# Análisis del índice bursátil S&P 500 y predicción del Stock de Apple

Didac Elgueta

# Índice

Inicialmente me platee hacer un proyecto sobre la predicción del precio de la vivienda en Barcelona. A través de la API de Idealista recogí los datos que me ofrecía dicho servicio. Analizando lo que disponía vi que no era suficiente para el desarrollo del proyecto que quería dado que no podía saber exactamente el precio al que se vendía el inmueble. Con este escenario me propuse cambiar de proyecto y allí es donde surgió la idea de analizar un índice bursátil con posterior predicción del precio de un stock. Para el desarrollo del proyecto escogí el índice americano S&P 500 dado que es engloba muchas de las grandes empresas americanas. Para el propósito del análisis predictivo del stock financiero he escogido Apple ya que es una de las empresas más conocidas a nivel internacional.

## Objetivo:

El propósito del proyecto es desarrollar un estudio del índice bursátil americano S&P 500 y la posterior predicción del stock de Apple para el periodo del 2015 al 2020.
Para la primera parte se han recogido los datos de todas las empresas que forman el índice. He agrupado las empresas por sectores para el análisis de los mismos.
En un primer lugar se muestra un top 5 de las compañías con mayor crecimiento por los distintos años, posteriormente analizo el crecimiento medio por sector y finalmente se muestra la diferencia entre la tasa de crecimiento mayor y menor.
Para la segunda parte del trabajo uso diferentes modelos para predecir el precio del stock de Apple. Se usan los modelos de medias móviles, regresión lineal, modelo ARIMA y finalmente la red neuronal recurrente LTSM.

## Análisis Exploratorio:

Para realizar este análisis he usado el paquete “datapackage” con el que me he descargado la tabla con todas las empresas que forman el índice bursátil dónde se puede observar el sector al que pertenecen. Posteriormente he creado una tabla para cada sector y lo he guardado en diferentes csv.

Para el sector tecnológico:

| ==========2016========= | ==========2017========= |
|---|---|
| AMD 2.951220 | ANET 1.434432 |
| NVDA 2.238471 | IPGP 1.169284 |
| DXC 0.818237 | MU 0.875912 |
| AMAT 0.728441 | PYPL 0.865214 |
| MU 0.548023 | NVDA 0.812816 |

| ==========2018========= | ==========2019========= | ==========2020========= |
|---|---|---|
| AMD 0.795720 | AMD 1.484291 | NVDA 1.146749 |
| FTNT 0.612039 | PAYC 1.162189 | PYPL 0.889064 |
| ZBRA 0.534008 | LRCX 1.147316 | AMD 0.827301 |
| PAYC 0.524337 | KLAC 0.990949 | AAPL 0.702844 |
| KEYS 0.492308 | QRVO 0.913881 | NOW 0.702146 |

Para el sector Industrial

| =========2016========= | =========2017========= |
|---|---|
| PWR 0.720988 | BA 0.894335 |
| CMI 0.552892 | CAT 0.699159 |
| CPRT 0.457774 | URI 0.628244 |
| URI 0.455473 | CPRT 0.558924 |
| ODFL 0.452345 | ODFL 0.533395 |

| =========2018========= | =========2019========= | =========2020========= |
|---|---|---|
| UAL 0.242285 | CPRT 0.903307 | ROL 0.664053 |
| TDG 0.238293 | HWM 0.825030 | ODFL 0.573427 |
| GWW 0.195175 | IR 0.793643 | FDX 0.443092 |
| ROL 0.163765 | FBHS 0.719926 | UPS 0.368614 |
| TDY 0.143086 | TDY 0.673540 | FAST 0.308525 |

Para el sector Financiero

| =========2016========= | =========2017========= |
|---|---|
| CMA 0.628257 | CBOE 0.686155 |
| ZION 0.576557 | MSCI 0.606245 |
| RF 0.495833 | PGR 0.586479 |
| SIVB 0.443734 | SPGI 0.575228 |
| FRC 0.394793 | MCO 0.565822 |

| =========2018========= | =========2019========= | =========2020========= |
|---|---|---|
| CME 0.288052 | MKTX 0.794094 | MSCI 0.432683 |
| MSCI 0.165086 | MSCI 0.751204 | SPGI 0.333016 |
| AJG 0.164665 | MCO 0.695301 | MKTX 0.308855 |
| AON 0.084776 | SPGI 0.606743 | PGR 0.288990 |
| PGR 0.071200 | AMP 0.596052 | NDAQ 0.231093 |

Para el sector de la Salud

| =========2016========= | =========2017========= |
|---|---|
| IDXX 0.608201 | ALGN 1.311349 |
| ALGN 0.459833 | VRTX 1.034207 |
| WST 0.408668 | CNC 0.785171 |
| UNH 0.360422 | ISRG 0.726383 |
| BIO 0.314582 | ILMN 0.706420 |

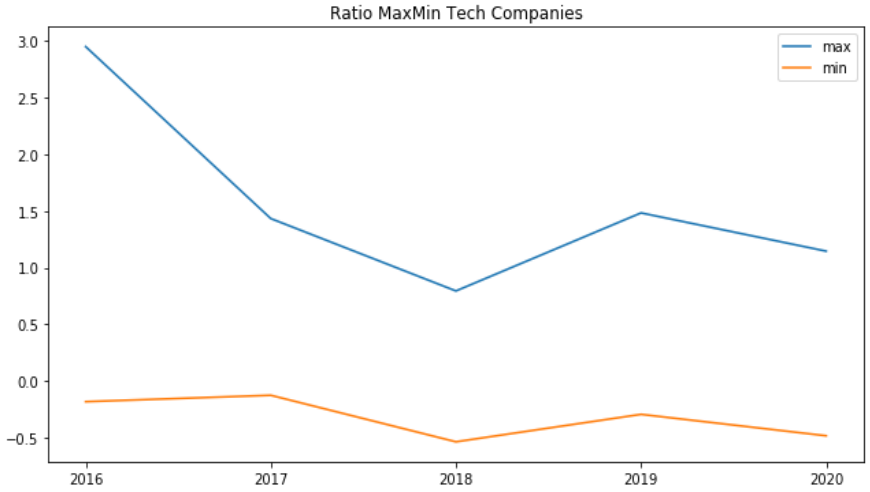
| =========2018========= | =========2019========= | =========2020========= |
|---|---|---|
| DXCM 1.087472 | DXCM 0.825876 | DXCM 0.955518 |
| ABMD 0.734379 | BIO 0.593446 | ABMD 0.823671 |
| BSX 0.425575 | ZTS 0.547229 | WST 0.822125 |
| HCA 0.416781 | WST 0.533510 | REGN 0.637823 |
| ILMN 0.372740 | EW 0.523079 | IDXX 0.479263 |

Para poder analizar mejor la tendencia de crecimiento anual que han seguido los diferentes sectores he hecho la media de todos los stocks que forman el sector.
A continuación, se presenta un gráfico con la media de los crecimientos anuales por sector.


Como se observa el gráfico se puede ver que los diferentes sectores siguen curvas de crecimiento parecidas a excepción del sector sanitario. Dicho sector no presenta variaciones tan pronunciadas cómo el resto.
Para el año 2018 la tasa de crecimiento medio cae entre un 0.3 y un 0.4 con un repunte en el año 2019 con tasas de crecimiento positivas similares a las caídas del año anterior. El 2019 se observa cómo las tasas de crecimiento son las mayores del periodo analizado. Finalmente, el 2020 vuelven caer motivado principalmente por la crisis sanitaria.
Durante el periodo analizado se observa como el sector tecnológico es el que experimenta mayores tasas de crecimiento medio a excepción del 2018 dónde es el sector sanitario el que presenta mayor tasa de crecimiento.

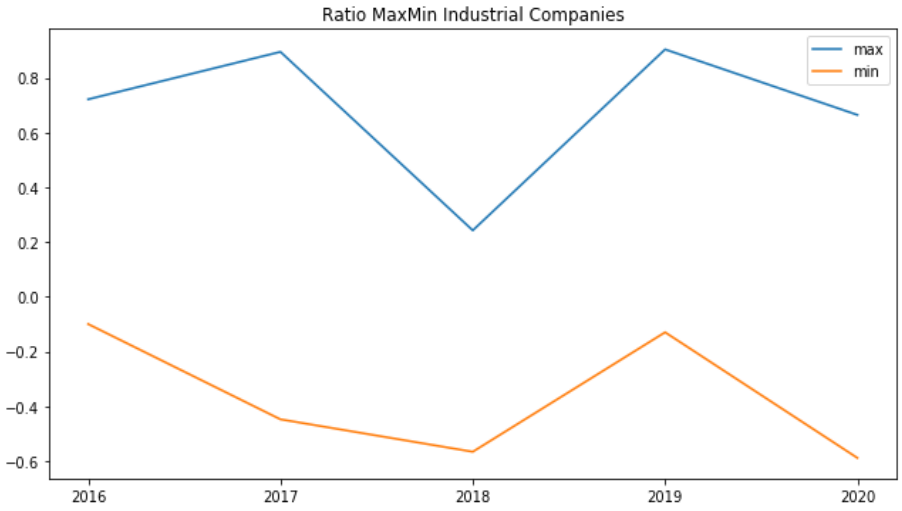
A continuación, se presentan gráficos dónde podemos ver la evolución de la compañía con mayores tasas de crecimiento con la de menor para los diferentes sectores.

Sector tecnológico



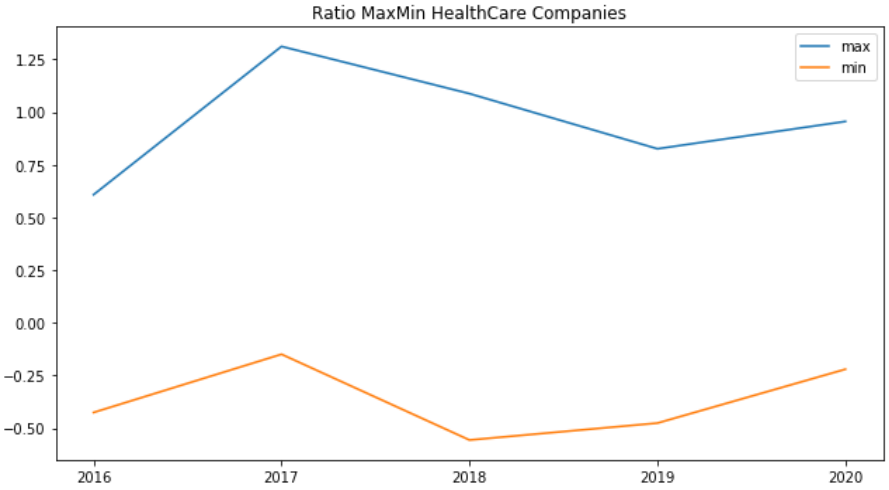

La evolución en el sector tecnológico se observa cómo a lo largo del periodo la diferencia entre la mayor y menor tasa de crecimiento de reduce pasando de una diferencia de 3 puntos a 1.5 puntos.

Sector industrial




Para el sector industrial se puedo ver cómo en 2017 la diferencia se amplía para volverse a reducir en el 2018 y mantenerse estable los dos últimos años de la serie.

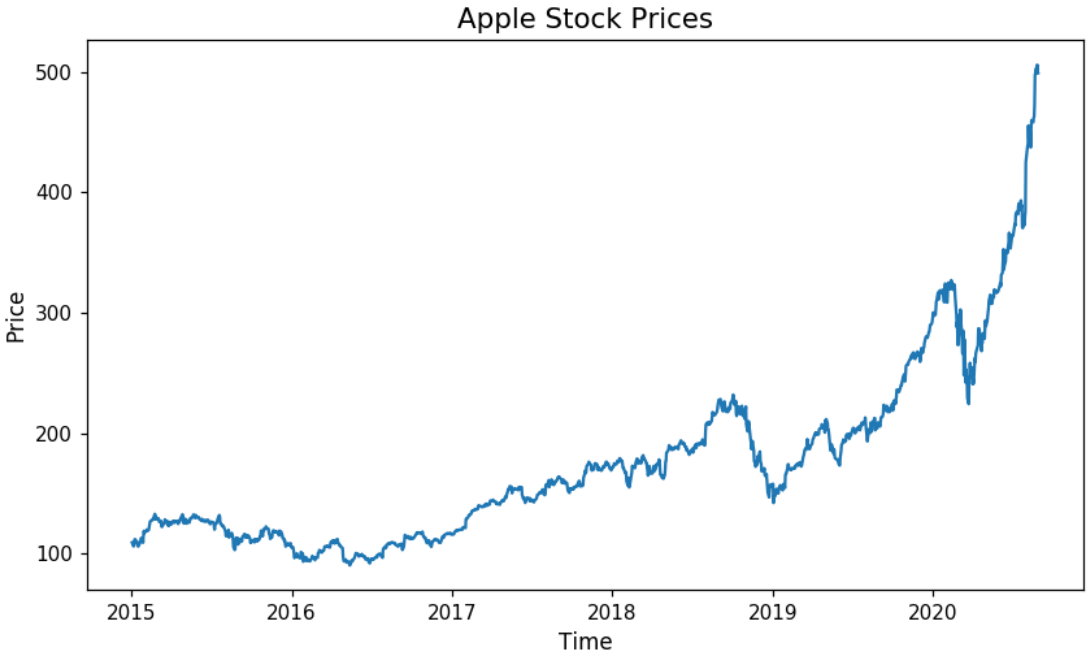
Para el sector sanitario




La evolución de las diferencias en las tasas de crecimiento se mantiene bastante similares, aunque se puede ver un ligero aumento para los años 2017 y 2018.

## Análisis del Stock de Apple:

Inicialmente antes de entrar con la predicción del precio del stock se hace un pequeño análisis exploratorio de la serie temporal.
A continuación, podemos ver la evolución que sigue la serie durante el periodo seleccionado.



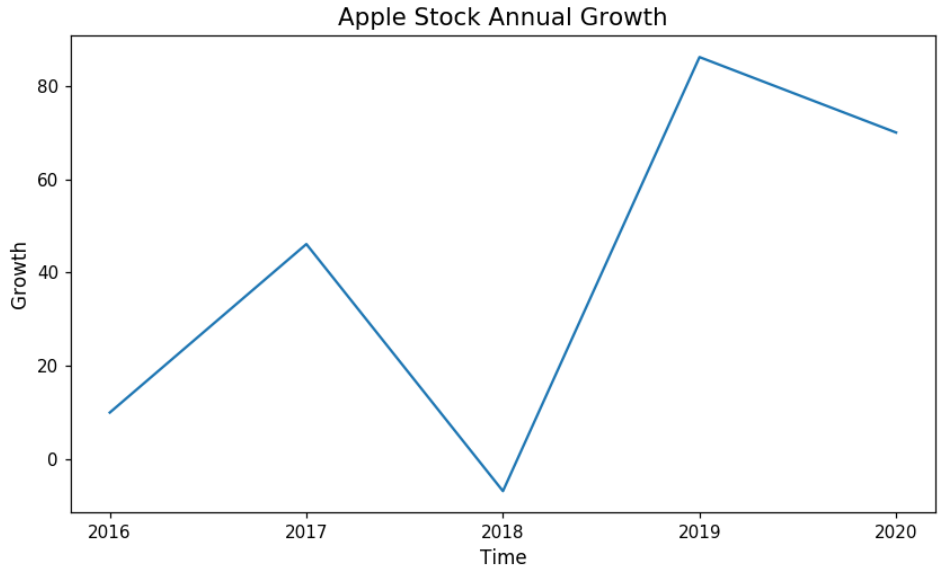
Podemos observar un lento crecimiento hasta finales del 2018 y un crecimiento más pronunciado los dos últimos años de la serie. Sólo se observan caídas pronunciadas a finales del año 2018 y principio del 2020.
El precio del stock en los 5 años observados pasa de 100$ a 500$.

A continuación, se presenta una tabla con las tasas de crecimiento anuales.

| Crecimiento | Año |
|---|---|
| 2016 | 10.032298 |
| 2017 | 46.114657 |
| 2018 | -6.789571 |
| 2019 | 86.160761 |
| 2020 | 70.008521 |

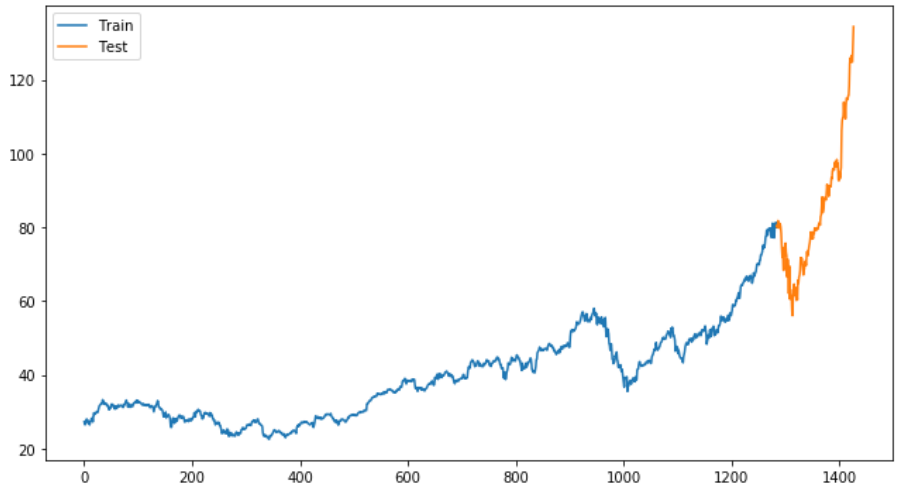
Durante toda la serie podemos ver tasas positivas a excepción del 2018 siendo el 2019 el año que mayor incremento experimenta.

En el próximo gráfico se ilustra las tasas de crecimiento.




## Aplicación de los modelos:

Para el desarrollo de los diferentes modelos he usado el 90% de los datos como entrenamiento y el 10% restante para validar los resultados de la estimación.



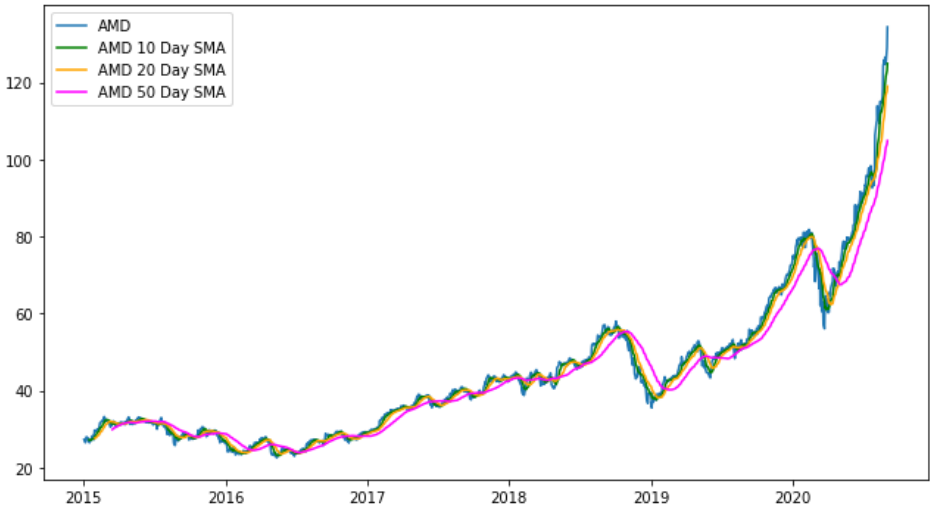

Para observar la precisión de los diferentes modelos he usado las métricas de Error mínimo cuadrático y porcentaje del error cuadrático medio.

MSE $$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

MAPE $$\mathrm{M} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|,$$

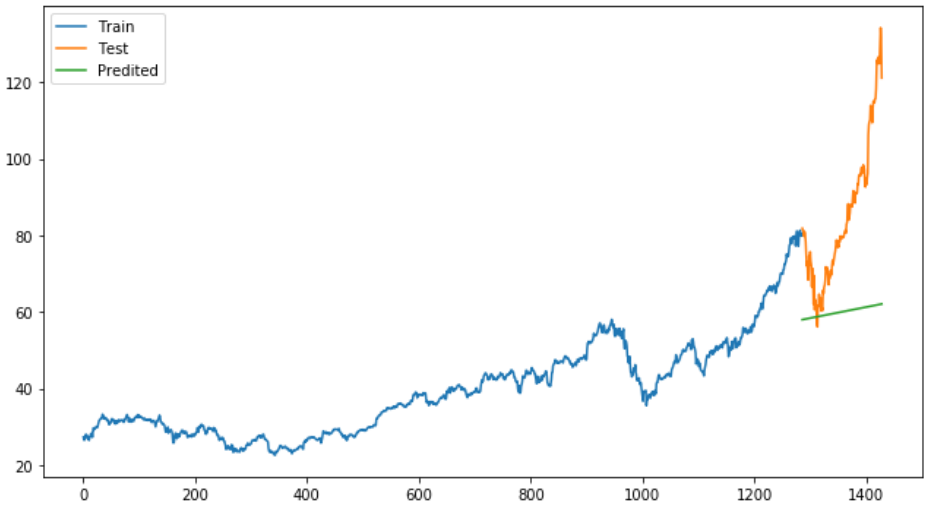
## Medias Móviles

Para las medias móviles he seleccionado ventanas de 10, 20 y 50 días. La que se ajusta más a la serie dando un error menor es la MA10 con un error medio cuadrático de 1.15




## Regresión Lineal

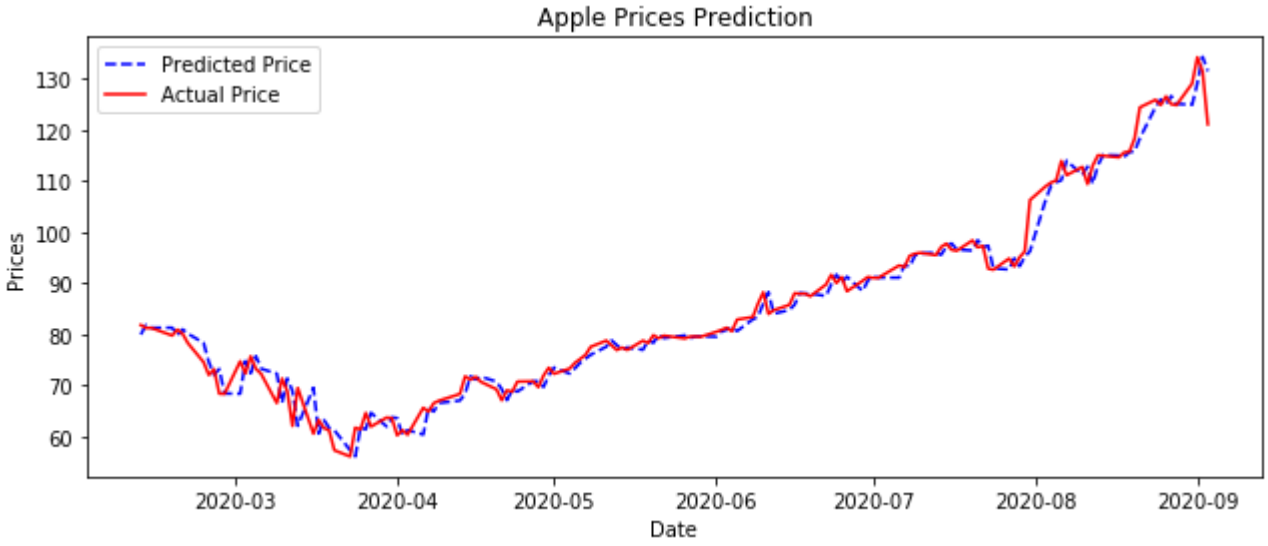
Cómo modelo de base he hecho una regresión lineal. Los resultados obtenidos son francamente malos dado que para los datos de test la serie se dispara y con una regresión lineal es imposible predecir dicho incremento.




## Modelo ARIMA

Para poder aplicar el modelo ARIMA antes se debe analizar si la serie es estacionaria. Para que así sea la serie debe presentar media y varianza constante y que la autocovariancia no dependa del tiempo. Para dicho análisis he usado el test de Dickey-Fuller Aumentado en él que la hipótesis nula representa que la serie no es estacionaria. Debemos observar un valor del p-value inferior a 0.05 para poder rechazar dicha hipótesis y concluir que la serie es estacionaria.
Para transformar la serie en estacionaria primero he eliminado la heteroscedasticidad aplicando el logaritmo. Posteriormente para eliminar la tendencia i la estacionalidad he aplicado la diferencia entre el valor anterior y el valor actual.
Con estas transformaciones se puede observar que la serie pasa a ser estacionaria ya que se obtiene un p-value del ADF test inferior a 0.05.
Una vez la serie es estacionaria para poder determinar los parámetros del modelo ARIMA debemos observar la autocorrelación parcial la cual nos dirá el grado que debemos aplicar a la parte autorregresiva del modelo y la autocorrelación para poder decir el grado que debemos aplicar al error.
Después de dicha observación podemos concluir que los parámetros del modelo que debemos aplicar al modelo son 0 para el coeficiente autorregresivo y 0 para el coeficiente de la media móvil. Con ello se aplica un modelo ARIMA(0,1,0).

Después de entrenar el modelo con los datos de entrenamiento lo validamos con los datos de test y podemos observar que la predicción se acerca bastante a los datos reales. En el siguiente gráfico podemos ver la curva de los precios reales de test con los precios predichos por el modelo.

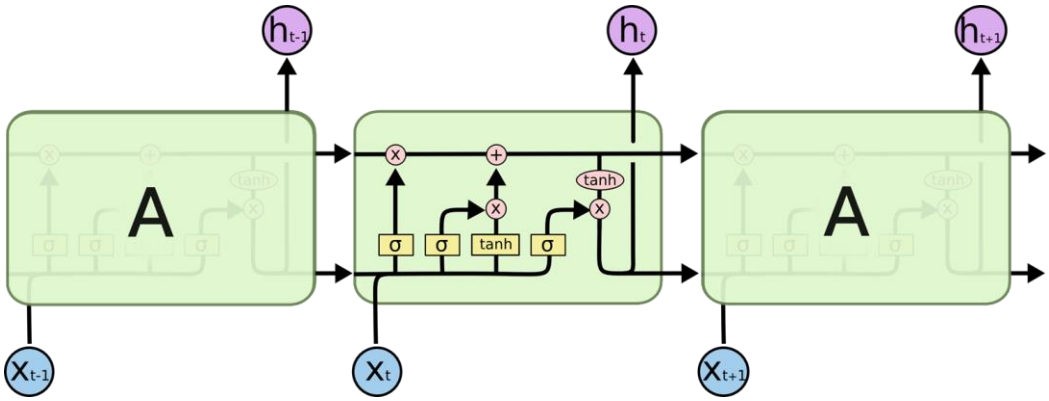



Aplicando las dos métricas antes mencionadas observamos que tenemos un error medio cuadrático bastante grande pero considerablemente inferior al obtenido anteriormente con la regresión lineal.

## Model LSTM (long short term memory)

Finalmente he considerado aplicar este modelo ya que permite hacer una predicción bastante ajustada. Esta red neuronal recurrente integra bucle de realimentación, permitiendo a través de ellos que la información persista durante algunos pasos o épocas de entrenamiento, a través de las salidas de las capas que "incrustan" sus resultados en los datos de entrada.
En el interior de cada neurona hay una función que nos dice que parte debemos olvidar otra en la cual le decimos que parte queremos recordar que es donde aplicamos los datos para ese periodo de tiempo y finalmente la última función es la que no define lo que le vamos a transmitir a la siguiente neurona.

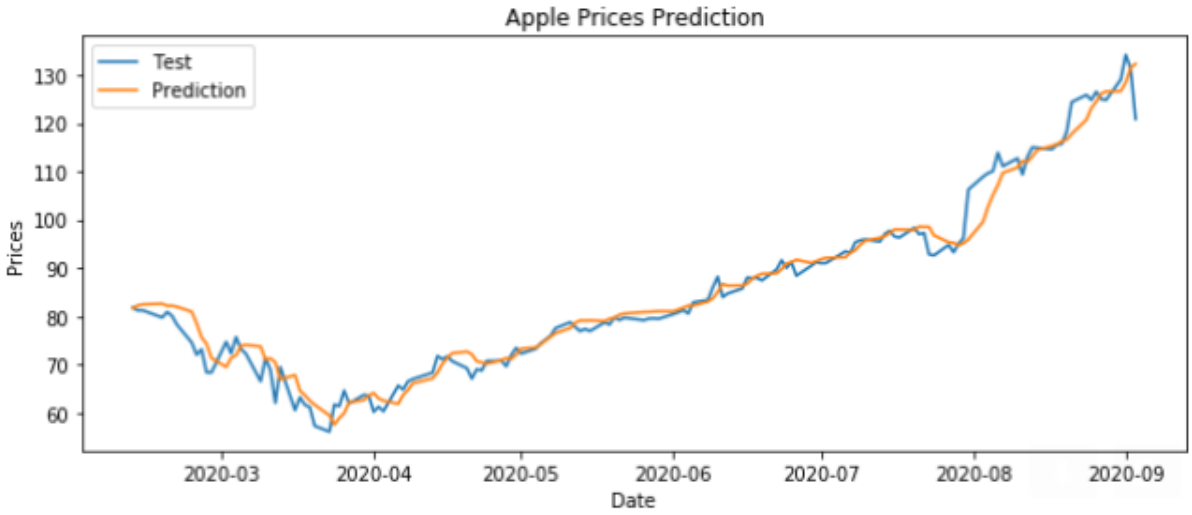
En la siguiente imagen se puede observar las cuatro funciones que actúan dentro de cada neuronal de la red LSTM.




Para poder introducir los datos a dicho modelo previamente debemos preparar dichos datos de la manera que la red los pide. Le debemos pasar una matriz de 3 dimensiones con los datos de la serie

temporal de entrenamiento, los datos que usamos para predecir el siguiente dato y finalmente el número de variables que usamos para la predicción.

En mi caso he dividido la secuencia usando los últimos 15 días para predecir el siguiente. Con ello lo he pasado una matriz al modelo de (1271, 15, 1).
El modelo ha sido construido con un a capa de 50 neuronas y compilado con un optimizador de Adam que es el encargado de actualizar los pesos de red al igual que la función de error de media cuadrática para también actualizar dichos pesos.




Los resultados obtenidos son los que se ajustan más a los datos de validación con un error de 9.28 para el error medio cuadrático y una 2.6 para el porcentaje del error absoluto medio.

## Conclusiones

Para la primera parte del proyecto podemos concluir que las empresas del sector tecnológico son las que más tasas de crecimiento positivo experimentan y las que mayor diferencia presentan entre la tasa más alta de crecimiento y la menor.
Por otro lado, hemos podido ver cómo las empresas del sector sanitario son las menos sensibles a cambios socioeconómicos ya que presentan una curva más estable de crecimiento durante el periodo analizado.
Un último dato importante a remarcar es la mayor convergencia entre las empresas que más crecen respeto a las que menos en el sector tecnológico dado que la tasa de productividad probablemente cae con el tiempo.

En cuanto a la parte de la predicción es evidente que solo con ver los datos de entrenamiento no es suficiente para poder hacer una buena predicción ya que los precios de los stocks evolucionan sin tener un patrón claro al que poder acogerse. Por ello la regresión lineal es un modelo que se aleja mucho de los datos reales.
En cuanto a los dos últimos modelos usados podemos ver que a la predicción mejora sensiblemente sobre todo con la red neuronal recurrente. El hecho de que la red permita recordar lo que ha sucedido anteriormente es de gran ayuda a la hora de predecir los siguientes valores.