# UADE

## CHALLENGE DE
## **INTELIGENCIA ARTIFICIAL APLICADA A FINANZAS**

Lic. en Finanzas junto al Departamento de Economía y Finanzas de la Universidad Argentina de la Empresa (UADE)

Septiembre de 2025


Autor:

**Matias Malo Medrano**

# Modelo de Estimación de Volatilidad basado en RF

***Clasificación del producto a entregar.***

El producto desarrollado es un modelo cuantitativo experimental con fines académicos, basado en inteligencia artificial (Random Forest).

***Objetivo del proyecto.***

El objetivo del proyecto es estimar la volatilidad realizada de un activo para poder compararla con la volatilidad implícita del mercado, ofreciendo una herramienta de apoyo rápido para la toma de decisiones en trading de opciones.

***Funcionalidad propuesta.***
El modelo permite seleccionar un ticker y entrenar un Random Forest Regressor que aprende de los patrones históricos. Utiliza como variables explicativas (X) los lags de la volatilidad de los últimos seis meses y las medias móviles de tres y seis meses, mientras que la variable objetivo (Y) es la volatilidad anualizada del mes siguiente. De esta forma, el modelo captura relaciones no lineales entre el comportamiento reciente de la volatilidad y su evolución futura, entregando al usuario una predicción junto con su interpretación relativa.

***Usuarios a quienes está dirigido.***

El modelo está orientado principalmente a operadores de opciones. Se trata de un público que requiere identificar oportunidades de trading, ya sea para detectar diferencias entre volatilidad implícita y realizada, o para gestionar riesgos de forma más eficiente. Por ejemplo, el valor de las opciones se ve afectado directamente por la volatilidad implícita: si esta cae, las primas de las opciones pierden valor, lo que puede generar pérdidas aun cuando la dirección del subyacente sea la correcta. Además, el modelo responde a la necesidad de los traders de contar con información simplificada, clara y sin ruido, que les permita tomar decisiones rápidas sin quedar atrapados en el exceso de datos del mercado.

***Análisis de mercado (benchmark).***

En el mercado internacional existen herramientas profesionales como Bloomberg o Refinitiv además de los sistemas propios de bancos de inversión que ofrecen este tipo de métricas. Sin embargo, estas plataformas suelen ser costosas y de difícil acceso. El presente modelo ofrece un enfoque académico y gratuito y reproducible, constituyéndose en una alternativa liviana y de menor escala a dichas soluciones.

***Tecnología involucrada***.

El núcleo del sistema es un algoritmo de Random Forest Regressor, una técnica de ensemble learning que combina múltiples árboles de decisión para mejorar la precisión de las predicciones y reducir el riesgo de sobreajuste. Cada árbol aprende a partir de subconjuntos distintos de los datos y variables, y la predicción final se obtiene como el promedio de todos los árboles. Esta metodología es especialmente útil para capturar relaciones no lineales y patrones complejos en la serie de datos. Además, se incluyen features construidos a partir de lags (volatilidades pasadas) y medias móviles, lo que permite que el modelo incorpore la memoria histórica y las tendencias recientes.

***Prototipo funcional y posibles mejoras.***

Es un prototipo funcional que cumple con su objetivo principal: estimar la volatilidad futura de un activo a partir de datos históricos. Como próximos pasos, resulta necesario explorar posibles mejoras para disminuir el MAE (Mean Absolute Error) el cual está ubicado en 15,32%. Asimismo, se plantea la integración de datos en tiempo real y la automatización de la selección de opciones ATM, de modo que el sistema compara la volatilidad implícita con la estimación realizada y genere de forma directa una señal de trading.

***Breve análisis sobre las limitaciones del proyecto o los posibles desafíos para implementarlo.***
Una de las principales limitaciones del proyecto radica en que la estimación generada por el modelo debe ser interpretada con cautela y considerada únicamente como un complemento dentro de un análisis integral, evitando que se convierta en el único criterio de decisión al momento de definir una estrategia de trading.

# Explicación Técnica del Código

***Parámetros.*** Se define el ticker del activo (desde yfinance). En el caso de BYMA, se debe agregar el sufijo ".BA" (ej.: GGAL.BA) para que la descarga funcione. Se elige la ventana histórica; uso 5 años porque aporta suficiente muestra para entrenar los árboles y captura el régimen reciente. Se fija 252 días como año bursátil para anualizar la volatilidad. Por último, se establece una fecha de corte para trabajar solo con meses completos (evita sesgos por meses parciales).

***Descarga.*** Se establece la fecha actual y la fecha de inicio del modelo (5 años atrás), contemplando los años bisiestos para mayor precisión. Luego, se descargan los precios históricos desde yfinance y se construye un DataFrame que conserva únicamente la columna de cierres ajustados. A partir de esta serie, se calculan los retornos diarios (variaciones porcentuales entre precios consecutivos), que se almacenan en una nueva columna para su posterior procesamiento.

***Recortes de meses completos.*** El recorte a meses completos se implementa para evitar sesgos en los cálculos. Si el usuario define una fecha de corte, el modelo trabaja directamente con ella; en caso contrario, se toma la última fecha disponible en los datos y se calcula el fin de ese mes. Si la última fecha es menor al último día del mes, se considera que el mes está incompleto y se utiliza el cierre del mes anterior; en cambio, si la última fecha coincide con el fin de mes, se asume que el mes está completo y se toma esa fecha como válida. De esta manera, el DataFrame final solo incluye meses completos, garantizando consistencia en el análisis.

***Cálculo de volatilidad.*** El cálculo de volatilidad se realiza agrupando los datos por meses y obteniendo el desvío estándar diario de los rendimientos. Luego, este valor se anualiza para que pueda ser comparable con la volatilidad implícita (IV) del mercado, expresada también en términos anuales. Finalmente, se eliminan los registros vacíos para garantizar la consistencia y calidad de la base de datos sobre la cual se entrena y evalúa el modelo.

$$\sigma_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(r_i - \bar{r})^2}$$

$$\sigma_a = \sigma_d \cdot \sqrt{252}$$

- ri = retorno diario
- $\bar{r}$ = retorno promedio
- n = número de observaciones

***Terciles.*** Ordenamos la volatilidad de mayor a menor y calculamos los terciles permitiéndonos categorizar la volatilidad en baja, media o alta.

***Variables explicativas y targets.*** En esta etapa se construye un DataFrame que contiene únicamente las columnas de los meses y de la volatilidad anualizada. A partir de allí se generan seis columnas adicionales que representan los lags, es decir, la volatilidad de los últimos seis meses, junto con dos columnas que reflejan medias móviles: una de tres períodos (ma3) y otra de seis períodos (ma6). Con estos datos se conforma la matriz de variables explicativas (X), que captura el comportamiento histórico reciente de la volatilidad, y la matriz objetivo (Y), que corresponde a la volatilidad realizada, sobre la cual el modelo buscará realizar sus predicciones.

***Random Forest.*** El modelo se configuró con 500 árboles de decisión, sin límite de profundidad para que cada árbol pueda capturar patrones complejos en los datos. Se estableció que cada hoja final debe contener un mínimo de 3 observaciones, lo cual reduce el riesgo de sobreajuste y asegura mayor robustez estadística. Además, se fijó una semilla aleatoria (random_state) para garantizar la reproducibilidad de los resultados. Con esta configuración, se entrenó el modelo utilizando las matrices previamente construidas: la matriz de variables explicativas (X) y la matriz de la variable objetivo (y).

Random forest es un algoritmo de ensemble learning que combina muchos árboles de decisión para hacer una predicción más robusta. Cada árbol estima una función $fb(X)$ a partir de los datos de entrenamiento.Un árbol divide el espacio de variables X (por ejemplo, lags y medias móviles de la volatilidad) en regiones $Rj$ y asigna en cada región un valor promedio de la variable objetivo:

$$f_b(X) = \sum_{j=1}^{M} c_j \cdot \mathbf{1}_{\{X \in R_j\}}$$

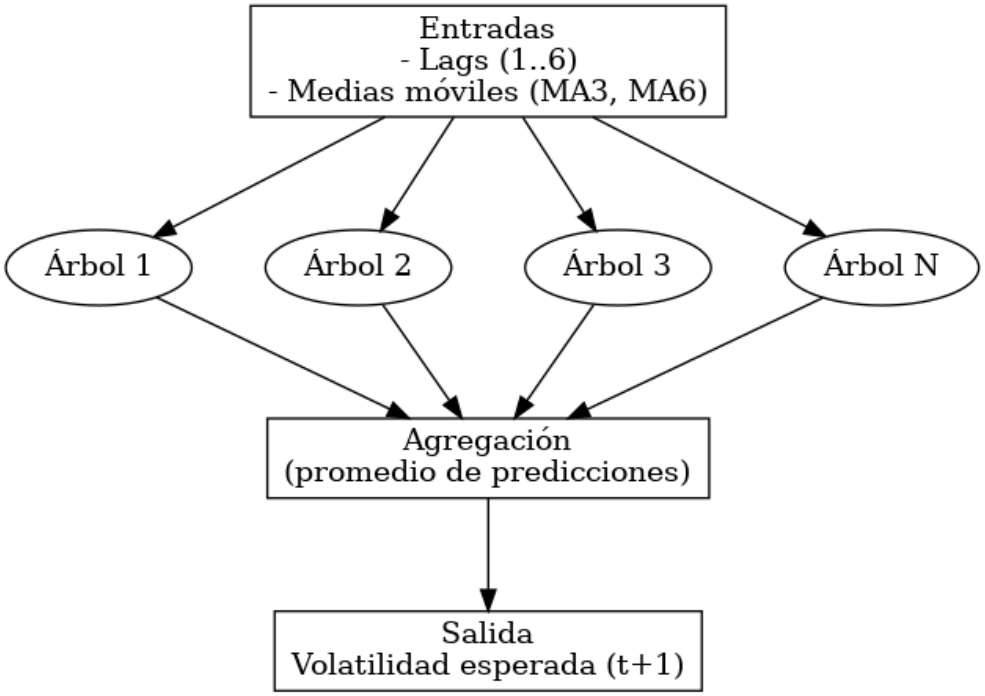
- M = número de hojas del árbol
- cj = promedio de los valores de entrenamiento en la regresión Rj
- $1\{X \in Rj\}$ = indicador que vale 1 si X cae en esa regresión

En vez de entrenar un solo árbol, Random Forest entrena B árboles, cada uno sobre un subconjunto de datos muestreo con reemplazo (bootstrap) y usando un subconjunto aleatorio de variables.

La predicción del bosque es el promedio de los B árboles:

$$\hat{y}(X) = \frac{1}{B} \sum_{b=1}^{B} f_b(X)$$

- ŷ(X) = predicción del modelo para la entrada X (volatilidad anualizada esperada par ael próximo mes)
- *f*b(X) = es la predicción del árbol b




***Estimación t+1.*** Una vez entrenado el modelo, se construye un nuevo registro con las variables explicativas más recientes (los seis últimos valores de volatilidad y las medias móviles de 3 y 6 meses). Es fundamental que estas variables coincidan en orden y formato con las utilizadas en el entrenamiento. Este registro se introduce en el modelo Random Forest previamente ajustado, el cual procesa la información y devuelve una estimación de la volatilidad anualizada esperada para el próximo mes (t+1).

***Entrenamiento y testeo.*** La base de datos se divide en dos subconjuntos: un 80% para entrenamiento y un 20% para prueba. Durante la etapa de entrenamiento, el modelo Random Forest Regressor ajusta múltiples árboles de decisión a

partir de los datos históricos. Posteriormente, en la fase de prueba, el modelo genera predicciones sobre observaciones no vistas y estas se comparan con los valores reales mediante el cálculo del Mean Absolute Error (MAE). En la estimación actual, el modelo presenta un MAE del 15,32%, lo que indica que, en promedio, la volatilidad predicha difiere en ese porcentaje respecto de la volatilidad efectivamente observada en el mercado.

***Aplicación práctica.*** El modelo predice para septiembre de 2025 una volatilidad esperada del 48,66%,. Supongamos que, al mismo tiempo, la volatilidad implícita ATM (IV) que muestra el mercado es de 60%. En este escenario, la IV está por encima de la RV proyectada, lo que indica que las opciones están relativamente caras. Por lo tanto, un operador podría implementar una estrategia de venta de volatilidad, como vender un straddle en torno al strike ATM, con la expectativa de que la volatilidad realizada sea más baja que la implícita y así capturar la prima extra que ofrece el mercado.

De forma contraria, si la IV ATM observada fuese de 30%, entonces estaría por debajo de la RV estimada del 48,66%. En ese caso, las opciones estarían relativamente baratas, y la señal sugeriría una estrategia de compra de volatilidad (por ejemplo, adquirir un strangle), buscando beneficiarse de que la volatilidad realizada supere a la implícita.

***Output.***

```
+-------------------------------------------------------------------------+
|  Modelo de Estimación de Volatilidad basado en RF                       |
|  Ticker: GGAL.BA                                                        |
+-------------------------------------------------------------------------+
|  Terciles (histórico):                                                  |
|    - p33 = 41.69%                                                       |
|    - p66 = 52.64%                                                       |
|    - p100 = 100.77%                                                     |
+-------------------------------------------------------------------------+
|  Últimos 6 meses observados:                                            |
|  Mes        Vol_Anual_%   Categoria                                     |
|  2025-03          51.65     Media                                       |
|  2025-04          83.74     Alta                                        |
|  2025-05          38.33     Baja                                        |
|  2025-06          43.02     Media                                       |
|  2025-07          33.04     Baja                                        |
|  2025-08          37.45     Baja                                        |
+-------------------------------------------------------------------------+
|  Volatilidad esperada para 2025-09 : 45.61% (Media)                     |
|  MAE del modelo: 15.32%                                                 |
+-------------------------------------------------------------------------+
```