Random Forest for Generating Recommendations for Predicting Copper Recovery by Flotation

Víctor Flores, Nicolas Henríquez, Edgardo Ortiz, Rafael Martínez-Peláez, and Claudio Leiva

***Abstract*— In the copper mining industry, Data Science (DS) techniques and Machine Learning (ML) methods are contributing to improve the prediction of results in industrial processes. In this paper, an experience of applying both DS techniques and a ML algorithm, using historical data from the flotation process is described. These data were collected using a prototype of flotation equipment developed at the Universidad Católica del Norte, in Antofagasta, Chile. To achieve the result an Extraction, Transformation and Load (ETL) process was made. Also, for both, improving the understanding of domain dynamics and selecting the most relevant predictive variables in the flotation process, a Random Forest (FR) model was developed. The combination of these previous results made it possible to generate recommendations on the management of predictor variables to improve copper recovery in the context of the flotation equipment prototype. In this document, the methodological details are presented, and the process used to obtain the aforementioned results is described. As progress was made through 2 iterations, the quality of the results obtained with the predictive model, generated by RF, was improving. At the end of the process, an accuracy of 94,44% was achieved, with an accuracy in each of the classes greater than 90%. These results demonstrate the effectiveness and outstanding performance of the predictive model. These values are highly competitive when compared to those obtained in other similar studies in the context of Industry 4.0.**

Link to graphical and video abstracts, and to code: https://latamt.ieeer9.org/index.php/transactions/article/view/8687

***Index Terms*— Machine learning, Data science, flotation, predictive model, industry 4.0**

I. Introducción

La industria 4.0 (I4.0) ha emergido como una oportunidad transformadora para el sector industrial, como un cambio significativo en sus operaciones, desde el diseño hasta la gestión de procesos. La integración de tecnologías como la Inteligencia Artificial (IA) y las Ciencias de Datos (CD) permite a las empresas mejorar su eficiencia y productividad [1], contribuyendo al avance de los Objetivos de Desarrollo Sostenible (ODS) para 2030, especialmente los 9 y 12 [2].

En este contexto, la implementación de la I4.0 no solo optimiza los procesos productivos, sino que también representa un medio para las empresas de alcanzar estos objetivos mencionados [2]. Este nuevo paradigma tecnológico no solo mejora la eficiencia operativa, sino que también promueve una mejor calidad de vida, alineándose con la visión global de un futuro más sostenible y equitativo [3]. Para el sector minero, la adopción de la I4.0 no solo implica una evolución tecnológica, sino también un campo de estudio con amplios beneficios potenciales, mediante la implementación de tecnologías avanzadas en la cadena de suministro minero y la adopción de prácticas sostenibles [3].

En particular, en la producción de cobre, la aplicación de IA y CD ofrece un terreno propicio para mejorar la eficiencia operativa. El presente trabajo se centra en la generación de un modelo predictivo para la variable dependiente Y, que representa el diámetro de Sauter (D32), en el proceso de flotación. Este modelo se basa en una metodología que integra la preparación y comprensión de los datos, el entrenamiento con el algoritmo Random Forest (RF), la validación del modelo y la interpretación de resultados, así como la generación de recomendaciones dirigidas al operador de la máquina de flotación [4]. A través del análisis de datos, buscamos identificar patrones y anticipar las mejores condiciones de separación de minerales, lo que proporciona una base sólida para la toma de decisiones informada, contribuyendo así a mejorar la eficacia y reducir los costos operativos [5].

La investigación aborda un escenario crucial en el contexto actual, donde la industria de producción de cobre está siendo impactada por una creciente demanda global. Este incremento se atribuye a diversos factores, entre ellos, el impulso proveniente de industrias en auge como las energías limpias [7] y la fabricación de vehículos eléctricos [8]. Este contexto resalta la importancia estratégica de nuestro trabajo, aprovechando tecnologías de vanguardia como la IA y CD.

II. Revisión de la Literatura

La industria minera ha evolucionado hacia la I4.0, que incluye características como automatización completa, flexibilidad multidisciplinaria, escalabilidad, agilidad y resiliencia [4, 5]. En este contexto, los sistemas de soporte a la decisión en la producción industrial de cobre están siendo integrados a herramientas computacionales que respaldan la toma de decisiones en diversas tareas [6,7].

A. Trabajos Relacionados

En [4], se presenta un análisis comparativo del rendimiento de algoritmos de Machine Learning (Máquina de Soporte Vectorial, Random Forest y Red Neuronal Artificial) en la generación de modelos predictivos de recuperación de cobre

V. Flores, N. Henriquez, E. Ortiz, R. Martínez-Peláez, and C. Leiva are with the Universidad Católica del Norte, Antofagasta, Chile (e-mails: vflores@ucn.cl, nicolas.henriquez@alumnos.ucn.cl, edgardo.ortiz@alumnos.ucn.cl, rafael.martinez@ucn.cl, and cleiva01@ucn.cl).

por lixiviación. En [8], se describe un sistema para respaldar la toma de decisiones en la fabricación de piezas metálicas basado en la filosofía de la I4.0. Este estudio utiliza reglas de producción para facilitar la selección óptima de materiales en la manufactura metálica aditiva.

Asimismo, [9] presenta una herramienta computacional que apoya la evaluación de los impactos del ciclo de vida y la potencial recuperación de recursos en relaves mineros. En [10], se describe un método para la selección de características que influyen en la recuperación de cobre, utilizando un proceso de regresión Gaussiana. Igualmente, [11] describe un trabajo de predicción de recuperación de cobre por lixiviación mediante datos operacionales y Random Forest, con métricas de calidad del modelo que indican una precisión del 97.72% en la clasificación resultante. En la literatura son escasos los trabajos similares, más recientemente en [12] se reporta el trabajo realizado para predecir la recuperación de cobre usando datos reales recolectados en cinco celdas de flotación de la empresa BHP Olympic Dam (en Australia), los modelos predictivos fueron implementados con los algoritmos SVM, Gaussian process regression (GPR), ANN, linear regression (LR) y RF.

En estos estudios, la selección y preparación adecuada de los datos de entrada es un factor crucial [13], y se realiza mediante técnicas como ETL [14, 15]. El ETL comprende el proceso en el cual los datos se extraen de una o varias fuentes, se transforman en un formato y estructura comunes, y se cargan en un sistema de destino para su análisis posterior [16]. Recientes trabajos resaltan la importancia del ETL, por ejemplo, en [15, 17], se destaca su relevancia en proyectos donde se preparan y manejan datos en el contexto de la I4.0. En otros contextos de la I4.0, el ETL se emplea como herramienta para la integración de datos, según informan estudios como [18, 19].

B. Algoritmo Random Forest en Mineria

Los algoritmos de Machine Learning (ML) se están integrando de manera prominente en la I4.0, especialmente en el ámbito minero del cobre [20, 21]. ML, se enfoca en desarrollar algoritmos y modelos que permiten a las computadoras aprender de los datos y mejorar su desempeño en tareas específicas [22].

Uno de los algoritmos de ML más utilizados para generar modelos predictivos es Random Forest (RF), presentado en [8, 21]. Este algoritmo ha ganado relevancia debido a su habilidad para realizar predicciones precisas en diversos contextos dinámicos [21-23]. RF opera mediante la construcción de un conjunto de árboles de decisión cuyas capacidades predictivas individuales se combinan para reducir la varianza y mejorar la precisión de las predicciones [21].

RF se basa en el concepto de Árbol de Decisión, utilizando un enfoque recursivo de división binaria para alcanzar nodos finales, y puede aplicarse tanto para clasificación como para regresión [24]. Sus ventajas incluyen la capacidad para producir errores bajos, un rendimiento óptimo en clasificación, eficiente manejo de grandes conjuntos de datos y una estrategia efectiva para lidiar con datos faltantes [22].

En la práctica, el proceso con RF implica la generación de múltiples árboles independientes mediante subconjuntos de datos seleccionados aleatoriamente del conjunto original. La información obtenida en estos entrenamientos parciales se combina para mejorar la calidad general del modelo, minimizando la función de pérdida en cada iteración [14, 20]. En [23,24] se destaca el buen desempeño de RF en términos de eficiencia en los ciclos de entrenamiento, su capacidad predictiva incluso en conjuntos de datos desequilibrados y la interpretación de los resultados en tareas como la predicción basada en datos.

C. Proceso de Flotación

El proceso de flotación implica la separación de partículas minerales valiosas contenidas en una dispersión sólido-líquido (pulpa) mediante su adhesión selectiva a burbujas de aire [25]. La pulpa se agita en un estanque o columna, lo que genera burbujas que atraviesan la pulpa y forman una espuma ascendente recolectada en una zona de rebosamiento. Este proceso ha sido detalladamente descrito en estudios previos como [25, 26].

Una variable crucial en el proceso de flotación es el diámetro de la burbuja (conocido como Sauter diameter o D32), el cual influye en la calidad de las burbujas que transportan partículas de cobre hacia la parte superior del equipo de flotación. Estas burbujas, integradas en la espuma, se recogen y se llevan a otra etapa del proceso donde se separa el cobre de la espuma. La generación de espuma implica el uso de aditivos o espumantes, el control de la viscosidad de la pulpa para la formación de burbujas y la manipulación de varias variables relacionadas para obtener una espuma con un D32 adecuado [26].

III. Metodología y Contexto de Trabajo

La metodología de este estudio se fundamenta en investigaciones previas, como se menciona en [4, 21, 27], y también incorpora elementos de otros trabajos, como [10]. Este enfoque metodológico se divide en cuatro pasos principales, como se muestra en la Fig. 1, y se detallan a continuación.

Paso 1: Preparación y comprensión de los datos-se utilizaron las variables descritas en la Tabla I. Esta fase se llevó a cabo mediante el proceso de extracción, transformación y carga de datos (ETL). Se emplearon datos crudos provenientes de ciclos de flotación obtenidos con un prototipo desarrollado en el Departamento de Ingeniería Química y Medio Ambiente (DIQMA) de la Universidad Católica del Norte (UCN) [9]. Los datos se prepararon según criterios específicos como los descritos en [28], incluyendo el balanceo de la variable objetivo según los umbrales definidos y la eliminación de registros con valores en blanco o negativos, así como aquellos que excedieran los umbrales definidos en la Tabla II.

Paso 2: Entrenamiento con RF-se empleó el algoritmo Random Forest (RF) para entrenar un modelo predictivo, utilizando el software RapidMiner Studio® (licencia de uso educacional) y el conjunto de datos mencionado anteriormente. El proceso de entrenamiento se realizó en dos iteraciones, variando los hiperparámetros (profundidad del árbol y número de árboles) para obtener el mejor modelo en cada iteración.

Paso 3: Evaluación del modelo e interpretación de resultados-se utilizaron métricas o indicadores de rendimiento que son estándares en la literatura de aprendizaje automático [1, 21, 29], concretamente son: recall (r), precision (p) y accuracy (acc).

Paso 4: Generación de recomendaciones-posterior a la validación de cada modelo predictivo de RF, se calcularon los valores de importancia de las variables predictoras sobre la variable objetivo (clase Y). Estos valores se interpretaron y se utilizaron para generar recomendaciones dirigidas al operador, teniendo en cuenta la combinación de los valores de las variables predictoras que influyen en la variable objetivo y el conocimiento previo sobre el proceso de flotación.

TABLA I
VARIABLES PREDICTORAS Y LA CLASE. LAS VARIABLES X1 A X9 SON LAS PREDICTORAS Y LA Y REPRESENTA LA VARIABLE DEPENDIENTE

Variable	Descripción y unidad de medida
X1	Velocidad del gas superficial [cm/s]
X2	Concentrador (o Holdup) [%]
X3	Velocidad del líquido superficial [cm/s]
X4	Diámetro equivalente de la columna
X5	Densidad de la pulpa [g/cm3]
X6	Densidad de la burbuja [g/cm3]
X7	Viscosidad [g/cm]
X8	Tipo de espumante (fronther)
X9	Concentración del espumante (ppm)
Y	Sauter diameter o D32 [mm]

TABLA III
ESTA TABLA CONTIENE LOS RANGOS DE VARIABLES PREDICTORAS USADOS PARA CALCULAR LOS VALORES DE LAS ETIQUETAS DE LA CLASE

Variable	Bajo	Medio	Alto
X1	<1.0	[1.0, 1.6]	>1.6
X2	<0.7	[0.7, 1.5]	>1.5
X3	<9.50	[9.5, 16.5]	>16.5
X4	<0.3	[0.3, 0.9]	>0.9
X5	<3.1	[3.1, 5.0]	>5.0
X6	<0.01	[0.01, 0.02]	>0.02
X7	<0.011	[0.011, 0.012]	>0.012
X8	<1.0	[1.0, 2.3]	>2.3
X9	<10.0	[10.0, 35.0]	>35.0
Y			

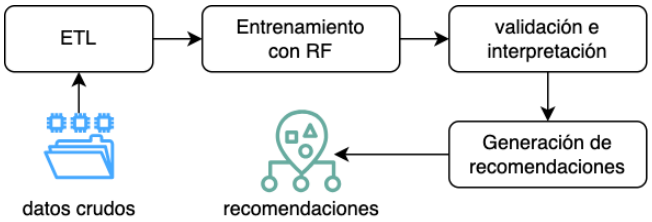


Fig. 1. Esquema de trabajo con la metodología planteada.

Para lo descrito en el Paso 3, en función a lo indicado en los Pasos 1 y 2, se preparó una matriz de confusión que resume los resultados. Esta matriz facilita el análisis necesario para determinar dónde se producen los errores de predicción. Utilizando esta matriz se calcularán los indicadores de rendimiento antes mencionados.

La matriz de confusión es una tabla que muestra la distribución de los errores por las distintas categorías [4, 14, 29]. Se detalla la matriz de confusión en la Tabla III, para el caso de dos clases (positiva y negativa en este ejemplo), siendo los conceptos directamente generalizables. La forma de calcular las métricas r, p y acc es mostrada en las ecuaciones (1), (2) y (3) respectivamente.

$$r = (a+d)/(a+c). \quad (1)$$

$$p = a/(a+b). \quad (2)$$

$$acc = (a+d)/(a+b+c+d). \quad (3)$$

TABLA IIIII
MATRIZ DE CONFUSIÓN ESQUEMÁTICA, MOSTRANDO LOS VALORES GENÉRICOS DE LAS POSIBLES CLASIFICACIONES

		Clases verdaderas	
		Positiva	Negativa
Clases predichas	Positiva	a	b
	Negativa	c	d

A. Breve Descripción del Diseño Experimental

Como se ha dicho previamente, los datos usados en este trabajo fueron generados en el contexto de un Capstone Project en el DIQMA de la UCN. Este proyecto se detalla en [9]. Para el desarrollo del sistema de la planta de flotación se implementó un controlador lógico programable (PLC) en forma de autómata programable capaz de controlar en tiempo real procesos secuenciales y de procesar la información recibida por sensores instalados en el prototipo.

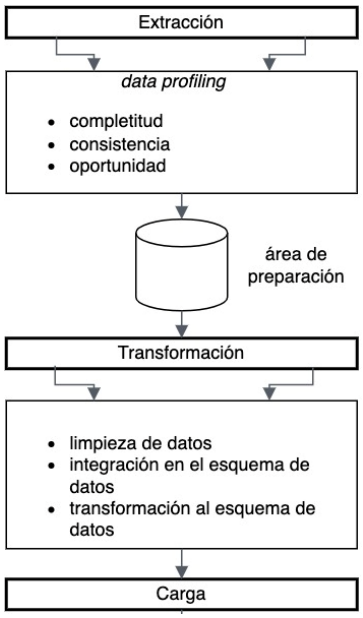


Fig. 2. Esquema ETL para preparar los datos del dataset.

La preparación se realizó siguiendo lo mostrado en la Fig. 2. La adquisición de datos numéricos de Y se realizó en ciclos de 2-3 hora de duración, donde se variaron para cada ciclo las configuraciones de X1, X3 y X8.

En todos los ciclos se mantuvo una única configuración para las variables X2, X4, X5 y X7. Existe una relación directa entre X8 y X9, es decir, dado un X8 se tiene un valor de X9. Se usaron solo dos tipos de X8 (solo uno en cada ciclo), con lo cual fueron igualmente dos posibles valores de X9. En la operación por cada ciclo se obtuvieron valores numéricos de Y a través de un balance de masa que permite obtener la cantidad de cobre recuperado para luego establecer una curva de recuperación. Al final de lo descrito, los datos se guardaron en un archivo plano en formato.csv.

B. Caso de Estudio

Para el caso de estudio se adquirieron datos según el Diseño Experimental del trabajo previo [9]. Se usaron 515 registros de datos operacionales para preparar 121 registros (el dataset) con datos balanceados, siguiendo lo descrito en el punto 1) de la Metodología.

Los valores numéricos del D32 (variable dependiente Y) fueron discretizados en 'Bajo', 'Medio' y 'Alto'. Para ello y de forma similar a [14], en esta discretización se consideró la desviación estándar σ; el valor 'Bajo' de Y corresponde a valores menores que -σ, el valor 'Medio' corresponde a valores que están en el intervalo [-σ, σ] y el valor 'Alto' corresponde a valores mayores que +σ. Una vez realizada la categorización según lo descrito previamente, se obtuvieron 29 registros en la categoría 'Alto', 85 en la categoría 'Medio' y 7 en la categoría 'Bajo'.

IV. Resultados y Discusión

En este trabajo se lograron dos resultados relevantes. El primero está relacionado con la generación del modelo predictivo y la identificación de las variables de mayor importancia sobre la clase Y. El segundo está relacionado con la generación de recomendaciones al operador de la máquina de flotación, para optimizar el proceso. A continuación, se detallan ambos resultados.

Para generar el modelo predictivo se usaron los pasos 1 y 2 de la metodología. La validación de la calidad del modelo se realizó con lo indicado en el paso 3. El Algoritmo 1 resume el pseudocódigo de estas tareas.

Para alcanzar el mejor modelo se realizaron dos iteraciones, dado que los resultados en la primera iteración no fueron muy óptmos según se explica más adelante. Las dos iteraciones se realizaron con configuraciones distintas respecto a los hiperparámetros del algoritmo RF número de árboles (num_arbol) y profundidad (prof). Para la primera iteración se crearon modelos con num_arbol in [50-60] y prof in [5-7]. Para la segunda se usaron num_arbol in [90-100] y prof in [7-9]. Para ambas iteraciones los datos del dataset fueron separados en 70% para entrenamiento y 30% para validación, el criterio de clasificación seleccionado fue information_gain.

Las predicciones de la primera iteración se detallan en la Tabla V. El mejor modelo se logró con num_arbol = 50 árbol y prof = 5. En este primer entrenamiento se logró una acc = 88.89% y la mayor precisión fue para la predicción de la etiqueta 'Medio'.

Del conjunto de entrenamiento, un total de 26 registros fueron categorizados con esta etiqueta, y se alcanzaron valores de p=89.66% y r=91.30%, lo que indica una buena proporción de casos positivos que fueron correctamente clasificados e identificados, respectivamente. Por otro lado, ningún caso fue clasificado como "Bajo", lo que lleva a pensar en la posibilidad de un mal entrenamiento con la configuración seleccionada o una mala selección de valores umbrales para crear la etiqueta.

Tabla IV
Valores Mínimos, Máximos, y Valores Medios de las Variables Predictoras Usados en el Análisis de Resultados

	X1	X2	X3	X4	X5	X6	X7	X8	X9
Max	2.2	2.5	32	1	10.31	0.02	0.012	3	75
Min	0.5	0.62	9.5	0.3	3.21	0.01	0.011	1	5
Media	1	0.98	12.4	0.91	3.81	0.01	0.011	1	10

Con los resultados de la primera iteración se realizó un análisis que sirve de 'base' para la siguiente iteración. Este análisis se incluyó el estudio de la pertinencia de las variables predictoras y los rangos de dichas variables (ver Tabla II), y el cálculo de la importancia de cada variable en el modelo. Con ambos aspectos se eliminaron las variables X6 y X7 del dataset.

Algoritmo 1 Entrenamiento y Generación de matriz confusión

```
SPLIT_DATASET(X)
  N1 ←Ramdom(X)*0.7
  N2 ←Ramdom((X-N1)*0.3
  Return(N1, N2)

TRAIN(N1)
  V1← RapidMiner®_RandomForest(N1)
  Return (V1)  //retorna valores de interés: a, b, c, d

TESTING(N2)
  V2←RapidMiner®_Performance(N2)
       // realiza procesos como CrossValidation
  Return (V2)  // retorna r, p, acc

PREDICT(X)
  SPLIT_DATASET(X)
  TRAIN(N1)
  TESTING(N2)
     Return(V1, V2)
```

A continuación se detalla el argumento para dicha acción.

Como muestra la Tabla IV, la variación entre los valores mínimo y máximo de las variables X6 y X7 son muy pequeños o coinciden con el valor medio. Estos valores tan poco significativos pueden influir en el modelo RF causando una mala clasificación de Y.

La Fig. 3 parte (a) muestra, en orden descendente, la importancia de cada variable predictora (rango de 0 a 1) para el modelo generado para la Iteración 1. Se puede observar que la variable X6 tiene un valor muy pequeño de significancia (0.001) y la variable X7 no tiene significancia, lo que significa que ambas no tienen pertinencia como variables predictoras en el modelo. Igualmente se puede ver en la figura 3 parte (a) que las cuatro variables que acumulan más del 0.5 son X3, X1, X4 y X5.

Tabla V
Matriz de Confusión con los Resultados de la Iteración 1. La Matriz Contiene la Cantidad de Instancias Clasificadas en cada uno de los Rangos Según las Etiquetas de la Clase

Accuracy: 88.89%	Verd. Medio	Verd. Alto	Verd. Bajo	precision
Pred. Medio	26	2	1	89.66%
Pred. Alto	1	6	0	85.71%
Pred. Bajo	0	0	0	0.00%
class recall	96.30%	75.00%	0.00%	

En base a lo descrito en los párrafos anteriores, el dataset fue modificado y se realizó el entrenamiento para la iteración 2. Las predicciones de la segunda iteración se detallan en la Tabla VI.

Tabla VI
Matriz de Confusión con los Resultados de la Iteración 2. La Matriz Contiene la Cantidad de Instancias Clasificadas en cada uno de los Rangos Según las Etiquetas de la Clase

Accuracy: 94.44%	Verd. Medio	Verd. Alto	Verd. Bajo	precision
Pred. Medio	26	0	1	89.66%
Pred. Alto	1	8	0	85.71%
Pred. Bajo	0	0	0	0.00%
Class recall	96.30%	100%	0.00%	

(a)

attribute	wei...
X3	0.287
X1	0.275
X4	0.108
X5	0.092
X2	0.073
X9	0.051
X8	0.033
X6	0.001
X7	0

(b)

attribute	wei...
X3	0.287
X1	0.275
X4	0.108
X5	0.092
X2	0.073
X9	0.051
X8	0.033

Fig. 3. Pesos de variables predictoras: parte (a) para la Iteración 1 y parte (b) para la Iteración 2.

El mejor modelo en la iteración 2 se logró con num_arbol = 50 árbol y prof = 5. En este segundo entrenamiento se obtuvo acc = 94.44%, lo que denota que la decisión de eliminar las variables X6 y X7 beneficia la exactitud del modelo. Del conjunto de entrenamiento, un total de 26 registros fueron categorizados como Medio y mejoró al 100% la categorización en Alto.

En este entrenamiento se alcanzaron valores de p = 92.66% y r=98.15%. que indican una buena proporción de casos correctamente clasificados y una excelente proporción de casos correctamente identificados, respectivamente.

Los resultados de esta iteración 2 fueron analizados en busca de patrones para crear recomendaciones que se puedan ofrecer al operador de la máquina de flotación. Para ello, en primer lugar, se generaron y analizaron los pesos por variable predictoras, luego se analizaron las distribuciones de valores de las variables con histogramas y también se analizaron las posibles correlaciones entre las variables. Estos análisis se detallan a continuación.

La Fig. 3 parte (b) muestra la importancia de cada variable predictora en el modelo para la Iteración 2. En esta parte de la Fig. 3 se puede observar que se mantiene tanto el orden como los valores individuales para las 4 variables más significativas.

Para identificar los rangos de valores de las variables significativas se generaron histogramas que son descritos en la Fig. 4. Como muestran las partes (a) y (b) de la Fig. 4, los valores de las variables X3 y X1 tienen una distribución más o menos homogénea en el rango de valores posibles, mientras que, como se puede observar en las partes (c) y (d), la mayoría de los valores de las variables X4 y X5 se ubican en los extremos del rango de posibles valores. Estos valores pueden indicar una necesidad de revisión de los rangos considerados para generar las etiquetas de Y.

La Fig. 5 muestra que la correlación entre las variables es relativamente baja, siendo la más alta entre las variables X1 y X9, y X8 y X9. Siendo que X1 es una de las variables que tiene más incidencia en el modelo (según se expresa en Fig. 3(b)), valdría la pena realizar un estudio más detallado para intentar disminuir esa correlación, y verificar su incidencia en el modelo predictivo.

El análisis descrito en los párrafos anteriores permitió deducir nuevo conocimiento que, unido al conocimiento previo, permitió generar las reglas de recomendaciones al operador de la máquina de flotación. Estas recomendaciones son conocimiento heurístico que puede disponerse en una aplicación software para ayudar a tomar decisiones para la mejor combinación de parámetros en la operación del prototipo de máquina de flotación para la predicción de Y.

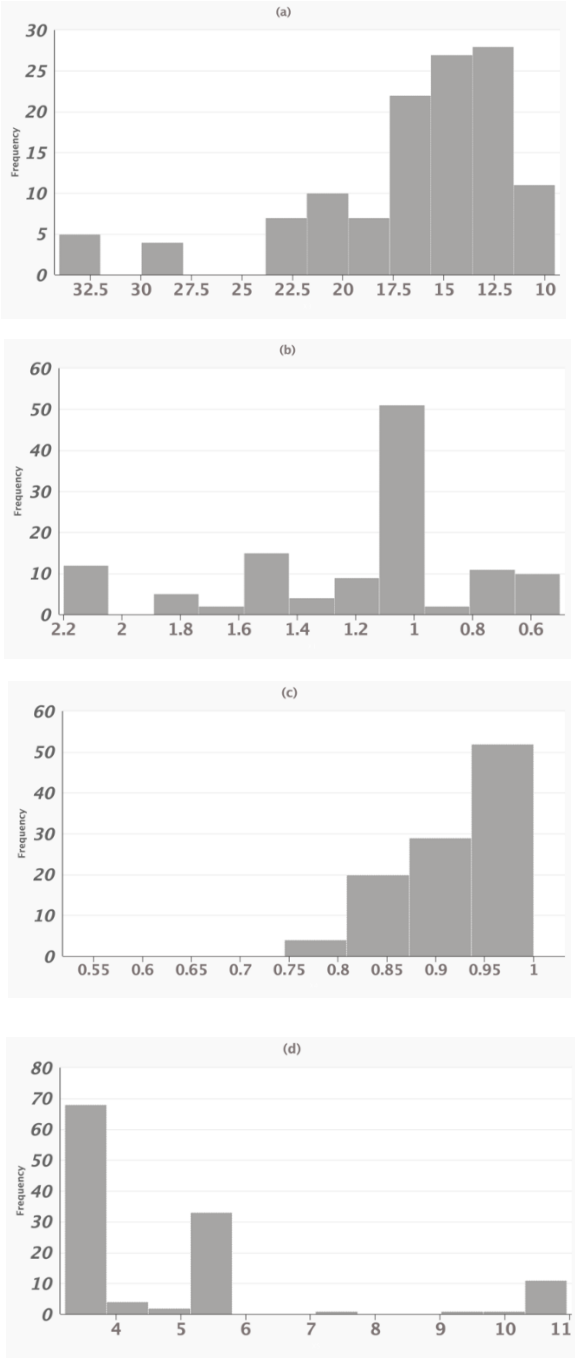


Fig. 4. Histogramas de las variables con mayor impacto en el modelo predictivo: parte (a) para la variable X3, parte (b) para la variable X1, parte (c) para la variable X4 y parte (d) para la variable X5.

La Fig. 6 muestra un ejemplo de combinación de valores de parámetros que dan origen a una recomendación como “Media” y “Alta”. Estas combinaciones pueden presentarse en formato texto en un prototipo de aplicación web. Estas recomendaciones son particulares al prototipo de máquina de flotación usado, que

han sido a su vez la entrada a los modelos predictivos descritos en este trabajo.

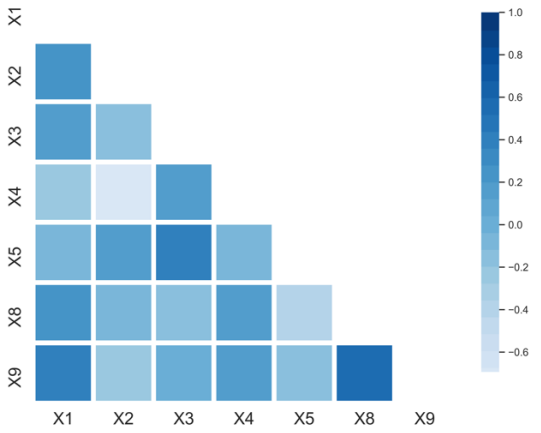


Fig. 5. Matriz diagonal de correlación entre las variables predictoras.

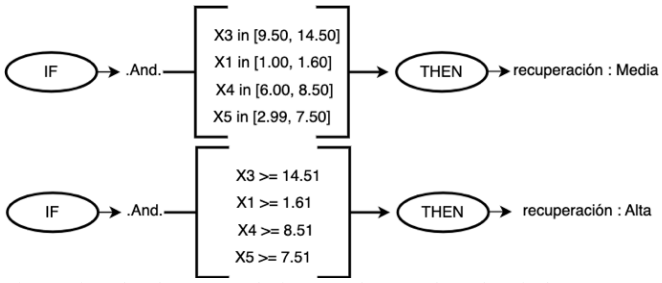


Fig. 6. Ejemplos de recomendaciones en formato de reglas If-Then.

V. Conclusiones y Líneas Futuras

En este estudio, se ha desarrollado una innovación que ofrece un enfoque novedoso para respaldar la toma de decisiones destinadas a mejorar los resultados en la recuperación de cobre mediante el uso de un equipo de flotación.

El proceso detallado en este trabajo abarca desde la preparación de datos utilizando ETL, hasta la generación de un modelo predictivo utilizando Random Forest, y la formulación de recomendaciones para optimizar la recuperación de cobre a través del proceso de flotación.

Es importante destacar que este tipo de innovaciones son escasas en la literatura existente, y hasta el momento no se ha encontrado evidencia de un sistema de recomendación en el ámbito específico abordado en este trabajo. Por lo tanto, los resultados obtenidos representan una contribución innovadora para el campo de la flotación y la aplicación de la Industria 4.0 en este sector.

Los datos utilizados fueron recopilados utilizando un prototipo de máquina de flotación desarrollado en el Departamento de Ingeniería Química y Medio Ambiente (DIQMA) de la Universidad Católica del Norte (UCN). Para la creación del modelo predictivo, se empleó el algoritmo Random Forest (RF), el cual fue sometido a dos iteraciones de entrenamiento. En la primera iteración, se identificaron variables de escaso impacto en el modelo, lo que condujo a ajustes para un segundo entrenamiento. Este segundo ciclo de entrenamiento permitió mejorar significativamente métricas de desempeño como la precisión en más de un 20%.

Además de la generación del modelo predictivo, se formularon recomendaciones específicas para el operador, basadas en el modelo generado y en el conocimiento previo del proceso. Estas recomendaciones, presentadas en forma de reglas, tienen el potencial de guiar la configuración de los parámetros del prototipo de flotación para mejorar la recuperación de cobre. En conjunto, estas contribuciones promueven el avance de la industria minera hacia la era de la Industria 4.0, al proporcionar herramientas y estrategias innovadoras para optimizar los procesos de recuperación de minerales.

En futuras líneas de trabajo, se plantea la validación y mejora del modelo mediante nuevas experimentaciones, así como la implementación de un sistema inteligente de recomendación en tiempo real para optimizar la operación de la máquina de flotación.

La principal limitación de la investigación radica en la falta de estudios previos que aborden específicamente el desarrollo de un sistema de recomendación en el dominio de la flotación de cobre. Esta carencia de evidencia en la literatura dificulta la comparación y validación de los resultados obtenidos en este trabajo, lo que resalta la necesidad de realizar más investigaciones para respaldar y validar la innovación propuesta.

Agradecimientos

El grupo de trabajo quiere gradecer el trabajo previo desarrollado en el Núcleo de investigación ‘Industria 4.0’ de la Universidad Católica del Norte (UCN), por las tareas previas de fabricación del prototipo de flotación y la recolección de datos que fueron usado como input al presente trabajo.

Referencias

[1] M. Abedini, M. Ziaii, T. Timkin, A. Pour, “Machine Learning (ML)-Based Copper Mineralization Prospectivity Mapping (MPM) Using Mining Geochemistry Method and Remote Sensing Satellite Data”. Remote Sensing, vol. 15, no. 15, pp. 1-12, 2023, doi: 10.3390/rs15153708.

[2] P. Katila, C.J. Colfer, W. De Jong, G.Galloway, P. Pacheco, G. Winkel. Sustainable development goals. Report 2022. Cambridge University Press. 2022, doi: 10.1017/9781009210058.

[3] X. Wang, P. Chen, C. Chow, D. Lau, "Artificial-intelligence-led revolution of construction materials: From molecules to Industry 4.0". Matter, vol. 6, no. 6, pp. 1831-1859, 2023, doi: 10.1016/j.matt.2023.04.016.

[4] V. Flores, C. Leiva, "A comparative study on supervised machine learning algorithms for copper recovery quality prediction in a leaching process". Sensors, vol. 21, no. 6, pp. 1-20. 2021, doi: 10.3390/s21062119

[5] N.C. Patel, A. Debnath, "Data Science with Semantic Technologies: New Trends and Future Developments". In: CRC Press ed. 2023, doi: 10.1201/9781003310785.

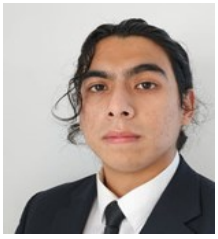
[6] Z. Wang, D. He, Z. Wang, Q. Li, "Timeliness and Stability-Based Operation Optimization for Copper Flotation Industrial Process" in IEEE Transactions on Instrumentation and Measurement, 72, 1-12. 2023, doi: 10.1109/TIM.2022.3225924.

[7] C. Li, T. Shen, Y. Zhou, "EMPC of aluminium wire and copper terminal for electric vehicles. Materials and Manufacturing Processes, vol. 38, no. 4, pp. 1-8. 2022, doi: 10.1080/10426914.2022.2105890.

[8] P. Putra, A. Azanuddin, B. Purba, Y. Dalimunthe. “Random Forest and decision tree algorithms for car price prediction". Journal Matematika Dan Ilmu Pengetahuan Alam LLDikti

Wilayah 1 (JUMPA), 3(2), 81-89. 2023. doi: 10.54076/jumpa.v3i2.305
[9] L. Pinto, C Rojas, L. Torres, "Operación y puesta en marcha de planta piloto de extracción por solvente mediante hollow drop", Capstone Project undergraduate thesis, Universidad Católica del Norte, Antofagasta. 2021.
[10] R. Parvanda, P. Kala, "Trends, opportunities, and challenges in the integration of the additive manufacturing with Industry 4.0.", Progress in Additive Manufacturing, vol. 8, no. 3, pp. 587-614. 2022, doi: 10.1007/s40964-022-00351-1.
[11] L. Adrianto, S. Ciacci, P. Pfister, S. Hellweg, "Toward sustainable reprocessing and valorization of sulfidic copper tailings: Scenarios and prospective LCA". Science of the Total Environment, vol. 871, pp. 1-13, 2023, doi: 10.1016/j.scitotenv.2023.162038.
[12] B. Amankwaa-Kyeremeh, C. McCamley, M. Zanin, C. Greet, K. Ehrig, R.K. Asamoah. “Prediction and Optimisation of Copper Recovery in the Rougher Flotation Circuit”. Minerals, 14(1), 36, 2023, pp. 1-31.
[13] M. Yaqot, B. Menezes, R. Franzoi, "Interplaying of industry 4.0 and circular economy in cyber-physical systems towards the mines of the future. In 32nd European Symposium on Computer Aided Process Engineering, vol. 15. Computer aided chemical engineering (Ed.), pp. 1609-1614, 2022, doi: 10.1016/B978-0-323-95879-0.50269-1.
[14] B. Amankwaa-Kyeremeh, J. Zhang, M. Zanin, W. Skinner, R.K. Asamoah. “Feature selection and Gaussian process prediction of rougher copper recovery”. Minerals Engineering, vol. 170, pp. 15-24, 2021, doi: 10.1016/j.mineng.2021.107041.
[15] V. Flores, B. Keith, C. Leiva, "Using artificial intelligence techniques to improve the prediction of copper recovery by leaching", Journal of Sensors, pp. 1-12. 2020, doi: 10.1155/2020/2454875.
[16] D. Seenivasan, "ETL (Extract, Transform, Load) Best Practices". International Journal of Computer Trends and Technology, vol. 71, no. 1, pp. 40-44, 2023, doi: 10.14445/22312803/IJCTT-V71I1P106.
[17] P. Gundarapu, "Industry 4.0: Data and Data Integration". In Big Data Applications in Industry 4.0, ch. 2, pp. 39-54, Auerbach Publications, 2022, doi: 10.1201/9781003175889-2.
[18] Shirazi, A. Hezarkhani, A. Shirazy, A. Pour, "Geochemical Modeling of Copper Mineralization Using Geostatistical and Machine Learning Algorithms in the Sahlabad Area, Iran". Minerals, vol. 13, no. 9, pp. 1-27, 2023, doi: 10.3390/min13091133.
[19] T.C. Uyan, K. Otto, M. Silva, E. Armakan, "Industry 4.0 foundry data management and supervised machine learning in low-pressure die casting quality improvement". International Journal of Metalcasting, vol. 17, pp. 414-429, 2023, doi: 10.1007/s40962-022-00783-z.
[20] D. Gyasi-Antwi, O. Apea, "Novel copper-sawdust nanocomposite preparation and evaluation". Results in Chemistry, vol. 5, pp. 1-14, 2023, doi: 10.1016/j.rechem.2022.100741.
[21] V. Flores, I. Bravo, M. Saavedra, M. “Water Quality Classification and Machine Learning Model for Predicting Water Quality Status-A Study on Loa River Located in an Extremely Arid Environment: Atacama Desert”. Water, vol. 15, no. 16, pp.1-18. 2023, doi: 10.3390/w15162868.
[22] H. Liang, C. Yang, K. Huang, D. Wu, W. Gui, "A transfer predictive control method based on inter-domain mapping learning with application to industrial roasting process". ISA transactions, vol. 134, 472-480. 2023, doi: 10.1016/j.isatra.2022.08.022.
[23] L. Breiman, "Random forests". Machine learning, vol. 45, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.
[24] F. Smarra, G. Di Girolamo, V. De Iuliis, A. Jain, R. Mangharam, A. D'Innocenzo, "Data-driven switching modeling for mpc using regression trees and random forests". Nonlinear Analysis: Hybrid Systems, vol. 36, pp. 1-12, 2020, doi: 10.1016/j.nahs.2020.100882.
[25] B.K. Loveday, C.J. Brouckaert, "An analysis of flotation circuit design principles". The Chemical Engineering Journal, vol. 59, pp. 15-21, 1995, doi: 10.1016/0923-0467(95)03001-8.
[26] S.R. Rao, "Surface Chemistry of Froth Flotation". In: Kluwer Academic/Plenum Publishers (Ed.), 2nd Rev Edition. Ch 1-3, 2003, doi: 10.1007/978-1-4615-7975-5.
[27] M. Saldaña, P. Neira, V. Flores, P. Robles, C. Moraga, "A decision support system for changes in operation modes of the copper heap leaching process". Metals, vol. 11, no. 7, pp. 1-12, 2021, doi: 10.3390/met11071025.
[28] Q. Qin, P. Qi, P. Shi, P. Scott, X. Jiang, "Selection of materials in metal additive manufacturing via three-way decision-making". The International Journal of Advanced Manufacturing Technology, vol. 126, no.3, 1293-1302. 2023. doi: 10.1007/s00170-023-10966-5.
[29] Sammut, C., & Webb, G. I. (Eds.). (2011). Encyclopedia of machine learning. Springer Science & Business Media.

Victor Flores received both Ph.D. degree (2011) and M.Sc degree on Software Engineering (2003) from the Technical University of Madrid. He recived a M.Sc. degree on Data science from the Postgraduate-Study-Center of Madrid (2023). He is professor in the Department of Systems Engineering and Computing, Universidad Católica del Norte. His research interests include Artificial intelligence, Data science, machine learning, Software engineering, and mineral processing. He is co-author of more than fifty scientific articles published in journals and conferences.



Nicolás Henríquez is a current fifth-year student in Civil Engineering with a specialization in Computation and Informatics at Universidad Católica del Norte (Antofagasta, Chile). Since his enrollment in 2019, he has fully immersed himself in the captivating realm of technology and informatics, honing his skills in programming and software development.

Edgardo Ortiz is a student of Civil Engineering in Computing and Informatics. He joined Universidad Católica del Norte (Antofagasta, Chile) in 2019 and is currently in his fifth year of studies, driven by his curiosity about the world of technology. Throughout his education, Edgardo's interest in artificial intelligence has steadily grown, immersing himself gradually in the world of data science.

Rafael Martínez-Peláez holds a PhD from the Polytechnic University of Catalonia and a Computer Systems Engineer from the Universidad del Valle de México in 2010 and 2003, respectively. Currently, he is a Professor at the Universidad Católica del Norte. He is co-author of more than fifty scientific articles published in journals and conferences.

Claudio Leiva recived a M.Sc. on Administration of Mining Contracts from Escuela Europea de Negocios. He is PhD candidate in Mineral Processing at University of Oulu (Finland). His interests are mainly focused on systems control, machine learning, and mineral processing. He is currently working at Universidad Católica del Norte (Antofagasta, Chile).