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.

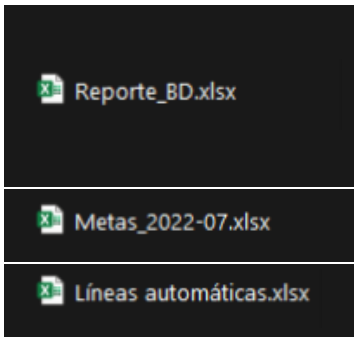
A pesar de sus diversas aplicaciones, KMeans, según Cambronero y Moreno (2006) presenta diversos inconvenientes. El primero es que se tiene realizar varias iteraciones para tener resultados óptimos, por lo que es computacionalmente costoso. Asimismo, es susceptible a valores extremos debido a que distorsionan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



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.

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.

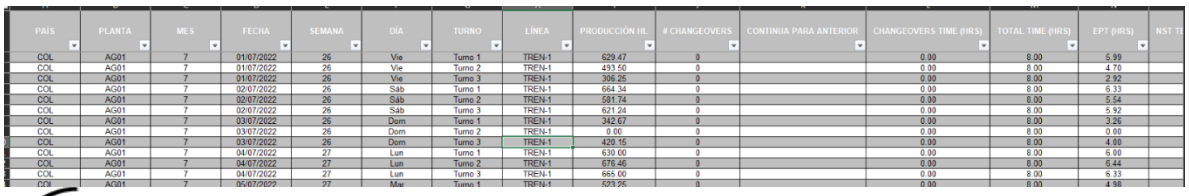
Se identifican la línea que extraen y reportan la información de tiempos de parada y producción de forma automática.

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.

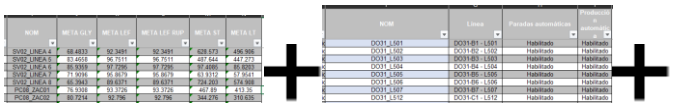


PAÍS	PLANTA	MES	FECHA	SEMANA	DÍA	TURNO	LÍNEA	PRODUCCIÓN HL	# CHANGEOVERS	CONTINUA PARA ANTERIOR	CHANGEOVERS TIME (HRS)	TOTAL TIME (HRS)	EPT (HRS)	NST TI
COL	AG01	7	01/07/2022	26	Vie	Turno 1	TREN-1	629.47	0		0.00	8.00	5.99	
COL	AG01	7	01/07/2022	26	Vie	Turno 2	TREN-1	493.50	0		0.00	8.00	4.70	
COL	AG01	7	01/07/2022	26	Vie	Turno 3	TREN-1	306.26	0		0.00	8.00	2.92	
COL	AG01	7	02/07/2022	26	Sáb	Turno 1	TREN-1	664.34	0		0.00	8.00	6.33	
COL	AG01	7	02/07/2022	26	Sáb	Turno 2	TREN-1	581.74	0		0.00	8.00	5.54	
COL	AG01	7	02/07/2022	26	Sáb	Turno 3	TREN-1	621.24	0		0.00	8.00	5.92	
COL	AG01	7	03/07/2022	26	Dom	Turno 1	TREN-1	342.67	0		0.00	8.00	3.26	
COL	AG01	7	03/07/2022	26	Dom	Turno 2	TREN-1	0.00	0		0.00	8.00	0.00	
COL	AG01	7	03/07/2022	26	Dom	Turno 3	TREN-1	420.15	0		0.00	8.00	4.00	
COL	AG01	7	04/07/2022	27	Lun	Turno 1	TREN-1	630.00	0		0.00	8.00	6.00	
COL	AG01	7	04/07/2022	27	Lun	Turno 2	TREN-1	676.46	0		0.00	8.00	6.44	
COL	AG01	7	04/07/2022	27	Lun	Turno 3	TREN-1	665.00	0		0.00	8.00	6.33	
COL	AG01	7	05/07/2022	27	Mar	Turno 1	TREN-1	593.25	0		0.00	8.00	4.98	

LINEA	T_Internal	T_Mtto	T_Aseos	T_CO&SUP	T_External	NST_Demand	NST_Materia	NST_Engeneerin	NST_Other	N_SKU	EPT	Regional	País	Auto
DO31_LS01	17.68527778	8.558055556	6.629444444	47.92444444	0.201666667	488.3858333	4.226388889	#N/D	47.92444444	7	159.9447222	R-CAC	DO	Habil
DO31_LS02	3.215277778	2.116666667	2.743055556	287.3422222	0.766666667	374.8891667	#N/D	#N/D	287.3422222	3	69.66083333	R-CAC	DO	Habil
DO31_LS03	24.15083333	20.74305556	10.10555556	36.57916667	6.268333333	437.1236111	#N/D	#N/D	36.57916667	3	199.7077778	R-CAC	DO	Habil
DO31_LS04	0	#N/D	#N/D	#N/D	0	744	#N/D	#N/D	#N/D	#N/D	0	R-CAC	DO	Habil
DO31_LS05	69.55027778	23.27861111	18.97805556	196.1386111	3.662222222	72.07194444	#N/D	#N/D	196.1386111	4	349.3927778	R-CAC	DO	Habil
DO31_LS06	36.04722222	14.16166667	12.265	219.6461111	8.904166667	159.2938889	12.18361111	#N/D	219.6461111	4	268.9672222	R-CAC	DO	Habil
DO31_LS07	26.40611111	1.961111111	13.40472222	267.6866667	14.70194444	222.6111111	#N/D	1.533333333	267.6866667	2	183.0916667	R-CAC	DO	Habil

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

LINEA	Regional	País	Automatic	GLY	LEF	OEE	META_GLY	META_LEF	META_ST	ST	TT	NST TOTA
DO31_L501	R-CAC	DO	Habilitado	79%	90%	21%	0	0	0	203.463333	744	540.536667
DO31_L502	R-CAC	DO	Habilitado	85%	96%	9%	0	0	0	81.7686111	744	662.231389
DO31_L503	R-CAC	DO	Habilitado	74%	89%	27%	60.8278	76.1587	432.186	270.297222	744	473.702778
DO31_L504	R-CAC	DO	Habilitado	#¡DIV/0!	#¡DIV/0!	0%	0	0	0	0	744	744
DO31_L505	R-CAC	DO	Habilitado	73%	83%	47%	60.67	77.765	418.625	475.789444	744	268.210556
DO31_L506	R-CAC	DO	Habilitado	76%	88%	36%	63.2839	80.7298	609.037	352.876389	744	391.123611
DO31_L507	R-CAC	DO	Habilitado	73%	87%	25%	71.8934	87.6464	517.433	252.168889	744	491.831111
DO31_L512	R-CAC	DO	Habilitado	90%	96%	40%	82.1357	95.6207	645.274	332.378889	744	411.621111
DO31_L521	R-CAC	DO	Habilitado	96%	100%	32%	85.7354	98.7675	560.831	244.799444	744	499.200556



LINEA	T_internal	T_Mtto	T_Aseos	T_CO&SUP
DO31_L501	17.68527778	8.558055556	6.629444444	47.92444444
DO31_L502	3.215277778	2.116666667	2.743055556	287.3422222
DO31_L503	24.15083333	20.74305556	10.10555556	36.57916667
DO31_L504	0	#N/D	#N/D	#N/D
DO31_L505	69.55027778	23.27861111	18.97805556	196.1386111
DO31_L506	36.04722222	14.16166667	12.265	219.6461111

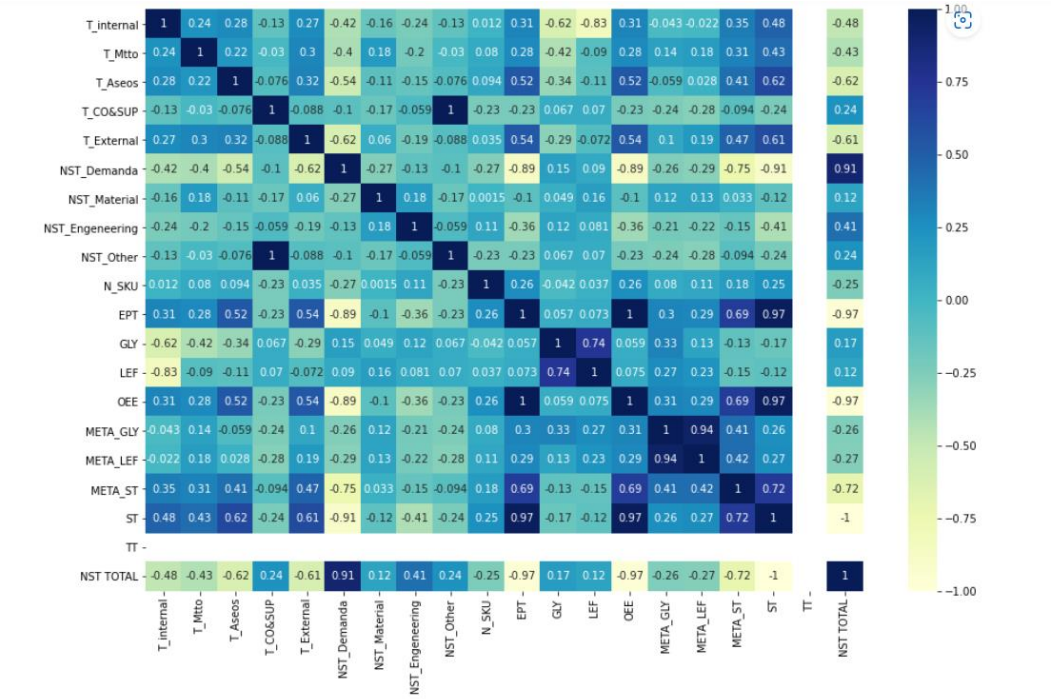
3. Se dará tratamiento a los valores nulos, considerando el contexto de la situación, en general es correcto remplazar estos valores por 0.
4. Se obtienen datos descriptivos del dataset.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
LINEA	176	176	BK01_LINEA 02	1							
T_internal	176				26	33	0	3	17	35.25	176
T_Mtto	176				15	16	0	0	12	24	92
T_Aseos	176				21	20	0	4	16.5	33	90
T_CO&SUP	176				42	72	0	0	8	51	396
T_External	176				13	14	0	1	9	21	61
NST_Demanda	176				265	242	0	47	191.5	438	744
NST_Material	176				12	53	0	0	0	0	485
NST_Engeneering	176				20	61	0	0	0	2	521
NST_Other	176				42	72	0	0	8	51	396
N_SKU	176				5	5	0	2	4	8	24
EPT	176				276	180	0	120	295	418.25	678
Regional	176	7	R-CAC	49							
País	176	14	MX_SUR	32							
Automatic	176	2	Habilitado	144							
GLY	176				1	0	0	1	1	1	1
LEF	176				1	0	0	1	1	1	1
OEE	176				0	0	0	0	0	1	1
META_GLY	176				68	24	0	66.75	74.5	81	100
META_LEF	176				83	28	0	85.75	93	96.25	100
META_ST	176				404	230	0	243	460	590	744
ST	176				364	233	0	165	413.5	550.75	741
TT	176				702	173	0	744	744	744	744
NST TOTAL	176				338	230	0	160.5	293	523	744

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:

	T_internal	T_Mtto	T_Aseos	T_CO&SUP	T_External	NST_Demanda	NST_Material	NST_Engeneering	NST_Other	N_SKU	EPT
count	166.000000	125.000000	149.000000	121.000000	166.000000	155.000000	45.000000	50.000000	121.000000	157.000000	166.000000
mean	27.371550	21.230136	25.007330	60.463345	13.810527	301.195998	45.024549	68.712117	60.463345	6.121019	292.918362
std	33.643709	15.112169	18.889999	80.240779	14.000203	236.436864	97.719175	99.434219	80.240779	4.631761	171.326543
min	0.000000	0.100000	0.250000	0.250000	0.000000	0.533333	0.016667	0.679444	0.250000	1.000000	0.000000
25%	4.362500	10.197222	10.919167	6.974167	1.786736	96.666667	3.504167	4.410417	6.974167	2.000000	165.924028
50%	17.783889	19.427500	19.750833	24.162500	9.669861	240.000000	9.000000	28.725556	24.162500	5.000000	313.506111
75%	36.753681	27.567778	34.611944	80.000000	21.590694	492.403333	24.000000	95.103125	80.000000	8.000000	423.612986
max	176.083333	91.924167	90.143611	395.813611	61.049167	744.000000	484.809444	520.530000	395.813611	24.000000	677.846389

GLY	LEF	OEE	META_GLY	META_LEF	META_ST	ST	TT	NST TOTAL
157.000000	157.000000	166.000000	169.000000	169.000000	169.000000	166.000000	166.0	166.000000
0.768217	0.918280	0.393313	71.023548	86.415803	420.962318	385.788223	744.0	358.211777
0.105706	0.082929	0.230047	20.121679	22.693592	218.643885	221.018830	0.0	221.018830
0.430000	0.590000	0.000000	0.000000	0.000000	0.000000	0.000000	744.0	2.883889
0.710000	0.900000	0.222500	68.130000	87.200000	286.025000	207.842639	744.0	179.757986
0.780000	0.940000	0.425000	75.245100	93.515400	467.890000	435.319722	744.0	308.680278
0.820000	0.970000	0.567500	81.500000	96.598800	591.000000	564.242014	744.0	536.157361
1.000000	1.000000	0.910000	100.000000	100.000000	744.000000	741.116111	744.0	744.000000



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 y K-Means 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.

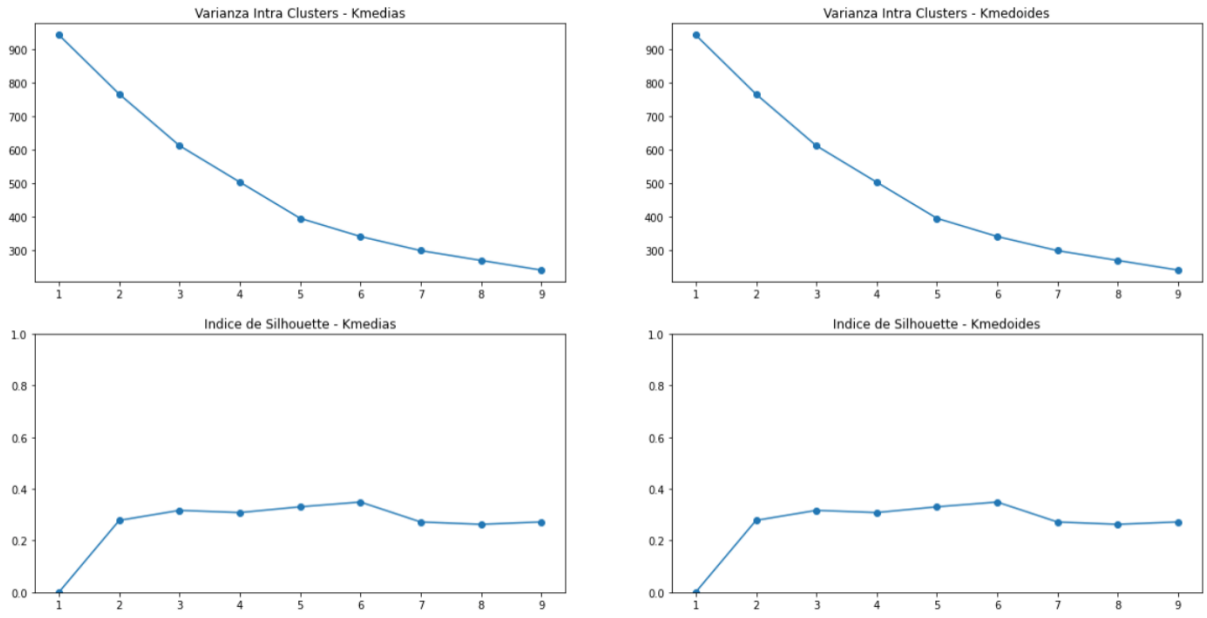
Materiales y Métodos

A partir de los datos analizados, se pudo identificar que existían variables que estaban completamente correlacionadas. Asimismo, se analizó el dataset con analistas de negocio de la empresa y se decidió realizar el análisis enfocándolo solamente en 5 variables:

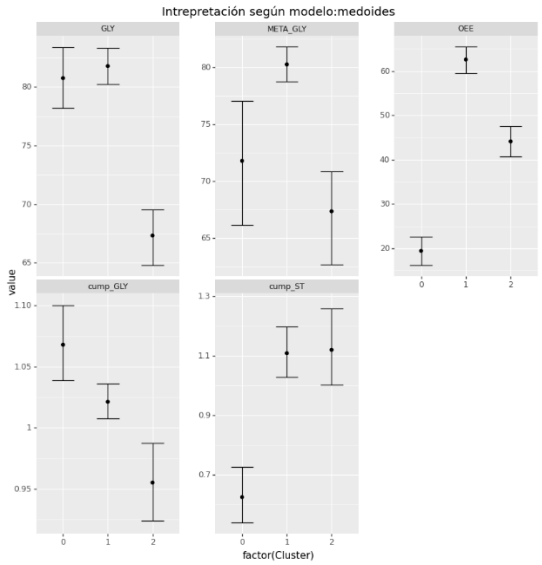
Variables	Descripción
GLY	Indicador de productividad calculado sobre el tiempo programado.
OEE	Indicador de productividad calculado sobre la capacidad instalada de cada línea.
META_GLY	Compromiso adquirido por la planta para el indicador de GLY durante el mes
Cump_GLY	Variable calculada a partir de la división de GLY / META_GLY
Cump_ST	Variable calculada a partir de la división de ST/META_ST

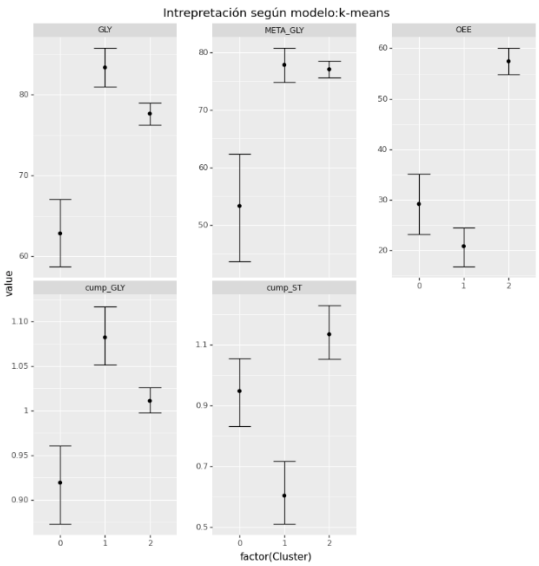
Al ser solamente 05 variables, se decidió no realizar PCA. Para entrenar a los algoritmos, se realizó el escalamiento estándar para asegurar que todas las variables tengan media 0 y desviación estándar 1.

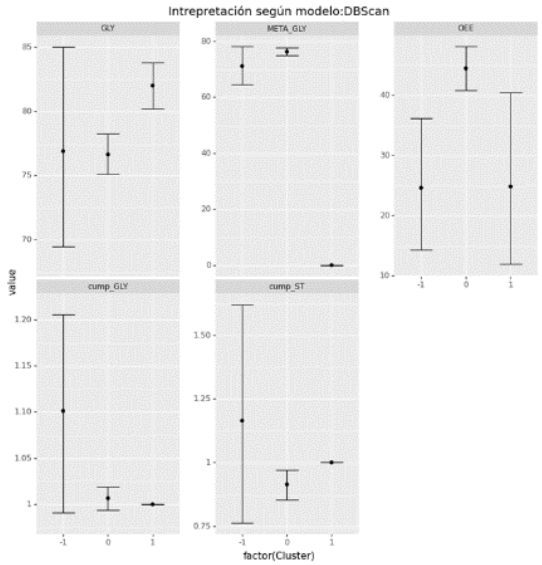
Para los algoritmos K-Medoides y K-Means, el número de clústeres se escogió en función al coeficiente de Silhouette y la varianza interclústeres, además de la interpretabilidad que se le puede dar a cada uno de los clústeres. Según este criterio, se escogieron 3 clústeres para cada uno de estos algoritmos, debido a la interpretación que se le puede dar a cada uno de los clústeres y al coeficiente Silhouette (0.3163 para ambos algoritmos)

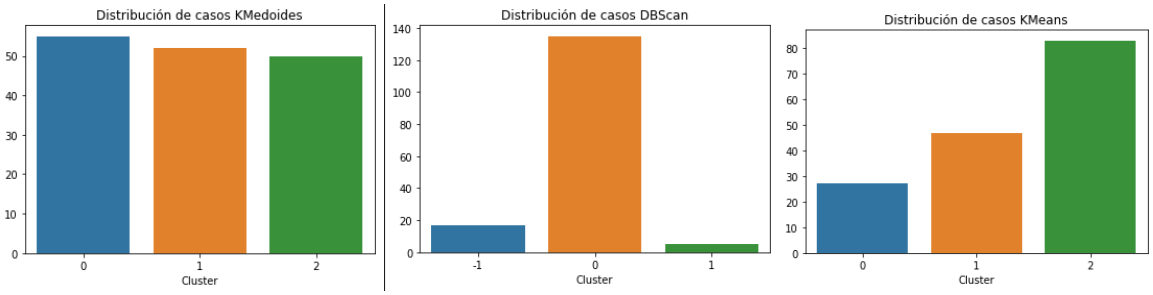


Para el algoritmo DBScan no fue necesario escoger el número de clústeres debido a que el mismo algoritmo determina la cantidad de clústeres. Se calcularon los parámetros eps = 1.298967377 como resultado del algoritmo NearestNeighbors y min_samples = 5 como sugerencia de parte del analista de negocio. Con estos parámetros, se generaron 2 clústeres. A continuación, se presenta un resumen de los clústeres generados en función a las variables del dataset y la distribución de los datos para cada clúster.









Finalmente, se utilizó el algoritmo KMedoides debido a que mostró un mejor desempeño en cuanto a interpretabilidad y los resultados hicieron sentido al momento de contrastarlos con el analista de negocio.

Conclusiones

- Cada una de las líneas de la zona se categoriza dentro de uno de tres posibles clusters basado en su desempeño de productividad y sus metas definidas para el mes especificado.
- El algoritmo de clustering que mostró mejor desempeño entre los utilizados en el proyecto fue el de Kmedoids, evaluado en las métricas de Silhouette y la varianza Inter-clústeres, así como la validación de los resultados por parte de un Stakeholder de la organización.
- A través del análisis de medidas de dispersión de las diferentes variables para cada uno de los Clústeres generados, se puede realizar una aproximación de a qué corresponde en el contexto de la organización cada uno de los clústeres.

Cluster	GLY	OEE	META GLY	CUMPL GLY	CUMP ST
0	ALTO	BAJO	ALTO	ALTO	BAJO
1	ALTO	ALTO	ALTO	MEDIO	ALTO
2	MEDIO	MEDIO	MEDIO	MEDIO	ALTO

- A partir de las agrupaciones obtenidas la organización puede tomar decisiones más acertadas para mejorar el desempeño y la programación de producción basándose en las diferentes características de cada una de las líneas y diferentes variables del negocio como demanda, disponibilidad de material y personal, en las diferentes zonas geográficas donde opera la compañía.

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.