Análisis de Modelos de Machine Learning para proceso de obtención concentrado de hierro

Integrantes: Felipe Cavieres Gárnica

Roberto Miranda Yáñez

Asignatura: Electivo Profesional Data Science

Profesor: Felipe Peña

Fecha: 11 de Enero del 2021

Índice

[Introducción 3](#_Toc61219475)

[Descripción del Problema 4](#_Toc61219476)

[Objetivos 4](#_Toc61219477)

[Objetivo General 4](#_Toc61219478)

[Objetivos Específicos 4](#_Toc61219479)

[Análisis del Data Set 5](#_Toc61219480)

[Desarrollo de Modelos 7](#_Toc61219481)

[Conclusiones 10](#_Toc61219482)

# Índice Ilustraciones

[Ilustración 1: Mapa de Calor de Correlación de Variables Data Set Inicial 4](#_Toc61391570)

[Ilustración 2: Mapa de Calor de Correlación de Variables Data Set Agrupado 5](#_Toc61391571)

[Ilustración 3: Método del Codo Clustering K-MEANS 6](#_Toc61391572)

[Ilustración 4: Gráfico de predicciones v/s y\_test Regresión Lasso 8](#_Toc61391573)

[Ilustración 5: Gráfico de predicciones v/s y\_test Grid Search CV 8](#_Toc61391574)

# Introducción

En el presente informe se analizará el proceso de obtención del mineral de hierro a través de modelos de Machine Learning con el objetivo de predecir el nivel de impureza presente en el concentrado de hierro. Esto ayudaría a brindar una información temprana a los ingenieros y así estos podrán tomar una correcta decisión al momento de mejorar sus procesos.

Para implementar modelos de Machine Learning se debe seguir una serie de pasos. El primero es recolectar los datos, en nuestro caso estos se obtuvieron de un Dataset el cual contaba con la información de cuanto porcentaje de hierro e impureza existía al comienzo, la cantidad de reactivos, el caudal de aire y el nivel de espuma que se tenían en el proceso de flotación y finalmente el porcentaje final obtenido de tanto de concentrado de hierro como de concentrado de sílice. El segundo paso es analizar los datos, donde primero se visualizaron para luego hacer un analizar de correlación y gráficas como pairplot, boxplot y heat map. El siguiente paso, fue elegir el modelo de Machine Learning que se aplicaría al modelo para poder predecir el nivel de impureza que se obtiene en el proceso de obtención del hierro. Pero antes se escalaron los datos y una vez hecho esto se aplicaron modelos de Clustering para lograr el agrupamiento de conjuntos de datos no etiquetados y luego se emplearon diferentes modelos de regresión lineal acompañado del enfoque grid. Cabe destacar que se debe entrenar el modelo por ello se utilizó el set de datos de entrenamiento para ejecutar nuestra máquina y se debe ver una mejora incremental (para la predicción). Para lograr esto se debe inicializar los “pesos” de nuestro modelo aleatoriamente, los pesos son los valores que multiplican o afectan a las relaciones entre las entradas y las salidas, estos se irán ajustando automáticamente por el algoritmo seleccionado cuanto más se entrena. Finalmente como último paso se evaluaron los modelos aplicados a través de métricas con el objetivo de determinar cuál elegir para predecir.

# Descripción del Problema

La principal problemática que se presenta es como gestionar de manera adecuada el proceso de flotación del hierro en base al nivel de impureza obtenido al final de este. Cabe destacar que la flotación es un proceso físico-químico de separación de minerales o compuestos que vienen del proceso de molienda. Para lograr una buena obtención de concentrado de hierro se deben considerar varios factores dentro de la flotación como por ejemplo la cantidad de reactivos a utilizar, los ciclos que se deben realizar y la cantidad de aire y burbujas que se deben generar en cada celda.

Si se logra predecir el nivel de impureza que se obtiene al final de este proceso junto a la ayuda de diferentes modelos de Machine Learning, se podrá definir de manera correcta los factores anteriormente mencionados para así obtener un mayor concentrado de hierro y un menor concentrado de sílice.

# Objetivos

## Objetivo General

Desarrollar diferentes modelos de Machine Learning mediante el análisis de los factores de temperatura, presión u otros en el proceso de flotación de la extracción del hierro con el objetivo de predecir el nivel de impureza que existe en el concentrado.

## Objetivos Específicos

* Análisis del data set del proceso de producción de concentrado de hierro.
* Interpretar los datos presentes a través de gráficas y análisis de correlación.
* Aplicar los modelos de Machine Learning más adecuados en base al análisis realizado con anterioridad.
* Concluir en base a los resultados obtenidos.

# Análisis del Data Set

En primera instancia se hizo un análisis de correlación utilizando todas las variables presentes en el data set, donde se obtuvo como resultado correlaciones no muy significativas, lo que no contribuía con el objetivo planteado en este proyecto.

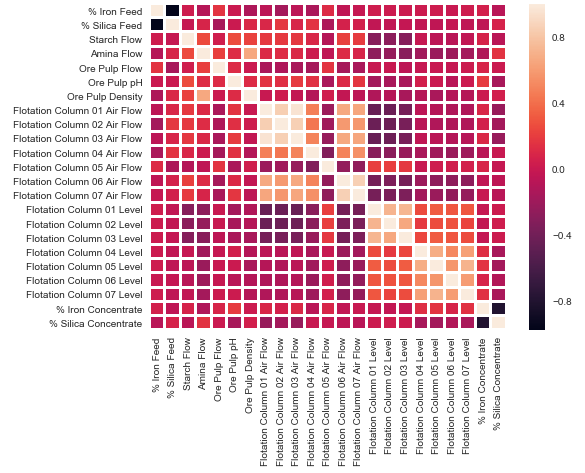


Ilustración 1: Mapa de Calor de Correlación de Variables Data Set Inicial

De esta forma no dimos cuenta que el data set necesitaba ser trabajado o implementar una estrategia que permitiera encontrar correlaciones significativas. Por lo tanto se identificó que la columna “date” que hace alusión a la fecha en la que fue registrada la información debía ser descompuesta para extraer información relacionada con el día, mes, año y hora. Una vez que se añadieron estas variables al data set se agrupó el promedio de la información en base a los criterios hora y mes. Esto no afectaba demasiado al data set ya que las variables significativas no tenían una gran desviación alrededor de su media. A continuación se presenta el mapa de calor con los datos agrupados.

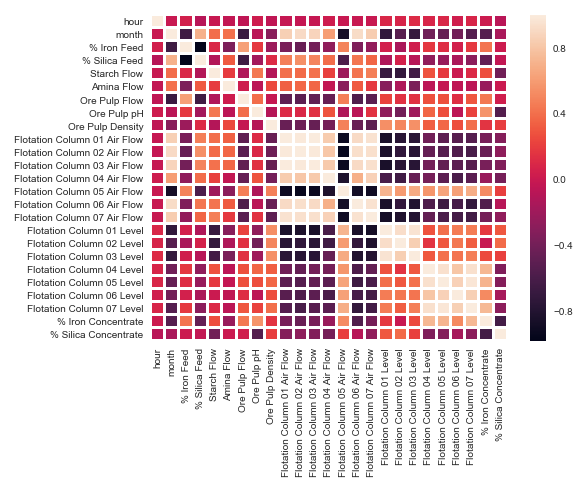


Ilustración 2: Mapa de Calor de Correlación de Variables Data Set Agrupado

Es posible identificar que esto mejoró bastante el modelo y se puede visualizar correlaciones bastante significativas, tanto positivas como negativas. Esta información se complementó mediante gráficas de dispersión.

También se realizaron Gráficos de Caja para analizar la distribución de los datos y existencia de outliers en la información que se tenía de cada variable del Data Set. Este análisis fue positivo ya que no se identificó outliers y por lo tanto no se incurrió en un proceso de limpieza de datos.

# Desarrollo de Modelos

**Clustering**

En primera instancia se realizaron las gráficas de dispersión de cada variable con respecto al % Silicate Concentrate (% Impureza), y posteriormente se decidió aplicar tres modelos de clustering para determinar cuál realizaba una mejor agrupación de los datos.

* K-MEANS: Se utilizó el Método del Codo para obtener el número óptimo de clusters. Sin embargo fue pertinente probar el código iterando diferentes valores del parámetro n\_clusters, ya que efectivamente este dependía de la variable que se analizara.
* DBSCAN: Fue pertinente probar el código iterando diferentes valores del parámetro eps, ya que efectivamente este dependía de la variable que se analizara.
* Agglomerative Clustering: Fue pertinente probar el código iterando diferentes valores del parámetro n\_clusters, ya que efectivamente este dependía de la variable que se analizara.

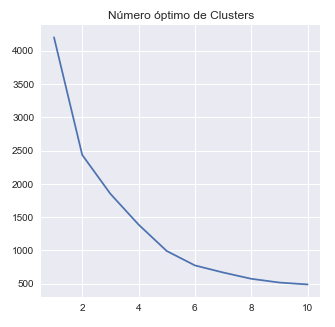


Ilustración 3: Método del Codo Clustering K-MEANS

El método que permitió realizar mejores clusters fue Agglomerative Clustering. Las variables en las que se identificó la presencia de clusters bien definidos junto con el número de clusters apropiado se presenta a continuación:

* month: 7
* Starch Flow: 2
* Ore Pulp Flow: 3
* Flotation Column 06 Air Flow: 3
* Flotation Column 07 Air Flow: 2
* Flotation Column 01 Level: 2
* Flotation Column 02 Level: 2
* Flotation Column 03 Level: 2
* Flotation Column 06 Level: 5
* Flotation Column 07 Level: 5

**Regresión**

Se aplicó modelos de regresión con el fin de lograr el objetivo principal de este proyecto, predecir la variable de respuesta % Silica Concentrate (% Concentrado Silicato) que corresponde al porcentaje obtenido de concentrado de mineral de baja ley en función de todos las variables que participan en el proceso de tratamiento de este y de la hora transcurrida. En primera instancia se “entrenó el modelo”, lo cual permitió determinar un porcentaje de datos de entrenamiento y otro de prueba. Luego se aplicaron 3 modelos considerando todas las variables en cada uno.

* Regresión Lineal: Este método de regresión es el más conocido y más popular. El R2 fue de 81% aproximadamente lo que indica un buen ajuste lineal. Además consideró apropiado asignar coeficientes (betas) a todas las variables.
* Regresión Ridge: En este método de regresión el R2 fue de 82,6% aproximadamente lo que indica un buen ajuste lineal. Además, al igual que en el caso anterior, consideró apropiado asignar coeficientes (betas) a todas las variables.
* Regresión Lasso: En este método de regresión el R2 fue de 84,6% aproximadamente lo que indica un buen ajuste lineal y el mejor entre los tres métodos utilizados. En este caso el método solo consideró apropiado asignar coeficientes (betas) a las siguientes variables:
* hour: -0,0336
* Ore Pulp Flow: 0,1773
* Ore Pulp pH: 0,036
* Ore Pulp Density: 0,0653
* Flotation Column 03 Air Flow: -0,46
* Flotation Column 05 Air Flow: 0,4122
* Flotation Column 06 Air Flow: 0,153
* Flotation Column 02 Level: -0,2267
* Flotation Column 03 Level: 0,1056
* Flotation Column 04 Level: 0,0431
* % Iron Concentrate: -1,0806

Comparando los métodos de regresión aplicados en este proyecto, se llegó a la conclusión de que el método de Regresión Lasso es el más apropiado ya que solo asigna pesos (betas) a variables que son verdaderamente significativas. Por otro lado, como nos dimos cuenta en la sección de Análisis del Data Set, con la agrupación por hora y mes, surgieron correlaciones bastante significativas entre las variables, por lo que Lasso, también soluciona el problema de autocorrelación, dejando solo una de dos variables significativas que se encuentran correlacionadas. De 25 variables, el modelo se Regresión Lasso redujo la importancia a solo 11 y lo más positivo es que sí consideró significativa la variables hora, que es la variable por la que cobre importancia este proyecto. Además, esta capacidad discriminativa del método Lasso, permite crear un modelo no supervisado, puesto que no es necesario analizar la correlación entre variables para eliminar la autocorrelación entre ellas eliminando algunas de estas antes de aplicar el modelo de regresión.

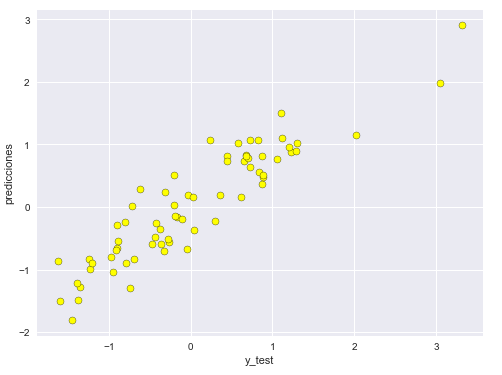


Ilustración 4: Gráfico de predicciones v/s y\_test Regresión Lasso

Finalmente, a modo de perfeccionamiento del modelo de Regresión Lasso, se procedió a aplicar Grid Search CV con valores de los parámetros verbose y cv de 3 y 5 respectivamente, y se concluyó que el mejor valor para los parámetros alpha y tol debián ser 0,005 y 0.001 respectivamente. Esto mejoraría el R2 a un valor de 84,9% aproximadamente, un valor no muy superior al obtenido en primera instancia sin utilizar Grid Search CV.

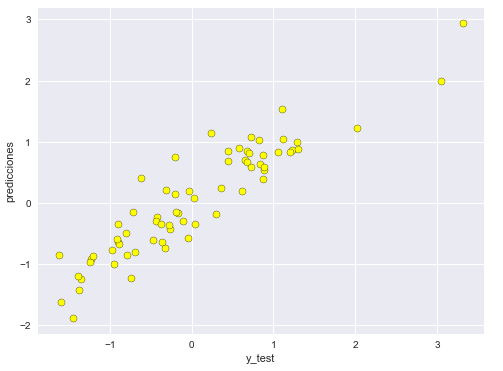


Ilustración 5: Gráfico de predicciones v/s y\_test Grid Search CV

# Conclusiones

Este trabajo fue muy significativo ya que permitió aplicar los conocimientos adquiridos durante el semestre en el electivo profesional de Data Science y obtener resultados bastante valioso para un Data Set con información referente al proceso de tratamiento de mineral en un proceso de producción de concentrado de hierro y silicato.

El primer gran aporte de este trabajo fue realizar estadística descriptiva de una gran cantidad de variables que forman parte del proceso de tratamiento del mineral.

El segundo hace referencia a la identificación de posibles clusters de agrupamiento de las variables, para categorizarlas y tener más menos una referencia de entre que valores varían y por lo tanto monitorear su comportamiento para evitar que estas se salgan del cluster correspondiente y predicho por el modelo. Esto se puede asociar al control de calidad del procesamiento del mineral y del valor que deben tomar las variables y factores que influyen en él para obtener un porcentaje de concentrado de Hierro y Silicato requerido por la administración de la planta de producción.

Por último, la regresión, nos permitió determinar un modelo a partir del cual será posible predecir el porcentaje de concentrado de Silicato en función de la hora y del valor que tomen 10 factores que participan en el proceso de tratamiento del mineral. Esto es muy importante ya que se podría implementar en la programación de operaciones que se realicen en la planta de concentrado para adecuar las condiciones del proceso en base a la hora, siendo esta un variable que se podría implementar en un modelo de programación dinámica, donde cada etapa de decisión sea la hora transcurrida en el proceso.