



PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE  
ESCUELA DE INGENIERIA

# **GENERACIÓN DE UN MODELO GEOMETALÚRGICO DE UN YACIMIENTO DE COBRE EN BASE A TÉCNICAS DE APRENDIZAJE DE MÁQUINAS**

**YUYANG MU**

Tesis para optar al grado de  
Magíster en Ciencias de la Ingeniería

Profesor Supervisor:  
**FELIPE NÚÑEZ**  
**JUAN CARLOS SALAS**

Santiago de Chile, (Junio, 2022)

© 2022, Yuyang Mu



PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE  
ESCUELA DE INGENIERIA

# GENERACIÓN DE UN MODELO GEOMETALÚRGICO DE UN YACIMIENTO DE COBRE EN BASE A TÉCNICAS DE APRENDIZAJE DE MÁQUINAS

YUYANG MU

Tesis presentada a la Comisión integrada por los profesores:

**FELIPE NÚÑEZ**

DocuSigned by:  
*Felipe Núñez*  
A0D15E927B2D4D4...

**JUAN CARLOS SALAS**

DocuSigned by:  
*Juan Carlos Salas*  
866D700CEE1E482...

**CARLOS MARQUARDT**

DocuSigned by:  
*Carlos Marquardt*  
F6ADA050565844E...

**ROBERTO PARRA**

DocuSigned by:  
*Roberto Parra*  
D34EA758387D418...

**DIEGO CELENTANO**

DocuSigned by:  
*Diego Celentano*  
91CFB35EC4C442F...

Para completar las exigencias del grado de  
Magíster en Ciencias de la Ingeniería

Santiago de Chile, (Junio, 2022)

*A mi familia, profesores y amigos.*

## **AGRADECIMIENTOS**

En primer lugar, me gustaría agradecer a mi familia, por el continuo apoyo en mis estudios.

En segundo lugar, a los profesores Juan Carlo Salas y Felipe Núñez, profesores guía, por su invitación al programa, la paciencia y constante apoyo durante todo el proceso de investigación, así como la motivación entregada, cuidado durante la pandemia y la gestión realizada para cumplir mis objetivos.

En tercer lugar, al Geólogo Ramon Freire y al Profesor Julio Beniscelli por su generoso apoyo y consejos en el desarrollo de esta tesis.

Quiero agradecer también a mis amigos, sus apoyos y acompañamiento durante este periodo importante de mi vida.

Finamente, agradezco a los Departamentos de Ingeniería de Minería e Ingeniería Eléctrica por las facilidades y espacios otorgados.

INDICE GENERAL

	Pág.
DEDICATORIA .....	ii
AGRADECIMIENTOS .....	iii
INDICE DE FIGURAS.....	vi
INDICE DE TABLAS .....	vii
ABSTRACT .....	viii
RESUMEN.....	ix
1. INTRODUCCIÓN.....	1
1.1. Geometalurgia y modelos geometalúrgicos .....	1
1.2. Estado del arte en la generación de modelos geometalúrgicos .....	3
1.2.1. Métodos tradicionales .....	4
1.2.2. Inteligencia artificial .....	7
1.3. Inteligencia artificial e industria minera chilena .....	15
1.4. Objetivos .....	16
2. BASES CONCEPTUALES .....	19
2.1. Técnicas y algoritmos .....	19
2.1.1. Preparación .....	19
2.1.2. Preprocesamiento.....	19
2.1.3. Reducción de dimensiones .....	20
2.1.4. Clasificación .....	23
2.2. Criterios de medición de desempeño comparativo entre modelos.....	31
2.2.1. Notación.....	31
2.2.2. Definición .....	31
2.3. Arquitectura del modelo.....	33
3. METODOLOGÍA DE PROCESAMIENTO DE DATOS .....	35
3.1. Bases de datos y software .....	35

3.2.	Preparación de datos de entrada.....	38
3.3.	Preprocesamiento.....	40
3.3.1.	<i>Clean</i> .....	40
3.3.2.	Normalización.....	41
3.4.	Reducción de dimensiones.....	41
3.4.1.	<i>Principal Component Analysis</i> (PCA).....	43
3.4.2.	<i>Autoencoder</i> .....	44
3.5.	Formalización de la base de entrada a la clasificación .....	47
3.6.	Clasificación e hiperparámetros.....	50
3.6.1.	<i>k-means</i> .....	50
3.6.2.	Agrupación jerárquica (AGG) .....	51
3.6.3.	<i>Density-based spatial clustering of applications with noise</i> (DBSCAN) 51	
3.6.4.	Mapas autoorganizados (SOM) .....	52
3.7.	Métricas.....	52
4.	RESULTADOS DEL MODELAMIENTO GEOMETALÚRGICO Y ANÁLISIS DE VALOR ECONÓMICO .....	54
4.1.	Resultados generales .....	54
4.2.	Análisis de sensibilidad.....	56
4.3.	Resultado final y análisis de varianza .....	59
4.4.	Evaluación económica .....	64
4.4.1.	Filtración.....	64
4.4.2.	Modelo de bloque .....	65
4.4.3.	Interpolación .....	66
4.4.4.	Modelo final y resultados económicos .....	70
5.	CONCLUSIONES.....	76
5.1.	Conclusiones .....	76
5.2.	Aplicabilidad.....	78
5.3.	Trabajo futuro.....	79
	BIBLIOGRAFIA .....	81

## INDICE DE FIGURAS

	Pág.
Figura 1-1. Modelo geometalúrgico.....	3
Figura 2-1. Arquitectura de <i>autoencoder</i> . ....	22
Figura 2-2. Ejemplo de <i>k-means</i> . ....	25
Figura 2-3. Ejemplo de agrupación jerárquica. ....	27
Figura 2-4. Ejemplo de DBSCAN. ....	29
Figura 2-5. Ejemplo de Mapa autoorganizado. Adaptado de quora.com.....	30
Figura 2-6. Arquitectura de modelos. ....	34
Figura 3-1. Grupos de variables. ....	40
Figura 3-2. Grupos de variables después de <i>clean</i> . ....	41
Figura 3-3. Flujo de PCA y PCA_inversa.....	46
Figura 3-4. Combinación de la base de datos de entrada. ....	48
Figura 3-5. Combinación de inclusión de las variables espaciales. ....	50
Figura 4-1. Análisis de sensibilidad en <i>k-means</i> con el indicador <i>Silhouette</i> . ....	57
Figura 4-2. Análisis de sensibilidad en <i>k-means</i> con el indicador <i>Calinski Harabasz</i> . ...	57
Figura 4-3. Visualización espacial de modelo geometalúrgico resultado – I. ....	60
Figura 4-4. Visualización espacial de modelo geometalúrgico resultado – II. ....	60
Figura 4-5. Diagrama de cajas – Cobre total en cada unidad geometalúrgica. ....	61
Figura 4-6. Diagrama de cajas – Recuperación metalúrgica en cada unidad geometalúrgica. ....	62
Figura 4-7. Diagrama de cajas – Índice de trabajo de Bond en cada unidad geometalúrgica. ....	62
Figura 4-8. Flujograma de filtro 6/7.....	65
Figura 4-9. Ejemplo de un casco convexo. ....	66
Figura 4-10. Visualización espacial de modelo geometalúrgico interpolado. ....	67
Figura 4-11. Visualización planta de modelo geometalúrgico interpolado. ....	67
Figura 4-12. Proceso de selección de algoritmo supervisado. ....	69
Figura 4-13. Visualización espacial límite de corte. ....	71
Figura 4-14. Visualización espacial modelo ajustado.....	71
Figura 4-15. Visualización planta modelo ajustado.....	72

## INDICE DE TABLAS

	Pág.
Tabla 1-1. Descripción de los modelos predictivos propuestos .....	6
Tabla 1-2. Resultado del coeficiente de correlación de Pearson para cada modelo .....	7
Tabla 1-3. Resultado de la validación cruzada quíntuple .....	10
Tabla 2-1. Función de activación. ....	23
Tabla 3-1. Estadísticas descriptivas para variables en la base de datos – I. ....	36
Tabla 3-2. Estadísticas descriptivas para variables en la base de datos – II. ....	36
Tabla 3-3. Estadísticas descriptivas para variables en la base de datos – III. ....	37
Tabla 3-4. Estadísticas descriptivas para variables en la base de datos – IV. ....	37
Tabla 3-5. Librerías de <i>python</i> utilizadas. ....	38
Tabla 3-6. Grupos de variables que se reducen a 1 dimensión. ....	42
Tabla 3-7. Información representada de la base original. ....	43
Tabla 3-8. Arquitectura de <i>autoencoder</i> . ....	45
Tabla 3-9. Información estadística del resultado de <i>autoencoder</i> . ....	45
Tabla 3-10. Información estadística del resultado de PCA inversa. ....	46
Tabla 3-11. Tabla comparativa de MSE. ....	47
Tabla 3-12. Hiperparámetros de <i>k-means</i> . ....	51
Tabla 3-13. Hiperparámetros de AGG. ....	51
Tabla 3-14. Hiperparámetros de DBSCAN. ....	52
Tabla 3-15. Hiperparámetros de SOM. ....	52
Tabla 4-1. Mejores resultados de <i>Silhouette</i> de cada algoritmo. ....	55
Tabla 4-2. Mejores resultados de <i>Calinski Harabasz</i> de cada algoritmo. ....	55
Tabla 4-3. Mejores resultados de la combinación de cada algoritmo. ....	56
Tabla 4-4. Análisis de sensibilidad de bases de entrada. ....	58
Tabla 4-5. Resultados de análisis de la varianza. ....	63
Tabla 4-6. Algoritmos utilizados para estimación de cobre total y recuperación metalúrgica. ....	68
Tabla 4-7. Hiperparámetros utilizados para la predicción. ....	70
Tabla 4-8. Parámetros para evaluación económica. ....	72
Tabla 4-9. Costos planta para evaluación económica. ....	73
Tabla 4-10. Costo de energía. ....	74
Tabla 4-11. Costo planta por escenario para evaluación económica. ....	74
Tabla 4-12. VPN según escenario. ....	75



## ABSTRACT

A great deal of parameters involved in a mining project causes higher complexity. The efficiency of the mining investment is a relevant issue in this industry, where the most common issues are exceeded estimated investment and/or overestimated processing capacity. Thus, the emerging concept of geometallurgical modeling was a probable solution proposed to improve decision making.

This thesis aims to design, implement, and validate a geometallurgical model of a copper deposit. For this purpose, the literature related to the generation and validation of the geometallurgical model is reviewed. An essential contribution of this work is application to a copper deposit and the impact evaluation on the project value.

A drilling database was used, containing 1112 points from 569 drill-hole core sample with spatial variables of location, grades of different elements, mineral processing parameters, mineralogy, and categorical geology variables. Various combinations of preprocessed and dimensionally reduced input data were tested with *k-means*, hierarchical clustering, DBSCAN, and self-organized map. The geometallurgical model quality was measured by the internal cluster validity index.

Finally, the main results demonstrated that it is possible to generate a geometallurgical model with machine learning techniques, and *k-means* has better performance. The best results indicated that spatial and geology categorical variables could be excluded. Besides, the application of the dimension reduction technique does not improve the performance of the geometallurgical model generation. Additionally, evaluating the impact on the project value, the application of the geometallurgical model impacts the extraction schedule and results in an improvement of 4% in the project value.

**Keywords:** Unsupervised learning; geometallurgy; geometallurgical modelling.

## RESUMEN

La gran cantidad de factores involucrados en los proyectos mineros es una de las causas de la complejidad que estos proyectos presentan. La eficiencia de la inversión minera es un tema crucial en esta industria, donde los problemas más comunes en la ejecución de los proyectos son el sobrepasar la inversión estimada y/o sobreestimar de la capacidad de procesamiento de la planta. El concepto emergente de modelo geometalúrgico de un yacimiento es una probable solución, incidiendo en mejoras en la toma de decisiones.

El objetivo general de esta tesis es diseñar, implementar y validar un modelo geometalúrgico de un yacimiento de cobre, utilizando técnicas de aprendizaje no supervisado. Una contribución importante de este trabajo es la aplicación en un yacimiento de cobre, junto con la evaluación del impacto al valor de proyecto.

Este estudio, utiliza una base de datos de 569 sondeos con 1112 vectores de datos, que contienen variables espaciales de ubicación, concentración de distintos elementos, datos de procesamiento metalúrgico, mineralogía y variables geológicas. Se evalúan distintas combinaciones de datos de entrada preprocesada y la reducción de dimensiones, junto con la aplicación de *k-means*, agrupación jerárquica (AGG), DBSCAN y mapa autoorganizado. Se mide la calidad del modelo geometalúrgico a través de indicadores de validación interna de agrupación.

Finalmente, el principal resultado obtenido es la generación de un modelo geometalúrgico con la técnica de aprendizaje de máquinas donde el algoritmo *k-means* tiene mejor desempeño. Los resultados indican que es posible excluir las variables espaciales y categóricas de geología. La aplicación de la técnica de reducción de dimensión no mejora el desempeño del modelo. Adicionalmente, en la evaluación del impacto al valor del proyecto, la aplicación del modelo geometalúrgico afecta la secuencia de extracción y presenta una mejora de orden de 4% en el valor.

**Palabras Claves:** Aprendizaje No supervisado, Geometalurgia, Modelo geometalúrgico.

## **1. INTRODUCCIÓN**

### **1.1. Geometalurgia y modelos geometalúrgicos**

Los proyectos mineros son sumamente complejos por la alta cantidad de variables involucradas, además de la relativa incertidumbre tanto sobre el recurso minero existente como sobre el comportamiento del mercado, entre otros. Durante las últimas décadas, la eficiencia de la inversión ha sido un tema muy relevante en la industria, ya que el sobrepasar el costo capital (CAPEX) planificado no ha sido la excepción en los proyectos. De hecho, Michaux & O'Connor (2020) en un estudio que considera 276 proyectos mineros, de los cuales 18 corresponden a proyectos recientes de la gran minería, encontró que 12 de ellos sobrepasaron la inversión inicial (CAPEX) estimada, y 14 no lograron llegar a la producción planificada. Además, de los 258 proyectos realizados entre los años 1980 y 2000, 86% de ellos sobrepasaron la inversión inicial con un promedio de 28% de mayor CAPEX.

Dada la complejidad propia de un yacimiento y la tendencia a utilizar escalas cada vez mayores de procesamiento, es más valioso contar con una buena estimación de la capacidad de la planta asociada. Dentro de este marco, surge el concepto geometalurgia, el cual fue definido alrededor del año 1970 (McQuiston & Bechaud, 1968; Hoal, 2008). En su definición más simple, la geometalurgia es la integración de la información geológica, metalúrgica y minera, con miras a mejorar

significativamente la certeza del valor del proyecto en la toma de decisiones en la planificación minera (Lamberg, 2011; Sepulveda et al., 2017; Ellefmo et al., 2019). Frente al desarrollo de la industria, el concepto de geometalurgia también ha ido evolucionando, incorporando conceptos como mina a planta (*mine to mill*) y variables relacionadas con aspectos ambientales, dándole una mirada aún más completa a la cadena de valor minero (Hunt & Berry, 2017; Fernandes & Leite, 2020). Con todo, el concepto no es algo nuevo. Su objetivo se mantiene, cuál es maximizar la captura de valor y minimizar el riesgo económico y operacional en la explotación de recursos mineros (Curtis et al., 2019; Mena et al., 2020; Koch & Rosenkranz, 2020).

Finalmente, la aplicación del concepto geometalurgia se representa en un modelo geometalúrgico como el de la Figura 1-1. Donde se observa un espacio tridimensional de bloques dividido en distintas unidades geometalúrgicas (cada color representa una unidad geometalúrgica), caracterizadas por ser un volumen espacial con un comportamiento similar en la extracción minera y metalúrgica (Rajabinasab & Asghari, 2019). Cabe destacar, un modelo geometalúrgico ideal debe ser generado a partir de informaciones geológicas, metalúrgicas, mineras y ambientales relacionadas al producto, sus subproductos y los efectos en la sustentabilidad de la operación (Dominy et al., 2018; Rajabinasab & Asghari 2019; Lishchuka et al., 2020)

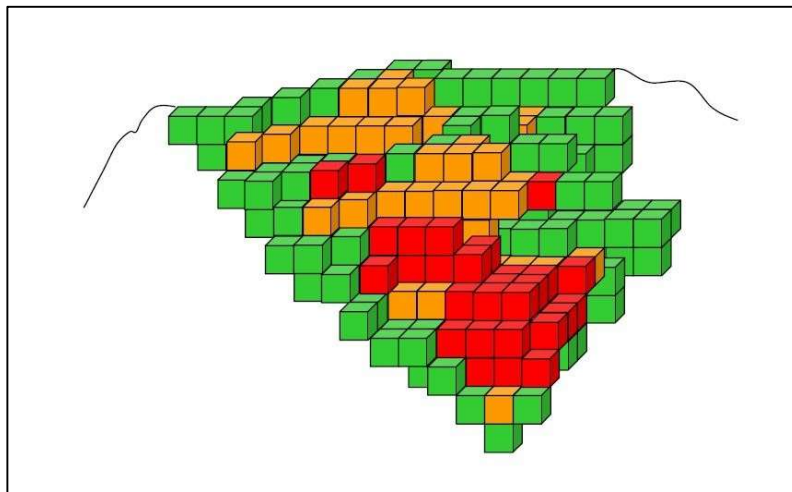


Figura 1-1. Modelo geometalúrgico. Adaptado de Canchaya (2008). Cada cubo es un bloque y el color indica distintas unidades geometalúrgicas.

### 1.2. Estado del arte en la generación de modelos geometalúrgicos

Actualmente, se identifican dos líneas de trabajo para la generación de modelos geometalúrgicos. La primera, que se podría denominar como “método tradicional” (Suazo et al., 2010; Boisvert et al., 2013; Rincon et al., 2018; Johnson et al., 2019), considera la utilización de la simulación de datos utilizando metodologías convencionales, como regresión múltiple u otros métodos de establecimiento de correlaciones entre de las variables de entrada y salida. La segunda línea, es la utilización de técnicas de inteligencia artificial, aplicando algoritmos supervisados y/o no supervisados que buscan extraer la máxima información de los datos disponibles (Koch et al., 2019; Fernandes & Leite, 2020; Mena et al., 2020; Rajabinasab & Asghari, 2018; Ordóñez-Calderón & Gelcich, 2018). A continuación,

se presentan las técnicas y las variables utilizadas en distintos estudios para la generación de modelos geometalúrgicos.

### **1.2.1. Métodos tradicionales**

Dentro de los métodos tradicionales, el más usual es la regresión lineal. En este caso se toman distintos tipos de variables de entrada para generar variables metalúrgicas de salida, para luego generar unidades geometalúrgicas a través de la agrupación en el espacio de los resultados anteriores. Las variables de entrada podrían ser geológicas, mineralógicas, o características extraídas de la imagen hiperespectral de muestras de minerales. Las variables de salida son ley, recuperación metalúrgica, índice de trabajo, también las otras variables relacionadas al proceso como índice de dureza de mineral al impacto y consumo de ácido.

Actualmente, algunos estudios que correlacionan distintas formas de variables mineralógicas con variables metalúrgicas a través del método de regresión lineal han tenido buenos resultados. Uno de ellos, Johnson et al., (2019) relaciona 260 datos de 41 variables de características mineralógicas extraídas con la técnica hiperespectral. El sistema de imágenes hiperespectral utiliza una cámara estacionaria que efectúa un escaneo lineal sobre una muestra a una velocidad controlada. Esto permite determinar el espectro de una muestra completa, la cual es analizada en 449 bandas que cubren el espectro electromagnético entre 500 y 2450 nm, permitiendo identificar especies mineralógicas específicas en cada región. Basado en un modelo de regresión lineal múltiple, se correlacionan estas variables características con la

capacidad de procesamiento esperada, la ley y la recuperación de cobre y oro. Como resultado final el índice de correlación varía entre 0.56 a 0.71 en las variables de salida.

En otro caso, Rincon et al., (2018) generaron un modelo lineal que establece correlación entre las variables de entrada geológica y las variables metalúrgicas de salida. Las 4 variables de entrada que utilizaron son: textura, tipo de mineralización, litología y ley de cobre, las variables de salida por determinar son índice de dureza de mineral por impacto (Axb) e índice de trabajo (OWi). Estos autores concluyeron que la presencia de Caolinita y Dickita se correlaciona directamente con el índice de trabajo (OWi) e inversamente con el índice de dureza de mineral por impacto (Axb).

En un caso que utilizó mayor cantidad de variables de entrada para generar las variables metalúrgicas, Boisvert et al. (2013) aplicaron un método similar de regresión con adición de selección de variables para la mina *Olympic Dam*. Empezaron con 204 variables de entrada disponibles, las cuales se dividieron en 4 categorías: retenidas (6 variables elegidas de leyes y mineralogías que consideraron más importantes), leyes, mineralogía y variables de asociación (describe área de contacto entre dos minerales dentro de un grano de material chanchado). En un primer paso, se redujeron variables por eliminación de variables poco relevantes, basado en apreciaciones según el conocimiento experto sobre el yacimiento. Ello habría permitido evitar el sobre entrenamiento. Después se realizaron dos etapas de

fusión de variables con el método de suposición gaussiana multivariante (*multivariate Gaussian assumption*), terminando en 4 variables que corresponde a una por categoría. Finalmente, se establecieron 3 modelos vía regresión lineal para predecir las 6 variables de rendimiento de la planta, los cuales son el índice de caída libre (*drop weight index*), índice de trabajo de Bond (*bond mill work index*), recuperación de Cu, recuperación de U<sub>3</sub>O<sub>8</sub>, consumo de ácido y recuperación neta. La recuperación neta es la diferencia entre el total del material extraído y la masa que se pierde por el procesamiento en la planta. Los tres modelos se diferencian por el número de variables de entrada (Tabla 1-1) y el resultado en coeficiente de correlación de Pearson entre valor real y predicción de cada variable (Tabla 1-2).

Tabla 1-1. Descripción de los modelos predictivos propuestos (Boisvert et al., 2013).

Modelo	Variables de entrada	Comentarios del autor
Completo	Retenidas. Leyes. Mineralogía. Variables de asociación.	Considera todas las variables del yacimiento.
Típico	Retenidas. Leyes. Mineralogía.	Modelo base, que considera datos adicionales al caso limitado de mayor disponibilidad.
Limitado	Retenidas. Algunas de leyes. Algunas de mineralogía.	Modelo que considera solo algunas variables de leyes y mineralogía que están usualmente disponibles.



Tabla 1-2. Resultado del coeficiente de correlación de Pearson para cada modelo (Boisvert et al., 2013)

<b>Variable de salida</b>	<b>Modelo completo</b>	<b>Modelo típico</b>	<b>Modelo limitado</b>
Índice de caída libre	0.80	0.77	0.75
Índice de trabajo de Bond	0.77	0.54	0.53
Recuperación de Cu	0.70	0.59	0.59
Recuperación de $U_3O_8$	0.65	0.62	0.61
Consumo de ácido	0.90	0.86	0.84
Recuperación neta	0.74	0.71	0.67

Finalmente, con los ejemplos mencionados anteriormente, se entiende que un modelo geometalúrgico determinado por el método de regresión es naturalmente función de los datos disponibles. Algunos utilizan mayor cantidad de variables de entrada que otros, pero su finalidad se mantiene la cual es correlacionarlas con las variables de salida. Estas últimas varían dependiendo del enfoque de cada proyecto. Por otro lado, se destaca que las regresiones tienen índice de correlación moderados, excepto en algunos casos que tienen valores mayores que 0.85. Sin embargo, esta línea de trabajo define la unidad geometalúrgica basada en el modelo geológico, en el cual tiene las unidades establecidas considerando la similitud geológica y se espera que los minerales que provienen de una misma unidad tengan comportamiento similar (Suazo et al., 2010 & Lamberg, 2011).

### **1.2.2. Inteligencia artificial**

Dentro del uso de las técnicas de clasificación utilizando inteligencia artificial se pueden clasificar en dos grandes familias: métodos de aprendizaje supervisado y

métodos de aprendizaje no supervisado. Los clasificadores no supervisados se aplican a una base de datos que aprenden de la misma base para la clasificación. Dicho de otra forma, agrupa los datos que tienen características similares e/o identifica datos atípicos (*outliers*). En cambio, un clasificador o regresor supervisado se puede entender como un mecanismo de aprendizaje autónomo: se aprende desde una base de datos con etiquetas, para que el clasificador identifique las diferencias entre distintas clases. En el caso de requerirse una estimación de algún regresor, debe utilizarse una base de datos con resultados previos para poder generar las estimaciones.

Cada familia de métodos tiene sus fortalezas y debilidades, dependiendo del escenario de uso. En el caso de los algoritmos supervisados, es indispensable tener datos históricos de los resultados metalúrgicos para poder hacer la predicción. En cambio, en el caso de un yacimiento nuevo, se requiere un trabajo de etiquetado en forma manual o usando un algoritmo de agrupamiento de la familia de aprendizajes no supervisados para generar etiquetas.

### **Aprendizaje supervisado**

Dentro de la familia de algoritmos supervisados, Mena et al. (2020) implementaron una combinación de modelos basados en redes neuronales con variables de entrada obtenidas de los análisis químicos por fluorescencia de rayos X, para predecir la mineralogía (porcentaje de minerales) y recuperación del proceso de separación

magnética (*Concentrate yield*) de un yacimiento de nefelina sienita ( $\text{Al}_2\text{KNaO}_8\text{Si}_2$ ). El primer paso de este modelo realiza un preprocesamiento para normalizar los datos de entrada. Luego, aplican redes neuronales prealimentadas con dos capas ocultas para predecir las variables de salida. El modelo genera una predicción con un coeficiente de correlación 0.9 con respecto a los valores reales. La base de datos utilizada son 1055 registros de sondajes. La división de datos en la etapa de entrenamiento, prueba y validación son 50%, 25% y 25%. Cabe mencionar que se determinaron la cantidad de capas ocultas a través de un análisis comparativo de error cuadrático entre los datos de entrenamiento, prueba y validación.

Asimismo, Fernandes & Leite, (2020) realizaron un estudio para una mina de zinc, se encontró que el uso de redes neuronales de 3 capas ocultas con variables de entrada como ley de zinc y plomo, más otras dos variables binarias que identifican la zona del yacimiento de la muestra, lograron predecir 6 variables de salida: recuperación másica de concentrado de Zn, recuperación másica de concentrado de Pb, recuperación másica de cola, Recuperación de Zn desde concentrado de Zn, recuperación de Zn desde concentrado de Pb y recuperación de Zn desde cola. El modelo fue entrenado con 140 muestras de sondajes, las cuales se dividieron en 112 y 28 para la etapa de entrenamiento y prueba. Se obtuvo excelentes resultados en algunas variables, como se muestra en la siguiente Tabla 1-3.

Tabla 1-3. Resultado de la validación cruzada quintuple (Fernandes & Leite, 2020).

Etapa	Variables	Coefficiente de correlación
Entrenamiento	Recuperación másica de concentrado de Zn.	
	Recuperación másica de concentrado de Pb.	$0.9905 \pm 0.0003$
	Recuperación másica de cola.	$0.2643 \pm 0.0096$
	Recuperación de Zn desde concentrado de Zn,	$0.9901 \pm 0.0003$
	Recuperación de Zn desde concentrado de Pb.	$0.8992 \pm 0.0034$
	Recuperación de Zn desde cola.	$0.0755 \pm 0.022$
Prueba	Recuperación másica de concentrado de Zn.	
	Recuperación másica de concentrado de Pb.	$0.9892 \pm 0.0013$
	Recuperación másica de cola.	$0.2249 \pm 0.0792$
	Recuperación de Zn desde concentrado de Zn.	$0.9888 \pm 0.0016$
	Recuperación de Zn desde concentrado de Pb.	$0.8908 \pm 0.0154$
	Recuperación de Zn desde cola.	$-0.0136 \pm 0.0766$

Además del uso de redes neuronales, varios estudios han tenido resultados con los algoritmos de la familia de aprendizaje de máquinas, como random forest, k vecinos más próximos y máquinas de vectores de soporte. Koch et al. (2019) utilizaron una base de datos de 1509 fotos de muestras de sondajes, aplicando la técnica de extracción de características para generar variables de entrada para predecir la mineralogía. El resultado sería utilizado a futuro en un modelo geometalúrgico enfocado en el proceso de conminución. Las técnicas de extracción de características de imágenes que utilizaron son GLCM (*grey-level co-occurrence matrix*) y LBP (*local binary patterns*). Los regresores que se probaron en esta oportunidad son random forest, k vecinos más próximos, máquinas de vectores de soporte y red neuronal artificial con tres capas ocultas. En este caso, random forest logró obtener

el mejor rendimiento con las distintas combinaciones de variables de entrada. El mejor resultado se obtuvo en una correlación igual a  $0.98 \pm 0.05$  en la predicción de textura con la base de entrada que incluye variables generadas por LBP y GLCM.

En otros casos similares, Guntoro et al. (2020) ocuparon variables características de fotos AIM (*association indicator matrix*) y LBP (*local binary pattern*) de fotos de sondajes para clasificar la textura de minerales, obteniendo un 84% de acierto con un algoritmo de clasificación random forest, con una base de entrada que tiene las variables de características AIM. En otra prueba tuvo 88% de acierto con el clasificador SVM y las variables características LBP. Se concluyó que ambas líneas requieren más estudio futuro con mayores pruebas de datos, pero podría ser una herramienta útil para utilizar en la construcción de modelo geometalúrgico.

Dentro de los estudios revisados, se identifica que el principal uso de algoritmos de aprendizaje supervisado está ligado a la predicción de variables metalúrgicas o a la clasificación de variables geológicas, las cuales se utilizan para la clasificación geometalúrgica. Todos los casos fueron estudiados con datos existentes que permiten evaluar el rendimiento del modelo establecido. Por lo tanto, esta técnica serviría para los yacimientos en operación o proyectos de expansión (*Brown Field*) y podría ser una alternativa a los ensayos de laboratorio, con un fin de ahorro de recursos.

### **Aprendizaje no supervisado**

Por otro lado, en el caso del uso del algoritmo no supervisado, Rajabinasab & Asghari (2018) probaron en un yacimiento de hierro con los algoritmos *k-means*, agrupamiento jerárquico (*hirarchical clustering*) y mapa autoorganizado (*self-organizing maps*) en la generación de unidades geometalúrgicas. Resultó mejor la clasificación usando *k-means* para definir zonas con baja, media y alta ley. Sus variables de entrada fueron: ley de Fe, ley de óxido de fierro, ley de S, susceptibilidad magnética (*magnetic susceptibility*) y las coordenadas de las muestras (X, Y, Z). Todas las variables son normalizadas antes de procesar. Dado que los algoritmos utilizados no pueden realizar el trabajo sin una asignación del número de clúster deseados, el índice de agrupaciones *Calinski–Harabasz* y *Silhouette* fueron utilizados para determinar dicho número. Finalmente, estos índices también fueron utilizados para determinar la mejor agrupación obtenida, lo cual determinó que los dos métodos recomendados de agrupación sean *k-means* y mapa autoorganizado, ya que obtuvieron similares valores en los dos índices.

También, Bhuiyan et al., (2019) aplicaron aprendizaje no supervisado a un yacimiento de oro, que tuvo resultados con la aplicación del método de K media. Su base de datos usó como variables de entrada el índice de resistencia a la carga puntual (*axial point load strength index*), calidad de roca (*Rock quality designation*), susceptibilidad magnética, índice de trabajo de Bond (*Bond ball mill work index*) y leyes de 20 elementos. Se aplicó un primer paso en reducción de dimensiones por

análisis de componentes principales (PCA: *Principal component analysis*), luego se aplica la agrupación con K media (*K-means*), donde el hiperparámetro K de este algoritmo se obtuvo con la relación entre K y la suma del error cuadrático dentro de cada agrupación, eligiendo el valor K cuando la curva tenga una pendiente cercana a -1. Luego se utilizó un método supervisado *RandomForest* para verificar la calidad de agrupación, enfocada al índice de trabajo de Bond para validación cruzada. Se obtuvo un 70% de acierto con la clasificación según el índice de trabajo de Bond.

En un ámbito no estrictamente geometalúrgico, en un estudio de un yacimiento de cobre, Ordóñez-Calderón & Gelcich (2018) comprobaron que el algoritmo K media es viable para generar facies de un depósito de skarn, donde facie es un conjunto de rocas sedimentarias o metamórficas con características determinadas. De similar forma, Fouedjio (2016) demostró la posibilidad de aplicar la técnica de agrupamiento para estudios geoestadísticos. En este caso se utilizó la ley de 8 elementos como variables de entrada, se probó 5 diferentes algoritmos y todos resultaron positivos en generar grupos con variables homogéneas. El mejor resultado según el autor fue un método propuesto consistente en una combinación de agrupamiento jerárquico y un estimador de núcleo no paramétrico con dependencia espacial. En una aplicación a un caso real, utilizando una base de datos de 1498 registros que incluye leyes de 8 elementos, se logró determinar dos agrupaciones espacialmente continuas, sin la inclusión de variables geométricas.

En un caso estudiado en Chile, Ortiz & Kracht (2020) utilizaron el algoritmo K media y las variables de entrada compuestas por 39 variables de análisis geoquímico, 3 variables de leyes de mineral de interés y 11 variables sintéticas creadas para representar ciertas características de los tipos de alteración geológica. El procesamiento se inicia con una etapa de selección de variables donde se analizan las que tienen mayor correlación con la respuesta final. Utilizando el método de selección hacia adelante (*Sequential Forward Selection*), se seleccionan las variables que generan mayor varianza entre etiquetas (tipos de alteración). Después en la clasificación, utilizaron un algoritmo no supervisado que identifica zona de alteración en el yacimiento. Finalmente, dada la existencia de datos de etiqueta según una clasificación manual por geólogos, se compara el resultado de la clasificación obtenido con el juicio experto, alcanzando un 77.1% de acierto con la clasificación de algoritmo.

Dentro de los estudios revisados con el método aprendizaje no supervisado, el algoritmo K media tuvo mayor frecuencia en uso. Las variables de entrada usuales son leyes de elementos deseados, el resto de las variables se seleccionaron dependiendo de la disponibilidad y decisión del investigador en consideración a la relevancia de la inclusión, ya que de los casos revisados cada uno tiene un enfoque particular o una aproximación propia al modelo geometalúrgico.



### **1.3. Inteligencia artificial e industria minera chilena**

El término inteligencia artificial, introducido por John McCarthy en el año 1955 junto con el desarrollo del hardware computacional y el desarrollo de los algoritmos, ha sido aplicado cada vez más en las distintas áreas del conocimiento.

La cuarta revolución industrial o Industria 4.0 es un concepto introducido por primera vez en el año 2011 en la feria de Hannover. Luego de casi una década de desarrollo, cada sector industrial tiene iniciativas en desarrollo alineadas a este concepto, pero la industria minera es una de las más atrasadas (Maniyika et al., 2015).

Efectivamente, en el sector industrial minero, el concepto Minería 4.0 es una versión de Industria 4.0 aplicado en minería. Un estudio de CESCO (2020) mostró el estado de avance de los proyectos de Minería 4.0 implementados por las grandes compañías mineras en Chile. Muchos proyectos relacionados han llevado en desarrollo y varios proyectos pasaron de la etapa de idea para llegar a la etapa de implementación.

En un futuro cercano, la Minería 4.0 construiría una nueva forma de trabajo que hará más uso de la automatización, incluyendo el uso de algoritmos inteligencia artificial. Con el objetivo permanente de maximizar la captura de valor y minimizar el riesgo, la aplicación del modelo geometalúrgico podrá ser parte de esta cuarta revolución industrial.

#### **1.4. Objetivos**

En este trabajo se plantea como hipótesis el que es factible desarrollar un modelo geometalúrgico en un yacimiento de cobre con inclusión de datos geológicos y metalúrgicos, basado en técnicas de la Inteligencia Artificial y más específicamente de Aprendizaje de Máquinas. De acuerdo con esta hipótesis, el objetivo general es diseñar, implementar y validar el modelo geometalúrgico.

Como se indicó anteriormente, un modelo geometalúrgico ideal debe ser generado a partir de todas las informaciones relacionadas a la cadena de valor de una mina. Sin embargo, de acuerdo con las investigaciones, aún no se tiene una arquitectura de los algoritmos que logran a un modelo ideal. Por lo tanto, este trabajo busca una mejora en este sentido.

Para cumplir el objetivo general, se establecieron los siguientes objetivos específicos:

- Realizar un análisis comparativo de algoritmos aprendizaje no supervisado para generar modelos geometalúrgicos, aplicando métricas de calidad de la clasificación de los algoritmos.
- Aplicar algoritmos para generar modelos geometalúrgicos.
- Evaluar el impacto del modelo geometalúrgico en el valor del proyecto.

Para alcanzar estos objetivos y generar un modelo geometalúrgico a partir de datos de sondaje, se realizan 5 etapas. La primera consiste en asegurar la consistencia de la base de datos, ya que contiene variables geológicas categóricas que deben convertirse a variables numéricas. En una segunda etapa se realiza un preprocesamiento de la base de datos, eliminando variables dependientes y constantes, y aplicando normalización para evitar que el modelo llegue a resultados subóptimos (García et al., 2015). En la tercera etapa se investiga la reducción de dimensiones de la base de datos, con el objeto de evaluar ventajas de reducción del tiempo de procesamiento y capacidad de memoria computacional durante la clasificación. Luego, se aplican técnicas de agrupación no supervisada para definir *clústeres* de datos con características similares, conformando el modelo geometalúrgico. Finalmente, se efectúa una evaluación del impacto al valor del proyecto, utilizando el modelo geometalúrgico para definir una secuencia de extracción minera que permite comparar los costos de explotación con un caso base.

El resultado esperado de esta tesis es un modelo cuya entrada son los datos disponibles de un yacimiento de cobre y como salida un modelo geometalúrgico del yacimiento. Sería uno de los primeros intentos de desarrollar un modelo geometalúrgico integral aplicado a la minería del cobre y con cuantificación económica del impacto. Este modelo es parte integrante de los desarrollos de la cuarta revolución industrial en minería y su motivación se deriva del mejoramiento esperado, tanto en calidad como en oportunidad de los modelos geometalúrgicos de

un yacimiento, con la consiguiente reducción de recursos y eventual reducción de la incertidumbre.

El resto del presente documento está dividido en cuatro capítulos.

En el capítulo 2, se presenta la arquitectura que se utiliza para generar el modelo geometalúrgico, detallando los fundamentos teóricos de las distintas metodologías a utilizar en cada etapa del modelo.

En el capítulo 3, se revisa en más detalle la metodología elegida en cada etapa del modelo, como son los valores de los hiperparámetros que se propone a utilizar y la métrica de evaluación para la elección final.

El capítulo 4, detalla los resultados obtenidos y una evaluación económica, que incluye un proceso de interpolación del espacio del yacimiento para establecer un modelo de bloques y un algoritmo de aprendizaje supervisado para la estimación de valores en este espacio. Se efectúa la evaluación del valor del proyecto comparando escenarios con y sin modelo geometalúrgico, estimando el beneficio de la aplicación de dicho modelo.

Finalmente, en el capítulo 5 se entregan las conclusiones, aplicabilidad y sugerencia de trabajo futuro para el mejoramiento del modelo.

## 2. BASES CONCEPTUALES

### 2.1. Técnicas y algoritmos

Junto con el requerimiento de un modelo geometalúrgico y las revisiones bibliográficas relacionadas, se plantea utilizar técnicas ligadas al área de aprendizaje de máquinas para cumplir la generación de las unidades geometalúrgicas. Se especifican 4 etapas de trabajo según el objetivo: preparación de datos, preprocesamiento de datos, reducción de dimensiones y clasificación. Cabe mencionar que todos los procesamientos se realizan en software Python y las librerías asociadas. Más adelante se ahondará en el uso de algoritmos en cada etapa.

#### 2.1.1. Preparación

Antes de realizar el preprocesamiento de datos, se debe revisar si existen variables categóricas, estas son variables no numéricas y determinadas por expertos en geología con códigos para indicar zonas con ciertas características geológicas. Para poder procesar estas variables, se propone utilizar una codificación *One-hot* (*One-Hot Encoding*), donde se generará N variables binarias, donde N corresponde a la cantidad de categorías; es decir, para cada categoría generará una variable, esta tiene valor 0 o 1, donde 1 indica que se pertenece a esta categoría y 0 el caso contrario (Müller & Guido, 2016).

#### 2.1.2. Preprocesamiento

Para mejorar el rendimiento de datos, se propone usar el algoritmo *Clean* de la librería *Pybalu*, en el que se revisa la correlación entre todas las variables e identifica

grupos de variables que tienen un coeficiente de correlación de Pearson mayor a 0.99. Luego se elige aleatoriamente una y elimina el resto. Dentro del mismo proceso *Clean*, se eliminan las variables que tienen desviación estándar menor a  $10^{-8}$ , estas variables se consideran como variables constantes.

Adicionalmente, una normalización de datos es recomendable, ya que los datos que provienen de distintas fuentes podrían tener distintas escalas por su naturaleza. Por lo tanto, un re-escalamiento es necesario para evitar que el modelo llegue a un subóptimo (García et al., 2015). Las técnicas de transformación de datos más comunes son la estandarización y la normalización, donde en la primera se llevan los datos a un rango de  $[-1, 1]$  con media igual a 0, mientras que en la normalización se transforma los datos a un rango de  $[0, 1]$ , ambos casos se mantienen la misma forma de distribución de datos.

### **2.1.3. Reducción de dimensiones**

La tercera etapa, corresponde a la reducción de dimensiones, donde existen dos alternativas. La primera es la transformación lineal, donde se proyectan los datos a una cantidad reducida de dimensiones. En este caso, el Análisis de componentes principales (PCA: *Principal component analysis*) es una técnica usual, la cual consiste en la proyección de los datos a nuevos ejes, donde los datos queden mejor representados en términos de mínimos cuadrados. El nuevo conjunto convertido (conjunto de variables artificiales) se llama componente principal y cada componente representa una cierta cantidad de información del dato original.

Finalmente se selecciona la cantidad de componentes de acuerdo con la cantidad de información representada del dato original (Müller & Guido, 2016).

En la otra línea se busca una transformación no lineal con la técnica de redes neuronales, esta se denomina como *autoencoder*, que busca la menor pérdida de informaciones ante una reducción de mayor dimensión (Chollet, 2017). En la Figura 2-1 se muestra una construcción básica de un modelo *autoencoder*, a partir de una base de entrada  $X$ , a través de una etapa de codificación. El producto de esta codificación se denomina espacio latente, y es una base de menor dimensión. Por otro lado, a partir del espacio latente se realiza una decodificación que busca reconstruir la base  $X$ . Pero como el proceso tiene pérdida, el producto de la decodificación no necesariamente logra generar la misma base  $X$ , y el resultado se denomina  $\hat{X}$ . Finalmente, para tener una evaluación de la pérdida de información en la reducción de dimensión (codificación), se define una función de pérdida entre las bases  $X$  y  $\hat{X}$ .

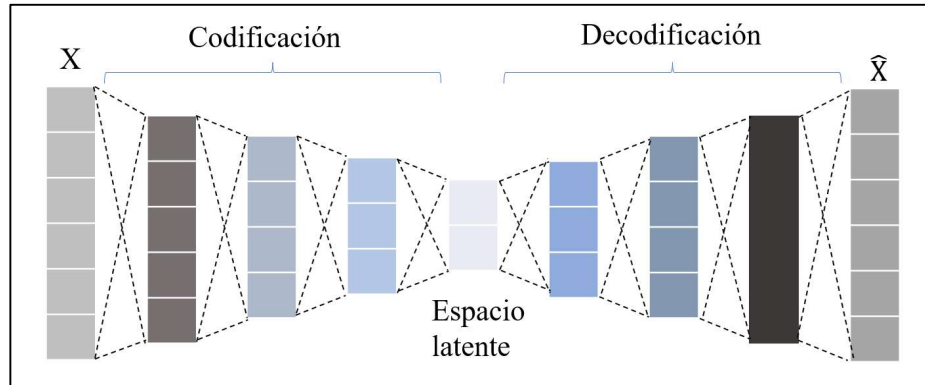


Figura 2-1. Arquitectura de *autoencoder*.

En las etapas de codificación y decodificación, se define una función  $h$ , también conocida como capa oculta de redes neuronales, que recibe información de la capa anterior y transmite información relevante para la siguiente etapa

$$h^{i+1} = \sigma^{i+1}(W^{i+1}h^i) \quad (2.1)$$

$W$  es una matriz de pesos asociada a las variables en cada etapa.  $\sigma$  es la función de activación, la cual recibe la información del nodo anterior y decide su activación para emitir la información para el siguiente nodo. La entrada de cada etapa es la información de salida de la etapa anterior (Kramer, 1991).

La función de activación podría tener muchas formas. Geron (2017) menciona que las funciones de activación más usuales y con mejor rendimiento son las de unidad lineal rectificada (ReLU: *Rectified Linear Unit*), función hiperbólica ( $\tanh$ ) y función logística (sigmoid). Estas funciones se muestran en la Tabla 2-1:



Tabla 2-1. Función de activación.

Función de activación	Función
ReLU	$f(x) = \begin{cases} 0, & \text{si } x < 0 \\ x, & \text{si } x > 0 \end{cases}$
tanh	$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$
Sigmoid	$f(x) = \frac{1}{1 + e^{-x}}$

Para la primera etapa, la entrada es la base de datos  $X$  (Kramer, 1991).

$$h^1 = \sigma^1(W^1X) \tag{2.2}$$

Para la última etapa, la salida es la base de datos generado  $\hat{X}$ , donde  $f$  es la cantidad de capas ocultas (Kramer, 1991).

$$\hat{X} = \sigma^f(W^f h^{f-1}) \tag{2.3}$$

Finalmente, para evaluar la pérdida de información se utiliza una función de pérdida (loss function)  $E$ , donde  $E$  podría ser en muchas formas, por ejemplo, error absoluto medio (MAE: *Mean absolute error*) o error cuadrático medio (MSE: *Mean squared error*) (Kramer, 1991).

$$loss = E(X - \hat{X}) \tag{2.4}$$

2.1.4. Clasificación

La última etapa corresponde a la clasificación para generar el modelo geometalúrgico, en la cual se considera que no existe modelo previo, por lo que, se utiliza la clasificación de la familia de aprendizaje no supervisado.

Existen varias alternativas disponibles, cada una con distintas lógicas de algoritmos, siendo *k-means* una de las más utilizadas según la revisión bibliográfica. Su esencia es generar  $K$  cantidad de centroides y agrupar los datos cercanos que se encuentran dentro de una distancia euclidiana desde los centros. Su algoritmo se detalla en cuatro pasos en la siguiente forma (Mackay, 2003) junto con un ejemplo de pasos que se muestra en la Figura 2-2.

Se supone que hay  $N$  cantidad de datos de  $I$  dimensiones, y se van a generar  $K$  agrupaciones.

Paso 1: Iniciación. Se asigna aleatoriamente el vector de agrupaciones  $m^K$ .

Paso 2: Asignación. A cada dato se le asigna a una agrupación o *cluster*, asignación asociada al centroide más cercano al dato.

$$\hat{k}^{(n)} = \arg \min (d(m^K, x^{(n)}))$$

$$r_k^{(n)} = \begin{cases} 1, & \text{si } \hat{k}^{(n)} = k \\ 0, & \text{si } \hat{k}^{(n)} \neq k \end{cases}$$

Paso 3: Actualización. Para cada agrupación, se calcula su media y error medio cuadrático.

$$m^K = \frac{\sum_n r_k^{(n)} x^{(n)}}{R^{(k)}}$$

$$R^{(k)} = \sum_n r_k^{(n)}$$

Paso 4: Repetir el paso 2 y 3, hasta que llega a un mínimo medio cuadrático.

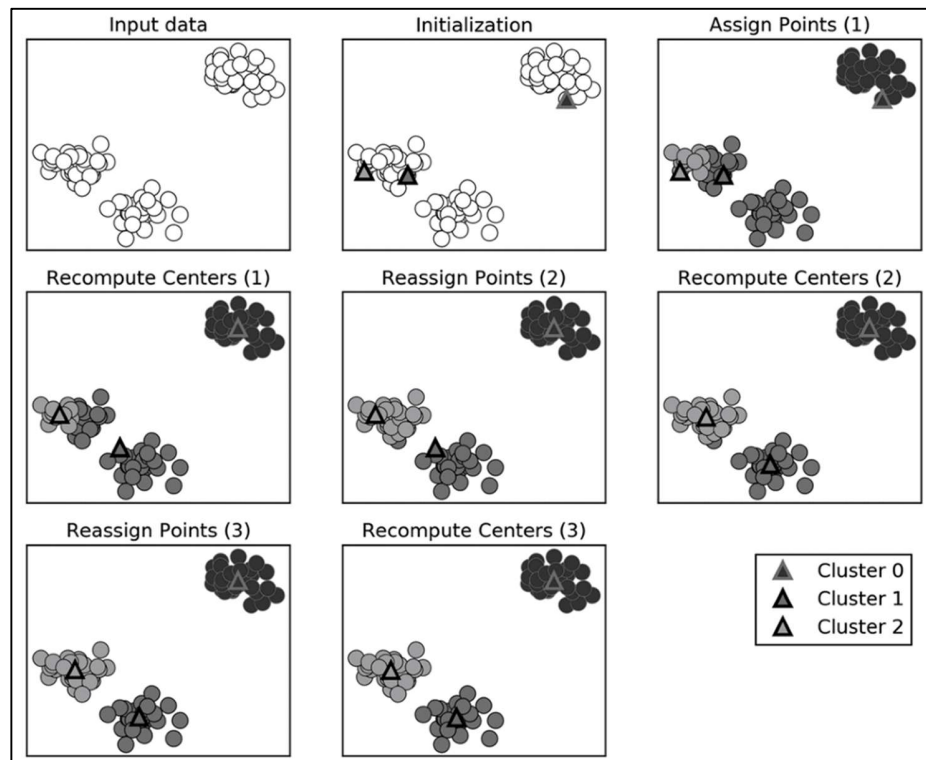


Figura 2-2. Ejemplo de *k-means*. Adaptado de Müller & Guido (2016).

La agrupación jerárquica (AGG, *Agglomerative Clustering*) tiene un objetivo similar que K media, pero funciona de manera distinta. En el inicio del proceso, se considera que cada dato es un grupo, luego se empiezan a agrupar los datos o los grupos cercanos con el algoritmo voraz con distinto criterio de fusión. El proceso termina cuando solo quedan K agrupaciones en el sistema. Los resultados del agrupamiento jerárquico son usualmente presentados en un dendrograma. Se detalla el algoritmo en el siguiente cuadro y un ejemplo en la Figura 2-3 (Dubien & Warde, 1979):

Se supone que hay  $N$  cantidad de datos de  $I$  dimensiones, y se van a generar  $K$  agrupaciones.

Paso 1: El proceso inicia con  $N$  agrupaciones.

Paso 2: Unir dos agrupaciones que cumplen con criterio de fusión (Müller, 2011).

1. *Ward*, fusionar los dos grupos que tienen varianzas mínimas.

$$\min \left( \sqrt{\frac{2 * n_A * n_B}{n_A + n_B} * \|\vec{c}_A - \vec{c}_B\|} \right), A, B \in 1..N$$

2. *Complete*, fusionar los dos grupos que tienen distancia máxima similares.

$$\max_{a \in A, b \in B} d(a, b)$$

3. *Single*, fusionar los dos grupos que tienen distancia mínima similares.

$$\min_{a \in A, b \in B} d(a, b)$$

4. *Average*, fusionar los dos grupos que tienen promedio de distancia intragrupo similares.

$$\frac{1}{n_A * n_b} \sum_{a \in A} \sum_{b \in B} d(a, b)$$

Leyenda:  $n_i$  es la cantidad de datos que pertenece a la agrupación  $i$ .  $\vec{c}_i$  es el centroide de la agrupación  $i$ .  $d(a, b)$  es la distancia euclidiana entre los puntos  $a$  y  $b$ .

Paso 3: Repetir paso 2 hasta que llega  $K$  cantidad de agrupaciones.

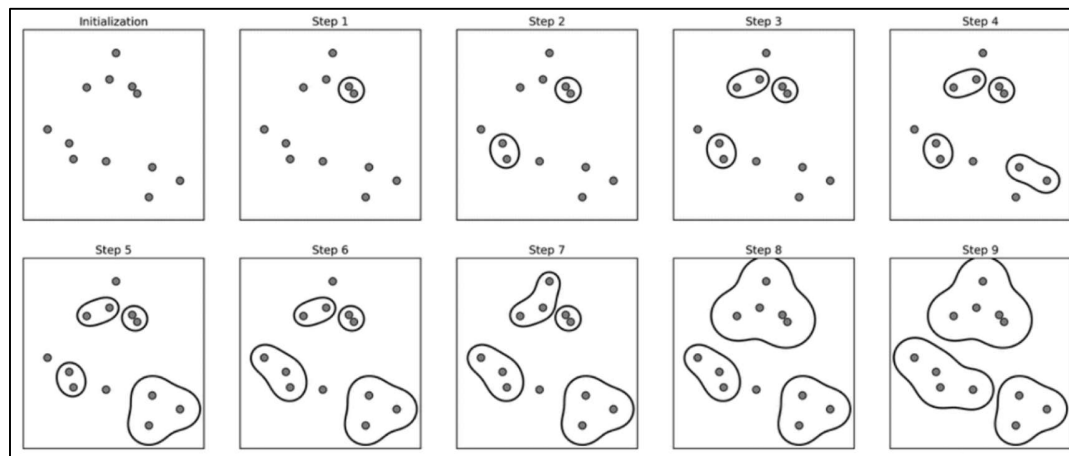


Figura 2-3. Ejemplo de agrupación jerárquica. Adaptado de Müller & Guido (2016).

*Density-based spatial clustering of applications with noise* (DBSCAN) es un método utilizado en las últimas décadas, que tiene su principal ventaja en generar agrupaciones de los datos según la densidad de estos, agrupando datos que son similares, pero puede estar distribuido en forma irregular (Müller & Guido, 2016). El algoritmo se ajusta con dos hiperparámetros, radio de búsqueda y cantidad mínima de datos en un grupo. El resultado de este algoritmo no es solo entregar las agrupaciones, sino también detectar los valores atípicos, como los datos que no son similares a otros datos agrupados. Se detalla los pasos del algoritmo de la siguiente forma (Ester et al., 1996; Schubert et al., 2017):

Se supone que hay  $N$  cantidad de datos de  $I$  dimensiones, el radio de búsqueda es  $r$  y la mínima cantidad de datos para formar una agrupación es  $MP$ .

Paso 1, revisar si dentro del radio de búsqueda desde el dato escogido existen más datos.

Paso 2, Si la respuesta del Paso 1 es positiva, saltar a los datos encontrados y realizar otras búsquedas con el mismo de radio de búsqueda. Marca los datos encontrados como estados visitados. Este paso se termina cuando no hay más datos por visitar dentro del radio de búsqueda.

Paso 3, ¿los datos visitados son mayores a  $MP$ ? Si la respuesta es positiva, marcan los datos visitados como una agrupación; en otro caso, marcan como datos atípicos.

Paso 4, seguir el paso 1, 2 y 3 con los datos no visitados.

En la Figura 2-4 se muestra una serie de ejemplos de los distintos resultados obtenidos con distintos ajustes del radio de búsqueda y la cantidad mínima de datos en un grupo. Se observa que cuando hay un radio de búsqueda mayor, todos los datos se agrupan como un solo grupo, perdiendo el objetivo de agrupación de los datos que tienen características similares. Sin embargo, cuando la cantidad mínima de datos en un grupo es mayor, se empieza a tener múltiples datos que fueron clasificados como valores atípicos. Lo mismo pasa cuando el radio de búsqueda es muy pequeño. Por lo tanto, para este algoritmo se debe encontrar un valor relativo de los dos hiperparámetros, que no se generen muchos datos de valor atípico ni formar todos los datos en un solo grupo.

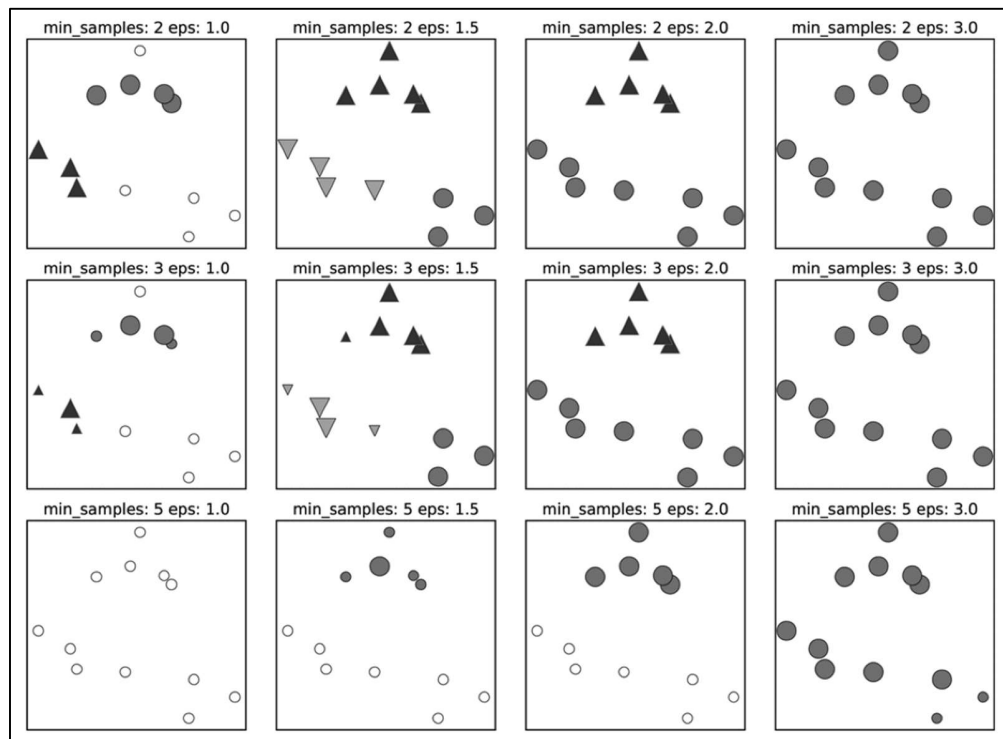


Figura 2-4. Ejemplo de DBSCAN. Adaptado de Müller & Guido (2016).

Otro algoritmo usado es el mapa autoorganizado (SOM: *self-organizing maps*). Este es un método de tipo red neuronal artificial, que podría tener mejor rendimiento que los algoritmos mencionados anteriormente de la familia de aprendizaje de máquinas, específicamente cuando se realizan agrupaciones para una base de datos con mayor dimensión. Se detallan los pasos del algoritmo en la siguiente forma y un ejemplo en la Figura 2-5 (Kohonen, 1998):

Se supone que hay  $N$  cantidad de datos de  $I$  dimensiones,

Paso 1, Hacer un mapa de vector de neuronas de pesos pequeños y aleatorios.

Paso 2, Tomar un vector de entrada  $D(t)$ .

Paso 3, Iterar por cada nodo de la neurona. Usar la distancia euclidiana para encontrar la similitud entre vector de entrada y los vectores de pesos de las neuronas del mapa. Luego encontrar el nodo que tiene menor distancia (este nodo es el que coincide más con el cuerpo, BMU (*best march unit*)).

$$d_v(X) = \operatorname{argmin} \left( \sum_{t=1}^N (D(t) - W_{vt}(s))^2 \right)$$

Paso 4, Actualizar los nodos en la vecindad del BMU, acercando al vector de entrada. Donde  $s$  es la iteración actual,  $t$  es el índice del vector de entrada del conjunto  $D$ ,  $v$  es el índice de un nodo en el mapa,  $W_v$  es el vector de peso del nodo  $v$ ,  $u$  es el índice del BMU en el mapa,  $\theta(u, v, s)$  es la función de vecindad,  $\alpha(s)$  es la restricción de aprendizaje según el progreso de las iteraciones.

$$W_v(s+1) = W_v(s) + \theta(u, v, s) * \alpha(s) * (D(t) - W_v(s))$$

Paso 5, Seguir iterando paso 2, 3 y 4 hasta llegar a la cantidad de iteración configurada o llegar a un BMU óptimo.

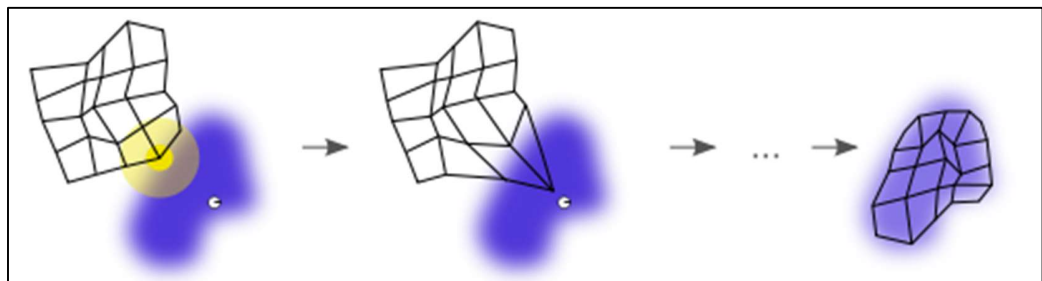


Figura 2-5. Ejemplo de Mapa autoorganizado. Adaptado de quora.com.



## 2.2. Criterios de medición de desempeño comparativo entre modelos

### 2.2.1. Notación

Para la definición de cada índice de validación de agrupación, se utilizará la siguiente notación. Se supone una base de datos  $X$  que tiene  $N$  datos u observaciones:  $X = \{x_1, x_2, x_3, \dots, x_N\}$ . Luego de la clasificación, resulta en un conjunto  $C$  que tiene  $K$  agrupaciones:  $C = \{c_1, c_2, c_3, \dots, c_K\}$ ,  $c_I$  es la agrupación  $I$ , donde  $I \in (1, \dots, K)$ . También debe indicarse que cada agrupación tiene  $M$  datos:  $M = \{m_1, m_2, m_3, \dots, m_K\}$ ,  $m_I$  es la cantidad de datos en la agrupación  $I$ , donde  $I \in (1, \dots, K)$ .  $x_i$  o  $x_j$  es el dato  $i$  o dato  $j$  de la base de datos.  $d_e(x_i, x_j)$  es la distancia euclidiana entre dato  $i$  y dato  $j$ .  $\bar{c}_I$  y  $\bar{c}_J$  es el centroide de la agrupación  $I$  y  $J$ .

### 2.2.2. Definición

Para evaluar la calidad de los resultados de agrupaciones obtenidas, se plantea medir con los índices de validación de agrupaciones (CVI: *Cluster validity index*). Cabe destacar que Fahad et al. (2014) en el estudio en que revisó 24 algoritmos de agrupación, concluyó que no existe un algoritmo de agrupación que pueda resultar mejor para todos los tipos de datos, ya que el rendimiento depende de la presencia de valores atípicos, densidad, monotonidad u otras características específicas de una base de datos. Por lo tanto, es importante escoger las métricas de calidad adecuadas para comprobar los resultados obtenidos.

Existen dos familias de índices de validación de agrupaciones, la interna (*Internal cluster validity index*) y la externa (*External cluster validity index*). La diferencia entre ellas radica en que en la validación externa se requieren etiquetas de cada agrupación, lo cual no está siempre disponible. El presente estudio plantea establecer un modelo para aplicar en etapas tempranas de la planificación minera, por lo mismo, no está disponible un modelo geometalúrgico previo, o dicho en otra forma, las etiquetas.

Rajabinasab & Asghari (2018) tuvieron éxito en validar las unidades geometalúrgicas generadas con los índices *Silhouette* (SI) (ecuación 2.5) y *Calinski Harabasz* (CH) (ecuación 2.9), pero estos fueron utilizados en un yacimiento de otro metal.

*Silhouette* (SI) mide la similitud dentro de una agrupación y la separabilidad entre agrupaciones mediante la medición de la distancia entre los puntos (datos).  $a(x_i)$  es el promedio de disimilitud entre objeto  $x_i$  con todos los otros objetos de la agrupación  $I$ , donde  $m_I$  es la cantidad de puntos en esta agrupación.  $b(x_i)$  es el promedio de disimilitud entre objeto  $x_i$  con los otros objetos que no pertenecen a la agrupación  $I$ ,  $N$  es la cantidad de puntos en la base de datos (Albelaitz et al., 2013; Rousseeuw, 1987).

$$SI = \frac{\sum_{c_I \in C} \sum_{x_i \in c_I} SI(x_i)}{N} \quad (2.5)$$

$$SI(x_i) = \frac{b(x_i) - a(x_i)}{\max\{a(x_i), b(x_i)\}} \quad (2.6)$$

$$b(x_i) = \min_{c_j \in C \setminus c_i} \frac{1}{m_j} \sum_{x_i \in c_i, x_j \in c_j} \{d_e(x_i, x_j)\} \quad (2.7)$$

$$a(x_i) = \frac{1}{m_i} \sum_{x_i, x_j \in c_i} d_e(x_i, x_j) \quad (2.8)$$

*Calinski Harabasz* (CH) es un índice de tipo ratio, que mide la compactibilidad de datos dentro de una agrupación relacionando la distancia de los puntos (datos) de cada agrupación y el centroide de ésta. También se mide la separabilidad de las agrupaciones a través de relacionar la distancia entre el centroide de cada agrupación y el centroide global. Se muestra la fórmula matemática de la siguiente manera, donde  $N$  es la cantidad de puntos en la base de datos y  $K$  es la cantidad de agrupaciones (Albelaitz et al., 2013; Calinski & Harabasz, 1974).

$$CH = \frac{N-K}{K-1} \frac{\sum_{c_i \in C} m_i d_e(\bar{c}_i, \bar{X})}{\sum_{c_i \in C} \sum_{x_i \in c_i} d_e(x_i, \bar{c}_i)} \quad (2.9)$$

Por otro lado, Liu et al. (2010) y Arbelaiz et al. (2013) demostraron que los índices *Silhouette*, *Calinski Harabasz* tuvieron buenos rendimientos en distintos escenarios planteados, como una base de datos con partición similares y diferenciada, de distintas cantidades de agrupación (2, 4 y 8), distintas cantidades de dimensiones (2, 4 y 8), datos con distintas densidades y presencia de valores atípicos. Por lo tanto, se propone estos dos índices para medir la calidad de agrupación.

### 2.3. Arquitectura del modelo

A partir de los procesos mencionados anteriormente, se plantean distintas combinaciones de algoritmos para poder determinar el modelo geometalúrgico.

Como se muestra en la Figura 2-6, el modelo empieza con una base de datos entregada como *input*, y a esta base de datos se le realiza el preprocesamiento en la primera etapa, a estos datos se les llamará datos preprocesados. Los datos preprocesados pasarán de forma directa como base de entrada para la clasificación, como también pasarán a la etapa de reducción de dimensión para pasar posteriormente a la etapa de clasificación. Los primeros serán llamados datos no reducidos, mientras que los segundos se catalogarán como datos reducidos. Es importante destacar que los datos preprocesados pasarán por estas dos vías, ya que lo que se quiere es verificar cuál generará un mejor rendimiento de clasificación

Ya teniendo ambas bases de entrada (datos reducidos como no reducidos) se llega a la clasificación, donde se puede clasificar con los métodos propuestos y comparar los resultados con los índices de calidad de agrupación mencionados.

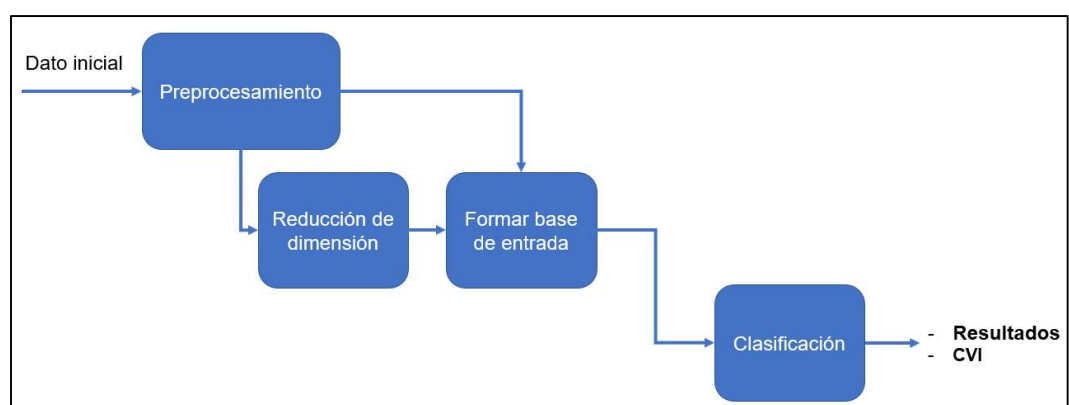


Figura 2-6. Arquitectura de modelos.

### **3. METODOLOGÍA DE PROCESAMIENTO DE DATOS**

#### **3.1. Bases de datos y software**

Michaux & O'Connor (2020) mencionaron en su caso de estudio que en la construcción de la base de datos para generar un modelo geometalúrgico, el número de variables de las muestras podría llegar a estar entre 150 y 350. Estos incluyen todos los análisis necesarios desde la etapa temprana de una mina como litología hasta la etapa final de producto como las pruebas de flotación.

La base de datos utilizada en este estudio incluye 1112 datos de sondeos con información geológica, mineralógica y metalúrgica de las muestras. Dentro de ésta se tienen variables cuantitativas tales como ubicación espacial (3 variables), concentración de elementos (9 variables), recuperación de cobre (1 variable), índice de trabajo de Bond (1 variable), índice de potencia específica (1 variable) y mineralógica (14 variables). Las variables categóricas de geologías son alteración (7 tipos), mineralización (14 tipos) y zonas minerales (3 tipos). Las Tablas 3-1 a 3-4 muestran las características principales de cada una de estas variables numéricas.

Tabla 3-1. Estadísticas descriptivas para variables en la base de datos – I.

	<b>Este</b> <b>[m]</b>	<b>Norte</b> <b>[m]</b>	<b>Cota</b> <b>[m]</b>	<b>CuT</b> <b>[%]</b>	<b>FeT</b> <b>[%]</b>	<b>Mo</b> <b>[%]</b>	<b>As</b> <b>[%]</b>
<b>Mínimo</b>	2841.0	2204.9	1448.3	0.13	0.02	0.00	0.00
<b>1er Cuartil</b>	3137.8	3586.7	2089.4	0.65	0.85	0.01	0.01
<b>2do Cuartil</b>	3294.2	4149.8	2221.9	0.87	1.26	0.03	0.01
<b>Media</b>	3347.6	4160.8	2223.7	1.04	1.72	0.05	0.04
<b>3er Cuartil</b>	3532.6	4782.3	2375.9	1.25	2.09	0.07	0.05
<b>Máximo</b>	4133.9	6038.2	2854.6	5.20	13.48	0.42	1.02
<b>Des. est,</b>	261.8	794.3	238.0	0.59	1.40	0.05	0.07

Tabla 3-2. Estadísticas descriptivas para variables en la base de datos –

## II.

	<b>Zn</b> <b>[%]</b>	<b>CuS</b> <b>[%]</b>	<b>SiO2</b> <b>[%]</b>	<b>Al2O3</b> <b>[%]</b>	<b>RecCu</b> <b>[%]</b>	<b>CuCo</b> <b>[%]</b>	<b>Wi</b> <b>[kWh/t]</b>	<b>SPI</b> <b>[kWh/t]</b>
<b>Mínimo</b>	0.00	0.01	8.60	4.00	53.60	21.70	10.03	0.46
<b>1er Cuartil</b>	0.01	0.02	66.10	12.80	85.60	28.60	13.01	2.32
<b>2do Cuartil</b>	0.02	0.04	70.90	14.40	88.70	31.40	14.11	3.23
<b>Media</b>	0.06	0.05	69.85	14.90	87.93	31.29	14.15	3.15
<b>3er Cuartil</b>	0.05	0.06	74.43	15.50	91.40	33.70	15.26	3.94
<b>Máximo</b>	2.15	0.40	94.20	76.90	97.00	42.90	19.77	8.36
<b>Des. est,</b>	0.12	0.04	10.09	7.88	4.97	3.94	1.60	1.17

CuS: cobre soluble, CuCo: Cobre en concentrado, Wi: Índice de trabajo de Bond, SPI: Índice de potencia específica.

Tabla 3-3. Estadísticas descriptivas para variables en la base de datos –

## III.

	<b>Cc</b>	<b>Dg</b>	<b>Cv</b>	<b>Bn</b>	<b>En</b>	<b>Cp</b>	<b>Py</b>
	<b>[%]</b>	<b>[%]</b>	<b>[%]</b>	<b>[%]</b>	<b>[%]</b>	<b>[%]</b>	<b>[%]</b>
<b>Mínimo</b>	0.0	0.0	0.0	0.0	0.0	0.0	0.0
<b>1er Cuartil</b>	0.0	0.0	0.0	0.0	0.0	0.0	15.9
<b>2do Cuartil</b>	0.0	2.9	3.2	1.6	0.0	20.2	40.0
<b>Media</b>	2.9	7.4	8.2	9.6	4.7	29.0	37.8
<b>3er Cuartil</b>	0.0	11.3	12.6	13.9	4.8	55.2	58.1
<b>Máximo</b>	97.0	50.0	61.8	77.4	72.9	100.0	93.3
<b>Des. est,</b>	10.8	9.7	11.2	15.0	9.6	30.1	24.4

Cc: Calcosina, Dg: Digenita, Cv: Covelina, Bn: Bornita,  
En: Enargita, Cp: Calcopirita, Py: Pirita.

Tabla 3-4. Estadísticas descriptivas para variables en la base de datos –

## IV.

	<b>Qz</b>	<b>QS</b>	<b>SVG</b>	<b>KSIL</b>	<b>PF</b>	<b>Clo</b>	<b>PIR</b>
	<b>[%]</b>	<b>[%]</b>	<b>[%]</b>	<b>[%]</b>	<b>[%]</b>	<b>[%]</b>	<b>[%]</b>
<b>Mínimo</b>	0.0	0.0	0.0	0.0	0.0	0.0	0.0
<b>1er Cuartil</b>	3.0	15.0	0.0	0.0	5.0	0.0	0.0
<b>2do Cuartil</b>	9.3	44.6	0.0	0.0	37.1	0.0	0.0
<b>Media</b>	15.1	51.4	4.0	7.1	36.9	0.6	3.2
<b>3er Cuartil</b>	22.0	94.7	4.6	6.6	62.4	0.0	3.7
<b>Máximo</b>	78.7	100.0	85.6	82.7	100.0	45.9	57.1
<b>Des. est,</b>	16.5	38.0	8.2	14.3	29.6	3.5	6.6

Qz: Cuarzo, QS: Cuarzo-sericita, SVG: Alteración sericita gris verde,  
KSIL: Alteración potásico intensivo, PF: Alteración pirofilita,  
Clo: Clorita, PIR: Alteración pirítica

Como plataforma de *software*, se utiliza Python para el procesamiento de datos y la generación del modelo final, utilizando todas las librerías necesarias. En la Tabla 3-5 se muestran las librerías utilizadas en cada etapa de modelo.

Tabla 3-5. Librerías de *python* utilizadas.

<b>Librería</b>	<b>Función</b>
Pybalu	Eliminar variables dependientes y/o constantes.
NumPy	Creación de matriz y vectores necesarios para el procesamiento de datos.
Pandas	Complementaria a Numpy para manipulación y análisis de datos.
Scikit-learn	Aprendizaje de máquinas, contiene algoritmo de aprendizaje supervisado y no supervisado.
Keras	Construcción de redes neuronales para ejecutar en Tensorflow
Tensorflow	Construcción y ejecución del sistema redes neuronales.
minisom	Construcción y ejecución del aprendizaje no supervisado mapa autoorganizado.
Matplotlib	Gráficos de los resultados

### 3.2. Preparación de datos de entrada

En esta etapa de preparación de datos de entrada, se dividen las variables en distintos grupos según algunos supuestos de la importancia de estas.

En el primer grupo, se priorizan las variables cobre total (CuT), cobre soluble (CuS), cobre en el concentrado (CuCO) y la recuperación metalúrgica. Éstas son esenciales para un proyecto de explotación de minerales de cobre, donde el cobre total tiene una implicancia directa en el valor del proyecto. El cobre soluble afecta la selección del proceso, debido a que éste indica la presencia de óxidos de cobre que preferentemente se procesan en plantas de lixiviación, y el cobre en el concentrado, indica la calidad del producto. Finalmente, la recuperación metalúrgica es otra



variable que tiene un efecto directo en el valor del proyecto, ya que dice relación con la productividad técnica del proceso.

En el segundo lugar, se incorporan variables de concentración reconocidas como “Otras leyes”, que incluyen Fe, Mo, As, Zn, SiO<sub>2</sub> y Al<sub>2</sub>O<sub>3</sub>, donde todas ellas se identifican como impurezas en el proceso.

El otro grupo de variables que se propone dice relación con variables de “Proceso”, el cual incluye el índice de trabajo de Bond y el índice de potencia específica.

Por último, el grupo “Mineralógica” son las 14 variables de porcentajes de minerales en cada muestra. El grupo “Geológicas” son las 24 variables binarias categóricas de geología.

La agrupación de variables se muestra gráficamente en la Figura 3-1.

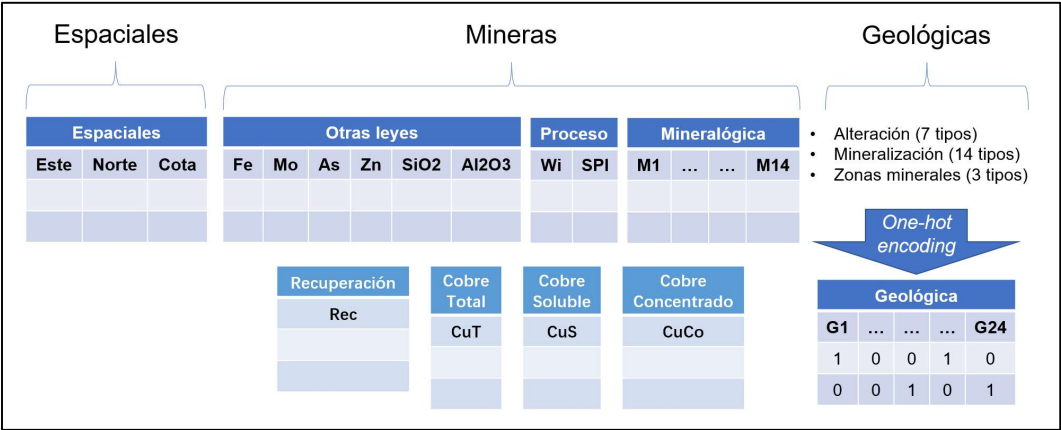


Figura 3-1. Grupos de variables.

Como se muestra en la figura 3-1, se tienen 3 variables espaciales, 6 de otras leyes, 2 de proceso, 14 mineralógicas y 4 variables por si solas, con un total de 29 variables cuantitativas. Por otro lado, las variables categóricas geológicas, después de realizar la codificación *One-hot*, corresponden a otras 24 variables binarias.

3.3. Preprocesamiento

En el preprocesamiento de los datos se busca minimizar el efecto de distorsión por distintos rangos de variables y eliminar variables que son dependientes.

3.3.1. Clean

Se aplica el *clean* para la base de datos propuesta en la sección 3.2, siendo necesario solo eliminar una variable binaria de la variable geológica zona mineral de código F, la cual indica zona secundaria fuerte. Finalmente, la base de datos queda con 52 variables, como se observa en la Figura 3-2:

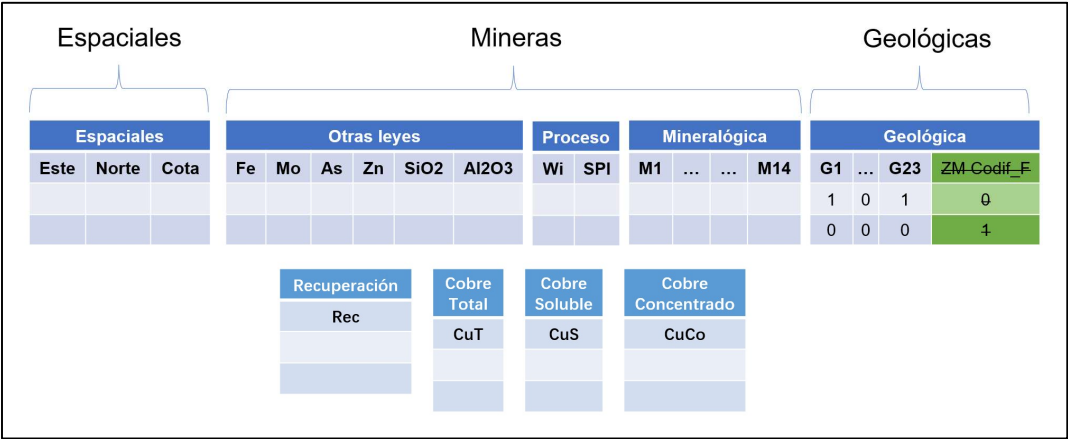


Figura 3-2. Grupos de variables después de *clean*.

3.3.2. Normalización

Dado que cada variable tiene distintas escalas debido a su naturaleza, que las variables geo-minero-metalúrgico no tienen valores negativos y la presencia de variables binarias, se propone utilizar normalización.

La transformación queda en la siguiente forma (Müller & Guido, 2016):

$$x_{i'} = \frac{x_i - \min(\vec{x})}{\max(\vec{x}) - \min(\vec{x})} \tag{3.1}$$

Este re-escalamiento permite generar las variables con una escala entre 0 y 1, y mantener la distribución de cada variable.

3.4. Reducción de dimensiones

Los principales beneficios de la reducción de dimensiones de las variables son: mejorar el rendimiento del modelo aprendizaje de máquinas, reducir el tiempo de

procesamiento y almacenamiento de datos y visualizar datos de mayor dimensión en 2 o 3 dimensiones (Müller & Guido, 2016).

En este estudio se propone una reducción de dimensiones por grupo de variables, en lugar de la forma habitual de reducir a partir de la base de datos en general. Esto debido a la consideración de que las variables cobre total (CuT), cobre soluble (CuS), cobre en concentrado (CuCO) y la recuperación metalúrgica de cobre tienen alta incidencia en la cadena de valor, por lo que es necesario preservar el peso de estas variables en el proceso de agrupación.

Los grupos de variables a reducir se muestran en la Tabla 3-6. Cada grupo se reduce a 1 dimensión.

Se considera que cada grupo tiene una implicancia similar a cobre total (CuT), cobre soluble (CuS), cobre en concentrado (CuCO) y la recuperación metalúrgica de cobre. La reducción se realiza a través de dos métodos y se compara la pérdida de información para elegir uno de los dos.

Tabla 3-6. Grupos de variables que se reducen a 1 dimensión.

Nombre de grupo		Variables
Espacial		Este, Norte, Cota
Otras leyes		Fe, Mo, As, Zn, SiO3, Al2O3
Proceso		Wi, SPI
Mineralógica	Cc, Dg, Cv, Bn, En, Cp, Py, Qz, Qs, SVG, KSIL, PF, Clo, PIR	

3.4.1. *Principal Component Analysis (PCA)*

El primer método aplicado es Análisis de Componentes Principales. En la Tabla 3-7 se presenta los resultados de la proporción de varianza explicada (*Explained variance ratio*) a los distintos grupos. La pérdida de información se calcula como la diferencia porcentual con respecto a la varianza explicada. En este estudio se propone reducir cada grupo de variable a 1 dimensión, es decir, el primer componente del análisis.

Tabla 3-7. Información representada de la base original.

Nombre de grupo	Proporción de varianza explicada [%]	Pérdida de información [%]
Espacial	72.71	27.29
Otras leyes	44.02	55.98
Proceso	80.47	19.53
Mineralógica	58.95	41.05

El resultado mostrado en la Tabla 3-7 indica que el grupo “Proceso” presenta más información explicada, 80.47%. Esto significa que tiene una pérdida de información igual a 19.53%. En cambio, los otros tres grupos tienen más pérdida de información, teniendo los grupos “Otras leyes” y “Mineralógica” pérdidas superiores a 40%, lo que los hace poco aceptables.

La cantidad inicial de las variables de cada grupo tiene implicancia en el resultado de la proporción de varianza explicada del primer componente (reducido a 1 dimensión). Los grupos que tienen mayor cantidad de variables tienden a tener

mayor pérdida de información. Por lo mismo, la limitación de aplicar PCA es que podría ser poco efectivo en el caso que exista una alta cantidad de variables y se busque reducir a 1 dimensión.

### 3.4.2. *Autoencoder*

El segundo método corresponde a utilizar un *autoencoder* para la reducción de dimensiones. Es importante destacar que para usar esta técnica, se debe tener una arquitectura variable en función de la cantidad de variables de cada grupo.

Para definir la mejor arquitectura de autoencoder para la reducción de dimensiones se utiliza el error medio cuadrático (MSE). La principal consideración de la elección es que en el MSE se penaliza en mayor medida la diferencia entre valor real y estimado. El MSE se define en la siguiente expresión, asumiendo que hay  $N$  variables en la base de entrada y  $M$  datos:

$$MSE = \frac{1}{NM} \sum_{i=1}^N \sum_{j=1}^M (X_{ij} - \hat{X}_{ij})^2 \quad (3.2)$$

Se probaron distintas combinaciones de función de activación mencionadas en la sección 2.1.3 y los mejores resultados se obtienen con las siguientes arquitecturas (Tabla 3-8). En la columna codificación y decodificación se muestran el número de neuronas activas en cada capa oculta y la función de activación en esta capa.

Tabla 3-8. Arquitectura de *autoencoder*.

Nombre de grupo	Codificación	Espacio latente	Decodificación
Espacial	3 ('tanh')	1	3 ('relu')
Otras leyes	6 ('sigmoid')	1	6 ('sigmoid')
Proceso	2 ('relu')	1	2 ('relu')
Mineralógica	14 ('tanh'), 9 ('relu'), 5 ('relu')	1	5 ('relu'), 9 ('relu'), 14 ('relu')

En la Tabla 3-9 se presenta el promedio y desviación estándar para cada grupo y de los resultados de decodificación, así como los valores de MSE. Los  $\mu$  y  $\sigma$  indican el promedio y desviación estándar. Las letras  $i$  y  $f$  diferencian que son datos iniciales ( $X_{ij}$ ) y el resultado de decodificación ( $\hat{X}_{ij}$ ). Se muestra que los valores de promedio y desviación estándar son muy similares y los valores de MSE son aceptable.

Tabla 3-9. Información estadística del resultado de *autoencoder*.

Nombre de grupo	$\mu_i$	$\sigma_i$	MSE	$\mu_f$	$\sigma_f$
Espacial	0.484	0.205	0.0104	0.483	0.18
Otras leyes	0.194	0.258	0.0044	0.197	0.257
Proceso	0.4	0.154	0.0046	0.4	0.138
Mineralógica	0.177	0.26	0.0157	0.169	0.243

Como métrica de comparación entre PCA y *autoencoder*, ya que no hay una forma específica para compararlos directamente, se plantea la comparación en MSE para realizar esta tarea. Donde se compara la diferencia entre la base de datos original y la base reconstruida. Para el caso de *autoencoder*, esta comparación está involucrada en la función de pérdida. Para el caso de PCA, se realiza una transformación inversa

de PCA para reconstruir la base desde el primer componente. Como se muestra en la Figura 3-3, el proceso de PCA reduce la base  $X$  a 1 dimensión ( $\tilde{X}$ ), luego se reconstruye la base a  $\hat{X}$  y se calcula el MSE entre  $X$  y  $\hat{X}$  como *autoencoder*.

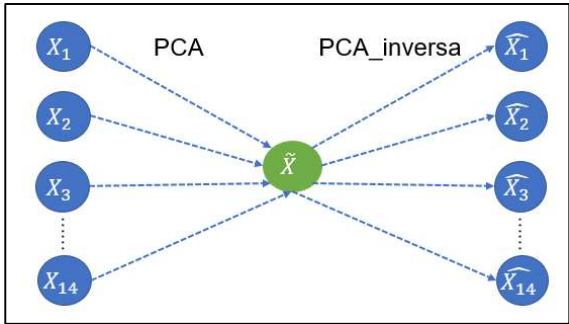


Figura 3-3. Flujo de PCA y PCA\_inversa.

En la Tabla 3-10 se observa que los promedios de las bases reconstruidas ( $\mu_f$ ) por PCA inversa son muy similares a los promedios de las bases originales ( $\mu_i$ ). Sin embargo, las desviaciones estándares tienden a ser menores en las bases reconstruidas ( $\sigma_f$ ). Esto indica que el proceso de reducción y reconstrucción disminuye la variabilidad de datos.

Tabla 3-10. Información estadística del resultado de PCA inversa.

Nombre de grupo	$\mu_i$	$\sigma_i$	MSE	$\mu_f$	$\sigma_f$
Espacial	0.484	0.205	0.0102	0.485	0.178
Otras leyes	0.194	0.258	0.0056	0.194	0.246
Proceso	0.4	0.154	0.0045	0.4	0.139
Mineralógica	0.177	0.26	0.0184	0.177	0.222



La Tabla 3-11 muestra el resultado de MSE para ambas técnicas de reducción de dimensión por grupo de variable. En los grupos de bajas dimensiones iniciales tienen pocas diferencias y el MSE de PCA son mejores que de *Autoencoder* en 2%. Sin embargo, tienen una disminución de MSE significativa en los grupos de mayores dimensiones, que son 21% para el caso de “Otras leyes” y 15% para “Mineralógica”. Por lo tanto, la técnica *Autoencoder*, tiene ventaja comparativa en reducción de dimensiones a partir de mayor cantidad de variables.

Tabla 3-11. Tabla comparativa de MSE.

Nombre de grupo	Dimensión inicial	PCA (a)	<i>Autoencoder</i> (b)	$\frac{a-b}{a} \cdot 100\%$
Espacial	3	0.0102	0.0104	-2
Otras leyes	6	0.0056	0.0044	21
Proceso	2	0.0045	0.0046	-2
Mineralógica	14	0.0184	0.0157	15

### 3.5. Formalización de la base de entrada a la clasificación

Se plantean distintas bases alternativas de entrada a la clasificación, cada una con un enfoque particular. La idea central es considerar las 4 variables cobre total (CuT), cobre soluble (CuS), cobre en concentrado (CuCO) y la recuperación metalúrgica de cobre como base, y combinarlas con distintas formas alternativas de otras variables para así tener distintas bases de entrada para la clasificación. Se muestra las combinaciones en la Figura 3-4.

Variable base	Variables geológicas	Otras leyes		Proceso			Mineralógica		Nombre de la base
		Original	Reducida	Original	Solo Wi	Reducida	Original	Reducida	
	■	■		■			■		Original
Cobre Total		■		■			■		Original sin geología
Cobre Soluble		■			■		■		Original sin geología ni SPI
Cobre Concentrado			■		■			■	Reducida parcial sin SPI
Recuperación			■	■				■	Reducida parcial con SPI
			■			■		■	Reducida completa

Figura 3-4. Combinación de la base de datos de entrada.

La base denominada “Original”, es la base original de 52 variables que contiene todas las 29 variables cuantitativas de la base de datos y 23 variables binarias geológicas, ya que una variable binaria geológica fue eliminada por el método *clean*.

La base “Original sin geología” es la base Original sin la inclusión de las variables binarias geológicas generadas a partir de las variables categóricas de geología (alteración, mineralización y zonas minerales), por consiguiente, tiene 29 variables.

La base “Original sin geología ni SPI”, es la base “Original sin geología” eliminando la variable SPI, se considera en caso de que una faena no tenga el equipo molino semi-autógeno.

La base “Reducida parcial con SPI” contiene las 4 variables no reducidas, cobre total (CuT), cobre soluble (CuS), cobre en concentrado (CuCO) y la recuperación

metalúrgica de cobre, adiciona las variables reducidas de los grupos “Otras leyes” y “Mineralogía”, además incluye las variables de procesos índice de trabajo de Bond ( $W_i$ ) e índice de potencia específica (SPI) no reducidas.

La base “Reducida parcial sin SPI”, es la misma base “Reducida parcial con SPI”, eliminando la variable índice de potencia específica (SPI), se considera en un caso de que una faena no tenga el equipo molino semi-autógeno.

Para el último caso, la base “Reducida completa” contiene las 4 variables no reducidas, cobre total (CuT), cobre soluble (CuS), cobre en concentrado (CuCO) y la recuperación metalúrgica de cobre, y adicionar las variables reducidas de los grupos “Otras leyes”, “Mineralógica” y “Proceso”.

Por último, para incluir las variables espaciales, se proponen 3 formas distintas. La primera es incluir las variables originales de 3 dimensiones (Este, Norte y Cota), la segunda es incluir la variable con reducción de dimensiones, esta tiene sola 1 dimensión, y la última forma es no incluir las variables espaciales. Esto se presenta en la Figura 3-5.

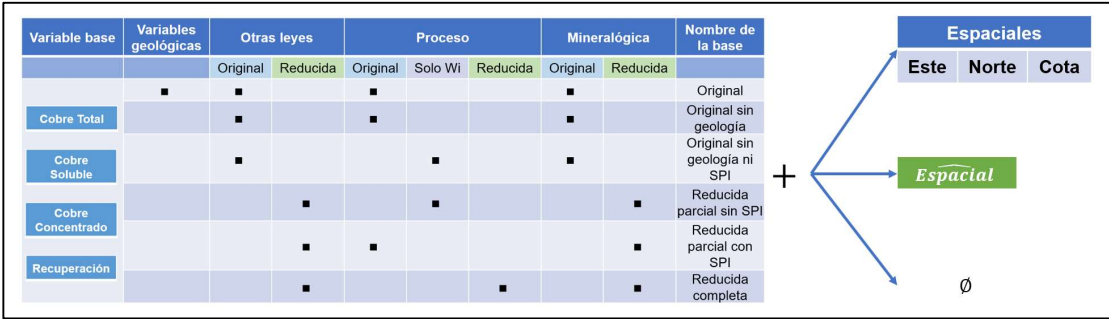


Figura 3-5. Combinación de inclusión de las variables espaciales.

3.6. Clasificación e hiperparámetros

Para cada algoritmo de agrupación se tienen distintos hiperparámetros, en esta sección se especificará la configuración de cada uno.

3.6.1. k-means

El único hiperparámetro del algoritmo *k-means* es la cantidad de agrupaciones requeridas. Sin embargo, existe otro hiperparámetro funcional que se debe tener en cuenta, este es el máximo número de iteración, el cual se debe configurar lo suficientemente grande para no llegar a un resultado subóptimo. Finalmente, otro hiperparámetro funcional que se debe tener en cuenta es la semilla de aleatoriedad donde se genera el centroide de inicialización. Se detalla la configuración de hiperparámetros en la siguiente tabla.

Tabla 3-12. Hiperparámetros de *k-means*.

Hiperparámetro	Valor
Cantidad de agrupación	[3, 50]
Máximo número de iteración	300

### 3.6.2. Agrupación jerárquica (AGG)

El primer hiperparámetro de AGG corresponde a la cantidad de agrupaciones requeridas, similar al hiperparámetro usado en *k-means*. El otro hiperparámetro es la forma de unión (*linkage*), el cual existe cuatros diferentes forma (Müller, 2011).

Tabla 3-13. Hiperparámetros de AGG.

Hiperparámetro	Valor
Cantidad de agrupación	[3, 50]
Forma de unión ( <i>linkage</i> )	‘ward’, ‘complete’, ‘single’, ‘average’

### 3.6.3. *Density-based spatial clustering of applications with noise* (DBSCAN)

Para el caso de DBSCAN, se tiene una mayor cantidad de hiperparámetros en comparación a los otros algoritmos. Estos hiperparámetros son radio de búsqueda y cantidad mínima de datos en un grupo. Con respecto a los demás hiperparámetros funcionales se dejaron por defecto de la librería.

Tabla 3-14. Hiperparámetros de DBSCAN.

Hiperparámetro	Valor
Radio de búsqueda	[0.05, 4] con salto de 0.05
Cantidad mínima de datos en un grupo	[2, 40]

### 3.6.4. Mapas autoorganizados (SOM)

Para el caso de los mapas autoorganizados, se tienen 4 hiperparámetros principales, los cuales son radio de avance de los nodos ( $\sigma$ ), tasa de aprendizaje, tamaño de malla (cantidad de nodos) y función de vecindad. En tamaño de malla se utiliza una función empírica que propuso Vesanto et al. (2000). Para la función de vecindad se ocupó la más usual de la librería minisom de Python.

$$\text{Cantidad de nodos por lado} = \text{round}(\sqrt{5 * \sqrt{N}}) \quad (3.3)$$

Tabla 3-15. Hiperparámetros de SOM.

Hiperparámetro	Valor
Radio de avance de los nodos ( $\sigma$ )	[0.1, 1, 10]
Tasa de aprendizaje	[0.0001, 0.001, 0.01, 0.1]
Tamaño de malla (cantidad de nodos)	13*13
Función de vecindad	['gaussian', 'mexican hat', 'bubble']

### 3.7. Métricas

Como se mencionó en la sección 2.2, no hay un solo índice de calidad de agrupación que sea mejor que otro (Fahad, 2014). Desde las bibliografías se deduce que *Silhouette* (SI) y *Calinski Harabasz* (CH) son los índices más estables (Liu et al.

2010 & Arbelaiz et al. 2013). El índice SI, varía entre -1 y 1, donde el valor 1 indica que los datos de la misma agrupación son similares y exhiben mayor diferencia con los de otras agrupaciones, mientras que el valor -1 indica lo contrario. El CH varía entre 0 e infinito, donde un mayor valor indica menor covarianza dentro de la agrupación y mayor covarianza entre agrupaciones.

Estos dos índices cumplen con el objetivo de que cada unidad geometalúrgica tenga su propia característica y se diferencie suficientemente entre unidades. Por otro lado, de los estudios revisados, en el caso aplicado a un yacimiento de hierro se utilizaron los mismos índices (Rajabinasab & Asghari, 2018). Por lo tanto, se propone utilizar estos dos índices para la definición del modelo geometalúrgico.

#### 4. RESULTADOS DEL MODELAMIENTO GEOMETALÚRGICO Y ANÁLISIS DE VALOR ECONÓMICO

A continuación, se describirá los resultados obtenidos. En primer lugar, se presenta el resultado según los índices (SI y CH). Posteriormente se analizarán los distintos resultados, y finalmente, se realizará una evaluación económica analizando la influencia de la secuencia de extracción incluida por el modelo sobre el valor del proyecto.

##### 4.1. Resultados generales

En la Tabla 4-1, se muestran 8 resultados, los cuales presentan los 2 mejores resultados de cada algoritmo de agrupación. La columna forma de variables espaciales tiene 3 posibilidades: 0 indica que no se incluye ninguna variable espacial; 1 indica que se incluye la variable espacial con reducción de dimensión a 1 y 3 indica que la base está formada con las variables espaciales de 3 dimensiones. Los resultados son evaluados con la base de datos “Original sin geología” con 3 variables espaciales.

Según el resultado en *Silhouette* (SI), el mejor algoritmo de agrupación es *k-means*. Por otro lado, AGG y SOM tienen indicadores SI 6% y 8% menor que *k-means*, mientras que DBSCAN obtiene los peores resultados. Una observación a destacar es que la base de datos “Originales sin geología ni SPI” y “Original sin geología” tienen mayor presencia en los mejores resultados, y el mejor resultado del índice SI es la base que no incluye variables espaciales tanto en *k-means* como AGG.



Tabla 4-1. Mejores resultados de *Silhouette* de cada algoritmo.

Algoritmo de agrupación	Base de datos de entrada	Forma de variables espaciales	SI
<i>k-means</i>	Original sin geología ni SPI	0	<b>0.264</b>
<i>k-means</i>	Original sin geología	0	<b>0.263</b>
AGG	Original sin geología ni SPI	0	<b>0.249</b>
AGG	Original sin geología	1	<b>0.249</b>
SOM	Reducida parcial sin SPI	3	<b>0.242</b>
SOM	Reducida parcial sin SPI	1	<b>0.227</b>
DBSCAN	Original sin geología	0	<b>0.205</b>
DBSCAN	Original sin geología	3	<b>0.205</b>

La Tabla 4-2 muestra los resultados según el indicador *Calinski Harabasz* (CH).

Resulta similar al indicador SI. El algoritmo *k-means* tiene el mejor resultado, AGG y SOM tienen menores indicadores en 6% y 14% respectivamente, mientras que DBSCAN tiene el peor resultado.

Tabla 4-2. Mejores resultados de *Calinski Harabasz* de cada algoritmo.

Algoritmo de agrupación	Base de datos de entrada	Forma de variables espaciales	CH
<i>k-means</i>	Original sin geología ni SPI	0	<b>485.21</b>
<i>k-means</i>	Original sin geología	0	<b>482.46</b>
AGG	Original sin geología	3	<b>455.42</b>
AGG	Original sin geología ni SPI	3	<b>452.46</b>
SOM	Original sin geología	0	<b>416.89</b>
SOM	Original sin geología	1	<b>415.83</b>
DBSCAN	Original sin geología	3	<b>288.40</b>
DBSCAN	Original sin geología ni SPI	0	<b>263.55</b>

La Tabla 4-3 muestra otra forma de resultado compuesto, cual es la multiplicación entre el indicador *Silhouette* (SI) y el *Calinski Harabasz* (CH). Ya que ambos

indicadores buscan mayor valor, en esta nueva forma los mayores valores indican mejores resultados. El resultado no es muy distinto que los obtenidos con los indicadores individuales, y el algoritmo *k-means* obtiene los mejores resultados.

Tabla 4-3. Mejores resultados de la combinación de cada algoritmo.

Algoritmo de agrupación	Base de datos de entrada	Forma de variables espaciales	SI*CH
<i>k-means</i>	Original sin geología ni SPI	0	<b>127.86</b>
<i>k-means</i>	Original sin geología	0	<b>126.70</b>
AGG	Original sin geología	3	<b>112.59</b>
AGG	Original sin geología ni SPI	3	<b>112.54</b>
SOM	Original sin geología	0	<b>85.78</b>
SOM	Original sin geología	1	<b>85.43</b>
DBSCAN	Original sin geología	3	<b>56.92</b>
DBSCAN	Original sin geología ni SPI	0	<b>48.49</b>

#### 4.2. Análisis de sensibilidad

En esta sección, se analizará la sensibilidad de desempeño de las distintas bases de entrada formadas en la sección 3.5 con el algoritmo de *k-means*. En las Figuras 4-1 y 4.2 se muestran la relación entre la cantidad de agrupaciones y los indicadores de calidad de agrupaciones. Se puede apreciar que cuando hay mayor cantidad de agrupaciones hay menor valor del indicador, esto significa que tienen peor calidad de agrupación. Desde otra perspectiva, se tiene que la base de entrada “original sin geología ni SPI” presenta mayor valor en los indicadores en la mayoría de los escenarios.

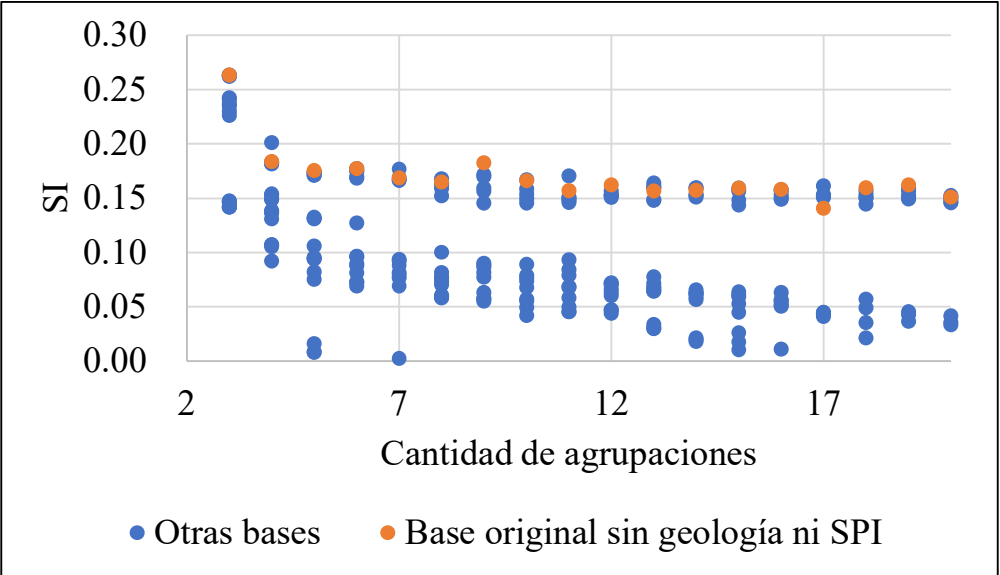


Figura 4-1. Análisis de sensibilidad en *k-means* con el indicador *Silhouette*.

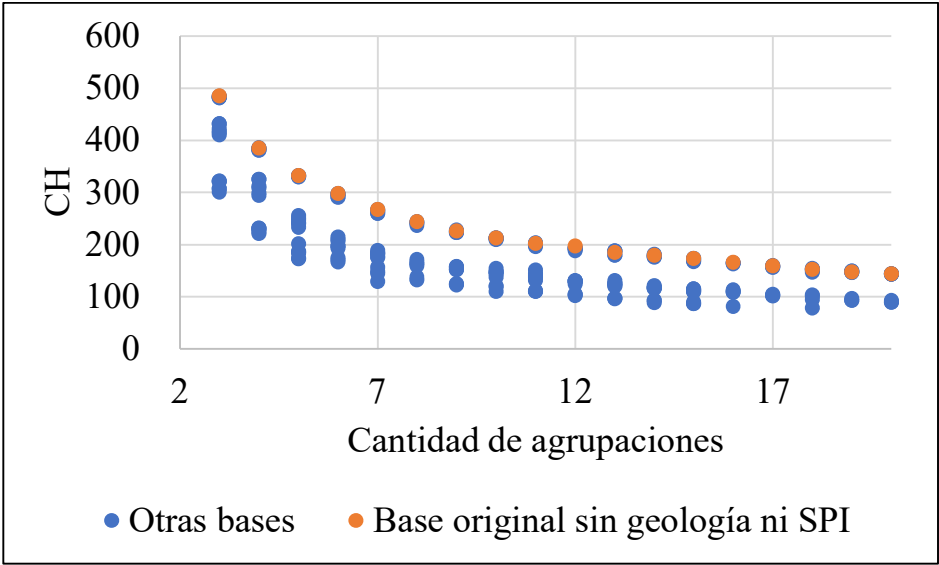


Figura 4-2. Análisis de sensibilidad en *k-means* con el indicador *Calinski Harabasz*.

En la Tabla 4-4 se muestra que para el mismo algoritmo de K-media con K igual 3, varía la calidad de agrupación. Se observa que las bases “Original sin geología” y “Original sin geología ni SPI” tienen similares resultados. Incluso no hay mayor diferencia en el valor de los indicadores con la inclusión de la variable espaciales. En el caso del indicador CH, se alcanzan mejores resultados con las bases donde no se incluyen las variables espaciales En el caso del indicador SI, los 6 primeros casos son muy similares.

Tabla 4-4. Análisis de sensibilidad de bases de entrada.

<b>Base de entrada</b>	<b>Forma de variable espacial</b>	<b>SI</b>	<b>CH</b>
Original sin geología ni SPI	0	0.264	485.21
Original sin geología	0	0.263	483.46
Original sin geología	3	0.263	481.68
Original sin geología ni SPI	3	0.263	481.57
Original sin geología ni SPI	1	0.263	483.59
Original sin geología	1	0.262	484.55
Reducida parcial sin SPI	3	0.243	431.68
Reducida parcial sin SPI	1	0.241	431.53
Reducida parcial con SPI	3	0.237	422.90
Reducida completa	3	0.235	417.52
Reducida parcial con SPI	1	0.230	414.20
Reducida completa	1	0.226	410.01
Reducida completa	0	0.147	307.46
Reducida parcial con SPI	0	0.147	307.46
Reducida parcial sin SPI	0	0.148	300.81
Original	1	0.142	322.25
Original	3	0.142	321.54
Original	0	0.142	320.61

La Tabla 4-4, se muestra que las bases originales incluyendo variables geológicas presentan peor calidad de agrupación, indicando que las variables geológicas

categorías no favorecen la eficiencia de agrupación. Por otro lado, las bases con reducción de dimensiones tampoco mejoran el desempeño de la agrupación. Aunque los 6 mejores resultados tienen indicadores muy similares, se escoge el mejor caso para análisis posteriores.

#### **4.3. Resultado final y análisis de varianza**

En las Figuras 4-3 y 4-4 se presenta la visualización espacial del modelo geometalúrgico generado por la base “Original sin geología ni SPI” y sin ninguna variable espacial, modelo que tiene el mayor valor en los indicadores. Las figuras muestran que en cada unidad geometalúrgica, los puntos (datos) se encuentran unidos espacialmente, con un borde claro entre unidades. Esto a pesar de que en la agrupación la base de entrada no incluye ninguna variable espacial. Este resultado valida el modelo obtenido y es consecuente con la distribución esperada en un yacimiento de cobre, ya que los cuerpos minerales son resultados de evoluciones geológicas en distintas etapas. Por lo tanto, la continuidad de una unidad geometalúrgica está íntimamente relacionada con los procesos de formación de yacimiento y no dado por las variables espaciales.

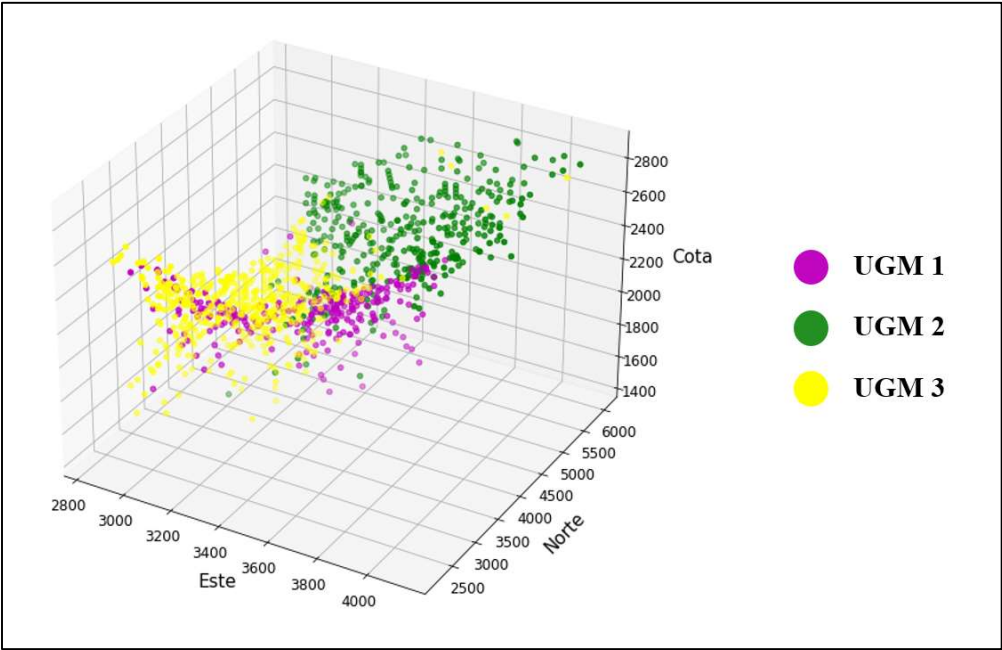


Figura 4-3. Visualización espacial de modelo geometalúrgico resultado – I.

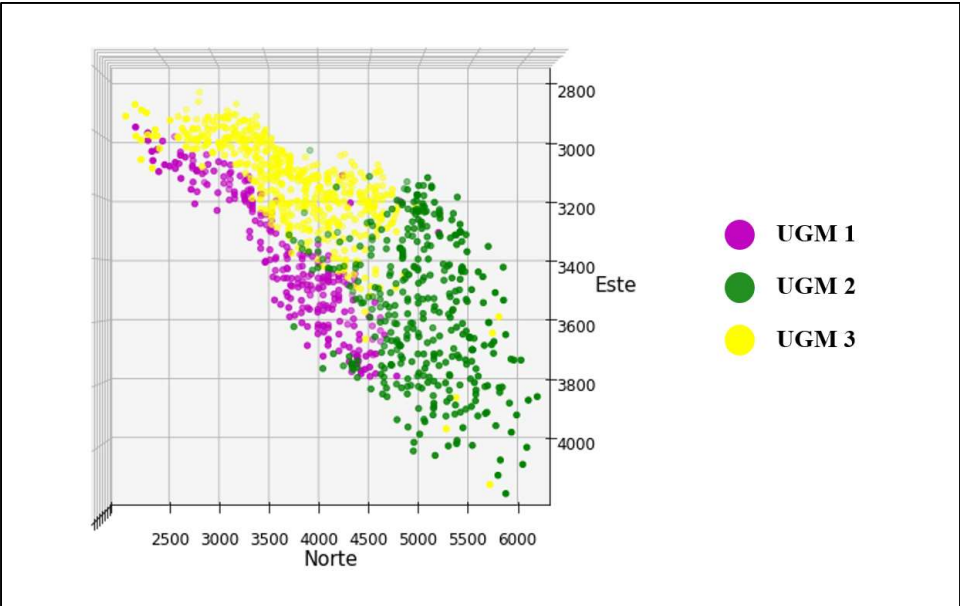


Figura 4-4. Visualización espacial de modelo geometalúrgico resultado – II.

En las Figuras 4-5, 4-6 y 4-7 se muestran diagramas de cajas para la variable cobre total, recuperación metalúrgica e índice de trabajo de Bond. En la variable de cobre total, la unidad geometalúrgica1 (UGM1) presenta un mayor promedio de la ley del cobre. Asimismo, la unidad geometalúrgica 1 (UGM1) tiene mayor recuperación metalúrgica, seguida de las unidades UGM3 y UGM2. Finalmente, el índice de trabajo de Bond se incrementa en el orden UGM1, UGM2 y UGM3, lo cual implica que la explotación de la zona de la unidad UGM3 induciría una menor capacidad de producción en la planta, si hay restricciones de capacidad instalada.

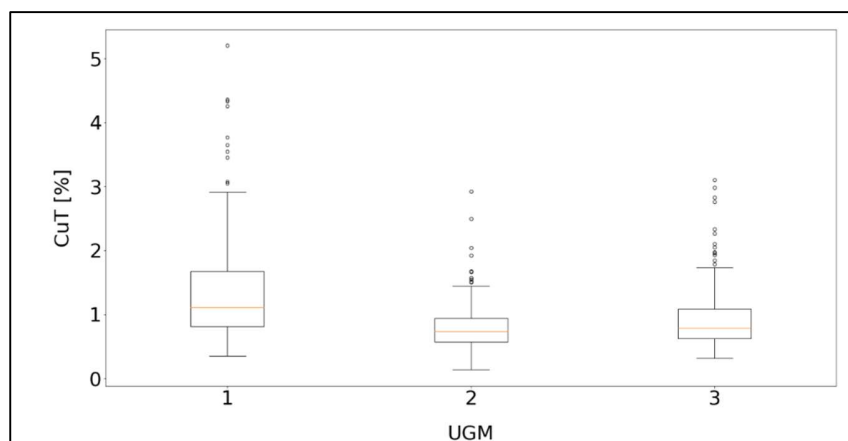


Figura 4-5. Diagrama de cajas – Cobre total en cada unidad geometalúrgica.

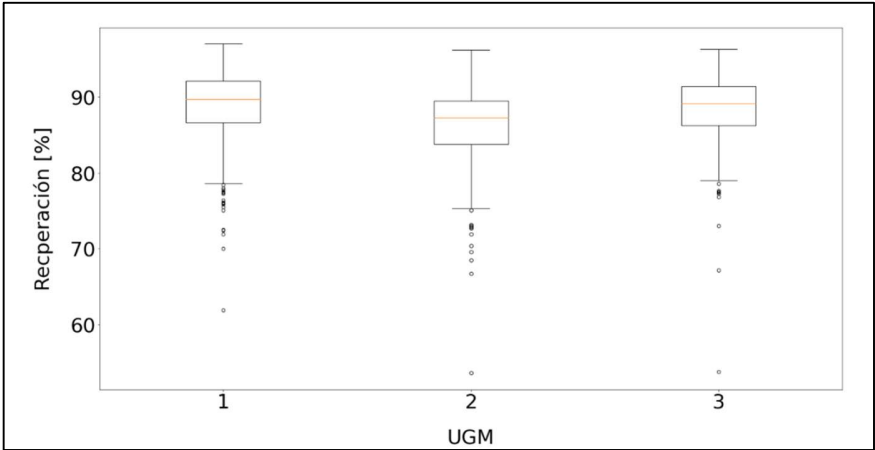


Figura 4-6. Diagrama de cajas – Recuperación metalúrgica en cada unidad geometalúrgica.

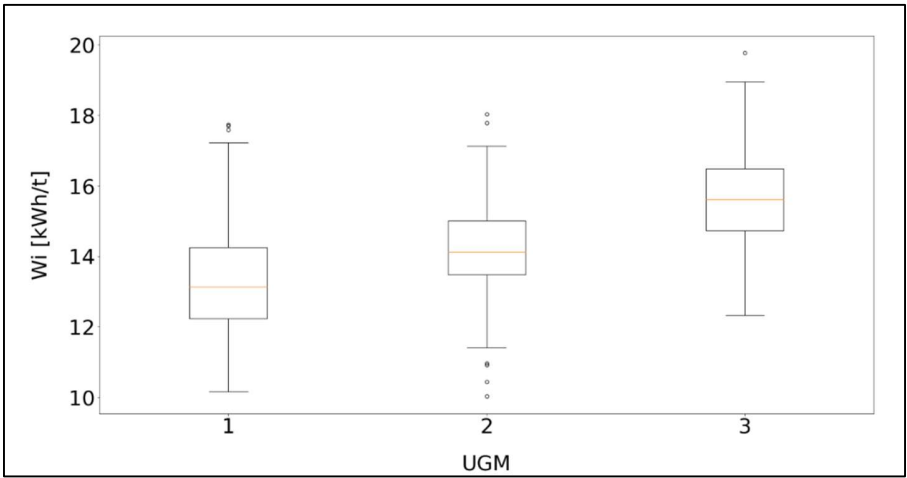


Figura 4-7. Diagrama de cajas – Índice de trabajo de Bond en cada unidad geometalúrgica.

Por otro lado, dado que las variables principales no son estadísticamente tan diferentes, se realiza un análisis de la varianza de un factor para todas las variables,



tal como se muestra en la Tabla 4-5. En todos los casos el valor F es mayor y el valor P menor a 0.01, por lo que se rechaza la hipótesis nula ( $H_0$ ) de que cada variable tiene media similar. Además, se aplicó la prueba de Tukey de comparaciones múltiples. En la columna “Pares similares” se indican los pares de variables que tienen similares medias.

Tabla 4-5. Resultados de análisis de la varianza.

<b>Variable</b>	<b>F</b>	<b>P</b>	<b>Pares similares</b>
CuT [%]	117.33	5.98E-47	-
FeT [%]	157.28	7.37E-61	-
Mo [%]	176.05	3.99E-67	-
As [%]	139.64	8.12E-55	UGM2 – UGM3
Zn [%]	16.62	7.73E-08	UGM1 – UGM2
CuS [%]	90.20	5.05E-37	UGM2 – UGM3
SiO2 [%]	23.01	1.61E-10	UGM2 – UGM3
Al2O3 [%]	37.18	2.35E-16	-
Recuperación [%]	29.80	2.48E-13	UGM1 – UGM3
CuCo [%]	138.39	2.22E-54	-
Wi [kWh/t]	266.63	2.82E-95	-
SPI [kWh/t]	185.93	2.31E-70	-
Cc [%]	19.49	4.82E-09	UGM2 – UGM3
Dg [%]	30.00	2.05E-13	UGM2 – UGM3
Cv [%]	108.17	1.22E-43	-
Bn [%]	294.65	2.35E-103	-
En [%]	153.65	1.26E-59	UGM2 – UGM3
Cp [%]	829.19	6.12E-221	-
Py [%]	820.51	2.01E-219	-
Veta Qz [%]	265.98	4.39E-95	-
QS [%]	2498.69	0	-
SVG [%]	85.26	3.61E-35	UGM2 – UGM3
KSIL [%]	394.30	4.46E-130	-
PF [%]	1107.01	5.30E-265	UGM2 – UGM3
Clo [%]	15.14	3.27E-07	UGM2 – UGM3
PIR [%]	50.85	7.44E-22	-

Como se observa en la Tabla 4-5, analizando la varianza, aproximadamente la mitad de las variables tienen pares de unidad geometalúrgica similares. Pero también hay otra mitad de las variables que no tienen similitud entre unidades. Por lo tanto, las unidades geometalúrgicas son diferenciables entre ellas y el modelo geometalúrgico logra el objetivo.

#### **4.4. Evaluación económica**

Para validar el resultado se propone una evaluación económica determinando el efecto del modelo en la secuencia de extracción y el valor del yacimiento. A continuación, se detallan los pasos.

##### **4.4.1. Filtración**

Como se puede ver en los resultados de las Figuras 4-3 y 4-4, dentro de los límites de UGMs generadas existen datos de otras agrupaciones. Por lo tanto, se realiza un filtrado según la ubicación espacial, dado que las unidades geometalúrgicas debe tener continuidad espacial.

El filtro propuesto revisa los 7 datos más cercanos a un punto determinado según la distancia euclidiana. Si hay más de 6 datos pertenecientes a la misma unidad geometalúrgica del dato en revisión, se considera como clasificado correctamente. En el caso de que el dato revisado y la mayoría de los datos cercanos no sean clasificados en una misma unidad, se cambia el código de este dato al código

mayoritario. La consideración de solo 6 datos se hace para clasificar adecuadamente los datos del borde entre unidades. El flujograma decisional en la Figura 4-8.

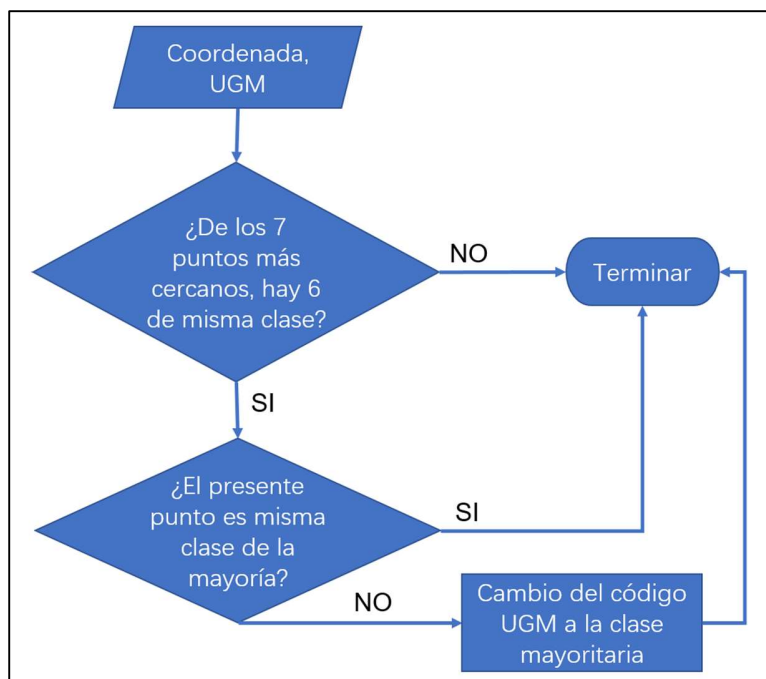


Figura 4-8. Flujograma de filtro 6/7.

#### 4.4.2. Modelo de bloque

Antes de realizar el cálculo de valor del yacimiento, se debe generar un modelo de bloques para poder realizar la planificación de la secuencia de extracción. Desde la base de datos obtenida, que es una base de sondaje, se proyecta un modelo de bloques, donde se asigna información de cada dato a cada bloque. Si en un bloque hay más de un dato, se le asigna el promedio de estos datos y aleatoriamente el código de uno de los datos. Cabe destacar que el tamaño del bloque es un cubo cuyas

aristas miden 45 metros. Además, dada la distribución espacial del yacimiento, se considera una topografía plana en la cota 2885.

#### 4.4.3. Interpolación

Con el modelo de bloque obtenido, se realizó una interpolación para definir el cuerpo mineral y realizar el cálculo de valor del yacimiento. Para eso se considera un método geométrico computacional nombrado *Alpha Shape* (Bellokk et al, 2018), el cual se realiza para llevar a cabo un casco convexo, es decir que busca generar bloques dentro de un área generada por los bloques del borde (Figura 4-9).

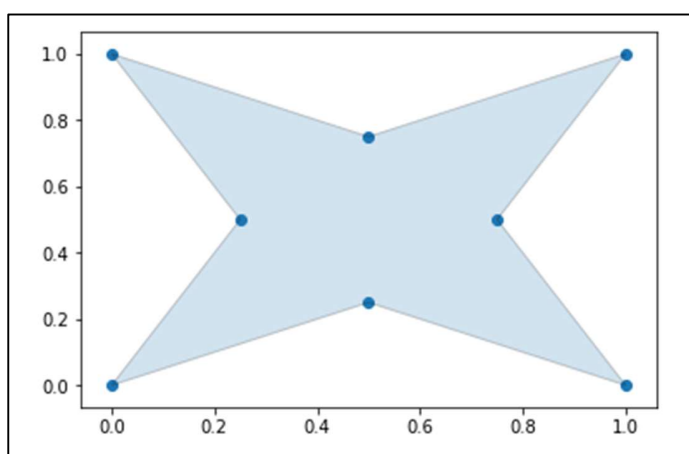


Figura 4-9. Ejemplo de un casco convexo. Adaptado de (Bellokk et al, 2018).

Se realizó *Alpha Shape* para la unidad geometalúrgica en cada plano en la cota que tiene al menos 3 bloques. El resultado se muestra en las Figuras 4-10 y 4-11. Cada punto es un bloque de cubo de 45 metros de aristas, donde los puntos más oscuros son datos de la base de datos y los puntos con tono más claro son puntos generados.

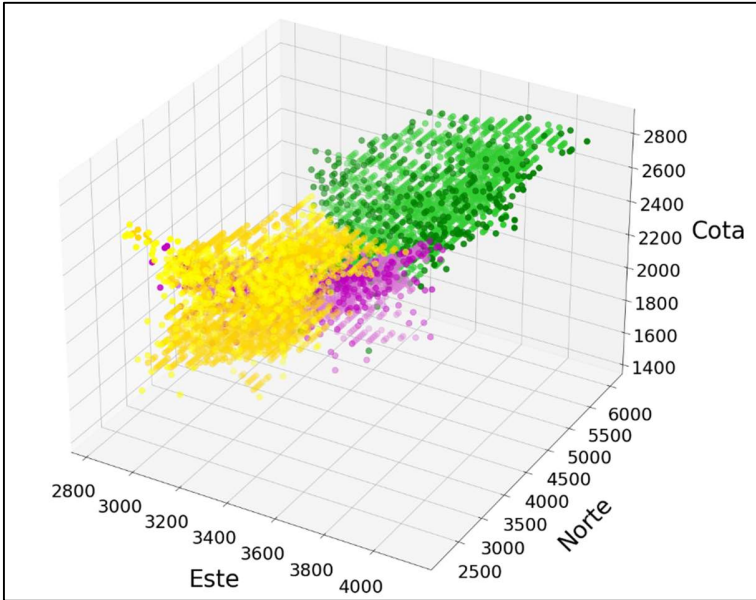


Figura 4-10. Visualización espacial de modelo geometalúrgico interpolado.

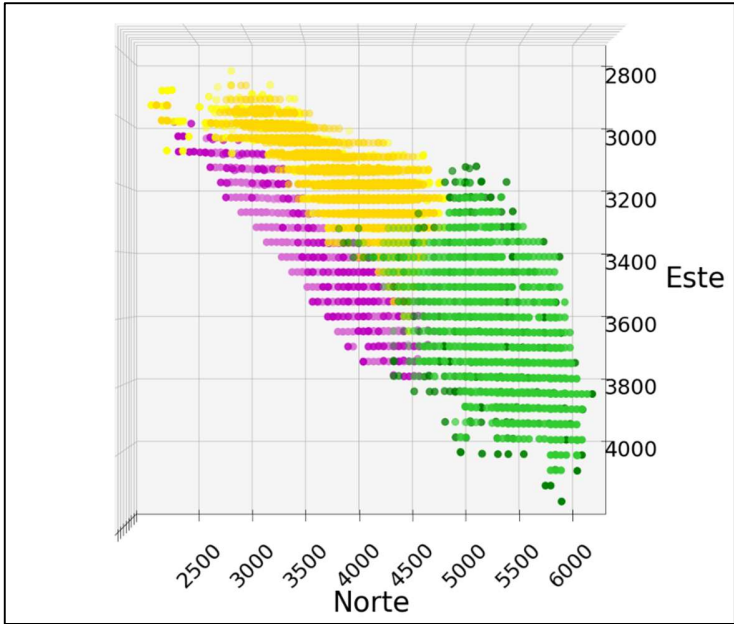


Figura 4-11. Visualización planta de modelo geometalúrgico interpolado.

Luego de tener los bloques del resultado de interpolación, se realiza una predicción con los datos generados. En este trabajo se realiza solo para las variables cobre total y recuperación metalúrgica, las cuales se van a necesitar para el cálculo del valor económico.

En primer lugar, se utilizan algoritmos de aprendizaje de máquinas y la configuración de hiperparámetros (Tabla 4-6). Se utilizó la base de datos del modelo geometalúrgico para entrenamiento, que consiste en una validación cruzada (*cross-validation*) asociada a una división de 5 series de la base de entrenamiento, donde se obtiene el resultado de error media cuadrática (MSE) entre valores reales y las estimaciones (Figura 4-12). Se selecciona el algoritmo con menor valor de MSE para realizar ajustes de los hiperparámetros más adelante. En este caso, el mejor rendimiento se dio con el algoritmo *Gradient boosting* para los 6 casos.

Tabla 4-6. Algoritmos utilizados para estimación de cobre total y recuperación metalúrgica.

Algoritmo de regresión	Hiperparámetros
k vecinos más próximos (n = 3)	n = 3
k vecinos más próximos (n = 5)	n = 5
k vecinos más próximos (n = 7)	n = 7
<i>Random Forest</i>	n_estimator = 100
Regresión lineal	fit_intercept = True
Máquinas de vectores de soporte	C = 1, kernel = ‘rbf’
Árbol de decisión	criterion = ‘squared_error’, max_depth = None
<i>Gradient boosting</i>	learning_rate = 0.1, n_estimator = 100
Redes neuronales con 2 capas ocultas	activation = ‘relu’

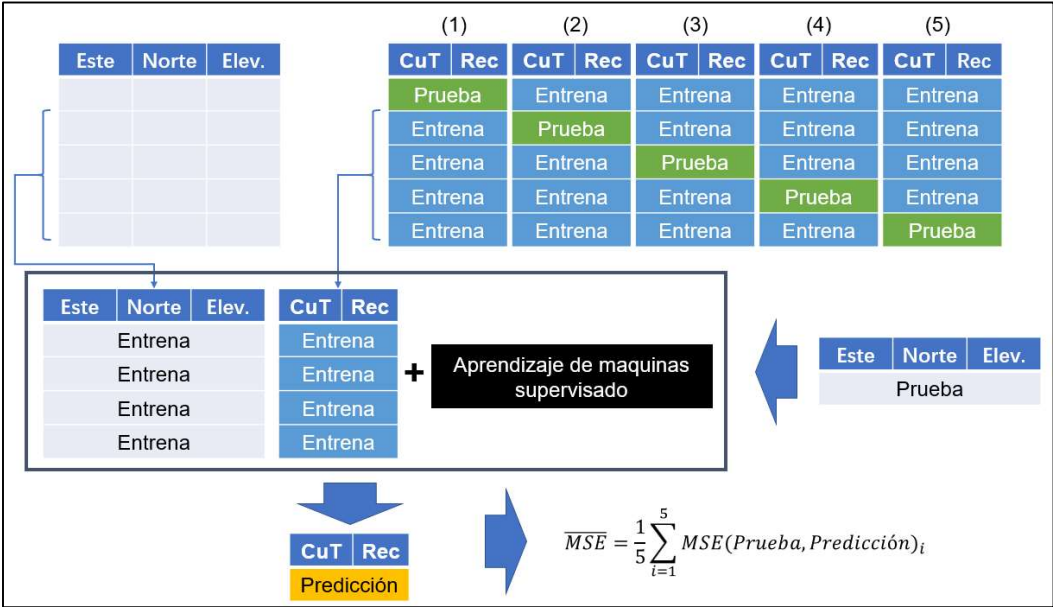


Figura 4-12. Proceso de selección de algoritmo supervisado.

*Gradient boosting* es una técnica de aprendizaje automático, la cual produce un modelo predictivo compuesto de un conjunto de submodelos de predicción débil. Los principales hiperparámetros son el número de estimadores y la tasa de aprendizaje. El número de estimadores es la cantidad máxima de iteración de los modelos débiles. La tasa de aprendizaje es el porcentaje de cambio con la actualización de los pesos en cada iteración. El ajuste de hiperparámetros se realiza en un conjunto de combinación de los dos hiperparámetros y define el modelo final según el menor valor de MSE obtenido en la validación cruzada. El resultado de ajuste de hiperparámetros se muestra en la Tabla 4-7.

Tabla 4.7. Hiperparámetros utilizados para la predicción.

UGM	Variable	Algoritmo elegido	Hiperparámetro	
			Número de estimadores	Tasa de aprendizaje
1	Cobre total (CuT)	<i>Gradient boosting</i>	500	0.01
	Recuperación metalúrgica	<i>Gradient boosting</i>	60	0.01
2	Cobre total (CuT)	<i>Gradient boosting</i>	230	0.01
	Recuperación metalúrgica	<i>Gradient boosting</i>	210	0.01
3	Cobre total (CuT)	<i>Gradient boosting</i>	700	0.01
	Recuperación metalúrgica	<i>Gradient boosting</i>	120	0.01

4.4.4. Modelo final y resultados económicos

Como se muestra en la Figura 4-10, el cuerpo mineral tiene una forma convexa, donde en el centro del plano este-norte del yacimiento existe una capa de estéril. Por lo tanto, la secuencia de extracción debe tener una orientación norte-sur ya que se explota menos estéril al inicio del período de extracción. Por ello, para efecto de la evaluación, se propone ajustar el yacimiento a un paralelepípedo como muestra en la Figura 4-13. El límite es entre 2800 y 3800 en el eje este, 2500 y 5200 en el eje norte y la cota entre 2000 y 2300. El resultado del corte se muestra en las Figuras 4-14 y 4-15.



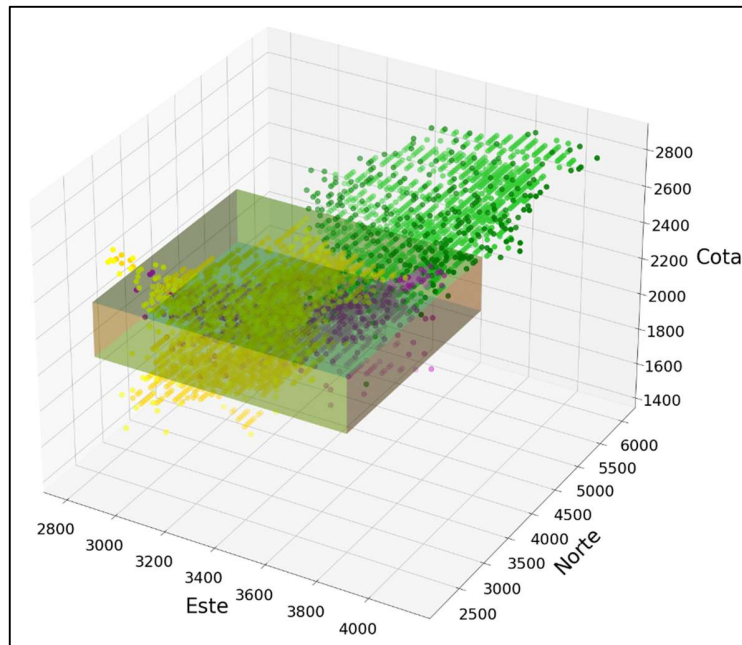


Figura 4-13. Visualización espacial límite de corte.

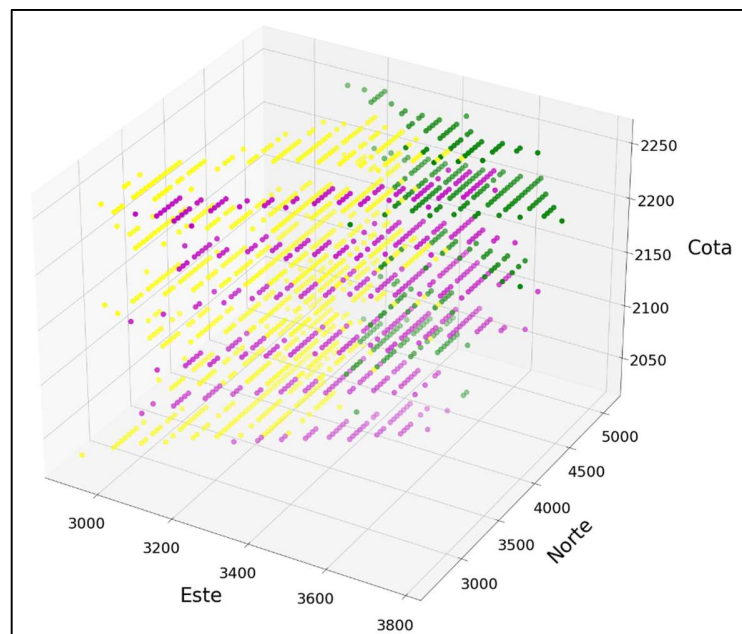


Figura 4-14. Visualización espacial modelo ajustado.

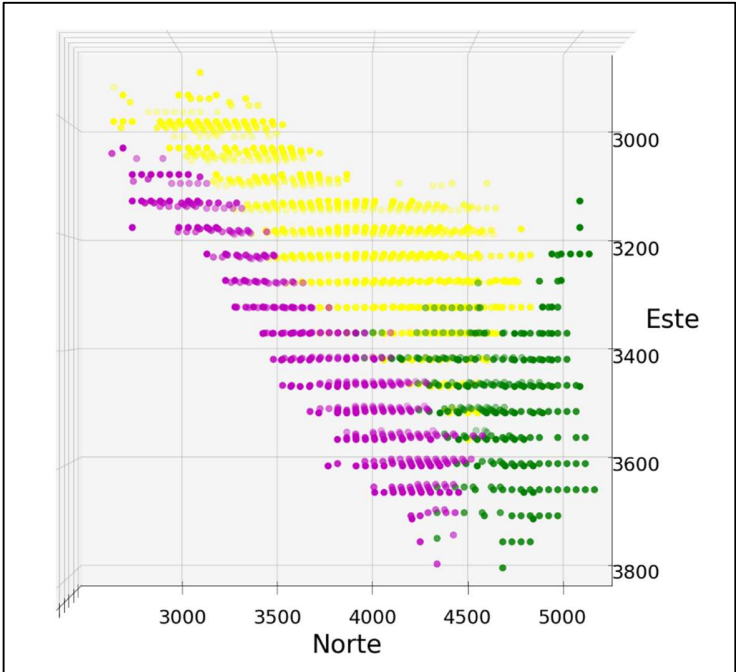


Figura 4-15. Visualización planta modelo ajustado.

Finalmente, con todos los datos procesados se realiza el cálculo de valor del yacimiento. Esto se lleva a cabo a través del modelo desarrollo por Anani et al. (2020) que busca determinar el valor de un proyecto minero mediante simulaciones. Para establecer el valor de un bloque del modelo, se utilizan los parámetros de las Tablas 4-8 y 4-9.

Tabla 4-8. Parámetros para evaluación económica.

Parámetro	Valor
Precio de cobre [US\$/lb]	3
Costo Mina [US\$/t]	3-3.95
Tasa de descuento [%]	10%
GAP [%]	0.4

El precio se considera razonable dentro del rango de variación en los últimos años. Se considera un costo mina base, el cual tiene un valor promedio de los informes técnicos de 3 yacimientos, y un costo incremental adicional debido al aumento de distancia de recorrido con la profundidad, lo que genera un mayor gasto energético. Este último se estima en 0.42 cUS\$/t mineral por cada metro de profundidad. Para la tasa de descuento se considera la más usual en la industria minera (Marinho & Nelson, 2017; Royo, 2018). El valor GAP es un índice del software de simulación *gurobi*, el cual indica la brecha entre la mejor solución conocida y la mejor solución límite. La determinación de este valor es una decisión propia de cada proyecto. Afum & Ben-Awuah (2019) indican un valor menor a 5% es tolerable, y Ben-Awuah et al. (2016) mencionan que para simulación en minería el GAP óptimo es menor a 1%.

Tabla 4-9. Costos planta para evaluación económica.

Escenario	Valor [US\$/t]
UGM 1	6.41
UGM 2	6.68
UGM 3	7.11
Caso base	6.67

Para determinar el costo planta promedio de cada unidad geometalúrgica, se utiliza la proporción del costo de energía dentro del costo planta en la industria. Utilizando un *benchmark* se obtiene que este valor tiene un promedio de 26% en la industria minera chilena. El costo de energía se estima a partir de los datos de *benchmark* de los informes técnicos recientes, y resulta en un costo de energía ( $C_{energía}$ ) 9.4

cUS/kWh (Tabla 4-10). Finalmente, se relaciona el costo planta ( $C_{planta}$ ) y el consumo de energía (E) en molienda (Tabla 4-11), por la ecuación 4.1, 4.2 y 4.3.

$$E \left[ \frac{kWh}{t \text{ mineral}} \right] * C_{energía} \left[ \frac{US\$}{kWh} \right] = 26\% * C_{planta} \left[ \frac{US\$}{t \text{ mineral}} \right] \quad (4.1)$$

$$E \left[ \frac{kWh}{t \text{ mineral}} \right] * 9.4 \left[ \frac{cUS\$}{kWh} \right] * \frac{1}{100} \left[ \frac{US\$}{cUS\$} \right] = 26\% * C_{planta} \left[ \frac{US\$}{t \text{ mineral}} \right] \quad (4.2)$$

$$C_{planta} \left[ \frac{US\$}{t \text{ mineral}} \right] = \frac{E \left[ \frac{kWh}{t \text{ mineral}} \right]}{26\%} * 9.4 \left[ \frac{cUS\$}{kWh} \right] * \frac{1}{100} \left[ \frac{US\$}{cUS\$} \right] \quad (4.3)$$

Tabla 4-10. Costo de energía.

	<b>Quebrada Blanca (2016)</b>	<b>Sierra Gorda (2009)</b>	<b>Santo Domingo (2018)</b>	<b>Promedio</b>
<b><i>Power cost</i> [cUS\$/kWh]</b>	11	10	7.2	9.4

Tabla 4-11. Costo planta por escenario para evaluación económica.

<b>UGM</b>	<b><math>\overline{Wt}</math> [kWh/t]</b>	<b>Consumo energético [kWh/t]</b>	<b>Costo de energía [US\$/t]</b>	<b>Costo planta [US\$/t]</b>	<b>Factor de producción</b>
1	13.31	17.73	1.67	6.41	1.09
2	14.22	18.48	1.74	6.68	0.97
3	15.66	19.66	1.85	7.11	0.88
Caso base	14.15	18.42	1.73	6.67	1

Para la evaluación económica, se plantea una capacidad de tratamiento de planta de 50,000 t/d para el caso base. Al aplicar el modelo geometalúrgico, la capacidad de producción se modifica por el factor de producción, asumiendo una potencia

instalada de 40.29 MW, calculada a través del índice de Bond promedio del yacimiento y la capacidad de tratamiento indicada anteriormente.

Se obtienen los valores del proyecto en distintos escenarios, los cuales se diferencian en los costos planta y la capacidad de alimentación según cada UGM, manteniendo el resto de las variables. El escenario base considera un costo planta constante para todo el yacimiento de 6.67 US\$/t, y se obtiene un VPN del proyecto igual a 3,746 millones de dólares. En cambio, el escenario considerando UGMs, tiene un costo planta según cada unidad geometalúrgica, obteniendo un VPN igual a 3,903 millones de dólares, que es alrededor del 4% mayor al caso base.

Tabla 4-12. VPN según escenario.

Escenario	VPN [Millones US\$]
Base	3,746
UGMs	3,903

## 5. CONCLUSIONES

### 5.1. Conclusiones

La investigación presentada busca establecer un modelo geometalúrgico con la inclusión de datos geológicos y metalúrgicos, mediante el uso de algunas técnicas de inteligencia artificial. El resultado final es generado con el uso de aprendizaje de máquinas.

De la revisión del estado del arte, se concluye que es viable generar un modelo geometalúrgico con el uso de algoritmos no supervisados, ya que existen estudios realizados en yacimientos de hierro y oro, donde se aplican tales técnicas. Los algoritmos más utilizados son *k-means*, agrupación jerárquica y mapa autoorganizado.

En este trabajo se evalúa desempeño de tres algoritmos mencionados de las literaturas, *k-means*, agrupación jerárquica y mapa autoorganizado, además se incorpora un algoritmo emergente en la última década, DBSCAN. Estos algoritmos se complementan con técnicas de reducción de dimensiones de los datos a modelar.

Se concluye que *k-means* tiene el mejor resultado según las métricas establecidas. El modelo obtenido identifica 3 unidades geometalúrgicas, con disimilitud

estadística según el análisis de varianza y la prueba de Tukey de comparaciones múltiples.

Se verifica que las variables espaciales pueden no ser relevantes para la generación de modelo geometalúrgico, obteniéndose unidades geometalúrgicas con continuidad especial excluyendo las variables espaciales. Esto indica que la variable espacial puede no ser un requisito específico para obtener un modelo geometalúrgico.

Las técnicas de reducción de dimensiones no son efectivas en mejorar la capacidad de generación de agrupación, pero puede ser útil para una base de datos con mayor cantidad de variables. Entre ellas, se destaca que el *autoencoder* tiene mejor desempeño que PCA con respecto a la pérdida de información. Pero *autoencoder* requiere una arquitectura única para cada grupo de variables.

Una evaluación económica aplicada a escenarios de explotación del yacimiento modelado, con y sin modelo geometalúrgico, indica que el modelo geometalúrgico permite una mejora del orden del 4% del valor del proyecto. Dentro del proceso de evaluación, la interpolación con *Alphashape* y la estimación de valores para los bloques interpolados usando *Gradient boosting* son adecuados para la generación del modelo de bloques.

En conclusión, se demuestra que es viable generar un modelo geometalúrgico en un yacimiento de cobre con un algoritmo de aprendizaje no supervisado, y su aplicación tiene impacto en el valor del proyecto. Por otro lado, el uso de herramientas de la ciencia de datos permite un empleo más eficiente de recursos en la planificación, facilita la actualización del modelo por mayor disponibilidad de datos o disponibilidad de nuevas variables geometalúrgicas.

## **5.2. Aplicabilidad**

La calidad del modelo geometalúrgico es directamente función de la calidad y cantidad de datos disponibles. Por lo tanto, dado las características del sector minero, donde existen economías de escala, podría haber una relación de la aplicabilidad con la escala de proyecto o faena.

Conceptualmente, la metodología desarrollada es factible de aplicar tanto en un yacimiento de sulfuros de cobre como de óxidos de cobre, aunque en el último caso es necesario ajustar las variables metalúrgicas.

El trabajo realizado utiliza la plataforma Python, con librerías abiertas y de libre acceso, permitiendo la implementación de algoritmos propios o particulares de los usuarios. Asimismo, no existe un requerimiento de hardware específico, por lo que es accesible para cualquier usuario o compañía minera.



### 5.3. Trabajo futuro

En primer lugar, en la presente investigación, el mejor resultado es obtenido con *k-means*. Sin embargo, con el desarrollo creciente en el mundo de la ciencia de datos, es esperable contar con mejores algoritmos para el modelamiento, especialmente en la agrupación de datos.

En segundo lugar, para tener una metodología más robusta, se requeriría aplicar la presente aproximación metodológica a bases de datos de distintos yacimientos. Esto conducirá a una metodología general que sea aplicable en cualquier yacimiento. También, la inclusión de mayor cantidad de variables de la cadena de valor del negocio minero, como por ejemplo las variables ambientales, sería de gran importancia, ya que amplía el alcance a elementos propios de la sostenibilidad del negocio.

En tercer lugar, el presente trabajo está enmarcado en las etapas tempranas de un proyecto minero, cuando los datos son escasos y no se cuenta con un modelo geometalúrgico previo. Por lo tanto, es importante definir metodologías para actualizar el modelo con datos provenientes de la explotación del yacimiento, para contar con mejor información en la toma de decisiones.

Finalmente, en el contexto de la Minería 4.0, una mejora adicional es complementar el modelo con capacidad de auto entrenamiento basado en aprendizaje profundo.

Esto permitiría mayor autonomía de la herramienta y mejor eficiencia en la actualización del modelo.

## BIBLIOGRAFIA

Anani, A., Li, H. & Ortiz, I.A. (2021). Conditions under which an integrated approach to the transition mine problem outperforms a disintegrated approach. APCOM 2021 online conference. South Africa: SAIMM.

Afum, B.O. & Ben-Awuah, E. (2019). MILP framework for open pit and underground mining transitions evaluation. *Mining Goes Digital* – Mueller et al. (Eds). Taylor & Francis Group, London. ISBN 978-0-367-33604-2

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J.M. & Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(2013)243-256, doi: 10.1016/j.patcog.2012.07.021.

Bellock, K., Godber, N. & Kahn, P. (2018). Alphashape, *github*, <https://github.com/bellockk/alphashape>.

Ben-Awuah, E., Richter, O. Elkington, T. & Pourrahimian, Y. (2016). Strategic mining options optimization: Open pit mining, underground mining or both. *International Journal of Mining Science and Technology*, 26(2016)1065-1071, doi: 10.1016/j.ijmst.2016.09.015.

Bhuiyan, M., Esmaili, K. & Ordóñez-Calderón, J.C. (2019). Application of Data Analytics Techniques to Establish Geometallurgical Relationships to Bond Work Index at the Paracutu Mine, Minas Gerais, Brazil. *Minerals*, doi: 10.3390/min9050302; 29p.

Boisvert, J.B., Rossi, M.E. & Ehrig, K. (2013). Geometallurgical Modeling at Olympic Dam Mine, South Australia. *Mathematical Geosciences*, doi: 10.1007/s11004-013-9462-5.

Calinsky, T. & Harabasz, J. (1974). A dendrite method for cluster analysis. *Commun. Statist.* 3(1): 1 – 27.

Canchaya, S. (2008). El modelo geometalúrgico. XIV Congreso Peruano de Geología – XIII Congreso Latinoamericano de Geología. Lima, Perú. Recuperado en: [https://www.sampling-ok.com/web/publicaciones/2013-01-31\\_WPR.pdf](https://www.sampling-ok.com/web/publicaciones/2013-01-31_WPR.pdf).

Chollet, F. (2017) Deep Learning with Python. *Manning* (1a ed.). Shelter Island, New York, EE.UU..

Curtis, J., David, B. & Neil, P. (2019). Hyperspectral Imaging Applications to Geometallurgy: Utilizing Blast Hole Mineralogy to Predict Au-Cu Recovery and Throughput at the Phoenix Mine, Nevada. *Economic Geology*, ISSN 0361-0128; doi:10.5382/econgeo.4684; 14 p.

Dubien, J.L. & Warde, W.D. (1979). A Mathematical Comparison of the Members of an Infinite Family of Agglomerative Clustering Algorithms. *The Canadian Journal of Statistics*. Vol. 7, No. 1, pp. 29-38. doi: 10.2307/3315012.

Dominy, S.C., O'Connor, L., Parbhakar-Fox, A., Glass, H. & Purevgerel, S. (2018). Geometallurgy—A Route to More Resilient Mine Operations. *Minerals*. 8, 560. doi:10.3390/min8120560.

Ellefmo, S., Aasly, K., Lang, A. Vezhapparambu, V. & Silva, C. (2019). Geometallurgical Concepts Used in Industrial Mineral Production. *Economic Geology*, ISSN 0361-0128; doi:10.5382/econgeo.4685; 12 p.

Ester, M., Kriegel, H.P., Sander, J. & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD-96, Volume 96, No.34*, pp.226-231. doi: 10.5555/3001460.3001507.

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A., Foufou S. & Bouras, A. (2014). A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis. *IEEE, Volume 2, No.3, September 2014*, 2166-6750.

Fernandes, F.G. & Leite, J.F.C. (2020). Mass and Metallurgical Balance Forecast for a Zinc Processing Plant Using Artificial Neural Networks. *Natural Resources Research*, Volume 29, Number 6. doi: 10.1007/s11053-020-09678-4.

Fouedjio, F. (2016). A hierarchical clustering method for multivariate geostatistical data. *Spatial Statistics*, 18 (2016) 333-351. doi: 10.1016/j.spasta.2016.07.003; 19 p.

García, S., Luengo, J. & Herrera, F. (2015). *Data Preprocessing in Data Mining*. (1a ed.). Switzerland: Springer International Publishing.

Geron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. EEUU: O'Reilly Media.

Guntoro, P.I., Ghorbani, Y., Butcher, A.R., Kuva, J. & Rosenkranz (2020). Textural Quantification and Classification of Drill Cores for Geometallurgy: Moving Toward 3D with X-ray Microcomputed Tomography (ICT). *Natural Resources Research*, Volume 29, Number 6. doi: 10.1007/s11053-020-09685-5.

Hunt, J.A. & Berry, R.F. (2017). Economic Geology Models 3. Geological Contributions to Geometallurgy: A Review. *Geoscience Canada*. Volume 44, Number 3. doi: 10.12789/geocanj.2017.44.121.

Hoal, K.O. (2008). Getting the geo into the geomet: Society of Economic Geologists Newsletter, no.73, p.1, 11-15.

Johnson, C.L., Browning, D.A. & Pendock, N.E. (2019). Hyperspectral Imaging Applications to Geometallurgy: Utilizing Blast Hole Mineralogy to Predict Au-Cu Recovery and Throughput at the Phoenix Mine, Nevada. *Economic Geology*, ISSN 0361-0128; doi:10.5382/econgeo.4684; 14 p.

Jubera, N., Joaquín, J., Louit, D., Barrientos, C., Carmona, S., Urzúa, O., Klein, P., Reyes, L. & Guzman, J.I. (2020). Hacia una minería 4.0 – Recomendaciones para impulsar una industria nacional inteligente. *Centro de estudios del cobre y la minería*. Recuperado en: <https://www.cesco.cl/wp-content/uploads/2020/06/Hacia-una-miner%C3%ADa-4.0.-Recomendaciones-para-impulsar-una-industria-nacional-inteligente-1-2.pdf>.

Koch, P. & Rosenkranz, J. (2020). Sequential decision-making in mining and processing based on geometallurgical inputs. *Minerals Engineering*, 149 (2020) 106262.

Koch, P., Lund, C. & Rosenkranz, J. (2019). Automated drill core mineralogical characterization method for texture classification and modal mineralogy estimation for geometallurgy. *Minerals Engineering*, 136 (2019) 99-109.

Kohonen, T. (1998). The self-organizing map, *Neurocomputing – Elsevier*, 21 (1998) 1-6.

Kramer, M.A. (1991). Nonlinear Principal Component Analysis Using Autoassociative Neural Network. *AIChE Journal*, 37(2):233-243. doi: 10.1002/aic.690370209.

Lishchuka, V., Koch, P., Ghorbani, Y. & Butcher, A. (2020). Towards integrated geometallurgical approach: Critical review of current practices and future trends. *Minerals Engineering*, 145 (2020) 106072.

Liu, Y., Li, Z., Xiong, H., Gao, X. & Wu, J. (2010). Understanding of Internal Clustering Validation Measures. *IEEE International Conference on Data Mining*.1550-4786/10. doi:10.119/ICDM.2010.35, 6 p.

Lamberg, P. (2011). Particles – the bridge between geology and metallurgy. Conference in Mineral Engineering, Luleå, Sweden, February 8-9, 2011, Proceedings, p. 1-16.

Mackay, D.J. (2003). An Example Inference Task: Clustering. *Information Theory, Inference and Learning Algorithms* (pp.284-292). Cambridge, UK. ISBN: 9780521642989.

Maite, J., Bouchard, K. & Bédard, L.P. (2019). Mineral grains recognition using computer vision and machine learning. *Computers and Geosciences*, 130 (2019) 84-93. doi: 10.1016/j.cageo.2019.05.009; 10 p.

Manyika, J., Ramaswamy, S., Khanna, S., Sarrazin, H., Pinkus, G., Sethupathy, G. & Yaffe, A. (2015). Digital America: a tale of the haves and have-mores. *Mckinsey Global Institute*. Recuperado en: <https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/digital-america-a-tale-of-the-haves-and-have-mores>.

Marinho, R. & Nelson, M. (2017). NI 43-101 Technical Report on Quebrada Blanca Phase 2. Tarapacá, Chile.

McQuiston, F.W., Jr. & Bechaud, L.J., Jr. (1968). Metallurgical sampling and testing, E.P., ed., Surface mining: New York, American Institute of Mining, Metallurgical, and Petroleum Engineers, P. 103-121.

Mena, C., Ellefmo, S., Sandøy, R., Sørensen, B. & Aasly, K. (2020). A neural network approach for spatial variation assessment – A nepheline syenite case study. *Minerals Engineering*, 149 (2020) 106178.

Michaux, S. & O'Conner, L. (2020). How to Set Up and Develop a Geometallurgical Program. *Geological Survey of Finland*. ISBN 978-952-217-409-3.

Müller, D. (2011). Modern hierarchical, agglomerative clustering algorithms, Stanford University, EE.UU.. arXiv:1109.2378.

Müller, A. & Guido, S. (2016) Introduction to Machine Learning with Python: A Guide for Data Scientists. *O'Reilly Media* (1a ed.). California, EE.UU..

Ordóñez-Calderón, J.C. & Gelcich, S. (2018). Machine learning strategies for classification and prediction of alteration facies: Examples from the Rosemont Cu-Mo-Ag skarn deposit, SE Tucson Arizona. *Journal of Geochemical Exploration*, 194 (2018) 167-188. doi: 10.1016/j.gexplo.2018.07.020; 22 p.

Ortiz, J.M. & Kracht W. (2020). Bia Data en geometalurgia. *Big Data en Minería* (pp.103-125). Santiago, CHILE.

Rajabinasab, B. & Asghari, O. (2019). Geometallurgical Domaining by Cluster Analysis: Iron Ore Deposit Case Study. *Natural Resources Research*, Vol. 28, No. 3, July 2019. <https://doi.org/10.1007/s11053-018-9411-6>.

Rayo, J. (2018). Estudio Factibilidad Rajo Inca (Proyecto Rajo Inca). Codelco, Chile.

Rincon, J., Gaydardzhiev, S. & Stamenov, L. (2019). Coupling comminution indices and mineralogical features as an approach to a geometallurgical characterization of a copper ore. *Minerals Engineering*, 130 (2019) 57-66.

Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20 (1987) 53-65.

Schubert, E., Sander, J., Ester, M., Kriegel, H.P. & Xu, X. (2017). DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *ACM Transactions on Database System*, Vol. 42, Issue 3. doi: /10.1145/3068335.

Sepulveda, E., Dowd, P.A., Xu, C. & Addo, E. (2017). Multivariate Modelling of Geometallurgical Variables by Projection Pursuit, *International Association for Mathematical Geosciences*, 49:121–143. doi: 10.1007/s11004-016-9660-z.

Sepulveda, E., Dowd, P.A. & Xu, C. (2018). Fuzzy Clustering with Spatial Correction and Its Application to Geometallurgical Domaining, *International Association for Mathematical Geosciences*, 50:895–925. doi: 10.1007/s11004-018-9751-0.

Suazo, C.J., Kracht, W. & Alruiz, O.M. (2010). Geometallurgical modelling of the Collahuasi flotation circuit, *Minerals Engineering*, 23 (2010) 137-142.

Vesanto, J. Himberg, J. Alhoniemi, E. & Parkhankangas, J. *SOM Toolbox for Matlab 5*, Report A57 2000; Neural Networks Research Centre, Helsinki University of Technology: Helsinki, Finland, 2000.