Institución Universitaria

Desarrollo metodológico para la obtención de modelos espaciales del subsuelo, de uso en prospección en geociencias, mediante métodos de machine learning con datos geológicos, geofísicos y geomecánicos

Franco Bertaiola Ríos

Instituto Tecnológico Metropolitano

Facultad de Ingenieros

Medellín, Colombia

2022

Desarrollo metodológico para la obtención de modelos espaciales del subsuelo, de uso en prospección en geociencias, mediante métodos de machine learning con datos geológicos, geofísicos y geomecánicos

Franco Bertaiola Ríos

Tesis presentada como requisito parcial para optar al título de:

Magister en Automatización y Control Industrial

Directores:

Ph. D. Andrés Mauricio Muñoz García

Ph. D. Moisés Oswaldo Bustamante Rúa

Grupos de investigación:

Grupo de Investigación Geofísica y Ciencias Computacionales (GGC3)– ITM

Instituto de Minerales CIMEX – UNAL

Instituto Tecnológico Metropolitano

Facultad de Ingenieros

Medellín, Colombia

2022

Dedicatoria

A mis familiares, amigos y colegas que me acompañaron, apoyaron y motivaron en este proceso de aprendizaje.

Agradecimientos

Quisiera agradecer inicialmente a mis asesores de tesis Andrés Mauricio Muñoz García, Moisés Oswaldo Bustamante Rúa y Luis Fernando Duque Gómez por los conocimientos transmitidos y las oportunidades brindadas.

También quiero agradecer a la empresa WGM por los datos suministrados sin los cuales esta investigación no sería posible.

Un agradecimiento especial a los miembros de grupo de investigación GGC3 y del Instituto de Minerales CIMEX por acompañarme en este proceso y brindarme su ayuda.

Resumen

La extracción de recursos minerales es una necesidad fundamental para el mundo. Sin su extracción, gran parte de los utensilios y herramientas a los cuales estamos acostumbrados en nuestra vida cotidiana dejarían de existir: Desde los simples utensilios de cocina, electrodomésticos, computadores, medios de transporte hasta el equipo hospitalario o los colisionadores de partículas requieren una basta cantidad de materias primas extraídas del subsuelo. Entre éstas se encuentra el oro, un metal bastante escaso pero muy codiciado por sus usos tanto industriales como domésticos.

En el departamento de Antioquia se concentra la mayor parte de la explotación de oro aluvial de Colombia, extrayéndose alrededor del 45 % del total del oro producido en el país, lo cual representa una importante fuente de ingresos para las empresas y familias dedicados a esta actividad. Sin embargo, la extracción de oro tiene muchos retos, entre ellos destaca la prospección y estimación de la cantidad de oro en una zona de interés. Esta tarea a nivel industrial requiere una inversión económica elevada en estudios del subsuelo, tales como perforaciones, técnicas de geofísica y geoquímica, entre otras, las cuales al final dan indicios de la viabilidad de un proyecto de extracción minera.

Debido al alto costo de estos estudios, a lo largo de los años se han desarrollado herramientas para extraer la mayor cantidad de información posible de los datos obtenidos. En el panorama del oro y otros elementos se suelen utilizar técnicas de interpolación espacial como Kriging, ponderación de distancia inversa (IDW, por sus siglas en inglés) e interpolación suave discreta, entre otras, con las cuales se obtienen modelos 2D y 3D del subsuelo, lo que permite establecer zonas de interés para la exploración y/o explotación minera. Con el aumento de la capacidad de cómputo y el refinamiento de técnicas de machine learning, se han explorado novedosas metodologías alrededor del mundo para apoyar la prospección mineral, entre ellas se encuentran: interpolación espacial mediante bosques aleatorios, máquinas de vectores de soporte, métodos de apilamiento, aprendizaje profundo, procesamiento de lenguaje natural, entre muchas otras.

Aunque existen estudios en los que se aplican técnicas de machine learning para apoyar la prospección minera en el mundo, estos estudios se encuentran alejados del panorama de la minería aurífera aluvial de Colombia. Por este motivo, para esta investigación, se seleccionó una mina de Caucasia, un municipio colombiano de tradición minera. Esta mina se explota a nivel industrial por la empresa WGM, la cual proporcionó, para fines de esta investigación, los datos de 147 perforaciones realizadas en las etapas iniciales de exploración. Estas perforaciones tienen asociadas descripciones textuales y análisis geoquímicos en intervalos de profundidad aproximados de 0.3 m, dando un total de 8642 muestras para el estudio.

Con la información suministrada se abordaron dos tareas, la primera es el procesamiento de los datos usando técnicas de procesamiento de lenguaje natural, conocimiento experto, reductores dimensionales, técnicas de clustering y álgebra lineal para generar 3 sets de datos, los cuales son utilizados en la segunda tarea, la cual consiste en el modelamiento prospectivo 2D y 3D mediante redes neuronales artificiales, interpolación lineal, interpolación Kriging y la reproducción de un estudio desarrollado por investigadores en Australia.

Como resultado se obtuvieron 3 rutas metodológicas aplicables al entorno de minería aluvial, las cuales permiten combinar información cualitativa y cuantitativa e integrarlas con modelos de interpolación espacial, para la generación de modelos 2D y 3D del subsuelo con distinto desempeño, siendo el caso de mejor resultado, un mapa prospectivo de zona de interés que concuerda con los hallazgos obtenidos de las explotaciones realizadas en la mina.

Palabras claves

Machine learning, mapeo prospectivo, minería aluvial, oro, Colombia, interpolación espacial

Abstract

The extraction of mineral resources is a fundamental necessity for the world. Without their extraction, many of the utensils and tools we are used to in our daily lives would cease to exist: from simple kitchen utensils, household appliances, computers, means of transportation to hospital equipment or particle colliders, they require a vast amount of raw materials extracted from the subsoil. Among these is gold, a metal that is quite scarce but highly coveted for its industrial and domestic uses.

Most of Colombia's alluvial gold mining is concentrated in the department of Antioquia, where about 45% of the total gold produced in the country is extracted, which represents an important source of income for the companies and families involved in this activity. However, gold mining has many challenges, including prospecting and estimating the amount of gold in an area of interest. This task at an industrial level requires a high economic investment in subsoil studies, such as drilling, geophysical and geochemical techniques, among others, which in the end give indications of the viability of a mining project.

Due to the high cost of these studies, tools have been developed over the years to extract as much information as possible from the data obtained. In the panorama of gold and other elements, spatial interpolation techniques such as Kriging, inverse distance weighting (IDW) and discrete smooth interpolation, among others, are often used to obtain 2D and 3D models of the subsurface, which allows the establishment of areas of interest for exploration and/or mining. With the increase in computational capacity and the refinement of machine learning techniques, novel methodologies have been explored around the world to support mineral prospecting, among them are: spatial interpolation using random forests, support vector machines, stacking methods, deep learning, natural language processing, among many others.

Although there are studies in which machine learning techniques are applied to support mining prospecting in the world, these studies are far from the Colombian alluvial gold mining scenario. For this reason, for this research, a mine in Caucasia, a Colombian municipality with a mining tradition, was selected. This mine is exploited at industrial level by the company WGM, which provided, for the purposes of this research, the data of 147 drillings carried out in the initial stages of exploration. These drill holes have associated textual descriptions and geochemical analyses at depth intervals of approximately 0.3 m, giving a total of 8642 samples for the study.

With the information provided two tasks were addressed, the first is the processing of the data using natural language processing techniques, expert knowledge, dimensional reducers, clustering techniques and linear algebra to generate 3 data sets, which are used in the second task, which consists of 2D and 3D prospective modeling using artificial neural networks, linear interpolation, Kriging interpolation and the reproduction of a study developed by researchers in Australia.

As a result, 3 methodological routes applicable to the alluvial mining environment were obtained, which allow combining qualitative and quantitative information and integrating them with spatial interpolation models, for the generation of 2D and 3D models of the subsoil with different performance, being the case of the best result, a prospective map of the area of interest that agrees with the findings obtained from the exploitations carried out in the mine.

Key words

Machine learning, prospective mapping, alluvial mining, gold, Colombia, spatial interpolation.

Tabla de contenido

Lista de Figuras

Lista de tablas

Lista de Ecuaciones

Lista de abreviaturas

Abreviatura	*Español*	*Inglés*
DEM	*Modelo Digital De Elevación*	*Digital Elevation Model*
GP	*Procesos Gaussianos*	*Gaussian Process*
IDW	*Ponderación De Distancia Inversa*	*Inverse Distance Weighting*
ML	*Aprendizaje Automático*	*Machine Learning*
NER	*Reconocimiento De Entidades Nombradas*	*Named Entity Recognition*
NLP	*Procesamiento De Leguaje Natural*	*Natural Language Processing*
PCA	*Análisis De Componentes Principales*	*Principal Component Analysis*
PUL	*Aprendizaje Positivo Y Sin Etiquetas*	*Positive And Unlabeled Learning*
RBF	*Función De Base Radial*	*Radial Basis Function*
RF	*Bosques Aleatorios*	*Random Forest*
RNA	*Redes Neuronales Artificiales*	*Artificial Neural Networks*
RT	*Arboles De Regresión*	*Regression Tree*
SVM	*Máquina De Vectores De Soporte*	*Support Vector Machine*
T-SNE	*Incrustación De Vecinos Estocásticos Distribuidos En T*	*T-Distributed Stochastic Neighbor Embedding*

1 Revisión de la literatura

1.1 Introducción

En Colombia la producción minera representa aproximadamente el 2 % del PIB del país (Arisi et al., 2017; Minenergía, 2020). Solo en el departamento de Antioquia, entre el 2004 y el 2016, se produjo alrededor del 48 % del total de oro nacional (Unidad de Planeación Minero Energética - UPME et al., 2017), donde la minería aluvial es la principal actividad productora de oro. Para el año 2021, Colombia contaba con aproximadamente 24677 ha con permisos técnicos y ambientales para explotación (Minenergía & UNODC, 2021).

Para un sector como la minería, en el ámbito mundial, el atractivo económico para los inversionistas reside en la prospección o potencial para encontrar yacimientos importantes (Martinez et al., 2021). Esto no es una tarea sencilla, dado que involucra varias actividades propias de las geociencias como estudios geológicos, geoquímicos y geofísicos con los que al final se espera obtener un modelo prospectivo del depósito mineral, con el cual se evalúe la viabilidad del proyecto minero. Estos estudios tienen relaciones muy complejas, lo cual los vuelve difíciles de modelar; por tal motivo con el auge de la captación de datos y la capacidad de cómputo se han explorado metodologías basadas en *Machine Learning (ML)* para abordar una gran variedad de problemáticas propias de las geociencias (Bergen et al., 2019). Entre estas se encuentran: la caracterización de formaciones (Hidalgo, 2020), la determinación de propiedades petrofísicas (Iturrarán-Viveros et al., 2018), la sismología (Luzón, 2018), la prospección minera (Fu et al., 2021; Kuhn et al., 2019; Xiong et al., 2018), entre otras.

Si bien en la literatura existen investigaciones en diversas zonas del mundo enfocadas en la prospección minera con ML, el entorno geológico de estos trabajos es diferente a uno de minería aluvial. Algunas investigaciones recientes sobre prospección minera en entornos no aluviales se realizan en: el trabajo de (Li et al., 2020) donde se aplica *transfer learning* y redes neuronales convolucionales para hacer un mapeo prospectivo de la mineralización aurífera en China; en (Sun et al., 2019) se crea un mapa prospectivo basado en Sistema de Información Geográfica (GIS) mediante la implementación y comparación de tres métodos de ML para un entorno de mineralización de cobre en China; además en (Yeomans et al., 2020) se emplean transformaciones con lógica difusa y un sistema de ML para apoyar el proceso de prospección en una mina de tungsteno en Inglaterra; finalmente en (Zhang et al., 2021) se aborda el problema de la prospección minera aurífera desde tres panoramas del ML, el aprendizaje supervisado, semi supervisado y no supervisado.

Por otro lado, en Colombia la escasez de estudios al 2022 en los que se evalúen metodologías de prospección minera basadas en *ML* y principalmente en entornos aluviales, da a entender que la prospección minera nacional tiene un vacío en esta tecnología emergente, por tal motivo en esta investigación se busca la creación de modelos prospectivos con ML para un entorno de minería aluvial en Colombia.

Para ello, con la información y los datos de prospección y exploración tomados en una mina del municipio de Caucasia en el departamento de Antioquia (Colombia), lugar en el cual realizaron 147 perforaciones y de éstas se obtuvieron 8642 descripciones litológicas y el respectivo análisis geoquímico. Con estos datos, se abordaron diversas técnicas para el agrupamiento y extracción de características tales como clasificación mediante criterio experto, extracción de características

mediante procesamiento de lenguaje natural relacionado con datos geoquímicos, reductores dimensionales y clasificación no supervisada (agrupamiento / clustering) *K-means* dando como resultado tres *sets* de datos, de los cuales, dos fueron usados para clasificación y uno para regresión.

Teniendo en cuenta los sets de datos resultantes, se abordaron las tareas de clasificación y regresión mediante diferentes modelos de ML. En la clasificación se implementó la metodología de procesamiento de lenguaje natural para la creación de modelos 3D del subsuelo propuesta por (Fuentes et al., 2020). En segundo lugar, se propuso la aplicación de una metodología que permitió integrar por medio de redes neuronales convolucionales la topografía del terreno con las descripciones litológicas, para crear modelos 3D del subsuelo con fines prospectivos. Por último, se implementó una interpolación lineal para utilizarla como punto de referencia a la hora de analizar los resultados. En la regresión se modificó la red neuronal propuesta con el fin de suplir la tarea de datos continuos. Para finalizar, se implementó un *Kriging* ordinario como punto de comparación para la red neuronal.

Como resultado de lo anterior, se identificaron las rutas de procesamiento de datos más adecuadas según la meta trazada, se compararon los diferentes modelos implementados según los sets de datos yse obtuvo sus representaciones 2D y 3D de dichos modelos.

1.2 Planteamiento del problema

Colombia es un país con potencial para la inversión minera. Solo el 3.17 % del territorio colombiano se dedica a la minería, actividad, que representó para el primer trimestre de 2019 el 1.72 % del PIB total del país, llegando a 3.50 billones de pesos, donde el carbón participa con el 68.43 %, seguido por los minerales metálicos con el 18.51 % (Fuentes López et al., 2021). Entre el 2004 y 2016 Antioquia ha sido uno de los principales productores de oro del país, extrayendo alrededor del 48 % de la producción total en ese periodo (Unidad de Planeación Minero Energética -UPME et al., 2017). El futuro de la explotación de recursos minerales se basa en la determinación precisa de reservas con garantías económicas para un inversionista (Martínez et al., 2021). Aún así, a pesar de la importancia de estos recursos minerales para Colombia y del oro en Antioquia, la ausencia de estudios, al 2022, en los que se evalúe el rendimiento de las nuevas metodologías prospectivas basadas en *ML* en el país, como las desarrolladas por (Fu et al., 2021; Kuhn et al., 2019; Xiong et al., 2018), da a entender que la evaluación de recursos minerales en Colombia tiene un vacío en esta tecnología emergente, la cual promete resultados más precisos, reproducibles y no empíricos en la evaluaciónde reservas.

En los últimos años el ML ha desempeñado un papel fundamental en las ciencias de la tierra (geociencias) debido a que ha permitido a los geocientíficos analizar y modelar grandes volúmenes de información con muy buenos resultados (Bergen et al., 2019). La eficiencia de estos algoritmos se ha evaluado en la caracterización de formaciones geológicas (Hidalgo, 2020), en estudios petrofísicos (Iturrarán-Viveros et al., 2018), sismología y vulcanología (Luzón, 2018), lo cual ha generado resultados prometedores.

No obstante, el ML presenta muchos retos en las geociencias relacionados con los datos. Su alta dimensionalidad, su distribución espacial en escenarios de geología compleja (como muchos lugares en Colombia) y la instrumentación empleada para adquirirlos, son solo algunos de los obstáculos para los diferentes métodos de ML (Karpatne et al., 2018). Por lo tanto, existe un gran reto, enfrentar dicha problemática y encontrar la arquitectura y la metodología de procesamiento de datos que mejor se adapte para su solución.

En el estudio de recursos minerales se han usado diferentes algoritmos de ML para la exploración de depósitos, como en el caso del uso de máquinas de soporte vectorial (SVMs, por sus siglas en inglés) para obtener un mapa de zonas prospectivas de un depósito de cobre localizado en Irán (Abedi et al., 2012) lo cual parte de técnicas geofísicas magnéticas y de resistividad, anomalías geoquímicas y singularidades geológicas. Redes neuronales artificiales (RNA), árboles de regresión (RTs, por sus siglas en inglés), bosques aleatorios (RF, por sus siglas en inglés), y SVMs, utilizadas para modelar la prospección mineral de oro epitermal del distrito de Rodalquilar en España (Rodriguez-Galiano et al., 2015) haciendo uso de 46 lugares de ocurrencia de oroconocidos, zonas de mineralización y datos fisicoquímicos. Se han empleado algoritmos de ML combinados con datos de gravimetría, magnetometría, electromagnéticos, mapeo geológico y resultados de ocurrencias minerales, para el mapeo de prospección mineral en Columbia Británica (Granek & Haber, 2015). Sin embargo, a pesar del auge del ML en las geociencias, los modelos de distribución de mineral y química elemental (área de interés de este trabajo) se siguen generando con métodos computacionales tradicionales, que emplean interpolación de *Kriging*, el índice local de Morgan, *inverse distance weighting* (IDW) e interpolación suave discreta (Gao et al., 2016; Jin et al., 2020), para determinar las ubicaciones de puntos de mayor concentración de los elementos de interés usando datos geoquímicos, testigos de perforaciones y anomalías geológicas.

Actualmente, existen publicaciones relacionadas con la creación de modelos 3D del subsuelo y/o prospección minera, tales como el trabajo de (Jin et al., 2020) en el que se realiza un modelado tridimensional mediante interpolación clásica para la evaluación de recursos auríferos en China, El trabajo de (Fuentes et al., 2020), en el que emplean un modelo de procesamiento de lenguaje natural para extraer características de más de 100 mil perforaciones y 800 mil descripciones para después crear un modelo litológico 3D de un fragmento de la zona de estudio mediante interpolación clásica en Australia. Finalmente, la publicación de (Fu et al., 2021) en la que concluyen que en Zhuxi (donde se encuentra el depósito de tungsteno más grande del mundo), los modelados 3D de prospección mineral basados en ML, junto con datos geológicos, geofísicos y geoquímicos son útiles en el descubrimiento de nuevos depósitos de mineral.

En el trabajo de (Jia et al., 2021) clasifican *voxeles* mediante apilamiento de métodos de ML mostrando una mejora en el rendimiento para datos multiclase y de alta dimensionalidad y una reducción de tiempo respecto a los métodos convencionales. (Li et al., 2020) utilizan la red neuronal de Google *Inception-v3* aplicando aprendizaje por transferencia para analizar 9 capas de geoquímica y clasificar zonas prospectivas, aplicando simultáneamente técnicas para el aumento de datos para tratar de solucionar una problemática común del ML en geociencias que es la escasez de datos. Para integrar el conocimiento experto y la capacidad de los métodos de ML, (Yeomans et al., 2020) utilizaron una transformación con lógica difusa la cual permitió, gracias al conocimiento de la zona, identificar nuevos puntos posibles de exploración que no se identificaron al aplicar únicamente ML.

En (Zhang et al., 2021) abordaron el problema del desbalance de etiquetas (falta de muestras positivas) en la clasificación de zonas de prospección minera utilizando el método de aprendizaje positivo sin etiqueta (PUL), el cual, para los datos usados y la naturaleza de su problema, demuestra ser más eficaz que los métodos convencionales utilizados para el mapeo mineral prospectivo 3D. En el trabajo realizado por (Kuhn et al., 2019) se exploró la eficiencia de los métodos de ML para la prospección minera según la etapa de producción de la mina; se destacaron dificultades para esta tarea como el desequilibrio de etiquetas y la incertidumbre que se le ingresa a los modelos de ML al utilizar datos interpretados para validación, los cuales dependen en gran medida del experto que los realiza. También hay trabajos que exploran la interpolación espacial con ML para otros fines como el monitoreo de cultivos, distribución de contaminantes del aire, variables ambientales, entre otros, los cuales pueden ser aplicados a la prospección mineral (Bromberg & Pérez, 2012; Ma et al., 2019; Sekulic et al., 2020; Zhu et al., 2020). Aunque estas metodologías aún están en desarrollo con variadas aplicaciones en algunos lugares del mundo, en Colombia es prácticamente inexistente, aún conociendo la historia minera que tiene el país y la importancia para su economía. Esto refuerza la relevancia que tiene el desarrollo metodológico de este trabajo, para el mapeo de zonas de mineralización usando ML, con fines prospectivos en el país.

En la Figura 1.1 se resumen brevemente los enfoques computacionales de prospección mineral empleados en los trabajos citados anteriormente.

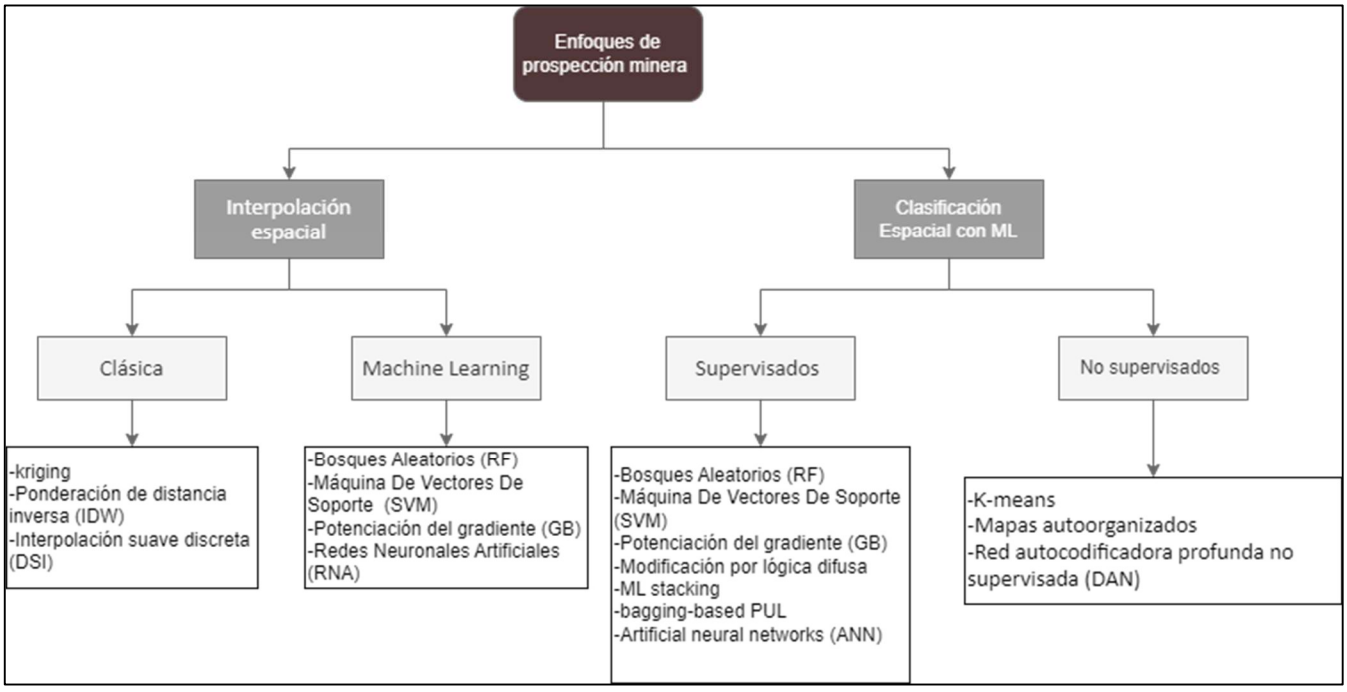


Figura 1.1. Resumen de rutas metodológicas que se emplean actualmente para realizar tareas de prospección minera según los trabajos citados.

Históricamente hablando, Colombia es un país con un gran potencial minero, a lo largo de los años la extracción de metales como el oro ha influido fuertemente en la economía nacional, el aprovechamiento de estos recursos depende en gran medida de la prospección, por tal motivo, en diversas partes del mundo han explorado tecnologías emergentes para reemplazar y/o mejorar las técnicas prospectivas tradicionales. Entre estas tecnologías emergentes se encuentran los algoritmos de Machine Learning, con los cuales se han abordado una gran variedad de escenarios en torno a la prospección, sin embargo, Colombia tiene una escasez en estudios relacionados y son prácticamente inexistentes (hasta donde pude

encontrar) los estudios de mapeo prospectivo con Machine Learning enfocado en la minería aurífera aluvial, por tal motivo es pertinente y novedoso el caso de estudio de esta investigación.

1.3 Hipótesis

La metodología para la generación de modelos relacionados con zonas de mineralización usando algoritmos de *machine learning* entrenados y validados con datos geológicos, geofísicos y geoquímicos apoyarán las exploraciones y explotaciones mineras en los procesos del encadenamiento productivo minero.

1.4 Objetivos

1.4.1 Objetivo general

Desarrollar una metodología para la obtención de modelos espaciales del subsuelo relacionados con depósitos minerales mediante métodos de analítica de datos y *machine learning* empleando datos geológicos, geofísicos y geoquímicos para apoyar la prospección minera.

1.4.2 Objetivos específicos

- Estructurar una base de datos geofísicos, geoquímicos y testigos de pozo mediante la obtención en campo, extracción y depuración de diversas fuentes de acceso público como insumo para el trabajo computacional cuyos resultados puedan ser validados en un contexto geológico específico.
- Determinar los parámetros de roca objetivos en cada set de datos mediante el análisis de las singularidades y el lugar donde fueron tomados, para la futura formación y validación de modelos de *machine learning* según el contexto geológico minero del lugar de estudio.
- Proponer una ruta computacional para el procesamiento, entrenamiento y validación de modelos de *machine learning* para la estimación de los parámetros de rocas de interés.

2 Marco teórico y metodológico

2.1 Minería

2.1.1 Conceptos

Minería aluvial:

La minería aluvial es la extracción de minerales valiosos u otros materiales geológicos de depósitos aluviales, que son sedimentos de arena, grava u otros materiales transportados y depositados por el agua (Rajapakse, 2016). Este tipo de minería generalmente se realiza en áreas donde los minerales o materiales se han sedimentado en lechos de ríos, llanuras aluviales u otras áreas que han sido afectadas por el flujo de agua, esto se debe a un fenómeno donde las corrientes de agua combinada con lodos, minerales y materia orgánica van erosionando el terreno por donde fluyen, lo que conlleva al aumento de la cantidad de material disuelto, el cual puede llegar a tener todo tipo de elementos, tales como el oro. Cuando el caudal del río disminuye, por diversos factores externos, los sedimentos más pesados se precipitan en su lecho.

El proceso de extracción generalmente implica el uso de equipos de excavación, como dragas Figura 2.1 o cajas de compuertas Figura 2.2, para extraer los materiales del depósito aluvial. Las dragas suelen ser grandes plataformas flotantes que están equipadas con una variedad de herramientas, como una manguera de succión, para extraer los materiales del depósito. Las cajas de compuertas se utilizan para separar los minerales o materiales valiosos del resto del yacimiento por medio de la sedimentación.

Figura 2.1. Foto de una draga tomada en salida de campo

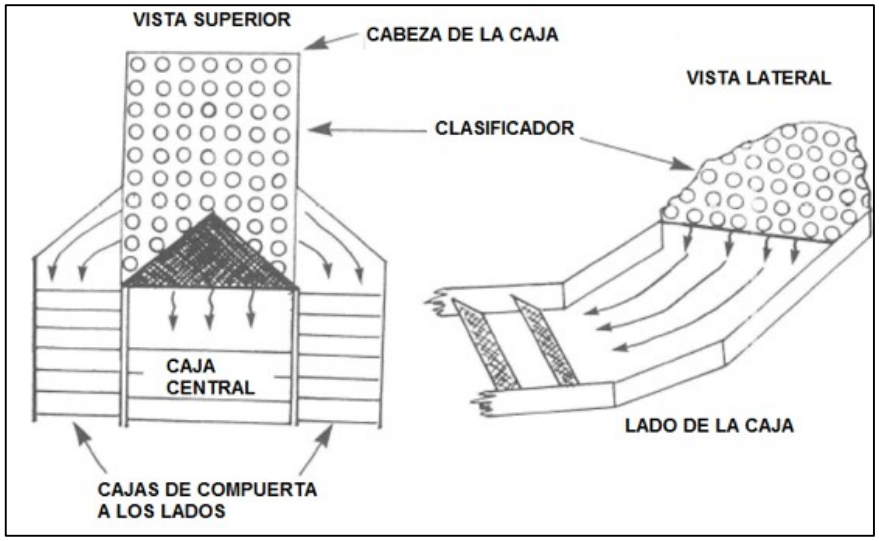


Figura 2.2. Esquema de caja de compuertas utilizada en minería aluvial. Imagen tomada de (David, 2021)

Trazadores de oro: Debido a las bajas concentraciones habituales de oro, en las tareas de exploración, se suelen identificar minerales y elementos asociados al oro, estas relaciones son propias de la zona y se les denomina trazadores. La identificación de trazadores es una técnica bastante extendida a lo largo del mundo y no se limita únicamente al oro; un ejemplo de su aplicabilidad en la identificación de oro y otros metales raros se encuentra en el trabajo realizado por (D'yachkov et al., 2021). En municipio de Caucasia los trazadores de oro son gravas, las cuales se asocian al lecho de los paleocanales (antiguo cauce del río), arenas negras asociadas a óxidos de hierro, los cuales por su alta densidad se sedimentan junto con el oro, y materia orgánica que suele coincidir con laparte central de los paleocanales donde usualmente se encuentra el oro.

Perforaciones – pozos: En las tareas de exploración del subsuelo, la perforación es una de las técnicas más utilizadas para identificar la composición y comportamiento del subsuelo (Saptawati & Nata, 2015). Esta técnica consiste en obtener datos del subsuelo mediante sensores incorporados en el taladro y/o en la extracción de muestras (testigos) para un posterior análisis en el laboratorio.

Modelo digital de elevación (DEM, por sus siglas en inglés): La geoforma del terreno es una información de gran valor para ciertos análisis y decisiones, para la identificación de montañas, cauces, valles, entre otros. Permite tener una perspectiva más amplia sobre los fenómenos que ocurren o han ocurrido en la zona. En la actualidad, una de las formas más precisas para representarlo, son los DEMs, estos son una representación 3D de la superficie de un terreno, representada como una cuadrícula de puntos de elevación espaciados uniformemente, los datos de elevación se pueden recopilar de varias fuentes, como levantamientos aéreos lidar, imágenes de radar o imágenes estereoscópicas satelitales, dicha cuadrícula de puntos de datos puede procesarse aún más para crear una superficie continua, usarse para crear representaciones 3D del terreno, también se pueden utilizar para generar líneas de contorno, mapas de sombras, pendientes y aspectos y otros tipos de datos geoespaciales (Polidori & El Hage, 2020).

2.1.2 Salida de campo

Entre el 16 y el 25 de junio de 2021 se realizó una de tres salidas de campo en el proyecto de explotación minera de la empresa WGM, entre los productos resultantes de dicha salida de campo se encuentra un informe elaborado por la geóloga Laura Alejandra Sánchez Giraldo dentro del

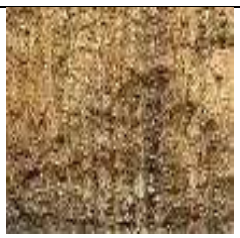
proyecto P20248, proyecto el cual abarca diferentes investigaciones en las que se incluye esta tesis y gracias al cual se conocieron los fenómenos propios de la zona de estudio y puntos de interés para la explotación. A continuación, se muestra un fragmento del informe en el que se ilustran la descripción litológica de la mina.

> Como se puede observar en la Figura 2.3 y en la *Tabla 1, "..... Se diferenciaron 12 estratos expuestos en el talud y 3 depósitos debido al colapso de estos pues el ángulo de corte no les permite ser lo suficientemente estables. También se observan 2 acumulaciones de agua, estas son producto del encuentro con el nivel freático a medida que avanza la actividad minera, para ello se utiliza una motobomba con la intensión de drenar el agua y sacarla de la zona de trabajo ..... "*

Figura 2.3. Identificación de estratos expuestos en frente de explotación de WGM. La inclusión de esta imagen es ilustrativa, el detalle de los estratos se describe en la Tabla 1.

Tabla 1. Estratos identificados en la salida de campo por la geóloga Laura Alejandra Sánchez Giraldo.

Estrato	Fotografía	Descripción
01		Es un estrato con sedimento de tamaño limo-arcilloso de color negro, en este se encuentran hojas y troncos, se define como materia orgánica. Su espesor aproximado es de 2 m.
02		Es un estrato de color gris oscuro, la matriz es de tamaño limo- arcillosa, se encuentran guijarros de diferentes tipos de roca,redondeados y de baja esfericidad, la selección general es media. Su plasticidad es baja y la humedad es media. La matriz corresponde a un 70 % y los guijarros al 30 %. Su espesor aproximado es de 70 cm. No tiene ningún tipo de orientación preferencial.
03		Es un estrato de color pardo oscuro, la matriz es de tamaño areno-limosa, se encuentran guijarros de diferentes tipos de roca, redondeados y de baja esfericidad, la selección general es buena.Su plasticidad es baja y la humedad es media. La matriz corresponde a un 80 % y los guijarros al 20 %. Su espesor aproximado es 5 m. No tiene ningún tipo de

		orientaciónpreferencial.
04		Es un estrato de color pardo claro a grisáceo, la matriz es de tamaño areno-limosa, se encuentran guijarros de diferentes tiposde roca, redondeados y de baja esfericidad, la selección general esbuena. Su plasticidad es baja y la humedad es media. La matriz corresponde a un 60 % y los guijarros al 40 %. Su espesor aproximado es de 2 m. No tiene ningún tipo de orientación preferencial.
05		Es un estrato de color crema a pardo claro, la matriz es de tamañolimo-arcillosa, se encuentran guijarros de diferentes tipos de roca,redondeados y de alta esfericidad, la selección general es media. Su plasticidad es alta y la humedad es media. La matriz corresponde a un 80 % y los guijarros al 20 %. Su espesor aproximado es de 2.5 m. No tiene ningún tipo de orientación preferencial.
06		Es un estrato de color pardo oscuro a naranja, la matriz es de tamaño limo-arcillosa, se encuentran guijarros de diferentes tiposde roca, redondeados y de alta esfericidad, la selección general esbuena. Su plasticidad es alta y la humedad es baja. La matriz corresponde a un 90 % y los guijarros al 10 %. Su espesor aproximado es de 3 m. No tiene ningún tipo de orientación preferencial.
07		Es un estrato de color pardo oscuro a naranja oscuro, la matriz esde tamaño limo-arcillosa, se encuentran guijarros de diferentes tipos de roca, redondeados y de alta esfericidad, la seleccióngeneral es alta. Su plasticidad es alta y la humedad es baja. La matriz corresponde a un 80 % y los guijarros al 20 %. Su espesor aproximado es de 3 m. No tiene ningún tipo de orientación preferencial.
08		Es un estrato de color pardo, la matriz es de tamaño arena fina a media, se encuentran guijarros de diferentes tipos de roca, redondeados y de baja esfericidad, la selección general es alta. Suplasticidad es baja y la humedad es media. La matriz correspondea un 60 % y los guijarros al 40 %. Su espesor aproximado es de 3 a 7 m. No tiene ningún tipo de orientación preferencial.
09		Es un estrato de color pardo oscuro, la matriz es de tamaño arenafina a media, se encuentran guijarros de diferentes tipos de roca, redondeados y de baja esfericidad, la selección general es alta. Suplasticidad es baja y la humedad es media. La matriz correspondea un 70 % y los guijarros al 30 %. Su espesor aproximado es de 2 m. No tiene ningún tipo de orientación preferencial.

10		Es un estrato de color gris, la matriz es de tamaño arena fina a media, se encuentran guijarros de diferentes tipos de roca, redondeados y de baja esfericidad, la selección general es alta. Suplasticidad es media y la humedad es alta. La matriz corresponde a un 70 % y los guijarros al 30 %. Su espesor aproximado es de 2 m. No tiene ningún tipo de orientación preferencial.
11		Es un estrato de color gris oscuro a negro, la matriz es de tamañoarena fina a media, se encuentran guijarros de diferentes tipos deroca, redondeados y de baja esfericidad, la selección general es alta. Su plasticidad es baja y la humedad es alta. La matriz corresponde a un 70 % y los guijarros al 30 %. Su espesor aproximado es de 2 m. No tiene ningún tipo de orientación preferencial. Se conoce como mina pues es el estrato con mayor contenido de oro según el conocimiento empírico de los trabajadores de la mina.
12		Es de color gris oscuro, se encuentran bloques, el tamaño de granoes arena, y tienen un alto contenido de cuarzo (50 %). Estos bloques son de baja esfericidad, y buena selección. Es un estrato estéril. Las explotaciones de oro aluvial en la zona finalizan en este estrato.

2.2 Analítica computacional

2.2.1 Machine learning (ML)

El ML o aprendizaje automático es un área donde convergen la estadística, la matemática y la computación, con el fin de crear modelos y/o sistemas con la capacidad de resolver una tarea sin estar programados explícitamente para resolverla (Mahesh, 2019). El ML abarca una gran cantidad de subramas de investigación, las dos más conocidas son:

- Aprendizaje supervisado: Busca patrones presentes en los datos, mediante una estructura entrada-objetivo (input-target) (A. Singh et al., 2016), donde se ingresan datosy la etiqueta esperada al modelo, con el fin de que busque los patrones que relacionan los datos de entrada con la etiqueta. Un ejemplo de esto es la tarea de clasificar imágenes, donde la entrada (input) es una imagen y el objetivo (target) consiste en identificar la etiqueta asociada a la imagen.
- Aprendizaje no supervisado: A diferencia del aprendizaje supervisado no cuenta con etiquetas, su objetivo principal es encontrar similitudes entre los datos de entrada para obtener los subgrupos y relaciones que permitan generar etiquetas (Gentleman & Carey, 2008). Entre los métodos de aprendizaje no supervisado más conocidos se encuentran los reductores dimensionales y algoritmos de *clustering.*

2.2.2 Redes neuronales artificiales

Las redes neuronales artificiales (RNA) son una rama del ML inspirada en el funcionamiento del cerebro, concretamente en la robustez de su aprendizaje y la conexión de las neuronas (Krogh, 2008). Las redes neuronales están formadas por elementos unitarios llamados perceptrones, el perceptrón es el homologo a una neurona individual y tiene tres componentes principales: multiplicación, suma y activación Figura 2.4, (Kanal, 2003), en la entrada del perceptrón, los datos se multiplican por los pesos "W", donde cada valor individual de entrada se multiplica por un peso independiente, luego se suman los valores resultantes de la multiplicación entre ellos junto con el sesgo "b", por último el resultado es transformado por una función de activación dando una salida "y", este proceso se resume en la Ecuación 1 (Suzuki, 2011). Con el paso del tiempo se conectaron varios perceptrones dando origen al *Deep Learning* (Shinde & Shah, 2018) donde las arquitecturas de RNA se volvieron más complejas y con mayor capacidad de representación.

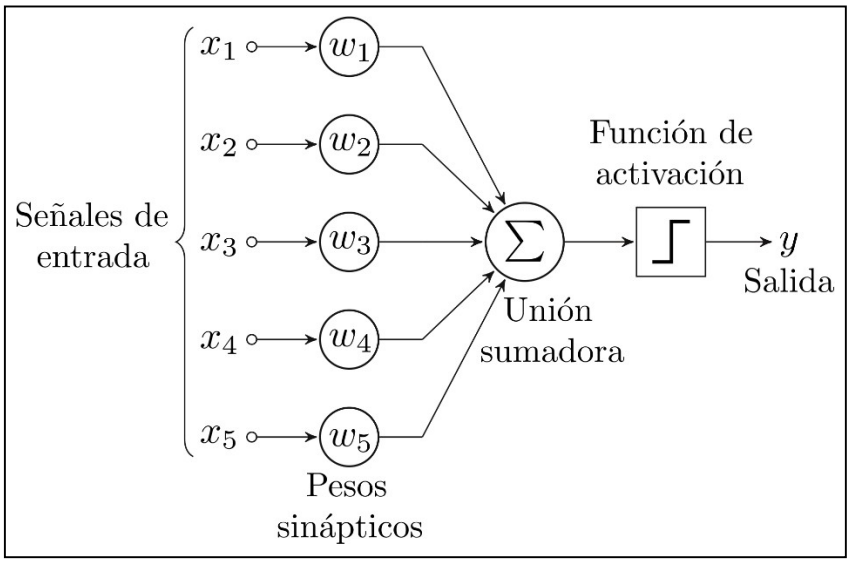


Figura 2.4 Esquema del perceptrón, tomado de («Perceptrón», 2021)

$$Y = F\left(\sum_{i=0}^{m} W_i * X_i + b\right)$$

Ecuación 1. Ecuación del perceptrón simple (Suzuki, 2011)

$$Y = F(y) = \begin{cases} 1 \; si \; y \geq threshold \\ 0 \; si \; y < threshold \end{cases}$$

Ecuación 2. Función de activación escalonada

Donde:

- Xi es el vector de entrada de longitud "m"
- Wi son los pesos del perceptrón
- b es el sesgo
- F es la función de activación

2.2.3 Redes neuronales convolucionales

Las redes neuronales convolucionales (CNN, por sus siglas en inglés, Figura 2.5) hacen parte del *Deep Learning*. Su principal campo de acción está en tareas que requieren extraer información de imágenes, tales como la clasificación, detección o segmentación mediante visión por computadora (Albawi et al., 2017). El campo de acción de éstas, al igual que de todo al área del *Deep Learning* es bastante amplio, pasando por la detección de objetos mediante imágenes satelitales (Guo et al., 2018), la clasificación de imágenes médicas (Anwar et al., 2018), reconocimiento de expresiones faciales (Pramerdorfer & Kampel, 2016), entre muchas otras.

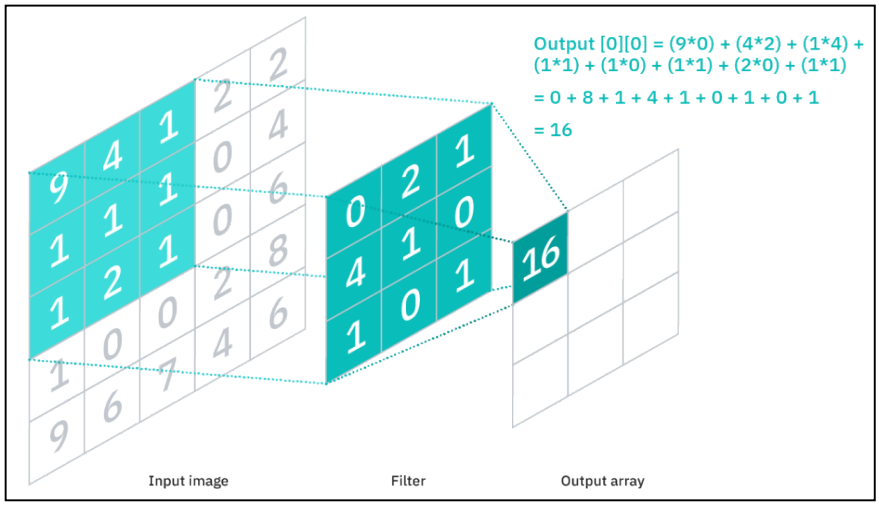


Figura 2.5. Ejemplo numérico de una Red Neuronal Convolucional. Imagen tomada de (*What Are Convolutional Neural Networks?*, 2021).

Estas redes se basan el filtros los cuales no son más que matrices, estos filtros se ubican sobre una sección de la representación matricial de la imagen de entrada donde se realiza un producto punto, el valor resultante indica la presencia o ausencia del patrón único del filtro en la respectiva sección de la imagen, posteriormente el filtro se desliza de izquierda a derecha y de arriba abajo hasta recorrer toda la imagen y obtener la nueva representación matricial de la entrada (Lopez Pinaya et al., 2020), a este deslizamiento se le denomina stride y equivale a la cantidad de celdas que se desplaza el filtro en cada iteración, esto se muestra en la Figura 2.6. Según las dimensiones del filtro y del stride vertical y horizontal, las dimensiones de la matriz de salida cambian, esto no siempre es deseado, por ello se emplea un relleno (padding) el cual es una margen de ceros alrededor de la matriz Figura 2.7, con esto y con los elementos mencionados, se puede emplear la Ecuación 3 para determinar las dimensiones de la matriz resultante.

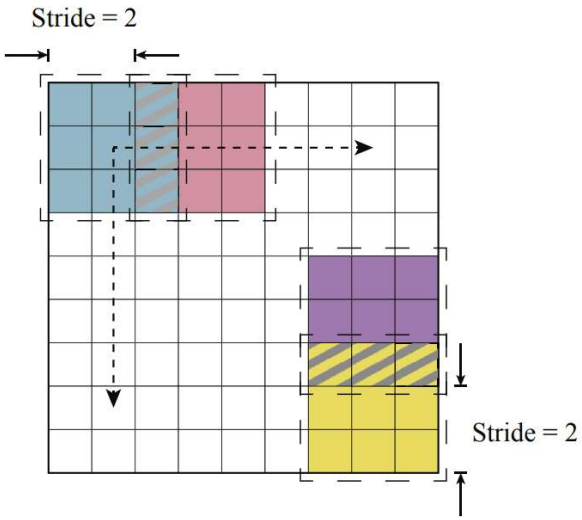


Figura 2.6. Ejemplo de convolución para una imagen de 9x9, un filtro 3x3 y un stride de , imagen tomada de (Lopez Pinaya et al., 2020)

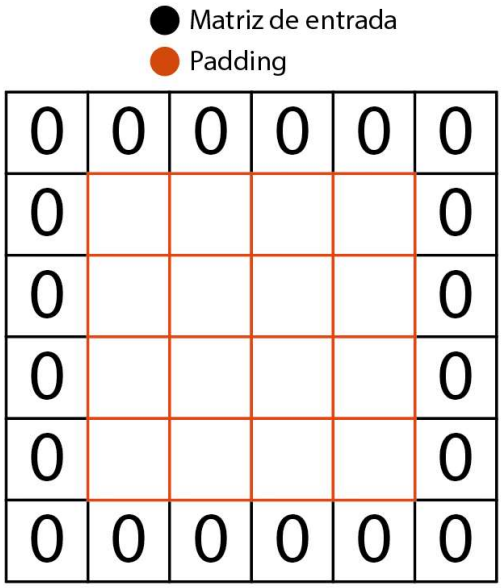


Figura 2.7. Representación de la implementación del Padding

$$Ancho\ salida\ = \frac{W\ -\ F_w\ +\ 2P}{S_w}\ +\ 1\ (A)$$

$$Alto\ salida\ = \frac{H\ -\ F_h\ +\ 2P}{S_h}\ +\ 1\ (B)$$

Ecuación 3. Ecuación que describe las dimensiones de salida de una convolución

Donde:

- W es el ancho de la matriz (imagen) de entrada
- H es el alto de la matriz (imagen) de entrada
- F_w ancho del filtro, F_h alto del filtro
- S_w stride horizontal, S_h stride vertical
- P es el relleno (Padding)

2.2.4 Procesamiento de lenguaje natural

El procesamiento de leguaje natural (NLP, por sus siglas en inglés) es un área de estudio que busca dotar a las computadoras la capacidad de procesar y entender la forma habitual en la que nos expresamos (Chowdhary, 2020). Entre las tareas más comunes del NLP se encuentran el análisis de sentimientos (Solangi et al., 2018) y reconocimiento de entidades nombradas (NER) (Sohrab & Miwa, 2018). Dentro de esta área existen diversas arquitecturas como las redes neuronales recurrentes (Jelodar et al., 2020) y las redes *Transformers* (Vaswani et al., 2017), las cuales dentro de su implementación hacen uso de las siguientes técnicas, las cuales fueron implementadas en el presente trabajo:

- **Tokenización:** Es una de las primeras fases del NLP; consiste en dividir cada oración y/o cadena de texto en tokens. Según la tarea, estos tokens pueden ser de varios tipos. Los más utilizados se presentan a nivel de carácter y a nivel palabra, por ejemplo: donde dada una oración "Hola mundo". Los tokens a nivel carácter son cada uno de los elementos de la oración ("H", "o", "l", "a", " ", "m", "u", "n", "d", "o"), los tokens a nivel de palabra ("Hola", "mundo") son cada una de las palabras que conforman la oración (Mielke et al., 2021; Webster & Kit, 1992).
  • **Embedding:** Los modelos de RNA requieren números para poder realizar sus operaciones matemáticas, sin embargo, los tokens son símbolos que no se pueden ingresar naturalmente a una operación matricial. Por esta razón existen los *embeddings*, los cuales son una forma de codificar vectorialmente cada token en un espacio "N" dimensional (ver Figura 2.8), donde la distancia entre tokens representa la similitud y relación entre ellos (Pennington et al., 2014).

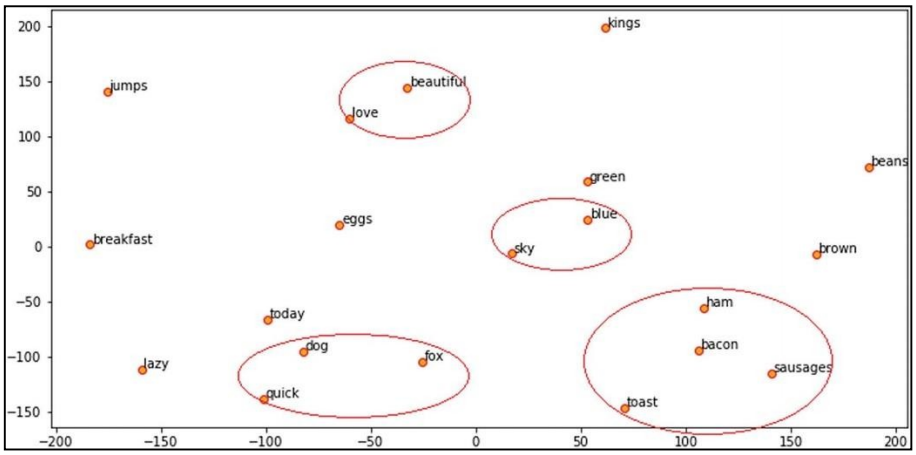


Figura 2.8. Ejemplo visualización de embedding de palabras GloVe en un espacio de dos dimensiones. Imagen tomada de (Sarkar, 2019).

El modelo GloVe en concreto es un algoritmo para la obtención de representaciones vectoriales de tokens, se basa en la observación de la probabilidad coocurrencias entre tokens respecto al contexto semántico con el cual se entrena el algoritmo. Un ejemplo de esto es la probabilidad de coocurrencia entre hielo (ice) y vapor (steam) de acuerdo a un contexto Figura 2.9, donde los valores del ratio $P(k|ice)/P(k|steam)$ cercanos a 1 hacen referencia a una característica compartida y los valores lejanos de 1 representan una diferencia entre los tokens, por ejemplo ambos tienen relación con el agua (wáter, 1.36) y tienen poca interacción en la moda (fashion, 0.96) en estos dos aspectos son similares, pero difieren en solido (solid, 8.9) y gaseoso (gas, $8.5 * 10^{-2}$) ya que poseen dierente estado de la materia (*GloVe: Global Vectors for Word Representation*, s. f.).

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

Figura 2.9. probabilidad de coocurrencia entre hielo (ice) y vapor (steam). Imagen tomada de (GloVe: Global Vectors for Word Representation, s. f.)

• **Codificador posicional:** El codificador posicional es una forma de representar la ubicación/posición de un elemento en una secuencia. En las redes transformers se expresa en forma de un vector con igual dimensión al *embedding*. Para generar este vector se mezcla codificación binaria traducida a frecuencia de senos y cosenos, creando un vector binario continuo (Kernes, 2021; Vaswani et al., 2017). Para llegar a este vector, se suele construir una matriz de "NxM" donde "N" es la posición máxima y "M" es la dimensión del embedding (ver Figura 2.10). El vector del codificador posicional para la posición "t" "P_t" está dado por la Ecuación 4 y se expresa de forma gráfica en la Figura 2.10

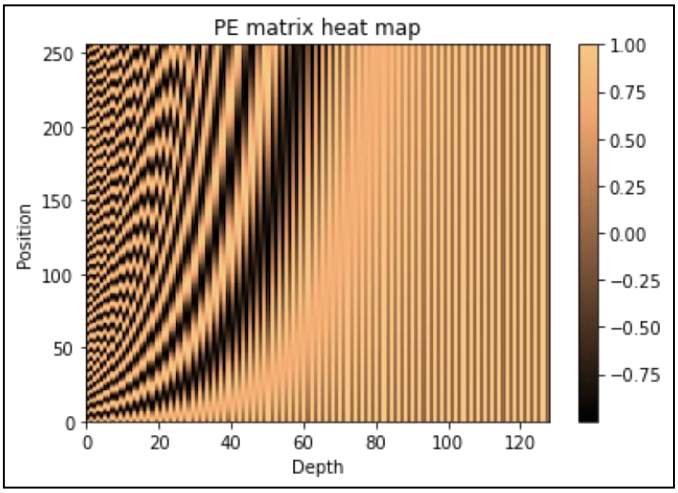


Figura 2.10. Representación matricial del codificador posicional, imagen tomada de (Kernes, 2021).

Ecuación 4. Descripción del enconder posicional. Tomado de (Kazemnejad, 2019)

$$\overrightarrow{p_t} = \begin{bmatrix} \sin(\omega_1 . t) \\ \cos(\omega_1 . t) \\ \\ \sin(\omega_2 . t) \\ \cos(\omega_2 . t) \\ \\ \vdots \\ \\ \sin(\omega_{d/2} . t) \\ \cos(\omega_{d/2} . t) \end{bmatrix}_{d \times 1}$$

Donde:

- i es la coordenada en el vector P_t
- Donde $k = [i/2]$
- d es la dimensión del embedcing
- t es la posición

$$W_k = \frac{1}{10000^{2k/d}}$$

2.2.5 Kriging

El método de interpolación Kriging (modelo de correlación espacial) hace parte de las técnicas de procesos gaussianos. Estas técnicas no paramétricas se utilizan para modelar y estimar funciones desconocidas partiendo de observaciones (datos conocidos) con las cuales se sintoniza una función probabilística. Por su naturaleza probabilística, además de la estimación del resultado, se obtiene un coeficiente que representa la varianza de la predicción (Schulz et al., 2018), esto se ve reflejado en la Figura 2.11. Una de las aplicaciones más conocidas de Kriging es la interpolación espacial, principalmente en geoestadística. Esto lo hace partiendo de la teoría de las variables regionalizadas y variogramas, donde se asume que la variación presenta una dependencia espacial entre los datos (Kleijnen, 2009; OLIVER & WEBSTER, 1990).

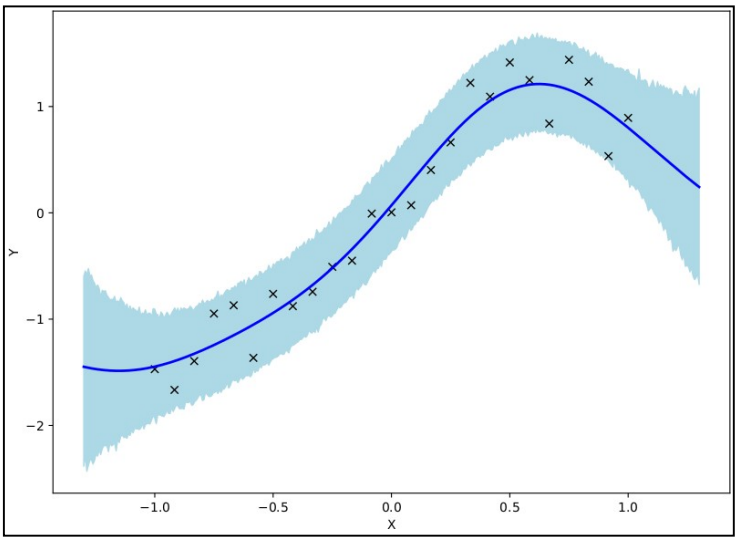


Figura 2.11. Ejemplo de una regresión gaussiana, donde las "x" son las observaciones, la línea azul es la media y la zona sombreada es la varianza. Imagen tomada de (*Example: Gaussian Process — NumPyro documentation*, s. f.).

2.2.6 K-means

Es un método de clustering el cual busca los patrones espaciales de un conjunto de datos para formar subgrupos con los elementos de mayor similitud, además de buscar la mayor diferencia entre distintos subgrupos (Figura 2.12). Una de las principales desventajas de este método es que se ve afectado por las condiciones iniciales, además es necesario ingresar el número de subgrupos que se desean formar (Sinaga & Yang, 2020). Este algoritmo se puede dividir en 5 pasos

1. Definir el número de centroides / subgrupos (k), donde k es menor al número de datos total.
2. Inicializar los centroides seleccionando k puntos aleatorios dentro del área del conjunto de datos
3. Luego a cada dato se le asigna el centroide más cercano, para esto se calcula la distancia euclidiana (Ecuación 5) de cada punto con todos los centroides y se selecciona el de la menor distancia
4. Se calcula la nueva posición del centroide promediando la posición de los puntos pertenecientes al subgrupo Ecuación 6
5. Repetir los puntos 3 y 4 hasta que los centroides sean estables

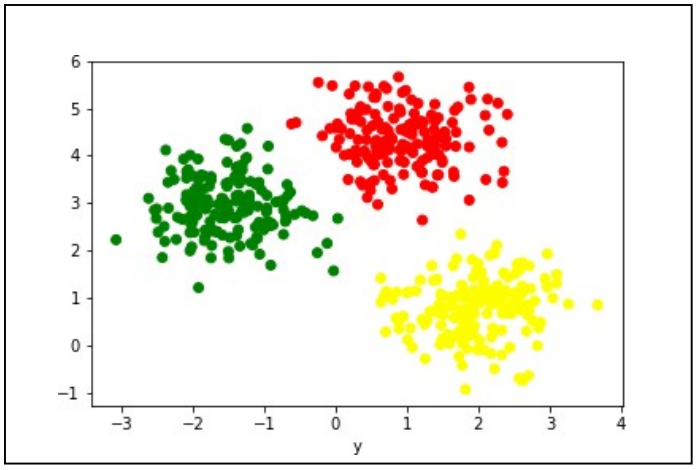


Figura 2.12. Ejemplo de k-means donde cada color representa un subgrupo, imagen tomada de (Vos, 2020).

$$dE(P, C) = \sqrt{\sum_{i=1}^{n} (p_i - c_i)^2} \qquad (A)$$

Ecuación 5. Distancia euclidiana utilizada en k-means. (Kanungo et al., 2002)

Donde:

- P es el punto en el espacio
- C es el centroide
- n es la dimensión del espacio vectorial

$$C_k = \frac{1}{m} \sum_{i=1}^{m} \overrightarrow{X_k} \qquad (B)$$

Ecuación 6. Ecuación para recalcular el centroide en k-means. (Kanungo et al., 2002)

Donde:

- C_k es el nuevo centroide
- m es el número de datos dentro del grupo del centroide
- X_k son los vectores pertenecientes al subgrupo K

2.2.7 Análisis de componentes principales

El análisis de componentes principales (PCA) es un método de reducción de dimensionalidad utilizado para representar un conjunto de datos proveniente de un espacio de "N" dimensiones en un espacio menor dimensión, conservando la mayor parte de la variabilidad de los datos mediante la identificación de los componentes principales (Ringnér, 2008). Este tipo de técnicas se suele utilizarpara llevar los datos a un espacio de 2 o 3 dimensiones con el fin de poder representarlos gráficamente e identificar la existencia de subgrupos y/o relaciones entre ellos.

2.3 Metodologías

En este trabajo se desarrolla una metodología para la generación de modelos relacionados con zonas de mineralización usando algoritmos de ML entrenados y validados con datos geológicos, geofísicos y modelos de física de rocas que se integre al planeamiento minero, que a mediano o largo plazo pueda estar al alcance de la pequeña y mediana minería. A continuación, se da un panorama de algunas de las investigaciones realizadas en diversas partes del mundo que han sido inspiración para la presente tesis.

2.3.1 **Contaminación por metales potencialmente peligrosos en el sistema suelo-arroz y su variación espacial en la ciudad de Shengzhou, China** (Gao et al., 2016)

Partiendo de una problemática de contaminación del suelo por metales pesados en Shengzhou, se analizaron 94 pares de muestras de arroz y tierra de la zona de estudio Figura 2.13, con la finalidad de determinar si estaban contaminando los cultivos, para esto, en los 94 pares de muestras de suelo y arroz se mide la concentración de metales tales como Cd, Cu, Ni, Pb, Zn distribuidos en la zona objetivo, luego se realizaron interpolaciones Kriging de la distribución de los metales con la finalidad de observar la distribución de su concentración en la zona Figura 2.14.

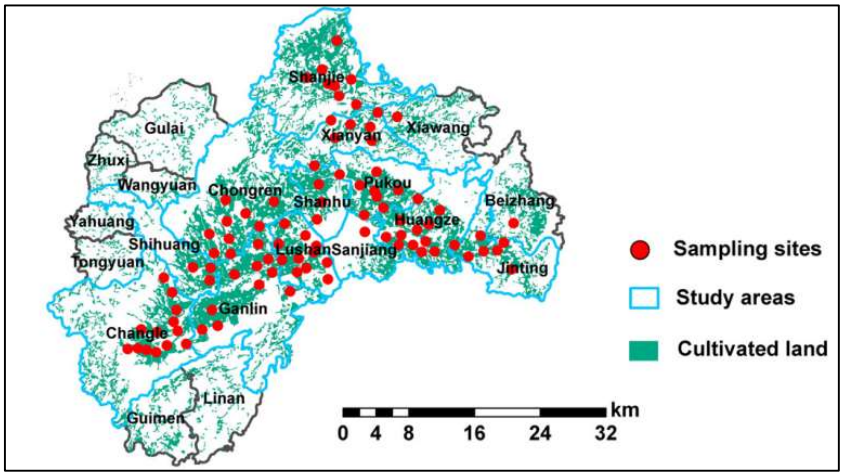


Figura 2.13. Puntos de muestreo en la zona de estudio. Imagen tomada de (Gao et al., 2016)

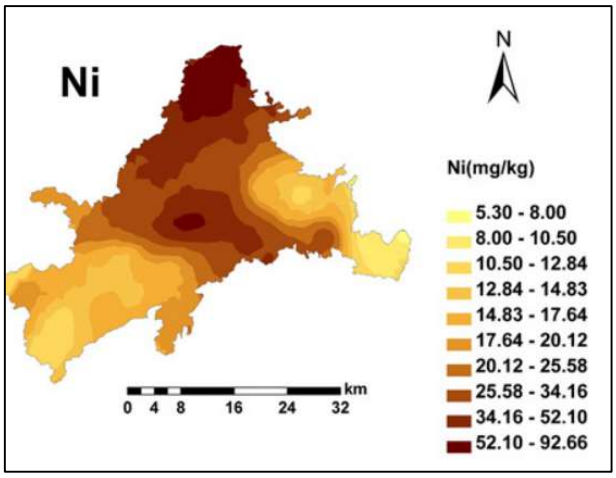


Figura 2.14. Resultado de la interpolación Kriging para Níquel. Imagen tomada de (Gao et al., 2016)

Se determinó que la concentración de Cd, Cu, Ni, Zn en el arroz no excede los reglamentos de China. Aunque no se superaron los umbrales, los resultados obtenidos son mucho mayores a los datos registrados por las entidades ambientales nacionales, lo que indica un aumento en la contaminación del suelo con el paso de los años. Como trabajos futuros se plantean tomar más muestras y repetir este tipo de estudios para datar la evolución de la concentración de estos metales en el suelo y en el arroz.

2.3.2 Un enfoque de aprendizaje automático para el modelado de prospectividad de tungsteno mediante la extracción de características basadas en el conocimiento y la confianza del modelo (Yeomans et al., 2020)

Se aborda la necesidad de apoyar el proceso de prospección minera de tungsteno en Inglaterra Figura 2.15, para esto se implementan transformaciones de lógica difusa y algoritmos de machine learning con el fin de identificar los criterios claves de exploración. El modelamiento prospectivo es basado en datos geofísicos radiométricos aerotransportados, geológicos y geoquímicos, entre los datos geológicos se encuentra una extensión mapeada de plutones de granito y la profundidad de capa de granito.

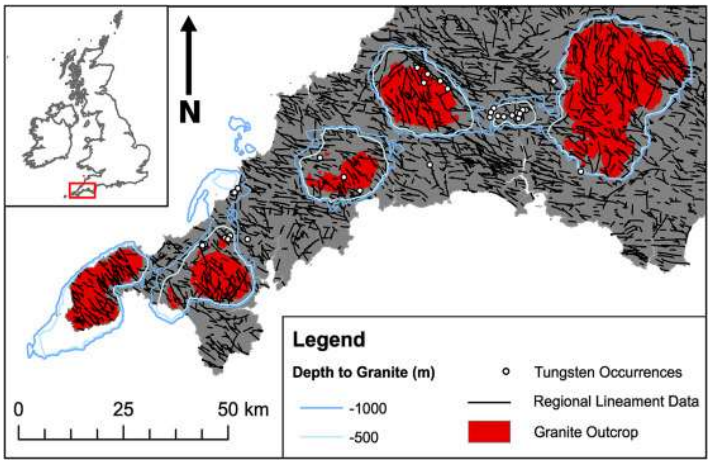


Figura 2.15. Resumen de la geología de la zona de estudio. Imagen tomada de (Yeomans et al., 2020)

Para desarrollar el mapa prospectivo, se obtienen una serie de características provenientes de las transformaciones difusas, luego se compara el rendimiento del método RF para clasificación entre los datos siendo transformados y los datos sin transformar, donde el objetivo del modelo de RF es determinar la probabilidad de mineralización de tungsteno de cada píxel, esta metodología se resume en la Figura 2.16. El resultado de la clasificación es en un plano 2D con la confianza y probabilidad de mineralización Figura 2.17.

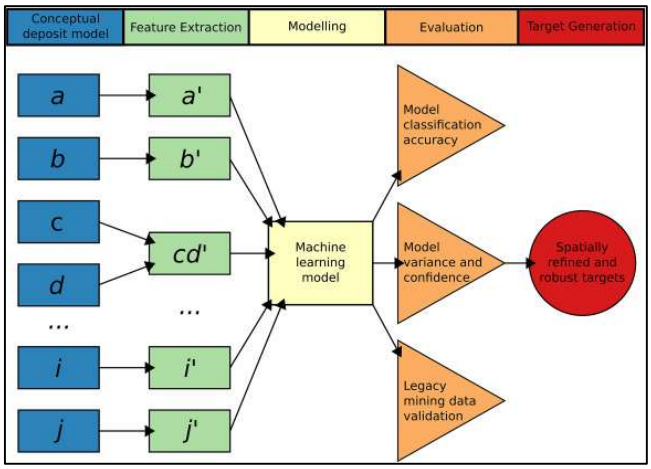


Figura 2.16. Resumen del flujo de trabajo propuesto por (Yeomans et al., 2020). Imagen tomada de (Yeomans et al., 2020)

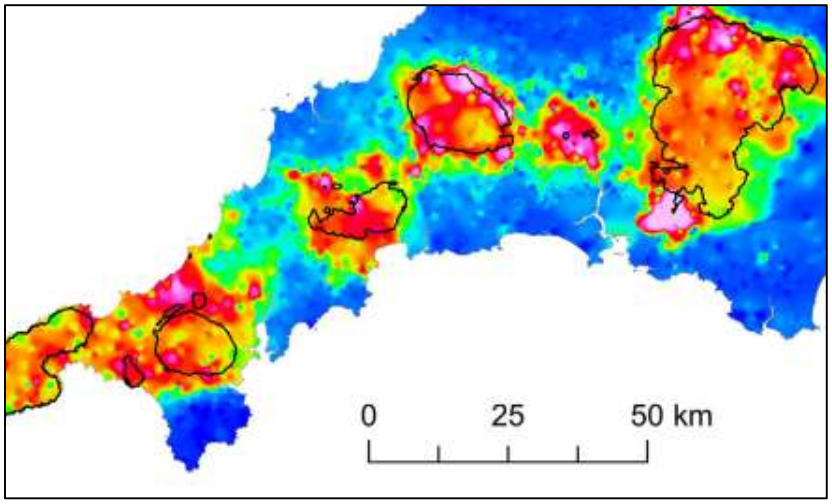


Figura 2.17. Modelo prospectivo geoquímico de tungsteno resultante de la metodología propuesta en (Yeomans et al., 2020). Imagen tomada de (Yeomans et al., 2020)

El método de RF presentó un buen rendimiento en los escenarios de lógica difusa y sin lógica difusa, se destaca que utilizando los datos sin la transformación difusa el modelo solo enmarcó las zonas conocidas, en cambio el modelo con las entradas provenientes de la transformación difusa enmarcó nuevas posibles zonas de explotación.

2.3.3 **Mapeo de prospectividad mineral a través de análisis de big data y un algoritmo de aprendizaje profundo** (Xiong et al., 2018)

Se busca identificar anomalías de mineralización para mapear la prospectividad minera de hierro tipo skarn en la zona metalogénica del suroeste de Fujian en China, mediante la implementación de una red autocodificadora profunda no supervisada (DAN). Para esto utilizaron 42 capas 2D, dos provenientes de la geología, una de la geofísica y 39 de la geoquímica de la zona.

El área de estudio comprende rocas intrusivas, formaciones sedimentarias, fallas y depósitos de hierro, los datos geoquímicos de sedimentos tomados a una densidad de 1 muestra por 4 km^2 se recopilaron a partir del proyecto de cartografía geoquímica nacional de China, Los datos geofísicos consisten en datos de intensidad magnética total en el aire muestreados a una resolución espacial de 2 km, proporcionados por el Servicio Geológico de China. Con estos datos se crearon rasters GIS de $(1km)^2$ para las 42 variables, posteriormente se crearon los vectores de entrada para el modelo DAN (Figura 2.18) el cual clasifica de 0 a 1 la probabilidad de mineral en la celda, dando como resultado la Figura 2.19 donde el modelo es capaz de identificar zonas ya conocidas.

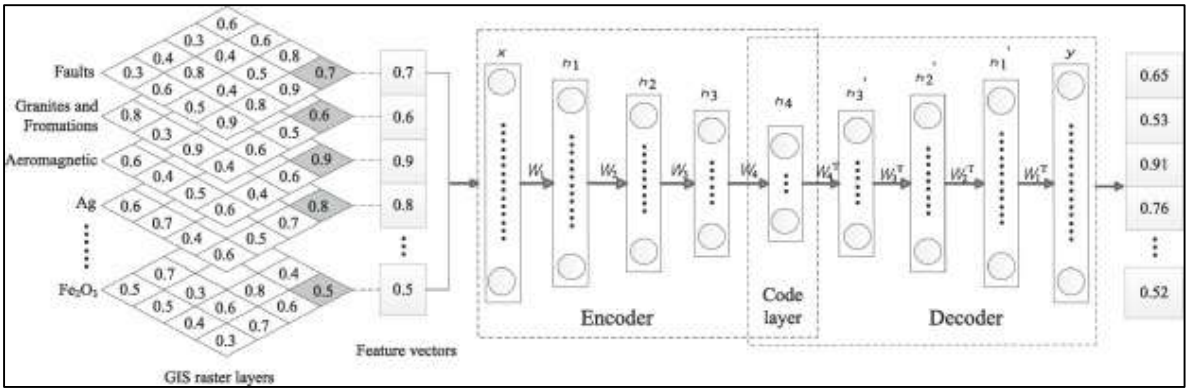


Figura 2.18. Esquema de la red autocodificadora profunda con sus 42 capas 2D de entrada. Imagen tomada de (Xiong et al., 2018)

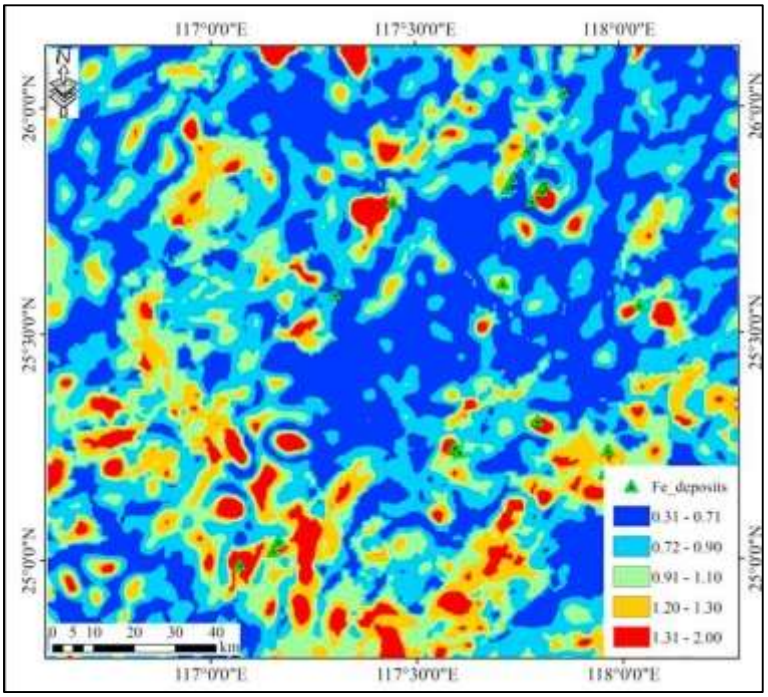


Figura 2.19. Mapa de mineralización de hierro resultante de la red autocodificadora profunda. Imagen tomada de (Xiong et al., 2018)

2.3.4 **Mapeo litológico en el cinturón de cobre de África Central utilizando Random Forests y clustering: estrategias para obtener resultados optimizados** (Kuhn et al., 2019)

Se aborda la tarea del mapeo litológico en el cinturón de cobre en África desde 4 perspectivas diferentes con el objetivo de producir y/o refinar mapas geológicos a partir de observaciones directas limitadas, para esto, se plantean 4 experimentos de ML, donde 3 experimentos son mediante la implementación de Random Forests y uno con algoritmos no supervisados k-means y Self-Organizing Maps. Los cuatro casos de estudio se describen a continuación.

- C1: Utiliza RF para el mapeo litológico utilizando muestras de afloramiento originales, emulando la etapa temprana de exploración y con desbalances de clases
- C2: Utiliza RF para el mapeo litológico utilizando muestras de afloramiento originales, emulando la etapa temprana de exploración con tamaño de muestras balanceadas por clase
- C3: Evaluar el desempeño de RF en la auditoria y mejora de mapas existentes, utilizando un

pequeño subconjunto de datos muestreados al azar de un mapa de interpretación geológica bien desarrollado para la generación de datos de entrenamiento (replicando una etapa más madura en la exploración)

- C4: evalúa la capacidad de los algoritmos de agrupamiento para producir una clasificación, en ausencia de cualquier entrada del usuario, que corresponda a la geología mapeada a la escala del proyecto

Para realizar estos experimentos se utilizaron datos geofísicos y geoquímicos del suelo proporcionados por la compañía que explota la zona, se obtuvieron conjuntos de datos geofísicos adicionales de estudios anteriores y se agregó el modelo digital de terreno (DEM), entre los datos geofísicos se encuentran aeromagnéticos, electromagnéticos aerotransportados, los datos totales comprenden aproximadamente 178.000 instancias, cada una con 59 variables. Se eliminaron datos con una correlación de Pearson mayor al 0.8, también datos ruidosos o con gran cantidad de datos faltantes, eliminando en total 15 variables, en el caso 1 se trabaja con muestras desequilibradas, en el caso 2 se utiliza muestreo Bootstrap y diezmado para corregir el desequilibrio, en el caso 3 se utiliza una información más extensa emulando una exploración madura, en el caso 4 se alimentó k-means con el total de las 178.000 instancias y se variaron las iteraciones con el número de grupos de 2 a 20

Utilizando RF con 500 árboles para C1, C2 y C3 se obtuvo una precisión en la validación cruzada C1=75.4%, C2=88.8% y C3=80.5%. C1 se muestra que el modelo de RF es dominado por el desequilibrio de las muestras, el resultado final tiene una similitud con el mapa interpretado de solo el 17%. C2 al igual que con C1, la coherencia con respecto al mapa de interpretación geológica fue deficiente, sin embargo, en este caso, aunque las etiquetas de litología siguen siendo inexactas, la muestra de entrenamiento equilibrada produjo resultados que se ajustan más a la geometría de los límites principales en el área del proyecto. El mapa resultante de C3 tuvo una consistencia del 67.2% con el mapa de geología interpretada debido a la información extra suministrada. En C4 ambos métodos (k-means y Self-Organizing Maps) mostraron una fuerte relación con los patrones de drenaje y lograron separar dos zonas de fenómenos litológicos definidos.

2.3.5 Modelado de prospectividad mineral en 3D basado en aprendizaje automático: un estudio de caso del depósito de tungsteno de Zhuxi en el noreste de la provincia de Jiangxi, sur de China (Fu et al., 2021)

Partiendo de datos del depósito de tungsteno de Zhuxi, el más grande del mundo, se plantea una metodología para el mapeo prospectivo 3D. Esta metodología toma como base la creación de un modelo 3D del subsuelo, partiendo de perforaciones, secciones geológicas, datos de perfil, muestras de roca y datos geofísicos como gravedad, magnetismo, magnetotelúricos y sísmicos (invertidos) haciendo uso de la interpolación provista por el software Geomodeller, esto da como resultado un área de superficie de aproximadamente 2694.14 km2 y una profundidad de 5 km.

Luego de tener el modelo 3D (Figura 2.20), se demarcan las zonas donde se conoce la presencia o ausencia de mineral, esto forma el set de datos para entrenar y evaluar el rendimiento de KNN, RNA y SVM en la tarea de prospección (1 para contenido mineral, 0 para ausencia de mineral), en el entrenamiento de los algoritmos de ML buscaron un balance en el número de muestras positivas y negativas para evitar sesgos, se utilizaron alrededor de 1500 datos (74%) para entrenar y 524 (26%) para testeo y una validación cruzada de 10 segmentos.

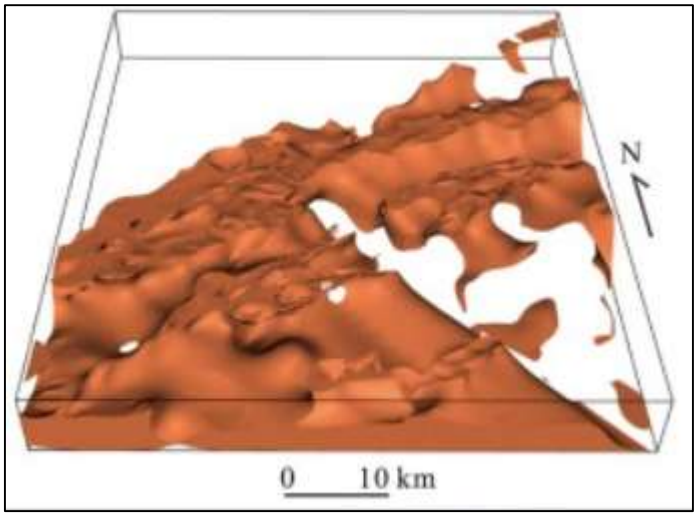


Figura 2.20. Ejemplo del resultado del modelado 3D representando la formación del Neoproterozoico. Imagen tomada de (Fu et al., 2021).

KNN presentó mejores resultados que RNA y SVM, aunque es muy posible que KNN tenga sobre ajuste, el problema principal es la escasez de datos en la zona, también lo difícil que es validar los resultados obtenidos, las zonas de mayor potencial provenientes de RNA Figura 2.21 (A) y SVM Figura 2.21 (B) tienen partes en común pero también difieren entre ellas en buena medida.

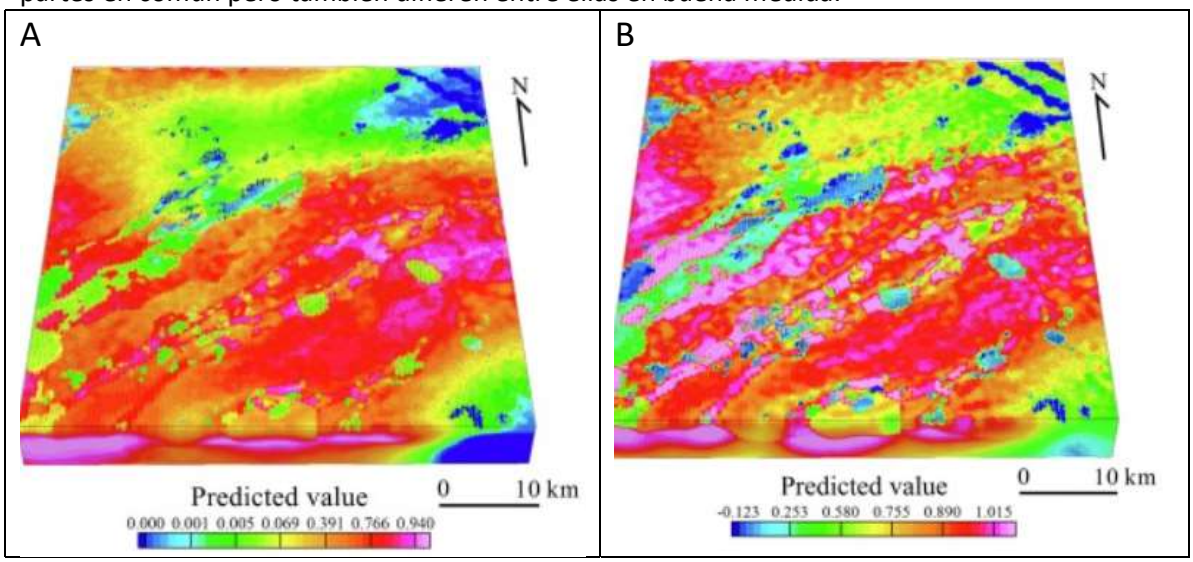


Figura 2.21. Resultado de los modelos desarrollados en (Fu et al., 2021). (A) Resultado de RNA. (B) Resultado de SVM. Imagen tomada de (Fu et al., 2021)

3 Base de datos

3.1 Descripción

La investigación se realizó en el departamento de Antioquia (Colombia), concretamente en el municipio de Caucasia, donde hay una fuerte tradición minera gracias a la influencia del río Cauca (García Jacome, 1978), segundo río más importante (de Colombia) y uno de los más ricos en contenido de oro del país. Dicha riqueza aurífera ha propiciado que en diversas zonas por las cuales alguna vez pasó el río Cauca, conocidas como paleocanales (Widera et al., 2019), se encuentre oro diseminado a distintas profundidades.

La extracción de oro en este tipo de zonas se conoce como minería aluvial (McGOWAN, 1996) y es una de las actividades económicas más importantes en la zona. Por tal motivo, empresas como WGM, han realizado importantes inversiones económicas en la tarea de exploración minera para determinar la viabilidad de un proyecto de explotación. De los diferentes estudios realizados en la zona de interés, dicha empresa proporcionó, para fines de esta investigación, los datos de las perforaciones realizadas y la información relacionada con la geología local del lugar. Estos datos, tomados a profundidad, corresponden a análisis geoquímico y descripciones litológicas elaboradas en campo por un experto.

3.1.1 Perforaciones

Los datos de estudio provienen de 147 perforaciones (Figura 3.1) que varían entre 8 y 40 m de profundidad, en las cuales se tomaron muestras en intervalos de 0.1 a 0.5 m, con un promedio de 58 datos por perforación y un total de 8642 descripciones litológicas. En dicho estudio se recuperaron testigos de perforación, a partir de los cuales se hizo la descripción de las características litológicas, la clasificación del tamaño de partícula del oro por tabla de colores y la cuantificación de la concentración de oro por cada rango de profundidad.

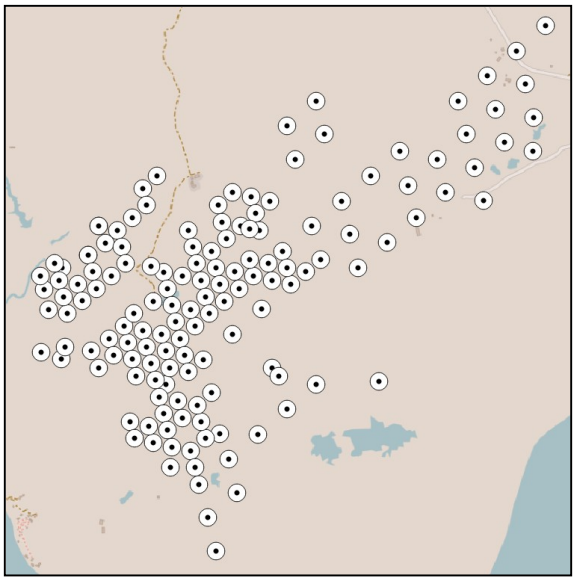


Figura 3.1. Zona de estudio, los puntos representan perforaciones realizadas en el proceso de exploración.

Cada muestra obtenida fue sometida a un análisis geoquímico y al análisis cualitativo de un geólogo experto. El análisis geoquímico consistió en concentrar las partículas de oro presentes en la muestra mediante gravimetría (Vega & Taboada, 2018), posteriormente las partículas obtenidas se clasificaron en seis grupos de acuerdo a su tamaño. A cada grupo (por convención se le denomina color al grupo) se le asignó un número como se muestra en la Tabla 2. Luego se procedió a contar las partículas en cada grupo para tenerunestimado del oro presente en cada perforación. El análisis cualitativo lo realizó un experto al momento de extraer la muestra, cuyo resultado fue una descripción litológica con base en la textura y el color del material encontrado en cada rango de profundidad. Tal es el caso de la identificación de arcillas, arenas, cuarzos, gravas, óxidos, entre otros.

Tabla 2. Convención de tamaño de partícula empleada en el análisis geoquímico.

Rango	> 2"	> 1"	> 3/8"	>1/8"	< 1/8" Arena	< 1/8" Limo
Colores	1	2	3	4	5	6
Malla #	5	10	18	35	80	200
Tamaños (mm)	4	2	1	0.5	0.177	0.074
Peso partícula oro del rango (mg)	647.4194	80.92742	10.11592	1.264491	0.05609	0.004099

Además de la geoquímica y de las descripciones litológicas, para cada pozo se registraron sus coordenadas, el ID y la profundidad de la muestra, dando como resultado la siguiente estructura de datos:

- **Descripción:** Representación textual realizada al tomar la muestra, dicha descripción depende del criterio del experto que la analiza.
- **Pozo:** El ID del pozo del que se tomó la muestra.
- **P1 – P6**: Corresponde a los 6 tamaños de partícula de oro provenientes del análisis geoquímico Tabla 2. Cada valor representa el número de partículas de un tamaño ***P*** que fueronencontradas en la muestra, donde P1 son las partículas de mayor tamaño y P6 las de menor tamaño.
- **Profundidad:** Distancia absoluta respecto a la superficie del punto donde fue tomada la muestra.
- **Coordenadas [X,Y]:** Coordenada del pozo según la proyección MAGNA-SIRGAS / Colombia Bogotá *zone.*

En la Tabla 3 se presenta un ejemplo de los datos mencionados, cabe resaltar que debido a que las descripciones fueron tomadas en campo, presentan errores ortográficos y de digitación, esto se aborda en la sección **3.1.2,** además con fines de protección de la información se ocultarán las coordenadas.

Tabla 3. Ejemplo de los datos suministrados por la empresa WGM. Solo se muestran 4 filas por simplicidad

Pozo	Profundidad	P1	P2	P3	P4	P5	P6	Descripción	X	Y
PM019	0.6	0	2	3	42	103	73	Materia orgánica, arcilla parda-roja, arenosa, pegajosa, gránulos de óxido de hierro. Compacto.	****	****
PM019	0.9	0	0	2	8	16	27	Arcilla parda-roja, arenosa, cuarzosa, pegajosa, gránulos de óxido de hierro. Compacto.	****	****
PM019	1.2	0	1	1	10	13	15	Arcilla amarilla, arenosa, cuarzosa, pegajosa,gránulos de óxido de hierro. Compacto.	****	****
PM019	1.5	0	0	0	6	5	8	Arcilla amarilla, arenosa, cuarzosa, pegajosa,gránulos de óxido de hierro. Grava fina. Compacto.	****	****

De los estudios mencionados con anterioridad, de tres salidas de campo y la información brindada por la empresa WGM, se clasificó el subsuelo en tres zonas como se muestra en la Figura 3.2. La primera es una capa rica en oro y de baja dureza que tiene en promedio 17 m de espesor y es la zona desde la cual se hace explotación minera en la región. La segunda es conocida por los lugareños como peña, es una capa muy dura, pobre en oro, se estima que posee entre 20 y 25 m de espesor. En la tercera y última zona se desconoce su espesor y contenido de oro, sin embargo, gracias a unas pocas perforaciones que lograron llegar hasta este punto se estima que puede ser rica en oro, pero por el momento es incierto y hacer explotación hasta esta profundidad se considera inviable económicamente.

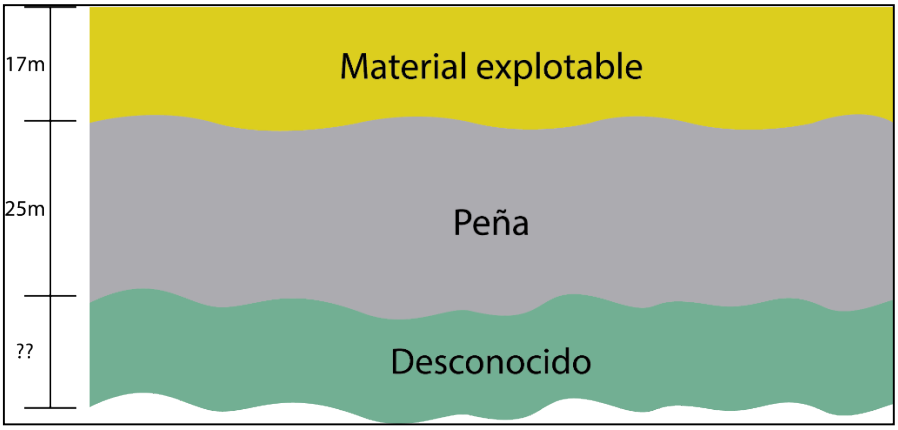


Figura 3.2. División del subsuelo de acuerdo con la información recopilada. En la denominada "Material explotable" se encuentran los estratos descritos en la Tabla 1, y la mayor parte de las descripciones litológicas en los registros de las 147 perforaciones.

3.1.2 Limpieza

Desde el año 2014 la empresa WGM ha extraído oro de la zona de estudio, basándose en gran medida en los datos provenientes de las perforaciones mencionadas. Sin embargo, la interpretación y análisis de éstos se ha hecho de forma manual, no han sometido los datos a un procesamiento computacional.

Al realizar en esta investigación, el primer acercamiento computacional al análisis de las 8642 descripciones litológicas pone en evidencia errores ortográficos y de digitación causados probablemente por las duras condiciones climáticas que por defecto representa el trabajo en campo y la manera convencional de registro de datos empleando bitácoras de campo y/o formatos no parametrizados. Para mencionar un ejemplo de algunos de estos hallazgos, se presentan errores como escribir "*mikasea*", "*mikacea*" y "*micácea*"; "*aren-a*", "*aren*" y "*arena*", los cuales representan 2246 descripciones diferentes. Luego de someter estas descripciones a una limpieza con expresiones regulares (Wang et al., 2019) y sustituciones manuales, la cantidad de descripciones diferentes se redujo a 1510. Posteriormente se eliminaron palabras vacías (*Stop words* (Sarica & Luo, 2021)) como "de", "muy", "en", entre otros; y luego se *tokenizaron* las palabras remanentes (P. Singh, 2019) (en la Tabla 4 se muestran algunos ejemplos). Esto dio como resultado 55 palabras que conforman el total de las descripciones. Esta representación en *tokens* es el punto de partida para el trabajo descrito en las secciones **3.2**, **3.4** y **4.2** donde se abordan los datos de diferentes formas para encontrar la ruta que mejor se adapte al mapeo prospectivo.

Tabla 4. Comparación entre las descripciones originales y el resultado de la limpieza.

Original	Resultado
Arcilla arenosa parda-roja-amarilla-gris, micásea, oxidada, grava fina, mediana y gruesa. Compacto.	arcilla, arenosa, parda, roja, amarilla, gris, micácea, oxidada, grava, fina, mediana, gruesa, compacto
Arena parda- gris, cuarzosa, mikasea, grava fina. Abundante arena negra.	arena, parda, gris, cuarzosa, micácea, grava, fina, abundante, arena, negra
Arena gris, cuarsosa, micásea,con gránulos de óxido de hierro, grava fina , mediana y gruesa. Compacto. Cementado.	arena, gris, cuarzosa, micácea, gránulos, óxido, hierro, grava, fina, mediana, gruesa, compacto, cementado

3.1.3 Modelo digital de elevación

Los paleocanales (Widera et al., 2019) son uno de los indicadores de depósitos de oro más importantes en la zona de estudio debido a que marcan las rutas por donde el río Cauca transitó y depositó sedimentos ricos en oro, con base en esto WGM ha orientado la exploración y explotación de oro. Partiendo de esta premisa se considera pertinente incluir esta información como uno de los parámetros de ingreso en la construcción de modelos del subsuelo, para lo cual se emplean los modelos digitales de elevación (DEM) (Wood, 1996). En este caso particular, el DEM codifica la topografía del terreno en pixeles con una resolución de 1 m^2 abarcando las 147 perforaciones en la zona de estudio. Un fragmento de este DEM se muestra en la *Figura 3.3*.

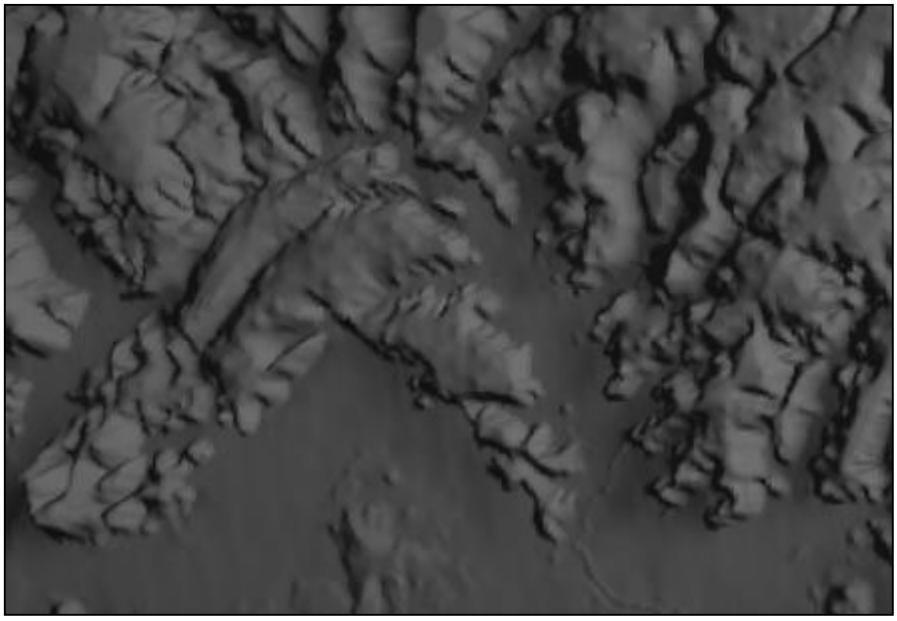


Figura 3.3. Fragmento del DEM representado en 3D, tiene una exageración vertical para mejorar el contraste en la visualización.

3.2 Clasificación manual

Partiendo de las 1510 descripciones de los pozos, producto del proceso de limpieza definido en la sección **3.1.2**, se procedió a realizar un análisis desde las geociencias, en el cual se identificaron entre las 55 palabras aquellas que dan la mayor información de la muestra en cuestión, es decir, qué palabras dentro de una descripción tienen mayor peso en la identidad de ésta; con el fin de poder identificar grupos litológicos asociados a la presencia o ausencia de oro.

Para llevar a cabo esta tarea, se realizó un análisis basado en el estudio previo del terreno, la existencia de paleocanales identificados en el DEM, tamaño de partícula identificadas en cada estrato y trazadores de oro como óxidos de hierro y arenas negras (D'yachkov et al., 2021). Como resultado se obtuvieron las siguientes 7 clases litológicas que agrupan las 8642 descripciones litológicas originales (sección **3.1.1**):

- **Materia orgánica:** Todas aquellas descripciones que contienen las palabras "materia" y "orgánica" fueron catalogadas en esta clase. Principalmente corresponde a las muestras tomadas cerca de la superficie o a paleocanales del río Cauca que todavía conservan materia orgánica.
- **Conglomerado:** En esta clase se agruparon los principales trazadores de oro como arenas negras, óxidos de hierro y conglomerados.
- **Arena arcillosa**: En esta clase se agruparon las descripciones relacionadas con arena fina a media de baja plasticidad.
- **Arcilla arenosa:** En esta clase se agruparon las descripciones relacionadas con arcilla arenosa, de plasticidad media, tamaño fino de cantos redondeados.
- **Arena:** En esta clase se agruparon las descripciones relacionadas con estratos con matriz de tamaño de arena con alto contenido de cuarzo.
- **Arcilla:** En esta clase se agruparon las descripciones relacionadas con los estratos de matriz arcillosa de alta plasticidad.

- **Limo**: En esta clase se agruparon las descripciones relacionadas con estratos de matriz limo arcillosa, de acuerdo con los hallazgos en los registros de perforación y el resultado del trabajo en campo de la Tabla 1.

3.3 Data augmentation

Con el fin de mejorar el proceso de entrenamiento de los modelos computacionales que serán descritos en la sección 4, se implementó *data augmentation* (Wen et al., 2021) de la siguiente manera. Se observó que las muestras de cada perforación no fueron tomadas exactamente al mismo intervalo de profundidad. Este delta de profundidad varía entre los 0.1 y los 0.5 m, comparando todas las perforaciones. Por tanto, se estandarizaron los datos para que el delta de profundidad fuera siempre de 0.1 m, esto se realizó asignando la misma descripción litológica a todos los elementosde un intervalo de profundidad, con lo cual se obtuvo la transformación que se muestra en la Figura 3.4, pasando de 8642 (total de descripciones litológicas) a 28652 datos.

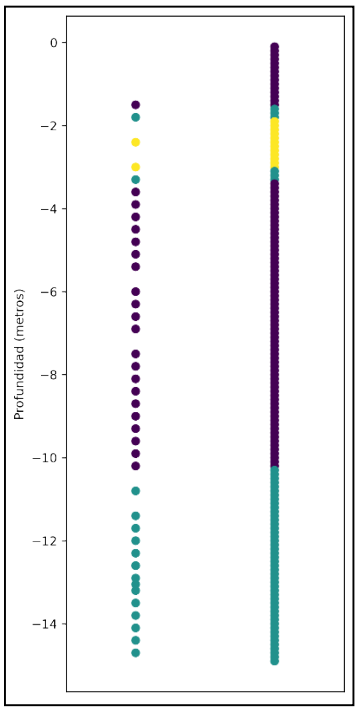


Figura 3.4. Ejemplo de la información de perforación original (izquierda) y el resultado de data augmentation (derecha).

Al unir el proceso de la clasificación manual y *data augmentation* se obtuvo como resultado un set de datos con 28652 muestras. La distribución por clase se encuentra en la Figura 3.5, donde se observa un desbalance de las etiquetas, “Arena arcillosa” contiene el 43 % de los datos seguida por“Arcilla arenosa” con 18 %, este es un reto que será tratado en próximas secciones. En la Figura 3.6 se representan los datos espacialmente después de aplicar la *data augmentation.* Debido a las diferencias en las profundidades máximas de cada perforación y a que la llamada peña se encuentra a una profundidad promedio de 17 m (según lo indicado en la sección **3.1.1**), se utilizaron solo los primeros 16 m de profundidad para la creación de los distintos modelos, puesto que las características físicas de la peña son muy diferentes a las del material en la capa suprayacente, además, son pocas las perforaciones que tienen datos relacionados con sus propiedades, por tal motivo, esta capa no se consideró en el modelamiento.

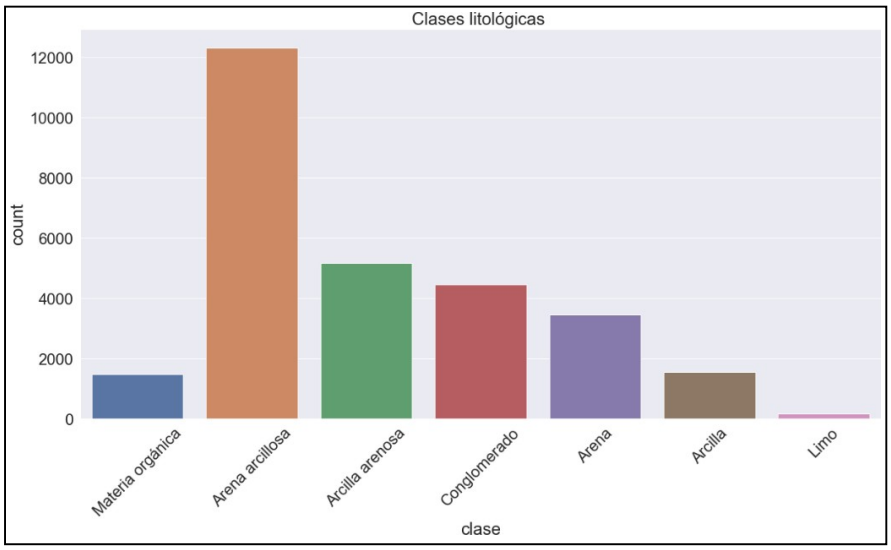


Figura 3.5. Distribución de los datos por clases después de la data augmentation y la clasificación manual.

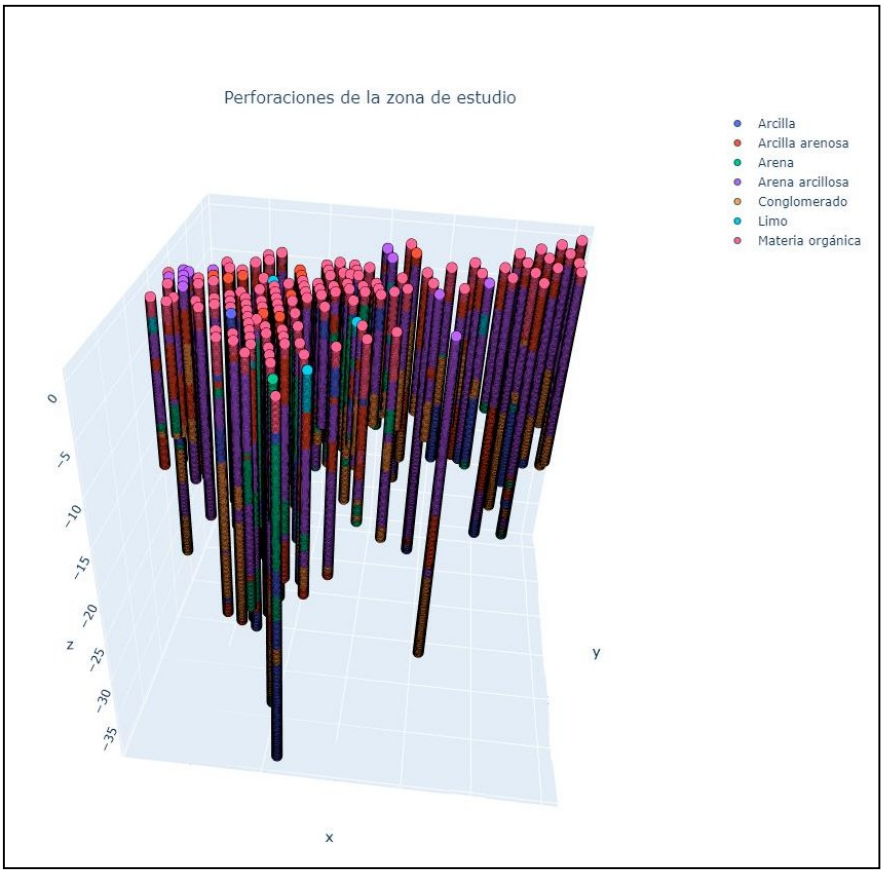


Figura 3.6. Representación 3D de las perforaciones en la zona de estudio, se han borrado las coordenadas X, Y para proteger la información. Representación de los datos espaciales después de la data augmentation.

3.4 Clasificación por tamaño de partícula

Los análisis cualitativos pueden llegar a ser muy subjetivos, dos expertos pueden analizar una misma muestra y/o descripción y dar conclusiones diferentes, estando ambas fuertemente justificadas. Esto puede ser ocasionado por muchos factores, como la falta de datos o la complejidad de los procesos geológicos (Parsa & Pour, 2021).

Por esto se explora una alternativa objetiva, donde los datos de las descripciones y del análisis geoquímico guíen la clasificación. Con este objetivo, se parte de la premisa de que todas las descripciones fueron tomadas en campo por un mismo experto y que en estas se codifica la naturaleza singular de la zona de estudio mediante patrones en las palabras utilizadas por el experto, patrones que se pueden decodificar utilizando los datos geoquímicos. Es decir que, aunque palabras como "óxido", "cuarzosa" o "azul" puedan tener múltiples interpretaciones en un contexto geocientífico, según el experto y/o la zona de estudio, si se correlaciona cada palabra con los datos cuantitativos de la geoquímica, se puede transformar cada palabra en un vector numérico y con esto obtener un vector que represente cada descripción litológica. El proceso para lograr esto se muestra a continuación.

3.4.1 Análisis semántico

Para encontrar la relación entre las descripciones y la presencia de partículas, se realizó un análisis estadístico tomando cada una de las 55 palabras que forman las 8642 descripciones asociadas al estudio geoquímico, para esto se calcularon las características presentes en la Tabla 5. El data set resultante se denomina *palabra-colores*, el cual contiene la relación estadística entre las palabras presentes en las descripciones y la cantidad de partículas de oro encontradas en las muestras según el tamaño. En la **¡Error! No se encuentra el origen de la referencia.** se encuentra la estructura del data set *palabra-colores.*

Tabla 5. Tabla de métricas utilizadas en la extracción de características

Características	Ecuación
Número de ocurrencias	$\sum Palabra$ *Ecuación 7. Cantidad de veces que aparece una palabra en el total de descripciones.*
% de ocurrencia	$\frac{Número\ de\ ocurrencias}{Número\ de\ descripciones} * 100$ *Ecuación 8. Frecuencia con la que aparece una palabra en el total de descripciones.*
% color n	$P(color\ n \mid palabra)$ *Ecuación 9. Es la probabilidad de que en la muestra exista una partícula de tamaño "n" dada una palabra.*
Promedio color n	$E(color\ n \mid palabra)$ *Ecuación 10. Es el promedio de partículas de tamaño "n" dada una palabra.*
Varianza color n	$V(color\ n \mid palabra)$ *Ecuación 11. Es la varianza del número de partículas de tamaño "n" dada una palabra*

Tabla 6. Estructura del data-set palabra-colores, va desde color 1 hasta color 6. Por temas de espacio solo se muestra el ejemplo para 5 palabras y del color 4.

Palabra	Número de ocurrencias	% de ocurrencia	% color 4	Promedio color 4	Varianza color 4
partículas	15	0.178	0	0	0
fina	6349	75.458	32.3	1,5	14.8
parda	6341	75.362	30.39	1,3	13.8
colas	354	4.207	26.84	1,0	4.9
azul	358	4.255	9.22	0,57	9.8

Con la extracción de características de la Tabla 6. se construyó la Figura 3.7, donde se muestra la probabilidad de ocurrencia de partículas de colores 3, 4, 5 y 6, para aquellas palabrasque aparecen por lo menos 100 veces (35 palabras). No se utilizó la probabilidad de color 1 y color 2 debido a que como se muestra en la Tabla 7, la cantidad de registros que contienen partículas de este tamaño no es significativo. En la Figura 3.7 se puede observar que las palabras “óxidos, hierro, gruesa, grava” se asocian con la presencia de partículas de oro, esto es buen indicio de la calidadde las muestras y del resultado, dado que los resultados concuerdan con los trazadores que se esperaría encontrar en un ambiente de oro aluvial (D’yachkov et al., 2021).

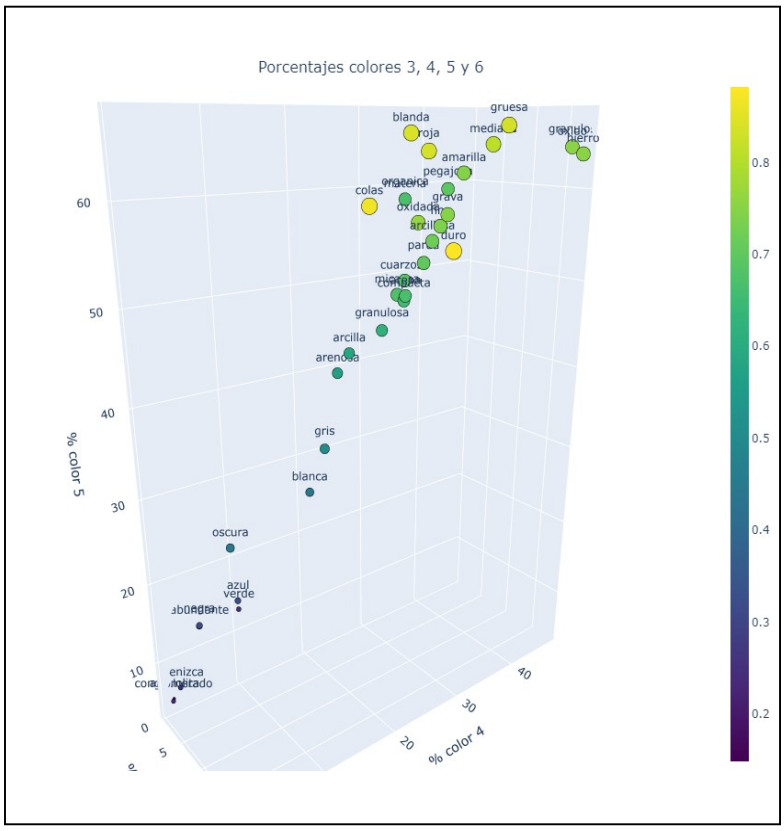


Figura 3.7. Relación entre % color con las palabras que aparecen más de 100 veces en las 8642 descripciones, donde las coordenadas X, Y, y Z están representadas por los % color 3, 4 y 5 respectivamente, el tamaño y coloración por el % color 6

Tabla 7. Cantidad de descripciones que tienen al menos una partícula de color "n". Se observa que de color 1, solo una descripción tiene registro, y de color 2 solo 71 de las 8642 tienen registro, esto representa menos del 1 % de las descripciones, por ello se descartaron para el análisis.

	Color 1	Color 2	Color 3	Color 4	Color 5	Color 6
Número de descripciones con almenos una partícula de color "n"	1	71	740	2176	3879	5257

La probabilidad de ocurrencia de una partícula de oro dada una palabra "% color" Tabla 8 se condensa en la Figura 3.8. En ésta se resume, mediante boxplot (Frigge et al., 1989), la distribución de las probabilidades de ocurrencia. Se puede observar que la data no presenta *outliers* (Hawkins, 1980) por lo que se puede afirmar que las diferentes palabras expresan de manera distribuida los rangos de probabilidad, donde se nota una tendencia creciente de la probabilidad de encontrar partículas cada vez más pequeñas. Esto resalta un fenómeno común en la zona de estudio: la presencia de partículas de oro diminutas diseminadas por toda área cercana al río Cauca. El inconveniente de este oro es que, al ser tan diminuto, la taza de recuperación es muy baja, es decir que mientras de 10 partículas de color 3 se logran captar 9, de 10 partículas de color 6 se pueden captar hasta 5. Sin embargo, se espera que, en muestras con gran cantidad de partículas pequeñas, la probabilidad y número de partículas de tamaños mayores aumente.

Tabla 8. Segmento de la tabla que resume la probabilidad de partículas dada una palabra, dicha tabla contiene las 35 palabras que aparecen más de 100 veces.

Palabra	**% color 1**	**% color 2**	**% color 3**	**% color 4**	**% color 5**	**% color 6**
pegajosa	0.00	1.32	11.26	33.11	60.26	69.54
oscura	0.00	0.00	2.59	8.62	22.84	44.83
arcillosa	0.02	1.10	10.89	31.31	55.03	72.06
gruesa	0.02	1.29	14.08	39.24	66.53	83.50
cuarzosa	0.02	0.87	9.54	28.52	51.00	68.80
blanda	0.00	1.69	5.62	32.58	65.73	84.27
oxido	0.00	2.33	17.05	46.51	64.34	75.19

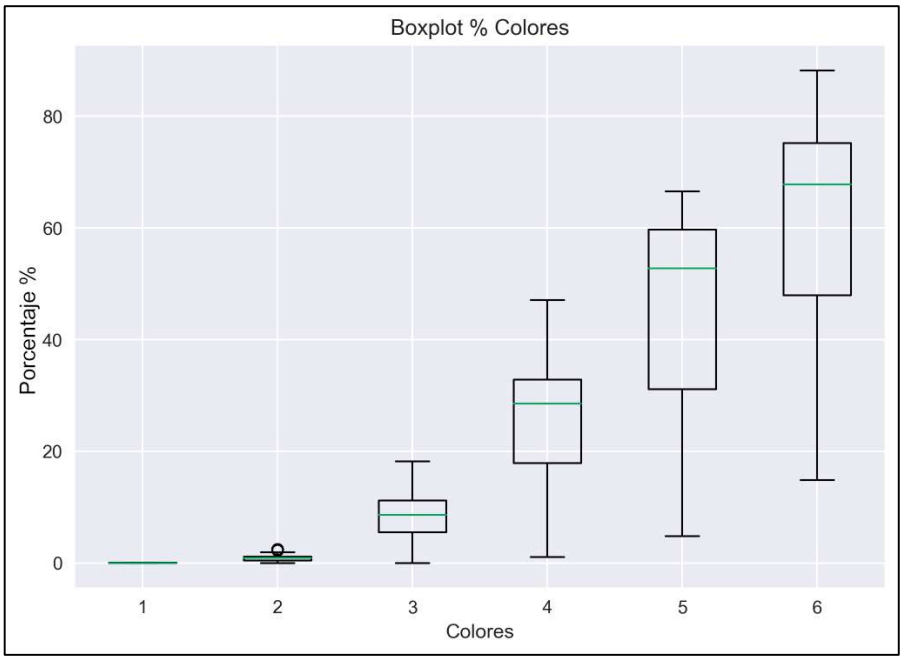


Figura 3.8. Boxplot que resume la distribución estadística de "% color" Tabla 4.

Mediante la técnica de matriz de correlaciones (Zhang et al., 2015) se obtuvo la Figura 3.9, en esta se observa que, en general, tanto los promedios como los porcentajes presentados en la **¡Error! No se encuentra el origen de la referencia.** se encuentran altamente relacionados entre sí, destacando el aumento de la correlación entre partículas de tamaño similar, probablemente por la génesis propia de estos depósitos aluviales. Esta alta correlación también puede estar influenciada por la forma en la que se extrajeron los "promedio color" y "% color #", dado que, para obtenerlos, se utilizaron descripciones completas, por lo que todas las palabras presentes en una descripción comparten información entre sí, lo que puede llegar a generar ruido a los modelos, principalmente por las palabras más frecuentes, sin embargo, este posible ruido se ve compensado por los buenos resultados expuestos hasta el momento en esta sección.

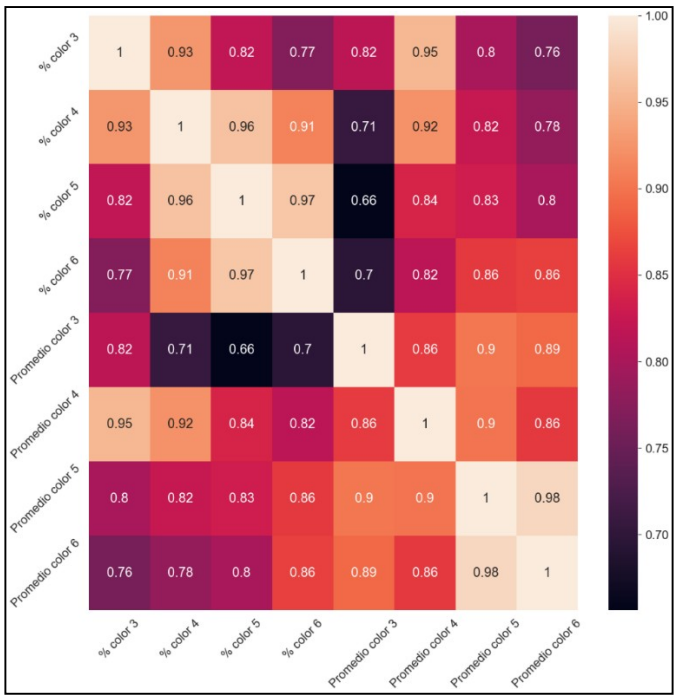


Figura 3.9. Matriz de correlación para el resultado de la sección 3.4.1

3.4.2 Reductores dimensionales PCA, t-SNE

Una de las ramas más importantes del Machine Learning es el aprendizaje no supervisado (Alloghani et al., 2020), esta busca encontrar patrones en los datos que permitan formar subgrupos de acuerdo con sus similitudes y diferencias. Entre los diversos algoritmos existentes para esta tarea, se encuentran los reductores dimensionales, tales como *análisis de componentes principales (PCA)* (Maćkiewicz & Ratajczak, 1993) y *la incrustación de vecinos estocásticos distribuidos en t* (t-SNE) (Linderman & Steinerberger, 2019). Estos algoritmos buscan reducir la dimensión de un set de datos, proyectando la distancia entre los datos en un espacio de menor dimensión, con el fin de encontrar patrones y/o formar subgrupos espacialmente separables, donde los elementos de cada subgrupo tienen características similares.

Con el objetivo de encontrar subgrupos intrínsecos en las descripciones litológicas, se utilizaron los datos provenientes de la sección **3.4.1,** para obtener una clasificación que no dependa del criterio de un experto. Se exploraron las reducciones dimensionales PCA y t-SNE en las 8 columnas donde se encuentran "% color" y "promedio color" para las partículas de color 3 al 6. Recordemos que estas columnas representan cada palabra, no obstante, el objetivo es representar cada descripción (conjunto de palabras), para lo cual se analizaron dos opciones, la suma y el promedio de vectores.

Partiendo de una descripción formada por "n" palabras, se representó cada una como un vector 1x8. Así, cada descripción se convirtió en una matriz nx8 (Tabla 9), luego se realizó la suma (Ecuación 12) o el promedio (Ecuación 13) de las filas según fuera el caso, de lo cual se obtuvo como resultado un vector 1x8. Estos vectores resultantes son la representación de las descripciones que son sometidos al proceso de reducción dimensional.

Tabla 9. Ejemplo de la representación de la descripción "arcilla parda roja".

Index	% color 3	Promedio color 3	% color 4	Promedio color 4	% color 5	Promedio color 5	% color 6	Promedio color 6
arcilla	6.22	0.12	23.23	0.83	43.23	2.97	58.56	5.21
parda	10.57	0.25	30.39	1.35	52.80	4.57	70.23	7.74
roja	7.13	0.15	33.93	1.13	63.92	3.84	83.76	7.19

$$suma(arcilla\ parda\ roja) = [(6.22 + 10.57 + 7.13), \dots, (5.21 + 7.74 + 7.19)] \quad (A)$$

$$suma(arcilla\ parda\ roja) = [23.92, 0.52, 87.55, 3.30, 159.94, 11.37, 212.55, 20.14] \quad (B)$$

Ecuación 12. Ejemplo del cálculo del vector suma proveniente de una descripción compuesta por 3 palabras

$$promedio(arcilla\ parda\ roja) = \frac{suma(arcilla\ parda\ roja)}{n} \quad (A)$$

$$promedio(arcilla\ parda\ roja) = \frac{[23.92, 0.52, 87.55, 3.30, 159.94, 11.37, 212.55, 20.14]}{3} \quad (B)$$

$$promedio(arcilla\ parda\ roja) = [7.97, 0.17, 29.18, 1.10, 53.31, 3.79, 70.85, 6.71] \quad (C)$$

Ecuación 13. Ejemplo del cálculo del vector promedio proveniente de una descripción compuesta por 3 palabras.

Tanto el vector de la suma como el del promedio tienen singularidades al momento de representar una descripción. El vector suma no penaliza las descripciones con palabras de poca relación con el oro, debido a que el aporte de cada palabra es independiente al resto, además favorece a las descripciones con muchas palabras. Por otro lado, el vector promedio penaliza las descripciones con palabras de poca relación con oro, ya que cada palabra aporta un porcentaje de la descripción, entonces al tener palabras de baja relación, el porcentaje de poca relación aumenta, además, esta misma razón evita que se favorezcan las descripciones de mayor tamaño.

Al ingresar los vectores resultantes de la suma a los reductores PCA y t-SNE, el resultado no fue satisfactorio, ya que no se logró formar grupos linealmente separables ni patrones que denotaran una relación que facilite una clasificación. En el caso de PCA, Figura 3.10, se obtuvo un único cúmulo de puntos, lo que puede significar que las descripciones tienen una alta similitud entre ellas. Probablemente sea debido a la forma en la que se extrajeron las características de cada palabra según los colores (tamaños de partícula). Por otro lado, para t-SNE se implementó *grid search* (Lerman, 1980), donde se aplicaron 200 combinaciones de hiperparámetros y se seleccionaron aquellos que formaban los mejores grupos, sin embargo, el resultado, Figura 3.11, no tiene patrones bien definidos. Los pocos cúmulos formados se deben a un sobre ajuste y no a patrones intrínsecos en los datos.

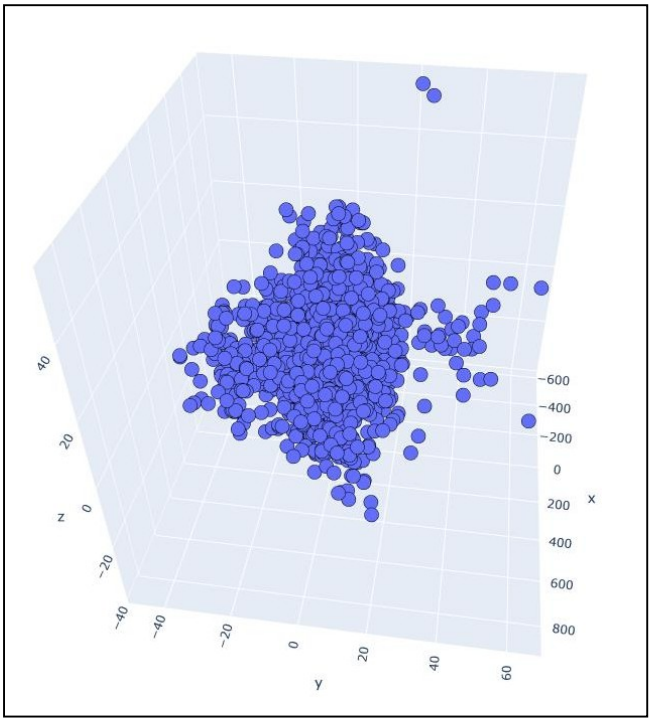


Figura 3.10. PCA para el vector proveniente de la suma, no se logran identificar patrones ni grupos linealmente separables.

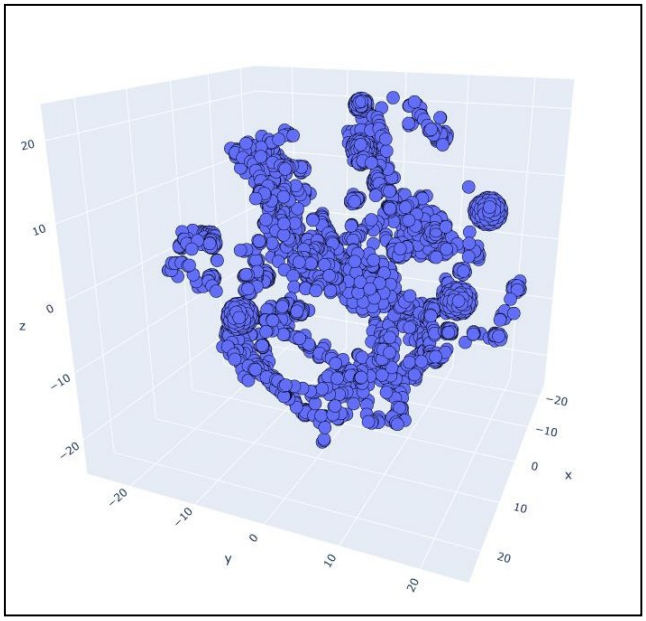


Figura 3.11. T-SNE para el vector proveniente de la suma, los pocos grupos y patrones formados se lograron después de un sobre ajuste, no son útiles para el objetivo de clasificación.

Para el vector proveniente del promedio sucedió igual que para el de la suma. El resultado de PCA, Figura 3.12, fue un único cúmulo sin patrones evidentes y el de t-SNE, Figura 3.13, los únicos grupos (poco delimitados) que se pudieron obtener fue debido a llevar el modelo a un sobreajuste con *grid search*.

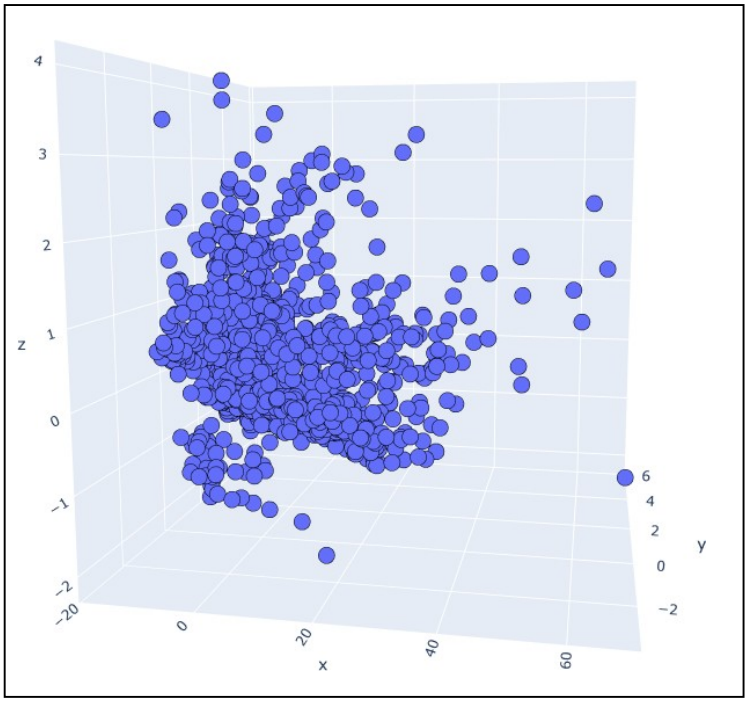


Figura 3.12. PCA para el vector proveniente del promedio, no se logran identificar patrones ni grupos linealmente separables.

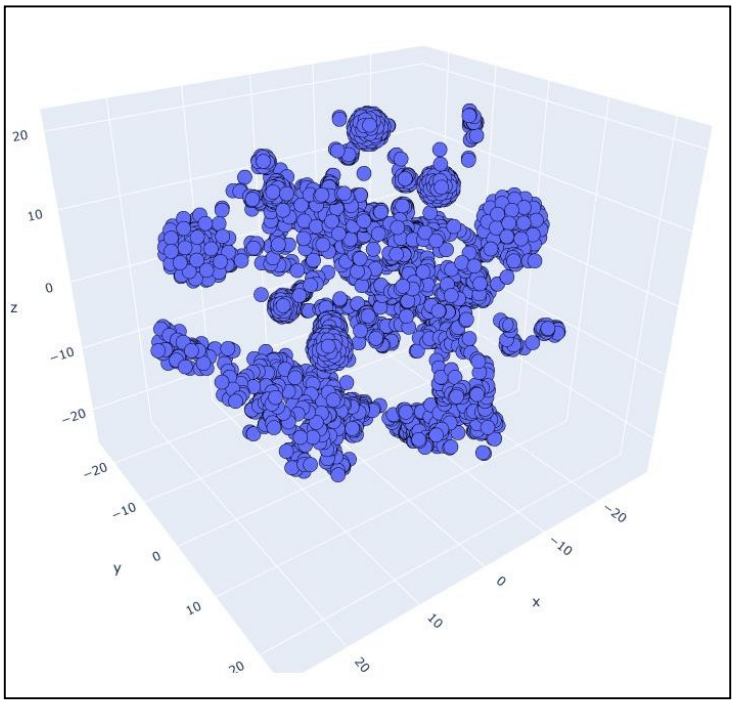


Figura 3.13. T-SNE para el vector proveniente del promedio, los pocos grupos y patrones formados se lograron después de un sobre ajuste, no son útiles para el objetivo de clasificación.

3.4.3 K-means

Debido a que el proceso de reductores dimensionales, sección **3.4.2,** no tuvo un resultado con el que se pudieran formar una nueva clasificación no supervisada (clustering), se abordó la tarea desde otro enfoque para representar cada palabra. Para esto se utilizaron solo 4 columnas "Promedio color 3 al 6". Por lo tanto, cada descripción de "n" palabras es una matriz nx4 (Tabla 10). Con esta matriz se calculó la suma de las filas (Ecuación 14) dando como resultado un vector 1x4.

Tabla 10. Ejemplo de la representación de la descripción "arcilla parda roja".

Index	Promedio color 3	Promedio color 4	Promedio color 5	Promedio color 6
arcilla	0.12	0.83	2.97	5.21
parda	0.25	1.35	4.57	7.74
roja	0.15	1.13	3.84	7.19

$$suma(arcilla\ parda\ roja) = [(0.12 + 0.25 + 0.15), \dots, (5.21 + 7.74 + 7.19)] \quad (A)$$

$$suma(arcilla\ parda\ roja) = [0.52, 3.31, 11.38, 20.14] \quad (B)$$

Ecuación 14. Ejemplo del cálculo del vector suma proveniente de una descripción compuesta por 3 palabras

En la Figura 3.14 se observa la representación espacial de las descripciones resultantes. Es evidente que los datos se distribuyen en una diagonal imaginaria que atraviesa al cubo 3D, permitiendo agrupar las descripciones según su posición en el espacio mediante el método de

clustering k-means (Hamerly & Elkan, 2003), concretamente en 3 clases que representan baja, media y alta probabilidad de oro, Figura 3.15. En la Figura 3.16 se presenta la distribución por clase. El problema del desbalance presente en la clasificación manual (sección **3.2**) ya no es tan marcado por lo que esta ruta es una buena alternativa a ese problema.

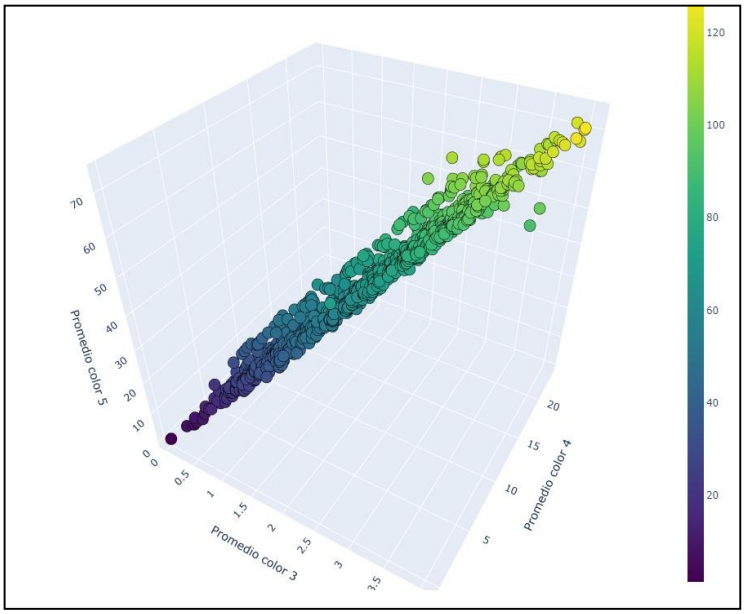


Figura 3.14. Representación de las descripciones provenientes de aplicar la Ecuación 14 a las columnas de la Tabla 10. La coloración es dada por "Promedio color 6"

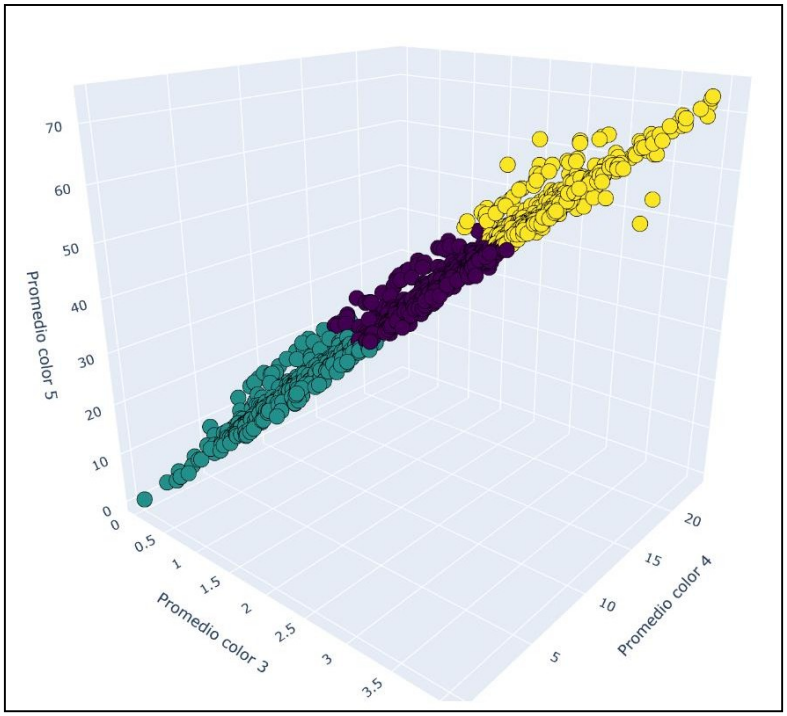


Figura 3.15. Resultado de aplicar k-mean a las descripciones mostradas en la Figura 3.14, donde amarillo es alta probabilidad de oro, morado es media y verde es baja

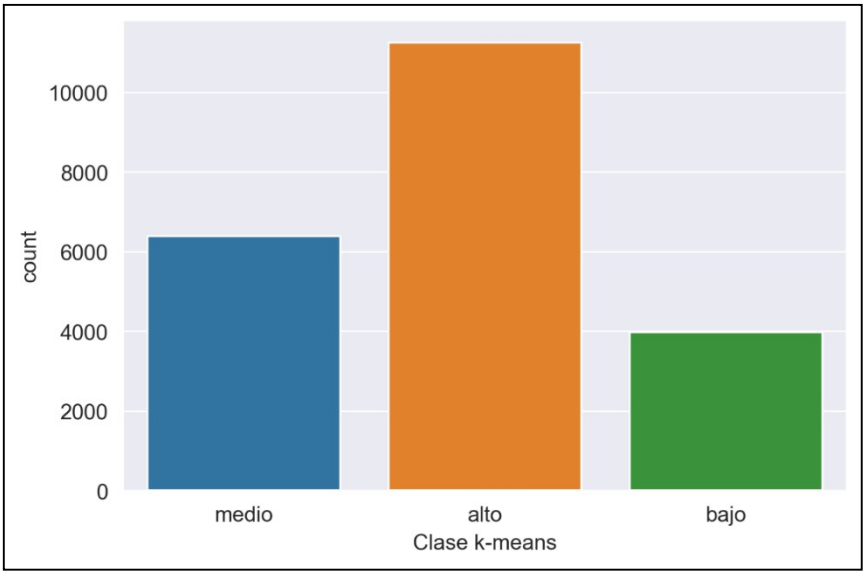


Figura 3.16. Cantidad de datos por etiqueta, "medio" 29 %, "alto" 52 % y bajo 18 %.

3.4.4 Norma vectorial

Hasta el momento las transformaciones y el tratamiento de los datos se han enfocado en llevar a los datos a un estado donde se puedan utilizar en el entrenamiento de modelos de clasificación debido a que es el enfoque principal de la investigación. No obstante, se reconoce la importancia de explorar modelos regresivos, ya que los modelos de clasificación, por su naturaleza discreta (Strum & Kirk, 1988) pierden información.

Para esto, se partió del resultado parcial de la sección **3.4.3,** concretamente de la representación delas descripciones de la Figura 3.14, donde cada descripción es un vector 1x4. A estos vectores se les calculó la norma vectorial para conocer su longitud, que es un escalar, el cual representa su asociación con el oro, entre mayor sea el valor, mayor será la relación con el oro. Alcombinar esto con la información de las perforaciones se obtuvo la tupla input-target para modelosde regresión. La distribución espacial de los datos en la zona de estudio se muestra en la Figura 3.17.

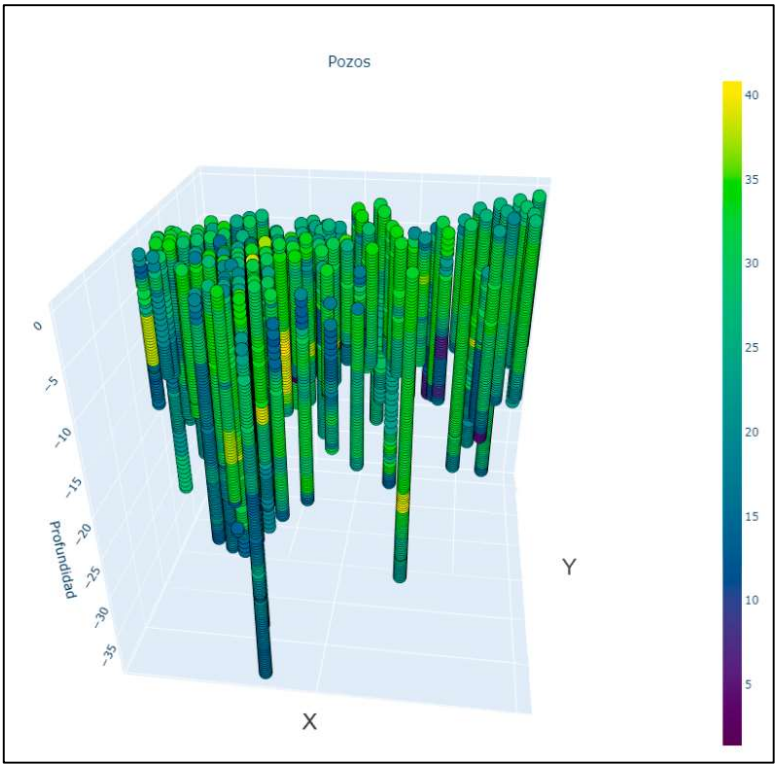


Figura 3.17. Representación 3D continua de las perforaciones en la zona de estudio. Se borraron las coordenadas X, Ypor cuestiones de protección de la información. La coloración representa la relación con el oro.

4 Modelos computacionales

4.1 Red convolucional híbrida

En la minería aluvial es importante la topografía del terreno. En el caso concreto de la zona de estudio, es relevante ubicar paleocanales (Widera et al., 2019), los cuales son las cuencas residuales por donde el río Cauca transitó en tiempos geológicos pasados. Estos se pueden ubicar de forma visual o mediante modelos digitales de elevación (DEMs) (Wood, 1996). Hasta el 2022 la exploración y explotación minera de la zona se ha enfocado en la extracción de material utilizando dichos paleocanales como trazadores de oro, lo cual ha dado resultados favorables. Por esta razón, para este tipo de minería y en este trabajo es pertinente ingresar información del terreno al modelo de *deep learning*. Para lograr esto se propuso un esquema de red neuronal artificial dividida en 4 etapas, como se muestra en la Figura 4.1.

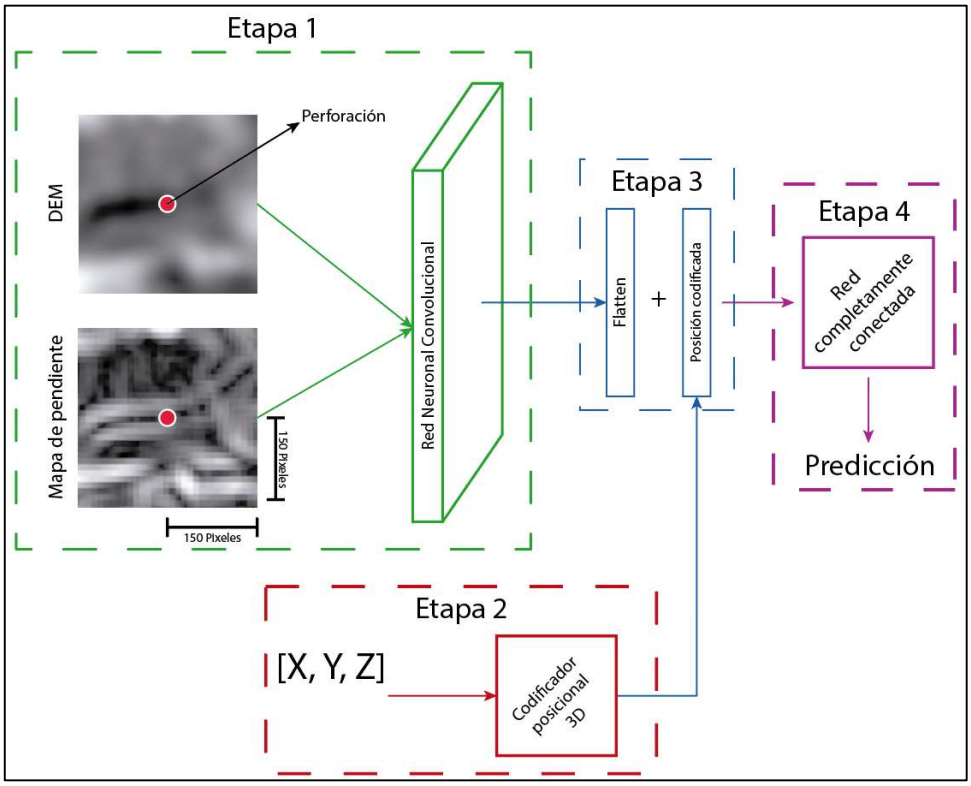


Figura 4.1. Esquema de arquitectura de RNA para interpolación espacial dividida en 4 etapas.

4.1.1 Etapa 1 (segmento convolucional)

En la etapa 1 (segmento convolucional), se utilizó el modelo digital de elevación (DEM) de la zona de estudio, en el que cada píxel equivale a un fragmento del terreno de 1 m^2. Con este DEM se calculó la pendiente del terreno mediante el *software QGis*. Esta representación resultante permitió aumentar la información de entrada para la RNA y reducir el número de iteraciones de entrenamiento necesarias para sintonizar el modelo, con las dos representaciones topográficas (DEM y pendiente) del área que abarcan todas las perforaciones. Se extrajeron recuadros de 300 x 300 pixeles alrededor de cada perforación, Figura 4.2. Los dos recuadros por perforación se unieron para formar una imagen de dos canales, la cuales la entrada de la etapa convolucional de la RNA (Albawi et al., 2017).

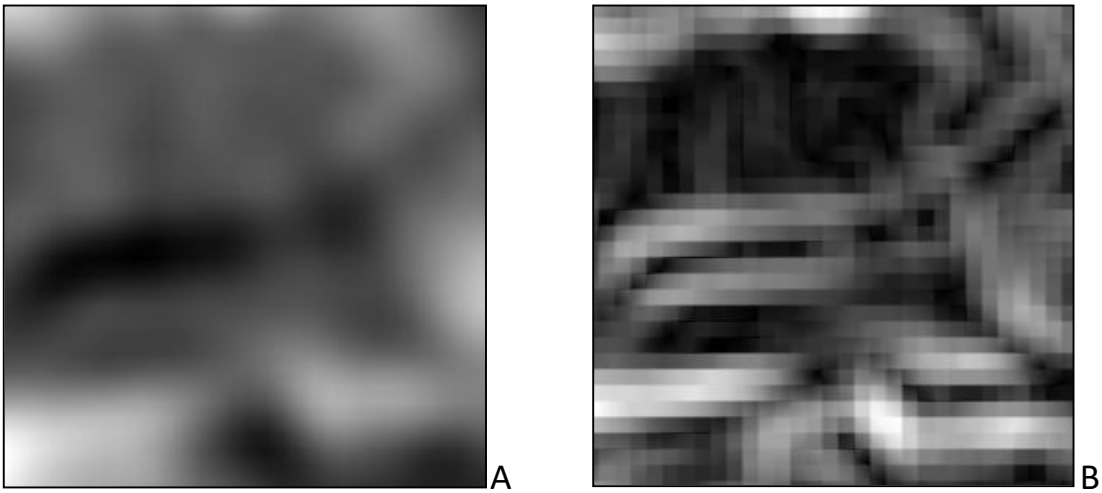


Figura 4.2. (A) Modelo digital de elevación alrededor de una perforación. (B) transformación a pendiente de (A)

4.1.2 Etapa 2 (codificador posicional)

La segunda etapa hereda un concepto de los *transformers*, denominado codificación posicional (Vaswani et al., 2017), el cual transforma una posición en el espacio en un vector, mediante una combinación de senos, cosenos y una representación binaria de la frecuencia (Kernes, 2021), lo que permite, en este caso, transformar un vector de 3 posiciones (X, Y, Z) en una señal (vector) codificada de N posiciones, ver Figura 4.3. Este codificador se implementó con 2 objetivos: el primero es evitar que los 3 valores de las coordenadas (X, Y, Z) sean opacados por el vector resultante de la convolución y el segundo es modificar el resultado de la convolución de la etapa 1, debido a que todos los datos de una misma perforación comparten el mismo DEM, mapa de pendiente, coordenada X y coordenada Y. Por lo tanto, es necesario aumentar la variabilidad de esta información para que el modelo pueda detectar los patrones presentes en los datos.

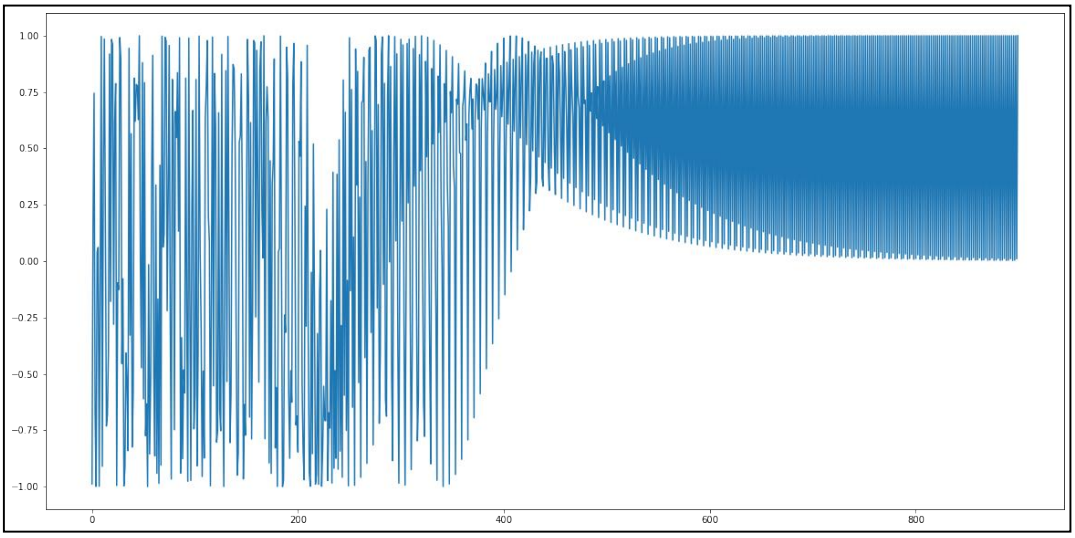


Figura 4.3. Resultado de codificar una coordenada X, Y, Z en un codificador posicional de 900 posiciones

4.1.3 Etapa 3 (suma vectorial)

En la etapa 3 se continuó con los preceptos de los Transformers (Vaswani et al., 2017). Originalmente, en procesamiento de lenguaje natural (NLP) (Qiu et al., 2020), después de obtener el vector proveniente de la codificación posicional se suma con el vector de embedding (Pennington et al., 2014). En este caso, no existe un embedding, existe la matriz resultante de la etapa convolucional (Etapa 1) a la que se le aplicó un proceso de *flattening* (Lin et al., 2020), y se obtuvo un vector de N posiciones. Este vector debe coincidir con el vector del codificador posicional (Etapa 2) para poderlos sumar y así obtener la nueva representación de los datos de entrada.

4.1.4 Etapa 4 (predicción)

La cuarta etapa recibe la suma de los dos vectores de la Etapa 3 para luego llevarlos a una red *Fully connected* (Basha et al., 2020), la cual predice la clase o valor del punto en el espacio de la entrada. Dependiendo de si el set de datos es *"***Clasificación manual***"* o ***"*K-means*"*,* la salida de la red neuronal será un vector de 7 o 3 posiciones respectivamente. En el caso de el set de datos *"***Normavectorial***"* la salida será un único valor debido a que es una tarea de regresión. En la Figura 4.4 se muestra la estructura interna completa de la red neuronal propuesta para el data set *"***Clasificaciónmanual***"*.

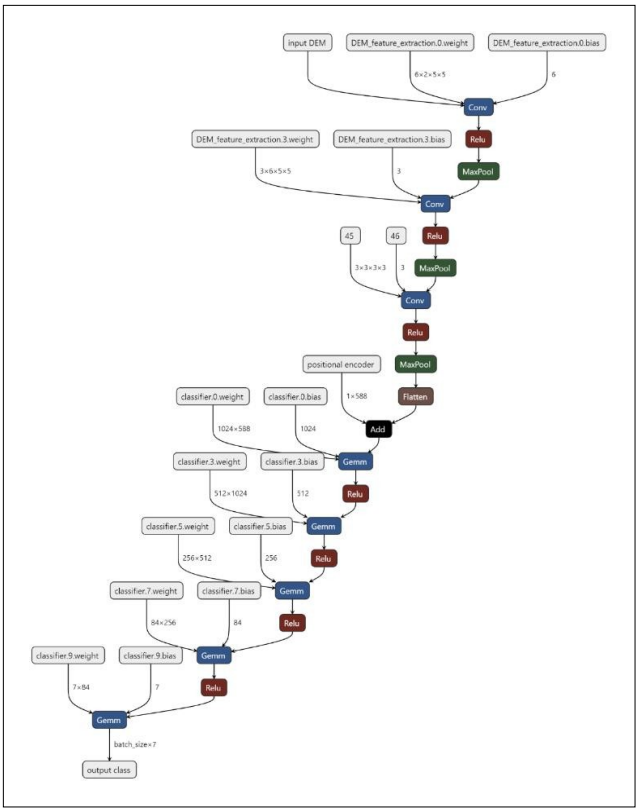


Figura 4.4. Arquitectura de la red neuronal propuesta implementada para el data set "3.2 Clasificación manual"

4.1.5 Aprendizaje por transferencia

En esta ruta experimental se exploraron dos variantes para la clasificación, en la primera se entrenó la red neuronal desde cero con la arquitectura propuesta en la Figura 4.4. En la segunda se exploró el paradigma del transfer learning (Agarwal et al., 2021). Concretamente, en la Etapa 1 de la RNA se utilizó la red preentrenada resnet18 (He et al., 2015) para extraer las características del DEM y del mapa de pendientes, después de esto, se continuó según lo propuesto, esto se puede observar en el esquema de la variante en la Figura 4.5.

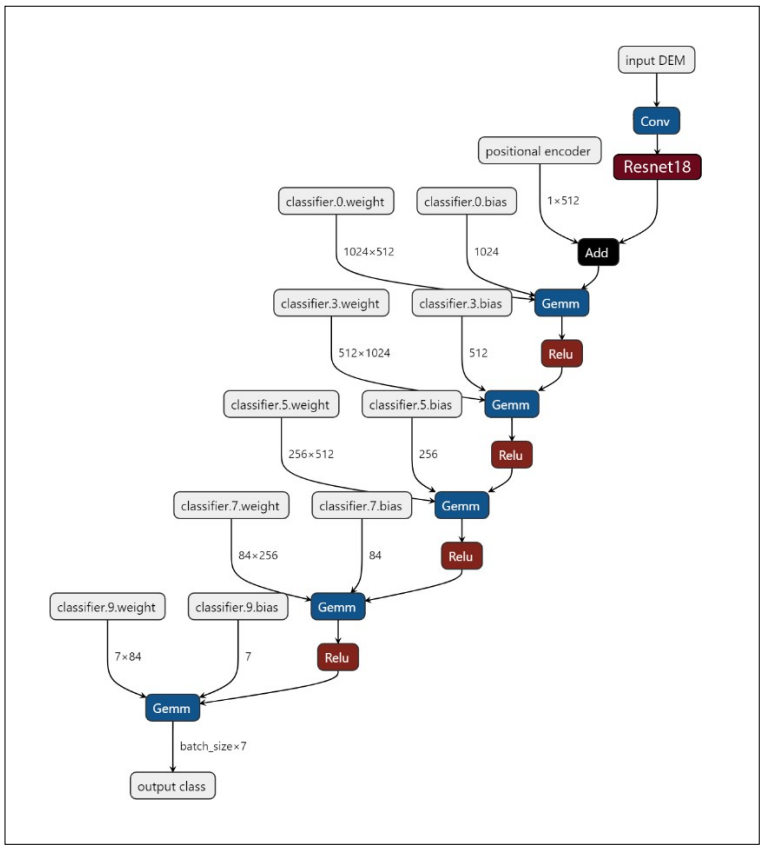


Figura 4.5. Esquema de la variante de red neuronal propuesta en la sección 4.1, en esta variante se implementa una red preentrenada llamada Resnet18 para extraer las características de los modelos digitales de elevación y del mapa de pendientes

4.2 Mapeo litológico 3D mediante procesamiento de leguaje natural

Para realizar uno de los modelos 3D de la zona se tuvo en cuenta el trabajo realizado por (Fuentes et al., 2020), en el cual se presenta una ruta metodológica para crear modelos litológicos tridimensionales partiendo de perforaciones y descripciones a profundidad utilizando procesamiento de lenguaje natural (NLP) e interpolación lineal. A continuación, se presenta una breve descripción de dicha metodología aplicada al sitio de interés en Caucasia.

4.2.1 Embedding

Se partió de las descripciones verbales de cada pozo. Cada descripción pertenece a un punto de la zona de estudio y está compuesta por varias palabras que representan lo que el geólogo notó en la muestra al momento de examinarla en campo según sus conocimientos. A estas descripciones se les realizó una limpieza para corregir errores ortográficos, de tipeo y de formato, y se obtuvo un set

de datos estandarizado (Sección **3.1.2**).

El set de datos estandarizado se ingresa a un procesamiento de lenguaje natural para extraer el ***lema*** de cada palabra (El lema es la forma que se acepta como representante de todas las formas de una misma palabra, ejemplo: el ***lema*** de *cantando* es *cantar*) (Khyani & B S, 2021) y luego traducirlas al inglés, dando como resultado la Ecuación 15

$$[organic, matter, red, \dots]$$

Ecuación 15. Fragmento de una descripción de pozo después de pasar por el primer procesamiento de lenguaje natural.

Para representar cada palabra y posteriormente cada descripción de una forma numérica, se hace uso de una herramienta del *deep learning* y del procesamiento de lenguaje natural llamada embedding o incrustación. Con este elemento se puede obtener la relación espacial entre miles de palabras, asignando una coordenada (vector) en un espacio de N dimensiones a cada una, en este caso se emplea un *GloVe embedding* (Pennington et al., 2014) que fue entrenado por (Padarian & Fuentes, 2019) con más de 300000 artículos científicos de geociencias en lengua inglesa, transformando cada palabra como se observa en la Ecuación 16

$$organic = [0.42, 0.63, \dots, N]$$

Ecuación 16. Representación vectorial de una palabra dentro del embedding. Las palabras se tradujeron al inglés para poder utilizar el embedding preentrenado.

Con cada palabra simbolizada como un vector numérico, se puede representar cada descripción litológica en un vector numérico. Para esto, se promedian los vectores de cada palabra que componen la descripción, y se obtiene un único vector que sintetiza la descripción completa (Ecuación 17).

$$\begin{gathered} organic = [0.42, 0.63, \dots, N_1] \\ matter = [0.25, 0.5, \dots, N_2] \\ red = [0.95, 0.34, \dots, N_3] \\ \hline Mean = [\frac{1.62}{3}, \frac{1.47}{3}, \dots, \frac{N_1 + N_2 + N_3}{3}] \end{gathered}$$

Ecuación 17. Promedio vectorial de una descripción, el color indica cuales columnas se promedian entre sí y el resultado.

4.2.2 Interpolación

Al combinar el resultado de la Ecuación 17 con las coordenadas X, Y, Z de cada descripción, se obtuvieron valores vectoriales asociados a un punto del espacio tridimensional, con el cual se realizó una interpolación lineal (Weisberg, 2005) con lo cual se obtuvo un estimado de los valores desconocidos delespacio. El resultado es un vector que representa dicho dato desconocido Figura 4.6.

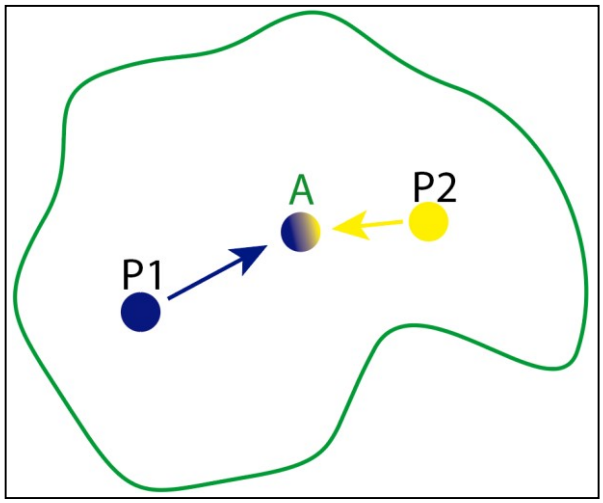


Figura 4.6. Ilustración de un proceso de interpolación lineal, donde P1 y P2 son vectores conocidos y "A" es el vector resultante de la interpolación del punto desconocido.

4.2.3 Clasificación

Para poder transformar los vectores numéricos que representan las descripciones litológicas (Ecuación 17 y Figura 4.6) en etiquetas, se entrenó una red neuronal donde las clases provenientes de *"***Clasificación manual***"* o *"***K-means***"*, según sea el caso, son el target y los vectores provenientes de la Ecuación 17 son el input. Ya con la red neuronal entrenada, se le ingresaron los valores interpolados a los que hace referencia la Figura 4.6 para determinar a qué etiqueta pertenecía el punto del espacio interpolado. Este proceso se resume en la Figura 4.7.

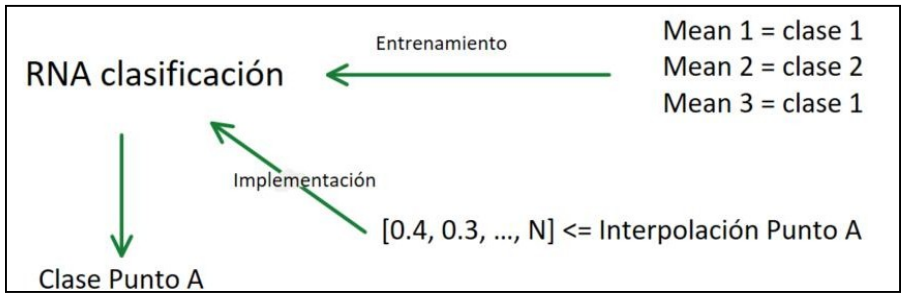


Figura 4.7. Esquema de la red neuronal artificial (RNA), entrenada con las etiquetas y los promedios vectoriales provenientes de las descripciones. La red entrenada se utiliza para clasificar los resultados de las interpolaciones espaciales.

4.3 Regresión lineal para clasificación

Para tener un punto de comparación del rendimiento de las metodologías de clasificación de las secciones **4.1** y **4.2**, se partió de las etiquetas provenientes de la clasificación manual y de K-means (secciones **3.2** y **3.4.3,** respectivamente), estas se codificaron en un vector OneHotEncoder, lo que dio como resultado la representación de la Figura 4.8, Estos vectores, junto a las coordenadas espaciales asociadas a cada muestra, se utilizaron para entrenar un modelo de interpolación lineal. Al modelo entrenado se le ingresó una coordenada de un punto en el espacio desconocido y devolvió un vector de N posiciones, donde la posición con el valor más alto representa la clase predicha, Figura 4.9.

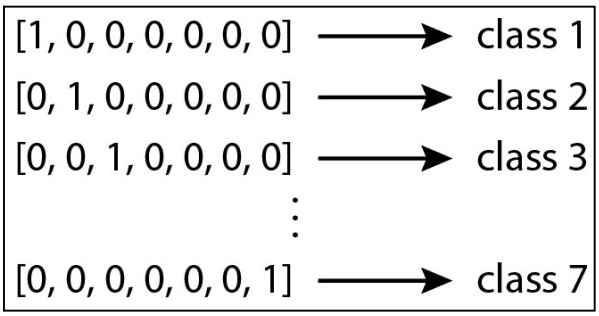


Figura 4.8. Resultado de la codificación OneHot para 7 clases.

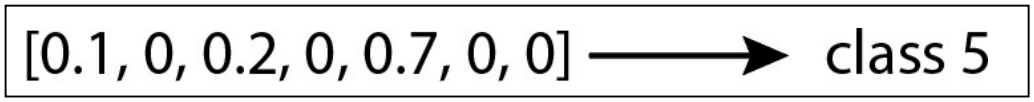


Figura 4.9. Ejemplo de un resultado de la interpolación lineal para 7 clases, donde la clase predominante es la número 5.

4.4 Interpolación Kriging

Kriging (Cressie, 1990) es uno de los métodos de interpolación espacial más utilizados en geociencias, hace parte de los llamados procesos gaussianos para regresión (Kleijnen, 2009; Schulz et al., 2018). Con este se busca expresar la variación espacial mediante variogramas y así reducir los errores de predicción (OLIVER & WEBSTER, 1990). Al ser un método tan ampliamente utilizado en geociencias, es importante comparar su rendimiento respecto a los modelos propuestos para la zona de estudio de minería aluvial. Debido a que Kriging es un método de regresión, solo se puede comparar con "**Red convolucional híbrida**" en su configuración de regresión con los datos resultantes de "**Norma vectorial**".

4.4.1 Configuración de los datos

Para el modelo de regresión mediante Kriging se partió de los datos provenientes de la sección "**3.4.4 Norma vectorial**" (Figura 4.10), los cuales representan la distribución de oro en un espacio 3D después de relacionar las descripciones litológicas con la geoquímica (para más detalles revisar la sección 3.4). Debido a que las principales aplicaciones de Kriging en geociencias es en espacios 2D (Dramsch, 2020), se promediaron los datos de las perforaciones hasta una profundidad de 16 m, Figura 4.10. Como resultado se obtuvo la representación 2D de la Figura 4.11(A). En la Figura 4.11(B) se observa que los datos 2D parecen seguir una distribución gaussiana (Goodman, 1963). Al realizar la prueba de normalidad de D'Agostino y Pearson (R. B. D'AGOSTINO, 1971; R. D'AGOSTINO & PEARSON, 1973) la distribución se cataloga como gaussiana con un p-value de 0.04 (Sullivan & Feinn, 2012), asimetría de -0.515 y una curtosis de 0.027.

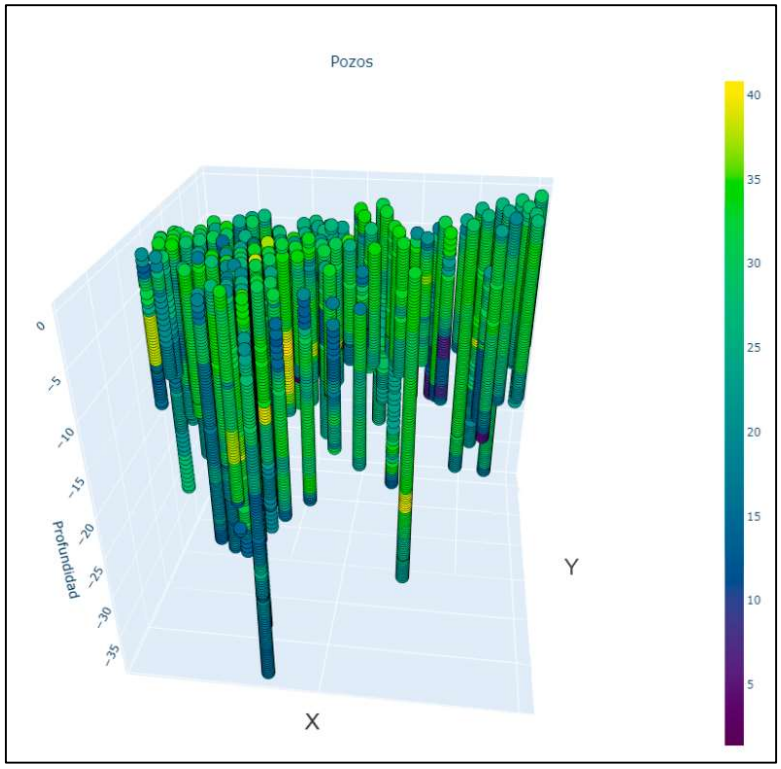


*Figura 4.10. Representación 3D proveniente de la sección **3.4.4**, se han borrado las coordenadas X, Y por cuestiones de protección de la información, la coloración representa la relación con el oro.*

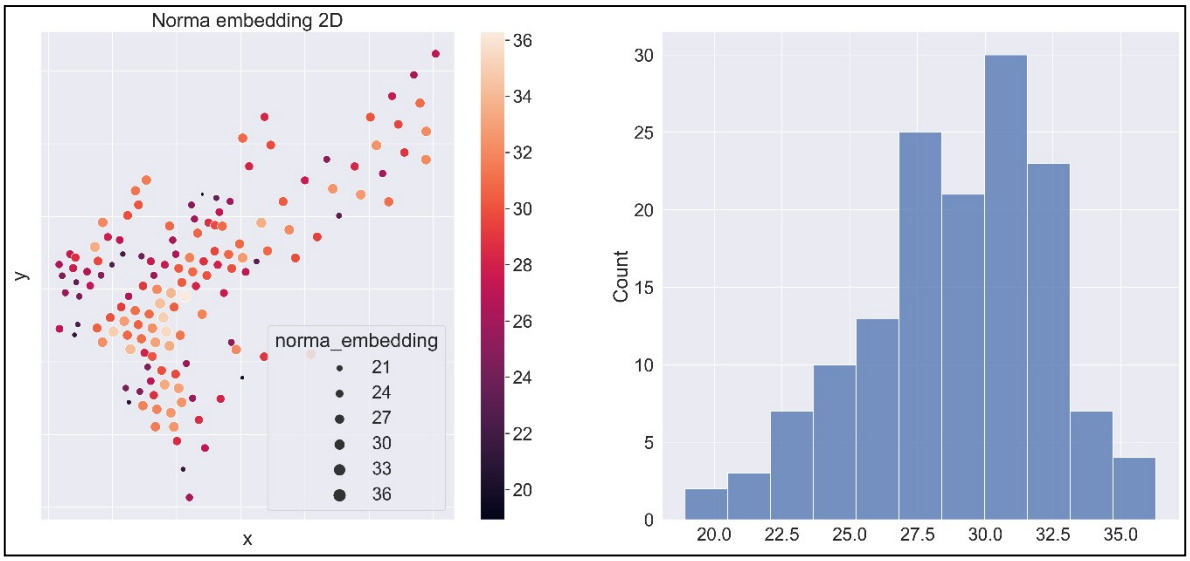


Figura 4.11. (Izquierda, A) Resultado de promediar los datos de las perforaciones hasta una profundidad de 16 m de la Figura 4.10 para llevar los datos a un espacio 2D. (Derecha, B) Distribución de los datos 2D mediate un histograma.

4.4.2 Modelo GPytorch

El modelo de regresión gaussiana o Kriging (Kleijnen, 2009; Schulz et al., 2018) fue implementado en el *GPytorch framework* (Gardner et al., 2021). Debido a que en la sección **4.4.1** se confirmó que

la data 2D sigue una distribución gaussiana, se le asignó al modelo una probabilidad (*likelihood*) gaussiana. Al sintonizar el modelo de forma heurística, el mejor rendimiento lo presentó con una media constante y un kernel de covarianza de función de base radial (RBF) (Schulz et al., 2018). Debido a los resultados que se presentan en la sección "**5.3.2** Interpolación Kriging", se descartó la implementación de este modelo en un espacio 3D por su bajo desempeño.

5 Resultados

A continuación, se describen los resultados de las diferentes rutas empleadas para desarrollar modelos de regresión y clasificación enfocados en el entorno de minería aluvial descrito anteriormente. Para los diferentes modelos siempre se utilizó la misma configuración de los datos de perforaciones para entrenamiento y validación, Figura 5.1, con lo cual se evitan sesgos provenientes de alguna configuración que favorezca a algún modelo. Para llegar a esta partición, se hicieron varias selecciones aleatorias y se escogió la que distribuyera los puntos de validación uniformemente por toda la zona, ya que en algunos casos la zona noreste, que tiene una densidad de pozos menor, no tenía puntos de validación.

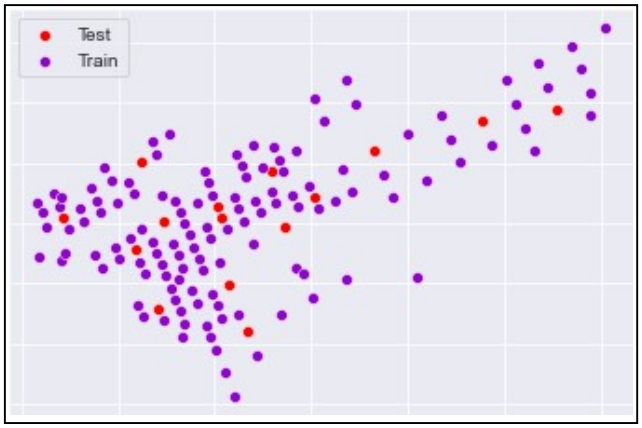


Figura 5.1. Distribución espacial de las perforaciones divididas en entrenamiento (morado) y validación (rojo)

5.1 Data set "clasificación manual"

Al momento de analizar el rendimiento de las metodologías aplicadas a los datos proveniente de la sección **3.2**, se debe tener en cuenta el desbalance. En la Tabla 11 se observa que, aunque existe un desbalance entre la proporción de las clases, este es proporcional entre los datos de entrenamiento y validación. Cabe resaltar que la clase litología de mayor interés es "conglomerado" debido a que en esta se encuentran asociados los trazadores del oro, por lo tanto, al momento de evaluar los modelos, el rendimiento para esta clase es un factor importante.

Tabla 11. Porcentaje de clases litológicas en el proceso de entrenamiento y validación, es esta se evidencia que el desbalance de clases permanece proporcional tanto en entrenamiento como en validación.

Clase	Entrenamiento (%)	Validación (%)
Arena arcillosa	47.9	51.9
Arcilla arenosa	19.1	14.9
Arena	11.6	14.8
Conglomerado	9.8	8.5
Materia orgánica	6.9	6.5
Arcilla	3.7	3.4
Limo	0.9	0.0

En la Tabla 12, se resumen los rendimientos de los 4 modelos de clasificación implementados con el data set "**Clasificación manual**", utilizando como medidas de rendimiento la precisión general, la precisión para "conglomerado" y el f1-score para "conglomerado". Se observa que "Red convolucional hibrida", "Procesamiento de lenguaje natural" y "Regresión lineal para clasificación" en el proceso de entrenamiento tuvieron métricas prácticamente iguales, además, aunque "Aprendizaje por transferencia" tuvo un resultado menor, se encuentra en el rango de las otras técnicas. Esto es un indicio de que estos modelos son capaces de representar y abstraer los datos espaciales que se le están ingresando. Ahora bien, en el proceso de validación se hace evidente que las capacidades para generalizar las zonas descocidas difieren entre modelos. Tanto el modelo de "Procesamiento de lenguaje natural" y "Regresión lineal para clasificación" tuvieron las precisiones más bajas, siendo el modelo de "Procesamiento de lenguaje natural" superior al de "Regresión lineal para clasificación", esto valida que la metodología propuesta por (Fuentes et al., 2020) si permite extraer información latente de las descripciones litológicas y es aplicable en un entorno de minería aluvial.

El modelo "Red convolucional hibrida" fue el que tuvo la mejor precisión general, sin embargo, el modelo "Aprendizaje por transferencia" tuvo el mejor desempeño (f1-score (Lipton et al., 2014)) respecto a "conglomerado". Dichos resultados para estos dos modelos se atribuyen principalmente a poder procesar información de la topografía del terreno en forma de DEMs, ya que ésta está muy relacionada con la minería aluvial de la zona.

Más adelante se abordará cada ruta experimental individualmente, y se dará más detalle de su desempeño y el resultado 3D.

Tabla 12. Resumen de los 4 modelos implementados, el modelo con el mejor desempeño general es "4.1 Red convolucional hibrida".

Método	Entrenamiento			Validación		
	ACC (%)	ACC Conglomerado (%)	f1 conglomerado (%)	ACC (%)	ACC Conglomerado (%)	f1 conglomerado (%)
4.1 Red convolucional hibrida	95	96	94	60	72	49
4.1.5 Aprendizaje por transferencia	93	86	91	57	59	55
4.2 Procesamiento de lenguaje natural	94	97	96	50	59	40
4.3 Regresión lineal para clasificación	96	97	96	46	44	32

5.1.1 Red convolucional híbrida clasificación

La "Red convolucional hibrida" entrenada desde cero, obtuvo una precisión para los datos de validación del 60 %, la cual fue la mayor precisión de los 4 modelos, lo que se atribuye a la información topográfica suministrada por los modelos digitales de elevación y la capacidad del segmento convolucional para extraer información de estos DEMs, ya que en este ambiente de minería aluvial las geoformas en superficie han sido uno de los principales factores para guiar la explotación minera de manera acertada.

En la Tabla 13 se presenta el rendimiento por clase litológica. En esta se puede observar que, en la etapa de validación, la precisión para “arcilla” y “limo” es de 0 %. Esto es ocasionado por el desbalance de etiquetas debido a que en el proceso de entrenamiento estas dos etiquetas apenas representaban el 3.7 % y el 0.9 % respectivamente, por lo que el modelo está ignorando estas clases. Aunque “arena” tiene un porcentaje de muestras mayor a “materia orgánica”, la precisión en “materia orgánica” esmucho mayor porque los datos presentes en esta clase se concentran en su mayoría en la superficiedel terreno, por lo que para el modelo es relativamente fácil encontrar el patrón característico. Aunque la precisión general de este modelo fue la mayor, el f1-score (Lipton et al., 2014) para conglomerado (litología de mayor interés) no supera el 50 %, aun así es 9 puntos porcentuales mayor que en “Procesamiento de lenguaje natural”, si a esto le sumamos que la precisión general de “Procesamiento de lenguaje natural” es inferior en 10 puntos porcentuales, se puede afirmar que el modelo “Red convolucional hibrida” tiene mejor desempeño que “Procesamiento de lenguaje natural”. En la Figura 5.2 se muestra el proceso de entrenamiento y validación, el mejor desempeño para los datos de validación se obtuvo en la época 7, con este se realizan los modelos 3D mostrados más adelante.

Tabla 13. Rendimiento de la Red convolucional híbrida desde cero para el proceso de entrenamiento y validación.

clase litológica	Entrenamiento		validación	
	Precisión (%)	f1-score (%)	Precisión (%)	f1-score (%)
arcilla	90	92	0	0
conglomerado	96	94	72	49
arena arcillosa	97	97	74	70
materia orgánica	100	95	89	83
arena	94	93	28	37
arcilla arenosa	90	93	37	45
limo	58	72	0	0

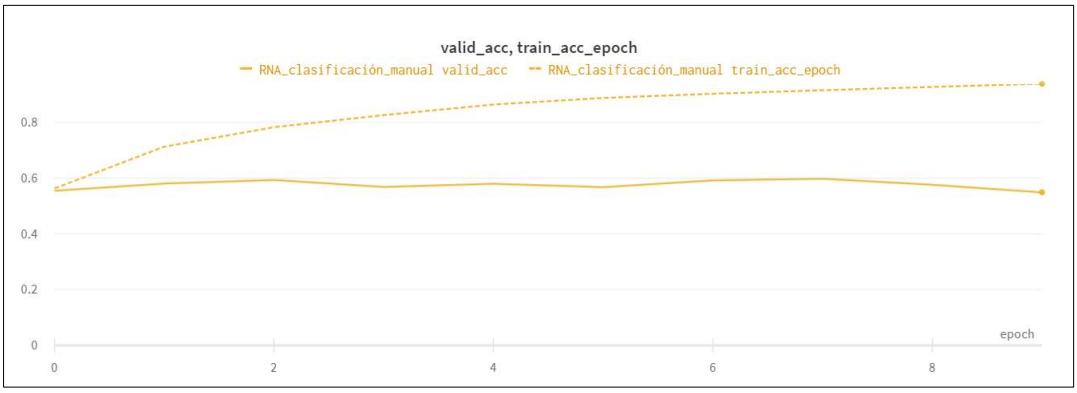


Figura 5.2. Proceso de entrenamiento y validación para “Red convolucional hibrida” con los datos de "Clasificación manual”

Al analizar la red neuronal con transfer learning, se puede concluir que se obtuvo la segunda mejor precisión general (57 %), sin embargo, en la
Tabla *14*, el f1-score de “conglomerado” es del 55 %. Esto es una mejora de aproximadamente 6 puntos porcentuales respecto a la red neuronal entrenada desde cero y 15 del modelo de “Procesamiento de lenguaje natural”. Debido a esto, aunque el modelo con transfer learning no tenga el mejor rendimiento general, si es el que mejor representa la litología de interés. Esta

mejora se debe al aprendizaje por transferencia, ya que el segmento convolucional proveniente de la resnet18 (He et al., 2015) abstrae mejor la información relacionada con “conglomerado” en comparación de la red entrenada desde cero sacrificando la precisión de “arcilla arenosa”. Para las clases de “limo”, “arcilla”, “arena” y “materia orgánica” ocurrió el mismo fenómeno asociado al desbalance que en la red entrenada desde cero. En la Figura 5.3 se muestra el proceso de entrenamiento y validación, el mejor desempeño para los datos de validación se obtuvo en la época 7, con este se realizan los modelos 3D mostrados más adelante.

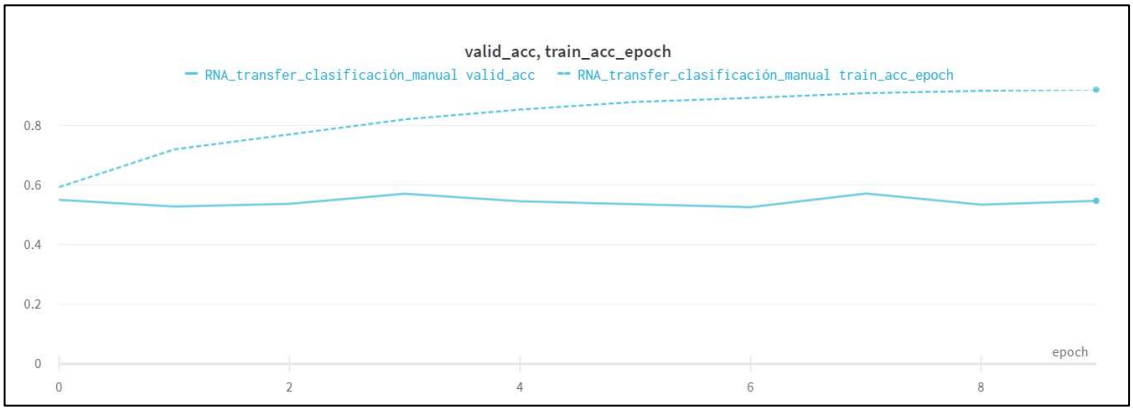


Figura 5.3. Proceso de entrenamiento y validación para “Red convolucional hibrida con transfer learning” con los datos de "Clasificación manual”

Tabla 14. Rendimiento de la Red convolucional híbrida con transfer learning para el proceso de entrenamiento y validación.

clase litológica	Entrenamiento		validación	
	Precisión (%)	f1-score (%)	Precisión (%)	f1-score (%)
arcilla	73	82	0	0
conglomerado	91	86	59	55
arena arcillosa	92	92	79	71
materia orgánica	87	92	94	84
arena	91	83	7	10
arcilla arenosa	85	85	28	32
limo	0	0	0	0

Los dos modelos 3D resultantes son bastante similares, Figura 5.4, donde el modelo con transfer learning es el que tiene una mayor fiabilidad en la distribución de “conglomerado”. En la Figura 5.5 se comparan los resultados en la zona para “conglomerado”. Los dos modelos tienen una distribución bastante similar con diferencias en el espesor de las capas de conglomerado. Se puede decir que ambos captaron y representaron la misma geoforma a pesar de las diferencias en las métricas.

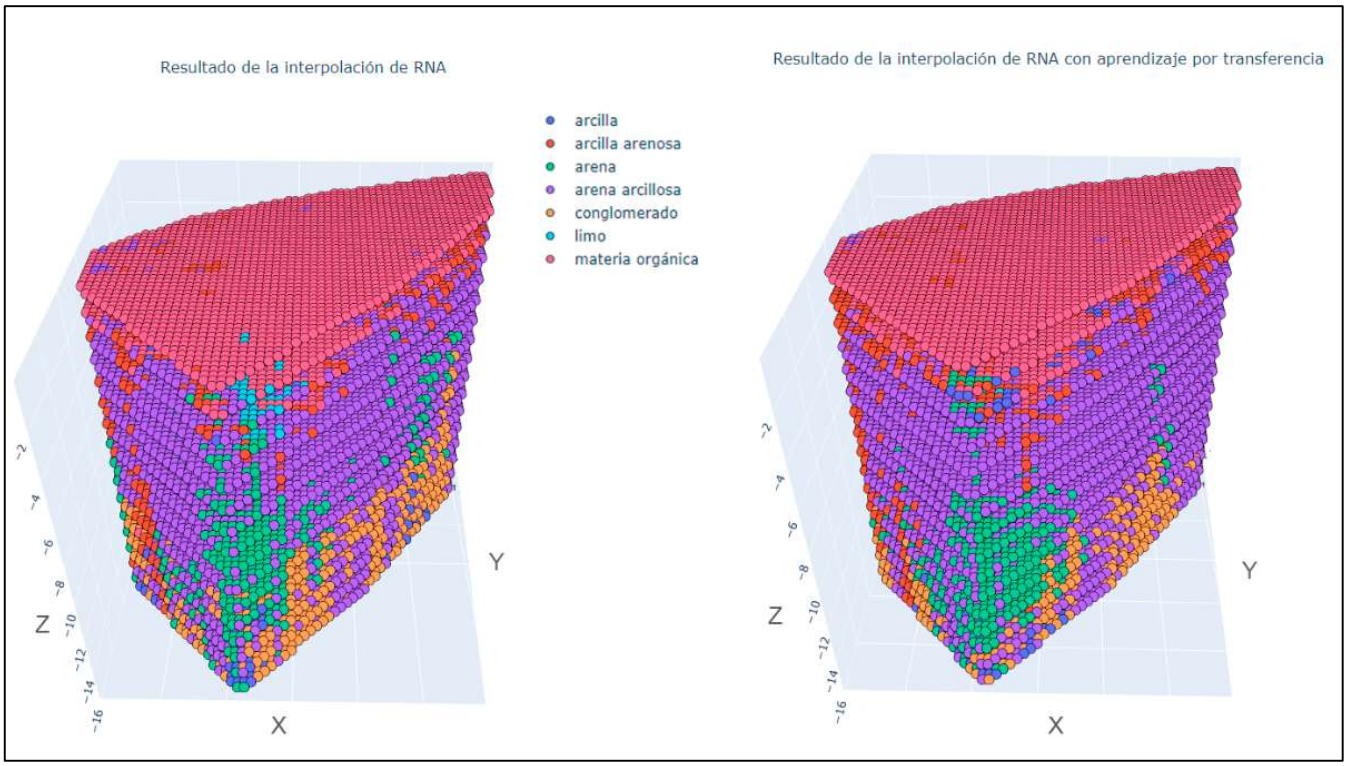


Figura 5.4. Resultado del modelo 3D proveniente de la red neuronal entrenada desde cero y la red a la que se le aplicó transfer learning.

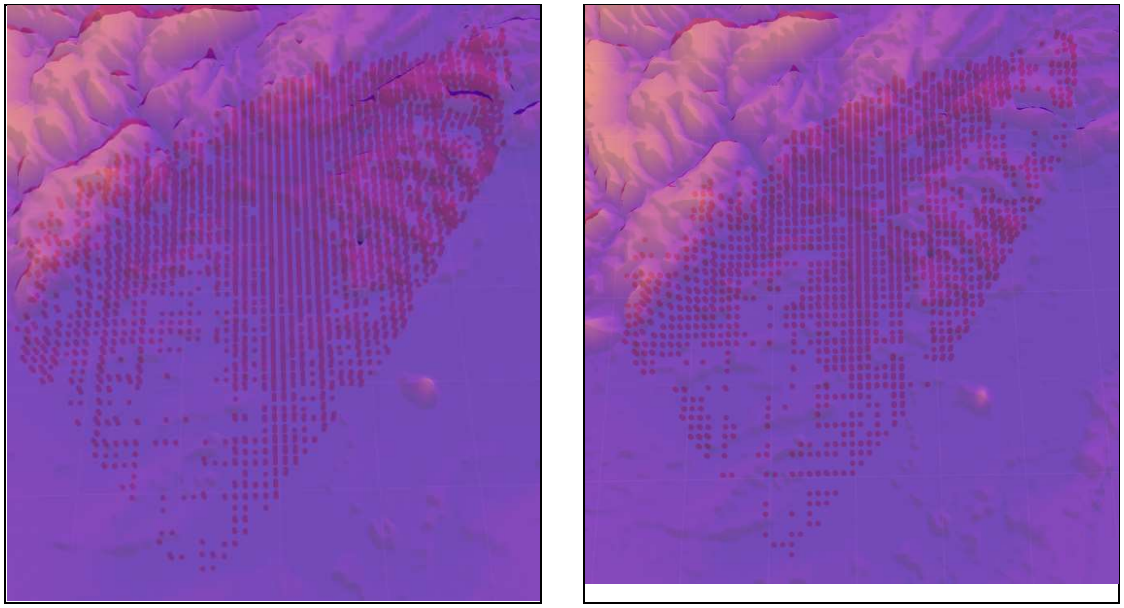


Figura 5.5. Comparación entre la distribución espacial de "conglomerado" para la red entrenada desde cero (izquierda) y la red con transfer learning (derecha).

5.1.2 Procesamiento de leguaje natural

La precisión para los datos de validación fue del 50 %, un rendimiento cercano al reportado en el trabajo de (Fuentes et al., 2020) en la zona de "Moree", lo cual es un buen indicio de la replicabilidad de la metodología que propusieron estos autores, a pesar de la gran diferencia entre la cantidad de datos, etiquetas y naturaleza de la zona de estudio. En la Tabla 15 se muestra el rendimiento por clase litológica. Al igual que en los resultados de la "Red convolucional hibrida"

(desde cero y con transfer learning) la precisión para “limo”, “arcilla” y “arena” estuvieron muy afectadas por el desbalance de etiquetas, dando como resultado valores cercanos a cero. En el caso de “materia orgánica” sucede igual a los modelos anteriores; esta clase litológica se distribuye principalmente en superficie, por lo que el modelo encuentra el patrón característico con facilidad a pesar de las pocas etiquetas. Una situación por destacar es que la precisión en “conglomerado” es 9 puntos porcentuales mayor a la precisión general del modelo, sin embargo, el f1-score es apenas del 40 %. Para “arena arcillosa” y “arcilla arenosa” el resultado estuvo por debajo con respecto al de la “Red convolucional hibrida” pero sin alejarse demasiado. En general, este modelo presentó un desempeño dentro del rango de la literatura. Para zonas llanas o donde la topografía a superficie no influya en el resultado, este modelo es una buena alternativa.

Tabla 15. Rendimiento para el modelo de procesamiento de leguaje natural para el proceso de entrenamiento y validación

clase litológica	Entrenamiento		Validación	
	Precisión (%)	f1-score (%)	Precisión (%)	f1-score (%)
arcilla	68	74	1	2
conglomerado	97	96	59	40
arena arcillosa	97	97	66	66
materia orgánica	96	95	88	82
arena	86	90	2	2
arcilla arenosa	93	92	32	30
limo	85	69	0	0

En la Figura 5.6 (A) se observa la distribución de las clases litológicas. La presencia de “arena arcillosa” a la profundidad de 10 m en adelante es mucho más marcada que en el modelo de la “Red convolucional hibrida”. En la Figura 5.6 (B) queda claro que las geoformas resultantes de las dos metodologías tienen diferencias significativas, sin embargo, coinciden en la ausencia y/o disminución de “conglomerado” en la zona demarcada en verde.

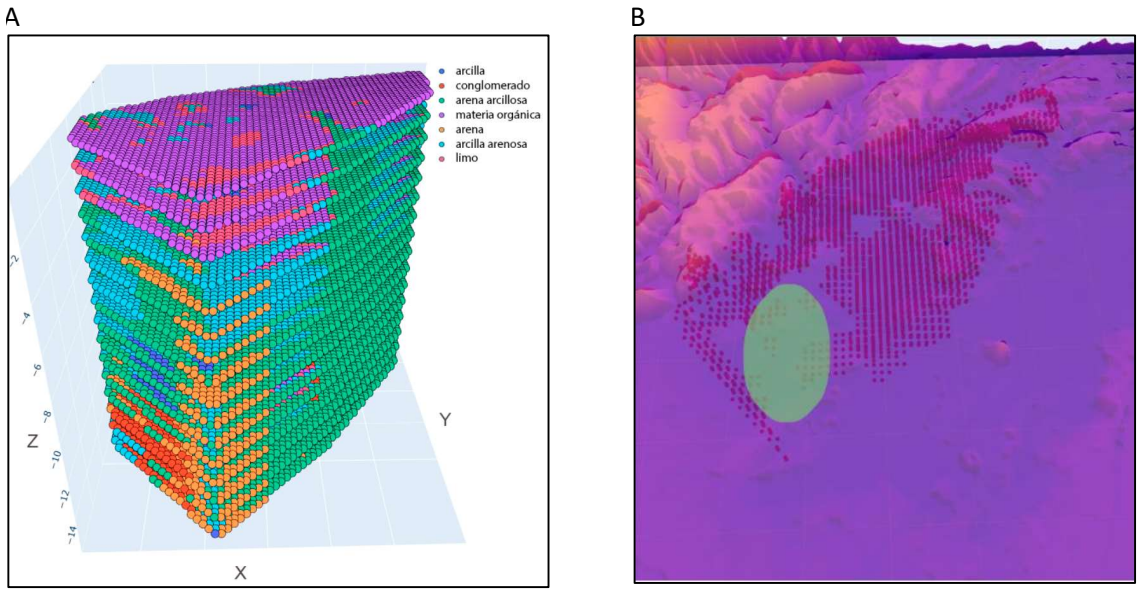


Figura 5.6. (A) Resultado del modelo 3D proveniente del modelo de procesamiento de lenguaje natural. (B) distribución de “conglomerado”, se demarca una zona verde donde el modelo de la “Red neuronal hibrida” y el modelo de

"Procesamiento de lenguaje natural" denotan una ausencia de "conglomerado"

5.1.3 Regresión lineal

Como era de esperarse el modelo de interpolación lineal fue el que presentó el menor rendimiento,tanto en la precisión global (46 %), como en la tarea de predecir la litología de interés (conglomerado), ver Tabla 16. Aun así, se obtuvo un buen desempeño con "materia orgánica", debido a que el 80 % de los datos de esta clase se encuentra a menos de 1.55 m de profundidad y el 75 % de los datos en el primer metro de profundidad es "materia orgánica" por lo que es relativamente sencillo representar esta relación con una regresión lineal. En general, tanto el modelo de la "Red neuronal híbrida" en sus dos variantes, como el modelo de "Procesamiento de lenguaje natural", presentaron mejores rendimientos respecto a este modelo. Sin embargo, el resultado de la distribución espacial de "conglomerado" fue muy similar al del modelo de "Procesamiento de lenguaje natural" (Figura 5.7), esto puede ser consecuencia de que los dos modelos utilizan una interpolación lineal para representar las relaciones espaciales.

Tabla 16. Rendimiento de la interpolación lineal para el proceso de entrenamiento y validación.

clase litológica	Entrenamiento		Validación	
	Precisión (%)	f1-score (%)	Precisión (%)	f1-score (%)
arcilla	96	95	4	3
conglomerado	97	96	44	32
arena arcillosa	96	97	63	63
materia orgánica	97	96	89	86
arena	96	96	5	5
arcilla arenosa	95	94	19	24
limo	91	94	0	0

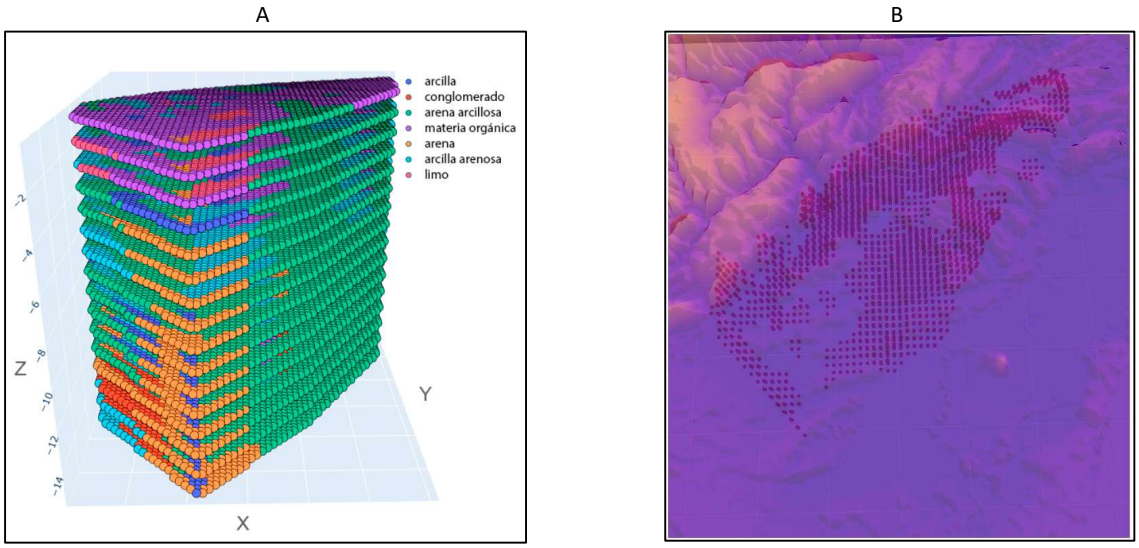


Figura 5.7. (A) Resultado del modelo 3D proveniente del modelo de 5.1.3 Regresión lineal. (B) distribución de "conglomerado", se hace evidente que el resultado de la geoforma producto de la interpolación es muy similar al del modelo de "Procesamiento de lenguaje natural".

5.2 Data set “clasificación por tamaño de partícula”

Como se mencionó en la sección **3.4.3 K-means** el desbalance de las clases presentes en este data set es menor al de la sección **3.2 Clasificación manual.** Comparando la clase predominante (“Alto”, 52 % de los datos), en “**3.4 Clasificación por tamaño de partícula**”, esta tiene aproximadamente 3 veces más registros que la clase menos frecuente (“Bajo”, 18 % de los datos); mientras que en el data set “**clasificación manual**”, la clase predominante (“arena arcillosa”, 43 % de los datos) tiene aproximadamente 2.4 veces más registros que la segunda clase más frecuente (“arcilla arenosa”, 18 % de los datos) y 3.8 veces más registros que las 3 clases menos frecuentes (“Materia orgánica” 5.2 %, “arcilla” 5.4 %, “limo” 0.6 %, entre las tres representan el 11.2 % de los datos). Esto es una reducción significativa en el desbalance de las clases, no obstante, no necesariamente garantiza un mejor resultado en las predicciones de los modelos que se muestran a continuación, pero puede facilitar la capacidad de representación. En la Tabla 17 se muestrala distribución porcentual de cada clase en la fase de entrenamiento y validación, estos provienen de los mismos pozos.

Tabla 17. Porcentaje de clases en el proceso de entrenamiento y validación.

Clase	Entrenamiento (%)	Validación (%)
Alto	52.9	44.1
Medio	28.5	39.3
Bajo	18.6	16.6

En la Tabla 18 se puede observar la precisión de los cuatro modelos empleados. Se concluye que el rendimiento en general fue similar, no hay diferencias notorias en primera instancia, dando un mal indicio sobre la utilidad de las transformaciones aplicadas a los datos. Una posible explicación es que el fenómeno que está ocurriendo es debido a la agrupación “k-means”, dado que, como se observa en la Figura 5.8, los puntos de unión entre dos clases no tienen separación alguna, por lo que al aplicar la agrupación “k-means” es posible que se pierdan las relaciones espaciales inherentes de los datos, lo que imposibilita encontrar los patrones que requieren los modelos para predecir las zonas desconocidas.

Tabla 18. Precisión de los 4 modelos computaciones para el data set “clasificación por tamaño de partícula”.

Método	Entrenamiento Precisión (%)	Validación Precisión (%)
4.1 Red convolucional hibrida	96	48
4.1.5 Aprendizaje por transferencia	91	50
4.2 Procesamiento de lenguaje natural	96	46
4.3 Regresión lineal para clasificación	97	46

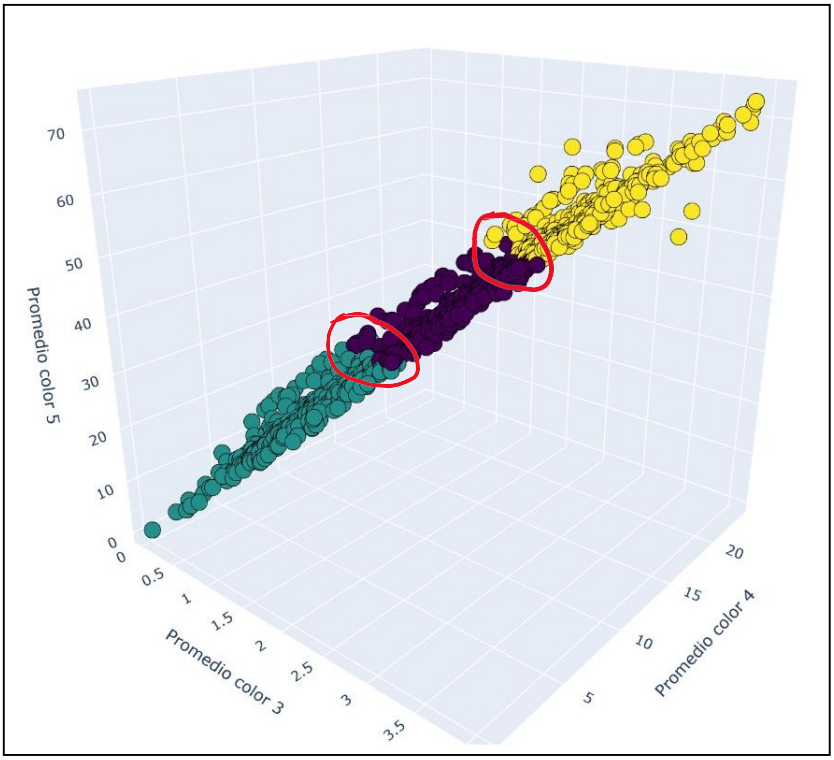


*Figura 5.8. Resultado de aplicar "k-means" en la sección **3.4.3.** Se señalan en rojo las zonas de transición entre una clase y otra sin ninguna distancia espacial entre grupos.*

5.2.1 Red convolucional híbrida clasificación

La precisión (accuracy) de los cuatro modelos en el proceso de entrenamiento y validación son prácticamente idénticas, la "red convolucional híbrida" entrenada desde cero no se destaca significativamente respecto a los otros modelos (Tabla 19). Su desempeño con la clase "Medio" fue levemente mejor a los otros modelos, sin embargo, los resultados de la clase "Bajo" fueron inferiores en comparación con la "red neuronal con transfer learning". En la Figura 5.9 se muestra el proceso de entrenamiento y validación, el mejor desempeño para los datos de validación se obtuvo en la época 6, con este se realizan los modelos 3D mostrados más adelante.

Tabla 19. métricas de rendimiento de la red convolucional hibrida entrenada desde cero.

Clase	Entrenamiento				validación			
	precisión	recall	f1-score	acc	precisión	recall	f1-score	acc
Alto	0.96	0.98	0.97		0.54	0.6	0.57	
Medio	0.93	0.94	0.93		0.48	0.43	0.38	
Bajo	0.98	0.92	0.95	0.96	0.29	0.27	0.46	0.48

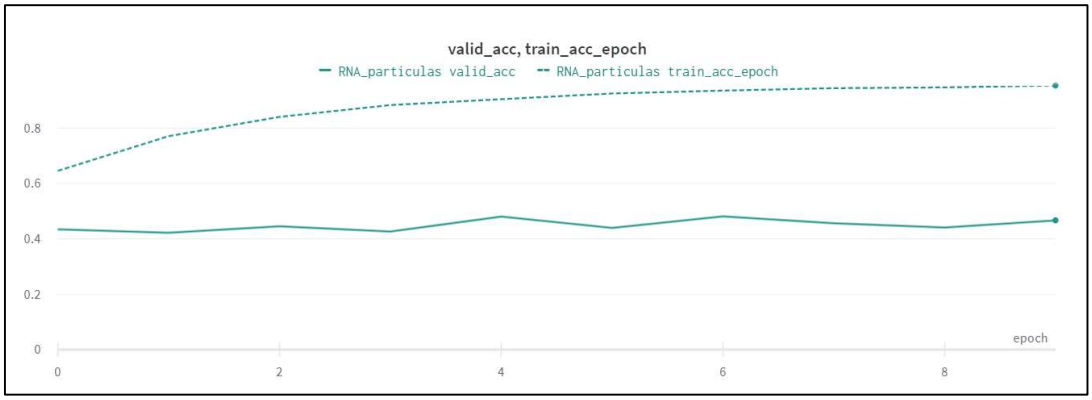


Figura 5.9. Proceso de entrenamiento y validación para "Red convolucional hibrida" con los datos de Clasificación por tamaño de partícula

La red neuronal con transfer learning tuvo el mejor "f1-score" en la clase "Bajo" (Tabla 20), aproximadamente 11 puntos porcentuales del siguiente mejor. Su rendimiento con la clase "Alto" fue levemente mejor al resto de los modelos, pero sin diferencias significativas. En la Figura 5.10 se muestra el proceso de entrenamiento y validación, el mejor desempeño para los datos de validación se obtuvo en la época 4, con este se realizan los modelos 3D mostrados más adelante.

Tabla 20. métricas de rendimiento de la red convolucional hibrida con transfer learning.

Clase	Entrenamiento				validación			
	precisión	recall	f1-score	acc	precisión	recall	f1-score	acc
Alto	0.90	0.97	0.94		0.58	0.57	0.58	
Medio	0.90	0.83	0.86		0.36	0.35	0.35	
Bajo	0.95	0.85	0.90	0.91	0.53	0.61	0.57	0.5

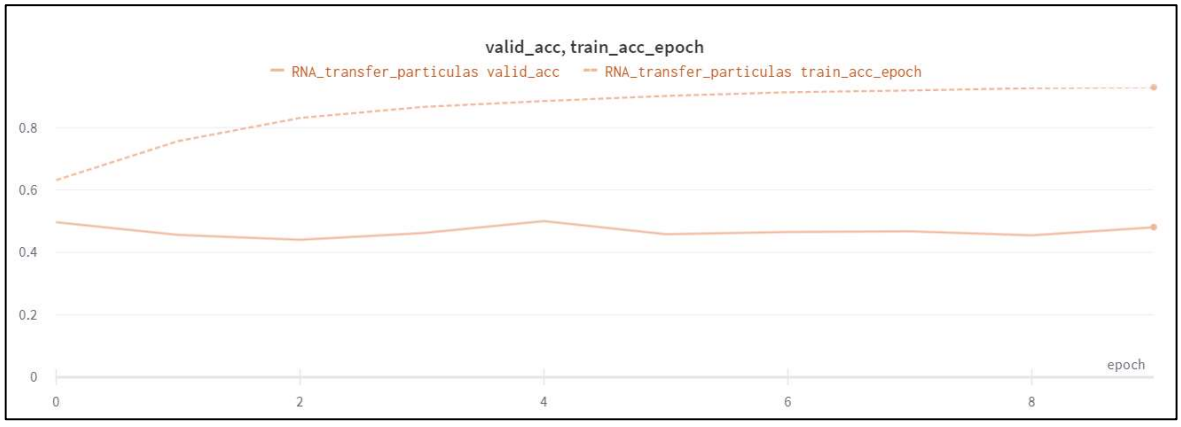


Figura 5.10. Proceso de entrenamiento y validación para "Red convolucional hibrida con transfer learning" con los datos de Clasificación por tamaño de partícula

En la Figura 5.11 se presentan los modelos 3D provenientes de la red neuronal entrenada desde cero (A) y la red a la que se le aplicó transfer learning (B). Se observa una gran cantidad de intrusiones o ruido, es decir que, dentro de una región donde predomina una clase, se encuentran puntos esparcidos de otra clase, esto se interpreta como la incapacidad de las redes neuronales para modelar esta versión de los datos. En contraste, esto no ocurrió con los datos de "clasificación manual".

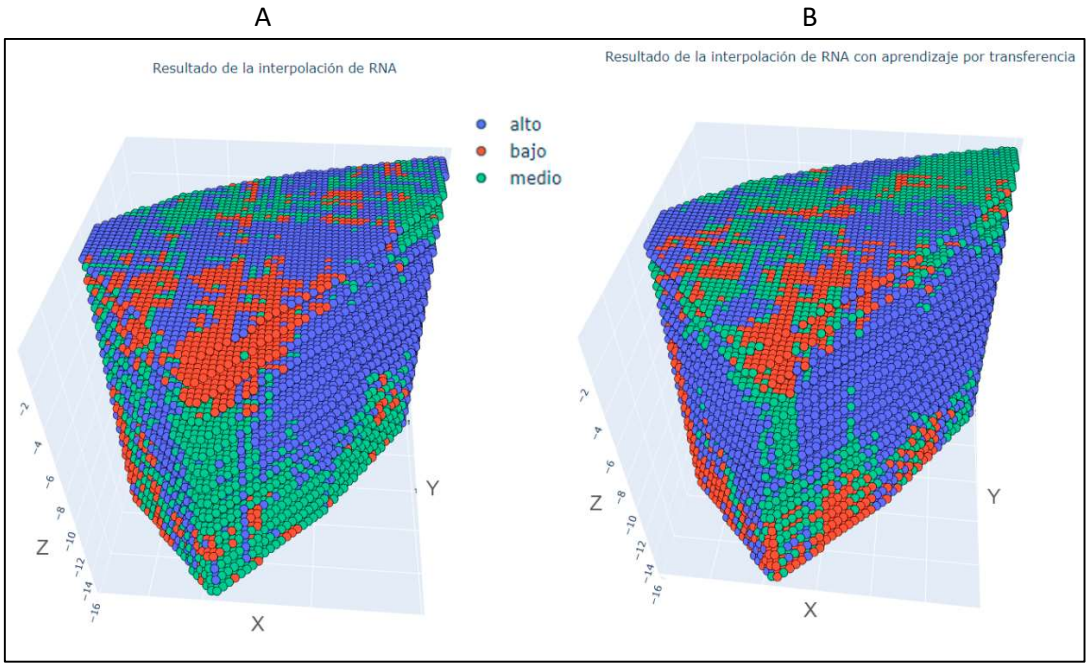


Figura 5.11. Resultado del modelo 3D proveniente de la red neuronal entrenada desde cero (A) y la red a la que se le aplicótransfer learning (B)

5.2.2 Procesamiento de leguaje natural

Este modelo no se destacó en ninguna clase del data set "**Clasificación por tamaño de partícula**"; incluso se puede decir que fue el modelo de menor desempeño debido a que obtuvo el "f1-score" (Tabla 21) más bajo para la clase "Bajo". No obstante, debido a las condiciones del data set "clasificación por tamaño de partícula", estos rendimientos no representan correctamente el alcance de los modelos.

Tabla 21. Rendimiento para el modelo de procesamiento de leguaje natural.

Clase	Entrenamiento				validación			
	precisión	recall	f1-score	acc	precisión	recall	f1-score	acc
Alto	0.98	0.97	0.98		0.71	0.48	0.57	
Medio	0.94	0.92	0.93		0.3	0.42	0.35	
Bajo	0.9	0.98	0.94	1	0.15	0.47	0.23	0.5

A diferencia del modelo 3D resultante de las redes neuronales para el data set "K-means", el resultado del modelo de "procesamiento de lenguaje natural", Figura 5.12, no presenta ruido o intrusiones, incluso se pueden apreciar zonas bien delimitadas. Un punto importante para destacar es que en la mayoría de las ocasiones entre las zonas de "Alto" y "Bajo" hay zonas de "Medio", esto es un indicador de que, para estos datos, este modelo representa mejor las geoformas que las redes neuronales propuestas.

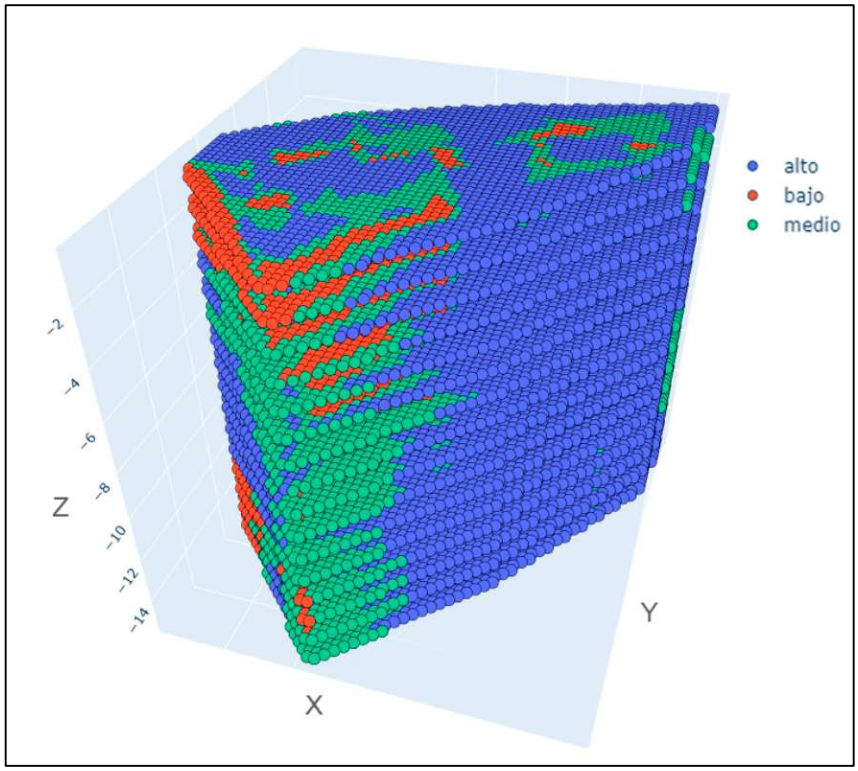


Figura 5.12. Resultado del modelo 3D proveniente del modelo de procesamiento de lenguaje natural.

5.2.3 Regresión lineal

En este trabajo la regresión lineal para clasificación se pensó desde un principio como un punto de control o referencia para comparar los demás modelos. En este caso, este modelo tuvo un desempeño bastante equilibrado, sin destacar por buen o mal rendimiento en ninguna clase, ver Tabla *22*. Al igual que en los modelos anteriores, los resultados asociados al data set proveniente de "K- means" pueden estar sesgados por la pérdida de información.

Tabla 22. Rendimiento para el modelo de regresión lineal.

	Entrenamiento				validación			
Clase	precisión	recall	f1-score	acc	precisión	recall	f1-score	acc
Alto	0.98	0.98	0.98		0.64	0.51	0.57	
Medio	0.96	0.95	0.95		0.32	0.45	0.37	
Bajo	0.96	0.97	0.97	1	0.34	0.34	0.34	0.5

A diferencia del modelo de "procesamiento de lenguaje natural", el resultado de la interpolaciónlineal si tiene bastantes zonas donde las clases "Alto" y "Bajo" se unen, Figura 5.13, esto significa que, la información de las descripciones ingresadas al modelo de "procesamiento de lenguaje natural" si generan una mejora respecto a la interpolación lineal.

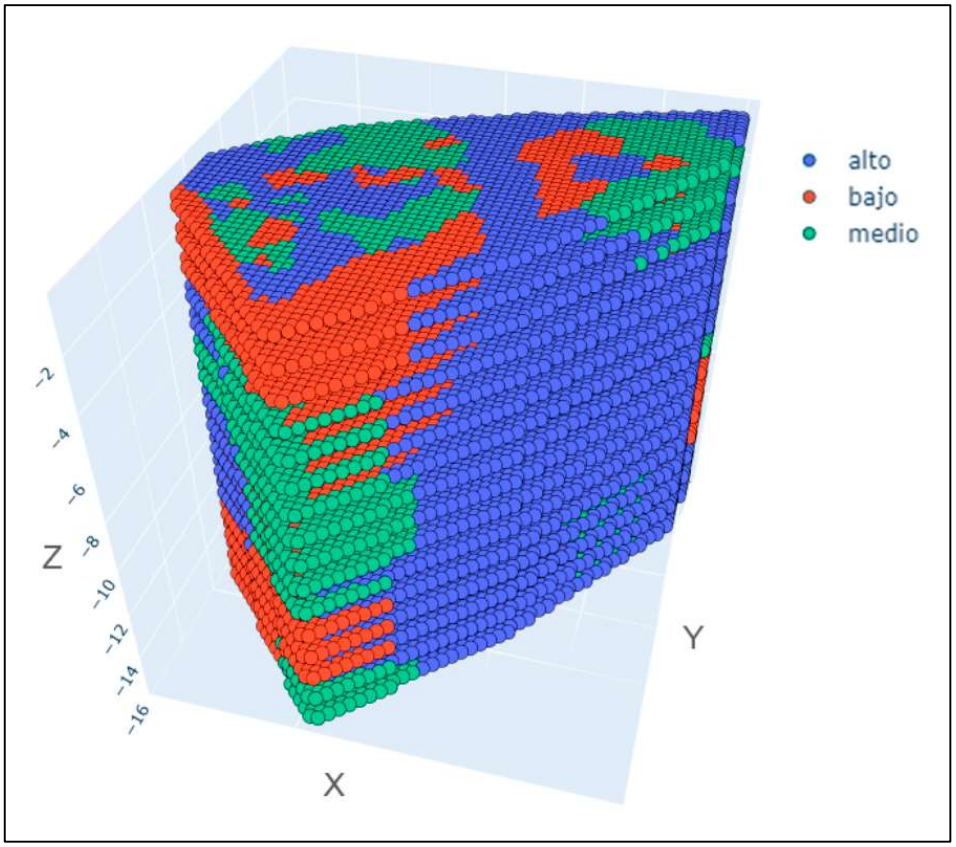


Figura 5.13. Resultado del modelo 3D proveniente de la interpolación lineal.

5.3 Data set “Norma vectorial”

En la sección **5.2** se habló del problema atribuido a agrupar mediante “**K-means**” los datos de la sección “**3.4 Clasificación por tamaño de partícula”.** Por esto es importante analizar el comportamiento de modelos regresivos, para determinar si el procesamiento propuesto en las secciones “**3.4.1 Análisis semántico**” y “**3.4.4 Norma vectorial**” genera valor.

5.3.1 Red convolucional híbrida regresión

Al igual que en los modelos anteriores se utilizó la misma combinación de pozos de entrenamiento y validación. En la Figura 5.14 se muestra la distribución de los valores resultantes de la norma vectorial. Aunque la diferencia entre la cantidad de entrenamiento y validación es alta, se puede apreciar que los datos de validación siguen una distribución similar a los de entrenamiento.

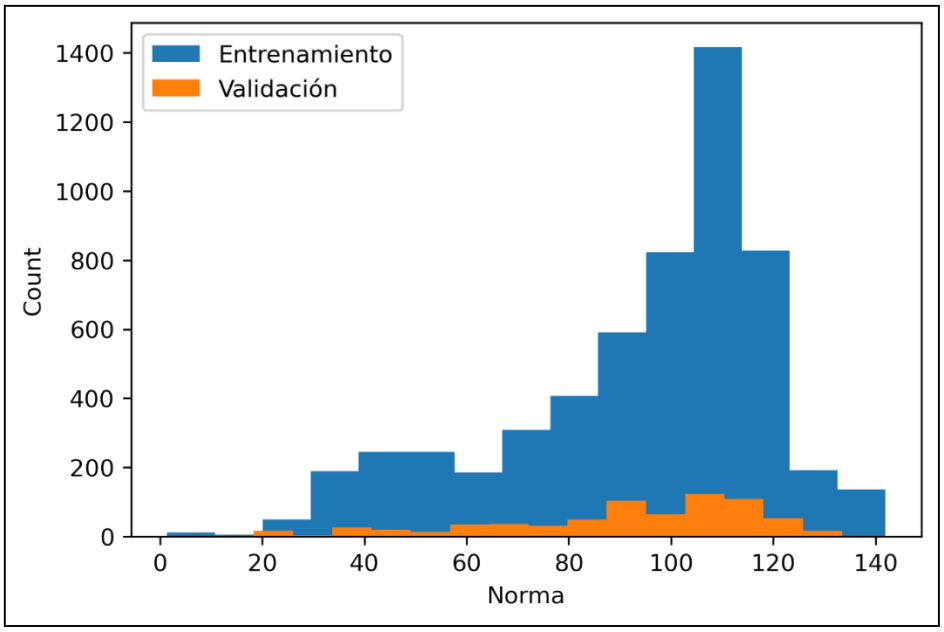


Figura 5.14. distribución de los datos de entrenamiento y validación para regresión 3D.

El proceso de entrenamiento y validación del modelo continuo fue bastante rápido, ya que llega al mejor desempeño de validación en la época 1 Figura 5.15, esto es debido a la naturaleza de los datos, el input de la red neuronal son imágenes satelitales, coordenadas “X”, “Y” y “Z”, de estos valores solo varía la “Z” (dentro del codificador posicional) a profundidad por lo que a la red recibe en una misma época varias veces información similar, esto combinado con un target continuo facilita la rápida convergencia del modelo. Este fenómeno no es tan notorio en las tareas de clasificación anteriores ya que las clases litológicas no tienen una continuidad explicita como lo es en el caso de una tarea de regresión. Este fenómeno es bastante evidente en la Figura 5.16, en esta podemos ver que las salidas del modelo continuo a diferentes profundidades tienen la misma geoforma, lo que cambia es la intensidad.

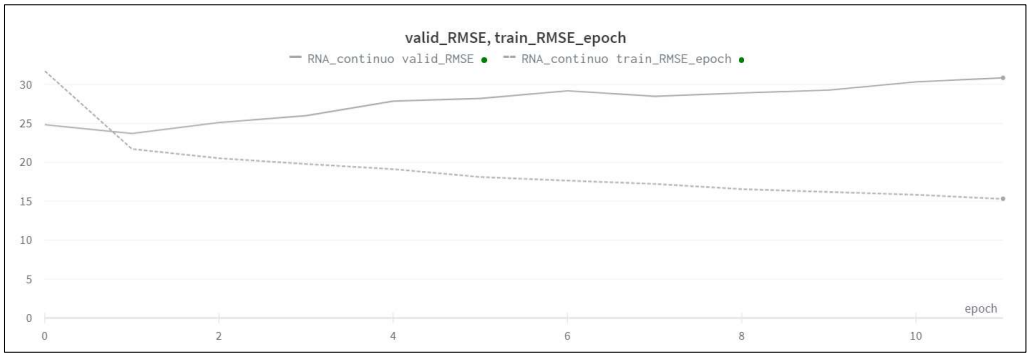


Figura 5.15. Proceso de entrenamiento y validación para la red convolucional híbrida regresión

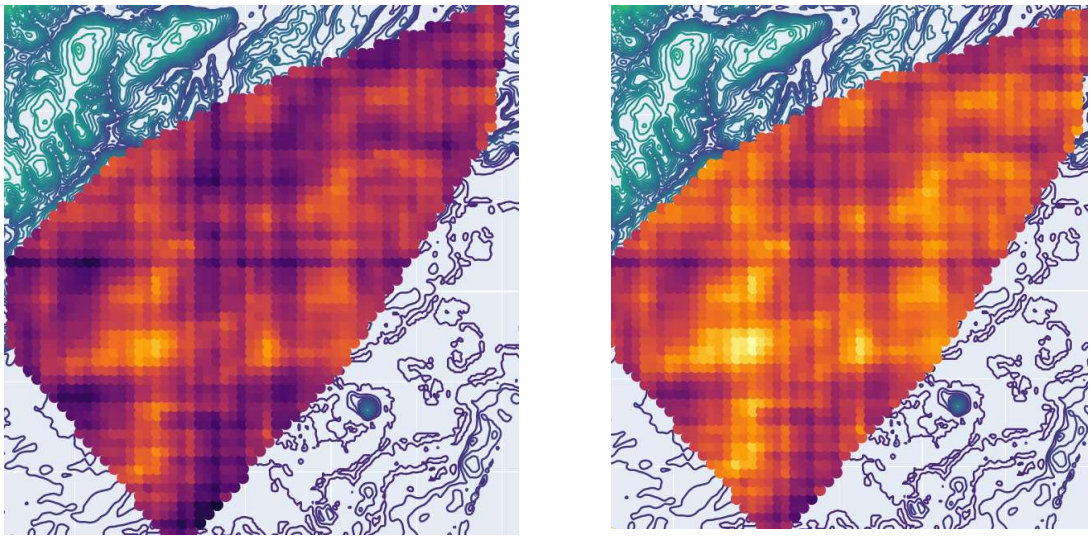


Figura 5.16. Resultado de la red convolucional hibrida para regresión donde se observan las mismas geoformas pero con diferencia de intensidad. (Izquierda) 2.6 m de profundidad. (Derecha) 7.58 m de profundidad

La red convolucional híbrida entrenada desde cero tuvo un RMSE para validación de 21, esto equivale a un error del 18 % del rango de variación de la norma (entre 20 y 140). En comparativa, es un resultado mejor que cualquiera de los modelos de clasificación. Aun así, el punto principal de este modelo continuo son las geoformas, dado que, al no perder información en la clasificación, se pueden demarcar las zonas de mayor interés con mayor precisión. En la Figura 5.17 se muestra una sección 2D obtenida de este modelo. Las zonas marcadas de azul ya fueron explotadas y concuerdan con altos contenidos de oro según lo indicado por WGM, lo cual, valida parcialmente los resultados obtenidos. Se desconocen los datos de explotación o cantidad de oro registrado de las otras zonas.

A B

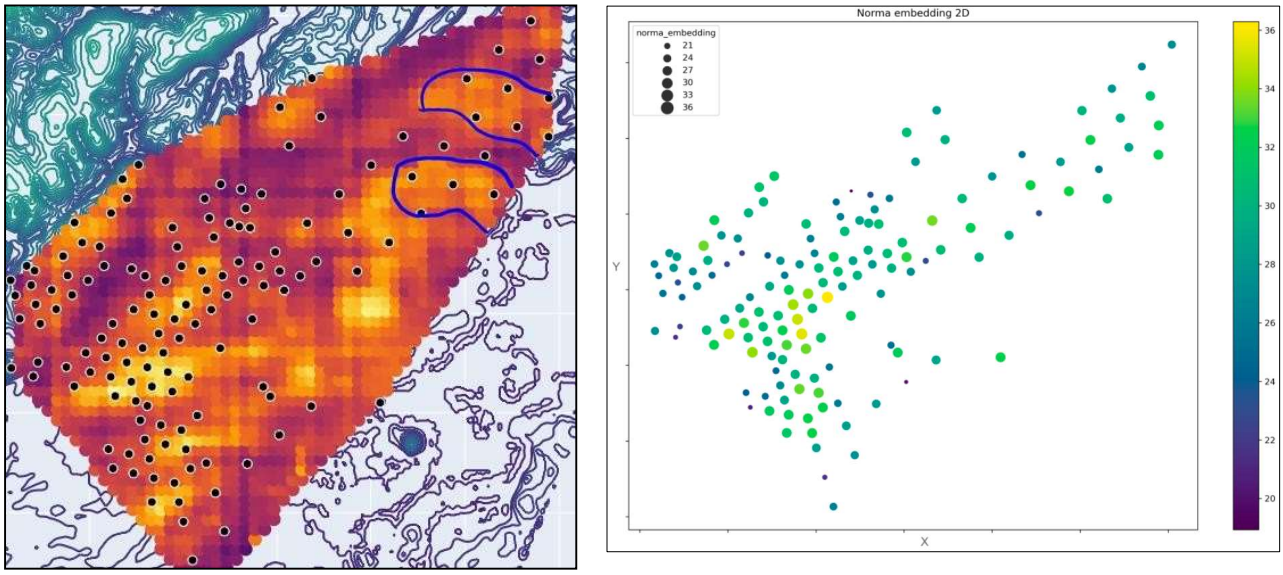


Figura 5.17. (A) Sección 2D a 6.5 metros de profundidad resultante de la red neuronal hibrida para regresión, los puntos negros son las perforaciones de donde provienen los datos. (B) Referencia 2D a superficie proveniente de Norma vectorial. Al comparar ambas imágenes se puede evidenciar que las zonas de interés aurífero coinciden

5.3.2 Interpolación Kriging

El modelo Kriging se planteó en este trabajo como un punto de comparación con la red neuronal, ya que es uno de los métodos más utilizados en geociencias para realizar interpolaciones espaciales. Este modelo no obtuvo un buen desempeño con los datos suministrados, aunque desde la métrica el resultado puede parecer prometedor, teniendo un RMSE de 3.4 para datos en un rango entre 20 y 35, equivalente a un error del 23%, al momento de representar el resultado 2D (Figura 5.18) es evidente que la capacidad de representación no se adapta muy bien a la zona de estudio.

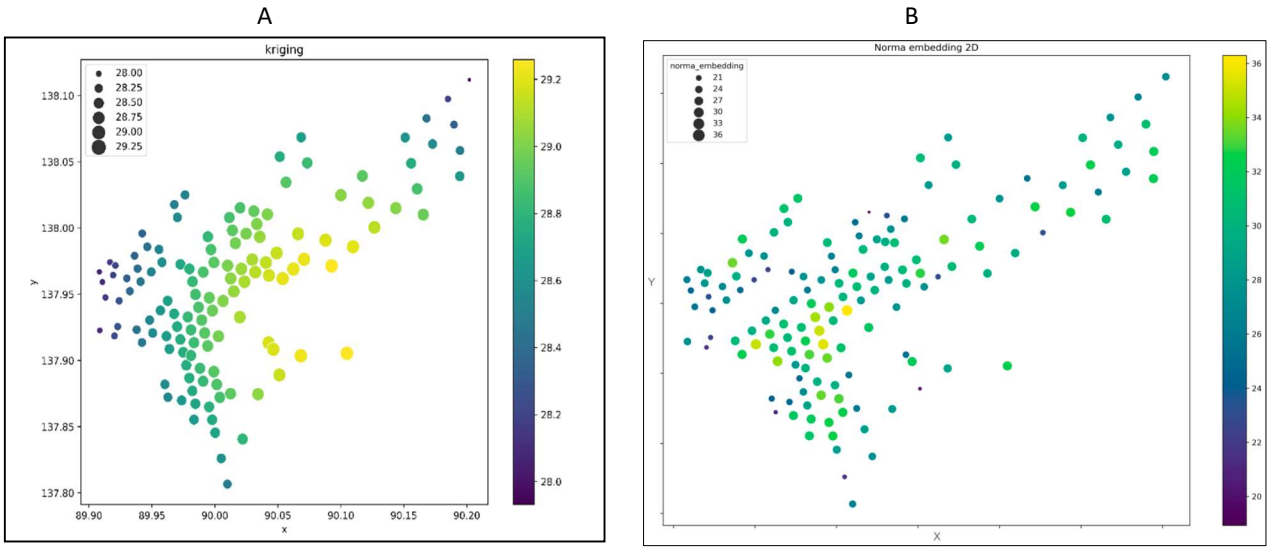


Figura 5.18. (A) Resultado de la interpolación Kriging, se observa que la distribución se concentra en el centro de la zona y no se identifican patrones complejos. (B) Datos reales de la zona

6 Conclusiones y trabajo futuro

En la presente tesis se abordaron diversas rutas para el procesamiento de datos pertenecientes a un entorno de minería aluvial y cómo a partir de estos se pueden implementar metodologías de machine learning para obtener modelos 2D y 3D del subsuelo con el objetivo de apoyar la prospección minera.

Los datos en geociencias suelen tener un alto costo, ya que las etapas de exploración inicial requieren inversiones significativas, sin embargo, el volumen de datos obtenidos es relativamente bajo, por este motivo, poder extraer información adicional de los datos originales es una apuesta muy llamativa. Sin embargo, los resultados de este trabajo se postulan como candidatos al aprovechamiento, procesado y/o reprocesado de información y datos históricos de campañas exploratorias, con herramientas computacionales, que permitan obtener resultados con mayores posibilidades de validación e interpretación de los modelos del suelo y del subsuelo construidos, y con ello orientar con mayor confianza la etapa de explotación de un proyecto minero.

En el caso de la metodología de procesamiento de lenguaje natural, transformar datos cualitativos en datos numéricos, como con las descripciones verbales, abre la posibilidad de implementar un gran abanico de herramientas computacionales. En este trabajo en específico, se desarrollaron modelos de machine learning, los cuales demostraron tener ciertas fortalezas para modelar los complejos sistemas presentes en el subsuelo.

Un resultado positivo fue poder relacionar las descripciones litológicas a nivel de palabra con la geoquímica, permitiendo identificar aquellas palabras con mayor relación con el oro y que estas coincidan con los trazadores de oro identificados en la zona.

No se obtuvo un resultado positivo de todos los experimentos. En el caso de los reductores dimensionales, el resultado no agregó valor a los datos iniciales. Sin embargo, no se descarta su uso con otros tipos de datos, ya que esta técnica puede ayudar a encontrar patrones que generen bastante valor en el entorno minero.

La clasificación de los datos mediante k-mean tuvo un buen resultado en la tarea de disminuir el desbalance a la hora de agrupar los datos, sin embargo, en este proceso se pierde información. En contra parte, obtener la norma vectorial para luego pasar a una tarea de regresión fue el procesamiento que representó mejor la zona de estudio, debido a que conserva los gradientes de transición y facilita a los modelos la tarea de representar la zona desconocida.

Un punto importante de este trabajo fue demostrar la aplicabilidad de la metodología propuesta por otros autores en un entorno de minería aluvial y en una zona con una topografía distinta al estudio original.

El resultado más importante de este trabajo proviene de la red neuronal híbrida para regresión, ya que, haciendo uso de los DEMs, coordenadas, descripciones textuales y geoquímica del oro, se pudo modelar la zona de estudio y coincidir con la presencia de oro en zonas que ya fueron exploradas. Es válido aclarar que esta información no se ingresó a la red de ninguna forma y que falta constatar amedida que avance la explotación si en las otras zonas también coincide.

Para trabajos futuros se plantea la implementación de estas técnicas en otros entornos mineros. Se espera que los resultados sean mejores con minerales que se encuentren en mayores concentraciones que el oro, ya que estas zonas tendrían una distribución más balanceada. De igual manera es necesario seguir refinando estas metodologías en entornos de minería aurífera aluvial

debido a su importancia para el país.

La metodología de trabajo podría ser escalada a otro tipo de depósitos minerales auríferos y de hecho a depósitos de otro tipo de minerales metálico y no metálicos.

7 Anexos

Para facilitar el análisis, entendimiento y replicación de los modelos implementados en esta investigación, se anexa el repositorio de github (https://github.com/francobr991/Mapeo_prospectivo) donde se encuentran los siguientes archivos.

- 1_positional_3D_encode.ipynb: Código del codificador posicional utilizado para transformar las coordenadas de los puntos espaciales en la sección 4.1.2
- 2_dataloader.ipynb: Esquema de dataloader desarrollado en Pytorch que muestra como le ingresan los datos a la red neuronal descrita en la sección 4.1
- 3_RNA.ipynb: Red neuronal descrita en la sección 4.1
- 4_RNA_transfer.ipynb: Red neuronal descrita en la sección 4.1.5
- 5_kriging.ipynb: Interpolador gaussiano desarrollado en GPytorch descrito en la sección 4.4.2
- 6_embdedding_and_linear_interpolation.ipynb: sección donde se aplica la metodología descrita en la sección 4.2 haciendo uso de *hydra.py*
- hydra.py: Código utilizado para interpolación lineal, graficas, manejo de DEM y manejos de datos, con la clase *get_model* y la función *interpolacion_lineal_basica* se realizan las secciones 4.2.2 y 4.3
- requirements.txt: Lista de paquetes y librerías con sus respectivas versiones presentes en el ambiente de desarrollo de esta investigación

8 Bibliografía

Abedi, M., Norouzi, G.-H., & Bahroudi, A. (2012). Support vector machine for multi-classification of mineral prospectivity areas. *Computers & Geosciences*, *46*, 272-283. https://doi.org/10.1016/j.cageo.2011.12.014

Agarwal, N., Sondhi, A., Chopra, K., & Singh, G. (2021). Transfer Learning: Survey and Classification. En S. Tiwari, M. C. Trivedi, K. K. Mishra, A. K. Misra, K. K. Kumar, & E. Suryani (Eds.), *Smart Innovations in Communication and Computational Sciences* (pp. 145-155). Springer. https://doi.org/10.1007/978-981-15-5345-5_13

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a convolutional neural network. *2017 International Conference on Engineering and Technology (ICET)*, 1-6. https://doi.org/10.1109/ICEngTechnol.2017.8308186

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. En M. W. Berry, A. Mohamed, & B. W. Yap (Eds.), *Supervised and Unsupervised Learning for Data Science* (pp. 3-21). Springer International Publishing. https://doi.org/10.1007/978-3-030-22475-2_1

Anwar, S. M., Majid, M., Qayyum, A., Awais, M., Alnowami, M., & Khan, M. K. (2018). Medical Image Analysis using Convolutional Neural Networks: A Review. *Journal of Medical Systems*, *42*(11), 226. https://doi.org/10.1007/s10916-018-1088-1

Arisi, D., Cortés, A., & Vieyra, J. C. (2017). *Colombia 2030: Mejorando la gestión del sector minero energético | Publications*. Inter-American Development Bank. https://publications.iadb.org/publications/spanish/document/Colombia-2030-Mejorando-la-gesti%C3%B3n-del-sector-minero-energ%C3%A9tico.pdf

Basha, S. H. S., Dubey, S. R., Pulabaigari, V., & Mukherjee, S. (2020). Impact of fully connected layers on performance of convolutional neural networks for image classification. *Neurocomputing*, *378*, 112-119. https://doi.org/10.1016/j.neucom.2019.10.008

Bergen, K. J., Johnson, P. A., Maarten, V., & Beroza, G. C. (2019). Machine learning for data-driven discovery in solid Earth geoscience. *Science*, *363*(6433).

Bromberg, F., & Pérez, D. (2012). *Interpolación espacial mediante aprendizaje de máquinas en viñedos de la Provincia de Mendoza, Argentina—13th Argentine Symposium on Artificial Intelligence, ASAI* (rayyan-666224532). *41*.

Chowdhary, K. R. (2020). Natural Language Processing. En K. R. Chowdhary (Ed.), *Fundamentals of Artificial Intelligence* (pp. 603-649). Springer India. https://doi.org/10.1007/978-81-322-3972-7_19

Cressie, N. (1990). The origins of kriging. *Mathematical Geology*, *22*(3), 239-252. https://doi.org/10.1007/BF00889887

D'AGOSTINO, R. B. (1971). An omnibus test of normality for moderate and large size samples. *Biometrika*, *58*(2), 341-348. https://doi.org/10.1093/biomet/58.2.341

D'AGOSTINO, R., & PEARSON, E. S. (1973). Tests for departure from normality. Empirical results for the distributions of b2 and √b1. *Biometrika*, *60*(3), 613-622. https://doi.org/10.1093/biomet/60.3.613

David. (2021, abril 19). CONCENTRADORES HIDRÁULICOS. *Foro por Metalurgista de 911Metallurgist*. https://www.911metallurgist.com/metalurgia/concentradores-hidraulicos/

Dramsch, J. S. (2020). Chapter One—70 years of machine learning in geoscience in review. En B. Moseley & L. Krischer (Eds.), *Advances in Geophysics* (Vol. 61, pp. 1-55). Elsevier. https://doi.org/10.1016/bs.agph.2020.08.002

D'yachkov, B. A., Bissatova, A. Y., Mizernaya, M. A., Khromykh, S. V., Oitseva, T. A., Kuzmina, O. N., Zimanovskaya, N. A., & Aitbayeva, S. S. (2021). Mineralogical Tracers of Gold and Rare-Metal Mineralization in Eastern Kazakhstan. *Minerals*, *11*(3), Art. 3. https://doi.org/10.3390/min11030253

Example: Gaussian Process—NumPyro documentation. (s. f.). Recuperado 11 de septiembre de 2022, de https://num.pyro.ai/en/stable/examples/gp.html

Frigge, M., Hoaglin, D. C., & Iglewicz, B. (1989). Some Implementations of the Boxplot. *The American Statistician*, *43*(1), 50-54. https://doi.org/10.1080/00031305.1989.10475612

Fu, G., Lü, Q., Yan, J., Farquharson, C. G., Qi, G., Zhang, K., Zhang, Y., Wang, H., & Luo, F. (2021). 3D mineral prospectivity modeling based on machine learning: A case study of the Zhuxi tungsten deposit in northeastern Jiangxi Province, South China. *Ore Geology Reviews*, *131*, 104010.

https://doi.org/10.1016/j.oregeorev.2021.104010

Fuentes, I., Padarian, J., Iwanaga, T., & Willem Vervoort, R. (2020). 3D lithological mapping of borehole descriptions using word embeddings. *Computers & Geosciences*, *141*, 104516. https://doi.org/10.1016/j.cageo.2020.104516

Fuentes López, H. J., Ferrucho Parra, C. C., & Martínez González, W. A. (2021). *La minería y su impacto en el desarrollo económico en Colombia*. https://revistas.uptc.edu.co/index.php/cenes/article/view/12225

Gao, Z., Fu, W., Zhang, M., Zhao, K., Tunney, H., & Guan, Y. (2016). Potentially hazardous metals contamination in soil-rice system and it's spatial variation in Shengzhou City, China. *Journal of Geochemical Exploration*, *167*, 62--69.

García Jacome, E. (1978). El oro en Colombia. *Sociedad Geográfica de Colombia*, *33*(113).

Gardner, J. R., Pleiss, G., Bindel, D., Weinberger, K. Q., & Wilson, A. G. (2021). *GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration* (arXiv:1809.11165). arXiv. http://arxiv.org/abs/1809.11165

Gentleman, R., & Carey, V. J. (2008). Unsupervised Machine Learning. En F. Hahne, W. Huber, R. Gentleman, & S. Falcon (Eds.), *Bioconductor Case Studies* (pp. 137-157). Springer. https://doi.org/10.1007/978-0-387-77240-0_10

GloVe: Global Vectors for Word Representation. (s. f.). Recuperado 22 de enero de 2023, de https://nlp.stanford.edu/projects/glove/

Goodman, N. R. (1963). Statistical Analysis Based on a Certain Multivariate Complex Gaussian Distribution (An Introduction). *The Annals of Mathematical Statistics*, *34*(1), 152-177.

Granek, J., & Haber, E. (2015). *Advanced Geoscience Targeting via Focused Machine Learning Applied to the QUEST Project Dataset, British Columbia*. 10.

Guo, W., Yang, W., Zhang, H., & Hua, G. (2018). Geospatial Object Detection in High Resolution Satellite Images Based on Multi-Scale Convolutional Neural Network. *Remote Sensing*, *10*(1), Art. 1. https://doi.org/10.3390/rs10010131

Hamerly, G., & Elkan, C. (2003). Learning the k in k-means. *Advances in Neural Information Processing Systems*,

16. https://proceedings.neurips.cc/paper/2003/hash/234833147b97bb6aed53a8f4f1c7a7d8-Abstract.html

Hawkins, D. M. (1980). *Identification of outliers* (Vol. 11). Springer.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition* (arXiv:1512.03385). arXiv. http://arxiv.org/abs/1512.03385

Hidalgo, J. (2020). *Caracterización de las Formaciones Quintuco y Vaca Muerta, a partir de datos sísmicos y de pozos, en el área Lindero Atravesado, Cuenca Neuquina*. http://rdi.uncoma.edu.ar/handle/123456789/15872

Iturrarán-Viveros, U., Muñoz-García, A. M., Parra, J. O., & Tago, J. (2018). Validated artificial neural networks in determining petrophysical properties: A case study from Colombia. *Interpretation*, *6*(4), T1067-T1080. https://doi.org/10.1190/INT-2018-0011.1

Jelodar, H., Wang, Y., Orji, R., & Huang, S. (2020). Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach. *IEEE Journal of Biomedical and Health Informatics*, *24*(10), 2733-2742. https://doi.org/10.1109/JBHI.2020.3001216

Jia, R., Lv, Y., Wang, G., Carranza, EmmanuelJohnM., Chen, Y., Wei, C., & Zhang, Z. (2021). A stacking methodology of machine learning for 3D geological modeling with geological-geophysical datasets, Laochang Sn camp, Gejiu (China). *Computers & Geosciences*, *151*, 104754. https://doi.org/10.1016/j.cageo.2021.104754

Jin, X., Wang, G., Tang, P., Hu, C., Liu, Y., & Zhang, S. (2020). 3D geological modelling and uncertainty analysis for 3D targeting in Shanggong gold deposit (China). *Journal of Geochemical Exploration*, *210*, 106442. https://doi.org/10.1016/j.gexplo.2019.106442

Kanal, L. N. (2003). Perceptron. En *Encyclopedia of Computer Science* (pp. 1383-1385). John Wiley and Sons Ltd.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 881-892. https://doi.org/10.1109/TPAMI.2002.1017616

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., & Kumar, V. (2018). Machine learning for the geosciences: Challenges and opportunities. *IEEE Transactions on Knowledge and Data Engineering*, *31*(8), 1544--1554.

Kazemnejad, A. (2019). Transformer Architecture: The Positional Encoding. *kazemnejad.com*. https://kazemnejad.com/blog/transformer_architecture_positional_encoding/

Kernes, J. (2021, marzo 5). *Master Positional Encoding: Part I*. Medium. https://towardsdatascience.com/master-positional-encoding-part-i-63c05d90a0c3

Khyani, D., & B S, S. (2021). An Interpretation of Lemmatization and Stemming in Natural Language Processing. *Shanghai Ligong Daxue Xuebao/Journal of University of Shanghai for Science and Technology*, *22*, 350-357.

Kleijnen, J. P. C. (2009). Kriging metamodeling in simulation: A review. *European Journal of Operational Research*, *192*(3), 707-716. https://doi.org/10.1016/j.ejor.2007.10.013

Krogh, A. (2008). What are artificial neural networks? *Nature Biotechnology*, *26*(2), Art. 2. https://doi.org/10.1038/nbt1386

Kuhn, S., Cracknell, M. J., & Reading, A. M. (2019). Lithological mapping in the Central African Copper Belt using Random Forests and clustering: Strategies for optimised results. *Ore Geology Reviews*, *112*, 103015. https://doi.org/10.1016/j.oregeorev.2019.103015

Lerman, P. M. (1980). Fitting Segmented Regression Models by Grid Search. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *29*(1), 77-84. https://doi.org/10.2307/2346413

Li, H., Li, X., Yuan, F., Jowitt, S. M., Zhang, M., Zhou, J., Zhou, T., Li, X., Ge, C., & Wu, B. (2020). Convolutional neural network and transfer learning based mineral prospectivity modeling for geochemical exploration of Au mineralization within the Guandian–Zhangbaling area, Anhui Province, China. *Applied Geochemistry*, *122*, 104747. https://doi.org/10.1016/j.apgeochem.2020.104747

Lin, Y., Yan, Z., Huang, H., Du, D., Liu, L., Cui, S., & Han, X. (2020). *FPConv: Learning Local Flattening for Point Convolution*. 4293-4302. https://openaccess.thecvf.com/content_CVPR_2020/html/Lin_FPConv_Learning_Local_Flattening_for_

Point_Convolution_CVPR_2020_paper.html

Linderman, G. C., & Steinerberger, S. (2019). Clustering with t-SNE, Provably. *SIAM Journal on Mathematics of Data Science*, *1*(2), 313-332. https://doi.org/10.1137/18M1216134

Lipton, Z. C., Elkan, C., & Narayanaswamy, B. (2014). Thresholding classifiers to maximize F1 score. *arXiv preprint arXiv:1402.1892*.

Lopez Pinaya, W. H., Vieira, S., Garcia-Dias, R., & Mechelli, A. (2020). Chapter 10—Convolutional neural networks. En A. Mechelli & S. Vieira (Eds.), *Machine Learning* (pp. 173-191). Academic Press. https://doi.org/10.1016/B978-0-12-815739-8.00010-9

Luzón, M. M. T. (2018). *Algoritmos avanzados de procesamiento de señal basados en técnicas de deep learning para descripción y caracterización de señales sismo-volcánicas* [Http://purl.org/dc/dcmitype/Text, Universidad de Granada]. https://dialnet.unirioja.es/servlet/tesis?codigo=216488

Ma, J., Ding, Y., Cheng, J. C., Jiang, F., & Wan, Z. (2019). A temporal-spatial interpolation and extrapolation method based on geographic Long Short-Term Memory neural network for PM2. 5. *Journal of Cleaner Production*, *237*, 117729.

Maćkiewicz, A., & Ratajczak, W. (1993). Principal components analysis (PCA). *Computers & Geosciences*, *19*(3), 303-342. https://doi.org/10.1016/0098-3004(93)90090-R

Mahesh, B. (2019). *Machine Learning Algorithms -A Review*. https://doi.org/10.21275/ART20203995

Martínez, A., López, E., & Zapata, N. (Asesor J. (2021). *El government take de la minería de oro en Colombia*. http://www.repository.fedesarrollo.org.co/handle/11445/4102

Martinez, G., Restrepo-Baena, O. J., & Veiga, M. M. (2021). The myth of gravity concentration to eliminate mercury use in artisanal gold mining. *The Extractive Industries and Society*, *8*(1), 477-485. https://doi.org/10.1016/j.exis.2021.01.002

McGOWAN, B. (1996). The Typology and Techniques of Alluvial Mining: The Example of the Shoalhaven and Mongarlowe Goldfields in Southern New South Wales. *Australasian Historical Archaeology*, *14*, 34-45.

Mielke, S. J., Alyafeai, Z., Salesky, E., Raffel, C., Dey, M., Gallé, M., Raja, A., Si, C., Lee, W. Y., Sagot, B., & Tan, S. (2021). *Between words and characters: A Brief History of Open-Vocabulary Modeling and Tokenization*

in NLP (arXiv:2112.10508). arXiv. http://arxiv.org/abs/2112.10508

Minenergía, M. de minas y energía. (2020). *Boletín 2020 Camino a la transparencia*. *2020*, 24.

Minenergía, M. de minas y energía, & UNODC, O. de las N. U. contra la D. y el D. (2021). *Colombia Explotacion de Oro de Aluvion EVOA Evidencias a partir de percepcion remota 2020*. https://www.unodc.org/documents/colombia/2021/Agosto/Colombia_Explotacion_de_Oro_de_Aluvion_EVOA_Evidencias_a_partir_de_percepcion_remota_2020.pdf

OLIVER, M. A., & WEBSTER, R. (1990). Kriging: A method of interpolation for geographical information systems. *International Journal of Geographical Information Systems*, *4*(3), 313-332. https://doi.org/10.1080/02693799008941549

Padarian, J., & Fuentes, I. (2019). Word embeddings for application in geosciences: Development, evaluation, and examples of soil-related concepts. *SOIL*, *5*(2), 177-187. https://doi.org/10.5194/soil-5-177-2019

Parsa, M., & Pour, A. B. (2021). A simulation-based framework for modulating the effects of subjectivity in greenfield Mineral Prospectivity Mapping with geochemical and geological data. *Journal of Geochemical Exploration*, *229*, 106838. https://doi.org/10.1016/j.gexplo.2021.106838

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532-1543. https://doi.org/10.3115/v1/D14-1162

Perceptrón. (2021). En *Wikipedia, la enciclopedia libre*. https://es.wikipedia.org/w/index.php?title=Perceptr%C3%B3n&oldid=139672359

Polidori, L., & El Hage, M. (2020). Digital Elevation Model Quality Assessment Methods: A Critical Review. *Remote Sensing*, *12*(21), Art. 21. https://doi.org/10.3390/rs12213522

Pramerdorfer, C., & Kampel, M. (2016). *Facial Expression Recognition using Convolutional Neural Networks: State of the Art* (arXiv:1612.02903). arXiv. https://doi.org/10.48550/arXiv.1612.02903

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, *63*(10), 1872-1897. https://doi.org/10.1007/s11431-020-1647-3

Rajapakse, R. A. (2016). *Pile design and construction rules of thumb*. Butterworth-Heinemann.

Ringnér, M. (2008). What is principal component analysis? *Nature Biotechnology*, *26*(3), Art. 3. https://doi.org/10.1038/nbt0308-303

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, *71*, 804-818. https://doi.org/10.1016/j.oregeorev.2015.01.001

Saptawati, G. A. P., & Nata, G. N. M. (2015). Knowledge discovery on drilling data to predict potential gold deposit. *2015 International Conference on Data and Software Engineering (ICoDSE)*, 143-147. https://doi.org/10.1109/ICODSE.2015.7436987

Sarica, S., & Luo, J. (2021). Stopwords in technical language processing. *PLOS ONE*, *16*(8), e0254937. https://doi.org/10.1371/journal.pone.0254937

Sarkar, D. (2019). Feature Engineering for Text Representation. En D. Sarkar (Ed.), *Text Analytics with Python: A Practitioner's Guide to Natural Language Processing* (pp. 201-273). Apress. https://doi.org/10.1007/978-1-4842-4354-1_4

Schulz, E., Speekenbrink, M., & Krause, A. (2018). A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, *85*, 1-16. https://doi.org/10.1016/j.jmp.2018.03.001

Sekulic, A., ar, Kilibarda, M., Heuvelink, G., Nikolic, M., & Bajat, B. (2020). Random forest spatial interpolation. *Remote Sensing*, *12*(10), 1687.

Shinde, P. P., & Shah, S. (2018). A Review of Machine Learning and Deep Learning Applications. *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, 1-6. https://doi.org/10.1109/ICCUBEA.2018.8697857

Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised K-Means Clustering Algorithm. *IEEE Access*, *8*, 80716-80727. https://doi.org/10.1109/ACCESS.2020.2988796

Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. *2016 3rd*

International Conference on Computing for Sustainable Global Development (INDIACom), 1310-1315.

Singh, P. (2019). Natural Language Processing. En P. Singh (Ed.), *Machine Learning with PySpark: With Natural Language Processing and Recommender Systems* (pp. 191-218). Apress. https://doi.org/10.1007/978-1-4842-4131-8_9

Sohrab, M. G., & Miwa, M. (2018). *Deep Exhaustive Model for Nested Named Entity Recognition*. 2843-2849. https://doi.org/10.18653/v1/D18-1309

Solangi, Y. A., Solangi, Z. A., Aarain, S., Abro, A., Mallah, G. A., & Shah, A. (2018). Review on Natural Language Processing (NLP) and Its Toolkits for Opinion Mining and Sentiment Analysis. *2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, 1-4. https://doi.org/10.1109/ICETAS.2018.8629198

Strum, R. D., & Kirk, D. E. (1988). *First principles of discrete systems and digital signal processing*. Addison-Wesley Longman Publishing Co., Inc.

Sullivan, G. M., & Feinn, R. (2012). Using Effect Size—Or Why the P Value Is Not Enough. *Journal of Graduate Medical Education*, *4*(3), 279-282. https://doi.org/10.4300/JGME-D-12-00156.1

Sun, T., Chen, F., Zhong, L., Liu, W., & Wang, Y. (2019). GIS-based mineral prospectivity mapping using machine learning methods: A case study from Tongling ore district, eastern China. *Ore Geology Reviews*, *109*, 26-49. https://doi.org/10.1016/j.oregeorev.2019.04.003

Suzuki, K. (Ed.). (2011). *Artificial Neural Networks—Methodological Advances and Biomedical Applications*. InTech. https://doi.org/10.5772/644

Unidad de Planeación Minero Energética -UPME, Ministerio de Minas y Energía, & Sistema de Información Minero Colombiano -SIMCO. (2017). *Producción de oro en kilogramos por departamentos. Años 2004-2016*. Antioquiadatos. http://www.antioquiadatos.gov.co/index.php/9-2-1-produccion-de-oro-en-kilogramos-por-departamentos-anos-2004-2016

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Vega, J., & Taboada, M. (2018). Factores de concentración gravimétrica centrífuga en la recuperación de oro de un mineral carbonáceo aurífero. *SCIÉNDO*, *21*(3), Art. 3. https://doi.org/10.17268/sciendo.2018.028

Vos, C. (2020, julio 31). K-Means Clustering Python Implementation. *Medium*. https://medium.com/@coopervos1/k-means-clustering-python-implementation-4ebec1b96c75

Wang, P., Bai, G. R., & Stolee, K. T. (2019). Exploring Regular Expression Evolution. *2019 IEEE 26th International Conference on Software Analysis, Evolution and Reengineering (SANER)*, 502-513. https://doi.org/10.1109/SANER.2019.8667972

Webster, J. J., & Kit, C. (1992). Tokenization as the Initial Phase in NLP. *COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics*. COLING 1992. https://aclanthology.org/C92-4173

Weisberg, S. (2005). *Applied Linear Regression*. John Wiley & Sons.

Wen, Q., Sun, L., Yang, F., Song, X., Gao, J., Wang, X., & Xu, H. (2021). Time Series Data Augmentation for Deep Learning: A Survey. *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 4653-4660. https://doi.org/10.24963/ijcai.2021/631

What are Convolutional Neural Networks? (2021, enero 6). https://www.ibm.com/cloud/learn/convolutional-neural-networks

Widera, M., Chomiak, L., & Zieliński, T. (2019). Sedimentary Facies, Processes and Paleochannel Pattern of an Anastomosing River System: An Example from the Upper Neogene of Central Poland. *Journal of Sedimentary Research*, *89*(6), 487-507. https://doi.org/10.2110/jsr.2019.28

Wood, J. (1996). *The geomorphological characterisation of digital elevation models*. University of Leicester (United Kingdom).

Xiong, Y., Zuo, R., & Carranza, E. J. M. (2018). Mapping mineral prospectivity through big data analytics and a deep learning algorithm. *Ore Geology Reviews*, *102*, 811-817. https://doi.org/10.1016/j.oregeorev.2018.10.006

Yeomans, C. M., Shail, R. K., Grebby, S., Nykänen, V., Middleton, M., & Lusty, P. A. J. (2020). A machine learning approach to tungsten prospectivity modelling using knowledge-driven feature extraction and model

confidence. *Geoscience Frontiers*, *11*(6), 2067-2081. https://doi.org/10.1016/j.gsf.2020.05.016

Zhang, Z., McDonnell, K. T., Zadok, E., & Mueller, K. (2015). Visual Correlation Analysis of Numerical and Categorical Data on the Correlation Map. *IEEE Transactions on Visualization and Computer Graphics*, *21*(2), 289-303. https://doi.org/10.1109/TVCG.2014.2350494

Zhang, Z., Wang, G., Liu, C., Cheng, L., & Sha, D. (2021). Bagging-based positive-unlabeled learning algorithm with Bayesian hyperparameter optimization for three-dimensional mineral potential mapping. *Computers & Geosciences*, *154*, 104817. https://doi.org/10.1016/j.cageo.2021.104817

Zhu, D., Cheng, X., Zhang, F., Yao, X., Gao, Y., & Liu, Y. (2020). Spatial interpolation using conditional generative adversarial neural networks. *International Journal of Geographical Information Science*, *34*(4), 735--758.