códigofacilito

Reporte final del reto

Francisco Leonid Gálvez Flores

3 de marzo de 2022

Índice

1. Conocimiento del negocio

1.1. Objetivos del Negocio

CEMEX busca reducir el gasto de energía manteniendo un nivel mínimo de calidad. Para esto, tienen en producción una combinación de maquinaria eléctrica y combustible, las máquinas eléctricas entregan un nivel de calidad superior, mientras que las de energía combustible son más baratas de operar. Además de la tasa de producción, se debe considerar la dureza del acero utilizado tanto para el gasto eléctrico como para la calidad del producto esperado.

1.2. Situación actual del Negocio

Actualmente están en busca de una combinación de energías eléctrica y calorífica que les permitan alcanzar los objetivos previamente planteados, por lo que están realizando pruebas y solicitando el apoyo a estudiantes de ingeniería en ciencia de datos y matemáticas. [Tecnológico de Monterrey, 2021]

1.3. Objetivos a nivel de minería de datos

- Optimizar el uso de energía eléctrica y fósil de la planta de producción de acero de CEMEX.
- Buscar correlaciones entre las variables disponibles
- Comprender cómo interactúan las variables entre ellas
- Crear un modelo que dadas x variables de entrada brinde como salida el uso óptimo de energía
- Minimizar el uso de energía eléctrica, pues es la que es más cara para la empresa, pero usar estas máquinas entrega una mejor calidad respecto a las de combustible fósil

1.4. Plan del proyecto

El plan del proyecto dependerá de cinco etapas claves, las cuales se definen a continuación:

1. Extracción de datos(Semana 1):

   Esta etapa se enfoca en la importación de datos

2. Exploración de datos(Semana 2):

   Estudiar las variables, definir cuales son importantes para el objetivo y cuáles no

Explorar el dataframe, definir puntos de gestión de calidad de los datos

3. Preprocesamiento de datos(Semana 3):

   Hacer eliminación de variables innecesarias y/o redundantes

   Eliminar datos con registros nulos

   Hacer eliminación de datos atípicos para no contar con anomalías en los datos

   Modificar el índice de la fecha, de manera que este sea un objeto date en lugar de un objeto genérico de Pandas

4. Modelación(Semana 4 y 5):

   Determinar cuáles serán las variables de las cuales generamos predicciones así como cuales usar para predecir estas.

   Determinar el conjunto de datos que optimizaran la energía para cada valor de las variables de producción

   Dividir el conjunto de datos anterior en subconjuntos de entrenamiento y de evaluación

   Crear 2 modelos de regresión

   Entrenar los conjuntos de regresión

5. Evaluación y creación de una aplicación web(Semana 5):

   Evaluar los modelos creados usando métodos de cross validation, determinar cual modelo entrega mejores resultados, asegurándose de que cada uno de estos modelos esté funcionando con los parámetros óptimos, evitando el overfitting

1.5. Antecedentes del negocio

CEMEX es una empresa líder global en su campo, la venta de cemento y productos industriales similares. Fundada en 1906 en la ciudad de Hidalgo, Nuevo León, ha crecido hasta rebasar los 40,000 empleados, y ha expandido sus operaciones a múltiples lugares, teniendo más de 250 plantas de concreto y 95 centros de distribución. [CEMEX, 2019].

Dentro de los múltiples desafíos que una empresa de este calibre encuentra, está la optimización de recursos energéticos. Para lograr este fin han llevado registro de distintas variables que describen este proceso, tales como Buscan reducir el gasto energético al procesar acero, esto deben hacerlo logrando una relación óptima entre máquinas que usan energía combustible y energía eléctrica. Se sabe que la energía combustible le cuesta 0.724 veces a la empresa lo que la energía eléctrica, pero las máquinas que usan energía eléctrica entregan un trabajo de mayor calidad. [Tecnológico de Monterrey, 2021]

2. Comprensión de los datos

2.1. Procesos de captura de datos

Los datos de la base de datos se recabaron desde el 1^{ro} de enero del año 1995 hasta el 30 de diciembre del año 2019, normalmente con entradas diarias en las cuales incluían: La fecha, dureza, tasa de producción, consumo de energía eléctrica, consumo de energía calórica, etc.

Para poder manipular los datos usamos la base de datos en formato CSV, así como la librería Pandas, esta sirve para manipular conjuntos de datos de una manera muy eficiente, permite operaciones de creación, lectura, modificación, eliminación, búsqueda, conteo, etc. [pandas development team, 2020]

2.2. Descripción de los datos

Q	Time	D	Tasa_Prod
Calidad	Timestamp	Dureza del acero	Tasa de producción

EC	EE	Asp
Energía combustible	Energía eléctrica	Se desconoce el contexto de la variable

Figura 1: Tabla 1. Se describen las variables que representan nuestros registros de datos.[Creación propia, 2021]

- El campo de calidad es un valor normalizado que se encuentra entre 0 y 1
- El campo time es una estampa de tiempo del momento en el que se capturaron los datos
- El campo D nos dice la dureza del acero usado, sus valores van de 80 a 112
- El campo Tasa_Prod indica la tasa de producción necesaria, sus valores van de 0 a 480
- No se conoce que representa la variable Asp

2.3. Exploración de los datos

Al momento de importar los datos, se observan 9389 entradas, de las cuales solo existe un dato nulo, el cual se encuentra en la columna de dureza. Todos los datos que se espera que sean numéricos lo son, no es necesario convertirlos, la única columna que requiere un cambio es la de Time, podemos convertir este

campo a date-time sin problema. Lo anterior nos refleja que los datos tienen bastante calidad y facilita la limpieza de los mismos.

2.4. Gestión de la calidad de datos

Los datos parecen encontrarse en buen estado. Para tener un conjunto de datos de calidad solo es necesario: transformar el campo Time a su formato correcto, se elimina el campo Asp y se eliminan los registros con valores nulos, además de la normalización de datos eliminando valores atípicos. Estas tareas son parte de la sección 3.2.

2.5. Artículos relacionados con el tema

2.5.1. Optimización aplicada a sistemas de bombeo

Sabogal Abril, B. R. , Palacios Peñaranda, J. A., & Pantoja Tovar, C. L. realizaron un proyecto similar para optimizar el uso de energía eléctrica en sistemas de bombeo, en este analizaron el impacto de usar frecuencias distintas con el objetivo de encontrar un modelo que les permitiera disminuir el uso de energía eléctrica. Tras recolectar una variedad de datos en distintas observaciones concluyeron que era posible un ahorro operando en un punto óptimo reduciendo la potencia necesaria mediante variaciones de velocidad. [Sabogal Abril et al., 2013]

2.5.2. Estimación de potencial fotovoltaico

En cuatro ciudades de Colombia utilizan la información recopilada en Bogotá, Cúcuta, Manizales y Pasto para con el empleo del software MATLAB se sometan los datos a dos algoritmos de comparación : K-means y Fuzzy C-means y uno de visualización: Análisis de componentes principales (PCA) todo esto con el fin de determinar la factibilidad de la implementación de las micro-redes. [Torres-Pinzón et al., 2019]

2.5.3. Predicción de costos de PYMES

Un algoritmo se encarga de predecir los costos del consumo eléctrico de las micro, pequeñas y medianas empresas del sector de alimentos. Utilizando el perfil de consumo y tarifa eléctrica para reproducir los gastos y escenarios del cambio de tarifa. De esta forma, se facilitó reconocer la tarifa más eficiente para cada una de las empresas considerando además la retroalimentación del usuario. Este es un ejemplo de cómo la metodología lleva a la productividad y competitividad de las empresas.[Cuisano et al., 2020]

2.5.4. Optimización de recursos hídricos

En el artículo se especifica que la Provincia de Mendoza emplea herramientas digitales para optimizar el uso del agua proyectando además, con la utilización de una máquina de Inteligencia Artificial, el impacto económico que provoca esta acción en su matriz productiva. [Cavaller Riva and Ortega Yubro, 2020]

2.5.5. Agricultura de precisión

Expone la propuesta de una red de sensores que es capaz de optimizar una plantación de café a través de una red sensorial con distintos nodos comunicados entre si usados para conocer los estados de temperatura, humedad, y radiación solar, y otras características de los cultivos, de esta manera se cuida el uso de recursos en la plantación teniendo información más profunda sobre su estado actual, ayudando a reducir costos. [Urbano-Molano, 2013]

3. Preparación de los datos

3.1. Establecer el universo de datos con los que trabajar

Como se mencionó en la sección 2.4, la única variable que debe ser eliminada es la de Asp, el resto de campos son útiles para el objetivo. En cuanto a los datos atípicos, se considerarán atípicos todos los datos que estén alejados más de 2 desviaciones estándar de la media.

3.2. Limpieza de datos

- Para eliminar los campos innecesarios usamos la función drop
- Para limpiar los datos vacíos se usa la función dropna
- Para limpiar los datos atípico se usa la función Zscore de la librería Scipy [Virtanen et al., 2020], para calcular el valor z de cada entrada por separado, y eliminar los valores cuyo valor z sea mayor a 2. Al usar la prueba Z, se están normalizando los datos sin necesidad de cambiar sus magnitudes

```
import pandas as pd

import scipy as sp

df=df.drop('Asp', axis = 1)

df.dropna()

sp.stats.zscore(df.<campo>))
```

3.3. Construir un juego de datos apto para ser usado para la modelación

Se busca determinar si existen correlaciones entre los datos, para esto se usa una matriz de correlaciones, la cual devuelve el coeficiente de correlación de Pearson de cada variable con otra. Usando un heatmap obtenemos una representación mas visual de estas correlaciones. En el libro *Handbook of Biological Statatistics* se menciona que este coeficiente funciona bien con variables que siguen una distribución normal, pero aún así es tan robusto que no se ve afectado en gran medida si hay perdida de normalidad[McDonald, 2009].

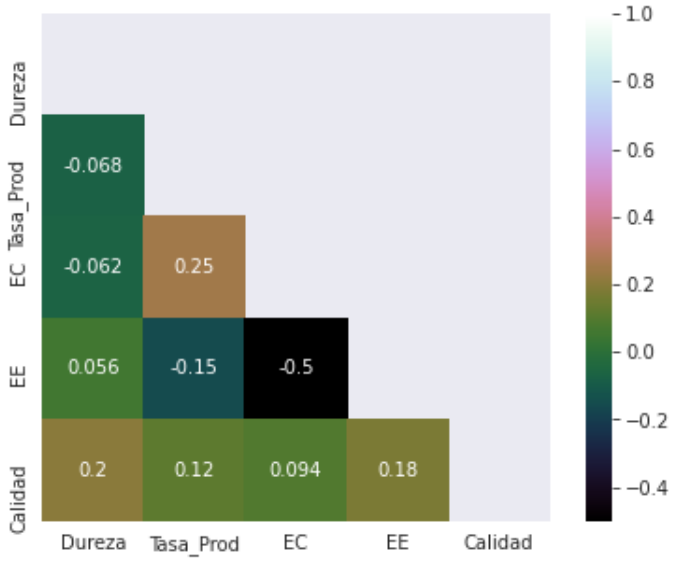

Figura 2: Heatmap de matriz de correlaciones con diagonal principal y simetría suprimida [Creación propia, 2021]

La distribución que siguen las variables es la siguiente:

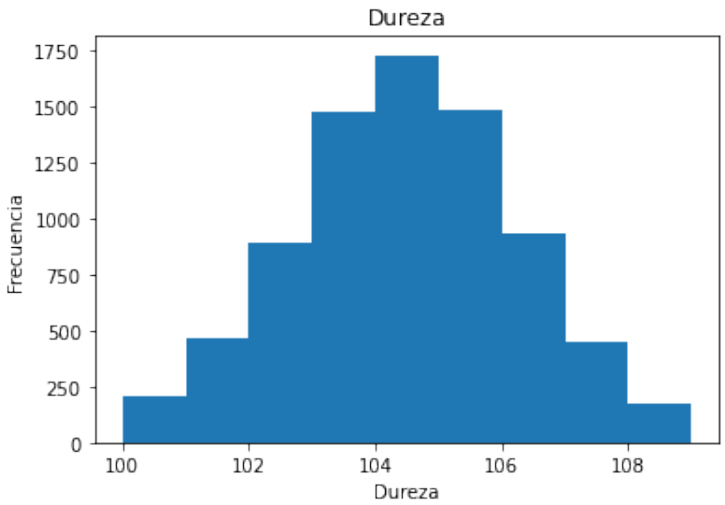

Figura 3: Distribución de valores de dureza [Creación propia, 2021]

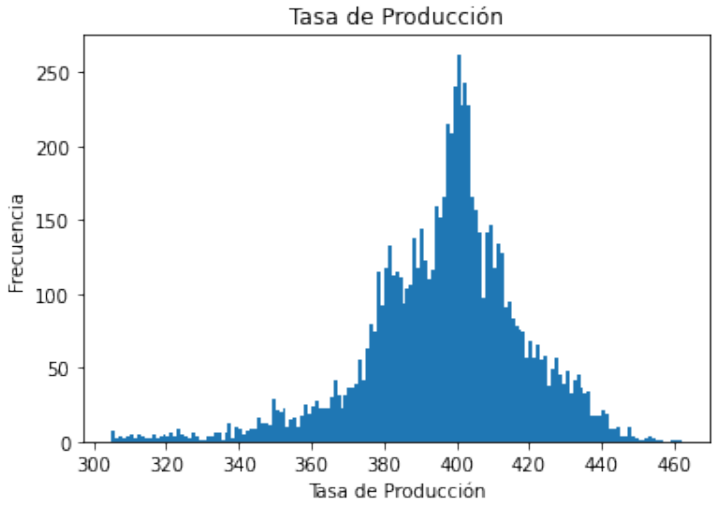

Figura 4: Distribución de valores de tasa de producción [Creación propia, 2021]

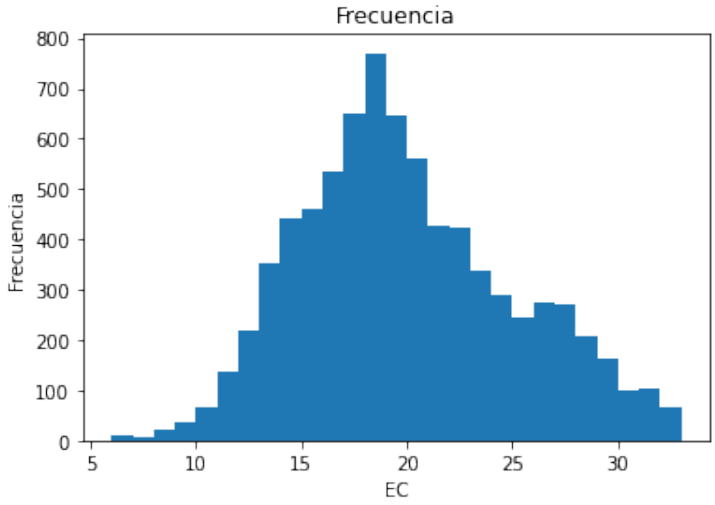

Figura 5: Distribución de valores de energía combustible [Creación propia, 2021]

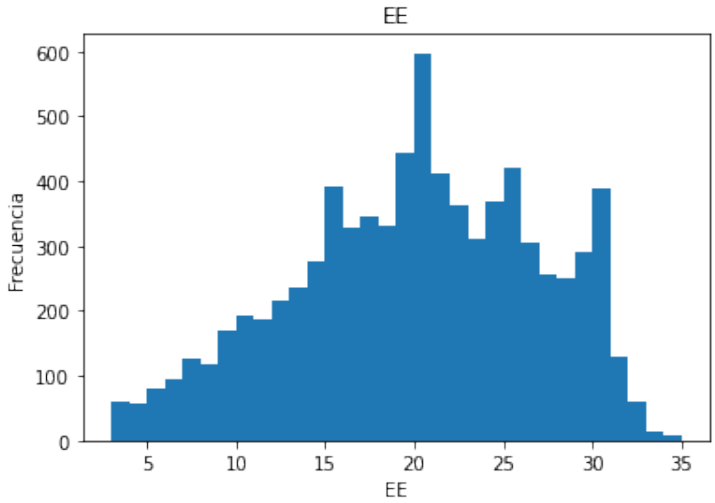

Figura 6: Distribución de valores de energía eléctrica [Creación propia, 2021]

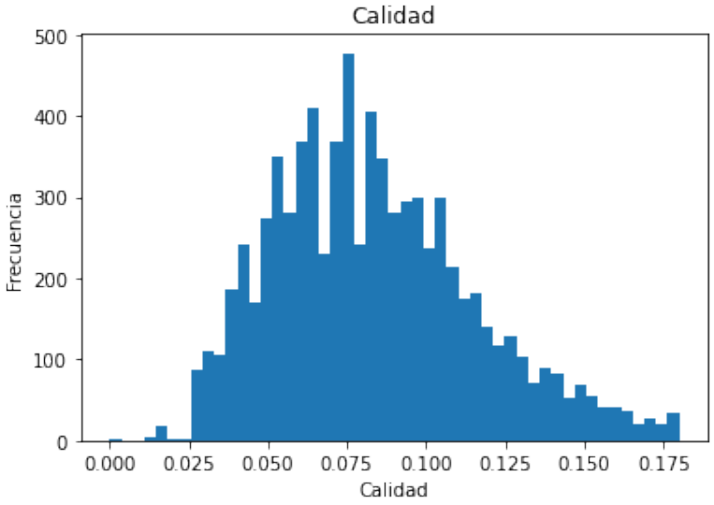

Figura 7: Distribución de valores de calidad [Creación propia, 2021]

Tras ver las distribuciones podemos argumentar que, aunque no es un ajuste perfecto, las variables sí se asemejan a una distribución normal, por lo que podemos confiar en la robustez de el coeficiente de correlación.

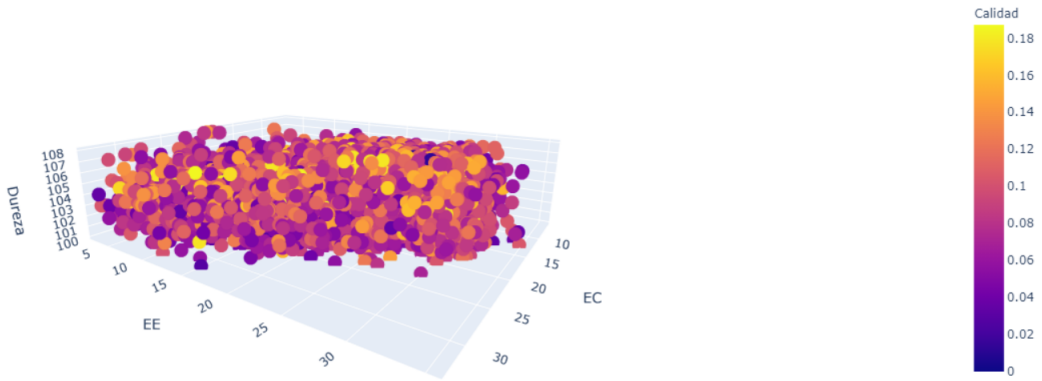

Figura 8: Dispersión EE, EC, Dureza [Creación propia, 2021]

En este punto se cuenta con un buen dataset en cuestión de calidad de datos, sin embargo, no está listo para cumplir el propósito de optimizar el gasto

energético. Para tener una idea de que datos son aptos para esto se calcula el costo y uso ponderado de la energía. Este costo se obtiene de la siguiente manera:

$$CostoPonderado_i = \frac{(EC_i)(0.724) + EE_i}{Tasa\_prod_i} \tag{1}$$

Donde i es la $i$esima entrada.

$$EEPonderdo_i = \frac{EE_i}{Tasa\_prod_i} \tag{2}$$

Donde i es la $i$esima entrada.

$$ECPonderdo_i = \frac{EC_i}{Tasa\_prod_i} \tag{3}$$

Donde i es la $i$esima entrada.

Una vez que se han obtenido los costos ponderados usando la ecuación 1, definimos el conjunto de datos final usando las entradas cuyo costo ponderado es menor o igual a 0.075. De esta manera el modelo será construido con los datos que tienen una relación mejor entre costo y tasa de producción en términos monetarios.

Los datos de entrenamiento y prueba se obtuvieron usando la función train_test_split de la librería Sklearn [Pedregosa et al., 2011]. Se usó la relación estándar 80 % datos de entrenamiento y 20 % datos de prueba. Al momento de separar los datos también se eliminó la columna de TIME pues es irrelevante para el modelado.

4. Modelación y evaluación de datos

4.1. Seleccionar las técnicas de modelado más adecuadas para nuestro juego de datos y nuestros objetivos

Derivado a la naturaleza de las predicciones, se espera obtener como resultado un modelo con múltiples salidas. Por ello, se eligieron los modelos de K-Nearest Neighbors y Linear Regression con Multi-Output. Para K-Nearest Neighbors se consideraron sus principales ventajas:

- No asume información sobre los datos.

  En la mayoría de los modelos se hacen suposiciones sobre los datos y sobre el tipo de distribución que estos llevan, sin embargo KNN evita este inconveniente.[Bonner, 2018]

- Se entrena en el momento de la predicción.

  KNN crea las clases que van a determinar la predicción simultáneamente al realizar esta, gracias a ello se pueden afectar nuevos datos sin que creen un desbalance.[Bonner, 2018]

- Fácil implementación para multivariables.

  Esto representa una ventaja pues el modelo va a tener tres variables de entrada. [Bonner, 2018]

Por su parte, la regresión lineal fue elegida para comparar los comportamientos de las predicciones utilizando un modelo lineal y un no-lineal.

4.2. Fijar una estrategia de verificación de la calidad del modelo

La primera métrica que se empleará es $R^2_{ajustada}$.Según el libro *Handbook of Applied Muiltivariate Statistics and Mathematical Modeling*, esta es una métrica que permite evaluar fácilmente distintos modelos de regresión que cuentan con un número distinto de estimadores. A diferencia de la R^2 simple, $R^2_{ajustada}$ agrega una penalización por cada estimador usado, este parte de la idea de que usar más estimadores no necesariamente es mejor [Tinsley Howard and Brown Steven, 2014], es por esto que, en nuestro caso, probablemente $R^2_{ajustada}$ sea bajo, pero eso no convierte al modelo en inservible. En la Figura 9 se observan sus componentes:

$$R^2_{ajust} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

Figura 9: Fórmula de $R^2_{ajustada}$.[Melillanca, 2018]

La segunda métrica por usar será el Error Absoluto Medio, en el libro *Machine Learning for Developers* se menciona que la métrica *Mean absolute error*, computa el error medio absoluto, que corresponde a el valor esperado de la perdida del error, es decir, cuanto se espera que el modelo se equivoque en cada predicción [Bonnin, 2018]. En la Figura 10 se observan los componentes:

$$\text{MAE} = \frac{\text{SAE}}{N} = \frac{\sum_{i=1}^{N} \left| x_i - \hat{x}_i \right|}{N}$$

Figura 10: Fórmula de Error Absoluto Medio.[Pro, 2016]

4.3. Construir un modelo a partir de la aplicación de las técnicas seleccionadas sobre el juego de datos

Primero se se separaron los datos en aquellos que servirían para la optimización, con este propósito se construyó una gráfica comparativa entre la Calidad y el Costo Ponderado. Se puede observar en la Figura 11

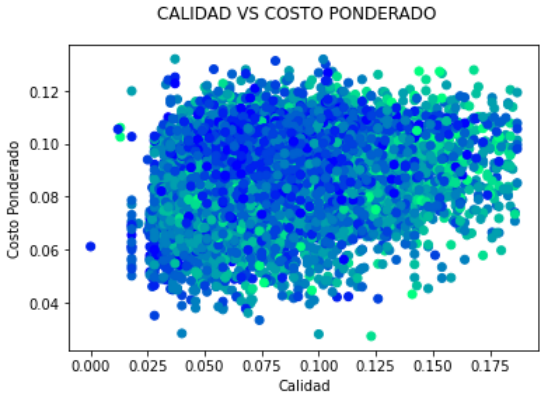

Figura 11: Representación de la relación entre costo ponderado y calidad. Distribución semejante a la normal con concentración de los datos en el centro.[Creación propia, 2021]

Tomando en cuenta la limpieza de calidad que se generó en los pasos anteriores, los puntos que se encuentran aquí cumplen con la calidad que se espera del modelo, derivado de esto, hace falta limitar el costo ponderado para así obtener los puntos óptimos para ello se seleccionaran tres cotas superiores para el costo ponderado formando tres subconjuntos de datos sobre los cuales aplicar los modelos.

4.3.1. Primer Conjunto de Datos

El primer grupo de variables que se formará será tomando como cota superior el valor de Costo Ponderado a 0.075. Es decir, elegiremos aquellos datos con un precio menor o igual a 0.075. Los datos se observan en la Figura 12:

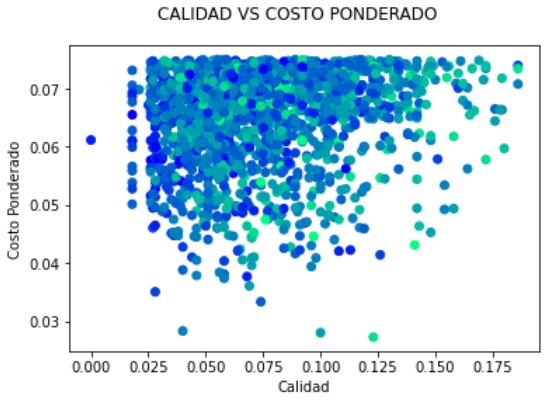

Figura 12: Datos que cumplen con las características de calidad con un costo menor o igual 0.075. [Creación propia].

Se eligió ese umbral ya que logra mantener un costo bajo, uno de nuestros objetivo, sin limitar excesivamente la cantidad de los datos evitando así el peligroso Overfitting.

Después de determinar el conjunto de datos sobre el que se trabajará, se procedió a repararlos en subconjuntos de entrenamiento (training) y prueba (test). Los valores que se utilizaron fueron 80 % y 20 %, así se aseguró una cantidad considerable para realizar el ajuste del modelo. Además, se seleccionaron aleatoriamente, con el fin de mantener una independencia entre los datos. Esta será también la manera en la que separaremos los siguiente dos subconjuntos.

Primero se aplicó el modelo K Nearest Neighbors. Se comenzó por seleccionar un K adecuada para el modelo. El valor K representa el número de valores cercanos que se eligen para determinar el promedio que al final será el valor asignado a la variable siendo predicha. El valor K óptimo depende de cada regresión, en general se utiliza aquel que comienza a dar un Error Medio Cuadrático Estable. Se advierte que al elegir un valor K bajo se puede caer en el Overfitting, en caso contrario al utilizar un valor K alto la modelación puede ser más acertada pero los valores en los extremos están muy indeterminados.[Miller, 2019]

Implementamos el valor K=12 pues es en ese momento cuando la curva en la Figura 13 comienza a volverse constante, además consideramos que es un número apropiado para evitar el Overfitting tomando en cuenta la cantidad de datos que tenemos.

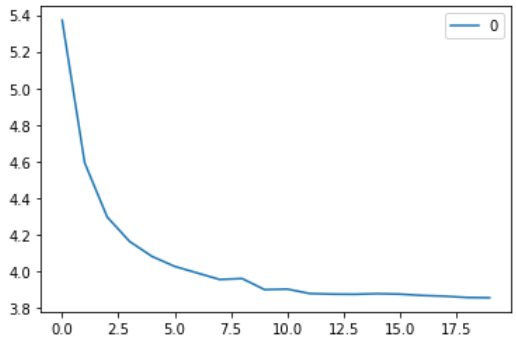

Figura 13: Error Cuadrático Medio, gráfica para identificar el valor óptimo de K. [Creación propia, 2021]

Al aplicar el modelo podemos obtener algunas predicciones como las de la Figura 14:

```
Predicción de EE:  15.133333333333333
Predicción de EC:  14.841666666666667
Predicción Costo Ponderado:  0.06355870562556387
```

Figura 14: Predicciones creadas por el modelo de KNN en el primer conjunto de datos.[Creación propia, 2021]

Y las gráficas que comparan los valores de los valor de Energía Eléctrica, Energía Combustible y Costo Ponderado predichos (de prueba) y los reales son:

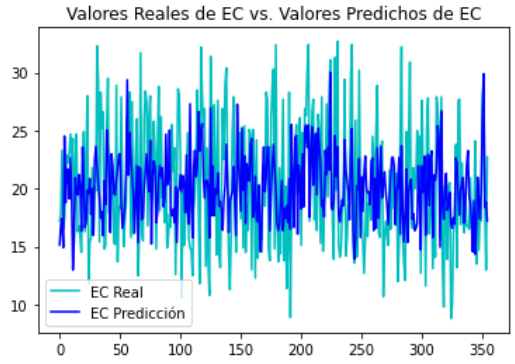

Figura 15: Comparación de los valores de EC vs. los predichos. Modelo KNN primer conjuntos de datos.[Creación propia, 2021]

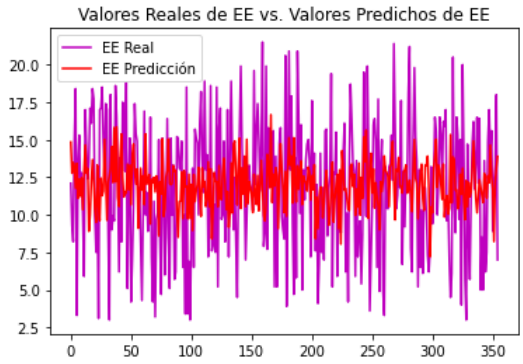

Figura 16: Comparación de los valores de EE vs. los predichos. Modelo KNN primer conjunto de datos.[Creación propia, 2021]

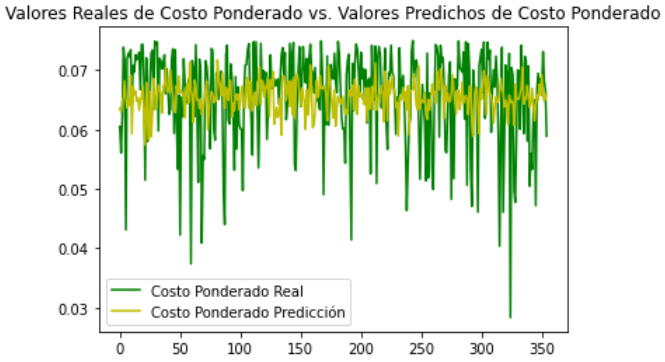

Figura 17: Comparación de los valores de Costo Ponderado vs. los predichos. Modelo KNN primer conjunto de datos.[Creación propia, 2021]

En las Figuras 15, 16 y 17 se busca observar que las variables realmente tengan un comportamiento semejante al de las originales, pues aunque no sean iguales los valores de lo real y lo predicho, mientras se comporten similarmente la predicción puede ser muy buena o aceptable. Como se puede observar en las figuras anteriores, los valores sí siguen un mismo patrón de comportamiento. Los ascensos y descensos tienen una ubicación similar para las gráficas de los valores reales y los predichos.

El segundo regresor será Multi Output Regressor con la Regresión Lineal. Se utiliza la librería Multi Output para obtener diversas variables de salida, sin embargo el concepto sigue siendo igual al de una Regresión Lineal Multivariada, donde en general se busca minimizar el error cuadrado.

Su aplicación es más sencilla pues únicamente se debe entrenar el modelo con

los subconjuntos de entrenamiento ya anteriormente seleccionados. Un ejemplo de las predicciones realizadas por el modelo son:

```
Predicción de EE:  19.227992518229268
Predicción de EC:  12.349155132746276
Predicción Costo Ponderado:  0.06472335764794526
```

Figura 18: Predicciones creadas por el modelo de Regresión Lineal en el primer conjunto de datos.[Creación propia, 2021]

Como se puede observar en la Figura 18 , no tiene mucho sentido que el Costo Ponderado sea tan bajo, apenas un poco más alto que el de KNN, mientras que la energía EE, aquella que implica un mayor gato, es tan elevada. Por ello podemos cuestionar la validez del modelo.

Las gráficas que comparan los valores reales y predichos son:

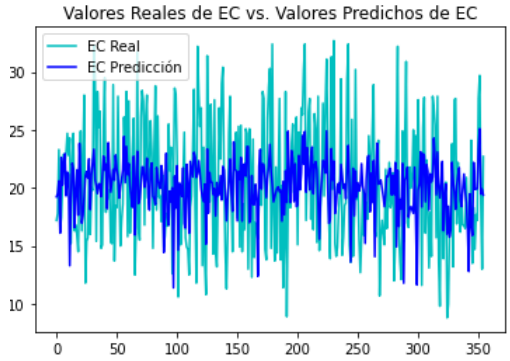

Figura 19: Comparación de los valores de EC vs. los predichos. Modelo Regresión Lineal primer conjunto de datos.[Creación propia, 2021]

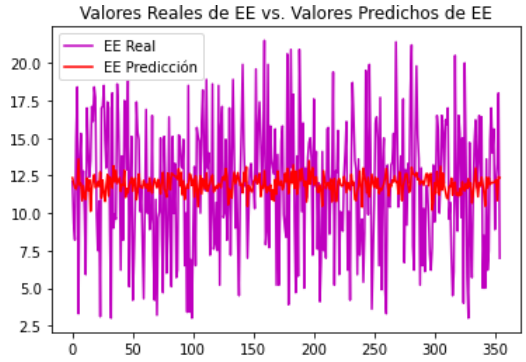

Figura 20: Comparación de los valores de EE vs. los predichos. Modelo Regresión Lineal primer conjunto de datos.[Creación propia, 2021]

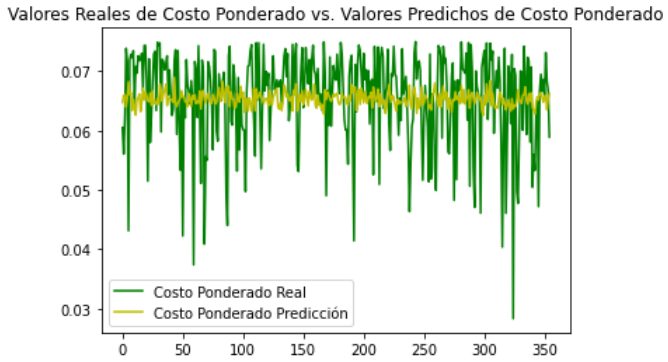

Figura 21: Comparación de los valores de Costo Ponderado vs. los predichos. Modelo Regresión Lineal primer conjunto de datos.[Creación propia, 2021]

Como se puede observar las Figuras 19, 20 y 20, aunque siguen un cierto patrón similar, lo cierto es que los valores predichos tienen un mayor número de depresiones que no suceden en los datos originales, de esta manera se determina que el modelo no sigue el mismo comportamiento de los datos originales.

4.3.2. Segundo Conjunto de Datos

En este caso tomamos un precio máximo de 0.06, como se observa en la Figura 22 esto puede limitar notablemente nuestra cantidad de datos, pero también representa una disminución buscada en el precio.

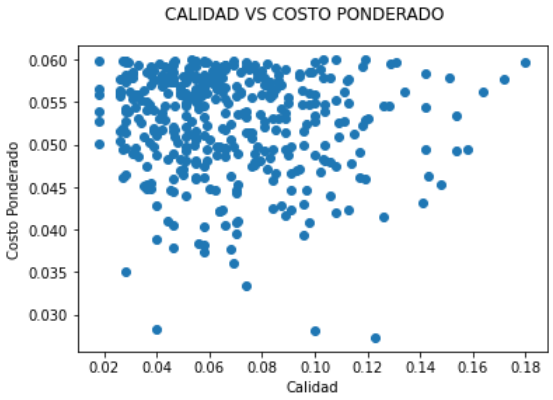

Figura 22: Datos que cumplen con las características de calidad con un costo menor o igual 0.06. [Creación propia].

Para aplicar el modelo K Nearest Neighbors se mantuvo el mismo procedimiento descrito anteriormente:

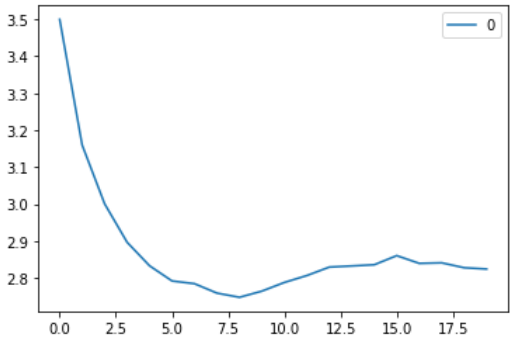

Figura 23: Error Cuadrático Medio, gráfica para identificar el valor óptimo de K. [Creación propia, 2021]

Se conserva la K=12 porque también en este caso es cuando la gráfica comienza a tomar valores constantes (Figura 23).

Se aplica el modelo y se obtienen los siguientes resultados:

Como se puede observar en la Figura 24 la combinación es mucho menos fiable que la del modelo anterior pues al tener una Energía Eléctrica tan elevada, es poco probable tener un Costo Ponderado tan bajo como 0.05.

```
Predicción de EE:  19.225
Predicción de EC:  6.866666666666667
Predicción Costo Ponderado:  0.05175854013082692
```

Figura 24: Predicciones creadas por el modelo de KNN en el segundo conjunto de datos.[Creación propia, 2021]

Las gráficas que nos ayudan a visualizar el comportamiento del modelo son:

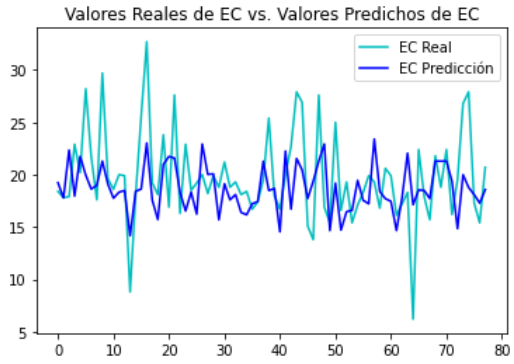

Figura 25: Comparación de los valores de EC vs. los predichos. Modelo KNN el segundo conjunto de datos.[Creación propia, 2021]

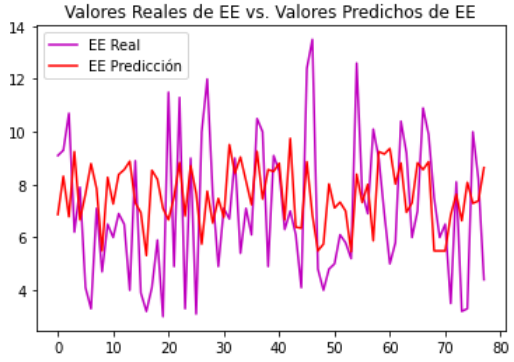

Figura 26: Comparación de los valores de EE vs. los predichos. Modelo KNN segundo conjunto de datos.[Creación propia, 2021]

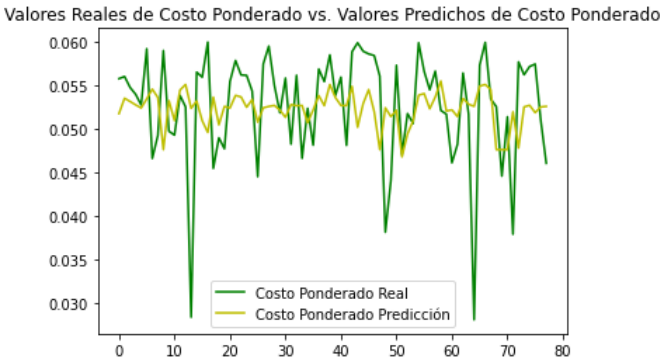

Figura 27: Comparación de los valores de Costo Ponderado vs. los predichos. Modelo KNN segundo conjunto de datos.[Creación propia, 2021]

Se observa en las Figuras 25, 26 y 27 que en este caso los datos siguen un patrón similar al de los datos originales, sin embargo, el modelo de KNN con el primer conjunto de datos continúa manteniendo un comportamiento todavía más similar.

Se procede con el modelo de regresión lineal que arroja las predicciones (Figura 28):

```
Predicción de EE:  18.743034033938464
Predicción de EC:  7.596704719767507
Predicción Costo Ponderado:  0.05263737806668102
```

Figura 28: Predicciones creadas por el modelo de Regresión Lineal en el segundo conjunto de datos.[Creación propia, 2021]

Los resultados son muy semejantes a los del modelo aplicado en el primer conjunto de datos, lo que indica que no es un modelo muy confiable.

Para analizar su comportamiento se presentan las Figuras 29, 30 y 31 :

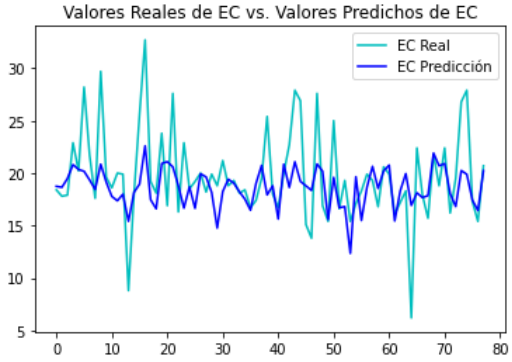

Figura 29: Comparación de los valores de EC vs. los predichos. Modelo Regresión Lineal segundo conjuntos de datos.[Creación propia, 2021]

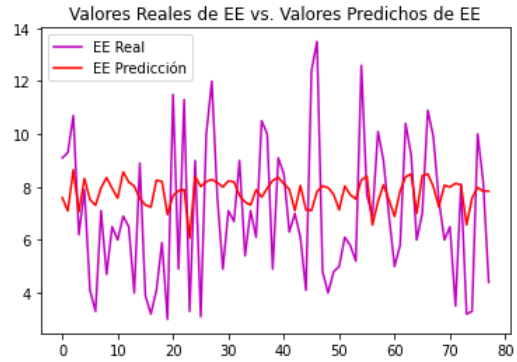

Figura 30: Comparación de los valores de EE vs. los predichos. Modelo Regresión Lineal segundo conjunto de datos.[Creación propia, 2021]

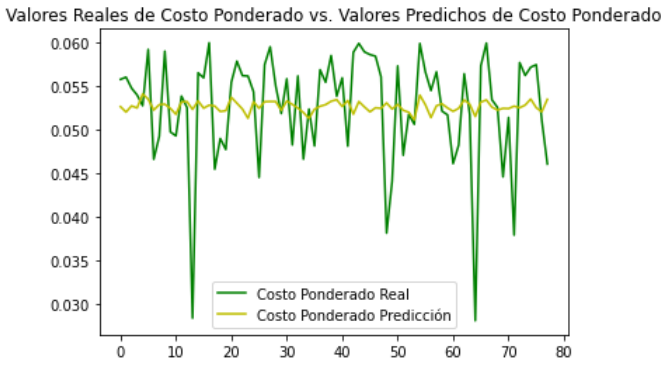

Figura 31: Comparación de los valores de Costo Ponderado vs. los predichos. Modelo Regresión Lineal segundo conjunto de datos.[Creación propia, 2021]

El patrón de comportamiento es todavía más inexacto pues este sigue una línea casi horizontal cuando el modelo original se compone por muchas crestas y depresiones.

4.3.3. Tercer Conjunto de Datos

Se redujo incluso más el grupo de datos, de forma que se logre ser más estrictos con la reducción de costos, sin embargo al mantener un conjunto tan pequeño, es mucho el riesgo de que el algoritmo no tenga suficiente capacidad para aprender correctamente los patrones.

El umbral que se utilizó fue de 0.055, como se observa en la Figura 32:

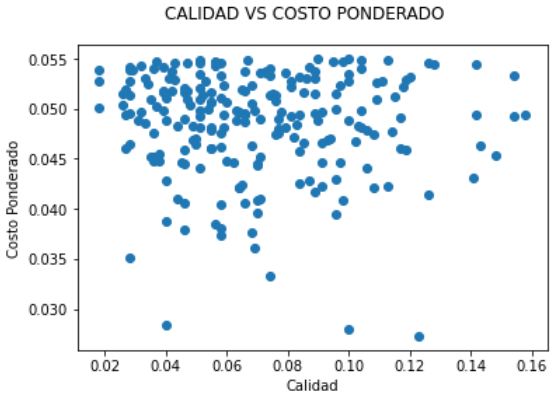

Figura 32: Datos que cumplen con las características de calidad con un costo menor o igual 0.055. [Creación propia].

Para aplicar KNN se busca la K:

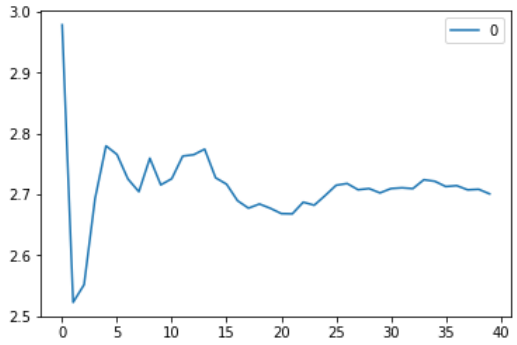

Figura 33: Error Cuadrático Medio, gráfica para identificar el valor óptimo de K. [Creación propia, 2021]

Se observa en la Figura 33 un comportamiento errático que dificulta la identificación del punto constante. Sin embargo, de acuerdo con nuestro criterio, el punto óptimo es K=20, pues la gráfica deja de tener cambios tan bruscos.

```
Predicción de EE:  19.32
Predicción de EC:  6.74
Predicción Costo Ponderado:  0.04886595868950783
```

Figura 34: Predicciones creadas por el modelo de KNN en el tercer conjunto de datos.[Creación propia, 2021]

Los valores predichos en la Figura 34 son todavía más improbables pues implican un nivel igual de elevado de la Energía Eléctrica con un Costo Ponderado mucho más bajo.

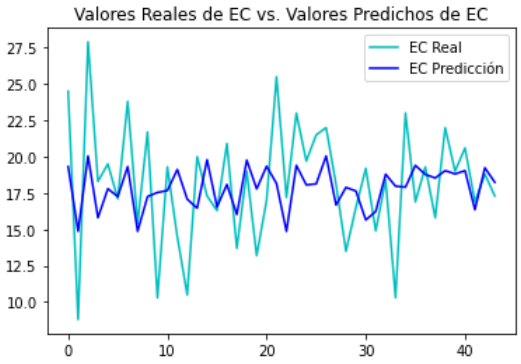

Figura 35: Comparación de los valores de EC vs. los predichos. Modelo KNN el tercer conjuntos de datos.[Creación propia, 2021]

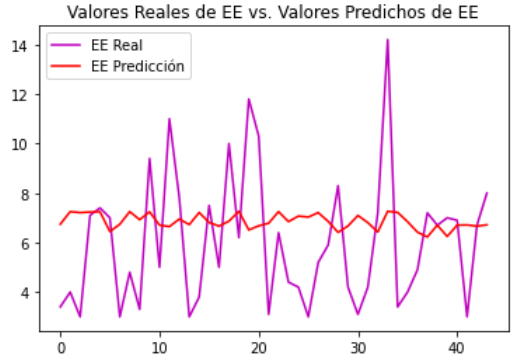

Figura 36: Comparación de los valores de EE vs. los predichos. Modelo KNN tercer conjunto de datos.[Creación propia, 2021]

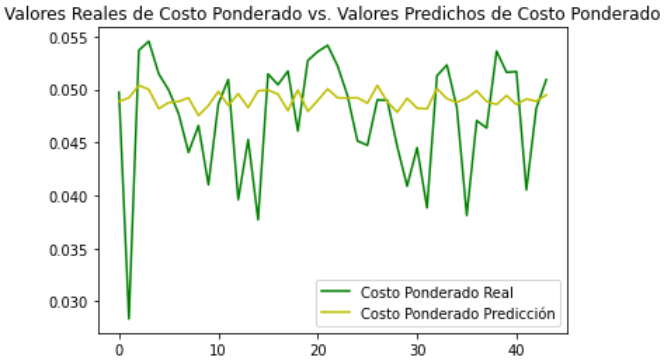

Figura 37: Comparación de los valores de Costo Ponderado vs. los predichos. Modelo KNN tercer conjunto de datos.[Creación propia, 2021]

Mientras que la Figura 35 asemeja ligeramente el comportamiento, el resto de las Figuras 36 y 37 distan mucho de tener unos valores predichos comparables con los valores reales.

En cuanto al modelo de Regresión Lineal, los valores obtenidos son (Figura 38):

```
Predicción de EE:  19.16439115729687
Predicción de EC:  6.780812010557291
Predicción Costo Ponderado:  0.04863947426992653
```

Figura 38: Predicciones creadas por el modelo de Regresión Lineal en el tercer conjunto de datos.[Creación propia, 2021]

Semejantes a los del modelo anterior, no tienen sentido por la misma razón de lo poco probable que es obtener un Costo Ponderado tan pequeño con ese uso de Energía Eléctrica.

Sus gráficas son las Figuras 39, 40 y 41:

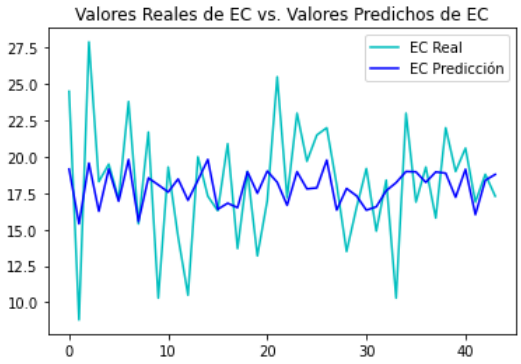

Figura 39: Comparación de los valores de EC vs. los predichos. Modelo Regresión Lineal tercer conjuntos de datos.[Creación propia, 2021]

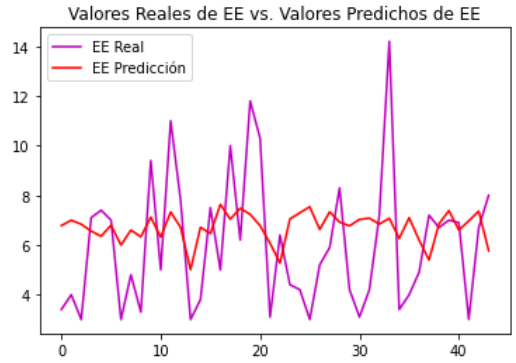

Figura 40: Comparación de los valores de EE vs. los predichos. Modelo Regresión Lineal tercer conjunto de datos.[Creación propia, 2021]

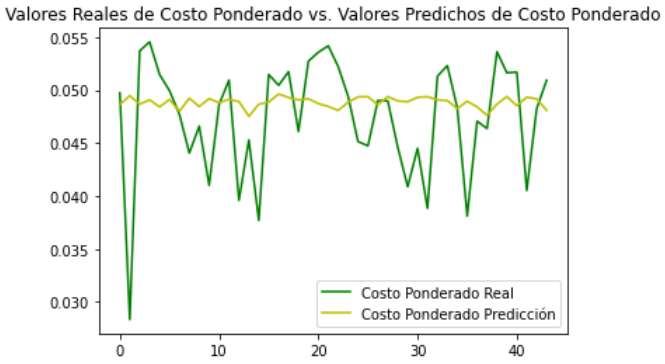

Figura 41: Comparación de los valores de Costo Ponderado vs. los predichos. Modelo Regresión Lineal tercer conjunto de datos.[Creación propia, 2021]

La Figura 39 sí mantiene el mismo comportamiento de los datos reales pero el resto de las figuras se desvían en gran medida.

4.4. Ajustar el modelo evaluando su fiabilidad y su impacto en los objetivos anteriormente establecidos

Utilizando las métricas anteriormente establecidas, evaluamos los modelos.

4.4.1. Primer conjunto de datos

Para el conjunto con una cota de 0.075 los resultados son del modelo KNN y la Regresión Lineal se encuentran en las Figuras 42 y 43 respectivamente:

```
R^2 Training: 0.1583054489271918
R^2 Testing: 0.03661177350306458
Mean absolute error training: 2.528914100737197
Mean absolute error testing: 2.61888625740626
```

Figura 42: El valor de $R^2_{ajustada}$ y el Error Absoluto Medio en modelo KNN primer conjunto. [Creación propia, 2021]

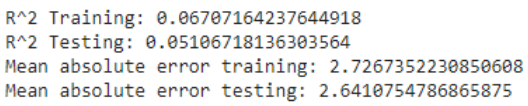

Figura 43: El valor de $R^2_{ajustada}$ y el Error Absoluto Medio en modelo Regresión Lineal primer conjunto. [Creación propia, 2021]

Una $R^2_{ajustada}$ tan pequeña es probablemente resultado de incluir tantos parámetros en la predicción. El Error Absoluto Medio comparado con la dimensión de los datos, está en un rango aceptable.

4.4.2. Segundo conjunto de datos

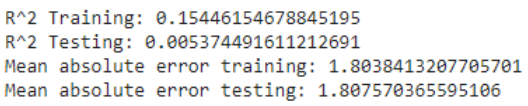

Figura 44: El valor de $R^2_{ajustada}$ y el Error Absoluto Medio en modelo KNN segundo conjunto. [Creación propia, 2021]

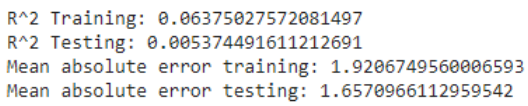

Figura 45: El valor de $R^2_{ajustada}$ y el Error Absoluto Medio en modelo Regresión Lineal segundo conjunto. [Creación propia, 2021]

Los valores del Error Absoluto Medio (Figuras 44 y 45) disminuyeron, sin embargo la $R^2_{ajustada}$ también disminuyó muy drásticamente.

4.4.3. Tercer conjunto de datos

En estos modelos ambas $R^2_{ajustada}$ (Figuras 46 y 47) son inferiores, en la Figura 46 incluso llega a ser negativa. Los Errores Medios Absolutos son semejantes a los del segundo conjunto.

El modelo de K Nearest Neighbors con K=12 evaluado en el primer conjunto de datos con cota superior de 0.075 es aquel que resulta óptimo pues tiene una $R^2_{ajustada}$ más elevada en comparación al resto y su Error Absoluto Medio entra en un rango aceptable tomando en cuenta todos los valores en los modelos.

```
R^2 Training: 0.08410945244332439
R^2 Testing: -0.009098740004624776
Mean absolute error training: 1.4281621005584582
Mean absolute error testing: 1.770233421732929
```

Figura 46: El valor de $R^2_{ajustada}$ y el Error Absoluto Medio en modelo KNN tercer conjunto. [Creación propia, 2021]

```
R^2 Training: 0.0809770688037218
R^2 Testing: 0.01649996582338568
Mean absolute error training: 1.4256433476271908
Mean absolute error testing: 1.7281166653031368
```

Figura 47: El valor de $R^2_{ajustada}$ y el Error Absoluto Medio en modelo Regresión Lineal tercer conjunto. [Creación propia, 2021]

4.4.4. Escalar Variables

El modelo más óptimo se escalará usando MinMaxScaler (Escalado de Variables), una función que se encarga de escalar las variables de manera que los nuevos valores se encuentren entre el rango dado, en este caso 0 y 1, pero conservando las proporciones. Se utiliza esta función para transformar las variables de entrada. La fórmula (Figura 48) que representa lo que ocurre es:

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Figura 48: Fórmula de Escalado de Variables.[Morante, 2018]

Se eligió este método pues no utiliza ni la desviación estándar ni la media, a diferencia del Escalado Estándar, evitando así cualquier sensibilidad que estás puedan tener. Es importante no utilizar MinMaxScaler cuando los valores son muy estables, sin embargo en este caso nuestras predicciones se caracterizan por sus altas caídas y subidas derivando en que no hay inconveniente al emplear este método.[Morante, 2018]

Es muy importante al utilizar machine learning escalar, pues hay algoritmos que tienen sensibilidad a los valores mayores, tal es el caso de KNN que, al basarse principalmente en distancias, puede cambiar la percepción de lo que es lejano y cercano al estar escaladas las variables de entrada. [Chaos, 2020]

Escalando los valores de entrada y aplicando el modelo obtenemos (Figura 49): En cuanto a su métrica (Figura 50):

```
Predicción de EE:  16.659999999999997
Predicción de EC:  15.187999999999999
Predicción Costo Ponderado:  0.06735419382290082
```

Figura 49: Valores predichos por el modelo escalado.[Creación propia, 2021]

```
R^2 Training: 0.1395926096876927
R^2 Testing: 0.04719966548506475
Mean absolute error training: 2.5906595251128373
Mean absolute error testing: 2.5915249946159413
```

Figura 50: Métricas del modelo escalado.[Creación propia, 2021]

Dado a que no hubo un cambio significativo en los valores que predijo el modelo y en cambio hubo un decremento en la $R^2_{ajustada}$, por ello se considera que es mejor conservar el modelo original.

5. Evaluación

5.1. Evaluar el modelo o modelos generados hasta el momento

Como se menciona en la sección 4.1, el conjunto de datos que mejores resultados entrega es el primero, el cual toma costos ponderados menores o iguales a 0.075. Usando la relación train-test definida en la sección 3.3, se evaluaron los modelos *K nearest neighbors* y *Multioutput linear regression*. Tras hacer esta evaluación se determinó que el modelo que mejores resultados entrega es *K nearest neighbors*, cuyo parámetro k se fijó como 12 (ver figura 13).

Usando las métricas de Sklearn, podemos evaluar fácilmente este modelo usando $R^2_{ajustada}$. Como se mencionó antes, $R^2_{ajustada}$ agrega una penalización por cada estimador usado, es por esto que, en nuestro caso, usando 3 estimadores, el valor de $R^2_{ajustada}$ sea bajo, esto no necesariamente significa que sea un mal modelo, al contrario, las figuras15, 16 y 17 nos muestran que, aunque la magnitud no es exacta, el modelo simula de manera correcta el comportamiento de los datos reales, por lo que podemos argumentar que en efecto es un modelo representativo para los datos de esta población. Los resultados de $R^2_{ajustada}$ obtenidos son los siguientes:

Por su parte, al computar el Error Absoluto Medio obtuvimos lo siguiente:

Esta métrica es sencilla de comprender pero hay que entender su contexto, estamos hablando de una métrica que nos indica que tanto se equivocan nuestras predicciones, pero no lo hace por separado (es decir por cada variable de salida), por lo que un valor alto podría ser preocupante si se hablara de costo ponderado, pues como se observa en la figura 17, sus valores son pequeños. Dicho esto, no

```
R^2 Training: 0.1583054489271918
R^2 Testing: 0.03661177350306458
```

Figura 51: Resultados de $R^2_{ajustada}$ obtenidos con el modelo *K nearest neighbors* con el primer conjunto de datos.[Creación propia, 2021]

```
Mean absolute error training: 2.528914100737197
Mean absolute error testing: 2.61888625740626
```

Figura 52: Resultados de *Mean absolute error* obtenidos con el modelo *K nearest neighbors* con el primer conjunto de datos.[Creación propia, 2021]

es el caso, si observamos con atención las figuras 15, 16 y 17 podemos observar que las 3 variables contribuyen de manera más o menos equitativa a este error, es decir, el *MAE* no está siendo sobrecargado por una sola variable con un error muy grande, sino que está distribuido en pequeños errores entre las 3 variables.

Todo lo anterior nos lleva a concluir que en efecto, aunque las métricas obtenidas no son sorprendentes, sin duda se ha logrado un buen modelo, y podemos entender las fuentes de error que se tienen.

5.2. Revisar todo el proceso de minería de datos que nos ha llevado hasta este punto

En la sección 1.3, se definieron los objetivos de minería de datos del proyecto. Llegada esta etapa podemos decir que se han cumplido con todos los puntos de la Metodología CRISP-DM. No se ha presentado ningún inconveniente para llevar a cabo los métodos y el análisis de los resultados. Se cuenta con un modelo multi entrada y multi salida que entrega resultados satisfactorios y ha sido evaluado con las métricas pertinentes. Además, las correlaciones entre variables fueron comprendidas y evaluadas en la sección 4.3.

5.3. Siguientes pasos

A pesar de que estamos satisfechos con los resultados obtenidos, hay que recordar que que este es un trabajo de investigación cuya mayor limitante era el tiempo, existen muchos próximos pasos que pueden darse para expandir el alcance de la investigación. A continuación se muestran 2 propuestas que darían una buena continuidad a este proyecto.

5.3.1. Conectar el modelo a un ecosistema MapReduce en Hadoop para obtener datos de entornos de producción en tiempo real

Una vez que se ha consolidado un modelo funcional que cumple con las expectativas del cliente, se puede conectar a un entorno de Hadoop data streaming [AWS documentation, NDb], de esta manera se pueden procesar cantidades enormes de información de una manera eficiente, sencilla, escalable y barata (gracias a las ventajas que ofrece la computación en la nube). Conforme se tengan más datos disponibles, se comprenderá de una mejor manera como se comportan y de esta manera se conseguirá un modelo que irá mejorando con el tiempo, y que eventualmente cumplirá la función de optimizar los costos de la mejor manera posible, sin reducir la calidad.

5.3.2. Usar Hyperparameter Tuning para obtener el modelo más óptimo posible, para su posterior implementación en producción

Si bien es cierto que se ha obtenido un buen modelo, y que la intervención humana puede ayudar a crear buenos modelos, la realidad es que hay límites en cuanto a el número de modelos que podemos probar. Es por eso que usar la técnica conocida como Hyperparameter tuning suena como una muy buena opción.

La documentación de Amazon Web Services menciona la búsqueda Bayesiana, esta funciona como una regresión. Consiste en optimizar un modelo para la métrica de nuestra elección adivinando que combinaciones de parametros son mas propensas a brindar mejores resultados, para posteriormente correr tareas de entrenamiento y prueba, para verificar los resultados obtenidos. [AWS documentation, NDa]

Esta técnica no requiere muchos recursos, basta con crear una instancia de *AWS Sage maker* para correr la optimización Bayesiana. Aunque es posible que esta tarde un tiempo considerable en encontrar los parámetros óptimos, considerando que las instancias de sage maker son económicas pues se cobran bajo el modelo *Paga por lo que usas*, la realidad es que valdrá la pena la inversión, sobre todo por que solo se pagaría por cada vez que se desee correr esta optimización.

6. Conclusión general

Dada la naturaleza de los datos y la cantidad de parámetros, es normal que el modelo no sea completamente exacto, sin embargo lo esencial es que sigue el mismo comportamiento que los datos originales, por ello podemos concluir que es un modelo funcional para optimizar el gasto de la energía eléctrica y el gasto de la energía combustible. Aumentar la $R^2_ajustada$ o disminuir el Error Absoluto Medio, lo convertiría en un modelo ideal sin embargo es posible que

sea necesario recabar más datos o utilizar menos parámetros pero tengan una correlación más alta con las variables de salida para lograrlo.

7. Bibliografía

Referencias

[AWS documentation, NDa] AWS documentation (N.Da). How hyperparameter tuning works. https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-how-it-works.html.

[AWS documentation, NDb] AWS documentation (N.Db). Process data with streaming. https://docs.aws.amazon.com/emr/latest/ReleaseGuide/UseCase_Streaming.html.

[Bonner, 2018] Bonner, L. (2018). K-nearest neighbors. https://www.fromthegenesis.com/pros-and-cons-of-k-nearest-neighbors/.

[Bonnin, 2018] Bonnin, R. (2018). *2.8.1.1 Mean Absolute Error.* Packt Publishing.

[Cavaller Riva and Ortega Yubro, 2020] Cavaller Riva, D. and Ortega Yubro, C. D. (2020). Data science in water resource optimization science of data on the water resource optimization.

[CEMEX, 2019] CEMEX (2019). Acerca de nosotros. https://www.cemexmexico.com/acerca-de-cemex.

[Chaos, 2020] Chaos, I. (2020). Escalado de datos. https://www.interactivechaos.com/es/manual/tutorial-de-machine-learning/escalado-de-datos.

[Cuisano et al., 2020] Cuisano, J. C., Chirinos, L. R., and Barrantes, E. J. (2020). Eficiencia energética en sistemas eléctricos de micro, pequeñas y medianas empresas del sector de alimentos. simulación para optimizar costos de consumo de energía eléctrica. *Información tecnológica*, 31(2):267–276.

[McDonald, 2009] McDonald, J. H. (2009). *Handbook of biological statistics*, volume 2. sparky house publishing Baltimore, MD.

[Melillanca, 2018] Melillanca, E. (2018). Coeficiente de determinación corregido o r-cuadrado ajustado. http://ericmelillanca.cl/content/coeficiente-determinaci-n-corregido-o-r-cuadrado-ajustado.

[Miller, 2019] Miller, M. (2019). The basics: Knn for classification and regression. https://towardsdatascience.com/the-basics-knn-for-classification-and-regression-c1e8a6c955.

[Morante, 2018] Morante, S. (2018). Precauciones a la hora de normalizar datos en data science. https://empresas.blogthinkbig.com/precauciones-la-hora-de-normalizar/.

[pandas development team, 2020] pandas development team, T. (2020). pandas-dev/pandas: Pandas.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Pro, 2016] Pro, N. (2016). Mae - error medio absoluto. https://support.numxl.com/hc/es/articles/215969423-MAE-Error-medio-absoluto.

[Sabogal Abril et al., 2013] Sabogal Abril, B. R., Palacios Peñaranda, J. A., and Pantoja Tovar, C. L. (2013). Optimización de energía en sistemas de bombeo. *Informador Técnico*, 77(1):47.

[Tecnológico de Monterrey, 2021] Tecnológico de Monterrey (2021). Optimización de escenarios en producción. https://experiencia21.tec.mx/courses/138974/pages/reto.

[Tinsley Howard and Brown Steven, 2014] Tinsley Howard, E. and Brown Steven, D. (2014). *6.1.1.4 Multiple R-Squared, Adjusted R-Squared, and Shrunken R-Squared.* Elsevier.

[Torres-Pinzón et al., 2019] Torres-Pinzón, C. A., Forero-García, E. F., et al. (2019). Estimación del potencial fotovoltaico mediante minería de datos en cuatro ciudades de colombia. *TecnoLógicas*, 22(46):77–97.

[Urbano-Molano, 2013] Urbano-Molano, F. A. (2013). Redes de sensores inalámbricos aplicadas a optimización en agricultura de precisión para cultivos de café en colombia. *Journal de Ciencia e Ingeniería*, 5(1):46–52.

[Virtanen et al., 2020] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.