Análisis de muestreos geoquímicos de suelos superficiales usando métodos no supervisados de aprendizaje automático como estrategia para la prospección de yacimientos minerales en Perú.

Integrantes: Diana Urbano, Edinson Fernandez, Daren Rodríguez

1.Resumen:

Entre el año 2000 al 2019 se realizaron diversos muestreos geoquímicos por el Instituto Geológico, Minero y Metalúrgico de Perú, de los cuales se tomó una base de datos con 536 muestras de suelos superficiales para la prospección de yacimientos minerales.

En este análisis se usaron técnicas de aprendizaje automático no supervisado hallando 9 componentes principales que explican el 80% de la varianza y usando el clúster aglomerativo con mejores agrupaciones que el DBSCAN, hallando 13 clústeres de los cuales 4 tienen correlaciones significativas con elementos indicadores de yacimientos minerales, con el clúster 6 indicando la presencia de un yacimiento, asociado a Au, Ag, Hg, Bi mientras que los otros 2 clústeres (8 y 10) se asocian con elementos de interés como: Au, Ag, Cu, Zn, Pb, In entre otros, y en menor medida el clúster 9 se asocia con Mo,V,Ni,Tl indicando yacimientos posiblemente hidrotermales o se sulfuros masivos. Estas agrupaciones delimitan nuevas zonas de prospección minera de 50 km entre la zona de Huancayo y Jauja siguiendo la rivera del rio Mantaro y otra zona puntual cerca a Cochamarca. Para estas dos zonas se recomienda realizar un mapeo geológico más detallado y decidir si se realizan estudios especializados.

2. Introducción:

La base de datos corresponde a 536 muestras geoquímicas de suelos superficiales efectuados en Perú. Para cada muestra se determinaron concentraciones de 58 elementos químicos y 11 óxidos mayores. Con este estudio se pretende generar una delimitación por agrupaciones asociadas a estos elementos químicos y óxidos y si estas agrupaciones pueden indicar o no algún yacimiento mineral, para esto se usan métodos no supervisados como análisis de componentes principales, conglomerados aglomerativos y el agrupamiento espacial basado en densidad de aplicaciones con ruido(DBSCAN) de esta manera se obtienen distribuciones espaciales que facilitan la visualización de áreas con minerales de interés en la prospección minera y permite tener una mejor comprensión del territorio al reducir las áreas de exploración a sectores más puntuales.

Muchos de estos elementos indicadores se conocen como “pathfinders” debido a que son elementos que comúnmente tienden a indicarnos la presencia de algún yacimiento como se observa en la Tabla 1.

Exploración geoquímica

“La exploración minera es un esfuerzo de equipo multidisciplinario que involucra a expertos de diferentes áreas, como la geología, la geofísica, la geoquímica, la petrología y la ingeniería, donde las técnicas geoquímicas, en particular, han contribuido significativamente a los descubrimientos de varios yacimientos minerales. Los datos estadísticos de China revelan que el 71% del total de los depósitos minerales fueron descubiertos por métodos geoquímicos durante el período 1981–2000” (Balaram y Sawant, 2021)

Deposits of Interest	Type of Deposit	Main Pathfinder Minerals	Main Pathfinder Elements
Gold		Pyrite, chalco-pyrite, arsenopyrite, bismuthinite magnetite, tellurides, tetrahedrite, pyrite, sphalerite, muscovite, monazite, bastnäsite, quartz, scheelite, wolframite, cassiterite.	Fe, Mn, Cu, Co, Ni, Sb, Zn, As, Bi, Te, Sn, Se, Tl, Ag, Hg, Pb, Mo and W.
REE	Carbonatite rocks	Bastnäsite group, ancylite, monazite, (fluor)apatite, pyrochlore, xenotime, florencite.	Na, Mg, Fe, P, Ba, F, S, Sr, Ca, Nb, Th, U, Zr, Cu, Ta, Ti, V, Mn, Pb.
	Igneous rocks (including hydrothermal upgrade)	Bastnäsite group, aegirine, eudialyte, loparite, allanite, monazite, fergusonite, zircon, xenotime, fluorapatite, ancylite, gadolinite, euxenite, mosandrite.	Na, K, Fe, Al, Zr, Ti, Nb, Ta, Li, F, Cl, Si, Th, U, P, Cs, Rb, Sn, W, Mo, Be, Ga, Hf, Mn, B.
	Placers and palaeoplacers	Monazite, xenotime, allanite, euxenite.	Ti, Nb, Zr, Au, Sn, Th, U, Pb, F.
	Laterites	Monazite, apatite, pyrochlore, crandallite group, bastnäsite group, churchite, rhabdophane, plumbogummite, zircon, florencite, xenotime, cerianite.	Fe, Al, Nb, Zr, Ti, Sn, Mn, P, low Si, negative Ce anomaly.
	Ion-adsorption	Clay minerals (mainly kaolinite and halloysite).	High Si (>75%), low P.
	Iron oxide-associated (including IOCG) deposits	Bastnäsite, synchysite, monazite, xenotime, florencite, britholite.	Fe, Cu, U, Au, Ag, Ba, F, P, S.
	Seafloor deposits, such as manganese nodules, ferromanganese crust, phosphorite.	Vernadite, todorokite, Fe-oxyhydroxide, carbonate fluorapatite, francolite.	Mn, Fe, P, Cu, Ni, Co.
Cu-Ni-PGE		pentlandite, chalcopyrite, pyrite, millerite, PGM, chromite, Cr-diopside, enstatite, olivine, Cr-andradite.	Ni, Cu, Pd, As, Cr, Co, S, PGE
Volcanogenic massive sulphide (VMS) deposits (Cu, Pb, Zn, Ag, Au)		Galena, sphalerite, chalcopyrite, pyrrhotite, gold, pyrite, gahnite, staurolite, cassiterite, spessartine, sillimanite, andalusite, beudantite, jarosite, barite, tourmaline, hogcomite, nigerite.	Cu, Zn, Pb, Ag, Mo, Sn, Ba As, Sb, In, Te, Bi, and Tl
W-Mo-Bi, and Sn-Zn-In deposits		Cassiterite, wolframite, molybdenite, topaz, chalcopyrite, galena, sphalerite, arsenopyrite, pyrite, loellingite, beudantite, anglesite, plumboferrite, plumbogummite.	Ag, As, Cd, Cu, Pb, Re, Te, Tl
Li		Spodumene, petalite, amblygonite, quartz, K-feldspar, albite, or montebrasite, lepidolite, zinnwaldite, eucryptite, cassiterite, lithiophilite, holmquisite, triphylite, quartz, muscovite, apatite, tourmaline tantalite-columbite, beryl.	K, Ca, Rb, Sr, Y, Nb, Sn, Cs, Ta, Sb, W, Bi, As, Ga, Tl, and the REE
Kimberlite-hosted diamonds	-	Cr-pyrope, Cr-diopside, eclogitic garnet, Mg-ilmenite, chromite, olivine, diamond.	C
U		Uraninite (pitchblende), thorianite, tourmaline, sulphides, monazite, allanite, zircon, baddelyite, niccolite, U-Th anatase, U-Th rutile, brannerite, magnetite.	Cu, Ag, As, Cr, Pb, Zn, Ni, Co, Re, Be, P, Mo, Mn, REE and radiogenic Pb isotopes

Tabla 1. Elementos indicadores de la presencia de yacimientos metálicos según los tipos de depósitos. (Balaram y Sawant, 2021)

Las anomalías geoquímicas del suelo se pueden utilizar para identificar “Pathfinders” en la exploración de depósitos minerales y apoyarse en métodos no supervisados para facilitar la búsqueda de yacimientos minerales como lo sugieren las siguientes investigaciones:

Levitan, y otros, (2015) realizan el análisis de datos de composición de suelos con métodos estadísticos multivariados usando conglomerados jerárquicos y análisis de componentes principales, para analizar los datos geoquímicos del suelo recopilados del depósito de uranio de Coles Hill, Virginia, EE. UU., para identificar los “Pathfinders” asociados con este depósito. Los resultados muestran que los minerales indicadores del depósito de Coles Hill incluyen elementos de tierras raras ligeras (La y Ce), que, cuando se normalizan por su contenido de Al, están correlacionados con U/Al, y valores elevados de Th/Al, que no están correlacionados con U/Al, apoyando el desacoplamiento de U de Th durante la generación del suelo. Estos resultados se pueden utilizar en modelos genéticos y de meteorización y aplicarlos a la prospección de yacimientos similares de Uranio en otras regiones.

Un trabajo similar al anterior se realizó por Nude, y otros, (2012) con un análisis estadístico multivariante en datos geoquímicos del suelo de múltiples elementos de los prospectos de oro Koda Hill-Bulenga en el cinturón de oro de Wa-Lawra, al noroeste de Ghana. El análisis factorial explicó el 79,093% de la varianza total de los datos a través de tres factores. El factor 3 presenta asociaciones de cobre, hierro, plomo y manganeso y explicando el 20.903% de la varianza total. A partir de la agrupación jerárquica, también se observó que el oro se agrupaba con plomo, cobre, arsénico y plata.

Debido a que la base de datos corresponde a información recopilada en Perú, el éxito de estos análisis puede ser de interés para Instituto Geológico, Minero y Metalúrgico del Perú, así como el gobierno y autoridades mineras que otorgan títulos mineros de acuerdo a las características geológicas y prospectivas de cada territorio. Por otra parte, las metodologías pueden ser replicadas en otros contextos geológicos y sirven de herramientas para las universidades, gobiernos, así como empresas privadas sin importar el país de origen, debido a que el éxito o fracaso de estas metodologías depende en gran medida de las condiciones geológicas de cada zona de estudio particular y la ocurrencia natural de yacimientos metálicos.

3. Materiales y Métodos

La base de datos unificada de suelos superficiales contiene inicialmente 540 registros de muestras geoquímicas de suelos superficiales tomadas entre el año 2000 al 2019 que posteriormente luego de la limpieza se reducen a 536 registros de muestras.

Cada muestra de suelo fue analizada en el laboratorio de INGEMMET y se determinaron concentraciones de 58 elementos químicos y 11 óxidos mayores. Las siglas ppm corresponden a partes por millón y pct corresponde al porcentaje.

La información se encuentra disponible para libre descarga por el Instituto Geológico, Minero y Metalúrgico de Perú en su plataforma GEOCATMIN y pertenecen a la serie B Prospección Geoquímica del Perú.

Enlaces de descarga:

https://www.gob.pe/institucion/ingemmet/informes-publicaciones/1423546-base-de-datos-de-geoquimica

Para verificar la ubicación espacial de las muestras se realiza una visualización inicial usando el software libre QGIS y con este se verifica que 3 muestras se encuentran por fuera del territorio de Perú y 1 muestra cae sobre el mar lo cual corresponde a un error en las coordenadas ya que las muestras corresponden a suelos y no sedimentos marinos, por lo cual se eliminan estos 4 registros.

Algunos símbolos como el "<" se refieren al mínimo de detección del instrumento por ejemplo <0.005 para el oro quiere decir que el instrumento no puede detectar valores de este elemento inferiores a 0.005 partes por millón por lo cual para este y los demás elementos se decide reemplazar estos valores por cero y de esta manera poder usar los datos de forma numérica para los análisis.

A continuación, se describen los métodos de limpieza que se realizaron en los datos inicialmente en la base de datos unificada en Excel ver Tabla 2:

Variables eliminadas sin valores registrados(N.R.)	**Variables eliminadas solo un valor valido**	**Valores reemplazados por cero (corresponden al valor mínimo de detección del instrumento)**	**Símbolos eliminados**	**Registros completos eliminados (filas)**
Au_ppb, Hg_ppm, Ag_ppb, Mn_ppm, P_ppm, Ti_ppm, B_ppm, Ge_ppm, Se_ppm, Sn_ppm, Te_ppm, S_pct, Re_ppm, F_ppm	Tl_2_ppm	<0.01, <0.005, <0.5, <2, <0.6, <0.1, <0.3, <0,15	* (se elimina el símbolo * de algunos valores altos de Hg_ppb)	Se eliminan 3 registros cuyas coordenadas no coinciden con el país de origen y una que cae en el mar.

Tabla 2. Limpieza de la base de datos.

La base de datos contiene 536 registros de muestras y 79 variables, de las cuales 10 son informativas y 69 variables corresponden a los elementos y óxidos de interés.

Los histogramas de algunos de algunos elementos muestran que la mayoría de los valores se ubican en cero o muy cerca a cero con muy pocos valores altos Figura 1.

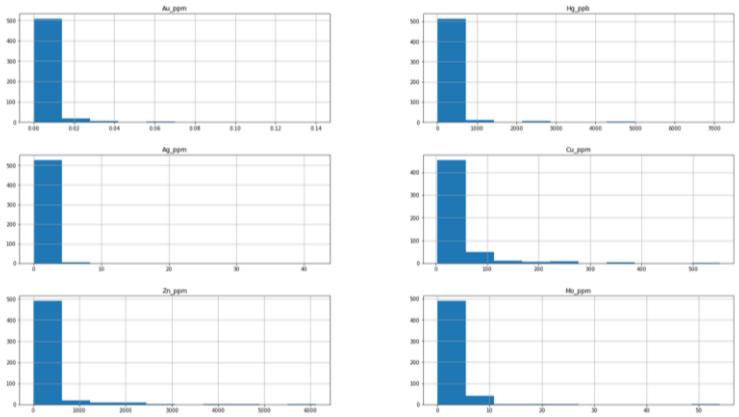

Figura 1. Histogramas para los elementos (Oro (Au), Mercurio(Hg), Plata (Ag), Cobre (Cu), Zinc(Zn) y Molibdeno(Mo).

La descripción estadística de algunos elementos y óxidos se presenta en la Tabla 3.

	Au_ppm	Hg_ppb	Ag_ppm	Al_pct	As_ppm	Ba_ppm	Bi_ppm	Ca_pct	Cd_ppm	Co_ppm	...	CaO_pct	Fe2O3_pct	K2O_pct	MgO_pct	MnO_pct	Na2O_pct	P2O5_pct	SiO2_pct	TiO2_pct	LOI_pct
count	536.000000	536.000000	536.000000	536.000000	536.000000	536.000000	536.000000	536.000000	536.000000	536.000000	...	536.000000	536.000000	536.000000	536.000000	536.000000	536.000000	536.000000	536.000000	536.000000	536.000000
mean	0.003165	207.603595	0.464829	6.522966	39.047795	384.513872	0.179910	2.704459	0.432213	14.133991	...	3.784085	5.010768	1.857802	1.104936	0.122382	0.743754	0.283475	59.939002	0.795169	13.505906
std	0.008886	572.940192	1.965958	1.260892	42.935075	154.948527	3.703582	3.107204	1.289702	4.054684	...	4.347238	1.213802	0.602705	0.626789	0.059605	0.706993	0.153848	7.979961	0.153032	4.936891
min	0.000000	0.000000	0.000000	2.070000	0.000000	139.519800	0.000000	0.030000	0.000000	2.000000	...	0.040000	0.700000	0.470000	0.090000	0.000000	0.000000	0.030000	27.840000	0.404484	2.960000
25%	0.000000	53.000000	0.000000	5.630354	20.985202	274.716000	0.000000	0.509683	0.000000	11.944100	...	0.710000	4.330601	1.400000	0.640000	0.088855	0.220000	0.200000	55.687821	0.692086	9.879711
50%	0.000000	81.000000	0.000000	6.546107	31.320670	359.299000	0.000000	1.550000	0.000000	14.000000	...	2.170000	4.871322	1.825663	0.930000	0.110000	0.513252	0.253939	60.405000	0.789230	12.958746
75%	0.005448	140.000000	0.682452	7.425000	42.205644	457.984663	0.000000	3.650586	0.000000	16.000000	...	5.107858	5.683218	2.290268	1.418924	0.142785	0.973827	0.330000	64.871488	0.890000	16.841557
max	0.140045	7137.500000	41.800000	13.740000	533.000000	991.000000	85.000000	17.510000	13.000000	35.000000	...	24.500000	9.750000	4.830000	4.792454	0.500000	3.670000	1.650000	89.660000	1.620000	30.300000

Tabla 3. Principales datos estadísticos de algunos elementos y óxidos.

Métodos no supervisados:

Se usan 3 métodos no supervisados que se describen a continuación:

El primer método de análisis de componentes principales(PCA) tiene como objetivo la reducción lineal de dimensionalidad. El algoritmo busca crear una representación reducida de los datos, mientras se conserva la mayor cantidad de información posible.

El método aborda la correlación entre las diferentes características. Si la correlación es muy alta entre un subconjunto de características, el método combina las características altamente correlacionadas y representa estos datos con un número menor de características linealmente no correlacionadas. El algoritmo sigue realizando esta reducción de correlación, encontrando las direcciones de máxima variación en los datos originales de alta dimensión y proyectándolos en un espacio dimensional más pequeño. Estos nuevos componentes generados se conocen como componentes principales (Patel, 2019).

El segundo método hace parte del aprendizaje no supervisado usando el análisis de clúster, que se encarga de formar grupos diferentes dentro de los datos. Al usar algoritmos de agrupamiento se encuentra la estructura en los datos, de manera que los elementos del mismo clúster o agrupación sean más similares entre sí que con los otros grupos generados (Román, 2019).

El clúster aglomerativo es un tipo de clúster jerárquico en el cual el agrupamiento se inicia con todas las observaciones separadas, cada una formando un clúster individual. Los clústeres se van combinado a medida que la estructura crece hasta converger en uno solo (Amat, 2020).

Para este método se usa el enlace o linkage Ward el cual considera la posibilidad de la unión de cada par de grupos y opta por la unión de aquellos dos grupos que menos incrementen la suma de los cuadrados de las desviaciones al unirse y además suele ser mas discriminativo en los niveles de agrupación que otros métodos.

El tercer método es el agrupamiento espacial basado en densidad con ruido(DBSCAN). Este se encarga de clasificar las observaciones en tres tipos:

Puntos core: son aquellos puntos que cumplen con las condiciones de densidad que hayamos fijado.

Puntos alcanzables: son aquellos puntos que, aun no cumplen con las condiciones de densidad, pero tienen cerca otros puntos core.

Ruido: son los puntos que no cumplen con las condiciones de densidad y, además, en su radio no tienen otros puntos.

Para el DBSCAN se calcula la matriz de distancias entre los distintos puntos. Generalmente se utiliza la distancia Euclídea, aunque se pueden usar otras. Teniendo en cuenta los parámetros del modelo, clasifica a cada punto entre punto core, punto frontera y ruido. En este sentido, puede que salgan diferentes puntos core ya que puede haber varias zonas de densidad. Cada uno de esos puntos core pertenecerá a un clúster y se asigna los núcleos alcanzables de cada clúster al clúster. (Fernández Jauregui, s.f.)

Para este algoritmo se usa el parámetro eps: en el cual dos puntos se consideran vecinos si la distancia entre los dos puntos está por debajo del umbral épsilon y min_samples: el número mínimo de vecinos que debe tener un punto dado para ser clasificado como un punto central.

Como distancia se usa la distancia euclídea.

4. Resultados y Discusión

Se encuentran 9 componentes principales, con lo cual se reducen las 69 variables a 9 conservando el 80% de la varianza de los datos.

Con los 9 componentes principales se genera un dendograma usando el método Ward y una distancia de 30, cortando 13 clústeres Figura 2.

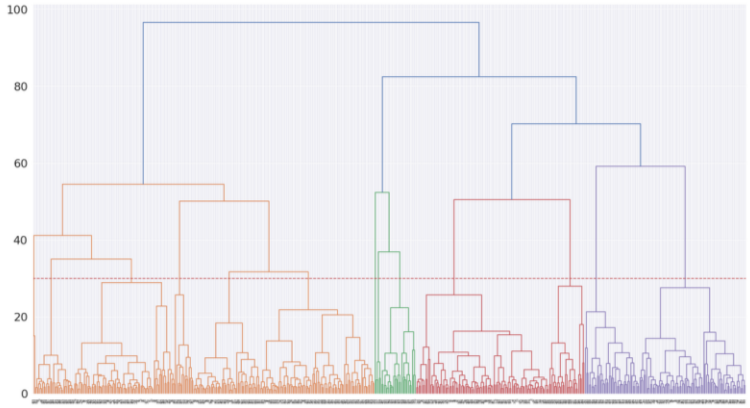

Figura 2. Dendograma usando las 9 componentes principales y una distancia de 30.

Usando el linkage Ward, los 9 componentes principales y 13 clústeres se desarrolla el algoritmo de clúster aglomerativo, Figura 3.

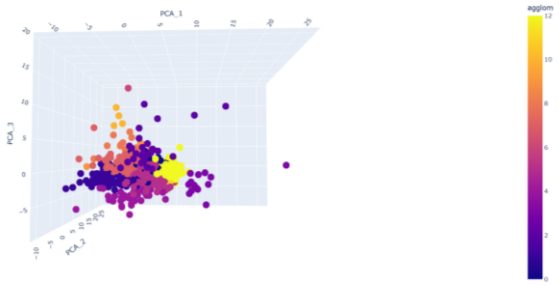

Figura 3. Agrupación de 13 clústeres usando el método aglomerativo.

Para el algoritmo DBSCAN se usa un eps de 4.48 usando la librería kneed para hallar los clústeres. Las gráficas muestran muy poca separación en los datos y tienden a agrupar todos los valores en un solo clúster por lo cual se descarta este método ya que el objetivo es ver diferencias entre los diferentes clústeres y su relación con yacimientos de interés, solo se logra una división en 3 clústeres usando min samples =2 Figura 4.

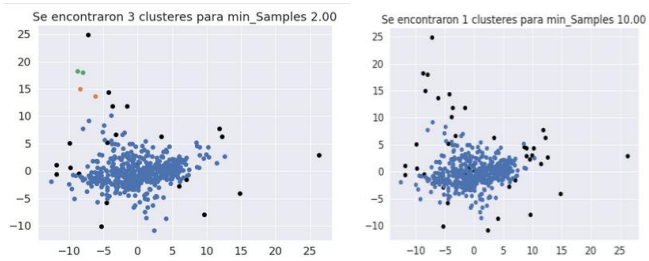

Figura 4. 3 clústeres usando el método DBSCAN para min samples =2.

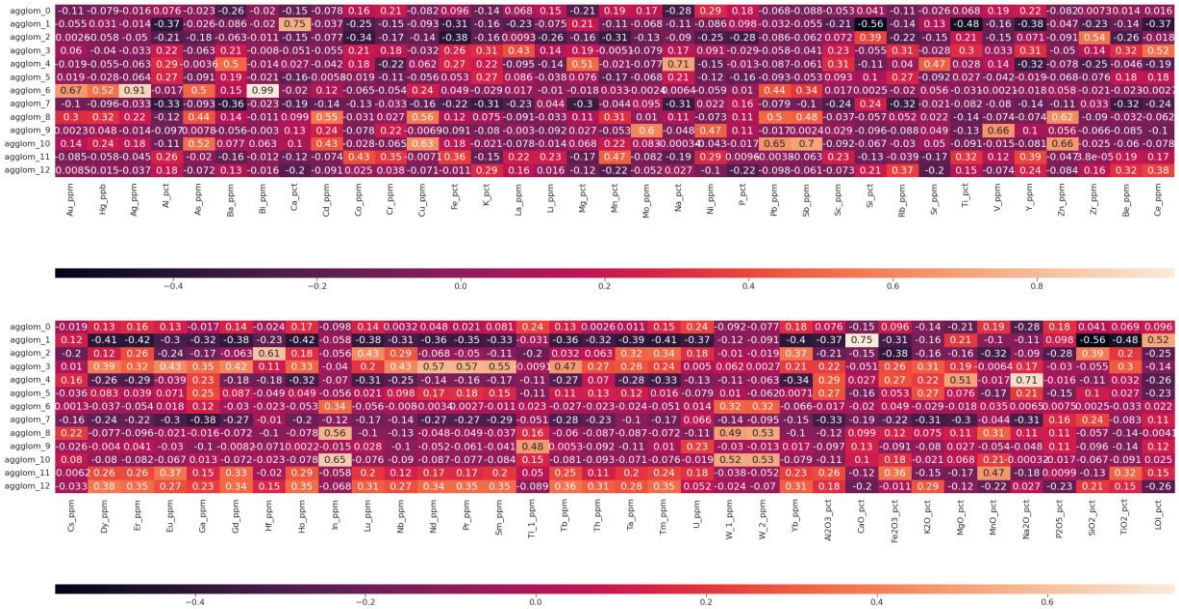

Figura 5. Correlación entre los 13 clúster con los diferentes elementos y óxidos.

En la Figura 5 se presenta la correlación entre los diferentes clústeres con los elementos indicadores y óxidos.

De los 13 clústeres que se usan con el método aglomerativo, 4 tienen resultados prometedores debido a su asociación con minerales indicadores. El clúster 6 con un punto tiene una correlación moderada a muy alta con los elementos Au, Ag, Hg y Bi por lo que esta zona entre el sector de Cochamarca y la reserva de Junín puede ser un buen prospecto para la búsqueda de yacimientos de oro hidrotermales o de sulfuros masivos en la zona (Punto rojo) Figura 6. Los otros 2 clústeres 8 y 10 se asocian a sulfuros masivos e hidrotermales con elementos de interés como: Oro, Plata, Cobre, Zinc, Plomo, indio entre otros. Estas agrupaciones delimitan nuevas zonas de prospección minera o nuevos "Target" de 50 km entre la zona de Huancayo y Jauja, siguiendo la rivera del rio Mantaro (puntos amarillos y magenta) Figura 6. El clúster 9 con dos puntos presenta baja a moderada correlación con Mo, V, Ni, Tl que se asocian con yacimientos de oro y de sulfuros masivos en esta misma franja al sur de Huancayo (puntos aguamarina)Figura 6.

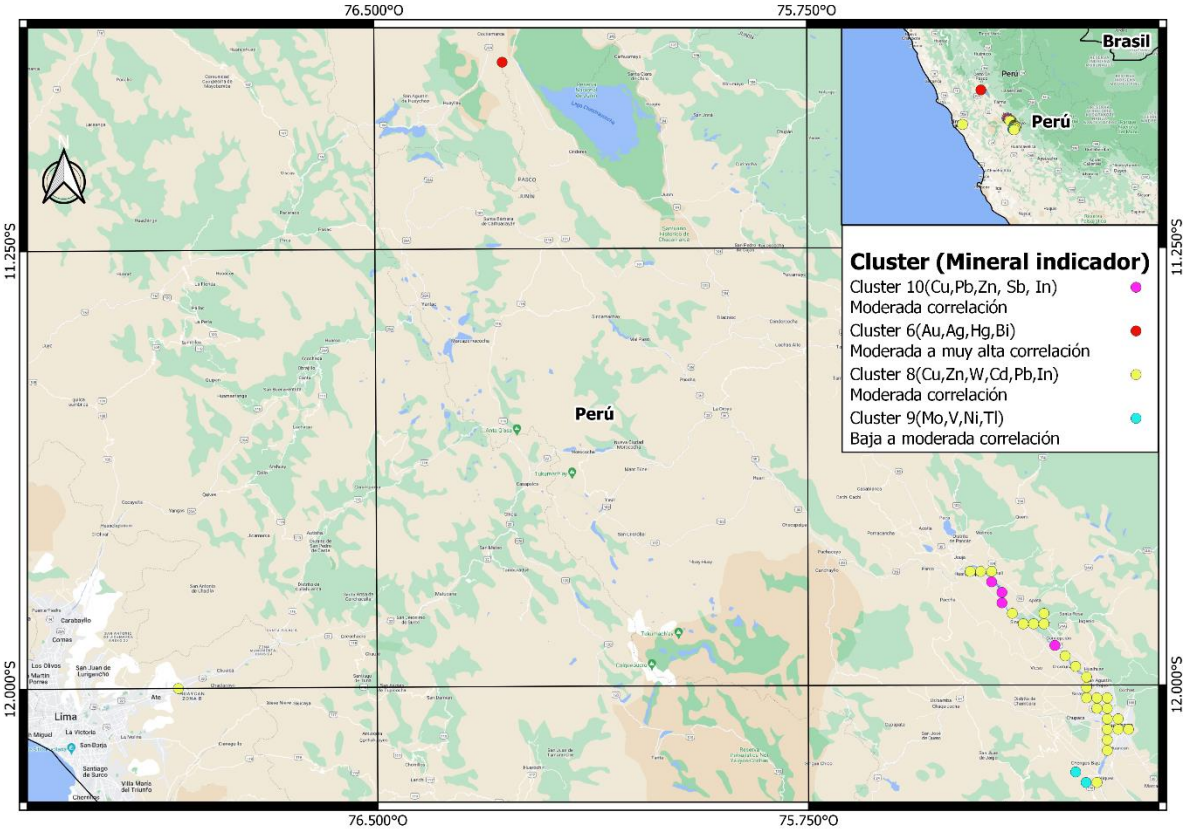

Figura 6. Los 4 clústeres que se pueden usar como indicadores de yacimientos minerales.

Para estas dos zonas se recomienda realizar un mapeo geológico más detallado preliminar y posteriormente decidir si ampliar el estudio con métodos geofísicos, sondajes exploratorios, así como campañas de muestreo más densas para delimitar aún más los posibles yacimientos minerales.

5. Conclusión

Se realizó el análisis de 536 muestras geoquímicas de suelo superficial en Perú con fines de prospección de yacimientos minerales, para este análisis se usaron técnicas de aprendizaje automático no supervisado hallando 9 componentes principales que explican el 80% de la varianza y usando el clúster aglomerativo con mejores agrupaciones que el DBSCAN, usando 13 clústeres de los cuales 4 tienen correlaciones significativas con elementos indicadores asociados a yacimientos minerales, de estos el clúster 6 puede indicar la presencia de un yacimiento, asociado a Au, Ag ,Hg, Bi, mientras que los otros 2 clústeres (8 y 10) se asocian con elementos de interés como: Au, Ag, Cu, Zn, Pb, In entre otros, y en menor medida el clúster 9 se asocia a Mo, V, Ni, Tl indicando todos estos clúster yacimientos posiblemente hidrotermales o de sulfuros masivos. Estas agrupaciones delimitan una nueva zona de prospección minera de 50 km entre la zona de Huancayo y Jauja siguiendo la rivera del rio Mantaro y otra zona puntual cerca a Cochamarca. Para estas dos zonas se recomienda realizar un mapeo geológico más detallado preliminar y posteriormente decidir si ampliar el estudio con métodos geofísicos, sondajes exploratorios, así como campañas de muestreo más densas para delimitar aún más los posibles yacimientos minerales.

6. Bibliografía

Amat, J. (2020). Clustering con Python. https://www.cienciadedatos.net/documentos/py20-clustering-con-python.html

Balaram, V.; Sawant, S.S. (2021). Indicator Minerals, Pathfinder Elements, and Portable Analytical Instruments in Mineral Exploration Studies. Minerals 2022, 12, 394.

Fernández Jauregui, A. (s.f.). *DBSCAN en Python: aprende cómo funciona*. Recuperado el 9 de 2022, de https://anderfernandez.com/blog/dbscan-py

Instituto Geológico, Minero y Metalúrgico de Perú. (2021). GEOCATMIN - Prospección Geoquímica del Perú, Serie B. Obtenido de http://metadatos.ingemmet.gob.pe:8080/geonetwork/srv/spa/catalog.search#/metadata/b1cc5e47-88c8-4c5e-832f-f6dcbd20211b

Levitan, D.M.; Zipper, C.E.; Donovan, P.; Schreiber, M.E.; Seal, R.R.; Engle, M.A.; Chermak, J.A.; Bodnar, R.J.; Johnson, D.K.; Aylor, J.G(2015). Statistical analysis of soil geochemical data to identify pathfinders associated with mineral deposits: An example from the Coles Hill uranium deposit, Virginia, USA. *Journal of Geochemical Exploration*, *154*, 238–251.

Nude, P. M., Asigri, J. M., Yidana, S. M., Arhin, E., Foli, G., & Kutu, J. M. (2012). Identifying pathfinder elements for gold in multi-element soil geochemical data from the Wa-Lawra belt, northwest Ghana: A multivariate statistical approach. International Journal of Geosciences, 3, 62-70.

Patel, A. (2019). *Hands-On unsupervised learning using Python: How to build applied Machine Learning Solutions from unlabeled data*. O'Reilly Media.

Román, V. (12 de junio de 2019). Aprendizaje no supervisado en Machine Learning: agrupación. https://medium.com/datos-y-ciencia/aprendizaje-no-supervisado-en-machine-learning-agrupaci%C3%B3n-bb8f25813edc.