PROYECTO DE INVESTIGACIÓN PARA LA FISCALIZACIÓN AMBIENTAL

Teledetección de rotura de diques de relave en minería utilizando imágenes satelitales y Machine Learning

INTEGRANTES:

APELLIDOS Y NOMBRES	ESPECIALIDAD Y CARRERA
Bocanegra Muñante, Fiorella Alessandra	Ing. Geográfica
Ortiz Guzman, Siwar Alvaro	Economía
Suni Idme, Maria Gabriela	Derecho
Vasquez Quispe, Carolay Zully	Economía
Vizcarra Guerreros Edwin Andrei	Ing. de Gestión Minera

NOMBRE DEL ASESOR: Frank Loroña Calderón

Tabla de contenido

1. Justificación de la investigación.

A nivel mundial, la adopción y adecuación de tecnologías informáticas para la identificación de impactos ambientales generados por actividades antrópicas crece inexorablemente, como señalan Shi *et al* (2019) en el contexto de China, que tiene extensas áreas contaminadas por extracción de tierras raras (*rare earths*). Esta tendencia ha sido consolidada con el advenimiento de la pandemia por Covid-19, pues las administraciones ambientales necesitan, más que nunca, utilizar técnicas de detección remota para mantener la eficacia de las supervisiones.

Ante ello, diversos países latinoamericanos, cada uno con sus competencias y normativas específicas a cargo de las instituciones responsables, han optado por la incorporación de una serie de herramientas y aplicaciones tecnológicas en las acciones de evaluación y seguimiento de las actividades bajo sus competencias, brindando un soporte novedoso, eficaz y eficiente para la continuidad de sus funciones y el cuidado del medio ambiente durante el actual contexto, dice Camacho-Hurtado (2019) sobre el contexto colombiano. Así, Colombia ha empezado a aplicar tecnologías geoespaciales en el área de evaluación y monitoreo ambiental (ANLA, 2020).

En el contexto nacional, el Organismo de Evaluación y Fiscalización Ambiental – OEFA publicó el "Reglamento de Acciones de Fiscalización Ambiental y seguimiento y verificación a Entidades de Fiscalización Ambiental del Organismo de Evaluación y Fiscalización Ambiental – OEFA durante el estado de Emergencia Sanitaria decretado en el país ante el brote del COVID-19", en el cual se establece que "*la función de supervisión a EFAs y administrados se ejerce de forma remota. Excepcionalmente, puede desarrollar acciones de supervisión in situ que tengan por finalidad verificar el cumplimiento de las funciones de fiscalización ambiental a cargo de las EFA, respecto al desarrollo de actividades esenciales*" (Resolución del Consejo Directivo N° 00008-2020-OEFA/CD).

En este sentido, el presente trabajo de investigación pretende generar información sobre los pasos a seguir y las opciones que existen para implementar un algoritmo de "aprendizaje supervisado" (*Stochastic Gradient Descent, Random Forests, etc.*) de Machine Learning capaz de clasificar imágenes satelitales de diques de relaves de acuerdo a la probabilidad de rotura. Esto se hará con la finalidad de reducir costos y tiempo en los procesos de evaluación o supervisión, específicamente en casos en los que el administrado no reporta la contingencia mencionada.

Así, el empleo de esta herramienta permitirá obtener indicios sobre el incumplimiento de una de las obligaciones de los administrados del OEFA: alertar en un plazo no mayor a 24 horas el acaecimiento de una contingencia como la rotura de un dique. De esta forma, contribuiría al proceso de evaluación y supervisión del OEFA.

El estudio generará información sobre las normas legales peruanas e internacionales aplicables a la teledetección y los módulos de *python* y pasos a seguir para que el OEFA pueda implementar un algoritmo de *Machine Learning* en la detección de rotura de presas de relave usando imágenes satelitales. Asimismo, se pretende dar cuenta del grado de precisión que tiene uno de estos algoritmos en la detección de la contingencia

señalada y describir la posible reducción de costos monetarios de la implementación del mismo.

2. Antecedentes y conceptos básicos

2.1. Marco Conceptual

Machine Learning: Se dice que un programa de computadora aprende de la experiencia E con respecto a alguna tarea T y alguna medida de desempeño P, si su desempeño en T, medido por P, mejora con la experiencia E (Mitchell, 1997).

Teledetección: Es una técnica que es utilizada para adquirir datos de la superficie terrestre sin que exista contacto material mediante sensores que permiten el escaneo o grabación a tiempo real como satélites o radares (Cede, 2011).

Imágenes satelitales: Fotografía que contiene datos geográficos de un territorio específico con parámetros y características específicas, tomada por un satélite artificial (Cede, 2011).

Diques: Obra de ingeniería el cual en su estructura contiene un muro para contener la fuerza del agua, embalsarla o reconducirla (Ledesma, 2018).

Relaves Minero: Son residuos sólidos y líquidos originados de la actividad extractiva minera, que en su mayor parte contienen subproductos de procesos mineros en la etapa de tratamiento del mineral (Ledesma, 2018).

2.2. Antecedentes

Nacional

Salas, Barboza y Oliva (2016) en su investigación lograron cuantificar la tasa de deforestación entre los años 1987 a 2000 y 2000 a 2013, en el distrito de Florida, departamento de Amazonas, mediante imágenes obtenidas del satélite Landsat 5 y 7. En base a los resultados obtenidos por el empleo de la teledetección, concluyeron que las coberturas de bosque se han alterado significativamente durante los dos periodos, detectandose una mayor deforestación. Se han identificado y zonificado siete clases de coberturas, siendo estas bosques, pastos y cultivos, arbustales y herbazales, zona pantanosa, cuerpos de agua, zona urbana, áreas quemadas, los cuales permitieron realizar el análisis de detección de cambios. La investigación es de gran importancia para hacer la identificación de las roturas de los diques de relave minero mediante las imágenes satelitales.

Ramirez y Villena (2017) elaboraron mapas temáticos que dieron como resultado la identificación de zonas donde existía mayor presencia de pasivos ambientales mineros y de esta forma lograron priorizar la remediación ambiental. Recomendaron el uso de estas técnicas, como herramientas previas de análisis y localización de pasivos ambientales que ayudarán a perfeccionar tiempos y costos tanto en la identificación como en la priorización de Pasivos Ambientales Mineros. El método descrito por

Ramirez y Villena (2017) servirá como guía para el método a desarrollar en la presente investigación.

Giraldo y Vasquez (2019) indican que haciendo uso de imágenes del satélite Landsat en el estudio de evaluación y medición de la expansión territorial de la minería informal generada por la explotación de oro de la minería informal en la cuenca alta del río Ramis, Puno, en el período de 1984 a 2015. Como resultado de la investigación se obtuvo que el uso de tecnologías de teledetección serían una herramienta eficiente y aplicable para evaluar la problemática existente de minería informal o ilegal en las actividades que implican algún tipo de modificación en el paisaje como el cambio de uso de suelos. El estudio de Ramirez y Villena (2017) y Giraldo y Vasquez (2019) han servido para constatar que si bien es cierto se han realizado estudios sobre teledetección en Perú, la mayoría aplica métodos de sistemas información geográfica (SIG) u otros utilizan imágenes satelitales pero no *Machine Learning*, por lo cual el presente estudio estaría a la vanguardia en el estado de conocimiento peruano.

Latinoamérica

Según Rojas (2019) la utilización de un procedimiento de clasificación supervisada mediante imágenes ópticas multiespectrales posibilita la categorización de patrones en las imágenes. Esta aplicación está diseñada para el uso de tierras la cual ofrece una herramienta para la rápida toma de decisiones y permite la probabilidad de escoger y modificar los límites que se tienen que considerar a lo largo del proceso de categorización. Esta aplicación es de bajo costo, ya que realizar una clasificación supervisada de imágenes multiespectrales no presenta gastos significativos comparado a otras técnicas utilizadas en esta rama de teledetección. Además, esta técnica permite la visualización y selección de muestras de las diversas coberturas del uso de tierras obteniendo una validación inmediata. En esencia, esta investigación aporta el uso de un método de clasificación para un área de estudio, mediante imágenes satelitales.

La investigación de Hanna (2017), estuvo centrada en el uso de tecnologías hiperespectrales, haciendo uso de discriminación espectral y Machine Learning, como método cuantitativo de la caracterización mineral de yacimientos para aplicaciones metalúrgicas en la minera Florida, Chile. En el estudio se obtuvo que la metodología propuesta de avance hacia una elaboración de una herramienta de detección de tiempo real para el monitoreo del mineral ingresado a las plantas de procesamiento mineral. La investigación aporta una propuesta metodológica inicial que relaciona el uso de imágenes hiperespectrales con técnicas de aprendizaje supervisado, además de las posibles limitaciones que se presenten al implementar la metodología, entre las más importantes se encuentra el control de las condiciones ambientales (nubosidad) durante la captura de imágenes hiperespectrales.

Jaramillo y Antunes (2018) hacen referencia al uso de tecnologías y métodos asociados al *Remote Sensing* en la detección de variaciones en la cobertura vegetal de áreas de gran extensión, poniendo como ejemplos a imágenes satelitales de alta resolución

espectral y espacial, y su importancia en proyectos orientados a la clasificación de cobertura vegetal y los cambios respecto al tiempo. La investigación está orientada a validar una metodología que permita identificar las variaciones en la cobertura vegetal, obteniendo que el uso de Redes Neuronales Artificiales (RNAs) son el mecanismo óptimo para ejecutar los procesos, esta investigación nos ayuda a identificar variaciones con patrones en imágenes satelitales, en menor tiempo.

Internacional

Shi et al. (2018) realizan el estudio de técnicas de teledetección y proximidad para el mapeo de la contaminación de suelos con metales pesados, considerando este tipo de contaminación es un problema ambiental severo a nivel mundial, y su mapeo es vital para que los administradores ambientales y los formuladores de políticas determinen sus distribuciones y puntos críticos. El artículo revisa la utilidad de la espectroscopía de detección remota y proximal múltiple para un método conveniente y económico de obtener espectroscopía de reflectancia del suelo. Concluye además, que los sistemas de imágenes basados en vehículos aéreos no tripulados y satelitales facilitará el desarrollo de una tecnología de observación hiperespectral de incorporación aeronáutica espacial con la capacidad para monitorear el entorno del suelo de manera rápida y precisa a gran escala. Esto permite a la presente investigación hacer uso de estas observaciones hiperespectral en las áreas mineras y poder monitorear las roturas de los diques.

Boehmke y Greenwell (2019) señalan que los algoritmos de Machine Learning generalmente buscan la representación óptima de datos utilizando una señal de retroalimentación en forma de función. Sin embargo, la mayoría de los algoritmos de Machine Learning solo tienen la capacidad de usar una o dos capas de transformación de datos para aprender la representación de salida. llaman a estos modelos superficiales ya que solo utilizan 1 o 2 representaciones del espacio de variables explicativas. El libro de estos autores aporta al entendimiento del funcionamiento y caracterización de los algoritmos de *Machine Learning* y la evidencia aportada se utilizará para el acápite 6 sobre el método de procesamiento de imágenes satelitales

De acuerdo con la investigación realizada por Maxwell et al. (2018), los algoritmos de Machine Learning ofrecen efectividad y eficiencia en la clasificación de imágenes de teledetección, es por ello la utilización de estas en el presente estudio. Se concluye que el Machine Learning sigue siendo un área activa de investigación en la teledetección y resalta la probabilidad que los nuevos desarrollos y algoritmos proporcionan una mejor funcionalidad para la clasificación a partir de datos de la detección remota; haciendo referencia al Deep Learning y las redes neuronales profundas, los cuales han demostrado ser muy prometedores en la detección remota y las implementaciones se han puesto a disposición en paquetes de software como MatLab, R y Python. La investigación nos ofrece las fortalezas del uso del Machine Learning en los estudios relacionados a la teledetección, y cómo el Deep Learning ofrece un área de investigación interesante para mejorar la capacidad de extracción de información a partir de datos de detección remota.

Las dos principales variables a estudiar son la precisión y los costos de implementación de un algoritmo de *Machine Learning* para la detección de rotura de diques de relave. La precisión es una medida de bondad de ajuste del algoritmo expresada en la probabilidad de identificar correctamente una imagen satelital que presenta rotura de dique (Géron, 2019). Los costos de implementación serán estimados de acuerdo al informe "Prevención del Riesgo Ambiental por Falla de Depósitos de Relaves mediante Teledetección y Sistemas de Información Geográfica" de la DSEM-CMIN (OEFA, 2020).

3. Preguntas, objetivos e hipótesis de la investigación

Formular la pregunta general y las preguntas específicas de investigación, con sus objetivos e hipótesis.

PREGUNTA GENERAL	OBJETIVO GENERAL	HIPÓTESIS GENERAL
¿Qué algoritmo de machine Learning permite identificar patrones en imágenes satelitales con menor costo y mayor precisión?	Seleccionar un algoritmo de Machine Learning que sea capaz de identificar patrones en imágenes satelitales con menor costo y mayor precisión	La precisión del algoritmo seleccionado, para la detección de rotura de diques, es mayor al 60% y genera un costo menor al actual

PREGUNTAS ESPECÍFICAS	OBJETIVOS ESPECÍFICOS	HIPÓTESIS ESPECÍFICAS
¿Cuál es el marco normativo del manejo de la teledetección en el Perú?	Describir el marco normativo de la actividad aeroespacial y la teledetección en el Perú	Se describe el marco normativo de la actividad aeroespacial y la teledetección existente en el Perú
¿Cuál es el marco normativo del manejo de la teledetección en el extranjero?	Describir el marco normativo de la actividad aeroespacial y la teledetección en el extranjero.	Se describe el marco normativo de la actividad aeroespacial y la teledetección existente en el extranjero

¿Cuáles son los algoritmos de *Machine Learning* que sirven para detectar rotura de diques de relave con pocas imágenes satelitales?	Identificar los algoritmos de *Machine Learning* que sirven para detectar rotura de diques de relave con pocas imágenes satelitales	Se identifica por lo menos un algoritmo de *Machine Learning* que sirve para detectar rotura de diques de relave con pocas imágenes satelitales
¿Cuál algoritmo de *Machine Learning* es más idóneo para detectar rotura de diques de relave con pocas imágenes satelitales?	Seleccionar y describir el algoritmo más idóneo para la detección de rotura de diques con pocas imágenes satelitales	Se selecciona y describe el algoritmo seleccionado como más idónea para la detección de rotura de diques con pocas imágenes satelitales
¿Qué costos se ahorraría en la fiscalización ambiental con el uso de la tecnología *Machine Learning*?	Describir los costos que se ahorrarían en la fiscalización ambiental con el uso de la tecnología *Machine Learning*	Se describen los costos que se ahorrarían en la fiscalización ambiental con el uso de la tecnología *Machine Learning*

4. Fuentes de información

Las imágenes satelitales del área de estudio fueron descargadas de la plataforma Google Earth Engine, la cual es una plataforma en la nube para realizar análisis científicos y visualización de datos geoespaciales. Esta plataforma trabaja con códigos de JavaScript.

Base de datos	**Información a extraer**
Google Earth Engine - Sentinel 2 MSI	Se han extraído 64 imágenes de Huancapeti para los años 2016 - 2021

Fuentes Bibliográficas	**Información a extraer**
(AMLA, 2020)	Identificar problema y justificar el mismo
Prevención del Riesgo Ambiental por Falla de Depósitos de Relaves mediante Teledetección y Sistemas de Información Geográfica (OEFA, 2020) y Autoridad Nacional de Licencias Ambientales. La experiencia de la ANLA en la	Justificar el problema por contextos internacionales, nacionales y locales

teledetección aplicada a la supervisión ambiental (ANLA, 2020)	
Teledetección de pasivos ambientales de origen químico utilizando imágenes satelitales LANDSAT 8 en la provincia de Hualgayoc - 2017 (Ramirez y Villena, 2019), *Evaluación y medición de la expansión territorial de la minería informal en la cuenca alta del Ramis, Puno, Perú, usando imágenes satelitales* (Giraldo y Vasquez, 2017), *Dinámica multitemporal de índices de deforestación en el distrito de Florida, departamento de Amazonas, Perú* (Salas, Barboza y Oliva, 2016)	Se extrae información sobre los métodos utilizados por los autores para la teledetección de impactos ambientales para los antecedentes nacionales.
Desarrollo y propuesta metodológica para el empleo de los Campos Aleatorios de Markov aplicados a técnicas de clasificación de coberturas en imágenes de la superficie terrestre (Rojas, 2019), *Evaluación de tecnologías hiperespectrales en la caracterización mineral de yacimientos para aplicaciones geometalúrgicas: Caso aplicado a mina Florida, distrito minero Alhué, Región Metropolitana, Chile* (Hanna, 2017) y *Change detection in vegetation cover through interpretation of Landsat images by artificial neural networks (ANN). Case study: Ecuadorian Amazon Region* (Jaramillo y Antunes, 2018)	Se utiliza la aplicación de teledetección aplicando Machine learning tanto en usos de tierra, cobertura vegetal y procesos mineros para plantear el procedimiento que se describe en este estudio.
Proximal and remote sensing techniques for mapping of soil contamination with heavy metals. Applied Spectroscopy Reviews (Shi et. al. , 2018), *Deep learning. En Hands-on machine learning with R* (Boehmke y Greenwell, 2019) y *Implementation of machine-learning classification in remote sensing: an applied review* (Maxwell et al. , 2018)	Antecedentes Internacionales
Metodología de la Investigación (Behar, 2008)	Sirvió para fundamentar la metodología de investigación

NORMAS LEGALES NACIONALES	¿QUÉ SE ESPERA OBTENER?
Decreto Ley N° 20643 Crea la Comisión Nacional de Investigación y Desarrollo Aeroespacial del Sector Aeronáutica CONIDA	Describir las funciones de CONIDA, ente rector de las actividades espaciales en el Perú, la cual suministra imágenes satelitales a personas naturales y jurídicas con fines de investigación.
Ley N° 28799 Ley que declara de interés nacional la creación, implementación y desarrollo de un "Centro Nacional de Operación de Imágenes Satelitales"	Describir las funciones del Centro Nacional de Operación de Imágenes satelitales, el cual monitorea el satélite Perú SAT-1 y suministra imágenes satelitales del Satélite Perú SAT-1 a dependencias del sector público y privado.
Ley N°28611 Ley General del Ambiente	Describir la normativa que asegura el ejercicio de un ambiente saludable, así como la Política Nacional del Ambiente y la Gestión Ambiental
Ley N° 29325 Ley del Sistema Nacional de Evaluación y Fiscalización Ambiental	Describir el Sistema Nacional de Evaluación y Fiscalización Ambiental el cual asegura el cumplimiento de la legislación ambiental vigente en el país.
Decreto Legislativo N° 109 Ley General de Minería	Describir las actividades enmarcadas en el sector minero.
Resolución Directoral Nº 2007-2019-OEFA-DFAI (Caso Minera Bateas S.A.C.)	Describir en el caso en concreto la utilización de imágenes satelitales de Google Earth por parte de la OEFA para determinar la responsabilidad del administrado.

Resolución N° 008-2020-OEFA/CD aprueba el Reglamento de acciones de fiscalización ambiental y seguimiento y verificación a Entidades de Fiscalización Ambiental del OEFA durante el Estado de Emergencia Sanitaria decretado en el páis ante el brote del COVID-19	Describir los criterios de fiscalización ambiental por parte de la OEFA en el contexto de la pandemia COVID -19.

NORMAS LEGALES INTERNACIONALES	**¿QUÉ SE ESPERA OBTENER?**
Condiciones de Uso de Google Earth Engine	Describir el uso del software, la política de privacidad, restricciones y cumplimiento de leyes y políticas de Google Earth sobre el uso de las imágenes satelitales obtenidas del satélite Sentinel 2 MSI
Reglamento UE N° 377/2014 del Parlamento Europeo y del Consejo.	Describir la finalidad del Programa Copernicus el cual proporciona servicios operativos en materia ambiental suministrando imágenes satelitales de Sentinel 2- MSI- L1
Ley N° 685 Código de Minas de Colombia	Describir y analizar mediante el derecho comparado la normativa vigente respecto a la explotación y aprovechamiento de recursos mineros
Ley N° 18248 Código de Minería de Chile	Describir y analizar mediante el derecho comparado la normativa vigente respecto a la administración y regulación de la actividad minera.
Ley N° 45 Ley de Minería de Ecuador	Describir y analizar mediante el derecho comparado la normativa vigente respecto a la actividad del sector minero.

5. Metodología de la investigación

El propósito de este estudio es generar información sobre los módulos de *python* y pasos a seguir para que el OEFA pueda implementar un algoritmo de *Machine Learning* en la detección de rotura de presas de relave usando imágenes satelitales. A diferencia del proyecto DAMSAT, que también utiliza algoritmos e imágenes satelitales para la teledetección, este proyecto de investigación tiene por finalidad que los funcionarios del OEFA tengan acceso al código fuente del algoritmo de detección. Asimismo, se pretende dar cuenta del grado de precisión que tiene uno de estos algoritmos en la detección de la contingencia señalada y describir la posible reducción de costos monetarios de la implementación del mismo.

En este sentido, el proyecto de investigación se enmarca dentro del **paradigma cuantitativo** de investigación, pues se pretende medir el nivel de precisión del algoritmo a implementar; En palabras de Hernandez-Sampieri, una investigación es de enfoque cuantitativo cuando se pretende medir la magnitud de una variable: la precisión en el presente estudio (2014, p. 5). El **alcance será descriptivo** debido a que el propósito del trabajo es describir cómo implementar uno de los algoritmos existentes para identificar patrones en imágenes y detectar rotura de diques; "una investigación es exploratoria cuando el objetivo es especificar procesos, objetos, personas, etc." (Hernandez-Sampieri, 2014, p. 92).

El **diseño será no experimental-longitudinal** pues el algoritmo analizará una serie de imágenes del período 2016-2021 de la unidad minera Huancapeti para determinar precisión de detección, aunque no se determinará tendencia alguna; según Hernandez-Sampieri, los diseños longitudinales son aquellos que recaban datos de de diferentes instantes de tiempo (2014, p. 159). El **método será inductivo** porque se pretende generalizar la precisión de detección obtenida del conjunto de datos de entrenamiento con una imagen de generalización externa a este conjunto; acorde a Behar (2008), el método inductivo consiste en elevar los postulados u observaciones de resultados puntuales a estatus de leyes generales o teorías.

Finalmente, el **principal criterio** para seleccionar el algoritmo a implementar es capacidad capacidad de ajuste a los datos con pocas observaciones (Hernandez-Sampieri, 2014). Este criterio es decisivo porque las observaciones de la matriz de regresores son 64 imágenes satelitales de 69x49 píxeles de dimensión, las cuales se descomponen en una base de datos cuyas filas representan una imágen y sus 69x49=3381 columnas, variables explicativas, corresponden a cada píxel de las imágenes. Esto implica que la matriz de datos posee más columnas que filas, lo que imposibilita resolver el sistema de ecuaciones subyacente, a la matriz, para algoritmos paramétricos (Géron, 2019). Por esta razón, algoritmos como *SGD* o *CNN*, que se mencionan en el siguiente acápite, fueron descartados y se eligió *Random Forest*, pues no es paramétrico (Géron, 2019).

6. Método de detección de imágenes y análisis económico

El método empleado estuvo basada en la aplicación de la Teledetección utilizando imágenes satelitales en la plataforma de Google Earth Engine, mediante códigos con el lenguaje JavaScript se pudo descargar imágenes Sentinel 2 del área exacta y el tamaño que uno requiera, estas imágenes tienen una resolución de 10 metros, también se pudo

añadir otras características como en porcentaje de nubes para tener imágenes más limpias de nubosidad y aprovechar mejor estas imágenes.

Código para la descarga de imágenes Sentinel 2 MSI:

```
var batch = require('users/fitoprincipe/geetools:batch')
var collection= ee.ImageCollection ('COPERNICUS/S2')
  .filterDate ('2015-07-01', '2021-02-04')

//fechas disponibles ('2015-07-01' - actualidad)

  .filterBounds (geometry)
  .filterMetadata ('CLOUDY_PIXEL_PERCENTAGE', 'Less_Than', 30);
var SentinelFiltro = ee.Image(collection.median());
var SentinelClip = SentinelFiltro.clip (geometry);
Map.addLayer (SentinelClip, {
  max: 4000,
  min: 0.0,
  gamma: 1.0,
  bands: ['B4','B3','B2']},
  'Imagen Sentinel 2');
print (SentinelFiltro);
Map.centerObject (SentinelClip);
var options={scale:10,
            name:'{system_date}',
            region: geometry};

batch.Download.ImageCollection.toDrive(collection,'HUANCAPETI',
options);
```

Posteriormente, estas imágenes pasaron por un proceso ejecutado con el lenguaje de programación R para extraer los valores numéricos de los píxeles (todo píxel tiene un valor numérico) y alinearlos todos en una base de datos formato *csv*. Se utilizó R por la facilidad que se tiene para manipular datos con este lenguaje. El código empleado es el siguiente.

```
library(raster)
library(rgdal)
# images
filenames<-
list.files("~/Documentos/python/machine_learning_oefa/datasets",
pattern = ".tif", full.names = TRUE)
lf <- lapply(filenames, raster)
lv <- lapply(lf, as.vector)
length(lv)
y <- rep(0, length(lv))
d <- data.frame(y)
colnames(d)[1] <- "Response"
for (i in seq(along = lv)) {
    if (i==1) {
        for (j in seq(along = lv[[i]])) {
            d[i, 1 + j] <- lv[[i]][j]
             colnames(d)[1 + j] <- paste("pixel", as.character(j), sep
= "")
        }
    } else {
```

```
        for (e in seq(along = lv[[i]])) {
            d[i, 1 + e] <- lv[[i]][e]
        }
    }
 }
d[is.na(d)] <- 0
write.csv(d,file=
"~/Documentos/python/machine_learning_oefa/datasets/images.csv")
```

Breve descripción de los algoritmos utilizados

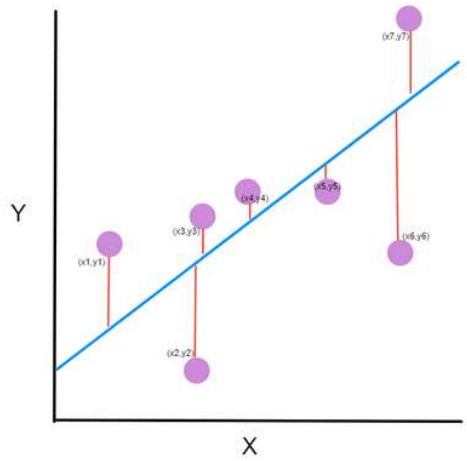
Los dos algoritmos elegidos, Stochastic Gradient Descent y Random Forests, para realizar la identificación de rotura de presas de relaves, usando imágenes satelitales, se clasifican dentro del grupo de algoritmos de aprendizaje supervisado. Se los llama así porque el conjunto de datos de entrenamiento que los alimenta incluye las soluciones deseadas, llamadas etiquetas (Mitchell, 1997). Asimismo, el modo de aprendizaje elegido fue aquel denominado "por lotes" que implica que el sistema es incapaz de aprender de forma incremental: debe entrenarse utilizando todos los datos disponibles, lo que generalmente se realiza sin conexión a Internet. Primero se entrena el sistema y luego se ejecuta sin más aprendizaje (Mitchell, 1997); simplemente aplica lo que ha aprendido.

Los algoritmos de aprendizaje supervisado existentes que pueden clasificar imágenes son *Convolutional Neural-Networks, Stochastic Gradient Descent, Decision Trees* y *Random Forest*, de los cuales se describirán dos a continuación.

Stochastic Gradient Descent

La lógica detrás del algoritmo *stochastic gradient descent* (SGD) es ajustar una línea recta en medio de las observaciones contenidas en la base de datos, de tal forma que se minimice la distancia entre los valores observados (los puntos morados en la Figura 1) y la línea recta que está ajustando el algoritmo (línea azul en la Figura 2). La distancia entre un punto morado (valor observado) y la línea recta se denomina error. La suma de todos los errores elevados al cuadrado (SEC) es aquello que el algoritmo busca minimizar. Para lograrlo, el algoritmo ajusta iterativamente la pendiente y el intercepto de la línea recta hasta encontrar el valor mínimo de la SEC. La pendiente y el intercepto estimados por el algoritmo son llamados parámetros. Una vez encontrado el valor mínimo, el algoritmo se detiene y puede ser usado para para predecir valores de *y* (la variable explicada, vea la Figura 1) en función a valores de la variable o las variables *x* (variable explicativa, vea la Figura 1).

Figura 1: Suma de los errores al cuadrado



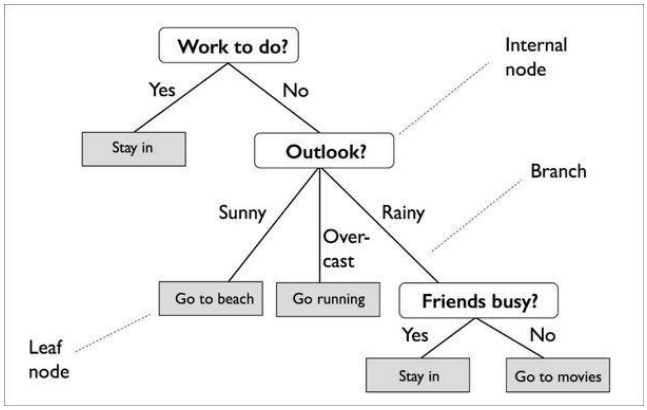
En problemas de clasificación, como el que intenta resolver este estudio, la variable explicada *y* suele tomar solo dos valores, "0" cuando no hay rotura de dique, o "1" cuando sí hay. Así, el modelo es capaz de predecir la probabilidad de que la variable explicada (también llamada variable respuesta) tome el valor de 1.

Random Forests

El algoritmo denominado bosques aleatorios (Random Forests) ajusta varios árboles de decisión, no correlacionados entre sí, a la base de datos que se le proporciona y luego compara las predicciones que realiza cada árbol para derivar predicciones más certeras (Boehmke and Greenwell, 2019). Los árboles de decisión (Decision Trees) son otro tipo de algoritmo.

Respecto a los árboles de decisión, la lógica detrás de este algoritmo es dividir las observaciones de la base de datos de forma recursiva hasta parar en un valor de la variable respuesta (hoja del árbol). Por ejemplo, en la Figura 2 se aprecia un árbol de decisión que empieza con un nodo raíz que contiene la pregunta ¿Hay trabajo que hacer? (*work to do?*). Este nodo raíz es lo que en el acápite anterior se denominó una variable explicativa. Esta variable explicativa sólo toma dos posibles valores "sí" o "no". Si la respuesta es "sí" llegamos a una de las hojas del árbol, es decir, a uno de los valores de nuestra variable respuesta: "Quedarse en casa" (*stay in*). Si la respuesta al nodo raíz fue "no" pasamos a un nodo intermedio que pregunta sobre el pronóstico del tiempo (*outlook?*). Si la respuesta a este nodo es "soleado" (*sunny*), llegamos a otro valor de la variable respuesta (hoja del árbol) que dice "Ir a la playa". En cambio, si la respuesta es "nublado", llegamos a otra hoja del árbol que dice "Ir a correr" (*go running*). Y de esta forma las observaciones de la base de datos pueden dividirse en varios subgrupos sucesivos hasta llegar a las últimas hojas del árbol de decisión.

Figura 2: Árbol de decisión (Decision tree)



Una vez terminada la clasificación del árbol de decisión, la probabilidad de ocurrencia de un valor de la variable respuesta se calcula respecto al total de hojas del árbol de decisión. Por ejemplo, en la Figura 2 hay cinco hojas y el valor "Quedarse en casa" se repite dos veces. La probabilidad de "Quedarse en casa" es entonces ⅖. lo que es igual a 40%. A diferencia del anterior algoritmo, éste no estima parámetros, lo que lo hace más idóneo para bases de datos con pocas observaciones, en el caso de este estudio con pocas imágenes.

Módulos utilizados e implementación del algoritmo

Para implementar estos algoritmos se utilizó el lenguaje de programación *python* y los módulos *Scikit-Learn*, *Matplotlib*, *Pandas* y *Numpy*, de dicho lenguaje. El primero es un módulo que contiene una colección de algoritmos de *Machine Learning* listos para implementar; el segundo contiene una colección de funciones para crear gráficos estáticos, animados e interactivos; el tercero provee estructuras flexibles para trabajar con bases de datos; y el cuarto permite implementar operaciones de álgebra lineal con los datos. Como se mencionó en el acápite anterior, el algoritmo *Random Forests* es el más idóneo cuando se tienen pocas observaciones, por lo cual éste será el que se implemente en este estudio.

El primer paso para implementar el algoritmo es instalar la última versión disponible de *python* que se encuentra en https://www.python.org/downloads/release. Luego hay que hacer clic en el icono de Windows de la esquina inferior izquierda, comenzar a escribir *PowerShell* e ingresar a la aplicación. Una vez dentro de la terminal (*PowerShell*) hay que asegurarse de tener instalada la última versión de *pip*, que es el gestor de paquetes de *python*, ejecutando la siguiente línea de código.

```
$ python3 -m pip install --user -U pip
```

Ahora se deben instalar los módulos mencionados líneas arriba, ejecutando también la siguiente línea de código.

```
$ python3 -m pip install -U jupyter matplotlib numpy pandas scipy
scikit-learn
```

Finalmente, se debe ingresar al ambiente de desarrollo de *python* llamado *jupyter*, ejecutando la siguiente línea.

```
$ jupyter notebook
```

Al ejecutar este último comando se abrirá una nueva pestaña en el navegador web de la computadora en el cual se podrán apreciar las carpetas y archivos del directorio "C:\users\Usuario\Documents" o "C:\users\Usuario". En este punto, hay que dirigirse a la carpeta en la que se descargó el documento "*OEFA.ipynb*", que es remitido junto con el informe de esta investigación. Una vez encontrado el documentos mencionado hay que abrirlo haciendo doble click y se podrá visualizar todo el código escrito para implementar el algoritmo utilizado en este trabajo.

Ahora se presenta el código implementado, con comentarios, para calcular la precisión de la clasificación de imágenes satelitales de los administrados Huancapeti y Cobriza realizada por el algoritmo.

```
# Cargando los módulos

>>> import pandas as pd

>>> import numpy as np

>>> import matplotlib as mpl

>>> import matplotlib.pyplot as plt

# Cargando la base de datos

>>> d = pd.read_csv("datasets/images.csv")

# Separando la base de datos en variable respuesta y variables explicativas

>>> y = d["Response"]

>>> X = d.iloc[:,1:]

# Graficando una de las imágenes

>>> some_digit = X.iloc[0].values

>>> some_digit_image = some_digit.reshape(116, 208)

>>> plt.imshow(some_digit_image)

>>> plt.axis("off")

>>> plt.show()
```

Figura 3

```
# Implementar el algoritmo

>>> from sklearn.ensemble import RandomForestClassifier

>>> from sklearn.model_selection import cross_val_predict

>>> forest_clf = RandomForestClassifier(random_state=42)

>>> forest_clf.fit(X, y)

# Generar las predicciones del modelo

>>> y_pred = forest_clf.predict(X)

# Importar función para calcular precisión de predicción y cálculo de
la misma

>>> from sklearn.metrics import precision_score, recall_score

>>> precision_score(y, y_pred)

1.0
```

Se aprecia que el resultado de la función `precicion_score()` es uno, es dicir, el algoritmo sería capaz de predecir si en una imagen se observa una rotura de dique de relaves con una precisión de 100%. La precisión es una medida de efectividad y se obtiene del ratio:

$$\frac{VP}{VP + FP}$$

Donde *VP* son los verdaderos positivos: las imágenes que el algoritmo clasifica correctamente como que presentan rotura de dique. *FP* son lo falsos positivos: las imágenes que el algoritmo clasifica incorrectamente como que presentan rotura de dique. Si bien es cierto que una alta precisión es deseable, en este caso se puede deber al hecho de que se le ha proporcionado muy pocas imágenes al algoritmo para poder entrenarse: tres. Esto implica que la alta tasa de precisión, en este caso, puede no estar reflejando la verdadera bondad de ajuste del algoritmo (Géron, 2019). Estudios similares han estimado una precisión entre 95% y 99%. Por esta razón, se decidió utilizar una imagen de generalización, esto es, comprobar que el algoritmo puede clasificar

correctamente imágenes fuera de la base de datos de entrenamiento. La siguiente línea de código realiza esta comprobación.

```
# imágen de generalización

>>> forest_clf.predict([some_image])

array([1], dtype=uint8)
```

El resultado señala que el algoritmo ha clasificado la imagen que se le ha proporcionado como que presenta rotura de dique, lo cual es correcto. Con esto se puede concluir que el algoritmo puede ser generalizable.

Respecto a las limitaciones, este trabajo no desarrolla una integración entre el código de implementación del algoritmo de *Machine Learning* y plataformas como *google earth engine* para poder automatizar la detección de rotura de diques de relave, es decir que la detección sea tiempo real y se envíen alertas si el *software* detectara una rotura. Esto no se realizó por la restricción de tiempo que se disponía para completar el estudio: tres semanas. No obstante, el principal componente para lograr dicha integración se encuentra en que el código del algoritmo implementado esté desarrollado y se haya probado su precisión, por lo cual el aporte de este estudio es importante para una siguiente investigación que culmine lo que esta no pudo.

Análisis económico de la implementación de un algoritmo de Machine Learning

De acuerdo al Informe de *Prevención del Riesgo Ambiental por Falla de Depósitos de Relaves mediante Teledetección y Sistemas de Información Geográfica* realizado por el OEFA (OEFA, 2020), se propone la identificación temprana de los peligros en los depósitos de relaves y el estudio de los materiales en éstos, así como el monitoreo remoto de espejos de agua.

Las principales oportunidades que ofrecerá el uso de un algoritmo de "aprendizaje supervisado" de Machine Learning capaz de clasificar imágenes satelitales de diques de relaves de acuerdo a la probabilidad de rotura, además de la libre disponibilidad de las imágenes satelitales del SENTINEL 2, serán la reducción del tiempo de detección de emergencias ambientales relacionadas a roturas de diques de relaves, así como la disminución de los costos estimados en dichas actividades.

Con el uso del algoritmo, en tan solo un (1) día, con el apoyo de un especialista en teledetección y un especialista ambiental, se podrán realizar los análisis y procesamiento de hasta cinco (5) imágenes satelitales SENTINEL 2 de diques de relaves que se encuentren en situación de riesgo alto o muy alto, ello contribuirá a optimizar el tiempo de detección de emergencias ambientales del rubro.

De igual manera, se analizaron los costos promedios de atención de emergencias ambientales por roturas de diques de relaves, los cuales ascienden a S/ 6,499.63 por intervenciones únicamente de la Coordinación de Supervisión Ambiental en Minería (CMIN) (OEFA, 2020). Mientras que la implementación de bases de datos Geoespaciales y modelamiento de Sistemas de Información Geográfica (SIG) reduciría los costos a S/ 4,143.81 (OEFA, 2020).

Se tendrá, por tanto, un beneficio económico promedio esperado de S/ 2,344.82 por cada intervención; además, se señala que cada año se realizan 25 intervenciones aproximadamente, de modo que, anualmente se esperaría ahorrar en promedio S/ 58,895.50.

En conclusión, la implementación de teledetección utilizando imágenes satelitales y algoritmos de *Machine Learning*, es considerada económicamente beneficiosa y generaría información crucial y actualizada ante las emergencias ambientales a causa de roturas de diques de relave en minería.

7. Bibliografía

ANLA (2020). Autoridad Nacional de Licencias Ambientales. Colombia. La experiencia de la ANLA en la teledetección aplicada a la supervisión ambiental. Noticias. Publicado el 23 de octubre de 2020. http://portal.anla.gov.co:93/noticias/experiencia-anla-teledeteccion-aplicada-supervision-ambiental

Behar, D. (2008). *Metodología de la Investigación*. Editorial Shalom.

Boehmke, B., & Greenwell, B. M. (2019). *Deep learning. En Hands-on machine learning with R*. CRC Press. https://bradleyboehmke.github.io/HOML/

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. 1st edn, vol. 1. https://www.scirp.org/(S(i43dyn45teexjx455qlt3d2q))/reference/ReferencesPapers.aspx?ReferenceID=6963

Camacho Hurtado, F. (2019). *Análisis multitemporal de las zonas mineras de los municipios de Lenguazaque y Guachetá para evidenciar impactos ambientales y cambios en la cobertura de suelos* [Tesis]. https://repositorio.uniandes.edu.co/handle/1992/45214

Cede, R., Cabrera, E., Vargas, D. M., Galindo, G., García, M. C., Ordoñez, M. F., & Alonso, F. (2011). Capítulo 10: Fundamentos de la teledetección. *Introducción a La Percepción Remota y Sus Aplicaciones Geologicas*, (4), 181–218. Retrieved from http://www.um.es/geograf/sigmur/%0Ahttp://www.ideam.gov.co/documents/13257/13817/Protocolo_para_la_cuantificación_Deforestación_Nivel_Nacional.pdf

Condiciones Legales para el uso del Software Google Maps/Google Earth y las API de Google Maps/Google Earth – Google. (2021). https://earth.google.com/intl/es/licensepro.html

Decreto Ley N° 20643. (1974) *Crea La Comisión Nacional de Investigación y Desarrollo Aeroespacial del Sector Aeronáutica CONIDA*.

https://www.peru.gob.pe/docs/PLANES/85/PLAN_85_2015_1.2.1.0._NORMA_DE_CREACION_DE_LA_ENTIIDAD_D.L._20643_11_JUN1974.PDF

Decreto Legislativo N° 109 (1991) *Ley General de Minería* https://www.peru.gob.pe/docs/PLANES/94/PLAN_94_DL%20N%C2%BA%20109_2008.pdf

Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media. https://www.knowledgeisle.com/wp-content/uploads/2019/12/2-Aur%C3%A9lien-G%C3%A9ron-Hands-On-Machine-Learning-with-Scikit-Learn-Keras-and-Tensorflow_-Concepts-Tools-and-Techniques-to-Build-Intelligent-Systems-O%E2%80%99Reilly-Media-2019.pdf

Giraldo, U., & Vasquez, P. (2019). *Evaluación y medición de la expansión territorial de la minería informal en la cuenca alta del Ramis, Puno, Perú, usando imágenes satelitales*. Espacio y Desarrollo N° 34, pp. 5-32 (ISSN 1016-9148) https://doi.org/10.18800/espacioydesarrollo.201902.001

Hanna, V. (2017). *Evaluación de tecnologías hiperespectrales en la caracterización mineral de yacimientos para aplicaciones geometalúrgicas: Caso aplicado a mina Florida, distrito minero Alhué, Región Metropolitana, Chile*. [Tesis para optar por el grado de Magíster en Ciencias, mención en Geología. Universidad de Chile]. http://repositorio.uchile.cl/handle/2250/149764

Jaramillo, L., & Antunes, A. (2018). *Change detection in vegetation cover through interpretation of Landsat images by artificial neural networks (ANN). Case study: Ecuadorian Amazon Region.* Revista de Teledetección, 51, 33-46. https://doi.org/10.4995/raet.2018.8995

Ledesma, W. (2018). *Propuesta de Tratamiento del Depósito de Relaves de Quiulacocha-Pasco para su Remediación Ambiental, Basa en Experiencias Exitosas en Empresas Mineras.* Tesis para optar el Grado Académico de Maestro. Universidad Nacional Daniel Alcides Carrión. http://repositorio.undac.edu.pe/bitstream/undac/878/1/TESIS%20MAESTRIA%20JLV%20-%202019.pdf

Ley N° 685 (2001) *Código de Minas de Colombia* https://www.minambiente.gov.co/images/normativa/leyes/2001/ley_0685_2001.pdf

Ley N° 18248 (1983) *Código de Minería de Chile* https://www.bcn.cl/leychile/navegar?idNorma=29668

Ley N° 28611 (2005) *Ley General del Ambiente* https://www.minam.gob.pe/wp-content/uploads/2017/04/Ley-N%C2%B0-28611.pdf

Ley N° 28799 (2006) *Ley que declara de interés nacional la creación, implementación y desarrollo de un Centro Nacional de Operación de imágenes satelitales*. https://docs.peru.justia.com/federales/leyes/28799-jul-19-2006.pdf

Ley N° 29325 (2009) *Ley del Sistema Nacional de Evaluación y Fiscalización Ambiental* https://www.oefa.gob.pe/?wpfb_dl=12165

Ley N° 45 (2009) *Ley de Minería de Ecuador https://www.oas.org/juridico/PDFs/mesicic4_ecu_mineria.pdf*

Maxwell, A., Warner, T., & Fang, F. (2018). *Implementation of machine-learning classification in remote sensing: an applied review.* International Journal of Remote Sensing, 39(9), 2784–2817. https://doi.org/10.1080/01431161.2018.1433343

Mitchell, T. M. (1997). *Does machine learning really work?.* AI magazine, 18(3):11–11. https://doi.org/10.1609/aimag.v18i3.1303

OEFA (2020). *Prevención del Riesgo Ambiental por Falla de Depósitos de Relaves mediante Teledetección y Sistemas de Información Geográfica*. PRASATG. DSEM-CMIN.

Organismo de Evaluación y Fiscalización Ambiental (2019) Resolución de Consejo Directivo N° 006-2019-OEFA/CD Reglamento de Supervisión. http://www.oefa.gob.pe/wp-content/uploads/2019/02/Resoluci%C3%B3n-N%C2%B0-006-2019-OEFA-CD.pdf

Ramirez E., & Villena, E. (2017) *Teledetección de pasivos ambientales de origen químico utilizando imágenes satelitales LANDSAT 8 en la provincia de Hualgayoc - 2017.* Tesis para optar el título profesional de Ingeniero de Minas. Universidad Privada del Norte. https://repositorio.upn.edu.pe/handle/11537/13249

Reglamento UE N° 377/2014 (2014) del Parlamento Europeo y del Consejo. https://eur-lex.europa.eu/legal-content/es/TXT/?uri=CELEX%3A32014R0377

Resolución del Consejo Directivo N° 00008-2020-OEFA/CD. *Reglamento de Acciones de Fiscalización Ambiental y seguimiento y verificación a Entidades de Fiscalización Ambiental del Organismo de Evaluación y Fiscalización Ambiental – OEFA durante el Estado de Emergencia Sanitaria decretado en el país ante el brote del COVID-19*. 5 de junio de 2020. https://busquedas.elperuano.pe/normaslegales/aprueban-el-reglamento-de-acciones-de-fiscalizacion-ambient-resolucion-no-00008-2020-oefacd-1867435-1/

Resolución Directoral Nº 2007-2019-OEFA-DFAI *(Caso Minera Bateas S.A.C.)* https://www.oefa.gob.pe/?wpfb_dl=38024

Rojas, S. (2019). *Desarrollo y propuesta metodológica para el empleo de los Campos Aleatorios de Markov aplicados a técnicas de clasificación de coberturas en imágenes de la superficie terrestre.* Tesis para optar al título de Magister en Ciencias de la

Información y las Comunicaciones. Universidad Distrital Francisco José de Caldas. https://repository.udistrital.edu.co/handle/11349/15645

Salas, R., Barboza, E. y Oliva, M., (2016) *Dinámica multitemporal de índices de deforestación en el distrito de Florida, departamento de Amazonas, Perú,* Indes, 2(1), 19-27. doi: 10.25127/indes.201401.002 http://revistas.untrm.edu.pe/index.php/INDES/article/view/59

Samuel, A. L. (1959). S*ome studies in machine learning using the game of checkers*. IBM Journal of research and development, 3(3):210–229. https://ieeexplore.ieee.org/abstract/document/5392560/

Shi, T., Guo, L., Chen, Y., Wang, W., Shi, Z., Li, Q., & Wu, G. (2018). *Proximal and remote sensing techniques for mapping of soil contamination with heavy metals*. Applied Spectroscopy Reviews, 53(10), 783–805. doi:10.1080/05704928.2018.1442346. https://www.tandfonline.com/doi/abs/10.1080/05704928.2018.1442346

Solicitar acceso al Portal de Imágenes Satelitales - COF. (2021). Retrieved February 12, 2021, from Www.gob.pe website: https://www.gob.pe/8392-solicitar-acceso-al-portal-de-imagenes-satelitales-cof