# Predicción & Clustering de Precios en Estaciones de Servicio

Francisco Chedufau, Alexander Lopez

UTN.BA
UNIVERSIDAD TECNOLÓGICA NACIONAL
FACULTAD REGIONAL BUENOS AIRES

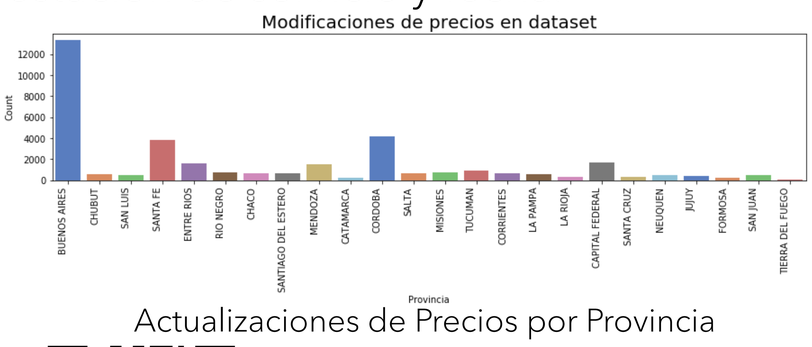
## Abstract

Con el objetivo de predecir el comportamiento de precios de combustibles en estaciones de servicios, se evaluaron tanto variables presentes en la información brindada por el Ministerio de Energía & Minería relativa a la actualización de precios en surtidor como la cotización histórica del dólar. A partir de ello, se utilizaron modelos de Machine Learning con el fin de realizar inferencias sobre el precio de los productos mencionados. Así, se creó un forecast que permite pronosticar el precio de combustibles más adecuado.

## Análisis Exploratorio de Datos

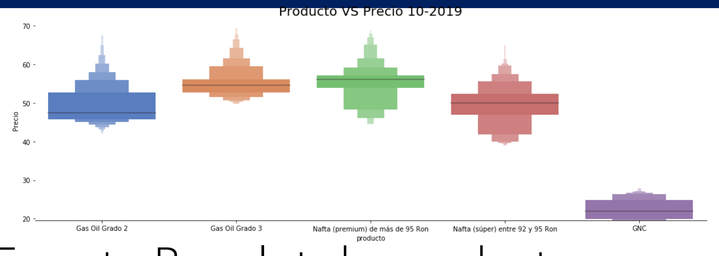
Respecto del análisis de datos, se logró visualizar el comportamiento del precio de todos los tipos de productos en venta, su distribución y aumento a través del tiempo.

En el dataset se presentaron las actualizaciones de precios de las empresas según la ubicación de cada estación de servicio y fecha.




Actualizaciones de Precios por Provincia

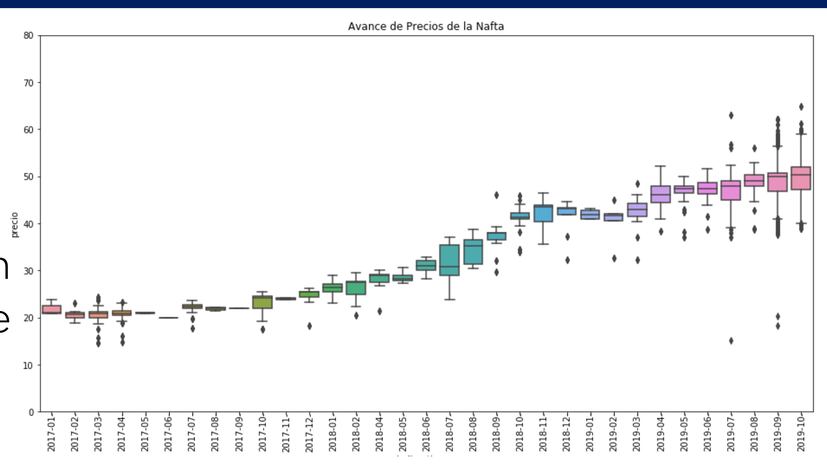
Scan QR for code




En este Boxplot de productos se analizan las actualizaciones de precios de octubre de 2019.

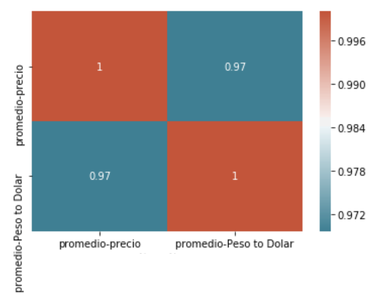
Para analizar la correlación del precio de la nafta con la cotización del dólar, se realizó un promedio de cada uno por mes. De esta manera, se obtuvo un precio promedio de nafta y cotización de dólar, por mes.

Para lograr una mejor visualización del aumento de precios de la nafta a nivel nacional, se utilizaron los boxplots de actualizaciones de precios por mes.




Correlación del precio promedio de nafta con cotización promedio del dólar por mes:

$$r = \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\left[\sum_i^n (x_i - \overline{x})^2 (y_i - \overline{y})^2\right]^{1/2}}$$




## Modelos

### Regresión

Los métodos utilizados para la regresión:

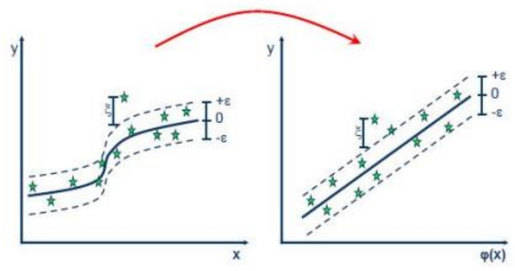
1. Regresión Lineal
Es una función lineal que se construye calculando parámetros "Beta" asociados a cada dimensión/feature.

$$\widehat{y} = f(x, \beta) \qquad \min_{\beta} \|X_w - y\|^2$$

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

2. KNN Regression
Es un módelo en el que se determina él valor según los k vecinos y los W "weights" según la interpolación utilizada, distancia en este caso.

3. Support Vector Regression (SVR)
construye una función lineal, mediante un hiperplano e hiperparametros.




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

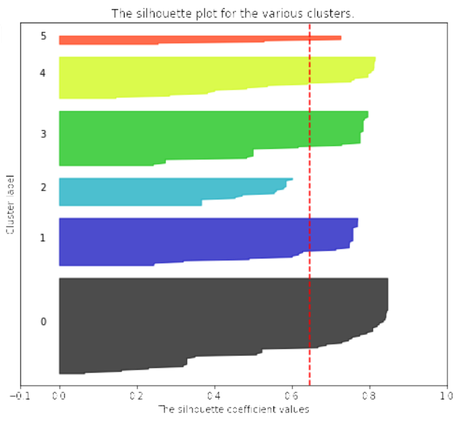
### Resultados

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

| | KNN | SVR | Regresión Lineal |
|---|---|---|---|
| RMSE | 8,084 | 4,252 | 4,133 |
| MSE | 17,085 | 18,084 | 17,085 |
| MAE | 3,285 | 3,297 | 3,53 |
| Score | 0,606 | 0,892 | |

### Clustering

Para el método de Clustering se utilizó el modelo de clasificación KNN generando 6 clusters del producto GNC en el año 2017. Siendo 242 muestras v obteniendo una clasificación




El silhouette score, que mide la similaridad (a) entre muestras de cada clase y la disimilaridad (b) respecto de otras clases, resultó ser de 0.645, siendo 1 la máxima.

$$a(x_i) = \frac{1}{n_k - 1} \sum_{x_j \in C_k, x_j \neq x_i} d(x_i, x_j)$$

$$b(x_i) = \min_{v=1,\ldots,K, v \neq k} \left[\frac{1}{n_v} \sum_{x_j \in C_v} d(x_i, x_j)\right]$$

$$S(x_i) = \frac{b(x_i) - a(x_i)}{\max[b(x_i), a(x_i)]} \qquad S_X = \frac{1}{K} \sum_{k=1}^{K} \left[\frac{1}{n_k} \sum_{x_i \in C_k} S(x_i)\right]$$

## Conclusiones

En este trabajo se evaluó la capacidad de predecir el valor de los productos mediante distintos modelos teniendo en cuenta variables como cotización del dólar, provincia y tipo de producto, alcanzando niveles significativos de accuracy y bajo error en el caso de Regresión.
Respecto del modelo de clustering se logró clasificar en 6 grupos con alta similaridad entre muestras de mismos clusters y disimilaridad respecto de otros, obteniendo así parámetros aceptables de clasificación. Estos modelos pueden ser de utilidad para calcular costos de transporte e insights competitivos para el mercado.