# Análisis de precios de combustibles con herramientas de Machine Learning

Francisco Chedufau y Alexander Lopez

**Resumen.** En este análisis se busca evaluar el comportamiento de precios de productos en surtidores según las variables que presenta el dataset y la correlación del precio respecto a la cotización del dólar.

## 1 Introducción

En este trabajo se abordará un dataset acerca de la modificación y actualización de precios del gobierno de la nación, con herramientas de Análisis Exploratorio de Datos y Machine Learning, utilizando Python. El principal objetivo es desarrollar un modelo que pueda describir el posicionamiento de precios y su clasificación, según su localización, categoría y relación respecto al tipo de cambio. Se van a utilizar modelos como Regresión Lineal [1], KNN Regression [2], Support Vector Regression (SVR) [3] en Jupyter Notebooks.

## 2 Metodología

El dataset utilizado se encuentra disponible en la página del Ministerio de Modernización [4].

En estos datos, cómo ya se mencionó anteriormente, se encuentran las actualizaciones de precios por surtidor, informándose en las 8 hs posteriores a la modificación del precio por parte de las empresas.

Los campos para actualizar son:

'indice_tiempo', 'idempresa', 'cuit', 'empresa', 'direccion', 'localidad', 'provincia', 'region', 'idproducto', 'producto', 'idtipohorario', 'tipohorario', 'precio', 'fecha_vigencia', 'idempresabandera', 'empresabandera', 'latitud', 'longitud', 'geojson'

Al cual le sumamos un dataset de la cotización del dólar respecto de cada fecha de

Investing [5]

El cual presenta los campos:

Fecha, Último, Máximo, Mínimo, %Var.

Para poder analizar la correlación entre los precios promedios de la nafta en este caso con la cotización promedio del dólar a la fecha.

En referencia a los modelos supervisados utilizados:

## 2.1 Regresión Lineal

Es una función que se construye calculando parámetros Beta asociados a cada variable.

$$\widehat{y} = f\left(x, \beta\right)$$

$$\hat{y}(x, w) = w_0 + w_1 x_1 + w_2 x_2 + ... + w_p x_p \quad (1)$$

$$\min_{\beta} \|X_w - y\|^2 \qquad \hat{\beta} = (X^T X)^{-1} X^T y \quad (2)$$

Para obtener los valores de los parámetros del modelo se utilizan mínimos cuadrados ordinarios y se obtiene una única solución.

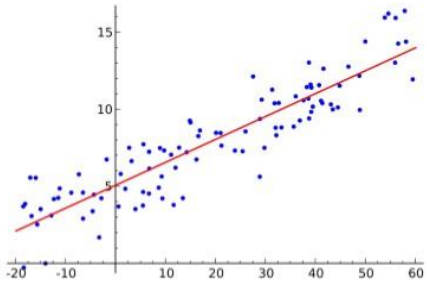


**Fig. 1.** Solución única de una regresión lineal.

## 2.2 KNN Regression

El Yi a predecir se determina por la interpolación de los Y en los K vecinos. En el training del modelo se determinan los k vecinos más cercanos por distancia euclídea.

$$d(x_a, x_b) = \sqrt{(x_{a1} - x_{b1})^2 + (x_{a2} - x_{b2})^2 + ... + (x_{ap} - x_{bp})^2} \quad (3)$$

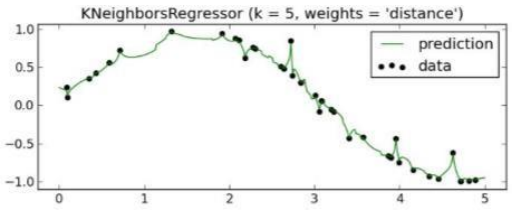


**Fig. 2.** Detalle de modelo de predicción KNN

### 2.3 SVR Support Vector Regression

Construye una función lineal y determina un radio (épsilon) como función de costo, de manera que todas las muestras estén dentro de este. Las muestras que no pertenezcan al sector determinado por el radio serán penalizadas.

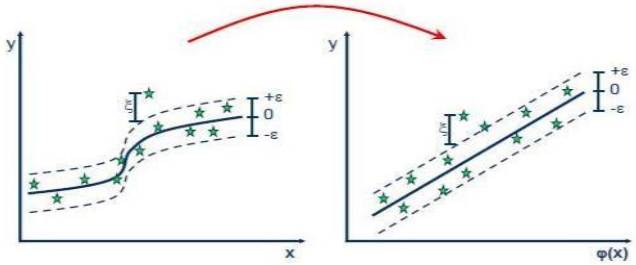


**Fig. 3.** Funcionamiento del modelo SVR con radio épsilon.

$$min\frac{1}{2}\|w\|^2 \qquad \begin{matrix} y - wx_i - b < \varepsilon \\ -y + wx_i + b < \varepsilon \end{matrix} \quad (3)$$

Solo algunas muestras definirán el radio y serán llamadas Support Vectors.

## 3 Resultados

Se realizó un Análisis Exploratorio de Datos para entender el comportamiento y las tendencias del dataset, de esta manera se puede abordar el modelo teniendo en cuenta la relación entre sus variables.

Respecto del análisis de datos logramos visualizar el comportamiento del precio de todos los tipos de productos en venta, su distribución y aumento a través del tiempo.

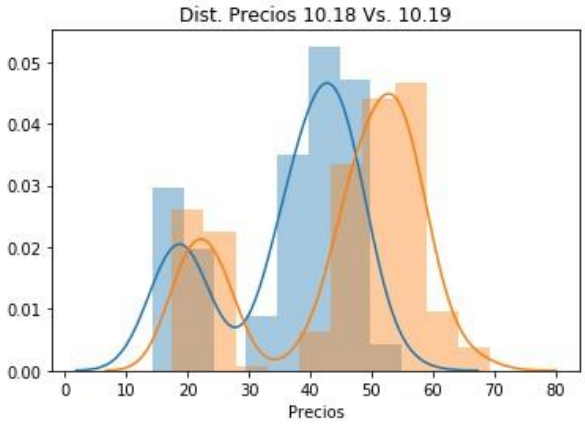


**Fig. 4.** Avance de distribución de precios de productos, actualizados del 2018 y 2019. 2018 azul, 2019 amarillo.

Finalmente, con la cotización del dólar se analizó la correlación entre el precio promedio de la nafta y la cotización del dólar promedio por mes, la cual resultó siendo de 0.97.

Respecto a los materiales, fueron nombrados anteriormente, se utilizó la agrupación por fecha, producto, provincia, precio promedio y cotización promedio del dólar a la fecha.

Regresión

Los métodos utilizados para la regresión son:

1. Regresión Lineal
2. KNN Regression
3. Support Vector Regression (SVR)

Se calcularon los errores, MAE y MSE, y resultados respecto a cada modelo.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}| \quad (4)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2 \quad (5)$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2} \quad (6)$$

**Tabla 1.**

| | KNN | SVR | Regresión Lineal |
|---|---|---|---|
| RMSE | 8,084 | 4,252 | 4,133 |
| MSE | 17,085 | 18,084 | 17,085 |
| MAE | 3,285 | 3,297 | 3,530 |
| Score | 0,606 | 0,892 | |

Por lo que se recomienda utilizar de acuerdo con los parámetros seleccionados el modelo SVR.

## 4 Discusión y Conclusiones

Respecto al manejo de datos, se puede generar un dataset de los precios vigentes por fecha de cada producto y empresa por provincia/región. De esta manera se lograrían hacer comparaciones y modelos de series de tiempos con mayor valor de predicción a futuro para dar utilidad ya sea a particulares o a personas jurídicas, para una planificación de costos logísticos posicionamiento de precios en el mercado o cualquiera sea la utilidad del modelo para el producto en cuestión. Teniendo en cuenta la volatilidad de la cotización del dólar en este país y su traslado al precio de productos respaldados por el valor de este, se considera el tipo de cambio una variable imprescindible.

Respecto a los modelos se utilizaron modelos utilizados en la cursada, siendo limitado a los recursos dados por la catedra y con la apertura a poder desarrollar mejores predicciones de precios con otros parámetros u otros modelos.

Las herramientas utilizadas son una manera de poder demostrar tendencias e información que no se alcanza a simple vista y con la aplicación de estas herramientas

y modelos, logran ser de gran utilidad para analizar el contexto de los precios de surtidor.

Se puede comparar las actualizaciones de precios respecto a la legislación Nacional que la regula y poder analizar de manera más intrínseca los cambios de precios.

Se tomó el valor del dólar como un recurso que impacta de manera directa el precio y su variación se traslada al corto plazo. No siendo así el precio de barril el cual está sujeto a relaciones político-económicas internacionales las cuales no se puede transferir su variabilidad de manera directa al precio de los productos abordados a nivel país y con una legislación interventora.

## 5 Conclusiones

En este trabajo se evaluó la capacidad de predecir el precio de los productos mediante distintos modelos teniendo en cuenta variables como cotización del dólar, provincia y tipo de producto, alcanzando niveles aceptables de evaluación en Regresión, con el modelo SVR.

## Referencias

1. Regresión lineal: "*Buckley, J., & James, I. (1979). Linear regression with censored da ta. Biometrika, 66(3), 429-436.*"

2. KNN Regression: "*Maltamo, M., & Kangas, A. (1998). Methods based on k-nearest neighbor regression in the prediction of basal area diameter distribution. Canadian Journal of Forest Research, 28(8), 1107-1115*."

3. Support Vector Regression: "Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J., & Vapnik, V. (1997). Support vector regression machines. In *Advances in neural information processing systems* (pp. 155-161)."

4. Ministerio de Modernización, disponible en internet: **https://datos.gob.ar/dataset/energia-precios-surtidor---resolucion-3142016**

5. Investing, disponible en internet: **https://es.investing.com/currencies/usd-arshistorical-data**