

# ***Kernel Ridge Regression para predecir IPC y variables climáticas***

*Equipo :*

- Cázares Trejo Leonardo Damián
- Malváez Flores Axel Daniel
- Peralta Rionda Gabriel Zadquiel

## **Introducción**

Este informe tiene como objetivo abordar un análisis integral para poder comprender las relaciones dinámicas entre diversas variables clave en México y poder explorar nuevas variables de impacto que muy posiblemente se están pasando por alto debido a su baja asociación por pertenecer a diferentes áreas de interés. Con datos mensuales que abarcan variables climáticas por estado (lluvia y temperatura), el número de automóviles por estado, el índice de Volúmen Físico (IVF) nacional desglosado por estado y el Índice de Precios y Cotizaciones (IPC), nuestro objetivo específico es realizar análisis de regresiones y evaluar su capacidad productiva mutua entre dichas variables.

La investigación se centra en dos enfoques: primero, determinar si las condiciones climáticas, la movilidad vehicular y el desempeño económico (IVF) pueden prever de manera significativa las variaciones en el IPC nacional. Segundo, evaluar si el IPC, el número de automóviles y el IVF pueden servir como indicadores predictivos para las variables climáticas. Además, se pretende explorar la posibilidad de hacer un análisis social que sugiera una relación entre las emisiones de carbono y el rendimiento financiero. Este enfoque busca analizar si es posible deducir la existencia de una "economía carbonizada" en México.

El análisis integral proporcionará una perspectiva valiosa sobre las interconexiones entre factores económicos y ambientales, ofreciendo *insights* cruciales para entender el impacto económico-ambiental en el contexto mexicano.

## **Metodología**

El método que usamos a lo largo de este proyecto es un algoritmo de regresión, es decir, un algoritmo que toma variables independientes y devuelve valores continuos para una variable de respuesta,

$$\hat{y} = f(x_1, x_2, \dots, x_n),$$

En particular el tipo de modelo que usaremos será uno llamado *Kernel Ridge Regression* (KRR), son varias las ventajas de este modelo sobre el resto de modelos de regresión estándar. La KRR es un algoritmo de regresión no-lineal, el cual usa funciones que

proyectan a los puntos de entrenamiento y de prueba en un espacio de alta dimensionalidad, donde se crean variables latentes que combinan las variables independientes originales, un proceso que *a priori*, sería tanto computacional como conceptualmente difícil de ejecutar por el alto número de posibles combinaciones que sería posible realizar.

Otra de las características de la *KRR* es que logra hacer predicciones sobre nuevos puntos del espacio de características de las variables independientes comparando qué tan similares son los valores de dichas variables independientes con respecto a los datos de entrenamiento, y es que la naturaleza de los kernels consiste en obtener un puntaje entre dos puntos cualesquiera, y en caso de obtener un valor grande como resultado de dicha evaluación, podemos concluir que dichos puntos son similares. Lo que hace el algoritmo de *KRR* es evaluar la similitud de todos los puntos de entrenamiento y construir una matriz de similitudes,

$$[K]_{ij} = \kappa(x_i, x_j)$$

donde la función de kernel puede ser de distinta naturaleza, dependiendo si queremos trabajar con cadenas, grafos, vectores, etc. O bien según queramos aprovechar propiedades de periodicidad, linealidad, etc. Para este trabajo usaremos con el kernel gaussiano,

$$\kappa(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$$

el cual tiene la particularidad de transformar nuestros datos a un espacio de infinita dimensionalidad, donde crea variables latentes con los polinomios de todos los grados sobre nuestras variables independientes, explorando todo el espacio de posibilidades, con sólo calcular distancia entre puntos y una función exponencial, que computacionalmente resulta simple.

Una vez que tenemos la matriz de similitudes es posible entrenar el modelo, resolviendo el sistema de ecuaciones,

$$(K + \lambda I)\vec{\alpha} = (y_1, \dots, y_n)^T$$

donde el término  $\lambda$  corresponde a un factor de regularización,  $\vec{\alpha} = (\alpha_1, \dots, \alpha_n)^T$  es un vector de pesos (un peso por cada punto de entrenamiento) y el vector del lado derecho de la ecuación corresponde a los valores reales (etiquetas) con los que entrenamos al modelo.

Una vez que el modelo fue entrenado, podemos hacer predicciones sobre nuevos puntos de interés, evaluando nuestra función de kernel sobre cada punto de entrenamiento y sobre el punto a predecir, ponderando la suma de las similitudes con ayuda de  $\vec{\alpha} = (\alpha_1, \dots, \alpha_n)^T$ , es decir,

$$\hat{y}(x) = \sum_{i=1}^n \alpha_i \kappa(x, x_i)$$

De manera que al momento de predecir sobre un nuevo punto, obtenemos la similitud respecto al resto de los puntos de entrenamiento, dándole a cada similitud el peso correspondiente a la importancia de su respectivo punto de entrenamiento dentro de la matriz de similitudes.

Por otro lado, una vez explicado el modelo, la estructura a seguir para la implementación de los modelos es con el conjunto de datos correspondiente a la tarea a hacer, separar los datos en un conjunto de entrenamiento y en un conjunto de prueba (con un tamaño de testing del 20% de los datos). Luego se hace una estandarización de los datos tanto para  $X$  como para  $y$ , dicha estandarización se hace restando a los datos su  $\mu$  (media) y su  $\sigma$  (desviación estándar), para los datos de testing se utilizan  $\mu$  y  $\sigma$  de los datos de prueba, así nos aseguramos de no sesgar el modelo.

## Conjuntos de datos

El conjunto de datos se conforma por diversas variables climáticas (lluvia y temperatura en el territorio nacional mexicano), variables económico-financieras (PIB, IPC, IVF), así como también variables constituyentes de datos estadísticos de los automóviles, camiones y autobuses en cada estado del territorio nacional. Se trabajará con los datos de una manera distinta dependiendo de cómo se esté abordando el problema y lo que esperamos predecir. Pues no tenemos la misma cantidad de datos para tratar de predecir una variable financiera, usando variables del clima y de los coches como independientes y viceversa.

En primera instancia, debido a que se desea comenzar con un análisis relación existente entre la variable del IPC con la variable PIB, estas variables se escogieron porque contamos con el IPC mensual, pero no contamos con el PIB mensual, por lo que si conseguimos dar una predicción con los algoritmos de estas variables, la falta de datos con el PIB puede dejar de ser un problema y se pueden dar estimaciones mensuales del PIB. El Índice de Precios y Cotizaciones (IPC) en México es un indicador bursátil que refleja la evolución de los precios de las acciones de las empresas más importantes que cotizan en la Bolsa Mexicana de Valores (BMV). Por otro lado, el Producto Interno Bruto (PIB) es una medida económica que representa el valor total de bienes y servicios producidos en un país durante un período específico, generalmente un trimestre o un año, sin considerar la variación de precios. Para realizar este análisis se contaban con los datos del PIB e IPC de manera trimestral desde 1980 - cuarto trimestre del año hasta el año 2023 - segundo trimestre del año. Por lo que se emparejaron los datos de PIB e IPC con respecto a sus respectivos años.

Por otro lado se desea realizar un análisis predictivo con dos enfoques similares y contrarios a la vez. En primer lugar, deseamos hacer un análisis donde utilizando variables climáticas y de transporte, deseamos poder predecir el valor del IPC. Para esto obtuvimos un conjunto

de datos comprendidos entre Enero de 1991 a Agosto de 2023, es decir 32 años de información mensual. Por otro lado, el segundo análisis requería de utilizar el conjunto de datos anterior concatenado a las variables del PIB mensual, sin embargo aunque no pudimos conseguir el conjunto de datos de PIB mensual, utilizamos el IVF por estado, no obstante debido a la falta de información del IVF en el último semestre del año, es por ello que ahora el conjunto de datos iba desde Enero de 1991 a Junio de 2023.

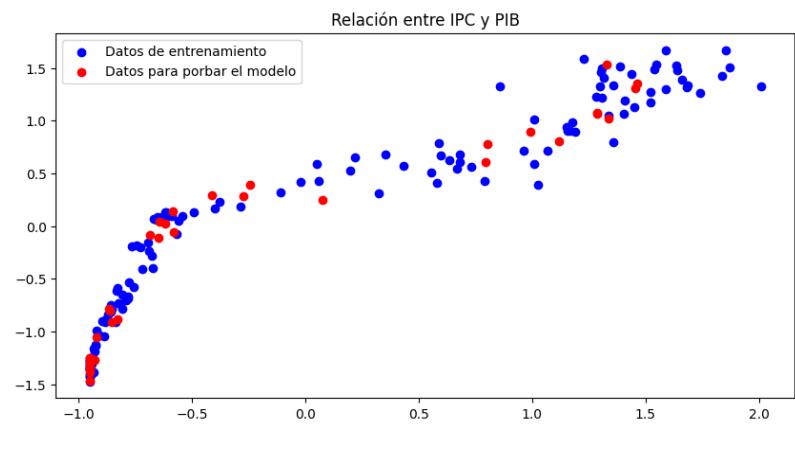
Los datos climáticos referentes a las lluvias y a las temperaturas fueron recolectados de la CONAGUA, mientras que los datos correspondientes al PIB e IPC fueron proporcionados por el profesor.

## Resultados

### PIB vs IPC

El primer análisis realizado fue el de comprobar si el IPC y el PIB (Nacional) eran variables que estaban altamente correlacionadas. Para esto lo que se hizo fue verificar el valor de correlación de pearson, la cuál efectivamente obtuvimos que  $\rho = 0.929$ , donde dicho resultado nos quiere decir que cuando una de las dos medidas aumenta la otra medida también tiende a aumentar y por el contrario, cuando una disminuye la otra también tiende a disminuir. Esta relación sugiere que a medida que la economía crece (PIB aumenta), también tiende a haber un aumento en los precios de bienes y servicios que consumen los hogares (IPC).

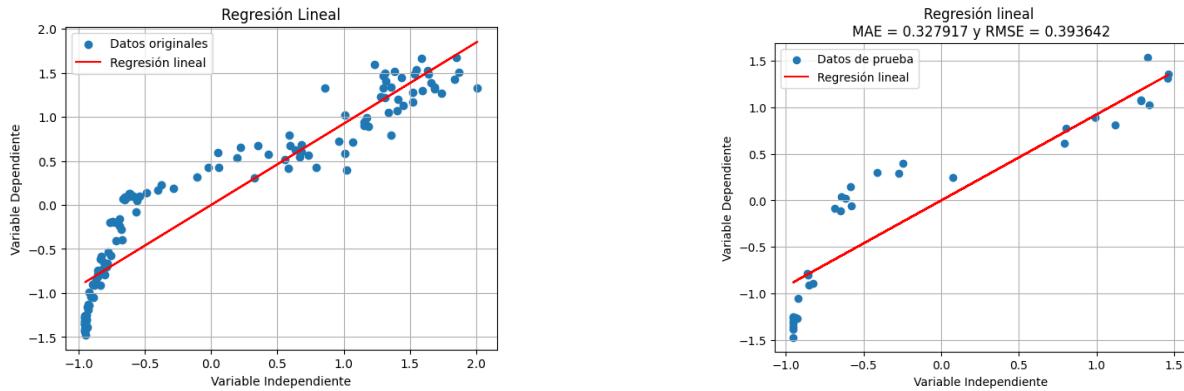
Sin embargo, es importante tener en cuenta que la correlación no implica causalidad. Es decir, el hecho de que estas dos variables estén altamente correlacionadas no significa necesariamente que un aumento en el PIB cause un aumento directo en el IPC o viceversa. Pueden estar influenciadas por otros factores o existir variables intermedias que afecten a ambas.



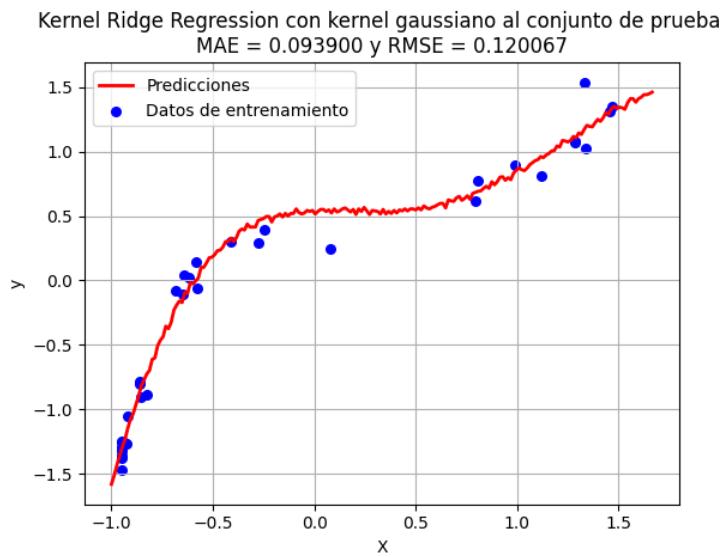
Pero en el caso de nuestro estudio nos concentraremos en generar modelos para ver si se puede dar una aproximación al resultado para poder dar estas predicciones. La relación que existe entre la variable IPC y PIB se encuentra en la Gráfica *Relación entre IPC y PIB*, en esta figura se refleja la correlación positiva que se había mencionado anteriormente.

Por lo que entonces primero probaremos viendo lo que se obtiene aplicando una regresión lineal. Dichos resultados se pueden visualizar en las Gráficas de *Regresión Lineal*. En la

primera se muestra el ajuste lineal que se presenta al conjunto de datos de entrenamiento, por otro lado en la segunda se muestra las predicciones que está realizando el modelo de regresión lineal, donde se está obteniendo que el error cuadrático medio es de .32 y la raíz del error cuadrático medio es de .39, lo cual nos está indicando que la aproximación que nos está presentando la regresión lineal esta bien.

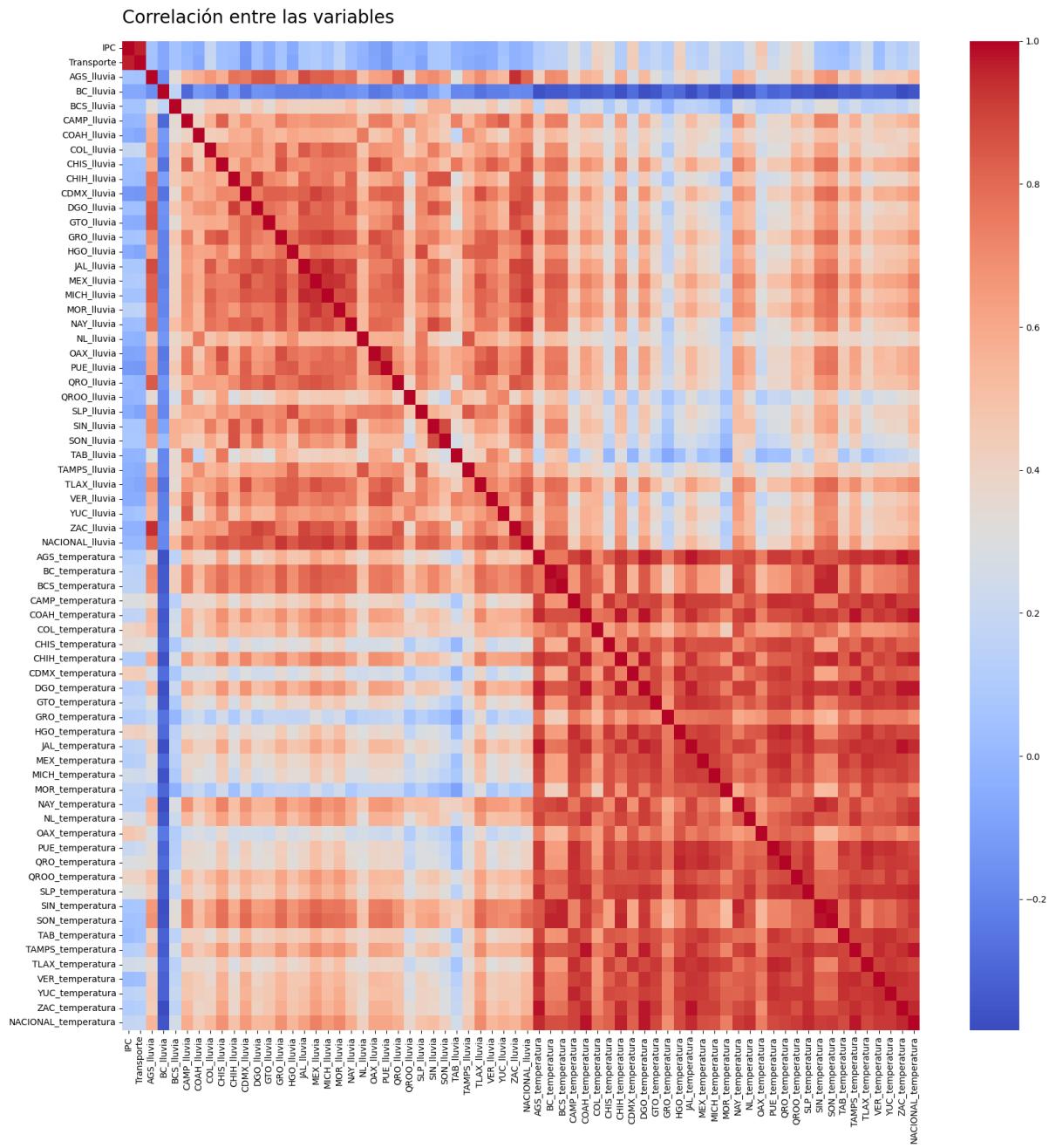


Además se realizaron pruebas estadísticas para ver que el modelo de regresión sea correcto, y se obtuvo que es válido el ajuste que se está realizando con la regresión lineal, el único problema que se tiene es que los errores siguen una distribución normal, lo cual disminuye el poder de predicción del algoritmo. Por ello es que como alternativa se utiliza el algoritmo de regresión Kernel, los resultados de esta regresión Kernel se encuentran en la Gráfica de *Kernel Ridge Regression con Kernel gaussiano al conjunto de prueba*, donde en esta figura se ve reflejado un mejor ajuste a la naturaleza de los datos, además esta regresión kernel presenta un error cuadrático medio sobre el conjunto de predicción de .093 y la raíz del error cuadrático medio de 0.12, por lo que este modelo muestra un mejor rendimiento sobre la predicción del PIP dada la variable de IPC, ya que los errores que está presentando son menores que el de la regresión lineal, lo cual nos indica que este tiene un mejor ajuste a las características de los datos.



## Análisis de correlaciones Clima y Transporte vs IPC

Primero, para este análisis comenzamos realizando un gráfico de correlaciones, para identificar posibles patrones dentro de nuestros datos y poder ver si alguno de ellos impactaba o no directamente con nuestra variable objetivo (IPC), sin embargo, reiteramos nuevamente que correlación no implica causalidad.



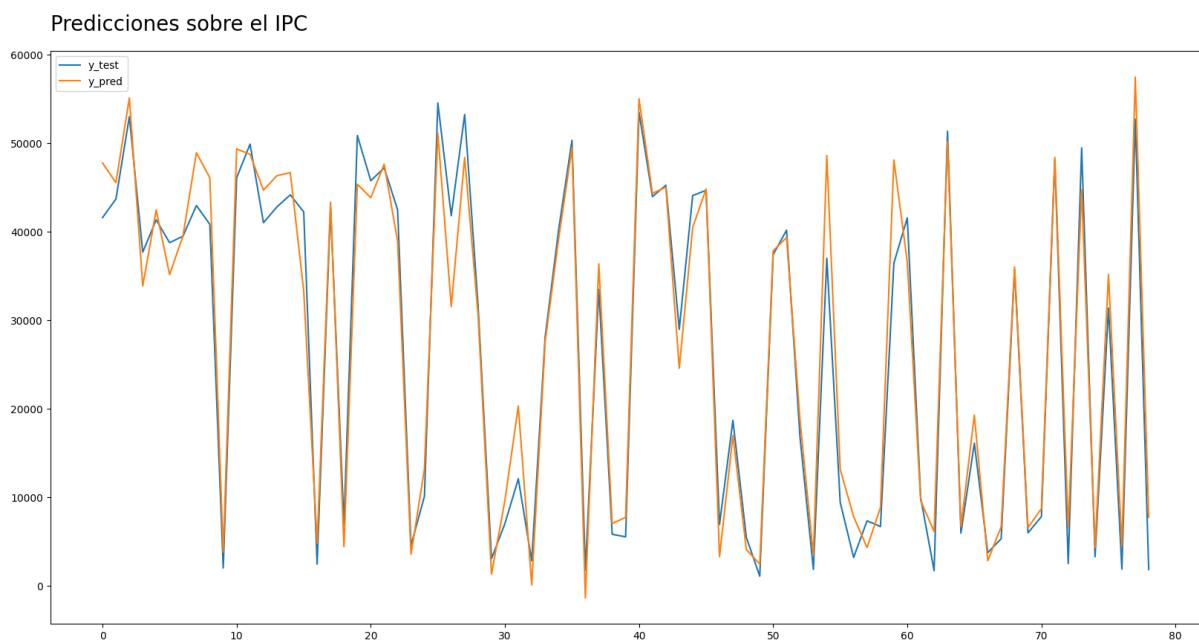
En esta gráfica es posible identificar variables que se encuentran correlacionadas en cierto grado con todas las otras demás variables. Es por ello que ajustaremos un modelo para tratar de predecir con todas estas variables (donde cada una tiene un impacto ya sea mayor o menor dependiendo de los pesos asignados por la regresión) el IPC.

## Clima y Transporte vs IPC

En el primer modelo dejamos cada una de las variables del clima de cada estado, junto a la cantidad de coches totales nacionales y así estas variables conforman a nuestros datos  $X$  y posteriormente obtenemos la variable IPC que corresponde a la variable objetivo  $y$ . Posteriormente siguiendo la estructura de preprocesamiento de los datos descrita en la parte final de la sección de la metodología obtenemos los datos separados y estandarizados y finalmente creamos y el entrenamiento del modelo en cuestión :

```
krr = KernelRidgeRegression(kernel='gaussian', params=0.01, reg_value=1e-4)
```

Posteriormente, realizamos las predicciones sobre el conjunto de prueba, obteniendo los siguientes resultados:



Dicha predicción se hace independientemente de la fecha en la que se realiza. El error obtenido fue  $MAE = 2895.239$  y  $RMSE = 3815.1218$ . Notemos que el error obtenido es bajo, basándonos en el rango en el que el IPC toma valores.

## IVF, IPC y Transporte vs Clima

En este siguiente análisis se busca poder obtener una predicción del clima para la precipitación de las lluvias y la temperatura dentro del territorio nacional, utilizando enteramente variables económicas (PIB), financieras (IPC) y demográficas (Transporte), siendo estos anteriores valores nuestros datos  $X$ , no obstante la variable objetivo es diversa pues nos enfrentamos al problema de tratar de predecir la lluvia o la temperatura a nivel nacional, o bien por cada uno de los estados de la república. Debido a que este análisis sería demasiado largo, únicamente obtuvimos los análisis de regresión con las variables objetivo y más importantes como son las lluvias y temperaturas nacionales, en conjunto con los estados con lluvias y temperaturas destacadas.

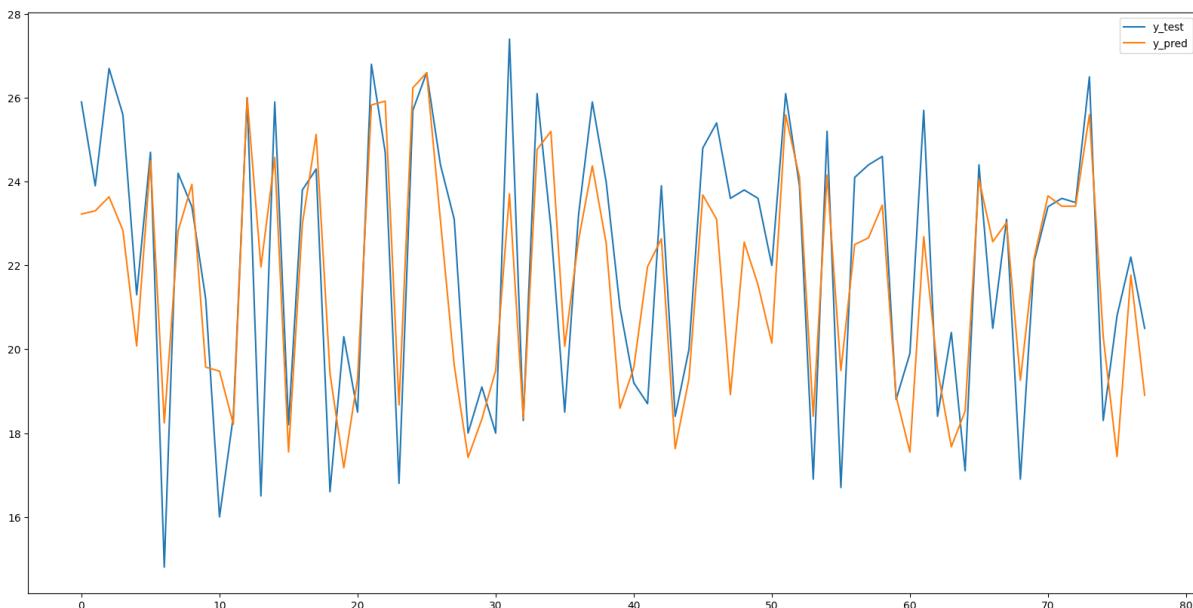
## IVF, IPC y coches vs Temperatura Nacional

En este modelo nuestros datos  $X$  se mantienen iguales, no obstante asignamos nuestra variable objetivo  $y$  como la Temperatura Nacional. Nuevamente con la estructura de preprocesamiento descrita anteriormente obtenemos los datos separados y estandarizados. Posteriormente hacemos una optimización de hiperparámetros haciendo *GridSearch*. Una vez que obtenemos los mejores hiperparámetros, el modelo creado y entrenado es el siguiente :

```
krr = KernelRidgeRegression(kernel='gaussian', params=0.01, reg_value=0.004)
```

Posteriormente, realizamos las predicciones sobre el conjunto de prueba, obteniendo los siguientes resultados:

Predicciones sobre la temperatura nacional



De nuevo, dicha predicción se hace independientemente de la fecha en la que se realiza. El error obtenido fue  $MAE = 1.5233759$  y  $RMSE = 1.9200169$ .

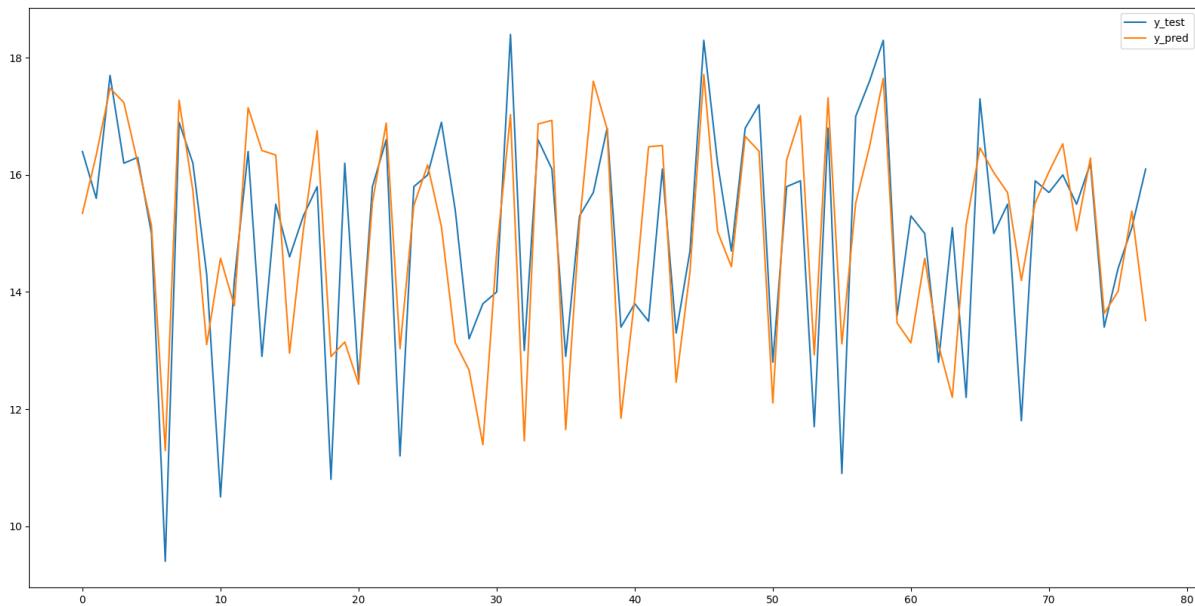
## IVF, IPC y coches vs Temperatura del Estado de México

En este otro modelo nuestros datos  $X$  se mantienen iguales, no obstante asignamos nuestra variable objetivo  $y$  como la Temperatura del Estado de México. Nuevamente con la estructura de preprocesamiento descrita anteriormente obtenemos los datos separados y estandarizados. Posteriormente hacemos una optimización de hiperparámetros haciendo *GridSearch*. Una vez que obtenemos los mejores hiperparámetros, el modelo creado y entrenado es el siguiente :

```
krr = KernelRidgeRegression(kernel='gaussian', params=0.1, reg_value=0.0001)
```

Posteriormente, realizamos las predicciones sobre el conjunto de prueba, obteniendo los siguientes resultados:

### Predicciones sobre la temperatura del Estado de México



De nuevo, dicha predicción se hace independientemente de la fecha en la que se realiza. El error obtenido fue  $MAE = 1.0264966$  y  $RMSE = 1.3838997$ .

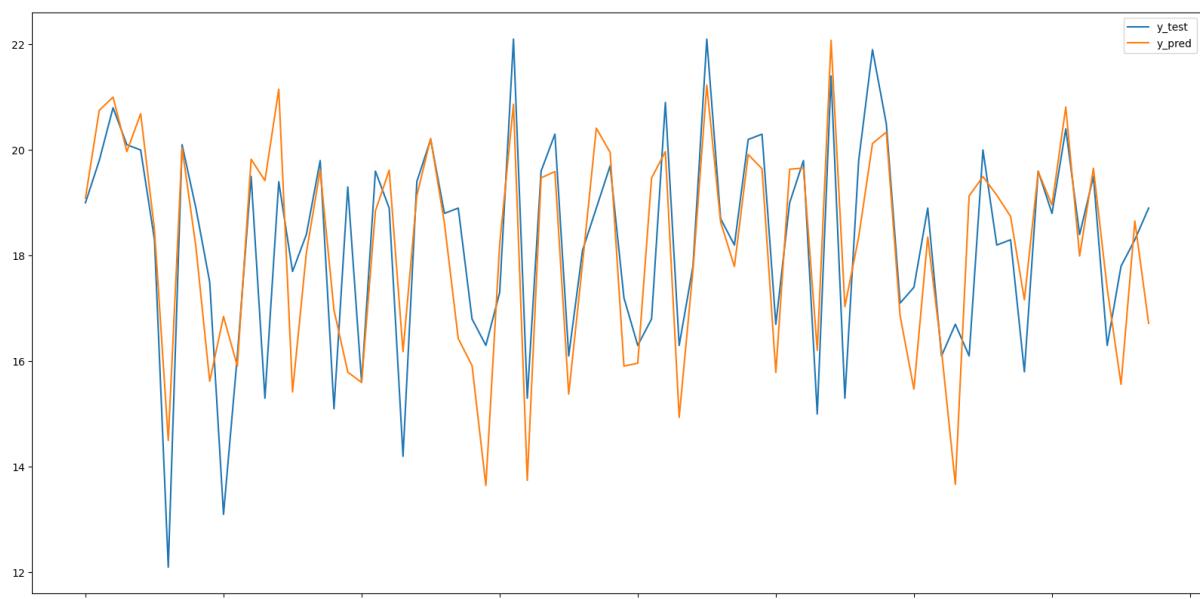
### IVF, IPC y coches vs Temperatura de Puebla

En este otro modelo nuestros datos  $X$  se mantienen iguales, no obstante asignamos nuestra variable objetivo  $y$  como la Temperatura de Puebla. Nuevamente con la estructura de preprocessamiento descrita anteriormente obtenemos los datos separados y estandarizados. Posteriormente hacemos una optimización de hiperparámetros haciendo *GridSearch*. Una vez que obtenemos los mejores hiperparámetros, el modelo creado y entrenado es el siguiente :

```
krr = KernelRidgeRegression(kernel='gaussian', params=0.3, reg_value=0.001)
```

Posteriormente, realizamos las predicciones sobre el conjunto de prueba, obteniendo los siguientes resultados:

### Predicciones sobre la temperatura de Puebla



De nuevo, dicha predicción se hace independientemente de la fecha en la que se realiza. El error obtenido fue  $MAE = 1.0208962$  y  $RMSE = 1.4125061$ .

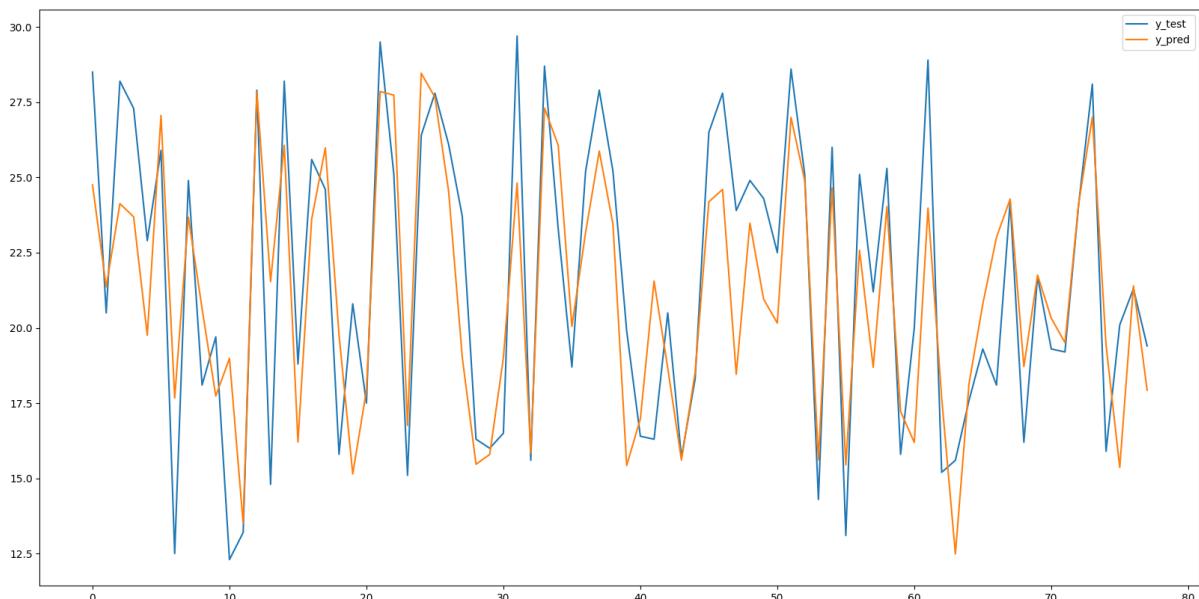
### IVF, IPC y coches vs Temperatura de Nuevo León

En este otro modelo nuestros datos  $X$  se mantienen iguales, no obstante asignamos nuestra variable objetivo  $y$  como la Temperatura de Nuevo León. Nuevamente con la estructura de preprocessamiento descrita anteriormente obtenemos los datos separados y estandarizados. Posteriormente hacemos una optimización de hiperparámetros haciendo *GridSearch*. Una vez que obtenemos los mejores hiperparámetros, el modelo creado y entrenado es el siguiente :

```
krr = KernelRidgeRegression(kernel='gaussian', params=0.01, reg_value=0.003)
```

Posteriormente, realizamos las predicciones sobre el conjunto de prueba, obteniendo los siguientes resultados:

Predicciones sobre la temperatura de Nuevo León



De nuevo, dicha predicción se hace independientemente de la fecha en la que se realiza. El error obtenido fue  $MAE = 2.244553$  y  $RMSE = 2.8061838$ .

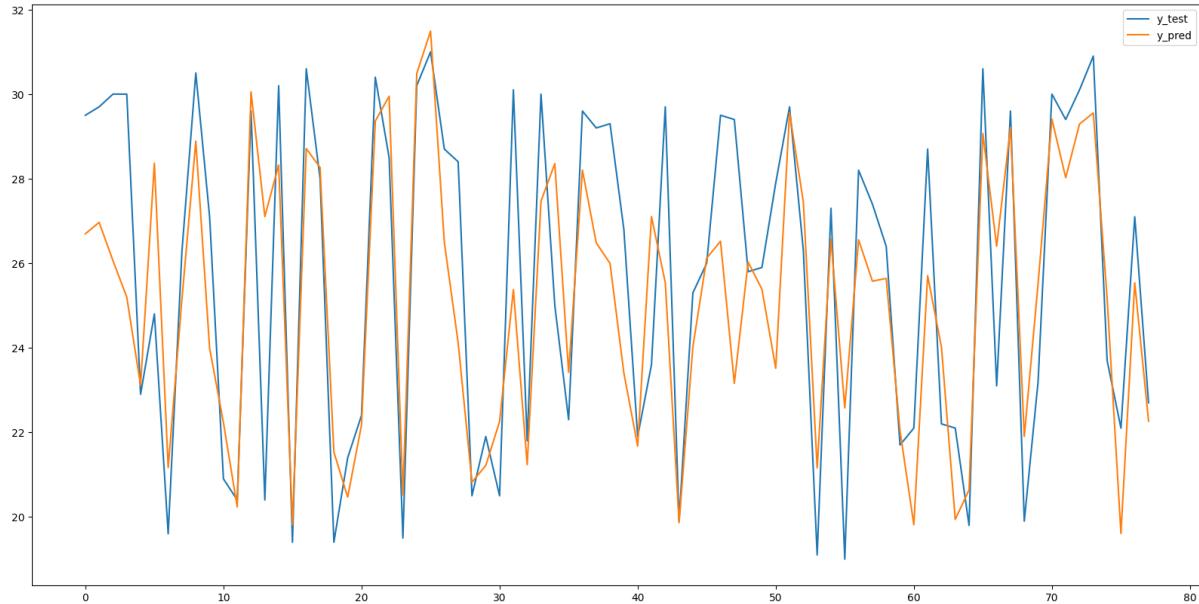
### IVF, IPC y coches vs Temperatura de Sinaloa

En este otro modelo nuestros datos  $X$  se mantienen iguales, no obstante asignamos nuestra variable objetivo  $y$  como la Temperatura de Sinaloa. Nuevamente con la estructura de preprocessamiento descrita anteriormente obtenemos los datos separados y estandarizados. Posteriormente hacemos una optimización de hiperparámetros haciendo *GridSearch*. Una vez que obtenemos los mejores hiperparámetros, el modelo creado y entrenado es el siguiente :

```
krr = KernelRidgeRegression(kernel='gaussian', params=0.01, reg_value=0.008)
```

Posteriormente, realizamos las predicciones sobre el conjunto de prueba, obteniendo los siguientes resultados:

Predicciones sobre la temperatura de Sinaloa



De nuevo, dicha predicción se hace independientemente de la fecha en la que se realiza. El error obtenido fue  $MAE = 1.8475219$  y  $RMSE = 2.3554316$ .

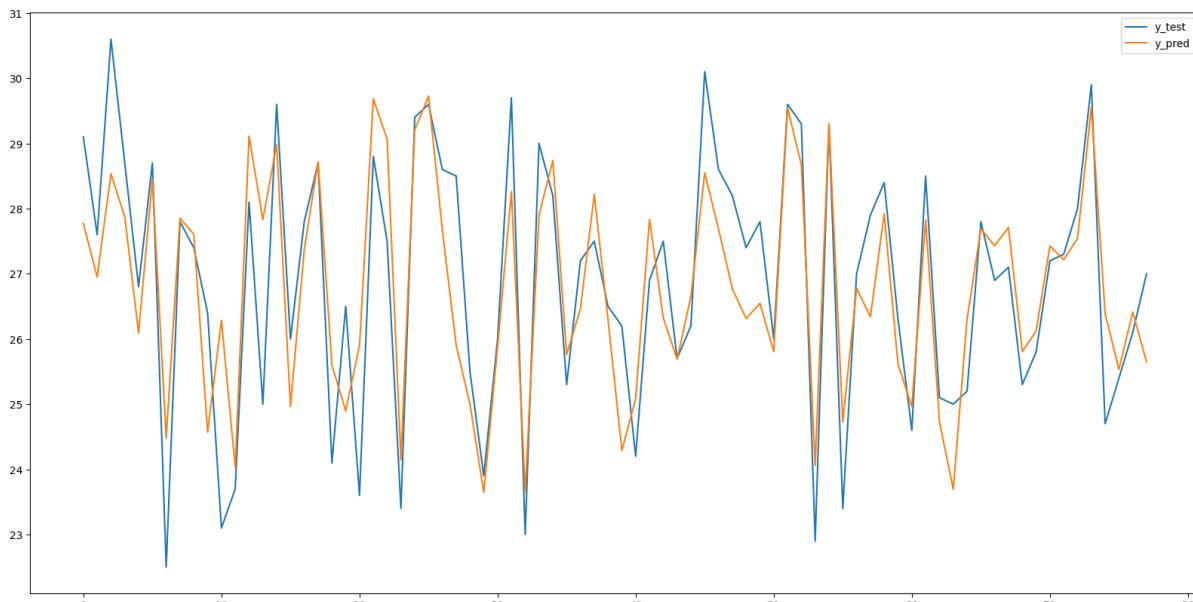
#### IVF, IPC y coches vs Temperatura de Quintana Roo

En este otro modelo nuestros datos  $X$  se mantienen iguales, no obstante asignamos nuestra variable objetivo  $y$  como la Temperatura de Quintana Roo. Nuevamente con la estructura de preprocesamiento descrita anteriormente obtenemos los datos separados y estandarizados. Posteriormente hacemos una optimización de hiperparámetros haciendo *GridSearch*. Una vez que obtenemos los mejores hiperparámetros, el modelo creado y entrenado es el siguiente :

```
krr = KernelRidgeRegression(kernel='gaussian', params=0.1, reg_value=0.04)
```

Posteriormente, realizamos las predicciones sobre el conjunto de prueba, obteniendo los siguientes resultados:

Predicciones sobre la temperatura de Quintana Roo



De nuevo, dicha predicción se hace independientemente de la fecha en la que se realiza. El error obtenido fue  $MAE = 0.85811955$  y  $RMSE = 1.1013886$ .

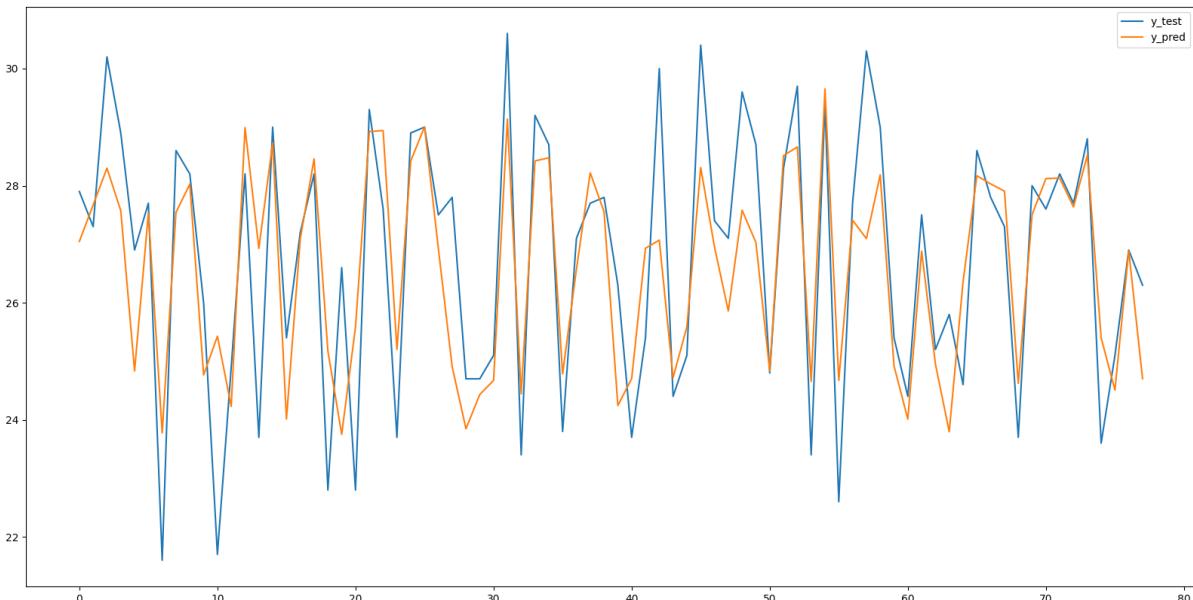
### IVF, IPC y coches vs Temperatura de Yucatán

En este otro modelo nuestros datos  $X$  se mantienen iguales, no obstante asignamos nuestra variable objetivo  $y$  como la Temperatura de Yucatán. Nuevamente con la estructura de preprocessamiento descrita anteriormente obtenemos los datos separados y estandarizados. Posteriormente hacemos una optimización de hiperparámetros haciendo *GridSearch*. Una vez que obtenemos los mejores hiperparámetros, el modelo creado y entrenado es el siguiente :

```
krr = KernelRidgeRegression(kernel='gaussian', params=0.1, reg_value=0.04)
```

Posteriormente, realizamos las predicciones sobre el conjunto de prueba, obteniendo los siguientes resultados:

Predicciones sobre la temperatura de Yucatán



De nuevo, dicha predicción se hace independientemente de la fecha en la que se realiza. El error obtenido fue  $MAE = 1.0562049$  y  $RMSE = 1.3877538$ .

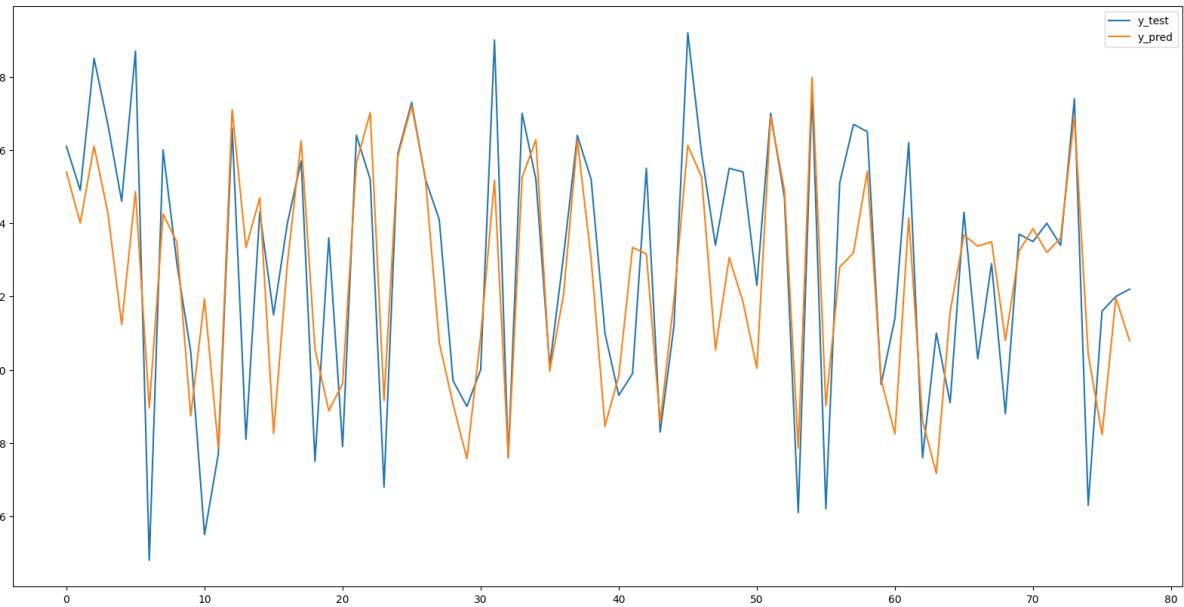
### IVF, IPC y coches vs Temperatura de San Luis Potosí

En este otro modelo nuestros datos  $X$  se mantienen iguales, no obstante asignamos nuestra variable objetivo  $y$  como la Temperatura de San Luis Potosí. Nuevamente con la estructura de preprocessamiento descrita anteriormente obtenemos los datos separados y estandarizados. Posteriormente hacemos una optimización de hiperparámetros haciendo *GridSearch*. Una vez que obtenemos los mejores hiperparámetros, el modelo creado y entrenado es el siguiente :

```
krr = KernelRidgeRegression(kernel='gaussian', params=0.1, reg_value=0.04)
```

Posteriormente, realizamos las predicciones sobre el conjunto de prueba, obteniendo los siguientes resultados:

Predicciones sobre la temperatura de San Luis Potosí



De nuevo, dicha predicción se hace independientemente de la fecha en la que se realiza. El error obtenido fue  $MAE = 1.7629331$  y  $RMSE = 2.2793207$ .

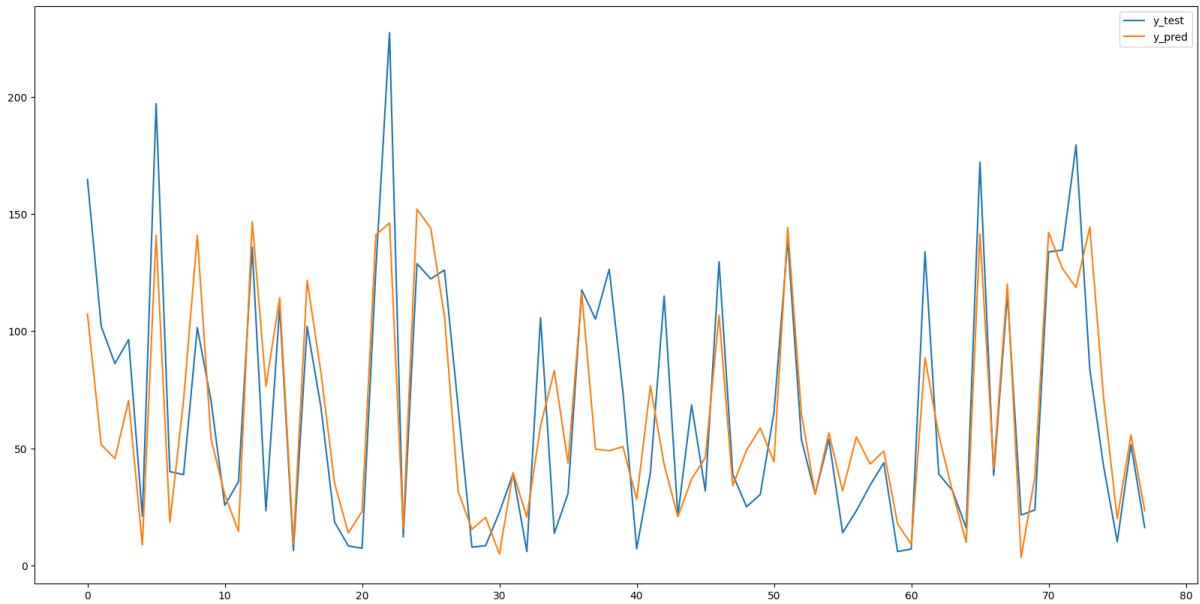
#### IVF, IPC y coches vs Lluvia Nacional

En este otro modelo nuestros datos  $X$  se mantienen iguales, no obstante asignamos nuestra variable objetivo  $y$  como la precipitación de lluvia Nacional. Nuevamente con la estructura de preprocesamiento descrita anteriormente obtenemos los datos separados y estandarizados. Posteriormente hacemos una optimización de hiperparámetros haciendo *GridSearch*. Una vez que obtenemos los mejores hiperparámetros, el modelo creado y entrenado es el siguiente :

```
krr = KernelRidgeRegression(kernel='gaussian', params=0.009, reg_value=0.008)
```

Posteriormente, realizamos las predicciones sobre el conjunto de prueba, obteniendo los siguientes resultados:

Predicciones sobre la lluvia nacional



De nuevo, dicha predicción se hace independientemente de la fecha en la que se realiza. El error obtenido fue  $MAE = 22.554369$  y  $RMSE = 30.15665$ .

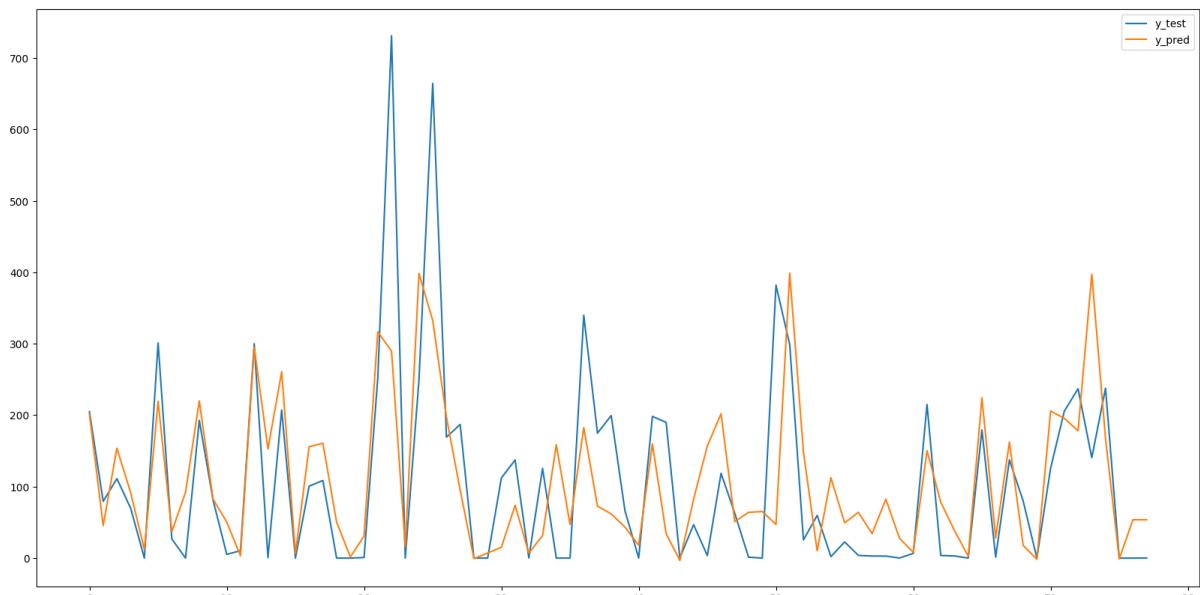
#### IVF, IPC y coches vs Lluvia Colima

En este otro modelo nuestros datos  $X$  se mantienen iguales, no obstante asignamos nuestra variable objetivo  $y$  como la precipitación de lluvia del estado de Colima. Nuevamente con la estructura de preprocesamiento descrita anteriormente obtenemos los datos separados y estandarizados. Posteriormente hacemos una optimización de hiperparámetros haciendo *GridSearch*. Una vez que obtenemos los mejores hiperparámetros, el modelo creado y entrenado es el siguiente :

```
krr = KernelRidgeRegression(kernel='gaussian', params=0.0001, reg_value=0.0001)
```

Posteriormente, realizamos las predicciones sobre el conjunto de prueba, obteniendo los siguientes resultados:

Predicciones sobre la lluvia de Colima



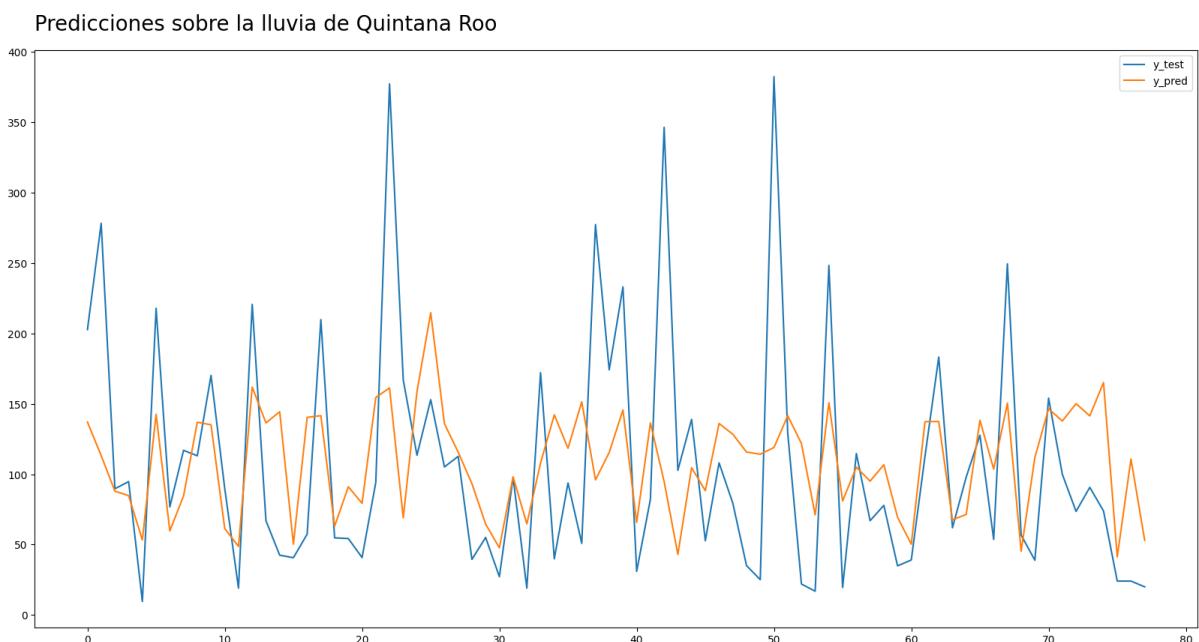
De nuevo, dicha predicción se hace independientemente de la fecha en la que se realiza. El error obtenido fue  $MAE = 67.91536$  y  $RMSE = 103.615074$ .

### IVF, IPC y coches vs Lluvia Quintana Roo

En este otro modelo nuestros datos  $X$  se mantienen iguales, no obstante asignamos nuestra variable objetivo  $y$  como la precipitación de lluvia del estado de Quintana Roo. Nuevamente con la estructura de preprocesamiento descrita anteriormente obtenemos los datos separados y estandarizados. Posteriormente hacemos una optimización de hiperparámetros haciendo *GridSearch*. Una vez que obtenemos los mejores hiperparámetros, el modelo creado y entrenado es el siguiente :

```
krr = KernelRidgeRegression(kernel='gaussian', params=0.001, reg_value=0.007)
```

Posteriormente, realizamos las predicciones sobre el conjunto de prueba, obteniendo los siguientes resultados:



De nuevo, dicha predicción se hace independientemente de la fecha en la que se realiza. El error obtenido fue  $MAE = 56.831257$  y  $RMSE = 76.56804$ .

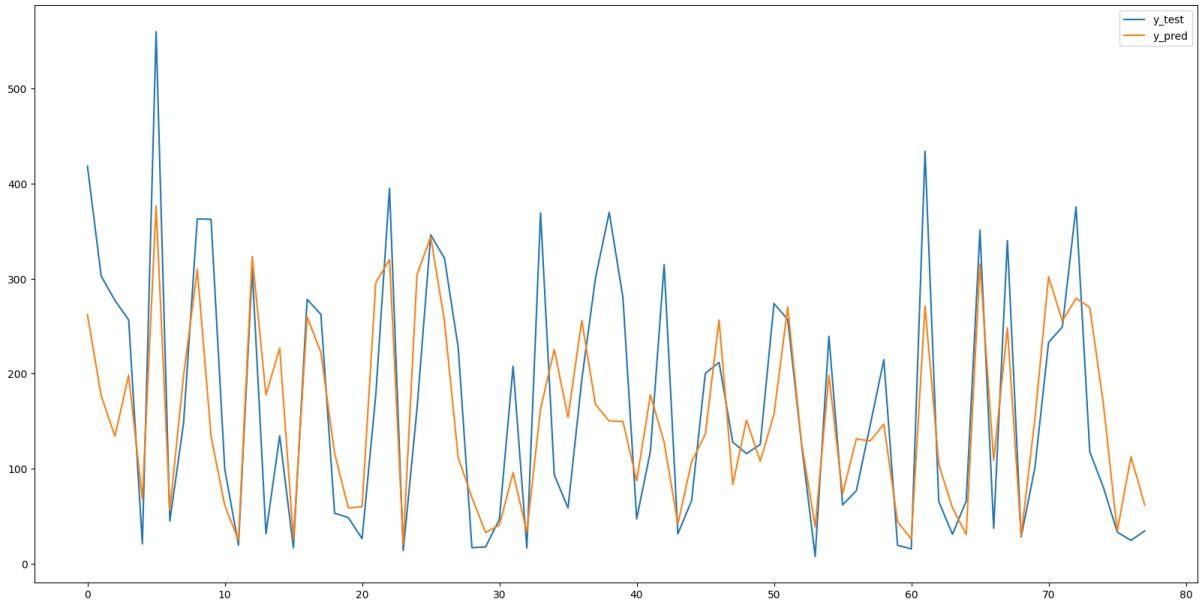
### IVF, IPC y coches vs Lluvia Chiapas

En este otro modelo nuestros datos  $X$  se mantienen iguales, no obstante asignamos nuestra variable objetivo  $y$  como la precipitación de lluvia del estado de Chiapas. Nuevamente con la estructura de preprocesamiento descrita anteriormente obtenemos los datos separados y estandarizados. Posteriormente hacemos una optimización de hiperparámetros haciendo *GridSearch*. Una vez que obtenemos los mejores hiperparámetros, el modelo creado y entrenado es el siguiente :

```
krr = KernelRidgeRegression(kernel='gaussian', params=0.009, reg_value=0.005)
```

Posteriormente, realizamos las predicciones sobre el conjunto de prueba, obteniendo los siguientes resultados:

Predicciones sobre la lluvia de Chiapas



De nuevo, dicha predicción se hace independientemente de la fecha en la que se realiza. El error obtenido fue  $MAE = 66.732635$  y  $RMSE = 87.82864$ .

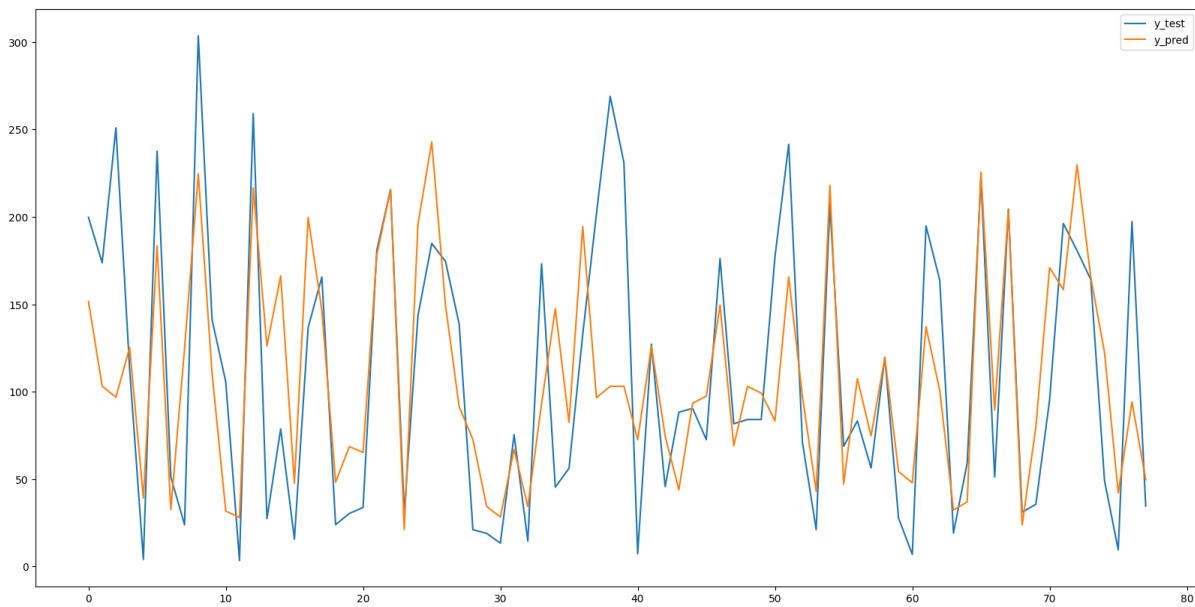
### IVF, IPC y coches vs Lluvia Campeche

En este otro modelo nuestros datos  $X$  se mantienen iguales, no obstante asignamos nuestra variable objetivo  $y$  como la precipitación de lluvia del estado de Campeche. Nuevamente con la estructura de preprocesamiento descrita anteriormente obtenemos los datos separados y estandarizados. Posteriormente hacemos una optimización de hiperparámetros haciendo *GridSearch*. Una vez que obtenemos los mejores hiperparámetros, el modelo creado y entrenado es el siguiente :

```
krr = KernelRidgeRegression(kernel='gaussian', params=0.009, reg_value=0.01)
```

Posteriormente, realizamos las predicciones sobre el conjunto de prueba, obteniendo los siguientes resultados:

Predicciones sobre la lluvia de Campeche



De nuevo, dicha predicción se hace independientemente de la fecha en la que se realiza. El error obtenido fue  $MAE = 42.962452$  y  $RMSE = 55.77219$ .

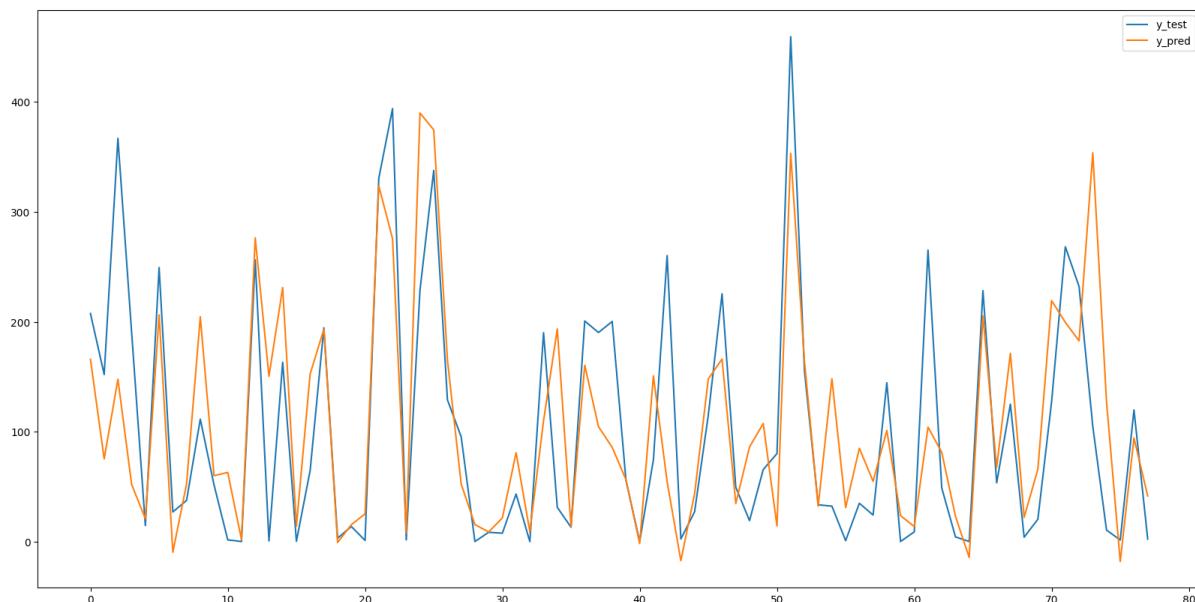
#### IVF, IPC y coches vs Lluvia Morelos

En este otro modelo nuestros datos  $X$  se mantienen iguales, no obstante asignamos nuestra variable objetivo  $y$  como la precipitación de lluvia del estado de Morelos. Nuevamente con la estructura de preprocessamiento descrita anteriormente obtenemos los datos separados y estandarizados. Posteriormente hacemos una optimización de hiperparámetros haciendo *GridSearch*. Una vez que obtenemos los mejores hiperparámetros, el modelo creado y entrenado es el siguiente :

```
krr = KernelRidgeRegression(kernel='gaussian', params=0.009, reg_value=0.01)
```

Posteriormente, realizamos las predicciones sobre el conjunto de prueba, obteniendo los siguientes resultados:

Predicciones sobre la lluvia de Morelos



De nuevo, dicha predicción se hace independientemente de la fecha en la que se realiza. El error obtenido fue  $MAE = 52.953743$  y  $RMSE = 75.98079$ .

## Discusión

Los resultados derivados del análisis de regresión lineal, enfocado en la relación entre el Índice de Precios al Consumidor (IPC) y el Producto Interno Bruto (PIB), ofrecen una perspectiva esclarecedora sobre la interdependencia entre estos dos indicadores fundamentales en el contexto financiero.

En primer lugar, los resultados de la regresión lineal revelaron una relación significativa entre el IPC y el PIB. Se evidenció una asociación sustancial entre estas variables, indicando que los cambios en el Índice de Precios al Consumidor tienen relación en el Producto Interno Bruto. Este hallazgo es crucial para comprender la dinámica económica, ya que el comportamiento del PIB está muy relacionado a las variaciones en los niveles de precios. Adicionalmente, se exploró una alternativa analítica mediante el empleo de la regresión con kernel Gaussiano. Esta metodología demostró un ajuste superior a los datos, superando los resultados obtenidos con la regresión lineal convencional. La capacidad de este enfoque para adaptarse con mayor precisión al conjunto de datos y para generar predicciones más certeras se reflejó en la reducción significativa de los errores predichos. La aplicación de la regresión del kernel Gaussiano se reveló como una estrategia más robusta y eficaz para modelar la relación entre el IPC y el PIB.

El análisis comparativo entre la regresión lineal y la regresión de kernel Gaussiano resalta la ventaja de este último en términos de precisión predictiva. La mayor flexibilidad y capacidad de capturar relaciones no lineales inherentes en los datos financieros posiciona a la regresión de kernel Gaussiano como un enfoque prometedor en la predicción de la relación entre el Índice de Precios al Consumidor y el Producto Interno Bruto.

El análisis de la relación entre el Índice de Precios al Consumidor (IPC) y las variables climáticas, específicamente la temperatura y la precipitación en México, junto con la incidencia de vehículos a nivel nacional, reveló perspectivas significativas sobre la dinámica económica.

En primer lugar, se observó una correlación notoria entre los datos de transporte, en particular la cantidad de automóviles, y la variable del IPC. Este descubrimiento subraya la relevancia de la presencia vehicular como una característica determinante para el modelo. La alta correlación entre estos datos de transporte y el IPC sugiere su importancia en la predicción del comportamiento del Índice de Precios al Consumidor.

En cuanto a las variables climáticas, se evidenció que estas no mantenían una correlación marcada con el IPC de manera individual. Sin embargo, al integrarlas en un modelo de regresión Kernel junto con los datos relativos a la cantidad de automóviles, se observaron resultados prometedores. La conjunción de los datos climáticos y las variables relacionadas con el parque vehicular demostró ser significativa en la capacidad predictiva del IPC. Este análisis combinativo condujo a un modelo con buenos resultados, reflejados en un error

cuadrático medio de las predicciones que no excedía límites significativos. Este indicador sugiere que el modelo posee un ajuste apropiado al conjunto de datos.

En la fase final de nuestro análisis, se implementó una regresión de Kernel Gaussiano considerando las variables climáticas a nivel nacional en México como la variable predicable, mientras que las variables independientes abarcaron el índice de volumen físico por estado, el Índice de Precios al Consumidor (IPC) a nivel nacional y la cantidad de automóviles en la República Mexicana. La complejidad de este análisis reside en la naturaleza multifacética de las variables involucradas, lo cual exige una atención detallada en la interpretación de los algoritmos empleados.

Al examinar específicamente el impacto de la temperatura nacional en relación con variables económicas y demográficas, se identificó una baja correlación inicial entre las variables climáticas y económicas. Sin embargo, mediante la adaptación de los algoritmos para explorar posibles relaciones no lineales, la regresión de Kernel reveló una conexión sustancial entre las variables económicas y demográficas con las condiciones climáticas. Esta asociación se cuantificó utilizando el error cuadrático medio como medida de evaluación.

El análisis del error cuadrático medio resultante reveló que las predicciones de las variables climáticas no se desviaron significativamente de sus valores reales. Específicamente, el error se mantuvo por debajo de 1.5 grados, indicando una aproximación precisa que refleja la relación intrincada entre las variables climáticas, económicas y demográficas. Este nivel de precisión subraya la capacidad del modelo para capturar de manera efectiva la complejidad de las interacciones entre estas diversas variables.

En las predicciones climáticas para cada uno de los estados se obtuvieron resultados interesantes, ya que estados como el Estado de México, Puebla, Quintana Roo y Yucatán tuvieron errores en predicción menores a 1.1 grados, lo cual es muy favorable para las predicciones proporcionadas por el modelo. Por otro lado, estados como Nuevo León, Sinaloa y San Luis Potosí tuvieron errores mayores a 1.7 grados, los cuales son más grandes de lo que nos gustaría tener.

También se obtuvieron buenos resultados con la predicción de la cantidad de lluvia a nivel nacional y a nivel estado, sin embargo en estas predicciones se tenía un mayor error cuadrático medio. Esto se debe a que las unidades que se presentan en la lluvia son mucho mayores, lo cual nos indica un mayor error, por las unidades, pero aun así puede ser un error que se pueda tolerar en las predicciones.

## Conclusión

La exploración de la interacción entre variables económicas y ambientales es una empresa desafiante que demanda un enfoque minucioso y algoritmos especializados para desentrañar las complejas relaciones presentes en conjuntos de datos de naturaleza tan distinta. Los métodos convencionales se quedan cortos al intentar capturar la sofisticación y la intrincada complejidad de esta tarea. No obstante, la regresión Kernel Gaussiano emerge como una herramienta prometedora, capaz de adaptarse de manera efectiva a la búsqueda y modelado de relaciones entre estos datos.

A pesar de la inicial baja correlación entre variables económicas y ambientales, la aplicación del enfoque de Kernel Gaussiano resultó en la identificación exitosa y modelado preciso de relaciones no lineales significativas. La robustez inherente al modelo se manifiesta en su capacidad para predecir con precisión variables climáticas en función de variables económicas y demográficas. Esta habilidad ofrece una valiosa comprensión de la intrincada red de interacciones presentes en estos conjuntos de datos diversos.

Es esencial continuar desarrollando metodologías capaces de abordar la modelación e identificación de relaciones complejas entre estos datos. Estas herramientas se perfilan como recursos clave para proyectar estimaciones precisas sobre posibles cambios climáticos en un escenario donde la economía ejerce una influencia significativa. El impacto potencial de la economía en el clima es un aspecto crucial a considerar: una economía carbonizada podría impactar drásticamente el entorno. Es fundamental aspirar a una producción económica próspera sin menoscabar el medio ambiente. En este sentido, el análisis de datos y las herramientas de ciencia de datos emergen como pilares para ofrecer estimaciones pertinentes sobre cómo gestionar la economía en los años venideros.

Esta convergencia entre la ciencia de datos y la comprensión de las relaciones entre variables económicas y ambientales promete no solo esclarecer dinámicas complejas, sino también brindar un marco sólido para tomar decisiones informadas que equilibren el crecimiento económico con la sostenibilidad ambiental en el horizonte del futuro.

## Agradecimientos

Agradecemos al Dr. Christian Pelagio por habernos acompañado durante un semestre y ofrecernos su inmenso apoyo y conocimiento. Aunado a su colaboración en ofrecernos los datos utilizados para este estudio.

## Referencias

- Murphy, K. P. (2013). Machine learning : a probabilistic perspective. Cambridge, Mass. [u.a.]: MIT Press. ISBN: 9780262018029 0262018020
- Bishop, C. M. (2007). Pattern Recognition and Machine Learning (Information Science and Statistics). Springer. ISBN: 0387310738  
<https://jax.readthedocs.io/en/latest/> (05/12/2023. 20:00hrs)
- Pedro Domingos. 2020. Every Model Learned by Gradient Descent Is Approximately a Kernel Machine.
- Chmiela, S., Tkatchenko, A., Sauceda, H. E., Poltavsky, I., Schütt, K. T., Müller, K.-R., *Science Advances*, 3(5), 2017, e1603015.
- Chmiela, S., Sauceda, H. E., Müller, K.-R., Tkatchenko, A., *Nature Communications*, 9(1), 2018, 3887.
- Chmiela, S., Sauceda, H. E., Tkatchenko, A., Müller, K.-R., In: *Machine Learning Meets Quantum Physics, Lecture Notes in Physics* (Springer), 968, 2020, pp. 129-154.
- Chmiela, S., Vassilev-Galindo, V., Unke, O. T., Kabylda, A., Sauceda, H. E., Tkatchenko, A., Müller, K.-R., *Science Advances*, 9(2), 2023, eadf0873.