Modelos de regresión lineal utilizando indicadores económicos

Introducción

El PIB (Producto Interno Bruto) es la suma del valor de todos los bienes y servicios de uso final que genera un país o entidad federativa durante un periodo (comúnmente un año o trimestre).

Es muy importante saber si la economía de un país está creciendo o no, es decir, si se produjo más o menos que el año anterior. El cambio del PIB a lo largo del tiempo es uno de los indicadores más importantes del crecimiento económico. **Un crecimiento en el PIB** significa que hay más dinero para construir edificios, casas o comprar maquinaria y que se producirán más bienes y servicios. Esto es beneficioso para todos porque habrá más empleo y más oportunidades para hacer negocios. Por el contrario, **una disminución en el PIB** significa que la producción y actividad económica del país disminuirá; en estas condiciones, es probable que haya desempleo y que esto afecte a muchas familias.

Una de las contribuciones al crecimiento/decrecimiento del PIB de México, se debe en gran parte a su actividad industrial; recordemos que la actividad industrial se define como la transformación de materias primas en productos de consumo final o intermedio. Las principales industrias en México son la automotriz, la petroquímica, la construcción y el cemento, la textil, la industria alimenticia y de bebidas, la minería y el turismo.

Cuando hablamos de actividad industrial, también hablamos de niveles de productividad; usualmente se nos da dicha información en unidades monetarias o en unidades porcentuales. México cuenta con una herramienta estadística llamada: **índice de volumen físico**, que se utiliza para medir los cambios en la cantidad de bienes y servicios producidos en una economía, independientemente de las fluctuaciones en los precios. Este índice es crucial para evaluar mensualmente el crecimiento económico real, ya que permite a los analistas y decisores entender cómo varía la producción sin que los resultados estén distorsionados por la inflación o deflación.

Para propósitos de este ejercicio. Solo se trabajó con la variable el **Índice del total de la actividad industrial mexicana** y la columna **Fecha.** Los datos de las variables se encuentran en las mismas unidades, por lo que no fue necesario normalizarlos, y fueron recopilados desde enero de 1993 hasta febrero del 2024 (índice base 2018 = 100) y descargados del banco de información económica (BIE) del Instituto Nacional de Estadística y Geografía (INEGI).

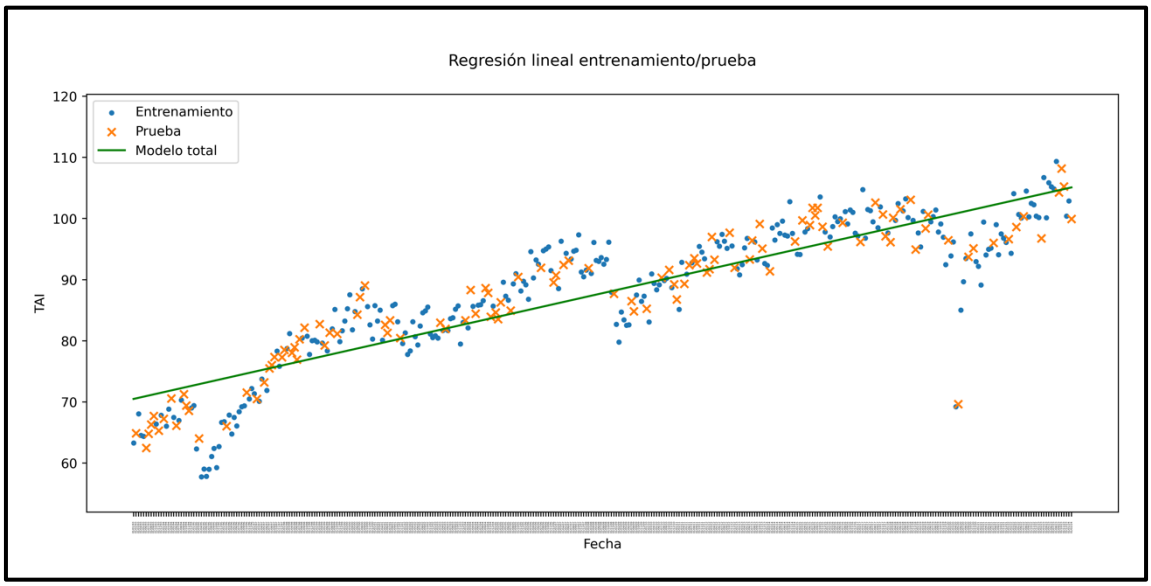
Modelos de regresión

$$\text{Índice base } 2018 = 100\ (\%\text{porcentaje}) \rightarrow \frac{\text{Cantidad producida en el año “n”}}{\text{Cantidad producida en el año } 2018} * 100$$

Variable	Representación	Tipo
TAI	Total de la actividad industrial (%porcentaje) Valores tomados: 5.788261 a 245.3426 puntos porcentuales	Continua
IEXPG	Índice de extracción de petróleo y gas (%porcentaje) Valores tomados: 5.788261 a 245.3426 puntos porcentuales	Continua
IMMN	Índice de minería de minerales metálicos y no metálicos excepto petróleo y gas (%porcentaje) Valores tomados: 5.788261 a 245.3426 puntos porcentuales	Continua
IE	Índice de edificación (%porcentaje) Valores tomados: 5.788261 a 245.3426 puntos porcentuales	Continua
ICOI	Índice de construcción de obras de ingeniería civil (%porcentaje) Valores tomados: 5.788261 a 245.3426 puntos porcentuales	Continua
IEA	Índice de elaboración de azúcares, chocolates, dulces y similares (%porcentaje) Valores tomados: 5.788261 a 245.3426 puntos porcentuales	Continua
IFVGO	Índice de conservación de frutas, verduras, guisos y otros alimentos preparados (%porcentaje) Valores tomados: 5.788261 a 245.3426 puntos porcentuales	Continua
IIB	Índice de industria de las bebidas (%porcentaje) Valores tomados: 5.788261 a 245.3426 puntos porcentuales	Continua
IIT	Índice de industria del tabaco (%porcentaje) Valores tomados: 5.788261 a 245.3426 puntos porcentuales	Continua

Cómo criterio de elección decidí utilizar el **coeficiente de determinación (R^2).** La razón por la que decidí usarlo es porque es el más utilizado en los lenguajes de programación con los que he trabajado. Excel, Rstudio y SPSS utilizan dicho criterio para analizar si el modelo aplicado explica la mayor parte de la variabilidad observada en los datos. En este caso, la variabilidad del "Total de la actividad industrial" con respecto al tiempo ("Fecha").

Modelo de regresión lineal



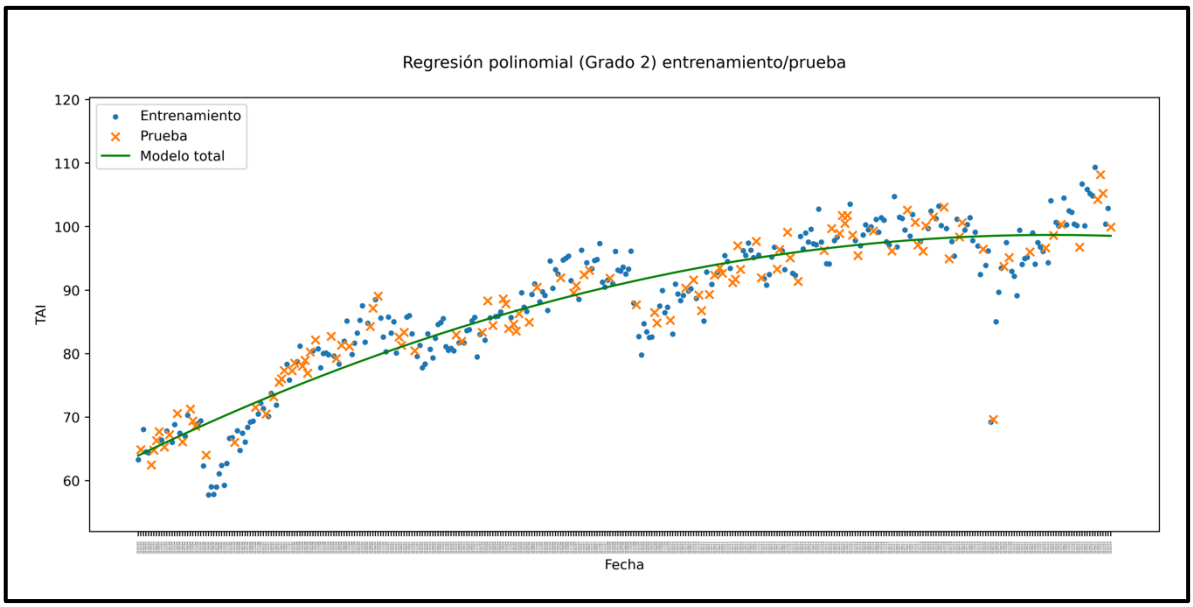
Coeficiente: 0.09366854522598557
Intercepto: 70.28165896604418
Coeficiente de determinación modelo entrenado (R^2) (entrenamiento): 0.762178927
Coeficiente de determinación modelo entrenado (R^2) (prueba): 0.800668902

Coeficiente: 0.09276052452212805
Intercepto: 70.49376046632901
Coeficiente de determinación modelo total (R^2) (entrenamiento): 0.762095124
Coeficiente de determinación modelo total (R^2) (prueba): 0.801176422

Interpretación:

Podemos ver que tanto en el modelo de entrenamiento como en el de prueba, los R^2 no padecen de algún sobreajuste o subajuste. Ya que la diferencia entre los ajustes porcentuales no tiene una diferencia significativa; no necesitamos recurrir a algún método de regularización.

Modelo de regresión polinomial



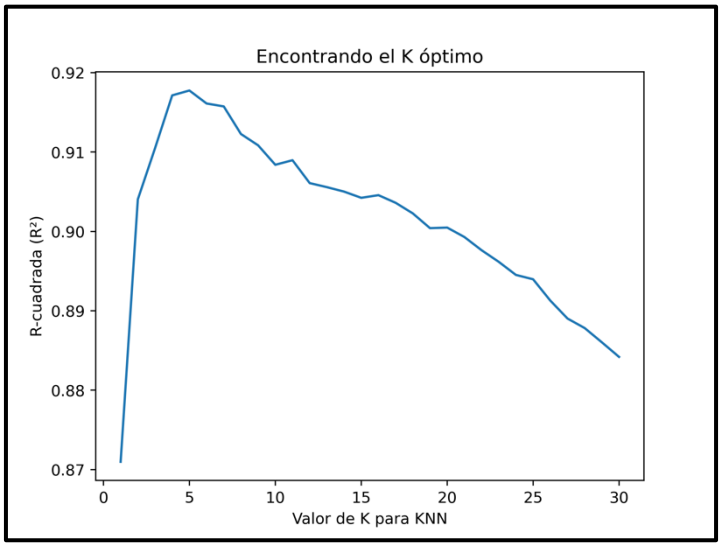
Coeficiente: [0. 0.21037883 -0.00030801]
Intercepto: 62.745477923759154
Coeficiente de determinación modelo entrenado (R^2) (entrenamiento): 0.838126089
Coeficiente de determinación modelo entrenado (R^2) (prueba): 0.842247915

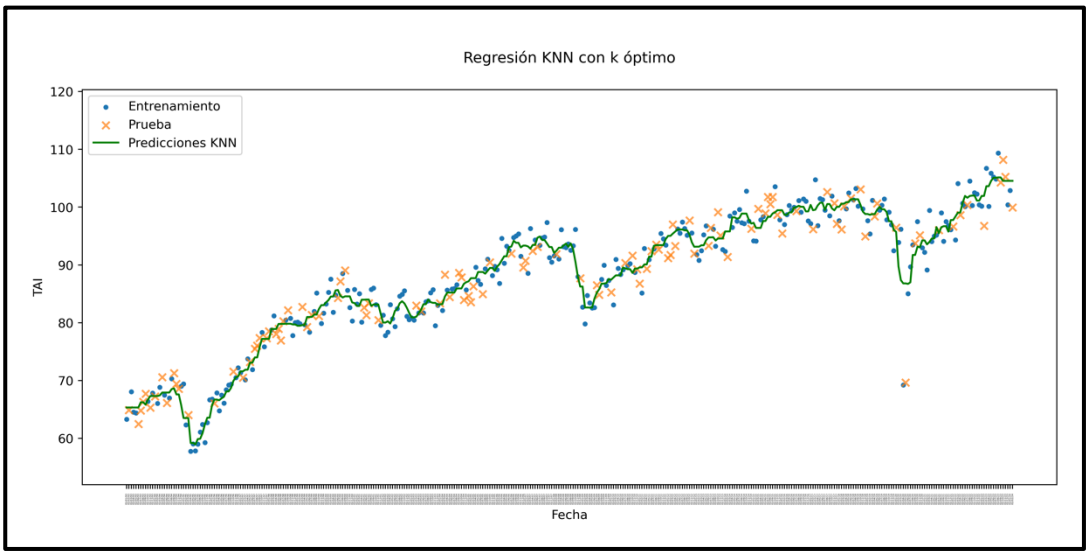
Coeficiente: [0. 0.198272 -0.00028287]
Intercepto: 63.95204869230575
Coeficiente de determinación modelo total (R^2) (entrenamiento): 0.836961550
Coeficiente de determinación modelo total (R^2) (prueba): 0.849568734

Interpretación:

Podemos ver que tanto en el modelo de entrenamiento como en el de prueba, los R^2 no padecen de algún sobreajuste o subajuste. Ya que la diferencia entre los ajustes porcentuales no tiene una diferencia significativa; no necesitamos recurrir a algún método de regularización.

Modelo de regresión lineal KNN





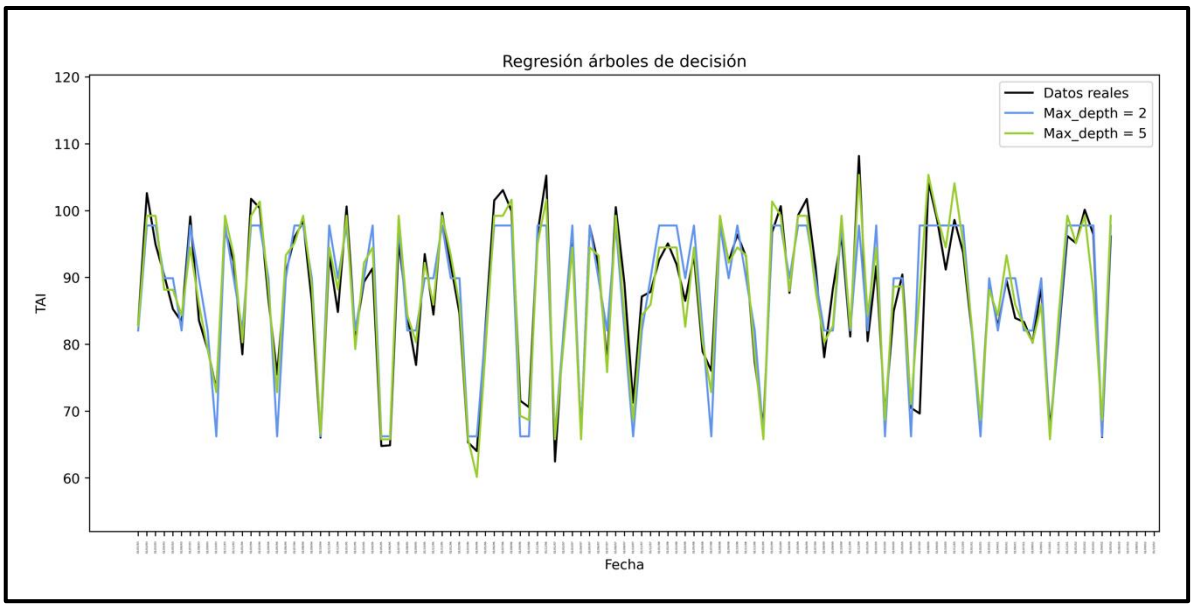
El valor óptima de k es: 5
Coeficiente de determinación (R^2) (entrenamiento): 0.96
Coeficiente de determinación (R^2) (prueba): 0.94

Coeficiente: 0.8730285947084788
Intercepto: -8.49135534912998e-17

Interpretación:

Podemos ver que tanto en el modelo de entrenamiento como en el de prueba, los R^2 no padecen de algún sobreajuste o subajuste. Ya que la diferencia entre los ajustes porcentuales no tiene una diferencia significativa; no necesitamos recurrir a algún método de regularización.

Modelo de regresión lineal árbol de decisión



Coeficiente de determinación (R^2) para max_depth = 2 (entrenamiento): 0.8672534783410406
Coeficiente de determinación (R^2) para max_depth = 2 (prueba): 0.833937723146652

Coeficiente de determinación (R^2) para max_depth = 5 (entrenamiento): 0.9582492698645353
Coeficiente de determinación (R^2) para max_depth = 5 (prueba): 0.9267310421384165

Interpretación:

Podemos ver que tanto en el modelo de entrenamiento como en el de prueba, los R^2 no padecen de algún sobreajuste o subajuste. Ya que la diferencia entre los ajustes porcentuales no tiene una diferencia significativa; no necesitamos recurrir a algún método de regularización.

Cross validation

El método de Cross validation nos permite decidir cuál de los 4 modelos anteriores de regresión es el mejor, a partir de los valores de la media y la desviación estándar.

Puntos a revisar:

1. Media baja: una media baja de los errores de predicción indica que, en promedio, el modelo está realizando predicciones cercanas a los valores reales. Esto es un buen indicio de que el modelo es preciso.

2. Desviación estándar baja: una desviación estándar baja indica que los errores de predicción están consistentemente cerca de la media del error. Esto sugiere que el modelo es fiable y consistente en sus predicciones.

Impresión de medias y desviaciones estándar

Regresión lineal [Media: -1.509, Desv: 0.449]
Regresión polinomial [Media: -0.350, Desv: 0.347]
Regresión KNN [Media: -1.137, Desv: 0.944]
Regresión árbol de decisión [Media: -0.774, Desv: 0.932]

Interpretación:

Al tener la media y desviación estándar más baja de -0.350 y 0.347, respectivamente. El mejor modelo de regresión es el **modelo polinomial de grado 2.**