# **Clústeres de Líneas de Producción de Cerveza (AB-Inbev)**

****Resumen****.

La compañía Cervecera **Ab-Inbev** opera dentro de diferentes zonas alrededor del mundo, una de ellas es MAZ (Middle Americas Zone) la cual está fragmentada en 5 regionales que a su vez, están divididas en países, donde se ubican las plantas que albergan las 176 líneas de envasado de la zona. Como parte de la estrategia de seguimiento de productividad de la compañía, cada una de estas líneas reporta turno a turno, sus tiempos productivos, es decir, aquellos que se reflejan en unidades de cerveza (y otras bebidas) y tiempos de parada, estos últimos detallando las causas que generaron dicha interrupción de la producción, especificando las posibles causas, sean actividades de aseo, mantenimiento, tiempos no programados, etc. Como es de esperar, existe variación entre las diferentes métricas de productividad y tiempos reportados por las diferentes líneas de envasado, dando como resultado líneas con alto desempeño y otras con oportunidades significativas en la eficiencia y cumplimiento de metas pactadas con la compañía.

El presente proyecto busca utilizar algoritmos de Aprendizaje no Supervisado para obtener agrupaciones de líneas cuyas características de tiempos perdidos, eficiencias y cumplimiento de metas durante el mes de julio de 2022 sean similares, así mismo busca identificar los factores de mayor influencia sobre los indicadores de productividad en las líneas de envasado, esto con el objetivo que la compañía direccione las correspondientes estrategias a cada una de las líneas dependiendo de las características del clúster al cual corresponde, con el fin de cumplir con las metas comprometidas por la zona con la compañía.

****Introducción****.

Durante la definición de metas de Supply en Ab-Inbev, el área de Packaging estableció y cascadeó los compromisos con cada una de las plantas y sus respectivas líneas de envasado para los indicadores de productividad del año 2022. Tras el cierre del mes de Julio el GLY, principal indicador para el seguimiento de la productividad de línea presentó una desviación respecto a la meta planteada del 3.1%, lo cual lleva a la necesidad de plantear estrategias específicas de acuerdo con las características de cada línea, con el fin de lograr el cierre de dicha brecha. Para esto, el equipo de directivo debe responder a la pregunta:

*¿Cuáles son las posibles agrupaciones de las líneas de envasado de la zona MAZ a partir de considerar información de tiempos de productividad y parada, cumplimiento de metas e indicadores de desempeño, así como variables geográficas de cada una de ellas en un tiempo dado?*

El equipo de este proyecto pretende encontrar patrones a través del proceso de clusterización que permitan implementar estrategias diferenciadas para cada una de las diferentes agrupaciones halladas, lo anterior con el fin de mejorar y focalizar esfuerzos que garanticen una óptima producción de las líneas.

**Revisión preliminar de la literatura**.

Sobre la utilización de métodos de Clustering para líneas de producción no se encuentra mucha literatura. Sin embargo, se tiene de referencia un artículo cuyo objetivo era “Predict Failures in Production Line” elaborado por Darui Zhang y Bin Xu (2016) donde los investigadores utilizaron métodos de clustering y de aprendizaje supervisado. Acá el proceso que utilizaron para llegar a los clusters fue: 1. Hacer reducción de dimensionalidad usando PCA, luego para definir el número de clusters con el algoritmo K-means fue ver con que cantidad de clustera la inercia dentro de los mismos dejaba de reducirse(el número fue 6). Por ende, se tomará este proceso de referencia para el proyecto que se va a realizar.

El algoritmo K-Means es uno de los más utilizados para realizar aplicaciones de clusterización o agrupación de observaciones. Según Cambronero y Moreno (2006), los algoritmos de clustering tienen como objetivo identificar cuál es la agrupación que tienen los datos por defecto. Según los mismos autores, las aplicaciones de los algoritmos de clusterización se pueden encontrar en diversas ramas como la biología, negocios, seguros, entre otras. Los autores proponen que un cluster es un conjunto con observaciones similares entre ellas, pero diferentes de las observaciones de otros grupos. KMeans fue creado por MacQueen en 1967 y es uno de los algoritmos más utilizados para realizar la clusterización. Lo que busca el algoritmo, según los autores, es encontrar un centroide, que vendría a ser la media de las observaciones.

A pesar de sus diversas aplicaciones, KMeans, según Cambronero y Moreno (2006) presenta diversos inconvenientes. El primero es que se tiene realizar varias iteraciones para tener resultados óptimos, por lo que es computacionalmente costoso. Asimismo, es susceptible a valores extremos debido a que distorsionan la distribución de los datos.

De la misma manera, Ahmed, Seraj y Shamsul (2007) encontraron que el algoritmo KMeans tiene aplicaciones en detección de rostros, segmentación de imágenes, procesamiento de texto, entre otras. Sin embargo, encontraron que existían ciertos problemas al momento de implementar este algoritmo como la incapacidad que tienen para manejar variables con diversos tipos de datos. A pesar de estos inconvenientes y aquellos como los que mencionaron Garcia y Gómez (2006), la experimentación y la investigación han demostrado que existen diversas formas de lidiar con estos inconvenientes, por lo que hacen que KMeans siga siendo un algoritmo robusto con gran aplicabilidad en diversos contextos.

Finalmente, DBScan ha surgido como una alternativa para solucionar algunos de los problemas que presenta KMeans, además de haber demostrado un gran desempeño en aplicaciones como reconocimiento de imágenes satelitales, extracción de patrones o detección de outliers. Sin embargo, a pesar de su fama ganada últimamente, no está exento de inconvenientes. Por una parte, requiere que los usuarios especifiquen los valores de los parámetros con los que opera, además, el algoritmo puede generar clústeres sin significado en datos con diversas densidades. Por estos inconvenientes, la investigación se ha centrado en mejorar el algoritmo generando derivados como VDBSCAN, FDBSCAN, DD\_DBSCAN, entre otros (Khan et al. 2014).

****Descripción detallada de los datos****.

Los datos a utilizar para el proyecto se descargarán de diferentes reportes, teniendo como resultado tres diferentes archivos de Excel. A partir de los cuales se construirá el dataset para el entrenamiento del modelo.

|  |  |
| --- | --- |
| Texto  Descripción generada automáticamente | Información para cada línea, fecha y turno, sobre los diferentes tiempos de paradas, asociados a múltiples causas, por ejemplo, fallas de máquinas, tiempos de aseos, mantenimientos, no programados, etc. Así como los tiempos productivos e indicadores de productividad de cada observación. |
| Texto  Descripción generada automáticamente | A partir de este archivo se obtienen las metas de productividad de cada línea pactadas con la compañía para el mes de Julio del 2022. |
| Texto  Descripción generada automáticamente | Se identifican la línea que extraen y reportan la información de tiempos de parada y producción de forma automática. |

El dataset deseado, se conformará por cada una de las líneas de la zona MAZ distribuido en las filas (como observaciones) y en las columnas (dimensiones) se tendrán los valores de tiempos perdidos asociados a cada categoría.

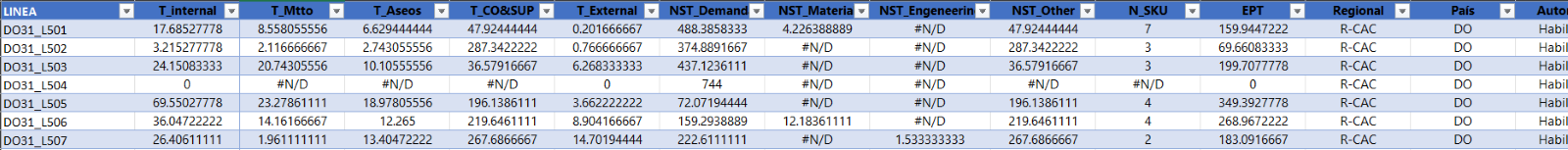
|  |  |
| --- | --- |
| T\_internal: Tiempos perdidos por fallas de máquina | Regional: Regional a la que pertenece la línea |
| T\_Mtto: Tiempos de paro por realización de actividades de mantenimiento | País: País donde se encuentra la línea |
| T\_Aseos: Tiempos de paro por realización de actividades de aseo | Automatic: 1 Si la línea captura datos de forma automática, 0 de lo contrario. |
| T\_ChangeO: Tiempos de paro por ralización de cambios de SKU | GLY: Indicador de productividad calculado sobre el tiempo programado. |
| T\_External: Tiempos de paro por causas externas al área de envasado | LEF: Indicador de productividad asociado a la eficiencia mecánica de las máquinas. |
| NST\_Demanda: Tiempos no programados por baja demanda | OEE: Indicador de productividad calculado sobre la capacidad instalada de cada línea. |
| NST\_Material: Tiempos no programados por disponibilidad de material. | META\_GLY: Compromiso adquirido por la planta para el indicador de GLY durante el mes |
| NST\_Engeneering: Tiempos no programados por ejecución de proyectos. | META\_LEF: Compromiso adquirido por la planta para el indicador de LEF durante el mes |
| NST\_Other: Tiempos no programados por otras causas | META\_ST: Compromiso adquirido por la planta para total de tiempo productivo |
| EPT: Tiempo efectivo de producción, calculado sobre la capacidad de línea. |  |
| N\_SKU: Cantidad de diferentes SKUs producidos en la línea durante el periodo. |  |

La consecución de los datos en el formato deseado requiere de diferentes transformaciones de cada tabla obtenido de las diferentes fuentes, y adicionalmente, la unión de estas en un único DataFrame. Los pasos requeridos para lograr dicho procedimiento se describen a continuación.

1. Pivotizar los datos para obtener la sumatoria del mes de todos los tiempos e indicadores definidos para cada observación.

FormaInterfaz de usuario gráfica

Descripción generada automáticamente con confianza media



1. Los datos obtenidos en el paso anterior se consolidan con la información disponible en las diferentes fuentes.

FormaInterfaz de usuario gráfica

Descripción generada automáticamente

Tabla

Descripción generada automáticamenteFormaTabla

Descripción generada automáticamenteUna captura de pantalla de un celular

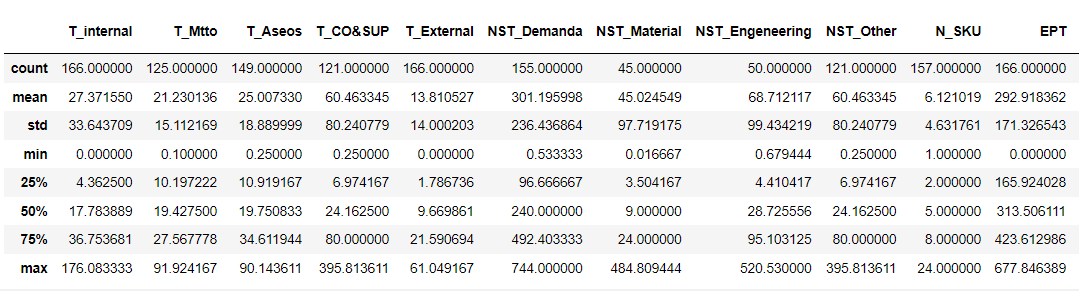
Descripción generada automáticamenteForma

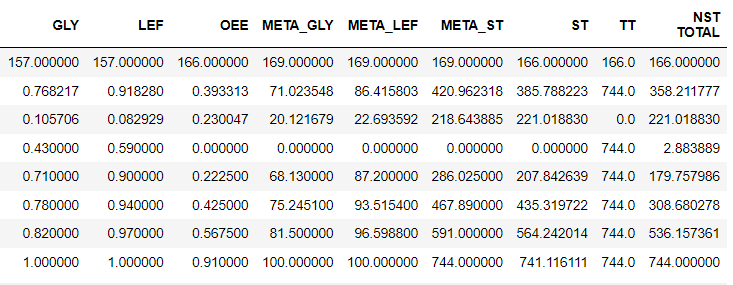
1. Se dará tratamiento a los valores nulos, considerando el contexto de la situación, en general es correcto remplazar estos valores por 0.
2. Se obtienen datos descriptivos del dataset.

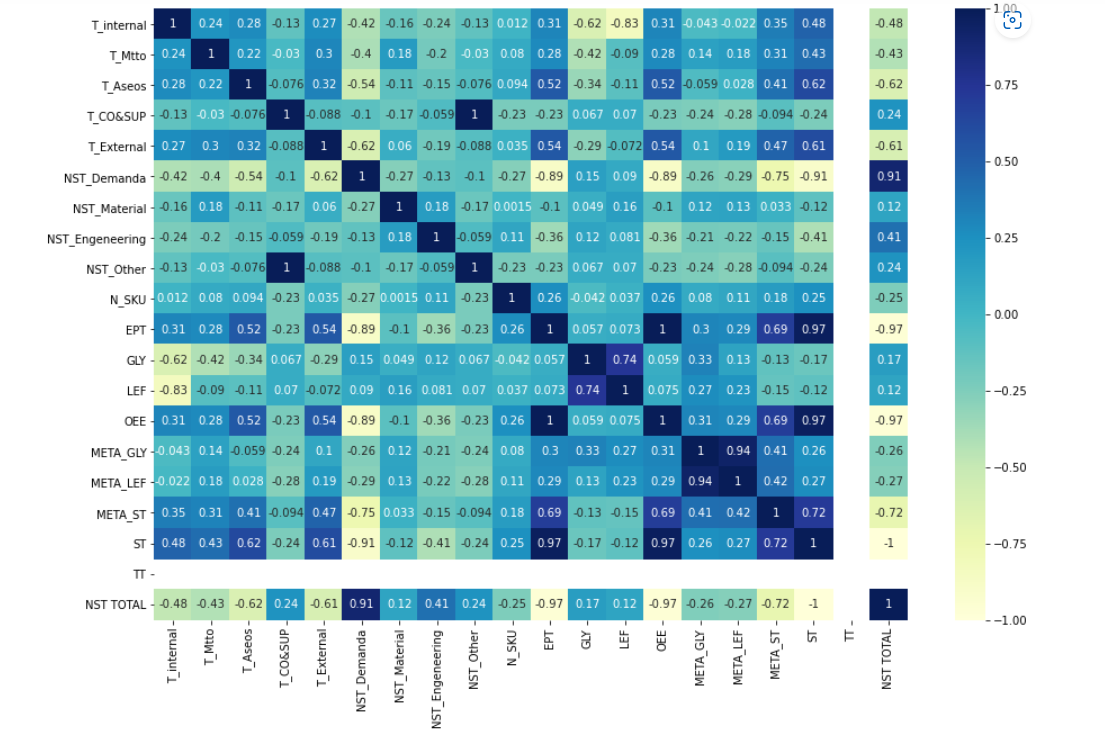
Interfaz de usuario gráfica, Aplicación, Tabla, Excel

Descripción generada automáticamente

Ahora es de nuestro interés obtener algunas estadísticas descriptivas de los datos y algunas visualizaciones como mapas de calor y pairplots que nos indiquen el comportamiento de los datos y el nivel de correlación entre cada una de las variables:







Con las estadísticas descriptivas de las variables presentadas en la tabla anterior podemos obtener información como la media, mediana y varianza de cada una de ellas, así como el rango que tiene cada variable. Por otra parte, en el mapa de calor se puede observar que hay una fuerte correlación entre algunas variables las cuales deben ser tratadas ya que pueden afectar el desempeño del algoritmo a implementar debido a la redundancia de información y de dimensionalidad.

****Propuesta Metodológica****

El problema planteado en este proyecto se espera resolver aplicando primero la reducción de dimensiones a través del procedimiento de componentes principales (PCA), posteriormente se implementará el algoritmo de K-Medoides teniendo en cuenta que el tamaño de la información no es relativamente grande (180 líneas de producción) y además contiene variables categóricas, adicionalmente de que este algoritmo es robusto a los datos atípicos (outliers) que puedan estar presentes en la data.

Adicionalmente se plantea como propuesta alternativa la implementación del algoritmo DBSCAN, lo anterior teniendo en cuenta que este presenta algunas ventajas de importancia como lo es que no fuerza todas las observaciones a pertenecer a un clúster en particular, lo que lo hace un óptimo excluyente de outliers, y además de que se enfoca en la identificación de clústeres a partir de la densidad de las observaciones del dataset.

****Bibliografía****

K. Khan, S. U. Rehman, K. Aziz, S. Fong and S. Sarasvady, "DBSCAN: Past, present and future," The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014), 2014, pp. 232-238, doi: 10.1109/ICADIWT.2014.6814687.

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, *9*(8), 1295.

Cambronero, C. G., & Moreno, I. G. (2006). Algoritmos de aprendizaje: knn & kmeans. *Intelgencia en Redes de Comunicación, Universidad Carlos III de Madrid*, *23*.

Zhang, Darui & Xu, Bin & Wood, Jasmine. (2016). Predict failures in production lines: A two-stage approach with clustering and supervised learning. 2070-2074. 10.1109/BigData.2016.7840832.