

Machine Learning aplicado a la exploración minera usando matriz de confusión

Machine Learning applied to mining exploration using confusion matrix

Marco A. Cotrina-Teatino^{1,*} , Alvaro I. Riquelme Sandoval² ; Jose A. Guartan Medina³ ; Jairo J. Marquina Araujo¹ 

¹ Facultad de Ingeniería, Universidad Nacional de Trujillo, Av. Juan Pablo II s/n – Ciudad Universitaria, Trujillo, Perú.

² Queen's University, Robert M. Buchan Department of Mining, Kingston, Ontario, Canada.

³ Departamento de Ingeniería Geológica y Minas. Universidad Católica de Loja, Loja, Ecuador.

* Autor correspondiente: mcotrinat@unitru.edu.pe (M. Cotrina-Teatino)

DOI: [10.17268/rev.cyt.2025.01.06](https://doi.org/10.17268/rev.cyt.2025.01.06)

RESUMEN

El objetivo del presente trabajo fue determinar cómo validar espacialmente conjuntos de datos geológicos obtenidos de la zona de Huancayo en el Perú, así como identificar las formas más comunes de evaluar el rendimiento de los modelos de clasificación. La metodología empleada consistió en el uso de algoritmos de aprendizaje automático aplicados a una matriz de confusión, cuyos resultados fueron representados mediante una curva ROC. Los resultados mostraron que el parámetro de exactitud (Accuracy) alcanzó un porcentaje máximo de 65.4%, la precisión fue del 80.4%, la recuperación del 96.2% y la métrica F1 obtuvo un valor de 96.1%. Se concluye que, al utilizar modelos de clasificación, es fundamental considerar métricas adicionales a la precisión para obtener una estimación más confiable.

Palabras claves: Exactitud; Precisión; Recuperación; Clasificación; Aprendizaje automático.

ABSTRACT

The objective of this study was to determine how to spatially validate geological datasets obtained from the Huancayo region in Peru, as well as to identify the most common methods for evaluating the performance of classification models. The methodology employed involved the use of machine learning algorithms applied to a confusion matrix, with results represented through a ROC curve. The findings indicated that the accuracy parameter reached a maximum percentage of 65.4%, precision was 80.4%, recall was 96.2%, and the F1-score achieved a value of 96.1%. It is concluded that when using classification models, it is essential to consider additional metrics beyond accuracy to obtain a more reliable estimation.

Keywords: Accuracy; Precision; Recall; Classification; Machine learning.

1. INTRODUCCIÓN

La zona de estudio donde tiene lugar esta investigación es en Huancayo en ese lugar se realizó un estudio de suelos donde se muestra que zonas contienen mineralización esa es la base de datos que se usará para realizar la predicción de estimación (Zuo & Carranza, 2023). Para ampliar el grado de precisión de los datos se delimitará la zona en 6 partes con la finalidad de analizar por separado cada zona mediante la librería pipeline que es una herramienta muy útil para la reproducibilidad y la prevención de la fuga de datos (Parsa et al., 2022). Al hacer uso de la matriz de confusión se pueden observar varios escenarios que permitieron tener una visión amplia acerca de la precisión de la estimación de los datos y con la finalidad de tener un análisis visual las curvas de ROC muestran el nivel de incertidumbre en el que está sometido cada campo (Sustaita, 2017). En el ámbito tecnológico, inteligencia artificial (IA) alude a la simulación de cognición humana ejecutada por sistemas computacionales, robots o dispositivos similares (Arango, 2021). Comúnmente, se entiende por IA la habilidad de un sistema para replicar habilidades humanas: aprender de situaciones, reconocer elementos, comprender e interactuar con lenguaje, decidir y solucionar problemáticas; y fusionar estas destrezas para ejecutar acciones humanas, como recibir a un visitante o manejar un vehículo (Pérez, 2014). Si bien en el pasado la IA estuvo circunscrita a la literatura futurista, en la actualidad es una realidad integrada en nuestra rutina diaria (Gutierrez & Vigo, 2021). El avance vertiginoso en IA ha sido impulsado por el acceso a voluminosos datos y el surgimiento de potentes sistemas capaces de gestionarlos con mayor eficiencia que el ser humano (Correa, 2010). El deep learning, en particular, puede emplear datos etiquetados, denominado



aprendizaje supervisado, para nutrir su algoritmo. Sin embargo, tiene la capacidad de procesar datos en su estado natural (como textos o gráficos) y discernir de manera autónoma las características esenciales que definen categorías de datos. A diferencia del aprendizaje automático, no precisa de mediación humana para el procesamiento de datos, permitiendo su expansión en modalidades más innovadoras (Julca, 2020). El aprendizaje automático es la rama de la informática que utiliza la experiencia pasada para aprender y utilizar sus conocimientos para tomar decisiones futuras. El aprendizaje automático está en la intersección de la informática, la ingeniería y la estadística. El objetivo del aprendizaje automático es para generalizar un patrón detectable o para crear una regla desconocida a partir de ejemplos dados (López, 2020).

2. MATERIALES Y MÉTODOS

2.1 Materiales

La base fundamental de esta investigación radicó en la implementación de algoritmos de machine learning, diseñados específicamente para trabajar con una matriz de confusión. Esta herramienta, reconocida en el ámbito de la inteligencia artificial, facilitó la evaluación del desempeño de algoritmos supervisados, presentando visualmente el grado de precisión y error en sus predicciones (Castillo, 2022). Además, se utilizó la curva ROC, una representación gráfica que mostró el rendimiento discriminativo de un modelo clasificador relacionando los verdaderos positivos con los falsos positivos a lo largo de diferentes umbrales de decisión (Cabanillas, 2022). Para facilitar este análisis y la manipulación de datos, se recurrió a Python, un lenguaje de programación versátil. En este contexto, las librerías Scikit-Learn, para herramientas de aprendizaje automático; NumPy, para operaciones matemáticas; y Pandas, para el manejo de datos, fueron esenciales (Jovic, 2019).

2.2 Procedimiento

Para construir y fundamentar este estudio, se accedió a una robusta base de datos proporcionada por el proyecto GE36b-5 (POI 2018), titulado “Geoquímica multipropósito de suelos en el área Huancayo - Acostambo”, el cual constituyó una fuente informativa con especificidades de 170 muestras de la capa superficial del suelo y 53 del subsuelo, situadas en las provincias de Huancavelica y Junín (véase Figura 1). Más específicamente, este repositorio incluyó concentraciones de 58 compuestos químicos y 11 óxidos predominantes. Esto abarcó desde metales preciosos como el oro y la plata hasta otros elementos de interés geológico como el cobre, el plomo, el zinc, entre otros. La riqueza y profundidad de esta base de datos (Joakin, 2021) sirvieron para una variedad de propósitos, desde la prospección minera y estudios geoambientales hasta el ordenamiento territorial (Li et al., 2023).

2.3 Método

Con el objetivo de realizar una estimación precisa y coherente sobre la base de datos, se recurrió a una herramienta especializada de Scikit-Learn, el Group Shuffle Split (GSS). Este método permitió dividir el conjunto de datos en función de una variable de grupo determinada (Hanuman, 2018). En el marco de este estudio, se creó una característica de grupo, denominada “Zonas”, que se basa en áreas geológicas específicas de la región de Huancayo, reflejadas en mapas geológicos proporcionados por INGEMMET. Mediante un enfoque sistemático, cada área geológica o de interés fue etiquetada con un número de zona correspondiente. Con esta organización en mente, se llevó a cabo la segmentación del conjunto de datos usando GSS, garantizando que los datos se agruparan y analizaran de acuerdo con estas zonas geológicas específicas (Ezequiel, 2019). Este enfoque meticuloso garantizó una comprensión más profunda y geográficamente relevante de la información (Chen et al., 2023) (Wang, 2022).

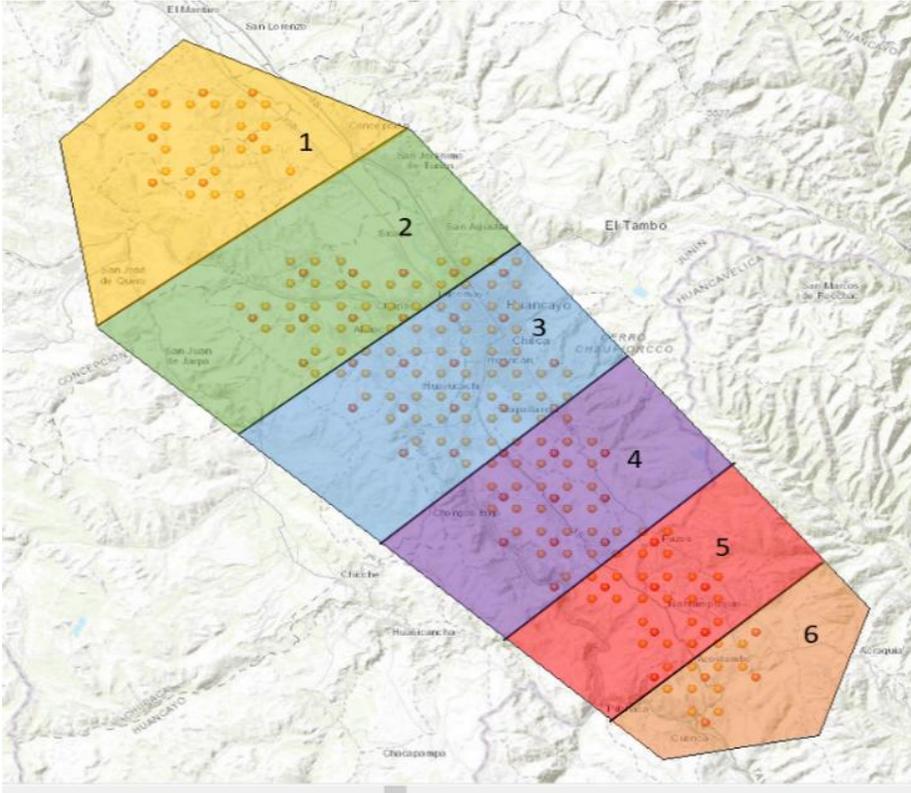


Figura 1. Delimitación del proyecto GE36b-5 (POI 2018) Huancayo - Acostambo

3. RESULTADOS Y DISCUSIÓN

3.1 Validación cruzada espacial

El área de estudio de este informe fue la zona de Huancayo, y la finalidad fue utilizar el aprendizaje automático para predecir áreas con posible mineralización de metales base. La base de datos consistió en una combinación de características espaciales geofísicas y geológicas obtenidas de INGEMMET. (Chen et al., 2023) definieron esta técnica como un método utilizado en el análisis espacial para evaluar la precisión de los modelos de interpolación y predicción. La técnica se basó en la división del conjunto de datos en subconjuntos de entrenamiento y prueba, donde los datos de prueba se utilizaron para evaluar la precisión del modelo. La variable objetivo se definió como un “Área de Interés” (AOI), y cada área conocida contenía el valor 0 para estéril o 1 para mineralizado. La mineralización, en este caso, se refirió a concentraciones elevadas de metales base, incluyendo Oro (Au), Cobre (Cu), Plomo (Pb), Zinc (Zn) y/o Plata (Ag).

Tabla 1. Resumen de la base de datos final

Áreas de interés	Log_Distance_T o_Faults	Log_Distance_to_Archae an_Contact	Log_Distance_to_Archae an_Contact	Log_TMI_Merged_Slope	Zonas
1	8.562915	8.344032	8.344032	1.923039	1
1	8.596839	9.729106	9.729106	2.304054	1
1	8.526839	9.659106	9.659106	2.234054	1
1	8.635839	9.768106	9.768106	2.343054	1
1	8.620839	9.753106	9.753106	2.328054	1
1	8.419839	9.552106	9.552106	2.127054	1
1	8.521839	9.654106	9.654106	2.229054	1
1	8.634839	9.767106	9.767106	2.342054	1
1	8.401839	9.534106	9.534106	2.109054	1
0	8.364839	9.497106	9.497106	2.072054	2
0	8.364839	9.497106	9.497106	2.072054	2
0	8.409839	9.542106	9.542106	2.117054	2
0	8.516839	9.649106	9.649106	2.224054	2

Áreas de interés	Log_Distance_To_Faults	Log_Distance_to_Archaean_Contact	Log_Distance_to_Archaean_Contact	Log_TMI_Merged_Slope	Zonas
0	8.533839	9.666106	9.666106	2.241054	2
0	8.414839	9.547106	9.547106	2.122054	2
0	8.418839	9.551106	9.551106	2.126054	2
0	8.371839	9.504106	9.504106	2.079054	2
0	8.489839	9.622106	9.622106	2.197054	2
0	9.613204	7.745506	7.745506	0.558646	2

Para superar los efectos de la autocorrelación espacial, se analizó la división de datos de entrenamiento y prueba en zonas espaciadas de manera separada. Se empleó el método Group Shuffle Split (GSS) de Scikit-Learn, el cual dividió el conjunto de datos en función de una variable de agrupación predefinida (Chen et al., 2023). Durante la construcción del conjunto de datos de características, se generó una variable de agrupación denominada “Zonas”.

3.2 Histograma de visualización de zonas

Uno de los desafíos identificados fue la variabilidad en el número de muestras por zona, lo que impidió mantener una proporción constante de datos en los conjuntos de validación cruzada (CV). Para visualizar la distribución de los datos en cada zona, se trazó un histograma. En la Figura 2, se observa una distribución de zonas mineralizadas, donde la zona con mayor cantidad de muestras fue la zona 6, mientras que la zona con menor mineralización correspondió a la zona 4. (Ezequiel, 2019) empleó un enfoque similar al aplicar Deep Learning para visualizar la distribución de zonas de interés, asignando valores de (1) y (0), técnica que se implementó en este estudio.

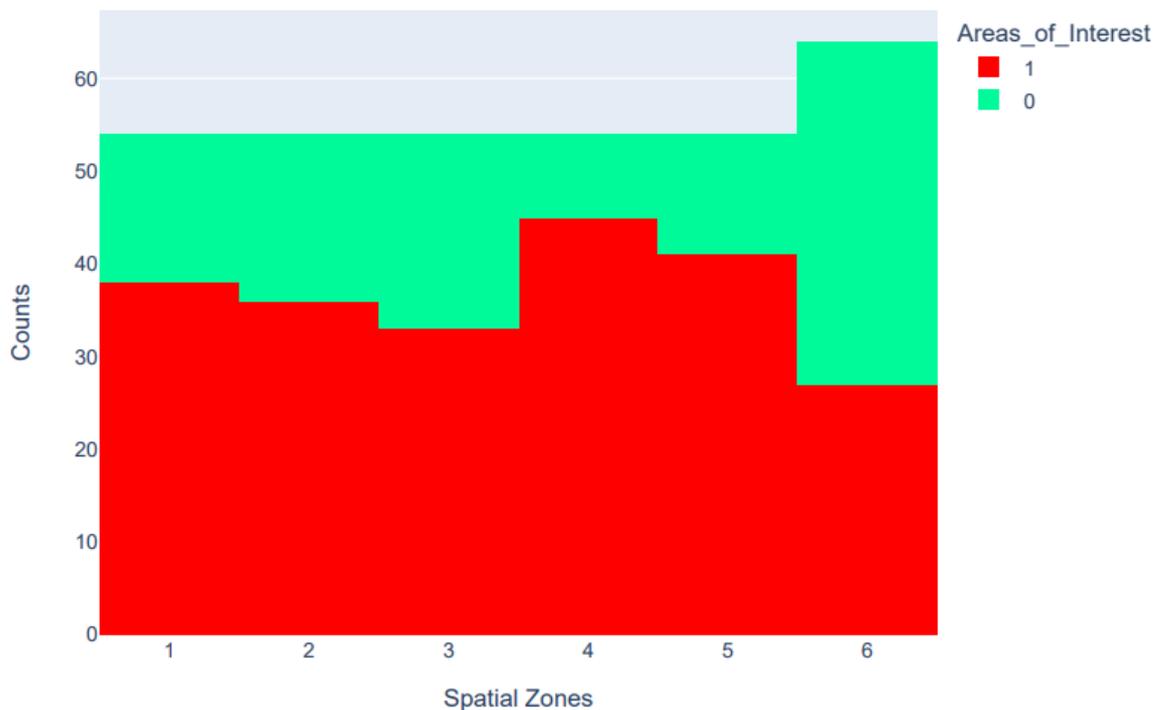


Figura 2. Histograma de visualización de zonas

3.3. División aleatoria de grupos

El uso del método GSS permitió dividir los datos en tres conjuntos distintos de validación cruzada, manteniendo una proporción de prueba lo más cercana posible al 70%-30%, tal como lo reportó (Cabanillas, 2022), quien logró una proporción de 80%-20%. En la Figura 3, se muestra la agrupación aleatoria de datos en tres conjuntos. Se observó que los conjuntos de prueba tuvieron una mayor ocurrencia, mientras que el conjunto de entrenamiento se concentró en las zonas 2 y 6.

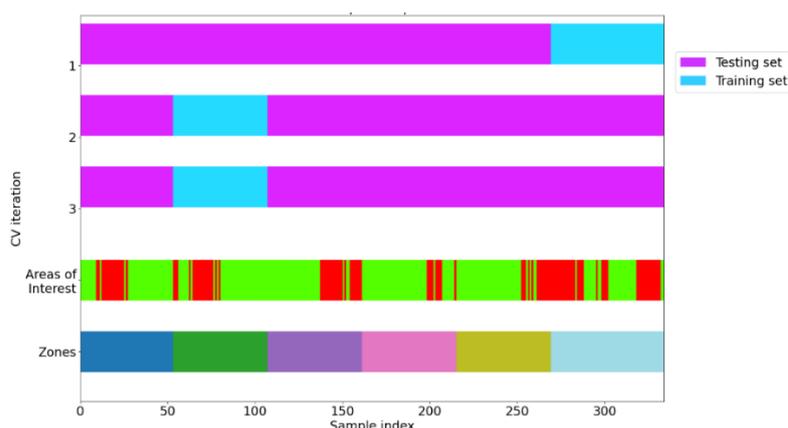


Figura 3. División aleatoria de grupos

En configuraciones alternativas de validación cruzada, se calculó el promedio de cada métrica CV para evaluar la solidez del modelo. Dado que los conjuntos de prueba y entrenamiento fueron espacialmente distintos y presentaron una variabilidad significativa en el tamaño de los datos, resultó fundamental analizar el comportamiento de cada conjunto CV de manera individual. En la Tabla 2, se presentan las zonas asignadas como datos de entrenamiento y prueba.

Tabla 2. Asignación de zonas para datos de entrenamiento y prueba en el análisis de validación cruzada

CV	Datos de entrenamiento	Datos de prueba
1	Zona 6	1, 2, 3, 4, 5, 6
2	Zona 2	1, 2, 3, 4, 5, 6
3	Zona 2	1, 2, 3, 4, 5, 6

3.4 Pipelines

Un pipeline se implementó como una herramienta para estructurar un flujo de procesamiento de datos en múltiples pasos. Cada paso de la canalización se aplicó en un orden predefinido para asegurar la coherencia en el análisis de los datos. Además de garantizar la reproducibilidad y consistencia en los flujos de trabajo, el uso de un pipeline permitió reducir la fuga de datos entre los conjuntos de entrenamiento y prueba (Hanuman, 2018). Esto se debió a que la canalización ajustó primero el escalar a los datos del conjunto de prueba después de establecer el conjunto de entrenamiento. Si todas las características se hubieran definido antes de separar los conjuntos de entrenamiento y prueba, la información del conjunto de prueba habría influido en el modelo, comprometiendo su validez.

3.5. Métricas de evaluación

Cuando se creó el modelo de clasificación de aprendizaje automático, se evaluó su rendimiento para determinar si predijo correctamente las clases. En particular, se analizó hasta qué punto el clasificador de Máquina de Vectores de Soporte identificó correctamente las zonas estériles y mineralizadas a partir de la base de datos (Cabanillas, 2022).

3.5.1. Exactitud

La exactitud de clasificación es una de las métricas más utilizadas para evaluar el rendimiento de un modelo. En este estudio, se calculó la precisión utilizando el procedimiento de validación cruzada representado en la Figura 4, que siguió tres pasos principales: División de datos: los datos se separaron en conjuntos de entrenamiento y prueba. Entrenamiento del modelo: el conjunto de entrenamiento se utilizó para ajustar el modelo, permitiéndole aprender a partir de los datos y ajustar sus parámetros. Evaluación del modelo: el modelo entrenado se utilizó para hacer predicciones sobre los datos de prueba. Estas predicciones se compararon con las etiquetas reales para evaluar la precisión del modelo.

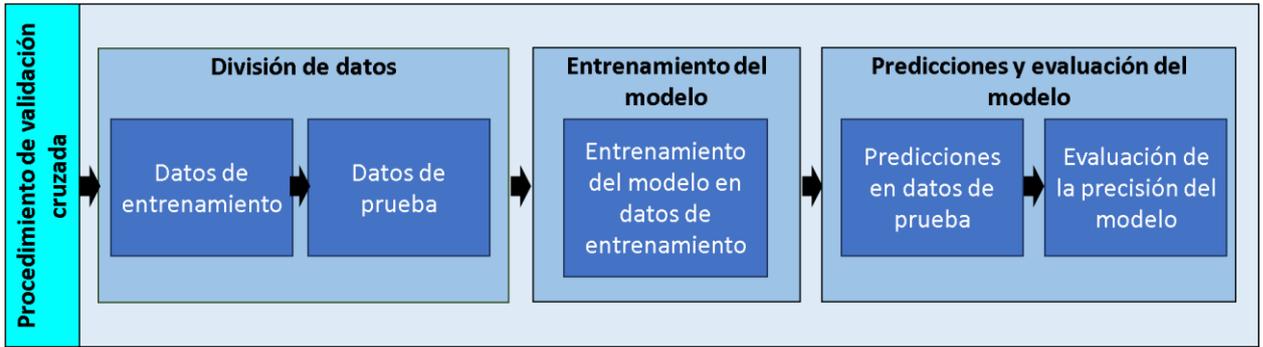


Figura 4. Diagrama del procedimiento de validación cruzada para evaluación de modelos de aprendizaje automático

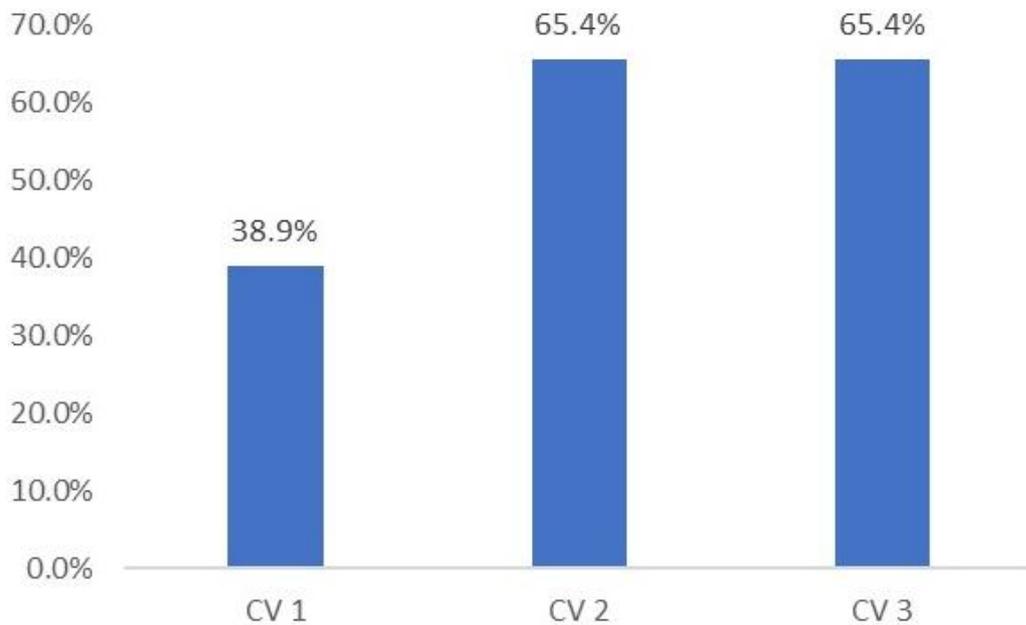


Figura 5 Resultados de precisión de clasificación en validación cruzada

Los resultados mostraron una gran variación en la exactitud de las diferentes divisiones de validación cruzada (CV). Esto se atribuyó a diferencias espaciales en los datos o a la variabilidad en el tamaño de los conjuntos de entrenamiento y prueba. Sin embargo, se llevó a cabo un análisis detallado de la representación de la precisión y su método de cálculo. (Julca, 2020) definió la exactitud como la proporción de datos clasificados correctamente respecto al total de datos. Para una mejor comprensión, se empleó la matriz de confusión.

3.5.2. Matriz de confusión

La matriz de confusión es una herramienta que permite visualizar el desempeño de un modelo al predecir cada clase dentro del conjunto de datos de prueba (Gutierrez & Vigo, 2021). En este estudio, se realizó una clasificación binaria para determinar si una zona era estéril o mineralizada, utilizando una matriz de confusión de 2x2. Idealmente, el modelo debería haber concentrado la mayor cantidad de puntos de datos en las celdas de Verdadero Negativo (TN) y Verdadero Positivo (TP), lo que indicaría que las predicciones fueron exitosas. (Cabanillas, 2022) señaló que, al evaluar la calidad de los modelos, la cantidad de predicciones correctas se califica con un valor de 1, mientras que las incorrectas se asignan un valor de 0. En su estudio, reportó que la predicción total fue correcta, y que la exactitud estuvo relacionada con el sesgo de la estimación. La ecuación utilizada para calcular la precisión en la matriz de confusión fue:

$$\text{Accuracy} = 2(\text{TN} + \text{TP}) / \text{TN} + \text{FN} + \text{TP}$$

Donde;

TN es una zona estéril clasificada como zona estéril.

FP es una zona estéril clasificada como zona mineralizada.

FN es una zona mineralizada clasificada como zona estéril.

TP es una zona mineralizada clasificada como zona mineralizada.

Teniendo esto en cuenta, se realizó la matriz de confusión para cada conjunto de validación cruzada.

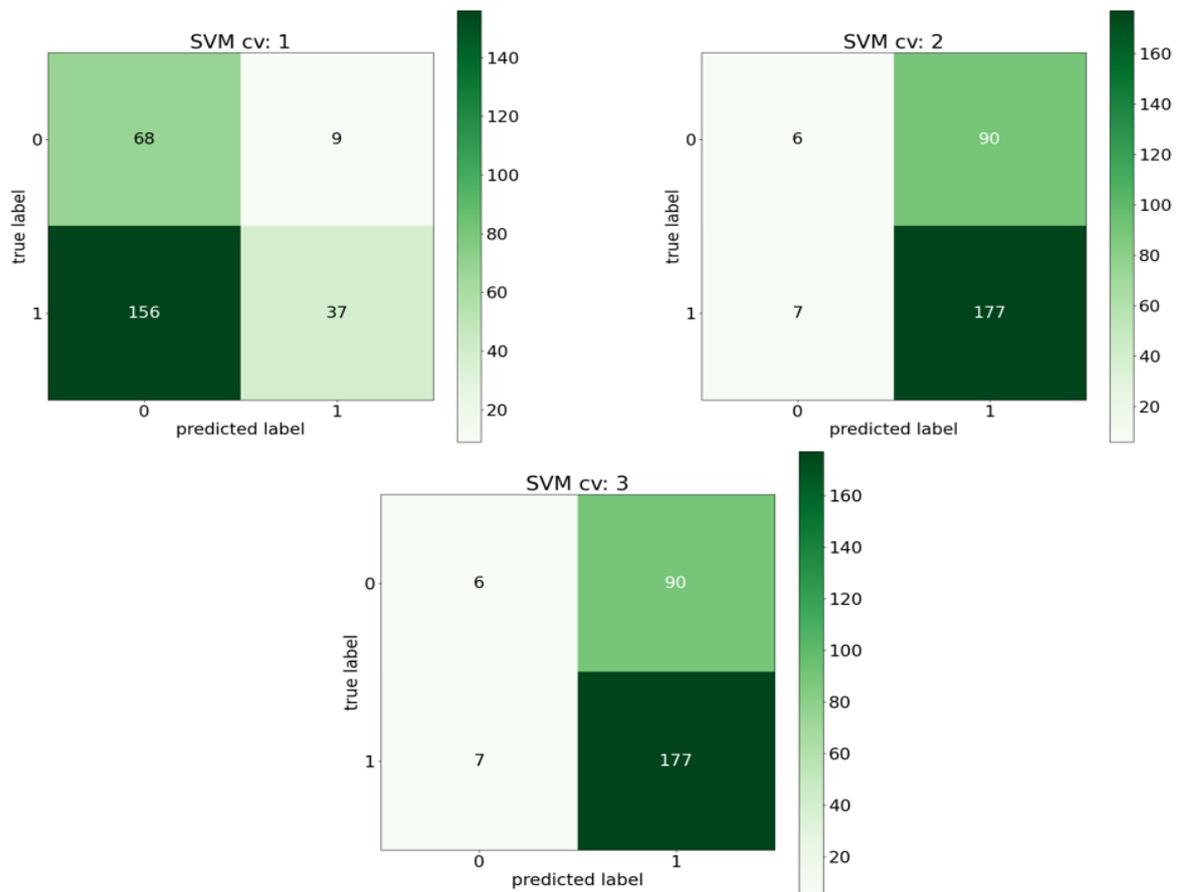


Figura 6. Matriz de confusión

En el cálculo de la exactitud de cada conjunto (Gutierrez & Vigo, 2021) indicó que en su investigación la mayoría de sus casos acertados estuvieron por encima del 97% de exactitud, en este caso el primer conjunto de CV arrojó una exactitud de 38.9%, al observar la matriz de confusión se entiende que esto es cierto puesto existe una gran cantidad de errores.

3.5.3. Precisión

Como métrica, la precisión es más efectiva cuando cada clase contiene un número similar de puntos de datos, pero se vuelve poco confiable cuando existe una gran diferencia en la cantidad de datos entre clases (como se observó en CV:1) (Cabanillas, 2022). Para evaluar el rendimiento del modelo en esta situación, se recurrió a diferentes métricas de evaluación, incluida la precisión. En la Figura 8, se observó que, aunque CV:1 presentó una exactitud razonablemente alta, obtuvo la puntuación de precisión más baja.

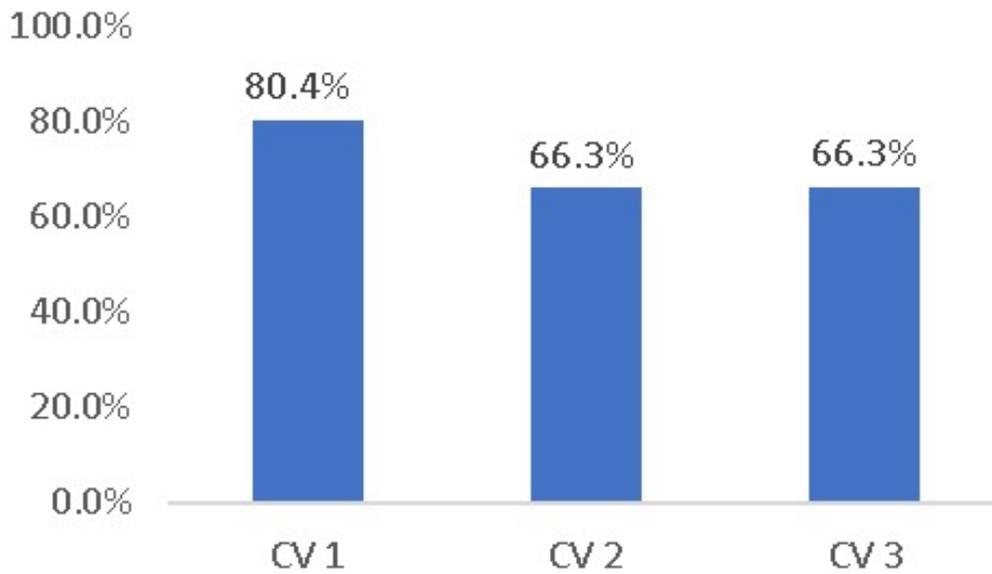


Figura 7. Precisión de la estimación

3.5.4. Recuperación

La precisión indica la cantidad de zonas mineralizadas correctamente clasificadas con respecto a la cantidad total de zonas mineralizadas reales. La recuperación fue un factor importante, aunque en este caso no resultó tan crucial como la precisión, ya que se priorizó clasificar las zonas mineralizadas como estériles antes que clasificar las zonas estériles como mineralizadas. Se obtuvieron porcentajes relativamente altos en términos de precisión, tal como lo reportó (Cabanillas, 2022)

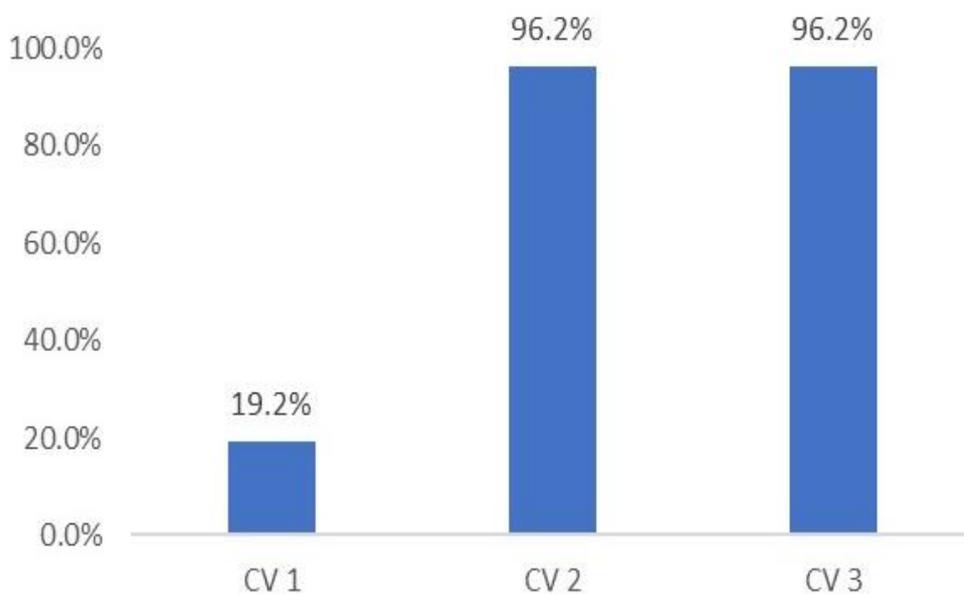


Figura 8. Recuperación de mineral

3.5.5. F1 Score

La puntuación F1 utiliza la media armónica de precisión y recall en lugar de la media aritmética, ya que, si uno de los valores es demasiado bajo, esto afecta significativamente la puntuación final (Sustaita, 2017). Se calculó la puntuación F1 para cada CV en Python, como se muestra en la Figura 9.

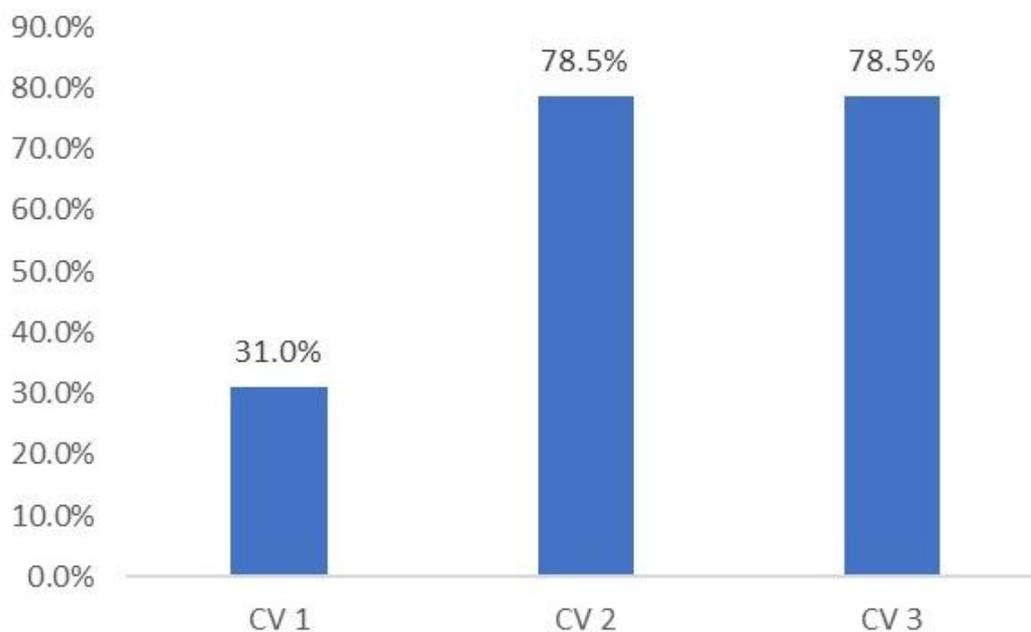


Figura 9. F1 Score

3.6. Informe de clasificación

El informe de clasificación proporcionó todas las métricas de evaluación previamente mencionadas para el modelo. En este caso de estudio, se analizó la primera división de validación cruzada y se recuperó el informe de clasificación utilizando Scikit-Learn.

Tabla 3 Informe de clasificación

SVM CV: 1 - Informe de resultados				
	precision	recall	f1 - score	Support
0	0.3	0.88	0.45	77
1	0.8	0.19	0.31	193
accuracy			0.39	270
macro avg	0.55	0.54	0.38	270
weighted avg	0.66	0.39	0.35	270

3.7. Probabilidad de que la zona se clasifique como mineralizada

En la Figura 10, el modelo SVM asignó probabilidades de predicción de mineralización al conjunto de pruebas CV:2, donde las muestras fueron coloreadas según sus valores reales (Castillo, 2022). Se observó que a la mayoría de las muestras mineralizadas se les asignaron altas probabilidades de predicción, lo cual fue una buena señal. Sin embargo, se detectaron algunas muestras negativas agrupadas entre ellas.

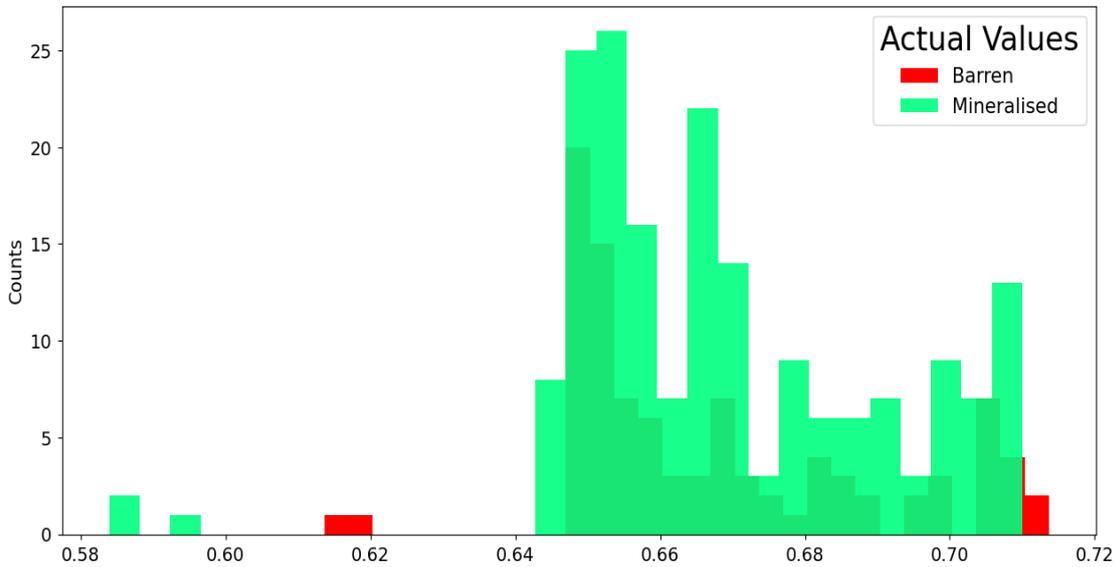


Figura 10. Probabilidad de que la zona se clasifique como mineralizada

3.8. Curva R

En la Figura 11 se observó un valor de AUC del 53.4 %, lo que indicó que la curva se encontraba en el límite de lo factible. Asimismo, el gráfico de la curva mostró una estructura escalonada. Esto se debió a que la cantidad de datos del conjunto de prueba fue bastante pequeña, y cada escalón de la curva representó un umbral de clasificación, tal como lo indicó (Cabanillas, 2022). Esto no implicó ninguna diferencia en la forma de calcular el AUC, sino que simplemente evidenció la insuficiencia de datos para obtener una curva ROC más suave. También se observó poca variación en el rendimiento del modelo SVM en los diferentes conjuntos de validación cruzada. En la curva ROC 3, se identificó el mayor desbalance en la base de datos, lo que ocurrió debido a la insuficiencia de datos en esa zona.

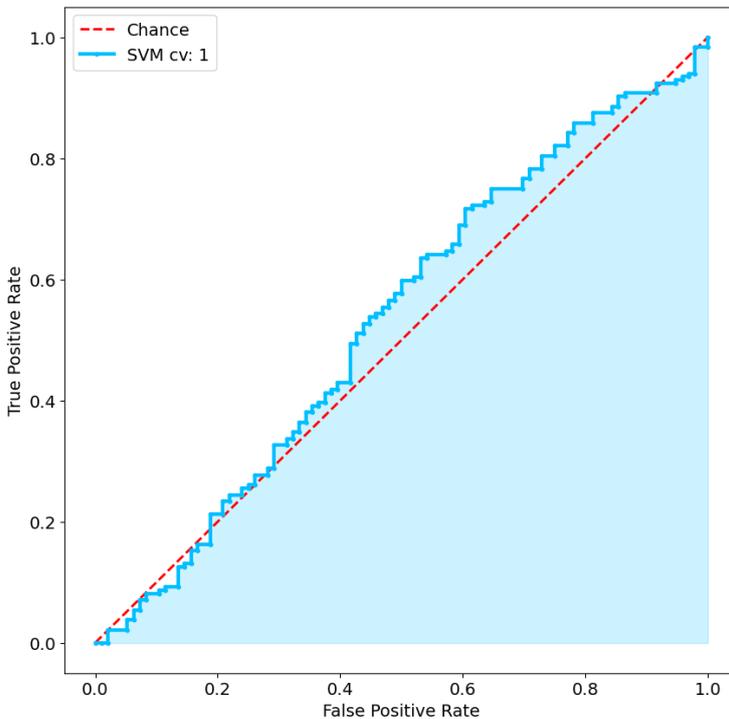


Figura 11. Curva R cv 1

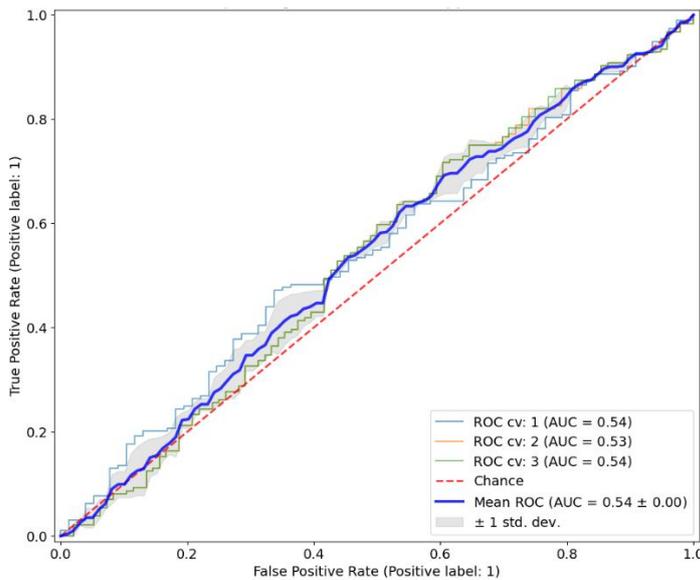


Figura 12. Característica operativa del receptor - Máquina de vectores soporte

4. CONCLUSIONES

Al analizar las representaciones gráficas de mineralización, se observó que todas las zonas delimitadas contienen recursos minerales. Sin embargo, la zona etiquetada con el número 6 destacó notablemente, mostrando el mayor contenido de mineralización entre las áreas estudiadas. Esta área específica alcanzó una probabilidad de mineralización del 66%, lo que indica un alto potencial para actividades mineras. A lo largo de este estudio, las librerías utilizadas en Python desempeñaron un papel crucial, sirviendo como base para todos los cálculos, la organización de datos y las visualizaciones gráficas. En particular, la librería “pipeline” resultó esencial para prevenir fugas de datos y garantizar la integridad del análisis. Para investigaciones futuras, se recomienda realizar estudios detallados sobre la zona 6, explorando sus características geológicas y económicas. Además, sería beneficioso integrar tecnologías emergentes y técnicas avanzadas de aprendizaje automático para mejorar la precisión de los modelos de predicción.

REFERENCIAS BIBLIOGRÁFICAS

- Arango, M. (2021). *Una librería en Python para el pronóstico de series de tiempo no lineales*. Medellín, Colombia: Universidad Nacional de Colombia.
- Cabanillas, J. (2022). *Eficacia de modelos Machine Learning para el pronóstico del riesgo crediticio*. TRUJILLO – PERÚ: UNIVERSIDAD NACIONAL DE TRUJILLO.
- Castillo, L. (2022). *Desarrollo de modelos predictivos de regresión en la industria minera mediante el uso de algoritmo de machine learning*. Lima, Perú: Universidad Nacional Mayor de San Marcos.
- Chen, J., Lu, J., Zhu, X., & Zhang, L. (2023). Generative Semantic Segmentation. *Computer Vision and Pattern Recognition*. Obtenido de <https://doi.org/10.48550/arXiv.2303.11316>.
- Correa, M. (2010). *Inteligencia artificial para la predicción y control del acabado superficial en procesos de fresado a alta velocidad*. Madrid: UNIVERSIDAD POLITÉCNICA DE MADRID.
- Ezequiel, A. (2019). *Análisis de la implementación de Machine Learning en el diagnóstico por imágenes*. Victoria, Buenos Aires: Universidad de San Andrés.
- Gutierrez, J., & Vigo, V. (2021). *Modelo de aprendizaje automatizado del proceso de venta de productos financieros en un call center*. Lima – Perú: Universidad de Lima.
- Hanuman, V. (2018). Python Libraries, Development Frameworks and Algorithms for Machine Learning Applications. *International Journal of Engineering Research & Technology (IJERT)*, 7 (4).
- Joakin, I. (2021). *Aplicación de tecnologías de aprendizaje automático para predecir negocios y tomar decisiones empresariales*. UNIVERSIDAD NACIONAL DE LA PLATA.
- Jovic, A. (2019). *An overview and comparison of free Python libraries for data mining and big data analysis*. *IEEE*. Obtenido de <https://doi.org/10.23919/MIPRO.2019.8757088>.

- Julca, B. (2020). *Aplicación de Deep Learning sobre imágenes topografica para mejorar la precisión del diagnóstico de Queratocono en una clinica de Lima*. Lima - Perú: Universidad de San Ignacio de Loyola.
- Li, C., Xin, D., Chen, T., Fang, X., Zhao, Q., & Chen, H. *Research on top-level planning for sustainable high-quality development of China's largest scale deep coal mining area*. *Front. Earth Sci*, 11. <https://doi.org/10.3389/feart.2023.1274369>.
- López, G. (2020). *Sistema inteligente para la detección automática de competencia y monitoreo de productos en mercado libre (mlcompet)*. Acapulco: Instituto Tecnológico De Acapulco.
- Parsa, M., Harris, J., & Sherlock, R. (2022). Improving mineral prospectivity model generalization: an example from orogenic gold mineralization of the Sturgeon Lake Transect, Ontario, Canada. *Ore Geology Reviews*, 104 -117.
- Pérez, A. (2014). *Maestria en Inteligencia Artificial*. . Mexico: Universidad Veracruzana.
- Sustaita, L. (2017). *Visualización de datos a partir de patrones de despliegue en lenguaje de programación R*. Zacatecas, Zacatecas.: Centro de Investigacion en Matematicas.
- Wang, C. (2022). Efficient customer segmentation in digital marketing using deep learning with swarm intelligence approach. *Information Processing & Management*, 59 (6).
- Zuo, R., & Carranza, E. (2023). Machine Learning-Based Mapping for Mineral Exploration. *Mathematical Geosciences*, (7), 891-895.