



Reto CEMEX Ventures

Aldo Sandoval ¹, Carlos Latorre ², Sebastian Neri ³

¹A01751137, ²A01379354, ³A01750190

Palabras clave: Regresión, Optimización, Metalurgia.

10 de Marzo de 2021

Resumen

Se busca poder minimizar el costo por tasa de producción con la combinación de la energía eléctrica y calorífica utilizada por un proceso de metalurgia, a partir de la calidad, dureza y tasa de producción de un material. Dadas las propiedades de los datos, se clusterizaron para optimizar su rendimiento en los modelos de regresión no lineal y se modelaron con las librerías de python Scipy y Scikitlearn para obtener la calidad máxima con el costo mínimo.

1. Introducción

Cemex cuenta con una fábrica que produce soportes metálicos con los que abastece un mercado. Para llevar a cabo el proceso de fabricación se utiliza maquinaria con Diesel y maquinaria con energía eléctrica. Se busca encontrar la producción en la que se utilice la menor cantidad de energía, tanto calórica como eléctrica, sin que se vea afectada la calidad. No se busca aumentar o disminuir la calidad, se busca que sabiendo la dureza del material crudo, la tasa de producción y la calidad dentro de un rango determinado, se encuentre las condiciones óptimas en la energía utilizada. Se sabe que una energía, la energía calórica es un 30% más económica que la energía eléctrica, sin embargo se necesita de ambos tipos de energía para llevar a cabo el proceso de producción, ya que se llevan a cabo procesos tanto de maquinaria antigua que funciona con Diesel, como también maquinaria moderna que requiere de energía eléctrica. La razón por la cual ambos tipos de máquinas son utilizadas, es por la eficiencia de dichas máquinas, siendo la maquinaria más cara (Energía eléctrica) también la más eficiente en la tasa de producción. Esto se debe a los materiales utilizados para este tipo de maquinaria.



Fig. 1. Ejemplos de configuraciones de maquinaria.

2. Exploración y entendimiento de los datos

2.1. Estadística descriptiva

Para comenzar con el análisis, se calcularon algunas medidas de tendencia central y dispersión de las variables numéricas.

	Dureza	Tasa_Prod	Asp	EC	EE	Calidad
count	9391.000000	9392.000000	9391.000000	9392.000000	9392.000000	9392.000000
mean	104.028644	391.005111	3.152306	19.362425	19.059135	0.089891
std	2.049060	43.352777	0.375251	6.698657	8.035162	0.048819
min	80.000000	0.000000	0.090000	0.000000	0.000000	0.000000
25%	103.000000	383.000000	3.040000	15.900000	14.200000	0.061000
50%	104.000000	398.000000	3.260000	19.200000	20.000000	0.081000
75%	105.000000	408.000000	3.380000	23.500000	25.200000	0.107000
max	112.000000	480.000000	3.520000	40.400000	35.300000	1.000000

Fig. 2. Tabla de descripciones estadísticas de los datos iniciales.

2.2. Valores nulos

Para el correcto análisis, se buscaron los valores nulos que contenía nuestra tabla, dado que representaba una porción grande de los datos, se decidió eliminar las filas que contenían dichos valores.

```
TIME      0
Dureza    1
Tasa_Prod 0
Asp       1
EC        0
EE        0
Calidad   0
dtype: int64
```

Fig. 3. Tabla de valores nulos por columna.

2.3. Correlación

Como se sabe que las variables son cuantitativas, se obtuvo un mapa de calor de la correlación entre variables mediante el coeficiente de correlación de Spearman.

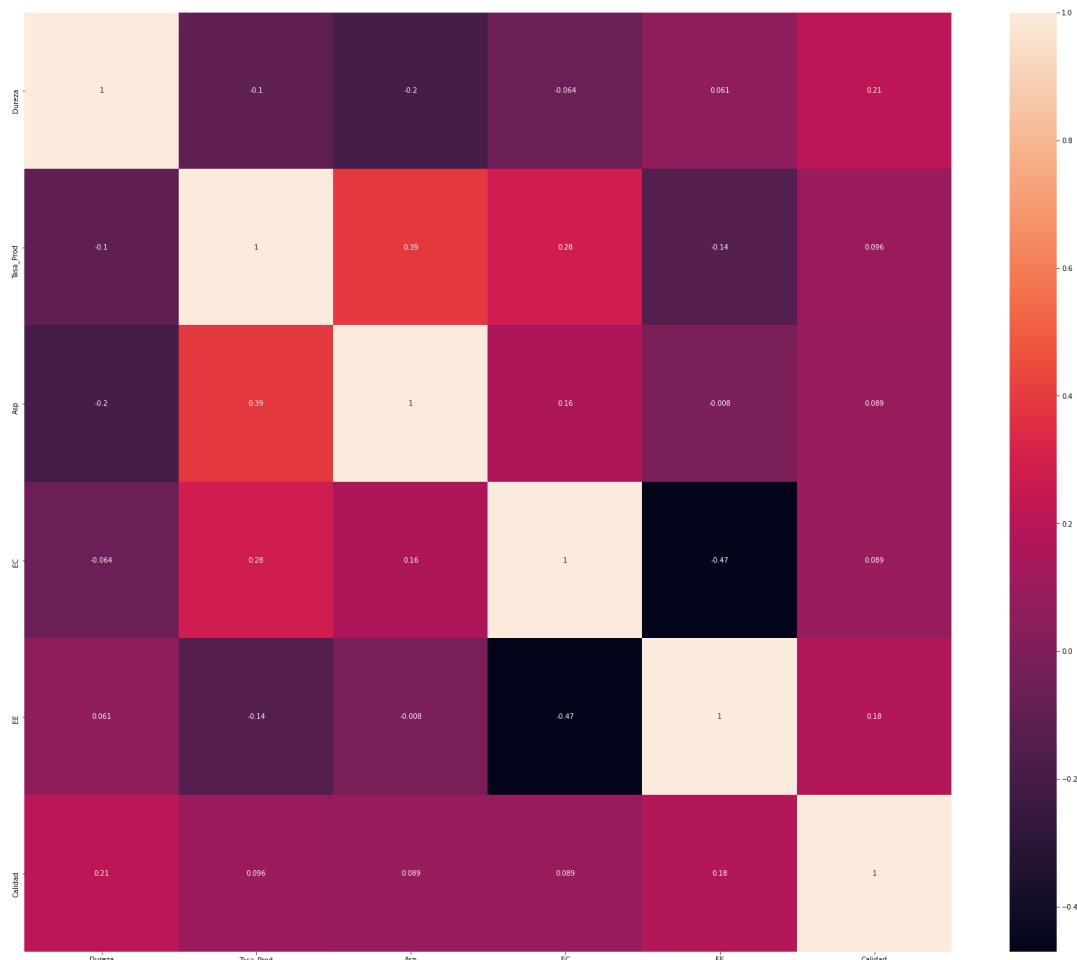


Fig. 4. Mapa de calor inicial que representa las correlaciones entre columnas.

2.4 Selección de Variables

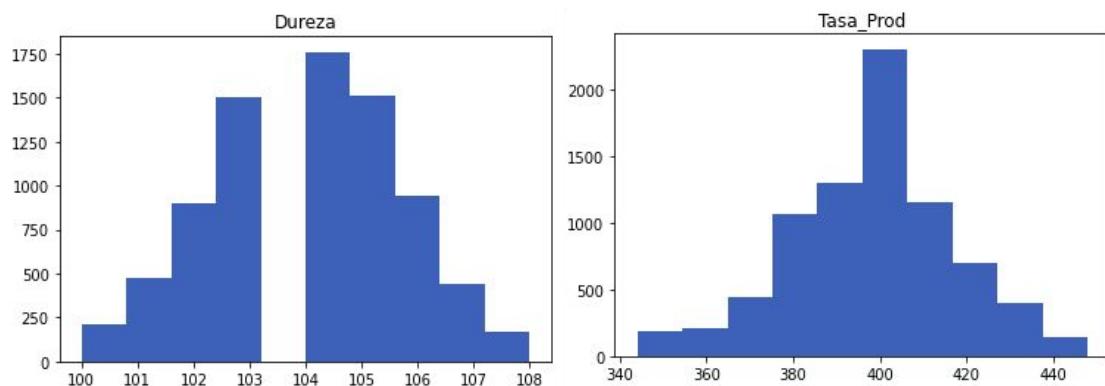
Dada la naturaleza de la regresión se eliminaron las columnas “TIME” y “Asp”, consideradas irrelevantes para nuestro análisis, dado que la columna “TIME” es una variable cualitativa, sinedo inservible para los propósitos de este análisis, a su vez, la columna “Asp” fue eliminada dado que presentaba una correlación fuerte con la variable de la tasa de producción, lo que podría empobrecer y sesgar el análisis. Una vez seleccionadas las variables más relevantes, se agregaron 4 columnas, “EE_TP”, “EC_TP”, “Costo” y “Costo_TP”, que son la energía por tasa de producción, la energía calórica por tasa de producción, el costo total de la energía total utilizada en el proceso de producción y el costo total por tasa de producción, respectivamente.

```
Data columns (total 9 columns):
 #   Column      Non-Null Count  Dtype  
 --- 
 0   Dureza       7913 non-null    float64
 1   Tasa_Prod   7913 non-null    int64  
 2   EC           7913 non-null    float64
 3   EE           7913 non-null    float64
 4   Calidad      7913 non-null    float64
 5   Costo        7913 non-null    float64
 6   Costo_TP    7913 non-null    float64
 7   EE_TP        7913 non-null    float64
 8   EC_TP        7913 non-null    float64
 dtypes: float64(8), int64(1)
 memory usage: 618.2 KB
```

Fig. 5. Tabla de columnas finales del dataframe.

2.5. Análisis de Distribución

Para conocer la naturaleza de los datos, se generaron los histogramas de cada una de las variables.



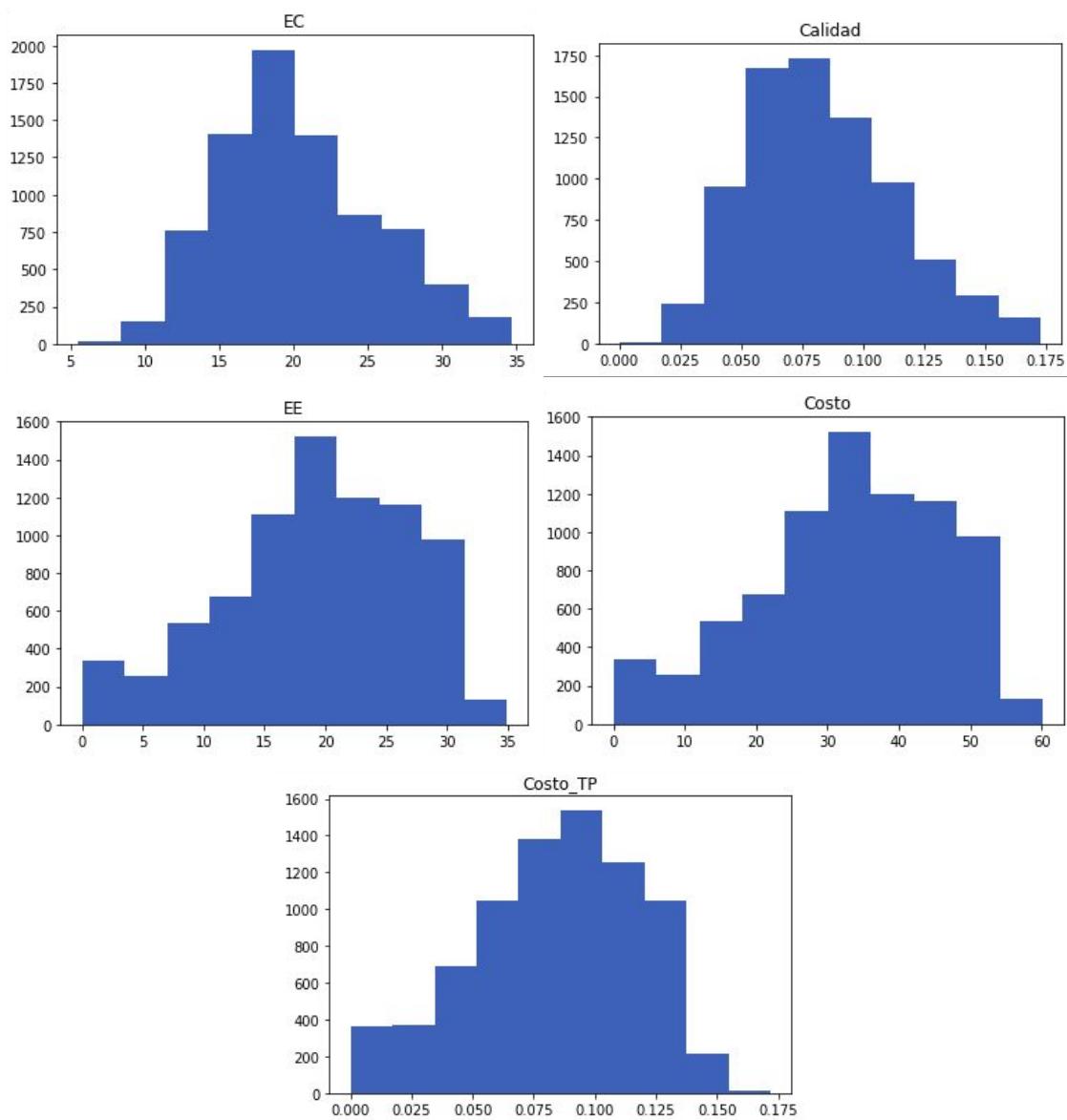


Fig. 6. Histogramas de cada columna.

El diagrama anterior denota que todas las variables son aleatorias, por lo tanto se considerará que su distribución es normal para el futuro uso de las propiedades estadísticas de estas distribuciones.

2.6. Outliers

Con las variables seleccionadas, se debe evaluar si la tabla contiene registros atípicos para eliminarlos y de esta manera evitar un error considerable en los modelos y resultados obtenidos. Para la valoración de los valores atípicos se utilizaron gráficas de caja para identificar dichos valores de una manera visual.

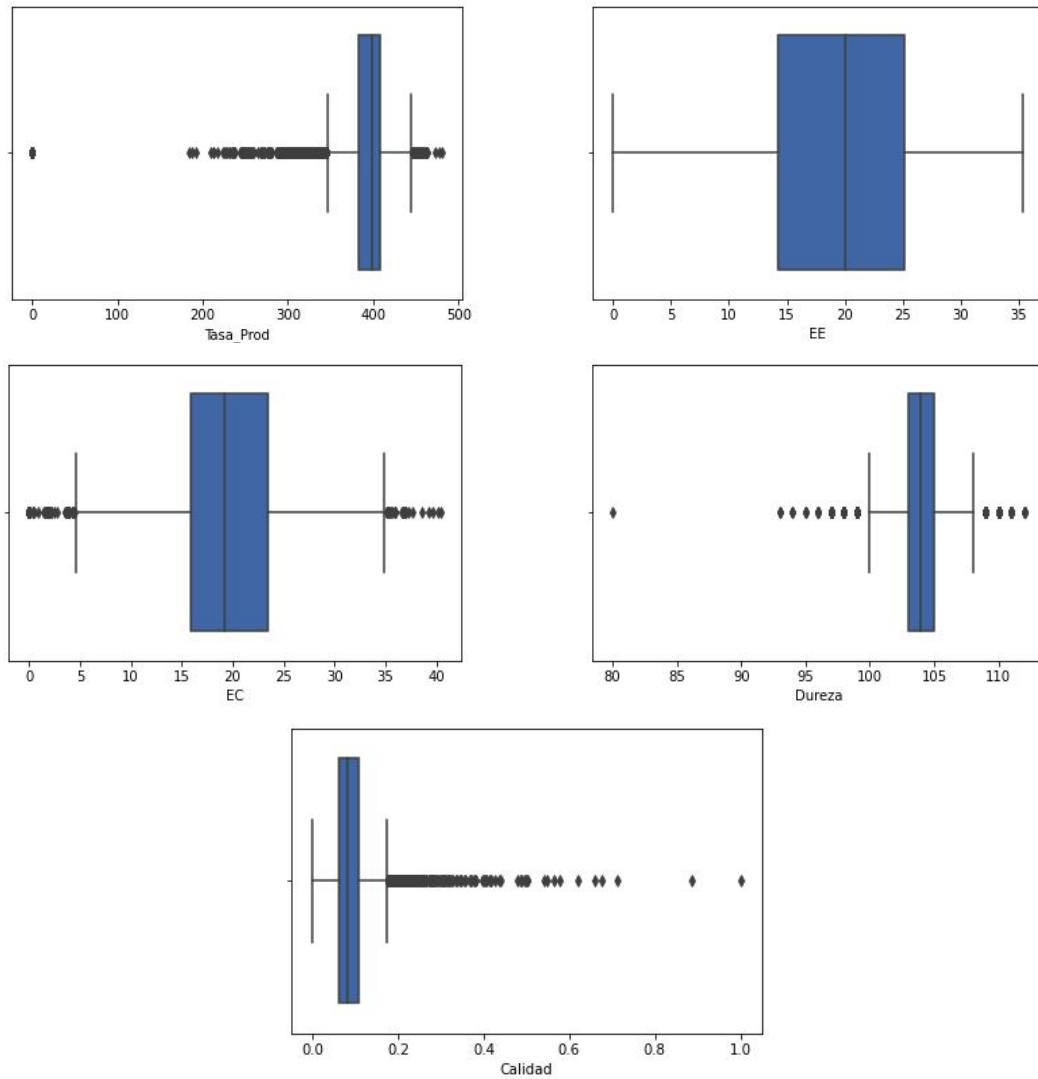


Fig. 7. Gráficos de caja de cada columna.

Una vez identificados los valores atípicos, se empleó el rango intercuartil como herramienta para eliminar estos datos atípicos, este proceso se repitió varias veces hasta ya no aparecieran outliers significativos en las gráficas de caja, esto debido a que cada vez que se eliminaban los outliers, aparecían más debido al desplazamiento de las medidas de posición. Finalmente, se pudo notar una mejora en la correlación de nuestros datos como lo indica la figura 8.

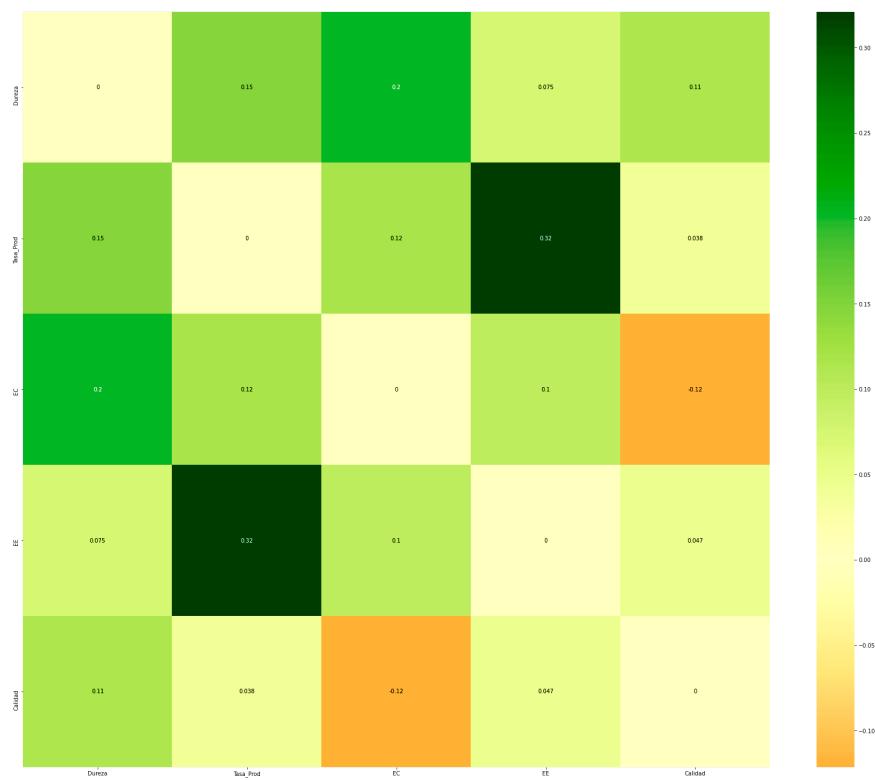


Fig. 8. Mapa de calor del cambio porcentual en la correlación después de eliminar valores atípicos.

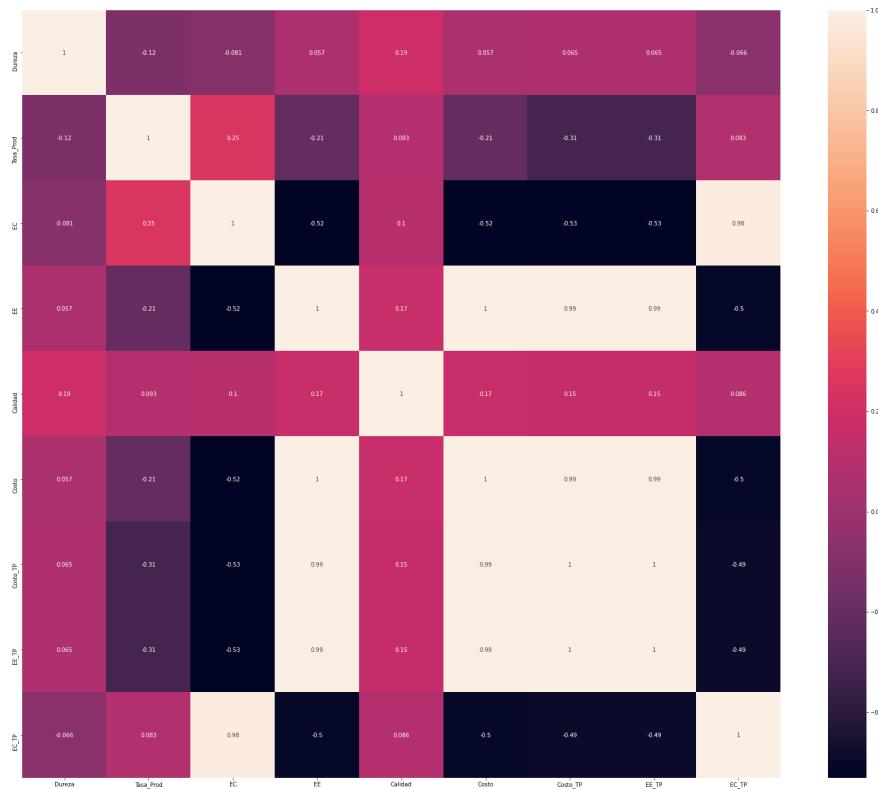


Fig. 9. Mapa de calor de cambio en la correlación después de eliminar valores atípicos.

2.7. Análisis de dispersión

Los diagramas de dispersión que se encuentran en la figura 10 muestran un comportamiento singular por el tipo de problema que se aborda, dado que no demuestra ningún tipo de relación lineal, sino que muestran un comportamiento aglomerado, por lo que un modelo de regresión lineal queda descartado directamente, por otro lado, debido al tamaño de la tabla es probable que muchas características estén llenas de ruido, que serían muy útiles para el modelo.

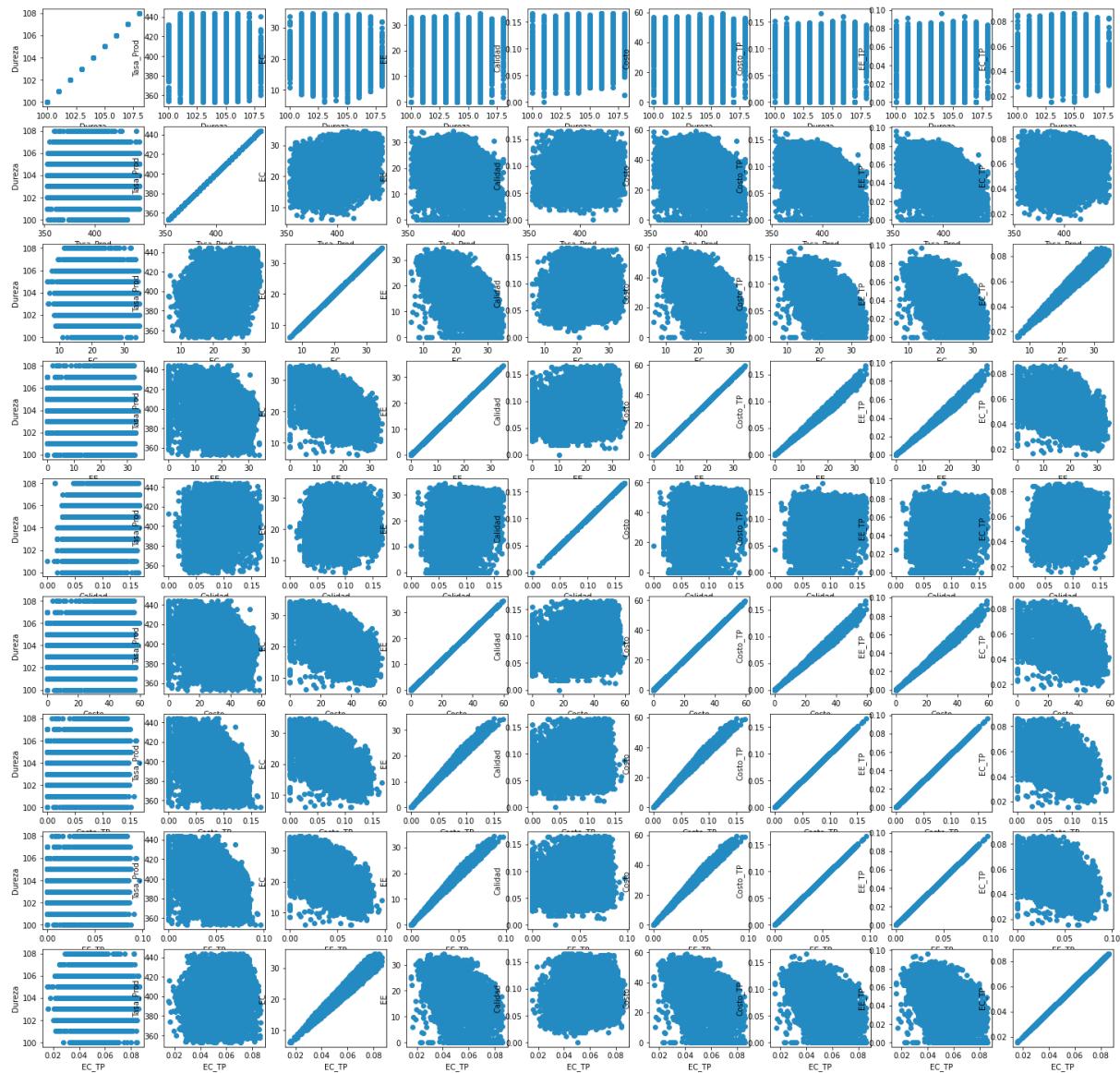
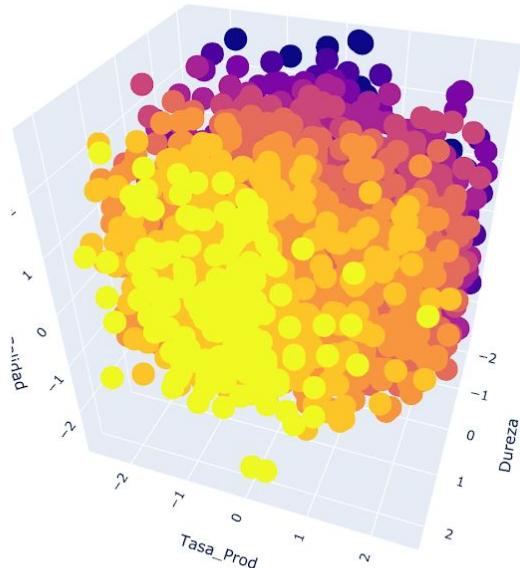


Fig. 10. Gráficas de dispersión de todas las columnas.

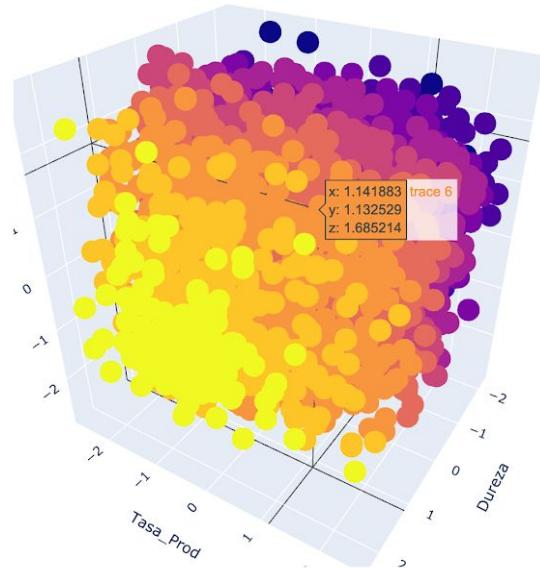
2.7.1 Análisis de Dispersión 3D

También se realizó una visualización en 3D de todas las variables para buscar alguna relación que pueda ser evidente en tercera dimensión.

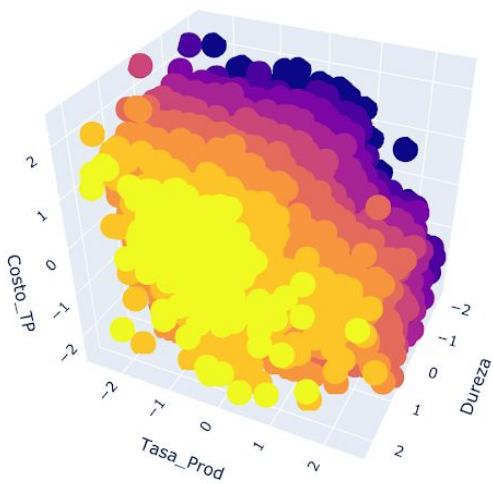
Dureza vs Tasa_Prod vs Calidad



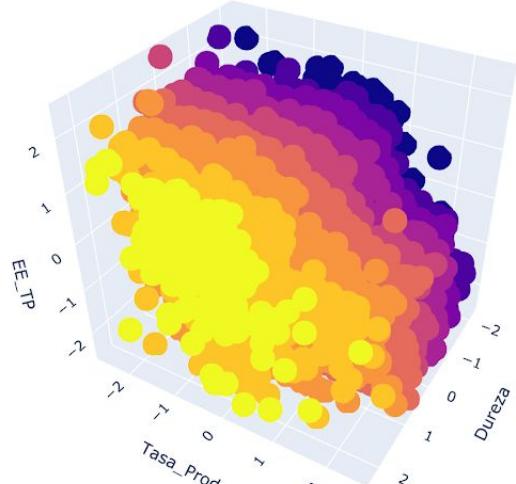
Dureza vs Tasa_Prod vs EC_TP



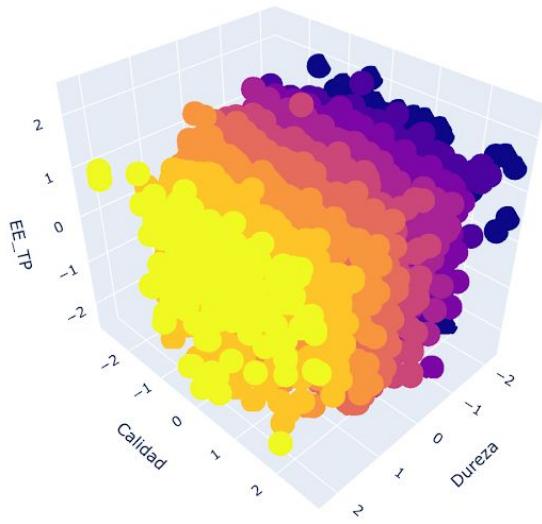
Dureza vs Tasa_Prod vs Costo_TP



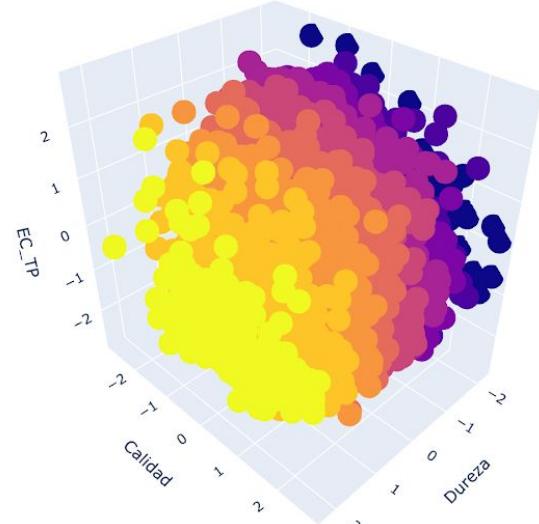
Dureza vs Tasa_Prod vs EE_TP



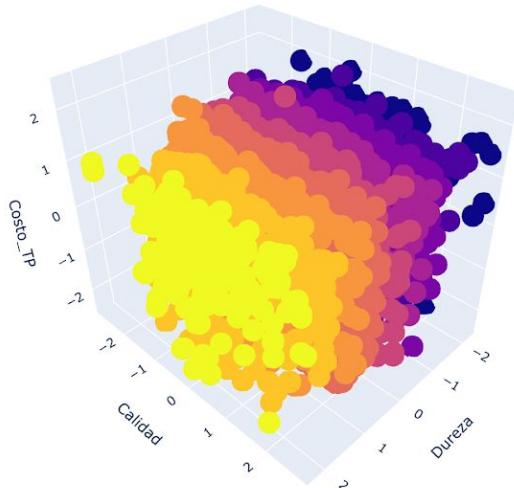
Dureza vs Calidad vs EE_TP



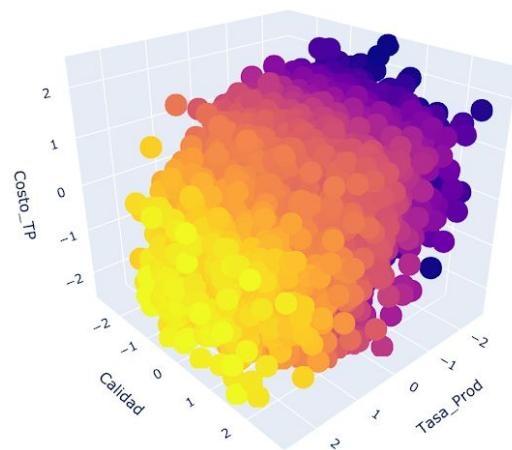
Dureza vs Calidad vs EC_TP



Dureza vs Calidad vs Costo_TP



Tasa_Prod vs Calidad vs Costo_TP



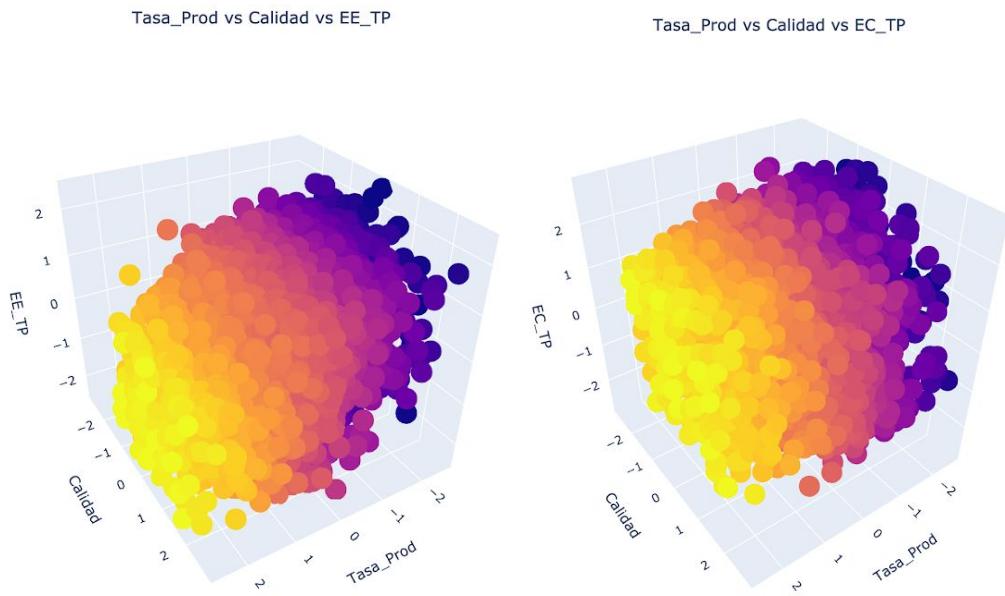


Fig. 11. Gráficas de dispersión en 3D de todas las variables.

3. Preparación de los datos

3.1. Estandarización y separación de los datos

Para el manejo de los datos, es necesarios escalarlos, para su correcto análisis en la visualización de datos, además es necesario conservar la relación que existe entre las variables de la tabla, es por eso que se estandarizaron, siguiendo la siguiente fórmula.

$$f'_{:,i} = \frac{f_{:,i} - \text{mean}(f_{:,i})}{\text{std}(f_{:,i})} \quad (1)$$

La cual nos permitirá conservar dicha información entre las variables, a diferencia de otros escaladores.

	Dureza	Tasa_Prod	EC	EE	Calidad	ET	Costo	Costo_TP
count	7.678000e+03							
mean	-1.717766e-15	-2.993900e-16	1.117185e-15	2.518688e-15	-9.728550e-17	5.847107e-16	2.032743e-15	4.404092e-17
std	1.000065e+00							
min	-2.246123e+00	-2.599661e+00	-2.828919e+00	-2.504670e+00	-2.756374e+00	-4.678898e+00	-2.504670e+00	-2.451745e+00
25%	-5.520431e-01	-6.766207e-01	-7.072298e-01	-5.930152e-01	-7.278231e-01	-6.116133e-01	-5.930152e-01	-6.322123e-01
50%	1.265007e-02	2.099454e-03	-1.338002e-01	1.092252e-01	-9.597933e-02	1.114596e-01	1.092252e-01	8.665602e-02
75%	5.773432e-01	6.242596e-01	6.498868e-01	7.594478e-01	6.688841e-01	7.290844e-01	7.594478e-01	7.682792e-01
max	2.271423e+00	2.547300e+00	2.752462e+00	1.981866e+00	2.763945e+00	2.612087e+00	1.981866e+00	2.427830e+00

Fig. 12. Tabla de descripción estadística después de la estandarización.

3.2. Clusterización

Dada la naturaleza de los datos, se infirió que no existía una linealidad en ellos, por lo que trabajar los datos de manera general sería un conflicto para el modelado ya que estaría alimentado de un sesgo que se puede minimizar. En este caso se planteó la siguiente tesis:

“Un conjunto de D con n elementos de características homogéneas, puede ser segmentado un en un conjunto de clusters K , donde cada elemento i_j del cluster K_i , tendrá características heterogéneas de dicho cluster, por lo que los clusters tendrán características lineales mejores de manera individual que de manera holística”

Para obtener el número de clusters óptimo para la base de datos se utilizó un dendrograma figura 13 y el criterio del número de líneas por las que pasa el trazo de una línea perpendicular a la rama con la distancia más larga, midiendo de manera secuencial las distancias entre líneas horizontales. A pesar de que la distancia más larga se encuentra arriba de donde se representa en la figura 13, el número de clusters se consideró muy pequeño para englobar de mejor manera a los clusters, por eso se escogió la segunda más larga. Esta observación nos permitió deducir que el número óptimo de clusters son cinco, lo que se puede apreciar en la figura 14.

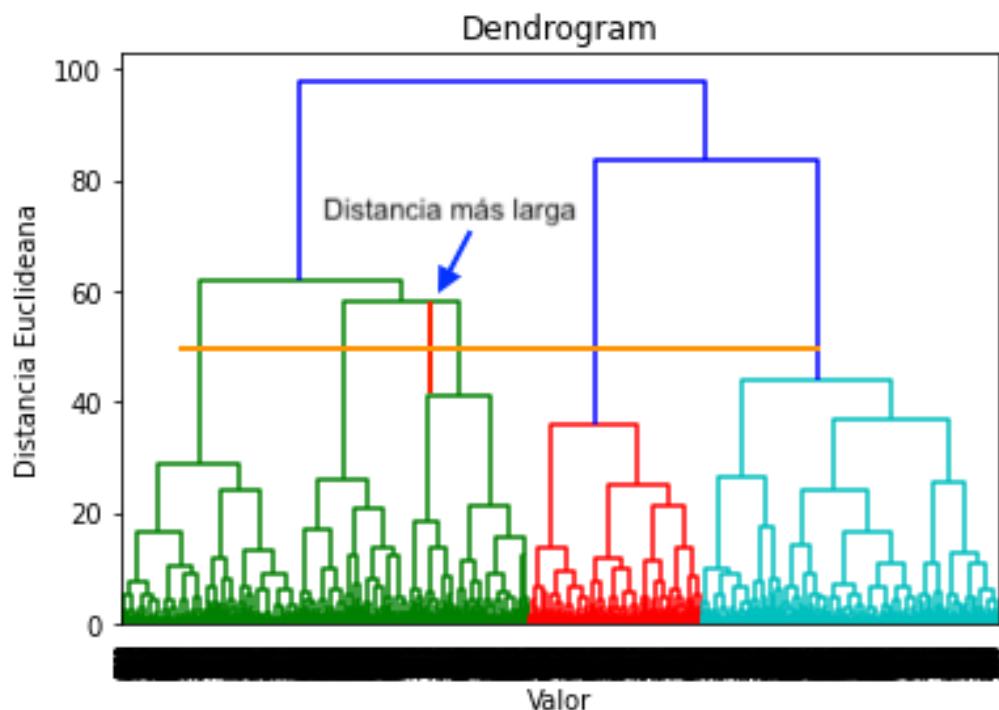


Fig. 13 Dendrograma de las variables “Calidad”, “Tasa_Prod”, “Dureza”

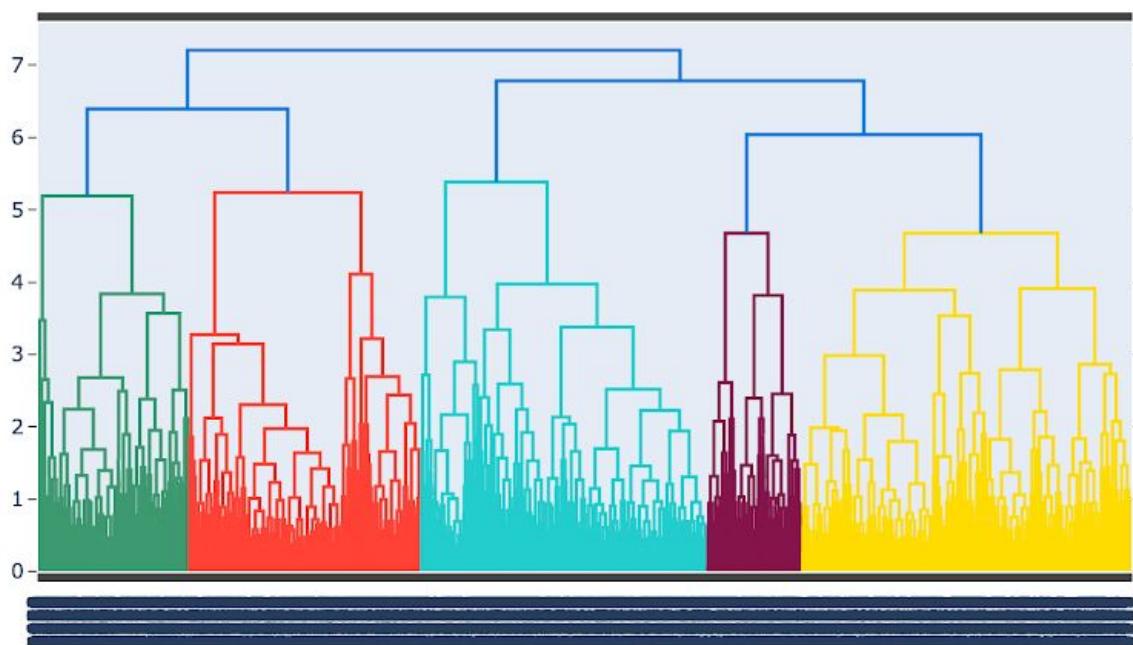


Fig. 14. Dendrograma de las variables “Calidad”, “Tasa_Prod”, “Dureza” agrupados en 5 clusters

3.3 Evaluación de los Clusters

Para la confirmación o denegación de nuestra tesis, se computó el cambio porcentual de la correlación de spearman de los cluster con las variables de interés respecto de la tabla original, las cuales son “EE”, “EC”, “Costo_TP”. Se obtuvieron los siguientes resultados.

Cluster 1

	Dureza	Calidad	Tasa de Producción
Energía Calórica	+83%	-120%	+140%
Energía Eléctrica	+6.4%	+49%	-180%
Costo por Tasa de Producción	+6.4%	+49%	-180%

Tabla. 1. Tabla de cambios porcentuales en el cluster 1.

Cluster 2

	Dureza	Calidad	Tasa de Producción
Energía Calórica	+57%	+57%	+79%
Energía Eléctrica	-7.8%	-40%	+41%
Costo por Tasa de Producción	-7.8%	-40%	+41%

Tabla. 2. Tabla de cambios porcentuales en el cluster 2.

Cluster 3

	Dureza	Calidad	Tasa de Producción
Energía Calórica	-140%	-57%	+36%
Energía Eléctrica	+53%	-39%	+50%
Costo por Tasa de Producción	+53%	-39%	+50%

Tabla. 3. Tabla de cambios porcentuales en el cluster 3.

Cluster 4

	Dureza	Calidad	Tasa de Producción
Energía Calórica	-200%	+24%	+9.8%
Energía Eléctrica	+120%	+0.2%	-1.8%
Costo por Tasa de Producción	+120%	+0.21%	-1.8%

Tabla. 4. Tabla de cambios porcentuales en el cluster 4.

Cluster 5

	Dureza	Calidad	Tasa de Producción
Energía Calórica	-200%	-69%	+16%
Energía Eléctrica	+44%	-4%	+82%
Costo por Tasa de Producción	+44%	+4%	+82%

Tabla. 5. Tabla de cambios porcentuales en el cluster 5.

Para el primer clúster, hubo un leve aumento de la dureza para las 3 variables de interés, siendo 83%, 6.4% y 6.4% para la energía calórica, eléctrica y costo por tasa de producción respectivamente. Para la calidad su correlación aumentó de manera proporcional para las mismas variables exceptuando la energía calórica, con un aumento de 49% y una disminución de -120% y la tasa de producción tuvo un aumento del 140% para la energía calórica y una disminución para 180% en la energía eléctrica y el costo por tasa de producción.

En el segundo cluster, hubo una disminución en la dureza para la energía eléctrica y el costo por costo de tasa de producción de 7.8%, sin embargo para la energía calórica hubo un aumento de 57%. De la misma forma en la calidad, la energía eléctrica y el costo por tasa de producción tuvieron una disminución de 40%, mientras que la energía calórica aumentó en 57%. En el caso de tasa de producción hubo un aumento del 41% en la energía calórica y costo por tasa de producción y aumentó 79% en la energía eléctrica.

En el caso del cluster número 3, la dureza disminuyó 140% en energía eléctrica, la energía calórica y el costo por tasa de producción aumentaron 53%. La calidad disminuyó de manera proporcional un promedio de 45%. Para la tasa de producción ocurrió lo opuesto, ya que aumentó casi proporcionalmente un 45.3%

En la dureza del cuarto cluster, hubo una disminución significativa de energía eléctrica de 200% y un gran aumento de 120% en la energía calórica y el costo por tasa. En el caso de la calidad hubo un ligero aumento en promedio de 24% y una disminución 1.8% en la energía calórica y visitó por tasa de producción en la tasa de producción y un aumento de 9.8%

Para el quinto y último cluster, la dureza tuvo una gran disminución de 200% en la energía eléctrica y un aumento del 44% para la energía calórica y el costo por tasa de producción, mientras que la calidad disminuyó 69% y 4% en ambas energías, con un ligero aumento de 4% en el costo por tasa de producción. La tasa de producción tuvo un aumento de 16% para la energía eléctrica y un 82% en la energía calórica y el costo por tasa de producción.

Dados los resultados anteriores, se puede concluir que el clusterizado tuvo un impacto positivo en las correlaciones, dado que aumentan en mayor medida que lo que disminuyen, por lo tanto se confirma la tesis planteada.

4. Optimización de los datos

El objetivo del análisis es obtener el costo mínimo obteniendo la máxima calidad posible, por lo tanto se requiere de un proceso de optimización de los datos, en este caso, maximizar la calidad y minimizar el costo. Para obtener los registros más óptimos se utilizará una métrica estadística para dicha optimización. Como ya se estableció, las variables que se han trabajado siguen una distribución normal, por lo tanto, dado esto, se obtendrá el 35% de las calidades más altas de todo el dataset, considerandolas como las calidades más óptimas para el análisis, dicha relación se planteó de la siguiente manera.

$$P(X \geq z) = 0.65 \quad (2)$$

$$P(X \geq 0.38532046640756773) = 0.65 \quad (3)$$

$$0.38532046640756773 * \sigma + \mu = Q_{max} \quad (4)$$

Una vez obtenido el valor de Q_{max} , se dividió la tabla para obtener las calidades mayores a dicho valor. Esto tuvo como resultado la creación de una nueva tabla, por lo que la distribución de las demás variables de la tabla será diferente, ahora, se obtuvieron los valores mínimos del costo dada esta máxima calidad, los cuales fueron representados como el 35% de los valores más pequeños de las calidades máximas.

$$P(X < z) = 0.35 \quad (5)$$

$$P(X < -0.38532046640756773) = 0.35 \quad (5)$$

$$0.38532046640756773 * \sigma + \mu = C_{min} \quad (6)$$

Siendo el valor de C_{min} el valor máximo de los costos mínimos de la distribución. Una vez obtenidos estos valores, se fraccionó la tabla con los valores mínimos posibles, construyendo así el intervalo de los valores máximos y mínimos de la calidad tanto como del Costo ponderado.

$$QI_{max} = [0.0373012491249344, 2.7629481314369153] \quad (7)$$

$$CI_{min} = [-2.460056324496983, 0.1277750310487141] \quad (8)$$

Los intervalos de las expresiones (7 y 8) fueron utilizados para partir las tablas de los clusters, esto debido a que tienen aproximadamente los mismos valores posibles del costo, puesto que vienen de una distribución normal, lo que preserva sus características de esta variable además de que el clustering no fue hecho a base de estos costos también.

Una vez separados estos datos, se calculó un intervalo de confianza al 95% de la media muestral, dicho intervalo fue utilizado para extraer los registros dentro de ese intervalo de la calidad. Una vez separados, se calcularon de nuevo las medidas de dispersión poblacional para esta nueva tabla.

$$\sigma_{95} = 0.010307674395704056 \quad (9)$$

$$\mu_{95} = -0.10148808359436017 \quad (10)$$

Siendo que se tiene una tabla y se quiere optimizar los datos de esa tabla, se aplicará el mismo criterio que anteriormente se había mencionado. Los intervalos de la nueva tabla fueron.

$$QI_{95_{max}} = [-0.09565713537808918, 1.931958228293018] \quad (11)$$

$$CI_{95_{min}} = [-2.460056324496983, 0.09089830998447562] \quad (12)$$

Dichos intervalos serán de ayuda al modelado para obtener un costo mínimo óptimo y un costo mínimo promedio con un fundamento estadístico fuerte.

5. Modelación

Dado el propósito del análisis, para obtener las métricas que se requieren como lo son el conjunto de energías y el costo por tasa de producción, se necesita plantear un modelo de regresión, en este caso al ser 3 energías las que se requieren calcular, se plantea el siguiente modelo de regresión.

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 = Y_1, Y_2, Y_3$$

Donde:

- X_1 : Tasa de producción
- X_2 : Dureza
- X_3 : Calidad del producto
- Y_1 : Energía eléctrica por tasa de producción
- Y_2 : Energía calórica por tasa de producción
- Y_3 : Costo por tasa de producción

Para este problema se consideraron 4 modelos esenciales dadas las características de los datos, pues teniendo que lidiar con no-linealidades los regresores que mejor desempeño tendrían dadas estas condiciones son: árboles de decisión, máquinas de vector de soporte, bosques aleatorios y red neuronal. Para optimizar los parámetros de los regresores se implementó una simulación para crear n número de modelos por cluster, arrojando un intervalo de confianza al 99% de los errores medios y los parámetros más óptimos de los modelos.

	Costo_TP	EC_TP	EE_TP
SVMR	0.4962125135930522, 0.5333855698546388	0.925209932346431, 1.0226742485116396	0.4962125135907768, 0.5333855698532585
Random Forest	0.567667485618338, 0.6187171120548963	1.0873585664660408, 1.2188150947652794	0.5676674856183385, 0.6187171120548968
Decision Tree	0.9916807496809212, 1.1097408057253164	1.8191768308408913, 2.0378461255629805	0.9916807496809215, 1.1097408057253166
Artificial Neural Network	0.624289425965621, 0.6804192928506433	1.538346188747686, 1.776374598193067	0.6242894259656213, 0.6804192928506436

Tabla 6. Resultados de la tabla general

	Costo_TP	EC_TP	EE_TP
SVMR	0.4730949745428932, 0.6529189746067281	0.8374681091040123, 0.9269201168691027	0.47309497454290295, 0.6529189746067385
Random Forest	0.5957400765524122, 0.7812476913851659	0.966867355459585, 1.1133521696916033	0.5957400765524125, 0.7812476913851661
Decision Tree	0.9957437648368312, 1.2092979091066822	1.3748490413779653, 1.8250960138621144	0.9957437648368314, 1.2092979091066822
Artificial Neural Network	0.6097309728705136, 0.847504418838726	1.5683900393680235, 1.862370556161606	0.609730972870514, 0.8475044188387264

Tabla 7. Errores cuadráticos medios del Cluster 1 por modelo

	Costo_TP	EC_TP	EE_TP
SVMR	0.31920399778231756, 0.7211151418018067	0.559199768207611, 1.0853160888555462	0.31920399778231767, 0.7211151418018072
Random Forest	0.4206507419723271, 0.7996094947102046	0.7489782275879502, 1.3238845966580424	0.4206507419723272, 0.7996094947102049
Decision Tree	0.6110692092709522, 1.0046558018873972	0.7541899356794393, 2.0163903993862533	0.6110692092709528, 1.004655801887398
Artificial Neural Network	0.5258484665674215, 0.9245658448863887	1.0703529695829594, 1.703094573957109	0.5258484665674217, 0.9245658448863892

Tabla 8. Errores cuadráticos medios del Cluster 2 por modelo

	Costo_TP	EC_TP	EE_TP
SVMR	0.15514248420938423, 0.239227613777845	0.7966697319425813, 1.0419342847742143	0.155142484209383, 0.23922761377784554
Random Forest	0.19419883313617026, 0.27361237096992364	0.9184414882070701, 1.1335636393059803	0.1941988331361704, 0.27361237096992375
Decision Tree	0.3499750767740897, 0.48124052462939915	1.470330897719665, 1.981164686177309	0.3499750767740899, 0.48124052462939937
Artificial Neural Network	0.20840952937845766, 0.30613250520705093	0.9261559399823435, 1.186665364281914	0.20840952937845775, 0.3061325052070511

Tabla 9. Errores cuadráticos medios del Cluster 3 por modelo

	Costo_TP	EC_TP	EE_TP
SVMR	0.4642093912320096, 0.8906427164227269	1.034979560031072, 1.244593421590519	0.46420939123201055, 0.8906427164227321
Random Forest	0.5662261841665852, 0.8938051803465281	1.1643400582984433, 1.4810299555580075	0.5662261841665852, 0.8938051803465281
Decision Tree	0.92746519774663, 1.2885987408550763	1.807903643999936, 2.6719211681041686	0.92746519774663, 1.2885987408550765
Artificial Neural Network	0.539120717715317, 1.0842095555301419	1.3987751732929745, 1.8086918957724476	0.5391207177153174, 1.0842095555301423

Tabla 10. Errores cuadráticos medios del Cluster 4 por modelo

	Costo_TP	EC_TP	EE_TP
SVMR	0.39592914402023127, 0.5009720377448628	0.7672654060778321, 0.9185749584578229	0.3959291440202342, 0.5009720377448634
Random Forest	0.4876139020472989, 0.5635176111370241	0.9485883168508641, 1.140638297549971	0.487613902047299, 0.5635176111370243
Decision Tree	0.807421673345504, 0.99368297205364	1.4544596743069376, 1.8296396939884099	0.8074216733455045, 0.9936829720536405
Artificial Neural Network	0.5299690706083979, 0.6674038065187282	1.2863845181281885, 1.630755397421667	0.5299690706083984, 0.6674038065187284

Tabla 11. Errores cuadráticos medios del Cluster 5 por modelo

5.1 Análisis de Modelación

Como se puede ver en las tablas, para todos los clusters, los modelos con el intervalo del menor error cuadrático medio fueron las máquinas de vector de soporte. Finalmente, se escogieron 2 modelos de regresión dado que las máquinas de vector de soporte interpolan a cada variable por sí misma, más no a las 3 que se quieren predecir, por lo que se decidió incluir un segundo modelo que por su naturaleza pueda predecir las 3 salidas entrenando al mismo tiempo, ergo, se considerará también los árboles aleatorios, cuyo intervalo del error cuadrático medio es el segundo menor de todos los demás.

Dados los argumentos anteriores, se define a los modelos de máquinas de vector de soporte y los árboles aleatorios como los mejores modelos para este análisis.

6. Evaluación

Para la evaluación de los modelos, lo que se busca es comprender cómo ambos algoritmos modelan a las 3 variables planteadas, es por eso que se analizarán de manera individual los resultados de cada modelo por variable.

Modelo General

Tabla general

EE_TP

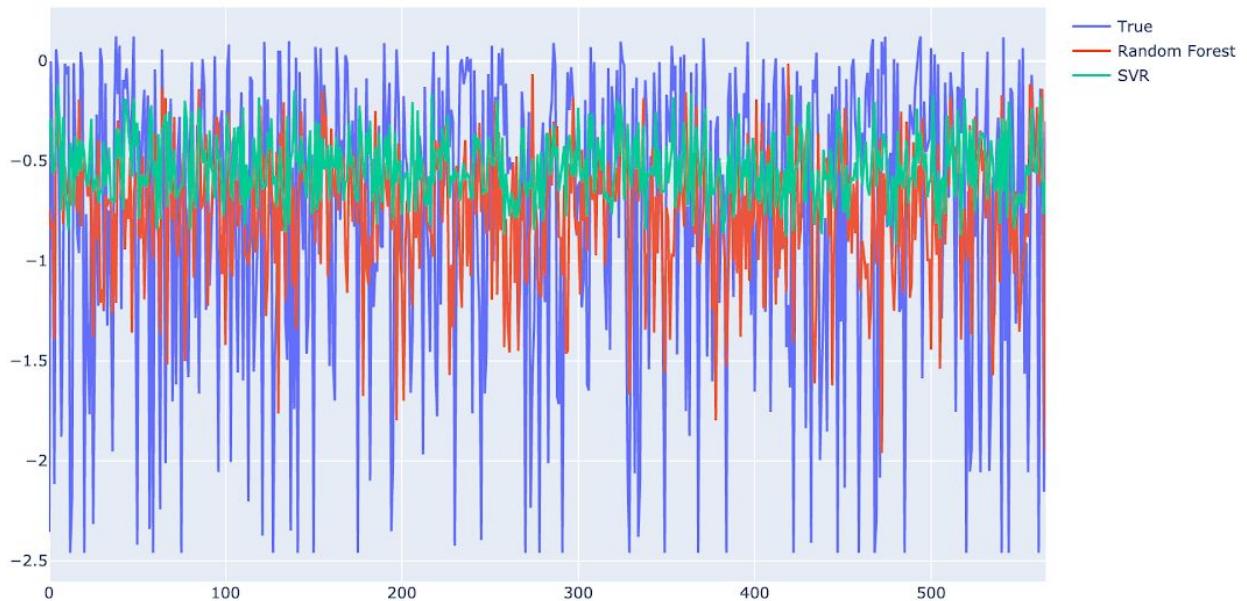


Fig. 15. Comparación de las predicciones de los modelos SVMR y Random Forest contra los valores reales de la energía eléctrica por tasa de producción de la tabla general.

EC_TP

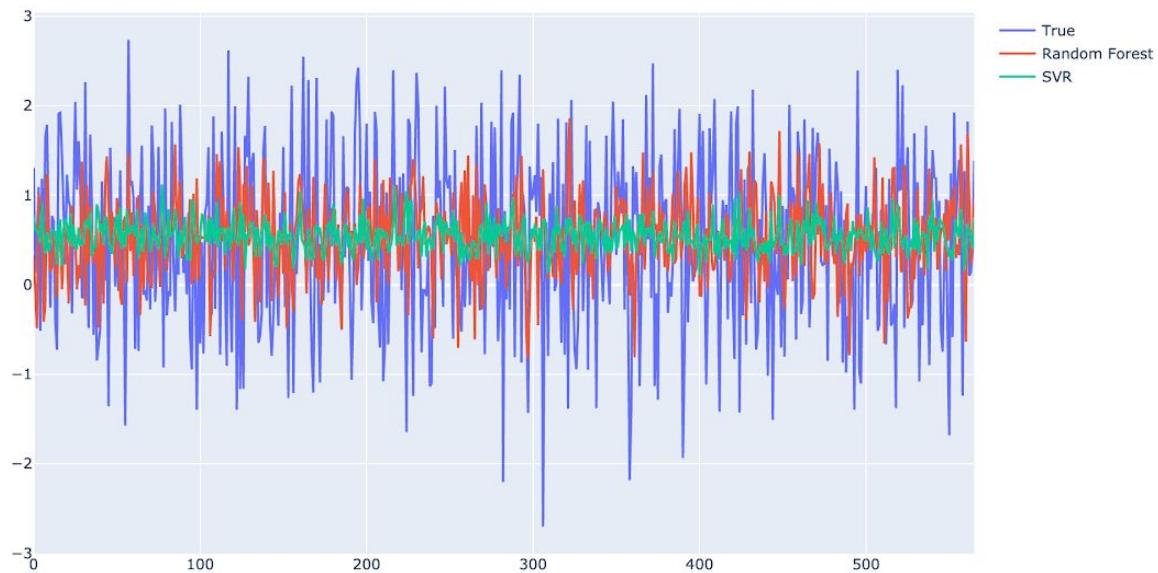


Fig. 16. Comparación de las predicciones de los modelos SVMR y Random Forest contra los valores reales de la energía calórica por tasa de producción de la tabla general.

Costo_TP

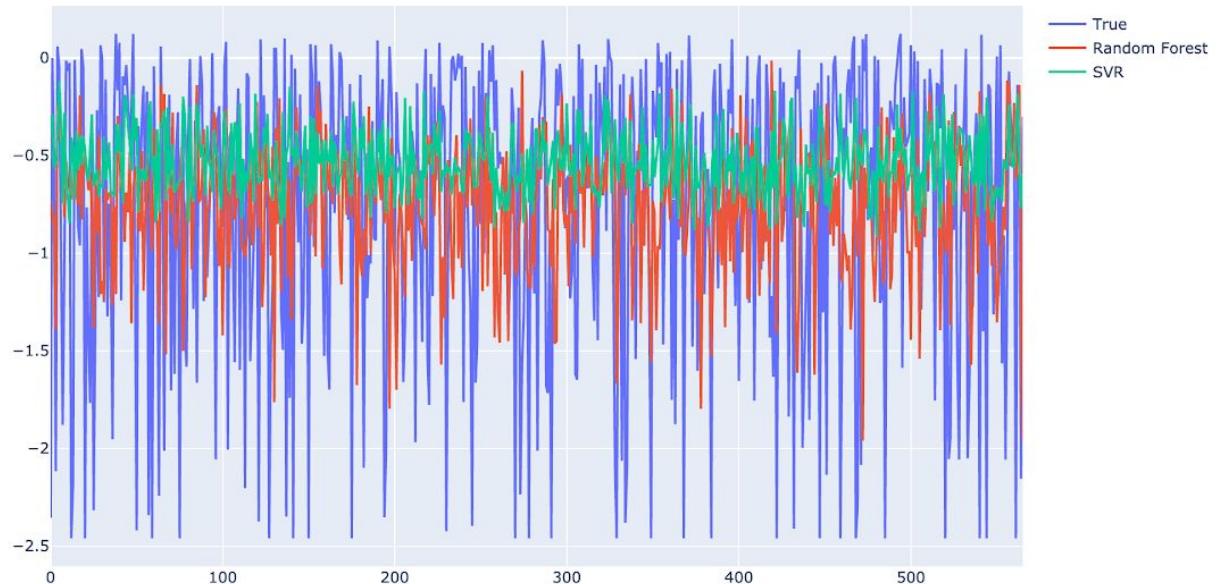


Fig. 17. Comparación de las predicciones de los modelos SVMR y Random Forest contra los valores reales de el costo por tasa de producción de la tabla general.

Cluster 1

EE_TP



Fig. 18. Comparación de las predicciones de los modelos SVMR y Random Forest contra los valores reales de la energía eléctrica por tasa de producción del cluster 1.

EC_TP



Fig. 19. Comparación de las predicciones de los modelos SVMR y Random Forest contra los valores reales de la energía calórica por tasa de producción del cluster 1.

Costo_TP

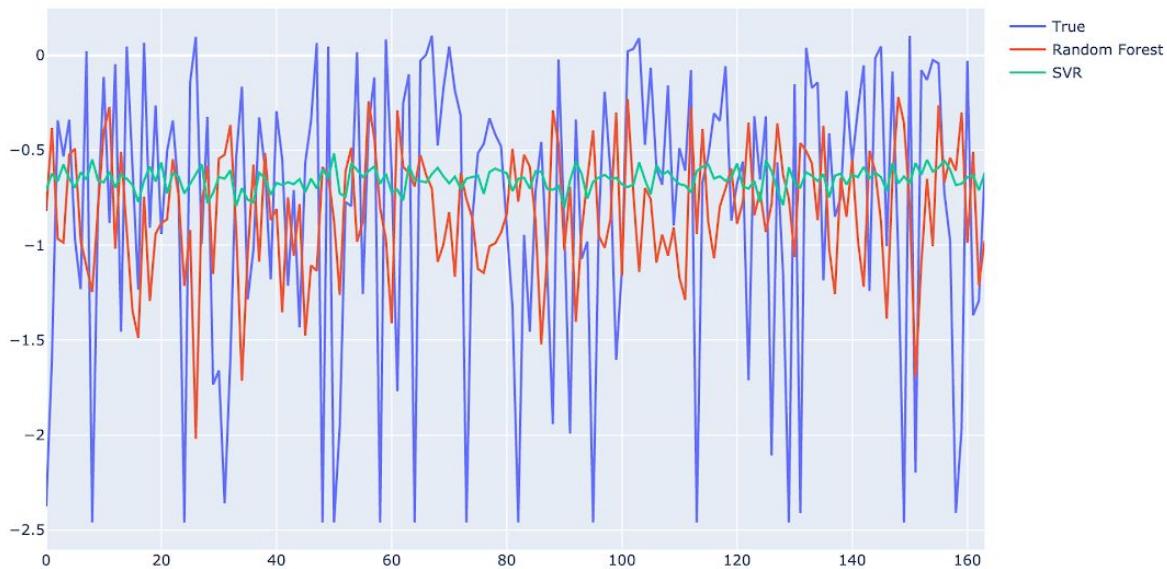


Fig. 20. Comparación de las predicciones de los modelos SVMR y Random Forest contra los valores reales del costo por tasa de producción del cluster 1.

Cluster 2

EE_TP



Fig. 21. Comparación de las predicciones de los modelos SVMR y Random Forest contra los valores reales de la energía eléctrica por tasa de producción del cluster 2.

EC_TP

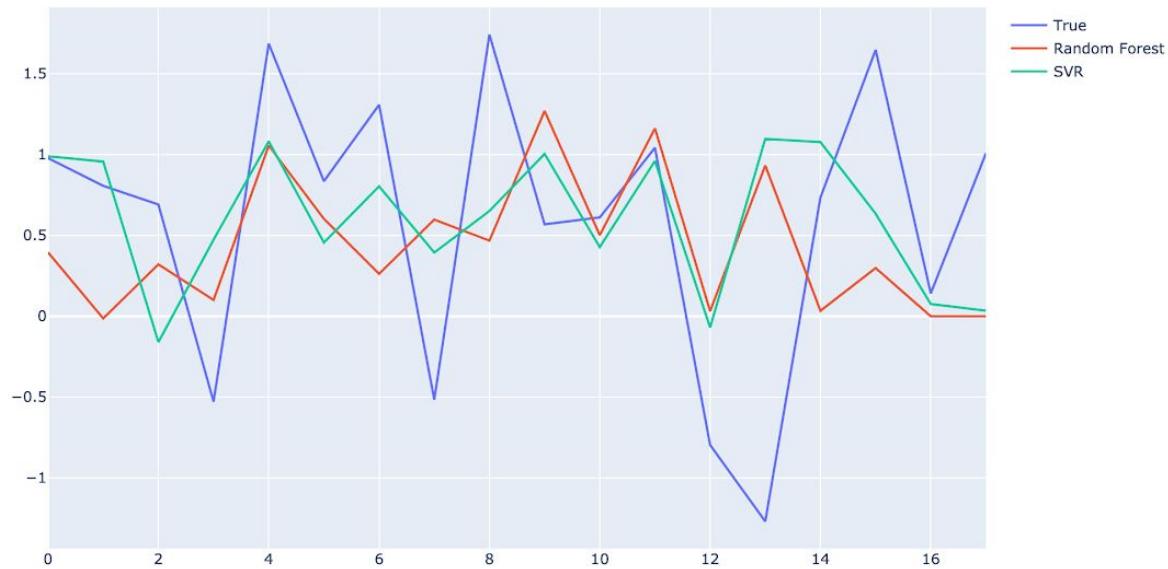


Fig. 22. Comparación de las predicciones de los modelos SVMR y Random Forest contra los valores reales de la energía calórica por tasa de producción del cluster 2.

Costo_TP



Fig. 23. Comparación de las predicciones de los modelos SVMR y Random Forest contra los valores reales del costo por tasa de producción del cluster 2.

Cluster 3



Fig. 24. Comparación de las predicciones de los modelos SVMR y Random Forest contra los valores reales de la energía eléctrica por tasa de producción del cluster 3.

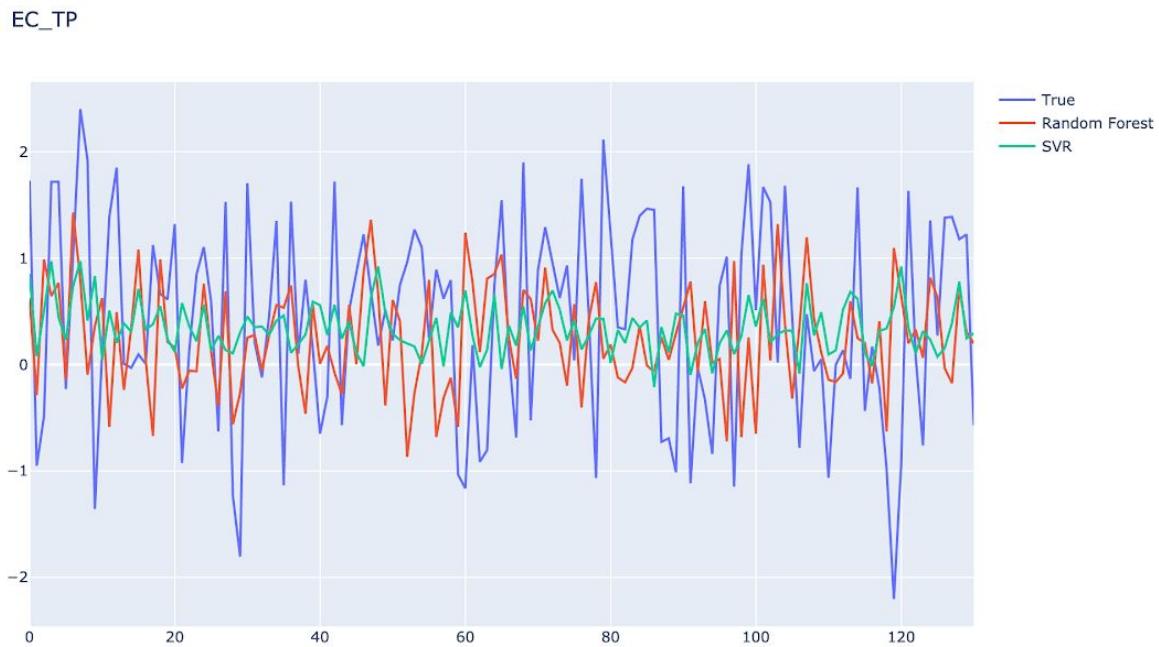


Fig. 25. Comparación de las predicciones de los modelos SVMR y Random Forest contra los valores reales de la energía calórica por tasa de producción del cluster 3.

Costo_TP

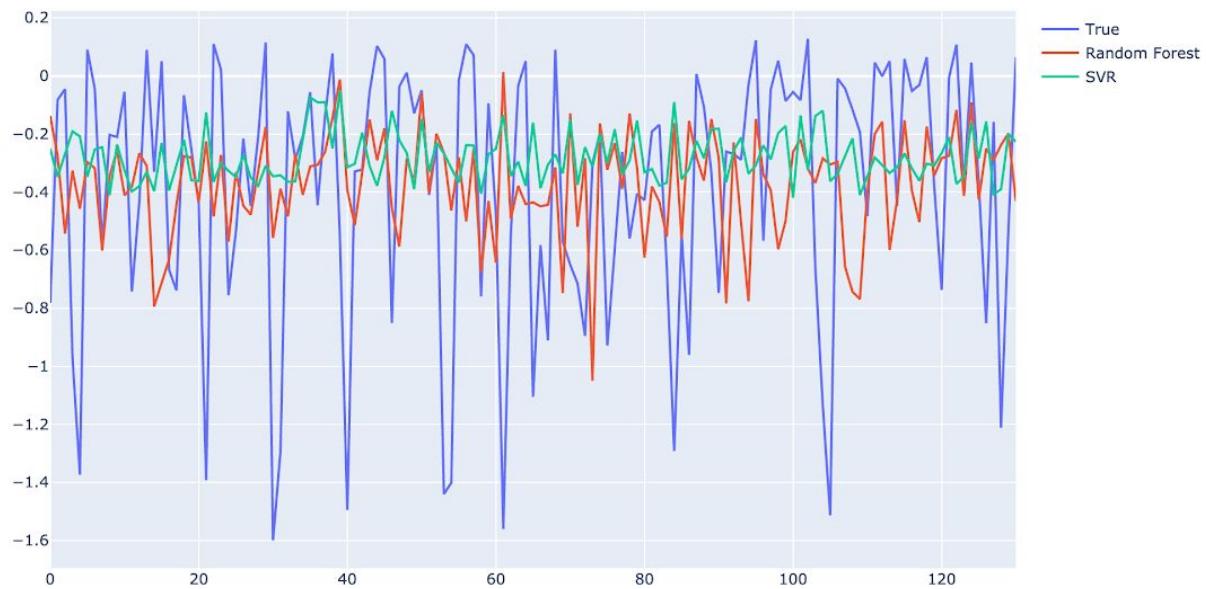


Fig. 26. Comparación de las predicciones de los modelos SVMR y Random Forest contra los valores reales del costo por tasa de producción del cluster 3.

Cluster 4

EC_TP

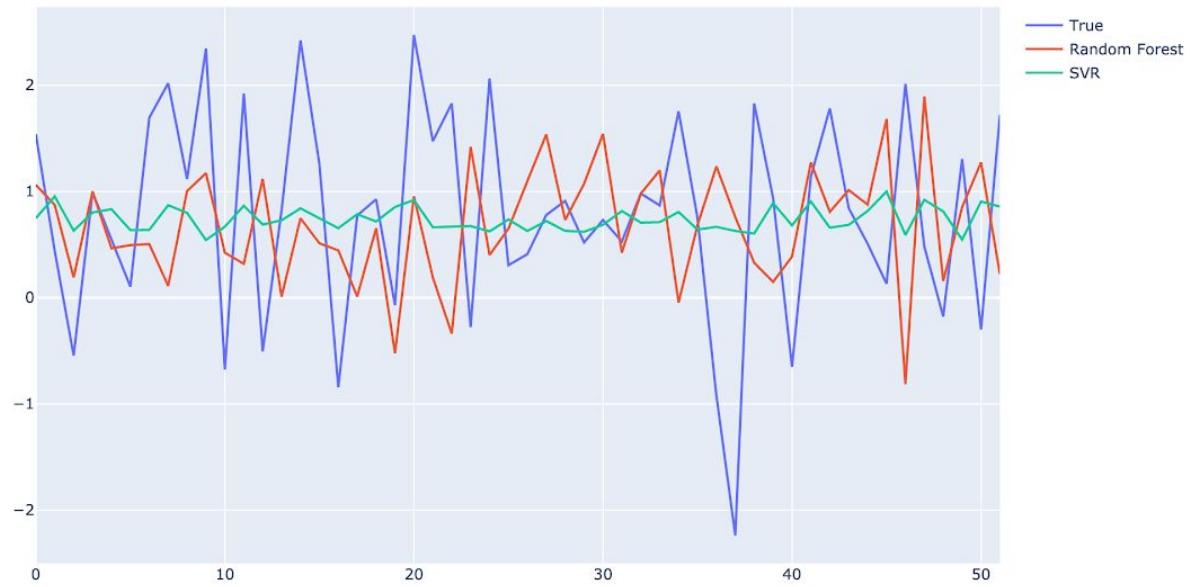


Fig. 27. Comparación de las predicciones de los modelos SVMR y Random Forest contra los valores reales de la energía calórica por tasa de producción del cluster 4.

EE_TP



Fig. 28. Comparación de las predicciones de los modelos SVMR y Random Forest contra los valores reales de la energía eléctrica por tasa de producción del cluster 4.

Costo_TP

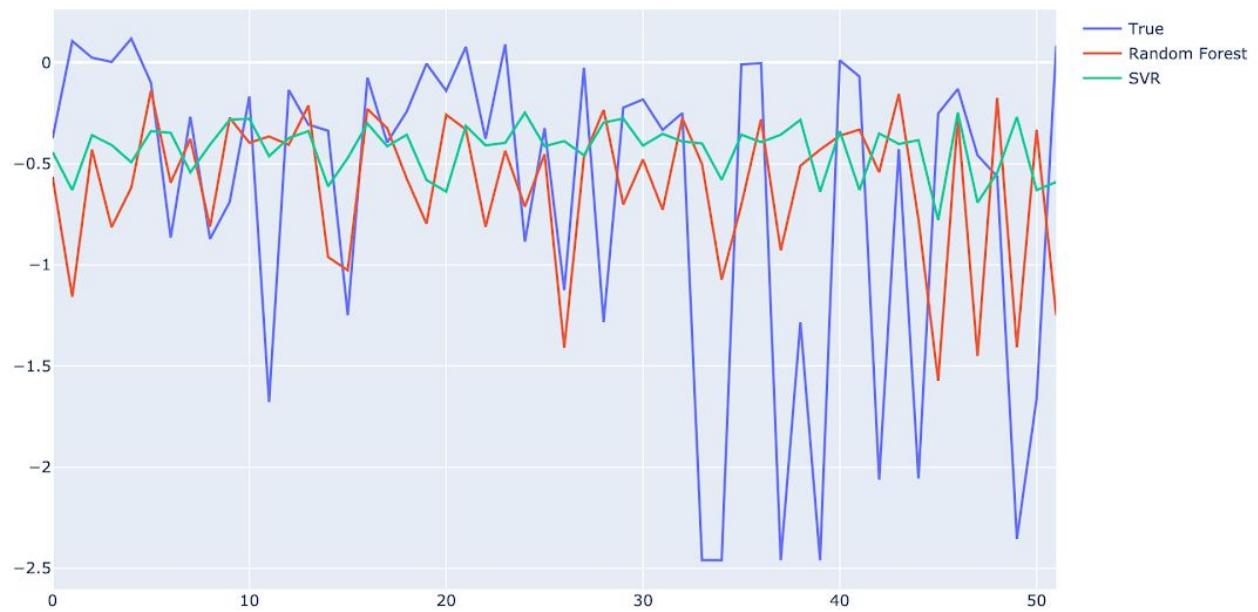


Fig. 29. Comparación de las predicciones de los modelos SVMR y Random Forest contra los valores reales del costo por tasa de producción del cluster 4.

Cluster 5

EE_TP

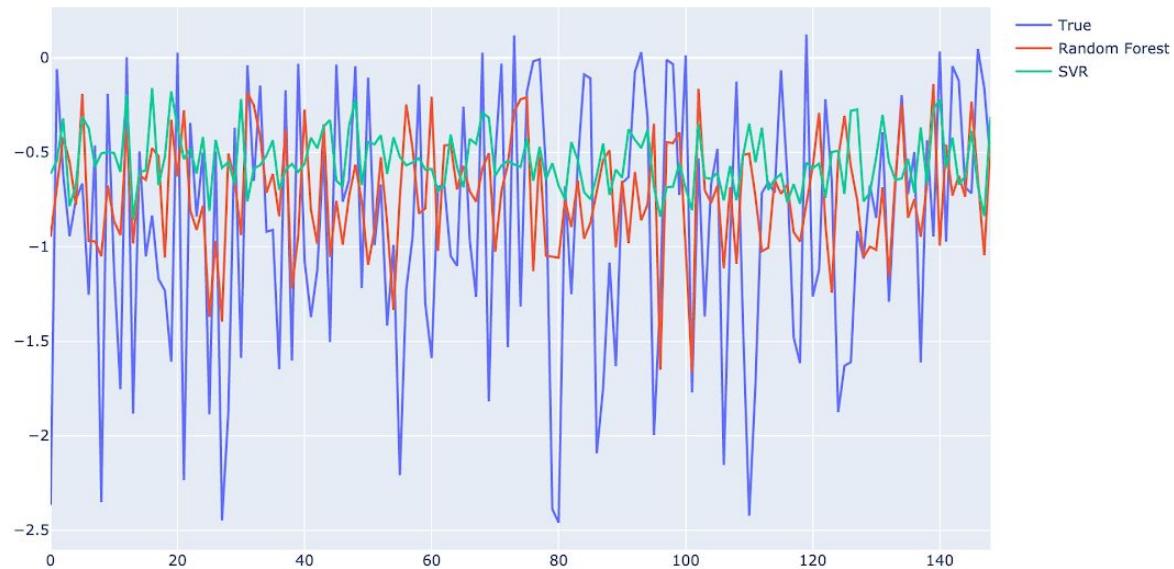


Fig. 30. Comparación de las predicciones de los modelos SVMR y Random Forest contra los valores reales de la energía eléctrica por tasa de producción del cluster 5.

EC_TP



Fig. 31. Comparación de las predicciones de los modelos SVMR y Random Forest contra los valores reales de la energía calórica por tasa de producción del cluster 4.

Costo_TP

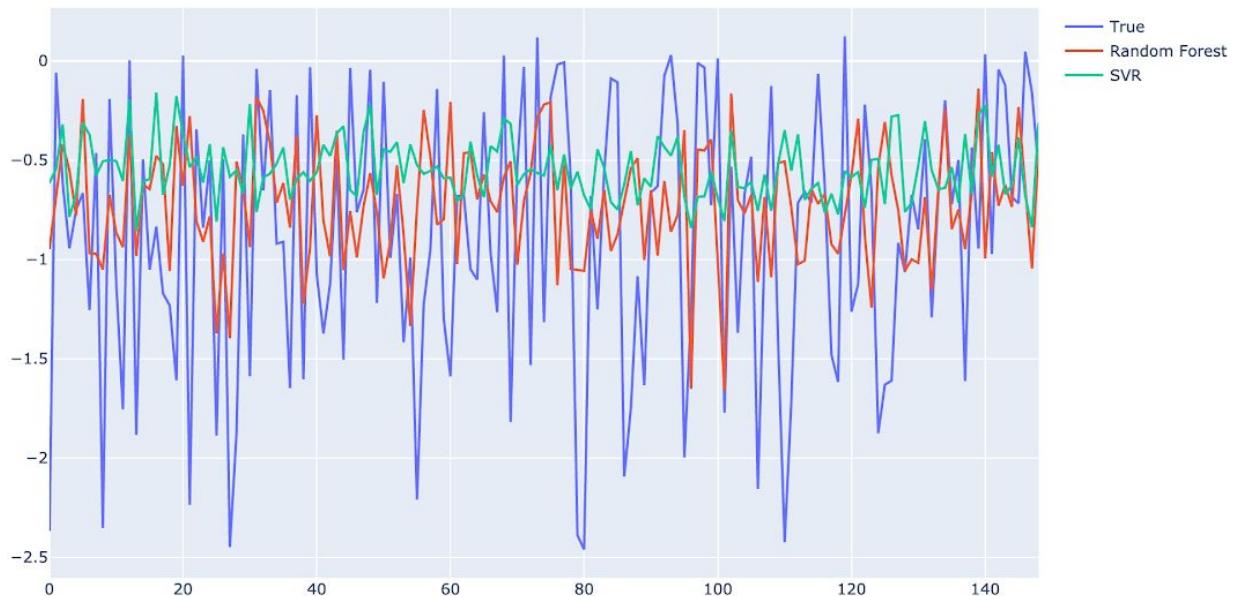


Fig. 32. Comparación de las predicciones de los modelos SVMR y Random Forest contra los valores reales del costo por tasa de producción del cluster 5.

De las gráficas anteriores, se puede observar que el algoritmo de random forest es el que, aunque tenga mayor error que las máquinas de vector de soporte, extrapolan de manera más exacta el comportamiento de las energías y el costo por tasa de producción, pues la SVM's extrapolan los datos de manera más cerrada, esto quiere decir que el error es menor pero no modelan de manera correcta el comportamiento, como si de los promedios se tratase, puesto que la regresión modela a los datos desde la media hacia afuera de los datos, por eso se ve que las SVM tienen menos variación que los bosques aleatorios y un comportamiento más tímido, pero no por eso deja de ser útil. Este comportamiento en específico es de mucha ayuda, ya que la ventaja de esto es que debido a la alta variación de los precios puede darnos una métrica muy óptima de lo que sería el costo mínimo, pues el intervalo de confianza al 99% del error muestra un error promedio bastante reducido.

Por otro lado, el comportamiento del algoritmo de random forest es de destacar, ya que está extrapolado de una manera bastante similar a lo que son los datos originales y todo eso por un error considerablemente bajo para este tipo de fluctuaciones en los valores de los costos mínimos, lo que lo hace el mejor regresor práctico de todos y se ha comprobado que los resultados en el campo serán verdaderamente óptimos.

En pocas palabras, el modelo de random forest tiene un error mayor al regresor de SVM, pero extrapola de mejor manera a los datos, pero las máquinas de vector de soporte, no la extrapola tan bien, pero las estimaciones encuentran en un rango más “promedio” de la extrapolación. Por lo tanto ambos modelos son útiles debido a este comportamiento.

Modelo al 95% de Confianza

Tabla general

EE_TP

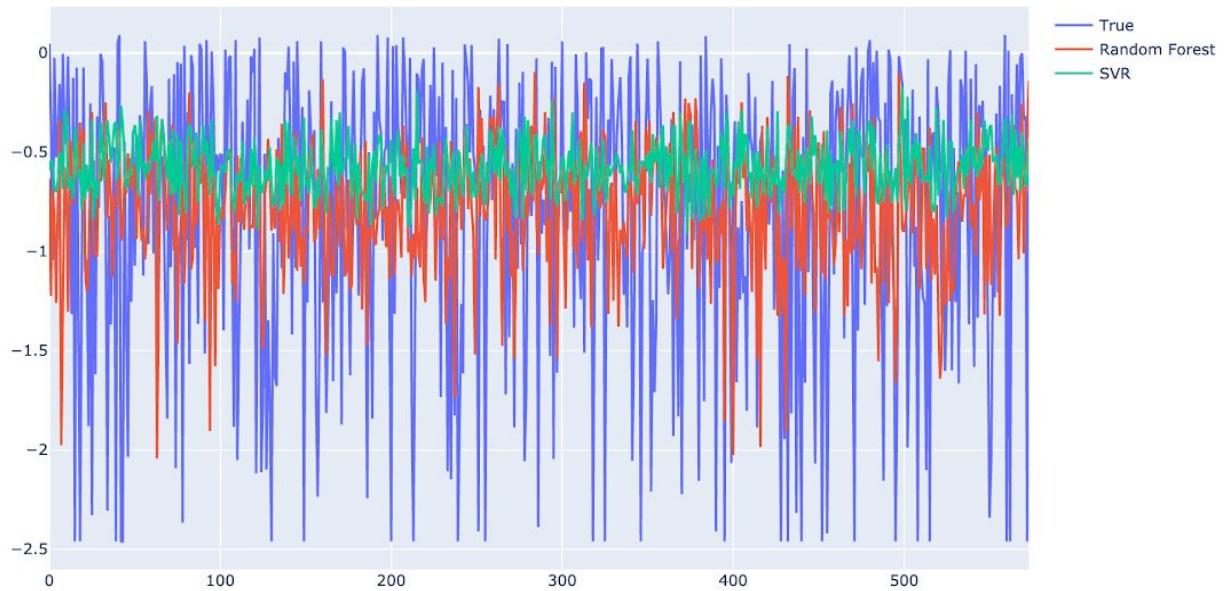


Fig. 33. Comparación de las predicciones al 95% de confianza de los modelos SVMR y Random Forest contra los valores reales de la energía eléctrica por tasa de producción de la tabla general.

EC_TP



Fig. 34. Comparación de las predicciones al 95% de confianza de los modelos SVMR y Random Forest contra los valores reales de la energía calórica por tasa de producción de la tabla general.

Costo_TP

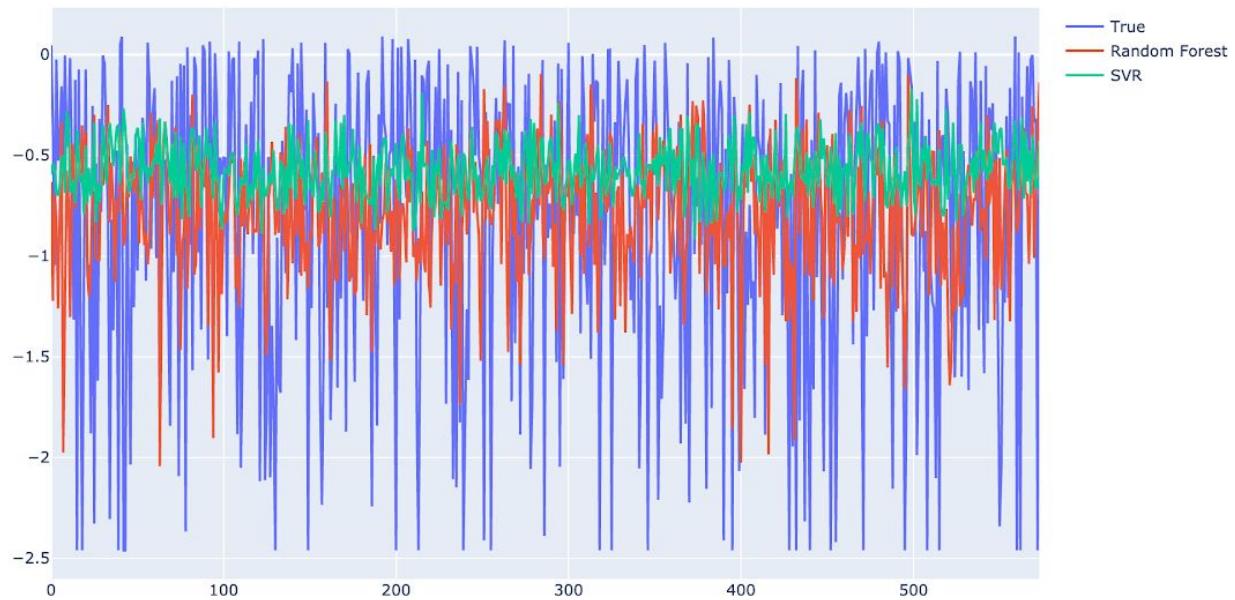


Fig. 35. Comparación de las predicciones al 95% de confianza de los modelos SVMR y Random Forest contra los valores reales del costo por tasa de producción de la tabla general.

Cluster 1

EE_TP

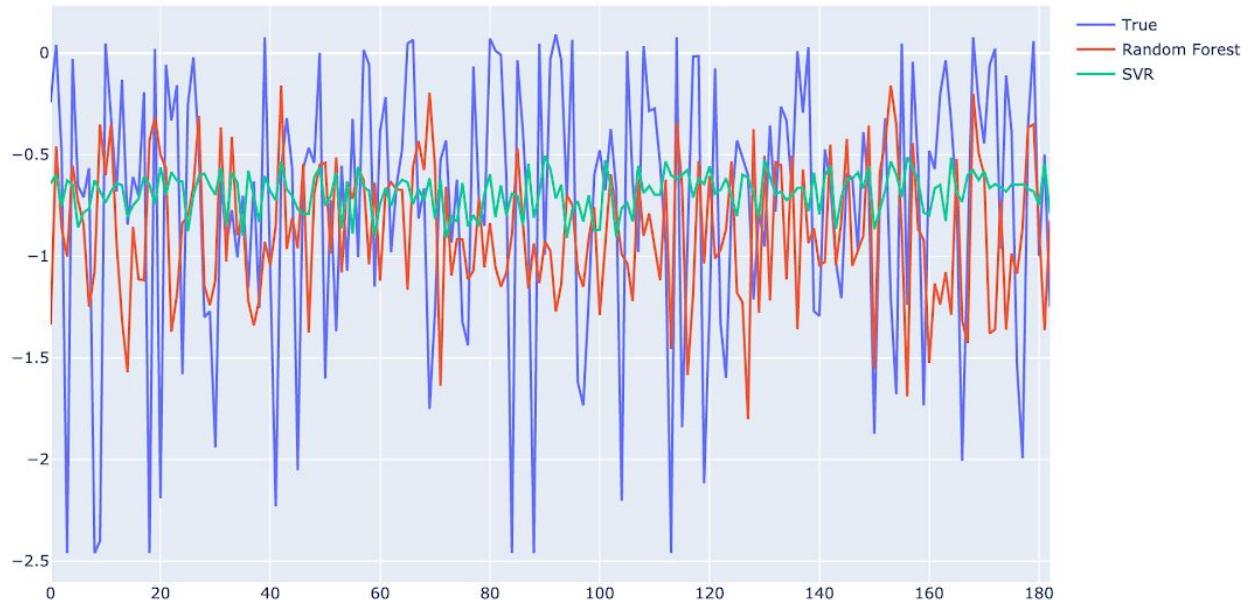


Fig. 36. Comparación de las predicciones al 95% de confianza de los modelos SVMR y Random Forest contra los valores reales de la energía eléctrica por tasa de producción del cluster 1.

Costo_TP



Fig. 37. Comparación de las predicciones al 95% de confianza de los modelos SVMR y Random Forest contra los valores reales del costo por tasa de producción del cluster 1.

EC_TP

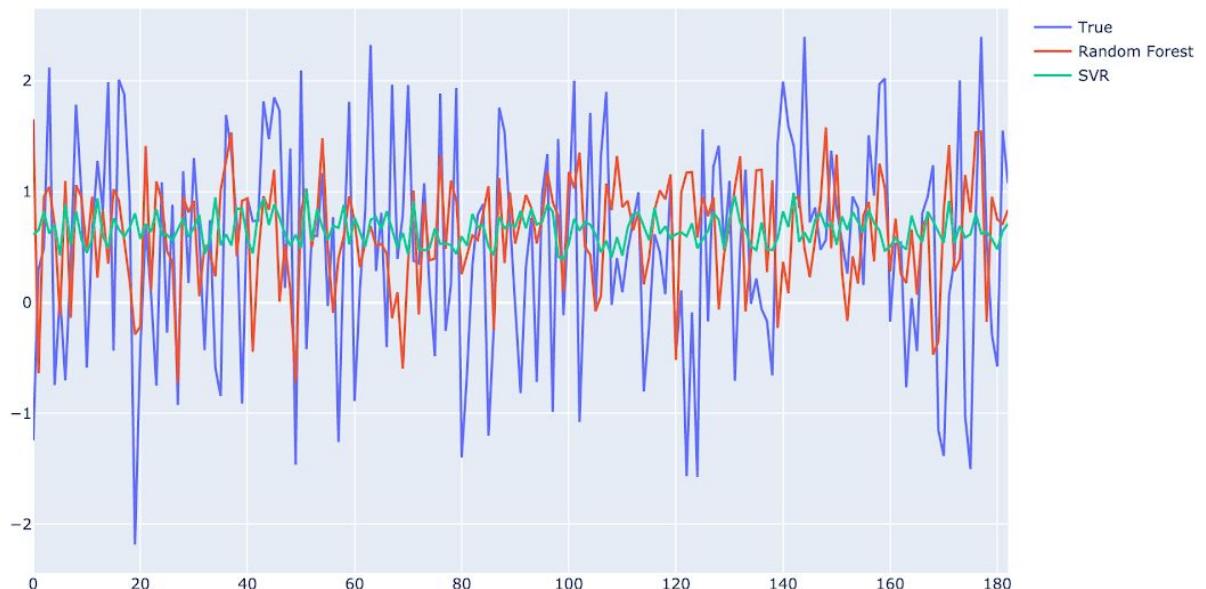


Fig. 38. Comparación de las predicciones al 95% de confianza de los modelos SVMR y Random Forest contra los valores reales de la energía calórica por tasa de producción del cluster 1.

Cluster 2



Fig. 39. Comparación de las predicciones al 95% de confianza de los modelos SVMR y Random Forest contra los valores reales de la energía calórica por tasa de producción del cluster 2.

EE_TP



Fig. 40. Comparación de las predicciones al 95% de confianza de los modelos SVMR y Random Forest contra los valores reales de la energía eléctrica por tasa de producción del cluster 2

Costo_TP



Fig. 41. Comparación de las predicciones al 95% de confianza de los modelos SVMR y Random Forest contra los valores reales del costo por tasa de producción del cluster 2.

Cluster 3

EC_TP



Fig. 42. Comparación de las predicciones al 95% de confianza de los modelos SVMR y Random Forest contra los valores reales de la energía calórica por tasa de producción del cluster 3.

EE_TP



Fig. 43. Comparación de las predicciones al 95% de confianza de los modelos SVMR y Random Forest contra los valores reales de la energía eléctrica por tasa de producción del cluster 3.

Costo_TP



Fig. 44. Comparación de las predicciones al 95% de confianza de los modelos SVMR y Random Forest contra los valores reales del costo por tasa de producción del cluster 3.

Cluster 4

EC_TP



Fig. 45. Comparación de las predicciones al 95% de confianza de los modelos SVMR y Random Forest contra los valores reales de la energía calórica por tasa de producción del cluster 4.

EE_TP

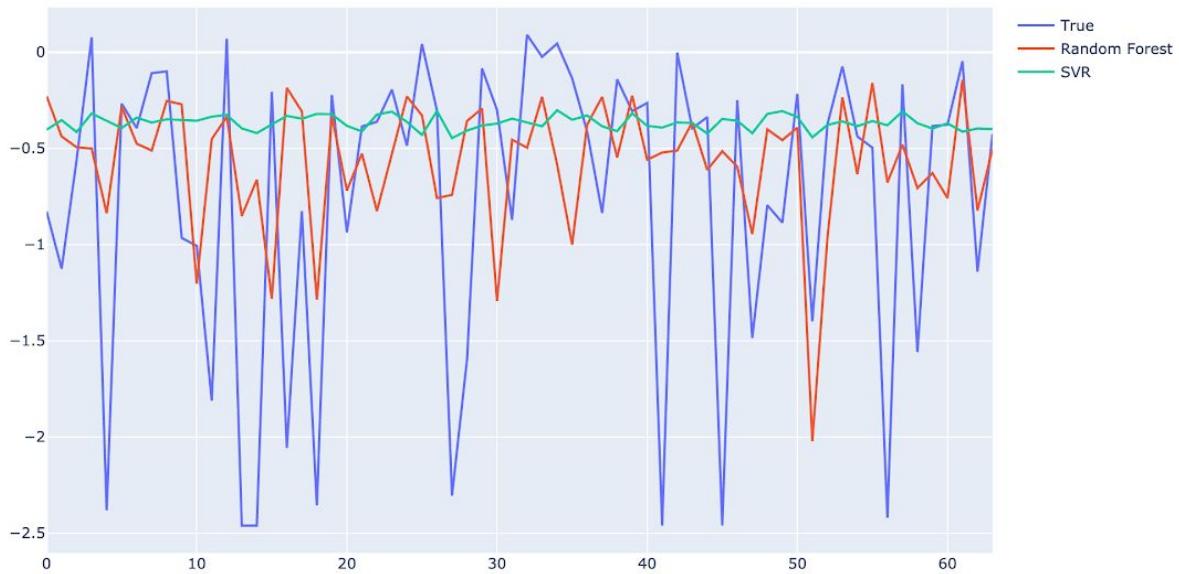


Fig. 46. Comparación de las predicciones al 95% de confianza de los modelos SVMR y Random Forest contra los valores reales de la energía eléctrica por tasa de producción del cluster 4.

Costo_TP

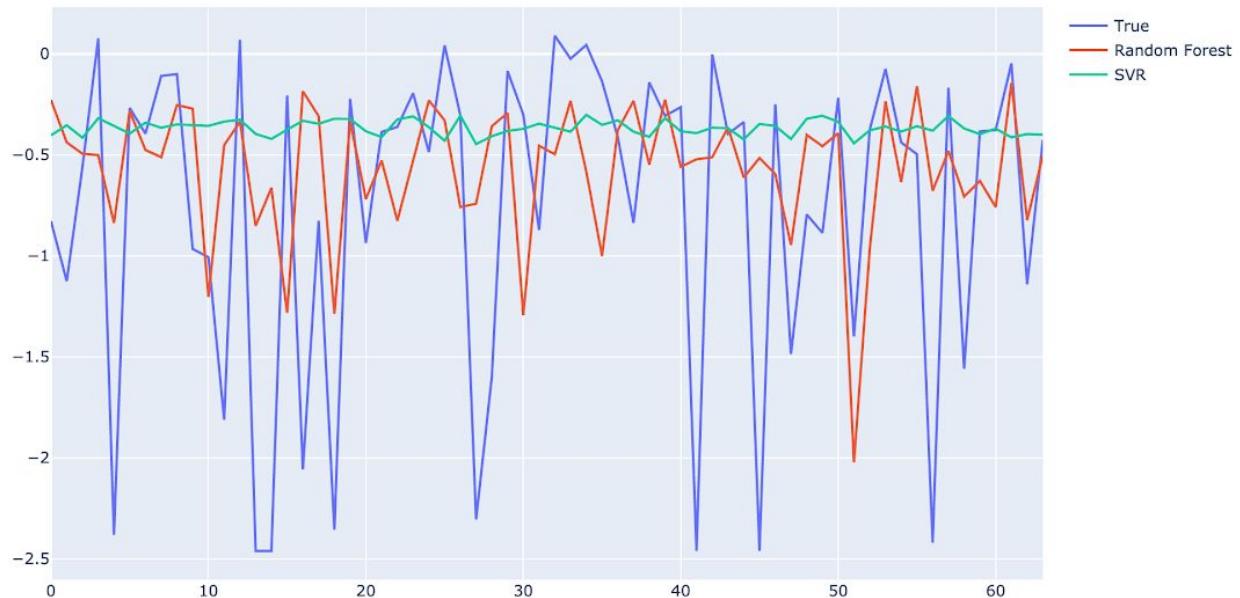


Fig. 47. Comparación de las predicciones al 95% de confianza de los modelos SVMR y Random Forest contra los valores reales del costo por tasa de producción del cluster 4.

Cluster 5

EC_TP

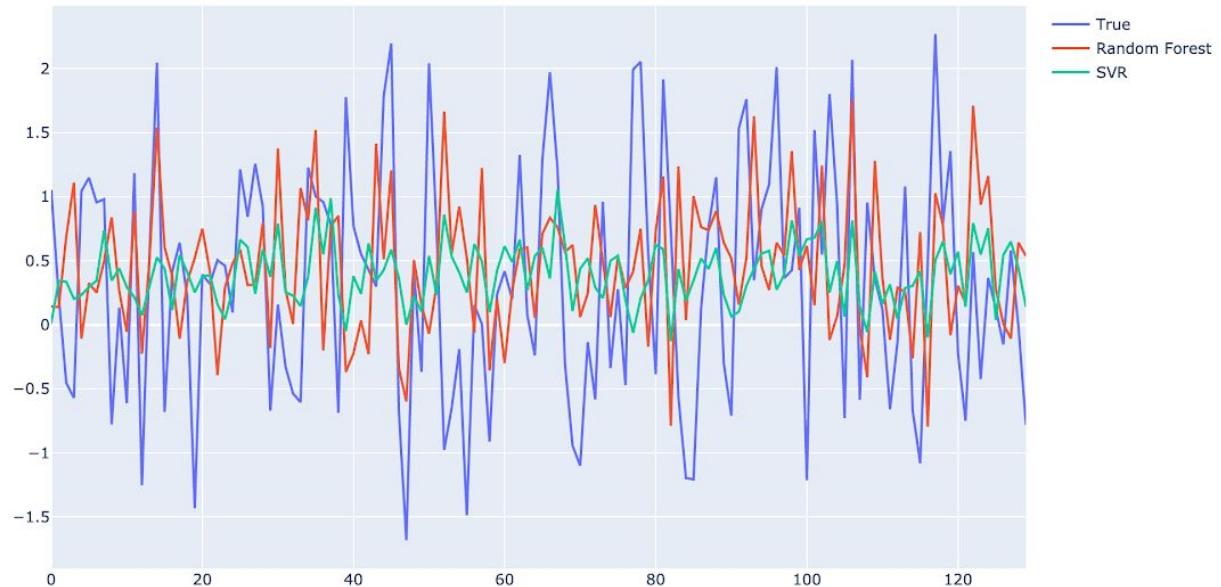


Fig. 48. Comparación de las predicciones al 95% de confianza de los modelos SVMR y Random Forest contra los valores reales de la energía calórica por tasa de producción del cluster 5.

EE_TP



Fig. 49. Comparación de las predicciones al 95% de confianza de los modelos SVMR y Random Forest contra los valores reales de la energía eléctrica por tasa de producción del cluster 5.

Costo_TP

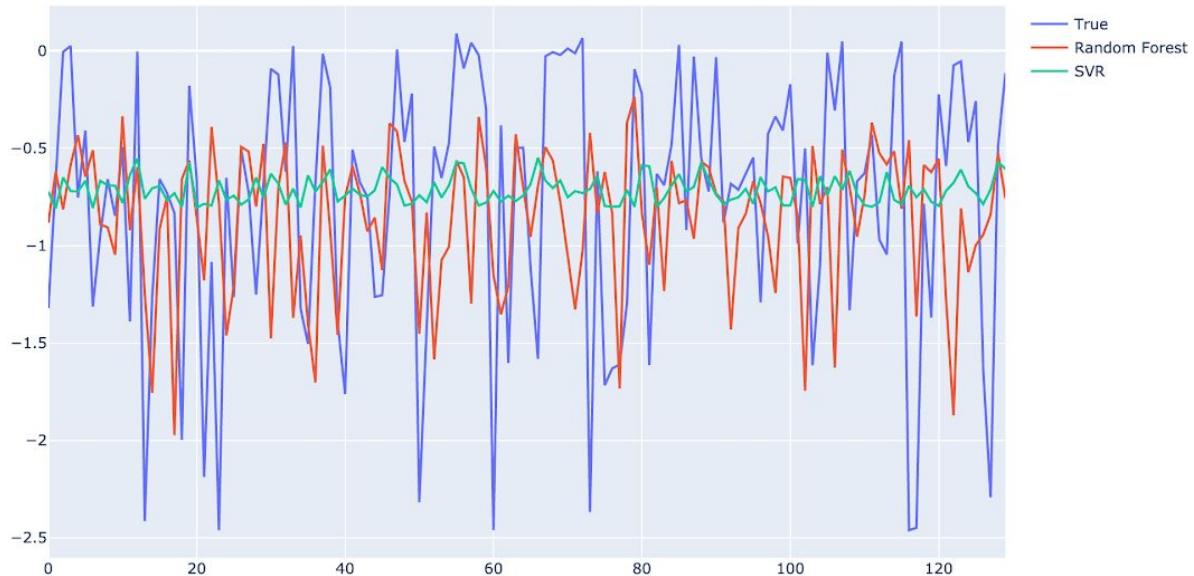


Fig. 50. Comparación de las predicciones al 95% de confianza de los modelos SVMR y Random Forest contra los valores reales del costo por tasa de producción del cluster 5.

Para los modelos con un intervalo de confianza del 95% se puede observar un comportamiento bastante similar que en el anterior, pues los modelos siguen el mismo principio. Es por eso que ambos modelos, tanto para el intervalo de confianza como para la distribución general, se encuentran útiles y aptos para el análisis, pues las características que ambos brindan a la aproximación del costo mínimo con la máxima calidad, hacen que el dúo de ambos modelos sean una herramienta muy poderosa.

7. Implementación

7.1. Aplicación web

Para consumir el modelo realizado se implementó una aplicación web en google colab, la cual, recibe como parámetros la dureza, calidad y tasa de producción que se desee usar y devuelve las estimaciones del costo mínimo óptimo (CMO) y costo mínimo promedio (CMP) por algoritmo junto con un set de visualizaciones, en las que se incluye una visualización 3D del clúster al que pertenece dicha estimación e interpretaciones. La aplicación se encuentra en el siguiente enlace: <https://docs.google.com/document/d/14SBKIHQjGQHfo2LjbMnWoJp1GAPtu-VHbqVI2KEFm8c/edit?usp=sharing>

7.2. REST API

A su vez, también se implementó una REST API con la finalidad de que otros servicios puedan interactuar con nuestro modelo, igualmente se encuentra en etapa de desarrollo, la documentación y enlace actualizados se puede encontrar en el enlace que se encuentra en el punto 6.1.

La API se basa en los principios REST. Se accede a los recursos de datos a través de solicitudes HTTP estándar en formato UTF-8 al endpoint de la API que se encuentra en la documentación.

Dada la naturaleza de los modelos, solamente se aceptan peticiones con el verbo GET.

Parámetro	Descripción	Ejemplo
Dureza	Dureza del material a trabajar	dureza=99
Tasa de producción	Tasa de producción deseada	tasaprod=368
Calidad	Calidad esperada por unidad producida	calidad=0.035
Token	Json Web Token para autenticación	token=FUD34...

Cada petición se responde con un formato JSON que contiene la información obtenida de los diferentes modelos implementados, nuevamente, la documentación detallada está disponible en el enlace del punto 7.1.