Optimización de escenarios en producción

Socio Formador: CEMEX Ventures

Por Ernesto Borbón A01701515, Gerardo Villegas A00571388, Jesús Gutiérrez A01637812, Felipe Villaseñor A01023976

Entendimiento del negocio

Entre las actividades de la empresa CEMEX Ventures, una de sus fábricas produce soportes metálicos. En su producción se utilizan dos maquinarias distintas. Por un lado, la máquina antigua es alimentada con energía calórica (EC) en forma de diesel, mientras que la máquina moderna se alimenta con energía eléctrica (EE). La máquina moderna es más eficiente, produciendo más unidades y de mayor calidad, pero requiriendo un mayor costo de producción. Se sabe que el costo de la energía calórica es el 72.4 % del costo de la eléctrica. Cabe destacar que ambas maquinas se utilizan al mismo tiempo en diferentes cantidades.

Además, los materiales que se emplean para la construcción de los soportes son laminas y perfiles metálicos, los cuales provienen de distintos proveedores y en consecuencia tienen distintas durezas y resistencias. La empresa registra diariamente la calidad, dureza, tasa de producción, energía eléctrica, energía calórica y un valor de aspiración. Donde la calidad de los soportes metálicos que se fabrican es evaluada por un equipo de la fábrica, el cual le asigna un valor normalizado (entre 0 y 1). Las calidades menores a 0 son inaceptables.

El **objetivo** de este proyecto es encontrar la combinación óptima de ambas maquinarias para la producción de un día, de tal forma que se minimice el gasto energético. Como acercamiento de solución, se espera producir un modelo de regresión.

En él se proveería la calidad deseada, la dureza de los materiales y la tasa de producción del día y se calcularía la cantidad óptima de EE y EC que minimiza el costo ponderado unitario. Para la producción de este modelo se poseen datos de producción históricos y se escribirá en Python.

La metodología con la que se dirigirá el proyecto es la *Cross Industry Standard Process for Data Mining*.

Hipótesis

A pesar de que hay registros con características similares respecto a calidad, dureza y tasa de producción, sus gastos energéticos son muy variados, y esto es porque se invirtieron diferentes combinaciones de energías. Sin embargo, de todas esas combinaciones, existe una que devolvió el costo más bajo.

Es decir, si dos registros son lo suficientemente similares, entonces su costo óptimo debería ser aproximadamente el mismo.

En la siguiente tabla se muestra un ejemplo dos registros distintos con Calidad, Dureza y Tasa de producción similares, pero tiene un costo de producción distinto, estos al tener valores de entradas tan parecidos pueden tener casi mismo costo óptimo.

Calidad	Dureza	Tasa de producción	Costo ponderado por unidad
0.21	105	450	0.05623
0.2	105	435	0 0.110414

Así se propone la creación de grupos de datos para la extracción de su respectivo costo ponderado por unidad mínimo.

El meollo es cómo se define la similaridad entre registros, o bien, cómo se agrupan. En el desarrollo del proyecto se trabajó con dos vertientes, agrupación por solo calidad, y agrupación por algo que se denominó “categorías 3D”.

Se hipotetiza que lo más idóneo es crear categorías “cúbicas” tras dividir los registros por contenedores de Calidad, Dureza y Tasa de producción.

La intersección de estos contenedores formaría a “los cubos”, y de cada cubo se extraería el costo ponderado unitario mínimo y se asignaría a todos los datos dentro de la categoría.

Es con estos datos que se entrenaría el regresor. De esta forma se cree que se estarían considerando las tres variables en el modelo.

Se comprobaría que la hipótesis es correcta si el gasto energético es directamente proporcional a la calidad, dureza y tasa de producción.

Entendimiento de los datos

Para la realización de este modelo, se cuenta con un archivo csv con los datos diarios de producción desde 1995. El total de registros es de 9392, y sus atributos numeran 7, los cuales son:

- TIME: Día de producción.
- Dureza: Resistencia de los materiales a ser manipulados.
- Tasa_Prod: Cantidad de soportes producidos en el día en toneladas por día.
- Asp: Valor de aspiración asociado a las virutas de metal aspiradas en el proceso.
- EC: Cantidad de energía calórica utilizada en Kcal.
- EE: Cantidad de energía eléctrica utilizada en Kcal.
- Calidad: Métrica que evalúa la calidad de los soportes metálicos (entre 0 y 1).

	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

Figura 1: Data Frame original

En el archivo original existe un dato nulo en la columna del valor de aspereza, y otro en la columna de dureza.

Se utilizaron gráficas de caja para observar la distribución y el comportamiento general de los datos.

En primer lugar, en la gráfica de caja de la dureza (Figura 2) se puede notar que la mayoría de las durezas se encuentran entre 100 y 108. Por otro lado, la gráfica de caja de la tasa de producción (Figura 3) muestra que la mayoría de las tasas de producción se concentran entre 350 y 450. Además es muy notoria la presencia de registros con 0 en tasa de producción.

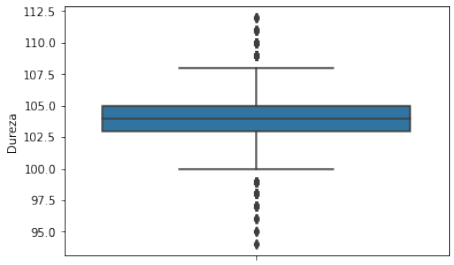

Figura 2: Boxplot dureza

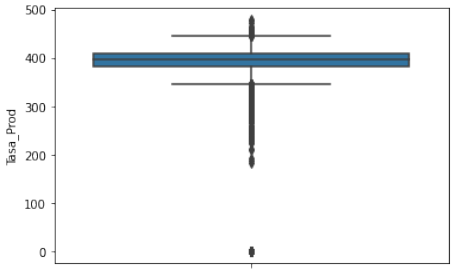

Figura 3: Boxplot tasa de producción

La gráfica de caja del valor de aspiración (Figura 4) muestra que la mayoría de los datos se distribuyen entre 2.5 y 3.5. Mientras que a través de la gráfica en la Figura 5 se puede notar que no hay calidades por debajo de 0. Es decir, no hay soportes con calidad inaceptable. Además de que la mayoría de las calidades se distribuyen entre 0 y 0.2.

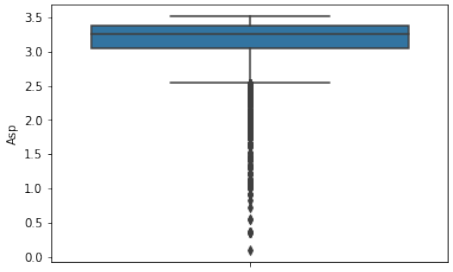

Figura 4: Boxplot valor de aspiración

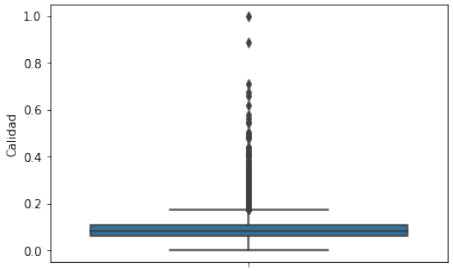

Figura 5: Boxplot calidad

Finalmente, las gráficas de caja de EC y EE muestran distribuciones muy parecidas. La de la energía calórica (Figura 6) muestra que los datos se distribuyen entre 0 y 40, aunque la mayoría se encuentran entre 5 y 35. Similarmente, la de la energía eléctrica (Figura 7) muestra que los datos se distribuyen entre 0 y 35, pero no tiene valores atípicos.

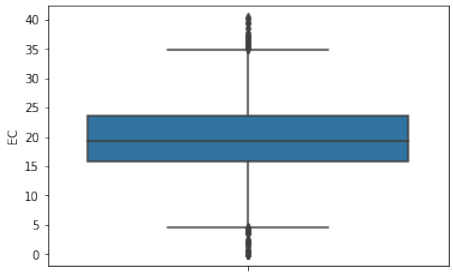

Figura 6: Boxplot EC

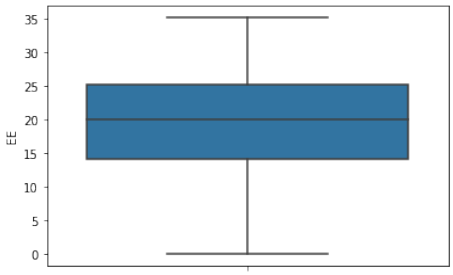

Figura 7: Boxplot EE

Adicionalmente se analizan las correlaciones entre las variables a través de un mapa de calor (Figura 8). Es notable que ningún par de variables está fuertemente correlacionada. La mayor correlación es entre las variables EE y EC y es negativa. También existe cierta correlación entre la EC y la tasa de producción y, entre el valor de aspiración y la tasa de producción.

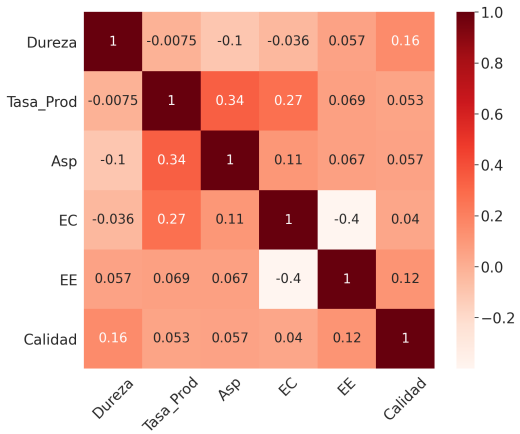

Figura 8: Mapa de calor de correlaciones

Preparación de datos

Como los registros con valores NaN son muy pocos en comparación al total de registros, se optó por eliminarlos. Se verificó también que no existieran fechas repetidas, pues solo se debe tomar una medida por día.

Siguiendo con la hipótesis de las categorías 3D, es de alta importancia prescindir de aquellas con muy pocos datos, pues pudieran extraer un mínimo no representativo y entorpecer el futuro modelo. Con esto en consideración, se realizaron los siguientes filtros:

Variable	Criterio de eliminación
Calidad	Registros con calidad mayor de 0.3
Tasa_Prod	Tasa de producción menor de 250
Dureza	Valores menores a 97 y mayores a 112
Asp	Registros menores a 1
Energías	Suma de las energías menor a 10

Luego de eliminar todo aquel dato considerado atípico se crearon dos variables: costo_ponderado y costo_pon_un. La primera describe el costo energético total en un día (medido en costo relativo a la EE) y la segunda el costo por unidad al día (ponderado con el costo de la EE).

- costo_ponderado: EE+(EC*0.724).
- costo_pon_un: (costo_ponderado)/(Tasa_Prod).

Se realizo una gráfica de dispersión en donde se pueden ver las variables de entrada Calidad, Dureza y Tasa de producción y el costo ponderado unitario es representado mediante los colores, como se puede observar en la Figura 9 para registros con calidades, durezas y tasa de producción similares los costos son muy distintos.

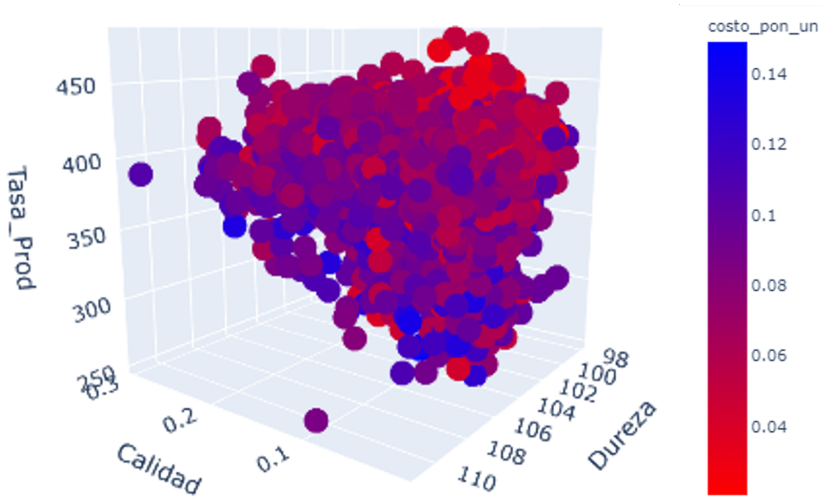

Figura 9: Visualización por calidad, dureza y tasa de producción

Para la implementación de los contenedores o *bins*, se necesitó la adición de otras 3 variables:

- Calidad_cut: representa al contenedor al que pertenece el registro según su calidad, puede tomar un valor entre 0 y 5.
- Dureza_cut: representa al contenedor al que pertenece el registro según su dureza, puede tomar un valor entre 0 y 12.
- TP_cut: representa al contenedor al que pertenece el registro según su dureza, puede tomar un valor entre 0 y 7.

Mediante la combinación de estas 3 variables se creó una cuarta llamada “cartesian”, que indica la posición de cada valor dentro del dataset.

Ya creada la variable cartesian se buscan los índices de los registros con menor costo ponderado unitario de cada bin 3D.

Con estos registros se creó una base de datos nueva añadiendo los valores óptimos de la energía eléctrica, calórica, costo ponderado y costo ponderado unitario, en columnas de nombre EE_op, EC_op, costo_pon_op y costo_op respectivamente.

Al obtener los valores óptimos por registro según su categoría, se obtiene la información valiosa para entrenar al regresor.

	cartesian	Calidad	Dureza	EC	EC_op	EE	EE_op	TIME	Tasa_Prod	costo_op	costo_pon_op	costo_pon_un	costo_ponderado
0	(0, 0, 1)	0.045	98.0	24.8	13.2	9.4	18.2	2000-07-08	290	0.093144	27.7568	0.094328	27.3552
1	(0, 0, 1)	0.032	98.0	13.2	13.2	18.2	18.2	2007-12-04	298	0.093144	27.7568	0.093144	27.7568
2	(0, 0, 1)	0.031	98.0	7.1	13.2	23.6	18.2	2007-04-24	299	0.093144	27.7568	0.096122	28.7404
3	(0, 0, 4)	0.043	99.0	20.0	24.1	21.6	0.0	1995-02-04	379	0.045676	17.4484	0.095198	36.0800
4	(0, 0, 4)	0.038	99.0	24.5	24.1	17.9	0.0	2000-06-28	372	0.045676	17.4484	0.095801	35.6380

Figura 10: Base de datos nueva

Modelado

El modelo de regresión tiene como entrada calidad, dureza y la tasa de producción y como salida los óptimos de las energías y el costo ponderado unitario.

Por lo cual para el modelo se asignaron X (variables independientes) y Y (variables dependientes) de la siguiente manera:

- X: con los valores de Calidad, Dureza y Tasa_prod.
- Y: con los registros de EE_op, EC_op y costo_op.

Después de definir las variables de entrada y salida, se separó la base de datos en los conjuntos de entrenamiento y prueba, con 10 % para la prueba y 90 % para el entrenamiento. Al tener ya separada la base de datos se entrenaron varios modelos de regresión con el propósito de poder encontrar aquel que predice de mejor manera. Los modelos empleados son los siguientes:

Modelos	Acercamiento de Multi Output Regression
Linear Regression	Direct Multioutput Regerssion
Decision Tree Regressor	Direct Multioutput Regression
Random Forest Regressor	Direct Multioutput Regression
XGB Regressor	Direct Multioutput Regression

Evaluación

Los modelos propuestos fueron evaluados con distintas métricas, las cuales son el coeficiente de determinación (R^2), error logarítmico cuadrado medio (MSLE), error cuadrado medio (MSE) y error absoluto medio (MAE). (Figura 11).

	Regresor	Puntaje_r2	MSLE	MSE	MAE
2	Random forest	0.927309	0.035598	1.876807	0.253724
3	XGBRegressor	0.924644	NaN	2.464481	0.232161
1	Árbol de decisión	0.902771	0.030366	2.514609	0.172662
0	Regresión lineal múltiple	0.149565	1.128577	37.456786	4.195652

Figura 11: Métricas de evaluación

Se observa que el árbol de decisión tiene mejor desempeño considerando MAE y MSLE, mientras que el random forest sobresale en R2 Y MSE.

Finalmente, se optó por el modelo de random forest, con 0.9273 de R2 y 1.8768 de MSE.

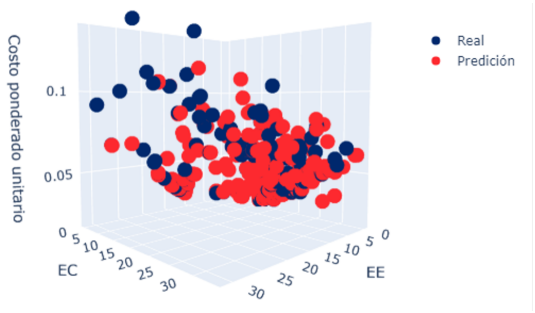

Figura 12: Visualización de error por calidad, dureza y tasa de producción

Al hacer pruebas de predicción con el modelo, se cumple que el costo y gasto energético en la salida son directamente proporcionales a la calidad, dureza y tasa de producción, lo que refuerza la hipótesis. Esto no se habría cumplido si por ejemplo solo se hubieran agrupado los registros por calidad y de cada grupo se extrajera el mínimo, pues en ese caso los cambios en las entradas de dureza y tasa de producción no se verían reflejados en las salidas del regresor.

Este modelo, al emplearlo en los valores históricos refleja que habría ahorrado en promedio 18.5 unidades de costo ponderado por día o 0.047 unidades de costo ponderado al día por tonelada.

A pesar de hiperparametrizar el regresor de XGBoost y utilizar detención temprana, no se logró que superara al random forest o al árbol de decisión en cuanto a las métricas de desempeño.

Los modelos con la base de datos escalada logran mejorar por muy poco el desempeño, y su implementación causa que el resto del proceso sea mas tardado. Es por ello que dicha técnica no fue empleada para obtener la solución.

Deployment

Trabajar el modelo solamente con la separación de la calidad y el costo ponderado unitario resultó no ser funcional debido a que al ingresar los valores en el regresor los valores óptimos estaban dados solo por la calidad y las demás entradas no cambiaban el resultado. Por el otro lado, el acercamiento de las categorías tridimensionales o cubos, rindió frutos satisfactoriamente.

Se ha logrado un modelo que determina combinación óptima de ambas maquinarias que reduce el gasto energético de la producción de soportes.

Este modelo arroja coeficiente de determinación R2 de 0.9273 y un error cuadrado medio de 1.8768, y se estima que ahorraría en promedio 18.5 unidades de costo ponderado por día, o bien 0.047 unidades de costo ponderado al día por tonelada.

Para dar acceso a esta solución, se elaboró con Flask una página web interactiva y fácil de usar y desplegada en Google Cloud Platform. Disponible en https://e1-cemex-app.nn.r.appspot.com

Fue necesario serializar el regresor con pickle para facilitar su implementación en la aplicación. Navegar por la página es sencillo: en la página principal se ingresa la Calidad, Dureza y Tasa de producción deseada y regresa los datos óptimos de las energías y el costo ponderado unitario en una nueva página, en la cual también se puede ver la gráfica 3D animada que visualiza el error y un video que explica de manera general el proceso seguido de la propuesta de valor. Finalmente desde todas las páginas se puede acceder a un dashboard donde se pueden visualizar una variedad de gráficas de utilidad.

Cuando se entra a la aplicación web se muestra la página de inicio (Figura 13) con un botón de continuar que te lleva a la siguiente página.

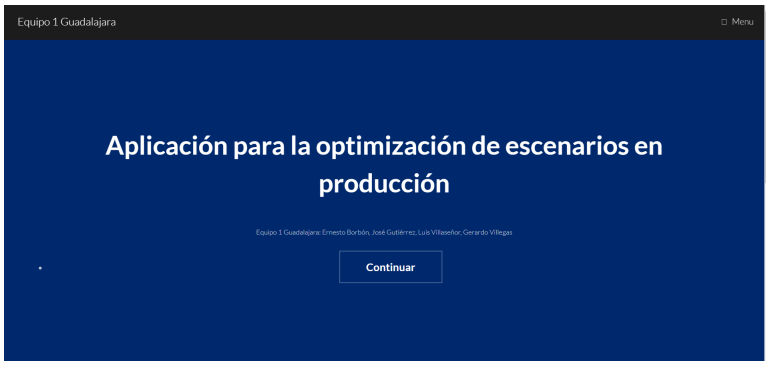

Figura 13: Página de inicio

Dentro de esta página se encuentran 3 cajas de entrada donde se escribe la Calidad, Dureza y Tasa de producción deseada para la producción del día.

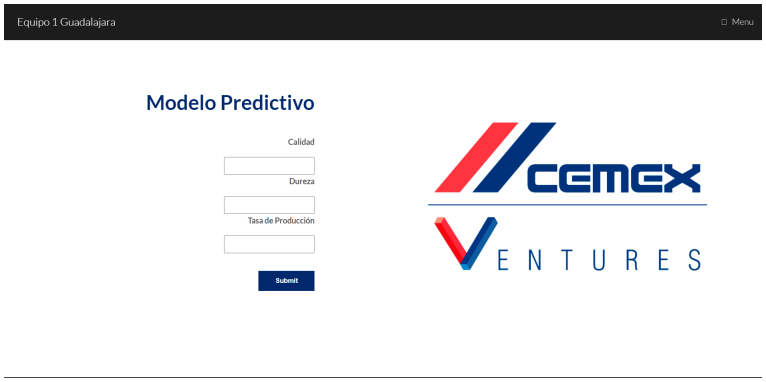

Figura 14: Página de entrada

Una vez ingresado los datos de entrada se pasa a otra página donde nos muestra un banner, el nombre de la empresa.

Figura 15: Banner de la página de resultados

Al desplazarse hacia abajo se muestran los valores de Calidad, Dureza y Tasa de producción que se ingresaron, al igual que los valores óptimos de las energías y el costo ponderado unitario resultante.

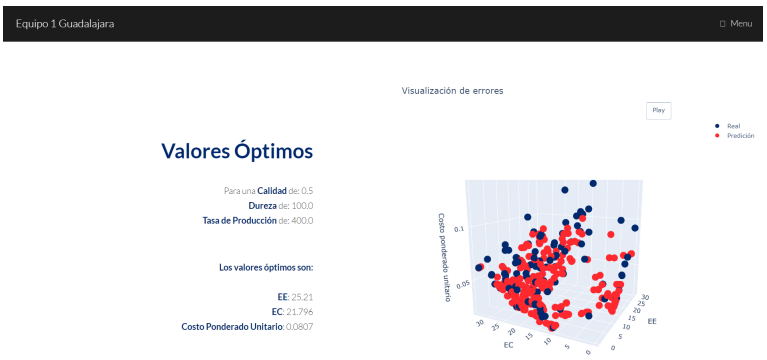

Figura 16: Página de resultados

Además de las páginas de entrada y resultados, la aplicación dispone de un menú en el cual se puede navegar a la página de inicio y a una página donde se muestra el dashboard.

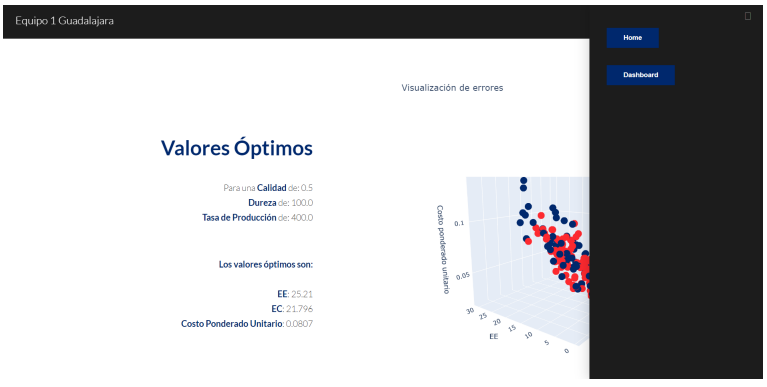

Figura 17: Menú de la página web

Página del dashboard, donde es más fácil visualizar el costo total, el promedio de uso de las energías, la combinación promedio de energías por calidad entre otras visualizaciones.

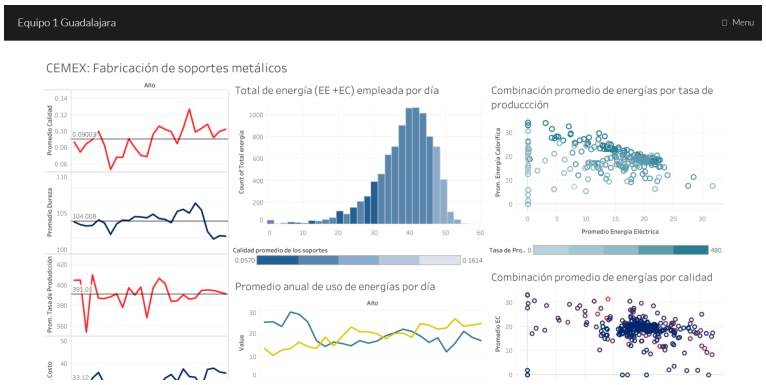

Figura 18: Dashboard

Si bien se considera al modelo propuesto en este proyecto como un buen aliado para reducir los costos de gasto energético, es de reconocerse que el modelo podría tener más precisión. Se sugiere el uso de técnicas de optimización de hiperparámetros para mejorar el regresor de XGBoost, como *grid search*.