Modelos de clasificación 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. Se analizará, junto con la columna **Fecha**, la relación entre el **Índice del total de la actividad industrial mexicana** y 8 subactividades industriales que se sospechan son predictivas para el total. Las subactividades seleccionadas son: **extracción de petróleo y gas, minería de minerales metálicos y no metálicos excepto petróleo y gas, edificación, construcción de obras de ingeniería civil, elaboración de azúcares, chocolates, dulces y similares, conservación de frutas, verduras, guisos y otros alimentos preparados, industria de las bebidas, e industria del tabaco.** 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).

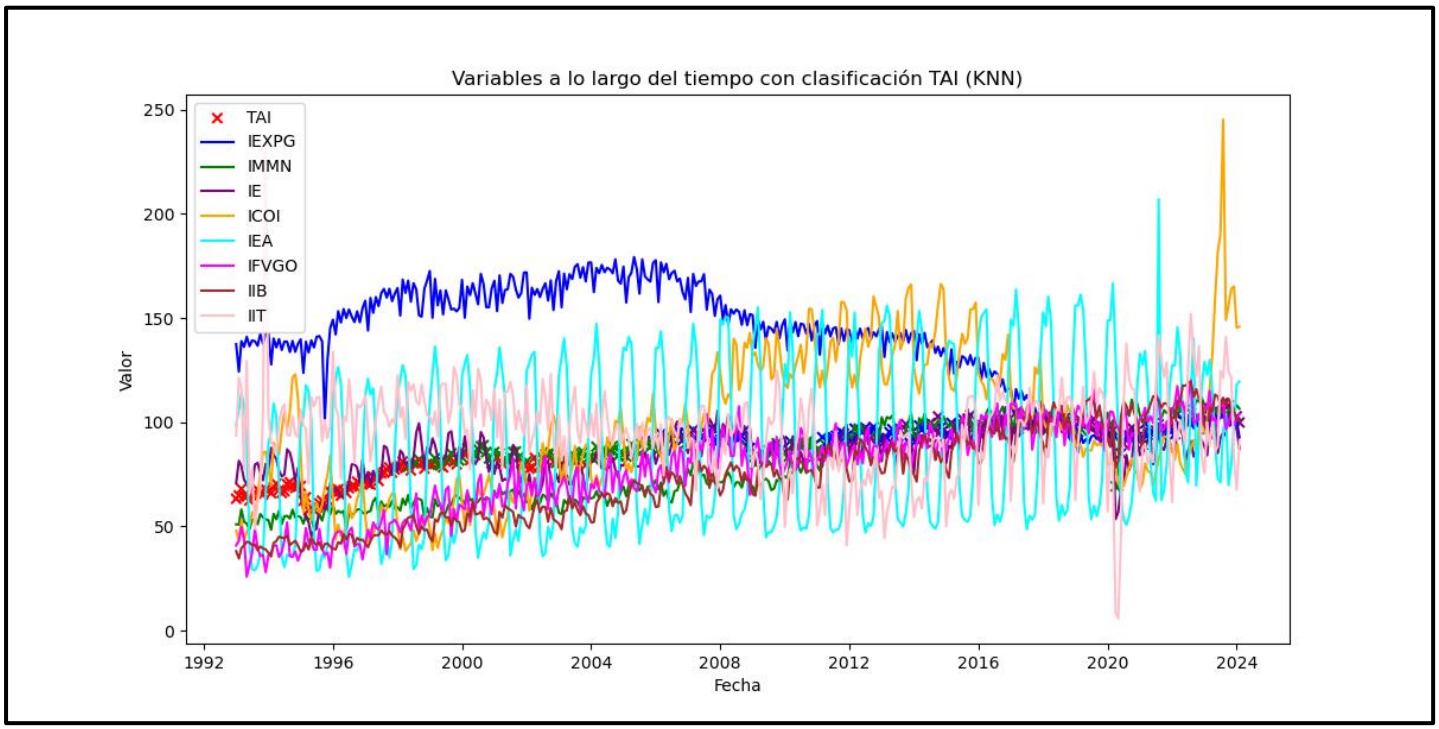
Modelos de clasificación

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

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 **Accuracy**. La razón por la que decidí usarlo es porque es el más eficiente cuando las métricas están equilibradas; lo cual es el caso, ya que todas las variables se encuentran estandarizadas y normalizadas. Además de que trabaja más con las predicciones correctas y positivas, a diferencia de los demás criterios.

Modelo de clasificación KNN



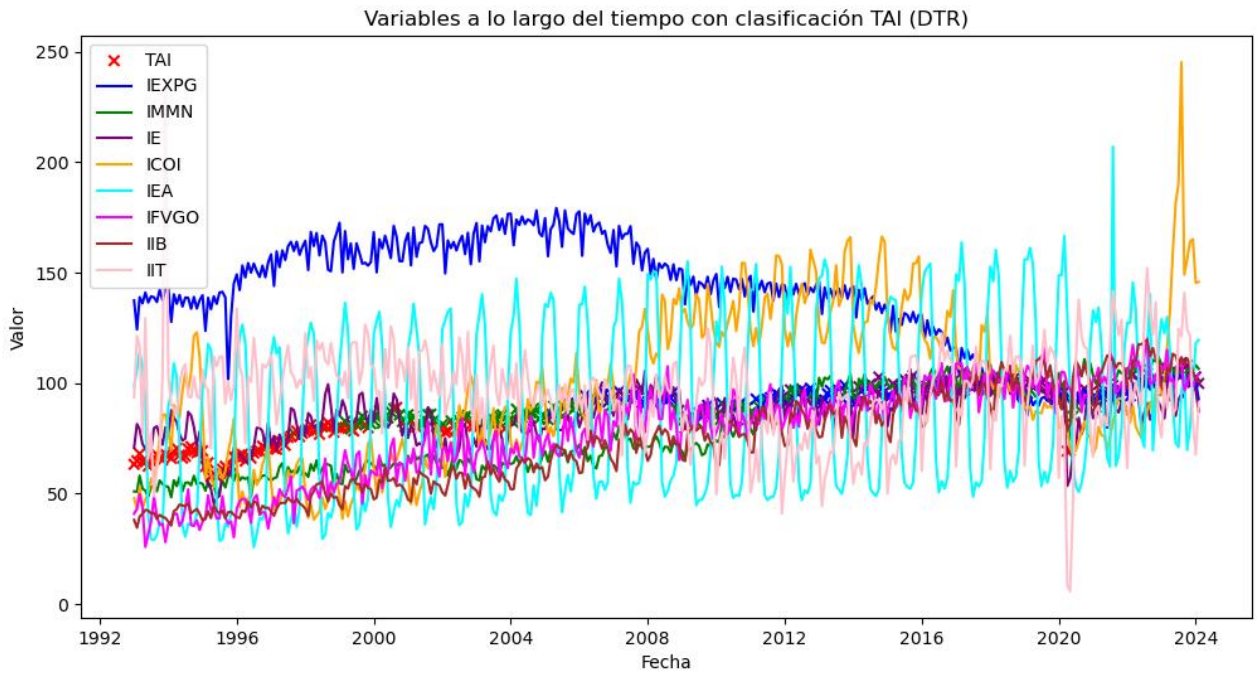
Accuracy (KNN): 0.7964601769911505

	precision	recall	f1-score	support
0	0.93	0.90	0.92	30
1	0.79	0.90	0.84	29
2	0.74	0.59	0.65	29
3	0.71	0.80	0.75	25
accuracy			0.80	113
macro avg	0.79	0.80	0.79	113
weighted avg	0.80	0.80	0.79	113

Interpretación:

El modelo nos indica que el 79.65% de las predicciones hechas son correctas. Es un buen modelo.

Modelo de clasificación árbol de decisión



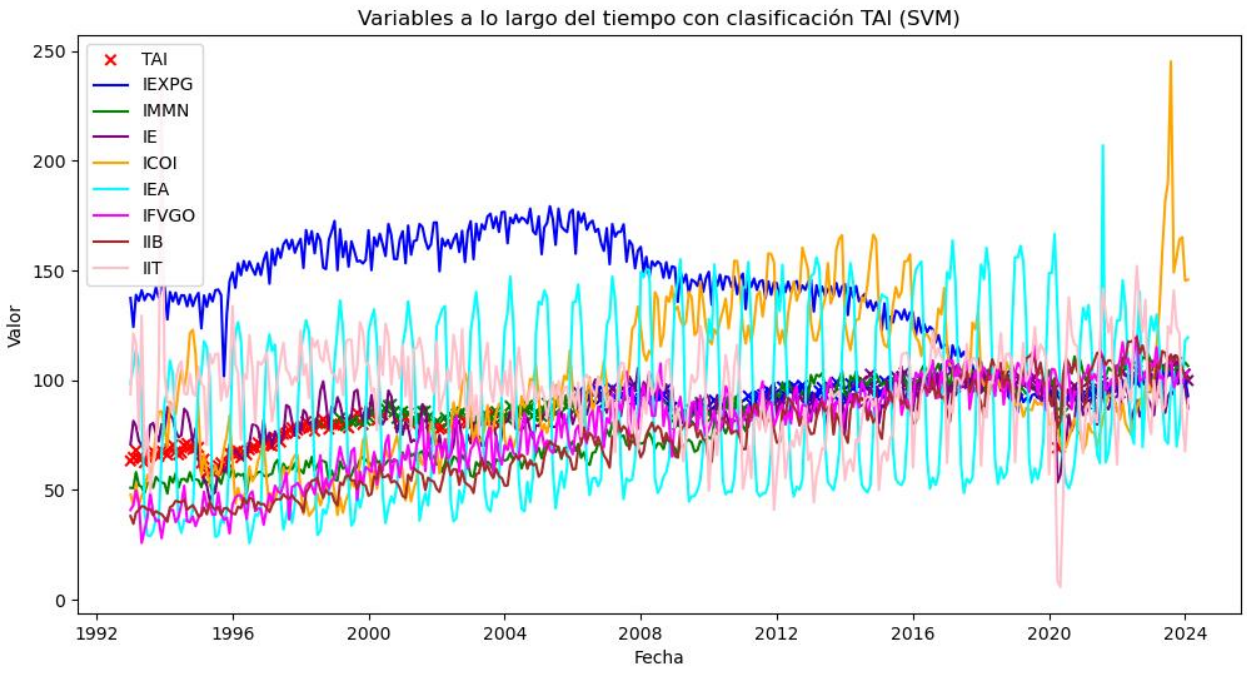
Accuracy (DTR): 0.7610619469026548

	precision	recall	f1-score	support
0	0.89	0.83	0.86	30
1	0.66	0.79	0.72	29
2	0.75	0.52	0.61	29
3	0.77	0.92	0.84	25
accuracy			0.76	113
macro avg	0.77	0.77	0.76	113
weighted avg	0.77	0.76	0.76	113

Interpretación:

El modelo nos indica que el 76.11% de las predicciones hechas son correctas. Es un buen modelo.

Modelo de clasificación SVM



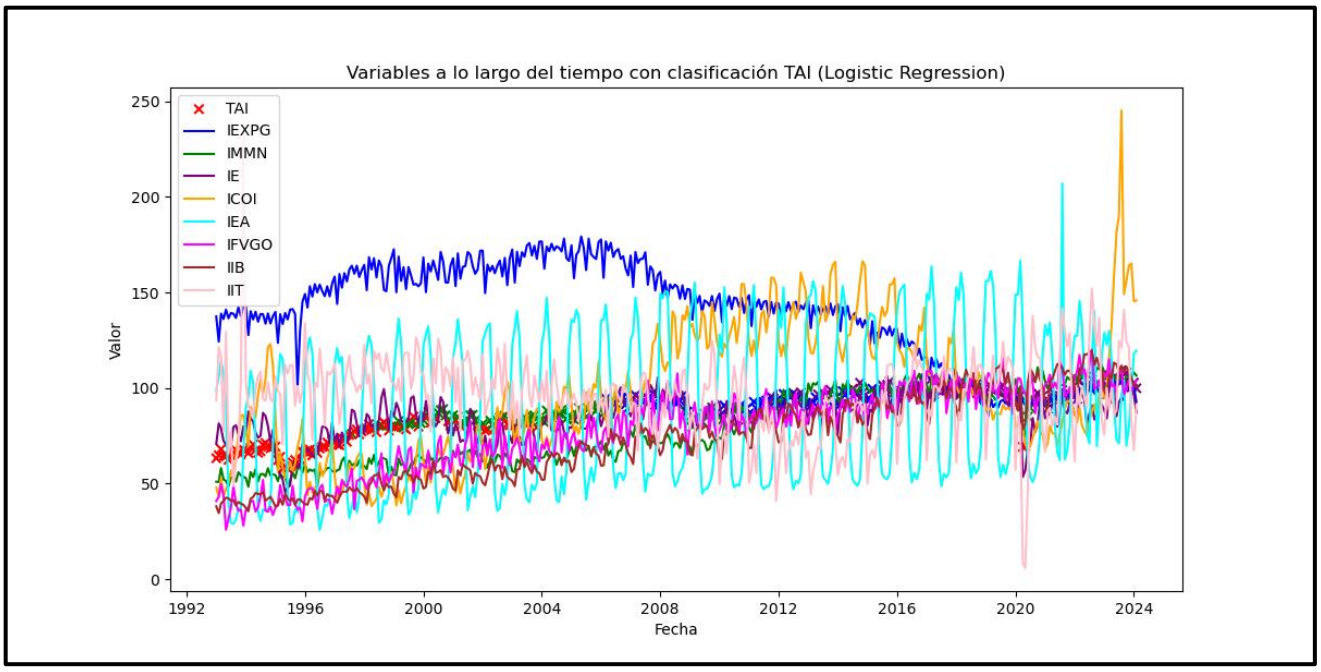
Accuracy (SVM): 0.7787610619469026

	precision	recall	f1-score	support
0	0.93	0.93	0.93	30
1	0.79	0.79	0.79	29
2	0.65	0.59	0.62	29
3	0.71	0.80	0.75	25
accuracy			0.78	113
macro avg	0.77	0.78	0.77	113
weighted avg	0.78	0.78	0.78	113

Interpretación:

El modelo nos indica que el 77.88% de las predicciones hechas son correctas. Es un buen modelo.

Modelo de clasificación Logistic Regression



Accuracy (Logistic Regression): 0.7433628318584071

	precision	recall	f1-score	support
0	0.93	0.87	0.90	30
1	0.71	0.69	0.70	29
2	0.62	0.55	0.58	29
3	0.71	0.88	0.79	25
accuracy			0.74	113
macro avg	0.74	0.75	0.74	113
weighted avg	0.74	0.74	0.74	113

Interpretación:

El modelo nos indica que el 74.34% de las predicciones hechas son correctas. Es un buen modelo.

Nota:

No es casualidad que las 4 gráficas se parezcan mucho, ya que se está trabajando con los mismos datos. Además, los datos predichos por todos los modelos (TAI) se parecen entre sí debido a que las diferencias entre los accuracy son muy mínimas.

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 del accuracy 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 el accuracy está 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

KNN [Media: 0.537, Desv: 0.187]
DTR [Media: 0.513, Desv: 0.151]
SVM [Media: 0.569, Desv: 0.187]
Logistic Regression [Media: 0.577, Desv: 0.217]

Interpretación:

Al tener la media y desviación estándar más baja de 0.513 y 0.151, respectivamente. El mejor modelo de clasificación es el **modelo árbol de decisión**