UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

OPTIMIZACIÓN DEL PROCESO DE PRODUCCIÓN DEL NITRATO DE POTASIO

MEMORIA PARA OPTAR AL TÍTULO DE

INGENIERO CIVIL INDUSTRIAL

GABRIEL BERNABÉ PALOMINOS MIRANDA

PROFESOR GUÍA:

PATRICIO CONCA KEHL

MIEMBROS DE LA COMISIÓN:

RODOLFO URRUTIA URIBE

GERARDO DIAZ RÓDENAS

SANTIAGO DE CHILE

2015

RESUMEN DE LA MEMORIA PARA OPTAR AL TÍTULO DE: Ingeniero Civil Industrial
POR: Gabriel Bernabé Palominos Miranda
FECHA: 9/12/2014
PROFESOR GUÍA: Patricio Conca Kehl

OTIMIZACIÓN DEL PROCESO DE PRODUCCIÓN DEL NITRATO DE POTASIO

El presente trabajo de memoria trata acerca del mejoramiento de los estándares de pureza de material elaborados en una planta de producción de nitrato de potasio (KNO_3) de una empresa particular del rubro de la minería no metálica.

El nitrato de potasio es un compuesto químico usado en distintas industrias, como por ejemplo, farmacéutica y de explosivos. En particular, el producto desarrollado por la empresa corresponde a la industria agrícola, donde es usado como fertilizante. Según información de la Comisión Chilena del Cobre, en los últimos 10 años, ha disminuido un 45,8% su producción nacional de nitrato llegando a 759,4 toneladas al año el 2013, mientras que su precio ha ido al alza llegando a US$ 800 por tonelada el 2013, es decir, un aumento del 179,6% .

Se abre una gran oportunidad de negocios, dado el alza en los precios, que esta memoria pretende aprovechar, sin embargo, no toda la producción cumple con los estándares de pureza que se requieren para su comercialización. Es por ello que este trabajo se enfoca en la disminución del porcentaje de producción defectuosa del nitrato de potasio.

El proceso productivo para este compuesto consta de 4 plantas de procesamiento. En cada planta están involucradas variables químicas y de control. Las plantas están conectadas en serie, donde el flujo de material pasa en cada una y luego de un tiempo determinado pasa a la siguiente. En total se demora 36 horas generando dos flujos de producto final y flujos de descarte o deshecho.

Para realizar el trabajo, se usa el campo de la minería de datos, el cual, mediante la metodología *Knowledge Discovery in Databases* para elaborar modelos explicativos y predictivos, y así descubrir relaciones entre las variables que interactúan en la producción. A la vez, se utiliza dos tipos de modelos, de clasificación y de regresión, de manera de comparar y decidir qué forma de predicción es más conveniente, estimar la categoría de pureza o predecir el nivel de contaminación.

Se generaron 4 ecuaciones para describir los contaminantes $KClO_4$ y $NaCl$ en los dos flujos de salida que tiene el proceso. Se implementaron 6 algoritmos de data mining, 3 de clasificación y 3 de regresión, de los cuales Support Vector Machine y Random Forest Regression se ajustaron mejor. Los modelos de regresiones dieron mejores resultados que los de clasificación dado que los primeros, a pesar de subestimar la producción de material impuro, presentan mayor precisión a la hora de predecir la categoría del producto.

El modelamiento permite disminuir la impureza histórica de cada contaminante, en cada corriente, desde un 11% a un 7% promedio. Además, se generan rangos donde las variables pueden variar un 4,5% sin que la impureza del sistema aumente. Finalmente, a nivel económico, los modelos permiten generar 6.833 toneladas anuales extras, equivalentes a unos 4,5 millones de dólares al año.

A mis queridos abuelos, Rolando y Uberlinda, hijos del salitre y del Norte. Que Dios los tenga en Su Santo Reino.

AGRADECIMIENTOS

Ya se va completando una etapa y los caminos que se inician son para algunos lógicos, para otros inimaginables, para mí, un poco de ambos. Pero independiente de donde deparen mis pasos ahora, me gustaría agradecer a cada persona que estuvo presente en este tiempo, vinculado tanto directa como indirectamente a mi carrera.

Quiero agradecer al primero de todos, a Dios, quien es el Compañero y Amigo de los caminos que he recorrido y quien tiene una Misericordia y Amor mucho más grande que el que uno se pueda imaginar. Es el tesoro de mi alma.

Infaltable el agradecer a mi familia, a cada uno, quienes me han visto crecer y me han formado: A mi papá, de quien admiro la justica de sus acciones, a mi mamá, de quien admiro su dulzura de corazón, a mi hermano Jorge, de quien admiro su sensibilidad para entender al otro y a mi hermano Gerardo, de quien admiro su espíritu curioso y determinado.

Dentro de la universidad un grupo muy importante fue la pastoral de Ingeniería, ya que sin ellos, la universidad habría sido un vacío del conocimiento, y son ellos los que le dieron parte del sentido que le faltaba. Agradezco a Patricia, a Francisca, a Mar, a Elizabeth, a Paula, a Esteban, a Gabriel Q.E.P.D., a Mauricio, a Andrés, a Claudio, a Miguel, a Fermín y a Mariano.

Agradezco a los amigos que fui haciendo en los distintos ramos y semestres. Ellos fueron parte de las alegrías y penas compartidas en tantos momentos de estudio y de ocio, además de aprender de cada uno y reconocer el gran valor humano que tienen. El amigo que lea este agradecimiento sabrá mi gratitud hacia él o ella.

También quiero agradecer a tres mis amigos muy especiales. A Patricia, con quien hemos compartido alegres y tristes momentos y nunca perdemos oportunidad para reír. Agradezco su sinceridad y amistad. También agradecer a Viviana con las gratas conversaciones que hemos tenido en los ratos que nos encontramos. Su lado humano y cristiano ha sido un ejemplo el cual agradezco de todo corazón. Y agradecer también a Cristián, quien me ha enseñado la importancia de la gratuidad en la vida y de quien agradezco su fraternidad, sencillez y humildad. A ellos gracias por su apoyo.

Finalmente, agradezco a todos mis familiares y antepasados, porque no vine solo al mundo, pues ellos han formado el espacio donde mis abuelos vivieron, mis padres viven y yo junto a mis hermanos seguimos viviendo, pues, al fin de cuentas, uno ha crecido y avanzado sobre lo que los ancestros han construido.

TABLA DE CONTENIDO

INDICE DE TABLAS

INDICE DE GRÁFICOS

1 INTRODUCCIÓN

1.1 Consideraciones Preliminares

Este trabajo de memoria se desarrolla en el contexto de la minería no metálica enfocado en el proceso productivo del nitrato de potasio, de una empresa particular del norte de Chile. Se trabajará los datos de una empresa en particular, pero por motivos de confidencialidad, se omitirá su nombre y se mencionará como *La Empresa*.

La industria minera se puede dividir en 2 categorías: minería metálica, la cual trabaja con productos metálicos tales como el cobre, el oro, la plata, entre otros, y minería no metálica, la cual trabaja con minerales como el litio, el potasio o los nitratos. También, en ambas categorías, los procesos se clasifican en 2 macro etapas:

1. Minera Extractiva: conocida comúnmente como proceso minero, el cual consiste en extraer los compuestos químicos del mineral desde los yacimientos naturales hasta dejarlos convertidos en insumos trabajables, ya sean material granulado, polvo fino, pulpa de concentración.
2. Proceso Metalúrgico: es la etapa de procesamiento del material, que consiste en la combinación de las materias primas obtenidas en Minería Extractiva para la obtención de compuestos más elaborados o de mejor niveles de concentración. Esta macro etapa está más presente en la minería no metálica.

En Chile, la explotación de los yacimientos mineros de salitre recopila principalmente el mineral comúnmente conocido como "caliche", que es una mezcla de nitrato sódico y potásico, además de otros subproductos, como yodo o litio. Chile posee enormes reservas de nitratos, que según Ericksen (1983), se estiman en 250 millones de toneladas de salitre. Los campos de nitratos, corresponden a un complejo de sales entre los cuales se encuentran sulfatos, cloruros, cloratos, nitratos, carbonatos de sodio, potasio, magnesio, yodatos, percloratos, entre otros. El uso de estos compuesto son demandados por una amplia diversidad de industrias, desde farmacéutica a de energía solar, pero en esta memora se enfocará en la industria agrícola al cual pertenece el compuesto KNO_3, nitrato de potasio, producido por *la Empresa*.

En la actualidad, la industria alimentaria se enfrenta a un aumento de su demanda dado factores tanto poblacionales, el crecimiento demográfico de distintos países, como climáticos, por ejemplo el calentamiento global que dificulta el desarrollo de la agricultura de en diversas zonas geográficas del planeta. Ante este escenario los fertilizantes son una potencial herramienta para mejorar la producción agrícola, ayudando a incrementar la productividad de los cultivos en cuanto a cantidad de frutas o verduras y mejorando las condiciones del suelo para plantaciones en sectores que a priori no son recomendables.

El precio del nitrato de potasio ha experimentado un alza del 179,6% en los últimos 10 años llegando a US$ 800 por tonelada el 2013, volviendose un producto de alto precio y con alta potencialidad de demanda. Entonces, ¿por qué no incrementar su producción directamente?

Dos de los problemas que presenta *la Empresa* para tomar esta decisión se deben a dificultades por limitaciones de producción y control de calidad. Este primer problema tiene su raíz en que la empresa produce tanto nitrato de potasio como el sustituto de este, el cloruro de potasio, el cual es más barato y presenta mayor demanda, y como ambos utilizan en su procesamiento el material extraído del caliche, aumentar la producción de uno perjudica al otro. El segundo problema hace referencia al control que se tiene sobre el proceso productivo. La Empresa tiene 4 fábricas de procesamiento de nitrato de potasio, de entre los cuales 3 tienen implementados sensores que reportan indicadores del flujo de material en tiempo continuo, permitiendo sacar el máximo rendimiento al mineral utilizado. No es el caso de una de las plantas, la cual fue construida hace más de 30 años donde el control de calidad del proceso es realizada periódicamente con análisis químicos realizados manualmente por los técnicos y operarios. Por lo mismo, en ésta última no todo el material utilizado es aprovechado 100%.

¿Qué tan necesario es el monitoreo del proceso? Uno podría pensar que basta con encontrar una configuración óptima de las variables de control involucradas en el proceso (ya sean las temperaturas en que cada máquina debe funcionar, la cantidad de volumen que se debe ingresar, etc…) para obtener el máximo rendimiento del mineral utilizado, pero eso ignora un tercer factor que añade la complejidad al proceso, el producto final debe cumplir con ciertos niveles de concentración de algunos compuestos para poder comercializarse. Debido al proceso químico involucrado para tratar el material, con el producto final, se genera también un producto de descarte que no cumple con los niveles mencionados, el que algunas veces puede ser reutilizado como flujo auxiliar en el proceso o en otras simplemente desecharlo. Actualmente hay tres compuestos que son considerados contaminantes del producto final: Perclorato de Potasio ($KClO_4$), Cloruro de Sodio ($NaCl$) y Boro (B). Su presencia es soportada hasta ciertos niveles de concentración, donde por ejemplo, si el cloruro de sodio está en mayor o igual concentración del 0,95 %p/p[1], el producto final se considera impuro y no puede ser comercializado directamente. Cabe añadir que basta con que uno de los contaminantes supere los límites de concentración (la sección 4.2 aborda este tema) para que el producto entero se catalogue como impuro.

Frente a la complejidad del escenario, ¿es razonable pensar en aumentar el nivel productivo? La hipótesis que esta memoria plantea es que es posible incrementar la cantidad de nitrato de potasio a través de la disminución de los niveles de contaminantes que el sistema genera, de esta forma, con la misma cantidad de material introducido en el proceso, se aumenta el porcentaje de producto comercializable y se disminuye el flujo de descarte. Como esta memoria plantea el problema desde la ingeniería civil industrial, no pretende mejorar el proceso químico, sino la gestión de las variables de control del sistema. ¿Y esta hipótesis responde a los problemas de limitación de producción y control de calidad? Como busca mejorar la cantidad de nitrato de potasio manteniendo el mismo nivel de material utilizado, las limitaciones de producción dados por la elaboración de otros

[1] %p/p hace referencia al porcentaje peso-peso del compuesto, es decir, la cantidad en masa del compuesto dividido por la cantidad en masa de la solución.

compuestos, como el cloruro de potasio, no es un impedimento dado que no requiere disminuir una línea de producción para aumentar la otra. Por el lado del control de calidad, el foco está en mejorar la gestión que se hace en la planta más antigua ya que es solicitud expresa de *La Empresa* es buscar políticas de producción que disminuyas las ineficiencias operativas al no contar con un sistema de control continuo.

La fábrica de manufacturación del nitrato de potasio, como se verá más adelante, comprende 4 plantas por donde es tratado el material. La primera es la planta de tratamiento de sales donde se disuelven las sales del mineral ingresado y se demora aproximadamente 3,8 horas. La segunda es la planta de Muriato que añade al flujo KCl (conocido como Cloruro de Potasio o Muriato) para enriquecer la mezcla del sistema con potasio y demora aproximadamente 1,4 horas. La tercera es la planta Dual que separa el nitrato de sodio, resultante en las reacciones químicas, del nitrato de potasio, demorando aproximadamente 25 horas. Por último está la planta de cristalización, que como su nombre lo indica, busca obtener nitrato de potasio de forma cristalizada y demora aproximadamente 5,8 horas. Como resultado de todo el proceso aparecen dos flujos de salida finales: un primer flujo de nitrato de potasio denominado corriente L, y un segundo flujo, denominado corriente M, que intenta reaprovecharlos residuos que quedaron de la corriente L.

El sistema de plantas puede caracterizarse a través de las mediciones químicas que se realizan periódicamente, que se entenderán como *variables químicas*, y las decisiones que se hacen sobre la cantidad de material ingresado, volumen de agua utilizado, temperatura programada en las máquinas, entre otros, que se entenderán como *variables de control*. Es importante mencionar que sobre las variables químicas se tiene poco capacidad de gestión, dado que estas son el resultado de reacciones que dependen tanto de las condiciones químicas del material ingresado, como de las variables de control. Por el contrario, las variables de control, son los atributos que se pueden gestión directamente. En total el sistema contempla 310 variables, donde 156 son de control y 154 de tipo químico.

1.2 Justificación

La importancia creciente de los productos de nutrición vegetal de especialidad, tales como nitrato de potasio, se refleja en el hecho de que su consumo mundial se ha más que doblado durante la década pasada. El contexto mundial se ha vuelto más favorable a este tipo de producto debido a distintas condiciones[2]:

- Escasez de agua, disponible para la agricultura lo que genera un llamado a aumentar la eficiencia en el uso del agua.
- Creciente competencia por el uso de la tierra, entre vivienda, industria, naturaleza y agricultura.

[2] (Potassium Nitrate Association, 2010)

- Reducción de la disponibilidad de tierra arable con un alto costo por área, y altos costos de nutrientes y energía, promueven el uso de la fertirrigación, así como el uso eficiente del agua, nutrientes y energía.
- Incremento en el consumo de hortalizas per cápita.
- Creciente demanda por alimentos de alta calidad.

Frente a este escenario, el nitrato de potasio es un fertilizante idóneo para abordar el consumo agrícola dado que, por sus características descritas en el punto 3.4, permite abordar estas condiciones. Como se puede observar en el *Gráfico 1.2.1*, dentro de los usos que se le da al nitrato de potasio en el segmento de cultivos Premium de fertilizantes, un 41% corresponde a Vegetales. Con un crecimiento de la demanda entre un 2-3% el 2013 y una elasticidades del precio menor a la del cloruro de potasio[3], aprovechar este mercado es un oportunidad de negocio favorable.

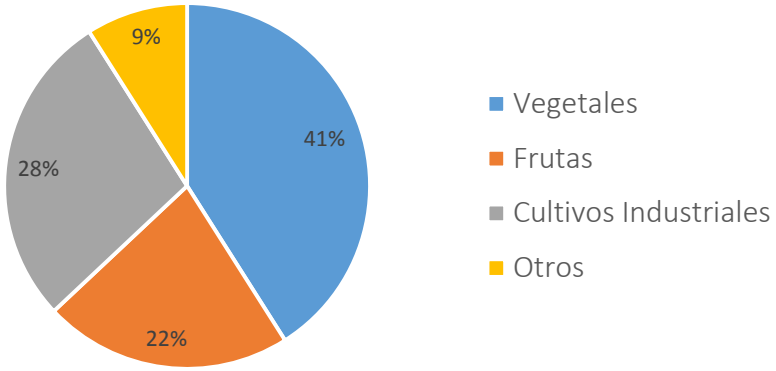


Gráfico 1.2.1: Porcentaje de usos del KNO3 en cultivos Premium

Fuente: Elaboración de *la Empresa*.

Si vemos el rendimiento de *La Empresa* en sus distintas líneas de negocio, como se puede ver en la *Tabla 1.2.1*, la contribución al margen bruto presentó su mayor crecimiento en el segmento de Nutrición Vegetal de especialidad, donde se encuentra el $\boldsymbol{KNO_3}$, con una variación del 3% versus el cloruro de potasio que sólo aumentó un 1%. El segmento de Nutrición vegetal de especialidad, representa el 33% de ingresos totales de *La Empresa*, siendo la línea con mayor contribución al ingreso y la segunda en contribución al margen bruto.

[3] Estimaciones de *La Empresa*

SEGMENTOS	Market Share	Producción (Mt*)	Ingresos (MMUS$)	% Ingresos	Contribución al Margen Bruto 2014	Variación vs 2013
Nutrición Vegetal de Especialidad[4]	48%	841	688	33%	25%	3%
Yodo y Derivados	28%	9	390	19%	31%	-4%
Litio y Derivados	27%	39	208	10%	13%	1%
Cloruro de Potasio y Sulfato de Potasio	50%	127	105	5%	7%	1%
Químicos Industriales		1.591	589	28%	23%	0%

*Mt es la abreviación de Megatoneladas que corresponde a un millón de toneladas
Fuente: La Empresa.

Tabla 1.2.1: Contribución al Margen Bruto por segmento

Actualmente existe un 20% de material que no cumple los límites de los contaminantes $KClO_4$ y $NaCl$, clasificándose como impuro y no aprovechable comercialmente. Sin embargo, durante el 2012, el primer contaminante mostró una moda por debajo del límite establecido, lo que abre la posibilidad de estudiar el comportamiento del proceso que dio este buen resultado y así replicarlo para el futuro.

En vista a lo descrito, el problema de negocios se concretiza en aprovechar la potencial demanda de unos de los segmentos que más contribución al margen bruto presenta, para así mejorar la rentabilidad de la producción de *La Empresa* incrementando la producción neta.

[4] El segmento de Nutrición Vegetal de Especialidad integra una serie de productos para nutrir el suelo agrícola, entre ellos el Nitrato de Potasio.

2 OBJETIVOS Y METODOLOGÍA

2.1 Objetivos

Una vez descrito la justificación del problema, se detalla a continuación los objetivos que se buscan alcanzar para dar solución a este problema de negocios. El elemento clave a analizar es la pureza con la que sale el producto final, pues este debe cumplir con los límites establecidos para el perclorato de potasio ($KClO_4$) y cloruro de sodio ($NaCl$) para ser vendido como abono a la agricultura. Entre los años 2006 y 2012, aproximadamente el 80% de la producción de nitrato potásico manufacturado por *La* Empresa, logró los estándares de pureza, es decir, un 20% del material no puede ser aprovechado quedando como desecho o como residuo reprocesado en el sistema.

2.1.1 OBJETIVO GENERAL

Optimizar la calidad de producción del nitrato de potasio en su proceso metalúrgico de fabricación, mediante el mejoramiento de sus políticas de operaciones a fin de incrementar el porcentaje de pureza usando reglas de decisión que sean integrables al funcionamiento actual de la planta.

2.1.2 OBJETIVOS ESPECÍFICOS

1. Caracterizar la pureza del nitrato de potasio mediante la selección de variables, presentes en el proceso, que tengan una capacidad explicativa de su resultado.

2. Desarrollar un modelo matemático que integre poder explicativo y predictivo de la pureza del nitrato de potasio como consecuencia de la combinación de las variables de control del proceso.

3. Análisis descriptivo y comparativo del modelamiento del proceso para comprender el meta-aprendizaje que el modelo otorga a la producción.

4. Análisis comparativo de las posibles mejoras que se pueden llevar en base al modelamiento en contraste con el proceso actual, para medir la creación de valor que permite este enfoque de trabajo.

2.1.3 ALCANCES

Los alcances del proyecto contemplan la realización de políticas de operación que pueden ser seguidas por los operarios de las plantas con tal de asegurar los estándares de calidad requeridos. En contraste, este proyecto no pretende evaluar en la práctica cómo el seguimiento de esta política contribuye a la rentabilidad de La Empresa, o el de verificar la

efectividad en terreno. La memoria queda enmarcada por el modelamiento del funcionamiento y sus proyecciones de mejora.

Las proyecciones esperadas en este trabajo integran aspectos operativos y aspectos teóricos. Por el lado operativo, se espera desarrollar un manual de políticas que pueda seguirse para regular el funcionamiento de las instalaciones de forma tal de asegurar la producción del nitrato de potasio a niveles de pureza altos a mayor frecuencia. Además, se pretende mostrar un modelo predictivo y explicativo ajustado al proceso que pueda ser gestionado por el área operativa de *La Empresa* para así obtener nuevas políticas integrando los cambios del entorno que puedan darse debido a la dinámica del mercado (disminuir la producción, funcionar a menor escala, balancear el proceso por alguna falla, entre otros).

Por el lado teórico, la memoria pretende evaluar distintos tipos de algoritmos de data mining, por ejemplo Random Forest o SVM de forma aplicada en el campo de la minería no metálica. Se espera poder contrastar estos modelos predictivos y servir de referencia para futuras investigaciones.

2.2 Metodología

Para la realización de esta memoria se utiliza, como marco metodológico, las etapas propuestas por el proceso KDD[5], adaptándolo a las condiciones específicas que se requieren en este tema en vista del objetivo general y los objetivos específicos. Por esto, se desarrollan 3 etapas del trabajo:

- **Estudio del proceso y variables implicadas**

Enfocada en la caracterización del primer objetivo específico, esta etapa busca trabajar la estructuración de la información, utilizando los pasos 3 y 4[6] de la metodología KDD. Para ello se describe la cantidad de data factible a utilizar, se estandariza la unidad temporal de estudio, se selecciona las variables implicadas en la pureza del resultado del KNO3 y se completan los valores perdidos.

- **Modelamiento del Proceso**

Enfocada en el segundo objetivo específico, esta etapa busca desarrollar un modelo que permita explicar y predecir la pureza del nitrato de potasio, como resultado de las variables implicadas, utilizando los pasos 5,6 y 7 de la metodología KDD. Para ello se contempla el estudio e implementación de modelos de data mining de clasificación y regresión que mejor se ajusten a la base de datos, para su posterior contraste y elección.

---

[5] Correspondientes a los mencionados en punto 5.1.1.

[6] Los pasos 1 y 2 se consideran abordados en el capítulo 3 y 4.

- **Diseño de pruebas y validación del modelo**

Enfocado en el tercer objetivo específico, esta etapa busca comprender y comparar el meta-aprendizaje que otorgue el modelo elegido y evaluar su poder explicativo y predictivo, utilizando el paso 8 de la metodología KDD. Para ello se describe el ajuste del modelo con la data histórica, si existe sobreajuste, los posibles sesgos que puedan existir y nivel de predicción.

- **Análisis de sensibilidad de resultados**

Enfocado en el cuarto objetivo específico, esta etapa busca determinar y evaluar las mejorar que el modelo encontrado otorgue al proceso, utilizando el paso 9 de la metodología KDD. Para ello, se analiza la importancia de las distintas variables, su impacto en el nivel de pureza del proceso y se comparan rangos operacionales que mejoren la producción.

3 INDUSTRIA DEL NITRATO

3.1 Contexto Industrial del Nitrato

Dentro de la historia de Chile, se encuentra la historia del Salitre, mineral propio de las tierras nortinas que significó el florecer de todo un país en torno a él, tanto como la imagen de un país minero como la movilización de familias que dejaban su tierras natales en el sur de Chile buscando nuevas oportunidades de trabajo en las salitreras del norte. Pero la aparición del salitre sintético implicó el decaimiento de la industria nacional.

El salitre es un mineral que dada su riqueza en compuestos como Cloruro de Potasio o Nitrato de potasio es un potencial abono para la agroindustria, pero requiere de una serie de tratamientos químicos para lograr separarlo en compuestos aprovechables, puesto que también trae otros elementos o impurezas que no contribuyen a la agricultura, como por ejemplo el boro. Lograr extraer el mayor número de impurezas del mineral es un trabajo que varias empresas han realizado y perfeccionado a lo largo del tiempo. Esta memoria, busca contribuir a este mismo objetivo, enfocado en uno de los subproductos del salitre, el nitrato de potasio o $\boldsymbol{KNO_3}$.

En la *Tabla 3.1.1* se puede apreciar los distintos rubros en donde los nitratos tienen utilidad.

Industrias	**Agricultura**	Fertilizante soluble y una fuente de nitrógeno nítrico y potasio virtualmente libre de cloruro.
	Industria	Diversas aplicaciones, incluyendo manufactura de vidrio, explosivos para minería y obras civiles, entre otros.
	Alimento	Una vía para cuidar y preservar comidas contra el ataque de agentes microbiales.
	Farmacéutica	Ingrediente común en pastas dentales sofisticadas.
	Plantas de Concentración de Energía Solar	Almacenaje para acumular energía térmica en las plantas de concentración solar de potencia.

Tabla 3.1.1: Industrias relevantes para el uso del KNO3

Fuente: www.kno3.org

3.2 Producción de Nitratos

Según información estadística de los últimos 10 años entregada por Cochilco (Comisión Chilena del Cobre), la producción nacional ha presentado una disminución del 45,8% respecto al 2004, llegando a 759,4 Toneladas el 2013. De la misma forma, la exportación también ha presentado una baja de un 20,7% respecto al 2004, llegando a 354 toneladas el

2013. Por otro el valor unitario ha experimentado un alza del 179,6% con respecto al 2004, llegando a US$ 834,7 por tonelada. En el *Gráfico 3.2.1* se puede observar la evolución de estos indicadores entre 2004 y 2013.

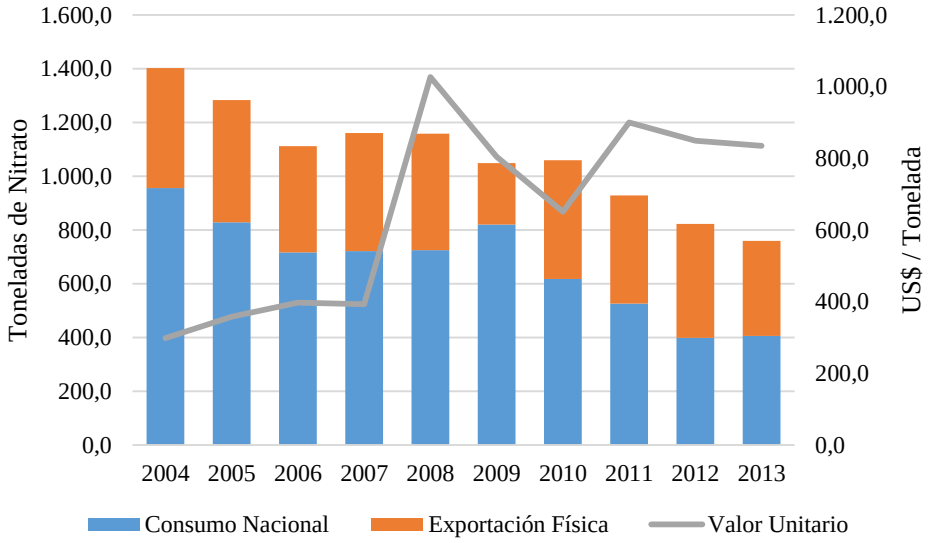


Gráfico 3.2.1: Producción, Exportación y precio de nitratos entre 2004-2013

Fuente: Cochilco, 2014.

Las causas de las disminuciones en la producción y exportación nacional pueden tener su origen tanto en los problemas operacionales que se han registrado en la industria en los últimos años como a condiciones del mercado al favorecer la producción de otros productos nítricos. Cabe señalar que los nitratos naturales de Chile, en el marco de la agroindustria, compiten con productos similares de origen sintético, tales como el nitrato de calcio, el nitrato de amonio o la urea que tienen un valor menor. En este último punto se acentúa el cloruro de potasio (***KCl***), otro fertilizante usando en la industria, que ha experimentado un auge en los últimos años. Como se puede observar en el *Gráfico 3.2.2,* la producción de ***KCl*** ha aumentado un 147,6% respecto al 2004, llegando a 1.838 mil toneladas y el precio ha caído un 22%, llegando a US$ 369,1 por tonelada. De esta forma se vuelve un sustituto más económico con mayor oferta. Por último, es bueno hacer notar que el foco de mercado del $\boldsymbol{KNO_3}$ son cultivos de alta calidad, pues el mayor valor unitario del fertilizante requiere cultivos que reporten un margen mayor para hacer rentable la inversión[7].

[7] (Comisión Chilena del Cobre, 2013)

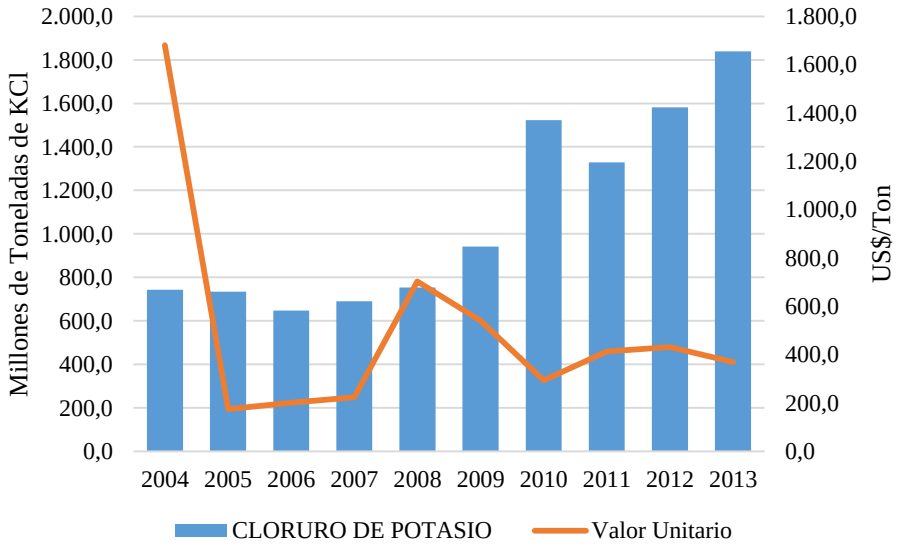


Gráfico 3.2.2: Producción y precio del Cloruro de Potasio

Fuente: Cochilco, 2014.

3.3 Empresas

La industria nacional de los nitratos es reducida a pesar de que se han agregado nuevos actores al sector, "el nivel de producción se mantiene estancado. Según expresiones de ejecutivos de la empresa SQM (se describe más adelante), el negocio de los nitratos propiamente tal no se sostendría sin la coproducción de yodo"[8].

Según la ODEPA[9], actualmente la industria nacional de nitratos concentra únicamente a sólo 3 empresas:

1. **SQM S.A.:** es una empresa minera privada que se dedica a la explotación, procesamiento y comercialización del salitre, iodo y litio en Chile. Dispone de una capacidad instalada de 950 mil toneladas por año para producir nitratos, sin embargo, se estima que alrededor del 32% de esa capacidad está destinada exclusivamente a la producción de nitrato de potasio. La extracción minera la realiza en tres faenas ubicadas en la Región de Antofagasta (Pedro de Valdivia, María Elena y Pampa Blanca) más una faena en la Región de Tarapacá (Nueva Victoria).
2. **Cosayach Nitratos S.A.:** El año 2002 puso en marcha en Cala Cala una planta con capacidad potencial de producción de 200 mil toneladas por año de nitrato de potasio, a partir de nitrato de sodio cristalizado en sus tres faenas, que también

[8] Competitividad Chilena En Los Recursos Salinos, 2012

[9] Oficina de Estudios y Políticas Agrarias, 2010. Ministerio de Agricultura.

coproducen yodo, ubicadas en la Región de Tarapacá (Cala Cala, Negreiros y Soledad).

3. **ACF Minera S.A.:** compañía chilena dedicada a la explotación y producción de yodo y nitratos naturales. dispone de una capacidad instalada de 20 mil toneladas por año de nitrato de sodio en su faena ubicada en la Región de Tarapacá (Laguna), donde coproduce yodo.

Mientras a nivel nacional la producción es de nitrato de potasio natural, a nivel internacional, la industria compite con la producción sintética del fertilizante. Los mayores competidores internacionales se detallan a continuación y el Gráfico 3.3.1 resume su relación con las empresas chilenas.

1. **Haifa Chemicals Ltd.:** corporación internacional privada que fabrica principalmente fertilizantes para la agricultura y los productos químicos para la industria alimentaria. Dispone una capacidad instalada de 500 mil toneladas por año de nitrato de potasio sintético y tiene 2 plantas productivas ubicadas en Israel y 1 en Francia.
2. **Kemapco Arab Fertilizers & Chemicals Industries LTD:** empresa árabe, cuenta con una capacidad de producción de la planta de 150 mil toneladas por año de nitrato de potasio sintético y hasta 60 mil toneladas por año en lotes de fosfato dicálcico (DCP) aditivo para la alimentación animal, además de ácido nítrico que se utiliza actualmente para su propio consumo. Las instalaciones de producción se encuentran en Aqaba - Jordania.

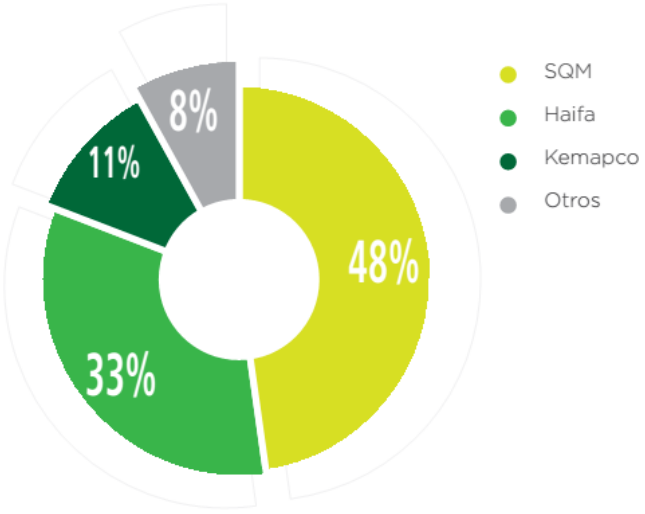


Gráfico 3.3.1: Competencia en la Industrial de Nitrato de Potasio

Fuente: Memoria 2014, SQM.

3.4 Nitrato de Potasio

El nitrato potásico ($\boldsymbol{KNO_3}$) es un producto cristalino, total y rápidamente soluble en agua. Ocasionalmente se presenta en la naturaleza en estado puro en depósitos de sales, pero normalmente se encuentra en forma de sales dobles minerales, en combinación con nitratos de calcio, magnesio y sodio. Este puede ser nitrato de potasio natural o sintético, que se diferencian en que el primero usa como insumo principal minerales ricos en nitratos desde yacimientos mineros, mientras que el segundo no requiere de estos yacimientos ya que usa combinaciones industriales de sus elementos para producirlo.

El compuesto KNO_3 se caracteriza por ser un fertilizante natural con gran potencial dado sus atributos[10]:

1. **Proporciona potasio exento de cloro:** Una abundancia de cloro puede disminuir los rendimientos y deteriorar la calidad de muchos cultivos. Esto es una ventaja del Nitrato de potasio (tanto natural cómo sintético) frente al cloruro de potasio, pues mientras el segundo contiene 47% de cloro, el primero contiene menos del 1% de cloro. Cantidades perjudiciales de cloro para los cultivos pueden presentarse como consecuencia de un elevado aporte de este elemento por parte de los fertilizantes, suelos, agua de riego y algunos plaguicidas. El estiércol y materias orgánicas también aportan cloruro y sales a los suelos. El único síntoma visual general de exceso de cloro consiste en hojas de tamaño reducido y una menor velocidad de crecimiento. En algunas especies se producen síntomas específicos, como quemaduras en las puntas o márgenes de las hojas, color tabaco y caída prematura de ellas y, en algunos casos, clorosis. El nitrato de potasio está prácticamente libre de cloro, ya que su contenido es menor al 1%.
2. **Es altamente soluble en agua:** Esta cualidad presenta ventaja al poder ser aplicado disuelto en agua. Su solubilidad es mucho mayor que la del sulfato de potasio, que es la otra fuente de potasio exenta de cloro.
3. **No incluye elementos innecesarios:** El potasio y el nitrógeno nítrico pueden ser absorbidos en su totalidad por las plantas, sin dejar residuos que puedan provocar acumulación de sales en el suelo o en otro medio de cultivo.
4. **El nitrógeno está en forma de nitrato:** El nitrógeno en forma de nitrato es inmediatamente aprovechable por las plantas sin necesidad de transformaciones previas. Ello aún en suelos finos, mojados, ácidos y fumigados, como también en periodos de tiempo semi-secos. El nitrato tiende a estimular la absorción de potasio, magnesio y calcio y a deprimir aquellas de cloruro, mientras que el amonio presenta los efectos opuestos.
5. **Es un fertilizante doble:** El KNO3, aporta dos de los tres elementos nutritivos minerales primarios de las plantas: Nitrógeno Nítrico y Potasio. La presencia del ion nitrato ($\boldsymbol{NO_3}$-) estimula la absorción de potasio, magnesio y calcio, por lo tanto, la aplicación simultánea de potasio y nitrógeno nítrico a través del KNO3, presenta una mejor eficiencia en la absorción del potasio y otros elementos secundarios.

[10] (Garcés Millas, 2000)

Los nitratos, tales como el nitrato de potasio, el nitrato de sodio, nitrato de litio, entre otros, se obtienen en Chile a partir de la explotación de los campos de nitratos, que se localizan en una franja de aproximadamente 700 km de largo por 30 a 50 km de ancho, que se ubica en el norte de Chile, al este de la Cordillera de la Costa, en las regiones de Tarapacá y de Antofagasta. Ésta es la única zona del mundo donde los depósitos de nitratos naturales tienen reservas y recursos con contenidos económicos[11], y cuya mena, denominada caliche, permite obtener diferentes productos como nitratos de sodio, nitratos de potasio, yodo y sulfato de sodio. El caliche se presenta, preferentemente, como una capa densa y dura de arenas y gravas cementadas con sales, con espesores variables entre 0,5 metros y 5 metros. La producción de nitratos para el año 2012, corresponde a 822.584 toneladas, de las cuales el 98% se originó en la Región de Antofagasta y solamente un 2% en la Región de Tarapacá[12].

[11] Existen yacimientos en distintas partes del mundo, pero por su baja concentración, no permiten hacer de la extracción y producción del nitrato de potasio natural un negocio sustentable en tiempo.

[12] (Servicio Nacional de Geología y Minería, 2013)

4 DESCRIPCIÓN DEL PROCESO

4.1 Plantas

La memoria analiza una de las centrales de producción de nitrato de potasio que tiene *La Empresa*, la cual involucra una serie de plantas de tratamiento donde el sistema de control de gestión está basado en registros y estadísticas de distintas variables en diferentes momentos del día.

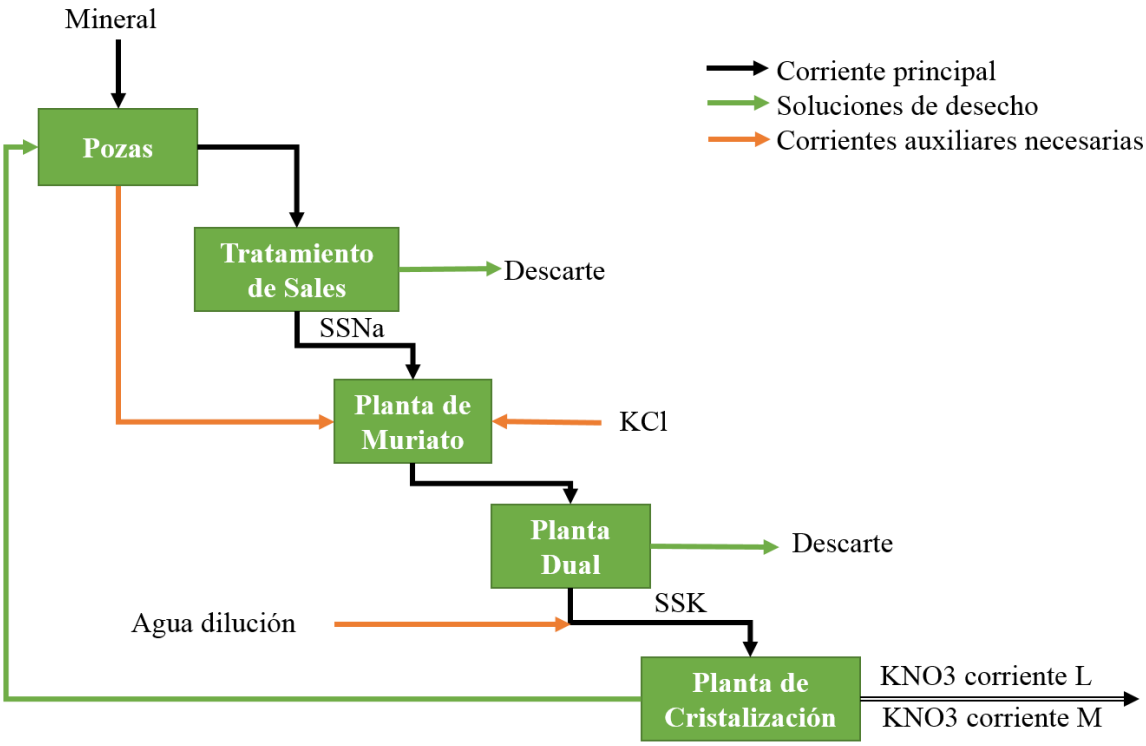


Ilustración 4.1: Procesamiento del KNO3[13]

Fuente: La Empresa

El proceso consta de flujos de entrada y flujos de salida de material, denominados "Corrientes", en cada una de las etapas, como se observa en la Ilustración 4.1. Este trabaja por lotes de material y consta de 5 etapas que se dan en distintas plantas:

1. **Pozas:** En esta etapa se procesa el caliche y se extraen las corrientes necesarias para los distintos productos de *La Empresa*. Como ésta es una etapa nodo entre otras plantas y no de exclusividad del proceso productivo del nitrato de potasio, se deja fuera del modelamiento del trabajo, usando la corriente de salida correspondiente como el input del sistema buscado.

[13] El KCl es conocido en la industria química como Muriato

2. **Planta de Tratamiento de Sales:** Etapa donde se produce la disolución de las sales y demora aproximadamente 3,8 horas en el proceso. De esta planta salen dos corrientes principales: la primera compuesta principalmente por sales insolubles que se desechan, denominada "Descarte", que en promedio hacen 450 toneladas mensuales en flujo de salida, y una solución rica en NaNO3 denominada SSNa, con una concentración promedio de 450 gramos por litro.

3. **Planta de Muriato:** alimentada con la solución SSNa, se une con sales de KCl que se agregan de fuentes externas al sistema a un tasa aproximada de 24 toneladas por hora, en donde ocurre la transformación del NaNO3 de acuerdo a la siguiente reacción química:

$$\boldsymbol{NaNO_3 + KCl \leftrightarrow KNO_3 + NaCl}$$

   Demora aproximadamente 1,4 hrs. en el proceso y genera como corriente de salida una pulpa formada por cristales de NaCl y una solución de KNO3, con una concentración promedio de potasio de 147,3 gramos por litro y un promedio diario de 1.897,1 toneladas.

4. **Planta Dual:** En esta etapa, el NaCl cristalizado es separado de la solución que contiene el KNO3 en el espesador y demora aproximadamente 25 hrs en el proceso. De la etapa dual salen dos corrientes, una de sales de desecho o descarte (principalmente NaCl) y la otra una solución de Nitrato de Potasio llamada SSK, la cual tiene una concentración promedio de potasio de 147 gramos por litro.

5. **Planta de Cristalización:** La solución SSK se diluye con el fin de evitar la cristalización del NaCl. Se obtienen como producto final del proceso, cristales de KNO3 por enfriamiento de la solución diluida y que salen en dos tipos de corrientes distintas:
   - **Corriente L:** Primer flujo de salida, que contiene el material con menores índices de contaminantes. Se estima su producción promedio en 1.044 toneladas diarias.
   - **Corriente M:** Segundo flujo de salida, que consiste en reaprovechar los residuos que quedaron de la primera corriente. Por lo mismo, contiene mayores índices de contaminantes. Se estima su producción promedio en 721 toneladas diarias.

   Demora aproximadamente 5,8 hrs. en el proceso y el desecho de la segunda corriente se recircula al sistema de pozas para su uso en el proceso.

En resumen, el proceso completo contempla 36 horas de producción, consierando 4 plantas de tratamiento.

4.2 Contaminantes

El proceso en cuestión también genera residuos de elementos no eliminados en el producto final, lo que agrega el grado de pureza o impureza al resultado y determina su posibilidad de comercialización. Existen 4 residuos que afectan la pureza del material: Nitrógeno (N), Boro (B), Perclorato de Potasio ($KClO_4$) y Cloruro de Sodio ($NaCl$). Mientras que el primero se necesita en cantidades mínimas para certificar la calidad del producto, los 3 restantes deben estar en concentraciones por debajo del especificado (véase Tabla 4.2.1**)**. Las concentraciones se miden en porcentaje peso-peso (%p/p) que significa la cantidad de masa del soluto por cantidad de masa de la solución.

	Corriente L		Corriente M	
	Condición	$\% p/p$	Condición	$\% p/p$
$NaCl$	Máx.	0,95%	Máx.	0,95%
$KClO_4$	Máx.	0,24%	Máx.	0,95%
B	Máx.	0,05%	Máx.	0,10%
N	Mín.	13,50%	Mín.	13,50%

Tabla 4.2.1: Concentraciones de compuestos que describen calidad

Fuente: La Empresa.

Sin embargo, el registro que lleva *La Empresa* no dispone de una medición frecuente de las concentraciones del boro y del nitrógeno, tanto de la corriente L como de la M.

	Corrientes L		Corriente M		
	# Datos	*# Vacíos*	*# Datos*	*# Vacíos*	Total
$NaCl$	78%	22%	78%	22%	31.389
$KClO_4$	81%	19%	81%	19%	
B	1%	99%	3%	97%	
N	0%	100%	0%	0%	

Tabla 4.2.2: Cantidad de datos de los contaminantes por corriente.

Fuente: Elaboración del autor

El trabajo se realiza bajo el supuesto de que el boro y el nitrógeno, debido al bajo nivel de medición que realiza *La Empresa*, son elementos que no afectan en demasía la comercialización del producto final, por lo que el foco de atención será en los contaminantes $NaCl$ y el $KClO_4$. Se deja abierto para alternativas de estudio, analizar el impacto de estos elementos en la producción de nitrato de potasio.

En la *Tabla 4.2.3* y la Tabla 4.2.4 se puede apreciar la presencia histórica de los contaminantes en las corrientes de salida. Debido a la diferencias en la frecuencia de medición[14], el eje horizontal se describe como número de observación, siendo un total de 31.389 ordenadas cronológicamente. Se añade además, una leyenda que describe a la

[14] La naturaleza de la frecuencia de medición es discutido más adelante, en el punto 6.1 Determinación de la Unidad de Tiempo.

concentración del contaminante como puro (color negro) o impuro (color rojo) y respectivo porcentaje de presencia en la muestra. De la Tabla 4.2.1 se extraen los límites en los cuales los niveles de concentración de cada contaminante, $KClO_4$ y $NaCl$, generan que la producción sea catalogada como impura, siendo 0,24%p/p y 0,95%p/p en la corriente L y 0,95%p/p y 0,95%p/p en la corriente M, de cada contaminante respectivamente. Para ver el detalle de los datos véase ANEXO A .

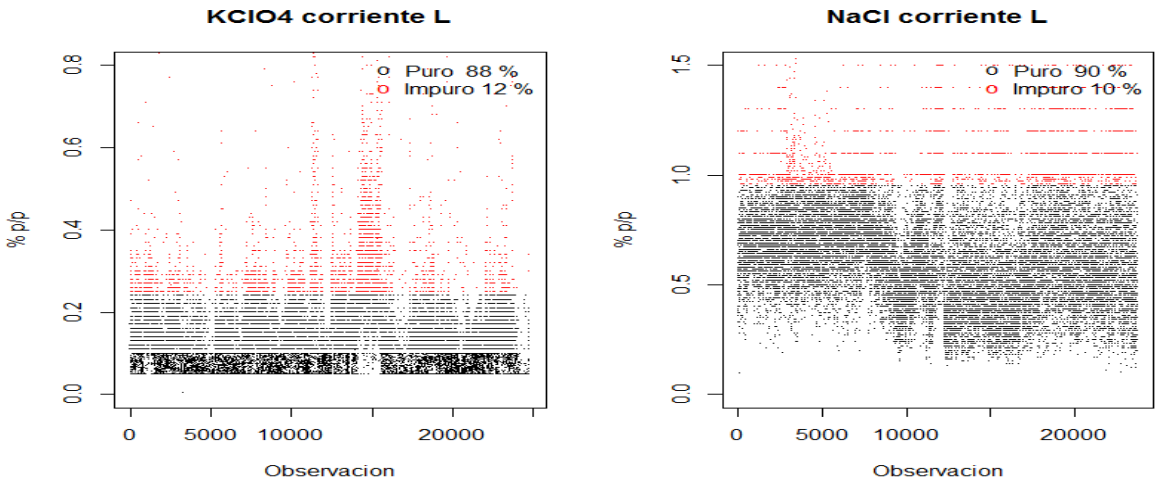


Tabla 4.2.3: Corriente L entre los años 2006-2012

Fuente: Elaboración del autor.

Como se puede apreciar, el porcentaje de observaciones "*puro*" es más elevado en la corriente L que en la M y sus concentraciones no superan los 1,5 %p/p en contraste con el 2,5% que alcanza el $NaCl$ en la corriente M.

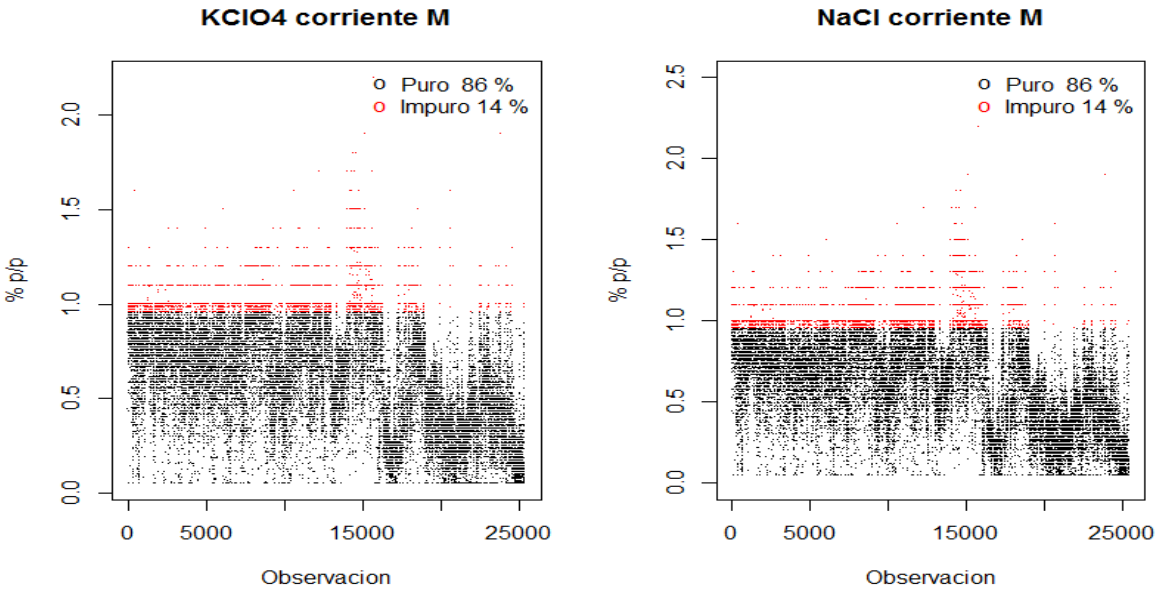


Tabla 4.2.4: Corriente M entre los años 2006-2012

Fuente: Elaboración del autor.

El principal enfoque de esta memoria es lograr reducir los porcentajes de impureza de estos contaminantes en las corrientes L y M.

4.3 Descripción de la Base de Datos

La base de datos está compuesta por el registro manual de los operadores que trabajan en la planta en distintos turnos y tiempos del proceso, por lo que no todos los registros presentan la misma unida temporal. Así, como se observa en la *Ilustración 4.3*, el tiempo de cada etapa difiere y las unidades de medición pueden ser por turnos, días, horas, bi-horas, entre otros.

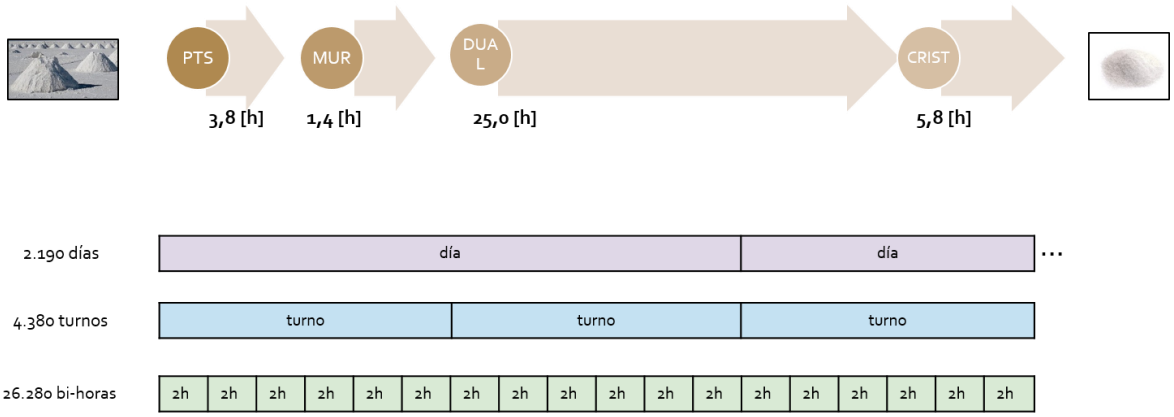


Ilustración 4.3: Descripción temporal de las mediciones

Fuente: La Empresa

La información disponible se puede clasificar en dos tipos de variables: variables químicas de las corrientes involucradas en el proceso y las variables de control del sistema.

4.4 Variables Químicas

Existen 28 flujos de material en el proceso, en las cuales se realizan distintos tipos de mediciones químicas y en cada una se realizan diferentes análisis dependiendo de los compuestos que se estén trabajando (véase ANEXO B para ver el listado de mediciones y sus correspondiente etapa). Con estas mediciones se construyen las 154 variables químicas presentes en el trabajo. Por ejemplo se puede nombrar la variable *Concentración del* $NaNO_3$ *en la corriente SSK* la cual hace referencia a la medición de la concentración del compuesto $NaNO_3$ que se hace en la corriente de nombre SSK que sale de la Planta Dual. Estos datos se miden en porcentajes de concentración masa/masa[15].

[15] El porcentaje masa/masa se define como la masa de soluto (sustancia que se disuelve) por cada 100 unidades de masa de la solución

La cantidad de información que tenga registrada cada variable química será un factor clave a la hora de elegir cuáles de estas utilizar en el modelamiento[16]. Otro factor determinante es la temporalidad de los datos de cada muestra. Como se observa en la *Tabla 4.4.1*, la variedad de tipos de medición hará necesaria la elección de una unidad estándar de tiempo para general el modelo y su predicción.

Frecuencia	Cantidad de Flujos
Cada 3 hrs.	1
Cada 12 hrs.	1
Bihoraria	2
Sin Información	2
TOTAL	6

Tabla 4.4.1: Frecuencia de mediciones Planta de Cristalización

Fuente: Elaboración del autor.

4.5 Variables de Control

Existen 156 variables operacionales dentro de las 4 etapas del proceso productivo. Como se observa en la *Tabla 4.5.1*, cada una presenta su descripción, unidad de medida y frecuencia de medición respectiva (véase ANEXO C para ver el listado de variables de control y sus correspondiente etapa).

PLANTA	Grupo	Descripción	Unidad	Periodo
PTS	Temperaturas	T. Entrada a Planta P.T.S.	C	TURNO
	Flujos Planta	Agua Centrifuga	Lt	TURNO
MURIATO	Entradas	KCl Agregado Minsal Bajo Cal.	TON	TURNO
	Soluciones (Flowmeter)	Poza 10 a Muriato	M3	TURNO
DUAL	Parámetros DELKOR	Densidad Pulpa	GC3	HORA
	Mediciones DELKOR	Agua Dulce	M3	TURNO
CRISTALIZACION	Medidores M3	Soluc. Tratada SS N° 1	M3	TURNO
	Producción Bruta Desglosada	Prod. de Refinado	TON	HORA
	DESCARGAS	Descarga - Disponibilidad MLT 1	%	HORA

Tabla 4.5.1: Ejemplo de variables de control

Fuente: La Empresa

.

[16] Uno de los problemas que se verá en el capítulo 6.3 será la presencia de valores ausentes en a base de datos

Así como en las variables químicas, las variables de control presentas distintas unidades de medición. La *Tabla 4.5.2* muestra el ejemplo de la planta de cristalización.

Frecuencia	Nº Corrientes
Hora	27
Turno	17
TOTAL	44

Tabla 4.5.2: Frecuencia de mediciones de control en Planta de Cristalización

Fuente: La Empresa.

5 MARCO TEÓRICO

5.1 Minería De Datos

La minería de datos o exploración de datos es un campo de las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos, que implica la inferencia de los algoritmos que exploran los datos, desarrollar el modelo y descubrir patrones previamente desconocidos. El modelo se utiliza para la comprensión de los fenómenos de los datos, análisis y predicción. Se enmarca dentro de la metodología *"Knowledge Discovery in Databases"* (KDD), la cual se explica a continuación.

5.1.1 METODOLOGÍA KDD

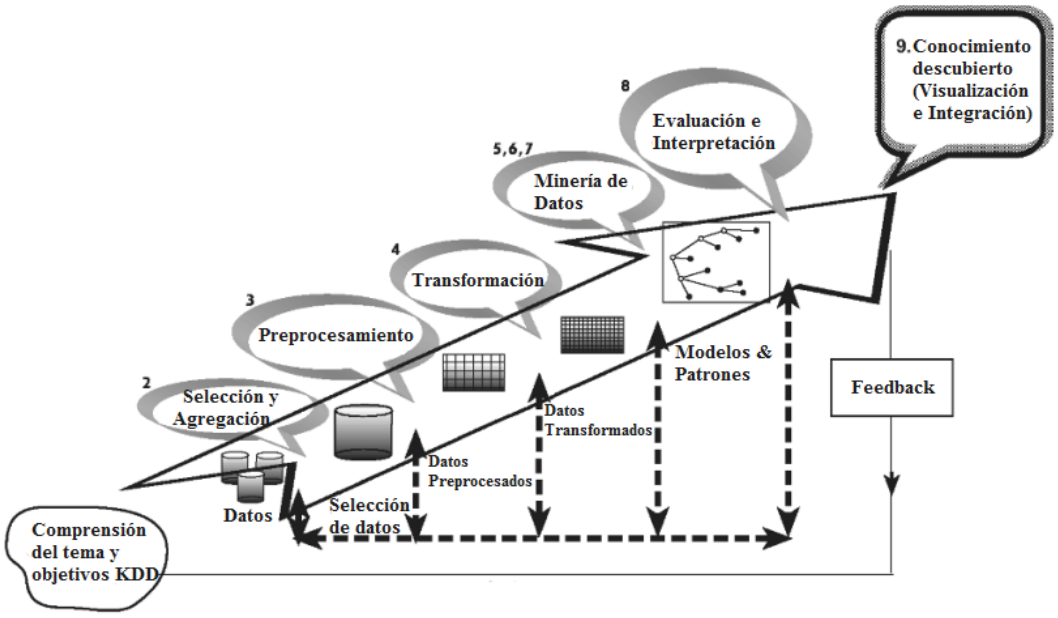


Ilustración 5.1: Metodología KDD en la base de datos

Fuente: Maimon & Rokach (2010). Traducción del autor.

La metodología *Knowledge Discovery in Databases* consta de una serie de pasos que van, desde la conceptualización del tema y sus objetivos, hasta el descubrimiento e integración de conocimiento. La literatura no establece un número estándar de pasos, por lo que a continuación se presentan las 9 etapas sugerida por Maimon & Rokach (2010):

1. **Desarrollo de una comprensión del dominio de aplicación:** Este es el paso preparatorio inicial. Se busca la comprensión de lo que se debe hacer con las alternativas de decisión (transformaciones, algoritmos, representaciones, entre otros).

En un proyecto de KDD, se necesita entender y definir los objetivos del usuario final y del contexto en el que el proceso de descubrimiento de conocimiento se llevará a cabo (incluyendo el conocimiento previo relevante). Por ser una etapa de partida, puede requerir revisiones y ajustes.

2. **Selección y creación del conjunto de datos de trabajo:** Habiendo definido los objetivos, se deben determinar los datos que se utilizarán .Esto incluye averiguar la disponibilidad de la información, la obtención de datos adicionales necesarios, para su posterior integración en una base única. Esta debe incluir los atributos importantes del proceso, puesto que su ausencia afecta el poder explicativo y predictivo del modelo.

   Para el éxito del proceso, es bueno considerar el mayor número posible de atributos en esta etapa. Se inicia con el mayor conjunto de datos disponibles y luego se reduce su tamaño para evitar costos por complejidad en la manipulación de la información.

3. **Preprocesamiento y limpieza:** En esta etapa, se vuelve a revisa la fiabilidad de la información y se incluye la limpieza de datos, tales como el manejo de los valores perdidos y la eliminación de ruido o valores atípicos.

4. **Transformación de datos:** En esta etapa, se prepara y se adapta los datos para simplificar su manejo y modelamiento. Aquí se ven métodos como la reducción de la dimensión (por ejemplo, funciones de selección y extracción, o la iteración del muestreo), y la transformación de atributos (como la discretización de sus valores numéricos o la transformación funcional). Sin embargo, incluso si se desconoce transformaciones útiles desde el principio, el proceso KDD mismo conduce a una comprensión de la transformación necesaria, posibilitando su aplicación en pasos posteriores.

Después de haber completado los cuatro pasos anteriores, los siguientes cuatro pasos están relacionados con la parte de minería de datos, donde la atención se centra en los aspectos algorítmicos empleadas para cada proyecto.

5. **Elección de un task apropiado:** El task hace referencia al tipo de respuesta que se busca para el conjunto de información. Esta puede ser, por ejemplo, una clasificación, una regresión, o una agrupación, dependiendo en gran medida de los objetivos KDD, y también en los pasos anteriores.

   Hay dos objetivos principales en Minería de Datos: la predicción y la descripción. Predicción a menudo se refiere a modelos supervisados (se explicará en el punto 5.1.2), mientras que descriptiva incluye modelos no supervisados y de visualización. La mayoría de las técnicas de minería de datos se basan en el aprendizaje inductivo, donde se construye un modelo de manera explícita o implícita por generalizar a partir de un número suficiente de datos de entrenamiento. El supuesto subyacente del enfoque inductivo es que el modelo entrenado es aplicable a los casos futuros.

6. **Elección de algoritmos:** Esta etapa se enfoca en seleccionar métodos específico de data mining que se utilizará para la búsqueda de patrones. La decisión se relaciona con el meta-aprendizaje de los algoritmos, esto quiere decir, en la explicación de lo que hace que un algoritmo de minería de datos para tener éxito o no en un problema particular. Por ejemplo, en la consideración de precisión frente a la comprensibilidad, la primera es mejor con redes neuronales, mientras que el segundo es mejor con árboles de decisión. Así, tratar de comprender las condiciones en las que un algoritmo de minería de datos es el más apropiado.

7. **Implementación del Algoritmo:** En este paso se aplica el o los algoritmos seleccionados a los datos. Puede ser que se necesite emplear el o los algoritmos reiteradas veces para obtener resultados satisfactorios, por ejemplo, mediante el ajuste de parámetros de control de los modelos.

8. **Evaluación:** En esta etapa se evalúa e interpretar los patrones encontrados en la etapa anterior con respecto a los objetivos definidos en el primer paso. Aquí se debe tener en cuenta los pasos de preprocesamiento para explicar los efecto sobre los resultados de el o los algoritmos. Este paso se centra en la comprensión y la utilidad del modelo inducido y se documenta el aprendizaje para su posterior uso.

9. **Uso del conocimiento descubierto**: Se incorpora el conocimiento al sistema para la acción futura. El conocimiento se vuelve activo en el sentido de que es posible realizar cambios en el sistema y medir los efectos. El éxito de este paso determina la eficacia de todo el proceso de KDD.

5.1.2 MÉTODOS SUPERVISADOS[17]

Métodos supervisados son métodos que tratan de descubrir la relación entre atributos de entrada (a veces llamadas variables independientes) y un atributo de destino (a veces referido como una variable dependiente). La relación descubierta está representada en una estructura conocida como un modelo. Por lo general, los modelos describen y explican fenómenos, que están ocultos en el conjunto de datos y pueden ser utilizados para predecir el valor del atributo de destino conocer los valores de los atributos de entrada. Los métodos supervisados se pueden implementar en una variedad de dominios tales como el marketing, las finanzas y la manufactura.

Es útil distinguir entre dos modelos supervisados principales: *modelos de clasificación* (clasificadores) y *modelos de Regresión* (regresor). Los modelos de regresión mapean desde el espacio de entrada en un dominio de valor real. Por ejemplo, un regresor puede predecir la demanda de un determinado producto dadas sus características. Por otro lado, los clasificadores mapean desde el espacio de entrada en clases predefinidas. Por ejemplo, los clasificadores se pueden utilizar para segmentara los consumidores hipotecarios como

[17] (Maimon & Rokach, 2010)

buenos (amortización totalmente la hipoteca a tiempo) y malos (en diferido de amortización).

Para la aplicación de estos modelos se particiona la base de datos en dos grupos, un *conjunto de entrenamiento* y un *conjunto de testeo*. El primero se utiliza para conseguir la descripción del modelo y el segundo se utiliza para ver el rendimiento de este, su nivel de sobreajuste, su capacidad predictiva, entre otros.

5.1.3 REGRESIÓN LOGÍSTICA[18]

La regresión logística (Ruczinski, Kooperberg, & LeBlanc, 2003) es una metodología para identificar combinaciones lógicas de las variables que mejor predicen la variable respuesta según un determinado modelo de regresión.

Más concretamente, este método explora modelos de regresión del tipo:

$$g\big(E(Y)\big) = \beta_0 + \sum_{j=1}^{t} \beta_j * \mathrm{I}_{\{L_j \; es \; cierta\}}$$

donde Y es la variable respuesta, g es una función adecuada y L_j es una expresión lógica de covariables[19].

El objetivo del método es encontrar el mejor modelo de regresión, es decir, determinar los L_j y las β_j que permiten una mayor capacidad predictiva. Para ello se utiliza un algoritmo de exploración que permite no tener que comprobar todas las posibles expresiones lógicas, puesto que el espacio de combinaciones lógicas posibles crece de forma exponencial a medida que aumenta el número de variables, y se hace computacionalmente inexplorable en su totalidad.

La regresión logística está implementada como un método iterativo en el que cada estado viene dado por un modelo de regresión, en el que los predictores son expresiones lógicas o también llamados árboles lógicos. La idea básica de este método es realizar, a cada iteración, un cambio en alguno de los árboles lógicos del modelo actual, de forma que la capacidad predictiva del nuevo modelo sea mejor que la del anterior. En el caso de modelos con varios árboles lógicos, a cada iteración se puede realizar un cambio en uno de los árboles, o un cambio en cada uno de ellos.

La estrategia que utiliza este método se basa en tres puntos: (a) una función para evaluar la capacidad de predicción, (b) la generación de los árboles lógicos a considerar, y (c) el algoritmo de búsqueda utilizado.

---

[18] (Gales, 2009)

[19] Variables continuas independientes que junto a una o más variables grupo de tratamiento sirven para explicar una variable respuesta continua.

(a) Para la función que tiene que evaluar la capacidad predictiva de cada modelo se proponen varias opciones según la naturaleza del modelo:

- En el caso de respuesta dicotómica y un único árbol lógico L, esta expresión lógica clasifica cada individuo en una de las clases si la condición se cumple, y en la otra si no. Para este caso, se utiliza la proporción de mal clasificados como medida de evaluación.

$$Y = \mathrm{I}_{\{L\ es\ cierta\}}$$

- En el caso de modelos de regresión con respuesta continua, el modelo se ajusta por el método de mínimos cuadrados, y la función que se utiliza para evaluar es la suma de cuadrados residuales.

$$Y = \beta_0 + \beta_1 * \mathrm{I}_{\{L_1\ es\ cierta\}} + \cdots + \beta_p * \mathrm{I}_{\{L_p\ es\ cierta\}} + \epsilon$$
$$con\ \epsilon \sim N(0, \sigma^2)$$

- Para modelos de regresión logística,

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 * \mathrm{I}_{\{L_1\ es\ cierta\}} + \cdots + \beta_p * \mathrm{I}_{\{L_p\ es\ cierta\}}$$

se utiliza la varianza.

(b) Dado un árbol concreto, se definen los árboles vecinos como el conjunto de todos los árboles generados a partir de realizar un sólo cambio en el árbol inicial. Estos cambios pueden ser cambiar, eliminar o agregar una variable, expandir o podar una rama, o intercambiar operadores lógicos.

(c) También existen varios tipos de implementación del algoritmo de búsqueda. Un primer tipo de búsqueda consiste en construir, a cada iteración, todos los árboles vecinos al actual por medio de los movimientos básicos definidos y seleccionar como nuevo árbol aquel que presente un valor menor para la función de evaluación. Este algoritmo continúa hasta que no se encuentra ningún árbol vecino que tenga un valor de la función de evaluación menor que el árbol considerado, en ese momento se detiene el proceso.

5.1.4 SUPPORT VECTOR MACHINE

Máquinas de Vectores de Soporte (SVMs) son un conjunto de métodos relacionados para aprendizaje supervisado, aplicables tanto a los problemas de clasificación y regresión. Desde la introducción del clasificador SVM hace una década (Vapnik & Cortes, 1995), SVM ganado popularidad debido a su sólida base teórica.

Supongamos, por el momento, que el conjunto de entrenamiento es separable por un hiperplano. Se ha demostrado (Vapnik & Cortes, 1995) que para la clase de hiperplanos, la complejidad del hiperplano puede ser limitada en términos de otra cantidad, el margen. El margen se define como la distancia mínima de una muestra para una superficie de decisión.

Por lo tanto, si acotamos al margen de una función de clasificación desde abajo, podemos controlar su complejidad.

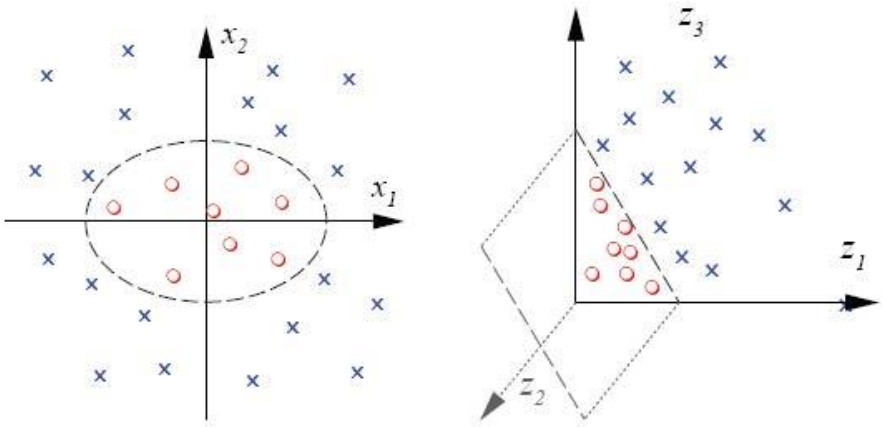


Gráfico 5.1.1: Transformación de espacios

Fuente: Elaboración del autor.

SVM implementa la idea que el riesgo de error se minimiza cuando se maximiza el margen. Una SVM elige un hiperplano de máximo margen que se encuentra en un espacio de entrada transformado (*Gráfico 5.1.1*) y divide las clases de la muestra. Los parámetros del hiperplano solución se derivan de un problema de optimización de programación cuadrática (*Ecuación 5.1*).

$$\underset{\{w,b\}}{\mathbf{Min}} \frac{\mathbf{1}}{\mathbf{2}} \|\boldsymbol{w}\|^{\mathbf{2}}$$

$$\boldsymbol{s.a.\, y_i} * (\boldsymbol{w} * \boldsymbol{\phi}(\boldsymbol{x_i}) + \boldsymbol{b}) \geq \mathbf{1}\ \forall \boldsymbol{i} = \mathbf{1}..\boldsymbol{n}$$

Ecuación 5.1: SVM

La elección de la transformación lineal que hace al espacio separable $\boldsymbol{\phi_w}(\,)$ parece ser de alta complejidad. Afortunadamente, Vapnik & Cortes (1995) proponen una técnica, denominada "*truco kernel*", que permite determinar la relación con el hiperplano de máximo.

Se entiende por *kernel* a una función $k(\cdot,\cdot)$ que permite mapear los datos desde su dominio hasta al espacio Imagen de la transformación lineal, como se puede ver en la Figura 5.1. Para más detalles del sustento teórico del método kernel, véase Vapnik & Cortes (1995).

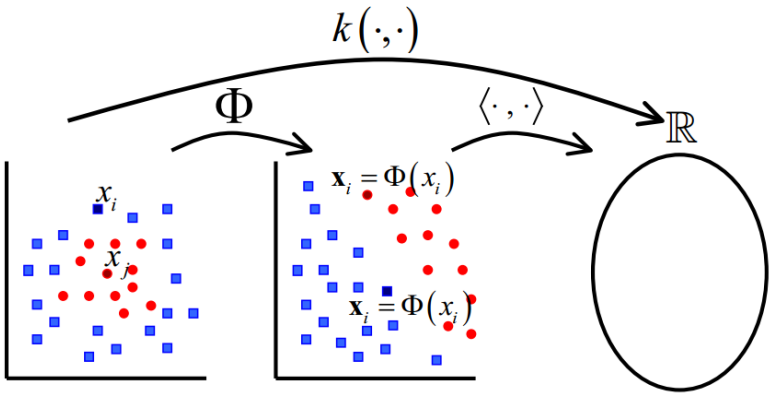


Figura 5.1: Relación entre espacios de dominio e imagen.

Fuente: Elaboración del autor.

La *Tabla 3.1.1* resume los tipos de kernel más comunes a usar:

Kernel	Fórmula
Lineal	$\boldsymbol{x^T y + c}$
Polinomial	$\boldsymbol{(\alpha x^T y + c)^d}$
Gaussiano	$\boldsymbol{\exp(-\frac{\|x-y\|^2}{2\sigma^2})}$
Anova	$\boldsymbol{\left(\sum_{k=1}^{n} \exp\left(-\sigma\left(x^k - y^k\right)^2\right)\right)^d}$
Multicuadratico	$\boldsymbol{\sqrt{\|x-y\|^2 + c^2}}$

Tabla 5.1.1: Kernel más comunes

Fuente: Elaboración del autor.

Sin embargo, el supuesto que la muestra es estrictamente separable por un hiperplano es muy fuerte dado que, por ejemplo, un alto nivel de ruido podría provocar cierta superposición de las clases. El uso de SVM, en la forma anterior, no podría minimizar el riesgo empírico. Para subsanar este problema, se generaron variaciones al modelo original que busca integrar esta posible superposición. A continuación se listan algunas de ellas:

- Soft Margin Support Vector Classifiers

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$$
$$y_i\left(\left(w * \phi(x_i)\right) + b\right) \geq 1 - \xi_i \; \forall i = 1,..,n$$
$$\xi_i \geq 0 \; \forall i = 1,..,n$$

Ecuación 5.2

- Support Vector Regression

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} |y_i - f(x)|_\varepsilon$$

Ecuación 5.3

- SVM-like Models

$$\min_{w,b,e} \frac{1}{2}\|w\|^2 + \gamma\frac{1}{2}\sum_{i=1}^{n} e_i^2$$
$$y_i\left(\left(w * \phi(x_i)\right) + b\right) = 1 - e_i \;\; \forall i = 1,..,n$$

Ecuación 5.4

5.1.5 RANDOM FOREST

Breiman (2001) propuso este modelo de data mining, el cual que consiste en la construcción de árboles de decisión, utilizando distintos muestreos tanto de los datos como de la cantidad de variables, para luego predecir la variable dependiente en base a los "votos"[20] de cada uno de los árboles.

Cada árbol se construye de la siguiente forma[21]:

- Siendo N el número de datos, se generan varios conjuntos de entrenamiento a través de muestras aleatorias, con reemplazo, de $n \ll N$ datos. En cada conjunto se construye un árbol de decisión.
- Si M es el número de variables independientes, un número $m \ll M$ es especificado para que, en cada nodo, m variables sean seleccionadas al azar y se escoja el mejor split del nodo en base a ellas. Este número se mantiene durante la construcción de los árboles de decisión. De esta forma se incluye una mayor variabilidad de árboles y se reduce la dependencia del resultado con las divisiones precedentes.
- Cada árbol es construido lo más extenso posible sin incurrir en poda.
- En cada árbol, los $(N - n)$ datos restantes se utilizan para estimar el error de predicción. Estos datos se conocen como Out-of-Bag (OOB).

[20] Para un dato, cada árbol de decisión genera una predicción de categoría. La categoría con mayor frecuencia será la predicción final de ese dato.

[21] (Breiman, Classification/Clustering, 2014)

Las principales ventajas de esta metodología son:

- Proporciona una buena capacidad predictiva incluso cuando hay más variables que observaciones y cuando la mayoría de las variables son ruido.
- Proporciona un ranking de importancia de las variables.
- No sobre ajusta los datos
- Contempla el uso de pocos parámetros: el número de variables en el subconjunto aleatorio en cada nodo y el número de árboles en el bosque[22].

5.1.6 VALORES PERDIDOS

El problema de la falta de valores de atributos es tan importante para la minería de datos como lo es para el razonamiento estadístico. En ambas disciplinas hay métodos para hacer frente a los valores de atributos que faltan. En general, los métodos para manejar la falta de valores de atributos pertenecen tanto métodos secuenciales (llamados también métodos de pre-procesamiento) como métodos ortogonales (métodos en los que se tienen en cuenta los valores de atributos desaparecidos durante el proceso principal de la adquisición de conocimientos).

Los métodos secuenciales incluyen técnicas basadas en la supresión de los casos con valores de atributo que faltan, la sustitución de un valor de atributo faltante por el valor más común de ese atributo, la asignación de todos los valores posibles para el valor del atributo que falta, la sustitución de un valor de atributo que falta por la media para los atributos numéricos, asignando a un valor de atributo que falta el valor correspondiente al caso más parecido, o la sustitución de un valor de atributo que falta por un nuevo valor, calculado a partir de un nuevo conjunto de datos, teniendo en cuenta el atributo original como una decisión.

El segundo grupo de métodos para manejar la falta de valores de atributos, en el que se tienen en cuenta los valores de atributos desaparecidos durante el proceso principal de la adquisición de conocimiento está representado, por ejemplo, mediante una modificación de la LEM2, inducción de reglas algoritmo, en el que las reglas de forma indujeron el conjunto de datos original, con la falta de valores de los atributos que se consideran "no importantes".

Los valores perdidos se clasifican en 3 tipos de acuerdo a su naturaleza:

1. **Missing Completely at Random (MCAR):** Son aquellos datos perdidos donde su patrón de comportamiento no se relaciona tanto con la variable dependiente como con las variables independientes.

[22] Se denomina “bosque” al conjunto completo de árboles de decisión construidos.

2. **Missing at Random (MAR):** Son aquellos datos perdidos donde su patrón de comportamiento no se relaciona con la variable dependiente, pero puede permitir relaciones cruzadas con las otras variables independientes.

3. **Not Missing at Random (NMAR):** Son aquellos datos perdidos donde su patrón de comportamiento se relaciona con la variable dependiente, es decir, poseen algún grado de explicación en la respuesta de esa variable.

5.1.6.1 Tratamientos de Valores perdidos

En su trabajo, Tsikriktsis (2005), desarrolla una lista de las técnicas de tratamiento de valores perdidos más comunes en la literatura, describiendo, además, las ventajas y desventajas de su uso, lo que se puede ver en la Tabla 5.1.2:

Tipo	Técnicas	Descripción	Ventajas	Desventajas
Basadas en eliminación	Eliminación Listwise	Elimina la mayor proporción de los casos con valores perdidos	Simple de implementar	Sacrifica un gran tamaño de la información que puede impactar negativamente en el poder estadístico4
	Eliminación Pairwise	Elimina sólo los casos de los análisis estadísticos que requieres información. Por ejemplo, datos altamente correlacionados.	Conserva más datos que la técnica anterior y es más preciso.	Las correlaciones y covarianzas pueden estar sesgadas
Basados en reemplazo	Sustitución por la media	Los valores perdidos son reemplazados por el valor promedio de sus variables	Conserva la información y es simple de implementar	Impacta negativamente en la estimación de varianza y los grados de libertad
	Sustitución por la media total	Similar al anterior, con la salvedad que el promedio es calculado sin los valores ausentes (los cuales valen cero)	Conserva la información y es simple de implementar	Subestima la varianza y covarianza
	Sustitución por promedio	Los valores perdidos son	Entrega mejores	Subestima la varianza y la definición de

Tipo	Técnicas	Descripción	Ventajas	Desventajas
	de subgrupos	reemplazados por el promedio del grupo al cual pertenecen	estimaciones comparada con las técnicas anteriores	subgrupos es arbitraria
	Imputación por regresión	Estima la relación entre las variables y usa los coeficientes de una regresión para calcular los valores perdidos.	Estima la información conservando la desviación de la media y la forma de la distribución	Distorsiona el número de grados de libertad y puede artificialmente aumentar la correlación entre variables
	Imputación Hot-deck	Reemplaza los valores perdidos por el valor actual de un caso similar de la base de datos	Los valores perdidos son reemplazados con valores reales que permiten mantenerse cerca de la distribución original de los datos	Existen problemas de aplicación cuando no existen casos suficientemente similares con los que reemplazar.
Basados en modelos	Máxima Verosimilitud	Se estiman los parámetros de una distribución para la información disponible y se estiman los valores perdidos en base a estos.	Incrementa Accuracy[23] si el modelo es correcto	La distribución supuesta requerida por esta técnica es relativamente estricta.
	Maximización esperada	Similar al anterior, con la salvedad que calcula los parámetros mediante	Incrementa Accuracy[17] si el modelo es correcto	El algoritmo toma tiempo en converge y es más complejo de implementar.

[23] Se define en el punto 7.1.

Tipo	Técnicas	Descripción	Ventajas	Desventajas
		iteraciones en vez de maximización.		

Fuente: Tsikriktsis (2005). Traducción del autor.

Tabla 5.1.2: Técnicas de tratamiento de valores perdidos más comunes

5.2 Series De Tiempo[24]

5.2.1 Definición

Se llama Series de Tiempo a un conjunto de observaciones sobre valores que toma una variable (cuantitativa) en diferentes momentos del tiempo. Los datos se pueden comportar de diferentes formas a través del tiempo, puede que se presente una tendencia, un ciclo; no tener una forma definida o aleatoria, variaciones estacionales (anual, semestral, etc…). Las observaciones de una serie de tiempo serán denotadas por $Y_1, Y_2, \dots, Y_T$, donde Y_t es el valor tomado por el proceso en el instante t.

Los modelos de series de tiempo tienen un enfoque netamente predictivo y en ellos los pronósticos se elaborarán sólo con base al comportamiento pasado de la variable de interés. Se puede distinguir dos tipos de modelos de series de tiempo:

- **Modelos deterministas:** se trata de métodos de extrapolación sencillos en los que no se hace referencia a las fuentes o naturaleza de la aleatoriedad subyacente en la serie. Su simplicidad relativa generalmente va acompañada de menor precisión. Ejemplo de modelos deterministas son los modelos de promedio móvil en los que se calcula el pronóstico de la variable a partir de un promedio de los "n" valores inmediatamente anteriores.
- **Modelos estocásticos:** se basan en la descripción simplificada del proceso aleatorio subyacente en la serie. En términos sencillos, se asume que la serie observada$Y_1, Y_2, \dots, Y_T$ se extrae de un grupo de variables aleatorias con una cierta distribución conjunta difícil de determinar, por lo que se construyen modelos aproximados que sean útiles para la generación de pronósticos.

Una serie $\{Y_t\}_{t=1}^{T}$ se puede clasificar en estacionaria o no estacionaria:

- **Serie no estacionaria:** es aquella cuyas características de media, varianza y covarianza cambian a través del tiempo lo que dificulta su modelamiento. Sin embargo, en muchas ocasiones, si dicha serie es diferenciada[25] una o más veces la serie resultante ser estacionaria (procesos no estacionarios homogéneos).

[24]Apuntes de Tópicos de Minería de Datos. FCFM, Universidad de Chile (Hurtado & Ríos, 2008).

[25] La serie diferenciada *n* veces de $\{Y_t\}_{t=1}^{T}$ es la serie $\{\hat{Y}_t\}_{t=1}^{T}$, donde $\hat{Y}_t = Y_t - Y_{t-n}$

- **Serie estacionaria:** es aquella cuya media y varianza no cambian a través del tiempo y cuya covarianza sólo es función del rezago. Gracias a estas características se puede modelar el proceso subyacente a través de una ecuación con coeficientes fijos, estimados a partir de los datos pasados. Matemáticamente, presenta las siguientes características:

  - Media: $E(Y_t) = E(Y_{t+m})\ \forall t \in \aleph, m \in \aleph\ tq\ t + m \leq T$
  - Varianza: $\sigma(Y_t) = \sigma(Y_{t+m})\ \ \forall t \in \aleph, m \in \aleph\ tq\ t + m \leq T$
  - Covarianza: $cov(Y_t, Y_{t+k}) = cov(Y_{t+m}, Y_{t+m+k})\ \ \forall t \in \aleph, m \in \aleph, k \in \aleph\ \ tq\ t + m \leq T$

  Para determinar la estacionalidad de una serie se usan distintos test estadísticos, siendo, uno de los más conocidos, la prueba de Dickey-Fuller aumentada (ADF). En este test, al rechazar la hipótesis nula, permite determinar la presencia de estacionalidad[26].

5.2.2 METODOLOGÍA DE BOX-JENKINS

El enfoque de Box-Jenkins es una de las metodologías de uso más amplio para el modelamiento estocástico de series de tiempo. Es popular debido a su generalidad, ya que puede manejar cualquier serie, estacionaria o no estacionaria, y por haber sido implementad o en numerosos programas computacionales.

Los pasos básicos de la metodología de Box-Jenkins son:

1. Verificar la estacionalidad de la serie. Si ésta no es estacionaria, diferenciarla hasta alcanzar estacionalidad.
2. Identificar un modelo tentativo.
3. Estimar el modelo.
4. Verificar el diagnóstico (si este no es adecuado, volver al paso 2).
5. Usar el modelo para pronosticar.

Lo que se trata es de identificar el proceso estocástico que ha generado los datos, estimar los parámetros que caracterizan dicho proceso, verificar que se cumplan las hipótesis que han permitido la estimación de dichos parámetros. Si dichos supuestos no se cumplieran, la fase de verificación sirve como retroalimentación para una nueva fase de identificación.

[26] La explicación de este test estadístico contempla conocimientos más profundos de la teoría de series de tiempo. Aquí, sólo se ha expuesto los elementos prácticos que servirán para el trabajo de esta memoria. Para conocer más al respecto, el lector puede consultar Cheung & Lai (1995).

Cuando se satisfagan las condiciones de partida, se puede utilizar el modelo para pronosticar.

Box y Jenkins han desarrollado modelos estadísticos que tienen en cuenta la dependencia existente entre los datos. Cada observación en un momento dado es modelada en función de los valores anteriores. Algunas funciones de modelamiento son los modelos: AR (Autoregresive), MA (Moving Average), ARMA (Autorregresive Moving Average) y ARIMA (Autorregresive Integrate Moving Average).

5.2.3 PROCESOS ESTOCÁSTICOS

5.2.3.1 Modelos de Media Movil, $MA(q)$

$$Y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{q-1}$$

En los modelos de media móvil, el proceso se representa como una suma ponderada de errores actuales y anteriores. El número de rezagos del error considerados (q) determina el orden del modelo de media móvil.

5.2.3.2 Modelos Autorregresivos, $AR(p)$

$$Y_t = \delta + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t$$

En los modelos autorregresivos, el proceso se representa como una suma ponderada de observaciones pasadas de la variable. El número de rezagos (p) determina el orden del modelo autorregresivo.

5.2.3.3 Modelos Mixtos Autorregresivos - Media Móvil, $ARMA(p, q)$

$$Y_t = \delta + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{q-1}$$

En estos modelos, el proceso se representa en función de observaciones pasadas de la variable y de los valores actuales y rezagados del error. El número de rezagos de la variable de interés (p) y el número de rezagos del error (q) determinan el orden del modelo mixto.

5.2.3.4 Mo delos Autorregresivos Integrados de Promedio Móvil, $ARIMA(p, d, q)$

Muchas series de tiempo no son estacionarias, por ejemplo el Producto Nacional Bruto o la Producción Industrial. Un tipo especial de series no estacionarias, son las no estacionarias homogéneas que se caracterizan porque, al ser diferenciadas una o más veces, se vuelven estacionarias.

La serie Y_t será no estacionaria homogénea de orden d si $W_t = \Delta^d Y_t$ es estacionaria, donde:

- $\Delta Y_t = Y_t - Y_{t-1}$
- $\Delta^{n+1} Y_t = \Delta^n Y_t - \Delta^n Y_{t-1}$

Si después de haber diferenciado la serie Y_t se consigue una serie estacionaria W_t, y dicha serie obedece a un proceso $ARMA(p,q)$, se dice que Y_t responde a un proceso $ARIMA(p,d,q)$:

$$W_t = \delta + \phi_1 W_{t-1} + \cdots + \phi_p W_{t-p} + \varepsilon_t + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{q-1}$$

5.2.4 AUTOCORRELACIÓN

Si se pretende establecer un modelo para una serie estacionaria, un paso usual, luego de eliminar componentes estacionales y tendencias, es estudiar la correlación entre una observación de la serie y las observaciones previas. La presencia de correlaciones altas entre observaciones de la serie (autocorrelaciones) puede ser consecuencia de un comportamiento lineal del fenómeno a través del tiempo y da una idea del tipo de modelo apropiado.

5.2.4.1 Definición

Dado un proceso estocástico $\{X_t\}_{t=1}^{T}$, se define la función de Autocovarianza como la función que relaciona los valores en diferentes instantes:

$$\gamma(s,t) = cov(X_s, X_t) = E[(X_s - E(X_s))(X_t - E(X_t))]$$

Un proceso estocástico se dirá *estrictamente estacionario* cuando la distribución conjunta es invariante ante traslaciones, o sea la función de distribución conjunta de cualquier subconjunto de variables es invariante respecto a un desplazamiento en el tiempo.

Un proceso estocástico $\{X_t\}_{t=1}^{T}$ se dirá estacionario si, para todo $s, t, r \in \aleph$ tal que $s + r \leq T \wedge t + r \leq T$, se cumplen las siguientes condiciones:

- $\gamma(s,t) = \gamma(s+r, t+r)$
- $E(X_t) = C$
- $E(X_t^2) < \infty$

Cuando el proceso $\{X_t\}_{t=1}^{T}$ es estacionario, es común definir la autocovarianza en función de un desplazamiento h sobre sí misma:

$$\gamma(h) \triangleq \gamma(t,h) = cov(X_t, X_{t+h})$$

5.2.5 CORRELOGRAMAS

Una forma visual de estudiar las autocorrelaciones es a través de la función de correlación, el cual muestra la correlación entre observaciones separadas por k intervalos de tiempo o rezagos. El proceso para calcular la autocorrelación particiona las observaciones de la serie en dos grupos, $\{Y_1, Y_2, \ldots, Y_{t-q}\}$ y $\{Y_{1+q}, Y_{2+q}, \ldots, Y_t\}$, para luego estimar la correlación entre los dos conjuntos. El grafo generado por esta función se denomina correlograma.

Para que un rezago sea significativo distintos de cero, se define un nivel de confianza α y se determina el intervalo un confianza para esta hipótesis, que viene a estar dada por $[-\frac{z_{1-\alpha}}{\sqrt{n}}, \frac{z_{1-\alpha}}{\sqrt{n}}]$, donde $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$ y Φ es la función inversa de la distribución normal estándar. En el Gráfico 5.2.1, Gráfico 5.2.2 y Gráfico 5.2.3 se presenta el intervalo de confianza como una línea azul segmentada.

Existen 3 tipos de funciones para generar correlogramas:

- **Autocorrelation Function (ACF):** Función de la autocorrelación $\rho(\cdot)$. Muestra la asociación entre valores de la misma variable en diferentes periodos de tiempo k (no aleatoria). La altura de las líneas en el correlogramas representa la correlación entre las observaciones que están separadas por la cantidad de unidades de tiempo que aparecen en el eje horizontal. La correlación para el primer rezago siempre es uno por lo que no deben tomarse en cuenta en las interpretaciones. Se utiliza para determinar el proceso MA. La autocorrelación corresponde a la autocovarianza normalizada por la varianza de $\{X_t\}_{t=1}^{T}$:

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)}$$

Ecuación 5.5: Fórmula de ACF

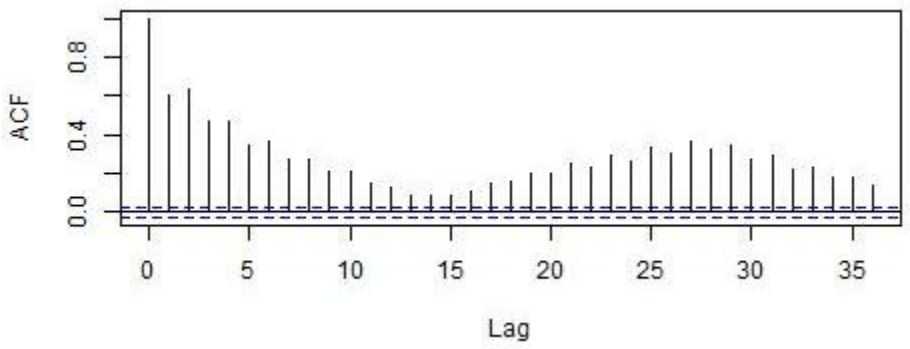


Gráfico 5.2.1: Ejemplo de gráfico ACF

Fuente: Elaboración del autor.

- **Partial Autocorrelation Function (PACF):** Función de la autocorrelación parcial $\phi_{kk}(\cdot)$. Identifica la relación entre los valores actuales y los k valores anteriores de la serie cronológica original, después de quitar los efectos de las autocorrelaciones de orden inferior. Se utiliza para determinar el proceso AR. Se puede calcular recursiva de la siguiente forma:

$$\phi_{kk}(k) = \frac{Cov((x_t - \hat{x}_t), (x_{t-k} - \hat{x}_{t-k}))}{\sqrt{Var(x_t - \hat{x}_t) Var(x_{t-k} - \hat{x}_{t-k})}}$$

Ecuación 5.6: Fórmula de PACF

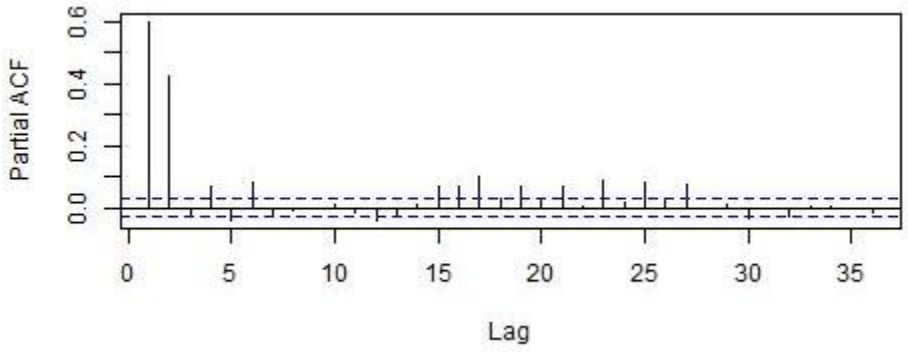


Gráfico 5.2.2: Ejemplo de función PACF

Fuente: Elaboración del autor.

- **Cross Correlation function (CCF):** Función de la correlación cruzada $\rho_{XY}(\cdot)$. Mide no solamente la fortaleza de la relación entre dos series de tiempo, $\{X_t\}_{t=1}^{T}$ e $\{Y_t\}_{t=1}^{T}$, sino también su dirección en que dependen una de otra. Esta última propiedad es útil para identificar variables causales. Por esta razón, es importante examinar la CCF tanto para los valores positivos de k como para los negativos. Para valores negativos de k, la CCF describe la influencia lineal de los valores pasados de $\{Y_t\}_{t=1}^{T}$ sobre $\{X_t\}_{t=1}^{T}$. Para valores positivos de k, la CCF indica la influencia lineal de los valores pasados de $\{X_t\}_{t=1}^{T}$ sobre $\{Y_t\}_{t=1}^{T}$.

$$\rho_{XY}(k) = \frac{E[(X_t - E(X_t))(Y_{t-k} - E(Y_t))]}{\sqrt{VAR(X_t)}\sqrt{VAR(Y_t)}}$$

Ecuación 5.7: Fórmula de CCF

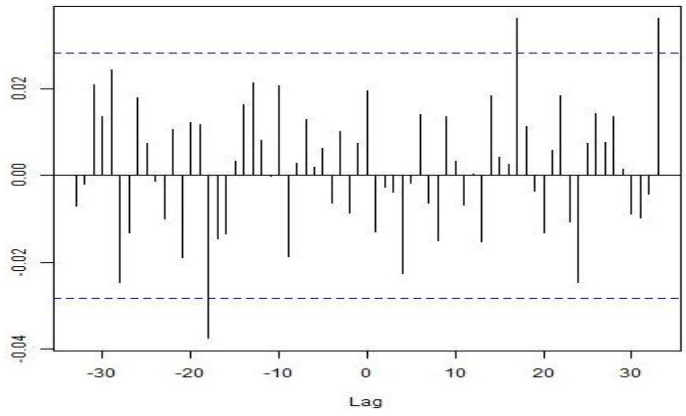


Gráfico 5.2.3: Ejemplo de función CCF

Fuente: Elaboración del autor.

Sin embargo, una dificultad para trabaja con la CCF es que ésta se ve afectada por la estructura de series de tiempo de la variable $\{X_t\}_{t=1}^{T}$ y cualquier tendencias comunes entre las series $\{X_t\}_{t=1}^{T}$ e $\{Y_t\}_{t=1}^{T}$ pueden tener con el tiempo.

Una de las estrategias para hacer frente a esta dificultad se llama Pre-blanqueamiento, dado que busca reducir las series de tiempo a ruido blanco[27], de forma tal de eliminar cualquier efecto de tendencia dentro de las series. La metodología se estructura de la siguiente manera:

Sean $X = \{X_t\}_{t=1}^{T}$ e $Y = \{Y_t\}_{t=1}^{T}$

1. Determinar un modelo $ARMA$ para X y almacenar los residuos de este modelo (R_X).
2. Se filtra la serie Y. Esto quiere decir, que se describe la serie Y utilizando el modelo de la serie X (usando los coeficientes estimados del paso 1). En este paso se almacena las diferencias entre los valores de Y con su serie filtrada $(fil(Y) - Y)$.
3. Se analiza el CCF entre R_X y $(fil(Y) - Y)$. Este CCF se puede utilizar para identificar las los rezagos significativos para el modelamiento.

5.2.5.1 Criterios

Para la interpretación de los correlogramas y la decisión de qué rezagos ingresar al estudio, la literatura sugiere los siguientes criterios.

- Si ninguna de las autocorrelaciones es significativamente diferente de cero, la serie es esencialmente ruido blanco.
- Si las autocorrelaciones decrecen linealmente, pasando por el cero, o muestra un patrón cíclico, pasando por cero varias veces, la serie no es estacionaria. Se tendrá que diferenciarla una o más veces antes de modelarla.
- Si las autocorrelaciones muestran estacionalidad, o se tiene un a alza cada periodo (cada 12 meses, por ejemplo), la serie no es estacionaria y hay que diferenciarla con un salto igual al periodo.
- Si las auto correlaciones decrecen exponencialmente hacia cero y las autocorrelaciones parciales son significativamente no nulas sobre un pequeño número de rezagos, se puede usar un modelo autoregresivo.
- Si las autocorrelaciones parciales decrecen exponencialmente hacia cero y las autocorrelaciones son significativamente no nulas sobre un pequeño número de rezagos, se puede usar un modelo de me días móviles.
- Si las autocorrelaciones simples y parciales de crecen lentamente hacia cero, pero sin alcanzar el cero, se puede usar un modelo autoregresivo combinado con medias móviles.

[27] Se entiende por ruido blanco a un proceso estocástico que se caracteriza por el hecho de que sus valores de señal en dos tiempos diferentes no guardan correlación estadística.

6 ESTUDIO DEL PROCESO Y VARIABLES IMPLICADAS

A continuación se expone la forma en que se preparan las variables para su posterior modelamiento. Se identifica la unidad temporal de estudio, las variables incompletas, la forma de tratar los valores perdidos, la selección de variables y su relación temporal. El trasfondo se sustenta en las etapas 3 y 4 de la metodología KDD: Preprocesamiento y limpieza y transformación de datos.

Para el análisis se debe tener en consideración el objetivo buscado en este apartado, donde se pretende lograr desarrollar 4 modelos predictivos. Cada uno de ellos contempla las mismas variables independientes pero cambia su variable dependiente, siendo estas los contaminantes de las corrientes de salida del proceso: Perclorato de potasio ($KClO_4$) y el Cloruro de Sodio ($NaCl$). Ambos contaminantes se miden en su porcentaje peso peso ($\%p/p$), esto es, la masa de soluto en masa de solución. A continuación se muestran las relaciones buscadas:

$$\mathbf{KClO_4\_L} = \boldsymbol{f_1}(\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n})$$

$$\mathbf{NaCl\_L} = \boldsymbol{f_2}(\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n})$$

$$\mathbf{KClO_4\_M} = \boldsymbol{f_3}(\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n})$$

$$\mathbf{NaCl\_M} = \boldsymbol{f_4}(\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n})$$

Ecuación 6.1: Modelos relacionales

Donde:

- **$\mathbf{KClO_4}$_L** es el $KClO_4$ en la corriente L medida en %p/p.
- **NaCl_L** es el $NaCl$ en la corriente L medida en %p/p.
- **$\mathbf{KClO_4}$_M** es el $KClO_4$ en la corriente M medida en %p/p.
- **NaCl_M** es el $NaCl$ en la corriente M medida en %p/p.
- $\boldsymbol{x_i}$ son las variables significativas para el proceso que se pretenden determinar, con $i = 1..n$.
- $\boldsymbol{f_j}(\,)$ es la relación funcional que se estudiará en el punto 7, con $j = 1..4$.

En adelante, cuando se haga referencia a estas relaciones, se nombrarán como "las ecuaciones", o se hará referencia al número de Ecuación 6.1.

6.1 Determinación de la Unidad de Tiempo

Uno de los problemas presentes en la información disponible es la multiplicidad de las horas en que se realizan toma de datos de cada una de las 310 variables, es decir, su frecuencia de medición. Existen 7 categorías de este tipo:

1 Turno : La toma de datos se da a las 8:00 y 20:00 hrs. de cada día.

2 Día : La toma de datos se da una vez al día, puede ser a las 8:00, 12:00

			o 20:00 hrs.
3	**Irregular**	:	Hace referencia a un patrón de datos sin un comportamiento explicable, en otras palabras, la cantidad de datos ausentes y la tomas de datos se dan en horarios que varían en el tiempo y no es posible determinar una categoría única.
4	**Cada 3 horas**	:	La toma de datos se da desde las 0:00, luego a las 3:00, a las 6:00 y así sucesivamente.
5	**Cada 12 horas**	:	La toma de datos se da a las 0:00 y a las 12:00 hrs de cada día.
6	**Cada 2 horas**	:	La toma de datos se da desde las 0:00, luego a las 2:00, a las 4:00 y así sucesivamente.
7	**Hora**	:	La toma de datos se da cada hora el día.

La *Tabla 6.1.1* resume la cantidad de variables que se agrupan en las distintas categorías dentro de cada etapa del proceso.

FRECUENCIA	PTS	MURIATO	DUAL	CRISTAL	SUBTOTAL
Turno	40	33	7	17	97
Día	5	6	2	0	13
Irregular	56	10	7	47	120
Cada 3 horas	0	0	1	1	2
Cada 12 horas	0	0	0	1	1
Cada 2 horas	0	0	0	2	2
Hora	6	35	7	27	75
SUBTOTAL	107	84	24	95	310

Tabla 6.1.1: Número de variables por unidad de frecuencia

Fuente: Elaboración del autor.

Como se puede apreciar en la Tabla 6.1.1, las frecuencias que concentran más variables son *Hora* y *Turno.* Cabe destacar que, para la decisión de unidad de tiempo, no se consideran variables con frecuencia irregular.

Por la naturaleza del proceso se requiere un nivel desagregado de información por Hora, sin embargo, sólo 75 variable presentan esta estructura. Si se escoge esa unidad de tiempo, se ve necesario modelar cada una de las 115 variables restantes[28] para estimar sus valores por hora, lo que introduce más ruido desde los errores de estimación de estos modelos. Si se escoge *turno* como unidad de tiempo, se ve necesario disminuir la frecuencia de 80 variables (pasar de hora a turno por ejemplo), lo que genera pérdida de información a

[28] Se recuerda la exclusión de 120 variables con frecuencia irregular para la decisión de unidad de tiempo. Una vez elegida la temporalidad, se procede a transformarlas.

considerar datos agregados. Además, es necesario modelar 13 variables de frecuencia *día* para estimar sus valores por turno.

Se decide como unidad de tiempo el *turno*, debido a que, para la implementación futura de los modelos de data mining, una modelación previa de las variables genera un mayor sesgo de la predicción que la agregación de datos (que genera pérdida de información). Así, de las 310, se transforman 13 variables de frecuencia *día*, 80 variables de frecuencia tipo 4, 5, 6 y 7, y 120 variables de frecuencia *irregular*. Las 97 de frecuencia *turno* se conservan sin transformación,

Para las variables de frecuencia *día*, se decide repetir el valor diario en las 8:00 y 20:00 hrs. para generar la frecuencia *turno*. Para las variables con frecuencia tipo 4, 5, 6 y 7 y las variables con frecuencia *irregular*, se considera el valor de las 8:00 como el promedio simple de las horas entre 21:00 (día anterior) y 8:00, mientras que para el valor de las 20:00 se asigna el promedio simple de las horas entre 9:00 y 20:00 del mismo día[29].

6.2 Reducción de Variables por Completitud de los Datos

Una vez transformada las frecuencias de las variables, se procede a realizar los estudios de la completitud de la información, es decir, cantidad de datos presentes versus datos faltantes.

Como se puede apreciar en el *Gráfico 6.2.1*, casi la mitad de las variables tienen más de un 90% de valores perdidos, lo que justifica su eliminación por no entregar la suficiente información. Según Malhotra (1987), ya a un nivel superior al 30% se requiere métodos más complejos para estimar los valores perdidos. Por esta razón, se conservan en la base de datos sólo las variables que estén bajo ese porcentaje, dejando un total de 68.

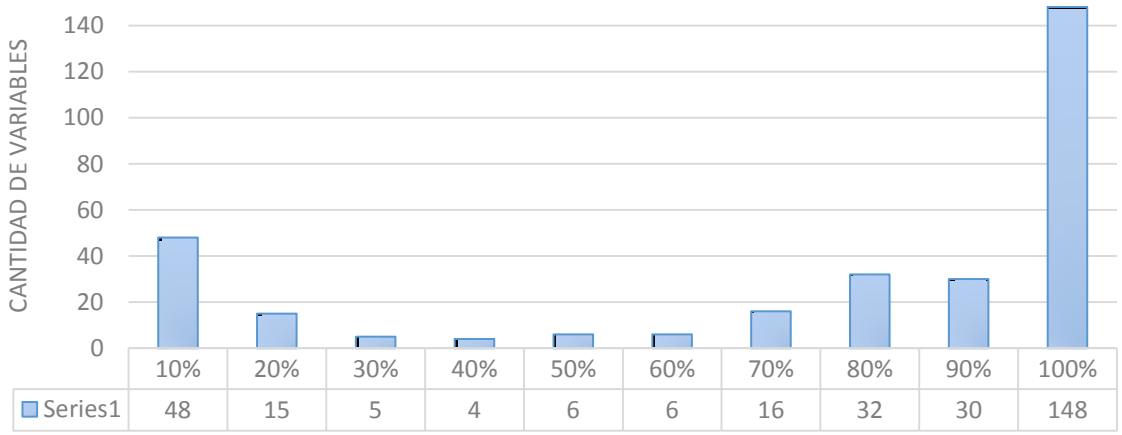


Gráfico 6.2.1: Distribución de valores perdidos

[29] Considerar que el cálculo del promedio simple excluye los valores ausentes, es decir, si de 8 datos, 3 son ausentes, el promedio simple se calcula en base a los 5 datos existentes.

Fuente: Elaboración del autor.

6.3 Tratamiento de los Valores Perdidos

Reducida las variables a 68, de tratan los valores perdidos. La cantidad de éstos limitan el modelamiento de las variables independientes con las dependientes. Para poder subsanar ese inconveniente, Nikos Tsikriktsis (2005) desarrolla en su trabajo una clasificación de herramientas a utilizar según cantidad y patrón que describen los valores perdidos (véase Tabla 6.3.1).

Suggested missing data techniques according to amount and pattern of missing data

Amount of missing data	Pattern		
	Missing completely at random	Missing at random	Non-missing at random
Less that 10%	1) Pairwise 2) Regression or hot-deck	1) Hot-deck 2) ML 3) Regression	1) ML 2) Hot-deck or regression
More than 10%	1) Pairwise 2) Regression or hot-deck	1) Hot-deck 2) ML	1) ML

Notes: (a) the preferred order of the missing data techniques is denoted by the number in front of each technique; (b) the above table is based on original work by Roth (1994).

Tabla 6.3.1: Tratamiento de datos perdidos[30]

Fuente: Tsikriktsis (2005).

Para llevar a cabo la completación, se ve necesario determinar el patrón que siguen los valores perdidos. Como se puede ver en el punto Valores Perdidos 5.1.6, existen 3 patrones: MCAR (Missing Completely at Random), MAR (Missing at Random) y NMAR (Not Missing at Random). Chisholm (2013), en su libro, desarrolla una serie de pasos para clasificar en estas categorías:

1. Se construye una matriz indicadora $\boldsymbol{N}$: sea $\boldsymbol{V} = \{\boldsymbol{V_j}\}_{\boldsymbol{j \in J}}$ la base de datos a trabajar, con $\boldsymbol{J}$ el total de variables. Para la variable $\boldsymbol{j}$, $\boldsymbol{V_j} = \{\boldsymbol{v_i}\}_{\boldsymbol{i \in I}}$, se define su indicador como $\mathbb{I}_{\boldsymbol{j}} = \{\boldsymbol{x_i}\}_{\boldsymbol{i \in I}}$, con I el total de datos, donde $\boldsymbol{x_i}$ toma valor 1 si la coordenada i tiene un valor válido, o si es un valor perdido. La matriz indicadora queda definida como $\boldsymbol{N} = \{\mathbb{I}_{\boldsymbol{j}}\}_{\boldsymbol{j \in J}}$.
2. Se calculan los coeficientes de correlación entre la matriz $\boldsymbol{V}$ y $\boldsymbol{N}$
3. Se analiza estos coeficientes y la distribución cruzada de las variables para concluir el patrón que siguen los valores perdidos.

La metodología aplicada mostró un bajo índice de correlación entre variables independientes con las dependientes, como se puede ver en la *Tabla 6.3.3* , lo que permite suponer que el patrón de valores perdidos no se relaciona con los resultados de las variables dependientes, por lo que se puede descartar la categoría NMAR.

[30] La descripción de las técnicas enunciadas están detalladas en el punto 5.1.6.1, donde ML significa Máxima Verosimilitud

	Sales Agregadas Pampa Blanca AL	*T. Reactor 2*	*KCl Agregado Minsal Bajo Cal.*	*Soluci⍰ Alimentaci⍰ Planta*	*T− Entrada Trona 1*	*T− Entrada Espesador Dual*	*T− Entada Trona 6*	*T− Entrada Trona 3*	*T− Salida Trona 5*	*T− Salida Trona 8*
Indicador Sales Agregadas Pampa Blanca AL	0,00	-0,37	0,32	0,03	0,09	-0,22	-0,21	-0,212	-0,213	-0,214
Indicador T. Reactor 2	-0,19	0,00	0,03	0,05	-0,02	-0,03	-0,03	-0,03	-0,03	-0,03
Indicador KCl Agregado Minsal Bajo Cal.	0,42	0,10	0,00	-0,06	0,23	0,04	0,05	0,05	0,05	0,05
Indicador Soluci⍰ Alimentaci⍰ Planta	0,35	0,09	0,08	0,00	-0,02	0,00	0,01	0,00	0,00	0,00
Indicador T− Entrada Trona 1	0,36	0,02	0,08	-0,30	0,00	-0,08	-0,08	-0,08	-0,08	-0,08
Indicador T− Entrada Espesador Dual	0,33	0,09	0,15	0,04	0,02	0,00	-0,01	-0,01	-0,01	-0,01
Indicador T− Entada Trona 6	0,33	0,09	0,14	0,04	0,02	0,00	0,00	0,00	0,00	0,00
Indicador T− Entrada Trona 3	0,33	0,09	0,15	0,04	0,02	0,00	0,00	0,00	0,00	0,00
Indicador T− Salida Trona 5	0,33	0,09	0,15	0,04	0,02	0,00	0,00	0,00	0,00	0,00
Indicador T− Salida Trona 8	0,33	0,09	0,15	0,04	0,02	0,00	0,00	0,00	0,00	0,00

Tabla 6.3.2: Subconjunto de variables analizadas para el paso 2

Fuente: Elaboración del autor.

Como se puede ver en la *Tabla 6.3.3*, el 90% de los pares de variables tienen coeficiente de correlación dentro del rango [0, 0,4], lo que permite suponer una distribución MCAR de los patrones de valores perdidos. Dado esto, junto a que el nivel de incompletitud es mayor al 10%, se ve necesario una completación por las técnicas Pairwise, Regresión o Hot-deck[31].

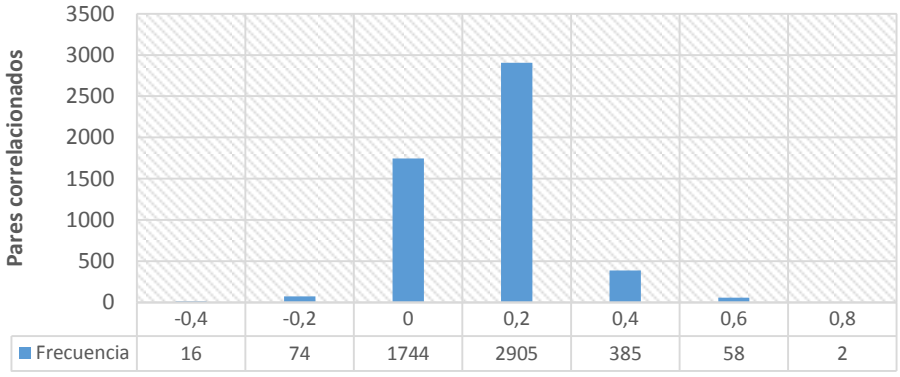


Tabla 6.3.3: Distribución de pares correlacionados

Fuente: Elaboración del autor.

Se escoge la técnica de regresión por ser simple de implementar[32] y no afectar tanto el promedio y desviación estándar de las variables como las técnicas Pairwise o Hot-Deck. Para la implementación, se estiman los valores perdidos de acuerdo a una interpolación lineal entre los valores válidos más cercanos. Logrado esto, se procede a seleccionar las variables de acuerdo a sus estadísticos de desviación estándar y correlación.

[31] Descripción de las técnicas en el punto 5.1.6.1.

[32] En el software *R Project*, el paquete *zoo* contiene una función que permite realizar la interpolación fácilmente.

6.4 Selección de Variables

A partir de las 68 variables, en esta sección se descartan variables que estén: altamente correlacionadas con la variable dependiente y variables que tengan desviación estándar nula. La primera condición permite eludir el problema de heterogeneidad dentro del modelo. Éste redujo la base a 31 variables independientes útiles, sin considerar las 4 variables dependientes descritas en la Ecuación 6.1, mientras que la segunda condición, quitar variables que no entreguen más información de la que pueda entregar un valor constante, no eliminó ninguna. El Gráfico 6.4.1 describe el tipo y procedencia de las variables conservadas (véase ANEXO D para la lista de variables)

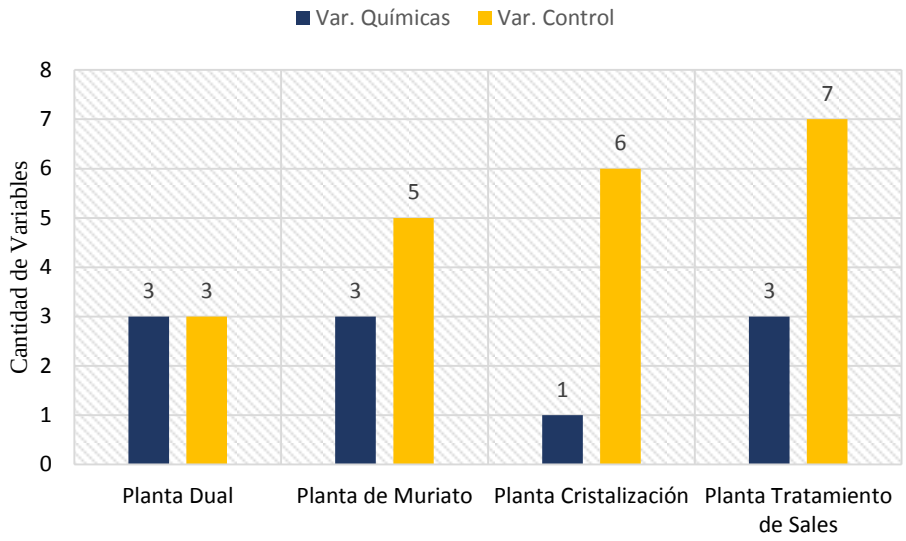


Gráfico 6.4.1: Número de Variables por Planta y tipo

Fuente: Elaboración del autor.

6.5 Relación Temporal de Variables

Con las variables dependientes reducidas a 31 y la unidad de tiempo estructurada en turnos de 8 horas, se procede a estudiar la relación temporal entre ellas. Debido a que los datos se registran temporalmente, una medición ocurrida en la etapa 1 en el turno $\boldsymbol{t_1}$ corresponde a una unidad de material distinta a la medida en la etapa 4 en el mismo turno $\boldsymbol{t_1}$. Por lo anterior, se ve necesario determinar la relación temporal entre las variables dependientes para así estructurar la base de datos. Para esto se recurre al análisis de series de tiempo y unidades de rezagos de la metodología Box-Jenkins que permite encontrar significancias de efectos correlacionados entre dos series de tiempo. En el punto 5.2.2 se describe los pasos de la metodología, sin embargo, como el fin no es pronosticar mediante este método, sólo se consideran los tres primeros pasos:

a. Verificar la estacionalidad de la serie. Si ésta no es estacionaria, diferenciarla hasta alcanzar estacionalidad.

b. Identificar un modelo tentativo.

c. Estimar el modelo.

Aplicando esto a cada una de los modelos de la Ecuación 6.1: Modelos relacionales, se busca determinar qué rezagos de cada variable tienen consecuencias significativas. Sin embargo, por posibles efectos de tendencias en las series de tiempo de cada variable, la metodología sugiere preblanquear[33] cada una. Recordando al lector, este consiste en:

1. Considerar dos series de tiempo, X e Y.
2. Determinar un modelo $ARMA$ para X y almacenar los residuos de este modelo (R_X).
3. Se filtra la serie Y. Esto quiere decir, que se describe la serie Y utilizando el modelo de la serie X (usando los coeficientes estimados del paso 1). En este paso se almacena las diferencias entre los valores de Y con su serie filtrada $(fil(Y) - Y)$.
4. Se analiza el CCF entre R_X y $(fil(Y) - Y)$. Este CCF se puede utilizar para identificar las los rezagos significativos para el modelamiento.

En el primer paso se consideró los pares ordenandos (X_i, Y_j), donde X_i representa las variables independientes de las ecuaciones e Y_j representa a las variables dependientes $KClO_4$_L, NaCl_L, $KClO_4$_M y NaCL_M.

En el segundo paso, se aplica la metodología de Box-Jenkins:

a. Se utilizó la *prueba de Dickey-Fuller aumentada*[34] para determinar la no estacionalidad de las series. En todos los casos el p-valor fue inferior a 0.01 lo que posibilita rechazar la hipótesis nula y afirmar estacionalidad.

b. Se estudiaron los correlogramas de cada serie para determinar sus estructuras $ARMA(p, q)$. La *Figura 6.1* es representativa de todas las variables analizadas. Como se puede observar allí, en el gráfico de la función de autocorrelación (ACF) que representa el proceso de media móvil, todos los rezagos analizados parecen ser significativos (fuera del intervalo de confianza descrito por la línea punteada azul) en contraste con el gráfico de la función de autocorrelación parcial (PACF), que representa el proceso autorregresivo, donde sólo los primeros parecen ser significativos. Como se establece en la metodología, la existencia de un patrón decreciente de rezagos como el ACF describe la ausencia del proceso, mientras que un decaimiento exponencial como el descrito en el gráfico PACF da cuenta de la existencia del proceso. Desde el punto de vista práctico, parece lógico pensar que un proceso industrial metalúrgico, como el que se está estudiando, dependa más del

[33] Véase punto 5.2.5.

[34] Véase punto 5.2.1.

valor que tomen sus condiciones en etapas anteriores que efectos aleatorios previos. De esta forma, sólo se estudian procesos del tipo AR.

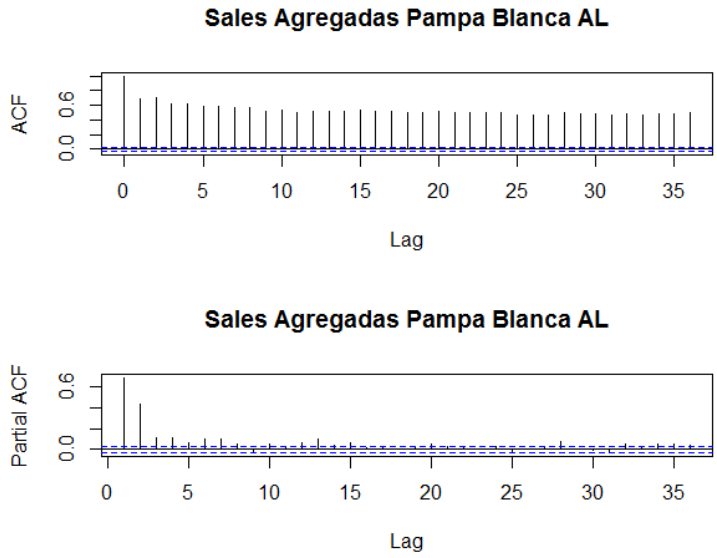


Figura 6.1: ACF y PACF de la variable Sales Agregadas Pampa Blanca AL

Fuente: Elaboración del autor.

c. Para determinar el parámetro p del proceso AR, se establece como límite el número de turnos hacia atrás que permite el proceso. Como se busca describir la concentración de los contaminantes que se miden en la etapa final de cristalización, las variables independientes se estructuran como realizaciones en el pasado que afectaron el estado presente de la concentración. Así, por ejemplo, si la variable $KClO_4\_L$ dio una concentración pura en el turno t_i, las variables de la planta Dual debieron haber tomado ciertos valores en turnos pasados. La *Tabla 6.5.1* describe cuantos turnos en el pasado deben considerarse para las variables de cada planta. Estos turnos son determinados por el tiempo que demora cada etapa.

	TURNOS PASADOS	
Planta	Mínimos	Máximos
Tratamiento de Sales	4	5
Dual	4	4
Muriato	1	4
Cristalización	0	1

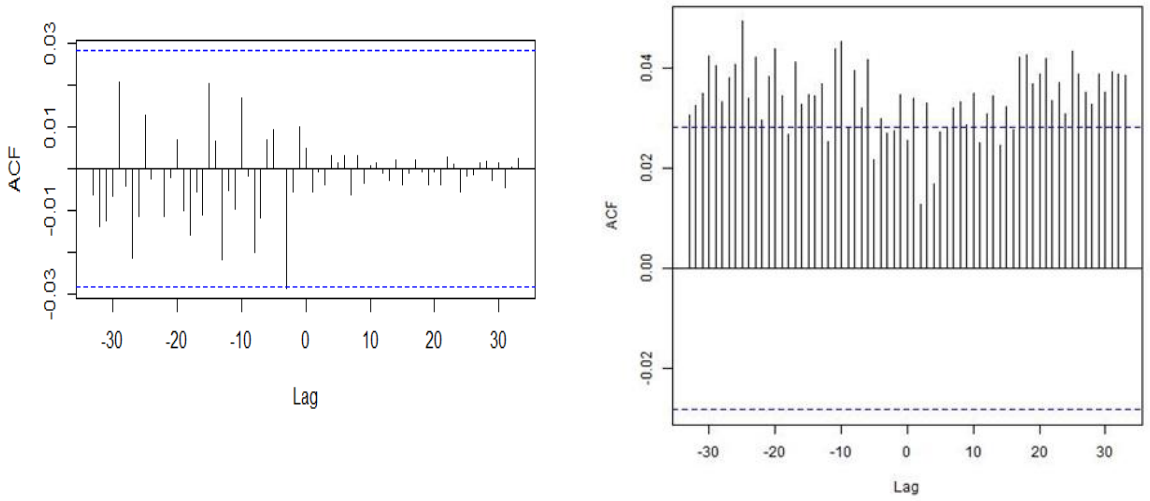
Tabla 6.5.1: Referencias de rezagos

Fuente: Elaboración del autor.

Al realizar los cálculos para determinar el parámetro p, se encontró una solución borde, donde los parámetros de cada proceso AR coincidía con su valor máximo posible. Siendo una solución válida, se acepta y se continúa con los pasos siguientes.

En el tercer paso, se calcula la serie filtrada de cada variable dependiente y se almacena la diferencia descrita en la metodología.

Al realizar el cuarto paso, se encontraron relaciones fuera de los límites permitidos. Los correlogramas cruzadas, que determina los rezagos de las variables independientes que afectan la realización de la variable dependiente, en algunos casos excedían los 5 turnos hacia atrás (el máximo permitido para cada ciclo de producción). Se exponen dos gráficos como ejemplo de los correlogramas encontrados. En el *Gráfico 6.5.1***,** la figura (a) muestra que la variable "*concentración de potasio en la corriente SSK*" presenta el cuarto turno del pasado como variable que afecta significativamente a la variable dependiente $KClO_4\_M$, mientras que en la figura (b), la variable "*concentración de potasio en la corriente MLR"* muestra que el rezago del 24° turno en el pasado es el que más afectaba en la variable dependiente $NaCl\_M$ (24 turnos equivaldrían a 5 ciclos productivos aproximadamente). Por ende, mediante este procedimiento no se logra determinar rezagos que vayan en la lógica del proceso, por lo que en su lugar se considera como aceptable el uso de los turnos mínimos expuesto en la *Tabla 6.5.1* para relacionar temporalmente las variables.



(a) $KClO_4\_M$ vs $variable$ 363106 $SSK\ K$ (b) $NaCl\_L$ vs $variable$ 335170 $MLR\ K$

Gráfico 6.5.1: Correlogramas

Fuente: Elaboración del autor.

7 MODELAMIENTO DEL PROCESO

En esta sección se busca determinar el modelo que relaciona las variables independientes con las dependientes, es decir, la función $\boldsymbol{f_j}(\,)$ de la *Ecuación 6.1*. Para ello se pretende usar algoritmos supervisados y contrastar su comportamiento a través de medidas de rendimiento. Se usa como conjunto de entrenamiento el 70% de la base de datos y como conjunto de testeo el 30% restante. Para este modelamiento se utilizó el software estadístico *R project* 3.1.2, el cual contiene una gran variedad de paquetes de datos donde vienen implementados los modelos que se pretenden utilizar. En particular se trabaja con el paquete *MLR (Machine Learners for R).*

Una pregunta interesante que también se busca responder es si es mejor utilizar un modelo de clasificación, el cual sólo predice la categoría de la variable dependiente, versus un modelo de regresión, el cual predice el valor de la variable dependiente. Debido a que en la literatura no hay información determinante que posicione a una sobre otra, se construirán ambos tipos de modelos.

Para este capítulo, se considerará la base de datos que contempla todas las modificaciones realizadas en el capítulo 6. Como los algoritmos a utilizar no presentan una formulación analítica sencilla, se detallarán los parámetros que usan en el ANEXO F.

7.1 Medidas de Rendimiento

Dado los dos tipos de modelo a estudiar, de clasificación y de regresión, se presentan dos grupos de indicadores de rendimiento, los cuales serán usan en cada una de la ecuaciones de la Ecuación 6.1 de forma independiente.

Para clasificación, la matriz de confusión (Tabla 7.1.1) entrega los aciertos y errores que comete cada modelo, donde ($\boldsymbol{a}$) corresponde a los valores verdaderos positivos, ($\boldsymbol{b}$) los falsos positivos, ($\boldsymbol{c}$) los falsos negativos y ($\boldsymbol{d}$) los verdaderos negativos. De esta tabla se obtienen los 3 indicadores de rendimiento del modelo: Accuracy, Precision y Recall.

		Valores Reales	
		Categoría Positiva	Categoría Negativa
Valores Predichos	Categoría Positiva	a	b
	Categoría Negativa	c	d

Tabla 7.1.1: Matriz de Confusión

Fuente: Elaboración del autor.

1. **Accuracy:** Porcentaje de predicciones correctas en el total de datos.

$$\boldsymbol{Ac = \frac{a + b}{a + b + c + d}}$$

2. **Precision**[35]: Para cada predicción, se define como el porcentaje de predicciones correctas dentro de la cantidad de datos de la categoría.

$$\boldsymbol{P_{positivo} = \frac{a}{a + b}, P_{negativo} = \frac{d}{c + d}}$$

3. **Recall:** Por cada valor real, se define como el porcentaje de predicciones correctas en el universo de valores de la categoría.

$$\boldsymbol{R_{positivo} = \frac{a}{a + c}, R_{negativo} = \frac{b}{b + d}}$$

4. **F-score:** Indicador que representa un balance entre la información entregada por *Precision* y *Recall.*

$$\boldsymbol{F_{score} = 2 * \frac{Precision * recall}{Precision + recall}}$$

Para regresión, se contrasta los valores observados con los predichos, para luego estimar el error que comete el modelo al predecir. Para ello se utilizarán 3 indicadores:

1. **MAE:** promedio de los errores absolutos que comete el modelo al momento de predecir. En otras palabras, representa cuánto se equivoca el modelo en términos absolutos.

$$\boldsymbol{MAE = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i|}$$

2. **MAPE:** promedio de los errores absolutos porcentuales que comete el modelo al momento de predecir. En otras palabras, representa cuánto se equivoca el modelo en términos porcentuales.

$$\boldsymbol{MAPE = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{\hat{y}_i - y_i}{y_i}\right|}$$

3. **MSE:** promedio del error cuadrático que comete el modelo al momento de predecir.

$$\boldsymbol{MSE = \frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}$$

7.2 Modelos de Clasificación

[35] Las siglas están en inglés, por lo que se omite la tilde en toda la memoria.

Se procedió a trabajar con 3 algoritmos de clasificación: Regresión Logística, Random Forest y Support Vector Machine (SVM). En la Tabla 7.2.1 se describen los parámetros que se requieren ajustar para conseguir el modelamiento de los datos.

Algoritmo	Parámetros
Regresión Logística	-
Random Forest	Mtry Ntree
Support Vector Machine	C Kernel

Tabla 7.2.1: Modelos y sus parámetros

Fuente: Elaboración del autor.

Cabe destacar que uno de los parámetros es el tipo de Kernel[36], el cual también presenta parámetros que pueden adaptarse a los datos. *R project* entrega valores default de éstos (Tabla 7.2.2), para así reducir la complejidad computacional que puede generar el cálculo de un ajuste del kernel y su modelo a la vez.

Kernel	Parámetros	Default
Lineal	No tiene	-
Polinomial	d	1
	α	1
	c	1
Gaussiano	σ	1
Anova	σ	1
	d	1
Multicuadratico	σ	1
	c	1

Tabla 7.2.2: Kernels y sus parámetros

Fuente: Elaboración del autor.

Eligiendo como categoría positiva la clase "impuro"[37], se ajustan los 3 tipos de algoritmos para cada una de las ecuaciones de estudio (Ecuación 6.1: Modelos relacionales), buscando el mejor indicador F_{score}, pero esto puede llevar a un sobreajuste en los datos. Para evitarlo, se estudia las medidas de rendimiento tanto en el conjunto de entrenamiento como en el

[36] Véase punto 5.1.4.

[37] En la Tabla 7.1.1, la categoría positiva es "impuro" y la categoría negativa es "puro". Los indicadores Precision y Recall se calculan sólo con la categoría positiva.

conjunto de validación. Si existe sobreajuste, se debería observar medidas de rendimientos muy altas en el primer conjunto y menores en el segundo.

7.2.1 Regresión Logística

La Tabla 7.2.3 resume el mejor ajuste de una Regresión Logística para cada ecuación, según F_{score} y menor sobreajuste.

	Conjunto de Entrenamiento				Conjunto de Testeo			
Ecuación	F_{score}	Accuracy	Precision	Recall	F_{score}	Accuracy	Precision	Recall
$KClO_4\_L$	0,40	0,91	0,29	0,66	0,38	0,90	0,27	0,61
$NaCl\_L$	0,10	0,95	0,06	0,31	0,13	0,95	0,08	0,35
$KClO_4\_M$	0,05	0,90	0,03	0,23	0,04	0,91	0,02	0,15
$NaCl\_M$	0,02	0,82	0,01	0,71	0,00	0,82	0,00	0,00

Tabla 7.2.3: Performance de una Regresión Logística en Clasificación

Fuente: Elaboración del autor.

Como se puede apreciar, los indicadores de cada ecuación son muy similares entre sus conjuntos de entrenamiento y de testeo, con la salvedad de la cuarta ecuación, $NaCl\_M$, que mostró una diferencia notoria, dado que sus indicadores de testeo, precisión y Recall son nulos. Sin embargo, esto no cae en el problema de sobreajuste, dado que de existir, los indicadores del conjunto de entrenamiento también deberían ser altos, pero el F_{score}y $Precision$ son relativamente bajos (cercanos al 1%).

7.2.2 Support Vector Machine

La Tabla 7.2.4 resume el mejor ajuste de SVM para cada ecuación, según F_{score} y menor sobreajuste.

	Conjunto de Entrenamiento				Conjunto de Testeo			
Ecuación	F_{score}	Accuracy	Precision	Recall	F_{score}	Accuracy	Precision	Recall
$KClO_4\_L$	0,65	0,94	0,50	0,93	0,52	0,92	0,38	0,81
$NaCl\_L$	0,92	0,99	0,85	1,00	0,17	0,94	0,13	0,24
$KClO_4\_M$	0,90	0,98	0,83	0,98	0,38	0,90	0,34	0,45
$NaCl\_M$	0,87	0,96	0,79	0,98	0,30	0,78	0,26	0,36

Tabla 7.2.4: Performance de SVM en clasificación

Fuente: Elaboración del autor.

La primera ecuación muestra ausencia de sobreajuste dado que sus indicadores de los conjuntos de entrenamiento y testeo son similares. En el resto de las ecuaciones, a pesar que el $Accuracy$ del conjunto de entrenamiento y de testeo son relativamente similares, el resto de los indicadores, $Precision\ y\ Recall$, en ambos conjuntos difiere mucho. El ajuste

con distintos parámetros no logró mejorar la relación, sin embargo, al generar nuevos conjuntos de testeo a través de muestreos aleatorios con reposición[38] de la misma base de datos, se observa que el promedio de las medidas de rendimiento se mantienen con una diferencia máxima de 6 puntos porcentuales (véase Tabla 7.2.5), lo cual es bajo. Por ende se aceptan los ajuste original.

	Conjunto de Testeo Original				Nuevo Conjunto de Testeo			
Ecuación	F_{score}	Accuracy	Precision	Recall	F_{score}	Accuracy	Precision	Recall
$KClO_4\_L$	0,52	0,92	0,38	0,81	0,50	0,92	0,37	0,80
$NaCl\_L$	0,17	0,94	0,13	0,24	0,18	0,93	0,18	0,19
$KClO_4\_M$	0,38	0,90	0,34	0,45	0,39	0,90	0,34	0,45
$NaCl\_M$	0,30	0,78	0,26	0,36	0,31	0,74	0,32	0,29

Tabla 7.2.5: Comparación conjunto de testeo

Fuente: Elaboración del autor.

7.2.3 RANDOM FOREST

La Tabla 7.2.6 resume el mejor ajuste de Random Forest para cada ecuación, según F_{score} y menor sobreajuste.

	Conjunto de Entrenamiento				Conjunto de Testeo			
Ecuación	F_{score}	Accuracy	Precision	Recall	F_{score}	Accuracy	Precision	Recall
$KClO_4\_L$	0,48	0,92	0,83	0,34	0,48	0,92	0,78	0,34
$NaCl\_L$	0,04	0,96	0,50	0,02	0,04	0,95	1,00	0,02
$KClO_4\_M$	0,28	0,91	0,78	0,17	0,27	0,92	0,65	0,17
$NaCl\_M$	0,03	0,82	0,67	0,02	0,03	0,83	0,80	0,02

Tabla 7.2.6: Performance de Random Forest en clasificación

Fuente: Elaboración del autor.

Como se puede observar, a diferencia de SVM, el error de sobreajuste es mucho menor, pues los valores de las medidas de rendimiento entre los conjuntos de entrenamiento y testeo son muy similares (a excepción de la precisión en la ecuación $NaCl\_L$). Esto es de esperarse, puesto que el algoritmo implementa una validación de datos de forma interna, de forma tal de evitar sobreajustes, así, la Tabla 7.2.6 es un reflejo de que el algoritmo convergió exitosamente.

7.2.4 CONTRASTE DE CLASIFICACIÓN Y SELECCIÓN DE ALGORITMO

[38] El muestreo aleatorio con reposición consiste es extraer muestras aleatorias del conjunto completo, sin reducirlo por las anteriores extracciones.

La Tabla 7.2.7 resume la información resultante de la implementación de los algoritmos. A pesar de que Random Forest demostró, en todas las ecuaciones, tener mayor certeza al predecir puros e impuros (Accuracy), su indicador F_{score} fue bajo debido a su bajo poder para clasificar correctamente la categoría impuro (Recall). Por el contrario, Support Vector Machine, mostró equilibrar de mejor forma los indicadores Precision y Recall, superando también a la regresión logística en todos los casos. Por tanto, se considera como modelos para las ecuaciones 1, 2, 3 y 4 el algoritmo Support Vector Machine.

		Conjunto de Testeo			
Ecuación	Algoritmo	F_{score}	Accuracy	Precision	Recall
$KClO_4\_L$	RL	0,38	0,90	0,27	0,61
	SVM	**0,51**	**0,92**	**0,38**	**0,80**
	RF	0,47	0,91	0,78	0,34
$NaCl\_L$	RL	0,13	0,95	0,08	0,35
	SVM	**0,16**	**0,94**	**0,13**	**0,24**
	RF	0,04	0,95	1,00	0,02
$KClO_4\_M$	RL	0,04	0,91	0,02	0,15
	SVM	**0,38**	**0,89**	**0,34**	**0,44**
	RF	0,27	0,92	0,65	0,17
$NaCl\_M$	RL	0,00	0,82	0,00	0,00
	SVM	**0,30**	**0,78**	**0,26**	**0,35**
	RF	0,03	0,82	0,80	0,01

Tabla 7.2.7: Contraste modelos de clasificación

*Algoritmo seleccionado se destaca en negrita.
Fuente: Elaboración del autor.

7.3 Modelos de Regresión

En la sección anterior se trabajó con tres algoritmos cuya finalidad es clasificar en la categoría impura, mientras que ahora se usarán tres algoritmos para predecir el valor de la concentración de cada contaminante en cada ecuación. Dos de los tres algoritmos tiene una adaptación para pasar del tipo clasificación al tipo regresión, estos son Random Forest y Support Vector Machine, cuyas adaptaciones son Random Forest Regression y Support Vector Regression. Además, para comparar los resultados con los algoritmos antes mencionados, se incluirá una regresión lineal cuyo fin es servir de comparación, dado que contribuye con el modelamiento simple del sistema. Regresión Lineal, Random Forest Regression y Support Vector Regression tienen las mismas especificaciones de parámetros (Tabla 7.2.1) y del kernel (Tabla 7.2.2) que en los modelos de clasificación (salvo para el modelo de regresión lineal que no tiene más parámetros externos). Para cada modelo primero se calcula las diferentes medidas de rendimiento: $MAE, MAPE\ y\ MSE$. Además, como se busca poder contrastar su rendimiento con los modelos de clasificación, se procede

también a estimar la categoría impura desde estas regresiones para luego calcular las medidas de rendimiento de clasificación $Accuracy, Precision, Recall\ y\ F_{score}$.

7.3.1 REGRESIÓN LINEAL

La Tabla 7.3.1 resume el ajuste de una regresión lineal en cada ecuación según las medidas de rendimiento de regresión y su posterior aplicación de medidas de rendimiento de clasificación.

	Regresión			Clasificación			
Ecuación	MAE	MAPE	MSE	Accuracy	Precision	Recall	F_{score}
$KClO_4\_L$	6,18%	53,59%	0,86%	0,95	1,00	0,0	0,1
$NaCl\_L$	12,15%	21,23%	2,49%	0,95	0,50	0,0	0,1
$KClO_4\_M$	15,49%	34,23%	3,81%	0,90	0,37	0,4	0,6
$NaCl\_M$	20,35%	31,18%	8,46%	0,82	0,21	0,0	0,1

Tabla 7.3.1: Performance de una Regresión Lineal

Fuente: Elaboración del autor.

Como se puede observar, como regresión presenta un MAPE que varía entre el 20% y 54%, lo cual lo describe como bajo poder explicativo y predictivo. A la vez, como clasificador, su indicador F_{score} es bajo, salvo la tercera ecuación. Era de esperarse que una regresión lineal simple no ajustase correctamente a los datos, dado que por el nivel de interrelaciones entre variables de control y químicas, un modelo lineal no podría describir bien el comportamiento del proceso.

7.3.2 SUPPORT VECTOR REGRESSION

La Tabla 7.3.2 resume el ajuste de Support Vector Regression en cada ecuación según las medidas de rendimiento de regresión y su posterior aplicación de medidas de rendimiento de clasificación.

	Regresión		Clasificación			
Ecuación	MAE	MSE	Accuracy	Precision	Recall	F_{score}
$KClO_4\_L$	0,05	0,01	0,91	0,69	0,36	0,47
$NaCl\_L$	0,11	0,03	0,95	0,60	0,04	0,08
$KClO_4\_M$	0,13	0,03	0,92	0,57	0,17	0,27
$NaCl\_M$	0,12	0,03	0,82	0,40	0,01	0,02

Tabla 7.3.2: Performance de Support Vector Regression

Fuente: Elaboración del autor.

Se puede observar que el indicador MAE y MSE no superan el 0,15 y el 0,5 respectivamente, lo que describe el buen ajuste de los modelo. Por el lado de los indicadores de clasificación, presentan un Accuracy sobre el 82%, lo que significa que clasifica correctamente el conjunto de categorías puro e impuro. Analizando sólo la categoría impuro, se puede apreciar que el indicador Precision es más elevado en las primeras 3 ecuaciones, pero indicador Recall, es bajo en los 3 últimos. Esto se traduce como que los modelos tienen un grado de poder de clasificación dentro de la categoría (Precision), pero sus predicciones capturan un porcentaje bajo del universo de observaciones impuros (Recall).

7.3.3 RANDOM FOREST REGRESSION

La Tabla 7.3.3 resume el ajuste de Random Forest Regression en cada ecuación según las medidas de rendimiento de regresión y su posterior aplicación de medidas de rendimiento de clasificación.

	Regresión			Clasificación			
Ecuación	MAE	MAPE	MSE	Accuracy	Precision	Recall	F_{score}
$KClO_4\_L$	0,02	18%	0,00	0,93	0,78	0,40	0,53
$NaCl\_L$	0,04	7%	0,00	0,96	0,60	0,05	0,09
$KClO_4\_M$	0,05	11%	0,00	0,91	0,92	0,08	0,14
$NaCl\_M$	0,07	11%	0,01	0,82	0,57	0,08	0,14

Tabla 7.3.3: Performance de Random Forest Regression

Fuente: Elaboración del autor.

7.3.4 CONTRASTE DE CLASIFICACIÓN Y SELECCIÓN DE ALGORITMO

La Tabla 7.3.4 resume la información resultante de la implementación de los algoritmos. En indicadores de regresión, como se puede observar, la regresión lineal tiene el peor ajuste, lo que es de esperarse como se comentó en el punto 7.3.1. Por otro lado, Random Forest fue el algoritmo mejor evaluado, tanto en MAE y MSE de regresión, como F_{score} en clasificación (siendo superado sólo en la tercera ecuación por SVR). Como algoritmo, es robusto en predicción y permite establecer un ranking de variables según su importancia de incidencia en la variable dependiente. Por tanto, se considera como modelos para las ecuaciones 1, 2 y 4, los algoritmos Random Forest, mientras que para la 3 se modela con Support Vector Regression.

Ecuación	Algoritmo	Regresión		Clasificación		
		MAE	MSE	F_{score}	Precision	Recall
$KClO_4\_L$	RL	0,06	0,01	0,1	1,00	0,00
	SVR	0,05	0,01	0,47	0,69	0,36
	RFR	**0,02**	**0,00**	**0,53**	**0,78**	**0,40**
$NaCl\_L$	RL	0,12	0,03	0,1	0,50	0,00
	SVR	0,11	0,03	0,08	0,60	0,04
	RFR	**0,04**	**0,00**	**0,09**	**0,60**	**0,05**
$KClO_4\_M$	RL	0,15	0,04	0,6	0,37	0,40
	SVR	**0,13**	**0,03**	**0,27**	**0,57**	**0,17**
	RFR	0,05	0,00	0,14	0,92	0,08
$NaCl\_M$	RL	0,20	0,09	0,1	0,21	0,00
	SVR	0,12	0,03	0,02	0,40	0,01
	RFR	**0,07**	**0,01**	**0,14**	**0,57**	**0,08**

Tabla 7.3.4: Contraste de modelos de Regresión

*Algoritmo elegido está en negrita.
Fuente: Elaboración del autor.

7.4 Elección del Modelo

Como se menciona en la sección anterior, se intenta comparar los modelos de clasificación y regresión dado que en la literatura no hay recomendaciones de cuál alternativa usar. Si bien, por ser un caso donde se busca predecir un estado de la muestra, los modelos de clasificación aparecen de forma natural. Sin embargo, estos modelos carecen de la capacidad de estudiar la sensibilidad del proceso ante perturbaciones, por lo que recurren a otros mecanismos, como Simulaciones de MonteCarlo, para hacerlo. Por otro lado, los modelos de regresión tienen esta capacidad, pero sólo son útiles cuando la variable dependiente es continua. En este trabajo, gracias a la naturaleza del proceso, es posible aplicar los dos modelos. Se procede a contrastar el poder de clasificación de ambos, sin embargo, se conservan los dos tipos por para aprovechar las ventajas de cada uno en la interpretación de los datos.

La Tabla 7.4.1 resume los algoritmos mejor ajustados según la ecuación de análisis. Aquí se puede observar que el poder de clasificación de los algoritmos de regresión es inferior, valga la redundancia, a los de clasificación, dado su bajo indicador F_{score}. Es interesante el caso de la primera ecuación, el contaminante $KClO_4$en la corriente L, donde se dio el caso contrario, Random Forest Regression superó a Support Vector Machine. Si se analiza en detalle el contraste, se observa que los modelos regresivos poseen un mayor Precision, mientras los de clasificación mejoran en Recall. En otras palabras, los primeros clasifican mejor las observaciones en sus respectivas categorías, y los segundos logran generar categorías más representativas del universo de observaciones. Se abre como línea de

estudio para otros trabajos, el análisis del rendimiento usando modelos híbridos que integren regresiones y clasificaciones.

Ecuación	Algoritmo	Tipo	F_{score}	Precision	Recall
$KClO_4\_L$	SVM	Clasificación	0,51	0,38	0,8
	RFR	Regresión	0,53	0,78	0,4
$NaCl\_L$	SVM	Clasificación	0,16	0,13	0,24
	RFR	Regresión	0,09	0,60	0,05
$KClO_4\_M$	SVM	Clasificación	0,38	0,34	0,44
	SVR	Regresión	0,27	0,57	0,17
$NaCl\_M$	SVM	Clasificación	0,30	0,26	0,35
	RFR	Regresión	0,14	0,57	0,08

Tabla 7.4.1: Elección de modelo

Fuente: Elaboración del autor.

8 DISEÑO DE PRUEBAS Y VALIDACIÓN DEL MODELO

A continuación se analiza los modelos de clasificación y regresión que mejor ajustaron para las 4 ecuaciones de estudio. Así, en esta sección, cuando se hable de los modelos de clasificación, se entenderá por el algoritmo de Support Vector Machine aplicado en $KClO_4\_L$, $NaCl\_L$, $KClO_4\_M$ y en $NaCl\_M$, y cuando se hable de los modelos de regresión, se entenderá por el algoritmo de Random Forest Regression aplicado en $KClO_4\_L$, $NaCl\_L$, $NaCl\_M$ y el algoritmo de Support Vector Regression aplicado en $KClO_4\_M$. Se procede a describir el comportamiento que predicen del proceso, las variables que afectan mayoritariamente a la realización de cada variable dependiente y comentar mínimos operacionales resultantes de estos modelos. La unidad temporal trabajada en cada ecuación es el turno como se determinó en el punto 6.1.

8.1 Descripción del Modelamiento

Una vez obtenido los mejores modelos de clasificación y de regresión que se ajustan en cada ecuación, se realiza un análisis de la forma en que se describe el proceso desde las predicciones, su nivel de impureza estimado y los errores que cometen los modelos. Es bueno recordar que una observación se considera impura si supera el límite de concentración de un contaminante, siendo 0,24% de concentración peso/peso el límite para el $KClO_4$ en la corriente L, mientras que para el $NaCl$ en la misma corriente el límite es 0,95% de concentración peso/peso. Para la corriente M, el límite para el $KClO_4$ y $NaCl$ es de 0,95% de concentración peso/peso en ambos casos.

En el proceso de producción del nitrato de potasio, de los 4815 turnos ocurridos entre los años 2006 y 2012 que contempla el análisis, la corriente L presentó 518 observaciones[39] impuras (10.8%) en $KClO4$ y 218 observaciones impuras (4,5%) en $NaCl$, mientras que en la corriente M se dieron 450 (9,3%) y 860 (17,9%) de observaciones impuras del primer y segundo contaminante respectivamente.

8.1.1 Descripción Con Modelos De Clasificación

En modelos de clasificación, un buen ajuste debiese discriminar un alto porcentaje de las producciones impurezas clasificadas como tales, con un bajo nivel de mal clasificados (clasificados como puros), mientras que un mal ajuste debiese discriminar ambas categorías en el mismo porcentaje. Uno podría pensar que un mal ajuste discrimina un bajo porcentaje de datos impuros como tales, pero invirtiendo la etiqueta de predicción se puede convertir ese "bajo porcentaje de impuros" como un "bajo porcentajes de puros", corrigiendo el ajuste. No poder discriminar se entiende como no permitir un grado de certeza sobre la

[39] Se entenderá por *observación* el producto obtenido en un turno

predicción, y de esta forma, un modelo que acierta en la mitad de las observaciones y yerra en las otras es un mal ajuste.

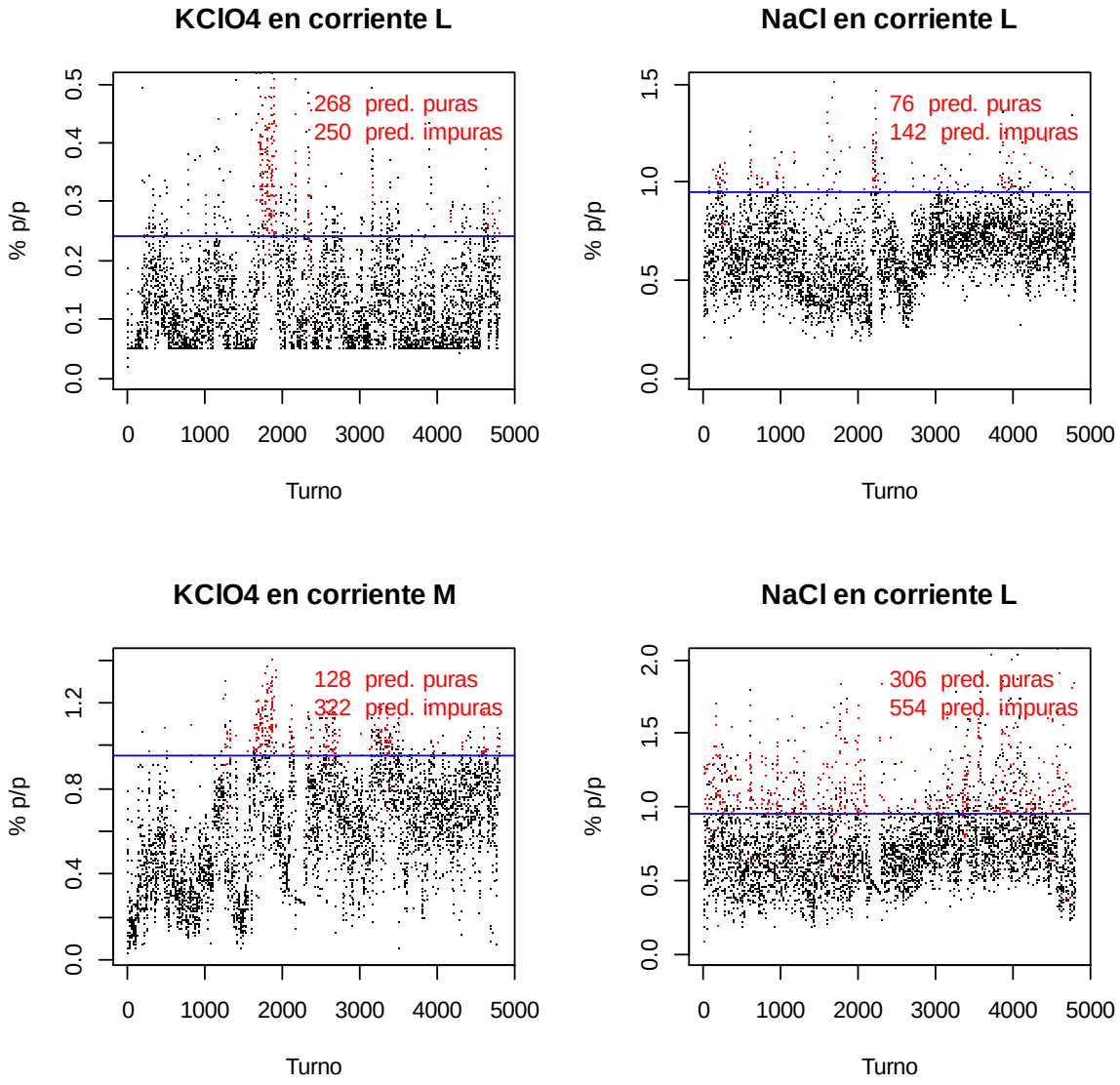


Gráfico 8.1.1: Ajuste de los modelos de clasificación para los 4815 turnos ocurridos entre 2006 y 2012

*En la parte superior derecha se muestra el número de predicciones puras e impuras que hace cada modelo sobre el universo de observaciones impuras.

Fuente: Elaboración del autor.

En el Gráfico 8.1.1 se puede apreciar de forma visual el ajuste que permiten los modelos de clasificación sobre la data histórica. En este se grafican los turnos analizados y el nivel de concentración que lograron los contaminantes en cada corriente. Luego se etiquetan en rojo las observaciones que los modelos clasifican como impuras y en negro las que no. Se añade una línea azul que representa el límite de concentración desde el cual una observación es catalogada como impura (sobre la línea) y una leyenda en la parte superior izquierda, que

describe el número de predicciones puras e impuras que hace cada modelo sobre el universo de observaciones impuras. Así, por ejemplo, en el gráfico de la parte superior izquierda, se ve que de las 518 observaciones impuras, el algoritmo de Support Vector Machine predice aproximadamente la mitad como puros y la mitad como impuros, lo que cataloga a este como un mal ajuste.

Del gráfico mencionado se puede inferir que, en la mayoría de las ecuaciones, los modelos no alcanzaron un nivel de clasificación que permitiera certeza de la categoría *impuro*, pues, mientras que en el gráfico superior izquierdo el porcentaje de observaciones impuras bien clasificadas fue de un 48%, en el resto de los gráficos varía entre un 64% y un 72%, lo que genera que en promedio un 30% de las observaciones se estén perdiendo.

8.1.2 Descripción Con Modelos De Regresión

En un modelo regresivo con buen nivel de ajuste, se esperaría que un gráfico de los valores reales versus las predicciones describieran una recta, sin embargo, se puede ver, en la *Gráfico 8.1.2*, que las predicciones no presentan una relación tan directa con el valor real, aunque se puede esbozar una tendencia lineal positiva. En la tercera ecuación ($KClO_4$ en corriente M) se perfila de mejor forma un ajuste lineal con una correlación de 0.87.

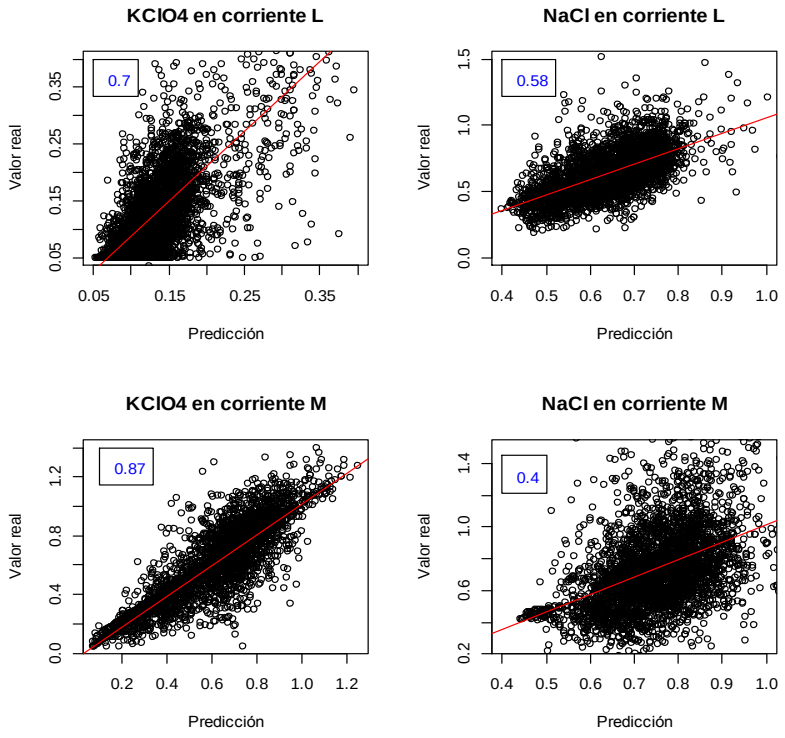


Gráfico 8.1.2: Valor real vs Predicción de la concentración de contaminantes para los 4815 turnos ocurridos entre 2006 y 2012

*En la parte superior izquierda se muestra el índice de correación de una regresión lineal simple (línea roja).
Fuente: Elaboración del autor

.

El grado de dispersión que presentan las predicción versus los valores reales puede explicarse por el sesgo que tienen los modelos al no incluir variables relevantes para explicar la concentración de los contaminantes, los cuales pudieron ser omitidos al no cumplir con el porcentaje de datos faltantes máximo que sugiere la metodología. De 310 variables existentes, el 89% de ellas quedó fuera de la base de datos consolidada, y debido a la ausencia de información que reportaron, no fue posible hacer una selección por capacidad predictiva de cada variable, lo que pudiese haber mejorado el ajuste del modelo. Por otra parte, un aspecto positivo es la comparación del nivel de concentración de datos que se puede observar en el Gráfico 8.1.2, lo que permite inferir que los modelos encontrados predicen la concentración en rangos cercanos a los reales.

Un resultado distinto es el que se consigue al estudiar la capacidad de los modelos regresivos de discriminar entre productos puros e impuros. El Gráfico 8.1.3, valga la redundancia, grafica las predicciones de cada modelo y las clasifica: en rojo si la observación original era impura, y lo contrario en negro. Se agrega en azul la concentración del contaminante que discrimina entre producto puro e impuro y una leyenda descriptiva que da cuanta de cuantas observaciones puras e impuras quedan sobre la línea azul (es decir, clasificadas en la categoría impura).

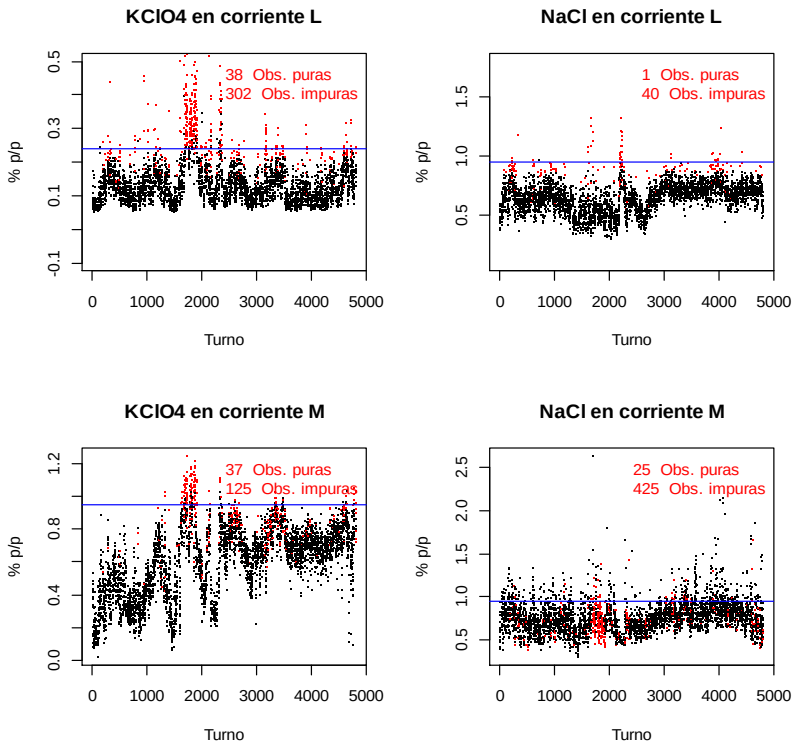


Gráfico 8.1.3: Capacidad de clasificación de modelos regresivos para los 4815 turnos ocurridos entre 2006 y 2012

*En la parte superior derecha se muestra el número de predicciones puras e impuras que hace cada modelo sobre el universo de observaciones impuras.
Fuente: Elaboración del autor.

Las observaciones impuras que se han registrado entre el 2006 y el 2012, han sido 518 y 218 para el $KCl0_4$ y el $NaCl$ en la corriente L, y 450 y 860 para el $KCl0_4$ y el $NaCl$ en la corriente M. Como se puede ver en el Gráfico 8.1.3, la primera ecuación (figura superior izquierda) muestra que del total de observaciones impuras, el modelo logra clasificar correctamente el 58,3%.

8.2 Importancia de variables

Para el análisis de importancia de variables se estudia el impacto que tiene retirar una variable independiente sobre los indicadores de rendimiento del modelamiento, para identificar qué variables son más relevantes para poder describir y predecir la información.

Para los modelos de clasificación se midió cuánto disminuye el indicador F_{score} al retirar una variable, debido a que, una disminución de este, implica que el modelo pierde poder para balancear su poder de clasificar correctamente la categoría impuro (Precision) como su capacidad de capturar una muestra representativa del universo de observaciones impuras (Recall). Para los modelos de regresión se midió el incremento del indicador MSE, que describe cuanto aumenta el error de predicción del modelo versus las observaciones reales.

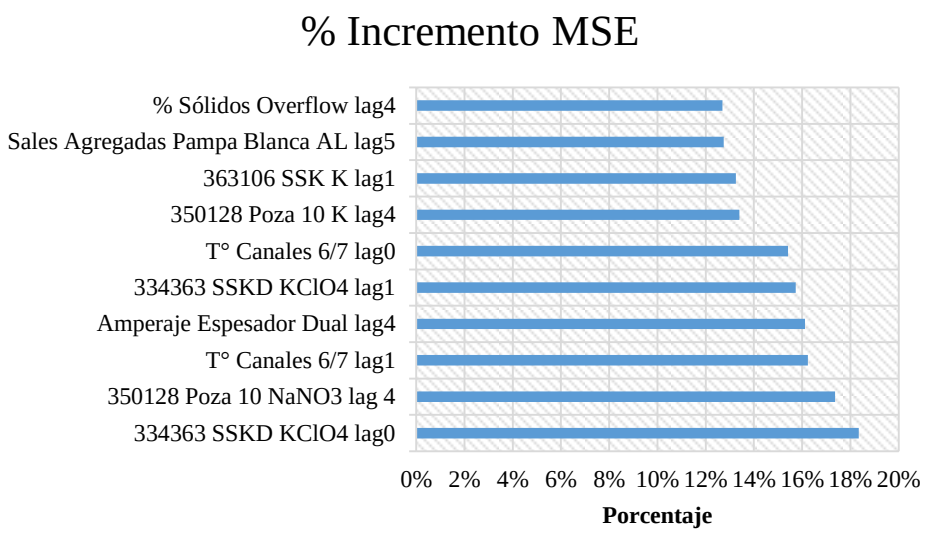
Además, se agrega la partición del tipo de variable, esto es, la cantidad de variables químicas y variables de control que componen la lista. Es importante mencionar que sobre las variables químicas se tiene poco capacidad de manejo, dado que estas son el resultado de reacciones químicas, lo que sí se puede manejar son las condiciones sobre las cuales esas reacciones ocurren; temperatura, volumen, corriente eléctrica, entre otros. Estas últimas condiciones son resumidas en el tipo "variable de control".

Cabe mencionar que las ecuaciones contemplan las mismas variables independientes (y por ende, la misma cantidad de variables químicas y de control), pero su importancia depende de que variable dependiente estén describiendo. Por lo ello, se describirá a continuación las 10 variables independientes que más importancia tienen en cada caso buscando caracterizar las ecuaciones a través de un conjunto reducido de variables.

8.2.1 RANKING DE VARIABLES EN MODELOS DE CLASIFICACIÓN

Mediante el análisis de los modelos de clasificación se puede obtener un listado de la importancia de cada variable en el poder de balanceo de Precision y Recall. Esto se calcula estudiando cuánto disminuye porcentualmente el indicador F_{score} si se extrae de la base la

i-ésima variable independiente. Repitiendo esto para las 66 variables se construye el listado



de

Gráfico 8.2.5, Gráfico 8.2.6, Gráfico 8.2.7 y Gráfico 8.2.8 con las 10 variables más relevantes.

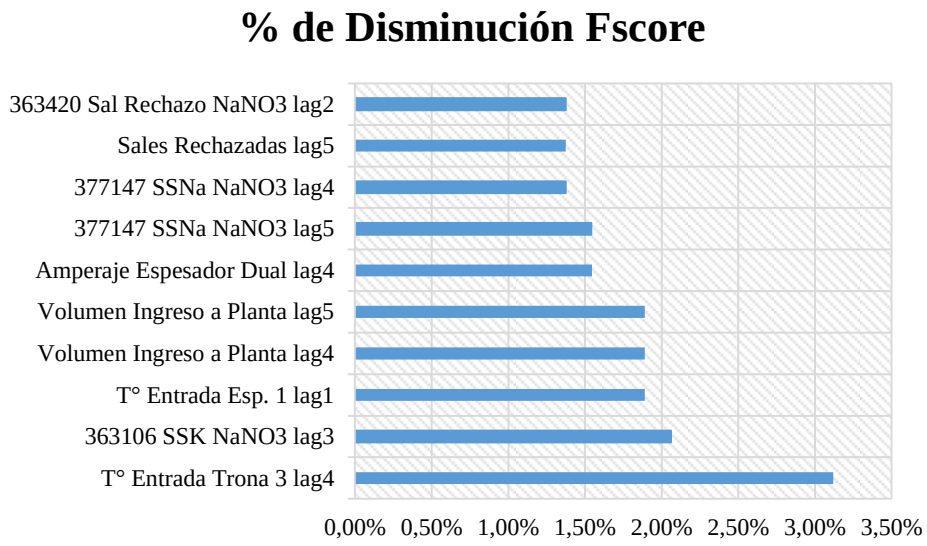


Gráfico 8.2.1: Ranking en KClO4_L por clasificación

Fuente: Elaboración del autor.

El Gráfico 8.2.1 muestra el ranking de importancia de variables para el modelo del contaminante $KClO_4$ en la corriente L. Como se puede observar, la variable más importante es la *T° Entrada Trona 3 lag4*, que representa el cuarto turno en el pasado de la temperatura de entrenada en la tercera mezcladora de material (trona), la cual pertenece a la planta de Muriato, lo que refleja que la concentración de este contaminante en el producto final

depende de la temperatura con la cual se mezcle la corriente que proviene de la planta de tratamiento de Sales con el KCl[40]. Sin embargo, sólo dos variables, la mencionada y el *Amperaje Espesador Dual lag4*, pertenecen a la planta de Muriato, mientras que 5 provienen de la planta de tratamiento de sales, lo cual es esperable puesto que las condiciones iniciales del proceso impactan significativamente en los resultados finales. Otro aspecto a mencionar de este ranking es que 4 variables corresponden a mediciones químicas, dejando 6 variables de control, posibilitando la existencia de manejo de este proceso.

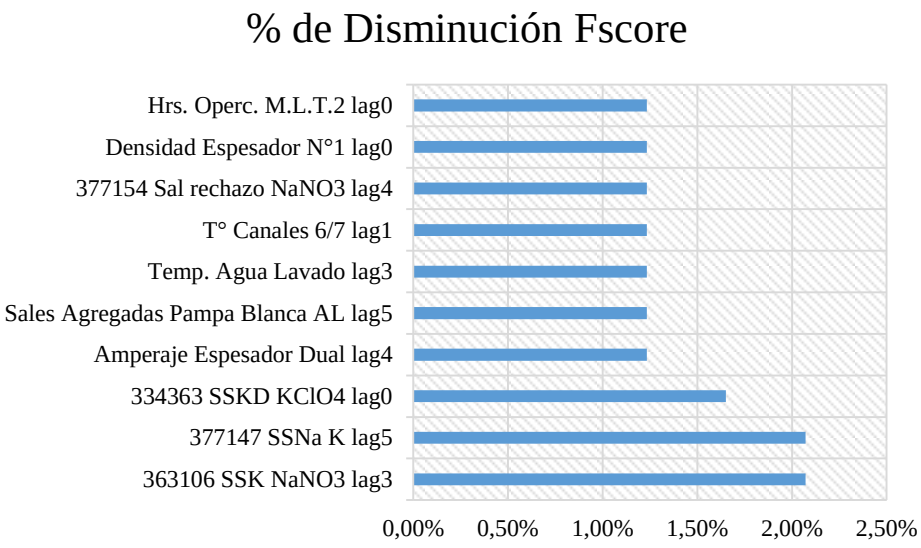


Gráfico 8.2.2: Ranking en NaCl_L por clasificación

Fuente: Elaboración del autor.

El Gráfico 8.2.2 muestra el ranking de importancia de variables para el modelo del contaminante $NaCl$ en la corriente L. Se puede observar que las tres variables más relevantes corresponden a variables químicas que representan concentraciones de K, $NaNO_3$ y $KClO_4$ en distintas corrientes, lo que genera que el producto final depende más de variables que no tiene mucho grado de control. Sin embargo, resultaron 6 variables de control de este ranking, permitiendo compensar esta dependencia a las primeras variables químicas. De la procedencia de las variables, se puede ver que de la planta de tratamiento de Sales vienen 3, de la Planta de Muriato vienen 2, de la Planta Dual vienen 1 y de la planta de Cristalización vienen 4, lo que deja ver que las condiciones más determinantes se dan al principio y final del proceso.

[40] Se recuerda que la planta de Muriato es donde se agrega, desde una corriente externa, el cloruro de potasio (KCl).

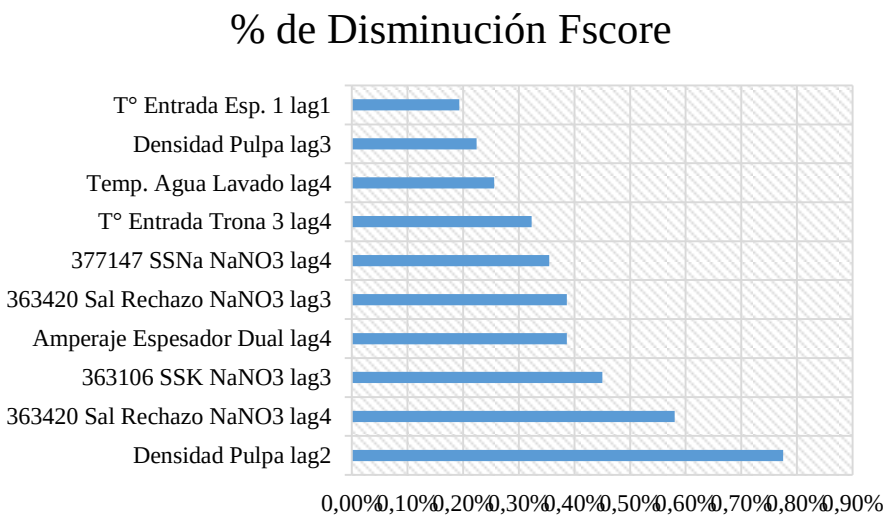


Gráfico 8.2.3: Ranking en KClO4_M por clasificación

Fuente: Elaboración del autor.

El Gráfico 8.2.3 muestra el ranking de importancia de variables para el modelo del contaminante $KClO_4$ en la corriente M. Interesante ver que, a diferencia de las ecuaciones anteriores, la variable más relevante, Densidad Pulpa lag2, proviene de la planta Dual junto a otras 5 variables, es decir, esta planta es determinante para la concentración de este contaminante en la corriente. No es claro una explicación directa de esta relación, una posibilidad es que esta planta es la que entrega el primer concentrado de potasio al sistema, por lo que su importancia radicaría en que a mayores concentraciones más posibilidades de generar más desechos. La partición de variables también generó 6 variables de control dentro del ranking.

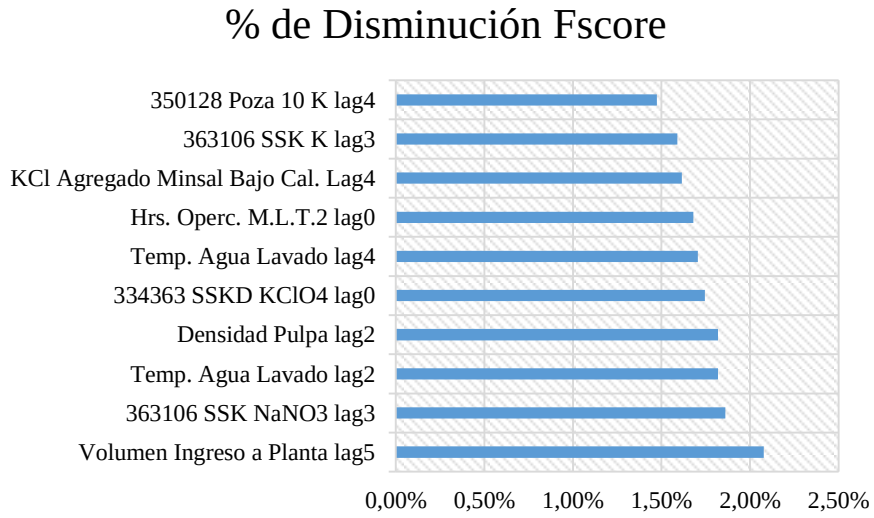


Gráfico 8.2.4: Ranking en NaCl_M por clasificación

Fuente: Elaboración del autor.

El Gráfico 8.2.8 muestra el ranking de importancia de variables para el modelo del contaminante $NaCl$ en la corriente M. Al igual que el caso anterior, aquí la variable más relevante, *Volumen Ingreso a Planta lag5*, también resultó provenir de la planta Dual. Sumado a lo anterior, se puede ver que el proceso de la segunda corriente depende más de la planta Dual, lo que puede deberse a que la corriente M, como es producto de un reprocesamiento del flujo de la corriente L, depende mayormente de la corriente de descarte que se genere aquí, de forma tal, que un mayor desecho, produce mejoras en la calidad del flujo final. La partición de variables también generó 6 variables de control dentro del ranking.

La **Tabla 8.2.1** muestra un resumen de los tipos de variables y su procedencia encontrada para cada ecuación.

	Plantas				Tipo	
Ecuación	PTS	MUR	DUAL	CRIST	CONTROL	AQ
$KClO_4\_L$	2	1	2	**5**	6	4
$NaCl\_L$	1	**4**	2	3	6	4
$KClO_4\_M$	2	1	**6**	1	6	4
$NaCl\_M$	2	2	**5**	1	6	4

Tabla 8.2.1: Número de variables relevantes por planta y tipo
*PTS: Planta de Tratamiento de Sales, MUR: Planta de Muriato, DUAL: Planta Dual, CRIST: Planta de Cristalización, CONTROL: Variables de Control, AQ: Variables Químicas.
Fuente: Elaboración del autor.

Se puede observar, en la Tabla 8.2.1**: Número de variables relevantes por planta y tipo**, que la distribución del tipo de variables (dentro de las 10 variables más importantes) es idéntica en cada ecuación, i.e., 6 de control y 4 químicas, lo que abre la posibilidad de mayor manejo del sistema al existir más variables de control. Resumiendo los párrafos anteriores, la primera ecuación es explicada en mayor medida a través de las variables de la planta de Cristalización, la segunda ecuación es explicada en mayor medida por las variables de la planta de Muriato y la tercera y cuarta ecuación son explicadas mayormente por las variables de la planta Dual. Una pregunta interesante es si existen variables transversales a las ecuaciones, es decir, estén dentro de las 10 más importantes independiente de la ecuación. Se encontró sólo una variable transversal, la variable *363106 SSK NaNO3 lag3* que corresponde a una variable química perteneciente a la planta Dual, representando la concentración %p/p del Nitrato de Sodio ($NaNO_3$) en el flujo del sistema. Es lógico entender esta variable como trascendental debido a que refleja uno de los elementos necesarios en el producto final, el nitrato, el cual, en bajas concentración no logra reaccionar con el cloruro de potasio generando pocas cantidades de nitrato de potasio. Además de esta, existen dos variables que podemos catalogar como transversales, pues pese

a que está presente en 3 de las 4 ecuaciones, están dentro de las 10 más importantes, estas son *Amperaje Espesador Dual lag4* (no es importante para la ecuación 4) *y Temperatura Agua Lavado lag2* (no es importante para la ecuación 1), las cuales pertenecen a las plantas de Muriato y Dual respectivamente. El análisis más operativo la existencia de estas variables requiere una interpretación más química del estudio, cosa que está fuera del alcance de esta memoria, pero se deja abierta la posibilidad a *La Empresa* de indagar más detalladamente este aspecto.

8.2.2 RANKING DE VARIABLES EN MODELOS DE REGRESIÓN

Mediante el análisis de los modelos regresivos se puede obtener un listado de la importancia de cada variable en el poder predictivo. Esto se calcula estudiando cuánto aumenta porcentualmente el indicador MSE si se extrae de la base la i-ésima variable independiente. Repitiendo esto para las 66 variables se construye el listado de los

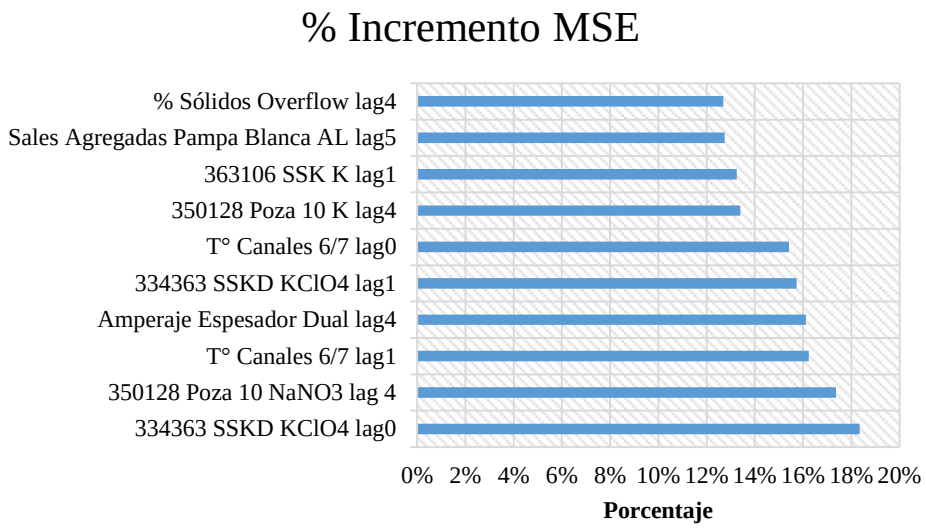


Gráfico 8.2.5, Gráfico 8.2.6,Gráfico 8.2.7 y Gráfico 8.2.8 con las 10 variables más relevantes.

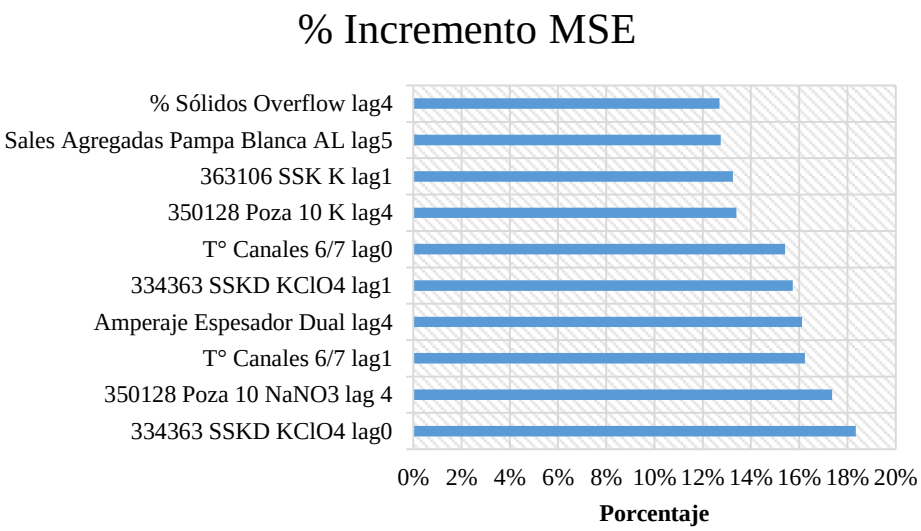
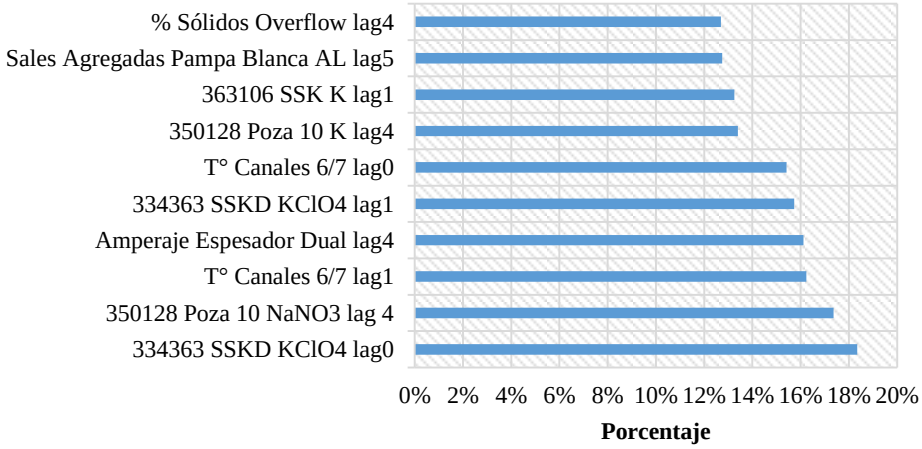


Gráfico 8.2.5: Ranking en KClO4_L por regresión

Fuente: Elaboración del autor.



En el

Gráfico 8.2.5 se puede observar que la variable que más afecta al MSE del modelo es el *334363 SSKD KClO4 lag0*, es cual corresponde a una variable química de la planta Cristalización. Esto es esperable, puesto que SSKD es la corriente que entra en la 4 planta, y la cantidad de $KClO_4$ que entre tendrá incidencia directa en la pureza del nitrato de potasio. Es importante hacer notar que este compuesto es un contaminante en el producto final, pero no lo es en el proceso, ya que es un insumo necesario para la producción. Los restos de este compuesto que no puedan ser aprovechados por el proceso son lo que quedan catalogados como contaminante. Como esta variable es química, no hay mucha capacidad de manejarla directamente, y recién la tercera variable relevante es de tipo "control". A nivel general, de las 10 variables relevantes, 5 son de control y 5 químicas.

Si se analiza la planta de procedencia de estas variables se puede ver que 4 pertenecen a la planta de Cristalización, 1 de la planta Dual, 3 de la planta de Muriato y 2 de la planta de Tratamiento de Sales. Es lógico pensar que la planta de la que depende más la pureza del producto es la planta final de cristalización donde se obtiene el nitrato de potasio por precipitación del material mediante enfriamiento, debido a que no lograr una buena separación de compuestos genera más presencia de contaminantes. Es más, la tercera variable relevante es *T° Canales 6/7 lag1*, que corresponde a la temperatura usada en el proceso en el turno anterior.

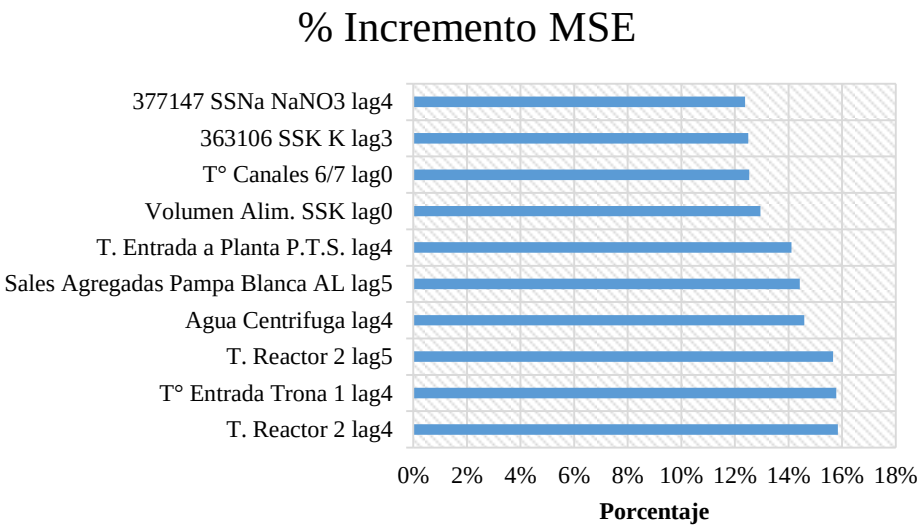


Gráfico 8.2.6: Ranking en NaCl_L por regresión

Fuente: Elaboración del autor.

Para la ecuación $NaCl\_L$, las primeras tres variables más importantes corresponden a variables de control relacionadas con la temperatura del proceso (Gráfico 8.2.6). Sin embargo, a diferencia de la ecuación $KClO_4\_L$, estas variables pertenecen a las plantas de Muriato y de Tratamiento de Sales, lo que puede explicarse por el tipo de contaminante que es más trabajado en una u otra planta. De esta forma, la concentración de cloruro de sodio en la corriente L depende más de las condiciones iniciales generadas en las primeras plantas que el perclorato de potasio que depende de la última.

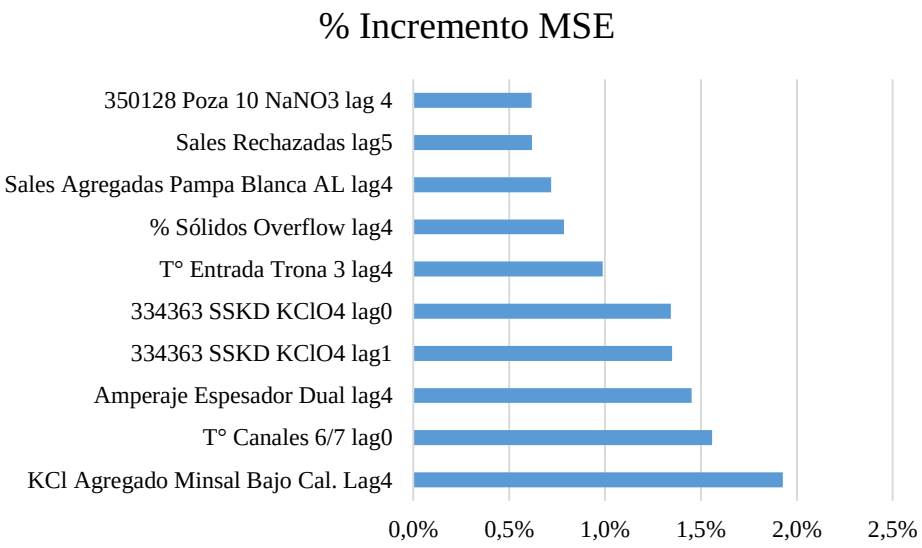


Gráfico 8.2.7: Ranking en KClO4_M por regresión

Fuente: Elaboración del autor.

Para la ecuación $KClO_4\_M$, la concentración de perclorato de potasio depende en mayor medida de la variable *KCL Agregado Minsal Bajo Cal. Lag4*, el cual corresponde a la cantidad de cloruro de potasio agregado en la planta de Muriato al flujo del proceso 4 turnos atrás. Como la corriente M es la segunda corriente de salida que aprovecha los residuos de la primera corriente, ésta recibe un flujo empobrecido de potasio, lo que hace que los niveles de pureza que logre dependa más del enriquecimiento de compuestos que reciba en la vuelta al proceso. Por otro lado, al igual que la primera ecuación, una de las primeras variables relevantes resultó ser la temperatura de canales en la planta de cristalización, salvo que esta vez en el rezago 1 (*T° Canales 6/7 lag1*). De esta forma, se puede inferir que ésta temperatura juega un rol crucial para disminuir el contaminante $KClO_4$ en el producto final.

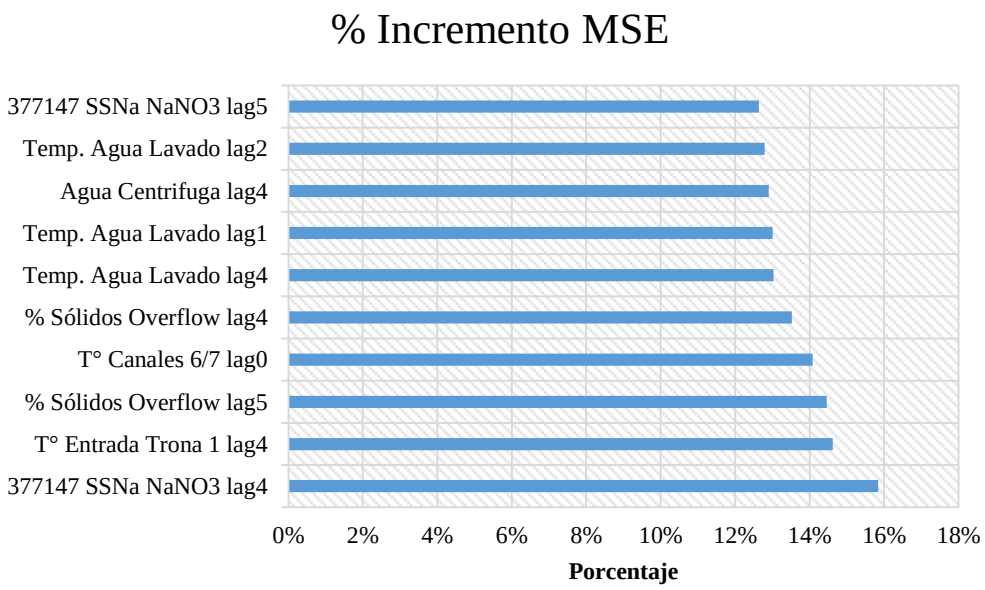


Gráfico 8.2.8: Ranking en NaCl_M por regresión

Fuente: Elaboración del autor.

Así como la concentración del cloruro de sodio en la corriente L, las primeras 3 variables de mayor relevancia en la ecuación $NaCl\_M$ también pertenecen a las plantas de Muriato y Tratamiento de Sales, es decir, la concentración del cloruro de sodio depende más de las condiciones iniciales del proceso (Gráfico 8.2.8). A pesar de que las primeras 3 no son las mismas, la lista contempla varias variables de temperatura del proceso tal como en el Gráfico 8.2.6, lo que permite suponer dependencias a la temperatura similares para el contaminante cloruro de sodio en ambas corrientes.

Es interesante ver el contraste entre el listado de variables de cada ecuación. La Tabla 8.2.2 muestra un resumen del tipo de variable y la planta a la que pertenecen. Un punto relevante a hacer notar es que la partición de tipo de variables permite ver la posibilidad de control sobre el proceso, si hubiese resultado una mayor proporción de variables químicas, la capacidad de gestionar y mejorar el proceso habría sido muy limitadas.

	Plantas				Tipo	
Ecuación	PTS	MUR	DUAL	CRIST	CONTROL	AQ
$KClO_4\_L$	2	3	1	**4**	5	5
$NaCl\_L$	**6**	1	1	2	8	2
$KClO_4\_M$	3	**4**	0	3	7	3
$NaCl\_M$	**5**	1	3	1	8	2

Tabla 8.2.2: Número de variables relevantes por planta y tipo

*PTS: Planta de Tratamiento de Sales, MUR: Planta de Muriato, DUAL: Planta Dual, CRIST: Planta de Cristalización, CONTROL: Variables de Control, AQ: Variables Químicas.

Fuente: Elaboración del autor.

Se puede observar, en la Tabla 8.2.2, la distribución del tipo de variable (dentro de las 10 variables más importantes) difiere según la ecuación, desde cantidades iguales en la primera ecuación, hasta 8 de control y 2 químicas en las ecuaciones de la corriente M, lo que abre la posibilidad de mayor manejo del sistema al existir más variables de control. Resumiendo los párrafos anteriores, la primera ecuación es explicada en mayor medida a través de las variables de la planta de Cristalización, la segunda y la cuarta ecuación son explicadas en mayor medida por las variables de la planta de Tratamiento de Sales y la tercera ecuación es explicada mayormente por las variables de la planta de Muriato. En el punto 8.2.1 se analiza la existencia de variables transversales que, recordando al lector, son las variables que están entre las 10 más importantes repitiéndose en cada ecuación. Se encontró sólo una variable transversal, la variable de control *T° Canales 6/7 lag0* que corresponde a la temperatura de los canales 6 y 7 de la planta de Cristalización. Es lógico entender esta variable como transversal, debido a que en la última planta última, la reacción química para generar la cristalización del nitrato de potasio depende directamente de la temperatura. Sin embargo, sólo es posible controlar esta variable al final del proceso, lo que limita la posibilidad de anticiparse cuando el flujo viene en condiciones no óptimas.

Para finalizar esta sección de Importancia de Variables cabe preguntarse cómo se comparan las variables importantes obtenidas por modelos de clasificación versus modelos de regresión. Para cada ecuación, se determinaron las variables que se repetían en su modelamiento regresivo como en el clasificatorio:

Ecuación	Variable	Tipo de Variable	Planta de procedencia
$KClO_4\_L$	Amperaje Espesador Dual lag4	Control	Muriato
$NaCl\_L$	Sales Agregadas Pampa Blanca AL lag5 T° Canales 6/7 lag1	Control	Tratamiento de Sales Cristalización
$KClO_4\_M$	Amperaje Espesador Dual lag4 T° Entrada Trona 3 lag4	Control	Muriato
$NaCl\_M$	T° Agua Lavado lag2 T° Agua Lavado lag4	Control	Dual

Tabla 8.2.3: Variables Transversales En El Modelamiento Regresivo Y Clasificatorio

Fuente: Elaboración del autor.

Interesante notar que dentro de estas variables ninguna pertenece al tipo químico, lo que permite suponer la relevancia de las variables de control independiente del tipo de modelamiento que se haga. Otro aspecto que se observó es la procedencia de estas variables, que en conjunto abarcan las 4 plantas, por lo que no se puede determinar la importancia de una sola planta en la calidad final del producto.

8.3 Contribución Marginal de las Variables

Una de las ventajas que presenta el software *R* al momento de trabaja con Random Forest Regression, es que permite analizar los cambios en la predicción del modelo variando una única variable a la vez. Como las ecuaciones $KClO_4\_L$, $NaCl\_L$ y $NaCl\_M$ se modelan con

este algoritmo, es posible aplicar este análisis. Para la ecuación $KClO_4\_M$ que usa Support Vector Regression, replicar esto requiere un mayor conocimiento teórico del algoritmo, el cual no es el foco de la memoria, por lo que se omite su estudio. Sin embargo, se espera no incurrir en inconvenientes del objetivo buscado, ya que como la ecuación omitida del análisis pertenece a la corriente M, la cual aprovecha los restos de la corriente L, una mejora en esta primera corriente debería traer consigo mejores desempeños de la segunda corriente. Esta sección sólo contempla el análisis desde modelos regresivos, dado que los modelos de clasificación de dos categorías no permiten estudiar el impacto que tiene el incremento o decremento de las variables independientes en las variables dependientes.

Se entiende por contribución marginal de una variable al efecto que genera en el valor de la predicción la variación de dicha variable. Se realizó este análisis en las ecuaciones descritas en el párrafo anterior con las 66 variables en juego y se determinó en qué rangos se consigue minimizar la concentración de los contaminantes. Cabe mencionar que, a pesar que en la realidad se tiene poco poder de gestión sobre las variables químicas, se analizó de igual forma su contribución marginal, puesto serán necesarias para el capítulo siguiente.

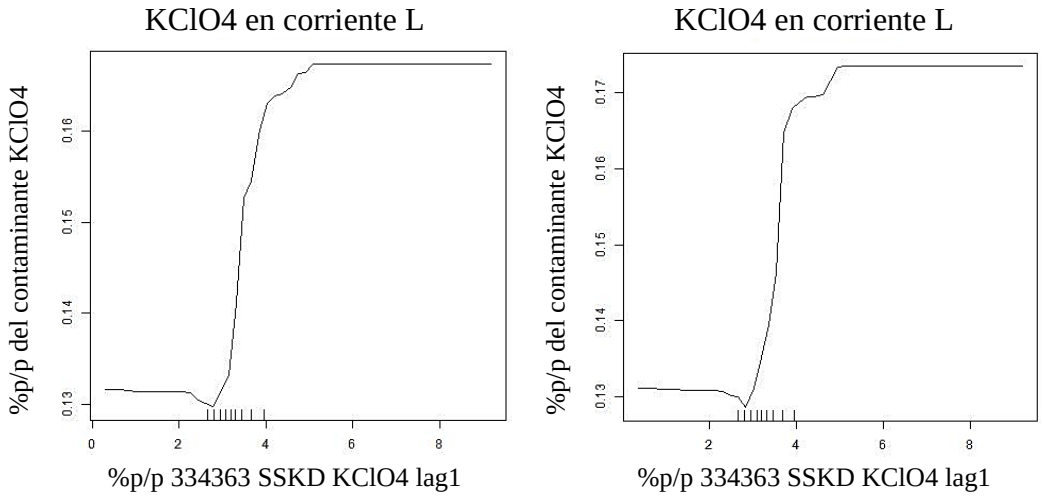


Gráfico 8.3.1: Comparación de la contribución Marginal de una misma variable en los 2 rezagos que usa el modelamiento.

Fuente: Elaboración del autor.

Es bueno recordar que de las 66 variables que son usadas en el modelamiento, algunas son el rezago de otra, por ejemplo de la variable *% Sólidos Overflow* se usa el rezago 4 (4 turno en el pasado) y el rezago 5 (5 turnos en el pasado), por ello es que una de las relaciones esperables es que, para los distintos rezagos de una misma variable, sus rangos de operación, que minimizan la concentración del contaminante, deberían ser similares. Satisfactoriamente se llegó a esta conclusión, lo que se ejemplifica con el Gráfico 8.3.1, donde, para los rezagos 0 y 1 de la variable *334363 SSKD KClO4*, el comportamiento de las curvas son parecidas y el mínimo se alcanza cuando los rezagos toman los valores de %p/p 1,63 y 1,65 respectivamente (véase ANEXO D para el detalle los mínimo óptimo de cada variable en cada ecuación).

Es importante estudiar el efecto que tiene el disminuir la impureza de un contaminante en una de las corrientes sobre la concentración del resto de los contaminantes en las corrientes. Así, por ejemplo, puede ocurrir que disminuir la impureza del $NaCl$ en la corriente L genere que la corriente M presente mayor concentración de $NaCl$. La Tabla 8.3.1 describe este análisis, donde cada columna indica qué ecuación se está mejorando y las filas el resultado que tiene la mejora en cada ecuación. Se añade el mínimo histórico que se ha logrado en cada ecuación para tener un contraste real de la predicción que logran los modelos.

Ecuaciones	Mejorar $KClO_4\_L$	Mejorar $NaCl\_L$	Mejorar $NaCl\_M$	Mínimo Histórico	Límite de Concentración*
$KClO_4\_L$	**0,07**	0,27	0,22	0,02	0,24
$NaCl\_L$	0,66	**0,49**	0,64	0,19	0,95
$KClO_4\_M$	0,43	0,49	0,49	0,03	0,95
$NaCl\_M$	0,74	0,75	**0,55**	0,09	0,95

(*)Es el límite en que se clasifica una producción como pura (bajo el valor) o impura (sobre el valor).

Tabla 8.3.1: Comparación de mínimos alcanzados

Fuente: Elaboración del autor.

Como se puede apreciar en la Tabla 8.3.1, mejorar cualquiera de las 3 ecuaciones (a excepción de la ecuación $KClO_4\_L$ con la mejora de la ecuación $NaCl\_L$) logra que los resultados mínimos de cada contaminante estén por debajo de su límite de concentración, generando productos de categoría puro. Cabe mencionar que, si bien los resultados esperados no mejoran en comparación con los mínimos alcanzado históricamente, si logra asegurar que en conjunto, todos los contaminantes estén por debajo de su límite (la columna *Mínimo Histórico* muestra mínimos alcanzados en ocurrencias distintas).

9 ANÁLISIS DE SENSIBILIDAD DE RESULTADOS

En esta sección se discute el efecto que tiene la implementación de rangos de variación sobre las variables de los modelos desarrollados en la sección anterior, para analizar cómo es afectada la pureza del producto final a través de su nivel de contaminación. Para ello, se usarán los resultados obtenidos de la sección 8.3 Contribución Marginal de las Variables[41] las variaciones correspondientes.

9.1 Introducción

Una pregunta simple que nace del punto *8.3 Contribución Marginal de las Variables* es qué sucede si se construyen rangos de variación de las variables definiendo que su contribución marginal no supere el límite de pureza, es decir, si se considera la variable *Volumen Ingreso a Planta lag4* del Gráfico 9.1.1 y se fija un límite de pureza del 0.1405. ¿Qué efectos se consiguen a nivel general considerando la variable restringida en el intervalo (404.44,1820)? Determinando estos rangos para cada variable, se simulan 1000 turnos de operación con estas restricciones para analizar el porcentaje de pureza que logra el proceso en cada corriente, para cada contaminante. De la misma forma que la sección anterior, se realiza un análisis cruzado del efecto de las restricciones de una ecuación sobre el porcentaje de impureza en el resto de las ecuaciones.

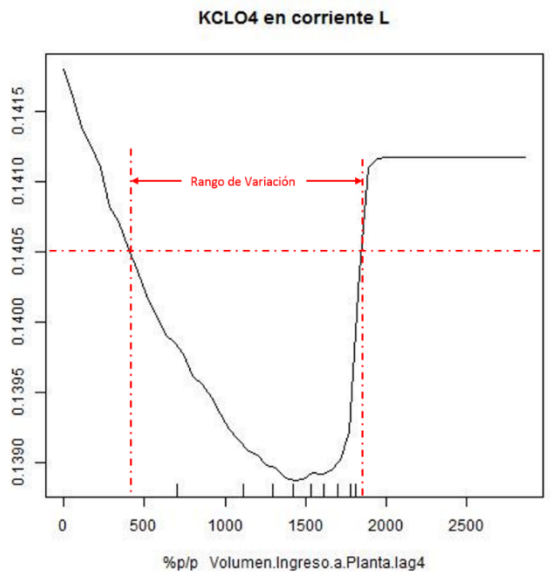


Gráfico 9.1.1: Ejemplo de determinación de rangos de variación usando el análisis de contribución marginal de la variable Volumen Ingreso a Planta lag4

Fuente: Elaboración del autor.

[41] Por la misma razón dada en la sección 8.3, sólo se trabajará con las predicciones de los modelos regresivos dado que los de clasificación no permiten el estudio de la contribución marginal de las variables.

La Tabla 9.1.1 resume la información de la simulación de 1000 turnos bajo las condiciones mencionadas en el párrafo anterior usando como límite de pureza el límite usual del sistema[42]. Como se puede observar, tanto la primera ecuación como la cuarta, los porcentajes de impureza alcanzados son enormes, lo que significa que hay factores que no se están considerando en el análisis. Una de las relaciones que no se muestran en los modelos es las interdependencias de las variables entre sí, dado que es un proceso metalúrgico, las variables químicas dependen de las de control y las de control dependen entre ellas (la temperatura de un proceso y el volumen que trabajan están correlacionados). Otro efecto no considerado es que se asume que se pueden restringir las variables químicas y las de control, siendo que las primeras no cuentas con capacidad de gestión directa. Es por ello que se analiza y compara el efecto de distintos rangos de variación junto al efecto de controlar las variables químicas o no.

Ecuaciones	Mejoras* en $KClO_4\_L$	Mejoras en $NaCl\_L$	Mejoras en $NaCl\_M$
$KClO_4\_L$	99%	99%	99%
$NaCl\_L$	3%	0%	0%
$KClO_4\_M$	0%	0%	0%
$NaCl\_M$	88%	86%	87%

Tabla 9.1.1: Impureza estimada

*Las Mejoras en $KClO_4\_L$ hacen referencia a la simulación en torno a los óptimos encontrados para la ecuación $KClO_4\_L$, así cada fila explica que pasa con el resto de las ecuaciones usando dichos óptimos. La misma idea se repite con cada columna.
Fuente: Elaboración del autor.

En esta sección se analizarán 3[43] tipos de políticas que se pueden implementar en el proceso productivo, estas hacen referencia la sensibilidad en que los óptimos encontrados en cada ecuación estudiando la contribución marginal se pueden aumentar o disminuir sin que perjudique la pureza de la producción. En otras palabras, sea $\bar{x}_j^{ecuación\ i}$ el valor óptimo de la variable j que minimiza la contribución marginal de la impureza en la ecuación i, se busca determinar los parámetros $a_j^{ecuación\ i}$ y $b_j^{ecuación\ i}$ para cada variable $j = 1..6$ y ecuación $i \in \{1,2,4\}$ tal que cualquier combinación de variables $\{x_j\}_{j=1}^{66}$ que cumpla

[42] Como se ha mencionado en las secciones anteriores, el límite de pureza para las ecuaciones $KClO_4\_L$, $NaCl\_L$, $KClO_4\_M$ y $NaCl\_M$ son 0,24, 0,95, 0,95 y 0,95 p/p respectivamente.

[43] Se omite la política sobre la ecuación $KClO_4\_M$, puesto no se logró desarrollar su análisis de contribución marginal de las variables dado que usaba el algoritmo Support Vector Regression y no Random Forest Regression que sí lo tiene implementado.

$$\{x_j\}_{j=1}^{66} \in \prod_{i=1}^{66}\left(\bar{x}_i^{ecuación\ i} - a_i^{ecuación\ i}, \bar{x}_i^{ecuación\ i} + b_i^{ecuación\ i}\right)$$

Y satisfaga que en simulación no contribuya a aumentar el porcentaje de impureza producido. La simulación a usar consiste en realizar combinaciones aleatorias uniformes de los valores que cada variable puede tomar entre los valores máximos y mínimos determinados por el intervalo buscado.

Para simplificar la notación, se usará $a_j^{ecuación\ i} = -b_j^{ecuación\ i} = p^{ecuación\ i} * \max\{x_j\}$, donde $p^{ecuación\ i}$ es el porcentaje, para la ecuación i, en que cada variable independiente puede variar con referencia a su valor máximo histórico.

De esta forma, se definen las políticas como:

1. **Primera política de rangos:** correspondiente al intervalo en que las variables independientes se pueden mover, en torno a los valores óptimos encontrados al analizar la contribución marginal que reduce la impureza de la ecuación $KClO_4\_L$.
2. **Segunda política de rangos:** correspondiente al intervalo en que las variables independientes se pueden mover, en torno a los valores óptimos encontrados al analizar la contribución marginal que reduce la impureza de la ecuación $NaCl\_L$.
3. **Tercera política de rangos:** correspondiente al intervalo en que las variables independientes se pueden mover, en torno a los valores óptimos encontrados al analizar la contribución marginal que reduce la impureza de la ecuación $NaCl\_M$.

Para determinar el rango de variación en cada política, se usará el análisis de contribución marginal, visto en la sección anterior, para definir los límites de cada variable, luego se simularán con ellos 1000 turnos de producción registrando el porcentaje de turnos que resultaron con producción impura. Además, se analiza la situación de ver el impacto que tiene el bajo poder de manejo de las variables químicas a través de un contraste en la simulación usando límites en todas las variables versus solamente en las variables de control. Para hacer este último caso, se evaluó hacer la simulación dejando que las variables químicas variaran usando sus máximos históricos, sin embargo, esto anularía el efecto que tiene las variables de control sobre ellas. Por ello, y para no agregar mayor complejidad a la simulación, se optó por fijar las variables químicas en sus valores moda, para representar los casos más usuales que uno esperaría. Finalmente, se analizará el efecto en el tonelaje e ingresos de la producción.

El análisis sobre la ecuación 3, $KClO_4\_M$ no se detallará en este capítulo dado que la simulación sobre ella dio un nivel de tolerancia de 40% de variación sobre las variables en cada una de las políticas mencionadas, rango que es mayor a cada uno de los analizados más adelante, permitiendo que sólo sea necesario focalizarse en el resto de las ecuaciones para determinar los rango de variación de cada política.

9.2 Primera Política de Rangos

Se procede a analizar el caso de los rangos de variación de la primera ecuación y su efecto sobre el resto de las ecuaciones. Primero se analizará el efecto que tiene restringir las variables químicas o no, para así definir los rangos de variación y luego estudiar su impacto en el resto de las ecuaciones.

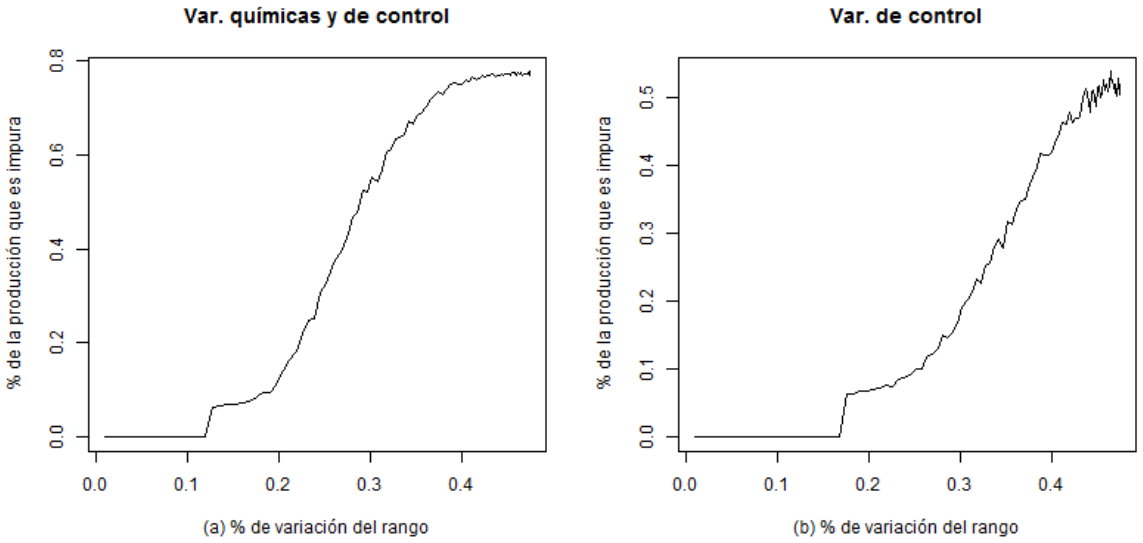


Gráfico 9.2.1: Efecto de la política 1 en la ecuación KClO4_L

Fuente: Elaboración del autor.

El Gráfico 9.2.1 muestra como varía el porcentaje de impureza de la corriente L, dado por el contaminante $KClO_4$, cuando aumenta el porcentaje de variación en torno a los mínimos óptimos de sus variables. En la figura (a) se observa el caso cuando todas las variables son controladas bajo rangos de variación. Aquí, el porcentaje de impureza simulado no aumenta hasta que la variación del rango no llegue a un 12% promedio en torno a los mínimos de cada variable. Después de esto, el porcentaje simulado empieza a incrementarse rápidamente hasta alcanzar un 78% de impureza cuando el rango varía un 47% promedio, lo cual es esperable dado que una mayor varianza en las variables del sistema genera más tipos de combinaciones no puras. Es interesante ver el contraste que se da cuando no se impone rangos a la variación de las variables químicas, lo que se ve en la figura (b). Para este caso, las variables químicas están fijas en su valor moda, por lo que el efecto de aumentar el rango no perjudica la producción hasta los 17,5% promedio de variación. Como era de esperar, si se controla el efecto de las variables químicas (éstas están fijas en su valor moda), el porcentaje de turnos con producciones impuras debería disminuir, lo que se observa en la figura (a) y (b), donde de un máximo de 78% baja a un 54%.

De esta forma, la política 1 se definiría como permitir una variación del 12% promedio en las variables de las ecuaciones en torno a los mínimos logrados para la ecuación $KClO_4\_L$.

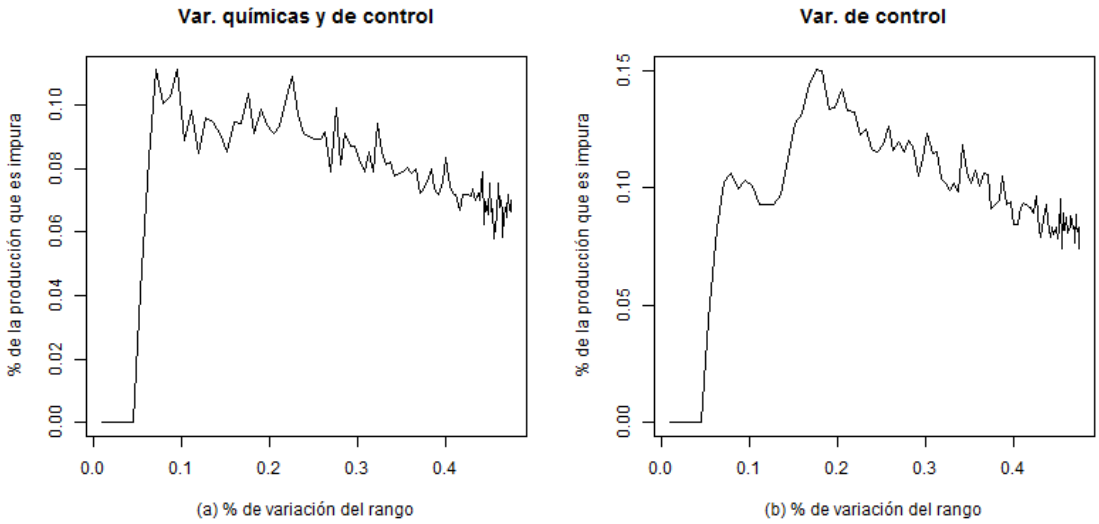


Gráfico 9.2.2: Efecto de la política 1 en la ecuación NaCl_L

Fuente: Elaboración del autor.

El Gráfico 9.2.2 muestra como varía el porcentaje de impureza de la corriente L, dado por el contaminante $NaCl$, cuando aumenta el porcentaje de variación en torno a los mínimos de las variables de la ecuación $KClO_4$. Se puede apreciar, que el comportamiento logrado es muy similar considerando variaciones en las variables químicas y de control o sólo en estas últimas. En el gráfico se puede ver un rápido ascenso del porcentaje de impureza cuando se deja un rango mayor al 4,5% promedio de variación de sus variables, llenando a un porcentaje de impureza aproximado del 11%, valor superior al porcentaje de impureza histórico para el $NaCl$ en la corriente L, correspondiente al 4,5%. También se puede apreciar una tendencia a la baja a medida que aumenta el rango de variación, esto puede explicarse por la aparición de efectos de compensación, es decir, la mayor variación de una variable mejora la contribución marginal a la pureza de otra variable. Por último mencionar que el límite de porcentaje de impureza simulado es mayor en la figura (b), lo que permite inferir que la variación de las variables químicas es un factor relevante a la hora de reducir la impureza provocada por este contaminante en la corriente L.

Con lo anterior, la política 1 aumentaría la concentración del contaminante $NaCl$ en la corriente L dado que permite variaciones del 12% promedio sobre sus variables. Para lograr evitar este perjuicio, se debe ajustar el rango de variación al 4,5%.

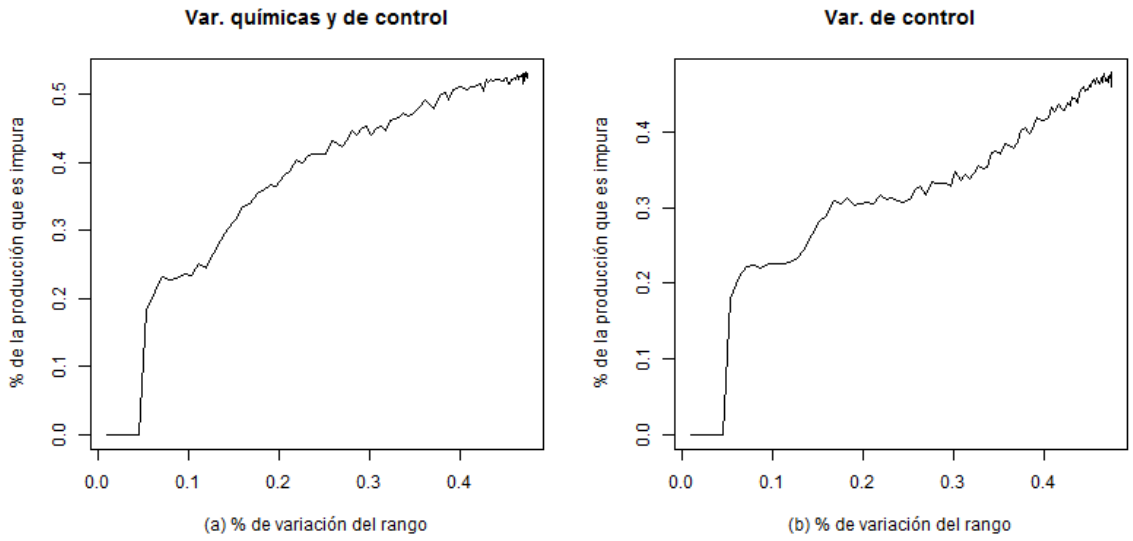


Gráfico 9.2.3: Efecto de la política 1 en la ecuación NaCl_M

Fuente: Elaboración del autor.

El Gráfico 9.2.3 muestra como varía el porcentaje de impureza de la corriente M, dado por el contaminante $NaCl$, cuando aumenta el porcentaje de variación en torno a los mínimos de las variables de la ecuación $KClO_4$ Como se puede apreciar, variando ambos tipos de variables como sólo las de control, el comportamiento descrito es similar, con un aumento del porcentaje de impureza a medida que el rango de variación sobre los 4,5% promedio. El máximo de impureza alcanzado es levemente mayor en la figura (a) (53%) que en la (b) (48%), dado que en la segunda, la estabilidad de las variables químicas favorece el resultado.

En este caso, la política 1, ajustada con la ecuación $NaCl\_L$, es decir, con un rango de variación de 4,5%, evita el incremento de la impureza simulada de ambos contaminantes en la corriente L y M.

9.3 Segunda Política de Rangos

La segunda política de rangos está definida la variación de las variables sobre los puntos mínimos que mejorar la pureza de la corriente L dado por la concentración del contaminante $NaCl$. Primero se analizará el efecto que tiene restringir las variables químicas o no, para así definir los rangos de variación y luego estudiar su impacto en el resto de las ecuaciones.

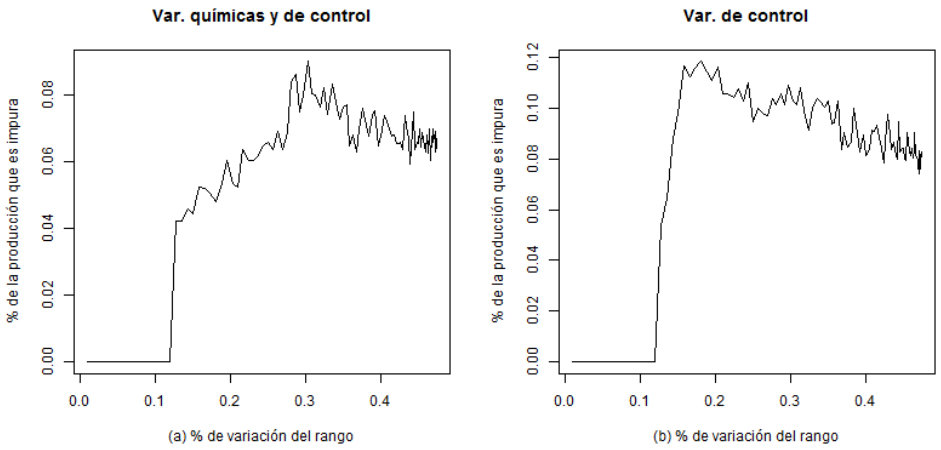


Gráfico 9.3.1: Efecto de la política 2 en la ecuación NaCl_L

Fuente: Elaboración del autor.

El Gráfico 9.3.1 muestra el efecto de esta política sobre la ecuación desde la cual fue definida, es decir, sobre la ecuación $NaCl\_L$. Como se puede ver, con un rango de variación menor al 12% promedio, la impureza simulada no aumenta. También se observa una leve diferencia en el comportamiento de ambos gráficos, pues mientras la figura (b) alcanza rápidamente 12% de impureza al variar los rangos de las variables en un 18% promedio, la figura (a) asciende a 4% de impureza para luego crecer más paulatinamente, alcanzando recién el 8,9% cuando el rango de variación alcanza un 30,3% promedio, lo que da cuenta de que en esta ecuación, la variación de las variables químicas sí contribuye a disminuir los porcentajes de impureza.

De esta forma, la política 2 se definiría como permitir una variación del 12% promedio en las variables de las ecuaciones en torno a los mínimos logrados para la ecuación $NaCl\_L$.

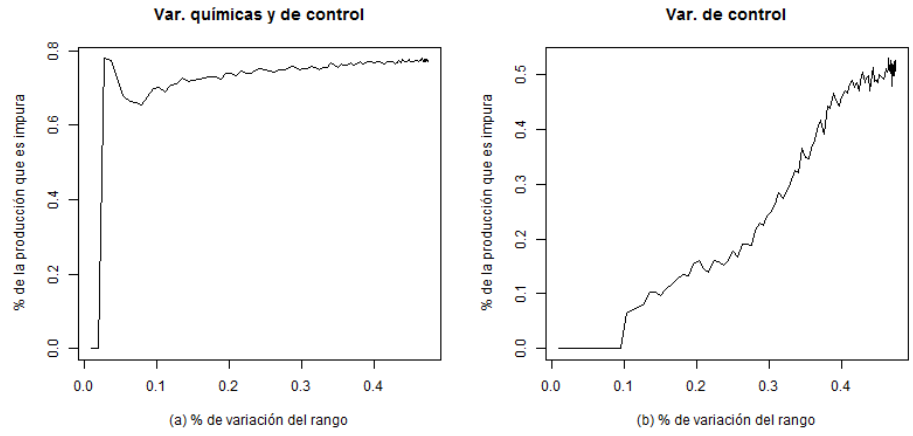


Gráfico 9.3.2: Efecto de la política 2 en la ecuación KClO4_L

Fuente: Elaboración del autor.

El Gráfico 9.3.2 muestra como varía el porcentaje de impureza de la corriente L, dado por el contaminante $KClO_4\_L$, cuando aumenta el porcentaje de variación en torno a los mínimos de las variables de la ecuación $NaCl\_L$. Aquí hay una gran diferencia entre el comportamiento aplicando variaciones a ambas variables versus aplicando sólo en las variables de control. Como se observa, en la figura (a), sólo es tolerado un bajo nivel de variación, entorno a los 1,8% promedio, para no generar impureza en el sistema, luego del cual asciende rápidamente llegando a porcentajes de impureza próximos al 80% con variación de 2,7%. Un escenario totalmente distinto es el descrito en la figura (b), donde a mayores rangos de variación, de 9,5% promedios, el sistema no presenta impurezas. Esto da evidencias de que, en la segunda política, la estabilidad de las variables químicas es relevante para la predicción de pureza del contaminante $KClO_4$.

Con lo anterior, la política 2 aumentaría la concentración del contaminante $KClO_4$ en la corriente L dado que permite variaciones del 12% promedio sobre sus variables. Para lograr evitar este perjuicio, se debe ajustar el rango de variación al 1,8%, lo cual deja poco margen de flexibilidad.

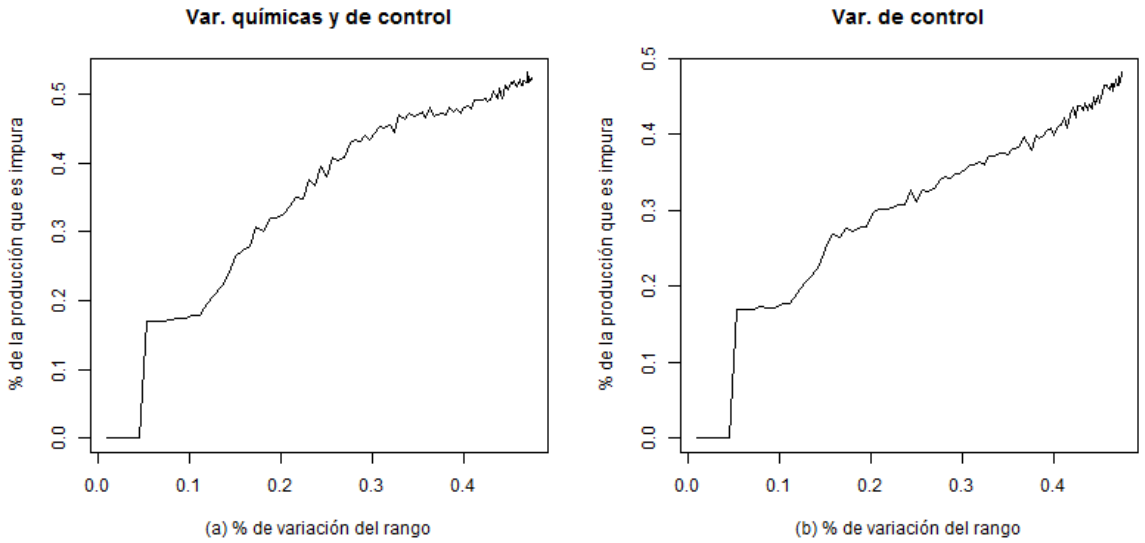


Gráfico 9.3.3: Efecto de la política 2 en la ecuación NaCl

Fuente: Elaboración del autor.

En el Gráfico 9.3.3, muestra como varía el porcentaje de impureza de la corriente M, dado por el contaminante $NaCl$, cuando aumenta el porcentaje de variación en torno a los mínimos de las variables de la ecuación $NaCl\_L$. Como se puede observar, el comportamiento de las curvas es similar, tanto en crecimiento como máximos alcanzados. Hay leves diferencias en la pendiente de las curvas al aumentar los puntos que se consideran en los rangos de variación, pero en general son muy parecidos, lo que permite suponer que la variación de las variables químicas es menos relevante que las variables de control para mejorar los índices de pureza de la producción. En cada figura, mantener los

rangos de variación bajo un 4,5% promedio en cada variable permite no aumentar el nivel de impureza en la producción.

En este caso, la política 2, ajustada con la ecuación $KClO_4\_L$, es decir, con un rango de variación de 1,8%, evita el incremento de la impureza simulada de ambos contaminantes en la corriente L y M. Sin embargo, como se hizo notar anteriormente, este rango es muy restringido si se busca flexibilidad a la hora de operar con las variables del sistema.

9.4 Tercera Política de Rangos

La tercera política de rangos está definida la variación de las variables sobre los puntos mínimos que mejorar la pureza de la corriente M dado por la concentración del contaminante $NaCl$. Primero se analizará el efecto que tiene restringir las variables químicas o no, para así definir los rangos de variación y luego estudiar su impacto en el resto de las ecuaciones.

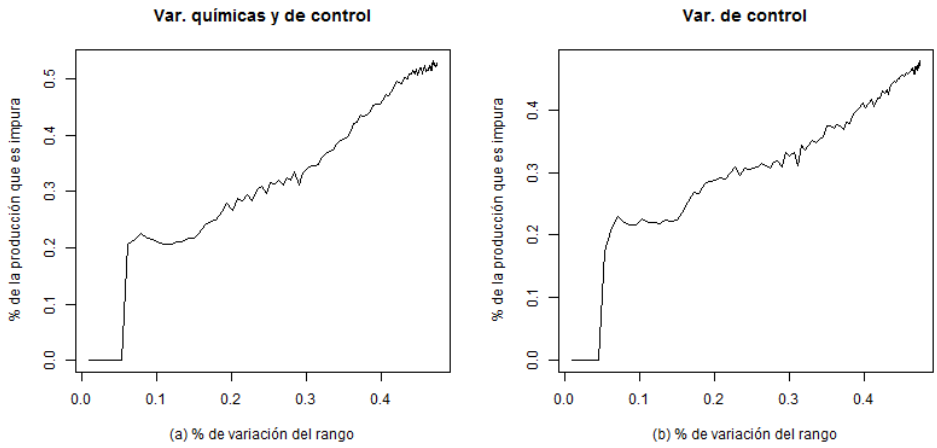


Gráfico 9.4.1: Efecto de la política 3 en la ecuación NaCl_M

Fuente: Elaboración del autor.

El Gráfico 9.4.1 muestra el efecto de esta política sobre la ecuación desde la cual fue definida, es decir, sobre la ecuación $NaCl\_M$. En el contraste de variación de variables químicas y de control versus solo de control se puede observar que el comportamiento es similar, una tendencia creciente, y máximos alcanzados parecidos. La simulación sugiere que bajo un 4,5% promedio de variación de cada variable, la producción impura es nula. De esta forma, la política 3 se puede definir como permitir una variación del 4,5% promedio en las variables de las ecuaciones en torno a los mínimos logrados para la ecuación $NaCl\_M$.

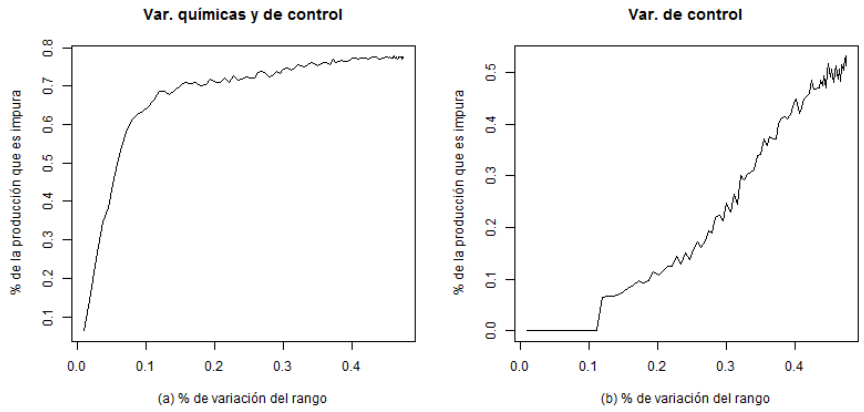


Gráfico 9.4.2: Efecto de la política 3 en la ecuación KClO4_L

Fuente: Elaboración del autor.

El Gráfico 9.4.2 muestra como varía el porcentaje de impureza de la corriente M, dado por el contaminante $KClO_4\_L$, cuando aumenta el porcentaje de variación en torno a los mínimos de las variables de la ecuación $NaCl\_M$. Aquí se observa una diferencia notable cuando se aplica variación a ambos tipos de variables versus sólo del tipo control. En la figura (a) se ve que a medida que aumenta el rango de variación, el crecimiento se dispara, sobrepasando el 70% de impureza con un 15% de variación en las variables. Por otro lado, si sólo se varían las variables de control, manteniendo estables las químicas, se consigue que el porcentaje de impureza no aumente a pesar de superar el 11% en variación de las variables, incluso llegando a un 40% de variación, el porcentaje de impureza alcanzado no supera el 45%.

En esta situación, es difícil definir un rango de variación para no contribuir a la impureza del sistema. Más que nada se puede elegir un rango de acuerdo a no superar el porcentaje de impureza histórico de este contaminante en esta corriente, es decir 10,8%. Con este supuesto, sólo un rango de variación del 1,6% permite no excederlo, lo que daría como siguiente paso ajustar la política 3 a este nuevo rango.

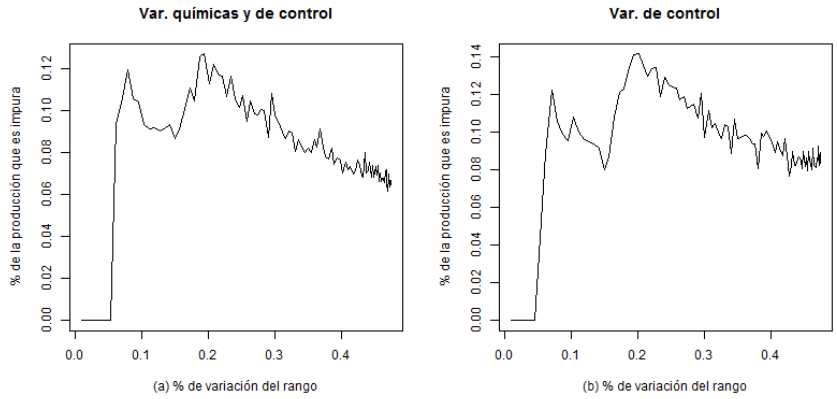


Gráfico 9.4.3: Efecto de la política 3 en la ecuación NaCl _L

Fuente: Elaboración del autor.

El Gráfico 9.4.3 muestra como varía el porcentaje de impureza de la corriente M, dado por el contaminante $NaCl$, cuando aumenta el porcentaje de variación en torno a los mínimos de las variables de la ecuación $NaCl\_M$. Como se puede observar, el comportamiento es similar considerando variaciones en ambos tipos de variables o sólo en las de control, donde a rangos menores a 4,5% de variación, no se experimenta incremento de la impureza simulada. También se observa un decaimiento de las curvas una vez alcanzados sus máximos, lo que hace supones que a mayores rangos de variación aparecen efectos de compensación que reducen la impureza.

En este caso, la política 4, ajustada con la ecuación $KClO_4\_L$, es decir, con un rango de variación de 1,8%, evita el incremento de la impureza simulada de ambos contaminantes en la corriente L y M. Sin embargo, como se hizo notar anteriormente, este rango es muy restringido si se busca flexibilidad a la hora de operar con las variables del sistema.

9.5 Mejoras en la producción

En esta sección se analiza el impacto que tiene la implementación del modelamiento a la producción de *La Empresa*, es decir, la contribución que hace mejorando la predicción de la producción impura.

Corriente	Promedio	Moda	Mínimo	Máximo	Desviación Estándar
L	422.86	269	11	861	158.51
M	381.83	492	20	1111	145.5

Tabla 9.5.1: Estadística descriptiva de las corrientes entre 2006 y 2012 (unidades en Toneladas)

Fuente: Elaboración del autor.

El tonelaje de la producción de *La Empresa* entre los años 2006 y 2012 está caracterizada en la Tabla 9.5.1. Ahí se puede apreciar el amplio rango de variación que ha tenido la producción entre el periodo de estudio (2006 al 2012), donde la corriente L ha producido en promedio mayores cantidades de material que la corriente M a pesar que esta última posee una moda de producción mayor a la primera. La evolución de estas condiciones queda más en evidencia en el Gráfico 9.5.1: Producción histórica por turnos donde la producción impura representa el 11,9% y el 28,6% de la producción de las corrientes L y M respectivamente. Como se puede observar, en la primera corriente, la producción no aprovechable mayormente ocurrió entre los turnos 750 y 1250, luego de los cuales, la producción es en su mayoría pura. Es caso contrario se aprecia en la corriente M, donde las producciones impuras se distribuyen en todo el historial de producción.

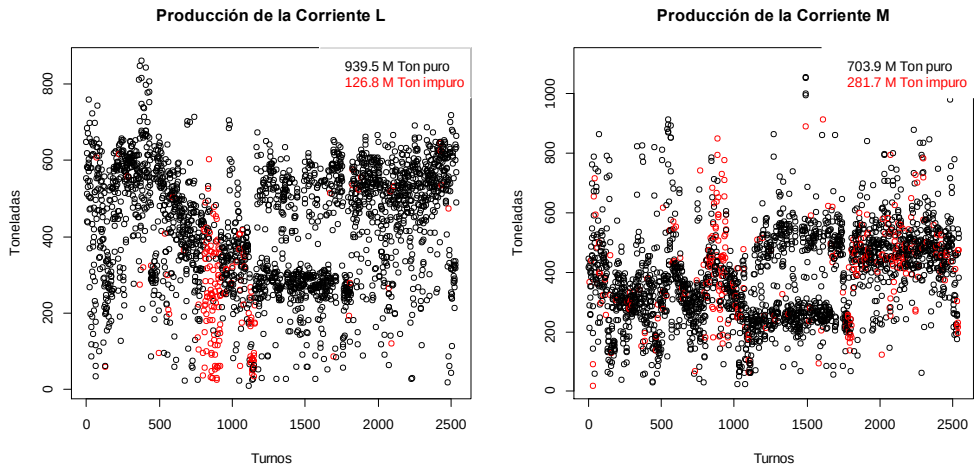


Gráfico 9.5.1: Producción histórica por turnos

Fuente: Elaboración del autor.

En los capítulos anteriores se describieron 4 ecuaciones, dos para cada corriente, clasificando a cada una en su nivel de pureza tanto para el $KClO_4$ como para el $NaCl$. En esta sección, se conjugan las ecuaciones de tal forma que, para cada corriente, la producción se cataloga como pura si cumple los límites de pureza para cada contaminante a la vez, e impura en caso contrario. Así, por ejemplo, si una producción proveniente de la corriente L se catalogó pura bajo el $KClO_4$ (concentración del contaminante inferior o igual a los 0,24% p/p) e impura bajo el $NaCl$ (concentración del contaminante superior a los 0,95% p/p), esa producción cae en la categoría "impura".

Se procede a analizar el contraste entre el uso del modelamiento para predecir la impureza versus la data histórica, para estudiar si efectivamente su aplicación contribuye o no a mejorar los niveles productivos de *La Empresa*. Para esto se clasifica la producción usando la predicción de los modelos y se analiza el tonelaje en su categoría real, lo cual queda descrito en el Diagrama 9.5.1 y el Diagrama 9.5.2.

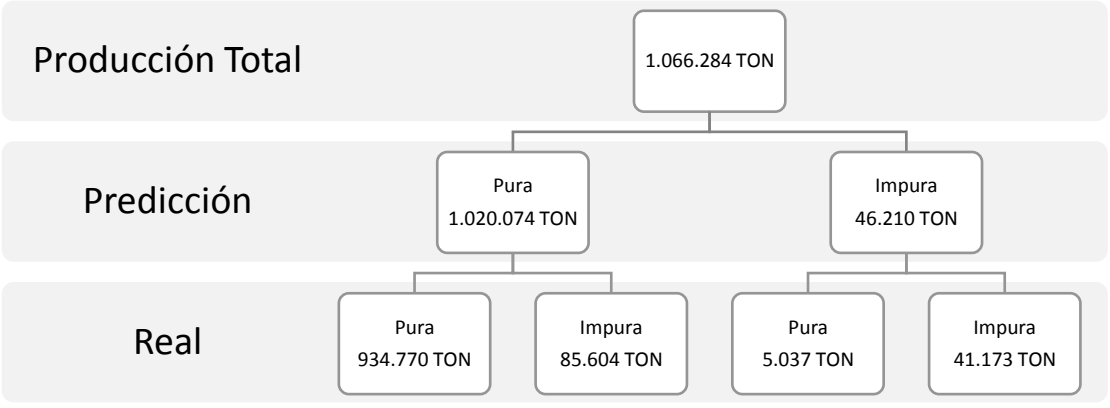


Diagrama 9.5.1: Impacto en la corriente L que tendría el modelo predictivo sobre el periodo de estudio

Fuente: Elaboración del autor.

En el Diagrama 9.5.1 se puede observar que la predicción sobreestima el tonelaje de la categoría puro y subestima el tonelaje de la categoría impuro en la corriente L, dado que en la data estos corresponden a 939.507 y 126.777 toneladas respectivamente. Sin embargo, vemos que la precisión del modelo es buena, puesto que de las categorías predichas, yerra un 8% en la categoría pura y un 12% en la impura. Con esto, se podría reaprovechar 41.173 toneladas de material dado que se puede predecir que será impuro, pero a la vez se pierden 90.641 toneladas de material que los modelos mal clasifican.

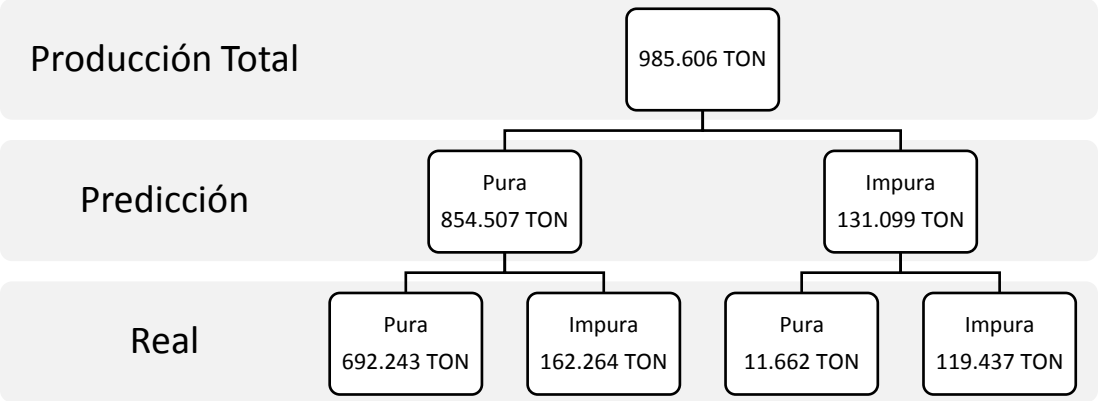


Diagrama 9.5.2: Impacto en la corriente M que tendría el modelo predictivo sobre el periodo de estudio

Fuente: Elaboración del autor.

En el Diagrama 9.5.2 se puede observar que la predicción, a diferencia de la corriente L subestima el tonelaje de la categoría puro y sobreestima el tonelaje de la categoría impuro en la corriente M, dado que en la data estos corresponden a 703.905 y 281.701 toneladas respectivamente. Se ve que la precisión del modelo es buena en el caso de la categoría impura, donde alcanza un 91% de clasificación correcta, mientras que la precisión para la categoría pura es inferior, llegando a un 81% de bien clasificados. Con esto, se podría reaprovechar 119.437 toneladas de material dado que se puede predecir que será impuro, pero a la vez se pierden 162.264 toneladas de material que los modelos mal clasifican, generando que la implementación en este esta corriente presenta mayores pérdidas que beneficios.

Durante el periodo de la información, el precio del nitrato de potasio fue en promedio de US$ 651,16 por tonelada, con esto se elabora la Tabla 9.5.2 que contrasta la producción de cada corriente en dólares, de sus valores reales como de sus predicciones. Para los valores reales, se muestra la ganancia que genera por la cantidad de material puro producido, la pérdida de material por la cantidad material impuro producido. Para los valores predichos, se muestra la ganancia por la producción de material puro que es correctamente clasificado, la perdida de material por las predicciones de ambas categorías incorrectamente clasificadas y el material impuro que es posible de recuperar a través de la predicción.

Corriente	Valor	Producción	Ganancia	Pérdida	Recuperación
L	Predicción	$ 694	$ 608	$ 59	$ 27
	Real	$ 694	$ 612	$ 83	
M	Predicción	$ 642	$ 451	$ 113	$ 78
	Real	$ 642	$ 547	$ 95	

Tabla 9.5.2: Comparación económica (unidades en millones de dólares)

Fuente: Elaboración del autor.

Como se puede observar, la predicción en la corriente L permite aprovechar US$ 27 millones en material impuro, disminuyendo los niveles de material perdido en US$ 24 millones. Sin embargo, en la corriente M, el monto en material recuperable es inferior al monto en material perdido, US$ 78 millones versus US$ 113 millones, lo que hace que implementar el modelamiento para esta corriente no favorable para el resultado económico de *La Empresa.* De esta forma, sólo es recomendable el control de la corriente L.

No se desarrolla un estudio por simulación de las predicciones económicas del proceso, dado que los modelos sólo predicen grados de concentración de contaminantes y no de toneladas. Se deja abierto una línea de investigación para predecir el nivel de producción pura de la planta dado los modelos de pureza desarrollados en esta memoria.

10 CONCLUSIONES

Del objetivo principal de reducir el porcentaje de impureza del proceso, dado por la presencia de los contaminante $KClO_4$ y $NaCl$, se lograron resultados satisfactorios que dan cuenta de la posibilidad de gestión de políticas de operación para mejorar la calidad del producto. De partida, al modelar los dos contaminantes en las dos corrientes de salida del proceso con las mismas variables, se facilita la comparación de los efectos en el nivel de pureza general.

Para modelar el proceso se usaron dos tipos de modelos, de clasificación y regresión, en cada cual se aplicaron 3 algoritmos distintos, siendo uno en cada caso un algoritmo estándar de clasificación y regresión que servía como benchmarking de los otros. Para los modelos del tipo clasificación, se usó la regresión logística como benchmarking y se comparó con Random Forest (RF) y Support Vector Machine (SVM). Como era de esperarse, sus rendimientos fueron superiores a la regresión logística, donde destacó SVM como mejor ajuste en cada ecuación. Por el lado de los modelos regresivos se usó como benchmarking una regresión lineal simple en contraste con Random Forest Regression (RFR) y Support Vector Regression (SVR). Ambos últimos superaron el benchmarking, destacando a RFR como mejor ajuste para las ecuaciones 1, 2 y 4, mientras SVM se ajustó mejor en la tercera.

Sorprendentemente, los modelos encontrados que dieron mejor explicación al proceso fueron del tipo regresivos, siendo más lógico pensar que modelos del tipo clasificación eran más eficientes a la hora de trabajar con las categorías puro e impuro. Una de las razones por las cuales se dio este contraste fue que los modelos de regresión resultaron tener una mejor precisión a la hora de predecir los resultados de cada contaminante en cada corrientes, es decir, que dado la clasificación del material producido en base a las predicciones de concentraciones de contaminantes, se observó que los modelos regresivos tenían un menor porcentaje de elementos mal clasificados catalogadas que los logrados por los modelos de clasificación. Estos últimos solo tenían la ventaja de permitir capturar una muestra más representativa de la categoría impuro en sus clasificaciones, pero el nivel de precisión hacia que con poca certeza se asegurara que la categoría iba a ser cierta, así, por ejemplo, de la ecuación $NaCl\_L$, de las 860 observaciones impuras que se dieron en la realidad, 554 lograda catalogar correctamente, pero 306 lo hacía incorrectamente. Los modelos regresivos, por su parte, no lograban capturar una gran cantidad de observaciones impuras, pero de las que capturaban su nivel de certeza era mucho mayor, así, por ejemplo de las 41 observaciones que catalogó como impuras en la ecuación $NaCl\_L$, sólo 1 fue mal clasificada. De esta forma, los modelos regresivos favorecen más el control de los niveles de impureza al equivocarse menos que los modelos de clasificación, logrando, a través de simulación, disminuir los porcentajes de impureza histórica, desde 11% a 5% para el contaminante $KClO_4$ en la corriente L, desde un 5% a 4% para el contaminante $NaCl$ en la corriente L, desde 9% a 8% para el contaminante $KClO_4$ en la corriente M y desde un 18% a un 10% para el contaminante $NaCl$ en la corriente M.

También se estudió el efecto que tenían las variables en la predicción de impurezas, buscando detectar las 10 más relevantes, de forma tal de comprender qué plantas del proceso y que tipo de variable, química o de control, explica en gran medida el resultado. De los modelos de clasificación se encontró que las 4 ecuaciones presentaban 6 variables de tipo control dentro de las 10 más relevantes, lo que posibilitaba al sistema mayor manejo de los resultados. Para los contaminantes de la corriente L, las plantas que mayormente impactaba en la pureza del material correspondían a la planta de tratamiento de sales para el perclorato de potasio, y la planta de cristalización para el cloruro de sodio, dando a entender que las condiciones iniciales y finales del sistema eran más decisivas en los resultados. Caso contrario se dio para la corriente M, donde la planta más relevante fue la Dual, pues depende más de las corrientes de descarte que aquí se generen para limpiar el flujo reprocesado. Por el lado de los modelos de regresión se observó mayor cantidad de variables de control presente en las 10 variables más relevantes, excepto en la primera ecuación donde ocurrió que 5 variables eran químicas y 5 de control, pero explicando el mismo principio, la posibilidad de gestión sobre los resultados. Sin embargo, no hubo coincidencia en la relevancia de la planta de procedencia, pues mientras en la corriente M, los modelos de clasificación eran explicados por la planta de Dual, los modelos de regresión eran explicados mayormente por la planta de tratamiento de sales. Esta diferencia puede deberse a que, como los modelos de clasificación de dos categorías no permiten ver en detalle cuán puro o impura es la muestra clasificada, a nivel general, definiendo la procedencia con este tipo de modelo sea una primera aproximación, pero a nivel micro, los modelos de regresión otorguen mayor sensibilidad con el tipo de categoría y la importancia de la planta que ellos predicen sea más exacta.

Otro aspecto relevante que se pudo concluir es la posibilidad de relajar las restricciones que los modelos imponen a las variables del proceso para lograr los resultados de pureza esperada. Permitir mayor variación en los rangos de valores que estas pueden tomar, facilita a *La Empresa* a tomar decisiones más flexibles a la hora de producir, permitiendo adaptarse a eventualidades que no puedan ser predichas, como la disminución del caliche extraído, o la necesidad de producir otro producto químico. Para esto se evaluó 3 tipos de políticas de rangos dadas por las condiciones donde las ecuaciones 1, 2 y 4 se optimizaban, es decir, las variables lograban mínimos que favorecían a cada una. Una vez definidas las políticas, se realizó un análisis cruzado para ver el impacto de lo que sucedía con el resto de las ecuaciones cuando mejoraba una en particular. De la primera política se obtuvo que una variación hasta el 4,5% promedio en cada variable sobre el mínimo de la ecuación $KClO_4$ mejoraba también la el porcentaje de producción pura en las otras ecuaciones, mientras que en la segunda y tercera política, el rango de variación permitido era menor, un 1,8%, dejando menos flexibilidad a *La Empresa* para operar y adaptarse.

Por último se estudió como afectaba el modelamiento al tonelaje de producción. Para ello, se usó la data histórica de las toneladas producidas para ver cuánto se hubiera ganado o perdido operando con las predicciones que entregaban los modelos. Se pudo observar que, en la corriente L, de los 1,066 millones de toneladas producidas entre los años 2006 y 2012, el material perdido se podía reducir de 126 mil toneladas a 90 toneladas y reaprovechar 41

mil toneladas en el proceso para extraer de él algún beneficio. No así para la corriente M, donde de las 985 mil toneladas producidas en el mismo periodo de tiempo, las toneladas perdidas aumentaban desde 145 mil toneladas a 173 mil toneladas, pero pudiendo ser reaprovechadas 113 mil toneladas. Si se ve en términos económicos, de los $694 millones de dólares que significa la producción en la corriente L, con ganancias de $612 millones de dólares, estás se reducen en $4 millones por la predicción pero permiten tener una promesa de $27 millones en material reaprovechable. Mientras que, de los $642 millones de dólares que significa la producción en la corriente M, con ganancias de $547 millones de dólares, estas se reducen en $96 millones para permitir tener una promesa de 78 millones en material reaprovechable, cosa que no es beneficioso. Así, lo es aconsejable modelar la producción de la corriente L.

El trabajo presente mostró complicaciones por la calidad de la información con la que se contaba, puesto las variables que se tenían en un principio para hacer este estudio presentaron grandes vacíos de información que limitaron la gama de variables a utilizar. De las 310 variables que se manejaban, se tuvo que reducir a 68 por existencia de historia incompleta, además, los valores faltantes de las 68 variables se imputaron mediante regresiones lineales simples, lo que debilita la bondad del ajuste que los modelos tengan con la realidad. Por lo mismo, es recomendable hacer un trabajo de recopilación de información previo que contemple asegurar la calidad de la información para la mayoría de las variables, o en caso de no poder realizarse para todas, que asegura un cierto nivel de completitud por lo menos para las variables de entrada al sistema, debido a que es crucial poder contar con la información de la calidad del material que entra para poder predecir mejor la calidad del material que sale.

Como aspectos importantes respecto a la operativización de estas políticas de operación para *La Empresa,* la presente memoria muestra las variables, los valores óptimos de cada una y los rangos de variación en el que puede oscilar cada una. La propuesta contempla que se establezca esta configuración en las 4 plantas para mejorar la producción. También, en vista que a nivel logístico, mantener 66 variables, o por lo menos las 43 variables de control puede ser muy complejo, se deja a disposición el análisis hecho en la sección *8.2 Importancia de variables*, donde se focaliza el nivel de atención en 10 variables por planta que permiten mejorar también el nivel de pureza de la producción.

Finalmente, es bueno recalcar dos líneas de investigación que pueden abrirse en base a esta memoria para futuros trabajos o proyectos. Como primera recomendación, recordar que como este trabajo aborda la problemática desde la ingeniería civil industrial, es aconsejable su complementación con la ingeniería química de procesos industriales para estudiar el impacto que tendrías las variables de control y sus recomendaciones a nivel de la reacciones químicas involucradas, de tal forma de poder mejorar la precisión de los modelos para la predicción de producción impura. Una segunda recomendación es el modelamiento híbrido que se enuncia al final del capítulo 7, puesto que general dos modelamientos independientes, de clasificación y de regresión, pero cabe preguntarse si es posible modelarlos simultáneamente. Como se observó en dicho capítulo, cada tipo de modelamiento tiene sus ventajas y desventajas, por lo que un modelo híbrido, que modele

el proceso con una regresión a la que después se le aplica una clasificación, podría rescatar los beneficios de ambos modelos, por lo que es interesante y recomendable su investigación más acabada.

BIBLIOGRAFÍA

Box, G. E., & Jenkins, G. M. (1973). Some comments on a paper by Chatfield and Prothero and on a review by Kendall. *Journal of the Royal Statistical Society. Series A (General)*, 337-352.

Breiman, L. (2001). *Random Forests.* University of California.

Breiman, L. (16 de Noviembre de 2014). *Classification/Clustering*. Obtenido de Random Forests: Leo Breiman and Adele Cutler: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

Cheung, Y. W., & Lai, K. S. (1995). Lag order and critical values of the augmented Dickey–Fuller test. *Journal of Business & Economic Statistics, 13(3)*, 277-280.

Chisholm, A. (2013). *Exploring Data With RapidMiner.* West Midlands: Packt Publishing Ltd.

Comisión Chilena del Cobre. (2013). *Monitoreo de los minerales industriales de Chile - Análisis de los Recursos Salinos.* Santiago.

Competitividad chilena en los recursos salinos. (1 de Agosto de 2012). *Minería Chilena*.

Gales, V. U. (2009). *Detección de interacciones genéticas asociadas a enfermedades complejas. Aplicación al cáncer de vejiga.* Cataluña: Tesis de Magister.

Garcés Millas, I. (2000). *Nitrato de Potasio.* Antofagasta.

Hurtado, C., & Ríos, G. (2008). *Series de Tiempo.* Universidad de Chile, Ciencias de la Computación. Santiago: Facultad de Ciencias Físicas y Matemáticas.

Maimon, O., & Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook.* New York: Springer.

Malhotra, N. (1987). Analyzing marketing research data with incomplete information on the dependent variable. *Journal of Marketing Research 24*, 74-84.

Oficina de Estudios y Políticas Agrarias. (2010). *Estudio de Diagnóstico de Mercado y Estudio de la Cadena.* Santiago.

Potassium Nitrate Association. (22 de Junio de 2010). *Acerca del Nitrato de Potasio*. Obtenido de PNA Potassium Nitrate Association: http://www.kno3.org

Ruczinski, I., Kooperberg, C., & LeBlanc, M. (2003). Logic Regression. *Journal of the Computacional and Graphical Statistics, 12*, 475-511.

Servicio Nacional de Geología y Minería. (2013). *Anuario de la Minería de Chile.* Santiago.

Tsikriktsis, N. (2005). A review of techniques for treating missing data. *Journal of Operations Management*.

Useche Castro, L. M., & Mesa Ávila, D. M. (2006). Una Introducción a la Imputación de Valores Perdidos. *Tierra Nueva Etapa, XXII*(31), 127-151.

Vapnik, V., & Cortes, C. (1995). Support vector networks. Machine Learning. *Journal Machine Learning, Volume 20 Issue 3*, 273-297.

ANEXOS

ANEXO A.

DESCRIPCIÓN ESTADÍSTICA	Corriente L		Corriente M	
	KClO4	NaCl	KClO4	NaCl
Media	0,14	0,65	0,62	0,75
Moda	0,1	1	1	1
Desviación Estándar	0,12	0,27	0,29	0,42
Mínimo	0,01	0,1	0,05	0,09
Máximo	1,4	9,3	2,2	7,8
Pureza Histórica	15,92%	13,29%	17,81%	31,66%

CORRIENTE L							
KClO4	2006	2007	2008	2009	2010	2011	2012
Media	0,130	0,107	0,126	0,165	0,178	0,114	0,119
Moda	0,05	0,05	0,05	0,05	0,05	0,05	0,05
Desviación Estándar %	70%	74%	68%	94%	89%	76%	72%
Mínimo	0,005	0,05	0,05	0,05	0,05	0,05	0,05
Máximo	1,3	0,92	0,96	1,4	1,4	1,3	0,77
Límite	0,24	0,24	0,24	0,24	0,24	0,24	0,24
NaCl	2006	2007	2008	2009	2010	2011	2012
Media	0,701	0,736	0,689	0,589	0,536	0,668	0,640
Moda	0,69	0,68	0,72	1	0,31	1	1
Desviación Estándar %	26%	26%	30%	52%	61%	39%	44%
Mínimo	0,096	0,19	0,15	0,13	0,14	0,17	0,1
Máximo	2,2	1,8	2,6	6,2	9,3	2,2	2,2
Límite	0,95	0,95	0,95	0,95	0,95	0,95	0,95

CORRIENTE M							
KClO4	2006	2007	2008	2009	2010	2011	2012
Media	0,731	0,710	0,707	0,709	0,674	0,406	0,377
Moda	1	1	1,1	1	1,1	0,34	0,05
Desviación Estándar %	30%	33%	37%	35%	52%	51%	61%
Mínimo	0,05	0,05	0,05	0,05	0,05	0,05	0,05
Máximo	1,6	1,5	1,6	1,7	2,2	1,6	1,9
Límite	0,95	0,95	0,95	0,95	0,95	0,95	0,95
NaCl	2006	2007	2008	2009	2010	2011	2012
Media	0,781	0,871	0,748	0,705	0,698	0,697	0,776
Moda	1	1	1,1	1	1	1	1
Desviación Estándar %	65%	51%	38%	56%	62%	57%	55%
Mínimo	0,11	0,17	0,13	0,15	0,11	0,11	0,09
Máximo	6,7	7,8	5,7	7,6	7,7	4	3,1
Límite	0,95	0,95	0,95	0,95	0,95	0,95	0,95

ANEXO B.

Tipo de corriente	Corriente	Planta	Frecuencia Actual
Entradas Materias Primas a PTS	376124 (Sal NV)	PTS	Turno
	376800 (Silv.SinAgrPTS)	PTS	Turno
	376801 (Silv.SinRecPTS)	PTS	Turno
	376609 (Silv.RecPTS)	PTS	Día
	376608 (Agregado KCl)	PTS	Día
	376610 (Cont.No Apto PTS)	PTS	Día
	376604 (Granza Agregada PTS)	PTS	Día
	376806(Agregado CNPC-L1)	PTS	Día
	376803(Agregado Murito Salar)	PTS	Día
Salidas PTS	377147 (Sol.SalidaPTS)	PTS	Turno

Tipo de corriente	Corriente	Planta	Frecuencia Actual
	377154 (Sal Rechazo)	PTS	Turno
Entradas Materias Primas a Muriato	350128 (Poza 10)	Muriato	Turno
	350105(Purgas NPT III a Muri)	Muriato	Turno
	350802 (Muriato Agregado)	Muriato	Día
	350843(CSSA)	Muriato	Día
	350607(MOP-H BL)	Muriato	Día
	350803(KCl Minsal)	Muriato	Día
	350814(Granza KNO3)	Muriato	Día
Salidas Muriato	350004 (Sol. Colector Tronas)	Muriato	Cada 3 hrs
	350001 (Agua Riñihue)	Muriato	Turno
Salidas Dual	363106 (Rebalse Espesador)	Dual	Cada 3 hrs
	363420 (Rechazo Dual)	Dual	Día
	364001 (Lavado Delkor)	Dual	Día
Entrada Cristal	334363 (SSK D)	Cristal	Cada 12hrs
Salidas Cristal	335162 (MLK)	Cristal	Cada 3 hrs
	335681 (CNPC-L)	Cristal	Bihoraria
			Turno
	335688 (CNPC-M)	Cristal	Bihoraria
			Turno
Recirculación	335170 (MLR)	Cristal	

ANEXO C.

Variable	Grupo	Unidad	Periodo	PLANTA
T. Entrada a Planta P.T.S.	Temperaturas	C	TURNO	PTS
T. Reactor 1	Temperaturas	C	TURNO	PTS
T. Reactor 2	Temperaturas	C	TURNO	PTS
T. Reactor 3	Temperaturas	C	TURNO	PTS
T° Retorno Planta P.T.S.	Temperaturas	C	HORA	PTS
Volumen a Molino PTS	Flujos Planta	M3	TURNO	PTS
Volumen a TK Acondicionador	Flujos Planta	M3	TURNO	PTS
Volumen Ingreso a Planta	Flujos Planta	M3	TURNO	PTS
Volumen Retorno Planta	Flujos Planta	M3	TURNO	PTS

Variable	Grupo	Unidad	Periodo	PLANTA
Agua Centrifuga	Flujos Planta	Lt	TURNO	PTS
Agregado Granza KNO3	Sales Agregadas	TON	TURNO	PTS
% K en Granza	Sales Agregadas	%	DÍA	PTS
Sales Agregadas Cosecha Pozas	Sales Agregadas	TON	TURNO	PTS
Sales Agregadas Pampa Blanca AL	Sales Agregadas	TON	TURNO	PTS
Sales Rechazadas	Sales Agregadas	TON	TURNO	PTS
Sodico Recibido	Sales Agregadas	TON	TURNO	PTS
Contaminado Mezclas PTS	Sales Agregadas	TON	TURNO	PTS
Agregados Varios	Sales Agregadas	TON	TURNO	PTS
NaNO3 de Cosecha Pozas	Sales Agregadas	Gpl	TURNO	PTS
Agregado Costras PV	Sales Agregadas	TON	TURNO	PTS
% Sólidos Overflow	% Solidos Overflow	%	HORA	PTS
Lavado PTS a Planta Dual	Medidores M3	M3	TURNO	PTS
Muriato a Planta PTS	Medidores M3	M3	TURNO	PTS
Desplazamiento	Sales Agregadas	TON	TURNO	PTS
Microprill	Sales Agregadas	TON	TURNO	PTS
Ceniza de Soda Agregada	Sales Agregadas	TON	TURNO	PTS
Ceniza de Soda Recibida	Sales Agregadas	TON	TURNO	PTS
Muriato Salar Bajo Ca	Sales Agregadas	TON	TURNO	PTS
Halita	Sales Agregadas	TON	TURNO	PTS
Silvinita	Sales Agregadas	TON	TURNO	PTS
Muriato Salar Alto Magnesio	Sales Agregadas	TON	TURNO	PTS
Muriato Litio Reciclo	Sales Agregadas	TON	TURNO	PTS
Despacho Silvinita a NPT	Sales Agregadas	TON	TURNO	PTS
Agregado Sal NV	Sales Agregadas	TON	TURNO	PTS

Variable	Grupo	Unidad	Periodo	PLANTA
Agregado Granza KCl	Sales Agregadas	TON	TURNO	PTS
Agregado Mezclas NK	Sales Agregadas	TON	TURNO	PTS
Sales Agregadas Pampa Blanca BL	Sales Agregadas	TON	TURNO	PTS
KCL + Silvinita	Sales Agregadas	TON	TURNO	PTS
CSSA F/E PV	Sales Agregadas	TON	HORA	PTS
Prill SSI - Devolución Tocopilla	Sales Agregadas	TON	HORA	PTS
SSI-C - Devolución Tocopilla	Sales Agregadas	TON	HORA	PTS
Contaminado No Apto	Sales Agregadas	TON	HORA	PTS
T° Entrada Agua APV6	Temperaturas	°C	HORA	MUR
T° Entrada Solución APV4	Temperaturas	°C	HORA	MUR
T° Salida Solución APV6	Temperaturas	°C	HORA	MUR
T° Entrada Solución APV8	Temperaturas	°C	HORA	MUR
T° Entrada Solución APV9	Temperaturas	°C	HORA	MUR
T° Entrada Trona 1	Temperaturas	°C	HORA	MUR
T° Entrada Trona 3	Temperaturas	°C	HORA	MUR
T° Salida Trona 5	Temperaturas	°C	HORA	MUR
T° Entada Trona 6	Temperaturas	°C	HORA	MUR
T° Salida Trona 8	Temperaturas	°C	HORA	MUR
T° Rebalse Espesador Dual	Temperaturas	°C	HORA	MUR
KCl Agregado Minsal Bajo Cal.	Entradas	TON	TURNO	MUR
KCl Agregado Litio de Reciclo	Entradas	TON	TURNO	MUR
KCl (SOP) Bajo Ca Especial	Entradas	TON	TURNO	MUR
Granza KNO3 Secado KNO3	Entradas	TON	HORA	MUR
Granza KCl de Secado Muriato	Entradas	TON	TURNO	MUR
Granza KNO3 de Granulación	Entradas	TON	HORA	MUR

Variable	Grupo	Unidad	Periodo	PLANTA
Raspado Cancha+NPT KNO3	Entradas	TON	TURNO	MUR
Costras Agregadas	Entradas	TON	TURNO	MUR
Poza 10 a Muriato	Soluciones (Flowmeter)	M3	TURNO	MUR
Agua Dulce	Soluciones (Flowmeter)	M3	TURNO	MUR
Entrada Agua Salada	Soluciones (Flowmeter)	M3	TURNO	MUR
Salida Agua Salada	Soluciones (Flowmeter)	M3	TURNO	MUR
Solución Alimentación Planta	Soluciones (Flowmeter)	M3	HORA	MUR
Purga NPT	Soluciones (Flowmeter)	M3	TURNO	MUR
Linea de Fierro a Planta Cristal	Soluciones (Flowmeter)	M3	TURNO	MUR
NaNO3 (Gpl) (c/2hrs)	Análisis Bi-Horarias	Gpl	HORA	MUR
K (Gpl) (c/2hrs)	Análisis Bi-Horarias	Hr	HORA	MUR
Romana N° 1	Sólidos (Flowmeter)	TON	TURNO	MUR
Romana N° 2	Sólidos (Flowmeter)	TON	TURNO	MUR
Romana N° 3	Sólidos (Flowmeter)	TON	TURNO	MUR
K Objetivo en SSK	Control SSK	Gpl	TURNO	MUR
T° Entrada Solución APV5	Temperaturas	°C	HORA	MUR
T° Entrada Solución APV6	Temperaturas	°C	HORA	MUR
T° Salida Solución APV7	Temperaturas	°C	HORA	MUR
T° Salida Agua APV4	Temperaturas	°C	HORA	MUR
Amperaje Espesador Dual	Temperaturas	AMP	HORA	MUR
T° Entrada Espesador Dual	Temperaturas	°C	HORA	MUR
Pril SPO	Entradas	TON	TURNO	MUR
Sodico Contaminado ME	Entradas	TON	TURNO	MUR
Contaminado NPK Tocopilla	Entradas	TON	TURNO	MUR
Contaminado K-2	Entradas	TON	TURNO	MUR
Producto F/E Cancha 4	Entradas	TON	TURNO	MUR

Variable	Grupo	Unidad	Periodo	PLANTA
Desplazamiento	Entradas	TON	TURNO	MUR
Muriato Mezcla Reciclo	Entradas	TON	TURNO	MUR
Arrastre de K en Muriato Litio	Variables de Arrastre (No ingresar)	%	DÍA	MUR
Intercambiador Tubular	Temperaturas	°C	HORA	MUR
Pozo Lavado Maquinas a TK Brine	Soluciones (Flowmeter)	M3	TURNO	MUR
Linea Pecc a Plta Cristalización	Soluciones (Flowmeter)	M3	TURNO	MUR
T° Entrada Agua Caliente Alfa 1	Temperaturas	°C	HORA	MUR
T° Entrada Solución Alfa 1	Temperaturas	°C	HORA	MUR
T° Salida Solución Alfa 1	Temperaturas	°C	HORA	MUR
T° Entrada Agua Caliente Alfa 2	Temperaturas	°C	HORA	MUR
T° Entrada Solución Alfa 2	Temperaturas	°C	HORA	MUR
T° Salida Solución Alfa 2	Temperaturas	°C	HORA	MUR
T° Entrada Solución Alfa 3	Temperaturas	°C	HORA	MUR
T° Salida Solución Alfa 3	Temperaturas	°C	HORA	MUR
Agua Dilución	Soluciones (Flowmeter)	M3	TURNO	MUR
Poza 10 a Pts	Soluciones (Flowmeter)	M3	TURNO	MUR
Poza 10 a TK 700	Soluciones (Flowmeter)	M3	HORA	MUR
MOP-S Devolución Tocopilla	Entradas	TON	HORA	MUR
MOP-H-SS	Entradas	TON	HORA	MUR
Contaminado No Apto	Entradas	TON	TURNO	MUR
Densidad Pulpa	Parámetros DELKOR	GC3	HORA	DUAL
Espesor Cake	Parámetros DELKOR	Plg	HORA	DUAL
Temp. Agua Lavado	Parámetros DELKOR	C	HORA	DUAL

Variable	Grupo	Unidad	Periodo	PLANTA
Temp. Pulpa	Parámetros DELKOR	C	HORA	DUAL
Velocidad Cinta Filtro	Parámetros DELKOR	%	HORA	DUAL
Agua Entrada Plana	Agua General Planta y Reñihue	M3	TURNO	DUAL
Agua Lavado Cake	Mediciones DELKOR	M3	TURNO	DUAL
Agua Bba. Vacío	Mediciones DELKOR	M3	TURNO	DUAL
Alimentación Pulpa	Mediciones DELKOR	M3	TURNO	DUAL
Total Agua Riñihue	Agua General Planta y Reñihue	M3	HORA	DUAL
Solución Poza 10 a Espesador	Mediciones DELKOR	M3	TURNO	DUAL
Presión de Vacio	Parámetros DELKOR	MMHG	HORA	DUAL
APV. en Servicio	Mediciones por Turno	CAN	HORA	CRIST
APV. en Lavado	Mediciones por Turno	CAN	HORA	CRIST
Volumen Alim. SSK	Mediciones por Turno	M3	HORA	CRIST
Purga Proceso Pta. NPT	Mediciones por Turno	M3	TURNO	CRIST
T° Entr. Alim. SSK	Mediciones por Turno	C	HORA	CRIST
T° Salida Alim. SSK	Mediciones por Turno	C	HORA	CRIST
T° Canales 6/7	Mediciones por Turno	C	HORA	CRIST
T° Entrada Esp. 2	Mediciones por Turno	C	HORA	CRIST
T° Salida Esp. 2	Mediciones por Turno	C	HORA	CRIST
T° Entrada Esp. 1	Mediciones por Turno	C	HORA	CRIST
T° Salida Esp. 1	Mediciones por Turno	C	HORA	CRIST
T° Ent. APV. MLK	Mediciones por Turno	C	HORA	CRIST
T° Sal. MLK	Mediciones por Turno	C	HORA	CRIST
Hrs. Operc. M.L.T.1	Horas Operación	Hr	TURNO	CRIST
Hrs. Operc. M.L.T.2	Horas Operación	Hr	TURNO	CRIST

Variable	Grupo	Unidad	Periodo	PLANTA
Soluc. Tratada SS N° 1	Medidores M3	M3	TURNO	CRIST
Soluc. Tratada SS N° 2	Medidores M3	M3	TURNO	CRIST
Agua en Centrif.	Medidores M3	M3	TURNO	CRIST
Agua en Lavado TKs.	Medidores M3	M3	TURNO	CRIST
Med. Mangueras Aseo Pta.	Medidores M3	M3	TURNO	CRIST
Centrifuga Roberts N° 1 a 5	Medidores M3	M3	HORA	CRIST
Centrifuga Roberts N° 4 y 7	Medidores M3	M3	HORA	CRIST
Centrifuga Broadbent N° 5	Medidores M3	M3	TURNO	CRIST
Centrifuga Konturbex	Medidores M3	M3	TURNO	CRIST
Sol. TK 1000 SSK	Medidores M3	M3	TURNO	CRIST
MLK a Pozo	Medidores M3	M3	TURNO	CRIST
Prod. CNPC-L Bruta	Producción Bruta Desglosada	TON	HORA	CRIST
Prod. CNPC-M Bruta	Producción Bruta Desglosada	TON	HORA	CRIST
Densidad Espesador N°1	Densidad Pulpa	%	HORA	CRIST
Densidad Espesador N°2	Densidad Pulpa	%	HORA	CRIST
Densidad Espesador N°3	Densidad Pulpa	%	HORA	CRIST
Descarga - Disponibilidad MLT 1	DESCARGAS	%	HORA	CRIST
Descarga - Disponibilidad MLT 2	DESCARGAS	%	HORA	CRIST
Descarga - Disp. Hrs. MLT 1+2	DESCARGAS	%	HORA	CRIST
Descarga- Disponibilidad Hrs. TKs	DESCARGAS	%	HORA	CRIST
Descarga- Disponibilidad Area	DESCARGAS	%	HORA	CRIST
Descarga - Horas TKs Recup.	DESCARGAS	Hr	HORA	CRIST
Descarga - Horas TKs Refrig.	DESCARGAS	Hr	HORA	CRIST

Variable	Grupo	Unidad	Periodo	PLANTA
Arrastre Espesador 1	Mediciones por Turno	%	HORA	CRIST

ANEXO D.

Variable Original	Planta	Tipo
363106 SSK NaNO3	DUAL	AQ
363106 SSK K	DUAL	AQ
363420 Sal Rechazo NaNO3	DUAL	AQ
350128 Poza 10 NaNO3	MUR	AQ
350128 Poza 10 K	MUR	AQ
350001 Agua Riñihue NaNO3	MUR	AQ
334363 SSKD KClO4	CRIST	AQ
377147 SSNa NaNO3	PTS	AQ
377147 SSNa K	PTS	AQ
377154 Sal rechazo NaNO3	PTS	AQ
T° Entrada Trona 3	MUR	CONTROL
Amperaje Espesador Dual	MUR	CONTROL
T° Entrada Trona 1	MUR	CONTROL
Solución Alimentación Planta	MUR	CONTROL
KCl Agregado Minsal Bajo Cal.	MUR	CONTROL
Densidad Pulpa	DUAL	CONTROL
Espesor Cake	DUAL	CONTROL
Temp. Agua Lavado	DUAL	CONTROL
T. Entrada a Planta P.T.S.	PTS	CONTROL
% Sólidos Overflow	PTS	CONTROL
Volumen Ingreso a Planta	PTS	CONTROL
Sales Rechazadas	PTS	CONTROL
Agua Centrifuga	PTS	CONTROL
Sales Agregadas Pampa Blanca AL	PTS	CONTROL
T. Reactor 2	PTS	CONTROL
Densidad Espesador N°1	CRIST	CONTROL
Hrs. Operc. M.L.T.1	CRIST	CONTROL
Hrs. Operc. M.L.T.2	CRIST	CONTROL
T° Canales 6/7	CRIST	CONTROL
T° Entrada Esp. 1	CRIST	CONTROL
Volumen Alim. SSK	CRIST	CONTROL
335681 L NaCl	CRIST	AQ
335681 L KClO4	CRIST	AQ
335688 M NaCl	CRIST	AQ

Variable Original	Planta	Tipo
335688 M KClO4	CRIST	AQ

ANEXO E.

Variable	Mínimo ecuación1	Mínimo ecuación2	Mínimo ecuación4	%Desviación entre Mínimos
377147 SSNa NaNO3 lag4	343,135	282,185	337,04	10%
T° Entrada Trona 1 lag4	343,135	239,52	343,135	19%
% Sólidos Overflow lag5	361,42	215,14	337,04	26%
T° Canales 6/7 lag0	404,085	166,38	202,95	50%
% Sólidos Overflow lag4	131,08	102,26	125,84	13%
Temp. Agua Lavado lag4	123,22	104,88	131,08	11%
Temp. Agua Lavado lag1	133,7	112,74	131,08	9%
Agua Centrifuga lag4	128,46	107,5	57,72	37%
Temp. Agua Lavado lag2	1,4856	4,2168	1,4856	66%
377147 SSNa NaNO3 lag5	0,12	0,12	0,12	0%
Amperaje Espesador Dual lag4	0,12	1,4856	0,12	137%
363106 SSK NaNO3 lag3	1,4856	5,5824	0,12	119%
T° Entrada Esp. 1 lag1	223,6	343,3	303,4	21%
Densidad Pulpa lag1	54,808	48,112	54,808	7%
Sales Rechazadas lag4	70,2787	101,2759	54,7801	31%
363106 SSK NaNO3 lag2	2,42838518	0,65164444	3,13908148	62%
T° Entrada Esp. 1 lag0	2,47342222	0,70853333	2,64991111	55%
377147 SSNa K lag4	346,003333	436,396667	274,98	23%
Sales Agregadas Pampa Blanca AL lag5	433,026667	318,436667	263,87	26%
Sales Rechazadas lag5	99,2674	80,9197333	95,5978667	11%
363106 SSK NaNO3 lag4	102,936933	80,9197333	0,19	88%
350128 Poza 10 K lag4	0,05	0,05	0,05	0%
KCl Agregado Minsal Bajo Cal. Lag4	0,05	0,05	0,05	0%
T° Entrada Trona 3 lag4	47,4833333	46,2	32,0833333	20%
Temp. Agua Lavado lag3	2,24	4,48	2,352	42%
T° Canales 6/7 lag1	56	50	50	7%
Espesor Cake lag1	310,32	20	20	144%
T. Reactor 2 lag5	48	32	208	101%
377154 Sal rechazo NaNO3 lag4	1,57573333	1,65866667	1,70013333	4%
377147 SSNa K lag5	1,57573333	0,2488	1,57573333	68%
363106 SSK K lag3	1,53426667	1,45133333	1,65866667	7%
334363 SSKD KClO4 lag0	1,53426667	1,57573333	1,4928	3%

Variable	Mínimo ecuación1	Mínimo ecuación2	Mínimo ecuación4	%Desviación entre Mínimos
350128 Poza 10 NaNO3 lag 4	0,84	0,168	0,56	65%
T. Reactor 2 lag4	0,672	0,168	0,336	65%
Espesor Cake lag3	0,672	0,336	0,56	33%
363106 SSK K lag4	0,616	0,224	0,392	48%
T. Entrada a Planta P.T.S. lag5	51,563	58,217	56,553	6%
Sales Agregadas Pampa Blanca AL lag4	54,89	58,217	58,217	3%
Volumen Alim. SSK lag1	53,227	58,217	58,217	5%
T. Entrada a Planta P.T.S. lag4	53,227	46,573	58,217	11%
Volumen Alim. SSK lag0	63,18	0	61,56	87%
Agua Centrifuga lag5	66,42	58,32	61,56	7%
Densidad Pulpa lag2	14,025	14,025	1,87	70%
363420 Sal Rechazo NaNO3 lag4	14,96	14,96	10,285	20%
363106 SSK NaNO3 lag1	1201,2	1658,8	1201,2	20%
Densidad Pulpa lag3	1315,6	1716	1201,2	19%
Solución Alimentación Planta lag4	18,2	0	18,2	87%
Densidad Pulpa lag4	36,4	0	18,2	100%
363420 Sal Rechazo NaNO3 lag2	0	0	0	0%
363420 Sal Rechazo NaNO3 lag3	0	0	0	0%
363106 SSK K lag2	0	302,26	0	173%
377154 Sal rechazo NaNO3 lag5	0	480,06	0	173%
363420 Sal Rechazo NaNO3 lag1	30,24	51,84	43,2	26%
Espesor Cake lag2	44,64	50,4	46,08	6%
363106 SSK K lag1	1,57696	1,18272	1,54112	15%
334363 SSKD KClO4 lag1	1,57696	1,03936	1,57696	22%
Espesor Cake lag4	52066,52	0	52066,52	87%
Densidad Espesador N°1 lag1	0	0	0	0%
Volumen Ingreso a Planta lag4	0	53600,16	11166,7	131%
Densidad Espesador N°1 lag0	0	0	0	0%
Volumen Ingreso a Planta lag5	28,6133333	22,4266667	24,7466667	12%
350001 Agua Riñihue NaNO3 lag4	27,84	23,9733333	26,2933333	7%
Hrs. Operc. M.L.T.2 lag0	7,28	9,52	7,28	16%
Hrs. Operc. M.L.T.1 lag0	8,4	9,52	7,28	13%
Hrs. Operc. M.L.T.1 lag1	191,52	165,984	191,52	8%
Hrs. Operc. M.L.T.2 lag1	191,52	153,216	178,752	11%

ANEXO F.

Support Vector Machine

1. C=5.278032,kernel="rbfdot",sigma=0.006801176
2. C=1.470334e+02, kernel="rbfdot",sigma=0.006801176
3. C=147.0334,kernel="rbfdot",sigma=0.006801176
4. C=1.470334e+02,kernel="rbfdot",sigma=0.006801176

Random Forest

1. ntree=500, mtry=16
2. ntree=500, mtry=16
3. ntree=500, mtry=16
4. ntree=500, mtry=16

Support Vector Regression

1. C=5.278032,kernel="rbfdot",sigma=0.006801176
2. C=5.278032,kernel="rbfdot",sigma=0.006801176
3. C=5.278032,kernel="rbfdot",sigma=0.006801176
4. C=5.278032,kernel="rbfdot",sigma=0.006801176

Random Forest Regression

1. ntree=500, mtry=22
2. ntree=500, mtry=11
3. ntree=500, mtry=22
4. ntree=500, mtry=11