UNIVERSIDAD DE CONCEPCIÓN
Facultad de Ingeniería
Departamento de Ingeniería Metalúrgica

Profesor Guía
Dr. Leopoldo Gutiérrez Briones
Ingeniero Supervisor
Sergio Arellano Gorigoitía

MODELACIÓN DEL PORCENTAJE DE INSOLUBLES EN CONCENTRADO FINAL DE COBRE EN FUNCIÓN DE VARIABLES OPERACIONALES DEL PROCESO EN COMPAÑÍA MINERA DOÑA INÉS DE COLLAHUASI

PATRICIO ANDRÉS MAMANI MIRANDA

Informe de Memoria de Título
Para optar al título de
Ingeniero Civil Metalúrgico

Octubre, 2025

Dedicado a mi yo del pasado, para demostrarle que se puede lograr todo lo que te propongas

Y a NewJeans (NJZ)

Agradecimientos

Primero quiero agradecer a Dios por guiarme en mi etapa universitaria y permitirme realizar mi memoria de título en una gran compañía minera.

Me gustaría agradecer a muchas personas. Inicialmente a mi familia, por el apoyo constante que me han entregado en todos estos años, especialmente a mi mamá por apoyarme en mis estudios, los momentos de videollamadas que hacíamos y por motivarme a continuar estudiando la carrera ejercida en la Universidad de Concepción, a mi papá por trabajar para tener un arriendo en Concepción con buenas condiciones de vivienda y a mis hermanas y hermano por estar siempre apoyando en lo que sea y de sus infaltables chistes o payasadas que hacían reír a pesar de la distancia.

También agradecer el apoyo de mis amigos y amigas de la universidad, quienes conocí a lo largo de estos años de carrera ya sea en tiempos de pandemia o de manera presencial, a Sebastián, Valentina, Yuliza, Christian (Calama), Alonso, Benjamín, Gabi, Walter, Nico y así con varias personas más que me gustaría agregar, pero me faltaría espacio en esta página, pero cada uno sabe lo importante que ha sido en mi vida. Atesoro cada momento que compartí con ellos: aquellos que la pasábamos estudiando a través de *Discord* o en la central o en el DIMET, los ratos que estuvimos jugando *League of Legends* o GTA V, los momentos que salía su junta o cualquier momento '*random'* que valoro mucho mientras sea el compartir con ellos y pasar un buen momento.

Agradecer a las personas que conocí mientras realizaba mi memoria de título, porque gracias a ellos tuve una experiencia gratificante en la minera. Agradezco a mi tutor por la paciencia y la disposición a responder cualquier pregunta requerida y agradecer a mi compañera de memoria y mis compañeras que realizaron la práctica por los buenos momentos que compartíamos en la oficina y por los chismecitos que había cada semana.

A Compañía Minera Doña Inés de Collahuasi, por brindarme la oportunidad de realizar mi memoria en dicha compañía, especialmente a la Gerencia de Operaciones Plantas y Gerencia de Metalurgia, por apoyarme constantemente en el proceso y ayudarme respecto a las inquietudes generadas en los 6 meses, tanto para la memoria y para conocer los procesos que están involucrados en la planta.

Finalmente me gustaría agradecer a mi profesor guía por ser un gran profesor y apoyarme en poder desarrollar una buena memoria de título.

Resumen

Este informe presenta un estudio de la variación de porcentaje de insolubles presentes en el concentrado final de cobre a través del uso de modelos predictivos supervisados con la información recopilada de 3 años y de 6 meses, cada uno de distintos periodos. Para esto se seleccionan datos de variables que tengan significancia en la variación de insolubles, esto a través de una limpieza de datos, o sea eliminando datos anormales, atípicos y con valor 0, y luego se analizan estadísticamente, optando por variables que tienen un valor p menor a 0,05 y algunas variables que sean mayor a 0,05 y que impacten de gran manera con los insolubles, con el fin de trabajar con las variables escogidas y que beneficien a los modelos tanto teórica y estadísticamente.

Los modelos predictivos que se usaron fueron supervisados, los cuales trabajan a partir de una variable objetivo (en este caso, porcentaje de insolubles) respecto a las variables independientes, siendo estas obtenidas anteriormente. Se trabajan con modelo de tipo árboles (*XGBoost, Gradient Boosting y Random Forest*), modelo de redes neuronales (Redes Neuronales Profundas, DNN, y Redes Neuronales Artificiales, ANN) y con modelos que usan máquinas de vectores de soporte, que en este caso se usó la Regresión de Vectores de Soporte (SVR). Se construye una tabla de métricas de desempeño, en donde se comparan los modelos y se selecciona el modelo que entregue valores ideales en dicha tabla.

Los resultados obtenidos en el análisis de datos operacionales muestran que los algoritmos predictivos basados en árboles, especialmente XGBoost, presentan un mejor desempeño al trabajar con información histórica de tres años, debido a su capacidad para manejar grandes volúmenes de datos y relaciones complejas. En cambio, al utilizar un subconjunto reducido de seis meses, las Redes Neuronales Artificiales entregan mejores resultados en términos de ajuste y error promedio respecto a las otras técnicas evaluadas. Esto sugiere que, mientras los métodos tipo árbol se adaptan mejor a series extensas, las redes neuronales responden con mayor eficacia cuando se dispone de un conjunto de datos más limitado y reciente.

Abstract

This report examines the variation in the percentage of insoluble material in the final copper concentrate using supervised predictive models, based on data collected over two distinct periods: three years and six months. To this end, variables significantly influencing insoluble content were selected through data preprocessing, which included the removal of abnormal, outlier, and zero-value data, followed by statistical analysis. Variables with a p-value below 0.05 were prioritized, while some with higher values were retained due to their practical relevance, ensuring the inclusion of inputs with both statistical and operational significance.

The predictive models applied are supervised learning approaches, relying on a target variable (percentage of insoluble) and a set of independent variables derived from the processed dataset. These include tree-based algorithms (XGBoost, Gradient Boosting, and Random Forest), neural network architectures (Artificial Neural Networks, ANN, and Deep Neural Networks, DNN), and a support vector machine model, specifically Support Vector Regression (SVR). A comparative performance metrics table was constructed to identify the model achieving the most accurate and reliable results.

The findings indicate that tree-based models, particularly XGBoost, perform optimally with three years of historical data due to their capacity to manage large datasets and capture complex relationships. Conversely, with a smaller six-month dataset, the ANN model demonstrated superior performance in terms of fit and error reduction compared to other approaches. This suggests that tree-based algorithms are more suitable for extensive datasets, while neural networks offer greater adaptability and accuracy when working with shorter and more recent data series.

INDICE

Índice de Figuras

Índice de Tablas

1 Introducción

El procesamiento de minerales es una etapa fundamental en el ámbito minero, ya que busca procesar la roca proveniente de yacimientos mineros a rajo abierto o subterráneo para así procesarlo por distintos equipos para obtener un producto que se pueda comercializar.

Dependiendo de la minería es que comercializan distintos tipos de elementos: cobre, oro, plata, molibdeno, yodo, entre otros. Entre esos, la Compañía Minera Doña Inés de Collahuasi (CMDIC) se encarga del procesamiento del cobre en su forma de concentrado de cobre como producto principal y concentrado de molibdeno como producto secundario.

Sin embargo, las rocas que contienen los minerales de interés no son las únicas que se encuentran en los yacimientos donde se extraen, ya que también incluyen arcillas u otros materiales insolubles que interfieren en el proceso productivo del concentrado de cobre. Esto depende de la zona de extracción del yacimiento, y es un desafío que enfrenta Collahuasi en estos días, por la inestabilidad que genera el tipo de mineral que va ingresando a la planta concentradora y en consecuencia el cambio de los métodos a operar para una mejor control de producción.

Esta investigación propone la creación de modelos predictivos que ayuden a analizar el comportamiento de los insolubles en el concentrado final, esto a partir de parámetros operacionales que impacten al control de insolubles. El trabajo considera la recopilación de datos históricos de tres años y de seis meses, con parámetros operacionales que puedan ser de mayor o menor importancia al momento de analizarlos por softwares estadísticos. Luego se construirán modelos predictivos, con el fin de determinar el modelo que presente mejor desempeño para así observar qué variables son las más importantes y las que impacten al porcentaje de insolubles.

El predecir la cantidad de insolubles que se encuentren en el concentrado final de cobre es complejo, principalmente por la variabilidad de los minerales en los yacimientos, las condiciones operacionales respecto a lo que ingrese a la planta y del tipo de arcilla presente. La creación de modelos predictivos con *machine learning* (ML) ha tomado fuerza en estos tiempos gracias a la evolución de la inteligencia artificial y su implementación en el mundo minero. Con esto, gracias a ML, Collahuasi ha implementado modelos predictivos respecto a distintas variables objetivo, como el P80, y la creación de un modelo respecto al porcentaje de insolubles no solo ayudaría en el análisis de su comportamiento en tiempo real, sino también a entregar un apoyo efectivo a la hora de tomar decisiones operativas, entregando así resultados óptimos y generando estabilidad del proceso.

1.1 Objetivo General

- Proponer un modelo que permita estimar el porcentaje de insolubles presente en la concentración final de cobre a partir de variables operacionales del proceso en Compañía Minera Doña Inés de Collahuasi.

1.2 Objetivos Específicos

- Recopilar y depurar datos históricos operacionales de los últimos tres años, identificando las variables con mayor impacto sobre el porcentaje de insolubles mediante criterios estadísticos y fundamentos técnicos.
- Construir un modelo predictivo que relacione el comportamiento de los insolubles con las principales variables operacionales identificadas.
- Validar y evaluar el desempeño del modelo utilizando datos reales del proceso, verificando su capacidad de generalización y precisión.
- Ajustar y refinar el modelo utilizando datos operacionales recientes (últimos seis meses), con el fin de analizar la variabilidad de los insolubles y comparar los resultados con los obtenidos en la etapa inicial del modelo.

2 Antecedentes

2.1 Antecedentes de la planta

La Compañía Minera Doña Inés de Collahuasi, por su sigla CMDIC, se encuentra en la comuna de Pica, región de Tarapacá. Sus yacimientos se encuentran a más de 4400 m.s.n.m. y el campamento donde residen los trabajadores se encuentra a 3800 m.s.n.m. Es actualmente una de las seis productoras de cobre más grande del mundo y una de las cuatro principales productoras de molibdeno de Chile, además de que cuenta con un depósito de cobre de 10380 millones de toneladas (Collahuasi, 2025), siendo una de las más grandes del mundo. Cuenta con tres accionistas los cuales son Anglo American plc (44%), Glencore (44%) y Japan Collahuasi Resources B.V. (12%).

El objetivo de Collahuasi es obtener concentrado de cobre y molibdeno, este último como valor agregado, es por esto que de los minerales que se extraen se tratan inicialmente en Ujina, donde se encuentra la planta concentradora. El concentrado de cobre es transportado a través de un sistema de mineroductos que recorren 203 km desde la planta hacia el Terminal Marítimo Collahuasi, situado en Puerto Patache, a 65 km al sur de Iquique. Ahí se encuentra la planta de molibdeno y de filtrado de cobre, donde se separa el concentrado de cobre y del molibdeno por flotación selectiva. El molibdeno concentrado es filtrado y llevado a maxi sacos, mientras que el cobre filtrado es llevado a un stock pile, para que luego ambos productos sean embarcados y comercializados internacionalmente (Collahuasi, 2025). De manera esquemática, se aprecia en la Figura 2.1 el proceso en general, desde el inicio con la extracción hasta la exportación de los concentrados.

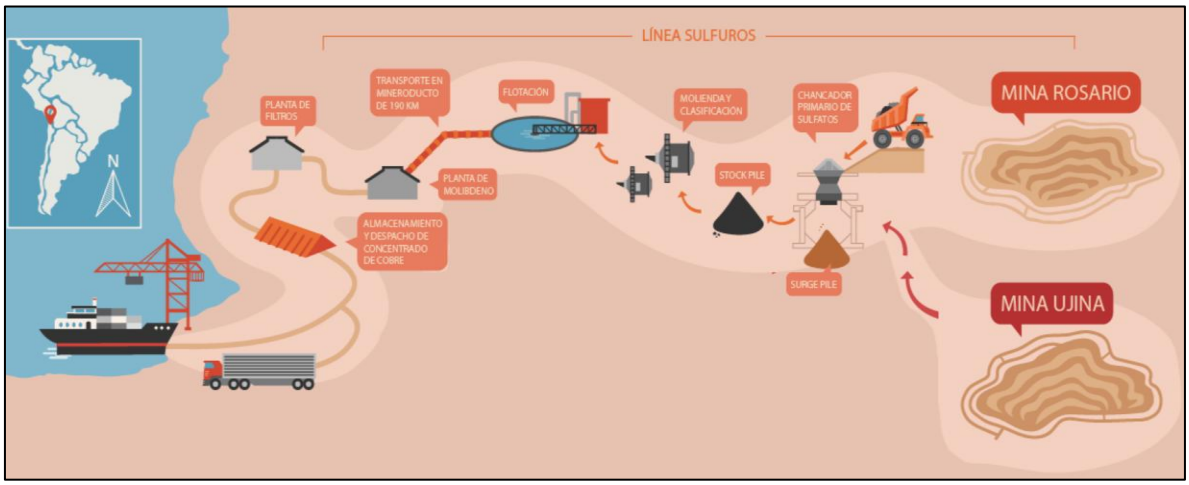

Figura 2.1 Diagrama de producción de línea de sulfuros CMDIC. (Fuente: Collahuasi.cl, 2025)

2.1.1 Chancado y Transporte

El circuito comienza en la etapa de chancado, en donde el fin de este proceso es reducir el tamaño de las grandes rocas provenientes de los yacimientos de Rosario y Ujina para tener una granulometría uniforme entre 6 a 8 pulgadas de diámetro y para poder liberar partículas de mineral, de manera que se logre exponer el metal de interés. Los equipos utilizados son chancadores de cono, uno ubicado en Ujina y otro en Rosario, los cuales reducen el tamaño a través de compresión gracias al movimiento excéntrico que se realiza, de manera que el mineral se tritura contra las paredes del cono.

El mineral reducido es llevado por medio de correas transportadoras hacia el *stock pile*, donde se almacena de manera temporal el producto de los chancadores, para luego ser llevado a la etapa de molienda, tal como se aprecia en la Figura 2.3. El stock pile tiene 8 chutes de descarga, donde cada uno va descargando a un *feeder* que saca el mineral desde la pila. Normalmente operan 4 de 8 chutes con el fin de asegurar el tonelaje de alimentación nominal al molino SAG y a la vez conseguir una distribución adecuada de mineral hablando granulométricamente (Wills & Finch, 2015). Al realizar esto, se busca un mezclado ideal, de manera que se reduzcan las fluctuaciones rápidas en las características del mineral y produce una operación más estable en la planta.

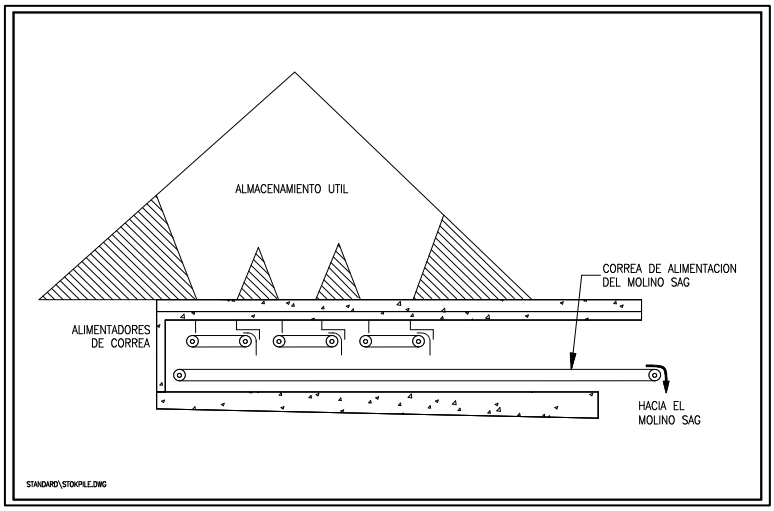

Figura 2.2 Pila de acopio de mineral grueso y alimentadores de correa (CMDIC, 2023)

2.1.2 Molienda

La segunda etapa de la planta concentradora es la de molienda. Lo que se busca es optimizar el proceso del chancado, con el fin de aumentar el grado de liberación de la partícula mineral y análogamente reducir el tamaño del mineral, pero que este tenga un tamaño granulométrico alrededor de 200 μm.

Actualmente la planta cuenta con 3 molinos SAG, 5 molinos de bolas (siendo el 5to molino puesto en marcha a finales del 2023) y 10 baterías de hidrociclones (BHC) en donde dos molinos de bolas son para el SAG 01 de la línea 1, dos para el SAG 02 de la línea 2, cuatro para el SAG 1011 de la línea 3 (por su mayor cantidad de tratamiento) y dos para el 5to molino, el cual recibe carga de los Pebbles y de transferencia de cada línea de un 8%, con el fin de poder moler las rocas con baja moliendabilidad y recibir la carga de distinta línea, con el fin de tener mayor capacidad de procesamiento. (CMDIC, 2025)

En el **Anexo 1: Dimensiones y tratamiento de molinos SAG y bolas.** se muestra las dimensiones de los equipos y su respectiva capacidad de tratamiento, demostrando que para el SAG 1011 se requieren más BHC por la gran capacidad de tratamiento que tiene el equipo.

El material que entra a los molinos se le añade agua, reactivos y se muele con distintos mecanismos, ya sea por la capacidad volumétrica del equipo y por la granulometría del mineral, donde en los SAG el mecanismo predominante es de impacto, porque se reduce el tamaño de las rocas a través de las bolas de acero y del mismo mineral. En los molinos hay harneros que separan la carga fina de la carga gruesa hay *trommels*, que separan la carga de la pulpa respecto de los Pebbles, mineral con granulometría que no se puede moler más pero tampoco que pasa a través del harnero, siendo dirigido al 5to molino. Luego dicha carga se dirige a un cajón de alimentación de BHC, para clasificar el tamaño granulométrico del mineral, el cual dentro de cada ciclón se separan las partículas finas que se dirigen a la zona superior del ciclón, o sea, por el *vortex Finder*. La carga que sale por esa dirección es llamada *overflow,* mientras que la carga que se dirige por la válvula inferior del ciclón, denominado como *apex*, es donde sale la carga con granulometría gruesa, conocido como *underflow* en las BHC, siendo necesario que se lleve a molienda secundaria para poder reducir más su tamaño y lograr un mayor grado de liberación (Concha & Bouso, 2021), tal como se aprecia en la Figura 2.4. En los molinos de bolas el mecanismo principal es el de abrasión. Lo que sale como producto es un mineral de un tamaño de 200 μm aproximadamente que se dirige al cajón de alimentación para llevar la carga fina a la etapa de flotación.

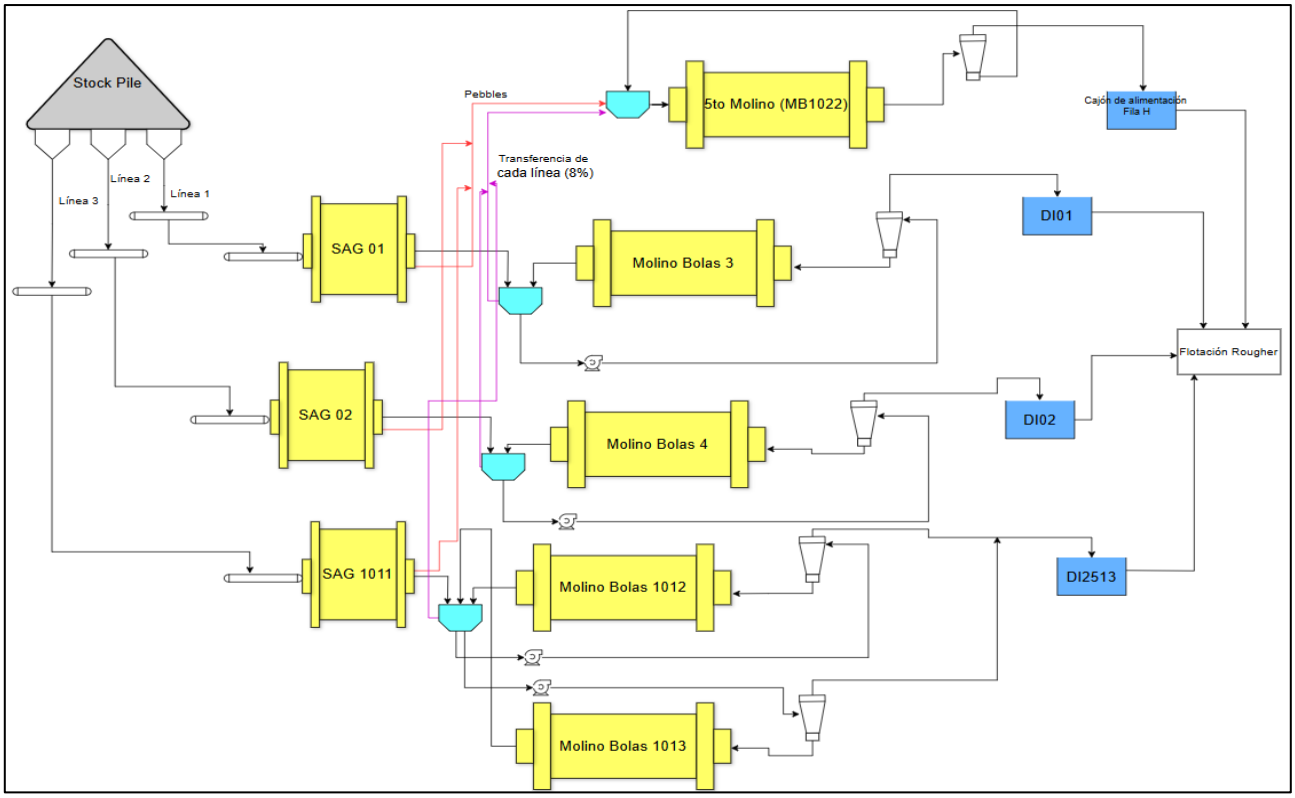

Figura 2.3 Diagrama de flujo de la etapa de molienda CMDIC (Elaboración propia, 2025)

2.1.3 Flotación

Luego de tener partículas de mineral con un bajo tamaño (200 – 250 µm) estas son dirigidas al proceso de flotación, el cual se basa en el principio fisicoquímico del tratamiento de la pulpa para darle propiedades hidrofóbicas a través de la adición de reactivos y generación de aire, para así lograr que se adhieran a una burbuja el mineral de interés (Wills & Finch, 2015), que en este caso es la de cobre y molibdeno. La planta cuenta con un arreglo de celdas de *tipo rougher,* limpieza (*cleaner*) y de barrido o *scavenger*, contando con 45 celdas Rougher para la línea 1 y 2, 45 celdas para el proceso *cleaner* y de barrido (contando con celdas que se pueden configurar como *cleaner* o barrido, dependiendo de la necesidad de la planta), y 24 celdas igualmente rougher, que tratan la carga proveniente de la línea 3. Mencionar también que a fines del 2024 empezó a operar una línea H, que opera como Rougher y recibe la carga del 5to molino.

En el **Anexo 2: Diagrama del proceso de flotación CMDIC** se detalla el diagrama de flujo para el proceso de flotación, donde la carga de la línea 1 y 2 se dirigen a la fila de celdas R1 a la R5, luego la carga concentrada se dirigen a un cajón de alimentación de BHC donde también se junta con la carga

de la línea 3, donde están la fila de celdas rougher D, E, F y G, para llevar la carga a la BHC que separa la carga fina de la gruesa, siendo la carga gruesa dirigida a los molinos verticales para liberar más el mineral y recircular en el cajón de alimentación, mientras que el material fino se dirige a un cajón de alimentación (cajón 1080) para alimentar la carga a las celdas *cleaner* y de barrido, las cuales son la 6, 7, A, B y C. Acá las filas de las celdas tenían configuradas ciertas celdas de tipo *cleaner*, otras de barrido y algunas celdas con configuración mixta, en donde depende de que si se busca aumentar la ley, se configuran como de limpieza, y si buscan recuperar, se configuran como de barrido. El concentrado de la carga del cleaner pasa a una segunda etapa *cleaner*, donde se encuentran las celdas columnares y la de los barridos pasa al cajón que alimenta la BHC. Los relaves tanto de las celdas *rougher*, *cleaner* y barridos se van a la canaleta de relaves que dirige la carga a los espesadores de relaves. (Observación directa, 2025)

La pulpa que se concentra en las celdas columnares se dirige a los espesadores de concentrado, mientras el relave generado es tratado en dos celdas Jameson, con el fin de concentrar lo más que se pueda y obtener cobre y molibdeno. La carga concentrada se lleva a los espesadores de concentrado y lo de descarte se devuelve al cajón de alimentación para las celdas de primera limpieza.

2.2 Antecedentes teóricos

2.2.1 Inteligencia Artificial

Se define en crear sistemas capaces de realizar tareas que normalmente requieren inteligencia humana, como razonar, aprender, resolver problemas y adaptarse a nuevas situaciones. Esta disciplina busca que las máquinas puedan tomar decisiones autónomas a partir de datos y patrones previos. Según Russell y Norvig (2020), la IA puede definirse como “el diseño de agentes inteligentes”, es decir, sistemas que perciben su entorno y toman acciones que maximizan sus posibilidades de éxito. La IA moderna abarca técnicas como el *machine learning*, visión computacional, procesamiento de lenguaje natural y robótica, y se aplica hoy en sectores tan variados como la medicina, finanzas, manufactura, minería, entre otras áreas más (Jung & Choi, 2021).

2.2.2 Machine Learning

Es una herramienta de la inteligencia artificial que permite que los computadores aprendan de los datos, sin que sea necesario programar cada paso que deben seguir. La idea es que, al reconocer patrones y adaptarse a la información que reciben, puedan mejorar su rendimiento por sí solos. Este concepto fue planteado hace décadas por Arthur Samuel (1959), quien definió el aprendizaje

automático como la capacidad de las máquinas para aprender sin estar explícitamente programadas. Hoy en día, *machine learning* se utiliza en casi todos los sectores: desde la medicina y la minería, hasta la industria financiera, porque permite generar modelos predictivos y de clasificación con gran precisión.

A diferencia de los enfoques tradicionales de programación, en los que se debe definir manualmente un conjunto de reglas o parámetros para resolver un problema, el aprendizaje automático permite que estas reglas se descubran directamente a partir de los datos. Tal como se puede ver en la Figura 2.5, cuando se aborda un problema de manera tradicional, normalmente se establecen instrucciones fijas y luego se evalúa el resultado para verificar si es correcto o requiere ajustes (Gerón, 2019). Este método no considera información extraída de datos reales, lo que implica que puede no adaptarse bien a situaciones complejas. El enfoque de *machine learning*, en cambio, busca justamente aprovechar esa data muestral para generar soluciones más flexibles, precisas y ajustadas al contexto.

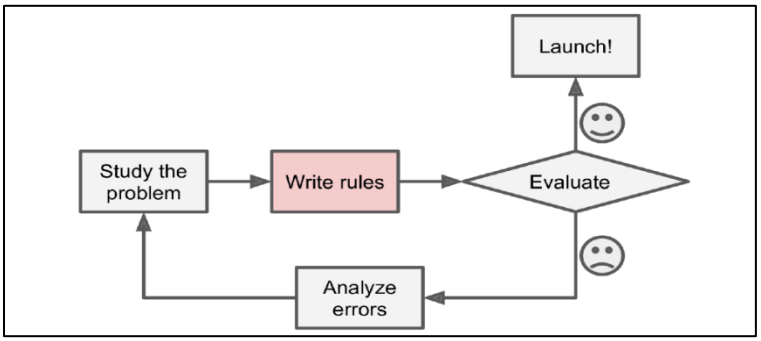

Figura 2.4 Esquema de evaluación de un problema tradicional (Gerón, 2019)

Por otro lado, en la Figura 2.6 definida por Gerón (2019) se aprecia un diagrama el cual se incluye el entrenamiento del algoritmo de las máquinas de aprendizaje, donde también se incluyen los datos recopilados, con el fin de que la máquina pueda hacer una evaluación con rangos de datos ya definidos o con data histórica de la situación y así tener un mejor rendimiento.

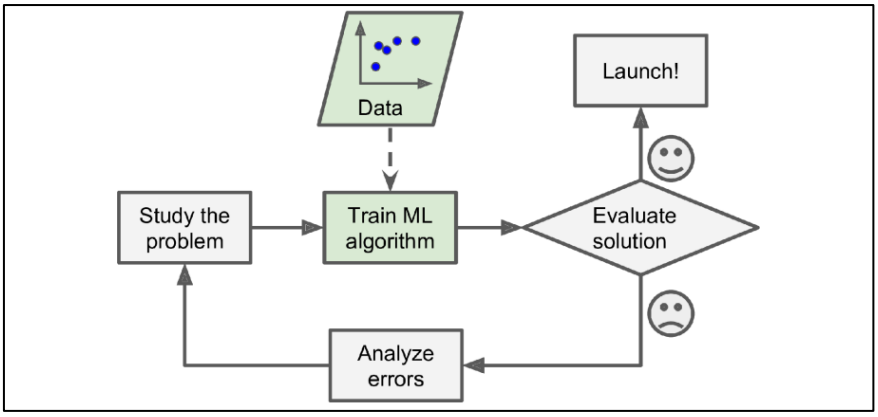

Figura 2.5 Enfoque de aprendizaje automático (Gerón, 2019)

2.2.3 Modelo de árboles

Este modelo predictivo funciona tomando decisiones a través de la división de los datos en función de las variables que mejor reduce los errores. Construyen predicciones a partir de muchos árboles simples, donde cada árbol busca corregir los errores realizados por los árboles anteriores. En este proceso se seleccionan ramas y divisiones que capturan patrones complejos en los datos, lo que permite obtener predicciones más precisas sin requerir transformaciones lineales previas (Rizkallah, 2025).

En este trabajo se busca emplear Random Forest, Gradient Boosting y XGBoost, dado que se encuentran entre los algoritmos más robustos para tareas supervisadas, presentando una alta capacidad de interpretación frente a grandes volúmenes de datos y ofrecen gran adaptabilidad frente a relaciones no lineales y de elevada complejidad (Imani et al., 2025).

2.2.4 Random Forest

Es un algoritmo de aprendizaje supervisado que extiende los árboles de decisión: crea un conjunto de árboles usando muestras *bootstrap* (método que consiste en crear copias del set de datos original tomando datos al azar y permitiendo repeticiones**)** y seleccionando al azar subconjuntos de variables para cada división, y la predicción final se obtiene mediante clasificación o regresión, lo que le otorga robustez ante sobreajuste y estabilidad en presencia de ruido (Schönlau & Zou, 2020).

2.2.5 Gradient Boosting

Gradient Boosting es una técnica que construye modelos de forma progresiva: comienza con un árbol sencillo y, en cada paso, añade uno nuevo que corrige los errores cometidos por los anteriores. De esta manera, cada árbol se enfoca en lo que el modelo anterior no logró predecir bien. Este método permite

capturar relaciones complejas entre las variables y mejorar la precisión en clasificación y regresión (Hein et al., 2019).

2.2.6 XGBoost

XGBoost es una versión mejorada de Gradient Boosting que ha sido optimizada para trabajar con grandes volúmenes de datos de manera rápida y eficiente. Además de construir árboles de manera secuencial, incluye mecanismos de regularización que ayudan a evitar el sobreajuste y aprovechar mejor la información. Esto lo convierte en uno de los algoritmos más utilizados hoy en día para tareas de predicción con datos complejos (Chen & Guestrin, 2016).

2.2.7 Máquinas de vectores supervisados

La Regresión de Vectores de Soporte (SVR) es una técnica que busca encontrar una función que logre aproximarse a los datos lo mejor posible, permitiendo cierto margen de error. A diferencia de otros métodos que intentan predecir cada punto con exactitud, el SVR se concentra en mantener la predicción dentro de un rango aceptable alrededor de los valores reales. Esto permite que el modelo sea más robusto frente a datos ruidosos. Además, gracias al uso de funciones kernel, se pueden capturar relaciones no lineales complejas entre las variables (Vapnik, 1995).

2.2.8 Redes Neuronales

Son un tipo de modelo que se basa en el funcionamiento de las neuronas del cerebro humano, donde múltiples capas de neuronas del modelo procesan la información de manera jerárquica. Debido a la gran cantidad de datos que se analiza, se sugiere el uso de Redes Neuronales Profundas (DNN) ya que, según Goodfellow et al. (2016) permiten extraer representaciones útiles de los datos de forma automática, siendo óptimo su rendimiento frente a métodos tradicionales, además de que puede identificar patrones no lineales y características ocultas en ciertos datos. Por otro lado, Las Redes Neuronales Artificiales (ANN) son modelos predictivos que pueden funcionar bien incluso con muestras de datos no muy abundantes, siempre que se controlen adecuadamente los hiperparámetros, y en esas condiciones tienden a evitar el sobreajuste (Wang, 2005).

Para conocer qué modelo predictivo es óptimo, se realiza una comparación estadística de acuerdo con los valores que entreguen las métricas de desempeño que ayudan a determinar el modelo que mejor responde frente al set de datos recopilados.

2.2.9 Coeficiente de correlación lineal (R)

Es una variable estadística que mide qué tanta relación existe entre dos variables, que en este caso sería la variable objetivo y cada variable independiente. La cuantificación de esta variable se expresa en un rango de -1 a 1. Un valor de -1 indica una correlación negativa perfecta, lo que significa que, al aumentar una variable, la otra disminuye proporcionalmente, y viceversa. Un valor de 1 representa una correlación positiva perfecta, donde ambas variables aumentan o disminuyen en la misma dirección. Por último, un valor de 0 indica que no existe correlación entre las dos variables, es decir, no hay una relación lineal entre ellas.

La ecuación 3.3 describe la obtención del valor R definida por Pearson en 1895, donde x_i y y_i, son los valores de las dos variables y $\bar{x}$ con $\bar{y}$ son sus respectivas medias (Stanton, 2001).

$$R = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 * \sum(y_i - \bar{y})^2}} \tag{3.3}$$

2.2.10 Coeficiente de determinación (R^2)

Es una medida estadística que indica qué tan cerca se encuentran los datos de la línea ideal de referencia. El valor del R^2 determina la proporción de la varianza total de los datos que se explica en base al modelo, siendo este un valor entre 0 y 1, donde 0 explica una variación de datos que no es representado de manera correcta, por lo que habrá una muestra dispersamente predicha y los valores predichos no serán los mismos que los valores experimentales. En cambio, cuando su valor es cercano a 1, indica que el modelo explica de manera correcta los datos, existiendo una relación entre los valores predichos y los valores experimentales (Wooldridge, 2016).

Sin embargo, un coeficiente de determinación alto no garantiza que el modelo sea bueno, especialmente si se obtiene en el conjunto de entrenamiento. Un valor muy alto de R^2 puede ser indicativo de *overfitting*, o sobreajuste, donde el modelo no solo aprende los patrones reales de los datos, sino también el ruido, lo cual deteriora su capacidad de generalización a nuevos datos (James et al., 2021). Para el modelo se busca un valor de R^2 de validación que se encuentre entre 0,7 a 0,8.

2.2.11 Raíz del error cuadrático medio (RMSE)

Esta métrica de evaluación indica la diferencia promedio entre los valores reales y los valores predichos por el modelo, tal como indica la ecuación 3.4, donde y_i es el valor real, $\hat{y}_i$ es el valor predicho y n es el número de observaciones.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3.4)$$

El objetivo es medir la precisión del modelo en términos absolutos, con el fin de conocer qué tan preciso es el modelo y qué tan equivocado está en comparación a los otros modelos. El RMSE es probablemente la métrica de error más utilizada para los modelos de regresión, ya que otorga una ponderación relativamente alta a los errores grandes y es fácil de interpretar, ya que tiene las mismas unidades que la variable predicha (Chai y Draxler ,2014). Un RMSE bajo indica que los valores predichos obtenidos están más cerca de los valores reales, siendo más precisa la predicción del modelo.

2.2.12 Error absoluto medio (MAE)

Indica cuanto es el promedio del valor absoluto de la diferencia entre lo predicho del modelo y de los valores observados, dando como resultado un valor en la misma unidad que la variable objetivo. En la ecuación 3.5 se indica la descripción matemática del modelo, donde y_i es el valor real, $\hat{y}_i$ es el valor predicho y n es el número de observaciones.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (3.5)$$

Al igual que el RMSE, un valor bajo de MAE indica que las predicciones del modelo están más cerca de los valores observados, sin embargo, es más preciso a la hora de evaluar errores del modelo, ya que todos los errores tienen la misma ponderación, haciéndolo más robusto ante valores atípicos y sin penalizar grandes desviaciones. Por otro lado, al tener un error absoluto medio alto es más probable que el modelo cometerá errores a la hora de predecir, lo cual es importante buscar un valor bajo de MAE entre los modelos.

Según Willmott y Matsuura (2005), el error absoluto medio es posiblemente la medida más natural de la magnitud del error promedio. Considera todos los errores con el mismo peso y es menos sensible a los valores atípicos que el RMSE.

2.2.13 Error absoluto medio porcentual (MAPE)

Esta última variable métrica mide el error promedio en porcentaje entre los valores predichos y valores observados. La ecuación 3.6 describe el MAPE, siendo y_i el valor real, $\hat{y}_i$ el valor predicho y n es el número de observaciones, además de multiplicar por 100% para dejarlo expresado en porcentaje la métrica.

$$MAPE = \frac{(100\%)}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (3.6)$$

El análisis de los valores de MAPE son sencillos, ya que expresa los errores como porcentaje, lo que facilita la interpretación. (Armstrong y Collopy, 1992). Esta métrica refleja qué tan alejados están, en promedio, los valores predichos respecto a los valores observados o valores experimentales. Sin embargo, la desventaja que presenta es que no es robusta cuando los valores reales son cercanos a cero, siendo relevante sólo para datos a escala de razón (Armstrong y Collopy, 1992).

De igual forma con las métricas anteriores, valores bajos de MAPE presentan errores bajos siendo más preciso el modelo predictivo, pero con valores altos indica que hay errores que pueden generar variabilidad entre los valores predichos y los datos experimentales.

Lo que se busca en el conjunto de validación, tanto para RMSE, MAE y MAPE es obtener valores bajos de error, ya que indican una mejor precisión en los modelos a la hora de su predicción y no genera gran dispersión de datos predichos respecto a los evaluados.

2.2.14 Variables Operacionales

Existen diversas variables operacionales que influyen de manera directa o indirecta en el control de insolubles. Su adecuada gestión permite minimizar el impacto de las arcillas y asegurar un procesamiento eficiente del mineral, lo cual es fundamental para cumplir con las metas de producción de cobre y, al mismo tiempo, diluir la ganga en el proceso de flotación. Acá se definen las variables que impactan en gran medida la variación de insolubles presentes en el proceso.

2.2.15 Mineralogía

La mineralogía de las arcillas es la rama de la mineralogía que estudia la composición, estructura, propiedades y abundancia relativa de los minerales del grupo de las arcillas presentes en una muestra, utilizando técnicas como la difracción de rayos X (DRX/XRD), microscopía electrónica y análisis térmico (Grim, 1968). Esto se realiza con el fin de determinar los tipos de arcillas, donde son los que ingresan a la planta concentradora y se considera su respectiva cantidad.

En CMDIC se analizan el porcentaje de arcillas presentes en la roca a través de cortadores metalúrgicos, siendo estas las siguientes:

- Pirofilita: Corresponde a un silicato de aluminio hidratado con estructura laminar. En el proceso de flotación, genera un aumento de la viscosidad de la pulpa, lo que favorece que las partículas insolubles se adhieran a las burbujas de aire y se dirijan al concentrado. Esto impacta

negativamente en la ley de cobre, dado que arrastra ganga y disminuye la selectividad del proceso (Cruz et al., 2015).

- Caolinita: Es un silicato de aluminio con estructura en capas y partículas muy finas. Una de sus principales características es el efecto de *slime coating*, que consiste en recubrir los minerales valiosos con lamas, reduciendo su hidrofobicidad y dificultando la acción de los reactivos colectores. Esto ocasiona que aumente la recuperación de insolubles en el concentrado, reduciendo la eficiencia de separación (Bulatovic, 2007).
- Ilita: Corresponde a una mica de grano fino con alto contenido de potasio. Su presencia genera un aumento de la viscosidad en la pulpa y una mayor estabilidad en la espuma, lo que produce arrastre mecánico de partículas de ganga hacia el concentrado. Además, interfiere con los reactivos de flotación, disminuyendo la efectividad de la recuperación de cobre (Ma et al., 2009).
- Moscovita: Es una mica dioctaédrica con estructura laminar muy estable. Al encontrarse en forma de partículas muy finas, aporta a la formación de lamas que afectan la selectividad de la flotación. Esto provoca que los insolubles se recuperen junto con el mineral valioso, reduciendo tanto la ley como la recuperación en el concentrado final (Cruz et al., 2015).

El control de arcillas en flotación se ha abordado mediante distintas estrategias. El uso de polímeros orgánicos como la goma guar permite flocular las partículas de arcillas y disminuir su efecto negativo sobre la recuperación (Jeldres Valenzuela, 2017). Por otro lado, los dispersantes aniónicos evitan la adhesión de arcillas a minerales valiosos, mejorando la selectividad del proceso, como se demostró en la mina Telfer en Australia (Seaman et al., 2012). Finalmente, el ajuste de pH con reactivos como la cal ayuda a modificar la estabilidad coloidal de minerales como la caolinita, reduciendo su interferencia en la flotación (Cruz et al., 2015). A pesar de los estudios para el tratamiento de las arcillas, la creación de un modelo predictivo complementaría con anticipación el control de estas frente a distintos escenarios y lograr un concentrado de cobre óptimo.

2.2.16 Celdas Columnares

Estas son celdas que se usan principalmente en el proceso de flotación, específicamente en la etapa de limpieza o *cleaner.* Las celdas columnares forman parte del desarrollo tecnológico de la flotación y se diferencian de las celdas mecánicas tradicionales por su diseño y mecanismo de aireación. Estos equipos están regidos por los principios fisicoquímicos que gobiernan el proceso de flotación:

interacción partícula–burbuja, formación y estabilidad de la espuma, y selectividad frente a la ganga. En este contexto, la presencia de arcillas como la pirofilita, caolinita, ilita y moscovita adquiere especial relevancia, ya que estas partículas finas tienden a estabilizar excesivamente la espuma o generar viscosidad en la pulpa, afectando el transporte y la selectividad en las columnas. Por esta razón, al abordar el uso de celdas columnares, resulta indispensable discutir cómo estos equipos enfrentan el desafío que imponen las arcillas en la flotación y qué ventajas operacionales presentan para mitigar el arrastre de insolubles.

La particularidad de estas celdas es que no necesitan un impulsor mecánico, sino un sistema de aireación que permita generar las burbujas y permita generar una zona de interfase, para poder extraer el mineral concentrado por la zona superior y separar la ganga a través de las colas por gravedad descendientes (Wills & Finch, 2015). Dichos equipos son la última línea de la planta concentradora para llevar el concentrado a las etapas posteriores y tener una pulpa con alta ley de cobre, por lo que es fundamental que estas celdas disuelvan lo mayor posible los insolubles presentes y así cumplir con el objetivo de producción minera.

2.2.17 Agua de lavado

Las celdas columnares tienen distintas zonas, una que es zona de limpieza y otra zona de colección. El agua de lavado crea una zona de limpieza en contraflujo que elimina impurezas sin afectar el transporte de las burbujas (Finch y Dobby, 1990), de manera que favorece la selectividad en el proceso y flote sólo el mineral de interés. La ventaja de esta variable operacional es que permite obtener concentrados de alta ley, a pesar de que haya presencia de minerales arcillosos y acomplejen el proceso, sin embargo, es fundamental tener un buen control de flujo de agua de lavado, con el fin de que cumpla su proceso y no permita que los insolubles se vayan con el mineral de interés.

2.2.18 Nivel de interfase

El nivel de interfase, que corresponde a la altura donde se separa la pulpa de la espuma, constituye un parámetro crítico para la operación. Mantenerlo estable asegura que las burbujas cargadas de mineral puedan estabilizarse antes de rebosar como concentrado. Un nivel demasiado alto puede provocar arrastre de ganga, mientras que un nivel muy bajo reduce la recuperación (Wills & Finch, 2015). En CMDIC, este control se realiza mediante sensores de presión diferencial, ajustando en tiempo real según el comportamiento de la pulpa.

2.2.19 Velocidad de rebose

La velocidad de rebose es el caudal con que el concentrado sale por la parte superior de la celda columnar. Esta velocidad está relacionada con el área superficial de la columna y el flujo de aire y de pulpa. Un rebose muy alto puede generar inestabilidad en la espuma y pérdida de mineral, mientras que uno muy bajo puede significar baja recuperación. Por eso es ideal tenerlo balanceado junto con el flujo de aire y el nivel de interfase (Wills & Finch, 2015).

2.2.20 Control de pH

El control de pH cumple un rol esencial en la flotación, ya que está directamente vinculado con la fundamentación fisicoquímica del proceso. El pH regula la carga superficial de los minerales y de las arcillas presentes en la pulpa, lo que determina la interacción de estas partículas con los reactivos colectores y depresores. A valores inadecuados, las arcillas mantienen una elevada dispersión, lo que incrementa su arrastre hacia la espuma y eleva la presencia de insolubles en el concentrado de cobre (Castro et al., 2022).

Desde la perspectiva reactiva, la adición de cal no solo regula el pH, sino que también cumple la función de deprimir minerales no deseados como la pirita. Sin embargo, un exceso de cal genera efectos contraproducentes, como la dispersión de arcillas y el aumento del consumo de reactivos, lo cual compromete la selectividad del proceso (Cisternas et al., 2019). Es por esto que el pH se convierte en una variable de equilibrio que debe mantenerse en rangos óptimos para favorecer la hidrofobicidad de la calcopirita y minimizar el arrastre de ganga.

Por lo tanto, el control de pH no solo se relaciona con un ajuste operacional, sino que constituye un factor clave en la estabilidad fisicoquímica del sistema de flotación. Su adecuada gestión permite mejorar la calidad del concentrado, reducir la influencia de los insolubles y anticipar comportamientos no deseados en la operación mediante herramientas predictivas.

En CMDIC el porcentaje de insoluble ha estado inestable, lo cual perjudica en el proceso de flotación, debido a que, al haber gran cantidad de insolubles, sus propiedades impactan de manera significativa. Por ejemplo, la pirofilita es un tipo de arcilla que genera viscosidad en las celdas de flotación, lo cual permite que los insolubles se adhieran a la burbuja y se dirijan a la canaleta de concentrado, afectando directamente al concentrado final de cobre. Es por esto que el pH en las celdas de flotación debe estar en un rango óptimo dependiendo de la cantidad de arcilla que va ingresando al proceso, priorizando la concentración de cobre y evitar que los insolubles sean arrastrados.

3 Desarrollo experimental

Inicialmente se analiza el problema de los insolubles, las variables que afectan al porcentaje del problema mencionado y la data histórica para observar el comportamiento del porcentaje de insolubles a lo largo del tiempo. Para esto, se incluye el uso de *machine learning* con *Python* e inteligencia artificial para crear modelos predictivos y de *softwares* que ayuden a realizar un buen análisis estadístico con el fin de obtener resultados efectivos.

La creación de un modelo predictivo para determinar el insoluble presente en el concentrado final de cobre es una manera que mezcla la estadística de data presente con programación, con el fin de poder llegar a un análisis del comportamiento de los insolubles frente a ciertas variables. Para esto, se hace una recopilación de datos históricos considerando tres años de antigüedad (desde el 10 de febrero del 2022 al 08 de febrero del 2025), seleccionando variables teóricas que impactan directa e indirectamente a los insolubles, como la presión de BHC, el pH, las variables operacionales de las celdas columnares (agua de lavado, nivel de interfase y velocidad de rebose) y mineralogía de las arcillas, donde son factores claves que permiten un mejor control al porcentaje de insolubles o bien, son los que tienen un mayor impacto a la hora de evaluar la cantidad de arcilla presente en el proceso de concentración en la planta. Claro está de que existe un sinfín de parámetros operacionales que coinciden con la descripción previa, sin embargo, su impacto es muy bajo o existen parámetros que aportan más información operacional. Adicionalmente, se considera la relevancia estadística de cada variable, buscando un equilibrio entre el respaldo teórico y la capacidad predictiva. Es decir, no solo se incorporaron variables mencionadas en la literatura especializada, sino también aquellas que en la práctica operativa demuestran correlaciones significativas con los niveles de insolubles. Esto permite reducir la probabilidad de introducir variables irrelevantes o redundantes, mejorando la robustez del modelo predictivo.

Luego se recopila data histórica de 6 meses de antigüedad (desde el 01 de diciembre del 2024 al 30 de mayo de 2025) con el fin de comparar los modelos de distinto rango de fechas y analizar su comportamiento respecto al ámbito operacional de los equipos y de la variación de insolubles, dado que la manera de operar cambia a través de los años, se busca una óptima operación con los equipos presentes y con el tipo de mineral que ingresa a la planta concentradora.

Tal como se ve en la Figura 3.1, se diseña un diagrama de flujo con la metodología del presente trabajo, en donde el enfoque principal es la construcción de los modelos a partir de datos de operación de Collahuasi.

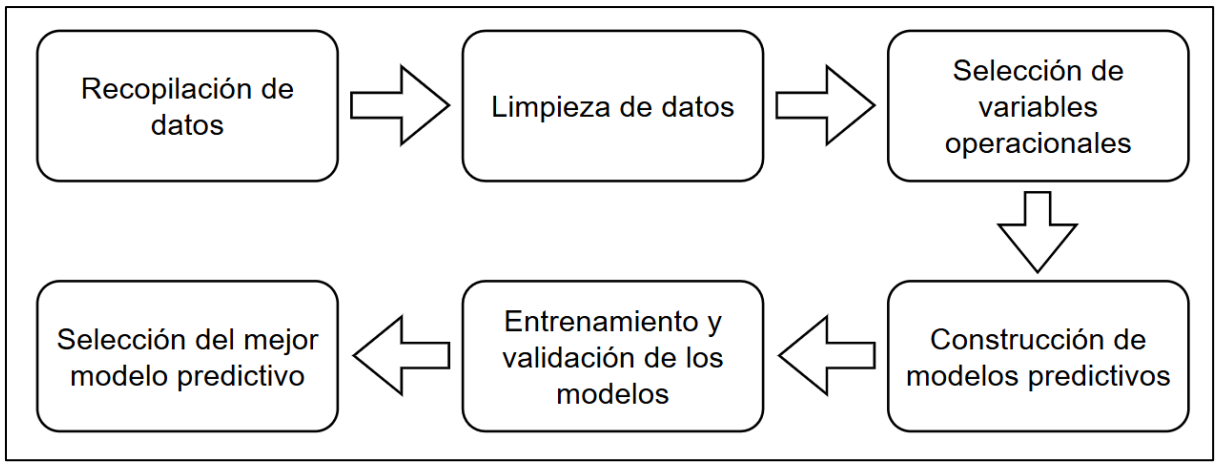

Figura 3.1. Metodología del trabajo para la construcción del modelo predictivo

La extracción de datos se realiza a partir de sábanas de datos, las cuales corresponden a planillas en formato Excel que se completan diariamente con información operacional de los equipos. Posteriormente, estas planillas se consolidan y almacenan de forma mensual para su análisis.

Se consideran dichos rangos de fecha para poder observar la evolución del porcentaje de insolubles a través de ese tiempo y determinar los momentos en que los rangos de insolubles se encuentran fuera del rango óptimo, donde se toma en cuenta un rango óptimo de insolubles menor a 8%.

Teniendo recopiladas las variables con las muestras de datos se procede a realizar una limpieza de datos en Excel. Esta limpieza consta de eliminar muestras de fechas que puedan interferir de manera negativa al modelo predictivo, siendo principalmente las fechas que tienen valor igual a 0, valores atípicos o anormales que puedan generar ruido.

Para eliminar los datos atípicos se utiliza el diagrama de caja y bigotes, desarrollado por Tukey, el cual resume la distribución de datos mediante los cuartiles, el rango intercuartílico (diferencia entre el 1er cuartil (Q1) y el 3er cuartil (Q3)) y los mínimos y máximos, siendo posible la identificación de dichos datos a través de los límites inferiores (L.I.) y superiores (L.S.) (Schwertman et al., 2004), tal como se aprecia en las ecuaciones 3.1 y 3.2, donde IQR viene siendo el rango intercuartílico.

$$L.I. = Q1 - 1{,}5 * IQR \tag{3.1}$$

$$L.S. = Q3 + 1{,}5 * IQR \tag{3.2}$$

Cuando se analiza un conjunto de valores, se pueden ver algunos datos alejarse mucho más allá de los demás (Tukey, 1977). Esto se puede observar cuando los datos no se encuentran entre el límite inferior

y límite superior, los cuales pueden generar incongruencias o ruidos en los valores obtenidos respecto al modelo.

Luego de recopilar los datos y limpiarlos, se analizan en el *software* llamado Minitab, el cual es una herramienta estadística que analiza datos tantos actuales como históricos, con el fin de encontrar variables estadísticas que permitan realizar un análisis óptimo para la selección de variables operacionales en el modelo. Los datos que se encuentran en Excel se exportan a dicho *software* y se ajusta al modelo de regresión, esto con el fin de determinar su valor p.

El valor p es una medida estadística que analiza la probabilidad de obtener un efecto igual o más extremo que el observado, considerando que la hipótesis nula es verdadera. Esto quiere decir que evalúa qué tan probable será el resultado obtenido si no hay alguna diferencia real entre las variables o los resultados se deban al azar (Biau et al., 2009).

Ronald Fisher fue quien introdujo el concepto de valor p, determinando que si se encontraba en un rango de 0,1 a 0,9 lo más probable es que las variables obtenidas respecto a las evaluadas se hayan determinado de manera aleatoria. Sin embargo, si este se encuentra menor a 0,05 es evidencia suficiente para cuestionar la hipótesis nula, dado que es conveniente tomar el 5% como un nivel estándar de significancia (Fisher, 1925).

Luego de obtener los valores p de cada variable, se procede a seleccionar las que tienen un valor menor a 0,05, las cuales vienen siendo las variables estadísticamente significantes, tal como se aprecia en el **Anexo 3: Variables con valores p menor a 0,05 en data de 3 años** para los 3 años de análisis.

Posterior a eso, se analizan las variables que más impactan a la generación de insolubles, independiente de su valor p, de manera que estén consideradas en el modelo y entreguen un mejor coeficiente de determinación, para así, dentro del conjunto de las variables, seleccionar las que más aportan estadística y teóricamente al modelo predictivo. En el **Anexo 4: Variables seleccionadas con valores p mayores a 0,05 en data de 3 años** indica las variables seleccionadas que son llevadas a análisis y que tienen un valor p mayor a 0,05.

Teniendo las variables listas, se comienza a construir los modelos predictivos, donde se consideraron las variables con valor p menor a 0,05 y ciertas variables con valor p mayor a 0,05 que, de manera teórica y práctica, afectan directamente al porcentaje de insolubles. Dichos modelos son supervisados, dado que el modelo aprende de la variable objetivo, que en este caso es el porcentaje de insolubles y de los datos de entrada, que son las variables que afectan a los insolubles. El fin de esto es predecir

los valores de insolubles presentes en el concentrado final a partir de los datos de entrada, siendo esta una caja negra, tal como se aprecia en la Figura 3.2

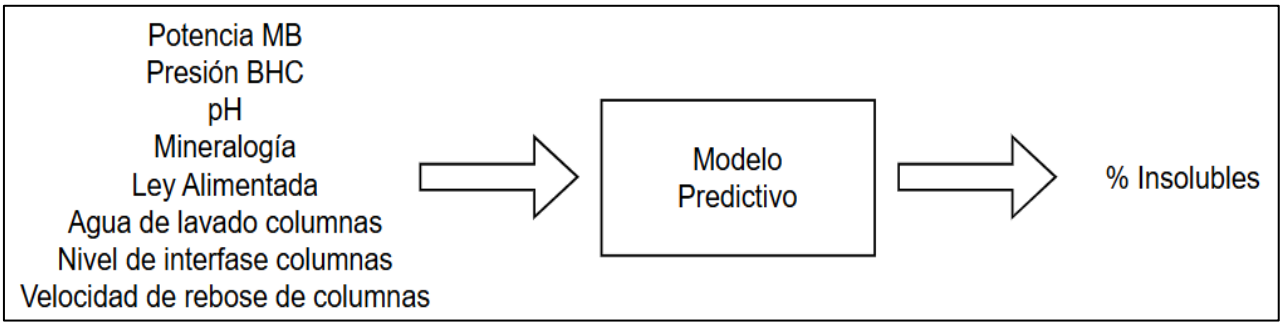

Figura 3.2 Caja negra del modelo predictivo para el análisis de insolubles

Los modelos que se están usando son en base a la cantidad de análisis de datos que deben de procesar y por la capacidad de interpretación, esto con el fin de encontrar un modelo que se ajuste al problema de la planta en CMDIC. Dichos modelos son de tipo árboles, de redes neuronales y de máquina de vectores de soporte, los cuales deciden las mejores opciones entre el desglose de la muestra de datos.

En cuanto al análisis de datos se sugiere subdividirse en 70% para el set de entrenamiento, ya que debe de recopilar una muestra grande para poder tener suficiente información al momento de aprender patrones representativos y se ajusten los parámetros internos del modelo, 15% para el set de validación para realizar ajustes a los hiperparámetros del modelo con el fin de tener mejores resultados estadísticos pero sin llevarlos a un sobreajuste y 15% para el set de pruebas, esto para evaluar el rendimiento del modelo. Los datos se dividen para evaluar correctamente el modelo sin evitar la generalización, ya que según Gerón (2019) esta práctica es esencial para evitar que el modelo realice un *overfitting*, o sea, que memorice los datos de entrenamiento y afecte al enfrentarse al mundo real.

Para cada modelo se busca obtener un coeficiente de determinación en validación (R^2) cercano a 1, pero no igual a 1, con el fin de evitar sobreajuste con la lectura de los datos. Para lograr esto, se debe de tener los datos limpios previamente y ajustar hiper parámetros que permitan tener un R^2 de validación cercano al valor deseado y que no genere sobreajuste en entrenamiento.

Para determinar el modelo predictivo que entrega un mejor análisis estadístico se construye una tabla comparativa de cada modelo en base a las métricas de desempeño, las cuales ayudan a medir la precisión y calidad de dichos modelos. Finalmente se escoge el modelo que entregue mejores resultados para así tener un mejor desempeño en cuanto al análisis de las variables tanto de entrada y de salida.

4 Resultados y discusiones

4.1 Análisis de 3 años

Inicialmente se analizan los modelos obtenidos de la data histórica de 3 años, en donde se compara cada uno con mejor desempeño tanto gráfica y estadísticamente a través de la métrica de desempeño de cada modelo. En la Tabla 4.1 se observan los resultados obtenidos de cada modelo a partir de sus métricas de desempeño en el conjunto de validación.

Tabla 4.1 Valores de métricas de desempeño en el conjunto de validación para cada modelo en 3 años

Modelo	RMSE	R^2	MAE	MAPE (%)
XGBoost	**1,20**	**0,77**	**0,94**	**12,90**
Random Forest	1,36	0,75	1,13	13,65
Gradient Boosting	1,35	0,75	1,09	13,28
SVR	1,44	0,72	1,19	14,62
DNN	1,50	0,69	1,24	15,33

El que presenta mejores resultados métricos es el modelo de *XGBoost,* ya que el valor de error de cada métrica (RMSE, MAE, MAPE) frente a otros modelos es más bajo, por ende, tiende a ajustar de mejor manera a los datos que se recolectaron y tiene una alta precisión en su respectiva predicción, además de tener un coeficiente de determinación mayor que los otros modelos, indicando que tiene menor variabilidad en los datos predichos respecto a los datos obtenidos.

Al comparar las gráficas de set de validación de cada modelo en las Figuras (4.1, 4.2, 4.3, 4.4 y 4.5), se aprecia que para *XGBoost* la línea de predicción se acerca a los puntos de datos experimentales de los insolubles, verificándose la precisión que tiene el modelo para predecir porcentajes de insolubles, no así como en el caso del modelo de Redes Neuronales Profundas (DNN), donde la dispersión de datos es levemente evidente respecto de la experimental en las respectivas gráficas de set de datos, lo cual no presenta una precisión adecuada presentado tanto gráfica como estadísticamente en la tabla de análisis de métrica.

Por otro lado, *Gradient Boosting* fue el segundo modelo con mejor equilibrio entre precisión y ajuste, destacando con un R^2 de 0,75 al igual que *Random Forest* que presenta un mismo coeficiente de

determinación, sin embargo, los errores de las métricas que presenta *Gradient Boosting* es levemente menor que *Random Forest,* lo cual quiere decir que dicho modelo comete menos errores a la hora de predecir en comparación al otro modelo. En cuanto a SVR, este modelo presentó resultados en R^2 (0,72) similares a los modelos previos, pero con RMSE, MAE y MAPE mayores que los modelos de tipo árboles, por ende su función predictiva no será precisa y cometerá errores a la hora de evaluarlo.

De manera resumida en la Tabla 4.2 se detallan descripciones visuales de los gráficos obtenidos en donde los modelos con mejores resultados presentan una mejor respuesta a la predicción, no así como los modelos que tienen un coeficiente de determinación más bajo que presentan mayor dispersión de datos.

Tabla 4.2 Descripción visual de los gráficos de distintos modelos en data de 3 años

Modelo	**Descripción visual figuras (a)**	**Descripción visual figuras (b)**
XGBoost ($R^2 = 0,77$)	Mejor alineación a la recta ideal respecto a las otras gráficas. Mayor precisión en rangos de insolubles entre 6 a 12%.	Las predicciones siguen bien la tendencia día a día, con poco atraso y buena captura de la variabilidad.
Random Forest ($R^2 = 0,75$)	Ajuste cercano, aunque con mayor dispersión en valores altos. Mejor precisión de insolubles entre 6 a 8%.	En el seguimiento diario reproduce la tendencia general, pero suaviza los cambios bruscos y tiende a subestimar los valores extremos.
Gradient Boosting ($R^2 = 0,75$)	Alineación aceptable, con ligera subestimación en valores altos.	En la validación diaria sigue la tendencia, aunque con cierto desfase y menor detalle en variaciones rápidas.
SVR ($R^2 = 0,72$)	Mayor dispersión en valores bajos y alejados de la línea ideal.	En el seguimiento diario la predicción es más rígida, con dificultad para reflejar cambios bruscos y variaciones más marcadas.
DNN ($R^2 = 0,69$)	Mayor dispersión global, sobre todo en valores altos.	En la validación diaria presenta desfase y errores sistemáticos, con predicciones menos estables.

(a) (b)

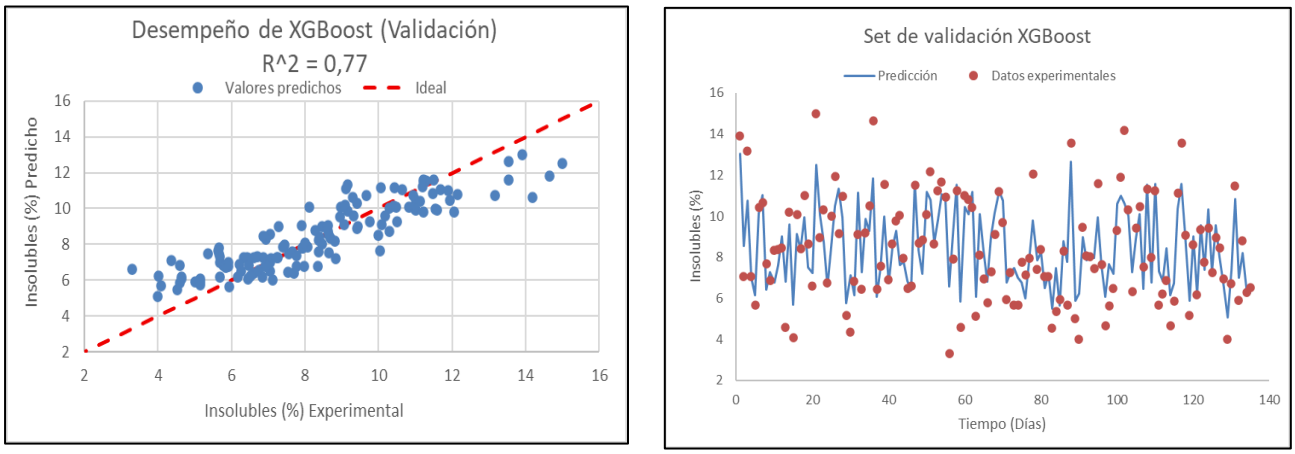

Figura 4.1 Resultados de *XGBoost* en datos de 3 años. (a) Ajuste del modelo, (b) Seguimiento de las predicciones a los datos experimentales.

(a) (b)

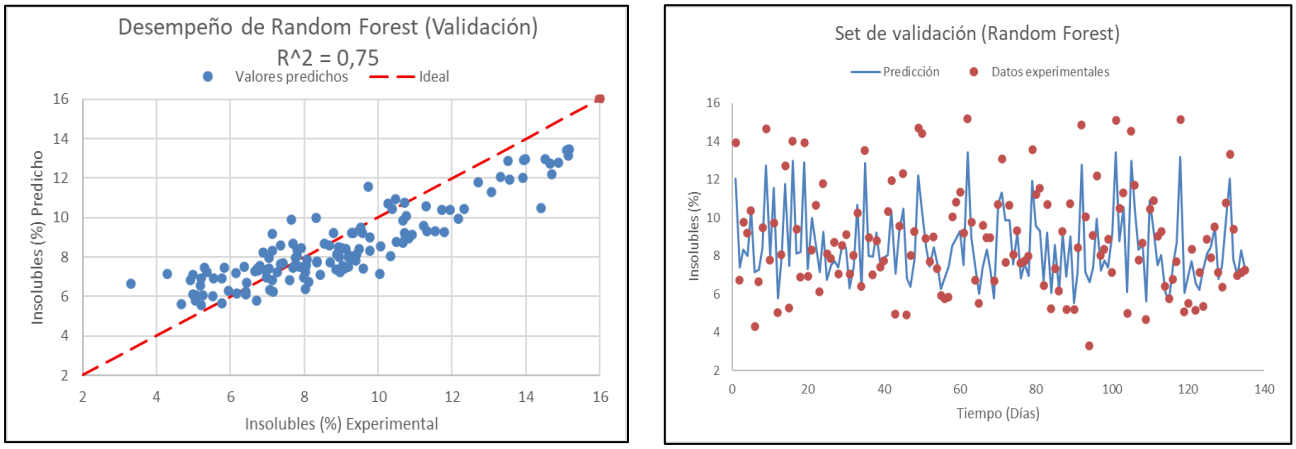

Figura 4.2 Resultados de *Random Forest* en datos de 3 años. (a) Ajuste del modelo, (b) Seguimiento de las predicciones a los datos experimentales.

(a) (b)

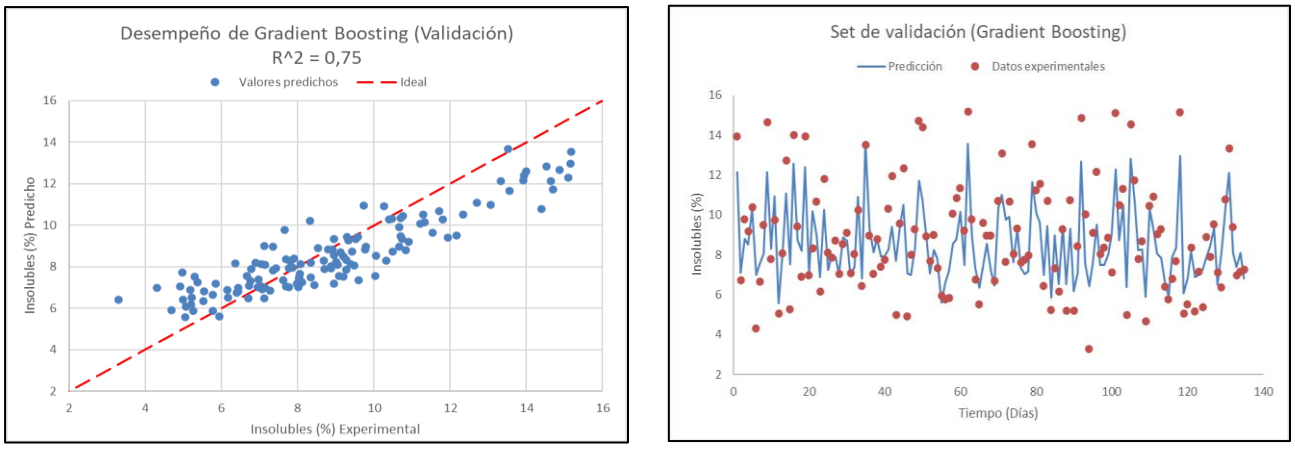

Figura 4.3 Resultados de *Gradient Boosting* en datos de 3 años. (a) Ajuste del modelo, (b) Seguimiento de las predicciones a los datos experimentales

(a) (b)

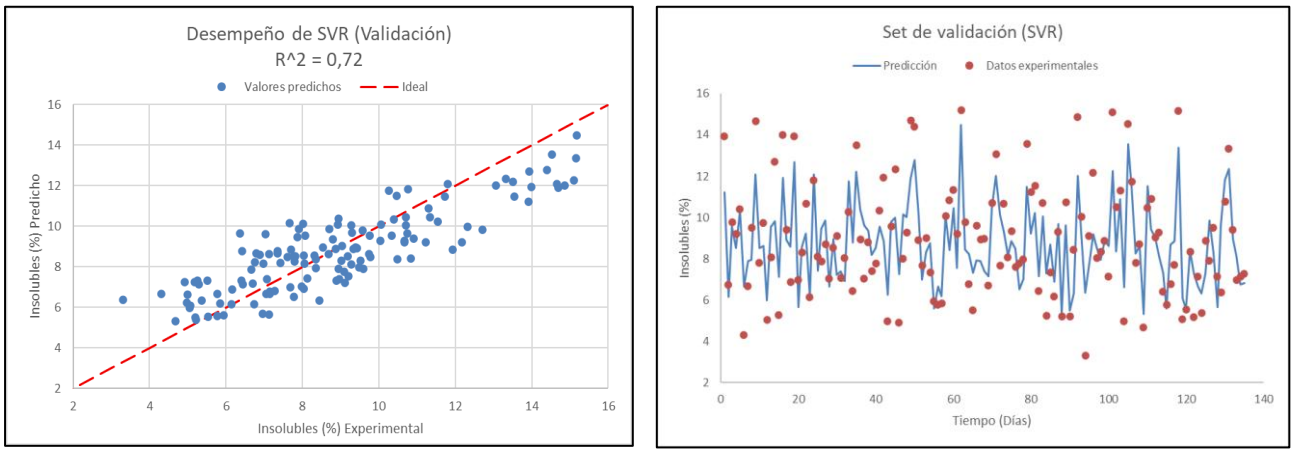

Figura 4.4 Resultados de SVR en datos de 3 años. (a) Ajuste del modelo, (b) Seguimiento de las predicciones a los datos experimentales

(a) (b)

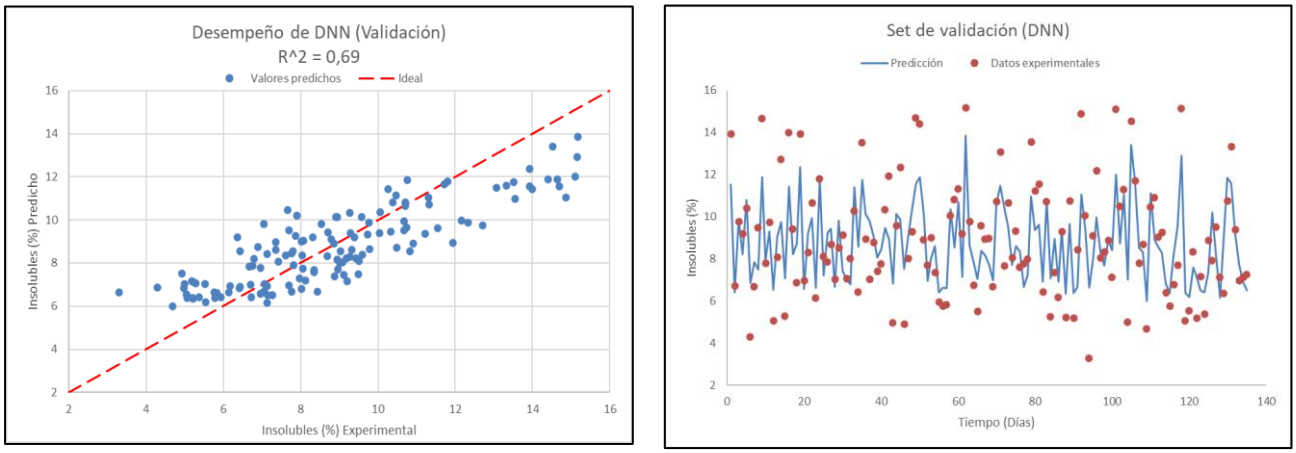

Figura 4.5 Resultados de DNN en datos de 3 años. (a) Ajuste del modelo, (b) Seguimiento de las predicciones a los datos experimentales

Luego se recopilaron los valores de las medidas de dispersión de cada variable, tal como se aprecia en el **Anexo 5: Valores estadísticos de variables seleccionadas en el rango de 3 años.** , donde se tienen los valores máximos, mínimos, promedios y su respectiva desviación estándar. Esta información es útil ya que permite analizar los valores de las variables seleccionadas y observar bajo qué rango operacional se encuentran operando tanto los equipos como las leyes de los minerales y de la mineralogía de las arcillas, y qué tanto varían los datos de cada variable.

Dado que se escogió el modelo de *XGBoost*, se analizan la importancia de las variables con el fin de observar qué parámetros operacionales son los que más impactan a la variabilidad del insoluble. Tal

como se aprecia en la Figura 4.6, las variables que más impactan son el pH de la línea 6 y línea 7 junto con la presión de BHC con el nivel de interfase de las celdas columnares.

El análisis de importancia de variables reveló que el pH de la línea 6 y 7 es el factor que más influye en la predicción del porcentaje de insolubles. Esta posibilidad existe debido a que esta etapa es la primera limpieza que pasa la pulpa, donde busca aumentar la ley de cobre y también el control y dilución de los insolubles. Por otro lado, al observar los valores p de los pH de cada celda en el **Anexo 3: Variables con valores p menor a 0,05 en data de 3 años,** esta variable arroja un valor 0 para L6-L7, donde las otras celdas tienen valores mayores, lo cual es un factor importante en el ámbito estadístico. Finalmente, al analizar la Tabla 8.5 del Anexo 5, el promedio de pH de dichas celdas es mayor respecto a las otras, donde las celdas *cleaner* se trabaja con un pH alto para que trabajen bien los reactivos, lo cual es un indicio de la importancia de dichas celdas para la variación de insolubles.

En segundo lugar, se observó una alta relevancia en la presión de la BHC CS003 proveniente de la carga que suministran los molinos de la línea 2, lo cual sugiere que esta variable tiene un impacto directo en la clasificación del mineral según su granulometría. También se ve en la gráfica que se encuentran 7 variables de la presión de equipos de BHC, siendo un total de 10 BHC en la etapa de molienda y 4 BHC en la etapa de flotación, por lo cual en los 3 años ocurridos los principales equipos con mayor importancia en cuanto a los insolubles son las baterías de hidrociclones.

Finalmente, el nivel de interfase en la celda columnar FT049 también fue una variable altamente significativa, y de igual forma su velocidad de rebose. Esta es la única celda que se encuentra presente en grado de importancia más alto respecto a las otras variables y con dos parámetros operacionales, lo cual indica que la variación de insolubles impacta significativamente en esta celda considerando los 3 años ocurridos.

La razón por la cual estas variables fueron más importantes que otras líneas o sensores puede deberse a que corresponden a rutas críticas de procesamiento en ese turno o periodo, donde la mineralogía o condiciones operativas particulares aumentaron su impacto en la generación de insolubles en el rango histórico de 3 años. Sin embargo, igual se destacan la importancia de otras variables operaciones de equipos específicos, donde si bien su importancia es levemente menor a comparación de las variables previamente mencionadas, igual generan un impacto para los insolubles, donde se destaca la presencia de la presión de distintas BHC, el nivel de interfase en las celdas columnares y el pH de distintas líneas de flotación.

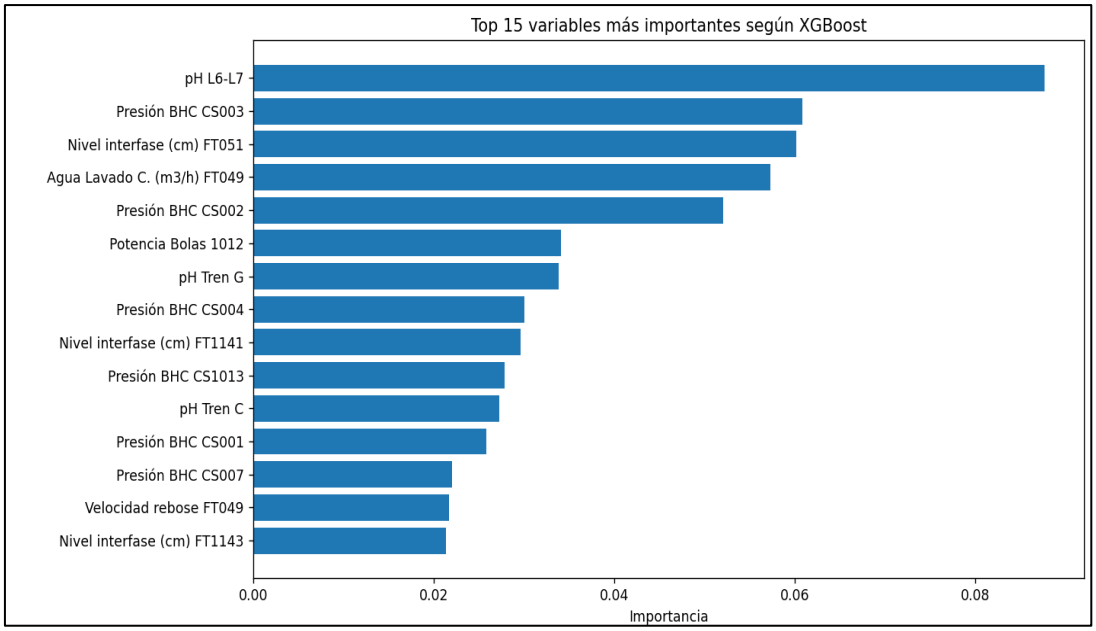

Figura 4.6 Parámetros operacionales más importantes del modelo *XGBoost* que impactan en el porcentaje de insolubles.

4.2 Análisis de 6 meses

En cuanto al análisis de data de los 6 meses, se considera la misma metodología usada para el análisis realizado en los 3 años, solo que en este caso se recopila la información de las variables que afectan a los insolubles de los últimos meses, con el fin de que se pueda comparar los resultados obtenidos entre los dos rangos de fecha y determinar si los modelos utilizados permiten presentar buenos modelos predictivos y las variables operacionales que influyen en la variación de insolubles para el concentrado final de cobre.

Así como se muestra en el **Anexo 6: Variables con valor p menor a 0,05 y mayor a 0,05 en data de 6 meses.**, se seleccionaron las variables que tienen un valor p menor a 0,05 y los que son mayor a 0,05 pero solo ciertas variables que afecten de manera importante a los insolubles, debido que no se busca saturar el modelo con una infinidad de parámetros operacionales.

Teniendo esto, se construyen los modelos predictivos, en donde se recopilan sus valores de métricas de desempeño en el conjunto de validación mostrados en la Tabla 4.3, para comparar entre modelos el que entregue un mejor desempeño estadístico predictivo.

Tabla 4.3 Métricas de desempeño en el conjunto de validación para 6 meses analizados

Modelo	RMSE	R^2	MAE	MAPE (%)
XGBoost	1,87	0,75	1,58	20,50
Random Forest	2,23	0,68	1,87	23,55
Gradient Boosting	2,03	0,74	1,70	18,21
SVR	2,24	0,68	1,77	21,79
ANN	**1,89**	**0,77**	**1,56**	**19,52**

Con los valores obtenidos, el modelo que presenta mejores resultados en el conjunto de 6 meses es el de ANN simple (*Artificial Neural Network*) por sus valores de métricas de desempeño. Si bien, este modelo presenta un RMSE similar al de *XGBoost*, lo que hace óptimo dicho modelo son sus valores de coeficiente de determinación y el error absoluto medio porcentual, ya que al tener un R^2 alto (sin llegar al valor de 1) y tener un MAPE bajo indica que al analizar nuevos datos no se generaría una gran dispersión entre los valores predichos hablando estadísticamente. Esta diferencia puede explicarse por el hecho de que las Redes Neuronales Artificiales pueden tratar de mejor manera un conjunto de datos que tiene pocas muestras con cierta cantidad de variables, no así los modelos tipo árboles (como *XGBoost*) que ocasionalmente presentan sobreajuste a la hora de tratar datos con pocas muestras.

En el caso de *Random Forest* y SVR, se aprecian coeficientes de determinación bajos y las otras métricas de desempeño con valores altos comparando con los otros modelos, lo que indica que estadísticamente no es factible apoyarse de dichos modelos para la modelación de los porcentaje de insolubles en el periodo previamente mencionado, dado que los datos estarán muy dispersos y habrá un margen de error de porcentaje de insolubles predichos que no estará ajustado a la idealidad de valores de insolubles experimentales respecto a los predichos.

Las figuras (4.7, 4.8, 4.9, 4.10 y 4.11) muestran la dispersión de datos en los distintos modelos predictivos, en donde se aprecia que el desempeño del modelo ANN en el conjunto de validación tiene una mejor precisión respecto a la línea ideal, lo cual tiene sentido considerando que su coeficiente de determinación es cercano a 1, mientras los modelos que tienen dicho parámetro más bajo tienen una dispersión de valores predichos frente a la línea ideal.

Por otro lado, en los sets de validación, para el ANN en los 6 meses tiene una predicción de insolubles que no tiene un rango de insolubles muy disperso frente a las otras gráficas, además que los datos experimentales coinciden con lo que se predice, evidenciando que el modelo funciona respecto a la tendencia de la predicción y de los valores recopilados de los datos históricos.

De manera resumida se describe en la Tabla 4.4 lo que se aprecia respecto a cada figura con el modelo utilizado, enfatizando en lo ajustado que se encuentran respecto a la línea ideal y en donde coinciden los datos de predicción respecto a los datos experimentales

Tabla 4.4 Descripción visual de los gráficos de distintos modelos en data de 6 meses

Modelo	**Descripción visual figuras (a)**	**Descripción visual figuras (b)**
Random Forest ($R^2 = 0,68$)	Ajuste moderado con dispersión evidente, sobre todo en valores altos.	A momentos sigue la tendencia general, pero responde con retraso y no alcanza bien los valores más extremos.
XGBoost ($R^2 = 0,75$)	Mejor alineación a la recta ideal, con menor dispersión global.	En temporal reproduce bien las variaciones, aunque suaviza algunos valores muy altos o bajos.
Gradient Boosting ($R^2 = 0,74$)	Ajuste cercano, aunque con ligera subestimación en valores altos.	Se ajustan los datos, pero presenta algo de atraso y menos detalle en las variaciones rápidas.
SVR ($R^2 = 0,68$)	Mayor dispersión, especialmente en valores bajos. En insolubles predichos sólo se queda en un rango de 6 a 14% en cualquier punto respecto al eje X.	La predicción es más rígida y no refleja bien los cambios bruscos de los datos reales.
ANN ($R^2 = 0,77$)	Mejor ajuste global, con puntos más cercanos a la recta ideal. Mayor precisión en rangos de insolubles entre 8 a 12%.	Refleja adecuadamente la tendencia y las variaciones, con menor desfase que los otros modelos.

(a) (b)

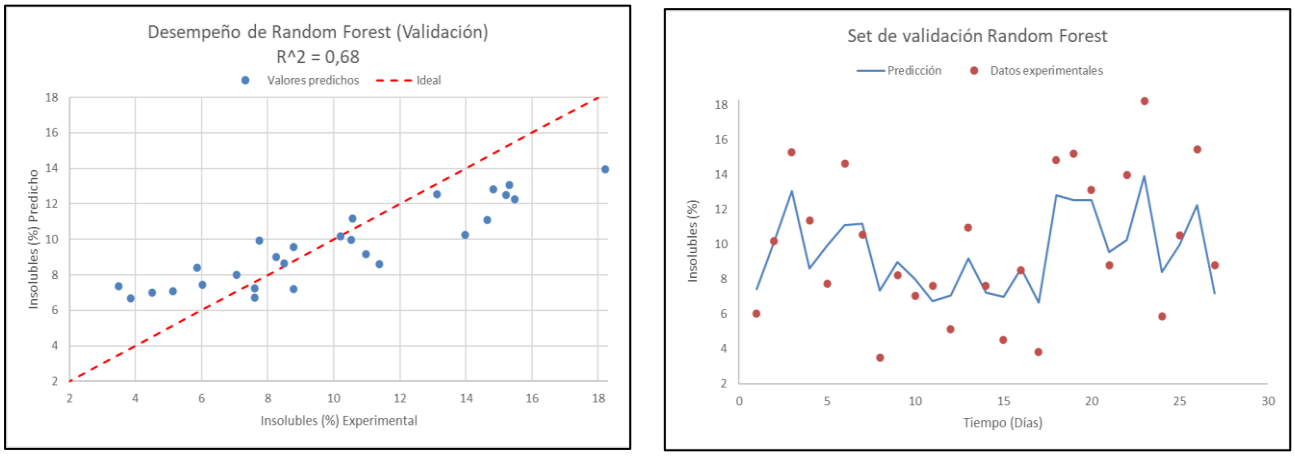

Figura 4.7 Resultados de *Random Forest* en datos de 6 meses. (a) Ajuste del modelo, (b) Seguimiento de las predicciones a los datos experimentales.

(a) (b)

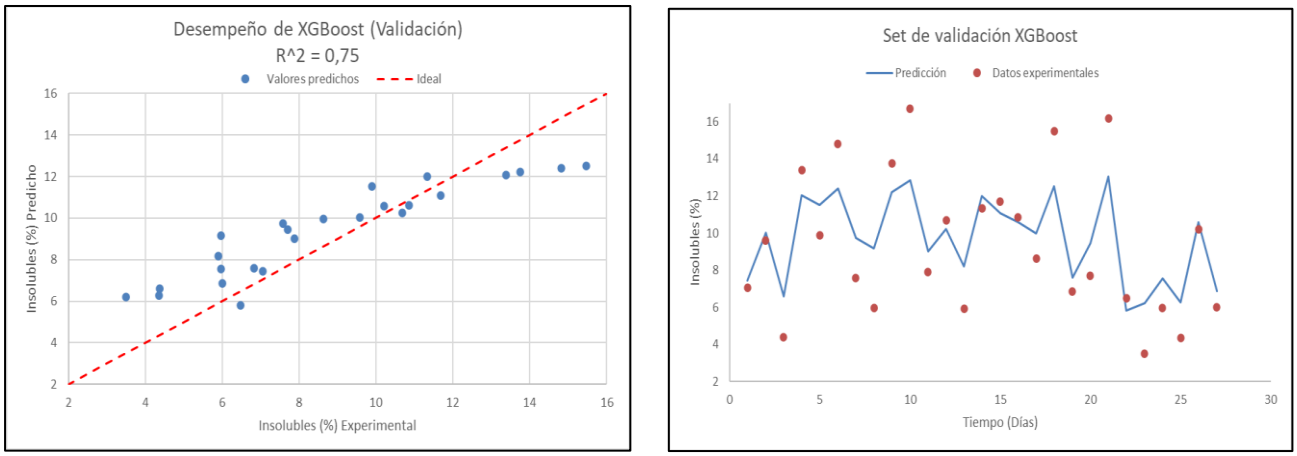

Figura 4.8 Resultados de *XGBoost* en datos de 6 meses. (a) Ajuste del modelo, (b) Seguimiento de las predicciones a los datos experimentales.

(a) (b)

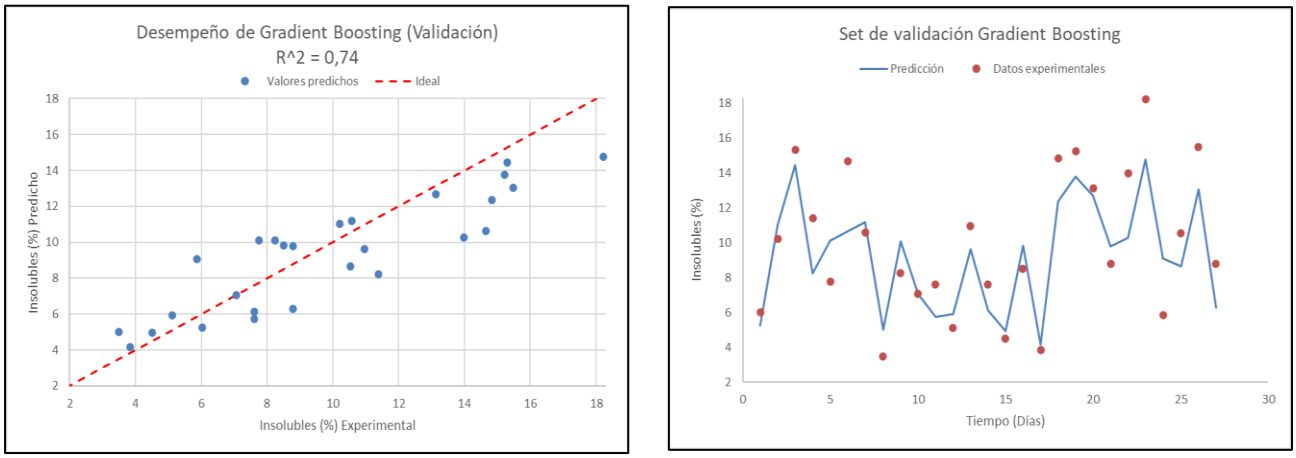

Figura 4.9 Resultados de *Gradient Boosting* en datos de 6 meses. (a) Ajuste del modelo, (b) Seguimiento de las predicciones a los datos experimentales.

(a) (b)

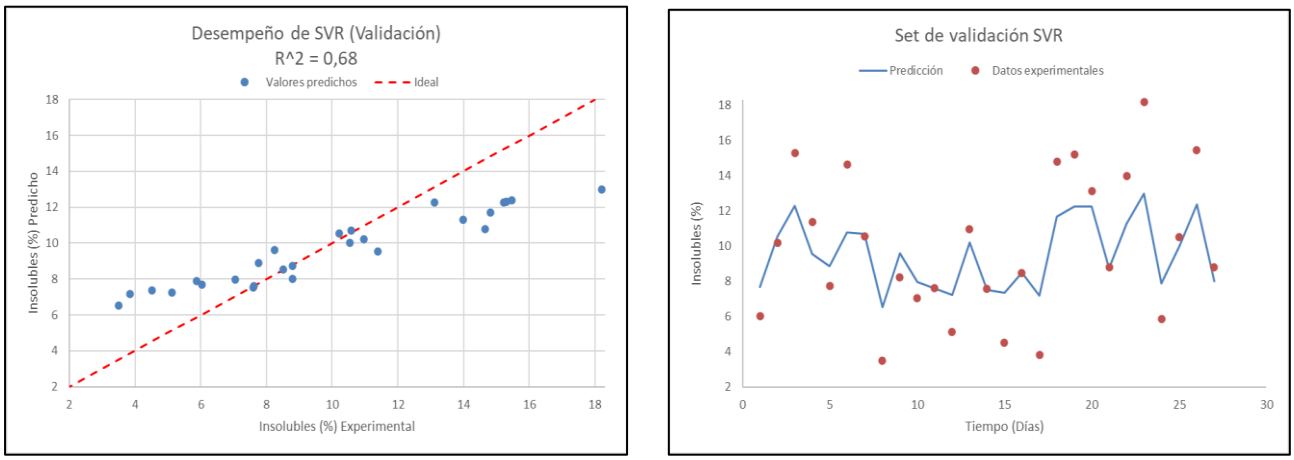

Figura 4.10 Resultados de SVR en datos de 6 meses. (a) Ajuste del modelo, (b) Seguimiento de las predicciones a los datos experimentales.

(a) (b)

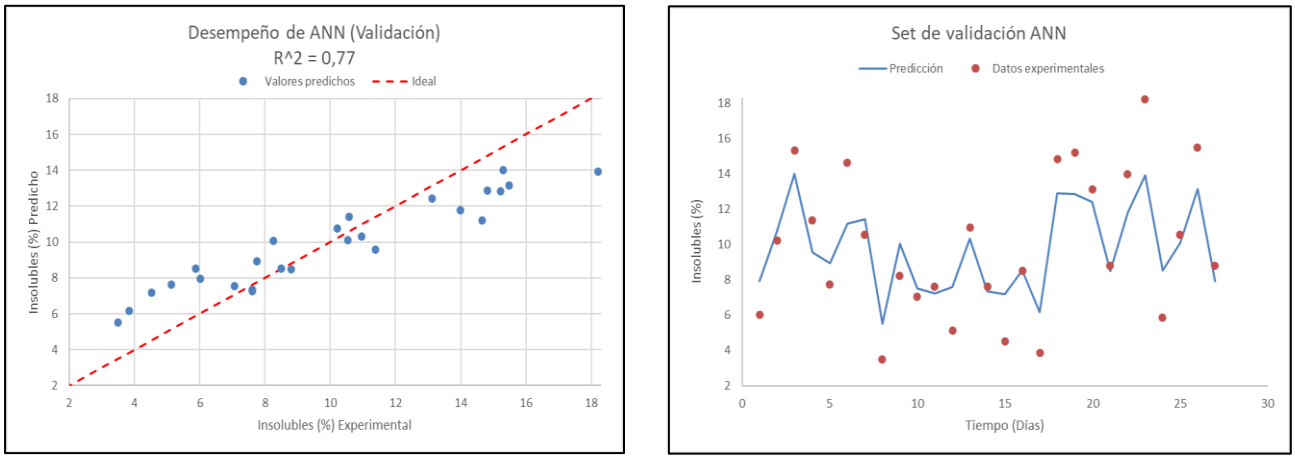

Figura 4.11 Resultados de ANN en datos de 6 meses. (a) Ajuste del modelo, (b) Seguimiento de las predicciones a los datos experimentales.

Teniendo los modelos definidos, se buscan los valores máximos, mínimos, promedio y desviación estándar de las variables que se agregaron al modelo (tanto las variables con valor p menor a 0,05 y ciertas variables con valor p mayor a 0,05) indicados en el **Anexo 7: Valores estadísticos de variables seleccionadas en el rango de 6 meses.** Este análisis permite caracterizar los parámetros operacionales, identificar sus rangos de operación y comparar las diferencias observadas entre variables y equipos.

Finalmente se obtienen las variables más importantes que se encuentran en el modelo respecto a la variación de insolubles. A diferencia del análisis de datos de 3 años, en donde hay una variedad de variables que influyen en la variación de los insolubles, para este caso las celdas columnares son las que tienen mayor presencia con los insolubles ya sea en un mejor control de los insolubles o una gran variabilidad de estos.

Las variables que encabezan la lista, como el nivel de interfase (FT050, FT051, FT052), se seleccionaron por su fuerte correlación operacional con la eficiencia del proceso de flotación respecto al mineral que va entrando. Su operación en esas celdas indica que hay que tener mayor control operacional respecto al nivel de interfase, con el fin de que no se genere mucha espuma o poco nivel de espuma para así no arrastrar insolubles.

El caudal de agua de lavado en las celdas columnares se encuentra presente en la lista por su relevancia práctica en la operación. El agua de lavado, al ser aplicada en la zona superior de las celdas, permite remover impurezas arrastradas por la espuma, donde si esta acción es deficiente o excesiva, se altera el perfil del concentrado, aumentando el porcentaje de insolubles. En la gráfica de las 15 variables más importantes se puede ver que se encuentran cinco celdas columnares, en donde la FT1143 y FT1144 son de tipo circular y las restantes son de tipo rectangular. El constante control operacional y limpieza del sistema de regio permite tener un control en cuanto a la dilución de los insolubles, pero también pueden ser las principales variables que alteren los insolubles. Sin embargo, la gráfica solo indica la importancia de dichos parámetros, y por esto es mejor comparar con las variables estadísticas obtenidas en el inicio.

Finalmente, se seleccionaron indicadores como el pH en diferentes etapas del proceso (L6-L7 y Tren C) al igual que en el análisis de los 3 años, lo cual evidentemente son las filas que mayor impacto tienen respecto a la variación de insolubles, principalmente porque son celdas de flotación de limpieza y no han estado por ciertos días sin operar, no así como el Tren A y el Tren B, los cuales también son celdas con configuración de limpieza, pero han tenido detenciones a lo largo de los 6 meses analizados,

ya sea para hacer una mantención a los equipos, porque falló una celda o alguna otra falla que no permita realizar su operación de manera correcta.

El análisis de importancia de variables tanto de la Figura 4.6 y 4.12 muestra que los factores con mayor incidencia en la variación de insolubles corresponden a parámetros operativos ajustables en planta, tales como pH, nivel de interfase de las celdas columnares, presión de las baterías de hidrociclones, entre otros. Esto implica que la predicción del modelo no solo tiene validez estadística, sino que ofrece puntos de control operacionales. Estos parámetros se pueden regular mediante sistemas de dosificación de cal y reactivos, control de válvulas y monitoreo automático de niveles, en donde a través de una sala de control se reciben la lectura de los datos operacionales a través de *PI System* y los trabajadores modifican dichos parámetros para poder llegar a un rango operacional óptimo, siendo en este caso para el control de los insolubles, o de igual manera se hace una inspección visual de los equipos en terreno, dado que hay situaciones en las que un equipo no opera de manera correcta y hay que notificar o suspender su operación para realizar el arreglo, sino puede realizar una mala operación y, en este caso, llegue a arrastrar insolubles a los procesos posteriores.

En la figura 4.12, se puede observar la importancia de las variables en el modelo seleccionado, donde se presentan las 15 variables más importantes en cuanto al impacto que generan en la variación de porcentajes de insolubles.

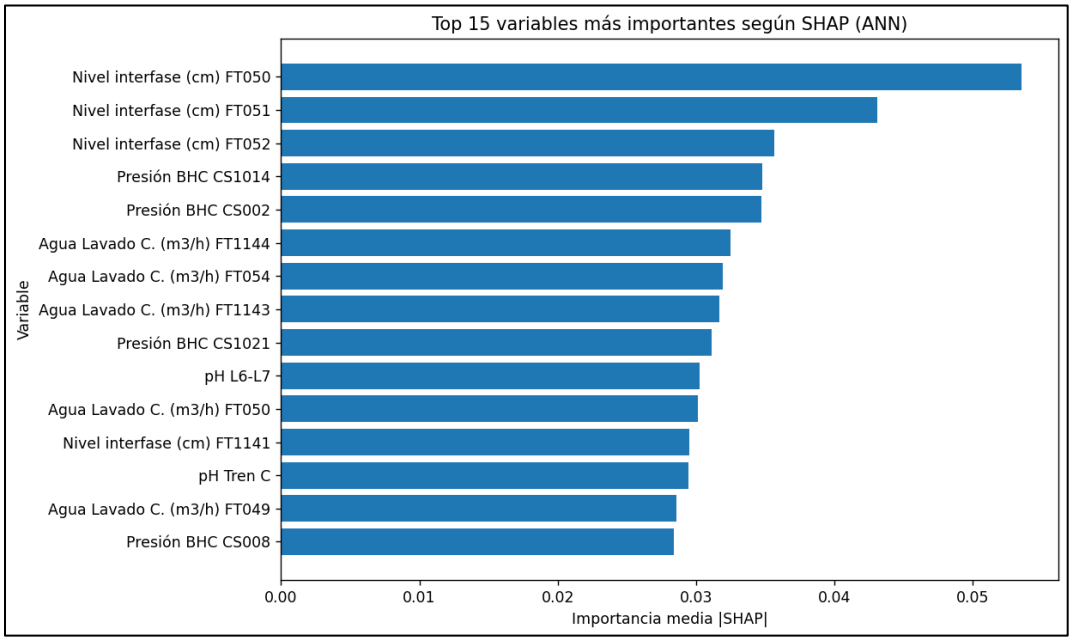

Figura 4.12 Variables más importantes que entrega el modelo ANN

Cabe destacar que el análisis que se hizo, tanto para 3 años y para 6 meses, se debe considerar que el mineral presente en los yacimientos ha cambiado, por lo cual los resultados obtenidos de los distintos análisis puede variar por el simple hecho del tipo de mineral que va entrando, además de la forma de operar, los nuevos equipos que ha incluido CMDIC a la planta concentradora, la mantención y antigüedad de ciertos equipos son aspectos que se consideran a la hora de analizar modelos predictivos estadísticos y las variables que se seleccionan para llevarlos ahí.

Para cada período de tiempo se obtienen distintas respuestas respecto a los modelos que mejor se ajustan y a los equipos operacionales que más inciden en la variación de insolubles. Por esto, en la Tabla 4.5 se describen las ventajas y desventajas respecto a llevar un estudio de un gran período de tiempo, siendo en este caso de tres años y de un período corto y más reciente, siendo el de 6 meses.

Tabla 4.5 Ventajas y desventajas de distintos períodos a analizar

Período	Ventajas	Desventajas
3 años	• Mayor cantidad de datos, lo cual habrá mejor robustez estadístico. • Permite entrenar modelos predictivos más estables y menos sensibles al ruido. • Mejor capacidad de análisis frente a distintos escenarios, ya sean operacionales, climatológicos o mineralógicos.	• Mayor probabilidad de obtener datos anormales que entreguen ciertas variables operacionales. • Puede incluir condiciones de operación que no sean representativas al período actual.
6 meses	• El modelo trabaja con datos actuales de la planta. • Responde a cambios recientes operacionales. • El modelo tendrá mejor respuesta analítica a corto plazo si se busca resolver un problema actual (como el desafío del control de los insolubles).	• Riesgo de sobreajuste en modelos que funcionen mejor con mayor cantidad de datos. • Se tiene registro de un período corto, lo cual no se puede tener un mejor análisis frente a distintas condiciones operacionales.

5 Conclusiones

Del estudio realizado con técnicas de *machine learning* y modelos estadísticamente predictivos, se concluye que para el análisis de datos de 3 años el modelo más adecuado es XGBoost. Su fortaleza radica en la capacidad de procesar grandes volúmenes de información histórica de la planta concentradora y capturar la variabilidad de insolubles presente en los yacimientos cupríferos entre 2022 y 2025. Los resultados en el conjunto de validación muestran métricas de buen desempeño con valores de RMSE (1,20), R^2 (0,77), MAE (0,94) y MAPE (12,90%), superando a los otros modelos.

Por otro lado, en el análisis de un periodo reducido de 6 meses, correspondiente a datos recientes de variables operacionales, las Redes Neuronales Artificiales (ANN) presentan un mejor rendimiento. Este modelo destaca al generalizar adecuadamente mediante técnicas como el *early stopping*, que evita sobreajuste y la captura de ruido en un conjunto de datos de menor tamaño. En términos comparativos, logra un ajuste más preciso que otros modelos cuando se trabaja con muestras acotadas.

Al contrastar los distintos enfoques, se observa que la relevancia de las variables depende del horizonte temporal de análisis. En los 3 años de información, aparecen múltiples factores asociados a los insolubles que afectan al concentrado de cobre; en cambio, en los 6 meses recientes, se evidencia un mayor peso de las variables asociadas al control de las celdas columnares. Esto sugiere que, en la actualidad, resulta clave optimizar la gestión de pulpa en dichas celdas de flotación para reducir la presencia de insolubles en el concentrado final.

En consecuencia, se recomienda utilizar modelos basados en árboles cuando se dispone de un historial extenso de datos, dada su robustez frente a ruido, valores faltantes y *outliers,* además de su capacidad para detectar interacciones complejas entre variables. En contraste, para periodos más cortos, las ANN se muestran como la alternativa más adecuada, ya que logran capturar patrones no lineales con buen equilibrio entre precisión y capacidad de generalización.

Finalmente, considerando que en Collahuasi opera un área DCS (*Distributed Control System*) que recopila y analiza información operacional en tiempo real, se sugiere implementar un modelo de XGBoost en este entorno. De esta manera, el sistema podría integrar el código desarrollado para el modelo predictivo, recoger lecturas de las variables seleccionadas y detectar variaciones de parámetros operacionales. Con ello, se facilitaría la entrega de alertas en línea al personal, apoyando

la toma de decisiones inmediatas para ajustar equipos o condiciones de proceso y alcanzar el porcentaje de insolubles deseado.

6 Recomendaciones para trabajos a futuro

Los insolubles son un tema que acompleja a Collahuasi en estos días, donde el principal afectado es el concentrado de cobre, es fundamental tener que hacer investigaciones y pruebas exhaustivas con el fin de poder diluir de mejor manera los insolubles y aumentar el concentrado de cobre. En esta investigación se consideran variables que afectan a la variación de insolubles (ya sea en mayor o menor cantidad) las cuales se llevan a modelos predictivos de manera que con la estadística y *machine learning* entreguen los mejores modelos con buenas predicciones, pero no siempre hay que guiarse plenamente a través de lo que indique la inteligencia artificial, también es necesario complementarlo con estudios en terreno (por ejemplo determinar las zonas de extracción del mineral), estudios con pruebas en laboratorio, hacer inspección en la planta (para ver si un equipo está operando correctamente), tener un control de lectura de las variables, y así con más variables que normalmente están fuera del alcance de la IA.

Es por esto que se entregan las siguientes recomendaciones con el fin de tenerlas en consideración por si se llega a realizar otro estudio de mejora para la compañía en el ámbito de producción con el uso de inteligencia artificial y *machine learning* y así descartar ciertos equipos o ciertas variables que afecten a la variable objetivo y tener en cuenta las que tienen mejor desempeño, de manera que se pueda tener una referencia y así una mejora con los equipos o variables descartadas.

- Tener un mejor control de operación con las celdas de flotación, donde se tome en cuenta la lectura correcta del equipo y que no realice *pulpeos* (que rebose mucha pulpa en la celda y arrastre insolubles)
- Mejorar la recopilación de datos de las variables que se encuentran en las celdas columnares (agua de lavado, nivel de interfase, velocidad de rebose, entre otros)
- Tomar muestras de datos de períodos estables, donde no haya ocurrido una gran variabilidad de datos con el fin de analizar si hay una mejor predicción para él % de insolubles o seleccionar variables estables.

7 Referencias

Bergh, L. G., & Yianatos, J. B. (1991). ALTERNATIVAS DE CONTROL DE COLUMNAS DE FLOTACION. In *Column Flotation'91 SME Annual Meeting 549* (Vol. 568).

Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2, pp. 39-44). Reading, MA: Addison-wesley.

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow* (2nd ed., pp. 30). O'Reilly Media.

Willmott, C. J., & Matsuura, K. (2005). *Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance*. *Climate Research*, 30(1), 79–82.

Armstrong, J. S., & Collopy, F. (1992). *Error measures for generalizing forecasting methods: Empirical comparisons*. *International Journal of Forecasting*, 8(1), 69–80.

Grim, R. E. (1968). *Clay mineralogy* (2ª ed.). McGraw-Hill.

Soto, H. (2009). *Fundamentos de Flotación de Minerales*. Universidad de Concepción.

Badillo, P. (2024*) Modelación del p80 del producto de la molienda secundaria en base a parámetros operacionales en Compañía Minera Doña Inés de Collahuasi.* Universidad Técnica Federico Santa María.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer.

Finch, J. A., & Dobby, G. S. (1990). *Column Flotation*. Pergamon Press.

Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.

Zhang, G., Eddy Patuwo, B., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, 14(1), 35–62.

Wills, B. A., & Finch, J. (2015). *Wills' Mineral Processing Technology: An Introduction to the Practical Aspects of Ore Treatment and Mineral Recovery* (8th ed.). Butterworth-Heinemann.

Collahuasi. (2025). *Nuestra Compañía*. Recuperado de https://www.collahuasi.cl/quienes-somos/la-compania/

CMDIC (2025). *Observación propia durante práctica profesional en Planta Concentradora*. Documento no publicado. Compañía Minera Doña Inés de Collahuasi, Tarapacá, Chile.

Jung, D., & Choi, Y. (2021). *Systematic Review of Machine Learning Applications in Mining: Exploration, Exploitation and Reclamation. Minerals.*

Rizkallah, L.W. (2025). *Enhancing the performance of gradient boosting trees on regression problems. J Big Data* 12, 35.

Imani, M., Beikmohammadi, A., & Arabnia, H. R. (2025). *Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels*. *Technologies, 13*(3), 88.

Wang, W. (2005). *Some Issues About the Generalization of Neural Networks for Time Series Prediction*. En *Artificial Neural Networks: Formal Models and Their Applications – ICANN 2005* (pp. 559-564). Springer.

Schwertman, N. C., Owens, M. A., & Adnan, R. (2004). *A simple more general boxplot method for identifying outliers*. Computational Statistics & Data Analysis.

Wooldridge, J. M. (2016). *Introductory Econometrics: A Modern Approach* (6th ed.). Cengage Learning.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R (2nd ed.).*

Stanton, J. M. (2001). *Galton, Pearson, and the peas: A brief history of linear regression for statistics instructors*. Journal of Statistics Education.

Concha, F., & Bouso, J. L. (2021). *Fluid Mechanics Fundamentals of Hydrocyclones and Its Applications in the Mining Industry*.

Hein, Z., Lin, D., Lau, T., & Wu, M. (2019). *Gradient Boosting Machine: A Survey*.

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System.*

Schönlau, M., & Zou, Y. Y. (2020). *The Random Forest algorithm for statistical learning*. *Journal Title*.

Bulatovic, S. M. (2007). *Handbook of Flotation Reagents: Chemistry, Theory and Practice*. Elsevier.

Cruz, N., Peng, Y., & Wightman, E. (2015). *Surface chemistry aspects of clay minerals in flotation*. *Minerals Engineering*, 66–68, 100–118.

Ma, M., Bruckard, W. J., & Holmes, R. J. (2009). *Effect of clay minerals on pulp rheology and the flotation of copper and gold ores*. *International Journal of Mineral Processing*, 93(2), 144–148.

Cruz, N., Peng, Y., Wightman, E., & Xu, N. (2015). The interaction of pH modifiers with kaolinite in copper–gold flotation. *Minerals Engineering, 84*, 27–34.

Jeldres, R. I. (2017). *Uso de goma guar como agente floculante de arcillas en flotación de calcopirita* [Tesis de Magíster, Universidad de Concepción]. Repositorio UdeC.

Seaman, D., Lauten, R. A., Kluck, S., & Stoitis, N. (2012). *Usage of anionic dispersants to reduce the impact of clay particles in flotation of copper and gold at the Telfer Mine*. *OneMine*.

Biau, D. J., & Jolles, B. M. (2009). *P value and the theory of hypothesis testing*. *Journal of Clinical Epidemiology*, 62(5), 493–498.

8 Anexos

8.1 Dimensiones y tratamiento de molinos SAG y bolas.

Tabla 8.1 Datos de molinos SAG y molino de bolas en CMDIC

Equipo	Fabricante	Cap. Tratamiento [ton]	Dimensión [pie]	Cap. Potencia [kW]
SAG #1 SAG #2	Fuller	1700	32x15	7992
MB003 MB004	Fuller		22x36	9695
SAG #1011	Metso	5500	40x24	21000
MB1012 MB1013	Metso		26x38	15500

8.2 Diagrama del proceso de flotación CMDIC

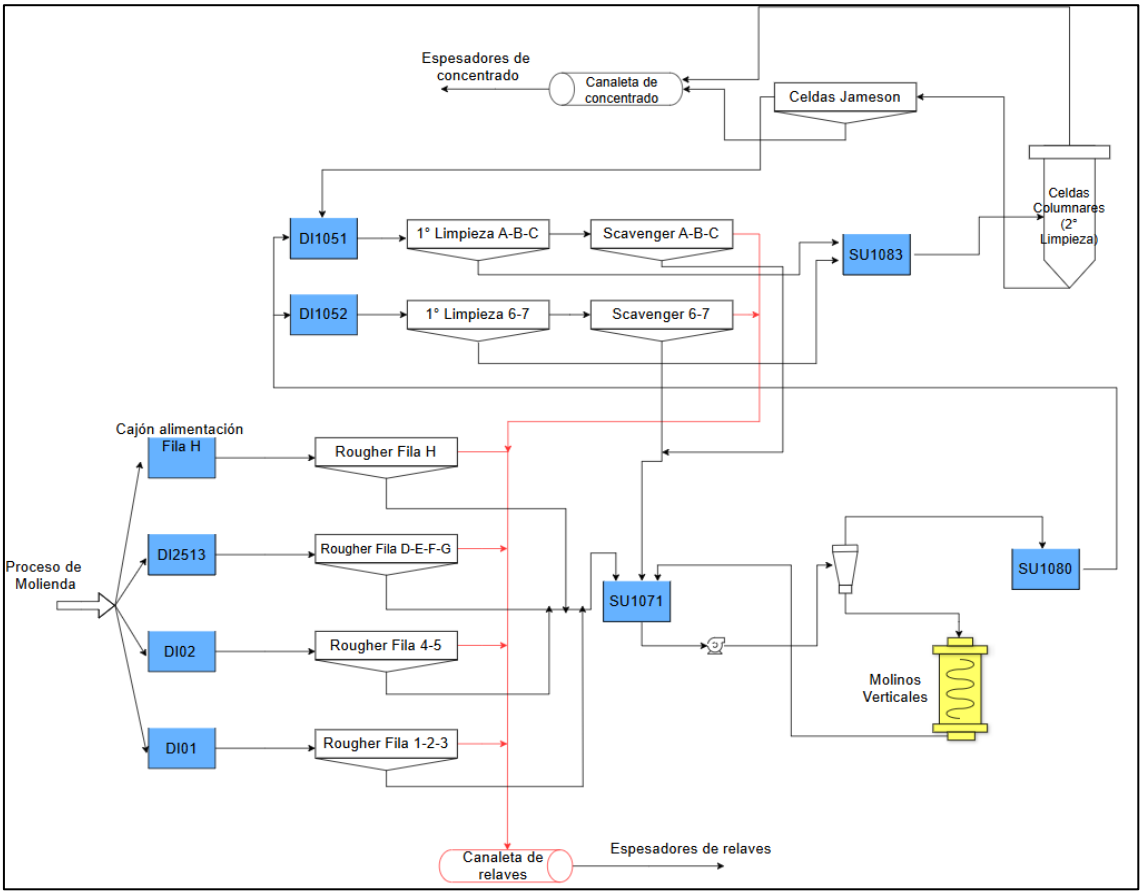

Figura 8.1 Diagrama del proceso de flotación en CMDIC

8.3 Variables con valores p menor a 0,05 en data de 3 años

Tabla 8.2 Variables con valores p menor a 0,05 seleccionadas en data de 3 años

Variable		Valor p
Potencia Molino de Bolas [kW]	MB004	0,023
	MB1012	0,034
Presión BHC [psi]	CS002	0,001
	CS003	0,001
	CS004	0,036
	CS1013	0,006
pH	L6-L7	0,000
	Tren A	0,012
	Tren C	0,008
	Tren D	0,012
Agua de lavado columnas [m3/h]	FT049	0,000
	FT052	0,002
	FT1143	0,038
	FT1144	0,040
Interfase de nivel de columnas [m3/h]	FT050	0,002
	FT051	0,000
	FT052	0,001
	FT053	0,006
	FT054	0,004
	FT1142	0,033

8.4 Variables seleccionadas con valores p mayores a 0,05 en data de 3 años

Tabla 8.3. Valores p mayor a 0,05 seleccionados en datos de 3 años de: potencia, presión de BHC y pH

Variable		Valor p
Potencia Molino de Bolas [kW]	MB003	0,546
	MB1013	0,401
	MB1022	0,666
Presión BHC [psi]	CS001	0,056
	CS1011	0,315
	CS1012	0,745
	CS1014	0,293
	CS1021	0,158
	CS1022	0,246
	CS005	0,247
	CS006	0,704
	CS007	0,991
	CS008	0,143
pH	R4-R5	0,273
	Tren B	0,821
	Tren E	0,493
	Tren F	0,923
	Tren G	0,074

Tabla 8.4. Valores p mayor a 0,05 seleccionados en datos de 3 años de: mineralogía, ley alimentada, agua de lavado de columnas, nivel de interfase de columnas y velocidad de rebose

Variable		Valor p
Mineralogía [%]	Pirofilita	0,257
	Caolinita	0,353
	Ilita	0,974
	Moscovita	0,257
Ley Alimentada [%]	CuT	0,155
	FeT	0,193
	As [ppm]	0,179
Agua de lavado columnas [m3/h]	FT050	0,393
	FT051	0,761
	FT053	0,697

	FT1141	0,383
	FT1142	0,453
Interfase de nivel de columnas [m3/h]	FT049	0,407
	FT1141	0,104
	FT1143	0,461
	FT1144	0,192
Velocidad de rebose [m3/h]	FT049	0,603
	FT054	0,868
	FT1141	0,284

8.5 Valores estadísticos de variables seleccionadas en el rango de 3 años.

Tabla 8.5 Valores máximo, mínimo, promedio y desviación estándar de variables seleccionadas en rango de 3 años

Variable		Máximo	Mínimo	Promedio	D. Estándar
Potencia Molino de Bolas [kW]	MB003	6262	5345	5813	174
	MB004	6468	5376	5926	217
	MB1012	18429	11130	14747	1235
	MB1013	17667	10904	14670	1241
	MB1022	19266	5838	13212	2532
Presión BHC [psi]	CS001	13,00	9,09	10,99	0,617
	CS002	12,95	9,12	10,88	0,614
	CS003	13,03	8,89	10,94	0,634
	CS004	12,97	9,32	11,03	0,604
	CS1011	13,32	9,22	11,18	0,639
	CS1012	13,09	8,72	11,27	0,693
	CS1013	12,40	9,34	10,99	0,582
	CS1014	12,83	9,04	10,82	0,677
	CS1021	13,20	9,33	11,19	0,652
	CS1022	13,06	8,94	10,98	0,654
	CS005	15,31	11,04	12,76	0,563
	CS006	15,00	11,42	12,96	0,653
	CS007	14,90	11,09	13,04	0,631
	CS008	15,21	11,29	12,73	0,599
pH	R4-R5	10,66	9,40	10,04	0,312
	L6-L7	13,11	10,28	11,55	0,509
	Tren A	11,20	9,71	10,45	0,372
	Tren B	11,00	9,60	10,30	0,342

	Tren C	10,90	9,50	10,18	0,356
	Tren D	10,50	9,30	9,89	0,294
	Tren E	10,39	9,20	9,79	0,294
	Tren F	10,30	9,10	9,69	0,299
	Tren G	10,20	9,00	9,58	0,287
Mineralogía [%]	Pirofilita	1,14	0,01	0,30	0,286
	Caolinita	1,22	0,10	0,55	0,252
	Ilita	13,04	4,97	8,98	1,464
	Moscovita	16,92	5,53	11,04	2,135
Ley Alimentada [%]	CuT	1,54	0,70	1,12	0,163
	FeT	5,36	1,89	3,32	0,747
	As [ppm]	319,34	16,30	118,02	73,157
Agua de lavado columnas [m3/h]	FT049	89,63	39,39	64,79	11,007
	FT050	87,45	41,66	64,70	9,859
	FT051	82,55	42,93	62,35	8,624
	FT052	89,43	36,67	61,02	11,547
	FT053	77,12	44,14	60,74	6,772
	FT054	73,37	50,18	62,10	5,480
	FT1141	80,42	45,42	64,42	7,424
	FT1142	96,80	34,20	64,70	13,435
	FT1143	72,35	41,60	56,81	6,143
	FT1144	75,48	48,45	63,59	5,897
Interfase de nivel de columnas [m3/h]	FT049	81,56	47,82	66,17	4,837
	FT050	80,68	35,80	58,23	7,002
	FT051	75,00	44,10	58,93	4,615
	FT052	86,94	36,43	61,79	7,899
	FT053	79,66	47,40	61,92	4,961
	FT054	91,75	34,14	63,86	9,249
	FT1141	81,25	45,74	64,40	5,494
	FT1142	77,76	41,25	57,64	4,919
	FT1143	85,79	38,99	62,61	8,039
	FT1144	81,79	41,03	61,62	6,815
Velocidad de rebose [m3/h]	FT049	484,99	93,04	227,64	88,775
	FT054	463,31	93,10	248,88	89,098
	FT1141	447,54	93,00	250,43	83,571

8.6 Variables con valor p menor a 0,05 y mayor a 0,05 en data de 6 meses

Tabla 8.6 Valores p menor a 0,05 seleccionados en data de 6 meses

Variable		Valor p
Presión BHC [psi]	CS002	0,000
	CS003	0,014
	CS004	0,041
	CS1011	0,034
	CS1014	0,048
	CS1021	0,001
	CS007	0,041
	CS008	0,006
pH	Tren A	0,020
	Tren B	0,039
	Tren E	0,009
Ley Alimentada [%]	FeT	0,029
Agua de lavado columnas [m3/h]	FT050	0,012
	FT053	0,046
	FT1144	0,006
Interfase de nivel de columnas [m3/h]	FT050	0,005
	FT051	0,039
	FT053	0,030
	FT1144	0,006

Tabla 8.7 Valores p mayor a 0,05 seleccionados en data de 6 meses de: potencia, presión BHC, pH y mineralogía

Variable		Valor p
Potencia Molino de Bolas [kW]	MB003	0,723
	MB004	0,330
	MB1012	0,722
	MB1013	0,337
	MB1022	0,084
Presión BHC [psi]	CS001	0,316
	CS1012	0,102
	CS1013	0,779
	CS1022	0,548

	CS005	0,587
	CS006	0,355
pH	L6-L7	0,880
	Tren C	0,753
	Tren D	0,153
	Tren F	0,257
	Tren G	0,499
Mineralogía [%]	Pirofilita	0,636
	Caolinita	0,532
	Ilita	0,945
	Moscovita	0,970
Ley Alimentada [%]	CuT	0,880
	As [ppm]	0,340

Tabla 8.8 Valores p mayor a 0,05 seleccionados en data de 6 meses de: ley alimentada, tiempo de residencia, agua de lavado de columnas, interfase de nivel de columnas y velocidad de rebose

Variable		**Valor p**
Ley Alimentada [%]	CuT	0,880
	As [ppm]	0,340
Tiempo de residencia [min]	L1-L2	0,592
	Cleaner	0,102
	L3	0,989
Agua de lavado columnas [m3/h]	FT049	0,149
	FT051	0,585
	FT052	0,143
	FT054	0,259
	FT1141	0,082
	FT1142	0,721
	FT1143	0,849
Interfase de nivel de columnas [m3/h]	FT049	0,251
	FT052	0,237
	FT054	0,652
	FT1141	0,104
	FT1142	0,731
	FT1143	0,643

Velocidad de rebose [m3/h]	FT050	0,197
	FT052	0,540
	FT053	0,316
	FT054	0,736
	FT1143	0,502
	FT1144	0,488

8.7 Valores estadísticos de variables seleccionadas en el rango de 6 meses.

Tabla 8.9 Valores máximo, mínimo, promedio y desviación estándar de variables seleccionadas en rango de 6 meses

Variable		Máximo	Mínimo	Promedio	D. Estándar
Potencia Molino de Bolas [kW]	MB003	6574	5346	6002	264
	MB004	6241	5213	5774	215
	MB1012	17990	12279	15020	1036
	MB1013	16460	12317	14721	885
	MB1022	20594	11719	16794	1950
Presión [psi]	CS001	10,52	9,29	9,87	0,276
	CS002	10,79	9,09	9,89	0,345
	CS003	11,60	9,30	10,55	0,518
	CS004	10,50	9,41	9,88	0,238
	CS1011	10,62	9,26	9,91	0,336
	CS1012	10,63	9,33	9,97	0,315
	CS1013	10,58	9,29	9,93	0,300
	CS1014	10,83	9,23	10,01	0,339
	CS1021	10,65	8,97	9,79	0,366
	CS1022	10,93	9,17	10,05	0,368
	CS005	15,29	11,67	13,43	0,821
	CS006	15,25	11,66	13,13	0,778
	CS007	15,24	9,15	11,81	1,341
	CS008	16,91	10,06	13,50	1,523
pH	L6-L7	11,76	10,16	10,94	0,379
	Tren A	11,95	10,37	11,06	0,345
	Tren B	11,72	9,62	10,56	0,471
	Tren C	11,72	8,86	10,28	0,706
	Tren D	10,56	9,17	9,95	0,320

	Tren E	11,23	8,75	9,90	0,540
	Tren F	10,75	8,60	9,73	0,517
	Tren G	10,76	9,05	9,98	0,407
Mineralogía [%]	Pirofilita	2,26	0,27	0,94	0,488
	Caolinita	1,37	0,23	0,67	0,256
	Ilita	15,70	3,03	9,26	2,488
	Moscovita	20,00	4,40	12,02	3,231
Ley Alimentada [%]	CuT	1,54	0,55	0,95	0,210
	FeT	5,96	2,60	4,04	0,719
	As [ppm]	252,79	47,15	130,28	46,364
Tiempo de residencia [min]	L1-L2	61,51	30,74	44,75	6,542
	Cleaner	32,00	17,00	22,27	3,369
	L3	56,44	29,52	42,71	4,862
Agua de lavado columnas [m3/h]	FT049	69,47	40,42	54,23	6,257
	FT050	69,35	40,87	54,64	6,172
	FT051	70,91	36,36	53,26	8,086
	FT052	68,12	40,25	53,83	6,063
	FT053	71,68	40,18	55,58	7,472
	FT054	73,58	41,61	55,99	7,381
	FT1141	74,23	41,84	57,84	7,351
	FT1142	68,59	40,13	53,72	6,597
	FT1143	72,92	41,23	56,56	7,759
	FT1144	70,18	40,00	54,35	6,497
Interfase de nivel de columnas [m3/h]	FT049	69,62	30,08	48,39	9,050
	FT050	65,88	32,27	48,88	8,410
	FT051	93,11	30,29	62,88	12,944
	FT052	96,66	38,15	64,45	12,322
	FT053	95,49	34,25	61,45	14,340
	FT054	92,32	30,81	61,15	13,471
	FT1141	86,35	36,04	58,17	11,924
	FT1142	87,43	33,22	59,73	11,947
	FT1143	93,78	30,01	60,89	14,634
	FT1144	90,49	33,03	59,98	13,693
Velocidad de rebose [m3/h]	FT050	342,59	30,74	142,70	79,597
	FT052	349,56	30,71	158,34	86,102
	FT053	349,94	32,83	169,89	86,032
	FT054	313,58	31,19	135,61	69,442

	FT1143	349,71	30,86	160,28	83,883
	FT1144	311,65	32,44	142,07	65,613

UNIVERSIDAD DE CONCEPCIÓN – FACULTAD DE INGENIERÍA

Departamento de Ingeniería Metalúrgica

Hoja Resumen Memoria de Título

Título: "Modelación del porcentaje de insolubles en concentrado final de cobre en función de variables operacionales del proceso en Compañía Minera Doña Inés de Collahuasi"
Nombre Memorista: Patricio Andrés Mamani Miranda

Modalidad	**Proyecto**	**Profesor(es) Guía (s)**
Concepto		
Calificación		
Fecha	03.10.2025	**Dr. Leopoldo Gutiérrez Briones**
Prof. Froilán Vergara G.		**Ingeniero Supervisor: Sergio Arellano Gorigoitía**
		Institución: Compañía Minera Doña Inés de Collahuasi

Comisión (Nombre y Firma)	
Prof. Andrés Ramírez	Prof. Luver Echeverry

Resumen

En esta memoria de título se crearon modelos predictivos supervisados con el fin de determinar el porcentaje de insolubles presentes en el concentrado final. Se analizaron variables operacionales que influyen directa o indirectamente a la variación de insolubles y se recopilaron datos de 3 años (10 de febrero de 2022 al 08 de febrero de 2025) y de 6 meses (01 de diciembre de 2024 al 30 de mayo de 2025), donde se eliminaron muestras de datos anormales y atípicos y se seleccionaron ciertas variables operacionales a los modelos predictivos. Se ajustaron los modelos y se escogieron los que presentaron las mejores métricas de desempeño de cada período de tiempo respectivo. Finalmente se destacaron las variables operacionales que más influyeron en la variación de los insolubles en concentrado final de cobre.