Entregable 3 GRUPO 8

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. Limpieza de datos

La base de datos unificada de suelos superficiales contiene inicialmente 540 registros que posteriormente luego de la limpieza se reducen a 536 registros de muestras.

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.

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 1:

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 1. Limpieza de la base de datos.

La base de datos limpia contiene 536 registros de muestras y 79 variables, de las cuales 10 son informativas y 58 variables corresponden a los elementos, 10 a óxidos y 1 a volátiles Tabla 2:

Variables base de datos original		Variables después de limpieza	Tipo de variable	Descripción
ID	**Sm_ppm**	Muestra	Categórica/alfanumérica	Código de la muestra
Código	**Tl_1_ppm**	Codigo_estandar	Categórica/alfanumérica	Código de la muestra según la región
Código estandarizado	**Tl_2_ppm**	Longitud	Numérica/float	Longitud en origen de coordenadas WGS84
Tipo de muestra	**Tb_ppm**	Latitud	Numérica/float	Latitud en origen de coordenadas WGS84
Longitud_X	**Th_ppm**	UTM_E	Numérica/float	Coordenada Este en coordenadas UTM
Latitud_Y	**Ta_ppm**	UTM_N	Numérica/float	Coordenada Norte en coordenadas UTM
UTM_E	**Tm_ppm**	Zona	Categórica/int	Zona origen coordenadas UTM
UTM_N	**U_ppm**	Region	Categórica/string	Región donde se tomó la muestra en Perú
Zona	**W_1_ppm**	Region_Hidrografica	Categórica/string	Región hidrográfica donde se tomó la muestra
Hoja	**W_2_ppm**	Cuenca	Categórica/Sting	Cuenca donde se tomó la muestra en Perú
Nombre del proyecto	**Yb_ppm**	Au_ppm	Numérica/float	Oro en Partes por millón
Año del proyecto	**Al_2O_3_pct**	Hg_ppb	Numérica/float	Mercurio en partes por millón
N° Boletín	**CaO_pct**	Ag_ppm	Numérica/float	Plata en partes por millón
Serie	**Fe_2O_3_pct**	Al_pct	Numérica/float	Aluminio en porcentaje
Año de Publicación	**K_2O_pct**	As_ppm	Numérica/float	Arsénico en partes por millón
Región	**MgO_pct**	Ba_ppm	Numérica/float	Bario en partes por millón
Región Hidrográfica	**MnO_pct**	Bi_ppm	Numérica/float	Bismuto en partes por millón
Cuenca	**Na_2O_pct**	Ca_pct	Numérica/float	Calcio en porcentaje
Franja metalogenética	**P_2O_5_pct**	Cd_ppm	Numérica/float	Cadmio en partes por millón
Laboratorio	**SiO_2_pct**	Co_ppm	Numérica/float	Cobalto en partes por millón
Roca total	**TiO_2_pct**	Cr_ppm	Numérica/float	Cromo en partes por millón
Multielemental	**LOI_pct**	Cu_ppm	Numérica/float	Cobre en partes por millón
Tierras raras (c)	**B_ppm**	Fe_pct	Numérica/float	Hierro en porcentaje
Análisis de oro por AAS	**Ge_ppm**	K_pct	Numérica/float	Potasio en porcentaje
Análisis de mercurio por vapor frío	**Se_ppm**	La_ppm	Numérica/float	Lantano en partes por millón
Au_ppb	**Sn_ppm**	Li_ppm	Numérica/float	Litio en partes por millón
Au_ppm	**Te_ppm**	Mg_pct	Numérica/float	Magnesio en porcentaje
Hg_ppb	**S_pct**	Mn_pct	Numérica/float	Manganeso en partes por millón
Hg_ppm	**Re_ppm**	Mo_ppm	Numérica/float	Molibdeno en partes por millón

Ag_ppb	**F_ppm**	Na_pct	Numérica/float	Sodio en porcentaje
Ag_ppm		Ni_ppm	Numérica/float	Níquel en partes por millón
Al_pct		P_pct	Numérica/float	Fosforo en partes por millón
As_ppm		Pb_ppm	Numérica/float	Plomo en partes por millón
Ba_ppm		Sb_ppm	Numérica/float	Antimonio en partes por millón
Bi_ppm		Sc_ppm	Numérica/float	Escandio en partes por millón
Ca_pct		Si_pct	Numérica/float	Silicio en porcentaje
Cd_ppm		Rb_ppm	Numérica/float	Rubidio en partes por millón
Co_ppm		Sr_ppm	Numérica/float	Estroncio en partes por millón
Cr_ppm		Ti_pct	Numérica/float	Titanio en porcentaje
Cu_ppm		V_ppm	Numérica/float	Vanadio en partes por millón
Fe_pct		Y_ppm	Numérica/float	Itrio en partes por millón
K_pct		Zn_ppm	Numérica/float	Zinc en partes por millón
La_ppm		Zr_ppm	Numérica/float	Circón en partes por millón
Li_ppm		Be_ppm	Numérica/float	Berilio en partes por millón
Mg_pct		Ce_ppm	Numérica/float	Cerio en partes por millón
Mn_pct		Cs_ppm	Numérica/float	Cesio en partes por millón
Mn_ppm		Dy_ppm	Numérica/float	Disprosio en partes por millón
Mo_ppm		Er_ppm	Numérica/float	Erbio en partes por millón
Na_pct		Eu_ppm	Numérica/float	Europio en partes por millón
Ni_ppm		Ga_ppm	Numérica/float	Galio en partes por millón
P_pct		Gd_ppm	Numérica/float	Gadolinio en partes por millón
P_ppm		Hf_ppm	Numérica/float	Hafnio en partes por millón
Pb_ppm		Ho_ppm	Numérica/float	Holmio en partes por millón
Sb_ppm		In_ppm	Numérica/float	Indio en partes por millón
Sc_ppm		Lu_ppm	Numérica/float	Lutecio en partes por millón
Si_pct		Nb_ppm	Numérica/float	Niobio en partes por millón
Rb_ppm		Nd_ppm	Numérica/float	Neodimio en partes por millón
Sr_ppm		Pr_ppm	Numérica/float	Praseodimio en partes por millón
Ti_ppm		Sm_ppm	Numérica/float	Samario en partes por millón
Ti_pct		Tl_1_ppm	Numérica/float	Talio en partes por millón
V_ppm		Tb_ppm	Numérica/float	Terbio en partes por millón
Y_ppm		Th_ppm	Numérica/float	Torio en partes por millón
Zn_ppm		Ta_ppm	Numérica/float	Tantalio en partes por millón
Zr_ppm		Tm_ppm	Numérica/float	Tulio en partes por millón
Be_ppm		U_ppm	Numérica/float	Uranio en partes por millón
Ce_ppm		W_1_ppm	Numérica/float	Wolframio tipo 1 en partes por millón
Cs_ppm		W_2_ppm	Numérica/float	Wolframio tipo 2 en partes por millón
Dy_ppm		Yb_ppm	Numérica/float	Iterbio en partes por millón
Er_ppm		Al2O3_pct	Numérica/float	Oxido de aluminio en porcentaje
Eu_ppm		CaO_pct	Numérica/float	Oxido de calcio en porcentaje
Ga_ppm		Fe2O3_pct	Numérica/float	Oxido de hierro(III) en porcentaje
Gd_ppm		K2O_pct	Numérica/float	Oxido de potasio en porcentaje
Hf_ppm		MgO_pct	Numérica/float	Oxido de magnesio en porcentaje
Ho_ppm		MnO_pct	Numérica/float	Oxido de manganeso en porcentaje
In_ppm		Na2O_pct	Numérica/float	Oxido de sodio en porcentaje
Lu_ppm		P2O5_pct	Numérica/float	Pentóxido de difósforo en porcentaje
Nb_ppm		SiO2_pct	Numérica/float	Sílice en porcentaje
Nd_ppm		TiO2_pct	Numérica/float	Oxido de titanio en porcentaje
Pr_ppm		LOI_pct	Numérica/float	Volátiles en porcentaje

Tabla 2. Comparativa de las variables antes y después de la limpieza, así como los tipos de datos en la base limpia.

A continuación, se describen las principales características estadísticas de algunos de los principales elementos (Oro (Au), Plata (Ag), Cobre (Cu), Zinc (Zn) y Molibdeno(Mo) Tabla 3:

	Au_ppm	**Ag_ppm**	**Cu_ppm**	**Zn_ppm**	**Mo_ppm**
count	536	536	536	536	536
mean	0.003165	1.764307e+08	1.394794e+08	1.141555e+08	116606.102612
std	0.008886	6.366175e+08	2.028466e+08	2.180378e+08	198725.586928
min	0.000000	0.000000e+00	3.000000e+00	1.300000e+01	0.000000
25%	0.000000	0.000000e+00	2.875000e+01	1.210000e+02	0.000000
50%	0.000000	0.000000e+00	5.650000e+01	2.330000e+02	2.000000
75%	0.005448	6.824524e-01	2.883085e+08	1.449035e+08	250222.500000
max	0.140045	5.231002e+09	9.586025e+08	9.966708e+08	823411.000000

Tabla 3. Descripción estadística general para los elementos (Oro (Au), Plata (Ag), Cobre (Cu), Zinc(Zn) y Molibdeno(Mo).

2. Propuesta metodológica:

Inicialmente se usan 2 métodos no supervisados usando aprendizaje automático con código en Python para generar las agrupaciones de interés en los datos de los muestreos geoquímicos de suelos superficiales.

El primer método corresponde al análisis de componentes principales que será el primer insumo al hacer una reducción lineal de la dimensionalidad para posteriormente aplicar el

método de clustering aglomerativo. Estos dos métodos se usarán inicialmente debido a que no se evidencian estudios de este tipo relacionados a estas muestras en el Perú y son métodos adecuados para análisis geoquímico de muestras de suelo asociadas a yacimientos minerales como lo citan los siguientes estudios: 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.

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. Los objetivos del estudio fueron definir las relaciones del oro con otros elementos traza para determinar posibles elementos “Pathfinders” del oro a partir de los datos geoquímicos del suelo. El estudio se centró en siete elementos, a saber, Au, Fe, Pb, Mn, Ag, As y Cu. Se realizaron análisis factoriales y análisis de conglomerados jerárquicos en las muestras analizadas.

2.1 Método PCA (análisis de componentes principales)

Este método de análisis de componentes principales 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).

2.2 Método clúster aglomerativo

El 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).

3. Implementación algoritmo

De acuerdo con la varianza acumulada se decide usar 17 componentes principales que explican el 80% de la variabilidad de los datos, posteriormente con estos 17 componentes se construye el modelo de cluster aglomerativo usando 13, 9 y 5 cluster y linkage Ward,

llegando a un muy buen resultado agrupando los diferentes grupos con 5 cluster como lo muestra la Figura 1.

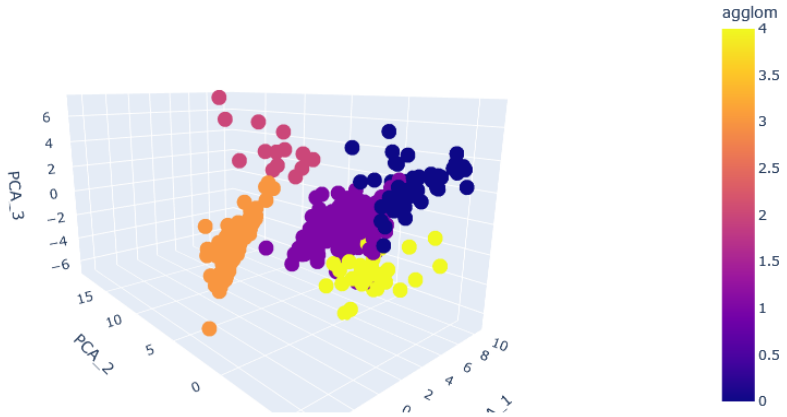

Figura 1. Se observan los 5 clúster usando las primeras 3 componentes principales.

Cluster	agglom_0		agglom_1		agglom_2		agglom_3		agglom_4
Sc_ppm	-34,27%	Al_pct	55,79%	Ag_ppm	77,92%	Al_pct	-92,50%	Cr_ppm	84,54%
Ho_ppm	57,75%	Sc_ppm	-58,27%	Cd_ppm	73,11%	Cu_ppm	-80,92%	Li_ppm	82,45%
Ta_ppm	72,28%	Si_pct	57,91%	In_ppm	79,27%	Fe_pct	-88,84%	Rb_ppm	59,84%
		Be_ppm	55,15%	W_1_ppm	66,68%	K_pct	-77,81%		
		Cs_ppm	55,40%	W_2_ppm	82,86%	La_ppm	-60,54%		
						Mo_ppm	-69,05%		
						Pb_ppm	-79,07%		
						Sb_ppm	-72,32%		
						Sc_ppm	89,88%		
						Si_pct	-94,10%		
						Y_ppm	-62,80%		
						Zn_ppm	-61,61%		
						Be_ppm	-88,21%		
						Ce_ppm	-88,84%		
						Cs_ppm	-80,43%		
						Dy_ppm	-88,86%		
						Er_ppm	-89,22%		
						Eu_ppm	-68,75%		
						Ga_ppm	-87,56%		
						Gd_ppm	-85,90%		
						Hf_ppm	-87,35%		
						Nb_ppm	-90,56%		
						Nd_ppm	-89,28%		
						Pr_ppm	-89,26%		
						Sm_ppm	-89,26%		
						Th_ppm	-91,09%		
						U_ppm	-86,81%		
						Yb_ppm	-84,81%		
						Fe2O3_pct	-86,32%		
						K2O_pct	-83,14%		
						SiO2_pct	-90,11%		

Tabla 4. Los 5 clúster con sus valores mas altos de correlación con diferentes elementos y óxidos.

En la Tabla 4 se muestran los 5 cluster y su correlación con diferentes elementos y óxidos, de esta manera el cluster agglom_3 tiene la mayor cantidad de elementos y óxidos con valores de moderados a muy altos teniendo correlación con 28 elementos y 3 óxidos, el cluster agglom_4 tiene valores moderados a muy altos de correlación con 3 elementos, para el caso del agglom_ 2 son 4 elementos con correlación moderada a alta, el agglom_0 tiene correlación baja a moderada con 3 elementos y el agglom_1 tiene correlación leve a moderada con 4 elementos.

El siguiente paso es asociar estas correlaciones con posibles “Pathfinders” o minerales indicadores de diferentes tipos de yacimientos minerales y contrarrestarlo con la información geológica del Perú y dar recomendaciones que puedan ser útiles parar la prospección minera.

Bibliografía

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

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.