con Bosque de Aislamiento**. El algoritmo Isolation Forest no utiliza medidas de distancia o densidad para detectar las anomalías lo cual lo convierte en un algoritmo mucho más eficiente ya que elimina el costo computacional de cálculo de distancias. Adicionalmente tiene la capacidad de manejar grandes set de datos y con alta dimensionalidad. (Liu et al., 2008)

Cargamos el set de datos y elegimos las variables con las cuales se va a trabajar, revisamos si existen valores nulos y los rellenamos.

**Figura 31.**

*Isolation Forest. Elección de Variables*

```
df = pd.read_parquet("bitacora_user3.parquet.gzip")
```

```
df1.isna().sum()
```

```
segundos          0
edad            169
long              0
lat               0
situacion_cod     0
sex_cod           0
diaSemana         0
hini1             0
ecivil_cod        0
dtype: int64
```

```
df1.fillna(value=0,inplace=True)
```

*Nota*. Selección de variables, análisis y llenado de valores nulos.

Dividimos set de datos en prueba, entrenamiento y validación.

**Figura 32.**

*Isolation Forest. División de datos*

```
from sklearn.model_selection import train_test_split
X_train, X_test = train_test_split(df1, test_size = 0.30)
```

*Nota*. División del set de datos en set de pruebas entrenamiento y validación para aplicar Isolation Forest.

Determinamos los parámetros óptimos.

**Figura 33.**

*Isolation Forest. Parámetro optimo*

```
from sklearn.model_selection import GridSearchCV
from sklearn import model_selection
model = IsolationForest(random_state=47)

param_grid = {'n_estimators': [1000, 1500],
              'max_samples': [10],
              'contamination': ['auto', 0.0001, 0.0002, 0.01],
              'max_features': [9, 15],
              'bootstrap': [True],
              'n_jobs': [-1]}

grid_search = model_selection.GridSearchCV(model,
                                           param_grid,
                                           scoring="neg_mean_squared_error",
                                           refit=True,
                                           cv=10,
                                           return_train_score=True)
grid_search.fit(X_train)

best_model = grid_search.fit(X_train)
print('Parametros óptimos', best_model.best_params_)
```

*Nota*. Determinación de parámetros óptimos.

Ajustamos parámetros y entrenamos el modelo de datos.

**Figura 34.**

*Entrenamiento Isolation Forest.*

```
modelo_isof = IsolationForest(
                n_estimators  = 1000,
                max_samples   ='auto',
                contamination = 0.01,
                n_jobs        = -1,
                random_state  = 123,
              )
```

```
modelo_isof.fit(X=X_train)
```

```
IsolationForest(contamination=0.01, n_estimators=1000, n_jobs=-1,
                random_state=123)
```

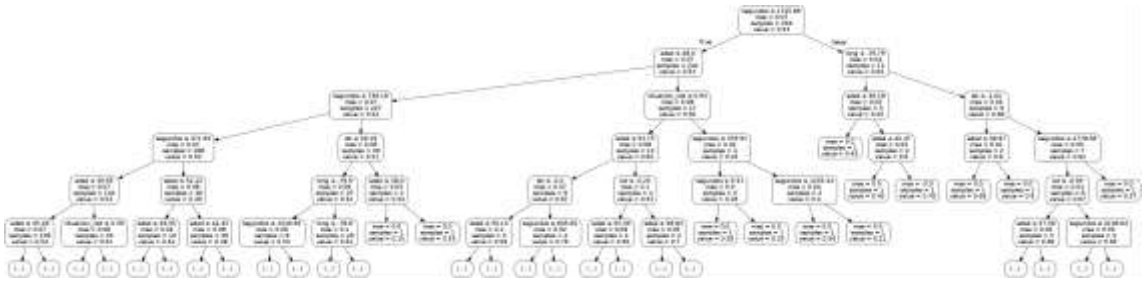
```
pred = modelo_isof.fit_predict(X_train)
```

*Nota*. Entrenamiento del modelo de datos con Isolation Forest.

Representación gráfica del modelo entrenado.

**Figura 35.**

*Diagrama Modelo Isolation Forest*



*Nota*. Diagrama de árbol del modelo entrenado con Isolation Forest.

Validamos el modelo con twin-sample. (Priyanshu Jain, 2020). La validación con twin-sample es un tipo de validación interna la cual consiste en:

Crear un set de datos gemelo del set de datos de entrenamiento.

Aplicar el algoritmo de aprendizaje no supervisado sobre el set de datos gemelo.

Importar resultados para el set de datos gemelos desde el set de datos de entrenamiento.

Calcular similitudes entre los dos set de resultados.

**Figura 36.**

*Entrenamiento Muestra Gemela*

```
X=X_train[["segundos","edad","long","lat","situacion_cod","sex_cod","diaSemana","hini1","ecivil_cod"]]
y=X_train[["anomaly_label"]]

X_twen,X_resto,Y_twen,Y_resto = train_test_split(X,y , test_size=0.50, random_state=1, stratify=y)
```

```
modelo_isof_twen = IsolationForest(
                n_estimators  = 1000,
                max_samples   ='auto',
                contamination = 0.01,
                n_jobs        = -1,
                random_state  = 123,
           )
```

```
modelo_isof_twen.fit(X=X_twen)
```

```
IsolationForest(contamination=0.01, n_estimators=1000, n_jobs=-1,
                random_state=123)
```

*Nota*. Aplicación de muestra gemela para validación de Isolation Forest.

Matriz de confusión.

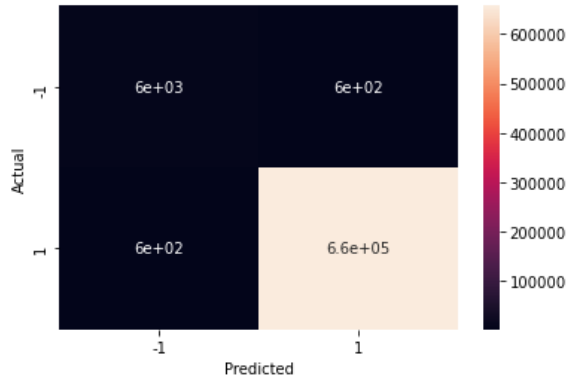
**Figura 37.**

*Validación Isolation Forest*

```
import seaborn as sn
import matplotlib.pyplot as plt

dfm = pd.DataFrame(X_twen, columns=['s','p'])
confusion_matrix = pd.crosstab(dfm['s'], dfm['p'], rownames=['Actual'], colnames=['Predicted'])

sn.heatmap(confusion_matrix, annot=True)
plt.show()
```




*Nota*. Mapa de calor de la matriz de confusión del modelo Isolation Forest.

Calculamos las métricas de validación.

**Figura 38.**

*Medidas Validación Isolation Forest*

```
from sklearn.metrics import classification_report
print (classification_report(dfm['s'], dfm['p']))
```

```
              precision    recall  f1-score   support

          -1       0.91      0.91      0.91      6652
           1       1.00      1.00      1.00    658457

    accuracy                           1.00    665109
   macro avg       0.95      0.95      0.95    665109
weighted avg       1.00      1.00      1.00    665109
```

*Nota*. Cálculo de medidas de validación del modelo Isolation Forest.

Comparamos las métricas obtenidas de los dos modelos aplicados.

**Tabla 11.**

*Comparación modelos.*

| **Árbol de Decisiones.** | | | | | **Isolation Forest** | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Precision | recall | F1-score | Support | | Precision | recall | F1-score | support |
| 1 (Anomalía) | 1.00 | 0.78 | 0.88 | 34481 | -1 (Anomalía) | 0.91 | 0.91 | 0.91 | 6652 |
| 0 (Normal) | 0.97 | 1.00 | 0.99 | 250566 | 1 (Normal) | 1.00 | 1.00 | 1.00 | 658457 |
| Accuracy | | | 0.97 | 285047 | | | | 1.00 | 665109 |
| Macro avg | 0.99 | 0.89 | 0.93 | 285047 | | 0.95 | 0.95 | 0.95 | 665109 |
| Weighted avg | 0.97 | 0.97 | 0.97 | 285047 | | 1.00 | 1.00 | 1.00 | 665109 |

*Nota*. Comparación de resultados entre modelos Isolation Forest y Árbol de Decisión.

***Fase V. Evaluación (obtención de resultados)***

Los resultados obtenidos permiten comparar las métricas de validación de los dos modelos entrenados, si bien es cierto ambos modelos nos proporcionan valores aceptables en

sus medidas, sin embargo se puede percibir una diferencia de mejora al aplicar el algoritmo Isolation Forest en la detección de anomalías.

Otro punto importante que debemos rescatar es que el algoritmo Isolation Forest no presentó inconvenientes al trabajar con variables adicionales como hini1 y diaSemana del set de datos elegido. A diferencia del algoritmo de Árbol de Decisión, el cual generó valores por debajo de los presentados en la Tabla No.10 cuando se adicionaron las variables hini1 y diaSemana.

En adición el algoritmo Isolation Forest, no requirió de preparación extra sobre los datos como: estandarización o normalización, agrupación, entre otros para su aplicación, lo cual si fue indispensable en el modelo de Árbol de Decisión para mejorar sus métricas de validación.

Debido a dichas observaciones, el modelo entrenado con el algoritmo Isolation Forest presenta mejores resultados en la consecución del objetivo del caso de estudio en cuestión.

Sin embargo se debe considerar que la falta de etiquetas en el set de datos limita la aplicación de algoritmos de aprendizaje supervisado, requiriendo de un trabajo arduo según el tamaño de datos con el que se esté trabajando, por lo que futuros estudios podrían ahondar en el uso de diversas técnicas de etiquetación de datos y así mejorar el set de datos para la aplicación de algoritmos supervisados.

De igual manera estudios posteriores pueden agregar dimensiones al set de datos o realizar una clasificación multiclase en lugar de una clasificación binaria.

***Fase VI. Despliegue (puesta en producción)***

El estudio pretende generar conocimiento a partir de la información que genera la plataforma tecnológica de la institución policial, dicho conocimiento será puesto a disposición de las autoridades para una toma de decisiones acertada.

La implementación en un ambiente de producción quedará a criterio de dichas autoridades, las cuales deberán analizar si es pertinente desplegar el caso de estudio en su

infraestructura o si lo toma como un punto de partida para generar nuevo conocimiento añadiendo información y variables que estén a su alcance.

Se sugiere considerar la disponibilidad de sus recursos y presupuesto para la adquisición de los datos de geolocalización por IP con el fin de mejorar el algoritmo por cuanto para la investigación del caso de estudio que se ha abordado se hizo uso de una versión gratuita. Se debe considerar además el consumo de recursos que se requerirá en caso de necesitar procesar una mayor cantidad de información.

Otro aspecto importante por la cual se deja a consideración de la entidad policial constituye los análisis de factibilidad, las autorizaciones correspondientes y demás requisitos normativos necesarios que son comúnmente llevados en una entidad de tipo estatal por parte de las autoridades encargadas de la administración de la plataforma tecnológica para implementar cualquier tipo de solución en un ambiente de producción.

**Elaboración del Reporte de Caso de Estudio**

El resultado del caso de estudio se expone en el siguiente capítulo donde se aborda los resultados de la investigación.

# Capítulo IV

## Resultados de la Investigación

**OE1-RQ1.1. ¿Cuáles son los estudios existentes sobre detección de accesos no autorizados a sistemas de información usando datos de pistas de auditoría?**

Según los datos extraídos del estado del arte, existen pocos estudios relacionados con el uso de herramientas de Aprendizaje Automático para la detección temprana de accesos no autorizados.

Se revisaron tres bibliotecas digitales: Springer, ACM Digital y IEEEXplorer. Se analizaron muchos estudios de aplicación del aprendizaje automático para la detección de intrusos enfocadas en accesos a la red que nos servirán de línea base para el caso de estudio que se investiga.

En la literatura no se ha identificado un estudio sobre la detección de accesos no autorizados a sistemas de información web, por lo que se considera será un aporte de nuestra investigación.

Sin embargo estudios realizados en otros ámbitos utilizan la detección de observaciones con valores que difieren del resto (outliers) para establecer anomalías. Este principio puede ser aplicado para detectar los accesos no autorizados o intrusiones a un sistema informático.

La revisión de la literatura indica que la detección de outliers se la puede alcanzar tanto con algoritmos supervisados como con no supervisados, para lograr dicho objetivo los algoritmos supervisados requieren de set de datos etiquetados en contraparte éstas no son necesarias para los algoritmos no supervisados. La literatura además indica que los algoritmos de clasificación brindan mejor ajuste para detectar anomalías

**OE1-RQ1.2. ¿Cuáles son las herramientas de ML Open Source que mejor se ajustan para el cumplimiento de los objetivos?**

Partiendo de las herramientas encontradas durante la revisión bibliográfica en fuentes de información confiables se eligió aquellas que nos permiten su uso de manera libre y un manejo y procesamiento aceptable de la cantidad de registros generados durante el año 2016 por el sistema con los dispositivos tecnológicos de hardware que se dispone.

Se descartó pySpark, a consecuencia del hardware que se dispone para el estudio, por cuanto pySpark y apache Spark trabajan mucho mejor con clúster de máquinas y requieren una mayor configuración y afinamiento para obtener el rendimiento deseado. Por lo tanto las herramientas que mejor se han ajustado a nuestro estudio y nos han brindado buenos resultados son las siguientes:

SckiLearn – Modelado Machine Learning.

Pandas – Análisis exploratorio de Datos (EDA)

Jupyter Notebook. Entorno de Desarrollo (IDE)

Python. -> Lenguaje de Programación.

**OE2-RQ2.1. ¿Cuáles son los algoritmos supervisados que mejor se ajustan al caso de estudio?**

El estudio del arte realizado nos arroja como resultado que los algoritmos de clasificación son los que mejor se ajustan para nuestro caso de estudio.

Sin embargo aquellos con los que se obtuvo mejor resultados para el objetivo propuesto fueron los siguientes:

Para el proceso de etiquetación de nuestro set de datos hemos optado por el uso de las técnicas de: etiquetación a través de clustering, eligiendo el algoritmo de clúster K-means y etiquetación a través del criterio de expertos. En este punto se utilizó estadística descriptiva,

Para el proceso de entrenamiento del modelo de clasificación se aplicó un algoritmo supervisado denominado Árbol de decisiones y un algoritmo no supervisado denominado Isolation Forest.

La aplicación del algoritmo supervisado Árbol de Decisión para la detección de accesos no autorizados (anomalías) de manera temprana a través del análisis de pistas de auditoria brinda un 88% de precisión-sensibilidad.

La aplicación del algoritmo no supervisado Isolation Forest para la detección de accesos no autorizados de manera temprana a través del análisis de pistas de auditoria brinda un 91% de precisión-sensibilidad.

Sin embargo ambos algoritmos detectan las instancias normales con porcentajes aceptables de precisión-sensibilidad, en un 99% el algoritmo de Árbol de Decisión y 100% el algoritmo Isolation Forest.

Al comparar los resultados de los dos algoritmos, un mejor resultado se obtuvo con el algoritmo Isolation Forest.

**OE2-RQ2.2. ¿Cuáles son las fuentes de datos que mejor se ajustan al caso de estudio?**

Se seleccionó los logs de auditoria producidos por el sistema SIIPNE3w relacionados al momento en que un usuario se conecta y desconecta al sistema.

Dichos logs de auditoria únicamente almacenan los intentos exitosos de conexión al sistema.

Con esta información se obtuvo los tiempos de conexión en segundos de cada usuario. Además se los complementó con datos adicionales de otras fuentes (Bases de Datos Personal Policial, Bases de Datos de Geolocalización de IP), como son: la edad, estado civil, situación en la institución, ciudad, país, latitud y longitud de donde se conecta. Por lo que en resumen las fuentes de datos que mejor se ajustan son:

- Logs de auditorías producidos por el sistema SIIPNE3w.
- Base de Datos Personal Policial
- Bases de Datos de Geolocalización de IP.

**OE3-RQ3.1. ¿Es posible la generación de una línea base para detectar accesos no autorizados mediante las herramientas de Business Intelligence que se va a utilizar?**

Mediante la aplicación del algoritmo de aprendizaje automático hemos logrado obtener una línea base, de la cual podrán partir los análisis de información por parte de las personas encargadas de llevar el control y seguridad de los datos de la plataforma SIIPNE3w.

Dicha línea base constituye un punto de partida que permitirá reducir tiempo de análisis pudiendo enfocar los esfuerzos en analizar información adicional para poder determinar los malos usos de la información y de la plataforma.

El estudio constituye un primer intento para clasificar la gran cantidad de información que recopila la plataforma relacionada con los accesos de los usuarios, de tal modo de filtrar de alguna manera aquellos que puedan ser catalogados como inusuales o que generen sospecha.

Como estudios posteriores se puede profundizar el estudio agregando dimensiones que permitan afinar el filtrado comparando ubicación de acceso y ubicación de lugar de trabajo, transacciones realizadas y sus permisos, horarios laborales y horarios de uso o acceso a la plataforma, de igual manera se podría ahondar en el uso de diversas técnicas de etiquetación de datos para mejorar el set de datos o realizar una clasificación multiclase en lugar de una clasificación binaria.

**OE3-RQ3.2. ¿Cuáles son las métricas usadas para evaluar la aplicación de algoritmos de aprendizaje automático?**

Las métricas de evaluación varían dependiendo del tipo de algoritmo de aprendizaje automático que se use. En virtud que para el estudio planteado se ha hecho uso de algoritmos de clasificación, las métricas usadas para evaluar los modelos han sido las siguientes:

Matriz de Confusión

Precisión.

Exactitud.

Sensibilidad

Especificidad.

F1 Score

Para la comparación de eficiencia de los modelos, se eligió los porcentajes de la métrica F1-score (precisión-sensibilidad) ya que ésta medida es muy útil cuando se trabaja con clases desbalanceadas, ésta métrica calcula la media armónica de la precisión y la sensibilidad. (Se utiliza la media armónica porque ambos valores son tasas). (Torres, 2020)

# Capítulo V

## Conclusiones y Recomendaciones

### Conclusiones

Los accesos no autorizados o intrusiones a un sistema informático pueden considerarse como anomalías y detectarse identificando las observaciones conocidas como outliers que son valores que difieren del resto.

La detección de outliers se la puede alcanzar tanto con algoritmos supervisados como con no supervisados. Los algoritmos supervisados requieren datos etiquetados de calidad para obtener buenos resultados, en cambio los algoritmos no supervisados, no.

Al comparar los resultados del presente trabajo, en la detección de outliers el algoritmo supervisado Árbol de Decisión brinda un 88% de precisión-sensibilidad, el algoritmo no supervisado Isolation Forest un porcentaje del 91% de precisión-sensibilidad, por lo que se rechaza la hipótesis nula planteada en el presente estudio de caso.

### Recomendaciones

Implementar mecanismos en la plataforma tecnológica SIIPNE3w que permitan de alguna manera almacenar información en los registros de accesos de los usuarios campos del tipo etiqueta, necesarias para una mejor aplicación de los algoritmos supervisados.

Mejorar la captura de datos relacionados a geolocalización del sistema SIIPNE3w en sus pistas de auditoria de los accesos de los usuarios a fin de que se cuente con datos más precisos para el análisis.

Utilizar el modelo elaborado como una línea de partida o primera aproximación para la generación de conocimiento relacionado a la detección de accesos no autorizados.

Mejorar los recursos tecnológicos de hardware que dispone la institución policial para la aplicación de algoritmos de aprendizaje automático con una mayor cantidad de datos.

**Trabajos Futuros.**

La falta de etiquetas en el set de datos limita la aplicación de algoritmos de aprendizaje supervisado, requiriendo de un trabajo arduo según el tamaño de datos con el que se esté trabajando, por lo que futuros estudios podrían ahondar en el uso de diversas técnicas de etiquetación de datos y así mejorar el set de datos para la aplicación de algoritmos supervisados.

Partiendo del presente trabajo estudios posteriores pueden profundizarlo agregando dimensiones que permitan afinar el filtrado comparando ubicación de acceso y ubicación de lugar de trabajo, transacciones realizadas y sus permisos, horarios laborales y horarios de uso o acceso a la plataforma.

El tipo de clasificación también puede motivar un trabajo futuro usando diversas técnicas para realizar una clasificación multiclase en lugar de una clasificación binaria.

**Bibliografía**

Alcalde, A. (2018, marzo 12). *Aprendizaje no Supervisado y Detección de Anomalías: ¿Qué es una Anomalía?* Aprendizaje no Supervisado y Detección de Anomalías: ¿Qué es una Anomalía? https://elbauldelprogramador.com/aprendizaje-nosupervisado-anomalias/

CODIGO ORGANICO INTEGRAL PENAL, COIP, COIP (2018). https://www.defensa.gob.ec/wp-content/uploads/downloads/2018/03/COIP_feb2018.pdf

Bo, L., Jinzhen, W., Ping, Z., Zhongjiang, Y., & Mao, Y. (2016). Research of Recognition System of Web Intrusion Detection Based on Storm. *Proceedings of the Fifth International Conference on Network, Communication and Computing*, 98-102. https://doi.org/10.1145/3033288.3033319

Chapple, M. J., Chawla, N., & Striegel, A. (2007). Authentication anomaly detection: A case study on a virtual private network. *Proceedings of the 3rd annual ACM workshop on Mining network data*, 17-22. https://doi.org/10.1145/1269880.1269886

DATTA. (2019, mayo). *DATTA BUSSINESS INNOVATION 322*. http://revista.datta.com.ec/publication/8f287c4a/

Endler, D. (1998). Intrusion detection. Applying machine learning to Solaris audit data. *Proceedings 14th Annual Computer Security Applications Conference (Cat. No.98EX217)*, 268-279. https://doi.org/10.1109/CSAC.1998.738647

Eric-Joel Blanco-Hermida Sanz. (2016). *Algoritmos de clustering y aprendizaje automático aplicados a Twitter*. https://upcommons.upc.edu/bitstream/handle/2117/82434/113257.pdf

*Free updated GeoIP legacy databases*. (2021, enero 13). https://mailfud.org/geoip-legacy/

Gliese710. (2019, julio 15). *Ejemplo de clustering con k-means en Python – Exponentis*. http://exponentis.es/ejemplo-de-clustering-con-k-means-en-python

Guido, S. E. (2012). LAS PISTAS DE AUDITORÍA. *Ciencias Económicas*, *1*, 16.

Hamid, Y., Sugumaran, M., & Journaux, L. (2016). *Machine Learning Techniques for Intrusion Detection: A Comparative Analysis*. 6.

Jackson, K. M., Hruska, J., & Parker, D. B. (1992). *Computer security reference book*. CRC. https://books.google.com.ec/books?id=mSYPAQAAMAAJ

Jiawei Han, Micheline Kamber, & Jian Pei. (2012). *The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf*.

Jordi Gironés Roig. (s. f.). *Metodologías y estándares*. Universitat Oberta de Catalunya.

Juan Ignacio Bagnato. (2018, abril 13). *Crea un Arbol de Decisión en Python | Aprende Machine Learning*. https://www.aprendemachinelearning.com/arbol-de-decision-en-python-clasificacion-y-prediccion/

Juan Ignacio Bagnato. (2019, mayo 16). *Clasificación con datos desbalanceados | Aprende Machine Learning*. https://www.aprendemachinelearning.com/clasificacion-con-datos-desbalanceados/

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. *2008 Eighth IEEE International Conference on Data Mining*, 413-422. https://doi.org/10.1109/ICDM.2008.17

Nguyen, G., Dlugolinsky, S., Bobák, M., Tran, V., López García, Á., Heredia, I., Malík, P., & Hluchý, L. (2019). Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: A survey. *Artificial Intelligence Review*, *52*(1), 77-124. https://doi.org/10.1007/s10462-018-09679-z

Palacio-Niño, J.-O., & Berzal, F. (2019). Evaluation Metrics for Unsupervised Learning Algorithms. *arXiv:1905.05667 [cs, stat]*. http://arxiv.org/abs/1905.05667

Pius Owoh, N., Mahinderjit Singh, M., & Zaaba, Z. (2018). Automatic Annotation of Unlabeled Data from Smartphone-Based Motion and Location Sensors. *Sensors*, *18*(7), 2134. https://doi.org/10.3390/s18072134

Priyanshu Jain. (2020, enero 27). *Unsupervised Machine Learning: Validation Techniques*. Guavus - Go Decisively. https://www.guavus.com/technical-blog/unsupervised-machine-learning-validation-techniques/

Robert K. Yin. (2009). *Case Study Research Design and Methods* (Cuarta Edisión).

Shipra Saxena. (2020, agosto 13). 8 Categorical Data Encoding Techniques to Boost your Model in Python! *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/

Torres, L. (2020, diciembre 20). Curva ROC y curva prec-recall: Evalúa tu clasificación. *The Machine Learners*. https://themachinelearners.com/curva-roc-vs-prec-recall/

vpnMentor. (2019, septiembre 11). *Report: Ecuadorian Breach Reveals Sensitive Personal Data*. VpnMentor. https://www.vpnmentor.com/blog/report-ecuador-leak/

Zaitsev, I. (2019, marzo 29). *The Best Format to Save Pandas Data*. Medium. https://towardsdatascience.com/the-best-format-to-save-pandas-data-414dca023e0d