INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY

Inteligencia artificial avanzada para la ciencia de datos I (Gpo 103)

“Entregable 3”

Jose Pablo Cobos Austria **A01274631**

Profesor:

Dr. Benjamín Valdés Aguirre

Santiago de Querétaro, 12 de septiembre del 2022

Selección de Implementación

La implementación seleccionada de machine learning fue la que se utilizan frameworks, ya que en lo personal considera que tiene un mejor desempeño que la que hice de forma manual.

A continuación explicaré un poco sobre la información que contiene nuestro dataframe que se utilizó en la práctica:

El conjunto de datos contiene 9568 puntos de datos recopilados de una central eléctrica de ciclo combinado durante 6 años (2006-2011), cuando la planta estaba configurada para funcionar a plena carga.
Las características consisten en variables ambientales promedio por hora:

- Temperatura (T)
- Presión ambiental (AP)
- Humedad relativa (RH)
- Vacío de escape (V)

Que se utilizaron para poder predecir la producción de energía eléctrica neta por hora (EP) de la planta.

Ya sabiendo esto, recordando que recientemente estuve teniendo prácticas en una fábrica, me dio mucha curiosidad cómo es que estás variables impactan al producir energía y cuáles de ellas son las que tienen mayor impacto en esta producción, sin embargo debido a que son muchas variables, seleccionaremos una relación simple y nos preguntaremos ¿Que tanto la temperatura impacta en la producción de energía ?

Datos Obtenidos

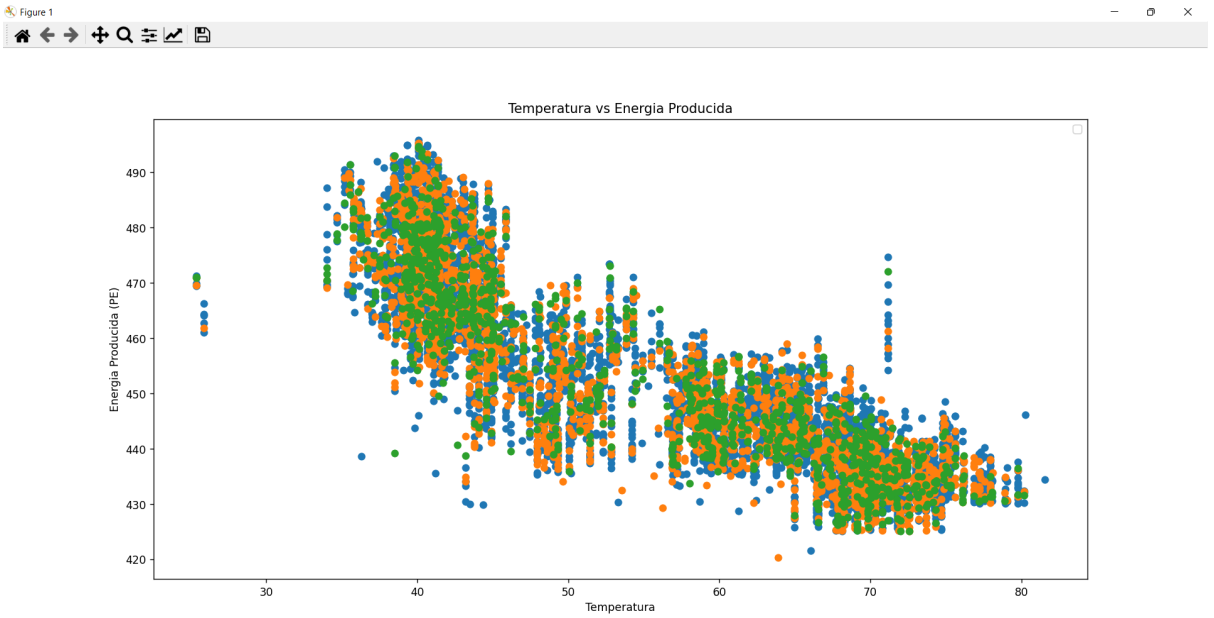
AT	V	AP	RH	PE
14.96	41.76	1024.07	73.17	463.26
25.18	62.96	1020.04	59.08	444.37
5.11	39.4	1012.16	92.14	488.56
20.86	57.32	1010.24	76.64	446.48

Split del modelo en sets y su validación

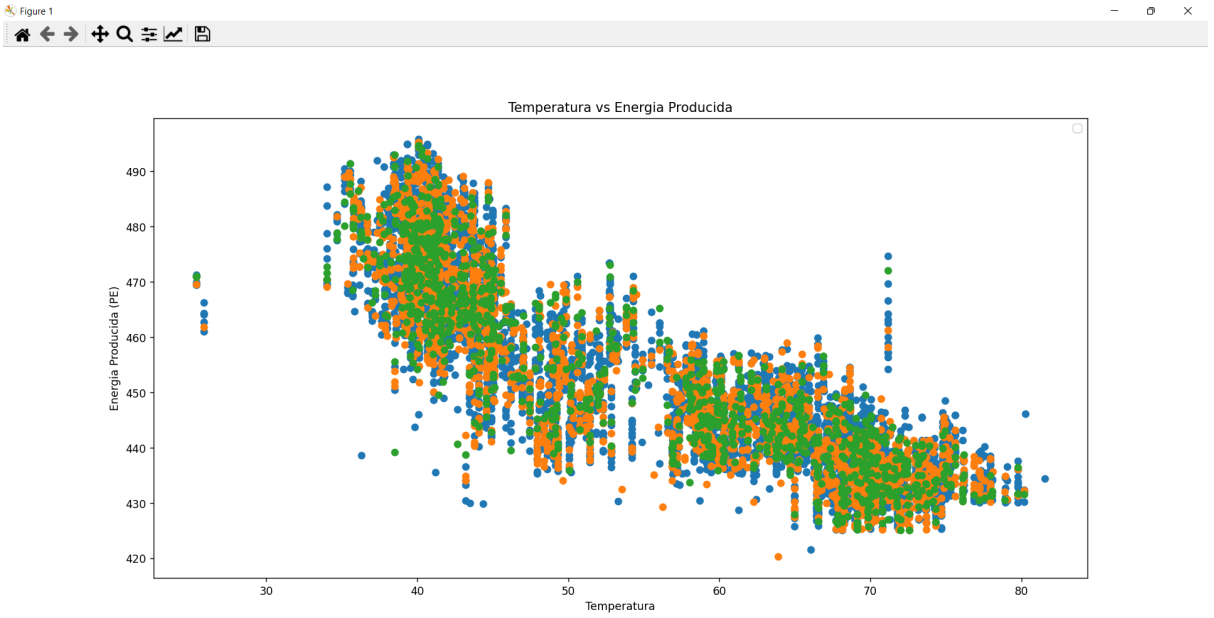
Para la evaluación del modelo, lo que se aplicó fueron conocimiento visto en clase, donde se seleccionaba una parte de parte de los datos, se le hacía split, una porcentaje para entrenar y otro porcentaje para poder probar que todo estuviera funcionando correctamente.

Para eso haremos uso de las herramientas de que nos presta la librería de sklearn, y haremos el siguiente split: 60% de los datos para pruebas, 15% para validación y 25% para entrenar nuestro modelo y los resultados fueron los siguientes:

- Datos importantes:
  Verde = valid set
  Naranja = test set
  Azul = train set



Gráfica 1. Comparación 1 entre cada uno de los sets realizados



Gráfica 2. Comparación 2 entre cada uno de los sets realizados

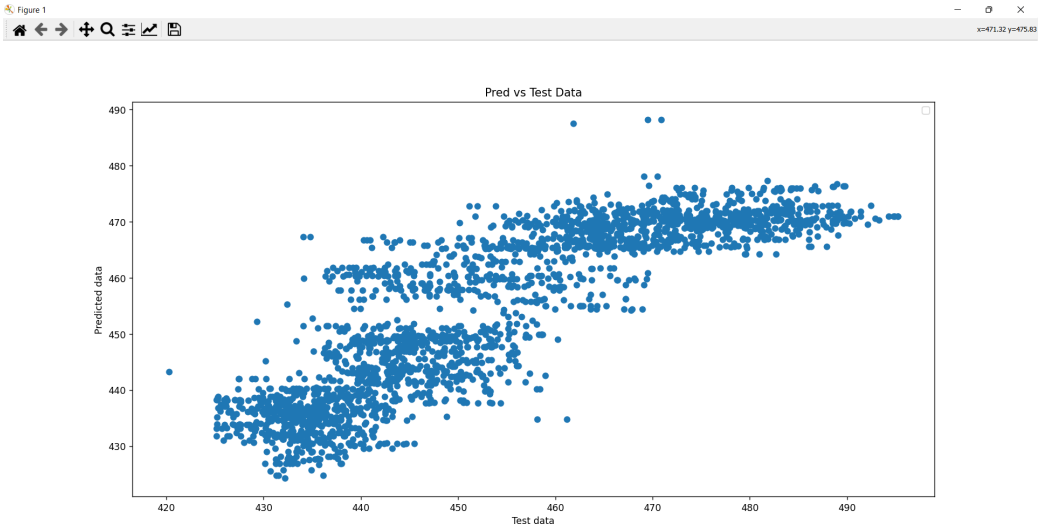
Diagnóstico

Tras haber realizado el primer análisis, vamos a observar que existe una relación entre las dos variables propuestas, incluso tiene un comportamiento extraño donde en un momento va bajando los valores de la variable dependiente conforme lo hacen los independiente y en esas bajadas de repente sube y vuelve a caer. Tomando esto en cuenta el siguiente paso a realizar puedes ir entrando a todo modelo y probando si se puede mejorar su eficiencia.

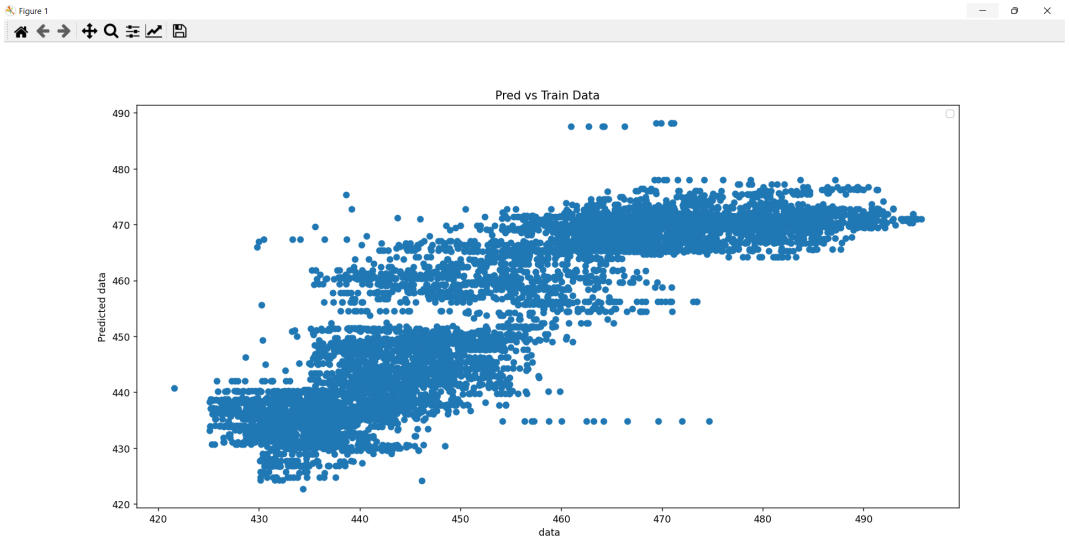
Se entrenó un modelo y se realizó la comparación entre ambos sets, tanto electrónico como el de test para poder medir la predicción que tenía con cada uno de estos.
:
Los resultados fueron los siguientes:
MSE de los datos train = 70.02
R^2 de los datos train = .758021

MSE de los datos test = 73.58
R^2 de los datos test = .75207

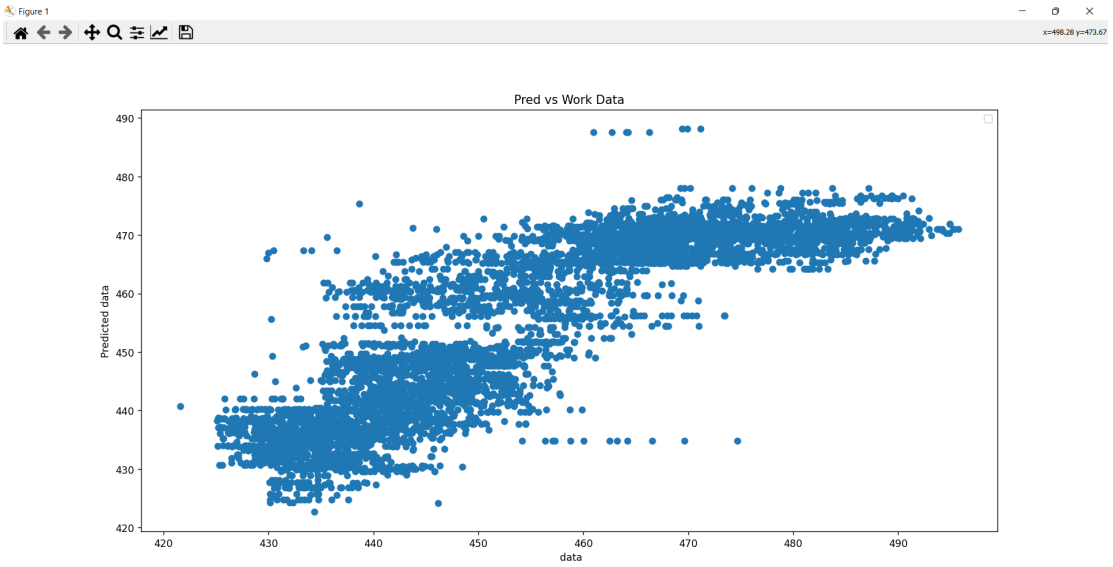
Esto nos dice que aunque su coeficiente de determinación es bastante similar con ambos test, puedes decir qué es cuestión de decimales la diferencia al momento de utilizar otro parámetro que es el MSE nos podemos dar cuenta que nuestra regresión posee un MSE muy grande que nos indica que fallas en el modelo.Y con el fin de poder comprender de mejor manera los parámetros se realiza una gráfica donde se explican los mismos.



Gráfica 3. Comparación entre los datos predecidos y los reales del test set



Gráfica 4. Comparación entre los datos predecidos y los reales del train set



Gráfica 5. Comparación entre los datos predecidos y los reales del train set

- # Grado de bias o sesgo:

- El bias es la precisión que llega a tener nuestro modelo para poder predecir valores que se acercan a los valores reales de nuestro dataset. Conociendo esto y habiendo un alisado de las gráficas anteriores, podemos decir que el bias es alto en nuestro modelo. Esto se puede decir por varios factores, uno de ellos porque simplemente es una regresión lineal y la manera en la que realiza la predicción puede tener errores con lo que se vio en las gráficas bastante grandes en comparación a otros modelos cómo le podría hacer un árbol de decisión por ejemplo. Además, otro. importante a entender es que con datos nuevos, le va a costar mucho tener una precisión mayor.

Resultados de las pruebas de precisión

```
R^2 con datos test = 0.7520706983009586
MSE con datos test = 73.58514960998848
R^2 con datos train = 0.7580216863798119
MSE con datos train = 70.02833775984391
R^2 con datos validation = 0.7562818366119196
MSE con datos validation = 70.95446704622196
```

- ## Grado de varianza

- Lo siguiente que toca por analizar sería el grado de varianza, que representa el cambio que hay en los datos, por lo que, en razón de lo mencionado anteriormente, de la manera en la que aprende el modelo, no hay mucho espacio para que las variables cambien mucho, por lo que tiene un grado de varianza bajo

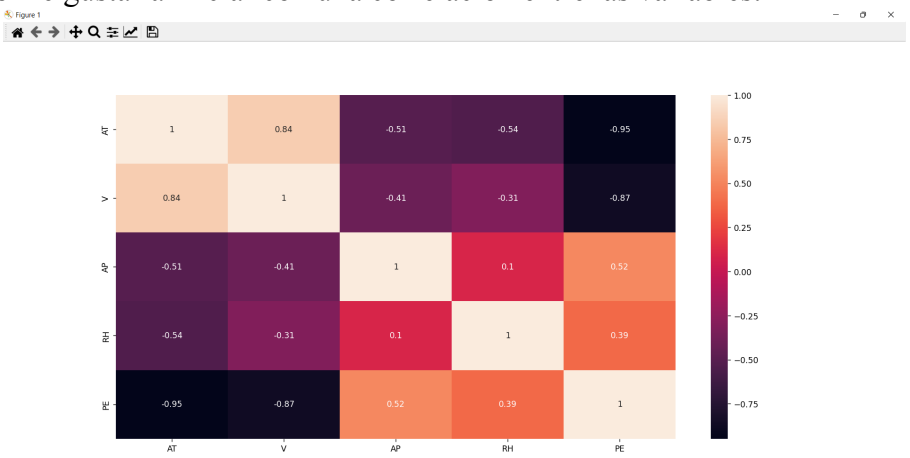
- ## Nivel de ajuste del modelo:

  El modelo presenta underfitting debido a su casi baja precisión con el uso de los datos de entrenamiento y no presentaba alguna mejora significativa al momento de realizar las pruebas correspondientes.

Aplicación de técnicas de regularización o ajuste de parámetros
Mejorar el desempeño de tu modelo

Ya teniendo listo nuestro modelo, habido en el estado de bias, la varianza, dado cuenta que nuestro nivel de ajuste el lo que podemos hacer es analizar con un poco más de detalle cuales son los datos, qué relación tiene entre ellos, y si podemos encontrar otro dato que nos pueda ayudar a tener una mejor precisión.

Entonces me gustaría iniciar con una correlación entre las variables:



Gráfica 6. Gráfica de correlación de las variables del dataframe

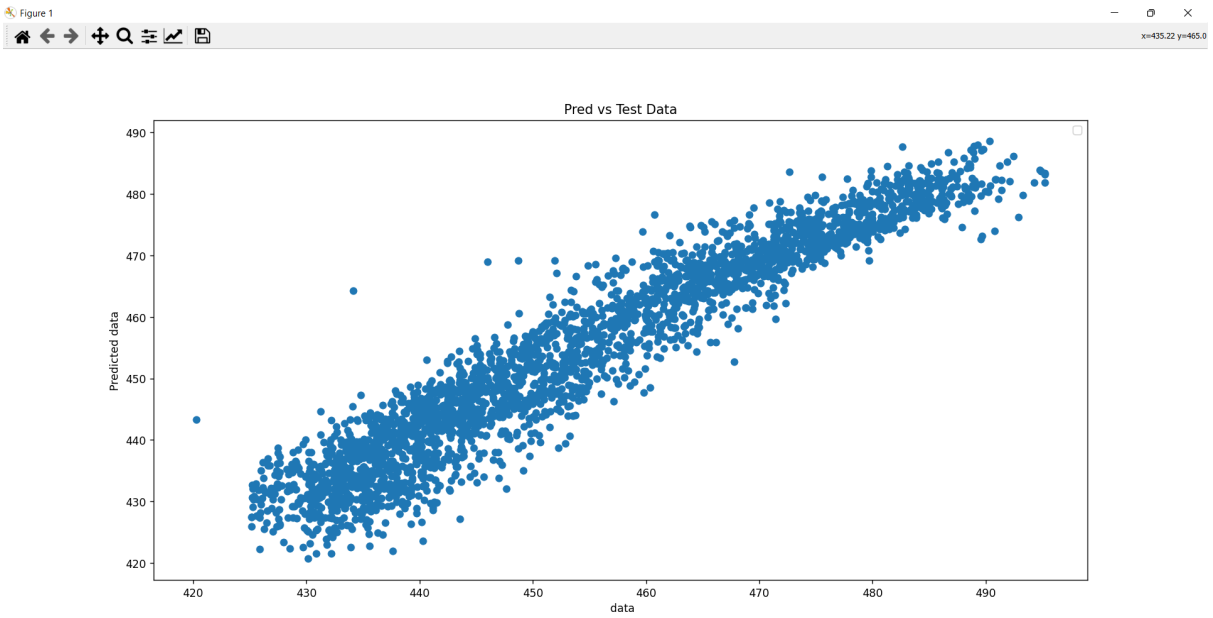
Como podemos observar existe un alto valor de correlación del AT (-.95) y V(-.87) por lo será bueno incluir la otra variable a nuestro modelo para observar si genera alguna mejora en el modelo.

Gracias a que se usó la librería de sklearn y pandas, simplemente tuvimos que seleccionar la otra columna de las variables sin necesidad de hacer muchos cambios en el código.

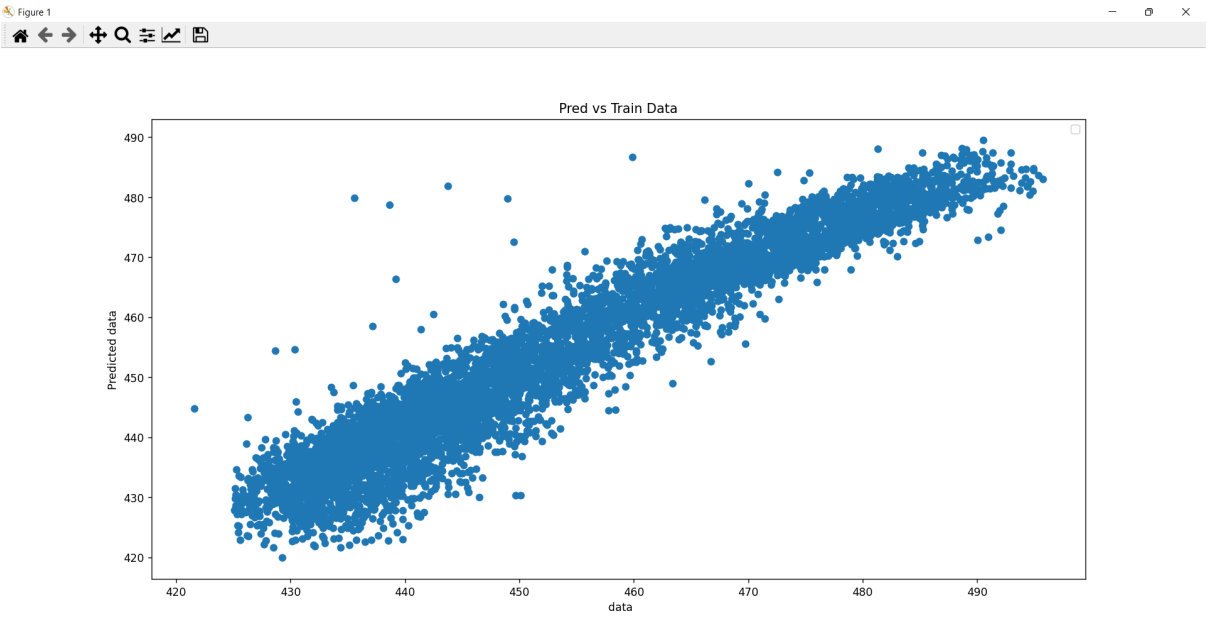
Ahora bien, cómo aumentamos otra variable, ya no podemos graficarla como lo hacíamos anteriormente porque se requeriría de una dimensión extra por lo que se graficaron la comparación entre los datos predecidos y los datos de entrenamiento, de test y de validation, incrementando un poco el porcentaje de los datos de train para ver qué resultados obtenemos

Resultados de implementación de los cambios

```
R^2 con datos test = 0.9168791371397308
MSE con datos test = 24.670182537395075
R^2 con datos train = 0.9153224318371884
MSE con datos train = 24.50562306709479
R^2 con datos validation = 0.9161687379368553
MSE con datos validation = 24.4060698587781 53
```



Gráfica 7. Gráfica de pred vs test data con las mejoras implementadas



Gráfica 8. Gráfica de pred vs test data con las mejoras implementadas

Finalmente como conclusión podemos decir que el agregar una variable más al modelo y aumentar los datos que se tomaban para el training de este mismo de verdad mostraron ser una mejora importante a nuestro modelo de regresión lineal, no obstante todavía tenemos el detalle que el MSE es bastante alto, lo que significa que sería bueno agregar las demás variables para ver cómo cambia nuestro modelo, si genera una mejora en la precisión y puede predecir mejor los datos nuevos.