



REPORTE FINAL – RETO MA1001B

Tec de Monterrey-CEM

GOLDEN DATA BATCH-KOF

Semestre Agosto-Diciembre 2021

Se analizaron los datos de Femsa-Coca Cola de la producción en la planta de Altamira del 6/07/2021 al 30/07/21. El reto se llevó a cabo en el periodo del 27/09/2021 al 22/09/2021.

Al mismo proyecto trabajaron también otros equipos de alumnos de los profesores María del Carmen Jiménez Hernández y Gabriel González González.

Los datos fueron entregados por KOF en formato .csv (la primera versión) y en formato Excel (la segunda versión creada por los alumnos de la Prof.ra María del Carmen Jiménez Hernández, con los datos transpuestos y sin datos faltantes). El objetivo del análisis es crear un modelo que explique la relación entre la energía necesaria para la cadena de producción y diferentes variables con la finalidad de minimizar el uso de energía.

Para los equipos de la Unidad de Formación MA1001B, de tercer semestre, fue posible desarrollar sólo una parte del estudio, debido al tiempo reducido de 5 semanas y compatiblemente a los temas de estadística aprendidos en la Unidad de Formación. El análisis que se llevó a cabo consistió en tres fases, cuyos resultados se presentan en seguida.

En seguida se presentan los resultados obtenidos por:

Equipo: Naranja

Integrantes del equipo:

Fernanda Jiménez Raya A01770109

Evelyn Geovanna Pérez Gómez A01368866

Adrián Landaverde Nava A01745052

Cristian Gonzaga López	A01745134
Naomi Padilla Mora	A01745914
Sabrina Nicole Rodríguez Salgado	A01745197

FASE 1: Análisis Exploratorio

Dada la complejidad de la base de datos, se hizo un análisis exploratorio para “orientarse” entre las variables. En seguida se muestran algunos de los gráficos obtenidos y su interpretación.

En primer lugar, antes de comenzar el análisis, se volvió a limpiar la base de datos. Ya que se pudieron identificar varias características de los datos. El primer proceso de limpieza consistió en eliminar la columna de EQUIPO, ya que este viene implícito en los nombres de las variables a medir. Una vez hecho esto, se pudo unir los datos que tuvieran la misma fecha, hora, y tipo de bebida producida, eliminando así muchos datos nulos y obteniendo un data frame más compacto. El segundo proceso de limpieza consistió en separar los datos por líneas. En esta fase nos enfocamos en analizar cada línea por separado, ya que cada línea tiene diferentes equipos a su disposición. Por un lado, de las siete líneas que hay (LINEA001, LINEA004, LINEA005, LINEA006, LINEA007, LINEA009, MULTI001) las dos últimas no cuentan con los equipos LLEN01 y CARBO. Además, las otras 5 líneas tienen diferentes variables a medir para el equipo LLEN01, por lo tanto, se decidió separar las líneas para esta fase. Una vez hecha esta limpieza, se procedió a realizar el análisis exploratorio.

Este análisis, para poder visualizarlo mejor para cualquier público, se decidió hacer un shiny dashboard en R, de tal forma que la información pueda ser accedida por la aplicación web generada por el código. Esta aplicación se encuentra en el siguiente link:

<https://adrian-landaverde.shinyapps.io/ProyectoCocaColaF1/> .

Como se observa en la aplicación, aquí se realizó un análisis exploratorio para observar el comportamiento en general de algunas variables para las diferentes líneas. La aplicación permite la visualización de gráficos de líneas de las variables de Potencia eléctrica, Temperatura de succión, temperatura de descarga, presión de succión, presión de descarga y presión de aceite en cada uno de los 10 compresores para cada una de las 7 líneas respecto al tiempo (Figura 1 a 3). Además, se realizaron gráficos para comparar la producción de

cajas y el flujo de bebida en el mezclador para las líneas 1,4,5,6 y 7 respecto al tiempo (Figura 4 y 5).

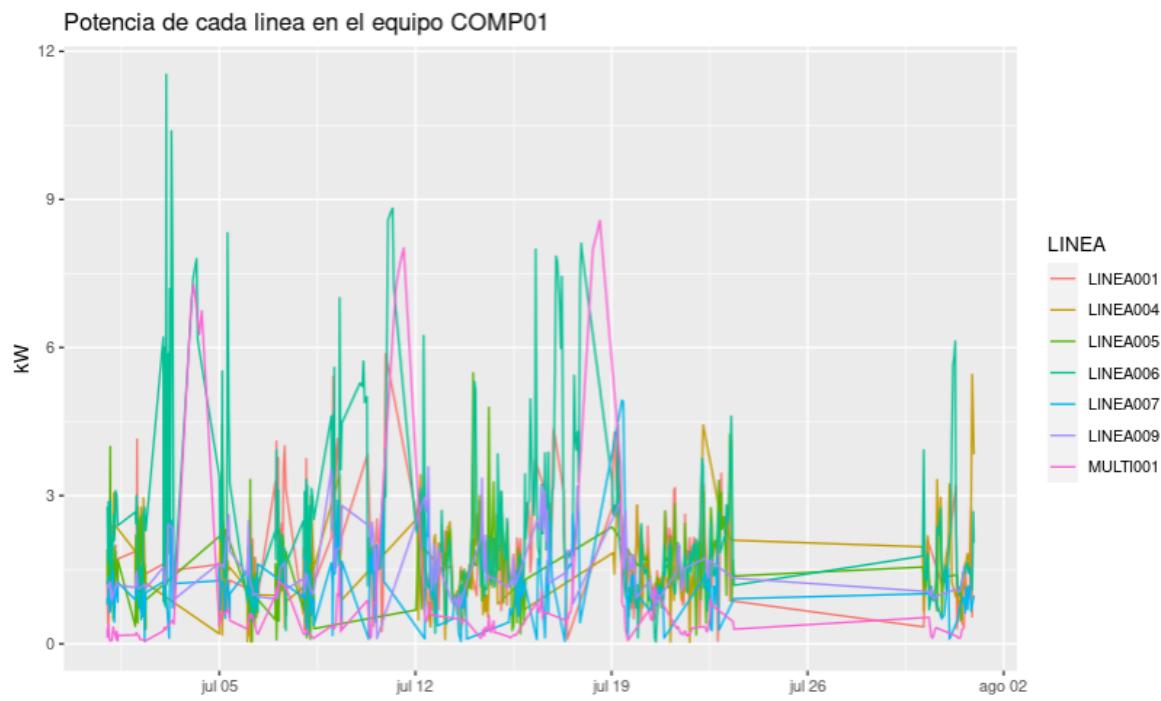


Figura 1. Gráfico de líneas de la potencia eléctrica por cada línea de producción en el compresor COMP01

PA de cada linea en el equipo COMP01

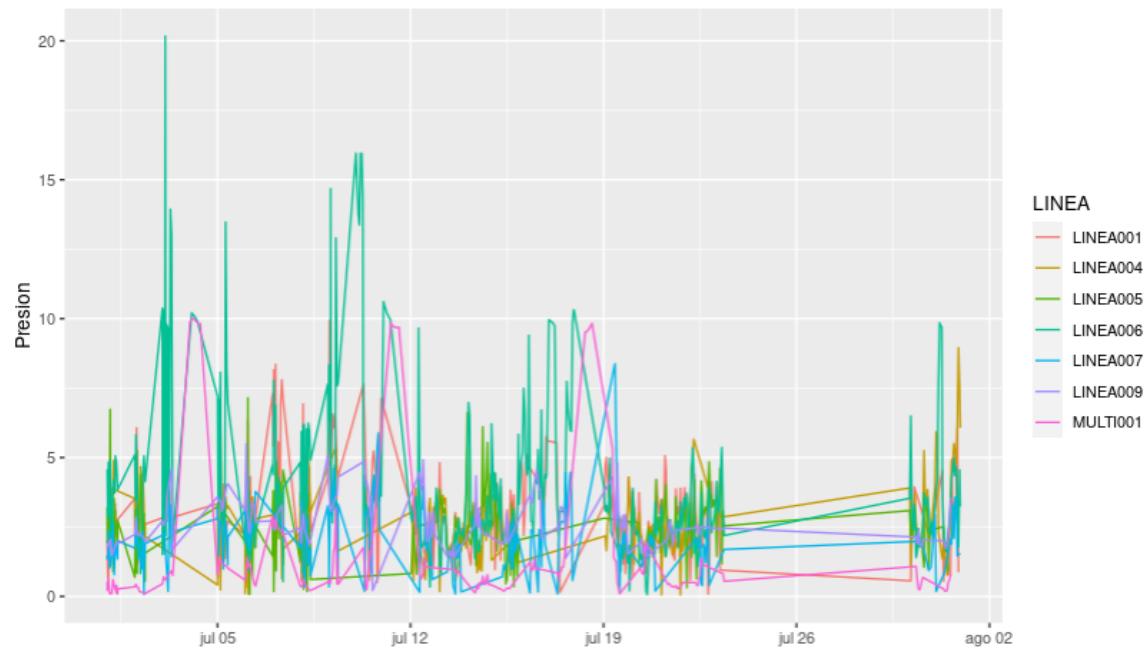


Figura 2. Gráfico de líneas de la presión de aceite por cada línea de producción en el compresor COMP01

TD de cada linea en el equipo COMP01

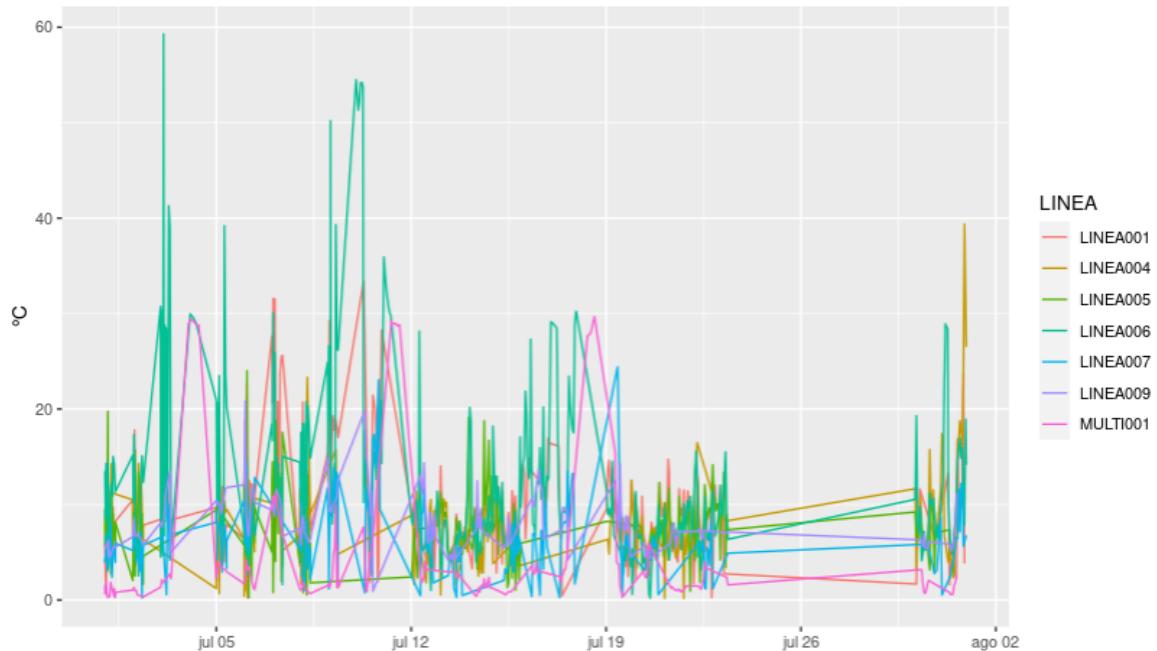


Figura 3. Gráfico de líneas de la temperatura de descarga por cada línea de producción en el compresor COMP01

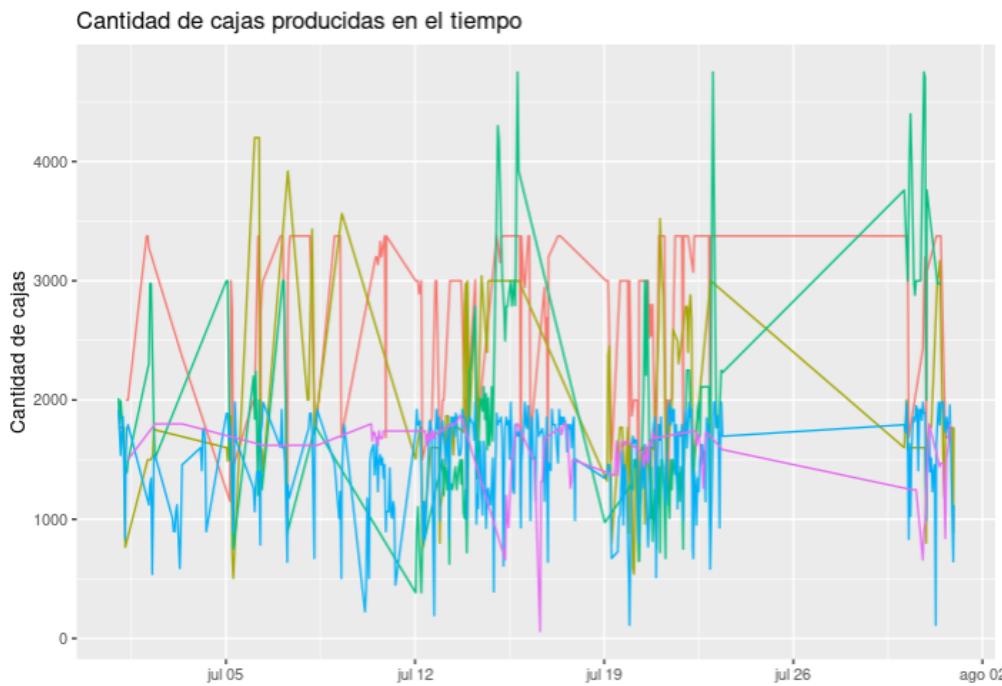


Figura 4. Gráfico de líneas de la cantidad de cajas producidas por cada línea de producción

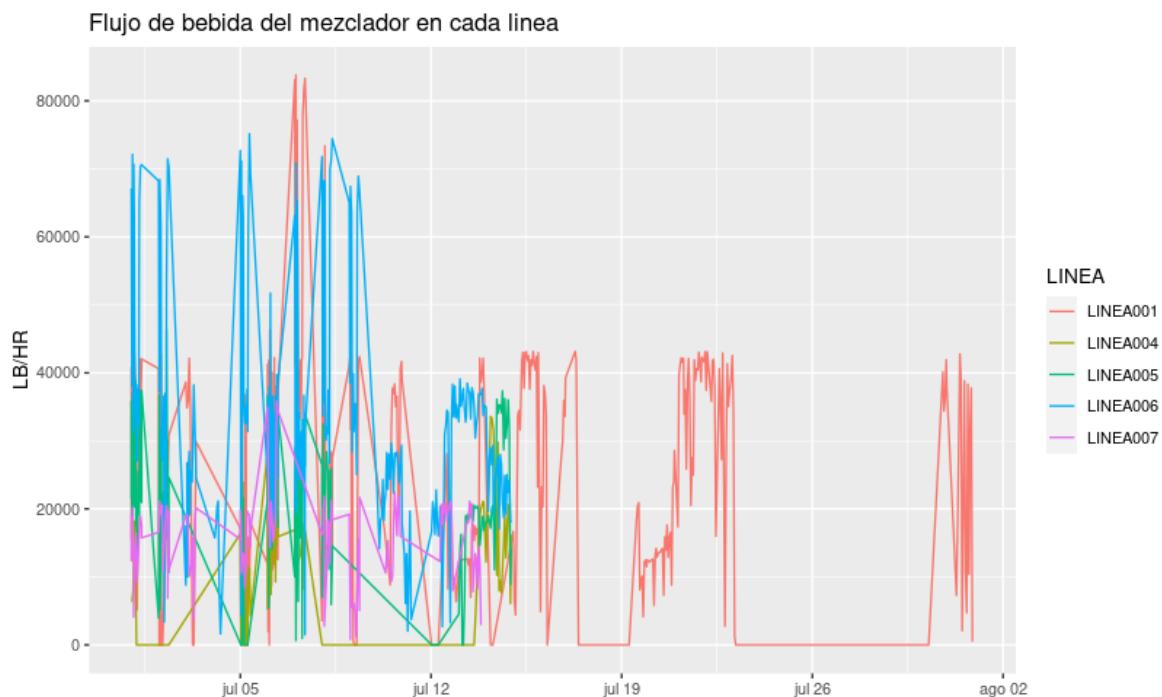


Figura 5. Gráfico de líneas del flujo de bebida en el mezclador por cada línea de producción

Asimismo, también se realizó un análisis del valor de estas variables, pero por bebidas en vez de líneas. Se encontró una gran cantidad de bebidas diferentes (15 diferentes BRAND), cada una con una diferente cantidad de datos, desde los 3 hasta los 300 datos. A grandes rasgos, estas bebidas mantienen hasta cierto punto un rango de datos estable cuando las bebidas pertenecen a una misma línea, pero cuando se produce la misma bebida en diferentes líneas, sus datos tienen una mayor variabilidad.

Para poder visualizar estos datos por bebida, primero se realizó un cuadro para obtener las medidas de tendencia central de estos tipos de bebidas (Figura 6). Y para observar qué tan variables eran los datos, se realizaron boxplots de las variables de los compresores. En uno de ellos se comparan todas las variables de un sólo compresor de una bebida (Figura 7). Mientras que en otro se compara 1 sola variable para todos los compresores de una bebida (Figura 8).

MIN	MEDIANA	MAX	PROMEDIO	DESVIACIONESTNDAR	VARIANZA	VARIABLE
0.000000	0.000000	376.524292	17.2841519994	57.686422644	3322.678226008028	MAAP_AUX_REFRIG_COMP01_EDIA
258.140228	5400.9248045	16925.927730	5899.8118577503	2563.221379108	6583545.365909105167	MAAP_AUX_REFRIG_COMP01_HM
0.109225	2.5080590	8.950888	2.7424089067	1.251278170	1.567629389244	MAAP_AUX_REFRIG_COMP01_PA
0.115417	2.6419644	8.858703	2.8464039730	1.255657208	1.579483075559	MAAP_AUX_REFRIG_COMP01_PD
0.062060	1.4392190	5.454117	1.5915415228	0.782517820	0.612453570574	MAAP_AUX_REFRIG_COMP01_PE
0.061872	1.4446865	5.518238	1.5959204700	0.788357387	0.621592596096	MAAP_AUX_REFRIG_COMP01_PS
0.320574	7.5176241	39.370129	8.4527036936	4.643886377	21.544607949561	MAAP_AUX_REFRIG_COMP01_TD
0.239670	6.5796900	16.522758	6.8136983809	2.989789589	8.955578210338	MAAP_AUX_REFRIG_COMP01_TS
0.000000	219.9942016	1154.977051	244.6599696234	220.334464909	48404.969907865241	MAAP_AUX_REFRIG_COMP02_EDIA
421.578705	8871.8442380	27922.121090	9676.9808714372	4206.485890442	17730556.936384133995	MAAP_AUX_REFRIG_COMP02_HM
0.182277	3.5379061	9.142625	3.7656355584	1.566746935	2.461717055661	MAAP_AUX_REFRIG_COMP02_PA
0.143541	2.9715384	8.831130	3.2170343909	1.358442575	1.850110376360	MAAP_AUX_REFRIG_COMP02_PD
0.000000	22.7041922	71.797913	21.1477593554	15.490689592	239.344093985871	MAAP_AUX_REFRIG_COMP02_PE
0.061971	1.4507915	5.553830	1.6020866505	0.792524152	0.628169777054	MAAP_AUX_REFRIG_COMP02_PS
0.630971	12.2882514	31.729275	13.3856791579	5.666590549	32.191250812339	MAAP_AUX_REFRIG_COMP02_TD
0.121658	3.1454355	22.933035	3.9762627160	2.902360703	8.402113544892	MAAP_AUX_REFRIG_COMP02_TS
0.000000	0.0000000	0.000000	0.0000000000	0.0000000000	0.000000000000	MAAP_AUX_REFRIG_COMP04_EDIA
272.124298	5689.2390140	17822.515630	6215.8694490765	2700.448907180	7307347.726458554156	MAAP_AUX_REFRIG_COMP04_HM
0.012483	0.2113960	0.579779	0.2342795184	0.109982078	0.012105873863	MAAP_AUX_REFRIG_COMP04_PA
0.013697	0.2362765	0.635519	0.2615511570	0.120809501	0.014610231304	MAAP_AUX_REFRIG_COMP04_PD

Figura 6. Primeras 20 variables de la tabla de medidas de tendencia central de las variables

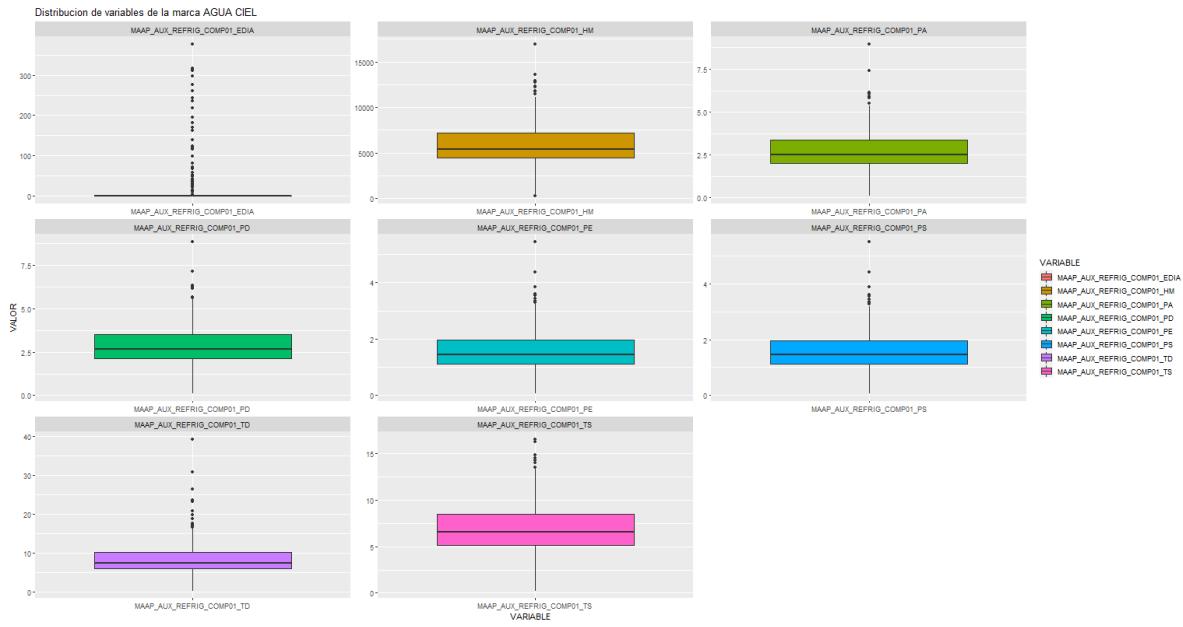


Figura 7. Boxplots de la distribución de las variables del compresor COMP01 para la bebida (BRAND) AGUA CIEL

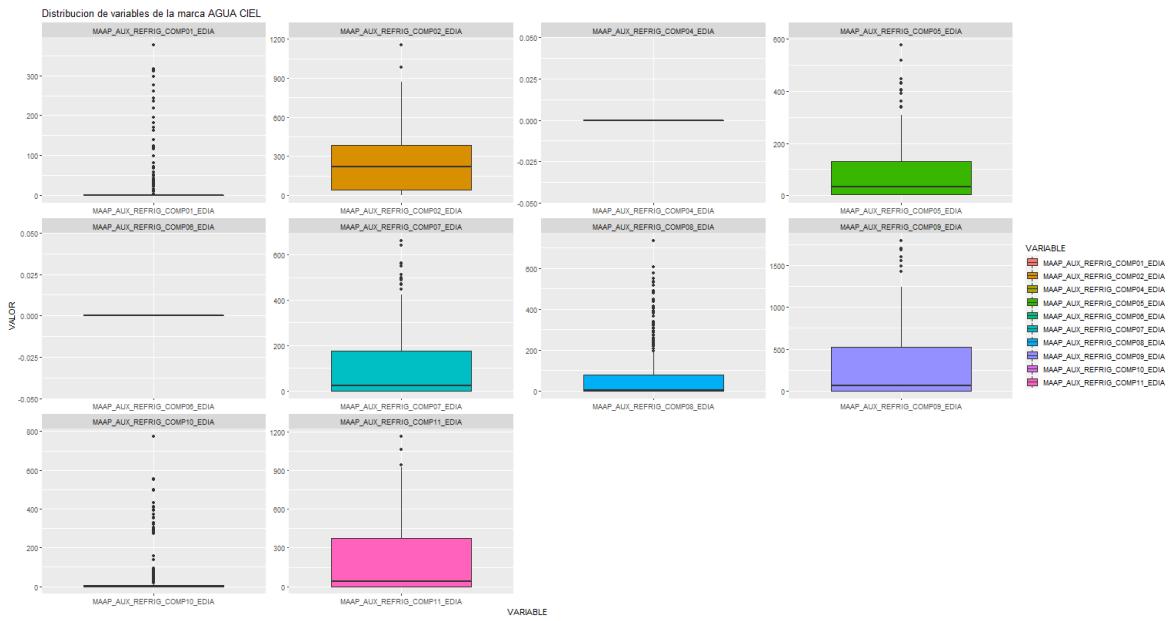


Figura 8. Boxplots de la distribución de la variable Energía por Día en cada compresor para la bebida (BRAND) AGUA CIEL

Con base en esta primera fase del proyecto, se encontró que la mayor cantidad de variables es producida por los compresores, y que a su vez estas variables son las más constantes, ya que se encuentran presentes en todas las líneas. Además, de manera general, las variables

de los compresores parecen estar correlacionadas, excepto la Energía por Día, que no sólo parece no tener relación con las otras variables, sino que también hay muchos datos donde su valor, por alguna razón, es 0, por lo que genera mucha variabilidad en los datos. Además, también se observó cierta tendencia entre los datos que fueran de la misma línea, ya que cuando son líneas diferentes, estos son muy diferentes.

FASE 2: Análisis de las variables de energía y potencia por SKU

El URL para tener acceso a todos los boxplots, gráficos y pruebas realizadas entre SKUs es el siguiente:

<https://adrian-landaverde.shinyapps.io/ProyectoCocaColaF2/>

En la fase de producción se define “SKU” la producción de una bebida definida por el “BRAND” (por ejemplo Coca-Cola, Fanta, Agua Ciel, ...), el sabor de la bebida “FLAVOR” (por ejemplo cola, naranja, agua mineral, ...), el tamaño de la botella “ITEM SIZE” (por ejemplo 1.5 litros, 600 ml, ...) y el tipo de material de la botella “MATERIAL TYPE” (PET, vidrio o garrafón). En la base de datos se identificaron 67 SKU diferentes. A cada integrante del equipo se le repartió una SKU diferente para analizar las variables relacionadas con energía y potencia de la línea donde se produjo la bebida.

Para nuestro equipo de trabajo, se usaron específicamente 6 SKU diferentes, los cuales se enlistan enseguida:

1. SKU 1: AGUA CIEL NATURAL GARRAFÓN 676.3 Onz / 20 LITROS
2. SKU 2: COCA-COLA COLA PET 33.8 Onz / 1 LITRO
3. SKU 3: COCA-COLA COLA PET 50.7 Onz / 1.5 LITROS
4. SKU 4: Coca-Cola Sin Azúcar COLA PET 20.3 Onz / 600 ML=CC
5. SKU 5: FANTA NARANJA PET 20.3 Onz / 600 ML=CC
6. SKU 6: Naranja y Nada NARANJA PET 20.3 Onz / 600 ML=CC

Debido a que se cuenta con 10 diferentes compresores, se incluyen a continuación únicamente los gráficos y los resultados de los intervalos de confianza y pruebas de hipótesis realizadas. Para los resultados restantes, se incluyó previamente la URL de la app en Shiny, con la oportunidad de poder visualizar los resultados y gráficos no sólo para estos 6 SKU, sino para los 67 identificados en la base de datos.

Mostrando gráficos de Boxplot para las variables EDIA (Energía por día) y PE (Potencia eléctrica) para los 6 SKU para los primeros 3 compresores (Figura 9 a figura 14):

Boxplots de la variable EDIA de los 6 SKU para el compresor COMP01

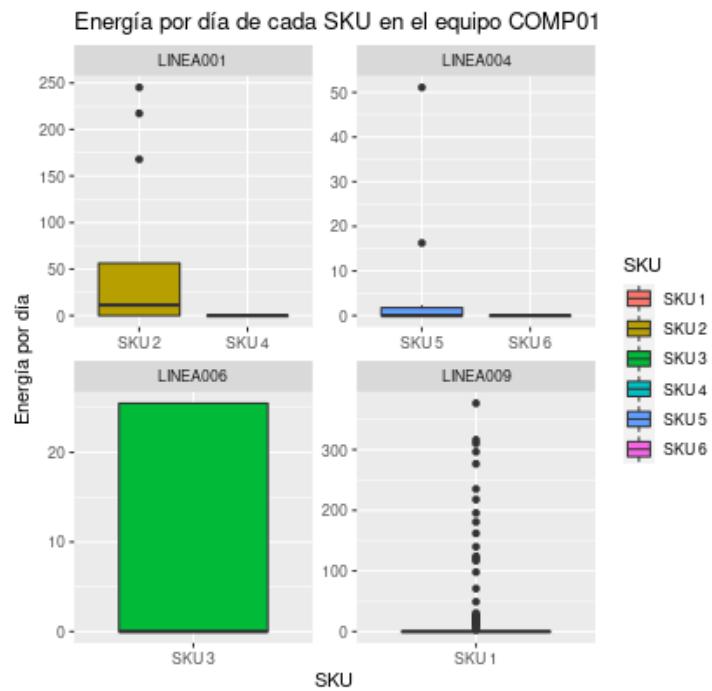


Figura 9. Boxplots de la distribución de la variable Energía por Día en el compresor 1 para cada SKU

Boxplots de la variable EDIA de los 6 SKU para el compresor COMP02

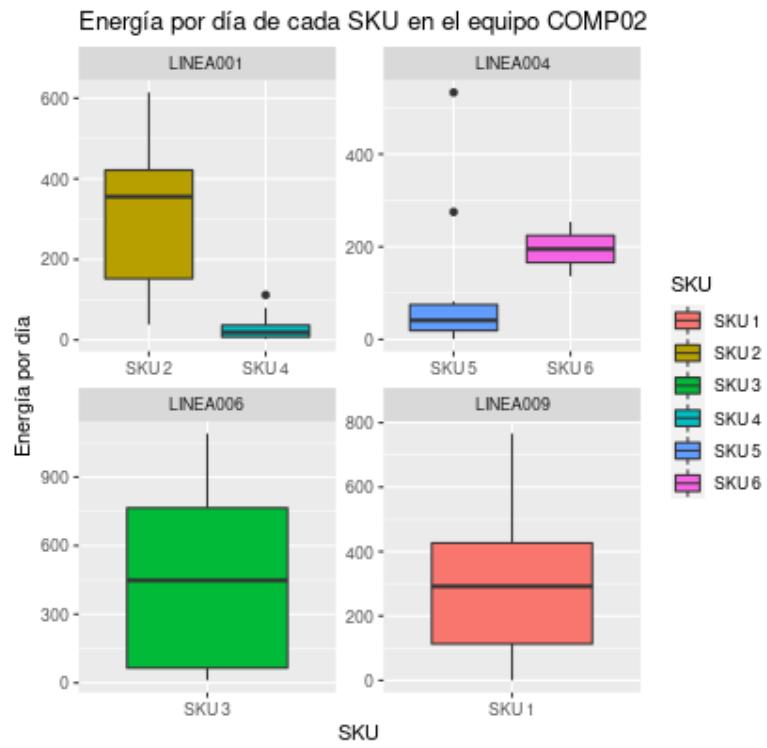


Figura 10. Boxplots de la distribución de la variable Energía por Día en el compresor 2 para cada SKU

Boxplots de la variable EDIA de los 6 SKU para el compresor COMP04

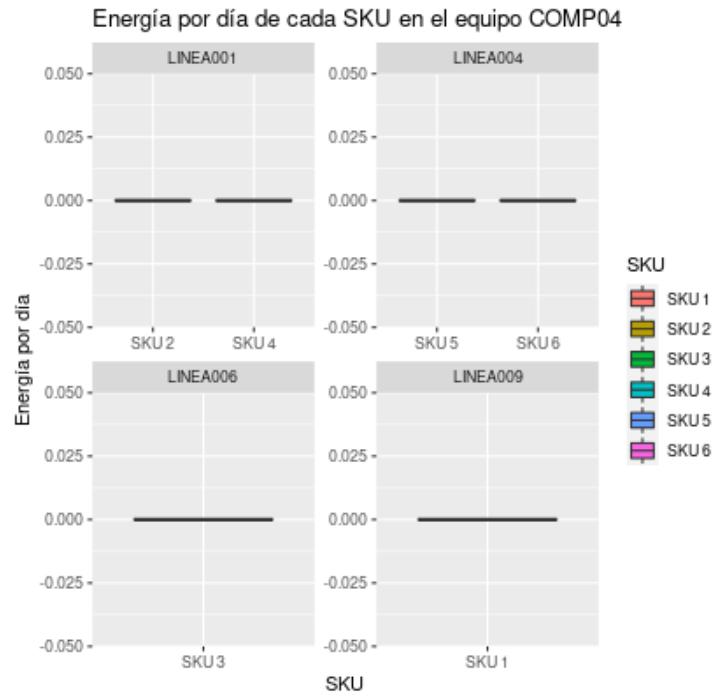


Figura 11. Boxplots de la distribución de la variable Energía por Día en el compresor 4 para cada SKU

Boxplots de la variable PE de los 6 SKU para el compresor COMP01

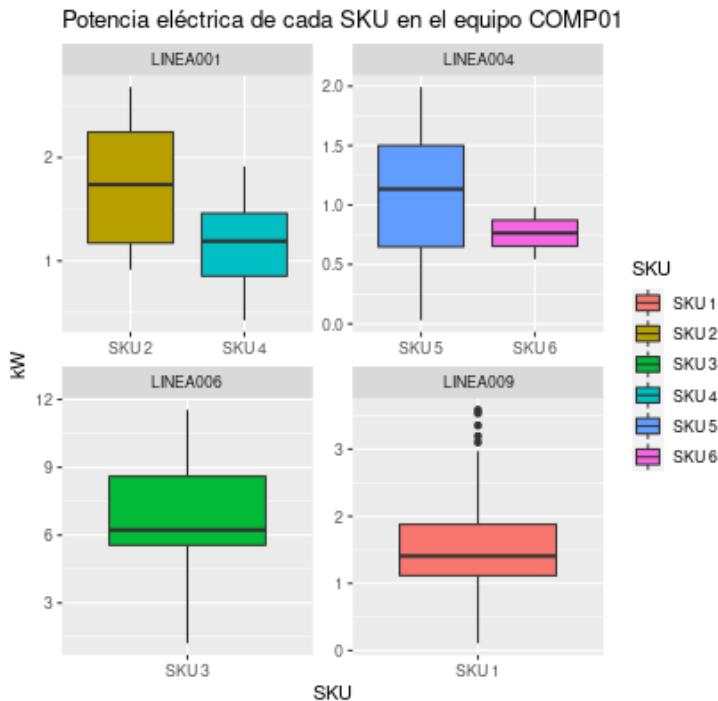


Figura 12. Boxplots de la distribución de la variable Potencia Eléctrica en el compresor 1 para cada SKU

Boxplots de la variable PE de los 6 SKU para el compresor COMP02

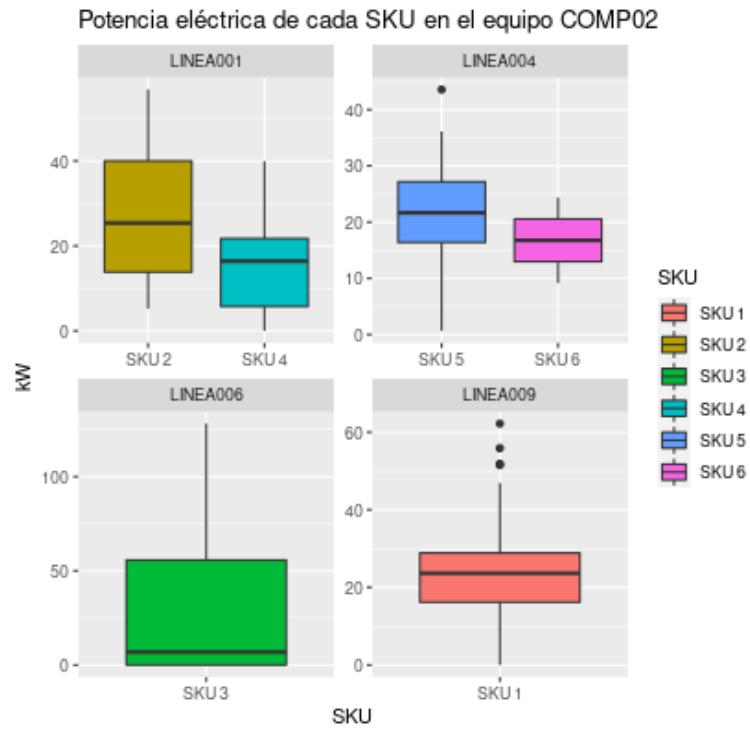


Figura 13. Boxplots de la distribución de la variable Potencia Eléctrica en el compresor 2 para cada SKU

Boxplots de la variable PE de los 6 SKU para el compresor COMP04

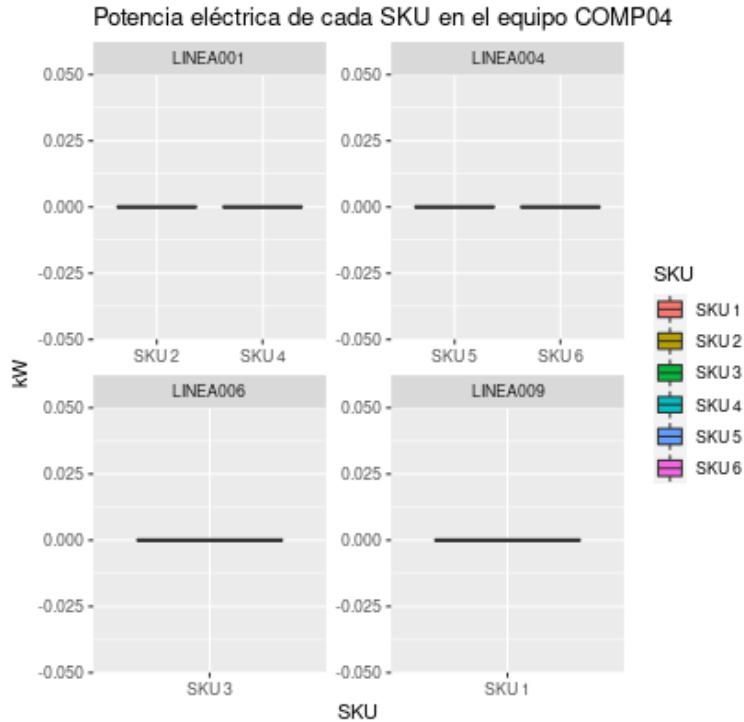


Figura 14. Boxplots de la distribución de la variable Potencia Eléctrica en el compresor 4 para cada SKU

A continuación se muestran los resultados de las pruebas de hipótesis de dos colas realizadas entre los 6 SKU, empleando el método de diferencia entre medias con un nivel de confianza del 95%, al igual que los intervalos de confianza calculados con el mismo porcentaje de confianza, y por lo tanto, una significancia de $\alpha = 0.05$.

Para la prueba de hipótesis:

Hipótesis	Prueba de 2 colas
H_0	$\mu_1 - \mu_2 = 0$
H_1	$\mu_1 - \mu_2 \neq 0$

Nuestra hipótesis nula será que las medias de las muestras de los sku y compresores analizados sean iguales, mientras que la hipótesis alternativa será que las medias de las muestras de los sku y compresores analizados sean diferentes. Si existe diferencia significativa, puede ser que se produzcan más botellas para cierta SKU, lo cual se puede controlar controlando el número de días u horas en que están activos los compresores.

Primero, se muestran los resultados únicamente para el primer compresor, con respecto a las variables de EDIA y PE (Figura 15 a figura 20):

Prueba de hipótesis e intervalo de confianza de SKU 1 y SKU 2 para el compresor COMP01 (EDIA)

```
Welch Two Sample t-test

data: VAR1 and VAR2
t = -1.3227, df = 11.778, p-value = 0.2111
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-96.25942 23.63121
sample estimates:
mean of x mean of y
21.59349 57.90760
```

Figura 15. Resultados de la prueba de hipótesis realizada para SKU 1 y SKU 2 con respecto a la variable de Energía por Día para el compresor 1

Se tiene que $p - value = 0.2111 > \alpha/2 = 0.05$, por lo que no rechazamos la hipótesis nula.

De igual manera, el intervalo de confianza calculado contiene al valor de 0.0. Por ello, tenemos evidencia estadística para no rechazar la hipótesis nula y decir que las medias de SKU 1 y SKU 2 para la variable EDIA en el compresor COMP01 son iguales.

Prueba de hipótesis e intervalo de confianza de SKU 3 y SKU 4 para el compresor COMP01 (EDIA)

```
Welch Two Sample t-test

data: VAR1 and VAR2
t = 3.1986, df = 19, p-value = 0.004729
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
3.080924 14.746691
sample estimates:
mean of x mean of y
8.913807 0.000000
```

Figura 16. Resultados de la prueba de hipótesis realizada para SKU 3 y SKU 4 con respecto a la variable de Energía por Día para el compresor 1

Se tiene que $p - value = 0.004729 < \alpha/2 = 0.05$, por lo que rechazamos la hipótesis nula.

De igual manera, el intervalo de confianza calculado no contiene al valor de 0.0. Por ello, tenemos evidencia estadística para poder rechazar la hipótesis nula y decir que las medias de SKU 3 y SKU 4 para la variable EDIA en el compresor COMP01 no son iguales.

Prueba de hipótesis e intervalo de confianza de SKU 5 y SKU 6 para el compresor COMP01 (EDIA)

```
Welch Two Sample t-test

data: VAR1 and VAR2
t = 1.3516, df = 9, p-value = 0.2095
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-4.698102 18.645834
sample estimates:
mean of x mean of y
6.973866 0.000000
```

Figura 17. Resultados de la prueba de hipótesis realizada para SKU 5 y SKU 6 con respecto a la variable de Energía por Día para el compresor 1

Se tiene que $p - value = 0.2095 > \alpha/2 = 0.05$, por lo que no rechazamos la hipótesis nula.

De igual manera, el intervalo de confianza calculado contiene al valor de 0.0. Por ello, tenemos evidencia estadística para poder no rechazar la hipótesis nula y decir que las medias de SKU 5 y SKU 6 para la variable EDIA en el compresor COMP01 son iguales.

Prueba de hipótesis e intervalo de confianza de SKU 1 y SKU 2 para el compresor COMP01 (PE)

```
Welch Two Sample t-test

data: VAR1 and VAR2
t = -0.76574, df = 12.588, p-value = 0.4579
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.5505491 0.2631078
sample estimates:
mean of x mean of y
1.566406 1.710126
```

Figura 18. Resultados de la prueba de hipótesis realizada para SKU 1 y SKU 2 con respecto a la variable de Potencia Eléctrica para el compresor 1

Se tiene que $p - value = 0.4579 > \alpha/2 = 0.05$, por lo que no rechazamos la hipótesis nula.

De igual manera, el intervalo de confianza calculado contiene al valor de 0.0. Por ello, tenemos evidencia estadística para poder no rechazar la hipótesis nula y decir que las medias de SKU 1 y SKU 2 para la variable PE en el compresor COMP01 son iguales.

Prueba de hipótesis e intervalo de confianza de SKU 3 Y SKU 4 para el compresor COMP01 (PE)

```
Welch Two Sample t-test

data: VAR1 and VAR2
t = 8.9261, df = 20.582, p-value = 1.617e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 4.000030 6.433973
sample estimates:
mean of x mean of y
 6.383758  1.166757
```

Figura 19. Resultados de la prueba de hipótesis realizada para SKU 3 y SKU 4 con respecto a la variable de Potencia Eléctrica para el compresor 1

Se tiene que $p - value = 1.617 \times 10^{-8} < \alpha/2 = 0.05$, por lo que rechazamos la hipótesis nula.

De igual manera, el intervalo de confianza calculado no contiene al valor de 0.0. Por ello, tenemos evidencia estadística para poder no rechazar la hipótesis nula y decir que las medias de SKU 3 y SKU 4 para la variable PE en el compresor COMP01 no son iguales.

Prueba de hipótesis e intervalo de confianza de SKU 5 y SKU 6 para el compresor COMP01 (PE)

```

Welch Two Sample t-test

data: VAR1 and VAR2
t = 0.97812, df = 3.6127, p-value = 0.3889
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.5952023 1.2017175
sample estimates:
mean of x mean of y
1.0672321 0.7639745

```

Figura 20. Resultados de la prueba de hipótesis realizada para SKU 5 y SKU 6 con respecto a la variable de Potencia Eléctrica para el compresor 1

Se tiene que $p - value = 0.3889 > \alpha/2 = 0.05$, por lo que no rechazamos la hipótesis nula.

De igual manera, el intervalo de confianza calculado contiene al valor de 0.0. Por ello, tenemos evidencia estadística para no rechazar la hipótesis nula y decir que las medias de SKU 5 y SKU 6 para la variable PE en el compresor COMP01 son iguales.

Con respecto a esta segunda fase del proyecto, pudimos notar que en la mayoría de los casos analizados anteriormente, se tienen promedios iguales para las muestras analizadas, por lo tanto, para los SKU analizados entre sí, se encuentran más resultados de diferencias no significativas entre sus medias que de diferencias significativas.

Análisis de las variables de energía y potencia: ANOVA de comparación entre SKU y pruebas para compresores

El URL para tener acceso a todas las pruebas de ANOVA realizadas en la app Shiny es el siguiente:

<https://adrian-landaverde.shinyapps.io/ProyectoCocaCola3/>

Se presentan los resultados de las pruebas de hipótesis realizadas e intervalos de confianza calculados entre los primeros 6 compresores, empleando el método de diferencia entre medias con un nivel de confianza del 95%, mostrando únicamente para el primer SKU (SKU 1), con respecto a las variables de EDIA y PE (Figura 21 a figura 26):

Prueba de hipótesis e intervalo de confianza de COMP01 y COMP02 para el SKU 1 (EDIA)

```
Welch Two Sample t-test

data: VAR1 and VAR2
t = -16.85, df = 203.9, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-298.7266 -236.1399
sample estimates:
mean of x mean of y
21.59349 289.02673
```

Figura 21. Resultados de la prueba de hipótesis realizada para el compresor 1 y el compresor 2 con respecto a la variable de Energía por Día para el SKU 1

Se tiene que $p - value = 2.2 \times 10^{-16} < \alpha/2 = 0.05$, por lo que rechazamos la hipótesis nula.

De igual manera, el intervalo de confianza calculado no contiene al valor de 0.0. Por ello, tenemos evidencia estadística para poder rechazar la hipótesis nula y decir que las medias de los compresores COMP01 y COMP02 para la variable EDIA con el SKU 1 no son iguales.

Prueba de hipótesis e intervalo de confianza de COMP04 y COMP05 para el SKU 1 (EDIA)

```
Welch Two Sample t-test

data: VAR1 and VAR2
t = -10.474, df = 167, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-105.6082 -72.1087
sample estimates:
mean of x mean of y
0.00000 88.85844
```

Figura 22. Resultados de la prueba de hipótesis realizada para el compresor 4 y el compresor 5 con respecto a la variable de Energía por Día para el SKU 1

Se tiene que $p - value = 2.2 \times 10^{-16} < \alpha/2 = 0.05$, por lo que rechazamos la hipótesis nula.

De igual manera, el intervalo de confianza calculado no contiene al valor de 0.0. Por ello, tenemos evidencia estadística para poder rechazar la hipótesis nula y decir que las medias de los compresores COMP04 y COMP05 para la variable EDIA con el SKU 1 no son iguales.

Prueba de hipótesis e intervalo de confianza de COMP06 y COMP07 para el SKU 1 (EDIA)

```
Welch Two Sample t-test

data: VAR1 and VAR2
t = -10.683, df = 167, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-141.91633 -97.64378
sample estimates:
mean of x mean of y
0.0000 119.7801
```

Figura 23. Resultados de la prueba de hipótesis realizada para el compresor 6 y el compresor 7 con respecto a la variable de Energía por Día para el SKU 1

Se tiene que $p - value = 2.2 \times 10^{-16} < \alpha/2 = 0.05$, por lo que rechazamos la hipótesis nula.

De igual manera, el intervalo de confianza calculado no contiene al valor de 0.0. Por ello, tenemos evidencia estadística para poder rechazar la hipótesis nula y decir que las medias de los compresores COMP06 y COMP07 para la variable EDIA con el SKU 1 no son iguales.

Prueba de hipótesis e intervalo de confianza de COMP01 y COMP02 para el SKU 1 (PE)

```

Welch Two Sample t-test

data: VAR1 and VAR2
t = -22.887, df = 167.91, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-22.88792 -19.25291
sample estimates:
mean of x mean of y
1.566406 22.636824

```

Figura 24. Resultados de la prueba de hipótesis realizada para el compresor 1 y el compresor 2 con respecto a la variable de Potencia Eléctrica para el SKU 1

Se tiene que $p - value = 2.2 \times 10^{-16} < \alpha/2 = 0.05$, por lo que rechazamos la hipótesis nula.

De igual manera, el intervalo de confianza calculado no contiene al valor de 0.0. Por ello, tenemos evidencia estadística para poder rechazar la hipótesis nula y decir que las medias de los compresores COMP01 y COMP02 para la variable PE con el SKU 1 no son iguales.

Prueba de hipótesis e intervalo de confianza de COMP04 y COMP05 para el SKU 1 (PE)

```

Welch Two Sample t-test

data: VAR1 and VAR2
t = -7.5601, df = 167, p-value = 2.544e-12
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-7.704339 -4.513676
sample estimates:
mean of x mean of y
0.000000 6.109007

```

Figura 25. Resultados de la prueba de hipótesis realizada para el compresor 4 y el compresor 5 con respecto a la variable de Potencia Eléctrica para el SKU 1

Se tiene que $p - value = 2.544 \times 10^{-12} < \alpha/2 = 0.05$, por lo que rechazamos la hipótesis nula.

De igual manera, el intervalo de confianza calculado no contiene al valor de 0.0. Por ello, tenemos evidencia estadística para poder rechazar la hipótesis nula y decir que las medias de los compresores COMP04 y COMP05 para la variable PE con el SKU 1 no son iguales.

Prueba de hipótesis e intervalo de confianza de COMP05 y COMP06 para el SKU 1 (PE)

```
Welch Two Sample t-test

data: VAR1 and VAR2
t = 7.5601, df = 167, p-value = 2.544e-12
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 4.513676 7.704339
sample estimates:
mean of x mean of y
 6.109007 0.000000
```

Figura 26. Resultados de la prueba de hipótesis realizada para el compresor 5 y el compresor 6 con respecto a la variable de Potencia Eléctrica para el SKU 1

Se tiene que $p - value = 2.544 \times 10^{-12} < \alpha/2 = 0.05$, por lo que rechazamos la hipótesis nula.

De igual manera, el intervalo de confianza calculado no contiene al valor de 0.0. Por ello, tenemos evidencia estadística para poder rechazar la hipótesis nula y decir que las medias de los compresores COMP05 y COMP06 para la variable PE con el SKU 1 no son iguales.

Se presentan a continuación los resultados obtenidos de la prueba ANOVA para los primeros 3 compresores, usando un nivel de 95% de confianza con respecto a las variables de Energía por Día y Potencia Eléctrica, comparando los 6 SKU entre sí (Figura 27 a figura 32):

Gráfico ANOVA de Energía por Día en el compresor COMP01

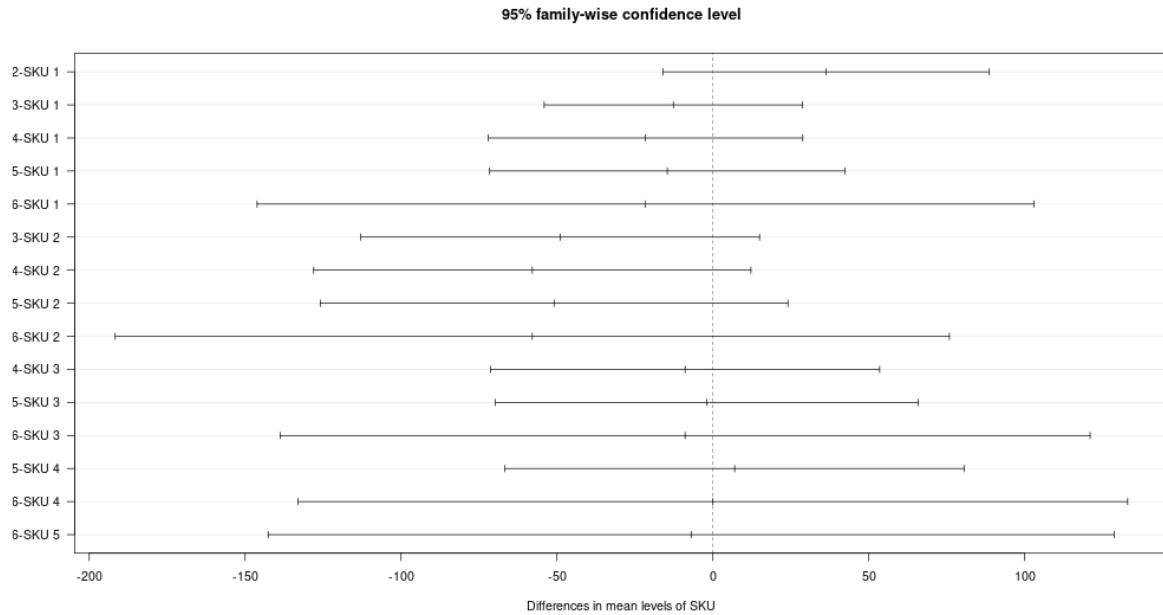


Figura 27. Resultados de ANOVA de la variable Energía por Día en el compresor 1

Observamos que para la variable EDIA en el compresor COMP01, no existe ningún SKU con diferencia significativa con respecto a ningún otro, dado que todos los intervalos calculados para las comparaciones entre los SKU entre sí incluye el valor 0.0.

Gráfico ANOVA de Potencia Eléctrica en el compresor COMP01

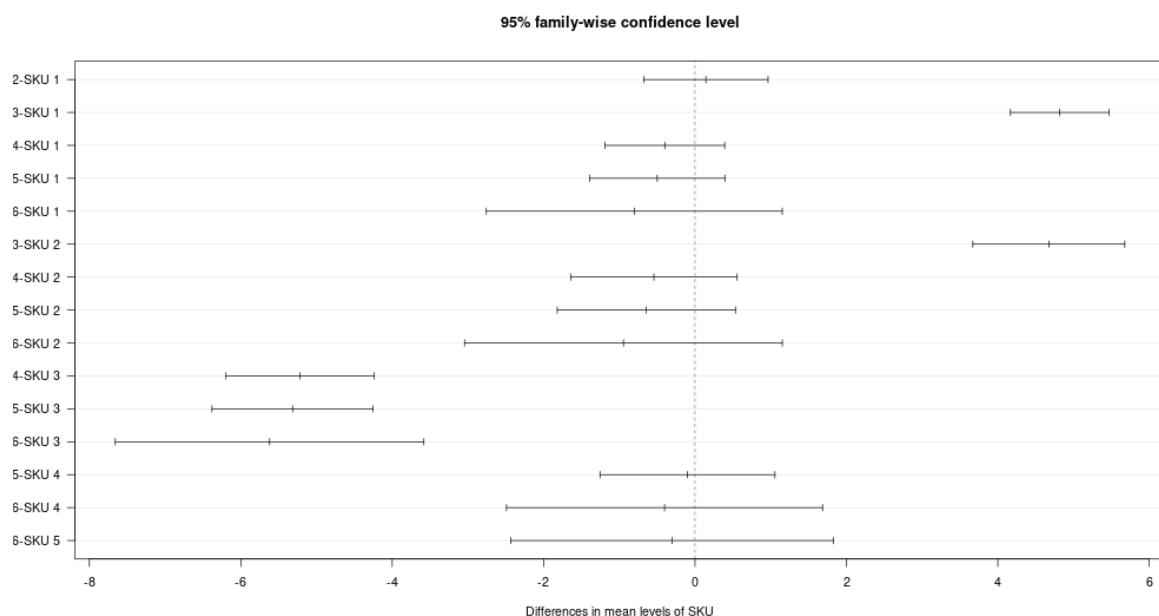


Figura 28. Resultados de ANOVA de la variable Potencia Eléctrica en el compresor 1

Observamos que para la variable PE en el compresor COMP01, los SKU entre los que se tiene una diferencia significativa para sus medias son SKU 3 Y SKU 1, SKU 3 y SKU 2, SKU 4 y SKU 3, SKU 5 y SKU 3, y por último, SKU 6 y SKU 3, debido a que los intervalos de confianza calculados para estos no incluyen al valor 0.0.

Gráfico ANOVA de Energía por Día en el compresor COMP02

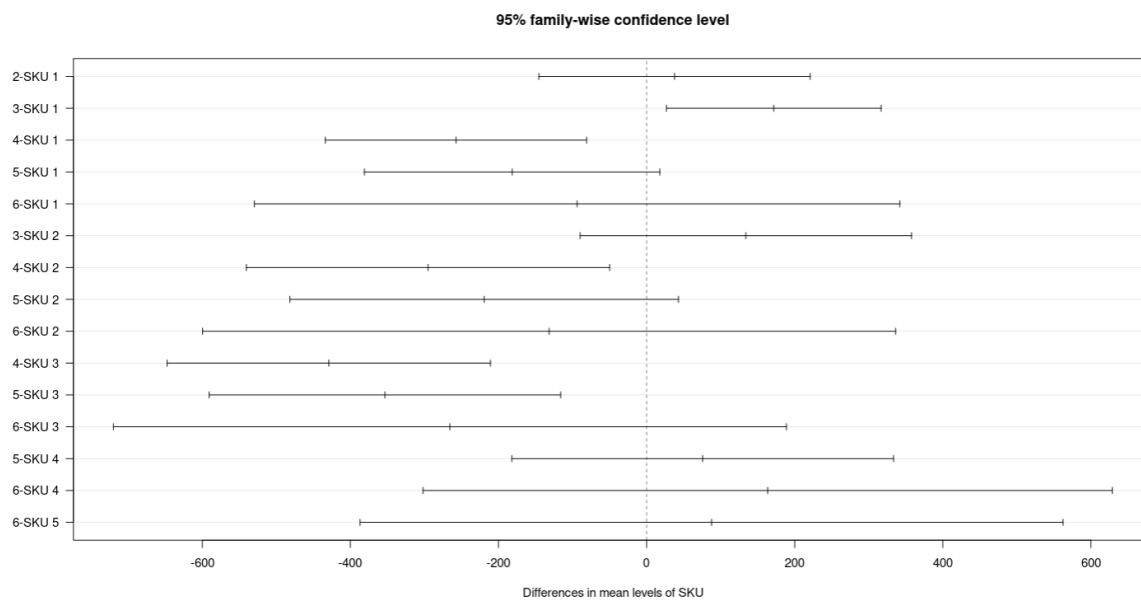


Figura 29. Resultados de ANOVA de la variable Energía por Día en el compresor 2

Observamos que para la variable EDIA en el compresor COMP02, los SKU entre los que se tiene una diferencia significativa para sus medias son SKU 3 y SKU 1, SKU 4 y SKU 1, SKU 4 y SKU 2, SKU 4 y SKU 3, y por último, SKU 5 y SKU 3, debido a que sus intervalos de confianza calculados para estos no incluyen al valor 0.0.

Gráfico ANOVA de Potencia Eléctrica en el compresor COMP02

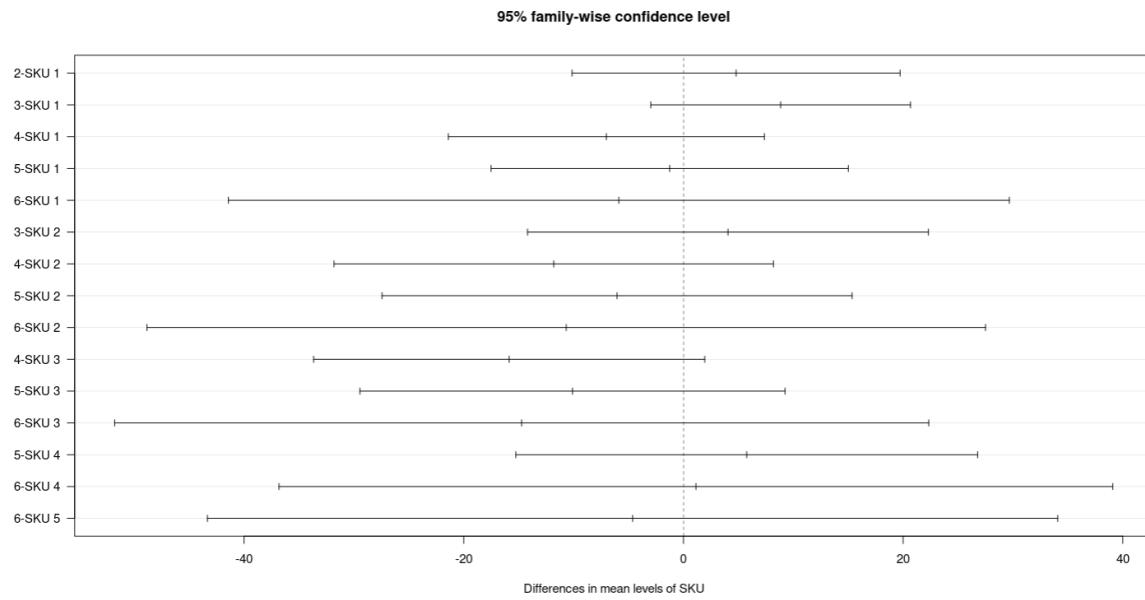


Figura 30. Resultados de ANOVA de la variable Potencia Eléctrica en el compresor 2

Observamos que para la variable PE en el compresor COMP02, no existe ningún SKU con diferencia significativa con respecto a ningún otro, dado que todos los intervalos calculados para las comparaciones entre los SKU entre sí incluye el valor 0.0.

Gráfico ANOVA de Energía por Día en el compresor COMP04

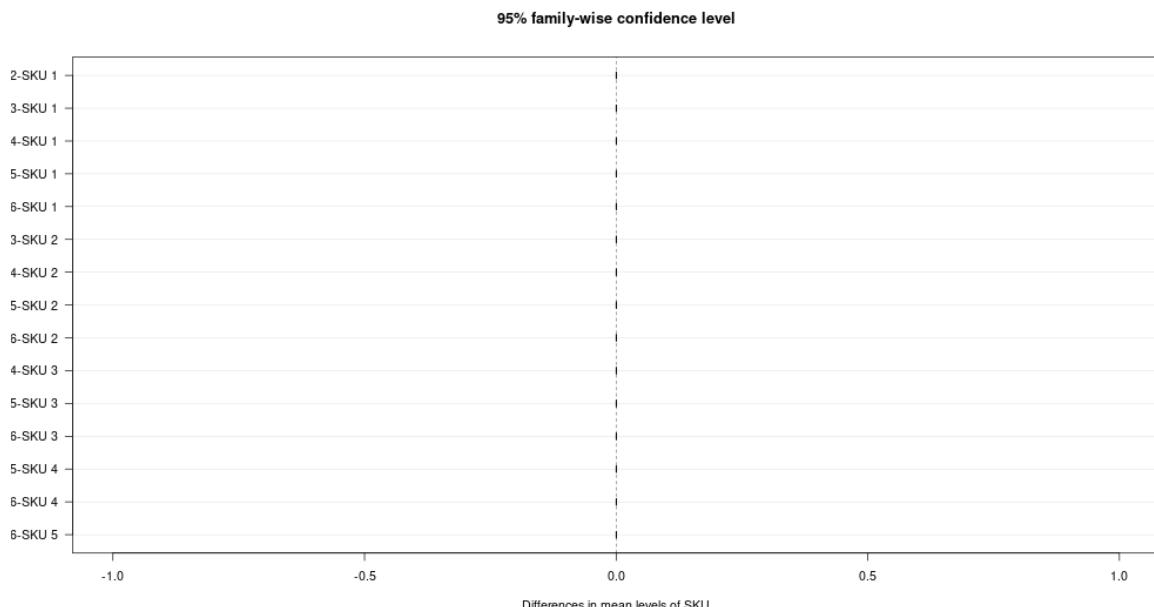


Figura 31. Resultados de ANOVA de la variable Energía por Día en el compresor 4

Para el compresor COMP04 no se tienen valores de la variable EDIA, por lo que no se puede realizar la comparación.

Gráfico ANOVA de Potencia Eléctrica en el compresor COMP04

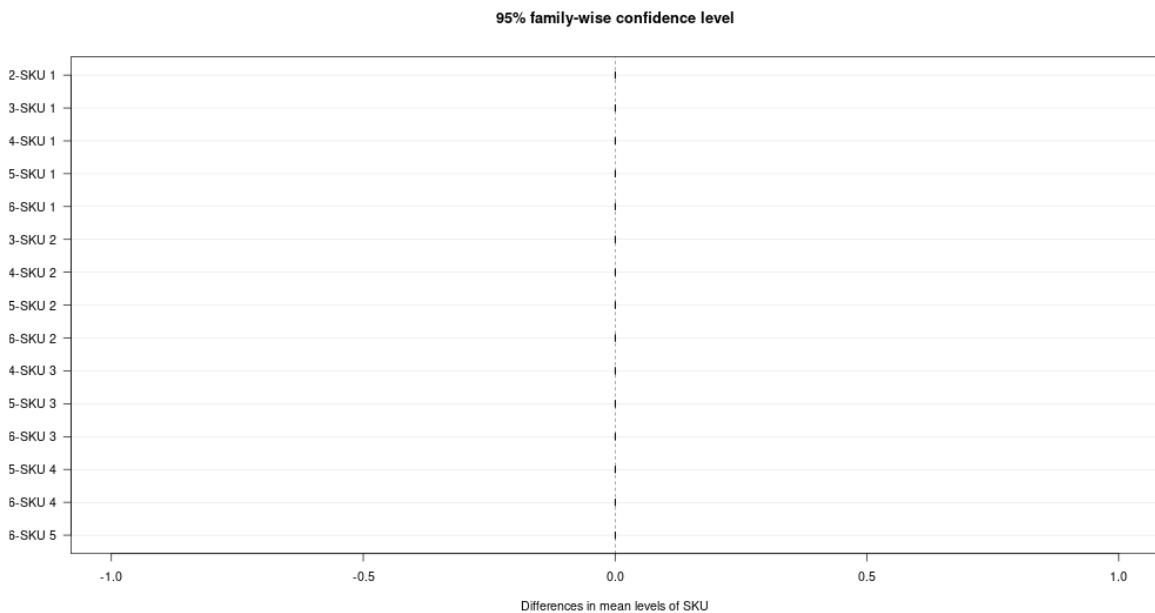


Figura 32. Resultados de ANOVA de la variable Potencia Eléctrica en el compresor 4

Para el compresor COMP04 no se tienen valores de la variable PE, por lo que no se puede realizar la comparación.

Análisis EXTRA

Para poder tener una respuesta con estos datos se optó por realizar algoritmos de Machine Learning para poder obtener alguna relación, estimación o predicción. Sin embargo, dado que estos algoritmos suelen ser más complicados que la estadística descriptiva, a partir de esta etapa se decidió usar Python, lenguaje de programación que, en conjunto con Scikit Learn, está optimizado para trabajar con Machine Learning en comparación con R. Los archivos para trabajar en Python se encuentran en la siguiente liga:

<https://drive.google.com/drive/folders/15LSyjcpoCVmrKryiqtrl-F6GJBlu0tT?usp=sharing>

En primer lugar, se optó por predecir los valores de Potencia Eléctrica de cada compresor. Y para esto, primero se omitieron los valores de la Energía por día, y los compresores 4 y 6. Lo primero debido a que la Energía por Día es muy variable y muchos de sus valores son 0, además de que desconocemos qué es exactamente la Energía por día, y la segunda acción se realizó porque para los compresores 4 y 6 la potencia eléctrica es 0 en casi todos sus datos.

Una vez esclarecido esto, se optó por realizar un algoritmo de Random Forest para predecir la potencia eléctrica de cada compresor en cada línea y con las líneas juntas. El siguiente procedimiento se encuentra en el archivo “Coca Cola Random Forest Loop.ipynb”. Y para comprobar la veracidad del modelo se calculó el Porcentaje de Error Absoluto Promedio (MAPE). Este porcentaje se observa en la figura 33.

	COMP01	COMP02	COMP05	COMP07	COMP08	COMP09	COMP10	COMP11
LINEA001	3.843437	27.436808	146.996636	29.957039	189.016878	4.472796	50.467753	16.775918
LINEA004	2.553214	21.087838	74.411128	27.110338	342.290743	22.467203	44.908405	19.431708
LINEA005	2.868967	17.955654	31.835187	38.088698	148.417604	4.678571	22.051449	48.313949
LINEA006	0.782131	79.939223	173.749594	36.436675	104.425787	27.386583	41.857207	7.544877
LINEA007	1.827036	54.225787	50.843061	41.289737	172.366129	8.673816	57.315550	187.535303
LINEA009	1.478911	13.989789	31.643856	60.946279	107.102660	10.647245	80.674056	3.541291
MULTI001	3.026704	25.080738	39.911213	77.459459	144.923856	171.779217	142.964784	89.821794
JUNTAS	0.583513	11.391338	54.655427	35.543585	87.058409	12.482571	113.373824	17.015543

Figura 33. Resultados del cálculo del Porcentaje de Error Absoluto Promedio (MAPE)

Además, se calculó la importancia de cada variable en su uso para este algoritmo, obteniendo así un gráfico de barras con la importancia de estas, como se muestra en la figura 34. Y al obtener estos gráficos, se encontró que de manera general, las variables que más contribuyen a este cálculo son la presión de aceite y la temperatura de succión.

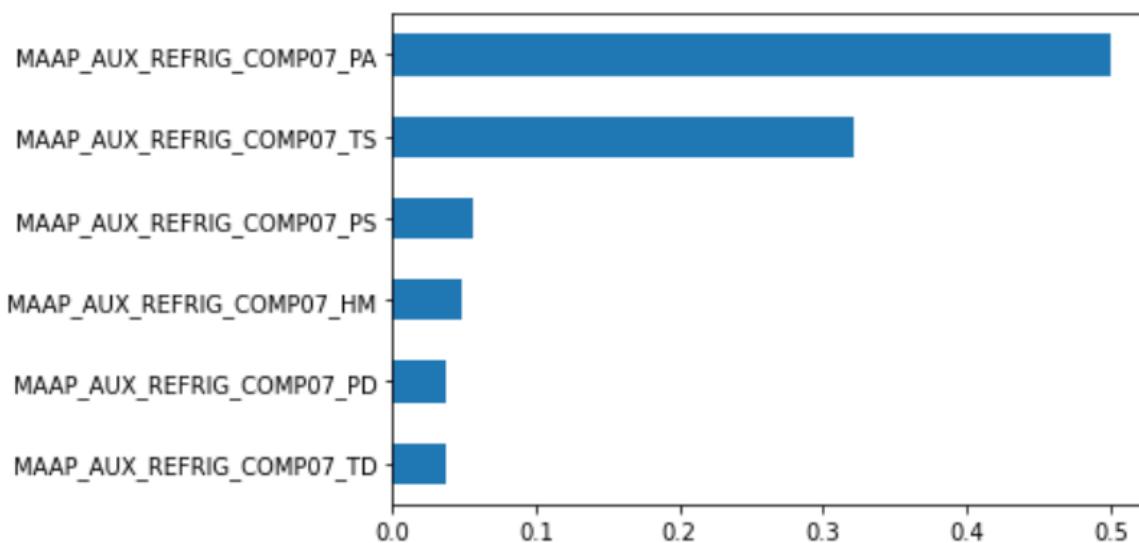


Figura 34. Gráfico de barras de la importancia de cada variable para el algoritmo.

Por lo tanto, con base en lo anterior, se encontró que este algoritmo de Random Forest con 100 árboles puede describir en gran medida la potencia eléctrica de los compresores COMP01, COMP02, COMP07, COMP09 y COMP11, siendo la Presión de Aceite y la Temperatura de succión las variables más dominantes para este cálculo.

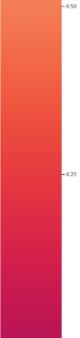
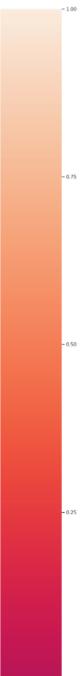
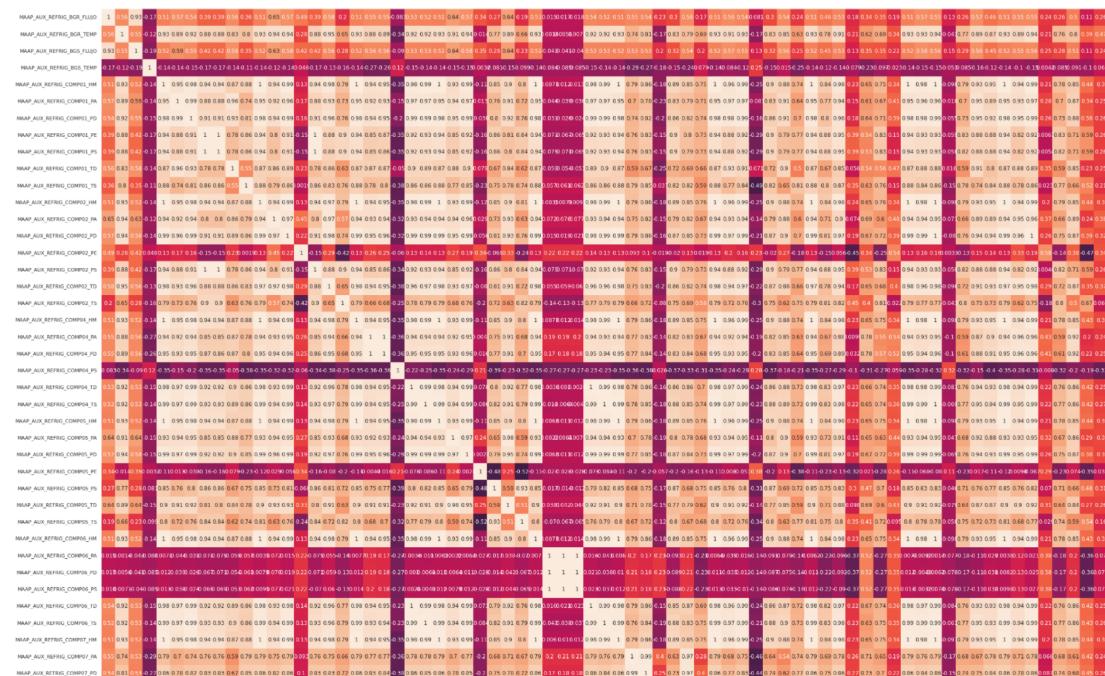
Por otro lado, para predecir la Energía por Día se tuvo que hacer otra selección de datos, ya que esta variable, al igual que la Potencia eléctrica, sólo tiene 0's en los compresores COMP04 y COMP06. Y además de eso, tiene más datos donde sus valores son 0. E incluso, esta energía dice ser “por día”, pero hay diferentes datos de energía por día en un mismo día. Por lo tanto, se intuyó que esta Energía por Día pertenece a la Línea en sí, y no a los compresores, por lo que la Energía por Día de 1 línea corresponde a la suma de Energía por Día de los compresores en un momento determinado. Por lo que se eliminaron las columnas de Energía por Día y se creó una columna de Energía Total.

Igual que para la potencia eléctrica, se decidió usar un algoritmo de Random Forest para predecir la Energía por día, pero en vez de usar como variables las de 1 compresor, se usaron todas las variables, por lo que en este caso, se separaron los datos por línea, ya que no todas las líneas tienen las mismas variables. El procedimiento se encuentra en el archivo “Random Forest EDIA Loop.ipynb”. Sin embargo, había datos de X donde este era 0, por lo que se realizó un proceso de imputación para cambiar estos valores por el promedio de esa columna. Al realizar este algoritmo en cada línea para con diferentes cantidades de árboles (de 100 a 1000), se obtuvo la tabla de valores de MAPE que se muestra en la figura 35.

	LINEA001	LINEA004	LINEA005	LINEA006	LINEA007	LINEA009	MULTI001
100	154.857292	153.475027	120.268661	139.714619	142.955713	51.661641	114.822235
200	152.807497	155.572954	119.612640	133.183863	140.775514	49.723193	117.442326
300	151.514054	153.354361	119.455488	132.612199	140.348403	48.374344	117.013944
400	150.550707	153.045110	118.648247	134.614093	139.050977	47.713863	116.042397
500	150.427782	153.763426	117.618311	134.970658	138.975057	47.607378	115.835559
600	149.905731	151.637353	116.822479	133.784355	137.989707	47.488506	116.501881
700	149.523190	150.526582	116.715094	135.685728	137.927148	47.926802	116.662039
800	149.522592	149.535529	116.537616	134.527555	137.684759	47.880147	116.914674
900	149.849853	149.003444	115.773641	133.672386	138.254277	47.814798	116.474337
1000	149.173886	148.212562	116.042658	134.156452	138.242484	47.822501	116.197361

Figura 35. Resultados del cálculo del Porcentaje de Error Absoluto Promedio (MAPE)

Sin embargo, como se puede observar, se obtuvieron valores de MAPE muy altos, superiores al 100%, excepto la línea 9 con un error de 50%. Por lo tanto, se decidió reestructurar las variables usando un algoritmo de Principal Component Analysis (PCA) para reducir las dimensiones de los datos. Este procedimiento se encuentra en el archivo “Exploración EDIA.ipynb”. Antes de esto, primero se realizó un gráfico de correlación de las variables, este se muestra en la figura 36, donde podemos apreciar que mientras más claro es el color, mayor es la correlación de las variables.



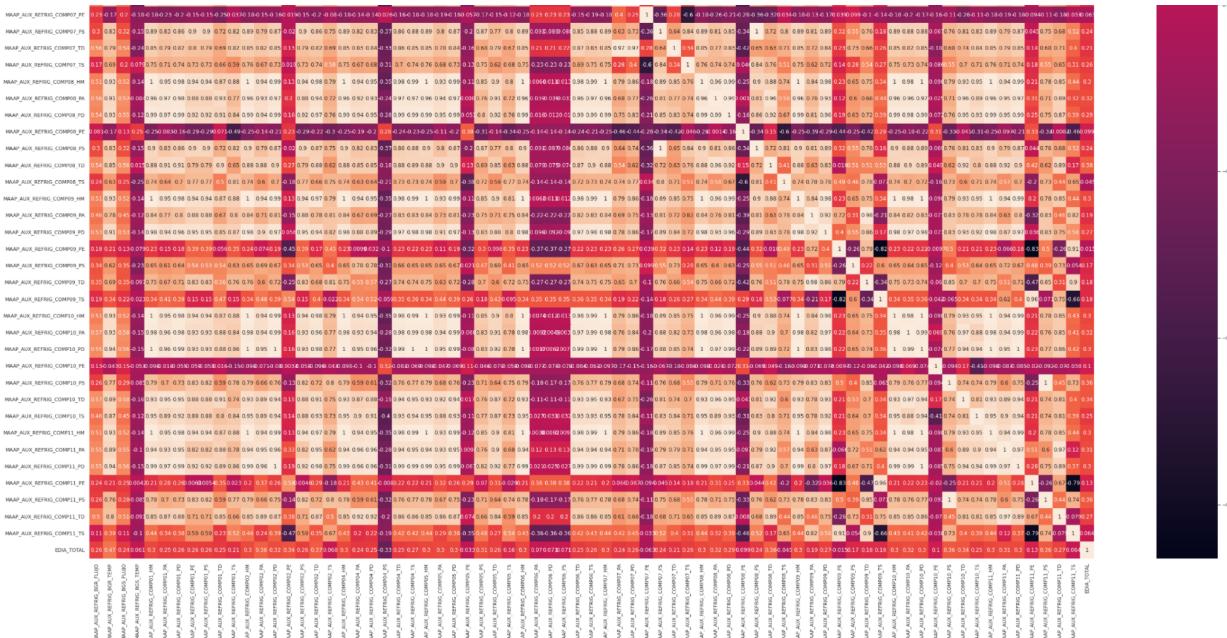


Figura 36. Gráfico de correlación entre las variables de la linea 1

Como se puede observar, muchas variables están correlacionadas entre sí, sobre todo los compresores, aunque con pequeñas excepciones, la mayoría de sus variables están correlacionadas, incluso con otros compresores. Es por esto que realizar el algoritmo de PCA es válido, ya que se reducirán estas correlaciones redundantes.

Al hacer el algoritmo de PCA se pueden reducir las variables al número que se quiera, por lo que pueden ser desde 1 variable hasta el mismo número de variables que tienen los datos. Estas nuevas dimensiones de los datos serán distintas entre sí, por lo que la correlación entre ellos será muy pequeña, como se observa en la Figura 37, donde se puede ver que la correlación entre las variables es muy pequeña.



Figura 37. Gráfico de correlación después de aplicar PCA

Una vez teniendo esto en cuenta, se volvió a realizar el algoritmo de Random Forest para cada línea con diferentes cantidades de variables resultantes (de 1 a 10 variables) y un valor fijo de 1000 árboles de decisión. Y se realizó una tabla con los valores de MAPE resultantes, misma que se muestra en la Figura 38.

	LINEA001	LINEA004	LINEA005	LINEA006	LINEA007	LINEA009	MULTI001
1	156.264385	407.770700	137.677792	376.590367	213.340096	100.736479	135.507609
2	154.432823	362.046731	121.261256	317.825764	176.778548	82.077479	139.925668
3	153.811009	294.135817	100.363051	247.837458	169.344229	64.811844	134.226554
4	150.714276	176.676938	106.389469	220.575432	157.525213	65.702744	113.874636
5	157.765167	166.273100	109.258391	161.976995	142.091519	60.920532	119.706589
6	154.110388	178.404103	110.841837	152.796977	135.189100	59.040168	115.312350
7	149.570608	179.056169	110.492292	141.428278	136.312525	56.054311	116.216370
8	151.441126	189.483723	109.591471	146.108274	136.886378	50.982551	124.775945
9	153.817244	210.291997	107.684980	141.451996	139.820560	53.867219	124.492817
10	146.138884	208.558275	107.905008	141.734884	143.016938	54.928986	124.496662

Figura 38. Resultados del cálculo del Porcentaje de Error Absoluto Promedio (MAPE)

Con base en estos resultados, se puede observar que la EDIA se comporta de una manera muy diferente al resto de las variables, ya que desde los gráficos de correlación (tanto antes como después de la PCA) muestran poca correlación con las diferentes variables, por lo que aunque se pudo aplicar la PCA correctamente, el modelo de Random Forest no se mejoró mucho, implicando que no se puede predecir adecuadamente la energía por día total mediante estas variables.

Resultados

Para los seis SKU analizados, y al hacer pruebas de hipótesis por pares de SKU, se notó un patrón: que no hay una diferencia significativa entre las medias de la variable EDIA entre todos estos SKU en el compresor 7. Esto significa que las seis SKU analizadas presentan valores similares de EDIA en dicho compresor. De hecho, al ver los *boxplot* para esta variable en cada compresor, es posible notar que los valores para cada SKU abarcan más o menos un rango similar, con valores que tienden más a ser bajos:

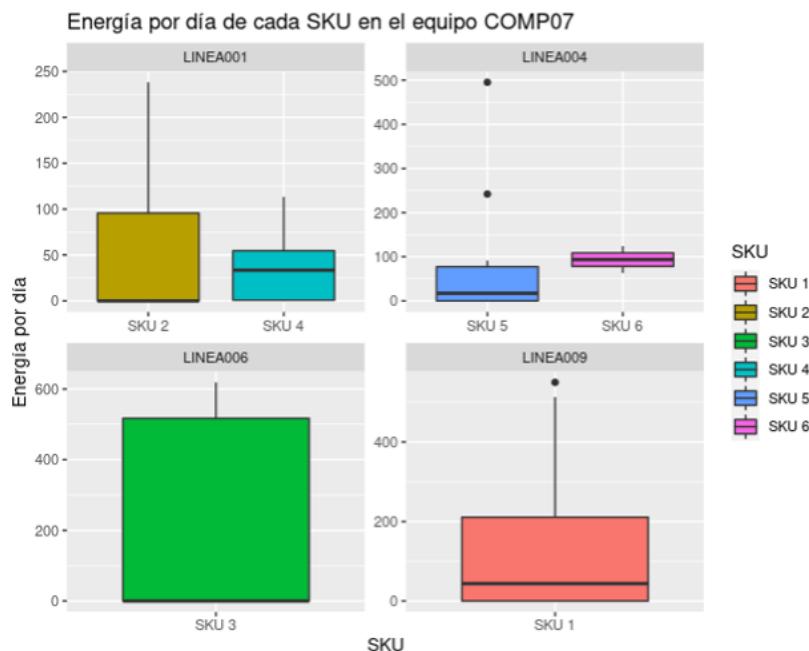


Figura 39. Gráficos boxplot para la variable EDIA, en el compresor 7, para los seis SKU analizados

Este es un patrón que se notó haciendo varias pruebas de hipótesis, lo que no es tan sencillo de hacer manualmente si se busca encontrar más patrones de este tipo, con tantas variables

y datos. No obstante, fue posible llegar a conclusiones más formales con el uso de algoritmos de Machine Learning.

En cuanto a la potencia eléctrica de cada compresor en cada línea fue posible predecir este valor usando un algoritmo de Random Forest de manera general, ya que el MAPE resultante fue relativamente bajo, aunque las líneas en que tuvo una menor eficiencia el modelo fue en los compresores 5, 8 y 10. Por lo tanto, se pudo obtener satisfactoriamente un modelo para predecir la potencia eléctrica en los compresores 2, 3, 7, 9, y 11.

En segundo lugar, en cuanto a la Energía por día, se tuvo que hacer una reducción de las dimensiones de los datos mediante PCA, ya que los resultados iniciales para predecir esta variable fueron poco eficientes. Sin embargo, al hacer esta reducción en las dimensiones, de todos modos seguimos obteniendo un valor muy grande de MAPE, por lo que aún eligiendo ciertas variables y aplicando procesos de limpieza de datos, el modelo no fue muy eficiente, siendo la línea 9 la que tuvo mayor eficiencia en el modelo con un 50% de porcentaje de error.

Además, en cuanto al análisis y la graficación de las variables que se nos fueron proporcionadas para los 6 SKUs con respecto a los compresores, se halló que para el compresor 1, en la mayoría de las variables (Horas de marcha, Potencia Eléctrica, Presión de Aceite, Presión de succión, Presión de Descarga, Temperatura de succión y Temperatura de Descarga) el SKU 3 “COCA-COLA COLA PET 50.7 Onz / 1.5 LITROS” es el que presenta y alcanza mayores valores. Esto se repite para los otros 9 compresores, teniendo excepciones únicamente para la Presión de succión en el compresor 4, la Potencia Eléctrica en el compresor 5, la Potencia Eléctrica en el compresor 8 y la Potencia Eléctrica en el compresor 10, además de la Energía por Día, la cual suele variar constantemente entre compresores, y para los cuales el SKU 3 registró los valores más altos sólo en el caso de los compresores 2, 7, 9 y 11 (4 de los 10 compresores).

Podríamos decir que esto significa que, este SKU, a pesar de no ser el que más activo se encuentra en cuanto a días, sí es el que más rendimiento presenta en cuanto a las variables consideradas y casi todos los compresores utilizados, y si se busca obtener un mejor rendimiento y calidad del producto final, es importante determinar de qué manera se puede hacer más eficiente y se puede ahorrar y/o reducir su impacto en el rendimiento y gasto o uso de elementos importantes como la energía, electricidad, etcétera.

Por otro lado, basándonos en los resultados obtenidos por las pruebas de hipótesis realizadas y los intervalos de confianza del 95% calculados, aquellos SKU para los que rechazamos la hipótesis nula y por ende se encontraron con una diferencia significativa entre sus medias para los distintos 10 compresores, son los siguientes:

- Compresor 1:
 - EDIA: Ninguno
 - PE: SKU 3 y 1, SKU 3 y 2, SKU 4 y 3, SKU 5 y 3, SKU 6 y 3
- Compresor 2:
 - EDIA: SKU 3 y 1, SKU 4 y 1, SKU 4 y 2, SKU 4 y 3, SKU 5 y 3
 - PE: Ninguno
- Compresor 4:
 - EDIA: Sin datos
 - PE: Sin datos
- Compresor 5:
 - EDIA: SKU 4 y 2, SKU 5 y 2
 - PE: SKU 2 y 1, SKU 3 y 2, SKU 4 y 2, SKU 5 y 2, SKU 6 y 2
- Compresor 6:
 - EDIA: Sin datos
 - PE: Sin datos
- Compresor 7:
 - EDIA: Ninguno
 - PE: Ninguno
- Compresor 8:
 - EDIA: SKU 2 y 1, SKU 3 y 2, SKU 4 y 2, SKU 5 y 2
 - PE: SKU 2 y 1, SKU 3 y 2, SKU 4 y 2, SKU 5 y 2, SKU 6 y 2
- Compresor 9:
 - EDIA: SKU 3 y 2, SKU 4 y 3
 - PE: Ninguno
- Compresor 10:
 - EDIA: Ninguno

PE: Ninguno

- Compresor 11:

EDIA: SKU 4 y 3, SKU 5 y 3

PE: SKU 3 y 1, SKU 4 y 3, SKU 5 y 3

Por lo que, para el caso de cada uno de estos SKUs en sus respectivos compresores, se producen más botellas para dichas SKUs, y esto se deberá controlar controlando el número de días u horas en que están activos los compresores.

Conclusión

Mejoras posibles y limitaciones del análisis

Lo que va a requerir el próximo paso del análisis es entender más a fondo cómo fue calculada la variante que representa a energía de los compresores y sería necesario conocer también la energía total consumida por día, o en dado caso la porción de energía consumida en las otras fases de producción, para poder estudiar cuales son las variables que más influye en la energía, para eventualmente poder construir un modelo.

Un ejemplo de las limitaciones del análisis que se hizo es que las muestras de datos para ciertos SKUs eran muy pequeñas (por ejemplo, 12 para 'FANTA NARANJA 20.3 Onz / 600 ML=CC PET', 13 para 'Coca-Cola Sin Azúcar COLA 20.3 Onz / 600 ML=CC PET', e incluso 2 para 'Naranja y Nada NARANJA 20.3 Onz / 600 ML=CC PET'). A pesar de lo anterior, se les aplicó la misma serie de métodos estadísticos, como la prueba de hipótesis y cálculo de intervalos de confianza para variables como EDIA, que a las muestras de tamaño adecuado. Estos datos finalmente se usaron para análisis posteriores, a pesar de que, desde el punto de vista estadístico, no es correcto tratar estas muestras pequeñas así.

Luego, se asumió normalidad para las muestras de los datos analizados, aunque no estamos completamente seguros de que este sea el caso. Lo anterior se hizo para poder llevar a cabo los tipos específicos de pruebas en todas las muestras. En ocasiones es necesario simplificar de esta forma.

Hay que interpretar los resultados para cada prueba teniendo en cuenta todo lo anterior. Esto lleva a otra de las limitaciones del análisis: los datos proporcionados eran poco consistentes, como en uno de los casos anteriores, para el número de datos muy bajo para algunas SKU. Si bien es trabajo de la ciencia de datos sacar conclusiones de un grupo de datos aparentemente aislados, si los datos son insuficientes, no se sabe interpretarlos o no

están recolectados adecuadamente y a lo largo de un periodo prolongado de tiempo, igualmente se pueden hacer análisis, pero éstos se vuelven menos confiables.

Por otra parte, respecto a las posibles mejoras del análisis, es importante y podría permitirnos hacer mejores interpretaciones de los resultados el tener conocimiento de qué significa a detalle cada variable analizada, al igual que por qué para algunas de ellas se tienen ciertos valores (o no se tienen ciertos valores), ya que esto repercute directamente en los resultados de las pruebas realizadas, pero no se puede establecer una razón aplicable a la situación analizada por la cual sucede, para brindar recomendaciones útiles para lo que se busca.

Además, el fijar una pregunta específica a responder (o un conjunto de ellas) tras realizar las diferentes pruebas, al igual que hipótesis concretas para rechazar o no rechazar, facilitaría la idea de lo que se busca y de si los resultados obtenidos se relacionan con el objetivo o no, incluso para qué, posteriormente, basándonos en estas conclusiones, consideremos la opción de solicitar que estos sean registrados o tomados de maneras distintas, o de que se incluya el registro de variables no consideradas que también pudieran estar involucradas y nos permitieran un mejor análisis, o una mejor comparación.

En conclusión, para poder mejorar este análisis es muy importante trabajar de la mano con las personas encargadas de la recolección de estos datos, o incluso estar en el lugar y tiempo en que se generan estos datos para entender cómo se relacionan entre sí, ya que aunque se puedan aplicar modelos muy complejos de machine learning, si no se tiene un buen entendimiento de los datos, nunca se podrán seleccionar las mejores variables que contribuyan de una mejor manera al modelo. Por lo tanto, no sólo es necesario tener los conocimientos técnicos de estadística y programación, sino también el conocimiento empírico de cómo funcionan los datos entre sí y así poder entender la relación entre los datos como un todo.