Pontificia Universidad Católica del Ecuador

Facultad de Ingeniería

Escuela de Sistemas de Información



TEMA:

Análisis y minería de datos del proceso de encolado de Tableros de MDF en la planta de aglomerados Cotopaxi

AUTOR:

Anahí Micaela Males Maldonado

TRABAJO PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO DE SISTEMAS Y COMPUTACIÓN

QUITO, 2023

DEDICATORIA

Quiero dedicar este trabajo a todas las almas valientes que han depositado su confianza en mí, incluso cuando yo mismo dudaba de mis capacidades. Agradezco a Dios por haberme brindado esta oportunidad y permitirme seguir un sueño que se ha convertido en realidad.

A mi padre, quien me ha inculcado valores sólidos y ha sido un apoyo importante en mi vida. A mi madre, cuyo amor incondicional ha sido mi ancla en los momentos más tempestuosos.

A mis hermanos, que han sido el combustible para seguir perseverando y quienes me han dado la fortaleza para mantenerme en pie. A las personas que Dios ha puesto en mi camino y me ayudaron a mantener el rumbo, a no rendirme y a continuar con determinación.

Finalmente, quiero expresar mi sincera gratitud a la Ingeniera Susana Masapanta, cuya guía y conocimientos han sido un faro en el desarrollo de este proyecto. Asimismo, agradezco a todos los profesores que me han brindado su apoyo durante la realización de este trabajo. Sin ustedes, este trabajo no sería lo que es hoy.

AGRADECIMIENTO

Agradezco a la Pontificia Universidad Católica del Ecuador, por haberme abierto sus puertas y permitirme cursar la carrera de mis sueños. Mi agradecimiento es profundo, pues la oportunidad de aprender y crecer aquí ha sido algo invaluable.

A mis queridos abuelitos maternos y paternos, quienes han dejado una huella profunda en mi vida con sus enseñanzas y amor. Cada lección, cada recuerdo, vive en mí y aprendo de ello cada día. Su legado continúa y sigue guiando mi camino aún en su ausencia física.

RESUMEN

En este proyecto de titulación, se realizó un análisis del subproceso de encolado dentro del proceso dosificado, empleando la metodología CRISP-DM. Para identificar las variables esenciales que tienen un impacto significativo de dicho subproceso. Estas variables se identificaron con la finalidad de ser utilizadas como Indicadores Clave de Rendimiento (KPIs) para monitorizar el proceso de producción.

A través de los modelos aplicados, se logró identificar las variables más críticas que influían en otros atributos relevantes dentro del subproceso de encolado. Estos indicadores se propusieron para mejorar la supervisión y control.

Además, se realizó un análisis para determinar qué atributos caracterizaban a los grupos con un mejor rendimiento en el proceso. Revelando que ciertos atributos, tenían correlaciones directas o inversas con las variables estudiadas.

Donde, algunos grupos que presentaban altos niveles de dichas características mostraban un mejor rendimiento en comparación con otros grupos. Estos hallazgos proporcionaron una perspectiva valiosa sobre el subproceso de encolado, mejorando la comprensión general de su rendimiento dentro del proceso de producción.

ÍNDICE

Contenido

ÍNDICE DE FIGURAS, GRÁFICOS Y TABLAS V

ÍNDICE DE FIGURAS V

ÍNDICE DE TABLAS VI

CAPÍTULO I: INTRODUCCIÓN 1

1.1. Justificación 1

1.2. Planteamiento del problema 1

1.3. Objetivo General 1

1.4. Objetivos Específicos 2

1.5. Alcance 2

CAPÍTULO II: FUNDAMENTACIÓN TEÓRICA 3

2.1. Análisis de datos 3

2.1.1. Método CRISP-DM 3

2.2. Tipos de aprendizaje 4

2.2.1. Aprendizaje supervisado 4

2.2.2. Aprendizaje no supervisado 5

2.3. Minería de datos 5

2.3.1. Técnicas de minería de datos 6

2.3.2. Modelos y algoritmos 8

2.3.3. Métricas comunes de minería de datos 8
2.4. *KPI's* 9
2.5. Bases de datos 10
2.6. Python 10
2.7. Herramientas y librerías 10
2.7.1. Google Colab 10
2.7.2. *Scikit-learn* 11
2.7.3. *Pandas* 11
2.7.4. *Pyspark* 11
2.8. Técnicas de exploración de los datos 11
2.8.1. Estadística descriptiva 11
2.8.2 Análisis de correlación 12
2.8.3. Análisis de valores atípicos 12
2.8.4. Análisis de valores faltantes 12
CAPÍTULO III: APLICACIÓN DEL MÉTODO CRISP-DM 13
3.1. Comprensión del negocio 13
3.1.1. Política, Visión, Misión y Objetivos del negocio 17
3.1.2. Recursos proporcionados 17
3.1.3. Determinación de los objetivos de minería de datos 18
3.2. Comprensión de los datos 19
3.2.1. Recopilación de los datos iniciales 19

3.2.2. Descripción de datos 20
3.2.3. Exploración de datos 22
3.3. Preparación de los datos 27
3.3.1. Selección de datos 27
3.3.3. Construcción de datos 27
CAPÍTULO IV: MODELADO 28
4.1. Selección de técnicas de modelado 28
4.2. Generación de un diseño de comprobación 28
4.3. Generación de los modelos 28
4.4. Evaluar los modelos 32
CAPÍTULO V: EVALUACIÓN 35
5.1. Evaluación de los resultados 35
5.2. Procesos de revisión 37
5.3. Desarrollo del documento con las recomendaciones 37
CONCLUSIONES Y RECOMENDACIONES 42
Conclusiones 42
Recomendaciones 43
BIBLIOGRFÍA 44
ANEXOS 47
Anexo A: Información de las columnas de la tabla de Proceso_Encolado_Dosificado 47
Anexo B: Matriz de correlación entre las columnas de Proceso_Encolado_Dosificado 47

Anexo C: Cuartiles y número de valores atípicos de las columnas de Proceso_Encolado_Dosificado 48

Anexo D: Información de las columnas de Sp_Vp_Encolado_Dosificado 49

Anexo E: Matriz de correlación entre las columnas de Sp_Vp_Encolado_Dosificado 50

Anexo F: Cuartiles y valores atípicos de Sp_Vp_Encolado_Dosificado 50

Anexo G: Número de nulos por columna de Sp_Vp_Encolado_Dosificado. 51

Anexo H: Número de ceros por columna de Sp_Vp_Encolado_Dosificado. 51

ÍNDICE DE FIGURAS, GRÁFICOS Y TABLAS

ÍNDICE DE FIGURAS

Figura 1 Metodología CRISP-DM 4

Figura 2 Clasificación de las técnicas de Data Mining 6

Figura 3 Fases de Fabricación de los Tableros Aglomerados Cotopaxi S.A 13

Figura 4 Coeficientes más influyentes del modelo de regresión lineal de los atributos: "Presion resina", "Consumo resina real", "Vp ENDURECEDOR" 30

Figura 5 I mportancias de las características del modelo Lasso ajustado de los atributos: "Presion resina", "Consumo resina real", "Vp ENDURECEDOR" 30

ÍNDICE DE TABLAS

Tabla 1 Descripción de los atributos de la tabla Proceso_Encolado_Dosificado 20

Tabla 2 Descripción de los atributos de la tabla Sp_Vp_Encolado_Dosificado 21

Tabla 3 Métricas de evaluación Regresión Lineal y LASSO de: "Presion resina", "Consumo resina real", "Vp ENDURECEDOR" 33

Tabla 4 Métricas de evaluación de K-means y DBSCAN 34

CAPÍTULO I: INTRODUCCIÓN

1.1. Justificación

El uso de minería de datos contribuye: a la administración de la organización, a la optimización de los productos en base a la información obtenida, a la determinación de los objetivos, al preprocesamiento de los datos y finalmente al análisis que se obtiene de los modelados aplicados al problema (Garosina, 2020).

Un ejemplo de éxito en el uso minería de datos es la empresa West Fraser. Mediante el análisis de los datos de producción y logística, West Fraser ha podido optimizar su planificación de la producción, reducir los costes de transporte y aumentar la eficacia de su cadena de suministro. Logrando posicionarse en el mercado mundial (West Fraser, 2023).

Retomando los anteriores puntos mencionados, el uso de minería de datos es una herramienta útil y poderosa, las organizaciones con una cantidad masiva de datos obtienen información valiosa para la toma de decisiones. Dado que, se pueden identificar patrones, tendencia y relaciones que apoyan a la empresa a entender mejor la información obtenida de los datos.

1.2. Planteamiento del problema

El proyecto tiene como objetivo presentar propuestas de mejora en el proceso de dosificación, mediante el análisis y uso de minería de datos. A pesar de que ya se cuenta con un sistema SCADA que captura grandes volúmenes de datos, no se aprovecha todo su potencial. Debido a que no se analizan, ni se aplican técnicas de minería de datos.

1.3. Objetivo General

Analizar los datos obtenidos del sistema SCADA, usando minería de datos para la elaboración de recomendaciones en el proceso de dosificación.

1.4. Objetivos Específicos

1) Comprender y definir los atributos claves con los que se trabajará en la minería de datos.

2) Preparar los datos de manera que haya valores claros, precisos y consistentes.

3) Elegir la técnica de modelado de acuerdo con los datos disponibles, objetivos de minería y los requisitos específicos del modelo seleccionado.

4) Seleccionar y generar las primeras pruebas a realizarse con los datos.

5) Realizar una retroalimentación después de cada prueba para identificar fallas, imprevistos, descubrimientos y resultados obtenidos.

1.5. Alcance

El objetivo principal del estudio es analizar los datos del subproceso de encolado en el proceso de dosificación de cola, a fin de identificar oportunidades de mejora. Las cuáles serán recopiladas y entregadas al departamento de calidad en un documento con las recomendaciones obtenidas, para mejorar el proceso mencionado.

Tratándose de un estudio teórico, este se centrará en el análisis de datos y la aplicación de técnicas y metodologías de la minería de datos.

CAPÍTULO II: FUNDAMENTACIÓN TEÓRICA

2.1. Análisis de datos

El análisis de datos es un proceso que consiste en extraer el significado de estos sometiéndolos a una serie de análisis, lecturas e interpretaciones. El proceso puede revelar problemas que afecten al objetivo inicial, por lo que se requiere una planificación y revisión constante. El análisis de datos puede ser cuantitativo o cualitativo, dependiendo de la información que se vaya a procesar y del objetivo de este. El analista debe ser riguroso para garantizar la fiabilidad (Peña, 2017).

2.1.1. Método CRISP-DM

CRISP-DM, acrónimo de *Cross-Industry Standard Process for Data Mining*, *CRISP-DM* es un modelo de proceso de minería de datos flexible que puede adaptarse a diferentes necesidades. Es iterativo y centrado en el negocio, y tiene varias ventajas, incluyendo una estructura clara y precisa (IBM Documentation, 2021).

Las fases del método CRISP-DM son:

Comprensión del negocio: en esta fase se establecen los objetivos del proyecto, se identifican variables relevantes y se recopila información necesaria.

Comprensión de los datos: se realiza una exploración de los datos para comprender su estructura, calidad y contenido. Se identifican posibles problemas en los mismos y se seleccionan las variables relevantes para el análisis.

Preparación de los datos: se realiza la limpieza de los datos, eliminando valores faltantes o erróneos y transformando las variables en formatos adecuados para su análisis. Se pueden crear nuevas variables a partir de las existentes para mejorar el análisis.

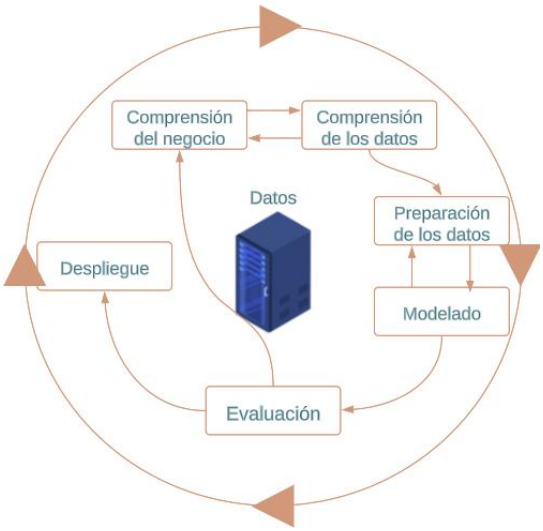
Modelado: se seleccionan los algoritmos adecuados para el análisis, y se aplican a los datos preparados para obtener modelos predictivos y descriptivos.

Evaluación: se evalúan los modelos obtenidos de la fase de modelado, comprobando su validez y eficacia en la resolución del problema planteado. Se pueden realizar ajustes y mejoras en los modelos según los resultados obtenidos.

Despliegue: Implementación de los modelos seleccionados, creación de la planificación de despliegue, la planificación del control y mantenimiento y elaborar el informe final (IBM Documentation, 2021).

En la Figura 1 se puede observar las fases que tiene *CRISP-DM* y cómo se puede avanzar o retroceder en ellas.

Figura 1 Metodología CRISP-DM



Autor: IBM. (2021, March 17). CRISP-DM Process. Recuperado de https://www.ibm.com/docs/en/SS3RA7_sub/modeler_crispdm_ddita/clementine/images/crisp_process.jpg

2.2. Tipos de aprendizaje

2.2.1. Aprendizaje supervisado

El aprendizaje supervisado requiere un conjunto de datos previamente etiquetados para indicar al modelo lo que debe aprender (Sánchez, 2020). Es un proceso que permite establecer

una relación entre las entradas y las salidas del sistema, basado en ejemplos en los que se induce un nuevo conocimiento a partir de una serie de casos y sus correspondientes etiquetas (Arteaga, 2015).

Por ejemplo, si tiene un conjunto de correos etiquetados como "problema" o "sin problema", podría entrenar un modelo de aprendizaje supervisado para clasificar los correos entrantes en una de esas dos categorías. Esto le permitiría automatizar la tarea de filtrar los correos electrónicos que requieren la atención del equipo de soporte o los que pueden ignorarse.

2.2.2. Aprendizaje no supervisado

El aprendizaje no supervisado se utiliza para extraer patrones ocultos de datos no etiquetados. También puede utilizarse para simplificar conjuntos de datos reduciendo su dimensionalidad. Para ello, el algoritmo debe definir una métrica de similitud o distancia que permita comparar los datos entre sí. (Sánchez, 2020).

Un ejemplo de aprendizaje no supervisado sería la segmentación de los pacientes que han sido atendidos en una sala de emergencias en grupos homogéneos, sin tener información previa sobre las características que deben definir a dichos grupos (Rueda, 2019).

2.3. Minería de datos

La minería de datos es un proceso que implica extraer información valiosa de grandes cantidades de datos. Se utiliza para predecir eventos futuros, identificar patrones y relaciones complejas, y tomar decisiones estratégicas informadas. Este proceso requiere conocimientos en matemáticas, informática, organización y negocios (Gironés Roig et al., 2017).

La minería de datos es un campo en constante desarrollo y cuenta con una amplia gama de técnicas y herramientas. Es importante que los profesionales de la minería de datos se mantengan actualizados con los últimos avances para ser efectivos en su trabajo (Hernández Orallo et al., 2007).

2.3.1. Técnicas de minería de datos

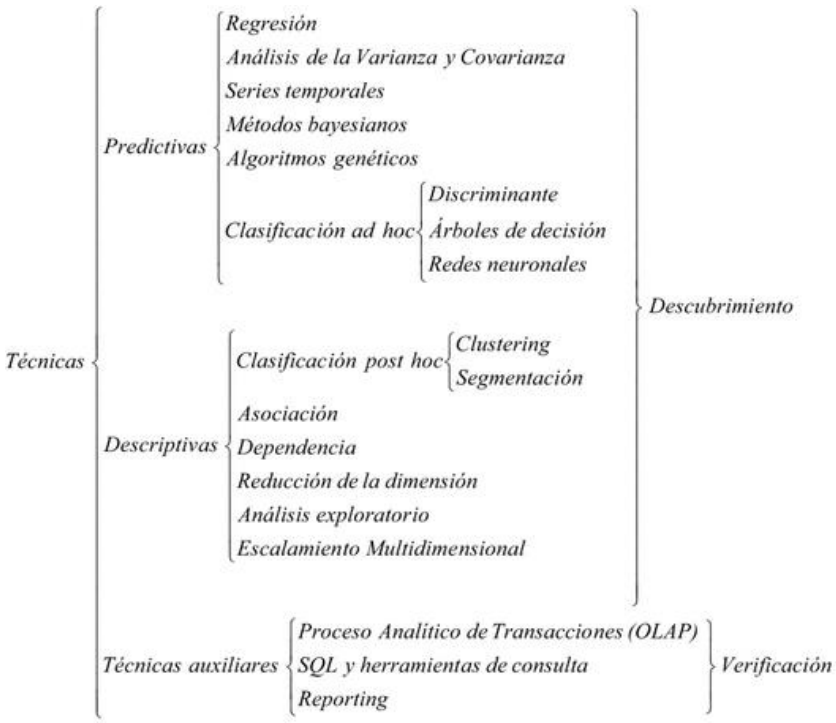
Las técnicas de minería de datos descubren automáticamente conocimientos a partir de grandes bases de datos (Pérez López & Santin González, 2007). Utilizando tecnologías avanzadas como el reconocimiento de patrones, las redes neuronales, la lógica difusa y los algoritmos genéticos para encontrar patrones, perfiles y tendencias en datos estructurados (Hernández Orallo et al., 2007).

Según Hernández et al. (2007), la minería de datos es un proceso complejo que requiere una gran inversión de esfuerzo, tiempo y dinero. Es importante destacar que los algoritmos no siempre pueden obtener buenos resultados, y eso no significa que el investigador sea incompetente.

Pérez López y Santin González (2007) dividen las técnicas de minería de datos en tres categorías principales: predictivas, descriptivas y auxiliares. Las técnicas predictivas separan las variables en dependientes e independientes, mientras que las técnicas descriptivas consideran todas las variables de la misma manera. Las técnicas auxiliares son complementarias a las técnicas predictivas y descriptivas.

A continuación, en la Figura 2 se indica la clasificación de las técnicas de minería.

***Figura** 2 Clasificación de las técnicas de Data Mining*



Autor: Pérez López, C., & Santin González, D. (2007). Minería de datos: Técnicas y herramientas (reimp.). Ediciones Paraninfo, S.A. [Imagen de la página 9].

Regresión lineal múltiple es un procedimiento estadístico que permite predecir el valor de una variable respuesta, basándose en dos o más variables independientes. Este método modela la conexión lineal entre las variables predictoras y una única variable respuesta (Mar casco, 2022). Existen dos o más variables predictoras y una única variable respuesta. Las predictoras corresponden a las variables independientes y la respuesta a la variable dependiente. La regresión lineal múltiple es frecuentemente empleada en el análisis de datos para anticipar el valor de una variable dependiente a partir de dos o más variables independientes (ProbabilidadyEstadística.net, 2023).

La regresión Lasso es una técnica de regularización que se emplea en los métodos de regresión con el objetivo de lograr predicciones más precisas. Este enfoque promueve la construcción de modelos más simples y dispersos, lo cual es beneficioso en situaciones de alta multicolinealidad o cuando se busca automatizar la selección de variables en el modelo (Castellano, 2023).

K-means es un algoritmo de agrupación que agrupa puntos de datos en k conglomerados en función de su similitud (Ramírez, 2023). Se utiliza para descubrir patrones en los datos que pueden no ser fácilmente discernibles (Sanz, 2023).

DBSCAN es un algoritmo de clustering que agrupa puntos de datos en clusters en función de su densidad local (Sharma, 2022). Resulta útil para identificar conglomerados en grandes conjuntos de datos espaciales y cuando no se conoce de antemano el número de conglomerados (Agarwal & Srivastava, 2014).

2.3.2. Modelos y algoritmos

Existen dos tipos de modelos: predictivos y descriptivos. (Agarwal & Srivastava, 2014). Los modelos predictivos estiman variables relevantes, mientras que los descriptivos identifican patrones en los datos. Por ejemplo, un modelo predictivo puede utilizarse para predecir el clima en función de su ubicación, mientras que un modelo descriptivo puede utilizarse para comprender cómo se comunican y relacionan las personas con diversos contenidos en línea utilizando datos de redes sociales (Hernández Orallo et al., 2007).

2.3.3. Métricas comunes de minería de datos

Las métricas que se utilizaron para evaluar los modelos son:

El coeficiente de determinación, también conocido como R cuadrado, es una medida que indica qué tan bien se ajusta un modelo a la variable que se intenta explicar. El coeficiente de determinación puede variar entre 0 y 1. Cuando se acerca a 1, significa que el modelo se ajusta de manera más adecuada a la variable en cuestión (Software DELSOL, 2019).

El error cuadrático medio (RMSE), es una medida de la distancia entre las predicciones de un modelo y los valores reales. Se calcula tomando la raíz cuadrada del error cuadrático medio entre las predicciones y los valores reales. Un valor más bajo indica un mejor ajuste del modelo.

El RMSE es una métrica útil para evaluar la precisión de las predicciones, especialmente cuando la predicción es el objetivo principal (González, 2022).

El error absoluto medio (MAE), es una medida de la proximidad entre la predicción del modelo y el valor real. Se calcula promediando la diferencia absoluta entre el valor previsto y el valor real. Un valor MAE más bajo indica un mejor ajuste del modelo (González, 2022).

WCSS es el acrónimo de *Within Clúster Sum of Squares*, en español conocido como la suma de cuadrados dentro del *clúster*. Esta métrica mide la proximidad entre los puntos de datos de un conglomerado. Se calcula sumando las distancias al cuadrado entre todos los puntos de datos de un conglomerado y el centroide del conglomerado. Un valor más bajo indica que los puntos de datos de cada cluster están bien agrupados (Akhtar, 2021).

El coeficiente de silueta se obtiene comparando las distancias medias dentro de un *clúster* y la distancia media más cercana a otros *clúster*es para cada muestra. Se calcula como (distancia más cercana a otros *clúster*es - distancia media dentro del *clúster*) dividido por el valor máximo entre esas dos distancias (González, 2021).

El índice Davies-Bouldin (DBI), se define como la medida media de la similitud entre cada conglomerado y su conglomerado más similar, donde la similitud es la relación entre la distancia dentro de un conglomerado y la distancia entre conglomerados. Por lo tanto, los conglomerados más distantes entre sí y menos desorganizados obtendrán una puntuación más alta (Scikit-learn contributors., 2023).

2.4. *KPI's*

Los indicadores clave de rendimiento (*KPI*) son herramientas esenciales para controlar y mejorar los procesos empresariales. Proporcionan mediciones cuantificables de factores importantes como el capital humano, la satisfacción del cliente y los costes (Marr B, 2012).

Hay dos tipos de *KPI*: operativos y estratégicos. Los *KPI* operativos miden el rendimiento de los procesos de la empresa, mientras que los *KPI* estratégicos miden el camino de la empresa hacia sus objetivos (Parmenter D, 2010).

Los *KPI* son herramientas fundamentales para estandarizar los procesos empresariales y medir el rendimiento. Deben expresarse en un lenguaje coloquial y disponer de los datos necesarios para alimentarlos y calcularlos. A la hora de plantear una estrategia de cambio de procesos, es importante tener en cuenta la rueda de la competitividad para establecer los *KPI* más pertinentes (Iván, 2019).

2.5. Bases de datos

La base de datos utilizada en el proyecto es SQL Server Management Studio 18, una herramienta de gestión de bases de datos desarrollada por Microsoft. SSMS permite a los usuarios crear, gestionar y buscar bases de datos mediante una interfaz gráfica. También incluye la posibilidad de solucionar errores de bases de datos (Santamaría & Hernández, 2016, p. 1).

2.6. Python

Python es un lenguaje de alto nivel, que se caracteriza por su simplicidad y la reducción del número de líneas de código necesario para realizar operaciones más simples. Es un lenguaje muy completo para el mundo de la programación, con un conjunto de librerías y módulos extensos que lo hacen adecuado para cualquier tipo de proyecto. Además, cuenta con una gran comunidad por detrás que lo mantiene actualizado (Moreno Muñoz & Córcoles, 2018).

2.7. Herramientas y librerías

2.7.1. Google Colab

Google Colab está basado en la nube, lo que permite a los usuarios acceder a él desde cualquier lugar. Además, ofrece acceso a hardware potente, como *GPU* (Unidades de Procesamiento Gráfico) y *TPU* (Unidades de Procesamiento Tensorial), lo cual puede acelerar

significativamente el entrenamiento y la ejecución de modelos de aprendizaje automático. (Redacción Tokio, 2022).

2.7.2. *Scikit-learn*

Scikit-Learn es una biblioteca para el aprendizaje automático que proporciona implementaciones eficientes de numerosos algoritmos comunes. Es fácil de usar y tiene una sintaxis consistente para diferentes tipos de modelos. *Scikit-Learn* también ofrece funcionalidades para realizar pasos de preprocesamiento (Rédac, 2022).

2.7.3. *Pandas*

Pandas es una librería para el análisis y la manipulación de datos. Dado que, es rápida, flexible y fácil de usar. Además de ser utiliza para tareas, como limpiar y visualizar los datos (Alexandre, 2022).

2.7.4. *Pyspark*

Apache Spark es una plataforma de código abierto que permite procesar de manera distribuida grandes conjuntos de datos. Puede utilizarse con *Python* a través de la biblioteca *PySpark*, que ofrece un rendimiento similar al de SCALA. *PySpark* es una alternativa adecuada a *Pandas* para grandes conjuntos de datos que requieren un tiempo de cálculo significativo (Rédac, 2022).

2.8. Técnicas de exploración de los datos

2.8.1. Estadística descriptiva

Se centra en la descripción cuantitativa de diversas propiedades de un conjunto de datos. Para proporcionar información objetiva y precisa sobre variables como el peso de una caja de verduras, la estatura de una persona o las ventas de una empresa. La estadística descriptiva utiliza métodos y herramientas estadísticas para analizar y representar cuantitativamente diversas características y patrones de los datos con el fin de facilitar la comprensión y explicación del fenómeno estudiado (López, 2022).

2.8.2 Análisis de correlación

Cuando se desea examinar la posible asociación entre variables, se recurre al análisis de correlación. Donde se distinguen las variables dependientes y las independientes. El objetivo de este análisis es determinar si cualquier cambio en la variable independiente provoca un cambio en la variable dependiente o si no existe una relación directa entre ambas. (DX Adobe, 2021).

2.8.3. Análisis de valores atípicos

Los valores atípicos son datos que difieren significativamente del resto. Pueden deberse a errores o a variaciones naturales. Es importante identificar los valores atípicos y decidir si mantenerlos o eliminarlos del conjunto de datos. (Yepes, 2022).

2.8.4. Análisis de valores faltantes

La falta de datos puede ser un problema en la investigación. Es importante tener en cuenta las razones por las que faltan datos y cómo pueden afectar a los resultados. Los programas pueden tratar los datos que faltan de forma incoherente, lo que puede dar lugar a errores en los resultados. También es habitual que los programas informáticos excluyan del análisis a los individuos con datos incompletos, lo que puede sesgar los resultados. (Bhandari, 2021).

CAPÍTULO III: APLICACIÓN DEL MÉTODO CRISP-DM

3.1. Comprensión del negocio

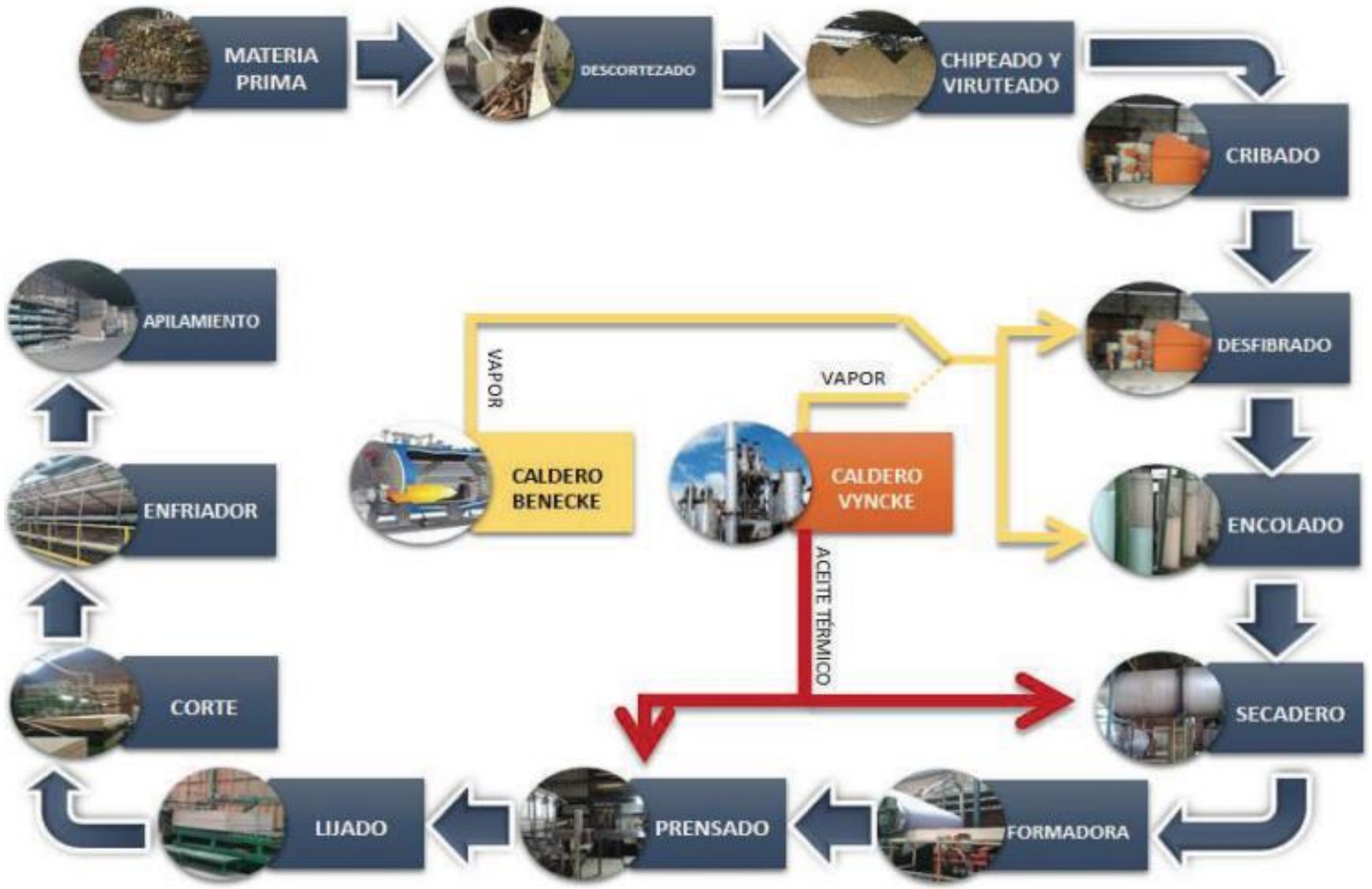
Aglomerados Cotopaxi S.A. es una compañía que se enfoca en actividades tanto industriales como forestales. Produce y vende tableros para su comercialización local e internacional, con tres líneas de producción diferentes: Aglomerados, MDF y Recubrimiento.

Tanto los tableros aglomerados como los tableros MDF están hechos de virutas de madera unidas con resina y prensadas bajo condiciones controladas. La diferencia principal reside en que el MDF utiliza partículas de madera de densidad media, lo que lo hace más estable frente a los cambios ambientales, mientras que el tablero aglomerado se compone de capas múltiples de partículas de madera, con las más gruesas en el centro y las más finas en la superficie.

Etapas del proceso de fabricación de tableros

La elaboración de tableros Aglomerados y MDF con o sin recubrimiento consta de etapas de proceso similares. En la empresa, se ha representado de forma resumida en la Figura 3 el proceso de producción para la fabricación de un tablero, abarcando desde la selección de la materia prima hasta el apilamiento del producto terminado.

Figura 3 Fases de Fabricación de los Tableros Aglomerados Cotopaxi S.A



Autor: Sarango, E. G. (2017). Modernización del sistema de control del caldero Vyncke utilizado para calentamiento de aceite térmico en la fábrica Aglomerados Cotopaxi S.A. (Trabajo de grado). Escuela Politécnica Nacional, Quito. [Imagen de la página 14].

1. **Materia prima:** Aglomerados Cotopaxi S.A. planta bosques de eucalipto y pino en sus propios terrenos y cosecha la madera después de 15 a 20 años. La madera se transporta a la planta industrial para fabricar diferentes tipos de aglomerado.

2. **Descortezado:** Los troncos se descortezan utilizando un tambor con cuchillas o una descortezadora. A continuación, los troncos se sierran y se separan en dos grupos: uno para tableros y vigas y otro para la producción de aglomerado.

3. **Chipeado y viruteado:** La madera obtenida en el aserradero se tritura para crear pequeñas partículas conocidas como astillas. Estas astillas se utilizan después para producir tableros aglomerados.

4. **Cribado:** Las partículas de madera se clasifican según su tamaño y características mediante una criba. Los residuos de este proceso se envían a la caldera de Vyncke como combustible.

5. **Desfibrado:** Proceso de eliminación de la anilina, un componente natural de la madera, para separar completamente las partículas de madera. Para ello, las astillas se hierven con vapor a alta presión en una caldera Benecke o Vyncke y, a continuación, se aprietan con un tornillo hasta eliminar completamente la anilina.

6. **Encolado:** El proceso de encolado es un paso importante en la producción de madera contrachapada, ya que garantiza que las capas de madera estén bien unidas entre sí. El tipo de cola utilizada variará en función del tipo de contrachapado que se esté fabricando, ya que algunos tipos de cola son más resistentes a la humedad o al calor que otros.

7. **Secadero:** Antes de continuar con la fabricación del tablero, la mezcla se seca calentando el aire a través de un serpentín con aceite térmico. A continuación, la mezcla se pesa en tolvas en función de la densidad del tablero que se va a fabricar.

8. **Formación:** Es la etapa en la que se da forma al tablero. Las partículas de madera se distribuyen mediante un control granulométrico y el proceso se realiza en una máquina denominada formadora. A continuación, el tablero o colchón se desplaza sobre una cinta transportadora antes de entrar en la prensa.

9. **Prensado:** En la fase de prensado, la planta dispone de prensas individuales para MDF y aglomerado, que controlan la presión, la temperatura y el tiempo para comprimir los colchones de partículas de madera y formar los tableros con el grosor requerido por el cliente. El aceite térmico utilizado para el funcionamiento de las prensas procede de la caldera Vyncke.

10. **Enfriamiento:** Tras el prensado, los tableros de partículas se inspeccionan y pesan para garantizar su calidad. Los que cumplen las normas pasan a la fase de enfriamiento, en la que se utiliza un volteador para facilitar el proceso.

11. **Corte:** Los tableros se cortan longitudinal y transversalmente con sierras de acuerdo con las dimensiones estándar para el tipo de tablero (aglomerado o MDF). Las dimensiones estándar para el aglomerado son 2150 x 2440 x 19 (mm) y para el MDF son 1830 x 2440 x 9 (mm).

12. **Lijadora:** El proceso de lijado consta de tres etapas para mejorar el pulido del tablero y reducir las virutas y los grumos. Tras la última etapa, el tablero se somete a una comprobación y control de calidad antes de ser recubierto según las especificaciones del cliente. Si no se solicita recubrimiento, los tableros se almacenan en depósitos tras el control de calidad.

13. **Recubrimiento:** Se utilizan tres prensas diferentes para aplicar papel melamínico a tableros de aglomerado o MDF. El papel se selecciona en función de la textura y el diseño especificados en el pedido y, a continuación, se adhiere al tablero mediante cola o papel con adhesivo. Tras el prensado, los tableros se enfrían y se apilan.

14. **Calderos:** Aglomerados Cotopaxi S.A. dispone de dos calderas, la Benecke y la Vyncke. La caldera Benecke suministra vapor para el desfibrado, el dimensionamiento y el tratamiento

del agua. La caldera Vyncke, con una capacidad de 10 Gcal/hora, se utiliza para calentar aceite térmico quemando biomasa.

3.1.1. Política, Visión, Misión y Objetivos del negocio

La Política del Sistema de Gestión Integrado SGI está definida en PO-001A que considera los aspectos y compromisos del Sistema de Gestión Integrado de Calidad, Ambiente, Seguridad y Salud Ocupacional, Control y Seguridad, y Certificación Forestal FSC.

Visión

"Ser líderes en la industria maderera regional, la mejor opción para nuestros clientes, y sinónimo de excelencia empresarial."

Misión

"Somos una industria forestal líder en el establecimiento de plantaciones, producción y comercialización de tableros de madera; comprometida con el desarrollo sustentable."

Objetivos del negocio

Aglomerados Cotopaxi busca el aseguramiento de la calidad y la mejora de los procesos y productos en todos los aspectos de sus operaciones. El cuidado de las personas y el medio ambiente se considera una parte integral del desarrollo de la empresa.

La información fue obtenida de documentos PDF de la página web de Aglomerados Cotopaxi S.A. y eco.business Fund

3.1.2. Recursos proporcionados

Para la realización del proyecto de minería, se tuvo los datos del sistema SCADA.

Inventario de recursos

Los recursos de hardware que se posee es una laptop con las siguientes características:

- **Marca:** Dell
- **Modelo:** XPS 15 7590
- **Procesador:** Intel(R) Core(TM) i7-9750H CPU @ 2.60GHz 2.59 GHz
- **Memoria instalada (RAM):** 8,00 GB
- **Tipo de sistema:** Sistema operativo de 64 bits, procesador x64
- **Sistema Operativo:** Microsoft Windows 11 Pro
- **Capacidad de almacenamiento:** 250GB
- **Tarjeta gráfica:** NVIDIA © GeForce 1650

Los recursos de software que se tiene son la base de datos *SQL Server Management Studio 18* y el lenguaje de programación (*Python*).

Requisitos, supuestos y restricciones

Debido a las restricciones de seguridad, no se tendrá acceso a todos los datos relacionados con la línea MDF. Sin embargo, se proporcionaron los datos necesarios para realizar correctamente el modelado.

3.1.3. Determinación de los objetivos de minería de datos

El objetivo principal en este proyecto es obtener información relevante del proceso de encolado para realizar las recomendaciones adecuadas. Utilizando las variables registradas en el subproceso de encolado, se plantean los siguientes objetivos de minería:

- Identificar y seleccionar las variables claves que influyen en el rendimiento del proceso de encolado, para establecer métricas para evaluar el impacto de cada variable.
- Definir y calcular las variables más influyentes en la predicción de las variables clave previamente calculadas.
- Usar técnicas de minería de datos para identificar patrones y brechas del proceso.

- Evaluar los modelos obtenidos para identificar cuál de ellos es el más adecuado y descartar los que no cumplen los requerimientos establecidos.

3.2. Comprensión de los datos

En esta fase, se accede a los datos y se exploran utilizando gráficos o tablas para evaluar su calidad. Se realiza una evaluación detallada para identificar cualquier problema o limitación que pueda afectar a su idoneidad para el análisis.

3.2.1. Recopilación de los datos iniciales

El sistema dispensador de cola recopila datos utilizando un PLC Siemens Simatic S7-300. Los datos se extraen y almacenan en una base de datos de Microsoft SQL Server mediante el software WinCC de Siemens. Para mejorar el acceso a los datos se utiliza un servidor de datos y un IoT 2040 de Siemens. Sin este paso, la base de datos WinCC sólo sería accesible a través de su propio software. Los datos se recopilaron durante una semana.

La máquina encoladora registró los datos cada segundo o cada dos segundos, estos están en formato UTF-16 LE BOM, por lo que se cambió el punto y coma por una coma y el punto decimal por un punto para facilitar el manejo de los datos.

Se facilitaron cuatro tablas: Visor_de_tablas_1_1, Visor_de_tablas_1_2, Visor_de_tablas_2_1 y Visor_de_tablas_2_2.

Las tablas “Visor_de_tablas_1_1” y “Visor_de_tablas_1_2” tienen los mismos campos, por lo que se fusionaron en una única tabla llamada **Proceso_Encolado_Dosificado**. La Tabla 1 muestra los detalles de los campos.

También, las tablas "Visor_de_tablas_2_1" y "Visor_de_tablas_2_2" tienen los mismos campos. Por consiguiente, se tomó la decisión de combinar estas tablas en una nueva llamada **Sp_Vp_Encolado_Dosificado**. Los detalles de los campos pueden encontrarse en la Tabla 2.

3.2.2. Descripción de datos

A continuación, se presentan los campos y sus descripciones de la Tabla 1 **Proceso_Encolado_Dosificado** y la Tabla 2 **Sp_Vp_Encolado_Dosificado**. Todos los datos son cuantitativos discretos, excepto la columna "columna tiempo" en ambas tablas, que es una cadena de caracteres.

Tabla 1 Descripción de los atributos de la tabla Proceso_Encolado_Dosificado

Nombre	Descripción
Columna de tiempo	Representa el momento en que se registraron los datos (hora, minutos, segundos, etc.).
Humedad	Representa la humedad en el ambiente durante el proceso de encolado.
Velocidad tornillo sin fin	Indica la velocidad del tornillo sin fin, que es un componente mecánico utilizado en el sistema de dosificación para mover y mezclar los materiales.
Presión resina	Indica la presión a la que se encuentra la resina durante el proceso de encolado.
Peso balanza	Representa el peso de los materiales en la balanza durante el proceso.
Peso tornillo	Indica el peso del tornillo sin fin en el proceso de dosificación.
Ciclo prensado	Representa el ciclo de prensado, que es el proceso en el que los materiales son comprimidos y compactados.
Consumo resina	Indica la cantidad de resina utilizada en el proceso de encolado.
Consumo resina real	Representa la cantidad real de resina utilizada en el proceso, posiblemente después de ajustes o correcciones.
kWh/Ton	Representa la cantidad de energía eléctrica (en kilovatios-hora) utilizada por tonelada de material procesado en el sistema.

Autor: Anahí Males

Tabla 2 Descripción de los atributos de la tabla Sp_Vp_Encolado_Dosificado

NOMBRE	**DESCRIPCIÓN**
Columna de tiempo	Representa el momento en que se registraron los datos (hora, minutos, segundos, etc.).
Vp COLA	Representa el valor actual del proceso de encolado (cola) en el sistema.
Sp Manual COLA	Representa el valor deseado (*set point*) manualmente establecido para la cola en el proceso.
Sp Auto COLA	Representa el valor deseado (*set point*) establecido automáticamente por el sistema para la cola en el proceso.
Vp UREA	Representa el valor actual del proceso de urea en el sistema.
Sp Manual UREA	Representa el valor deseado (*set point*) manualmente establecido para la urea en el proceso.
Sp Auto UREA	Representa el valor deseado (*set point*) establecido automáticamente por el sistema para la urea en el proceso.
Vp TINTE	Representa el valor actual del proceso de tinte en el sistema.
Sp Manual TINTE	Representa el valor deseado (*set point*) manualmente establecido para el tinte en el proceso.
Sp Auto TINTE	Representa el valor deseado (*set point*) establecido automáticamente por el sistema para el tinte en el proceso.
Vp EMULSION	Representa el valor actual del proceso de emulsión en el sistema.
Sp Manual EMULSION	Representa el valor deseado (*set point*) manualmente establecido para la emulsión en el proceso.
Sp Auto EMULSION	Representa el valor deseado (*set point*) establecido automáticamente por el sistema para la emulsión en el proceso.

Vp ENDURECEDOR	Representa el valor actual del proceso de endurecedor en el sistema.
Sp Manual ENDURECEDOR	Representa el valor deseado (*set point*) manualmente establecido para el endurecedor en el proceso.
Sp Auto ENDURECEDOR	Representa el valor deseado (*set point*) establecido automáticamente por el sistema para el endurecedor en el proceso.
Vp CERA	Representa el valor actual del proceso de cera en el sistema.
Sp Manual CERA	Representa el valor deseado (*set point*) manualmente establecido para la cera en el proceso.
Sp Auto CERA	Representa el valor deseado (*set point*) establecido automáticamente por el sistema para la cera en el proceso.

Autor: Anahí Males

3.2.3. Exploración de datos

Se usó estadística descriptiva, análisis de correlaciones, análisis de valores atípicos y análisis de valores perdidos para evaluar la coherencia e integridad de los datos de las Tablas 1 y 2.

Al crear las estructuras correspondientes, se detectó que algunos registros estaban vacíos. Por tanto, se decidió considerar cualquier celda con el valor "NA" o una celda vacía como un valor nulo.

3.2.3.1. Exploración de la Tabla 1

Estadística descriptiva

Se puede apreciar en el Anexo A, la información de las columnas de la tabla **Proceso_Encolado_Dosificado**, se incluye el número de valores por columna, la media, la desviación estándar, y los valores mínimo y máximo. Estos datos son útiles para comprender el proceso, identificar tendencias y detectar anomalías. Por lo que se puede afirmar que:

- El campo "Columna de tiempo", es tomado cada segundo.
- Las columnas: "Humedad" (4.45), "Velocidad tornillo sin fin" (15.41), "Peso tornillo" (30.32), "Consumo resina" (36.35), "Consumo resina real" (48.81), "kWh/Ton" (2345.24). Indican la variabilidad, desde la más baja hasta la más alta de la tabla.

Los valores mínimo y máximo de las columnas, como: "Peso tornillo" tiene un valor mínimo de 0.0 y un valor máximo de 105.96, los cuales son inusuales y requieren un análisis más detallado para determinar si se trata de errores, valores atípicos o condiciones extremas en el proceso de encolado. Esta evaluación identificará cualquier anomalía para poder tomar las medidas oportunas.

Análisis de correlación

La matriz de correlación de la tabla **Proceso_Encolado_Dosificado**. contiene atributos con una correlación positiva fuerte (con un valor entre 0.7 y 1). Algunos de estos atributos son: "Humedad" y "Peso balanza", "Humedad" y "Consumo resina", "Velocidad tornillo sin fin" y "Peso balanza". La columna "kWh/Ton" de la tabla **Proceso_Encolado_Dosificado**, se la excluyó debido a la falta de datos suficientes

Estas correlaciones ayudarán a elegir las variables más relevantes para predecir o analizar otras columnas. Para obtener más detalles, véase el Anexo B.

Análisis de valores atípicos

Con los valores obtenidos de las figuras en los Anexos A y C junto a los de la tabla **Proceso_Encolado_Dosificado**, se puede afirmar:

- El valor de la media (45.52) en la columna "Velocidad del tornillo sin fin" indica que la mayoría de las mediciones se encuentran cerca del valor máximo (53). Esto sugiere que el valor del primer cuartil tiene variabilidad a velocidades más bajas.

- La media y el promedio (9.46) en la columna “Presión resina” indica que las mediciones están bastante distribuidas en torno a estos valores.
- Los valores de la columna “Peso balanza” están relativamente cerca del promedio (43.97), revelando que la distribución no está muy sesgada hacia valores extremos o atípicos.
- En la columna “Peso tornillo”, aunque la mayoría de las mediciones están cerca del valor máximo, algunas de ellas tienen valores más bajos (0.0, 0.06, 0.14, 0.22, 0.3) y hay cierta dispersión en esos valores.
- Todos los valores de "Ciclo prensado" están en cero. Esto significa que no hay variación en esta característica. Por lo tanto, los datos de "Ciclo prensado" no son relevantes para el análisis.
- Las columnas: “Consumo resina” (94.78) y “Consumo resina real” (101.15) sugiere una distribución amplia en torno a su promedio.

Análisis de valores faltantes

En la columna “kWh/Ton” que tiene 679562 valores nulos, y en la columna "ciclo prensado", que tiene valores de cero, se eliminaron. Con esto se busca evitar distorsiones en los futuros resultados.

3.2.3.2. Exploración de la Tabla 2

Estadística descriptiva

Las tablas indicadas en la figura del Anexo D muestran el recuento, la media, la desviación típica y los valores mínimo y máximo de cada columna de la tabla **Sp_Vp_Encolado_Dosificado**. Esta información puede utilizarse para comprender el proceso, identificar patrones y detectar irregularidades. Por lo que se puede decir que:

- El campo "Columna de tiempo", es tomado cada dos segundo.

- En algunos atributos de la tabla como: "Vp COLA", "Sp Auto COLA", "Sp Auto UREA", "Sp Auto TINTE", "Sp Auto ENDURECEDOR", "Sp Auto CERA"; existen variaciones significativas en los datos, sobre todo en los extremos. Esto podría sugerir la presencia de valores atípicos.
- Algunas columnas como: "Vp UREA", "Vp TINTE", "Vp ENDURECEDOR", "Vp CERA"; muestran variabilidad, lo que indica una diversidad de valores dentro del conjunto de datos. Esto es una buena señal, ya que significa que los datos no son todos iguales.
- Otros atributos como: "Sp Auto EMULSION", "Sp Manual CERA", "Vp EMULSION", "Sp Manual TINTE", "Sp Manual ENDURECEDOR", "Sp Manual UREA"; tienen una varianza y una desviación estándar de cero, lo que podría indicar un posible error durante la recopilación de datos.

Análisis de correlación

La matriz de correlación de la tabla **Sp_Vp_Encolado_Dosificado** se muestra en la figura del Anexo E**,** donde existen columnas con una correlación positiva fuerte (valores), por ejemplo: "Vp COLA" y "Sp Auto COLA", "Vp ENDURECEDOR" y "Sp Auto ENDURECEDOR", "Vp CERA" y "Sp Auto CERA".

Indicando la fuerte correlación entre el valor deseado (Sp) en modo automático y el valor actual del proceso (Vp), lo que indica que el proceso de producción en modo automático funciona eficazmente en estos casos.

Análisis de valores atípicos

Con los resultados obtenidos de la figura del Anexo F, es posible agrupar los atributos para identificar si serán útiles para la siguiente fase, dado que presentan características similares, como:

- Columnas con valores atípicos, con una distribución sesgada hacia la derecha: “Vp UREA”, “Vp TINTE”, “Sp Auto TINTE”, “Vp ENDURECEDOR”, “Vp CERA”, “Sp Auto CERA”. Lo que apoyará a detectar patrones o anomalías.
- Columnas con límites normales y sin variabilidad, que no proporcionan información útil para el análisis: “Sp Manual COLA”, “Sp Manual UREA”, “Sp Manual TINTE”, “Vp EMULSION”, “Sp Auto EMULSION”, “Sp Manual CERA”, “Sp Manual EMULSION”, “Sp Manual ENDURECEDOR”.
- Columnas con valores dentro del rango esperado, lo que ayudará a la siguiente fase:
    a) Con una distribución ligeramente sesgada hacia la izquierda: “Vp COLA” y, “Sp Auto COLA”.
    b) Con una distribución sesgada hacia la derecha: “Sp Auto UREA” y “Sp Auto ENDURECEDOR”.

Los atributos que presentan una gran cantidad de valores atípicos en la tabla **Sp_Vp_Encolado_Dosificado** serán analizados para determinar si se trata de errores de medición o si representan condiciones específicas relevantes para el estudio.

Análisis de valores faltantes

La tabla **Sp_Vp_Encolado_Dosificado** tiene varios valores nulos en las columnas “Sp Manual” porque la máquina funciona principalmente en modo Automático. Estas columnas se eliminaron para evitar problemas en la siguiente fase del análisis, para más información, véase las figuras de los Anexos G y H.

Algunas columnas de la tabla **Sp_Vp_Encolado_Dosificado** tienen un elevado número de ceros, lo que indica una baja densidad de datos o una gran dispersión de la información. Esto es problemático para el análisis, ya que lleva a la pérdida de información importante y a la generación de sesgos en el modelo de análisis.

3.3. Preparación de los datos

Durante esta fase, se preparó los datos para la etapa de modelado, seleccionando y contrayendo nuevas estructuras según fuera necesario.

3.3.1. Selección de datos

En la selección de datos, se identifican y seleccionan los datos relevantes y necesarios para el análisis, descartando los que no son útiles o que pueden generar errores o sesgos en el modelo de análisis. Los campos elegidos para el análisis son los siguientes:

- En la tabla **Proceso_Encolado_Dosificado:** "Columna de tiempo", "Humedad", "Velocidad tornillo sin fin", "Presión resina", "Peso balanza", "Peso tornillo", "Consumo resina"," Consumo resina real".
- En la tabla **Sp_Vp_Encolado_Dosificado:** "Vp COLA", "Sp Auto COLA", "Vp UREA", "Sp Auto UREA", "Vp TINTE", "Sp Auto TINTE", "Vp ENDURECEDOR", "Sp Auto ENDURECEDOR", "Vp CERA", "Sp Auto CERA".

3.3.3. Construcción de datos

De acuerdo, con los campos seleccionados de las tablas **Proceso_Encolado_Dosificado** y **Sp_Vp_Encolado_Dosificado**, se combinaron utilizando el atributo "Columna de tiempo" para crear una nueva tabla denominada **Tabla_Encolado_Dosificado**. Donde, solo se consideraron los registros tomados después de dos segundos y que no tuvieran valores nulos en al menos tres columnas consecutivas.

CAPÍTULO IV: MODELADO

En esta fase, se seleccionan las técnicas para alcanzar los objetivos de minería propuestos. Asimismo, se genera un diseño de comprobación para aplicar las técnicas sobre los datos y generar el modelo seleccionado. Posteriormente, se evaluará el modelo para determinar si cumple con los criterios establecidos.

4.1. Selección de técnicas de modelado

Para el modelado, se utilizó la tabla **Tabla_Encolado_Dosificado** y *Python*, con la librería *sklearn*. Esta librería cuenta con algoritmos de clasificación, regresión y análisis de grupos, y está diseñada para trabajar en conjunto con otras librerías numéricas y científicas.

Debido a que el objetivo del proyecto es analizar los datos usando minería de datos para la elaboración de recomendaciones en el proceso de dosificación, se decidió escoger las siguientes técnicas: Regresión lineal, Regresión LASSO, *K-means* y *DBSCAN*.

4.2. Generación de un diseño de comprobación

El procedimiento que se usa para probar la validez del modelo implica el análisis de diversos indicadores, los cuales son:

- **Regresión lineal y LASSO:** El coeficiente de determinación (R^2), el error cuadrático medio (MSE) y el error absoluto medio (MAE).
- **K-means:** Se utilizarán las siguientes métricas para evaluar el rendimiento del clustering: WCSS, el coeficiente de Silueta y el Índice de Davies-Bouldin.
- **DBSCAN:** Al igual que en K-means, se tomará la Silueta y el Índice de Davies-Bouldin.

4.3. Generación de los modelos

A continuación, se realiza la implementación de los modelos escogidos, utilizando la librería Scikit-learn en Python, donde se importan las bibliotecas necesarias y se carga el conjunto de

datos (**Tabla_Encolado_Dosificado**) en un *DataFrame pandas*. La columna "Columna de tiempo" se cambia a un formato *datetime* y se mapea a un valor ordinal, que es una forma de convertir fechas a números que puedan ser interpretados por el modelo.

Regresión lineal

Se usó el algoritmo de regresión logística para crear el modelo, con diferentes tres variables dependientes: "Presión resina", "Consumo resina real" y "Vp ENDURECEDOR". También, se trabajó con los parámetros ya establecidos por la librería como:

- El conjunto de datos se dividió en conjuntos de entrenamiento y de prueba utilizando la función *train_test_split.*
- Se creó un objeto de regresión lineal utilizando la clase *LinearRegression*.
- El modelo de regresión lineal se entrenó utilizando el conjunto de entrenamiento (X_train e y_train).
- Los coeficientes del modelo se obtuvieron utilizando el atributo coef_ del objeto reg.
- Las predicciones se realizaron utilizando el conjunto de prueba (X_test).

Los coeficientes de regresión lineal para cada característica en el modelo entrenado se encuentran disponibles en la Figura 4.

Al interpretar los coeficientes, es importante considerar el signo y la magnitud para comprender cómo afectan las variables independientes cada variable dependiente en el contexto del modelo. Por ejemplo, en la figura 4 "VP UREA" tiene un coeficiente de -1.003, en relación con "Presión resina". Esto implica que, manteniendo todas las demás variables constantes, un aumento de una unidad en la variable "VP UREA" se asocia con una disminución de aproximadamente 1.003 unidades en la “Presión resina”.

Figura 4 Coeficientes más influyentes del modelo de regresión lineal de los atributos: "Presion resina", "Consumo resina real", "Vp ENDURECEDOR"



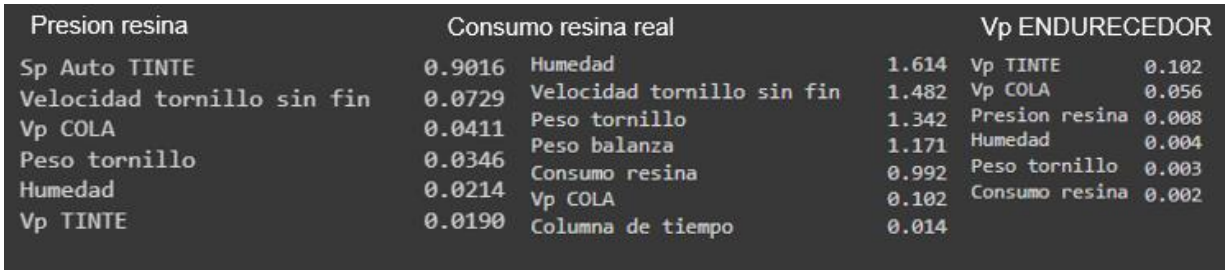
Fuente: Anahí Males

Regresión LASSO

Para la implementación se utilizó el algoritmo de regresión LassoCV para crear el modelo. En este caso, se empleó la validación cruzada para seleccionar el mejor parámetro de regularización.

Como resultado se obtuvo las características con mayor impacto en las variables seleccionadas. Por ejemplo, la variable "Sp Auto TINTE" tiene un coeficiente elevado que afecta a la variable "Presion resina". Esto significa que un aumento en la variable "Sp Auto TINTE" está asociado a un aumento de la "Presion resina".

Figura 5 Importancias de las características del modelo Lasso ajustado de los atributos: "Presion resina", "Consumo resina real", "Vp ENDURECEDOR"



Fuente: Anahí Males

Para los siguientes dos algoritmos, por la gran cantidad de registros se realizó un muestreo aleatorio del 10% de los datos del *DataFrame* por qué es un buen equilibrio entre obtener una muestra representativa de los datos y no tardar demasiado en procesarlos.

K-means

Para la implementación se usó el algoritmo *K-Means* y la agrupación de datos en tres *clúster*es. Donde el método del codo nos dio el número de *clúster*es. El resultado muestra que, el *clúster* 0 tiene15731, el *clúster* 1 tiene 15582 registros y el *clúster* 2 tiene 3241. Basándonos en las características de los tres *clúster*es obtenidos, podemos decir que:

***Clúster* 0:** Este grupo se caracteriza por una "humedad" y una "Presion resina" moderadas, un "Consumo resina" y un "Peso tornillo" elevados, y unos valores "Vp COLA" y "Sp Auto COLA" coherentes. También contiene el mayor número de registros, lo que sugiere una mayor frecuencia de operaciones o una distribución más amplia de los datos. Este *clúster* podría representar un conjunto de puntos de datos relacionados con un proceso de producción de alto rendimiento con alta eficiencia y producción.

***Clúster* 1:** Este *clúster* es similar al *Clúster* 0, con valores ligeramente superiores de "humedad", "consumo de resina", "peso del tornillo", "Vp COLA" y "Sp Auto COLA". También contiene un número similar de registros. Sin embargo, tiene un valor ligeramente inferior para "Vp COLA", lo que sugiere que el producto producido en este *clúster* es de un rendimiento ligeramente inferior.

***Clúster* 2:** Este *clúster* se caracteriza por una baja "Humedad", "Presion resina", consumo de resina, peso del tornillo y valores "Vp COLA" y "Sp Auto COLA". También contiene un menor número de registros que los demás *clúster*es, lo que sugiere una menor frecuencia de operaciones o una distribución más limitada de los datos. Esta agrupación podría representar un conjunto de puntos de datos relacionados con un proceso de producción de bajo rendimiento. El bajo nivel de eficiencia, la pequeña cantidad de producción y la baja calidad del producto sugieren que el proceso no está funcionando a un nivel óptimo.

DBSCAN

Para implementar el algoritmo de DBSCAN y crear el modelo. Se tuvieron que definir los parámetros del algoritmo. Donde, el número de *clúster*es encontrados fue 577.

El algoritmo encontró *clúster*es con un amplio rango de densidades. El *clúster* más denso tenía una densidad de 22, mientras que el *clúster* menos denso tenía una densidad de 5. Esto sugiere que los datos no están distribuidos uniformemente y puede que no sean adecuados para el clustering. Dado que, algunos conglomerados serán muy densos, mientras que otros serán muy dispersos. Esto puede dificultar la interpretación de los resultados de DBSCAN.

4.4. Evaluar los modelos

En esta sección se analiza los resultados obtenidos. Considerando las métricas de rendimiento y la capacidad de selección de las características, para más información de los valores obtenidos véase las tablas 3 y 4.

El modelo de regresión lineal con el atributo "Presion resina" captura una parte significativa de la variación en los datos, pero presenta cierto nivel de error en las predicciones.

Con el campo "Consumo resina real" muestra un ajuste razonable, explicando alrededor del 62.5% de la variación en los datos. Aunque el modelo presenta cierto error, tanto el MSE como el MAE indican que aún puede realizar predicciones precisas.

Con “Vp ENDURECEDOR”, R^2 indica que la variación de la variable dependiente puede explicarse por la variable independiente. El MSE y el MAE también son bajos, mostrando que el modelo realiza predicciones precisas. En general, el modelo se ajusta muy bien a los datos, además que funciona bien y es fiable.

El segundo modelo de regresión LASSO aplicado a "Presion resina" explica el 45.1% de la varianza, pero presenta un error alto (MSE) y no es muy preciso (MAE). En resumen, el modelo tiene limitaciones en precisión y ajuste.

Para el atributo "Consumo resina real" explica el 61.7% de la variación de los datos. Aunque hay margen de mejora, el modelo tiene un buen desempeño en la predicción de valores de resina en el rango medio.

Para Vp ENDURECEDOR muestra un ajuste muy bueno, ya que explica el 93.1% de la variación de los datos. Además, presenta bajos valores de MSE y MAE, lo que indica una baja cantidad de errores en las predicciones. En resumen, el modelo muestra una buena capacidad de predicción para los valores de Vp ENDURECEDOR.

Tabla 3 Métricas de evaluación Regresión Lineal y LASSO de: "Presion resina", "Consumo resina real", "Vp ENDURECEDOR"

Atributos	Métricas	Regresión Lineal	Regresión LASSO
Presion resina	R^2	46.1%	45.1%
	MSE	0.828	0.844
	MAE	0.404	0.422
Consumo resina real	R^2	62.5%	61.7%
	MSE	888.056	907.272
	MAE	12.351	12.177
Vp ENDURECEDOR	R^2	96.4%	93.1%
	MSE	0.008	0.016
	MAE	0.050	0.086

Fuente: Anahí Males

El algoritmo *k-means* indica que los puntos de datos están bien agrupados. El valor de la silueta también es relativamente alto, mostrando que los puntos de los datos están bien asignados a sus conglomerados. Sin embargo, el índice de Davies-Bouldin es alto, lo que significa que los conglomerados no están perfectamente separados.

Los resultados de DBSCAN indican que los datos no están perfectamente divididos en *clúster*es. Sin embargo, los conglomerados tampoco están demasiado mal separados. El valor

de silueta indica que los puntos de datos están algo relacionados con los demás puntos de datos de su *clúster*. El índice de Davies-Bouldin indica que los grupos están algo separados.

Tabla 4 Métricas de evaluación de K-means y DBSCAN

Métricas	K-means	DBSCAN
Número de *clúster*es	3	577
WCSS	118398359.401	
Silueta	0.493	0.408
Índice de Davies-Bouldin	0.734	0.435

Fuente: Anahí Males

CAPÍTULO V: EVALUACIÓN

5.1. Evaluación de los resultados

Durante esta etapa, se evalúa la calidad de los modelos de minería generados y se selecciona el modelo que mejor se ajuste a los objetivos de minería propuestos.

Modelo Regresión lineal: El modelo muestra un alto grado de precisión al predecir el atributo "Vp ENDURECEDOR", pero no para los atributos "Presion resina" y "Consumo resina real". Esto sugiere que el modelo puede tener dificultades para captar las relaciones complejas entre las características y la variable objetivo en estos atributos. Por lo que es posible que se requiera un modelo más sofisticado.

Modelo Regresión LASSO: Los resultados del modelo son mixtos. El "Vp ENDURECEDOR" se predice con precisión. Sin embargo, "Presion resina" y "Consumo resina real" no se predicen tan bien, con valores de R^2 de 45.1% y 61.7% respectivamente, y mayores valores de MSE. Esto puede deberse a la limitación del modelo para predecirlos.

Modelo K-means: Según estos resultados, el modelo K-means ha realizado un buen trabajo de agrupación de los datos. Los *clúster*es están muy agrupados, los puntos de datos están bien asignados a sus *clúster*es y los *clúster*es están espaciados uniformemente. Sin embargo, es importante tener en cuenta que estos resultados se basan en una muestra del conjunto de datos. Es posible que el modelo funcione de forma distinta en un conjunto de datos diferente.

Modelo DBSCAN: El modelo DBSCAN ha identificado 577 *clúster*es en los datos. El coeficiente de silueta es de 0.408, lo que indica que los conglomerados no están perfectamente separados. El índice de Davies-Bouldin es de 0.435, lo que indica que los conglomerados no están bien separados. En general, el modelo DBSCAN no parece ajustarse bien a estos datos.

Según estas métricas, parece que el modelo DBSCAN no se ajusta bien a los datos. Dado que, los conglomerados no están perfectamente separados. Esto podría deberse a la naturaleza de los datos.

Alineación entre los objetivos de minería y los modelos obtenidos

Objetivo 1: El modelo que más puede ser considerado factible es el modelo de Regresión LASSO. Dado que, sus métricas de evaluación están dentro del porcentaje aceptable (60%) y se obtuvo las variables clave que influyen en el proceso de encolado que son: “Consumo resina real” y “Vp ENDURECEDOR”

Objetivo 2: El modelo de Regresión LASSO es factible para identificar las variables más relevantes. Según el modelo LassoCV, 'Vp TINTE', 'Vp COLA' y 'Presion resina' son las variables más importantes para predecir 'Vp ENDURECEDOR'. Estas variables pueden ser consideradas como KPIs si son significativas y controlables para el proceso. Utilizar esta información en la definición de KPIs puede ayudar a monitorear y optimizar el proceso de producción.

El modelo LassoCV identificó las variables más relevantes para predecir 'Consumo resina real' como 'Humedad', 'Velocidad tornillo sin fin', 'Peso tornillo', 'Peso balanza' y 'Consumo resina'. Si estas variables son controlables y están relacionadas con los objetivos del negocio, podrían considerarse como KPIs. Por ejemplo, si se desea optimizar 'Vp ENDURECEDOR' y se tiene control sobre 'Vp TINTE', 'Vp COLA' y 'Presion resina', podría usarlos como KPIs y monitorearlos para asegurar el logro de tus objetivos.

El modelo LassoCV identificó las siguientes variables como las más relevantes para predecir la 'Presion resina': 'Sp Auto TINTE', 'Velocidad tornillo sin fin', 'Vp COLA', 'Peso tornillo' y 'Humedad'. Si estas variables son significativas para el proceso y se puede influir en ellas, podrían considerarse como KPIs importantes en el proceso.

Objetivo 3: El modelo aceptado es K-means, ya que sus métricas evaluadas superan las expectativas esperadas y se lograron identificar patrones y brechas del proceso, donde el *clúster* ideal es el 0, con menos desperdicios.

Objetivo 4: Dadas las métricas por cada modelo utilizado, se pudo identificar cuáles son los modelos que más se alineaban a los objetivos propuestos, que fueron: Regresión LASSO y K-means.

5.2. Procesos de revisión

EL proceso se ejecutó correctamente. Sin embargo, hubo algunas dificultades dentro de la fase 3 de la metodología, dado que había demasiada información que necesitaba ser procesada, consolidada y limpiada para la fase de modelado.

Además, se utilizó otras bibliotecas de Python durante la fase de preparación de los datos como: pyspark debido a la gran cantidad de información disponible, ya que en ocasiones el tiempo de procesamiento resultaba lento.

5.3. Desarrollo del documento con las recomendaciones

Aglomerados Cotopaxi S.A.

Fecha: 22 de mayo, 2023

Para: Aglomerados Cotopaxi S.A.

Asunto: Recomendaciones basadas en el proyecto de análisis de eficiencia operativa

Estimado [Nombre del destinatario],

Introducción

El propósito de este documento es presentar las recomendaciones derivadas del reciente proyecto de análisis y minería de datos del proceso de encolado de Tableros de MDF, llevado a cabo en su empresa durante los últimos cuatro meses.

Descripción detallada del proyecto

En el proyecto reciente, se aplicó minería de datos a los datos obtenidos del sistema SCADA en su proceso de dosificación, con un enfoque específico en el subproceso de encolado. Aunque su organización ya captura grandes volúmenes de datos a través del sistema SCADA, se identificó que este valioso recurso no se estaba utilizando plenamente, ya que los datos no se estaban analizando ni se estaban aplicando técnicas de minería de datos.

El análisis y la minería de datos brindaron respuestas valiosas para mejorar el proceso de encolado. Mediante el modelo Lasso, se aplicaron técnicas de regularización que permitieron seleccionar y reducir las variables predictoras, lo que mejoró la interpretación y generalización del modelo. Por otro lado, el algoritmo *K-means* se utilizó para agrupar los datos en *clúster*es, identificando patrones y características similares. Esto proporcionó una comprensión más profunda de las estructuras presentes en los datos. Ambos enfoques son herramientas poderosas que pueden contribuir significativamente al proceso de encolado.

Recomendaciones

Basándonos en los resultados obtenidos de la regresión Lasso y el agrupamiento K-means, sugerimos las siguientes estrategias para optimizar el proceso de encolado en su empresa:

Es de suma importancia realizar un seguimiento constante de los siguientes indicadores clave de rendimiento (KPI): Humedad, Velocidad del tornillo sinfín, Peso del tornillo, Peso de equilibrio, Consumo de resina. Si se detectan variaciones de estos KPI fuera del valor previsto, se deben aplicar medidas correctivas inmediatas para rectificar la situación.

A partir de los hallazgos obtenidos del método de agrupación k-means, proporcionamos las siguientes recomendaciones referentes a los valores ideales de los atributos antes mencionados:

El *clúster* 0 se caracteriza por ser un proceso de producción de alto rendimiento y eficiencia. Este agrupamiento se distingue por presentar una humedad y presión de resina moderadas, un consumo de resina y un peso del tornillo elevados, así como valores coherentes de Vp COLA y Sp Auto COLA. Los valores óptimos que se recomienda mantener son:

- Humedad: 15.10
- Velocidad tornillo sin fin: 49.42
- Presión resina: 9.77
- Peso balanza: 46.45
- Peso tornillo 102.88
- Consumo resina: 103.07
- Consumo resina real: 111.89
- Vp COLA: 16.43
- Sp Auto COLA: 15.79
- Vp UREA: 0.36
- Sp Auto UREA: 0.37
- Vp TINTE: 0.270
- Sp Auto TINTE: 0.274
- Vp ENDURECEDOR: 1.031

- Sp Auto ENDURECEDOR: 1.049
- Vp CERA: 0.64
- Sp Auto CERA: 0.65

Estos parámetros pueden ser referenciados para optimizar el proceso. Por ejemplo, incrementar la humedad a 15.10 podría mejorar la calidad del producto. Simultáneamente, disminuir la velocidad del husillo a 49.42 podría generar un ahorro en el consumo energético.

Es destacable que este *clúster* presenta un alto consumo de resina, indicando una gran utilización de materias primas. Esto puede optimizarse aplicando técnicas de producción más eficientes.

El *clúster* 2, por otro lado, representa un proceso de producción de bajo rendimiento y eficiencia. Donde se sugiere no llegar a estos valores:

- Humedad: 1.29
- Velocidad tornillo sin fin: 0.42
- Presión resina: 7.90
- Peso de la balanza: 16.27
- Peso tornillo: 1.42
- Consumo resina: 1.53
- Consumo resina real: 1.74
- Vp COLA : 0.026
- Sp Auto COLA: 12.58
- Vp UREA: 0.0
- Sp Auto UREA: 0.005319
- Vp TINTE: 0.0
- Sp Auto TINTE: 0.002069
- Vp ENDURECEDOR: 0.0
- Sp Auto ENDURECEDOR: 0.015505
- Vp CERA: 0.0
- Sp Auto CERA: 0.008673

Monitorear de cerca estos KPI y aplicar las acciones correctivas pertinentes permitirá optimizar el proceso de encolado para lograr los resultados deseados.

Siguiendo estas recomendaciones, la empresa puede mejorar el proceso de encolado y alcanzar sus objetivos empresariales.

Gracias por su tiempo y consideración. Si tiene alguna pregunta o necesita aclaraciones adicionales, no dude en ponerse en contacto conmigo.

ATENTAMENTE,

ANAHÍ MICAELA MALES MALDONADO

Estudiante de ingeniería en sistemas de la información

CONCLUSIONES Y RECOMENDACIONES

Conclusiones

- Se utilizó el modelo de regresión LASSO para analizar los datos obtenidos del sistema SCADA. El modelo pudo identificar las variables clave que influyen en el proceso de dimensionamiento, que son "Consumo real de resina" y "Vp de endurecedor". Esta información se utilizó para optimizar el proceso de dimensionamiento.
- Se evidenció al modelo K-means como una valiosa herramienta para poder mejorar la eficacia del proceso. Al identificar patrones y brechas del modelo puede utilizarse para realizar ajustes en el proceso con el fin de evitar problemas, mejorar la precisión y aumentar la eficiencia.
- En este proyecto, se pudo concluir que el uso de Google Colab y Python, junto con las librerías como Pandas, Scikit-Learn y PySpark, desempeñó un papel fundamental al maximizar el poder de cómputo disponible. Estas herramientas permitieron manejar eficientemente grandes volúmenes de datos y aplicar algoritmos avanzados de análisis.
- La exploración de los datos proporciono una comprensión profunda de los datos. Ayudó a identificar problemas de calidad, correlaciones y características relevantes de estos. Esta información fue utilizada para desarrollar modelos más precisos, predictivos e interpretables.
- La fase de evaluación determinó la efectividad del modelo, obteniendo conocimientos valiosos de los modelos de regresión LASSO y K-means. Por ejemplo, se identificaron los atributos que tienen una influencia significativa en otros dentro del subproceso de encolado, así como el grupo del conjunto de datos que mostró un mejor rendimiento en ese subproceso. Estos hallazgos son fundamentales, ya que proporcionan una base sólida para la formulación de posibles recomendaciones a la empresa.

- Durante la evaluación, se identificaron problemas o limitaciones en los modelos, revelando que no todos los modelos eran apropiados para un mismo problema debido a diversos factores. Estos factores incluyen la naturaleza del problema, la estructura y formato de los datos utilizados, así como consideraciones de eficiencia y escalabilidad computacional.

Recomendaciones

- Como trabajo futuro, se recomienda realizar un análisis de minería de datos utilizando un conjunto de datos recopilados durante un periodo más prolongado, como un mes completo. Ya que, al abarcar un periodo más largo, es posible identificar variaciones estacionales o cambios a lo largo del tiempo que podrían no ser evidentes en un análisis de datos más corto. Además de usar diferentes técnicas de minería para identificar si existe una mejor que las expuestas.
- Es fundamental que se comprenda el giro del negocio para lograr el éxito en cualquier proyecto de minería de datos. Al comprender el negocio, se podrán identificar los datos relevantes, aplicar las técnicas de minería adecuadas y comunicar los resultados de la mejor manera posible.
- Se recomienda usar los modelos de Regresión LASSO y K-means si los datos de algún proyecto tienen una naturaleza similar. Dado que, estas herramientas han demostrado ser valiosas para una posible mejorar en la eficiencia del proceso. Con estos modelos, se podrá determinar las variables clave que afectan el proceso y descubrir patrones y anomalías.
- Se sugiere optimizar el modelo de regresión LASSO para el proceso de dimensionamiento. Esto se logra al identificar las variables clave que afectan dicho proceso y realizar ajustes basados en la información proporcionada por el modelo. Esta optimización puede conducir a mejoras significativas en la calidad y eficacia de los resultados.

BIBLIOGRFÍA

1. DX Adobe. (2021, March 26). What is correlation analysis? Business Adobe. https://business.adobe.com/blog/basics/correlation-analysis
2. Agarwal, A., & Srivastava, J. (2014). Data mining: Concepts and techniques. Springer.
3. Aglomerados Cotopaxi. (s.f.). Plan de manejo forestal [Archivo PDF]. Recuperado de https://www.cotopaxi.com.ec/sites/default/files/2018-01/plan%20de%20manejo%20forestal.pdf
4. Akhtar, F. (2021, January). What is WCSS? CSIAS. https://www.csias.in/what-is-wcss/
5. Alexandre. (2022, diciembre 19). Pandas: La biblioteca de Python dedicada a la Data Science. Formation Data Science | DataScientest.com. https://datascientest.com/es/pandas-python
6. Arteaga, H. C. (2015). Técnicas de aprendizaje supervisado y no supervisado para el aprendizaje automatizado de computadoras. In Memorias del primer Congreso Internacional de Ciencias Pedagógicas: Por una educación integral, participativa e incluyente (pp. 549-564). Instituto Superior Tecnológico Bolivariano.
7. Bhandari, P. (2021, December 8). Missing Data. Scribbr. Obtenido de https://www.scribbr.com/statistics/missing-data/
8. Castellano, E. (2023, enero 14). ≫ ¿Qué es la definición de regresión LASSO, ejemplos y técnicas? Inteligencia Artificial. https://artificialgaming.net/que-es-la-definicion-de-regresion-lasso-ejemplos-y-tecnicas/
9. Castellano, J. (2023). Lasso Regression. In A. S. A. Abdel-Moniem (Ed.), Machine Learning: A Practical Approach (pp. 123-136). CRC Press.
10. Cómo calcular el porcentaje de error. (2019, March 4). Geniolandia. https://www.geniolandia.com/13075056/como-calcular-el-porcentaje-de-error
11. ECOBusiness Fund. (2020). Cotopaxi Client Story [Archivo PDF]. Recuperado de https://www.ecobusiness.fund/fileadmin/user_upload/Publications/spanish/Cotopaxi_Client_Story_ES_final.pdf
12. Ferri Ramírez, C., Hernández Orallo, J., & Ramírez Quintana, M. J. (2007). Introducción a la minería de datos. Pearson Educación.
13. Gironés Roig, J., Casas Roma, J., Minguillón Alfonso, J., & Caihuelas Quiles, R. (2017). Minería de datos modelos y algoritmos. Editorial UOC.
14. Hernández Orallo, J., De la Rosa, J. L., & Herrera, F. (2007). Data mining in environmental science: A review. Environmental Modelling & Software, 22(10), 1537-1554.
15. Hernández Orallo, J., Ramírez Quintana, M. J., & Ferri Ramírez, C. (2007). Introducción a la Minería de Datos. Pearson Educación.
16. IBM Documentation. (2021, agosto 17). https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview
17. IBM Documentation. (2021, March 17). CRISP-DM: Cross Industry Standard Process for Data Mining. Obtenido de https://www.ibm.com/docs/en/SS3RA7_sub/modeler_crispdm_ddita/clementine/crisp_help/crisp_overview.html
18. Ivan, R. J. O. (2019). Key Performance Indicators (KPI).

19. Marcasco. (2022, diciembre 1). Regresión múltiple: Definición, usos y 5 ejemplos - Historiadelaempresa.com. Historiadelaempresa.com. https://historiadelaempresa.com/regresion-multiple
20. Marr, B. (2012). KPIs: The Metrics That Drive Performance. Pearson Education.
21. Moreno Muñoz, A., & Córcoles Córcoles, S. (2018). Aprende Python en un Fin de Semana (Independently Published).
22. Parmenter, D. (2010). Key Performance Indicators: Developing, Implementing, and Using Winning KPIs. John Wiley & Sons.
23. Peña, J. (2017). El análisis de datos: Un enfoque sistemático. Investigación empresarial, 10(3), 1029-1048.
24. Peña, S. (2017). Análisis de datos. Bogotá: AREANDINA. Fundación Universitaria del Área Andina.
25. Pérez López, C., & Santin González, D. (2007). Minería de datos. Técnicas y herramientas: técnicas y herramientas. Ediciones Paraninfo, SA.
26. ProbabilidadyEstadística.net. (2023). Regresión lineal múltiple. Probabilidad Y Estadística. https://www.probabilidadyestadistica.net/regresion-lineal-multiple/
27. Ramirez, J. (2023). K-means clustering. In A. S. A. Abdel-Moniem (Ed.), Machine Learning: A Practical Approach (pp. 123-136). CRC Press.
28. Ramírez, L. (2023, enero 5). Algoritmo k-means: ¿Qué es y cómo funciona? Thinking for Innovation. https://www.iebschool.com/blog/algoritmo-k-means-que-es-y-como-funciona-big-data/
29. Rédac, T. (2022, August 2). PySpark: Todo sobre la biblioteca Python. Formation Data Science | DataScientest.com. https://datascientest.com/es/pyspark-todo-sobre-la-biblioteca-python
30. Rédac, T. (2022, September 1). Scikit-Learn: Descubre la biblioteca de Python dedicada al Machine Learning. Formation Data Science | DataScientest.com. https://datascientest.com/es/scikit-learn-decubre-la-biblioteca-python
31. Redacción Tokio. (2022, August 26). Google Colab, una nueva herramienta de Google pensaba para especialistas de IA y Data Analysis. Tokio School. https://www.tokioschool.com/noticias/google-colab/
32. Sánchez, J. A. (2020). ¿Cómo aprenden las máquinas? Machine Learning y sus diferentes tipos. datos. gob. es, 31.
33. Santamaría, J., & Hernández, J. (2016). Microsoft SQL Server. SQL SER vs MY SQL, 1-6.
34. Sanz, F. (2023). Algoritmo K-Means Clustering – aplicaciones y desventajas. The Machine Learners. https://www.themachinelearners.com/k-means/
35. Sanz, M. (2023). K-means clustering. In J. R. Quinlan (Ed.), Data Mining: Concepts and Techniques (3rd ed., pp. 253-276). Morgan Kaufmann.
36. Sharma, A. (2022). How to Master the Popular DBSCAN Clustering Algorithm for Machine Learning. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/09/how-dbscan-clustering-works/
37. Sharma, S. (2022). DBSCAN: A density-based clustering algorithm. In A. S. A. Abdel-Moniem (Ed.), Machine Learning: A Practical Approach (pp. 123-136). CRC Press.
38. Yepes, V. (2022, febrero 21). ¿Qué hacemos con los valores atípicos (outliers)? – El blog de Víctor Yepes. El blog de Víctor Yepes. https://victoryepes.blogs.upv.es/2022/02/21/que-hacemos-con-los-valores-atipicos-outliers/#:~:text=Un%20valor%20at%C3%ADpico%20(outlier%2C%20en,proceso%20o%20en%20una%20poblaci%C3%B3n.

39. Software DELSOL. (2019, June 3). ▷ Coeficiente de determinación ¿Qué es? https://www.sdelsol.com/glosario/coeficiente-de-determinacion/
40. González, L. (2022). Evaluando el error en los modelos de regresión. Aprende IA. https://aprendeia.com/evaluando-el-error-en-los-modelos-de-regresion/#:~:text=El%20error%20absoluto%20medio%20o,diferencia%20entre%205%20y%200.
41. González, L. (2021). Seleccionar el número adecuado de *clúster*es. Aprende IA. https://aprendeia.com/seleccionar-el-numero-adecuado-de-*clúster*es/
42. Davies, David L.; Bouldin, Donald W. (1979). "A Cluster Separation Measure." IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1 (2): 224-227
43. López, J. F. (2022). Estadística descriptiva. Economipedia. https://economipedia.com/definiciones/estadistica-descriptiva.html
44. Scikit-learn contributors. (2023, March 8). Davies-Bouldin score. In scikit-learn: Machine learning in Python (Version 1.2.2). Obtenido de https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html

ANEXOS

Anexo A: Información de las columnas de la tabla de Proceso_Encolado_Dosificado

summary	Columna de tiempo
count	691080
mean	null
stddev	null
min	01/03/2023 0:00:10
max	08/03/2023 9:59:59

summary	Humedad
count	691080
mean	14.173162123052169
stddev	4.447599620483017
min	0.07
max	20.03

summary	Velocidad tornillo sin fin
count	691080
mean	45.52205977599946
stddev	15.413131370020706
min	0.0
max	53.0

summary	Presion resina
count	691080
mean	9.461665248597027
stddev	1.2306310690502187
min	0.54
max	20.68

summary	Peso balanza
count	691080
mean	43.97295658969174
stddev	10.569003988098142
min	1.27
max	65.39

summary	Peso tornillo
count	691080
mean	94.09093722902071
stddev	30.32403686713818
min	0.0
max	105.96

summary	Ciclo prensado
count	691080
mean	0.0
stddev	0.0
min	0.0
max	0.0

summary	Consumo resina
count	691080
mean	94.77600250339873
stddev	36.3583304315657
min	0.0
max	257.48

summary	Consumo resina real
count	691080
mean	101.14891037229816
stddev	48.80593850466112
min	0.0
max	436.18

summary	KWh/Ton
count	11518
mean	182.38207818147217
stddev	2345.2437426360702
min	0.00
max	99.80

Fuente: Anahí Males

Anexo B: Matriz de correlación entre las columnas de Proceso_Encolado_Dosificado

	Humedad	Velocidad tornillo sin fin	Presion resina	Peso balanza
Humedad	1.000000	0.910668	0.419570	0.827363
Velocidad tornillo sin fin	0.910668	1.000000	0.452383	0.919057
Presion resina	0.419570	0.452383	1.000000	0.488954
Peso balanza	0.827363	0.919057	0.488954	1.000000
Peso tornillo	0.934923	0.985125	0.445192	0.904997
Ciclo prensado	NaN	NaN	NaN	NaN
Consumo resina	0.764367	0.750598	0.520314	0.791273
Consumo resina real	0.564511	0.550161	0.368035	0.544006

	Peso tornillo	Consumo resina	Consumo resina real
Humedad	0.934923	0.764367	0.564511
Velocidad tornillo sin fin	0.985125	0.750598	0.550161
Presion resina	0.445192	0.520314	0.368035
Peso balanza	0.904997	0.791273	0.544006
Peso tornillo	1.000000	0.810159	0.616766
Ciclo prensado	NaN	NaN	NaN
Consumo resina	0.810159	1.000000	0.763326
Consumo resina real	0.616766	0.763326	1.000000

Fuente: Anahí Males

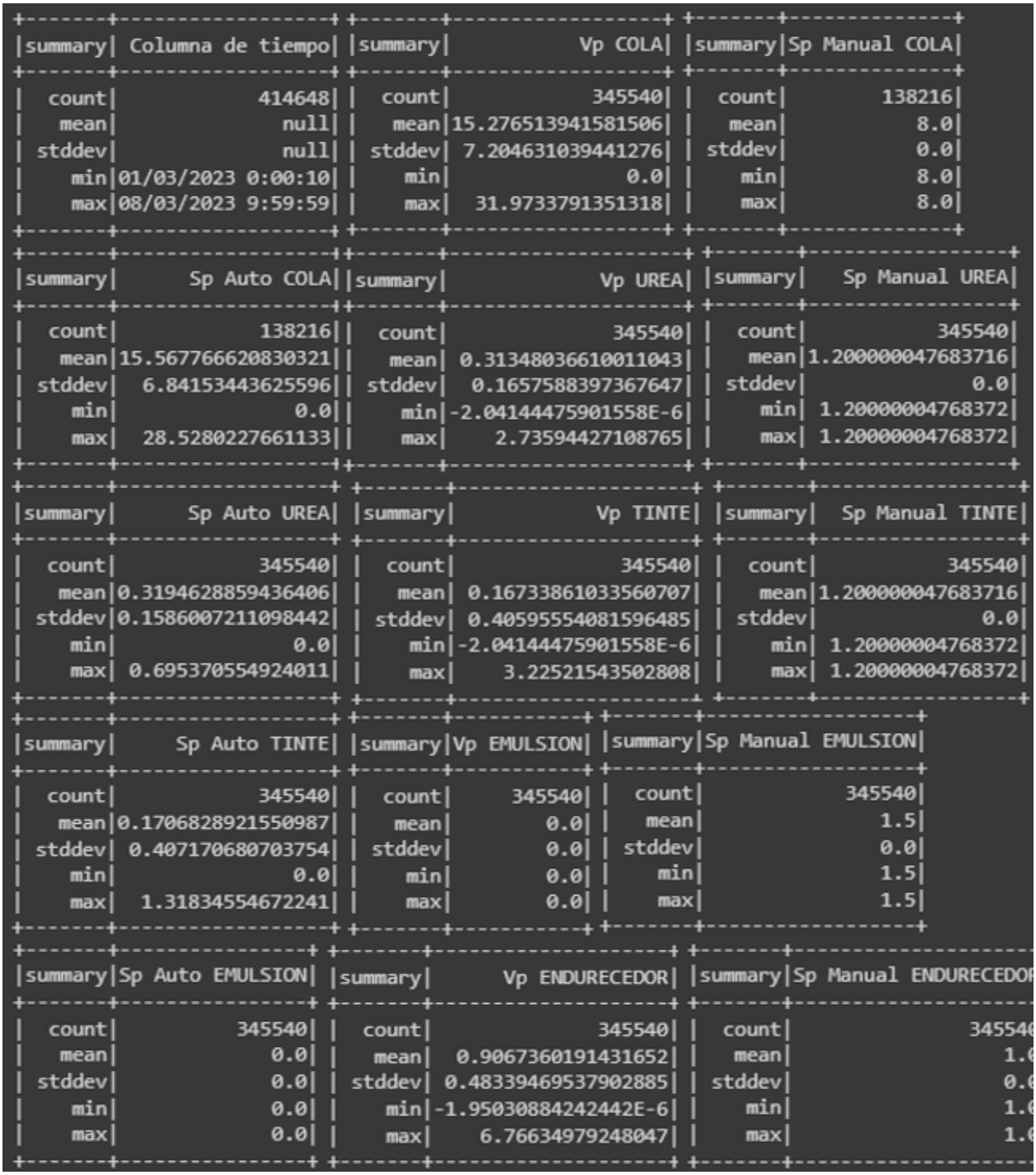
Anexo C: Cuartiles y número de valores atípicos de las columnas de Proceso_Encolado_Dosificado

	Humedad	Velocidad tornillo sin fin	Presion resina	Peso balanza
25%	14.400000	46.810000	9.100000	45.490000
50%	15.350000	53.000000	9.520000	47.340000
75%	16.420000	53.000000	9.970000	49.250000
Número de valores átipicos	73055	81340	44557	85624

	Peso tornillo	Ciclo prensado	Consumo resina	Consumo resina real
25%	102.080000	0.0	77.060000	77.520000
50%	105.960000	0.0	97.880000	99.420000
75%	105.960000	0.0	117.970000	118.720000
Número de valores átipicos	91137	0	61311	83919

Autor: Anahí Males

Anexo D: Información de las columnas de Sp_Vp_Encolado_Dosificado



summary	Columna de tiempo
count	414648
mean	null
stddev	null
min	01/03/2023 0:00:10
max	08/03/2023 9:59:59

summary	Vp COLA
count	345540
mean	15.276513941581506
stddev	7.204631039441276
min	0.0
max	31.9733791351318

summary	Sp Manual COLA
count	138216
mean	8.0
stddev	0.0
min	8.0
max	8.0

summary	Sp Auto COLA
count	138216
mean	15.567766620830321
stddev	6.84153443625596
min	0.0
max	28.5280227661133

summary	Vp UREA
count	345540
mean	0.3134803661001104
stddev	0.1657588397367647
min	-2.04144475901558E-6
max	2.73594427108765

summary	Sp Manual UREA
count	345540
mean	1.200000047683716
stddev	0.0
min	1.20000004768372
max	1.20000004768372

summary	Sp Auto UREA
count	345540
mean	0.3194628859436406
stddev	0.1586007211098442
min	0.0
max	0.695370554924011

summary	Vp TINTE
count	345540
mean	0.16733861033560707
stddev	0.40595554081596485
min	-2.04144475901558E-6
max	3.22521543502808

summary	Sp Manual TINTE
count	345540
mean	1.200000047683716
stddev	0.0
min	1.20000004768372
max	1.20000004768372

summary	Sp Auto TINTE
count	345540
mean	0.1706828921550987
stddev	0.407170680703754
min	0.0
max	1.31834554672241

summary	Vp EMULSION
count	345540
mean	0.0
stddev	0.0
min	0.0
max	0.0

summary	Sp Manual EMULSION
count	345540
mean	1.5
stddev	0.0
min	1.5
max	1.5

summary	Sp Auto EMULSION
count	345540
mean	0.0
stddev	0.0
min	0.0
max	0.0

summary	Vp ENDURECEDOR
count	345540
mean	0.9067360191431652
stddev	0.48339469537902885
min	-1.95030884242442E-6
max	6.76634979248047

summary	Sp Manual ENDURECEDOR
count	345540
mean	1.0
stddev	0.0
min	1.0
max	1.0

Autor: Anahí Males

Anexo E: Matriz de correlación entre las columnas de Sp_Vp_Encolado_Dosificado

	Vp COLA	Sp Auto COLA	Vp UREA	Sp Auto UREA	Vp TINTE	Vp ENDURECEDOR
Vp COLA	1.000	0.962	0.894	0.857	0.235	0.944
Sp Auto COLA	0.962	1.000	0.862	0.889	0.230	0.915
Vp UREA	0.894	0.862	1.000	0.963	0.360	0.942
Sp Auto UREA	0.857	0.889	0.963	1.000	0.362	0.915
Vp TINTE	0.235	0.230	0.360	0.362	1.000	0.409
Sp Auto TINTE	0.217	0.226	0.343	0.361	0.991	0.392
Vp ENDURECEDOR	0.944	0.915	0.942	0.915	0.409	1.000
Sp Auto ENDURECEDOR	0.919	0.956	0.925	0.959	0.420	0.951
Vp CERA	0.787	0.735	0.860	0.816	0.020	0.777
Sp Auto CERA	0.746	0.759	0.830	0.848	0.003	0.745

	Sp Auto TINTE	Sp Auto ENDURECEDOR	Vp CERA	Sp Auto CERA
Vp COLA	0.217	0.919	0.787	0.746
Sp Auto COLA	0.226	0.956	0.735	0.759
Vp UREA	0.343	0.925	0.860	0.830
Sp Auto UREA	0.361	0.959	0.816	0.848
Vp TINTE	0.991	0.420	0.020	0.003
Sp Auto TINTE	1.000	0.420	0.002	-0.003
Vp ENDURECEDOR	0.392	0.951	0.777	0.745
Sp Auto ENDURECEDOR	0.420	1.000	0.737	0.765
Vp CERA	0.002	0.737	1.000	0.965
Sp Auto CERA	-0.003	0.765	0.965	1.000

Fuente: Anahí Males

Anexo F: Cuartiles y valores atípicos de Sp_Vp_Encolado_Dosificado

	Vp COLA	Sp Manual COLA	Sp Auto COLA	Vp UREA
25%	10.243055	8.0	10.248710	0.228319
50%	16.174768	8.0	16.189800	0.258257
75%	20.500578	8.0	20.499556	0.445049

	Sp Manual UREA	Sp Auto UREA	Sp Manual TINTE	Sp Manual EMULSION
25%	1.200000e+00	0.231260	1.200000e+00	1.5
50%	1.200000e+00	0.252203	1.200000e+00	1.5
75%	1.200000e+00	0.448402	1.200000e+00	1.5

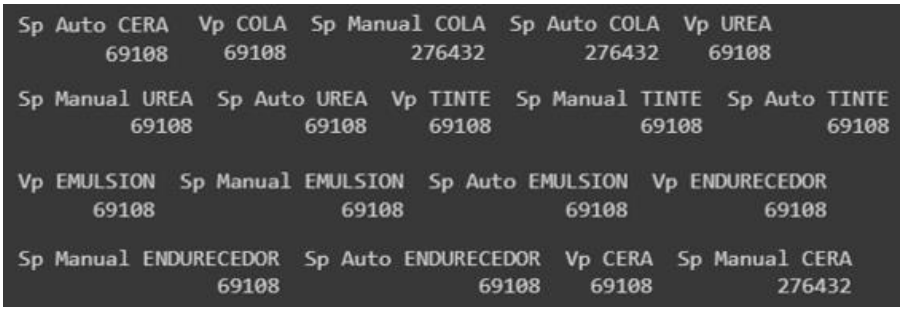
	Vp ENDURECEDOR	Sp Manual ENDURECEDOR	Sp Auto ENDURECEDOR
25%	0.547795	1.0	0.531396
50%	0.879100	1.0	0.851552
75%	1.279103	1.0	1.293787

	Vp CERA	Sp Auto CERA
25%	0.491479	0.493320
50%	0.529252	0.529806
75%	0.662474	0.662258

Número de valores átipicos			
Sp Auto TINTE	52419	Vp UREA	2
Vp ENDURECEDOR	510	Vp TINTE	50574
Sp Auto CERA	50263	Vp CERA	57984

Fuente: Anahí Males

Anexo G: Número de nulos por columna de Sp_Vp_Encolado_Dosificado.



```
Sp Auto CERA   Vp COLA  Sp Manual COLA  Sp Auto COLA  Vp UREA
       69108     69108          276432        276432    69108

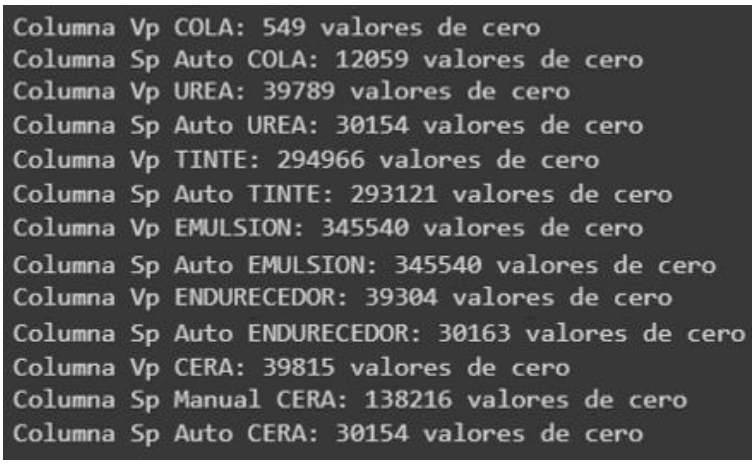
Sp Manual UREA  Sp Auto UREA  Vp TINTE  Sp Manual TINTE  Sp Auto TINTE
         69108         69108     69108            69108          69108

Vp EMULSION  Sp Manual EMULSION  Sp Auto EMULSION  Vp ENDURECEDOR
      69108               69108             69108           69108

Sp Manual ENDURECEDOR  Sp Auto ENDURECEDOR  Vp CERA  Sp Manual CERA
                69108                 69108    69108          276432
```

Fuente: Anahí Males

Anexo H: Número de ceros por columna de Sp_Vp_Encolado_Dosificado.



```
Columna Vp COLA: 549 valores de cero
Columna Sp Auto COLA: 12059 valores de cero
Columna Vp UREA: 39789 valores de cero
Columna Sp Auto UREA: 30154 valores de cero
Columna Vp TINTE: 294966 valores de cero
Columna Sp Auto TINTE: 293121 valores de cero
Columna Vp EMULSION: 345540 valores de cero
Columna Sp Auto EMULSION: 345540 valores de cero
Columna Vp ENDURECEDOR: 39304 valores de cero
Columna Sp Auto ENDURECEDOR: 30163 valores de cero
Columna Vp CERA: 39815 valores de cero
Columna Sp Manual CERA: 138216 valores de cero
Columna Sp Auto CERA: 30154 valores de cero
```

Fuente: Anahí Males