Universidad de Concepción
Facultad de Ingeniería
Departamento de Ingeniería
Industrial

Profesora Patrocinante
Dr. Lorena Pradenas

“PREDICCIÓN DE LA TEMPERATURA FUERA DEL RANGO DE CONTROL, DE LA CALDERA DE GASES EN HORNO DE FUSIÓN FLASH DE CONCENTRADO DE COBRE”

Esteban Alfonso Saravia Riffo

Informe de Tesis,

para optar al Grado de

MAGISTER EN INGENIERÍA INDUSTRIAL

Enero, 2022

Agradecimientos

Primeramente, agradecer a mi familia por haberme apoyado en este desafío desde principio a fin, espero haber estado a la altura de sus expectativas. Siempre serán mi motivación para seguir creciendo como persona y como profesional.

Agradecer a mis profesores Roberto Parra, quien me motivo para seguir estudiando y especialmente a Roberto Parada, por darme su confianza y apoyarme en todo el proceso de tesis, un abrazo para ambos. A mi profesora patrocinante, Lorena Pradenas, quien me ha tenido una paciencia increíble y se dio el tiempo de ser mi guía. También, al profesor Víctor Parada quien me dio el último empujón, con sus conocimientos.

Agradecer a Pablo Urzua y Leandro Meneses, quienes fueron los ingenieros de la fundición Chuquicamata, que me entregaron toda la información y apoyo teórico en el desarrollo del problema, eternamente agradecido de la oportunidad que me dieron.

Finalmente agradecer a la persona que se volvió mi compañera de vida, quien ha soportado mis peores momentos en el desarrollo de este estudio y también los mejores. Gracias por motivarme cuando más lo necesitaba. Este es un pequeño escalón dentro de una escalera inmensa que planeamos subir juntos.

A todos ustedes, muchas gracias.

Amulepe Taíñ Weichan!

Resumen

La temperatura de la primera pantalla en la caldera de gases, de un horno de fusión flash de concentrado de cobre, es una variable relevante dentro del control de proceso. Un aumento prolongado de esta puede provocar daños estructurales en la caldera como en equipos posteriores, baja calidad de los productos, entre otros.

El control operacional de la temperatura se basa en aumentar la recirculación de polvos y aumento de carga fría, bajar los flujos de oxígeno, entre otros. Sin embargo, el horno flash es una caja negra compleja que tiene muchas variables operacionales y estocásticas de entrada además, se hace imposible tomar mediciones en línea a mitad del proceso. Por lo que generar una predicción resulta complejo.

En este estudio, se presenta un modelo de predicción de la temperatura en la primera pantalla basado en algoritmos de *Machine Learning (ML)* utilizando, variables y factores que tienen mayor influencia en la temperatura como entrada del modelo predictivo. Se estudiaron diversos algoritmos predictivos: regresión logística, árboles de decisión, bosques aleatorios y máquinas de soporte vectorial. Se presenta un flujo de trabajo paso a paso sobre cómo tratar el conjunto de datos y como encontrar los mejores modelos variando, los hyperparametros del algoritmo y la selección, final del modelo predictivo.

En particular, los bosques aleatorios logran las mejores métricas de pronóstico con un *accuracy* del 82%. Por lo tanto, los modelos de *ML* permiten, modelar y predecir correctamente la temperatura de la primera pantalla. La metodología expuesta tiene potencial para extenderse a cualquier otro conjunto de datos y objetivos, en la industria de fundición.

Abstrac

In the gas boiler of a copper concentrate flash melting furnace, the temperature of the first screen is a relevant variable to the control process. A prolonged increase in this parameter can cause structural damage to the boiler and subsequent equipment, low product quality, among others.

The operational control of the temperature, increasing the recirculation of powders and increasing the cold load, lowering the oxygen flows based, among others. However, the flash furnace is a complex black box, has many operational and stochastic input variables, and makes it impossible to take mid-process online measurements. So, generating a prediction is a complex process.

In this study, a prediction model for first-screen temperature, at Machine Learning (ML) algorithms based, is presented using variables and factors with the greatest influence on the temperature in the input to the predictive model.

Various predictive algorithms: logistic regression, decision trees, random forests, and support vector machines were studied. A step-by-step workflow on how to treat the dataset and find the best models by varying the algorithm hyperparameters and the final selection of the predictive model is exposed.

In particular, random forests achieve the best forecast metrics with an accuracy of 82%. Therefore, ML models allow, correctly model, and predict the temperature of the first screen. The methodology exposed has the potential for any other set of data and objectives in the foundry industry to be extended.

Índice

1. Introducción y planteamiento de problemas....1
1.1 Contexto....1
1.2 Descripción del problema....3
2. Revisión de la literatura....4
2.1 Industria del cobre....5
2.2 Industria del acero....6
2.3 Industria de otros metales....7
3. Hipótesis y objetivos....9
3.1 Hipótesis del estudio....9
3.2 Objetivos....9
4. Proceso de fusión y proyecto potenciamiento Horno Flash DCh....10
4.1 Proceso de fusión de concentrado de cobre....10
4.2 Proceso de fusión Chuquicamata y proyecto potenciamiento....12
5. Métodos de *Machine learning*....15
5.1 Machine Learning definiciones....15
5.2 Métodos de *Machine learning* aplicados en esta investigación....16
5.2.1 Regresión logística....16
5.2.2 Máquinas de soporte vectorial (SVM)....17
5.2.3 Arboles de decisión....19
5.2.4 Bosques aleatorios....21
6. Metodología....22
6.1 Descripción general....22
6.2 Limpieza y preprocesamiento de datos....24
6.3 Selección datos de entramiento, validación y prueba....28
6.4 Entrenamiento, validación cruzada y prueba....28
7. Resultados....30
8. Conclusiones y Recomendaciones....37
8.1 Conclusiones....37
8.2 Recomendaciones....38
9. Anexos....40
9.1 Anexo 1: Histogramas por grupos, antes y después de limpieza....40

9.2 Anexo 2: Matriz de correlación por grupos..... 41
10. Bibliografía..... 43

Índice de figuras

Figura 1 diagrama general proceso fusión. Fuente: elaboración propia 10
Figura 2 esquema de insumos y productos del proceso de fusión de concentrado de cobre. Fuente: COCHILCO 11
Figura 3 diagrama del tren de gases Fuente: COCHILCO 11
Figura 4 diseño caldera. Fuente: CODELCO..... 12
Figura 5 diagrama fundición DCh antes del proyecto de potenciamiento. Fuente: CODELCO..... 13
Figura 6 diagrama fundición DCh después del proyecto de potenciamiento. Fuente: CODELCO 14
Figura 7 método general de aplicación de machine learning. Fuente: Géron, 2017 15
Figura 8 modelo regresión logística. Fuente: cienciadedato.net 17
Figura 9 modelo Support vector machine. Fuente: scikit-learn 18
Figura 10 aplicación del kernel en SVM. Fuente: www.javatpoint.com 19
Figura 11 modelo árbol de decisión. Fuente: aprendeia.com..... 20
Figura 12 modelo bosques aleatorios. Fuente: aprendeia.com..... 21
Figura 13 metodología general del estudio..... 23
Figura 14 matriz de confusión. Fuente: elaboración propia 23
Figura 15 histograma y grafico de caja, temperatura primera pantalla y tasa de fusión antes y después de limpieza. Fuente: elaboración propia..... 30
Figura 16 matriz de correlación entre las variables de alimentación. Fuente: elaboración propia..... 31
Figura 17 importancia de variables en modelo bosques aleatorios 36
Figura 18 histograma grupo alimentación, antes y después limpieza 40
Figura 19 histograma grupo quemador, antes y después limpieza 40
Figura 20 histograma grupo caldera, antes y después limpieza 41
Figura 21 matriz correlación grupo alimentación 41
Figura 22 matriz correlación grupo caldera 42

Índice de tablas

Tabla 1 Resumen revisión bibliográfica y GAP del conocimiento 8
Tabla 2 categorías variable objetivo 24
Tabla 3 variables grupo alimentación 25
Tabla 4 variables grupo quemador 25
Tabla 5 variables grupo caldera 25
Tabla 6 variables grupo precipitador electrostático (PP.EE) 26
Tabla 7 variables grupo eje-escoria 26
Tabla 8 variables grupo cámara de mezcla 26
Tabla 9 hyperparametros regresión logística 29
Tabla 10 hyperparametros máquinas de vectores de soporte 29
Tabla 11 hyperparametros árbol de decisión 29
Tabla 12 hyperparametros bosques aleatorios 29
Tabla 13 Correlación con la variable objetivo 32
Tabla 14 variables finales para generación de modelo 33
Tabla 15 evaluación entrenamiento 33
Tabla 16 matriz de confusión entrenamiento 34
Tabla 17 hyperparametros óptimos regresión logística 34
Tabla 18 hyperparametros óptimos árbol de decisión 34
Tabla 19 hyperparametros óptimos bosques aleatorios 34
Tabla 20 hyperparametros óptimos SVM 34
Tabla 21 evaluación testeo 35
Tabla 22 matriz de confusión testeo 35

1. Introducción y planteamiento del problema

1.1 Contexto

Para el año 2020 la industria del cobre representó, cerca del 15% del PIB chileno, el 60% de las exportaciones y el 20% de los ingresos de Chile provienen, de ésta industria (SERNAGEOMIN, 2021).

La industria de producción de cobre se basa en la extracción y posterior, conversión del mineral en metal fino mediante sucesivos procesos físicos y químicos. El cobre se obtiene a través de dos tipos de procesos: hidrometalúrgicos y pirometalúrgicos dependiendo, del tipo de mineral encontrado. El 80% del cobre fino producido a nivel mundial proviene de procesos pirometalúrgicos, lo que equivale a 42,3 millones de toneladas, al año (COCHILCO, 2021).

La fundición Chuquicamata perteneciente a la División del mismo nombre, del Distrito norte de Codelco es una de las principales y más antiguas fundiciones, de cobre del país, produciendo aproximadamente, 240,000 ton de ánodo al año 2018 (COCHILCO, 2021). En julio de 1988 la fundición inició, la operación de su horno de fusión flash de tecnología *Outotec* como su segunda línea de producción acompañando al convertidor teniente en operación (Brook Hunt & Associates Ltd, 2007). Desde entonces, la capacidad del horno flash ha ido en aumento en desmedro del convertidor teniente convirtiéndose, desde el año 2003 en el principal horno, de la planta metalúrgica (Wood Mackenzie , 2010).

Debido a: las exigencias de las comunidades, los hechos históricos relacionados a contaminación por parte de las industria de fundición y proyectando, el escenario mundial donde, las políticas públicas sobre medio ambiente apuntan a mayores restricciones, el año 2013 en Chile se establecieron, nuevas normativas ambientales aplicables a fundiciones de cobre, como es el D.S N° 28/2013 del Ministerio de Medio Ambiente que “Establece Norma de Emisión para Fundiciones de Cobre y Fuentes Emisoras de Arsénico” y que consisten, principalmente en el aumento de la captura de SO_2 y As al 95% con límite, de emisión de 49.700 ton/año y 476 ton/año respectivamente (Congreso nacional de Chile, 2013).

Para dar cumplimiento a las nuevas normativas ambientales y mejorar, su posición competitiva, la Gerencia de Proyectos de Codelco División Chuquicamata en su plan estratégico desarrolló, un proyecto de racionalización de procesos e instalaciones, denominado “Mejoramiento Fundición con Horno Flash Potenciado DCh.”. Consiste, en alcanzar una capacidad de fusión de 1.170 kton/a de

mezcla de concentrados de minas Chuquicamata y Radomiro Tomic (RT) fase 1 y calcinas, de la División Ministro Hales (DMH). Se estableció, una nueva configuración productiva con una línea única de fusión basada en el Horno Flash el cual, aumenta su capacidad desde 825 a 1.170 kton/a con potencial de lograr 1.400 kton/a dejando, fuera de operación al convertidor teniente (Amec Foster Wheeler, 2018).

El alcance del proyecto incluye: una nueva torre enchaquetada del horno; modificaciones en su *up-take* y caldera; mejoras en los sistemas de alimentación de cargas; manejo de polvos y sistemas de manejo de gases. Lo anterior, permitiría lograr una optimización de los costos de operación, incremento de la productividad laboral y evitar, las inversiones de aquellas instalaciones, que detienen su operación (Amec Foster Wheeler, 2018).

Completado el proyecto, pasaron dos meses de operación, con baja tasa de alimentación debido al retraso de las modificaciones, en la planta de ácido. Posteriormente, se notifica de desviaciones en la temperatura, de la primera pantalla en la caldera la cual, alcanzó valores sobre los 1000°C y según, criterios de diseño la temperatura máxima es de 920°C. El problema más importante, es que puede provocar diversos daños en la estructura de la caldera y en todos, los equipos posteriores desencadenando, en una posible y muy probable parada de operación implicando, menor producción por consiguiente, menos ingresos además, del gigantesco recurso humano requerido, para reparar y reiniciar operación.

El control operacional de la temperatura de la primera pantalla, según la Superintendencia de control de proceso se basa en el siguiente, programa estándar:

1. Aumentar la recirculación de polvos metalúrgicos.
2. Cuando la recirculación de polvos se encuentra en su máximo permitido, el siguiente paso es aumentar, la carga fría.
3. Cuando ya no es posible aumentar, la recirculación de polvo ni la carga fría se disminuye, el enriquecimiento de oxígeno aumentando, el flujo de aire del proceso con el fin de, aportar nitrógeno al flujo de gas y disminuir, la temperatura de los gases generados.
4. Si no se puede aumentar el flujo de aire y si, operacionalmente el proceso lo permite, se debe, bajar el coeficiente de oxígeno disminuir, el oxígeno del quemador. Esto implica, una disminución en la ley del eje generando, pérdida de cobre.
5. Si no se puede ejecutar ninguna de las opciones anteriores, se procede a bajar la tasa de alimentación provocando, disminución de la producción de cobre y por lo tanto, menores ingresos para la fundición.

El control de la temperatura en la primera pantalla es, por lo tanto, un tema de investigación importante, para la Superintendencia de control de proceso. La tasa de alimentación y el flujo de oxígeno, son las variables controlada de mayor interés y determinan, en gran medida la temperatura de los gases generados (Gaskell, 2003). Sin embargo, es común en el control de procesos industriales, que existan otras variables a considerar para asegurar, un adecuado funcionamiento de la planta.

1.2 Descripción del problema

La caldera es el primer equipo en recibir, los gases generados en el horno de fusión flash. Los problemas relacionados, con el control de temperatura se debe a que, el proceso no es lineal por lo tanto, la relación entre las variables de entrada y salida cambian, significativamente entre diferentes regímenes de operación además, cuando la alimentación y los flujos de oxígenos ingresan, hasta que los gases pasan por la primera pantalla existe, un tiempo de retraso que genera un impacto en el control de proceso y perdida en el rendimiento y/o en la calidad, del producto obtenido.

La principal complejidad que tiene la operación, es la imposibilidad de realizar mediciones intermedias a variables, del proceso que permitan identificar la eficiencia además, de un comportamiento sujeto a perturbaciones dinámicas y el cambio estocástico, en las propiedades de la materia prima. Como consecuencia, el control del proceso de la temperatura de la primera pantalla es limitado, a correcciones en la medida que la variable de salida, es obtenida.

Este estudio se centra, en un modelo predictivo de la temperatura en la primera pantalla que, está expuesta a todas las dificultades anteriormente mencionadas y de manera constante. Una vez que se compruebe, que este modelo predictivo obtiene una estimación acertada bajo algún criterio, será útil para disminuir la variabilidad de la temperatura, en la primera pantalla y encontrar, las condiciones óptimas, para las variables de entrada.

El primer paso para enunciar una hipótesis de investigación, que pueda ser probada, medida y validada consiste, en comprender las técnicas y enfoques actuales, de modelos predictivos.

2. Revisión de la literatura

En esta sección, se presenta una revisión de la literatura sobre modelos y control de temperatura de caldera, sus principales investigaciones y desarrollos además, los distintos algoritmos de *machine learning* aplicados, a la industria minero-metalúrgica.

La revisión se realizó en el buscador de *Clarivate Web Of Science*, las palabras claves utilizadas para la búsqueda son: “Control temperatura”, “Horno de fusión flash”, “Temperatura de caldera”, “Control temperatura caldera”, “*Machine Learning*”, “Modelos predictivos”. Los filtros utilizados, para la búsqueda fueron por categorías: “Metalurgia Ingeniería Metalúrgica”, “Mineralogía”, “Minería Procesamiento de Minerales” y por años de publicación: “1990 hasta, el 2021”.

La temperatura del flujo de gas esta predeterminada, en gran parte mediante el balance de materia y energía termodinámico. La mayor parte de los estudios, sobre modelos de control y simulación de caldera, se basan en este principio.

Así lo investigo Yang, 1996 donde, simulo el comportamiento del flujo de gas y la transferencia de calor, en la caldera del proceso de fundición flash de cobre, el modelo utilizado para la dinámica del fluido, fue un código comercial llamado *Phoenics*. Un estudio más actualizado, de Bezuidenhout, Yang, & Ekteen, 2008 también, desarrolló una simulación del comportamiento del fluido de gas en una caldera de calor residual, de una fundición flash de metales sulfurados el modelo utilizado, es un paquete CFD comercial, *Fluent 6.2.16*. El estudio de Khoshhal, Rahimi, Ghahramani, & Alsairafi, 2011 también, desarrollaron un modelo dinámico utilizando, el paquete CFD comercial, *Fluent 6.2.16* para el uso de la técnica de combustión de aire a alta temperatura, en una caldera recuperadora de calor, en el complejo petroquímico Fajr, Irán.

El uso, del software HYSYS, para la simulación de una caldera de recuperación de calor ha presentado, buenos resultados para el análisis de la influencia, de parámetros de operación (Montes de Oca, Dominguez, Días, López, & Tápanez, 2017). Otras investigaciones, desarrolladas con modelos CFD que aportan, a la investigación son de: Vázquez, Galindo, Mani, & Rossano, 2010 y Peñalba, 2004 que consideran, modelos matemáticos basados en balance de masa y energía.

Con el desarrollo de la tecnología de la información y la inmensa disponibilidad de datos operativos a recuperar, de forma fluida y en línea, es de interés utilizar registros históricos, para la identificación del sistema (Zhongsheng Hou, 2017). En otras palabras, construir un modelo del proceso a partir de la recopilación de información de entrada-salida generada, sin perturbar el funcionamiento normal. Lo anterior, se denomina, enfoque basado en datos.

El *Machine learning,* es una rama de la informática y la inteligencia artificial que trata los problemas de; clasificación, regresión, predicción y agrupamiento (Heaton, 2015) y mediante, el uso de conjuntos de datos organizados y algoritmos de entrenamiento se espera, que el programa, reconozca patrones en los datos y genere, una representación del modelo de esa estructura. El *Machine learning* es una buena alternativa, para la generación de un modelo utilizando datos operativos de un proceso industrial. Y es el principal foco del presente estudio.

Al realizar la búsqueda de modelos predictivos, asociados a calderas en conjunto con las herramientas de *machine learning* no se dispone aún, de investigaciones robustas pero, existen estudios como el de Dhanuskodi, y otros, 2015 que utiliza redes neuronales artificiales para predecir la temperatura, de la pared de calderas supercríticas y el estudio, de Tavoosi & Mohammadzadeh, 2021 que presenta un modelo de control predictivo, basado en red neuronal de funciones de base radial (RBFN-MPC) para controlar, la temperatura del vapor de una caldera de planta de energía.

A continuación, se detalla una revisión bibliográfica de los métodos, de *machine learning* aplicados a la industria minero-metalúrgica, estado del arte de esta herramienta aplicada a esta industria. A partir de las investigaciones encontradas sobre modelos de *machine learning,* estas se pueden clasificar en tres grupos: grupo pertenecientes a la industria del cobre, a la industria del acero y otras industrias de metales.

2.1 Industria del cobre

En la industria del cobre, la etapa de explotación y concentración, de mineral tienen la mayor investigación en *machine learning*. Modelar, optimizar y controlar, la recuperación metalúrgica ha sido el objetivo de la mayor parte, de las investigaciones.

Los algoritmos de clusterización fueron utilizados, por van Duijvenbode, Buxton, & Soleymani, 2020 y Lishchuk, Lund, & Yousef, 2019 para generar un modelo de optimización y mejoramiento, de la recuperación metalúrgica basado en “huellas dactilares” que contienen, variables geometalurgicas. El modelo de regresión lineal múltiple utilizado, por Bascur & Soudek, 2019 también, se basa en variables geometalurgicas para estimar el tamaño de partícula y así mejorar, la recuperación metalúrgica del proceso molienda-flotación.

Las fallas de operación, han sido otro objetivo de investigación. El modelo integrado basado, en redes neuronales convulocionadas (CNN) combinado, con el enfoque de aprendizaje por transferencia y máquina de vectores de soporte (SVM) de Li, Gui, & Zhu, 2019 utilizan, variables geometalurgicas para predecir, condiciones de flotación que indiquen fallas.

Reducir el consumo de energía, en el molino SAG fue estudiado por Avalos, Kracht, & Ortiz, 2020, utilizando: regresión polinomial, K vecinos más cercano, máquina de vectores de soporte, perceptrón multicapa, memoria a largo y corto plazo y unidades recurrentes cerradas, para predecir cuando exista un consumo alto de energía, basado en variables geometalurgicas.

En la industria de fundición, los estudios sobre herramientas de *machine learning* son acotados. El estudio de Schaaf, Gómez, & Cipriano, 2010 desarrolló un modelo empírico de control predictivo (MPC) en base a data operacional, que tiene por objetivo predecir: la concentración de cobre en metal blanco, concentración de magnetita en escoria, temperatura de metal blanco y nivel de metal blanco y escoria. Posterior Wang, 2018 desarrolló, un modelo dinámico que tiene por objetivo predecir parámetros, en línea como: grado de mata, razón Fe/SiO_2 en escoria, temperatura de fusión, nivel total y nivel de mata.

Posteriormente, Marija V, 2015 desarrolló, un modelo estadístico, de análisis de regresión lineal múltiple (MLRA), redes neuronales artificiales (ANN) y el sistema, de interferencia difusa basado en redes adaptativas (ANFIS), con el objetivo de definir correlaciones del grado de perdida de cobre en la escoria con respecto, a variables como el SiO_2, FeO, Fe_3O_4 y Al_2O_3 contenido, en la escoria y en la mata.

2.2 Industria del acero

En la industria del acero, la fusión en hornos eléctricos de alto consumo energético, se ha llevado la mayor parte de la investigación. Así lo demuestra Carlsson, Samuelsson, & Jonsson, 2019 con su estudio, donde utiliza modelos estadísticos para la predicción del consumo de energía eléctrica del horno de arco eléctrico y posteriormente, un modelo estadístico de redes neuronales artificiales (ANN), para predecir el consumo de energía eléctrica, de un horno de arco eléctrico que produce acero inoxidable.

Otros estudios, se basan en la predicción y control operacional del alto horno para la reducción del mineral de hierro. El modelo utilizado por Phull, Egas, Barui, Mukherjee, & Chattopadhyay, 2019, basado, en arboles de decisión y máquinas de vectores de soporte, tiene como objetivo predecir la desfosforación en BOF a través, de la relación de partición $(\%P_{escoria}/\%P_{acero})$.

La temperatura de los productos obtenidos, es otra variable de investigación. Así lo señala Bae, Li, Stahl, & Mathiason, 2020, con un modelo estándar de *machine learning* basado, en máquinas de vectores de soporte y redes neuronales, para predecir la ley del metal blanco y la temperatura del metal y de la escoria.

2.3 Industria de otros metales

En la industria de otros metales y minerales, la industria del aluminio tiene ventaja. Así lo demuestra Moon-Jo, Jong, Ji-Ba-Reum, Seung-Jun, & DongEung, 2020 con modelos, de *machine learning*: regresión lineal, árboles de regresión, regresión de proceso gaussiano, máquina de vectores de soporte (SVM) y conjuntos de árboles de regresión, para predecir la temperatura del aluminio líquido, en una fundición sobre variables operacionales

Mingwei, y otros, 2021 analizan, mediante un modelo de *machine learning* y modelo de máquina de vectores de soporte (SVM) la predicción, de propiedades mecánicas en las aleaciones de Al forjado, basado en variables operaciones.

Debido a la complejidad del sistema a representar: el número de variables involucradas; las perturbaciones operacionales no anticipadas y la necesidad, de resolver el problema en el menor tiempo posible para el control del proceso y en base a, la literatura revisada se decide, usar herramientas de *Machine learning* para la predicción de la temperatura en la primera pantalla.

De la revisión bibliográfica encontrada, se genera la tabla 1, que muestra los estudios encontrados, sus principal objetivo y tipo de modelos utilizados, en cada uno de estos.

Tabla 1 Resumen revisión bibliográfica y GAP del conocimiento

Paper	Características											
	Objetivos					Modelos						
	Recuperación metalúrgica	Consumo energía eléctrica	Temperatura calderas	Fallas operacionales	Propiedades físicas/químicas	Modelos CFD	Regresión logística	Árbol de decisión	Bosques aleatorios	SVM	Redes neuronales	Clustering
(Yang, 1996)			✓			✓						
(Bezuidenhout, Yang, & Ekteen, 2008)			✓			✓						
(Khoshhal, Rahimi, Ghahramani, & Alsairafi, 2011)			✓			✓						
(Montes de Oca, Dominguez, Días, López, & Tápanez, 2017)			✓			✓						
(Vázquez, Galindo, Mani, & Rossano, 2010)			✓			✓						
(Peñalba, 2004)			✓			✓						
(Dhanuskodi, y otros, 2015)			✓								✓	
(Tavoosi & Mohammadzadeh, 2021)			✓								✓	
(van Duijvenbode, Buxton, & Soleymani, 2020)	✓											✓
(Lishchuk, Lund, & Yousef, 2019)	✓											
(Bascur & Soudek, 2019)	✓						✓		✓	✓		
(Li, Gui, & Zhu, 2019)				✓						✓	✓	
(Avalos, Kracht, & Ortiz, 2020)		✓					✓			✓		
(Schaaf, Gómez, & Cipriano, 2010)	✓					✓						
(Marija, 2015)	✓										✓	
(Bin Wang, 2018)	✓					✓						
(Carlsson, Samuelsson, & Jonsson, 2019)		✓									✓	
(Phull, Egas, Barui, Mukherjee, & Chattopadhyay, 2019)					✓			✓		✓		
(Bae, Li, Stahl, & Mathiason, 2020)					✓					✓	✓	
(Moon-Jo, Jong, Ji-Ba-Reum, Seung-Jun, & DongEung, 2020)					✓				✓	✓		
(Mingwei, y otros, 2021)					✓					✓		
Este estudio			✓				✓	✓	✓	✓		

3. Hipótesis y objetivos

3.1 Hipótesis del estudio

Posterior a un tiempo de operación y dado que la temperatura de la primera pantalla, de la caldera se encuentra fuera de los rango de control, ocurren fenómenos físicos y químicos, en el proceso de fusión y que se reflejan en sus datos de operación. Un algoritmo de *Machine learning* podría caracterizar y modelar esta situación.

3.2 Objetivos

I. Objetivo general

Modelar y predecir la temperatura de la primera pantalla, en la caldera del horno de fusión flash mediante, herramientas de *Machine learning*.

II. Objetivos específicos

a. Revisar, la literatura actualizada del control de temperatura de caldera y herramientas, de *machine learning* aplicadas, a la industria minero-metalúrgica.
b. Estudiar, diferentes técnicas de *Machine learning* adecuadas, para la identificación de sistemas y selección de éstos y el desarrollo de un modelo de predicción de temperatura de caldera.
c. Evaluar, el rendimiento predictivo de los algoritmos de aprendizaje automático seleccionados, mediante simulación con datos, operativos actuales.
d. Generar recomendaciones para la mejora del modelo y potenciar, el control predictivo a otras variables deseadas.

4. Proceso de fusión y proyecto potenciamiento Horno Flash DCh.

Este capítulo, presenta el proceso de fusión de concentrado de cobre y el manejo de los gases generados. Se detalla el diseño de la fundición Chuquicamata y su proyecto de potenciamiento del Horno Hlash.

4.1 Proceso de fusión de concentrado de cobre

La industria de fundición de cobre consiste, en tres grandes etapas; fusión, conversión y refinación, la Figura 1 muestra, el diagrama general del proceso de fusión. La fusión del concentrado es la etapa principal del proceso y se lleva a cabo, en reactores denominados hornos de fusión los cuales funde el concentrado para separar el cobre del resto de elementos que contiene el mineral. Este proceso, se realiza a través de distintas reacciones químicas, que por medio de altas temperaturas generan, dos fases líquidas inmiscibles y una fase de gases (Riveros, 2009).

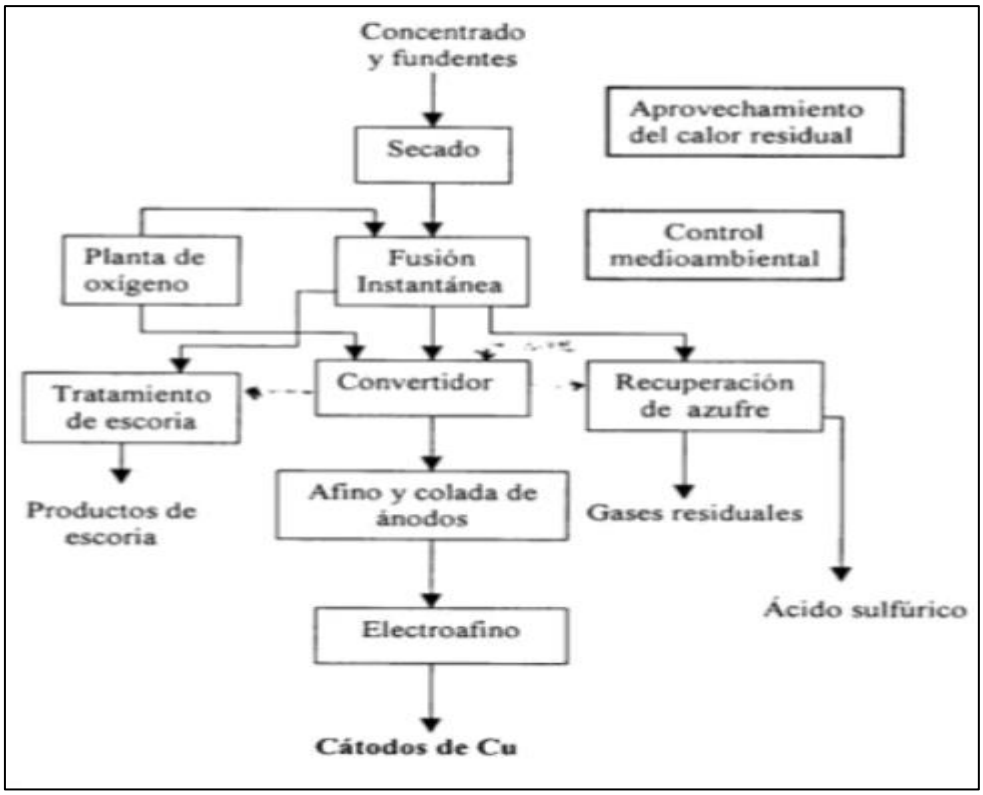

Figura 1 diagrama general proceso fusión. Fuente: elaboración propia

La Figura 2, presenta el esquema de insumos y productos de la fusión de concentrado. La fase liquida de sulfuros, denominada eje es donde, se encuentra el cobre, además, contiene los metales preciosos como: oro, plata e impurezas (arsénico, bismuto y antimonio). La fase liquida oxidada, denominada escoria, es constituida, principalmente por óxidos de hierro y silicio. Las fases liquidas, tienen características distintas entre sí, una de estas, es la densidad que permite la separación mediante gravedad (Gaskell, 2003). La fase de gases, se compone principalmente de dióxido de azufre y polvos de fundición, que son direccionados a la planta de tratamiento de gases, para la producción de ácido sulfúrico y los polvos son recirculados, al horno flash.

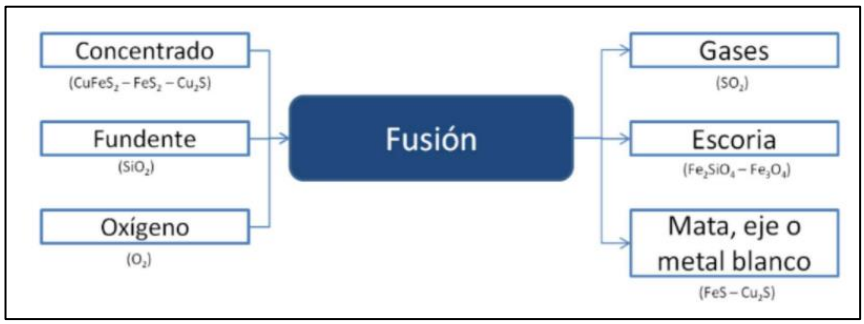

Figura 2 esquema de insumos y productos del proceso de fusión de concentrado de cobre. Fuente: COCHILCO

El control de los gases generados es importante ya que, las condiciones de: flujo, temperaturas y presiones están reguladas, por las operaciones aguas abajo. La caldera, el precipitador electroestático, la cámara de mezcla y la planta de ácido forman, una especie de botella donde el control de temperatura y presión, es crítico. Las Figuras 3 y 4 muestran, el diagrama del tren de gases y el diseño, de la caldera con sus respectivas divisiones.

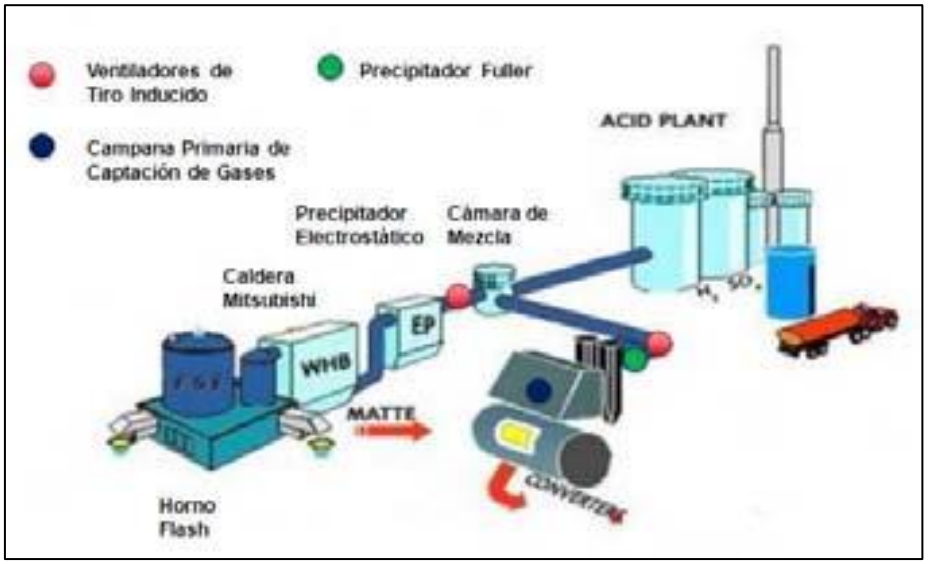

Figura 3 diagrama del tren de gases Fuente: COCHILCO

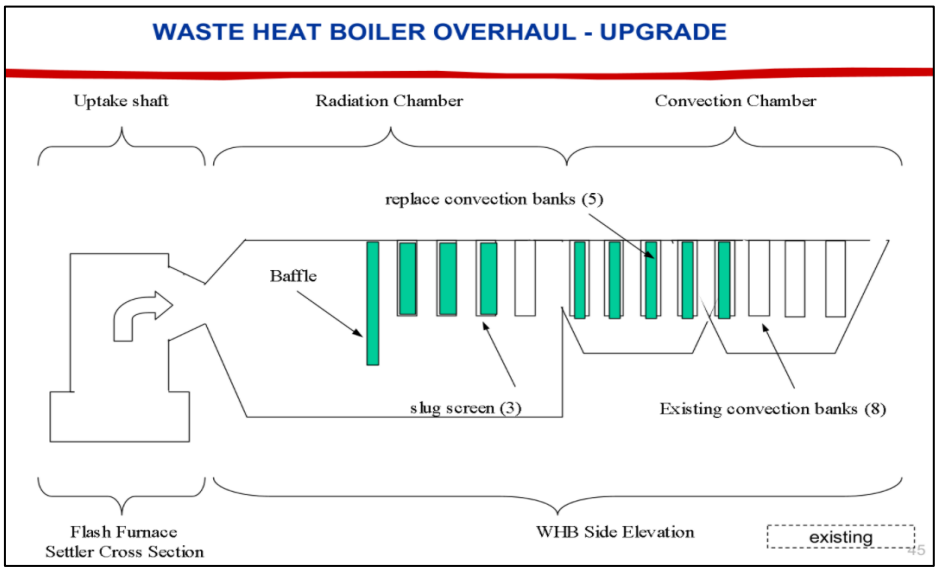

Figura 4 diseño caldera. Fuente: CODELCO

La caldera tiene como principal objetivo la recuperación de calor y se divide en dos zonas: de radiación y convección, en ambas zonas, existen tubos que recuperan el calor pero, en la zona de convección existe una densidad mayor debido al tipo de transferencia de calor. Se le denomina "primera pantalla", a los tubos que se encuentran en la zona de radiación. La temperatura de los gases en la salida del *Uptake* del horno es de 1300°C aproximadamente y en la salida de la caldera es de 350°C.

El precipitador electroestático, tiene por objetivo realizar la limpieza de los gases, capturando los polvos arrastrados por el flujo de gas y enviándolos a su tratamiento. Los gases entran con una temperatura de 350°C al PPEE y salen con 310°C aproximadamente.

En la cámara de mezcla se unen los gases proveniente del horno flash y de los convertidores pierce smith. La temperatura de los gases, en la salida de la cámara de mezcla, no puede superar los 300°C, debido a problemas en la planta de ácido para, la producción de ácido sulfúrico.

4.2 Proceso de fusión Chuquicamata y proyecto potenciamiento

La Figura 5, presenta el diseño y el diagrama de procesos de la fundición Chuquicamata, antes del proyecto de potenciamiento. El concentrado es llevado a la etapa de secado, para obtener una humedad adecuada para la fusión posterior, es enviado al Convertidor teniente (CT) y al Horno flash (HF), para la fusión del concentrado y el eje obtenido es enviado a los Convertidores Pierce Smith (CPS). La escoria es direccionada a la planta de tratamiento de escorias y los gases generados, son

enviados a la planta de ácido. El cobre blíster obtenido, de los CPS es enviado, a los hornos de refino y así finalmente, obtener ánodos de cobre de 99% de pureza.

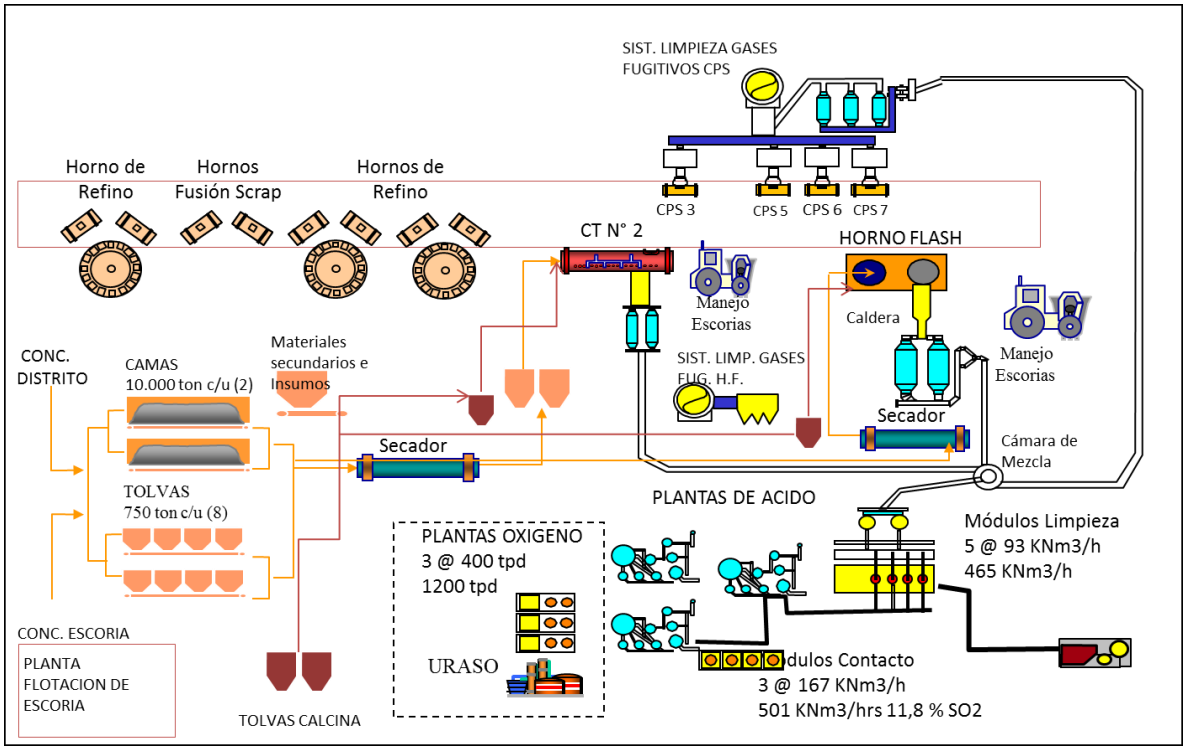

Figura 5 diagrama fundición DCh antes del proyecto de potenciamiento. Fuente: CODELCO

El proyecto realizado, por la empresa *Amec Foster Wheeler* desarrolló un nuevo esquema operacional de la Fundición Chuquicamata que se muestra en la Figura 6. Considera, al Horno Flash potenciado, como único equipo de fusión e incluye, los siguientes equipos y/o instalaciones principales (Amec Foster Wheeler, 2018):

- Un nuevo Secador a Vapor (Secador N° 6), en reemplazo del Secador N°5 existente.
- Un Horno Flash potenciado, para una capacidad de fusión nominal de 1.170 kton/a y capacidad potencial de diseño de 1.400 kton/a.
- Dos Convertidores Peirce Smith (CPS), en soplado simultáneo y 2 CPS en caliente y un quinto CPS, en mantención o stand-by.
- Seis hornos de ánodos, dos de 350 ton y cuatro de 250 ton, que operan, con tres ruedas de moldeo de 45 - 50 ton/h cada una.
- Cuatro plantas de limpieza de gases y dos plantas, de contacto.
- Una planta de tratamiento de efluentes.

Para operar con estas nuevas condiciones se requieren, cumplir restricciones que se detallen a continuación (Amec Foster Wheeler, 2018):

- Para efectos de capacidad, de fusión diaria se considera, que el HF opera: 330 días/año, 23 horas/día, equivalente a 7.590 horas/año.
- El factor de generación de Carga Fría, corresponde a un 12,5% respecto del concentrado fundido incluyendo, la calcina. La tasa máxima, de alimentación será de 300 tpd.
- La recirculación de polvo metalúrgico, del tren de gases del HF al mismo HF está limitada, a sus requerimientos térmicos. Serán recirculados solamente, los polvos metalúrgicos recuperados, de la Caldera del Horno Flash.
- La máxima generación de polvos metalúrgicos, no debe exceder el 6% del total de carga sólida alimentada, al horno.
- Las escorias del HF y la de soplado a escoria de CPS serán, procesadas mediante molienda-flotación obteniendo, concentrado de escoria que se alimenta, en su totalidad al HF junto a la mezcla de concentrados frescos (Chuquicamata y RT) y calcina.

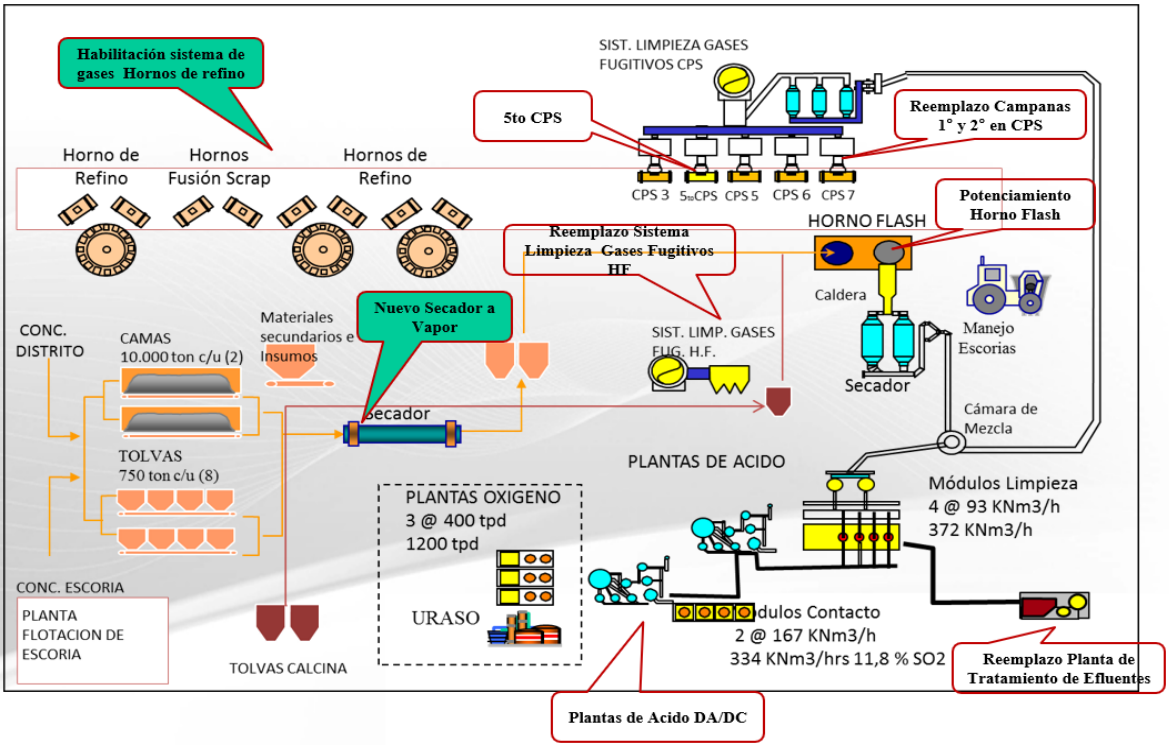

Figura 6 diagrama fundición DCh después del proyecto de potenciamiento. Fuente: CODELCO

El 15 de noviembre del 2018, se detiene la operación del horno flash para comenzar con sus arreglos y construcción según, el proyecto de potenciamiento y el 13 de diciembre del 2018 se detiene definitivamente, el convertidor teniente. Después, de seis meses de trabajo y sin producción de la planta. El 11 de mayo del 2019 se inicia nuevamente la operación del horno flash con sus mejoras.

5. Métodos de *Machine learning*

En este capítulo, se define que es *machine learning* y se presenta formalmente las herramientas de *machine learning* a usar: método de aprendizaje automático para la predicción de la temperatura.

5.1 Machine Learning definiciones

Machine learning es la ciencia y el arte de programar computadoras, que puedan aprender de los datos. Otra definición más general: es el campo de estudio, que brinda a las computadores la capacidad de aprender, sin ser programadas explícitamente, (Samuel, A 1959). Y una definición más orientada, a la Ingeniería es: Se dice que un programa de computadora aprende, de la experiencia *E* con respecto a alguna tarea *T* y alguna, medida de desempeño *P.* Si su desempeño en *T*, medido por *P*, mejora, con la experiencia *E*, (Mitchell, T 1997). La Figura 7 muestra el enfoque de la metodología, de las herramientas de *machine learning* (Géron, 2017).

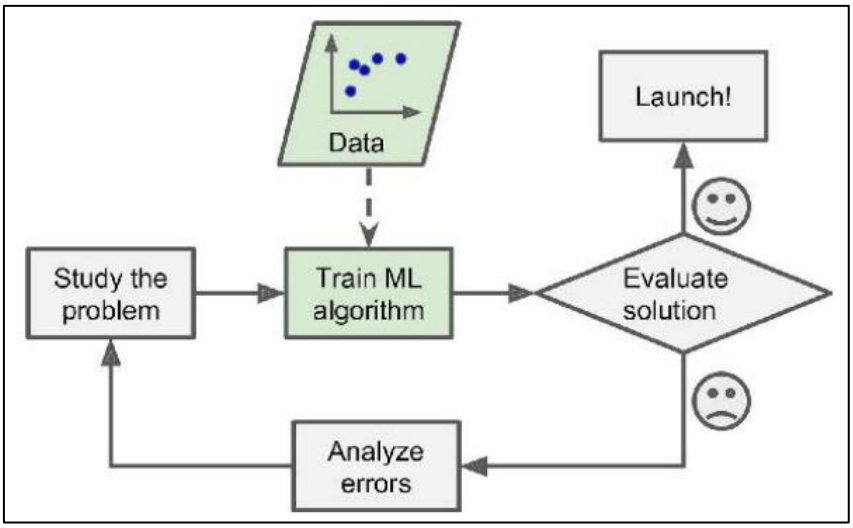

Figura 7 método general de aplicación de machine learning. Fuente: Géron, 2017

Los modelos de *machine learning,* se clasifican en (Géron, 2017):

- Supervisados: se entrenan (o aprenden) a partir de conjuntos de datos, que están asociados con un *target or label*.
- No Supervisados: se entrenan a partir de *sets* de datos, que no están asociados con un *target* o *label*. El sistema aprende, sin instructor.

Los modelos supervisados se pueden dividir en dos tipos: los de clasificación y de regresión. La diferencia corresponde en el tipo de resultado deseado, que queremos que la técnica de *machine learnign* produzca (Heras, 2020).

Cuando usamos modelos de clasificación, el resultado es una clase entre, un número limitado de clases. Ejemplos, de estos problemas son ¿comprará el cliente este producto? [si, no], ¿tipo de tumor? [maligno, benigno], ¿la temperatura estará, sobre el rango de control? [si, no], etc. Cuando usamos regresión, el resultado es un número. Es decir, el resultado será, un valor numérico dentro, de un conjunto infinito de posibles resultados. Ejemplos de estos podemos destacar predecir, por cuánto se va a vender una propiedad inmobiliaria predecir, cuánto tiempo va a permanecer un empleado en una empresa, etc. (Heras, 2020).

5.2 Métodos de *Machine learning* aplicados en esta investigación

Debido a la naturaleza y planteamiento del problema en estudio se considera, la *categoría de aprendizaje supervisado y de clasificación*. A continuación, se detallan los modelos de *machine learning* utilizados, en esta investigación.

5.2.1 Regresión logística

La regresión logística, utilizada en el núcleo del método, la “función *Sigmoide*”. Esta función es una curva en forma de *S* que puede tomar, cualquier número de valor real y asignar, un valor entre 0 y 1 (Gonzalez, 2019). Si la curva, va a infinito positivo la predicción se convertirá en 1 y si la curva pasa al infinito negativo, la predicción se convertirá, en 0. Si la salida de la función *Sigmoide,* es mayor a 0.5 podemos clasificar el resultado como 1. Si es menor que 0.5 podemos clasificarlo como 0. Si el resultado es 0.75 podemos, decir en términos de probabilidad que existe un 75% de probabilidades que el paciente, sufra de cáncer. La Figura 8, muestra la interpretación geométrica del modelo, para un problema binario.

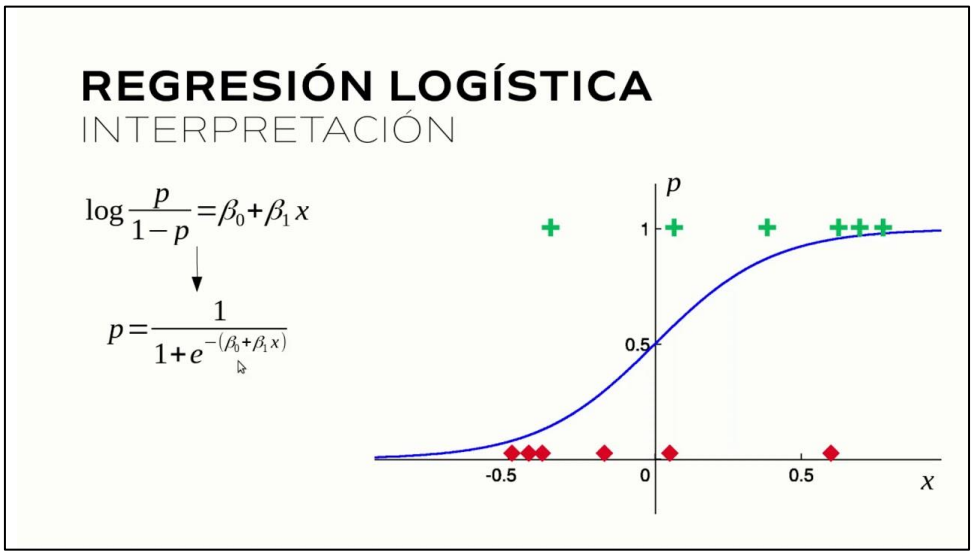

Figura 8 modelo regresión logística. Fuente: cienciadedato.net

Matemáticamente, el algoritmo se formula de la siguiente manera. La regresión lineal, tiene como función $y = a_1x_1 + a_2x_2 + a_nx_n + b$ donde, y es la variable dependiente y, $x_1, x_2 \dots x_n$ son, variables independientes. Por su parte, la ecuación de la función *Sigmoide,* es la siguiente: $p = \frac{1}{1+e^{-y}}$. Entonces, si aplicamos la función Sigmoide a la regresión lineal queda (De la Fuente Fernandez, 2015):

$$p = \frac{1}{1 + e^{-(a_1x_1+a_2x_2+\cdots+a_nx_n+b)}} \quad 5.1$$

El aprendizaje, se realiza con optimización numérica. No existe ninguna fórmula que nos de los coeficientes óptimos $a_{1,2\dots,n}$, sino que, tenemos que estimarlos. La función costo, que se optimiza es la siguiente:

$$J = \frac{1}{m}\sum_{i=1}^{m}\left[-y^i \ln\left(p(x^i)\right) - \left(1-y^i\right)\ln\left(1-p(x^i)\right)\right], \quad 5.2$$

$$x^i = x_1^i, x_2^i, \dots, x_n^i, \quad m = cantidad\ de\ datos$$

5.2.2 Máquinas de soporte vectorial (SVM)

Las máquinas de soporte vectorial ofrecen, una precisión muy alta en comparación con otros clasificadores como, la regresión logística. El SVM, por sus siglas en inglés, construye un hiperplano en un espacio multidimensional, para separar las diferentes clases. La idea central de SVM es

encontrar un hiperplano marginal máximo que mejor, divida el conjunto de datos, en clases (Gonzalez, 2019). La Figura 9, muestra la descripción geométrica del modelo.

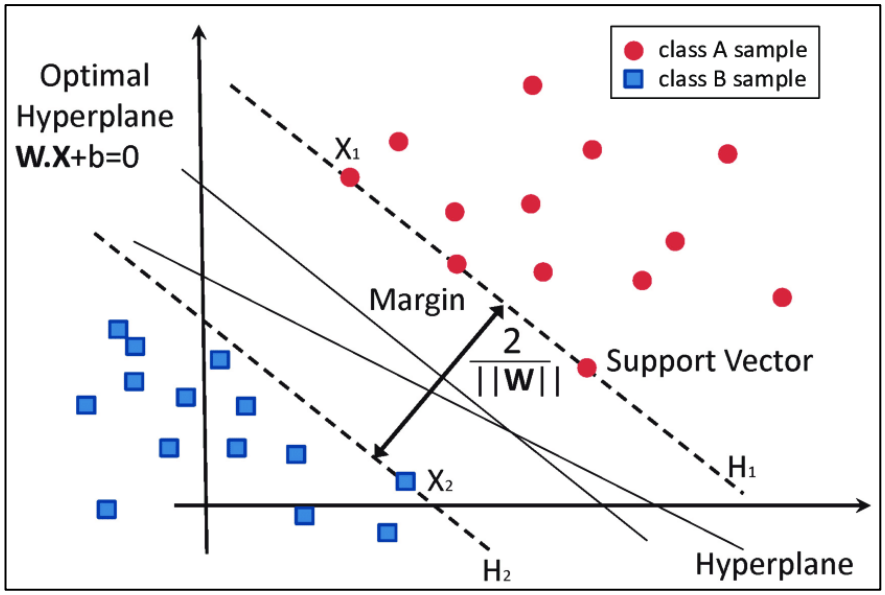

Figura 9 modelo Support vector machine. Fuente: scikit-learn

Cada uno de los elementos que compone esta figura son:

- Vectores de Soporte: son los puntos de datos más cercanos, al hiperplano. Estos puntos definen, la línea de separación calculando los márgenes.
- Hiperplano: es un plano de decisión que separa, entre un conjunto de objetivos que tienen membresía de clases diferente.
- Margen: es un espacio entre, dos líneas de los puntos más cercanos, de la clase. Se obtiene como, la distancia perpendicular desde, la línea hasta los vectores de soporte o puntos, más cercanos.

Matemáticamente, el algoritmo se formula de la siguiente manera: Supongamos que tenemos un problema de clasificación donde, la variable a predecir es binaria y tenemos n casos, de entrenamiento (x_i, y_i) para, $i = 1,2, \dots, n$ donde, $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$. Asumimos, que $y_i \in \{-1,1\}$ denota la etiqueta de clase. El hiperplano optimo, se puede escribir como: $w \cdot x + b = 0$ donde, w y b son parámetros del modelo (Vojslav, 2005).

- Teorema: w es perpendicular, al vector generado
- Teorema: El margen d, se puede calcular como $d = \frac{2}{||w||}$
- El problema de optimización, se representa como:

$$\min_{w} \frac{||w||^2}{2} \qquad 5.3$$

$$sujeto\ a: y_i(w \cdot x_i + b) \geq 1\ para\ i = 1,2,\dots,n$$

En el ejemplo, mostrado en la Figura 9, se resuelve con un hiperplano lineal pero, no siempre, esto se puede resolver fácilmente debido a que, los problemas reales tienen más de dos dimensiones o no son fáciles de separar, por lo tanto, aplicar un hiperplano lineal no siempre es posible. En tales situaciones, el algoritmo utiliza un estrategia del núcleo para transformar el espacio de entrada, en una especio dimensional superior como, se muestra en las Figura 10.

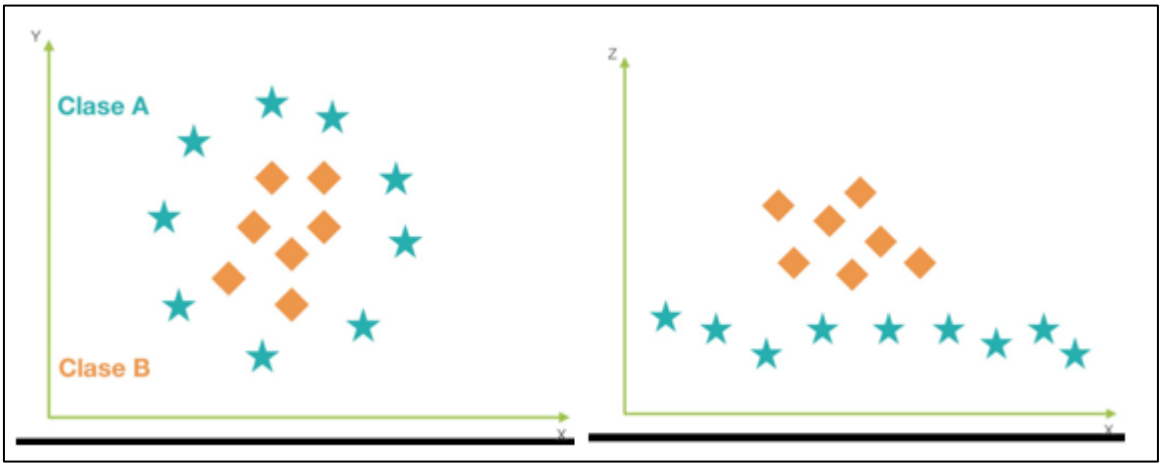

Figura 10 aplicación del kernel en SVM. Fuente: www.javatpoint.com

En la Figura 10 de la izquierda, se observa que no se puede utilizar un hiperplano lineal. Aplicando la estrategia del núcleo los puntos de datos se grafican en los ejes "x" y "z", en donde $z = x^2 + y^2$. Observado la imagen de la derecha, se puede separar fácilmente éstos puntos utilizando, la separación lineal. A este estrategia se le conoce como *kernel*. Un *kernel* transforma un espacio de datos de entradas en la forma requerida (Vojslav, 2005).

5.2.3 Arboles de decisión

Los árboles de decisión, son uno de los algoritmos de *machine learning* más usados, dado que son fácilmente visible para que un humano pueda entender lo que está sucediendo. La Figura 11, muestra el diseño de un árbol de decisión. Este tiene una estructura similar a un diagrama de flujo donde un nodo interno representa una característica o atributo, la rama representa, una regla de decisión y cada, nodo u hoja final representa el resultado. El nodo superior, de un árbol de decisión se conoce, como nodo raíz (Gonzalez, 2019).

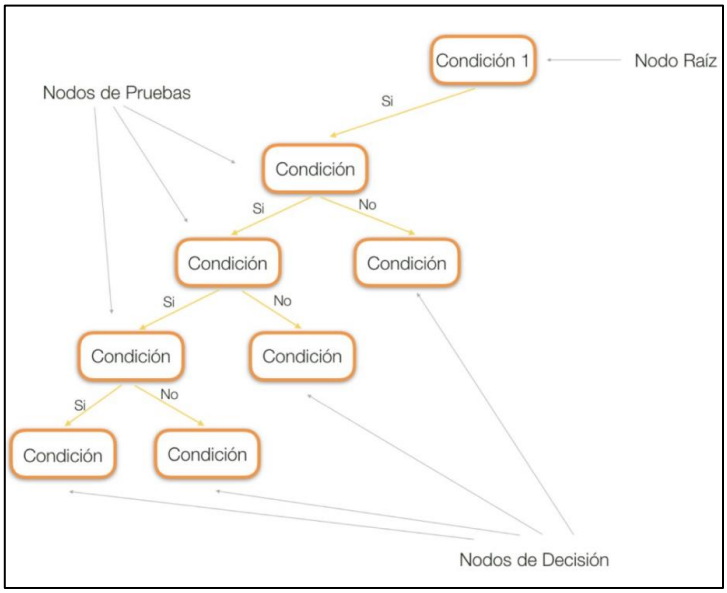

Figura 11 modelo árbol de decisión. Fuente: aprendeia.com

La idea básica, detrás de cualquier problema de árbol de decisión, es la siguiente:

Sea D_t el conjunto de registros, de entrenamiento en un nodo t dado. Sea $y_t = \{y_1, y_2, \dots, y_n\}$ el conjunto de etiquetas de las clases

Muchos algoritmos, usan una versión con un enfoque "*top-down*" o *dividir* y *conquistar*" conocido como, *Algoritmo de Hunt*. Si todos los registros D_t pertenecen, a la misma clase y_t, entonces t es, un nodo hoja, que se etiqueta como y_t. Si D_t contiene, registros que pertenecen a más de una clase se selecciona una variable atributo para dividir los datos en dos subconjuntos, más pequeños. Recursivamente se aplica, el procedimiento a cada subconjunto hasta que:

- Todas las variables pertenecen, al mismo "valor de atributo"
- Ya no quedan, más atributos
- No existen, otros casos.

La selección de atributo es una heurística que selecciona el criterio separando, los datos de la mejor manera posible. Esta medida, proporciona un rango a cada característica explicando, el conjunto de datos dados. El atributo de mejor puntuación, se selecciona como atributo, de división. En el caso de un atributo de valor continuo también, es necesario definir puntos de división, por las ramas. Las medidas de selección, más populares son: Error de Clasificación, el índice de *Gini* y la Entropía (Mitchell, 1997).

$p(j|t)$ = La probabilidad de pertenecer, a la clase j en el nodo t.

$$Error\ de\ clasificación: Error(t) = 1 - \max_j [p(j|t)] \quad 5.4$$

$$Índice\ de\ Gini: GINI(t) = 1 - \sum_j [p(j|t)]^2 \quad 5.5$$

$$Entropía: Entropía(t) = -\sum_j p(j|t) \log_2 p(j|t) \quad 5.6$$

5.2.4 Bosques aleatorios

Es el algoritmo más flexible y fácil, de usar. Un bosque está compuesto de árboles. Se dice que cuantos más árboles existan más robusto, será el bosque. Los bosques aleatorios, crean árboles de decisión a partir, de muestras de datos seleccionadas al azar, obteniendo predicciones, de cada árbol y selecciona la mejor solución mediante votación (Gonzalez, 2019). La Figura 12, muestra la estructura de un bosque aleatorio.

Técnicamente, es un método de conjunto basado, en el enfoque de *dividir* y *conquistar,* de árboles de decisión generados, en un conjunto de datos divididos al azar. Los árboles de decisión individual se generan utilizando, un indicador de selección de atributos como: la ganancia de información, la relación de ganancia y el índice de *Gini*, para cada atributo. Cada árbol, depende de una muestra aleatoria, independiente. Funciona en tres pasos (Gonzalez, 2019):

- Construir un árbol de decisión para cada muestra y obtener, un resultado de predicción de cada árbol de decisión.
- Realizar, una votación por cada resultado previsto.
- Seleccionar, el resultado de la predicción con más votos, como predicción final.

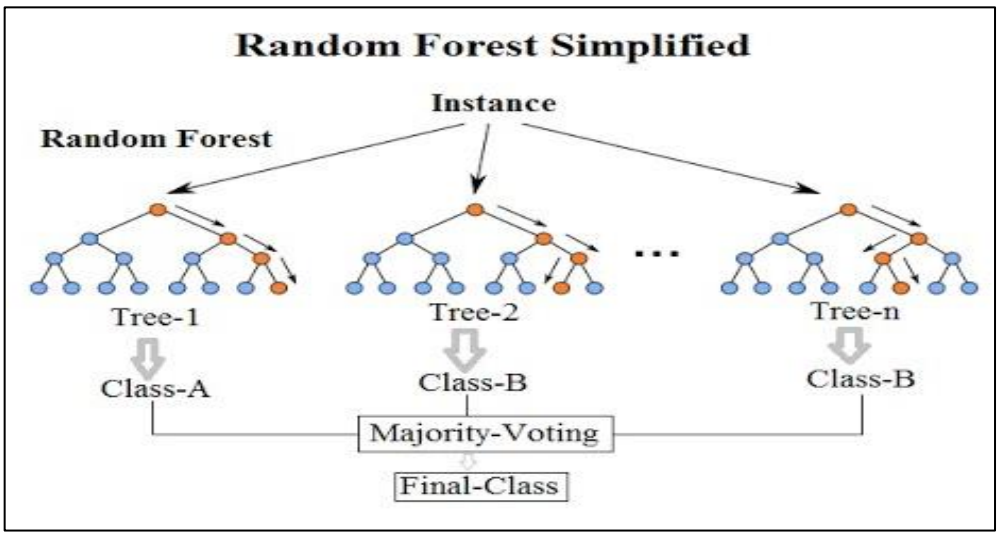

Figura 12 modelo bosques aleatorios. Fuente: aprendeia.com

Los árboles de decisión profundos, pueden sufrir de sobreajuste pero, los bosques aleatorios evitan el sobreajuste creando árboles en subconjuntos aleatorios. Los árboles de decisión, son computacionalmente más rápidos (Géron, 2017).

6. Metodología

En este capítulo, se describe la metodología que es usada en este estudio. Este método, tiene como objetivo analizar e identificar las principales variables, que participan en el proceso basándose ,en los datos históricos de la planta y así, construir el modelo predictivo mediante los algoritmos de *machine learning*. Para ésto es relevante considerar con un set importante de datos históricos de las variables en cuestión.

6.1 Descripción general

El manejo, tratamiento de la data, y la programación de los algoritmos se realizó en lenguaje de computación *Python 3,* a través del *Jupyter Notebook*.

El estudio, se plantea como un problema de supervisión y de clasificación, por lo tanto, primeramente, la variable objetivo se debe caracterizar en clases. Posteriormente, las variables independientes, deben ser agrupados dentro de su etapa operacional. En la práctica, esto permite un análisis particular para cada grupo.

La información en cada variable tiene datos de alta o baja frecuencia así como también, datos fuera de rango e incluso, que describen eventos de: mantenciones, paradas de emergencia, etc. Siendo necesario realizar una limpieza y pretratamiento de éstos, dejando solo aquellos datos representativos.

Teniendo la “data limpia”, se procede a la separación en datos de entrenamiento, validación y prueba. Con la data de entrenamiento, se entrenarán los algoritmos de *machine learning* seleccionados y se evalúa, para conocer a priori que método entrega mejores resultados. Se realiza una validación cruzada y búsqueda de hyperparametros, con la data de validación con el fin de comprobar, el error medido en el entrenamiento y encontrar los mejores hyperparametros para el modelo. Se vuelven a entrenar, los modelo con la data de entrenamiento ahora, con los mejores hyperparametros encontrados y se evalúa, el modelo con los datos de prueba. Finalmente, se entregan los resultados, conclusiones y sugerencias.

La Figura 13 muestra la metodología a seguir

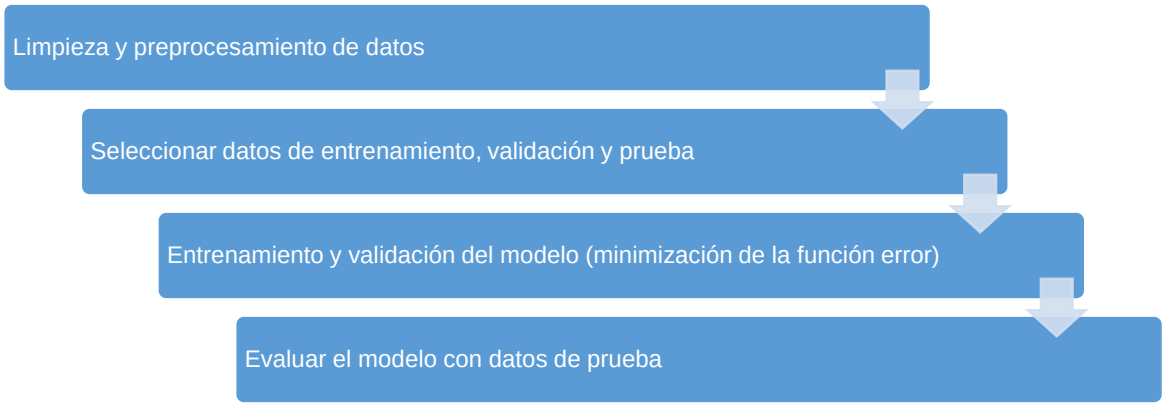

Figura 13 metodología general del estudio

Las métricas de evaluación, para los problemas de clasificación, se describen a continuación:

- Matriz de confusión: es una tabla que describe, el rendimiento de un modelo supervisado de *machine learning*. En la Figura 14, muestra las parte de la matriz
  - *Verdaderos Positivos (VP)*: la clase real del punto es verdadero y la predicha también
  - *Verdaderos Negativos (VN)*: la clase real del punto es falso y el pronóstico también
  - *Falsos Positivo (FP)*: la clase real del punto es falso y el pronóstico verdadero
  - *Falsos Negativos (FN)*: la clase real del punto es verdadero y pronóstico es falso

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Figura 14 matriz de confusión. Fuente: elaboración propia

- *Precision*: es utilizada para saber qué porcentaje de valores que se han clasificados como positivos son realmente positivos.

$$precision = \frac{TP}{(TP + FP)} \qquad 6.1$$

- *Recall*: conocida como el ratio de verdaderos positivos es utilizada, para saber cuántos valores positivos son, correctamente clasificados.

$$recall = \frac{TP}{(TP + FN)} \quad 6.2$$

- *Acurracy*: representa, el porcentaje total de valores correctamente clasificados tanto como positivos o negativos.

$$accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad 6.3$$

- *F*: esta métrica combina el *precision* y el *recall*, para obtener un valor más objetivo.

$$F = \frac{2 \times (recall \times precision)}{(recall + precision)} \quad 6.4$$

6.2 Limpieza y preprocesamiento de datos.

Se recopilan los diagramas de flujo, se identifican los procesos y sus etapas, las variables y principales ecuaciones. Así se tiene una idea, del comportamiento de cada variable sobre el proceso global y el impacto, que pudiesen tener. Las Tablas 3-8 muestran, las variables involucradas en cada grupo.

Los datos proporcionados por la Superintendencia de Control de Operaciones contienen información, desde el 01-03-2021 00:00 hasta el 30-08-2021 23:59, con rango de tiempo en cada un minuto, por lo que se tiene una data con 263.581 registros.

Primeramente, se debe categorizar la variable objetivo debido, al problema de clasificación planteado. La variable objetivo, está representada por la “Temp entrada radiación norte” del grupo Caldera. Para esto, se generan tres rangos de temperatura como, se muestra en la Tabla 2. Estos se definen según lo conversado con personal de operación y estadísticas descriptivas, de la variable. Lo ideal, para la operación es disponer, temperaturas dentro del rango 1

Tabla 2 categorías variable objetivo

Rango	**Categoría**
Bajo 940 °C	0
Entre 940 y 980 °C	1
Mayor a 980 °C	2

Tabla 3 variables grupo alimentación

Variable	Descripción	Unidad
Tasa de fusión	Alimentación concentrado + Calcina	kton
Ley Cu, Fe, S, As, Zn, Sb	Ley de elementos en alimentación	%
Tasa de fusión Line 1 y 2	Alimentación se divide en line (tornillos) 1 y 2	kton
Bin Weigh Line 1 y 2	Nivel de llenado de las tolvas	kton
Velocidad *Line* 1 y 2	Velocidad de los tornillos 1 y 2	RPM

Tabla 4 variables grupo quemador

Variable	Descripción	Unidad
Oxigeno quemador	Oxigeno que ingresa al quemador	Nm3/h
Aire quemador	Aire que ingresa al quemador	Nm3/h
Aire distribución	Aire que distribuye la alimentación en la torre de reacción	Nm3/h
Enriquecimiento	% oxígeno en el flujo total de gas	%
Coeficiente	Cantidad de oxígeno por tonelada de alimentación	Nm3/ton
Velocidad aire proceso	Velocidad a la cual ingresa el aire a la torre de reacción	m/s
Angulo de distribución	Angulo en cómo se distribuye el aire de distribución	°
Flujo de petróleo	Flujo de petróleo que aporta calor al sistema	kg/h
Aire atomización lanza	Aire que ingresa junto al flujo de petróleo	Nm3/h

Tabla 5 variables grupo caldera

Variables	Descripción	Unidad
Presión interna HF	Presión en el *settler* del horno flash	kpa
Aire sulfatación	Aire que ingresa al comienzo de la caldera	Nm3/h
Temp entrada radiación norte	Temperatura caldera radiación norte	°C
Temp entrada radiación sur	Temperatura caldera radiación sur	°C
Temp entrada convección	Temperatura caldera entrada convección	°C
Temp salida convección	Temperatura caldera salida convección	°C
Presión radiación	Presión en caldera zona radiación	kpa
Presión convección	Presión en cadera zona convección	kpa
Flujo vapor	Flujo de vapor generado por la transferencia de calor	t/h

Tabla 6 variables grupo precipitador electrostático (PP.EE)

Variables	Descripción	Unidad
Temp entrada PPEE 35 y 36	Temperatura entrada al precipitador electroestático 35 y 36	°C
Temp salida PPEE 35 y 36	Temperatura salida al precipitador electroestático 35 y 36	°C
Presión entrada PPEE 35 y 36	Presión entrada al precipitador electroestático 35 y 36	kpa
Presión salida PPEE 35 y 36	Presión salida al precipitador electroestático 35 y 36	kpa

Tabla 7 variables grupo eje-escoria

Variables	Descripción	Unidad
Temperatura eje y escoria	Temperatura del eje y escoria	°C
Ley Cu, Fe, S y Fe3O4 eje	Ley de elementos en eje	%
Ley Cu, Fe, S y Fe3O4 escoria	Ley de elementos en escoria	%

Tabla 8 variables grupo cámara de mezcla

Variable	Descripción	Unidad
Frecuencia VTI 55 y 56	Frecuencia a la cual funcionan los ventiladores 55 y 56	RPM
Presión salida VTI	Presión a la salida de los ventiladores	kpa
Presión entrada CM	Presión en la entrada de cámara de mezcla	kpa
Flujo medido después de VTI	Flujo de gas medido después de los ventiladores	kNm3/h

Se realizó, una inspección visual y estadística descriptiva, de cada variable generando gráficas, que muestren el comportamiento de éstas, considerando, histogramas y graficas de caja.

De la data se observa que existen muchos datos que no contienen operación de la fundición. Para el objetivo de esta investigación, éstos no son representativos, ya que se pretende modelar la temperatura de la primera pantalla cuando, el horno se encuentra en operación, por lo tanto, se eliminan todos estos registros.

Posterior, se deben eliminar y reemplazar, los registros donde existan datos nulos. En la data entregada, éstos datos están representados como “Bad Input” y se debe, a problemas con los TAG y/o por motivos de fuera de servicio debido a, mantención/arreglo. Para evitar, eliminar el registro y disminuir la data éste es reemplazado, utilizando el pronóstico de promedio ponderado, de los últimos diez datos no nulos, anteriores.

$$x_{t+1} = \frac{(x_t \cdot 9 + x_{t-1} \cdot 8 + x_{t-2} \cdot 7 + x_{t-3} \cdot 6 + x_{t-4} \cdot 5 + x_{t-5} \cdot 4 + x_{t-6} \cdot 3 + x_{t-7} \cdot 2 + x_{t-8})}{50} \quad 6.5$$

$$x_{t+1}: \text{dato nulo}, \quad x_{t+i}: dato\ no\ nulo, \quad i: 0, \dots, 8$$

El siguiente paso es examinar los datos atípicos, más conocidos como *outliers*. Los puntos se analizan y estudian, para comprenden su comportamiento. La solución para estos datos es realizar un pronóstico de promedio ponderado igual que, para los datos nulos.

Se observa además, que existen datos que no se representan físicamente, por ejemplos; flujos de oxígeno y alimentación, con valores negativos. Para éstos datos se analizan y estudian cada variable y se toman las siguientes decisiones:

i. Cuando los flujos de alimentación: oxigeno, aire, entre otros tienen, valores negativos, estos se reemplazan, por el valor 0.
ii. El enriquecimiento de oxígeno, cuando el flujo de oxígeno es igual a 0, éste es de 21% (Amec Foster Wheeler, 2018).
iii. Cuando no existe alimentación de un tornillo, la velocidad de éste queda estancada alrededor de 2 RPM. Todos los valores, que sean menor a 10 RPM son reemplazados por 0.

El siguiente paso, para el tratamiento de la data es realizar, correlaciones. Primeramente, entre las variables de cada grupo y posterior con la variable objetivo. Las correlaciones, entre las variables de cada grupo sirven, para encontrar comportamientos idénticos osea, cuando la correlación es muy cercana a 1 y así eliminar variables redundantes. La correlación entre la variable objetivo y las variables de entrada sirve, para comprender las variables que tienen mayor influencia en el objetivo y así también, eliminar las que tienen una correlación muy cercana a 0.

Una vez completada la etapa de limpieza, ésta nueva data es usada en la generación del modelo.

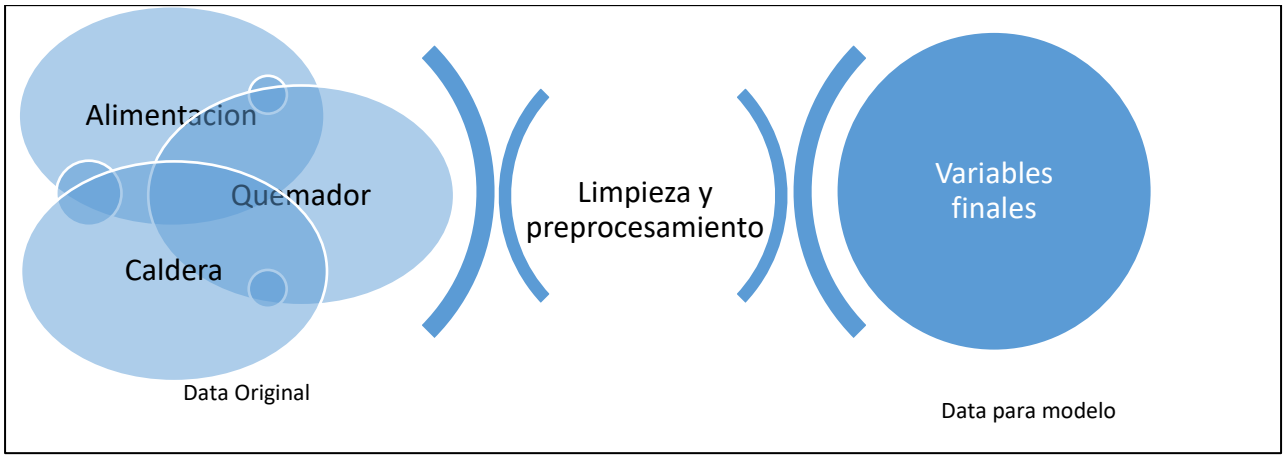

Ilustración 1 diagrama limpieza y preprocesamiento. Fuente: elaboración propia

6.3 Selección datos de entramiento, validación y prueba

La metodología general usa la misma data de entrenamiento para realizar la validación cruzada y búsqueda de hyperparametros, pero debido al tamaño de la data y el tiempo computacional requerido para la validación cruzada, se definió una data, especial para este proceso.

Para seleccionar los datos de entramiento, validación y prueba se importa, la función *train_test_split* de la librería *sklearn*. La separación, se hace en dos pasos:

i. La data, se separa en dos, con una relación de 80/20 y se realiza, de forma aleatoria:

$$X = X1 + X2, \qquad donde\ X1 = 0.8X\ y\ X2 = 0.2X, \qquad X: data\ total$$

ii. La data $X1$ generada, en el paso anterior se divide en dos nuevamente, con una relación de 80/20 y se realiza, de forma aleatoria.

$$X1 = X3 + X4, \qquad donde\ X3 = 0.8X1\ y\ X4 = 0.2X1$$

Finalmente, se tiene a $X3$, $X4$ y $X2$ como datos de: entrenamiento, validación y prueba respectivamente.

Antes de pasar a la siguiente etapa, del entrenamiento de los modelos, se debe normalizar, las variables de entrada, esto es acotarlas en un rango de [0-1]. Ya que teniendo variables con rangos de valores distintos como, es el caso de la leyes y los flujos de gas, los valores altos obtendrían una importancia exponencialmente mayor versus, los valores pequeños afectando, la construcción del modelo. Para realizar la normalización de los datos, se importa la función, *MinMaxScaler* de la librería *sklearn*.

6.4 Entrenamiento, validación cruzada y prueba

Con la data de entrenamiento se procede a entrenar los algoritmos seleccionados, en la sección 4.2. Se importan los algoritmos "*LogisticRegression", "DecisionTreeClassifier", "RandomForestClassifier" y "SVC"* de la librería *sklearn*. Los algoritmos utilizan los hyperparametros que vienen por defecto. Para la evaluación de los modelos, se importan las funciónes "*precision_score", "recall_score", "fbeta_score", "accuracy_score" y "confusión_matrix*".

Con la data de validación, se procede a realizar una validación cruzada y búsqueda de mejores hyperparametros, para la generación del modelo. Mediante la función *GridSearchCV* de la librería

sklearn. Las Tablas 9, 10, 11 y 12 muestran las opciones de hyperparametros buscados, para los algoritmos regresión logística, *SVM*, árbol de decisión y bosques aleatorios, respectivamente.

Tabla 9 hyperparametros regresión logística

Hyperparametro	Descripción	Opciones
Solver	Algoritmo encargado de resolver el problema de clasificación	"*newton-cg, sag, saga, lbfgs*"
C	Parámetro que aplica regularización con el objetivo de reducir el *overfitting*	1, 10, 100, 1000
max_iter	Número máximo de iteraciones para que converja el modelo	50, 100, 150, 200

Tabla 10 hyperparametros máquinas de vectores de soporte

Hyperparametro	Descripción	Opciones
Kernel	Especifica el tipo de *kernel* que utilizara el algoritmo	"*linear, poly, sigmoid*"
C	Parámetro que aplica regularización con el objetivo de reducir el *overfitting*	1, 10, 100, 1000
gamma	Coeficiente del núcleo para los tipos de *kernel*	0,001, 0,0001

Tabla 11 hyperparametros árbol de decisión

Hyperparametro	Descripción	Opciones
Criterion	Medida de selección para la división de los datos de la mejor forma	"gini, entropy"
min_samples_split	cantidad mínima de muestras de un nodo para poder subdividir	2,4,8,16
min_samples_leaf	Cantidad mínima que puede tener una hoja final	1,2,4
max_depth	Profundidad máxima del árbol	2,4,16,256

Tabla 12 hyperparametros bosques aleatorios

Hyperparametro	Descripción	Opciones
n_estimators	Número de árboles que el algoritmo construye	50,100,200,
Criterion	Medida de selección para la división de los datos de la mejor forma	"*gini, entropy*"
min_samples_split	cantidad mínima de muestras de un nodo para poder subdividir	2,4,8,16
min_samples_leaf	Cantidad mínima que puede tener una hoja final	1,2,4
max_depth	Profundidad máxima del árbol	2,4,16,256

La combinación de hyperparametros, que obtenga el mayor *accuracy* se consideran como los mejores hyperparametros para, la generación del modelo predictivo.

Posteriormente, el modelo se vuelve a entrenar, con los datos de entrenamiento y con los mejores hyperparametros encontrados, en la etapa de validación. Finalmente, se evalúa el modelo predictivo generado con los datos de prueba y se obtienen, los resultados.

7. Resultados.

Este capítulo tiene por objetivo presentar, todos los resultados obtenidos en cada etapa de la metodología anteriormente, descrita.

La Figura 15 muestra, el histograma y grafico de caja de las variable: “Temperatura primera pantalla” y “Tasa de fusión total”, antes y después, de la limpieza. Originalmente, la data tiene casi 70.000 datos donde, la alimentación es 0 t/h mostrando, un gráfico de caja estirado parecido, a un rectángulo. Estos datos coinciden, con la temperatura en la primera pantalla bajo los 400°C. Éstos datos no son representativos debido a que, no existe operación en esos momentos por lo que son, eliminados de la data. En el anexo 1, se muestra la representación para el resto de las variables.

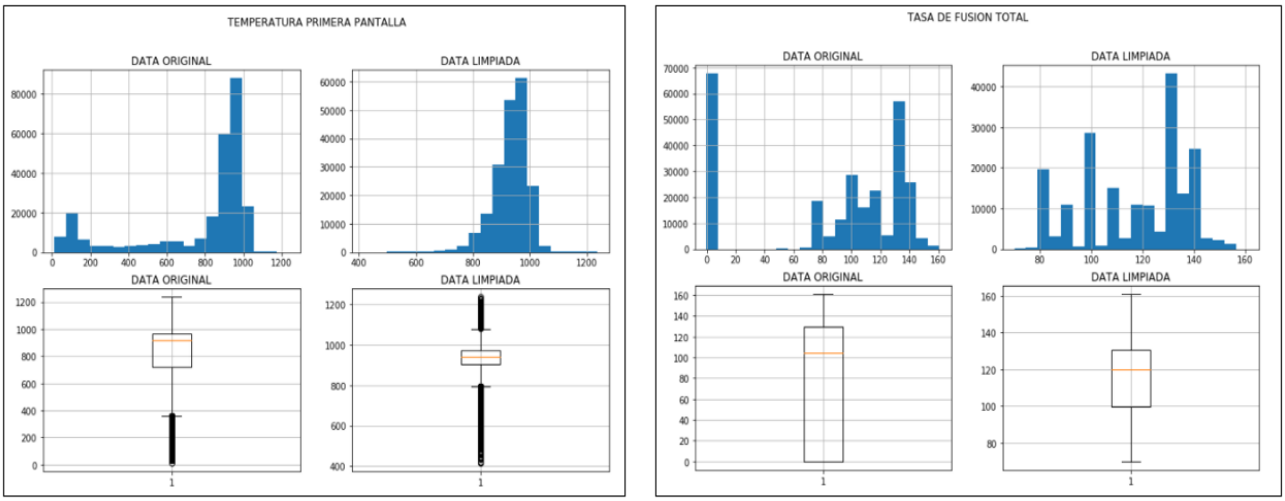

Figura 15 histograma y grafico de caja, temperatura primera pantalla y tasa de fusión antes y después de limpieza. Fuente: elaboración propia

La Figura 16, muestra la matriz de correlación del grupo quemador. Se observa que existe una alta correlación entre las variables: “oxigeno quemador”, “aire quemador” y “aire distribución”. Esto debido a la naturaleza del proceso, estos tres flujos regulan el oxígeno necesario, para la oxidación del mineral. Además, se observa una correlación negativa entre las variables: “aire quemador” y “aire distribución” versus el “enriquecimiento”. Esto debido, a que el aire solo tiene un 21% de

oxígeno, y el enriquecimiento, es el % de oxígeno total en el flujo de gas, al aumentar el aire, el enriquecimiento disminuirá.

La variable “flujo de petróleo”, tiene una alta correlación con el “aire de atomización”, dado que el flujo de petróleo es inyectado a través de la misma lanza (tubo). La variable “ángulo de distribución”, tiene una correlación negativa, con la “velocidad aire proceso”. Esto debido a que el inyector, por donde ingresa el aire de proceso al quemador, aumenta y disminuye su área. Al aumentar el área, el ángulo de distribución aumenta y la velocidad disminuye debido a que la presión que genera el aire, es menor. En los anexos 2 se muestran las matrices del grupo alimentación y caldera.

Figura 16 matriz de correlación entre las variables de alimentación. Fuente: elaboración propia

La correlación, entre la variable objetivo y las variables de entrada, se muestra en la Tabla 13. Se observa que las mayores correlaciones corresponden: al “aire de distribución”, “oxigeno quemador” y “tasa de fusión total”. Es interesante, notar que las variables: “aire de sulfatación”, “% arsénico” y “% antimonio” tienen, una correlación mayor por sobre variables que a priori, se considerarían más importantes como es el “coeficiente” y el “% de azufre”. Las variables: “enriquecimiento”, “Line 1 *bin weight”* y “Line 2 *bin weight”,* tienen una correlación muy cercana a 0 implicando, que no influyen en la variable objetivo (temperatura de primera pantalla).

Tabla 13 Correlación con la variable objetivo

Variables	Valor correlación
Aire distribución	0.350049
Oxígeno quemador	0.339619
Tasa de fusión total	0.317186
Line 2 speed	0.305032
Line 2 flow	0.299557
Line 1 flow	0.285051
Aire de sulfatación	0.283860
Line 1 speed	0.278662
Aire quemador	0.253848
Arsénico concentrado	0.193524
Antimonio concentrado	0.184999
Presión Interna Horno Flash	0.140215
Coeficiente	0.098000
Velocidad aire proceso	0.084580
Flujo petróleo (Lanza central)	0.069457
Aire atomización lanza	0.053370
Fierro concentrado	0.040486
Sílice concentrado	0.029230
Azufre concentrado	0.019060
Enriquecimiento	0.002159
Line 1 bin weight	-0.008835
Line 2 bin weight	-0.011676
Ángulo de distribución	-0.024644
Zing concentrado	-0.026683
Cobre concentrado	-0.125535

Finalmente, se eliminaron todas las variables del grupo PP.EE y cámara de mezcla además, algunas variables del grupo caldera tales como: “temp radiación sur”, “temp salida convección”, “presión radiación” y “presión convección”. Esto según lo conversado y analizado con personal de operaciones ya que, son variables posteriores a la temperatura de la primera pantalla, y por lo tanto son consecuencia de ésta.

Se eliminaron, las variables del grupo eje-escoria, ya que, representan una salida del horno al igual que los gases por lo que, su información para el control de la temperatura en la primera pantalla no tiene relevancia en el proceso.

Se eliminaron los datos, donde no existe alimentación, se suavizaron los registros nulos y *outliers*. De acuerdo con las matrices de correlación se eliminaron las variables: “enriquecimiento”, “Line 1 *bin weight”* y “Line 2 *bin weight”,* debido, a la baja correlación, en consideración a la temperatura de la primera pantalla.

Después de realizar, el análisis de limpieza y preprocesamiento de los datos, las variables finales, para la generación del modelo de predicción son las siguientes (Tabla 14):

Tabla 14 variables finales para generación de modelo

Variables	Unidad
Tasa de fusión total	kton
Ley Cu, Fe, S, As, Zn, Sb	%
Tasa de fusión line 1 y 2	Kton
Velocidad de alimentación line 1 y 2	RPM
Oxigeno quemador	Nm3/h
Aire quemador	Nm3/h
Aire distribución	Nm3/h
Coeficiente	Nm3/ton
Velocidad aire proceso	m/s
Angulo de distribución	°
Flujo de petróleo	Kg/h
Aire atomización lanza	Nm3/h
Presión interna horno flash	Kpa
Aire sulfatación	Nm3/h
Temperatura entrada termocupla radiación norte	°C

Finalmente, se comenzó con una data de 55 variables y 263.581 registro, luego de la limpieza y pretratamiento se tiene una data de 22 variables y 195.379 registros. La cual, se compone de 96.007 de la clase 0, 58.874 de la clase 1 y 40.498 de la clase 2.

La evaluación de los modelos predictivos entrenados, utilizando los datos de entrenamiento, como se muestran en las Tablas 15 y 16.

Tabla 15 evaluación entrenamiento

Algoritmo	***Precision***	***Recall***	***Acurracy***	***F***
Regresión logística	[0.63 – 0.44 – 0.52]	[0.85 – 0.29 – 0.36]	0.58	[0.72 – 0.35 – 0.43]
Árbol de decisión	[1 – 1 – 1]	[1 – 1 – 1]	1.0	[1 – 1 – 1]
Bosques aleatorios	[1 – 1 – 1]	[1 – 1 – 1]	1.0	[1 – 1 – 1]
SVM	[0.78 – 0.63 – 0.65]	[0.88 – 0.53 – 0.59]	0.72	[0.83 – 0.58 – 0.62]

Tabla 16 matriz de confusión entrenamiento

Regresión logística				**Árbol de decisión**				**Bosques aleatorios**				***SVM***		
52050	6425	2845		61320	0	0		61320	0	0		54141	5394	1785
21217	10896	5675		0	37788	0		0	37788	0		11237	20042	6509
9004	7543	9387		0	0	25934		0	0	25934		4190	6414	15330

Los algoritmos de árbol de decisión y bosques aleatorios tienen, los mejores resultados con un *accuracy* del 100% para las tres clases. El algoritmo *SVM* tiene un *accuracy* del 72% y la regresión logística de 58% siendo, el modelo con peor resultado. La clase 0, es la que obtiene los mejores resultados en los cuatro modelo, seguida de la clase 2 y de la clase 1, que obtiene los peores resultados, con un 63% de *precision* en el mejor de los casos.

La matriz de confusión de la Tabla 16 muestra cómo se distribuyen los datos predichos versus los reales. Se confirma el 100% de *precision* de los algoritmos: árbol de decisión y bosques aleatorios. También, se observa la dispersión de datos de los algoritmos *SVM* y regresión logística ambos, con importante números de datos falsos positivos para la clase 0.

Los resultados de la búsqueda de hyperparametros realizando, validación cruzada se muestran en las Tablas 17, 18, 19 y 20.

Tabla 17 hyperparametros óptimos regresión logística

C	***max_iter***	***solver***	***accuracy***
100	50	sag	0.56

Tabla 18 hyperparametros óptimos árbol de decisión

criterion	***max_depth***	***min_samples_leaf***	***min_samples_split***	***accuracy***
entropy	256	1	2	0.88

Tabla 19 hyperparametros óptimos bosques aleatorios

criterion	***max_depth***	***min_samples_leaf***	***min_samples_split***	***n_estimators***	***accuracy***
entropy	256	1	2	100	0.92

Tabla 20 hyperparametros óptimos SVM

C	***gamma***	***kernel***	***accuracy***
10	0.001	linear	0.56

La Tabla 17 muestra, los hyperparametros óptimos, para el modelo de regresión logística. Utilizando: *C* de 100; 50 iteraciones y función *sag* para el problema de optimización, se obtiene una *accuracy* del 56% variando, en 2% con respecto al entrenamiento.

La Tabla 18 muestra, los hyperparametros óptimos, para el modelo árbol de decisión. Utilizando: el criterio *entropy*; profundidad del árbol de 256; 1 dato mínimo para generar una hoja y 2 datos mínimo para generar un nodo, se obtiene un *accuracy* del 88%.

La Tabla 19 muestra, los hyperparametros óptimos para el modelo de bosques aleatorios. Utilizando: el criterio *entropy*; cantidad de 100 árboles; profundidad del árbol de 256; 1 dato mínimo para generar una hoja y 2 datos mínimo para generar un nodo, se obtiene una *accuracy* del 92%, siendo el más alto en esta etapa.

La Tabla 20 muestra los hyperparametros para el modelo *SVM*. Utilizando: *C* de 10, *gamma* de 0.001 y *kernel linear*, se obtiene una *accuracy* del 56% bajando, considerablemente su evaluación en relación con el entrenamiento. Obteniendo resultados similares a la regresión logística.

Con los hyperparametros óptimos, se entrenan los algoritmos de *machine learning* y se evalúan los modelos predictivos generados con los datos de testeo. Las Tablas 21 y 22 muestran los resultados.

Tabla 21 evaluación testeo

Algoritmo	***Precision***	***Recall***	***Acurracy***	***F***
Regresión logística	[0.78 – 0.35 – 0.37]	[0.46 – 0.31 – 0.78]	0.48	[0.58 – 0.33 – 0.50]
Árbol de decisión	[0.64 – 0.43 – 0.42]	[0.65 – 0.49 – 0.32]	0.54	[0.65 – 0.46 – 0.37]
Bosques aleatorios	[0.85 – 0.75 – 0.85]	[0.94 – 0.71 – 0.72]	0.82	[0.89 – 0.73 – 0.78]
SVM	[0.78 – 0.36 – 0.40]	[0.43 – 0.43 – 0.68]	0.49	[0.56 – 0.40 – 0.50]

Tabla 22 matriz de confusión testeo

Regresión logística				**Árbol de decisión**				**Bosques aleatorios**				***SVM***		
8998	5290	5061		12767	4702	1880		18103	1080	166		8346	6926	4077
1988	3624	5917		4242	5598	1689		2469	8191	869		1843	5000	4686
503	1236	6459		2915	2650	1633		652	1653	5893		530	2013	5655

El algoritmo de bosques aleatorios obtiene los mejores resultados con un *acurracy* de 82%. Lo sigue el algoritmo árbol de decisión, *SVM* y regresión logística.

La clase 0 tiene, la mejor predicción en todos los modelos según el criterio *F* seguido, de la clase 2 y la clase 1, excepto para el modelo árbol de decisión donde, que obtienen mejores resultados para la clase 1 en relación con la clase 2.

La matriz de confusión, de la Tabla 22 muestra que existen, muchos falsos positivos para la clase 2, en los algoritmos regresión logística y SVM y a la vez escazas predicciones para la clase 0.

La Figura 17 muestra, la importancia de las variables para el modelo predictivo generado, de bosques aleatorios utilizando, los mejores hyperparametros. Las variables de mayor importancia son: “% Zn en concentrado”, “aire de sulfatación”, “%As en concentrado” y “ángulo de distribución”. Las variables de menor importancia para el modelo son: “presión interna HF”, “aire de atomización” y “velocidad aire proceso”.

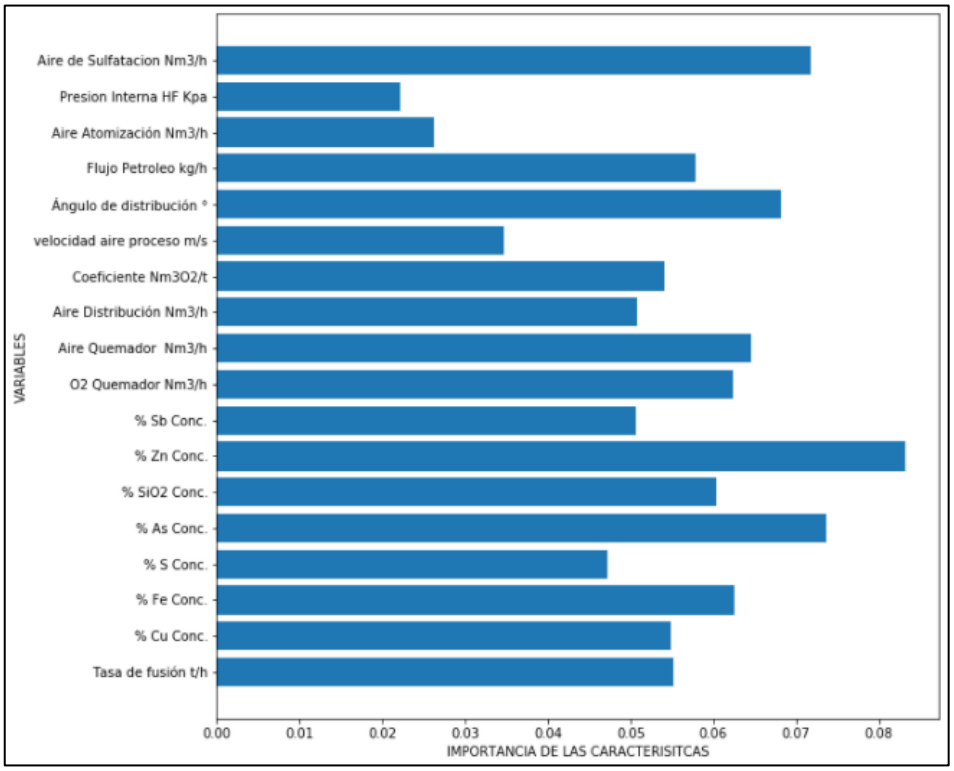

Figura 17 importancia de variables en modelo bosques aleatorios

8. Conclusiones y Recomendaciones

8.1 Conclusiones.

Del análisis de correlación, las variables que representan: flujos de aire, oxígeno y alimentación, son las que presentan mayor correlación con la variable objetivo (temperatura primera pantalla). Estas son las mismas variables que se modifican según, el control estándar del proceso indicado en el sección 1.1.

El estudio muestra, una variable más interés que es el “aire de sulfatación” ya que, su correlación es superior al “aire del quemador” y actualmente no es considerado en el control de proceso.

Las variables: “% de arsénico” y “% antimonio” son las que mayor correlación tienen con la temperatura de la primera pantalla, dentro de la leyes de elementos, incluso más que el “% azufre”. Por lo que se debe realizar un control preventivo, de la ley de estos elementos en la mezcla de concentrado-calcina.

De los algoritmos de *machine learning*, el algoritmo de bosques aleatorios, es el que obtiene los mejores resultados con un *accuracy* de 82%. El algoritmo de regresión logística, es el que peores resultados obteniendo en todas las etapas del estudio, tanto para el entrenamiento, validación y testeo, con un promedio de *accuracy* de 54%.

El algoritmo bosques aleatorios, al obtener un *acurracy* del 100% en el entrenamiento, el modelo se encuentra sobreajustado pero, al realizar la validación cruzada y evaluar con datos de prueba, el modelo con los mejores hyperparametros, se obtienen altas precisiones, eliminando la condición de sobre ajustado.

Las evaluaciones de los modelos, muestran que la clase 0, es la que mejor resultados obtiene seguido, de la clase 2 y de la clase 1. A pesar del desbalance, de la cantidad de datos por clase, los algoritmos predicen, de mejor manera para la clase 2 que, para la clase 1.

El algoritmo árbol de decisión obtiene una *accuracy* de 100% para la data de entrenamiento pero, en la etapa de validación y testeo muestra un *accuracy* de 88 y 54% respectivamente. Esto indica que el modelo esta sobreajustado con los datos de entrenamiento provocando, una mala predicción al evaluar, con los datos de prueba bajando, a la mitad el accuracy. El algoritmo bosques aleatorios evita, el sobreajuste mediante la generación, de árboles aleatorios.

El algoritmo *SVM* tiene un *accuracy* de 72% para el entrenamiento, pero los resultados, de la validación cruzada y prueba obtiene valores similares al algoritmo de regresión logística. El algoritmo *SVM* modela bien, los datos de entrenamiento pero, no predice correctamente nuevos datos por lo que, no se considera como un buen predictor para el presente estudio.

Finalmente, el algoritmo bosques aleatorios caracteriza y modela la temperatura de la primera pantalla con una *accuracy* del 82% por lo tanto, la hipótesis planteada en este estudio se valida y confirma. Esta herramienta permite, predecir con un amplio grado de certeza la temperatura de la primera pantalla.

Al analizar las variables, según el modelo de árbol de decisión generado con mayor *accuracy*, las principales variables de importancia son: “% Zn en el concentrado” y “aire de sulfatación”. Las de menor importancia: “presión interna del horno” y “aire de atomización”. Esto refleja que no es suficiente con la Tabla 13 de correlación para, obtener la importancia de las variables sobre el objetivo. Además, el modelo confirma la importancia, del “aire de sulfatación” y “% de arsénico”.

Por lo tanto, cuando el horno se encuentra en operación y se requiere realizar cambios en alguna de las variables pertenecientes al modelo se debe, procesar la información primero en el modelo predictivo generado de *machine learning*. Lo anterior, para predecir si estas nuevas condiciones permitirán operar con temperatura dentro, del rango de control.

Este modelo predictivo estando en operación, permitiría reducir los costos por mantenciones de emergencia, por lo consiguiente, disminuir el tiempo de no operación del horno, aumentando su producción promedio. Permitiría mantener una tasa de fusión mayormente constante, disminuyendo las desviaciones.

8.2 Recomendaciones.

Para el proceso de obtención de la data existen, otras variables operacionales y geometalurgicas que no están incluidas en el presente estudio pero, que son relevantes para el proceso, por ejemplo: “% de carga fría”, “% de polvos recirculados”, “razón entre calcina y concentrado”, entre otros, y que serían fundamentales, para el control estándar de operaciones. Se debe gestionar, un sistema de medición para estas variables y así incluirlas, en un posible modelo predictivo.

La eliminación y reemplazo de los datos nulos y *outliers*, se puede lograr mediante un algoritmo de predicción. Esto permitiría una mayor precisión para el modelo, en relación con la herramienta aplicada en este estudio (predicción con promedio ponderado)

El número de datos para cada clase se debe balancear, osino el modelo obtendrá mejores resultados para el que tenga más información. En este estudio, esto se corrige utilizando el hyperparametro "*class_weigth*", el cual internamente realiza un balance de cada clase para la construcción del modelo. Otra solución, es un método de generación de datos mediante, algoritmos que generan data a partir de otra fuente de datos.

Si las condiciones operacionales, cambian o si se genera un cambio en el proceso como: modificaciones en equipos, cambios en los circuitos de alimentación, cambios en los sistema de medición, entre otros, es conveniente entrenar el modelo con datos y condiciones actuales, en caso contrario el modelo realizará predicciones erróneamente. Dado que el modelo se entrenó con ciertas condiciones de borde, por lo que el modelo ya no responde a los requerimientos actuales.

Validada la herramienta de *machine learning*, se recomienda extrapolar este estudio a otros objetivos de interés, para la fundición, como, por ejemplo: "ley de Cu en el eje", "% de Fe_2O_3 en la escoria", entre otros. Estas variables son muy relevantes para la industria, si se pudiera predecirlas apropiadamente, constituiría un apoyo importante para el negocio.

Además, se recomienda modelar la variable objetivo (temperatura de primera pantalla) con otros algoritmos de *machine learning*, como: Redes neuronales, perceptrón multicapa entre, otros. Esto con el fin de comparar los modelos y obtener, el mejor que logre predecir la temperatura de la primera pantalla.

9. Anexos.

9.1 Anexo 1: Histogramas por grupos, antes y después de limpieza

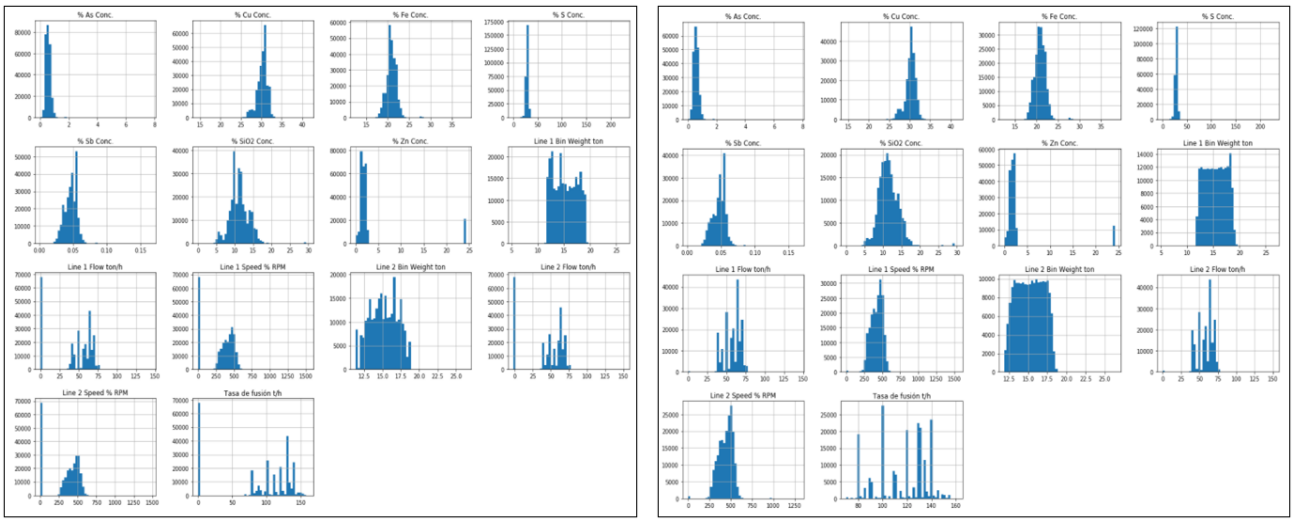

Figura 18 histograma grupo alimentación, antes y después limpieza

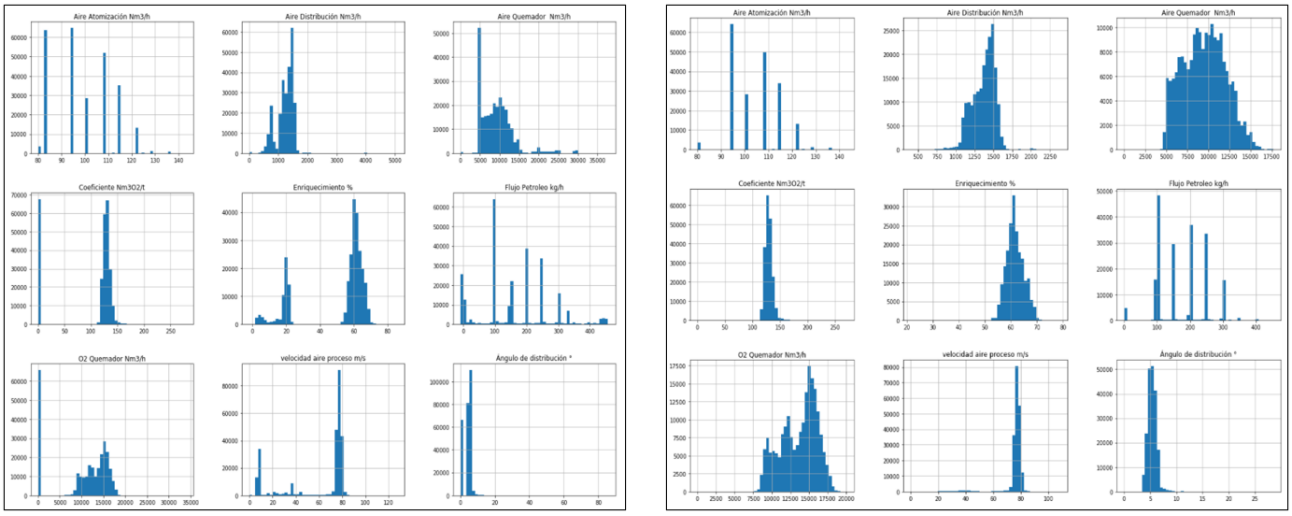

Figura 19 histograma grupo quemador, antes y después limpieza

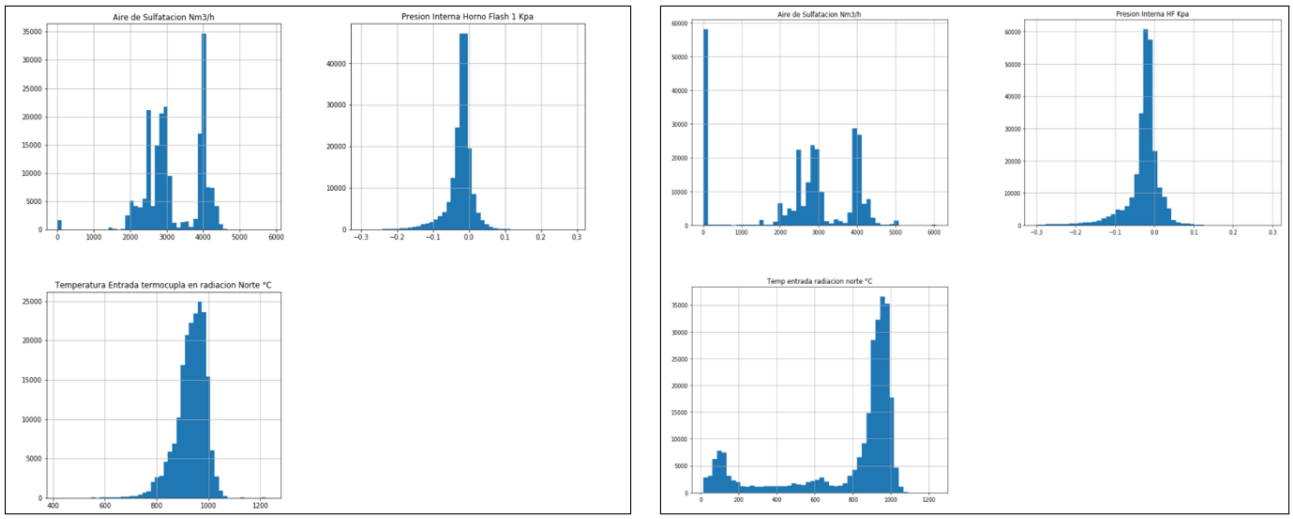

Figura 20 histograma grupo caldera, antes y después limpieza

9.2 Anexo 2: Matriz de correlación por grupos.

Figura 21 matriz correlación grupo alimentación

Figura 22 matriz correlación grupo caldera

10. Bibliografía

Amec Foster Wheeler. (2018). *Ingeniería de detalles potenciamiento Horno Flash fundición Chuquicamata.* Codelco División Chuquicamata Gerencia de Proyectos.

Avalos, S., Kracht, W., & Ortiz, J. (2020). Machine Learning and Deep Learning Methods in Mining Operations: a Data-Driven SAG Mill energy Consumption Prediction Application. *Mining, Metallurgy & Exploration*.

Bae, J., Li, Y., Stahl, N., & Mathiason, G. (2020). Using Machine Learning for Robust Target Prediction in a Basuc Oxygen Furnace System. *METALLURGICAL AND MATERIALS TRANSACTIONS*.

Bascur, O., & Soudek, A. (2019). Grinding and Flotation Optimization Using Operational Intelligence. *Mining, Metallurgy & Exploration*.

Bezuidenhout, J., Yang, Y., & Ekteen, J. (2008). Computational fluid dynamic modelling of a waste-heat boiler associated with flash smelting of base metal sulphides. *JOURNAL OF THE SOUTH AFRICAN INSTITUTE OF MINING AND METALLURGY, 108*, 179-188.

Brook Hunt & Associates Ltd. (2007). *Brook Hunt Smelter.*

Camacho, E. &. (2007). *Model predictive control (2nd ed., Vol. 1).* Springer.

Carlsson, L., Samuelsson, P., & Jonsson, P. (2019). Predicting the Electrical Energy Consumption of Electric Arc Furnaces Using Statistical Modeling. *Metals*.

Carlsson, L., Samuelsson, P., & Par, J. (2019). Using Statistical Modeling to Predict the Electrical Energy Consumption of an Electric Arc Furnace Producing Stainless Steel. *Metals*, 53-60.

COCHILCO. (2021). *Anuario de estadísticas del cobre y otros minerales 2001 - 2020.* Santiago.

Congreso nacional de Chile. (2013). *D.S 28. ESTABLECE NORMA DE EMISIÓN PARA FUNDICIONES DE COBRE Y FUENTES.* Valparaiso: Biblioteca del Congreso Nacional de Chile.

De la Fuente Fernandez, S. (2015). *Regresión Logística.* Universidad Autónoma de Madrid.

Dhanuskodi, R., Kaliappan, R., Suresh, S., Anantharaman, N., Arunagiri, A., & Krishnaiah, J. (2015). Artificial Neural Networks model for predicting wall temperature of supercritical boilers. *APPLIED THERMAL ENGINEERING*, 749-753.

Fernández, S. d. (s.f.). *Regresión Logística.* Facultad Ciencias Económicas y Empresariale.

Gaskell, D. R. (2003). *Introuction to the Thermodynamics of materials (Fourth edition).* Taylor & Francis.

Géron, A. (2017). *Hands-On Machine Learning with Sckit-Learn & TensorFlow.* O'Reilly Media, Inc.

Gonzalez, L. (28 de Junio de 2019). Obtenido de https://aprendeia.com/regresion-logistica-multiple-machine-learning-teoria/

Gonzalez, L. (16 de Agosto de 2019). Obtenido de https://aprendeia.com/maquinas-vectores-de-soporte-clasificacion-teoria/

Heaton, J. (2015). *Artificial Intelligence for Humans Volume 3: Deep Learning and Neural Networks.*

Heras, J. M. (29 de Septiembre de 2020). *www.iartificial.net*. Obtenido de https://www.iartificial.net/clasificacion-o-regresion/

Khoshhal, A., Rahimi, M., Ghahramani, A., & Alsairafi, A. (2011). Computational fluid dynamics modeling of high temperature air combustion. *KOREAN JOURNAL OF CHEMICAL ENGINEERING*, 1181-1187.

Li, Z.-m., Gui, W.-h., & Zhu, J.-y. (2019). Foult detenction in flotation processes based on deep learning and support vector machine. *Journal of Central South University*, 2504-2515.

Lishchuk, V., Lund, C., & Yousef, G. (2019). Evaluation and comparison of different machine-learning methods to. *Minerals Engineering*, 156-165.

Matich, D. J. (2001). *Redes Neuronales: Conceptos Básicos y Aplicaciones.* Rosario: Universidad Tecnológica Nacional.

Max, S., Muzaffer, A., Thomas, B., & Dirk, A. (2021). Review on model predictive control: an engineering perspective. *The International Journal of Advanced Manufacturing Technology*(117), 1327–1349.

Mingwei, H., Qiyang, T., Ruth, K., Sen, W., Xue, L., Tianqi, W., . . . Ming-Xing, Z. (2021). Prediction of Mechanical Properties of Wrought Aluminium alloys Using Feature Engineering Assited Machine LEarning Aproach. *METALLURGICAL AND MATERIALS TRANSACTIONS*.

Mitchell, T. (1997). *Machine Learning.* McGraw-Hill.

Montes de Oca, L., Dominguez, F., Días, Y., López, Y., & Tápanez, Á. (2017). SIMULACIÓN DE UNA CALDERA DE RECUPERACIÓN DECALOR UTILIZANDO EL SOFTWARE HYSYS. *Tecnología Quimica, 28*(1).

Moon-Jo, K., Jong, P. Y., Ji-Ba-Reum, Y., Seung-Jun, C., & DongEung, K. (2020). Prediction of the Temperature of Liquid Aluminum and the Dissolved Hydrogen Content in Liquid Aluminum with a Machine Learning Approach. *Metals*.

Peñalba, J. (2004). *Modelado y Simulación de una Caldera Convencional.* Universidad Rovira I Virgil.

Phull, J., Egas, J., Barui, S., Mukherjee, S., & Chattopadhyay, K. (2019). An Application of Decision Tree-Based Twin Support Vector MAchines to Classify Dephosphorization in BOF Steelmaking. *Metals*.

Riveros, G. (2009). *Curso M1 51 A Pirometalurgia.* Universidad de Chile, Departamento ingenieria de minas.

Sánchez, A., Fernández, F., Valero, C., Muñoz, M., Rodríguez, A. F., López, M., & Espejo, I. (2009). *Estadística Descriptiva y Probabilidad: (Teoría y problemas)*. Recuperado el 28 de 11 de 2021, de https://libros.metabiblioteca.org/handle/001/140

Schaaf, M., Gómez, Z., & Cipriano, A. (2010). Real-time hybrid predictive modeling of the Teniente Converter. *Journal of Process Control*, 3-17.

SERNAGEOMIN. (2021). *Anuario de la minería de Chile 2020.* Santiago.

Slingh, K., Vakkantham, P., Nistala, S. H., & Runkana, V. (2020). Multi-objetive Optimization of Integrated Iron Ore Sintering Process Using MAchine Learning and Evolutionary Algorithms. *Metallurgy, Materials Engineering*.

Tavoosi, J., & Mohammadzadeh, A. (2021). A New Recurrent Radial Basis Function Network-based Model Predictive Control for a Power Plant Boiler Temperature Control. *INTERNATIONAL JOURNAL OF ENGINEERING*, 667-675.

van Duijvenbode, J., Buxton, M., & Soleymani, M. (2020). Performance Improvements during Mineral. *Minerals, 10*(366).

Vázquez, A., Galindo, I., Mani, G., & Rossano, M. (2010). Simulación CFD, una alternativa para el análisis de emisiones. *VIII Congreso Internacional sobre Innovación y Desarrollo Tecnológico*.

Vojslav, K. (2005). *Support Vector Machines - An Introduction.* Auckland: The University of Auckland, School of Engineerng.

Wang, B. (2018). Dynamic model of bottom blown oxygen copper. *Int. J. Modelling, Identification and Control, 30*(2).

Wood Mackenzie . (2010). *Wood Mackenzie Smelter Chuquicamata.*

Yang, Y. (1996). Computer simulation of gas flow and heat transfer in waste-heat boilers of the outokumpu copper flash smelting process. *ACTA POLYTECHNICA SCANDINAVICA-CHEMICAL TECHNOLOGY SERIES*(242), 1-135.

Zhongsheng Hou, H. G. (Mayo de 2017). Data-Driven Control and Learning Systems. *IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, 64*(5), 4070-4075.