Nota de Aceptación

Aprobado por el Comité de Trabajo de Grado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana para optar el título de Ingeniero de Sistemas y computación.

Dr. Hernán Camilo Rocha Niño
Decano de la Facultad de Ingeniería

Dr. Gerardo Mauricio Sarria Montemiranda
Director Carrera Ingeniería de Sistemas y Computación.

Dr. Diego Luis Linares Ospina
Director del Trabajo

Dra. Gloria Inés Álvarez Vargas
Codirectora del Trabajo

Dr. Jorge Finke
Jurado 1

Mag. Jaime Alberto Reinoso
Jurado 2

Pontificia Universidad JAVERIANA Cali

Facultad de Ingeniería y Ciencias

Acta de Correcciones al Proyecto de Grado
Ingeniería de Sistemas y Computación

Fecha: 5/03/2022

Autores: Jeffrey Steven García Gallego; Jose David Gutiérrez Uribe

Nombre del Proyecto de Grado: Predicción del precio de acciones de la bolsa estadounidense utilizando técnicas de aprendizaje automático basadas en datos de análisis técnico y fundamental

Director: Diego Luis Linares Ospina

Codirectora: Gloria Inés Álvarez Vargas

Como indica el artículo 2.27 de las Directrices de Trabajo de Grado, he verificado que los estudiantes indicados arriba han implementado todas las correcciones que los Jurados del Proyecto de Grado definieron que se efectuaran, como consta en el Acta de Calificación correspondiente.

Firma de Director del Proyecto de Grado

Firma de Codirectora del Proyecto de Grado

Pontificia Universidad Javeriana Cali
Facultad de Ingeniería.
Ingeniería de Sistemas y Computación.
Trabajo de Grado.

Predicción del precio de acciones de la bolsa de valores estadounidense utilizando técnicas de aprendizaje automático basadas en datos de análisis técnico y fundamental

Jose David Gutiérrez Uribe
Jeffrey Steven García Gallego

Director: Dr. Diego Luis Linares Ospina
Directora: Dra. Gloria Inés Álvarez Vargas

03/12/2021

Santiago de Cali, 03/12/2021.

Señores
Pontificia Universidad Javeriana Cali.
Dr. Gerardo Mauricio Sarria
Director Carrera de Ingeniería de Sistemas y Computación.
Cali.

Cordial Saludo.

Por medio de la presente nos permitimos informarles que los estudiantes de Ingeniería de Sistemas y Computación Jose David Gutiérrez Uribe (cod: 8934841) y Jeffrey Steven García Gallego (cod: 8935155) trabajan bajo mi dirección en el trabajo de grado titulado "Predicción del precio de acciones de la bolsa de valores estadounidense utilizando técnicas de aprendizaje automático basadas en datos de análisis técnico y fundamental".

Atentamente,

Dr. Diego Luis Linares Ospina

Dra. Gloria Inés Álvarez Vargas

Santiago de Cali, 03/12/2021.

Señores
Pontificia Universidad Javeriana Cali.
Dr. Gerardo Mauricio Sarria
Director Carrera de Ingeniería de Sistemas y Computación.
Cali.

Cordial Saludo.

Nos permitimos presentar a su consideración el trabajo de grado titulado "Predicción del precio de acciones de la bolsa de valores estadounidense utilizando técnicas de aprendizaje automático basadas en datos de análisis técnico y fundamental" con el fin de cumplir con los requisitos exigidos por la Universidad para llevar a cabo el trabajo de grado y posteriormente optar al título de Ingeniero de Sistemas y Computación.

Al firmar aquí, damos fe que entendemos y conocemos las directrices para la presentación de trabajos de grado de la Facultad de Ingeniería aprobadas el 26 de Noviembre de 2009, donde se establecen los plazos y normas para el desarrollo del anteproyecto y del trabajo de grado.

Atentamente,

Jose David Gutiérrez Uribe
Código: 8934841

Jeffrey Steven García Gallego
Código: 8935155

Abstract

La predicción de acciones de la bolsa de valores ha sido una actividad que se ha venido realizando de distintas maneras desde la aparición de los mercados de acciones. En la actualidad, con la presente tendencia de la aplicación del aprendizaje automático en distintos campos, se avanza en la investigación de modelos de aprendizaje automático, atributos a tener en cuenta y datos utilizados para realizar una predicción sobre el precio o la volatilidad de una acción en específico. En lo que respecta a los datos utilizados, actualmente, los principales son los datos basados en análisis técnico, análisis fundamental y análisis de sentimientos. Según la literatura, la mayor parte de los estudios e investigaciones en este campo se basan en datos de análisis técnico. Por lo tanto, en este proyecto se buscó explorar el comportamiento de distintas técnicas de aprendizaje automático basadas en datos de análisis técnico y análisis fundamental utilizando un conjunto de acciones de la bolsa de valores estadounidense, pertenecientes a uno de los índices más importantes el S&P 500. También se exploraron distintos métodos como Análisis de Componentes Principales (PCA por sus siglas en ingles), ventana deslizante y una selección de atributos a través de la literatura. En este estudio se pusieron a prueba estos modelos a través de métricas como la del Error Cuadrático Medio y el Error Absoluto Medio. Estos modelos se sometieron a una prueba que simula una situación real de inversión conocida como backtesting, en el cual se hace uso de la estrategia de comprar bajo y vender alto. Se encontró que la técnica que presentó menor error para análisis técnico fue SVR y para análisis fundamental fue MLP. Sin embargo, en backtesting RF fue la que mayores beneficios obtuvo tanto para análisis técnico como para análisis fundamental. Se plantea que posiblemente las métricas de error en la predicción del precio de una acción no resultan lo suficientemente expresivas como para determinar el desempeño de un modelo en una situación real.

Palabras Clave: Bolsa de valores, acción, análisis fundamental, análisis técnico, aprendizaje automático, predicción.

Índice general

Introducción

El mercado de valores se define como un mercado público donde empresas ponen a la venta acciones que representan una parte de su valor total, y los inversionistas compran estas acciones esperando recibir un beneficio a corto, mediano o largo plazo en relación al comportamiento del precio de la acción de una empresa determinada. La importancia de este concepto reside en la gran capacidad que tiene de mover la economía de una nación, beneficiando en el proceso a los actores en este mercado, de los cuales -desde la visión de este proyecto- se cubren únicamente a empresas e inversionistas.

Con el tiempo se desarrollaron distintas técnicas de análisis con bases matemáticas para identificar el comportamiento de una acción obteniendo así información de gran relevancia para establecer futuras estrategias de inversión sobre dicha acción. Proceso el cual dejo de ser manual a pasar a desarrollarse en las máquinas de cómputo progresivamente, hasta el punto en el que a día de hoy es posible realizar predicción sobre distintos atributos de una acción, como puede ser el precio, utilizando aprendizaje de maquina.

Capítulo 1

Descripción del Problema

1.1. Planteamiento del Problema

El mercado de valores ha tenido sus altos y bajos en años como los de la gran depresión la inversión en la bolsa no era algo a lo que todos podían acceder, el proceso de invertir consumía más tiempo, era más caro, riesgoso y menos accesible en aquella época en comparación a lo que es en la actualidad. El surgimiento de nuevas tecnologías de cómputo elevó los niveles de participación en este mercado y redujo los costos en el proceso, cambios que llevaron a un nuevo paradigma de inversión [1].

Existen muchos factores que los inversionistas tienen en cuenta al elegir sobre qué acción realizar operaciones de compra o venta. De tal manera, existen dos corrientes de pensamiento predominantes para realizar análisis y predicción del precio de acciones a futuro:

El análisis fundamental es un método de evaluar una acción a través de medidas intrínsecas de ésta [2]. Estas medidas están relacionadas a la economía de la empresa y la condición de la industria en la que se encuentra, siendo variables como ganancias, gastos, pasivos, activos y ratios financieros, las más usadas para este tipo de análisis.

Por otro lado, el análisis técnico toma datos propios de transacciones realizadas en el mercado de valores, como puede ser el precio de la acción, el volumen o monto de dinero invertido en esa transacción, entre otros. El análisis técnico trata de buscar patrones y tendencias que indiquen el comportamiento de la acción a futuro [2].

Una de las principales problemáticas en las inversiones en la bolsa de valores es que según las estadísticas el 90 % de los inversionistas pierde el dinero en este mercado [3]. Es un negocio muy complejo, que requiere de muchas variables e información a tener en cuenta a la hora de establecer las estrategias de inversión basadas en los debidos análisis. De tal modo, es relativamente poca la participación de personas con pocos conocimientos sobre estos dominios en el mercado de valores o con poco presupuesto para invertir [4].

Muchas personas piensan que la bolsa de valores es un sistema caótico y casi imposible de predecir. Debido a esto existe cierta desconfianza de algunos inversionistas respecto al análisis técnico debido a que observando únicamente el comportamiento del precio no se puede obtener información de gran importancia sobre el estado económico de una empresa [5].

Son los avances computacionales los que originaron la revolución y la mayor participación en los mercados de valores, estando las transacciones de compra y venta a un clic de ser ejecutadas. Al comienzo fue la digitalización y computarización de las acciones, dando también un acceso más amplio a toda la población a través de internet, mejorando la participación y expandiendo este mercado. Posteriormente, se empezaron a implementar las técnicas de aprendizaje automático para

comprender mejor cómo se comporta el mercado, procesar cantidades de datos cada vez mayores e identificar patrones para así finalmente realizar una predicción sobre el precio futuro de la acción en base a datos de alguna de las dos corrientes predominantes de análisis bursátil.

La mayor parte de las investigaciones en el tema han sido realizadas en base a análisis técnico, debido a que este presenta un mayor volumen de datos accesibles y gratuitos desde internet. Sucede lo contrario con el análisis fundamental, siendo estos datos publicados trimestralmente por las empresas, y contando con una accesibilidad restringida respecto a los datos históricos en algunas ocasiones por medio suscripciones o pagos. Por lo tanto, resulta mucho más atractivo centrarse en este tipo de análisis, además de ser preferido por los inversionistas para beneficios a corto plazo. Sin embargo, se dejaría de lado la discusión existente sobre los resultados obtenidos tras utilizar análisis técnico o fundamental para la predicción de acciones en la bolsa. De este modo, este proyecto se centra en la implementación de modelos utilizando técnicas de aprendizaje automático, basados en datos tanto de análisis fundamental como de análisis técnico, observando, analizando y comparando los desempeños de cada modelo según el tipo de análisis en el que fue basado.

1.1.1. Formulación

¿Cuál sería el comportamiento de técnicas de aprendizaje automático aplicado al análisis técnico, fundamental en la predicción del precio de un conjunto de acciones de la bolsa de valores estadounidense?

1.1.2. Sistematización

- ¿Cuáles serán las fuentes de las que se obtendrán los datos de análisis técnico y análisis fundamental?
- ¿Cómo se prepararán los datos de entrada para los modelos?
- ¿Cuáles serán las variables o atributos que se utilizaran para la construcción de los modelos?
- ¿Cuáles técnicas de aprendizaje automático se tendrán en cuenta para realizar la predicción con base en análisis técnico y análisis fundamental?
- ¿Como se desarrollaría la implementación para cada uno de los modelos?
- ¿Cómo evaluar y comparar el comportamiento de cada modelo teniendo en cuenta los resultados obtenidos con base al tipo de análisis?

1.2. Objetivos

1.2.1. Objetivo General

Analizar el comportamiento de algunas técnicas de aprendizaje automático aplicado al análisis técnico y fundamental en la predicción del precio de un conjunto de acciones de la bolsa de valores estadounidense.

1.2.2. Objetivos Específicos

- Seleccionar un conjunto de fuentes que contengan datos de análisis técnico y análisis fundamental.
- Preparar el conjunto de datos para la implementación de los modelos.
- Identificar variables o atributos que influyen positivamente en la ejecución de los modelos.
- Seleccionar tres técnicas de aprendizaje automático para implementar los modelos de predicción basado en datos de análisis técnico y análisis fundamental.
- Estimar los parámetros que mejorarán el desempeño e implementar los modelos.
- Evaluar el comportamiento de cada modelo teniendo en cuenta el tipo de dato con los que están siendo entrenados (datos técnicos y fundamentales).

1.3. Justificación

Debido a la gran repercusión económica que trae el mercado de valores para un país, se han llevado a cabo gran cantidad de investigaciones al respecto, en las que se ven involucrados inversionistas, estados, científicos de la computación, matemáticos, entre otros.

Estas investigaciones se han centrado sobretodo en cómo generar confiabilidad en el proceso de inversión a través de estrategias de inversión o predicción de acciones guiadas por técnicas de aprendizaje automático. Se debe tener en cuenta que los inversionistas no son únicamente personas, también lo pueden ser entidades públicas o privadas. Un ejemplo de esto, son los fondos de inversión, los cuales pueden tomar dineros públicos o privados dependiendo si el fondo pertenece al estado o no. Por lo tanto, en este caso se estaría poniendo en riesgo el capital de todo un país, un tema delicado, que debe ser totalmente transparente [6, 7].

El proceso de inversión es visto como una actividad de alto riesgo, dado que existen diversos factores que afectan negativamente los posibles resultados de una inversión, algunos intrínsecos al proceso en sí, como puede ser la probabilidad -siempre existente- de no generar beneficios, o incluso generar pérdidas. Otros factores también presentes son los que tienen que ver con las emociones humanas. Al momento de invertir las personas tienden a dejarse llevar por distintas emociones como lo pueden ser miedo, incertidumbre, exaltación, ira, tristeza, entre otras. Esto puede conducir a estos inversionistas a llevar a cabo análisis o estrategias de inversión muy inestables aumentando así el riesgo de pérdida. Por estos motivos, es mucho más confiable basarse en la información que otorgan los modelos dedicados a estas tareas, puesto que así el proceso es mucho más formal, dado que sus resultados pueden ser medibles y mejorables. Estudios han mostrado una clara diferenciación en el desempeño obtenido por aplicación de técnicas de aprendizaje automático frente al desempeño que podría generar un inversionista promedio.

Los dos tipos de análisis que se tratan en este proyecto son diferentes entre sí. Existe cierta discusión entre los inversionistas sobre cuál es mejor utilizar dependiendo de diferentes situaciones.

La importancia de esta discusión reside en qué tipo de análisis genera mejores resultados respecto al otro o, mirando desde otro punto de vista, si es posible mejorar los resultados independientes realizando algún tipo de complementación entre los dos tipos de análisis.

Este proyecto busca aportar a la solución de las problemáticas descritas, relacionadas con el proceso de inversión, el impacto de la computación en este campo y la diferenciación de los dos tipos de análisis tratados en este proyecto. Por estos motivos con el proyecto se busca observar distintas técnicas de aprendizaje automático con base en los dos tipos de análisis bursátil, se dará un acercamiento a la comparación del desempeño de cada uno de estos. Lo que dará una visión desde otra perspectiva a trabajos previos explorando nuevas posibilidades de los modelos y del comportamiento de los dos tipos de análisis. De esta manera se aportará al avance y entendimiento en las investigaciones sobre este campo de las inversiones en el mercado de valores y, por lo tanto, se contribuirá a la construcción de una solución para los problemas mencionados que derivan del proceso de inversión.

1.4. Delimitaciones y Alcances

- Se emplearan 3 técnicas de aprendizaje automático.
- Se trabajará con acciones individualmente no con la conformación de portafolios.
- Se ejecutarán modelos que realizan predicciones sobre una acción sin implementar un sistema de trading automático.
- El proyecto solo abarca un conjunto de acciones de la bolsa de valores estadounidense pertenecientes al índice S&P 500.

1.4.1. Entregables

- Documento con toda la información relativa a las técnicas de aprendizaje automático utilizadas y sus respectivos resultados.
- Código fuente de los modelos implementados.

Capítulo 2

Marco teórico

2.1. Análisis bursátil

El análisis bursátil representa los métodos analíticos para obtener información relevante de un mercado de valores que pueda ayudar a la toma de decisiones de inversión. Se destacan principalmente dos tipos de análisis bursátil clásico, los cuales son: análisis técnico y análisis fundamental.

2.1.1. Análisis técnico

El análisis técnico es un tipo de análisis bursátil, que se encarga del estudio de los movimientos pasados de los precios con el objetivo de predecir los movimientos futuros de los precios a partir de los movimientos del pasado. Charles Dow fue el que sentó las bases del análisis técnico, Dow pensaba que las expectativas de la economía nacional se traducían en órdenes de mercado que hacían subir o bajar los precios de las acciones a largo plazo, normalmente antes de la evolución real de la economía, La Teoría de Dow supone que toda la información está descontada en las medias, por lo que no se necesita ninguna otra información para tomar decisiones comerciales [8].

El análisis técnico se enfoca principalmente en el estudio de datos históricos del mercado. Esta información es utilizada para realizar decisiones que afectan a los negocios de inversión e identificar oportunidades de inversión analizando tendencias estadísticas obtenidas por la actividad del mercado, como lo son el precio y el volumen [9].

Este tipo de análisis tiene su raíces en la teoría básica de la economía, donde en [9] establecen una serie de suposiciones como:

- Los precios de las acciones están determinados únicamente por las interacciones de oferta y demanda.
- Los precios de las acciones suelen moverse según tendencias.
- Intercambios entre oferta y demanda causa tendencias inversas.
- Los intercambios entre oferta y demanda pueden ser identificados en los gráficos técnicos.
- Los patrones que se encuentran en los gráficos tienden a repetirse.

2.1.2. Análisis fundamental

El análisis fundamental es el estudio del potencial financiero que tiene una compañía o empresa, este estudio se basa en datos históricos consignados en reportes periódicos; en el sector de la empresa

y su posicionamiento en la industria, así como también influyen factores administrativos, dividendos y capitalización para poder identificar un potencial de crecimiento a futuro. La combinación de estos datos representa información que no está directamente relacionada con el precio de la acción. Esta información obtenida después del proceso es utilizada para definir el momento y el monto de la inversión, así como también es utilizada como factor de comparación entre distintas acciones [10].

Este tipo de análisis bursátil determina el valor intrínseco de la empresa, midiendo de este modo su "fuerza", la eficiencia de su gestión y las oportunidades de negocio que ofrece. En el análisis fundamental se utilizan expresiones analíticas como lo son los ratios e indicadores los cuales se encargan de representar relaciones numéricas entre distintas variables haciendo mucho más compresible y manejable el análisis.

Dada la naturaleza no estructurada de los datos de variables fundamentales, la automatización del análisis fundamental resulta complejo y requiere de altos recursos de cómputo y tiempo. Sin embargo, desde el auge del aprendizaje automático los investigadores han podido realizar predicciones con este tipo de datos [11]. Se afirma que el análisis fundamental es útil para la predicción del movimiento de acciones a largo plazo pero no para predecir cambios en el precio en el corto plazo [12].

2.2. Predicción del mercado de valores

La predicción de acciones de la bolsa de valores es el proceso por el cuál se determina el valor futuro de la acción de una compañía o varias compañías (portafolio). Este ha sido un tema ampliamente debatido, lo que ha generado distintas posiciones en el ámbito académico.

En [13][14] y [15], los autores adoptan una posición en la cual destacan que el mercado de valores es estocástico, y, por lo tanto, es impredecible. A partir de esto surgieron las dos hipótesis de la predicción de acciones más famosas, llamadas, *random-walk hypothesis* (RWH) y *efficient market hypothesis* (EMH) [11].

La primera hipótesis *random-walk hypothesis*, desde la suposición de que el precio de una acción es fundamentalmente estocástico, se declara que toda iniciativa o esfuerzo por intentar predecir el precio futuro de una acción fallará sin lugar a dudas [16][11].

La segunda hipótesis la cual tiene la misma suposición de que el mercado es aleatorio y, por lo tanto no es predecible es la famosa EMH (efficient market hypothesis) planteada por Fama [13]. Esta dice que el mercado de valores es ïnformacionalmente eficiente", lo que significa que los precios de las acciones reflejan toda la información de la compañía, además hipotetiza que el mercado siempre comercializa al valor justo y correspondiente de la compañía, haciendo así imposible para los inversionistas comprar acciones devaluadas o vender acciones con precios inflados [11][17].

Claramente es observable que los mercados pueden ser ineficientes, y unos más que otros. Un mercado ineficiente es en el que el precio de sus acciones no reflejan su valor real. Por lo que, aceptar la veracidad de EMH es complicado, dado que no siempre ocurre que el precio de la acción refleje toda la información pública y privada de una compañía. Por lo tanto, en [14], el autor categoriza la hipótesis en eficiencia fuerte, semi-fuerte y débil teniendo en cuenta estos hechos [17].

Teniendo en cuenta lo mencionado anteriormente, existe posibilidad de predecir el mercado

basándose en conocimiento técnico y fundamental acerca del mercado. A pesar de estas hipótesis, una gran parte de los académicos dedicados a esta tarea sostienen que existen mercados, particularmente los mercados emergentes, que proporcionan mejores resultados que realizando selección aleatoria [18]. En [19] se sostiene que el mercado de valores es predecible hasta cierto punto cuando se tienen en cuenta comportamientos económicos y socio-económicos desde un punto punto de vista teórico-financiero.

2.3. Variables calculadas

A continuación se presentan las fórmulas utilizadas para calcular los atributos adicionales, que se obtienen a través de los atributos básicos o base.

2.3.0.1. Análisis fundamental

En la Tabla 2.1 se presentan las fórmulas utilizadas para calcular los atributos adicionales para datos de análisis fundamental, que se obtienen a través de los atributos básicos o base.

Tabla 2.1: Fórmulas de ratios calculados. Nota: Trailing Twelve Months (TTM), representa la sumatoria de una variable determinada durante doce meses consecutivos.

Ratio	Fórmula	Descripción
P/E	$\mathbf{P/E} = \frac{StockPrice}{EPS(TTM)}$	Price to Earnings Ratio muestra la relación del precio de la acción de una compañía con sus ganancias por acción. Provee información que le permite a los inversionistas determinar si una acción está devaluada o sobrevaluada en relación a otras del mismo sector.
PEG	$\mathbf{PEG} = \frac{P/E}{EPS(TTM)Growth\,\%}$	Projected Earnings Growth compensa la falta de visión a futuro de la que carece P/E, de este modo los analistas pueden estimar la tasa de crecimiento a futuro teniendo en cuenta su tasa de crecimiento histórica.

FCF	**FCF** = Net Cash From Operations (TTM) + Net Change Capital Expenditures (TTM)	Free Cash Flow representa el dinero resultante de una compañía una vez se pagan gastos operativos y gastos en capital ('operating expenses' y 'capital expenditures'). Determina básicamente la capacidad económica que tiene una compañía para pagar a sus accionistas.
P/FCF	$\mathbf{P/FCF} = \frac{StockPrice}{FCFpershare(TTM)}$ $FCFpershare(TTM) = \frac{FCF(TTM)}{SharesOutstanding}$	Price to Free Cash Flow es muy similar a lo que representa el ratio FCF, sin embargo, esta se considera una medida más exacta.
P/S	$\mathbf{P/S} = \frac{MarketCap}{Revenue(TTM)}$ Market Cap = Stock Price * Average Basic Shares Outstanding	Price to Sales Ratio es un indicador que ayuda a determinar el valor real o justo de una acción utilizando el Market Cap y el retorno de una compañía.
P/B	$\mathbf{P/B} = \frac{StockPrice}{BookValuepershare}$ $BookValuepershare = \frac{TotalAssets-TotalLiabilities}{AverageBasicSharesOutstanding}$	Price to Book este indicador ayuda a los inversionistas a determinar cuando una acción está devaluada o sobrevaluada en relación a su 'book value' y en comparación con el P/B de otra acción.
ROE	$\mathbf{ROE} = \frac{NetIncome(TTM)}{\frac{CurrentBookValue+PastYearBookValue}{2}} * 100$	Return on Equity representa la tasa de retorno que un accionista recive por la porción que invirtió en tal compañía. Este ratio mide qué tan buenos retornos genera una compañía para sus inversionistas.

2.3.0.2. Análisis técnico

En la Tabla 2.2 se presentan las fórmulas utilizadas para calcular los atributos adicionales para datos de análisis técnico, que se obtienen a través de los atributos básicos o base.

Tabla 2.2: Formulas indicadores técnicos

Indicador	Fórmula	Descripción
momentum	$price(i) - price(i-1)$ donde i=precio actual	El Momentum es una medida de la aceleración y desaceleración de los precios. Indica si los precios están aumentando a un ritmo creciente o disminuyendo a un ritmo decreciente
BOP	$\mathbf{B}OP = \frac{(Close - Open)}{(High - Low)}$	Es un oscilador que mide la fuerza de la presión de compra y venta.
EMA	$price(t) * k + EMA(y) + (1 - k)$ Donde: $t = preciodehoy$ $y = preciodeayer$ $N = numerodediasdelaEMA$ $k = 2/(N + 1)$	Una media móvil exponencial (EMA) es un tipo de media móvil que otorga un mayor peso e importancia a los puntos de datos más recientes.
MACD	macd=12periodEMA - 26periodEMA signal=9 Dias MA de MACD MA=moving average EMA=exponential moving average	Moving average convergence divergence (MACD) es un indicador de momentum que muestra la relación entre dos medias móviles del precio, se utiliza principalmente para operar tendencias.
BOLLINGER BANDS	upper=$MA(precioCierre, 20) + (2 * \sigma[precioCierre, 20])$ lower=$MA(PrecioCierre, 20) - (2 * \sigma[precioCierre, 20])$	Las bandas de Bollinger constan de tres líneas. La banda central es una media móvil simple (normalmente de 20 periodos) del precio. Las bandas superior(upper) e inferior(lower) son desviaciones estándar (normalmente 2) por encima y por debajo de la banda media. Las bandas se ensanchan y estrechan cuando la volatilidad del precio es mayor o menor, respectivamente.

2.4. Técnicas de aprendizaje automático utilizadas

Para este proyecto se decidió utilizar 3 técnicas de aprendizaje automático tanto para los datos de análisis técnico, así como para los datos de análisis fundamental. Las técnicas empleadas en los modelos fueron bosques aleatorios, Máquinas de vectores soporte y Perceptrón multicapa. Estas tres técnicas fueron seleccionadas con base a la información recopilada de distintos artículos en [11], donde destacan la participación de estas tres técnicas en la mayoría de los artículos revisados, siendo ANN y SVM los más utilizados, puesto que según [20][21], consiguen una generalización potencial mucho mayor que sus contrapartes.

2.4.1. Bosques aleatorios

El algoritmo de Bosques aleatorios es un algoritmo de aprendizaje supervisado basado en árboles de decisión, que puede ser utilizado tanto en tareas de clasificación como en tareas de regresión. Para generar los resultados el algoritmo realiza el promedio del valor de salida de cada uno de los árboles de decisión, en caso de que se trate un problema de regresión. De ser un problema de clasificación, el resultado se obtiene por medio de votación, es decir, la mayoría es la que decide a qué clase pertenece la observación [11].

Una característica importante en los bosques aleatorios es la ejecución del *bagging*, este proceso reduce la varianza de los estimadores base al aleatorizar, por ejemplo, cómo cada árbol crece para posteriormente promediar las predicciones y así lograr reducir el error de generalización. Este proceso es ampliamente utilizado para lidiar con el 'sobre-entrenamiento' [22].

Bosques aleatorios se basa en el modelo de árboles de decisión, los cuales son utilizados tanto para tareas de clasificación, así como de regresión. Dos características principales en la construcción de Bosques Aleatorios o Random Forests es el 'bagging' y la selección aleatoria en cada nodo.

2.4.2. Máquinas de vectores soporte

Las máquinas de vectores de soporte son modelos de aprendizaje supervisado con algoritmos de aprendizaje asociados que analizan los datos utilizados para la clasificación y el análisis de regresión. En la regresión de vectores de soporte, la línea recta que se requiere para ajustar los datos se denomina hiperplano.

Los hiperplanos son límites de decisión que se utilizan para predecir la salida continua. Los puntos de datos a cada lado del hiperplano que están más cerca del hiperplano se llaman vectores de soporte. Estos se utilizan para trazar la línea necesaria que muestra la salida predicha del algoritmo.

La regresión de vectores de soporte es un algoritmo de aprendizaje supervisado que se utiliza para predecir valores discretos. La regresión de vectores de soporte utiliza el mismo principio que las SVM. La idea básica de SVR es encontrar la línea de mejor ajuste. En SVR, la línea de mejor ajuste es el hiperplano que tiene el máximo número de puntos.

A diferencia de otros modelos de regresión que tratan de minimizar el error entre el valor real y el predicho, el SVR trata de ajustar la mejor línea dentro de un valor umbral. El valor del umbral

es la distancia entre el hiperplano y la línea límite. [23]

2.4.3. Perceptrón multicapa

El Perceptrón Multicapa o MLP, por sus siglas en inglés, es un modelo que agrega perceptrones, los cuales representan una unidad de cómputo asemejándose a una neurona, contando con una entrada, una salida y una función de activación. Un MLP está compuesto de una capa de entrada, una o más capas intermedias llamadas *capas escondidas* y una capa de salida. Las capas cercanas a la capa de entrada normalmente se les denomina capas inferiores y las más cercas de la capa de salida se les conoce como capas superiores. Cada capa está conectada completamente con la capa siguiente, exceptuando la capa de salida [24]. Las unidades de entrada no desempeñan ningún papel activo en el procesamiento del flujo de información, ya que se limitan a distribuir las señales a las unidades de la primera capa oculta. Todas las unidades ocultas funcionan de forma idéntica y la unidad de salida es una versión más simple de una unidad oculta. En un MLP, cada unidad oculta transforma las señales de la capa anterior en una señal de salida, que se distribuye a la capa siguiente. Cada unidad oculta tiene una función de activación, generalmente no lineal.

La salida de una unidad oculta está determinada por la suma ponderada de las señales de la capa anterior, que es transformada por la función de activación. En la unidad de salida la función de activación es la función de identidad [25].

2.5. Métricas utilizadas

Las métricas de error utilizadas en este proyecto nos ayudan a realizar comparaciones entre los resultados obtenidos, valorando el error que estos generan en comparación con el valor real que se intenta predecir.

Al tratarse de un problema de regresión, se eligieron las métricas *Root Mean Square Error* (RMSE) y *Mean Absolute Error* (MAE), las cuales están entre las más utilizadas para actividades de predicción de precios de acciones [11]. Estos errores conservan la unidad, que en este caso se trata de dólares estadounidenses (USD).

Estas métricas de error están correlacionadas, es decir, si uno aumenta, el otro también lo haría. Siendo el RMSE el que crece a mayor medida, puesto que utiliza el error al cuadrado, penalizando mucho más el error cometido. Cuando se calculan ambas métricas, el RMSE es, por definición, mayor que el MAE. Algunos estudios no sostienen que el RMSE sea más efectivo que el MAE. En cambio, sugieren una combinación de métricas tales como RMSE y MAE, a menudo, para evaluar el rendimiento del modelo [26].

El error cuadrático medio (RMSE) se ha utilizado como métrica estadística estándar para medir el rendimiento de los modelos en los estudios de meteorología, calidad del aire y clima, entre otros. El error medio error absoluto (MAE) es otra medida útil ampliamente utilizada en las evaluaciones de modelos [26].

2.5.1. RMSE

Root Mean Square Error (Error Cuadrático Medio) indica qué tan dispersos se encuentran los resultados de la media, representando la desviación estándar de las diferencias entre valores predichos y reales. Tiene la capacidad de ser interpretado con facilidad, dado que conserva las unidades de los datos originales.

Para calcular el RMSE se usa la siguiente fórmula:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Donde
$\hat{y}_i$ = valor de salida real
y_i = valor de salida predicción
n=número de puntos de datos

2.5.2. MAE

Mean Absolute Error (Error Medio Absoluto), como su nombre indica, representa la magnitud media de error generada por el modelo de regresión. Se obtiene al calcular la diferencia absoluta entre la predicción del modelo y el valor real. Al igual que RMSE, también puede ser fácilmente interpretado y las unidades de su resultado son las mismas que las unidades de los datos originales.

Para calcular el MAE se usa la siguiente fórmula:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i|$$

Donde
$\hat{y}_i$ = valor de salida real
y_i = valor de salida predicción
n=número de puntos de datos

2.6. Trabajos relacionados

Algunos trabajos relacionados que han servido como orientación para nuestro trabajo han sido los siguientes. Empezando con el artículo **Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques**, donde se utilizaron cuatro modelos de aprendizaje automático (ANN, SVM, RF, Naive-Bayes) para predecir el movimiento y el precio de dos acciones y dos índices del mercado de valores hindú. Se tomaron dos aproximaciones para los datos de entrada, siendo el primero índices obtenidos de datos técnicos y el segundo representar estos índices como una tendencia. Donde los resultados experimentales sugieren que el mejor desempeño entre todos los modelos fue conseguido con el algoritmo Random Forest (RF), también se llega a la conclusión de que todos los modelos mejoran con el segundo

acercamiento en la entrada de los datos, es decir, conviertiéndolos en datos determinísticos de tendencia [27].

Comparing Technical and Fundamental indicators in stock price forecasting. Este trabajo evalúa si el análisis fundamental o el análisis técnico produce mejores resultados a la hora de predecir los precios de las acciones con modelos de aprendizaje automático, además considera si el uso combinado de los dos tipo de análisis bursátil generaría un mejor resultado. En este trabajo también se hace uso de empresas del S&P 500 donde se tomaron 140 acciones dentro de este índice. Se encontró que los modelos que utilizan indicadores basados en el análisis fundamental superan a los que utilizan indicadores de análisis técnico, con un nivel de rendimiento superior que varía según los sectores a los que pertenezcan las empresas. Además, en más del 95 % de los casos, la utilización de indicadores combinados da lugar a un RMSE inferior al de los indicadores fundamentales o técnicos por separado [28].

Machine Learning for Stock Prediction Based on Fundamental Analysis. En este trabajo se han preparado 22 años de datos de los trimestres de diferentes acciones y se hizo uso de tres algoritmos de aprendizaje automático: red neuronal de avance (FNN), bosque aleatorio (RF) y el sistema de inferencia difusa neuronal adaptable (ANFIS) para la predicción de basada en datos de análisis fundamental. Además se aplicó la selección de características basadas en RF (Random Forest). Los resultados de este trabajo muestran que el modelo RF consigue los mejores resultados de predicción, y la selección de características es capaz de mejorar el rendimiento de las pruebas experimentales utilizando FNN y ANFIS. Además, el modelo agregado supera a todos los modelos de referencia, y también al índice DJIA (Dow Jones Industrial Average) el cual se uso como referencia [29].

A systematic review of fundamental and technical analysis of stock market predictions. Este trabajo llevo a cabo una revisión sistemática de 122 trabajos de investigación publicados en revistas académicas durante 11 años (2007-2018) en el área de la predicción del mercado de valores utilizando el aprendizaje automático. Las diversas técnicas identificadas en estos informes se agruparon en tres categorías, análisis técnicos, fundamentales y combinados. La agrupación se hizo sobre la base de los siguientes criterios: la naturaleza de un conjunto de datos y el número de fuentes de datos utilizadas, el marco temporal de los datos, los algoritmos de aprendizaje automático utilizados, la tarea de aprendizaje automático, las métricas de precisión y error utilizadas y los paquetes de software utilizados para la modelización. Los resultados revelaron que el 66 % de los documentos revisados se basaban en el análisis técnico, mientras que el 23 % y el 11 % se basaban en el análisis fundamental y los análisis combinados, respectivamente. En cuanto al número de fuentes de datos, el 89,34 % de los documentos revisados utilizaron una sola fuente, mientras que el 8,2 % y el 2,46 % utilizaron dos y tres fuentes, respectivamente. Los algoritmos de aprendizaje automático más utilizados para la predicción bursátil fueron las máquinas de vectores de soporte y las redes neuronales artificiales [11].

Capítulo 3

Análisis Técnico

3.1. Preparación de los datos

3.1.1. Recolección de los datos

En primera instancia, el proceso de recolección de los datos se abordó para los datos del análisis técnico, haciendo uso de Yahoo Finance. Los datos obtenidos hacen parte de una selección de empresas pertenecientes a el *Standard and Poor's 500* (S&P 500) el cual es uno de los índices bursátiles más importantes de Estados Unidos, es considerado el índice más representativo de la situación real del mercado. Este índice se compone de las 500 empresas con más poder económico que cotizan en las bolsas NYSE, NASDAQ y Cboe BZX Exchange. El índice S&P 500 captura aproximadamente el 80 % de la capitalización de mercado en los Estados Unidos. Las empresas que pertenecen a este índice deben cumplir una serie de características y requisitos para comprobar su poder financiero. De esta manera, la elección de acciones de empresas basadas en este índice garantiza consistencia y volumen en los datos.

Dentro de las empresas que componen este índice se seleccionaron 346 empresas que cumplen condiciones como estar dentro del índice antes o durante Marzo de 1985 y estar activa en el índice hasta Septiembre de 2020. Esto con el fin de abarcar una mayor cantidad de registros.

Los datos fueron obtenidos haciendo uso de la librería de Python para realizar consultas en Yahoo Finance, llamada *yfinance*. A través de los respectivos tickers de las acciones seleccionadas se obtuvieron los siguientes atributos: open, high, close, low, close adj y volume con una frecuencia diaria.

A partir de ese conjunto de datos construido, se eligen 5 acciones que representarán distintos sectores de la industria, Las cuales son: Industrial Business Machines (IBM), PepsiCo (PEP), Becton Dickinson and Co (BDX), Globe Life (GL), Norfolk Southern Corp (NSC). La mayor parte de las empresas mencionadas contizan en la bolsa NYSE a excepción de PEP, la cual cotiza en NASDAQ.

El criterio utilizado para seleccionar estas acciones fue acciones con el mínimo número de datos faltantes, ser acciones de empresas en distintos sectores de la industria.

3.1.2. Construcción de los conjuntos de datos utilizados

Los conjuntos de datos utilizados son los utilizados como materia prima para generar las diferentes variaciones que se tratarán en las siguiente secciones.

Con base en los datos provistos por la fuente ya mencionada, se calcularon nuevos atributos o variables llamados en el lenguaje del análisis técnico como indicadores.

Estos indicadores se calcularon a partir de los 6 atributos extraídos de la fuente: Apertura (Open), Mínimo (Low), Máximo (High), Cierre (Close), Volumen (Volume) y Cierre Siguiente o Predicción (CloseNext), los cuales hacen referencia al precio de apertura, el precio mas bajo y alto del día, el precio de cierre, el volumen y la variable a predecir la cual es el precio del día siguiente. Debido a que estos atributos por sí solos no aportan gran información al modelo se decide calcular una serie de indicadores [11] los cuales fueron un total de 10 atributos nuevos que se crearon a partir de estos para dar mayor expresividad a la información del conjunto de datos. Se incluyeron los siguientes atributos: *bop*, *ema20* *ema200*,*lower* (banda baja bollinger bands),*upper*(banda alta bollinger bands), *ma20*, *macd*, *momentum*, *signal*, *std20d* .

Para el conjunto de datos de PCA se usaron estos mismos atributos mencionados anteriormente, siendo para este conjunto un total de 16 atributos.

En la Figura 3.1 se puede observar el proceso llevado a cabo para llevar los datos obtenidos a través de la API hasta obtener los dos conjuntos de datos principales.

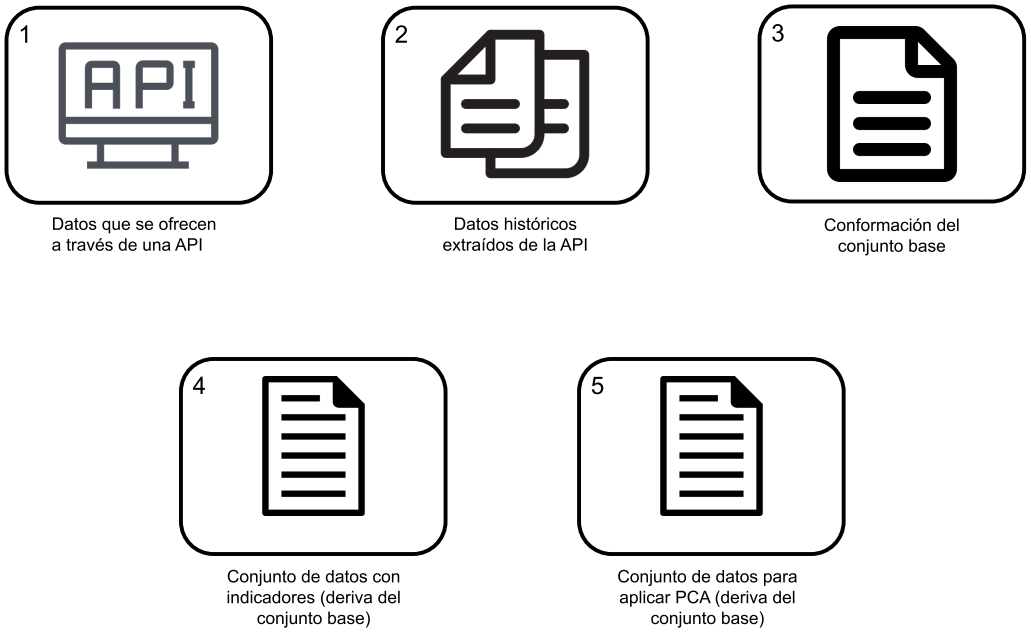

Figura 3.1: Esquema del tratamiento de los datos para análisis técnico

3.1.3. Análisis de los datos

Para el análisis técnico se obtuvieron un total de 9033 registros y 16 atributos por registro, los datos del análisis técnico son de frecuencia diaria. La fuente ofrece unos datos muy completos, es decir, sin valores nulos o vacíos en los atributos seleccionados. Y el conjunto generado para aplicar PCA cuenta con 9030 registros y 9 atributos por registro. Esto para las 5 acciones seleccionadas.

En los dos conjuntos se encontró que los datos faltantes por fila o periodo de tiempo se concentran sobre todo en los años iniciales debido a que algunos indicadores como la EMA de 20 toma los 20 datos anteriores entonces debido a esto se generan unos datos vacíos. Esto representa una cantidad de datos faltantes muy poco relevante y afecta casi nulamente a la totalidad de la información que presentan estos datos.

En el conjunto de datos para el análisis técnico se observan atributos con una correlación superando el 90 %, como se muestra en Figura 3.2. Dando lugar a procesos que se aclararán en las siguientes secciones. También se encontraron grandes diferencias entre las escalas de los distintos atributos como se aprecia en Figura 3.3

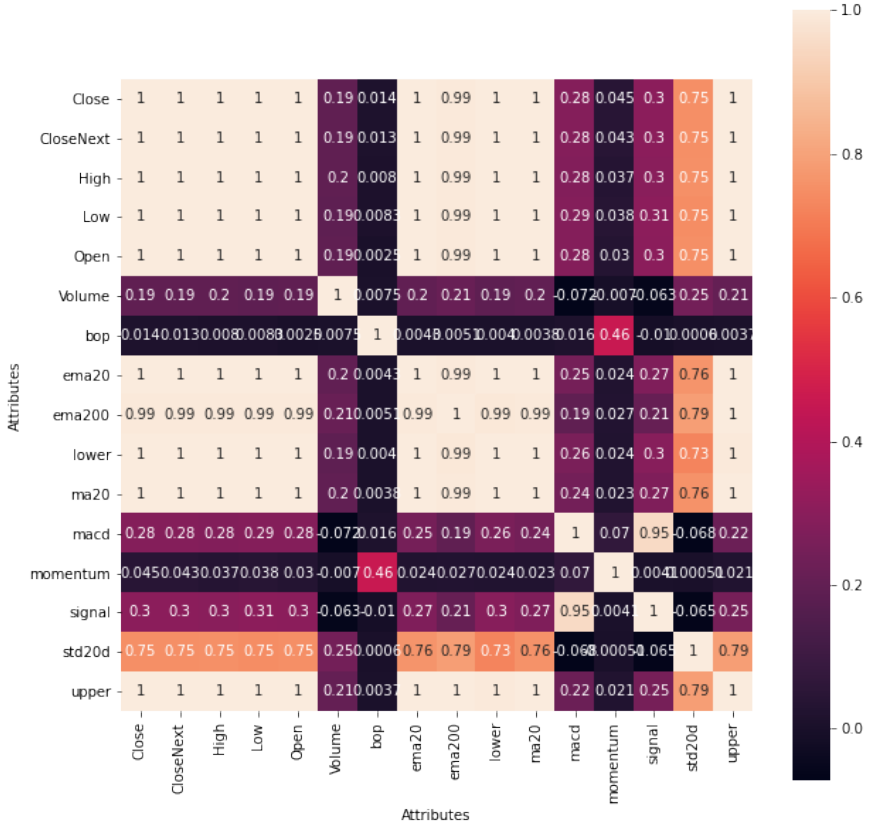

Figura 3.2: Matriz de correlación en etapa de análisis para el conjunto de datos técnicos. Acción: NSC

3.1.4. Procesamiento

Los conjuntos de datos generados fueron sometidos a una serie de procesos para llegar a los conjuntos finales con los que se alimentarán los distintos modelos de aprendizaje automático. Estos procesos tienen que ver con la limpieza de los datos y las distintas transformaciones realizadas como

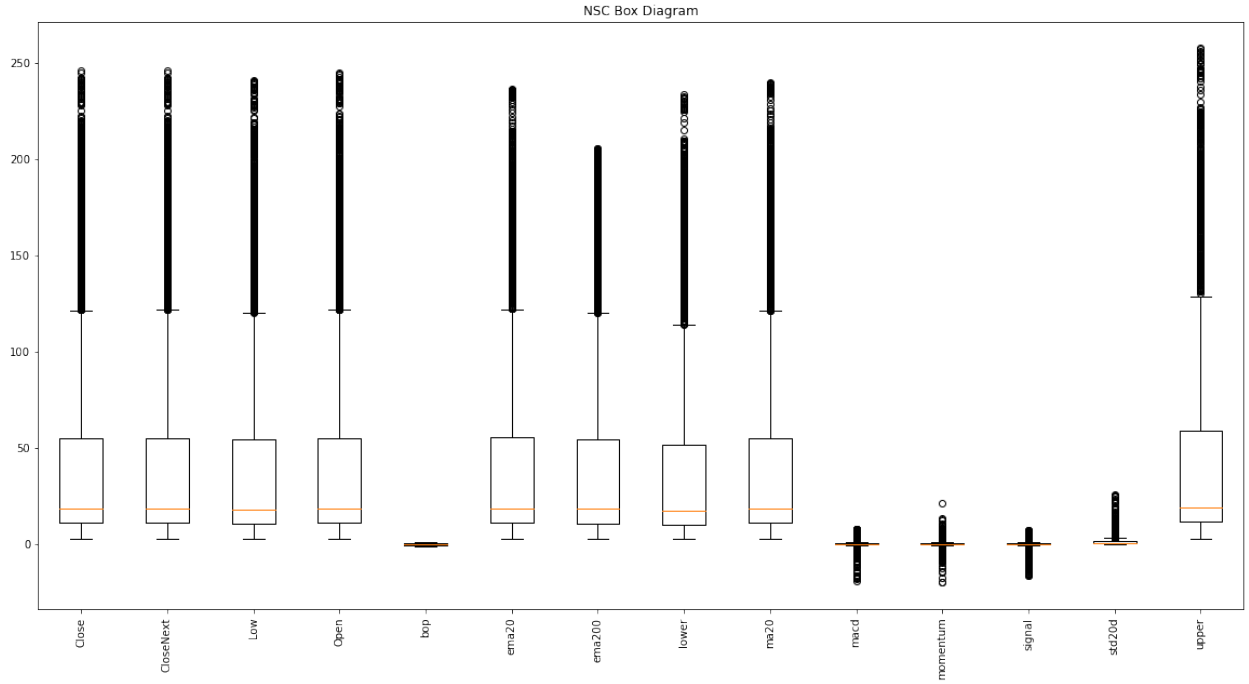

Figura 3.3: Diagrama de cajas y bigotes para la acción NSC para el conjunto de datos técnicos

experimentación sobre los datos y los posibles resultados que pueden generar.

3.1.4.1. Limpieza de los datos

Tanto para el conjunto de datos donde se utilizan únicamente los ratios así como también para el conjunto al que se aplicará PCA se realizaron las siguientes actividades de limpieza de datos.

En primera instancia se reemplazaron los valores nulos o vacíos aplicando el método "forward fill"haciendo que los registros nulos tomen el valor no nulo más reciente en el conjunto de datos hasta encontrar otro valor no nulo y repetir el proceso.

En los datos donde observáramos altas correlaciones con variables independientes, se deciden eliminar de esta manera, de los 16 atributos iniciales se obtienen 9 en total, tras finalizar el proceso de limpieza de datos, quedan como resultado los siguientes atributos Close, Volume, bop, ema200, macd, momentum, signal y std20. esto da como resultado una matriz de correlaciones como se muestra en la Figura 3.4 para la acción NSC a comparación de la matriz de correlaciones anterior a la limpieza Figura 3.2.

Puesto que los dos conjuntos tanto de indicadores como al que se le aplicará PCA son los mismos, las actividades de limpieza serán las mismas para ambos.

Figura 3.4: Matriz de correlación después de la etapa de limpieza para el conjunto de datos técnicos: Acción NSC

3.1.5. Transformaciones

Los conjuntos de datos tratados hasta ahora, fueron sometidos a una serie de transformaciones, las cuales en algunas se transformaron únicamente los datos de los conjuntos y otras dieron lugar a nuevos conjuntos. Estos nuevos conjuntos de datos se utilizarán para posteriormente realizar una extensa exploración en busca del conjunto de datos que tuviera un mejor desempeño con la técnica de aprendizaje automático utilizada.

En total se crearon cuatro variaciones de conjuntos construidos para análisis técnico, el primero siendo en el que se seleccionaron ratios o indicadores, el segundo en el que se aplica PCA al conjunto que se ha ido construyendo para aplicar este algoritmo y el conjunto de datos de experimentación utilizando el método de ventana deslizante tanto para el conjunto de ratios e indicadores como para el conjunto PCA, siendo estos la tercera y cuarta variante. Adicionalmente cada uno de estos conjuntos de datos posee una versión donde la variable de salida se encuentra estandarizada al igual que el resto del conjunto y otra variante donde la variable de salida no se estandariza pero las de entrada sí. Por lo tanto, se cuenta con un total de 8 conjuntos (como se puede observar en la Tabla 3.1) de datos para cada una de las 5 acciones.

Tabla 3.1: Características de los conjuntos generados

	Conjunto base		Ventana Deslizante		Estandarización	
Conjunto	Atrib. sel.	PCA	Sin VD	VD	X-estándar	XY-estándar
SX	x		x		x	
SXY	x		x			x
W-SX	x			x	x	
W-SXY	x			x		x
PCA-SX		x	x		x	
PCA-SXY		x	x			x
W-PCA-SX		x		x	x	
W-PCA-SXY		x		x		x

A continuación se detallará cómo fue el proceso de construcción para estos nuevos conjuntos de datos.

3.1.5.1. Estandarización

Se aplicó una transformación en relación a las escalas de los atributos de cada uno de los conjuntos generados. Teniendo que los atributos independientes estarían estandarizados para todos los conjuntos de datos realizados. Por otra parte, para el atributo dependiente o atributo de salida se decidió por tener dos variaciones, una siendo estandarizado este atributo y otra en la que conserva su escala original.

El método de escalado utilizado fue estandarización o 'Standard Scaler' como es denominado

en la librería *sci-kit learn* cuya fórmula calculada para el ejemplo x es:

$$z = \frac{(x - u)}{s}$$

Donde u es la media del conjunto de ejemplos, y s es la desviación estándar del conjunto de ejemplos.

3.1.5.2. Ventana deslizante

El proceso de generar los conjuntos con ventana deslizante se realizó tanto para los conjuntos de datos de indicadores tanto como para los que hacen uso de PCA. Para análisis técnico se utilizó un tamaño de ventana de 7, representando que con los datos de 7 días es posible predecir el siguiente día.

Para realizar esta transformación de los conjuntos de datos utilizados se diseñó un código que agrupa en un registro los datos de todos los atributos para el tamaño de ventana seleccionado añadiendo el atributo de salida, en este caso, el precio del siguiente día correspondiente al tamaño de la ventana.

3.1.5.3. Análisis de componentes principales (PCA)

En el método de PCA se utilizaron el total de 16 atributos, Se utilizó el algoritmo de PCA para obtener una "Varianza explicada media"del 95 %. De este modo, el algoritmo logró ajustarse a este valor con 6 componentes generados a partir de los atributos originales. Sin embargo, solo un componente de los 6 totales, comparte correlación con el atributo a predecir, situación que se puede evidenciar en la matriz de correlación generada para este conjunto de datos Figura 3.5

3.2. Desarrollo de los modelos

3.2.1. Herramientas utilizadas

Para el desarrollo de los modelos de aprendizaje automático se utilizó el lenguaje de programación Python, en su versión 3.6.12. En adición de librerías como scikit-learn 0.24.1, pandas 1.1.5, numpy, seaborn y matplotlib, para el manejo de los datos, el propio proceso de aprendizaje automático y la graficación.

3.2.2. Conformación del modelo base

El modelo base hace alusión al modelo de referencia que se utilizará a lo largo del análisis de los resultados para ser comparado con los resultados de los modelos que se construirán.

Para este proyecto, el modelo base consiste en una predicción básica del precio de una acción, la cual dice que el precio de mañana será igual al de hoy.

Para lograr este objetivo, se creó un pequeño algoritmo, el cual se encarga, en primera instancia, de realizar la separación de datos en entrenamiento y prueba, donde se utiliza un tamaño para el conjunto de prueba del 20 % del total del número de datos haciendo uso de una semilla que

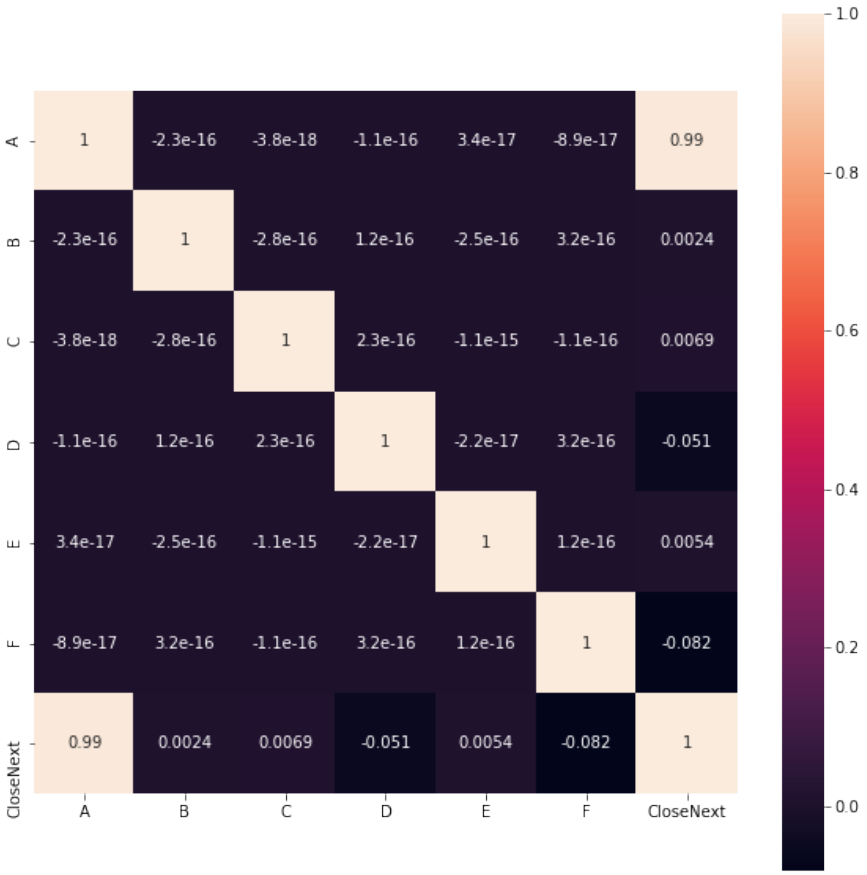

Figura 3.5: Matriz de correlación después de ejecutar PCA para el conjunto de datos tecnico. Acción: NSC

controla la aleatoriedad con la cual se seleccionan los ejemplos para incluirlos en el conjunto de entrenamiento o en el de prueba. Este es un elemento importante, dado que prepara un mejor terreno para la comparación con los modelos de aprendizaje automático, dado que comparten esta característica.

Posteriormente se genera un nuevo arreglo, el cuál representará la predicción y contendrá el precio del siguiente periodo en relación al precio que se encuentra en el arreglo que representa el conjunto de prueba. Este proceso se realiza para las 5 acciones seleccionadas, obteniendo un total de 5 modelos base.

3.2.3. Selección del conjunto de datos para cada técnica

Esta fue la primera etapa del desarrollo de los modelos, la cual consistió en la exploración de los distintos conjuntos de datos generados en la sección anterior Tabla 3.1 poniéndolos a prueba con cada una de las técnicas de aprendizaje automático seleccionadas. Donde el objetivo principal de este proceso es seleccionar el conjunto de datos que obtenga el mejor desempeño para cada algoritmo de aprendizaje automático utilizado.

En la selección de parámetros para la construcción de estos modelos se realizó un proceso de GridSearchCV para cada conjunto de datos teniendo en cuenta los siguientes hiper-parámetros para el analisis tecnico Tabla 3.2.

Como se mencionó anteriormente, para la división en conjunto de datos para entrenamiento y conjunto de prueba, se utiliza una semilla para mantener una uniformidad en el modo que se seleccionan estos ejemplos para ser comparados con mayor certeza con el modelo base.

En este proceso se generaron un total de 90 modelos de aprendizaje automático basados en cada una de las 3 técnicas seleccionadas alimentado con cada uno de los 8 conjuntos de datos para cada una de las 5 acciones.

3.2.4. Estimación de hiper-parámetros

La estimación de parámetros se realizó en dos etapas. Como se mencionó anteriormente la primera etapa de construcción de los modelos fue una etapa donde no se profundizó en la selección de los hiper-parámetros. Luego en la segunda etapa, en la cual ya se selecciona el conjunto de datos que tiene mayor desempeño con cada técnica. Se busca mejorar los modelos, aquí se decidió tomar unos rangos amplios de los valores de los parámetros, y con ayuda de grid search se observa cuáles son los valores para los hiper-parámetros que mejor comportamiento ofrecían y a medida que se realizaban varias ejecuciones se acotaba aún más la amplitud de los rangos de los parámetros.

Tabla 3.3 se describe el espacio de búsqueda utilizado por el algoritmo GridSearchCV para determinar los mejores valores para cada uno de los hiper-parámetros tratados. Después de haber ejecutado el proceso de búsqueda de parámetros para cada técnica se obtienen los valores resultantes que serán utilizados en el entrenamiento final de los modelos. Estos resultados se encuentran consignados en Tabla 3.4.

Tabla 3.2: Primera etapa de búsqueda de hiper-parámetros para las distintas técnicas de aprendizaje automático. Análisis técnico

Técnico	Hiper-parámetros	Valores Seleccionados
SVR	Kernel	Linear
SVR	C	[50, 80]
SVR	Gamma	[0.0001,0.01]
SVR	Epsilon	[0.001,0.1]
RF	n estimators	[100, 150, 200]
RF	min samples leaf	[1, 3]
RF	max features	['auto', 0.5]
RF	oob score	[True, False]
RF	max depth	[None, 7]
RF	criterion	mae
MLP	hidden layer sizes	[(100, 100), (50, 50,50),(100,100,100)]
MLP	activation	['identity','logistic']
MLP	solver	['sgd', 'adam']

Tabla 3.3: Segunda etapa de búsqueda de hiper-parámetros según la técnica de aprendizaje automático. Análisis técnico

Técnico	Hiper parámetros	Valores
SVR	Kernel	Linear
SVR	C	[50, 80, 10, default, 2.0, 5.0, 7.0,9.0]
SVR	Gamma	['auto',0.0001,0.01]
SVR	Epsilon	[0.001,0.1]
SVR	Degree	[3,5,10]
RF	n estimators	[100, 150, 200, 500, 700, 1000]
RF	min samples leaf	[1, 2, 3, 5, 10, 15, 20]
RF	max features	['auto', 0.5, 'log2', 'sqrt', 0.7]
RF	oob score	[True, False]
RF	max depth	[None, 7, 10, 15, 20]
RF	criterion	mae
MLP	hidden layer sizes	[(50, 50,100),(50,100,50),(50,50,50),(50), (200,200), (200,200,200)]
MLP	activation	['logistic','adaptive']
MLP	solver	['sgd', 'adam']
MLP	alpha	[0.001,0.01,0.0001, 0.1]
MLP	learning rate	adaptive

Tabla 3.4: Resultados de la búsqueda de hiper-parámetros para las distintas técnicas de aprendizaje automático. Análisis técnico.

Técnico	Hiper-parámetros	Valores Seleccionados
SVR	Kernel	Linear
SVR	C	5,0
SVR	Gamma	auto
SVR	Epsilon	0,001
SVR	Degree	[3,5,10]
RF	n estimators	1000
RF	min samples leaf	2
RF	max features	0.7
RF	oob score	True
RF	max depth	10
RF	criterion	mae
MLP	hidden layer sizes	(200,200,200)
MLP	activation	logistic
MLP	solver	sgd
MLP	alpha	0.001
MLP	learning rate	adaptive

3.2.5. Entrenamiento y prueba de los modelos

El proceso de entrenamiento y, posteriormente, prueba de modelos ocurre en dos momentos durante el desarrollo del proyecto. La primera, al realizar la exploración de los conjuntos de datos y la segunda, la cual se da una vez se determina el conjunto de datos a utilizar para cada técnica y se obtienen los hiper-parámetros que generan una mejor predicción, para la obtención de los parámetros se hizo uso de la técnica gridsearch.

Para el entrenamiento se hizo uso de la técnica holdout haciendo una división de un un 80 % de los datos totales para entrenamiento y un 20 % para prueba en todas las etapas, además se utiliza la misma semilla para la división de los datos para entrenamiento y prueba, permitiendo de esta manera obtener una comparación más justa de los resultados de los modelos de aprendizaje automático con el modelo base. Para cada técnica se utilizó el algoritmo correspondiente incluido en la librería *sci-kit learn*.

3.2.6. Backtesting

3.2.6.1. Descripción

El término 'backtesting' hace referencia al proceso de poner a prueba una estrategia de inversión. En este caso el objetivo de esta prueba es observar el desempeño de los modelos construidos en un terreno mucho más cercano a la realidad.

La idea principal que se tiene para esta prueba es comprar o vender una acción en base a la predicción que genere el modelo y de este modo, se busca obtener el mayor rendimiento posible en términos económicos.

Para llevar a cabo este proceso fue necesario de datos que no hubieran sido usado para entrenamiento y que fuesen consecutivos, por tal razón se tomaron desde 2020-09 hasta 2021-03. Este proceso se realizó para las acciones IBM y PEP, puesto que fueron la acción con mejor y peor desempeño respectivamente.

3.2.6.2. Estrategia

La estrategia utilizada para el backtesting realizado es conocida como c̈omprar bajo, vender alto.̈ Se trata de una de las estrategias de inversión más básicas, esta consiste en comprar una accion cuando el precio se considere bajo, en este caso, en comparación con la fecha anterior y vender cuando el precio sea alto, igualmente en comparación con el precio de la fecha anterior. La diferencia que se presenta con la estrategia clásica, es que la señal de compra o venta va a ser decidida en función del resultado que arroje la predicción de uno de los modelos construidos, y además esta decisión es tomada antes de que se conozca el precio de la fecha siguiente.

3.2.6.3. Desarrollo

Para implementar el backtesting, primero se obtuvieron las predicciones para las fechas indicadas, así como también su precio real y el dinero inicial del que se dispone. Con estos datos se calcula la diferencia entre la predicción del precio para el futuro y el precio real de la fecha actual, si este resultado es positivo, se compra o se mantiene la acción, de lo contrario, se procede a vender la acción o a permanecer inactivo en caso de no tener una acción para vender. Este proceso genera 3 variables de salida: posición (una posición se refiere a la compra de una acción), la cual indica el valor de la acción que se posee; dinero (cash, por su traducción en inglés), el cual indica el dinero no invertido disponible y el total, el cual es la suma de las dos variables anteriores.

En esta aproximación se trabaja únicamente con una posición para la acción de la empresa seleccionada.

3.3. Resultados y análisis

3.3.1. Resultados de la exploración de conjuntos de datos

Como bien se expresó en anteriores capítulos, se realizó una exploración sobre distintos conjuntos de datos construidos a partir de dos conjuntos principales, los cuales eran: conjunto indicadores y el conjunto sobre el que se aplicaría PCA. Esta exploración se realizó con el fin de determinar cuál de los distintos conjuntos Tabla 3.1 se desempañaba mejor con las distintas técnica de aprendizaje automático.

Por lo tanto, en esta sección se mostrarán los resultados obtenidos de esta exploración en comparación a los resultados arrojados por el modelo base. En general, para las tres técnicas

(SVR, MLP y RF) el conjunto de datos con mejores resultados fue el que agrupa indicadores. Sin embargo, se presentan ciertas variaciones en estos resultados, haciendo referencia al escalado o no de la variable a predecir.

La herramienta de comparación utilizada en este caso se trata de los diagramas de cajas y bigotes, puesto que proporciona una información mucho más integral que la que se puede obtener con un promedio. Estos diagramas se generan a partir de la métrica de error RMSE obtenida de las predicciones realizadas con estos modelos para todas las acciones. De este modo, se obtienen de todas las acciones el valor mínimo, máximo y los cuartiles 1 y 3 para generar el diagrama. La nomenclatura utilizada en los resultados de los modelos, va precedida por la técnica que se utiliza seguida de el conjunto de datos.

3.3.1.1. Resultados sobre Regresor de Vectores de Soporte

Los resultados obtenidos en Figura 3.6 muestran que los resultados de los conjuntos SX y SXY se asemejan mucho a los resultados del modelo base, siendo estos levemente inferiores al base. Comparando sus extremos tenemos que el mínimo y máximo de SX respectivamente son 0,6252277892 y 1,406490182, y de SXY son 0,6256099094 y 1,406145945. Observando que SX gana en mínimo y SXY gana en máximo. Por lo tanto la caja de SXY está ligeramente más abajo que la del conjunto SX. De este modo, se selecciona el conjunto escalado para la siguiente etapa.

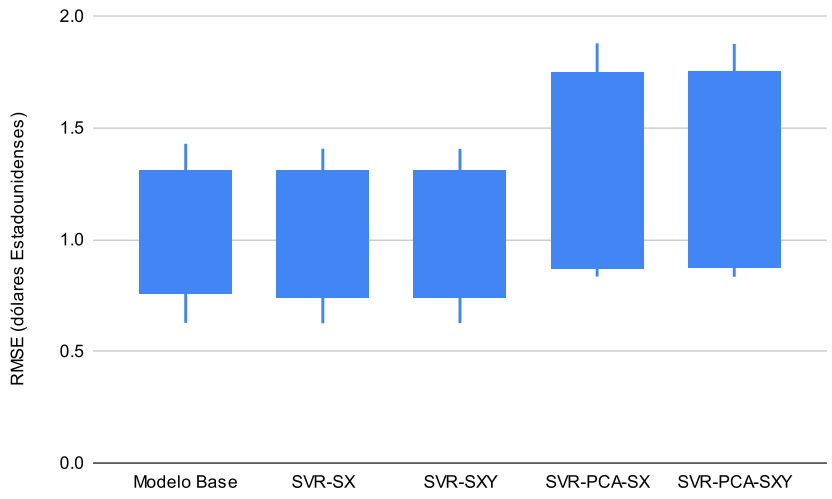

Figura 3.6: Analisis tecnico: Comparación de los distintos modelos

3.3.1.2. Resultados sobre Bosques Aleatorios

Los resultados mostrados en la Figura 3.7 para la técnica de Bosques Aleatorios muestra un desempeño similar entre los resultados del modelo base y los resultados de los conjuntos SX y SXY.

La caja generada para estos dos conjuntos de datos mencionados se extienden un poco más hacia arriba en comparación a la del modelo base. Siendo SXY la que menos extiende su punto máximo de las dos. Por lo tanto, se toma este conjunto para continuar con la siguiente etapa.

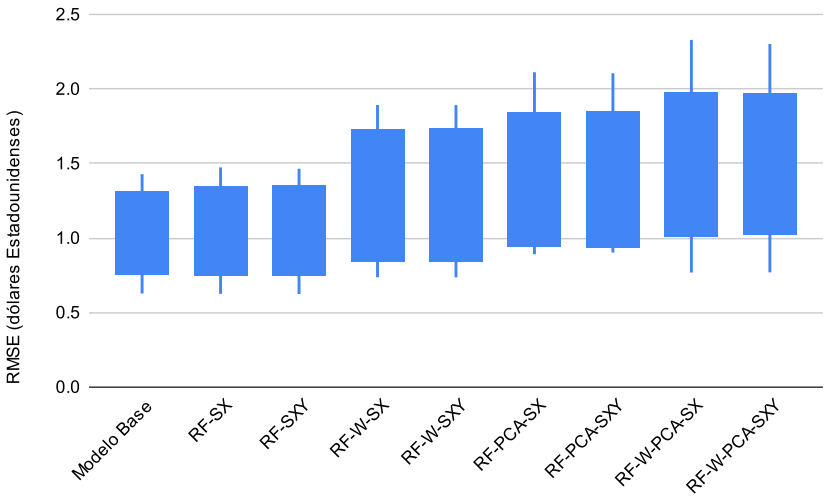

Figura 3.7: Analisis tecnico:Comparación de los distintos modelos

3.3.1.3. Resultados sobre Perceptrón Multicapa

Los resultados obtenidos con la técnica MLP indican sin duda alguna que el conjunto SX es el que mejor desempeño genera para esta técnica como se puede observar en la Figura 3.8. Puesto que los resultados obtenidos con el conjunto SXY se centran en unos valores más elevados. Sorprendentemente en los resultados de esta técnica se observa que los conjuntos de PCA sin ventana de desplazamiento pueden generar errores más pequeños (para algunas acciones) que el conjunto SXY.

3.3.1.4. Análisis de resultados

Según lo mencionado anteriormente, se dispone del conjunto de datos basado en ratios para construir los modelos de la siguiente etapa, especificando su variación en la Tabla 3.5. Y resultando las mismas variaciones para todas las técnicas para datos de análisis técnico, de cierto modo corroborando estas decisiones.

Como se ha observado en los resultados anteriores, los demás conjuntos de datos no suponen una competencia relevante para el conjunto de indicadores. Por lo tanto, para esta aproximación el algoritmo PCA no resultó siendo acertado con los componentes construidos, dado que estos no representaron correctamente la información original afectando así al modelo de predicción.

Así como tampoco se vio mejor desempeño en las predicciones aplicando la ventana deslizante

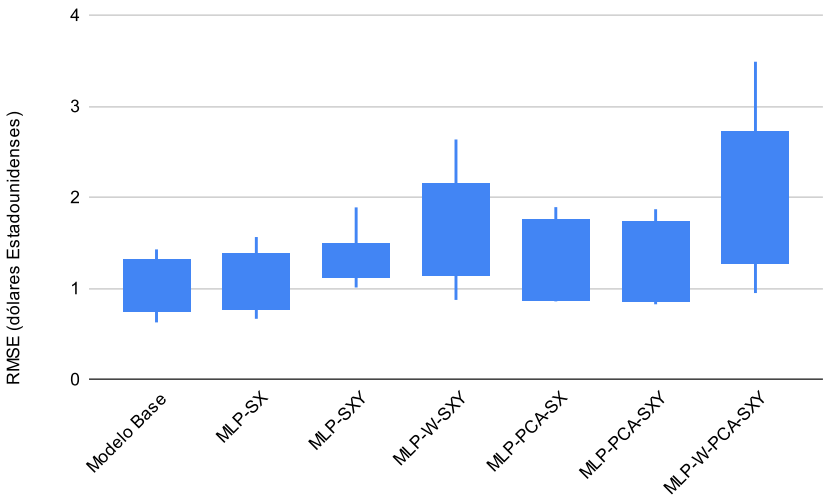

Figura 3.8: Analisis tecnico: comparación de los distintos modelos de la técnica MLP

Tabla 3.5: Conjuntos de datos seleccionados según técnica. Análisis fundamental y técnico

Técnica	Conjunto
SVR	SXY
RF	SXY
MLP	SX

tal como se especificó para este proyecto. Cumpliendo siempre la función de aumentar el error como se puede ver tanto para los conjuntos basados en ratios como los basados en PCA.

3.3.2. Resultados de los modelos de aprendizaje automático

Como se describió en la sección de "Desarrollo de los modelos", la segunda etapa respecto al desarrollo de los modelos consistió en realizar el entrenamiento y prueba de los mismos con los conjuntos de datos seleccionados de la etapa 1, los cuales son con los que se obtiene mejor desempeño en las predicciones.

En esta sección se mostrarán los resultados obtenidos por cada uno de los modelos construidos para cada una de las 3 técnicas y para cada una de las 5 acciones. Con el objetivo de conocer qué técnica tiene un error menor en comparación con las demás y con el modelo base.

Para cumplir con este objetivo, se utilizan diagramas de cajas y bigotes para realizar comparaciones en las predicciones, dos diagramas de barras, uno donde se muestra el promedio de los errores obtenidos paras las 5 acciones según el modelo utilizado para cada técnica, y el otro diagrama de barras por cada técnica donde se puede observar el comportamiento del modelo con cada una de las distintas acciones en comparación con el modelo base. Se utiliza la siguiente nomenclatura para representar el modelo de segunda etapa, primero antecede el nombre de la técnica, seguido por el conjunto de datos utilizados y por último la letra T significando prueba (test en inglés).

3.3.2.1. Resultados sobre Regresor de Vectores de Soporte

Los resultados agregados en formato de diagrama de cajas y bigotes como se observa en la Figura 3.10, muestran que el modelo SVR-SXY-T construido tiene un desempeño muy similar al modelo base pero con una ligera mejora. El modelo alcanzó valores de RMSE ligeramente más bajos que los del Modelo Base, esto se puede observar en que su mínimo y máximo se desplazaron un poco hacia abajo en comparación con el modelo base.

Por otro lado, en la Figura 3.9 se observa la comparación de los promedios resultantes del modelo base con los 2 modelos realizados con este conjunto de datos para la técnica de aprendizaje automático SVR. Observando en estas métricas una ligera disminución del error del modelo SVR-SXY-T con el modelo SVR-SXY. Superando así ligeramente, estos dos modelos, al modelo base en estas dos métricas de error presentadas.

En la Figura 3.11 se puede observar que en el modelo SVR-SXY-T se obtuvieron errores ligeramente menores frente al modelo base en todas las acciones con las que se trabajaron.

3.3.2.2. Resultados sobre Bosques Aleatorios

Los resultados agregados mostrados en la Figura 3.13 muestran que el modelo RF-SXY-T construido tiene un desempeño muy similar al modelo base, según las métricas obtenidas. El modelo alcanzó valores de RMSE ligeramente más bajos que los del Modelo Base, esto se puede observar en el mínimo y el cuartil 1 los cuales se observan desplazados ligeramente hacia abajo como nos

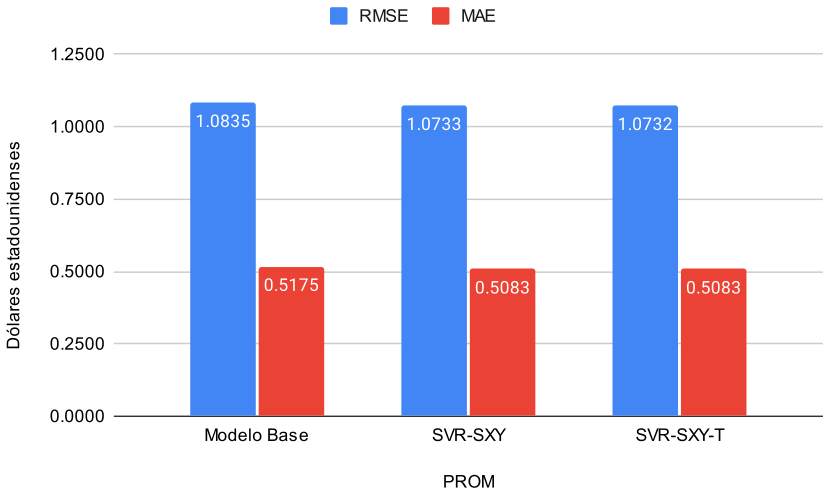

Figura 3.9: Análisis técnico: Promedio de errores RMSE y MAE para las 5 acciones para modelo base, modelo etapa 1 y modelo etapa 2. Técnica: SVR

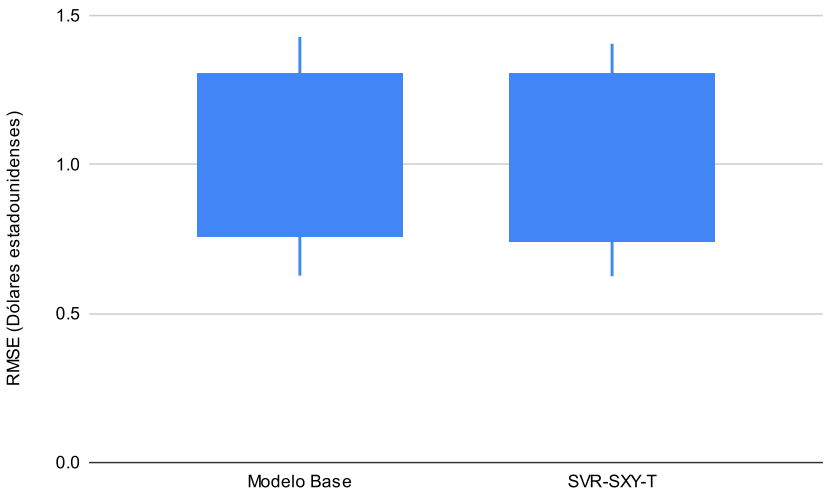

Figura 3.10: Análisis técnico: Diagrama de cajas y bigotes para el modelo SVR-SXY-T frente al modelo base respecto a la métrica RMSE

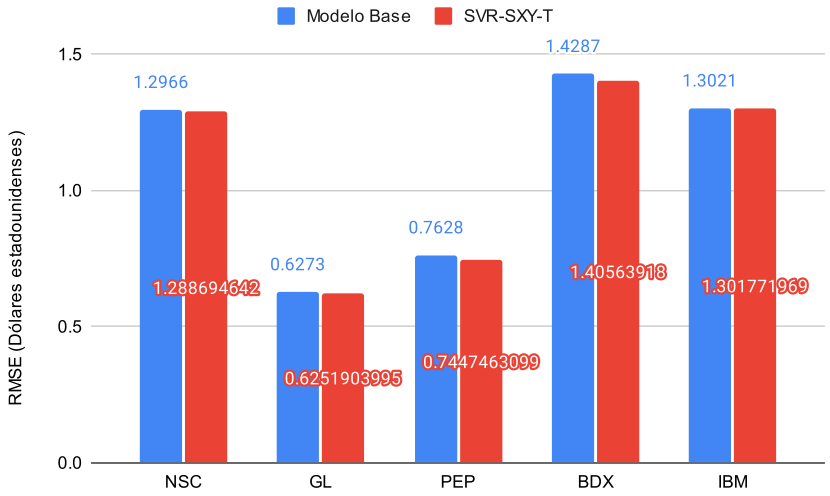

Figura 3.11: Análisis técnico: Comparación del desempeño del modelo SVR-SXY-T en las distintas acciones frente modelo base respecto a la métrica RMSE

muestra la Figura 3.13. Pero también alcanza valores sutilmente mayores de RMSE en comparación al modelo base, en el cuartil 3 y el valor máximo.

Por otro lado, en la Figura 3.9 se puede observar que, en promedio, a esta técnica de aprendizaje automático se le dificulta mejorar las métricas de error del modelo base. Sin embargo, se nota una mejoría entre el modelo de etapa 1 con el modelo de etapa 2, siendo este el que más se acerca al promedio alcanzado por el modelo base.

En la siguiente Figura 3.14 se puede observar que el modelo de RF se desempeña mejor en algunas de las acciones seleccionadas como GL y PEP, pero para acciones como NSC, BDX y IBM tiene un desempeño ligeramente peor que el conseguido por el modelo base.

3.3.2.3. Resultados sobre Perceptrón Multicapa

Los resultados agregados mostrados en la Figura 3.16 muestran que el modelo MLP-SX-T construido tiene un desempeño inferior comparado con el modelo base. Según las métricas obtenidas, el modelo alcanzó valores de RMSE ligeramente superiores a los del modelo base, situación que nos indica que su error tiende a estar más elevado en comparación al modelo base para algunas acciones.

En la Figura 3.9 se puede observar que el promedio de las métricas de error para el modelo de primera etapa es claramente superior al del modelo base, situación que mejora con la segunda etapa. Sin embargo, no se logra alcanzar los valores promedios conseguidos con el modelo base.

La Figura 3.17 muestra, como es de esperarse, un RMSE claramente mayor en las predicciones realizadas por el modelo MLP-SX-T en comparación al modelo base. Siendo esta técnica la que peor desempeño obtuvo de las 3 ejecutadas.

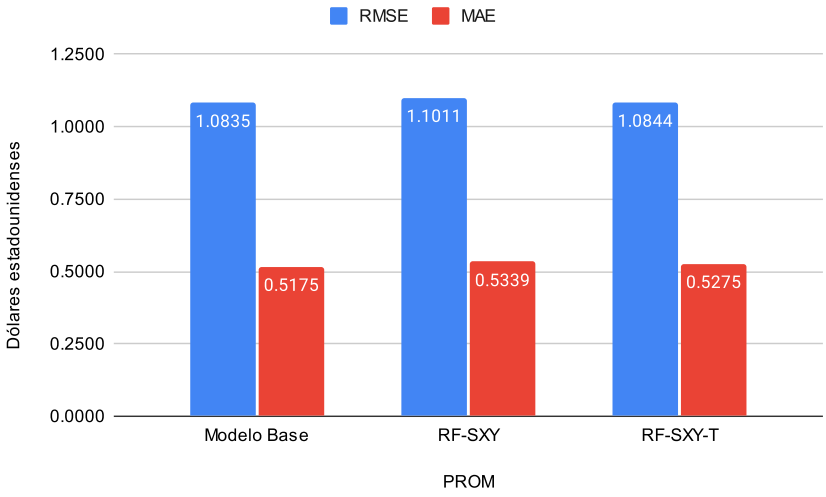

Figura 3.12: Análisis técnico: Promedio de errores RMSE y MAE para las 5 acciones para modelo base, modelo etapa 1 y modelo etapa 2. Técnica: RF

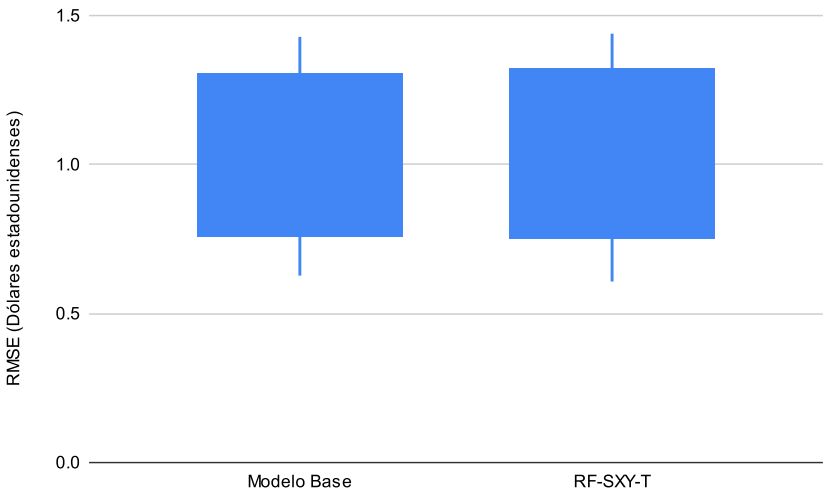

Figura 3.13: Análisis técnico: Diagrama de cajas y bigotes para el modelo RF-SXY-T frente al modelo base respecto a la métrica RMSE

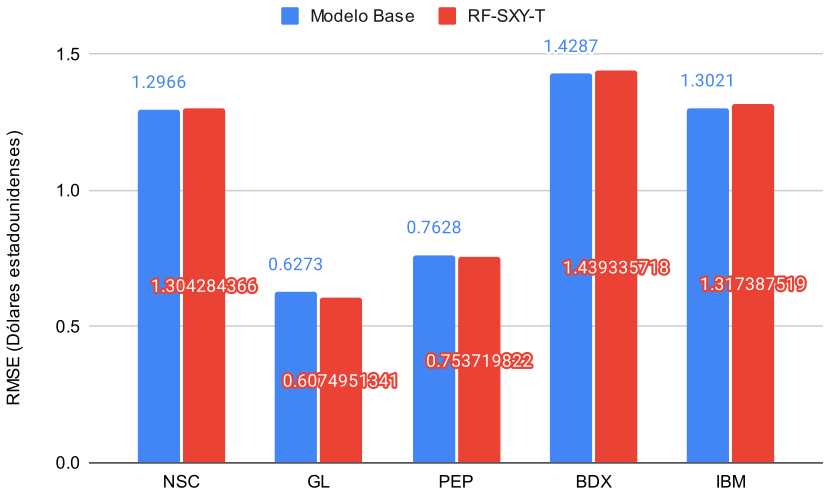

Figura 3.14: Análisis técnico: Comparación del desempeño del modelo RF-SXY-T en las distintas acciones frente modelo base respecto a la métrica RMSE

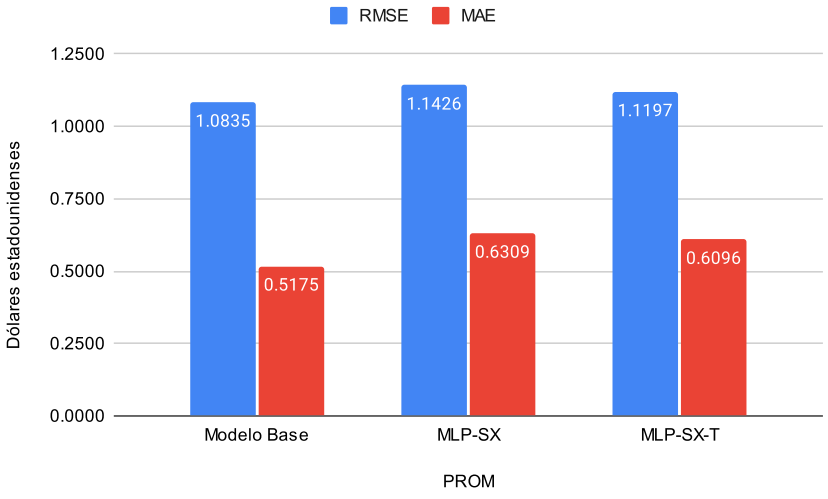

Figura 3.15: Análisis técnico: Promedio de errores RMSE y MAE para las 5 acciones para modelo base, modelo etapa 1 y modelo etapa 2. Técnica: MLP

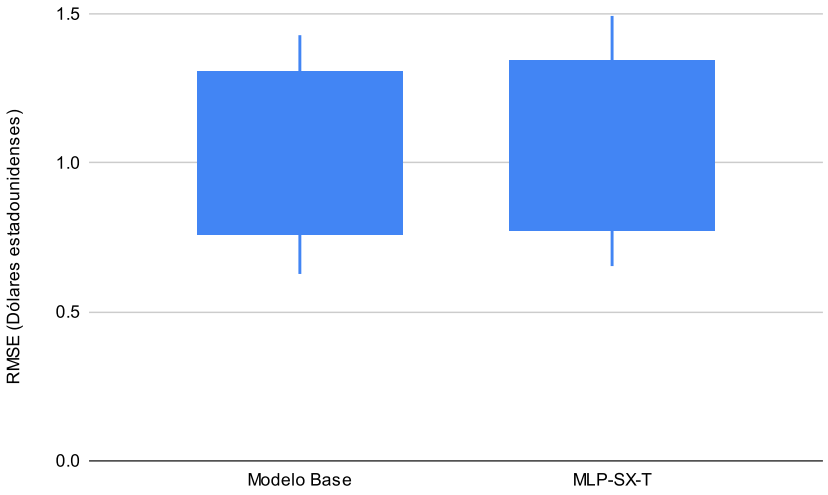

Figura 3.16: Análisis técnico: Diagrama de cajas y bigotes para el modelo MLP-SXY-T frente al modelo base respecto a la métrica RMSE

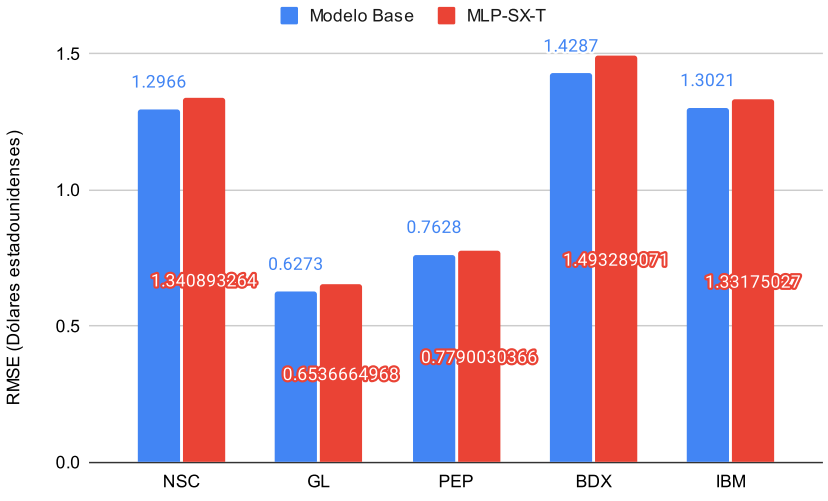

Figura 3.17: Análisis técnico: Comparación del desempeño del modelo MLP-SXY-T en las distintas acciones frente modelo base respecto a la métrica RMSE

3.3.2.4. Análisis de resultados

Es posible observar en los resultados presentados en la Tabla 3.6, que en la mayoría de las acciones no fue posible obtener modelos que sobrepasen la métrica RMSE del modelo base construido, el único modelo que en todas las acciones tuvo mejor comportamiento que el modelo base fue el modelo del SVR-SXY-T, seguido del RF-SXY-T, el cual en dos de las cinco acciones logró un mejor desempeño que el modelo base. Y dejando al modelo MLP-SX-T en último lugar, teniendo un error superior para la predicción en todas las acciones utilizadas. Esta misma información se presenta en la Figura 3.18, donde el orden sugerido por el promedio anteriormente se puede observar en este diagrama de cajas, viendo que el modelo que más se aproxima a los resultados del modelo base es SVR-SXY. Los otros dos modelos RF-SXY-T y MLP-SX-T, se encuentran ligeramente elevados uno del otro.

Este comportamiento de los modelos con datos de análisis técnico puede ser debido a diversas razones. Una de ellas es la dificultad existente en predecir el precio, lo cuál es mucho más complejo que predecir el movimiento del precio. Por otro lado, puede que los indicadores usados sean de utilidad para predecir la tendencia, momentum, volatilidad, volumen y la fuerza pero no sean tan efectivos en la ayuda de la predicción del precio.

Por el lado de los modelos, visualmente se puede explicar el mejor desempeño del modelo SVR-SXY-T, dado que pueden observar similitudes visualmente con las "predicciones"del modelo base, donde básicamente se predice que el precio de mañana será el mismo que el de hoy. De este modo, la gráfica generada por el modelo de SVR en segunda etapa se asemeja con esta lógica pero realizando unos pequeños ajustes a los valores.

Según se puede observar en los resultados, a los modelos de las otras técnicas se les dificulta más batir las métricas del modelo base. Este comportamiento puede ser posible a que los errores en el modelo base y el modelo SVR mantenían siempre una distancia uniforme, por ende la media no resultaba muy diferente. En cambio los modelos que utilizan RF y MLP intentan realizar una predicción más libre y menos ajustada al precio actual, generando una predicción mucho más alejada o cercana al valor real suponiendo una media mayor y esto se puede ver reflejado mucho más al utilizar la métrica de error RMSE, la cual penaliza más el error cometido.

En la Figura 3.19 y la Figura 3.20 es posible observar el comportamiento de las distintas técnicas seleccionadas para las acciones de PEP e IBM. Siendo estas las que tuvieron mejores y peores resultados respectivamente. Se puede corroborar lo que se estableció anteriormente sobre el modelo que utiliza SVR, las predicciones siguen muy de cerca, con una distancia uniforme, el valor real del precio de la acción. Se puede observar que en grandes cambios del precio los modelos MLP-SX-T y RF-SXY-T se desempeñan peor en su predicción, ocurriendo lo contrario en la situación donde el cambio del precio no tiene magnitudes tan grandes. Una causa del peor desempeño de los modelos para la acción IBM puede ser esta. Se visualiza en la gráfica que el movimiento del precio para los modelos que implementan MLP y RF no están tan ligados al precio actual como ocurre con SVR.

Tabla 3.6: Resultados de las distintas técnicas para el análisis técnico

NSC	Modelo Base	MLP-SX-T	RF-SXY-T	SVR-SXY-T
RMSE	1.296615916	1.340893264	1.304284366	1.288694642
MAE	0.5648853387	0.7055994873	0.5770870231	0.5547265749
GL	Modelo Base	MLP-SX-T	RF-SXY-T	SVR-SXY-T
RMSE	0.6272791587	0.6536664968	0.6074951341	0.6251903995
MAE	0.2734199314	0.3462974014	0.2740267249	0.2681700058
PEP	Modelo Base	MLP-SX-T	RF-SXY-T	SVR-SXY-T
RMSE	0.7628137758	0.7790030366	0.753719822	0.7447463099
MAE	0.3672300526	0.4183963892	0.3640867674	0.3590722026
BDX	Modelo Base	MLP-SX-T	RF-SXY-T	SVR-SXY-T
RMSE	1.428662787	1.493289071	1.439335718	1.40563918
MAE	0.6353006782	0.7811032267	0.6557515233	0.6172360993
IBM	Modelo Base	MLP-SX-T	RF-SXY-T	SVR-SXY-T
RMSE	1.302131596	1.33175027	1.317387519	1.301771969
MAE	0.7466066406	0.7964280808	0.7666667747	0.742073423
PROM	Modelo Base	MLP-SX-T	RF-SXY-T	SVR-SXY-T
RMSE	1.0835	1.33175027	1.0844	1.0732
MAE	0.5175	0.7964280808	0.5275	0.5083

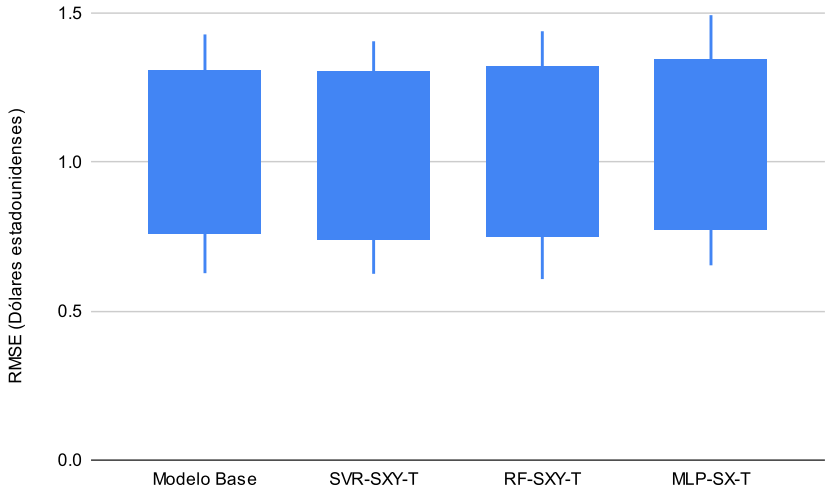

Figura 3.18: Análisis técnico: Diagrama de cajas del error obtenido (RMSE) por los modelos resultantes y el modelo base

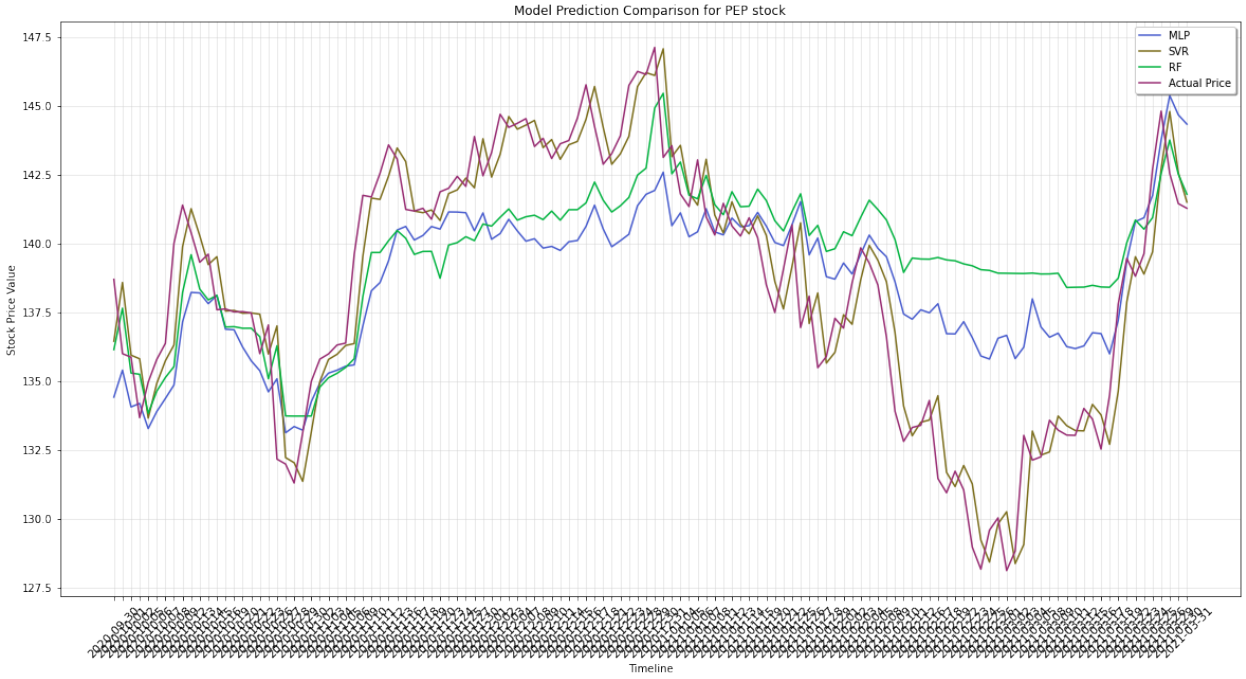

Figura 3.19: Análisis técnico: Comparación del desempeño del modelo MLP-SXY-T en las distintas acciones frente modelo base respecto a la métrica RMSE

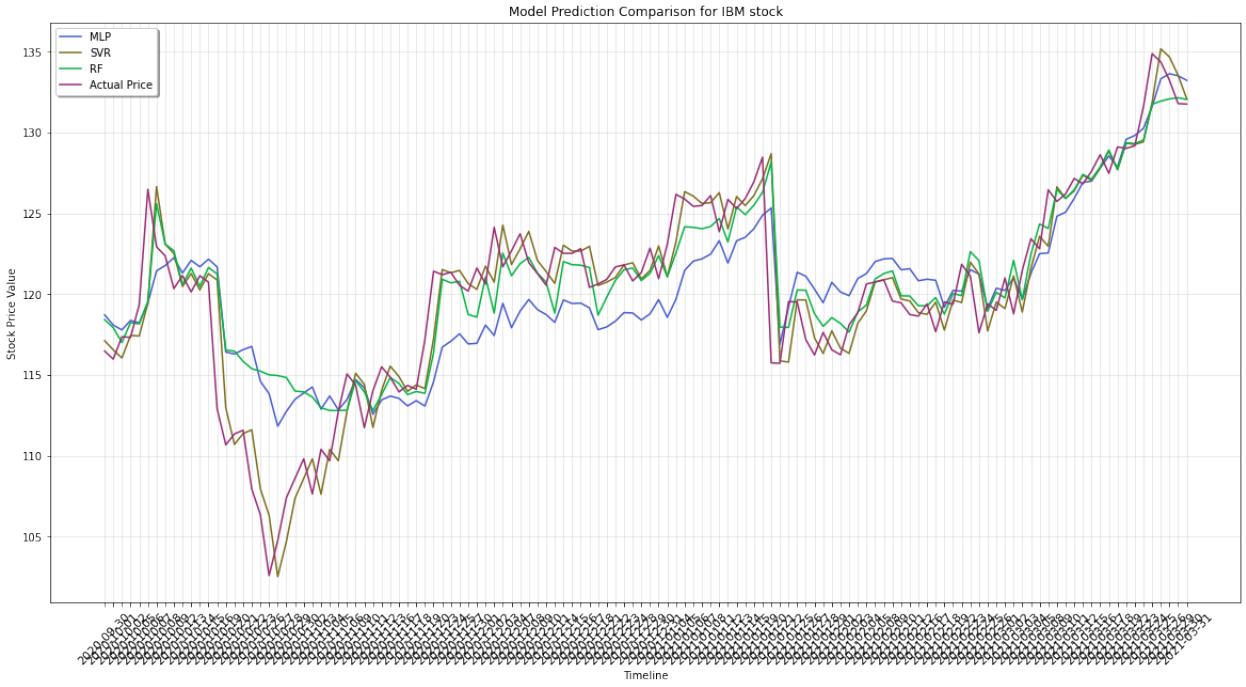

Figura 3.20: Análisis técnico: Comparación del desempeño del modelo MLP-SXY-T en las distintas acciones frente modelo base respecto a la métrica RMSE

3.3.3. Resultados del Backtesting

El proceso de backtesting consistió en poner a prueba las acciones PEP e IBM, las cuales representan la acción que mejor y peor desempeño tuvieron en las predicciones respectivamente. Para esto se aplica la estrategia descrita en el capítulo anterior. El criterio para valorar los resultados obtenidos del backtesting se encuentra en las ganancias en dólares estadounidenses obtenidas a lo largo de dos trimestres, desde 2020-09-30 hasta 2021-03-31, donde para análisis técnico representa cada día de este periodo de tiempo. Se utilizan los tres modelos construidos en la segunda etapa para realizar las predicciones.

Se hace uso de una gráfica que describe las acciones tomadas por el algoritmo de backtesting implementado. Las tres variables que se encuentran en la gráfica cambian de valor dependiendo de la acción tomada si es compra o venta.

Los resultados que se aprecian en la Tabla 3.7 muestran que el modelo que mejor desempeño obtuvo en el backtesting, es RF con una ganancia de $7,54 (siete dolares con cincuenta y cuatro centavos) en PEP y de $25,67 (Veinticinco con sesenta y siete centavos) en IBM. Mientras que el modelo que implementa SVR tuvo el peor desempeño de los tres modelos utilizados. Esta es una situación a destacar, dado que según los resultados mostrados hasta ahora con las métricas utilizadas (RMSE y MAE) el modelo de máquinas de vectores de soporte es el que presenta un mejor comportamiento en las predicciones, según las métricas. Sin embargo, estas métricas no evidencian unos mejores resultados en el backtesting. Este fenómeno podría deberse a lo descrito con anterioridad en el análisis de resultados, donde se decía que la predicción que realiza el modelo de SVR es similar a lo que propone el modelo base.

Tabla 3.7: Análisis técnico: Resultado del backtesting para las distintas técnicas

Análisis Técnico	Acción	Inicial	Ganancia	Valor final
RF	PEP	$1000	$7,54	$1007,54
RF	IBM	$1000	$25,67	$1025,67
SVR	PEP	$1000	$3,41	$1003,41
SVR	IBM	$1000	$6,17	$1006,17
MLP	PEP	$1000	$4,79	$1004,79
MLP	IBM	$1000	$20,67	$1020,67

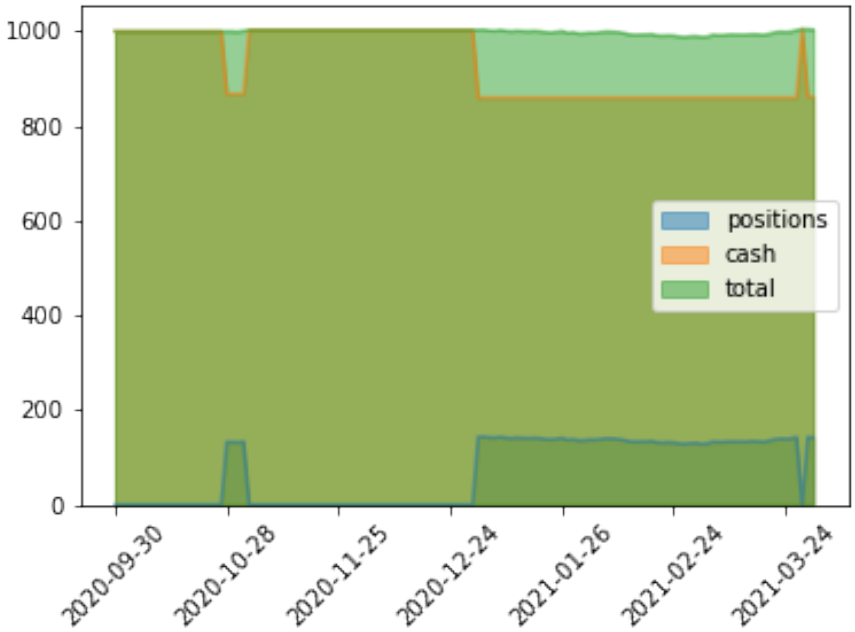

Figura 3.21: Acción PEP: Resultado backtesting técnica SVR análisis técnico

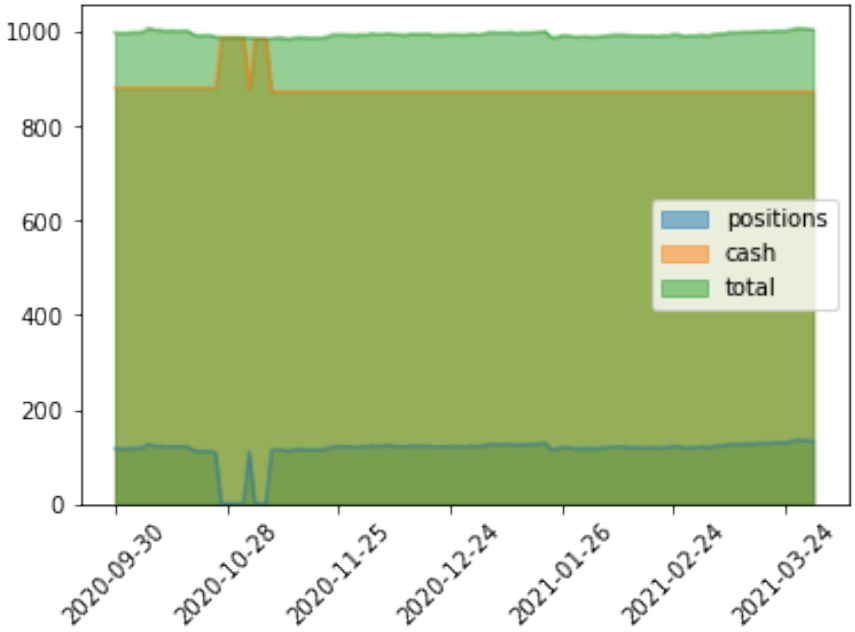

Figura 3.22: Acción IBM: Resultado backtesting técnica SVR análisis técnico

Capítulo 4

Análisis Fundamental

4.1. Preparación de los datos

4.1.1. Recolección de los datos

En primera instancia, el proceso de recolección de los datos se abordó para los informes económicos fundamentales, encontrados en el sitio web *ycharts.com*, de las empresas que conforman el S&P500.

Para obtener los informes del sitio web, se recurrió a programar una rutina de *webscrapping* para así obtener los datos necesarios. Para los datos de análisis fundamental se agregan dos atributos al conjunto de datos, los cuales se obtienen de datos técnicos y representan el precio de la acción y el precio del siguiente trimestre para cada uno de los registros en el conjunto de datos. A partir de ese conjunto de datos construido, se eligen las mismas 5 acciones seleccionadas en el análisis técnico

El criterio utilizado para seleccionar estas acciones fue de acciones con el mínimo número de datos faltantes, ser acciones de empresas en distintos sectores de la industria y que tuvieran el máximo número de atributos (98), dado que no todos los atributos se encontraron en todas las empresas o informes periódicos fundamentales.

4.1.2. Construcción de los conjuntos de datos utilizados

Los conjuntos de datos utilizados son los utilizados como materia prima para generar las diferentes variaciones que se tratarán en las siguiente secciones para los datos de análisis fundamental.

Con los datos recolectados para análisis fundamental se generaron nuevas variables, conocidas en el lenguaje del análisis fundamental como 'ratios', los cuales describen de una forma mucho más entendible para los seres humanos la información que se encuentra detrás de todas las variables que se presentan en los informes trimestrales de las empresas.

El conjunto de datos original cuenta con 143 registros y 98 atributos, como se describió en la sección anterior. A partir de algunos de estos atributos se calcularon 7 ratios, los cuales representan los más utilizados por las fuentes consultadas [11][29]. Estos son: P/E (Price/earnings ratio), PEG (Price/Earnings-to-Growth), FCF (Free Cash Flow), P/FCF (Price/Free Cash Flow), P/S (Price/sales ratio), P/B (Price/book ratio) y ROE (Return on equity). En este conjunto de datos original también se encuentra el atributo del precio de la acción de esa empresa para cada uno de los periodos abarcados y la predicción, que en este caso representa el precio de la acción para el siguiente trimestre.

Por otro lado, se designó un conjunto de datos fundamental aparte para aplicarle el algoritmo de PCA para la selección de atributos. Este conjunto de datos también se desprende de los datos

recolectados del sitio web y cuenta también con los atributos adicionales, precio actual y su predicción para el siguiente trimestre, así como también cuenta con los ratios calculados para el conjunto de datos anterior.

Para la formación de este conjunto de datos se tuvieron en cuenta todos los atributos presentes en la fuente, sobre la cual se aplicaron una serie de filtros para tener uniformidad en las dimensiones del conjunto de datos. Se aplicaron los siguientes criterios secuencialmente. El primero indica que el atributo debe estar presente en las 5 acciones seleccionadas. El segundo criterio trata sobre el número de datos faltantes, el cual debe ser menor a un umbral para ser seleccionado. Este umbral se calcula eliminando los datos atípicos y obteniendo la media sobre el número de datos faltantes para cada atributo. Obteniendo de esta manera los atributos mostrados en Tabla 4.1.

Tabla 4.1: Atributos presentes en el conjunto de datos para PCA

Nombre del atributo
AverageBasicSharesOutstanding
AverageDilutedSharesOutstanding
BeginningCash
CashandEquivalents
CashandShortTermInvestments
CashfromFinancing
CashfromInvesting
CashfromOperations
CashfromOperations(TTM)
ChangeinCash
CurrentDebt&CapitalLeaseObligation
CurrentPortionofLongTermDebt
DividendPerShare
EBIT
EPSBasic
EPSBasicfromContinuingOperations
EPSDiluted
EPSDiluted(TTM)
EPSDilutedfromContinuingOperations
EndingCash
FCF
IncomefromContinuingOperations
NetChangeinCapitalExpenditures
NetChangeinCapitalExpenditures(TTM)
NetCommonEquityIssued(Purchased)
NetDebtIssuance
NetIncome

NetIncome(TTM)
NetPP&E
Non-CurrentPortionofLTDandCapitalLeaseObligation
Non-CurrentPortionofLongTermDebt
NormalizedBasicEPS
NormalizedDilutedEPS
NormalizedIncome
OperatingRevenue
OtherIncomeandExpenses
P/B
P/E
P/FCF
P/S
PEG
Prediction
Pre-TaxIncome
PricePerShare
ProvisionforIncomeTaxes
ROE
RetainedEarnings
Revenue
Revenue(TTM)
ShareholdersEquity
TotalAssets
TotalCurrentAssets
TotalCurrentLiabilities
TotalDividendsPaid
TotalLiabilities
TotalLongTermAssets
TotalLongTermLiabilities
TotalNetChangeinInvestments
TotalOperatingExpenses
TotalRecievables

En Figura 4.1 se pueden observar las etapas descritas anteriormente para la conformación de estos dos conjuntos principales que se utilizarán posteriormente para la generación de distintas variaciones de los mismos.

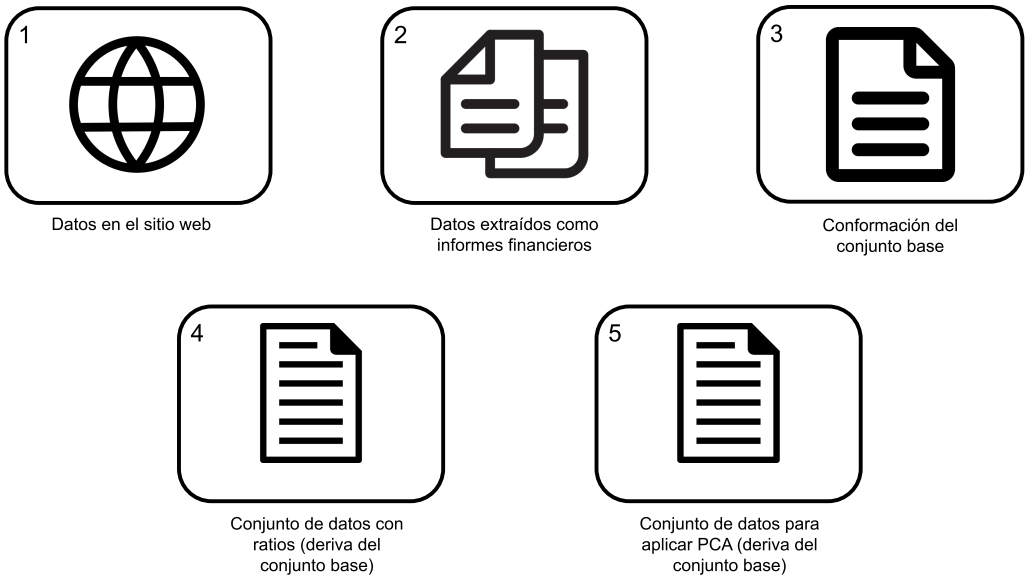

Figura 4.1: Esquema del tratamiento de los datos para análisis fundamental

4.1.3. Análisis de los datos

El conjunto de datos de análisis fundamental de ratios cuenta con 143 registros y 8 atributos por cada registro. Y el conjunto generado para aplicar PCA cuenta con 143 registros y 61 atributos por cada registro. Esto para las 5 acciones seleccionadas.

En los dos conjuntos se encontró que los datos faltantes por fila o periodo de tiempo se concentran sobre todo en los años iniciales, estando la mayor parte de los datos faltantes entre los años 1985 y 1990.

El mayor número de datos faltantes por atributo es 112 para el conjunto de ratios y 137 para el conjunto sobre el que se va a aplicar PCA.

Las gráficas de cajas y bigotes generadas para los dos conjuntos, como se puede observar en Figura 4.2 y Figura 4.3 nos proporcionan información acerca de la gran diferencia de escalas que se maneja entre atributos. A su vez, se observan también atributos con correlaciones muy altas, superando el 90 % de correlación con algunas columnas. En la Figura 4.4 se puede apreciar este fenómeno.

4.1.4. Procesamiento

Los conjuntos de datos generados fueron sometidos a una serie de procesos para llegar a los conjuntos finales con los que se alimentarán los distintos modelos de aprendizaje automático. Estos procesos tienen que ver con la limpieza de los datos y las distintas transformaciones realizadas como experimentación sobre los datos y los posibles resultados que pueden generar.

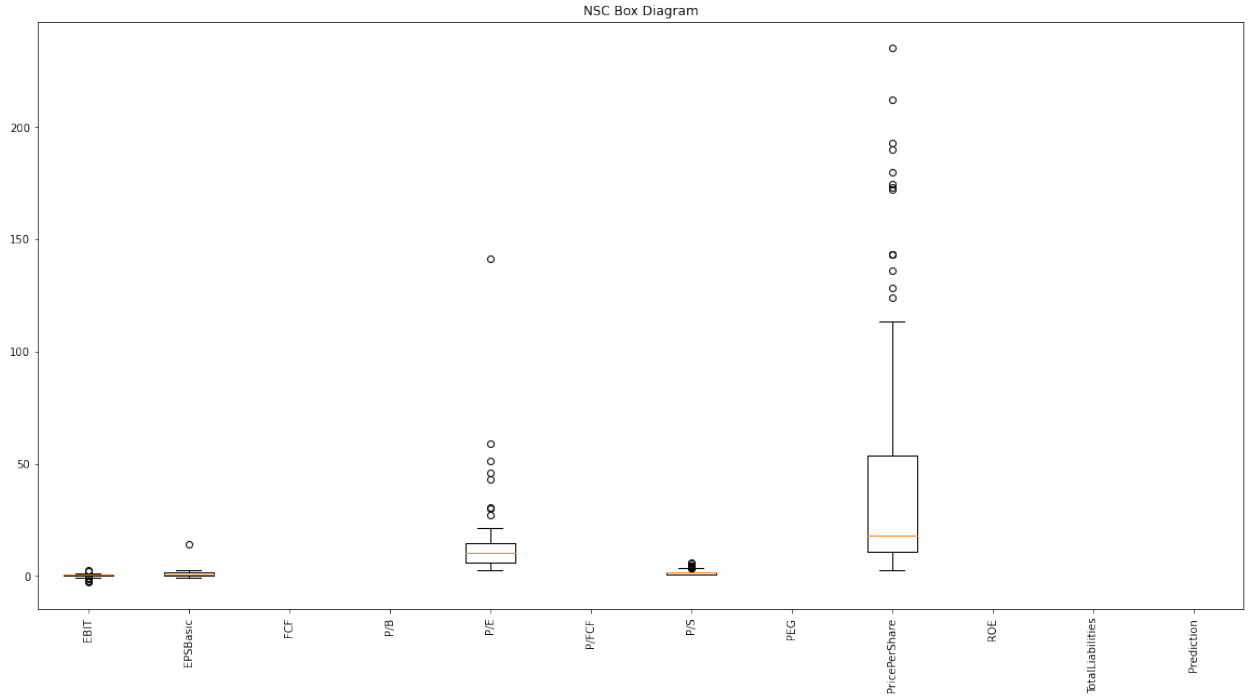

Figura 4.2: Diagrama de cajas y bigotes para la acción NSC para el conjunto de ratios

4.1.4.1. Limpieza de los datos

Tanto para el conjunto de datos donde se utilizan únicamente los ratios así como también para el conjunto al que se aplicará PCA se realizaron las siguientes actividades de limpieza de datos.

En primera instancia se reemplazaron los valores nulos o vacíos aplicando el método "forward fill"haciendo que los registros nulos tomen el valor no nulo más reciente en el conjunto de datos hasta encontrar otro valor no nulo y repetir el proceso.

Para el conjunto de datos de ratios se eliminaron los atributos P/S y P/B, dada su alta correlación con el atributo que indica el precio de la acción para un determinado periodo, siendo esta correlación entre variables independientes mayor al 95 %, indicando que hay redundancia en la información. Del mismo modo se eliminaron los atributos P/FCF, P/E, PEG, debido a su casi nula correlación con la variable a predecir. De este modo, para el conjunto de ratios se obtiene una matriz de correlación como se ejemplifica en la Figura 4.5, donde se observa una gran diferencia con la gráfica en la etapa de análisis.

En el caso de los datos para análisis fundamental es necesario reducir el número de características en el conjunto de datos para evitar incurrir en "La Maldición de la Dimensionalidad".

Finalmente, se obtiene un conjunto de datos de un total de 3 características, siendo estos: *Return on Equity*, *Free Cash Flow* y el precio de la acción al final del trimestre. Y, por otro lado, se tiene el conjunto de datos al que se le aplicará PCA, contando con los mismos 61 atributos seleccionados.

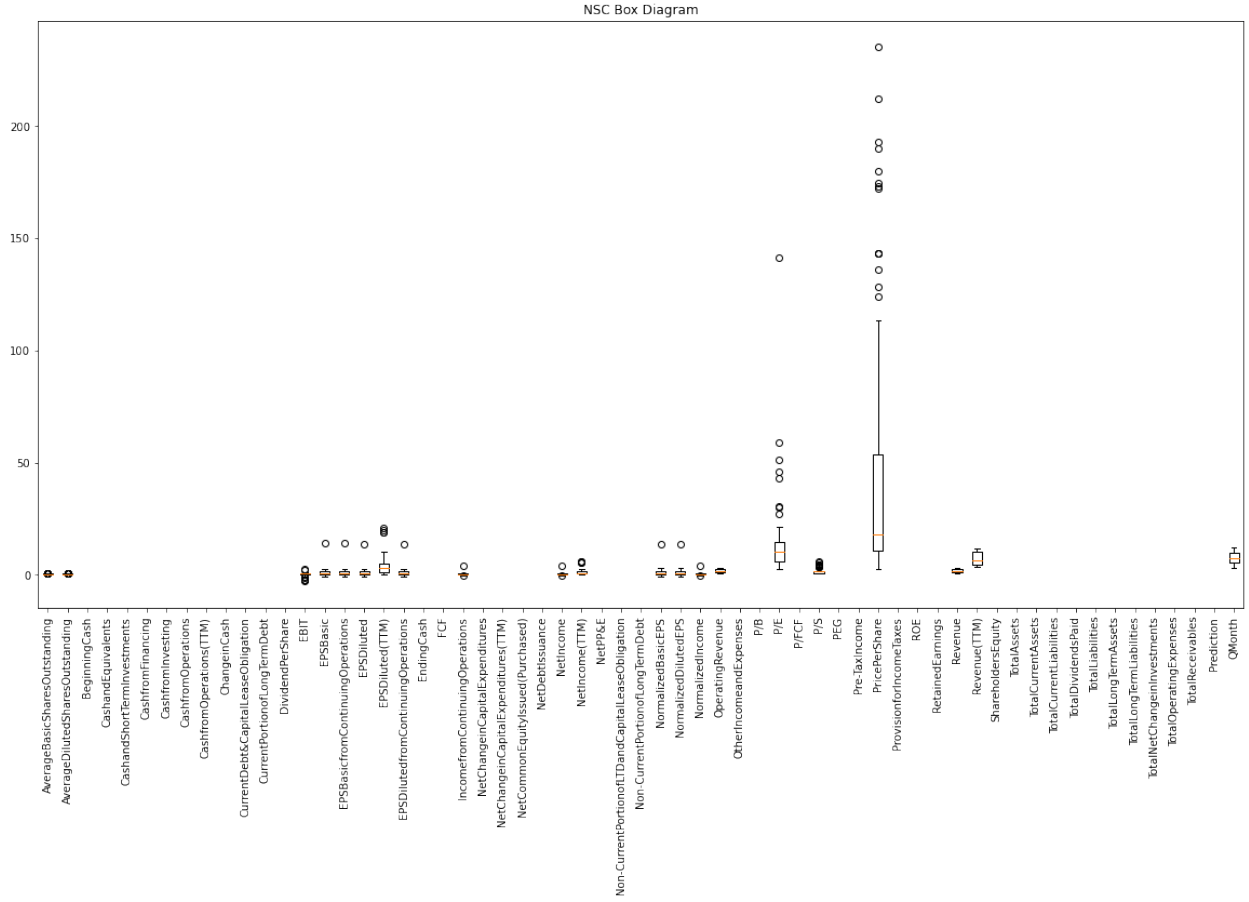

Figura 4.3: Diagrama de cajas y bigotes para la acción NSC para el conjunto para PCA de datos fundamentales

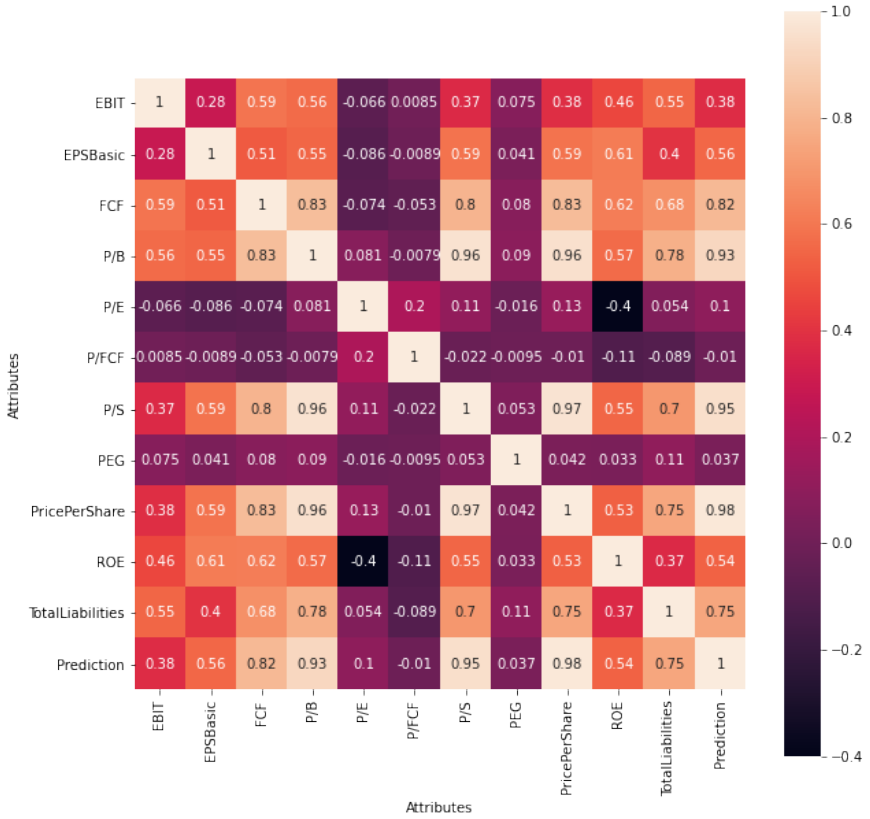

Figura 4.4: Matriz de correlación en etapa de análisis para el conjunto de ratios. Acción: NSC

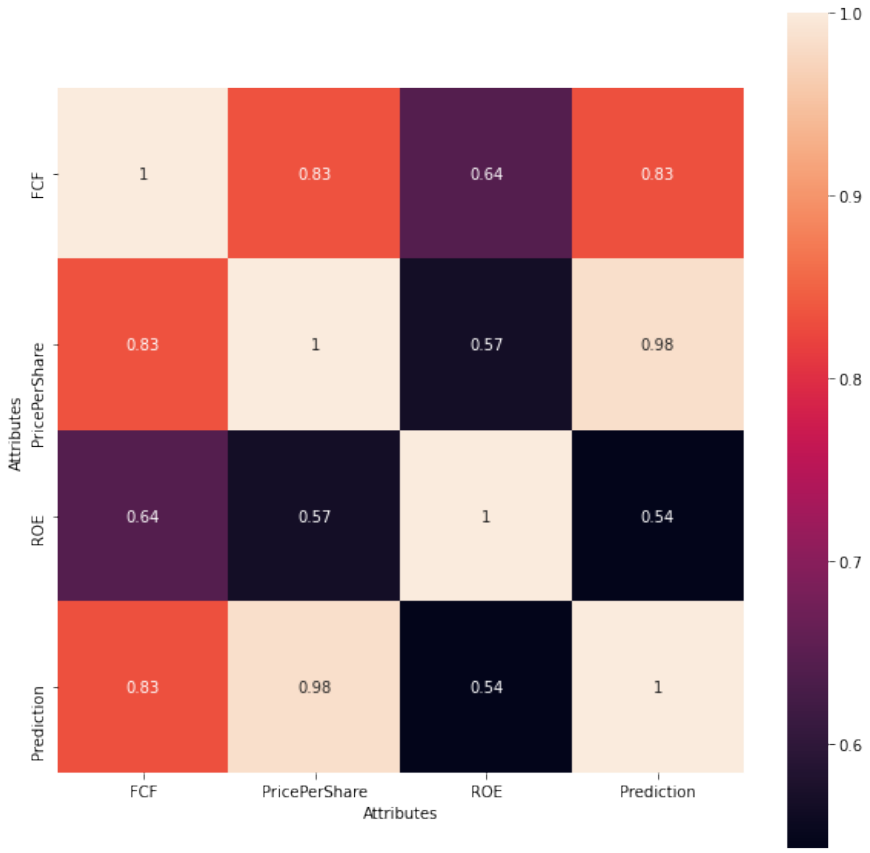

Figura 4.5: Matriz de correlación en etapa de limpieza para el conjunto de ratios. Acción: NSC

4.1.5. Transformaciones

Los conjuntos de datos tratados hasta fueron sometidos a una serie de transformaciones. Las cuales, en algunas se transformaron únicamente los datos de los conjuntos y otras dieron lugar a nuevos conjuntos. Estos nuevos conjuntos de datos se utilizarán para posteriormente realizar una extensa exploración en busca del conjunto de datos que tuviera un mejor desempeño con la técnica de aprendizaje automático utilizada.

En total se crearon cuatro variaciones de conjuntos, el primero siendo en el que se seleccionaron 'ratios', el segundo en el que se aplica PCA al conjunto que se ha ido construyendo para aplicar este algoritmo y el conjunto de datos de experimentación utilizando el método de ventana deslizante tanto para el conjunto de ratios como para el conjunto PCA, siendo estos la tercera y cuarta variante. Adicionalmente cada uno de estos conjuntos de datos posee una versión donde la variable de salida se encuentra estandarizada al igual que el resto del conjunto y otra variante donde la variable de salida no se estandariza pero las de entrada sí. Por lo tanto, se cuenta con un total de 8 conjuntos (como se puede observar en la Tabla 4.2) de datos para cada una de las 5 acciones.

Tabla 4.2: Características de los conjuntos generados

	Conjunto base		Ventana Deslizante		Estandarización	
Conjunto	Atrib. sel.	PCA	Sin VD	VD	X-estándar	XY-estándar
SX	x		x		x	
SXY	x		x			x
W-SX	x			x	x	
W-SXY	x			x		x
PCA-SX		x	x		x	
PCA-SXY		x	x			x
W-PCA-SX		x		x	x	
W-PCA-SXY		x		x		x

4.1.5.1. Estandarización

Se aplicó una transformación usando escalado estándar tal y como se menciono en el análisis técnico 3.1.5.1 .

4.1.5.2. Ventana deslizante

El proceso de generar los conjuntos con ventana deslizante se realizó tanto para los conjuntos de datos que utilizan los ratios o indicadores tanto como para los que hacen uso de PCA. Se utilizó un tamaño de ventana de 4, representando que con los datos fundamentales de 4 trimestres o 1 año se puede predecir el valor de la acción para el siguiente trimestre. Para análisis técnico se utilizó un tamaño de ventana de 7, representando que con los datos de 7 días es posible predecir el siguiente día.

Para realizar esta transformación de los conjuntos de datos utilizados se diseñó un código que agrupa en un registro los datos de todos los atributos para el tamaño de ventana seleccionado añadiendo el atributo de salida, en este caso, el precio del siguiente día o trimestre correspondiente al tamaño de la ventana.

4.1.5.3. Análisis de componentes principales (PCA)

A partir del conjunto de datos que se ha venido creando para aplicar PCA se utilizó finalmente el algoritmo de PCA para obtener un "Varianza explicada media"del 95 %. De este modo, el algoritmo logró ajustarse a este valor con 15 componentes generados a partir de los atributos originales. Sin embargo, solo un componente de los 15 totales, comparte correlación con la propiedad a predecir, situación que se puede evidenciar en la matriz de correlación generada para este conjunto de datos Figura 4.6.

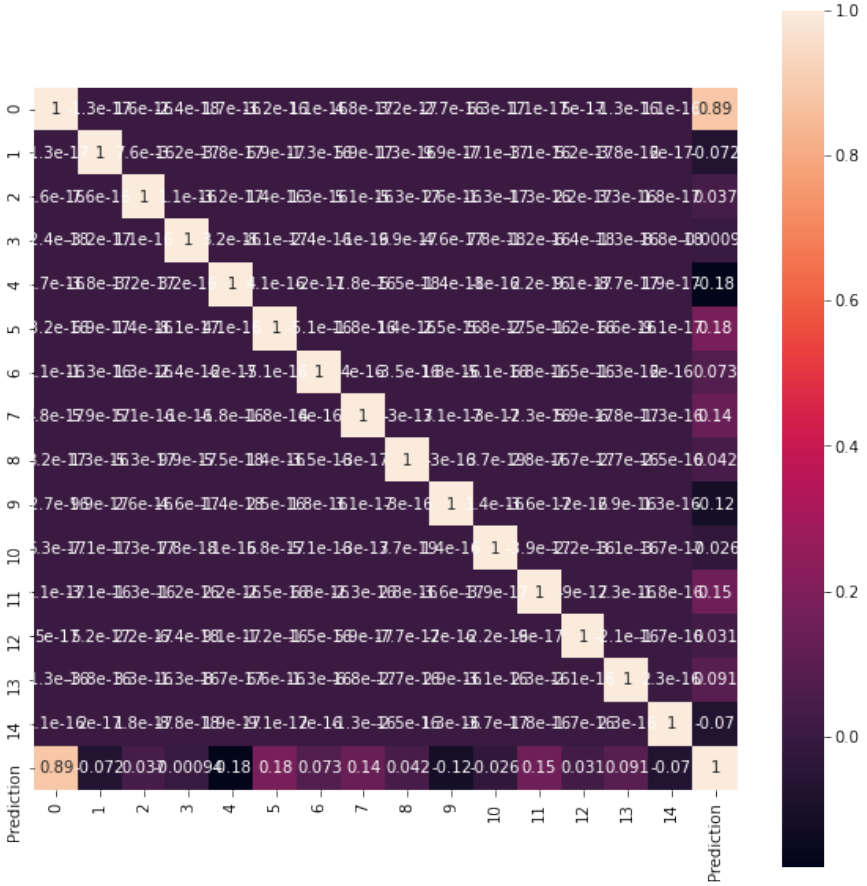

Figura 4.6: Matriz de correlación después de ejecutar PCA para el conjunto de datos fundamentales. Acción: NSC

4.2. Desarrollo de los modelos

4.2.1. Herramientas utilizadas

Para el desarrollo de los modelos de aprendizaje automático se utilizó el lenguaje de programación Python, en su versión 3.6.12. En adición de librerías como scikit-learn 0.24.1, pandas 1.1.5, numpy, seaborn y matplotlib, para el manejo de los datos, el propio proceso de aprendizaje automático y la graficación.

4.2.2. Conformación del modelo base

El modelo base hace alusión al modelo de referencia que se utilizará a lo largo del análisis de los resultados para ser comparado con los resultados de los modelos que se construirán.

Para este proyecto, el modelo base consiste en una predicción básica del precio de una acción, la cual dice que el precio del siguiente trimestre será igual al de este trimestre.

Para lograr este objetivo, se creó un pequeño algoritmo, el cual se encarga, en primera instancia, de realizar la separación de datos en entrenamiento y prueba, donde se utiliza un tamaño para el conjunto de prueba del 20 % del total del número de datos haciendo uso de una semilla que controla la aleatoriedad con la cual se seleccionan los ejemplos para incluirlos en el conjunto de entrenamiento o en el de prueba. Este es un elemento importante, dado que prepara un mejor terreno para la comparación con los modelos de aprendizaje automático, dado que comparten esta característica.

Posteriormente se genera un nuevo arreglo, el cuál representará la predicción y contendrá el precio del siguiente periodo en relación al precio se encuentra en el arreglo que representa el conjunto de prueba. Este proceso se realiza para las 5 acciones seleccionadas, obteniendo un total de 5 modelos base.

4.2.3. Selección del conjunto de datos para cada técnica

Esta fue la primera etapa del desarrollo de los modelos, la cual consistió en la exploración de los distintos conjuntos de datos generados en el capítulo anterior Tabla 4.2 poniéndolos a prueba con cada una de las técnicas de aprendizaje automático seleccionadas. Donde el objetivo principal de este proceso es seleccionar el conjunto de datos que obtenga el mejor desempeño para cada algoritmo de aprendizaje automático utilizado.

En la selección de parámetros para la construcción de estos modelos se realizó un proceso de GridSearchCV para cada conjunto de datos teniendo en cuenta los siguientes hiper-parámetros Tabla 4.3:

Como se mencionó anteriormente, para la división en conjunto de datos para entrenamiento y conjunto de prueba, se utiliza una semilla para mantener una uniformidad en el modo que se seleccionan estos ejemplos para ser comparados con mayor certeza con el modelo base.

En este proceso se generaron un total de 90 modelos de aprendizaje automático basados en cada una de las 3 técnicas seleccionadas alimentado con cada uno de los 8 conjuntos de datos para cada una de las 5 acciones.

Tabla 4.3: Primera etapa de búsqueda de hiper-parámetros para las distintas técnicas de aprendizaje automático. Análisis Fundamental

Fundamental	Hiper-parámetros	Valores Seleccionados
SVR	Kernel	['linear', 'poly', 'rbf', 'sigmoid']
SVR	C	[1, 3, 5, 7]
SVR	Gamma	[0.00009, 0.00008, 0.00007, 0.00006, 0.0001]
SVR	Epsilon	[0.0001, 0.0003, 0.0005, 0.0008, 0.01]
RF	n estimators	[10, 20, 30, 40, 50, 100]
RF	min samples leaf	[1,2,3,4]
RF	max features	['auto', 'sqrt', 'log2']
RF	oob score	[True,False]
RF	max depth	[3, None]
RF	criterion	mae
MLP	hidden layer sizes	[(20,), (30,), (50,), (20,20)]
MLP	activation	['identity', 'logistic', 'tanh', 'relu']
MLP	solver	['sgd', 'adam']

4.2.4. Estimación de hiper-parámetros

La estimación de parámetros se realizó en dos etapas. Como se mencionó anteriormente la primera etapa de construcción de los modelos fue una etapa donde no se profundizó en la selección de los hiper-parámetros. Luego en la segunda etapa, en la cual ya se selecciona el conjunto de datos que tiene mayor desempeño con cada técnica. Se busca mejorar los modelos, aquí se decidió tomar unos rangos amplios de los valores de los parámetros, y con ayuda de grid search se observa cuáles son los valores para los hiper-parámetros que mejor comportamiento ofrecían y a medida que se realizaban varias ejecuciones se acotaba aún más la amplitud de los rangos de los parámetros.

En la Tabla 4.4 se describe el espacio de búsqueda utilizado por el algoritmo GridSearchCV para determinar los mejores valores para cada uno de los hiper-parámetros tratados. Después de haber ejecutado el proceso de búsqueda de parámetros para cada técnica se obtienen los valores resultantes que serán utilizados en el entrenamiento final de los modelos. Estos resultados se encuentran consignados en Tabla 4.5.

4.2.5. Entrenamiento y prueba de los modelos

El entrenamiento y la prueba en el análisis fundamental se llevo a cabo de igual manera que en el análisis técnico 3.2.5 haciendo uso de técnicas como gridsearch para la búsqueda de hiper-parámetros y holdout para la división de los conjuntos de datos para entrenamiento y pruebas.

Tabla 4.4: Segunda etapa de búsqueda de hiper-parámetros según la técnica de aprendizaje automático. Análisis fundamental

Fundamental	Hiper-parámetros	Valores
SVR	Kernel	['linear', 'poly', 'rbf', 'sigmoid']
SVR	C	list(np.arange(0.1, 5, 0.1))
SVR	Gamma	['scale', 'auto']
SVR	Epsilon	list(np.arange(0.001, 1, 0.001))
RF	n estimators	list(range(40,121))
RF	min samples leaf	list(range(1,21))
RF	max features	['auto', 'log2', 'sqrt', 1, 2, 3]
RF	oob score	[True, False]
RF	max depth	[1,2,3,4,5,None]
MLP	hidden layer sizes	[(2,), (2,1), (2,2), (2,3), (2,4), (2,5), (2,2,2), (2,2,2,2), (3,), (3,3), (3,3,3)]
MLP	activation	['identity', 'logistic', 'tanh', 'relu']
MLP	solver	['lbfgs', 'sgd', 'adam']
MLP	alpha	[0.0000431, 0.00009, 0.0001, 0.0002, 0.0003, 0.0004, 0.0005, 0.0006, 0.0007, 0.0008, 0.0009]
MLP	learning rate	['constant', 'invscaling', 'adaptive']
MLP	learning rate init	[0.0245, 0.0246, ..., 0.0259] - [0.01,0.02,...,0.09]

Tabla 4.5: Resultados de la búsqueda de hiper-parámetros para las distintas técnicas de aprendizaje automático. Análisis fundamental

Fundamental	Hiper-parámetros	Valores seleccionados
SVR	Kernel	linear
SVR	C	1,8000000000000003
SVR	Gamma	auto
SVR	Epsilon	0,002
RF	n estimators	49
RF	min samples leaf	3
RF	max features	auto
RF	oob score	True
RF	max depth	None
MLP	hidden layer sizes	(3,3,3)
MLP	activation	relu
MLP	solver	lbfgs
MLP	alpha	0,0000431
MLP	learning rate	constant
MLP	learning rate init	0,0245

4.2.6. Backtesting

En el backtesting para el análisis fundamental se desarrolla de la misma manera que para el análisis técnico 3.2.6, se usan las dos mismas acciones IBM y PEP, y la misma estrategia de comprar bajo vender alto, con la diferencia de que los datos que se obtienen para el backtesting en el análisis fundamental, son datos con una temporalidad trimestral.

4.3. Resultados y análisis

4.3.1. Resultados de la exploración de conjuntos de datos

Como bien se expresó en anteriores secciones, se realizó una exploración sobre distintos conjuntos de datos construidos a partir de dos conjuntos principales, los cuales eran: conjunto de ratios o indicadores (según si se trata de análisis técnico o fundamental) y el conjunto sobre el que se aplicaría PCA. Esta exploración se realizó con el fin de determinar cuál de los distintos conjuntos Tabla 4.2 se desempañaba mejor con las distintas técnica de aprendizaje automático.

Por lo tanto, en esta sección se mostrarán los resultados obtenidos de esta exploración en comparación a los resultados arrojados por el modelo base. En general, para las tres técnicas (SVR, MLP y RF) el conjunto de datos con mejores resultados fue el que agrupa ratios e indicadores. Sin embargo, se presentan ciertas variaciones en estos resultados, haciendo referencia al escalado o no de la variable a predecir.

La herramienta de comparación utilizada en este caso se trata de los diagramas de cajas y bigotes, puesto que proporciona una información mucho más integral que la que se puede obtener con un promedio. Estos diagramas se generan a partir de la métrica de error RMSE obtenida de las predicciones realizadas con estos modelos para todas las acciones. De este modo, se obtienen de todas las acciones el valor mínimo, máximo y los cuartiles 1 y 3 para generar el diagrama. La nomenclatura utilizada en los resultados de los modelos, va precedida por la técnica que se utiliza seguida de el conjunto de datos.

4.3.1.1. Resultados sobre Regresor de Vectores de Soporte

Los resultados mostrados en la Figura 4.7 se observa a los conjuntos basados en ratios como los de mejor desempeño para esta técnica, marcando una clara diferencia con los demás conjuntos de datos. Además, las cajas de estos dos conjuntos se sitúan en valores menores a los que se observan en el modelo base. Finalmente se decide por continuar la siguiente etapa con el conjunto de datos SXY por su menor valor en el cuartil 3, lo cual hace que el tamaño de la caja para este conjunto sea menor que la de SX.

4.3.1.2. Resultados sobre Bosques Aleatorios

En los resultados que se muestran en Figura 4.8 se puede observar que las cajas que más se acercan a los valores del modelo base son las de los conjuntos SX y SXY. Superando estas al modelo base en su valor máximo y mínimo, sin embargo, sobrepasando el cuartil 3 del modelo base que

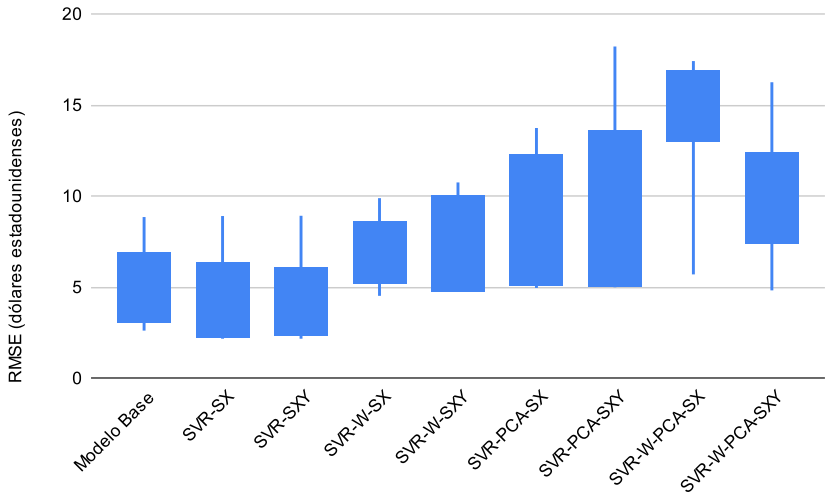

Figura 4.7: Análisis fundamental: comparación de los distintos modelos de la técnica SVR

se encuentra en un error de 6,858204683 dólares estadounidenses. Se selecciona el conjunto SXY, puesto que su máximo y su cuartil 3 se encuentran por debajo de los del conjunto SX.

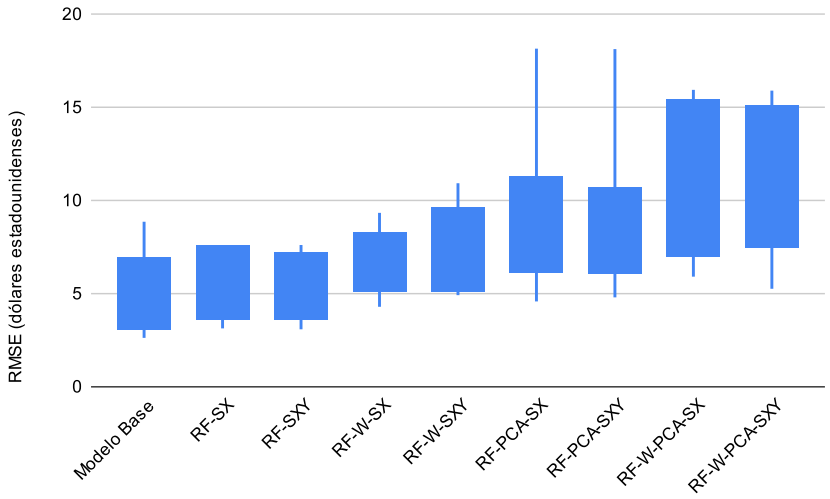

Figura 4.8: Análisis fundamental: comparación de los distintos modelos de la técnica RF

4.3.1.3. Resultados sobre Perceptrón Multicapa

Según se muestra en la Figura 4.9, se observa claramente que el conjunto con la variable de salida sin escalar es el que genera un menor error, siendo sus valores comparables con los del modelo base,

estando los valores del mínimo y cuartil 1 más bajos en comparación. Por otro lado, los resultados para el conjunto SXY muestran una extensión de la caja mayor en su parte superior llegando hasta un error máximo de casi 10 dólares.

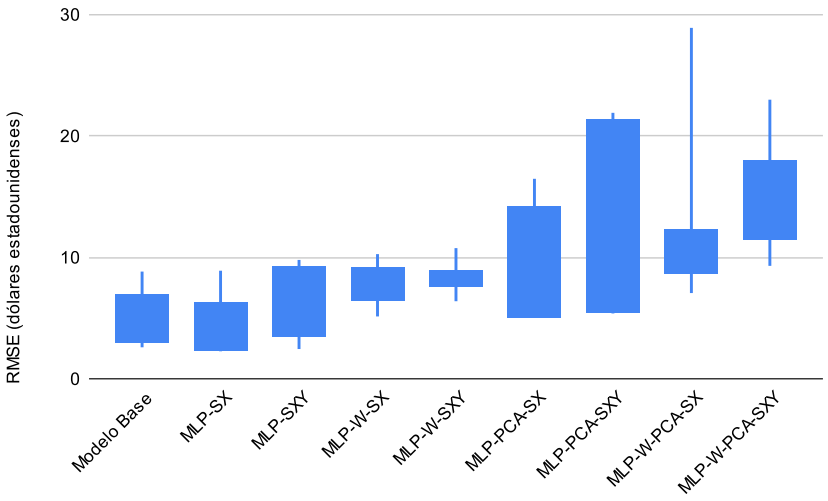

Figura 4.9: Análisis fundamental: comparación de los distintos modelos de la técnica MLP

4.3.1.4. Análisis de resultados

Según lo mencionado anteriormente, se dispone del conjunto de datos basado en ratios para construir los modelos de la siguiente etapa, especificando su variación en la Tabla 4.6. Y resultando las mismas variaciones para todas las técnicas tanto para datos de análisis técnico como para fundamental, de cierto modo corroborando estas decisiones.

Tabla 4.6: Conjuntos de datos seleccionados según técnica. Análisis fundamental y técnico

Técnica	Conjunto
SVR	SXY
RF	SXY
MLP	SX

Como se ha observado en los resultados anteriores, los demás conjuntos de datos no suponen una competencia relevante para el conjunto de indicadores. Por lo tanto, para esta aproximación el algoritmo PCA no resultó siendo acertado con los componentes construidos, dado que estos no representaron correctamente la información original afectando así al modelo de predicción.

Así como tampoco se vio mejor desempeño en las predicciones aplicando la ventana deslizante tal como se especificó para este proyecto. Cumpliendo siempre la función de aumentar el error como se puede ver tanto para los conjuntos basados en ratios como los basados en PCA.

4.3.2. Resultados de los modelos de aprendizaje automático

Como se describió en la sección "Desarrollo de los modelos", la segunda etapa respecto al desarrollo de los modelos consistió en realizar el entrenamiento y prueba de los mismos con los conjuntos de datos seleccionados de la etapa 1, los cuales son con los que se obtiene mejor desempeño en las predicciones.

En esta sección se mostrarán los resultados obtenidos por cada uno de los modelos construidos para cada una de las 3 técnicas y para cada una de las 5 acciones. Con el objetivo de conocer qué técnica tiene un error menor en comparación con las demás y con el modelo base.

Para cumplir con este objetivo, se utilizan diagramas de cajas y bigotes para realizar comparaciones en las predicciones, dos diagramas de barras, uno donde se muestra el promedio de los errores obtenidos paras las 5 acciones según el modelo utilizado para cada técnica, y el otro diagrama de barras por cada técnica donde se puede observar el comportamiento del modelo con cada una de las distintas acciones en comparación con el modelo base. Se utiliza la siguiente nomenclatura para representar el modelo de segunda etapa, primero antecede el nombre de la técnica, seguido por el conjunto de datos utilizados y por último la letra T significando prueba (test en inglés).

4.3.2.1. Resultados sobre Regresor de Vectores de Soporte

Los resultados agregados mostrados en la Figura 4.11 muestran que el modelo construido tiene, generalmente, un mejor desempeño que el modelo base. Únicamente el valor máximo del modelo construido muestra un error mayor. Sin embargo, es muy notorio el desplazamiento de la caja hacia abajo en referencia al modelo base.

En la Figura 4.10 se puede observar una gran diferencia entre los promedios de los errores entre el modelo base y el modelo de primera etapa (SVR-SXY). Esta mejora también se presenta entre el modelo SVR-SXY y el modelo SVR-SXY-T pero es mucho menos relevante que la comparación anterior.

Se observa en la Figura 4.12 que el modelo SVR obtenido presenta un RMSE menor al modelo base para todo el conjunto de acciones seleccionadas, a excepción de la acción de IBM, la cual supera por unas centésimas al RMSE obtenido en el modelo base para esta acción. En cambio, para las demás acciones hay una notoria diferencia entre el RMSE del modelo base y del modelo SVR-SXY-T.

4.3.2.2. Resultados sobre Bosques Aleatorios

Los resultados agregados mostrados en la Figura 4.14 muestran que el modelo RF-SXY-T construido presenta un RMSE mucho más concentrado en valores intermedios en referencia al modelo base. El modelo no alcanza valores de RMSE tan bajos como los del modelo base, debido a que su mínimo y su cuartil 1 se desplazaron hacia arriba.

En la Figura 4.10 se observa una leve mejoría en los valores de los modelos construidos en comparación al modelo base. Esto es debido a lo ya mencionado anteriormente en la Figura 4.14 donde se ve que la caja disminuye su tamaño centrándose en un rango de valores que recae dentro

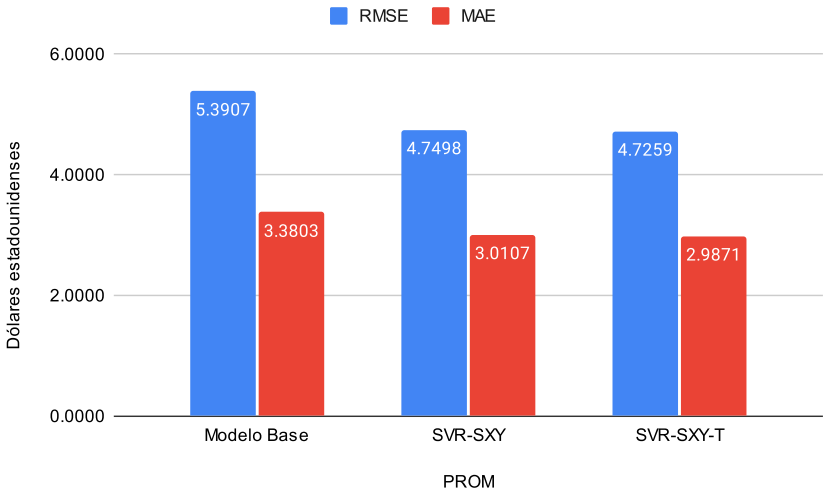

Figura 4.10: Análisis fundamental: Promedio de errores RMSE y MAE para las 5 acciones para modelo base, modelo etapa 1 y modelo etapa 2. Técnica: SVR

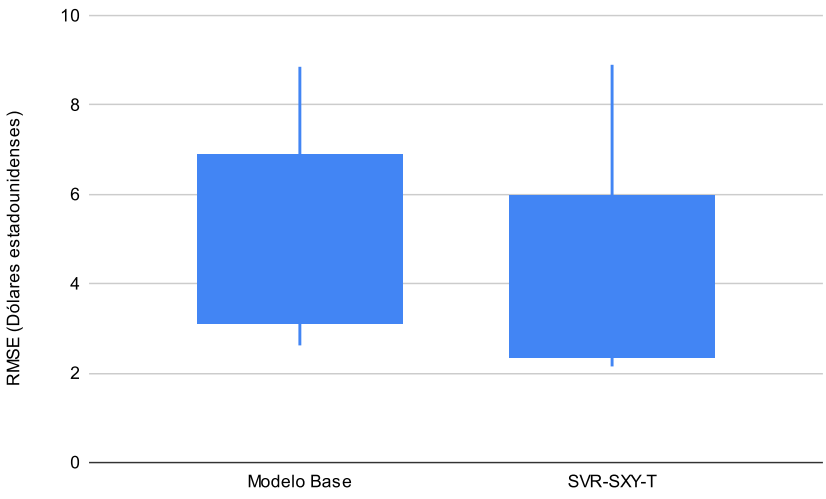

Figura 4.11: Análisis fundamental: Diagrama de cajas y bigotes para el modelo SVR-SXY-T frente al modelo base respecto a la métrica RMSE

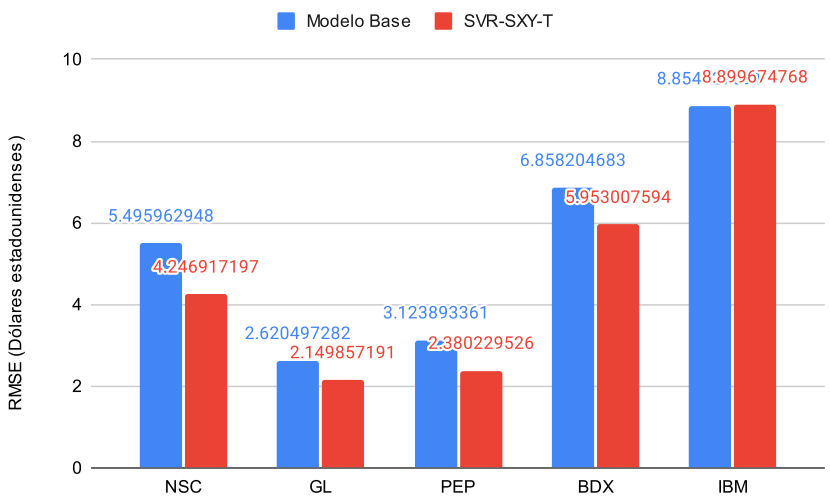

Figura 4.12: Análisis fundamental: Comparación del desempeño del modelo SVR-SXY-T en las distintas acciones frente modelo base respecto a la métrica RMSE

de la caja del modelo base, además que disminuye su máximo y cuartil 3. Por tal razón, se consigue una leve disminución del error promedio.

En la Figura 4.15 se observan resultados variados dependiendo de la acción que se observe. En GL y PEP se tiene un error mayor que en el modelo base. Sin embargo, en NSC, BDX e IBM se tienen valores que varían en distancia del error obtenido con el modelo base, pero nuca siendo mayor a este.

4.3.2.3. Resultados sobre Perceptrón Multicapa

Los resultados agregados que se muestran en la Figura 4.17 muestran claramente cómo se redujo el tamaño de la caja, y además se desplazó hacia abajo. El error que representa la caja es menor en todos los parámetros (mínimo, cuartil 1, cuartil 3 y máximo). El valor máximo sigue siendo elevado, sin embargo, este sigue por debajo del valor máximo en el modelo base.

En la Figura 4.10 se puede observar cómo en cada etapa mejoró el promedio de error tanto para RMSE como para MAE en comparación con el modelo base. Siendo una evolución positiva notoria para estas métricas de error.

La Figura 4.18 muestra un balance positivo para todas las acciones, puesto que en todas ellas se tiene un valor de RMSE inferior al del modelo base, siendo GL la más cercana al valor obtenido con la predicción del modelo base.

4.3.2.4. Análisis de resultados

Según la Tabla 4.7, se puede observar la comparación de cada modelo con las dos métricas de error utilizadas y para cada una de las 5 acciones, incluido el promedio de estas. Visualizando el

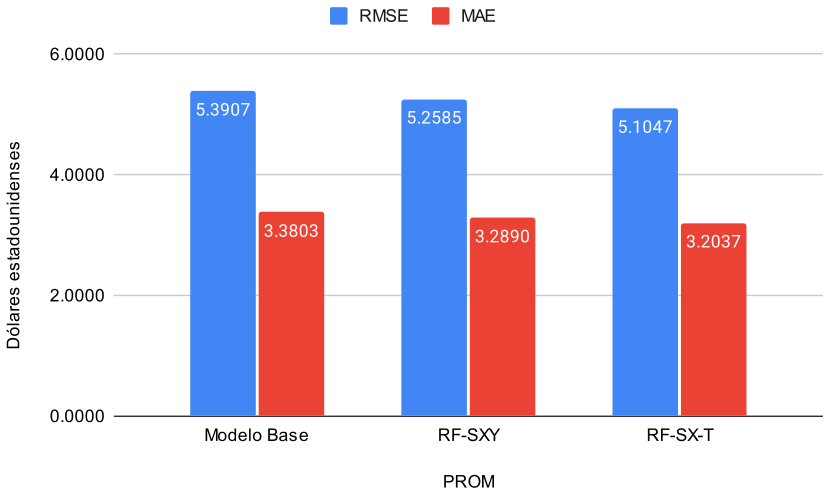

Figura 4.13: Análisis fundamental: Promedio de errores RMSE y MAE para las 5 acciones para modelo base, modelo etapa 1 y modelo etapa 2. Técnica: RF

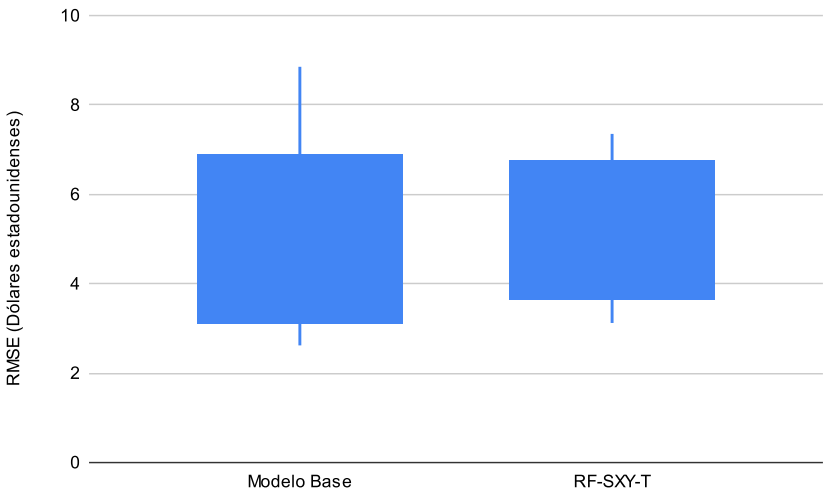

Figura 4.14: Análisis fundamental: Diagrama de cajas y bigotes para el modelo RF-SXY-T frente al modelo base respecto a la métrica RMSE

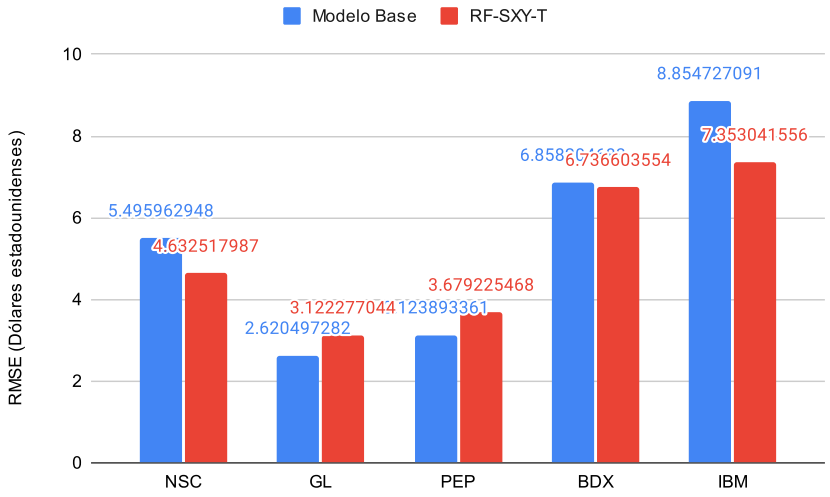

Figura 4.15: Análisis fundamental: Comparación del desempeño del modelo RF-SXY-T en las distintas acciones frente modelo base respecto a la métrica RMSE

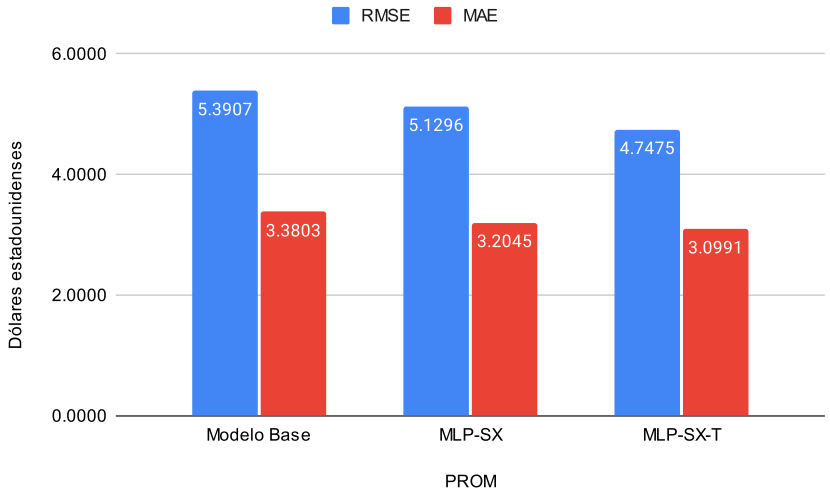

Figura 4.16: Análisis fundamental: Promedio de errores RMSE y MAE para las 5 acciones para modelo base, modelo etapa 1 y modelo etapa 2. Técnica: MLP

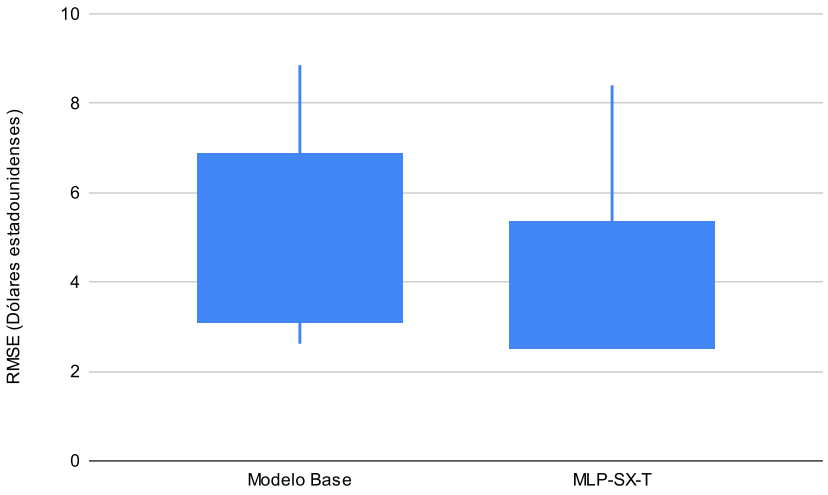

Figura 4.17: Análisis fundamental: Diagrama de cajas y bigotes para el modelo MLP-SXY-T frente al modelo base respecto a la métrica RMSE

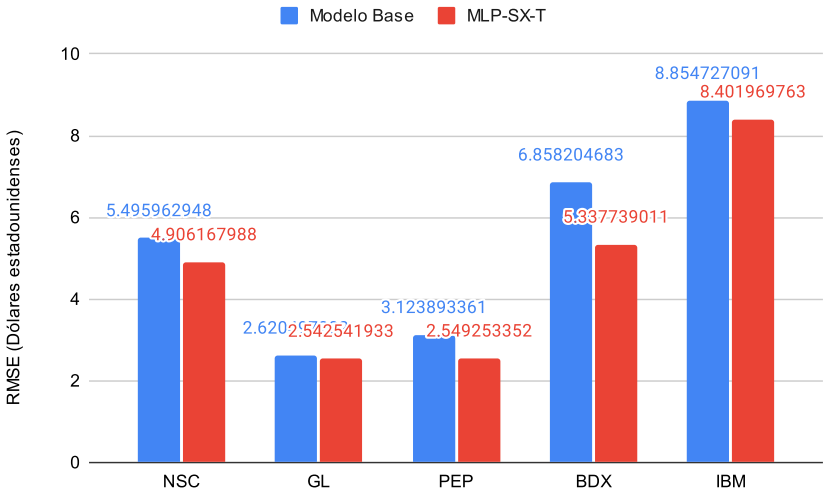

Figura 4.18: Análisis fundamental: Comparación del desempeño del modelo MLP-SXY-T en las distintas acciones frente modelo base respecto a la métrica RMSE

RMSE se muestra que, en promedio, el modelo SVR-SXY-T obtiene errores ligeramente menores de el modelo MLP-SX-T, donde se disputan la victoria en algunas acciones. En último lugar con un promedio notablemente mayor al de los otros dos modelos se tiene el modelo RF-SXY-T.

Sin embargo, al observar con mayor amplitud los resultados de todos los modelos como se muestra en Figura 4.19 es posible generar mayor claridad sobre los resultados que han tenido cada uno de los modelos. Se visualiza la caja generada por los resultados del modelo MLP-SX-T como la que se concentra en valores menores que los demás, aunque teniendo un valor máximo muy prominente. Sin embargo, este dato puede ser tratado como un dato atípico, por lo que sería la opción correcta escoger el modelo MLP-SX-T como el que mejores resultados sobre métricas de error respecta. Con respecto a los otros dos modelos construidos, se puede declarar al SVR-SXY-T como un modelo que genera menores errores que el modelo RF-SXY-T.

Es posible observar que los resultados han sido positivos para los datos de análisis fundamental. Como se muestra en la Figura 4.19 las cajas se mantienen por debajo del modelo base, lo cual es, a priori, un buen indicador del desempeño de estos modelos con datos de análisis fundamental.

Cabe la posibilidad que el uso de estos datos fundamentales haya potenciado la capacidad de predicción de los modelos, dada la información mucho más relacionadas a la realidad de la empresa que dan los ratios fundamentales. También es posible que, debido a la poca cantidad de datos como para entrenar modelos para estas técnicas, el conjunto de pruebas no sea lo suficientemente grande para tener mayor confiabilidad en las métricas de error presentadas.

En la Figura 4.20 y la Figura 4.21 se muestran las 30 últimas fechas de las acciones que PEP e IBM, las cuales representan la acción con mejores resultados y la de peores, respectivamente. En estas dos figuras podemos observar cómo las predicciones se ajustan más para acciones que no tienen precios tan fluctuantes como ocurre en IBM. También es posible observar, que aunque tienen mayor error en la predicción del precio, hacen un buen trabajo en predecir la tendencia que tendrá la acción a futuro. Cabe aclarar que entre estas 30 observaciones se encuentran datos de entrenamiento y prueba, dado que estos son obtenidos aleatoriamente de todo el universo de datos a la hora de realizar la separación.

Tabla 4.7: Resultados de las distintas técnicas para el análisis fundamental

NSC	Modelo Base	MLP-SX-T	RF-SXY-T	SVR-SXY-T
RMSE	5.495962948	4.906167988	4.632517987	4.246917197
MAE	3.360447431	3.05032089	3.032783214	2.650007312
GL	Modelo Base	MLP-SX-T	RF-SXY-T	SVR-SXY-T
RMSE	2.620497282	2.542541933	3.122277044	2.149857191
MAE	1.822139329	1.934540685	1.978948263	1.519792963
PEP	Modelo Base	MLP-SX-T	RF-SXY-T	SVR-SXY-T
RMSE	3.123893361	2.549253352	3.679225468	2.380229526
MAE	1.925721411	1.676717036	2.224793906	1.59863416
BDX	Modelo Base	MLP-SX-T	RF-SXY-T	SVR-SXY-T
RMSE	6.858204683	5.337739011	6.736603554	5.953007594

MAE	3.713997052	3.206606032	3.968482373	3.053908797
IBM	Modelo Base	MLP-SX-T	RF-SXY-T	SVR-SXY-T
RMSE	8.854727091	8.401969763	7.353041556	8.899674768
MAE	6.079072689	5.627548821	4.813265622	6.113194138
PROM	Modelo Base	MLP-SX-T	RF-SXY-T	SVR-SXY-T
RMSE	5.3907	4.7475	5.1047	4.7259
MAE	3.3803	3.0991	3.2037	2.9871

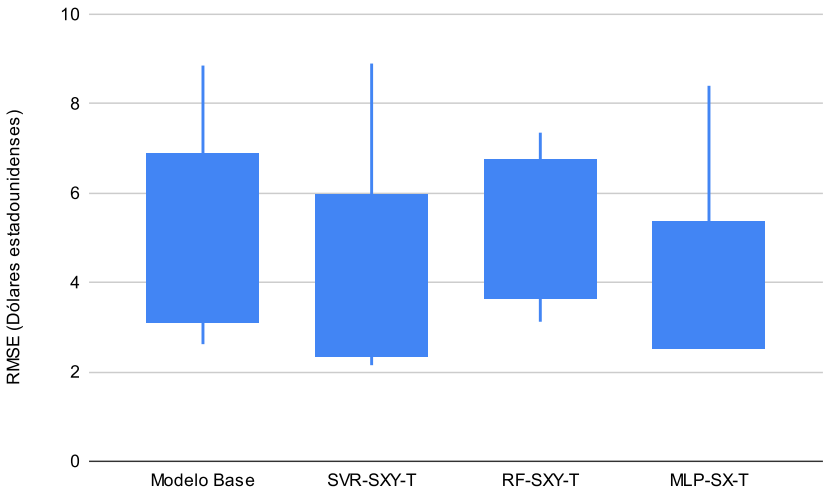

Figura 4.19: Análisis fundamental: Diagrama de cajas del error obtenido (RMSE) por los modelos resultantes y el modelo base

4.3.3. Resultados del Backtesting

El proceso de backtesting consistió en poner a prueba las acciones PEP e IBM, las cuales representan la acción que mejor y peor desempeño tuvieron en las predicciones respectivamente. Para esto se aplica la estrategia descrita en el capítulo anterior. El criterio para valorar los resultados obtenidos del backtesting se encuentra en las ganancias en dólares estadounidenses obtenidas a lo largo de dos trimestres, desde 2020-09-30 hasta 2021-03-31, donde para análisis técnico representa cada día de este periodo de tiempo y para el fundamental representa 3 periodos, es decir, se tienen 3 fechas. Se utilizan los tres modelos construidos en la segunda etapa para realizar las predicciones.

Se hace uso de una gráfica que describe las acciones tomadas por el algoritmo de backtesting implementado. Las tres variables que se encuentran en la gráfica cambian de valor dependiendo de la acción tomada si es compra o venta.

Los resultados que se aprecian en la Tabla 4.8 muestran que el modelo que mejor desempeño obtuvo en el backtesting, es RF con una ganancia de $10,59 (Diez con cincuenta y nueve centavos)

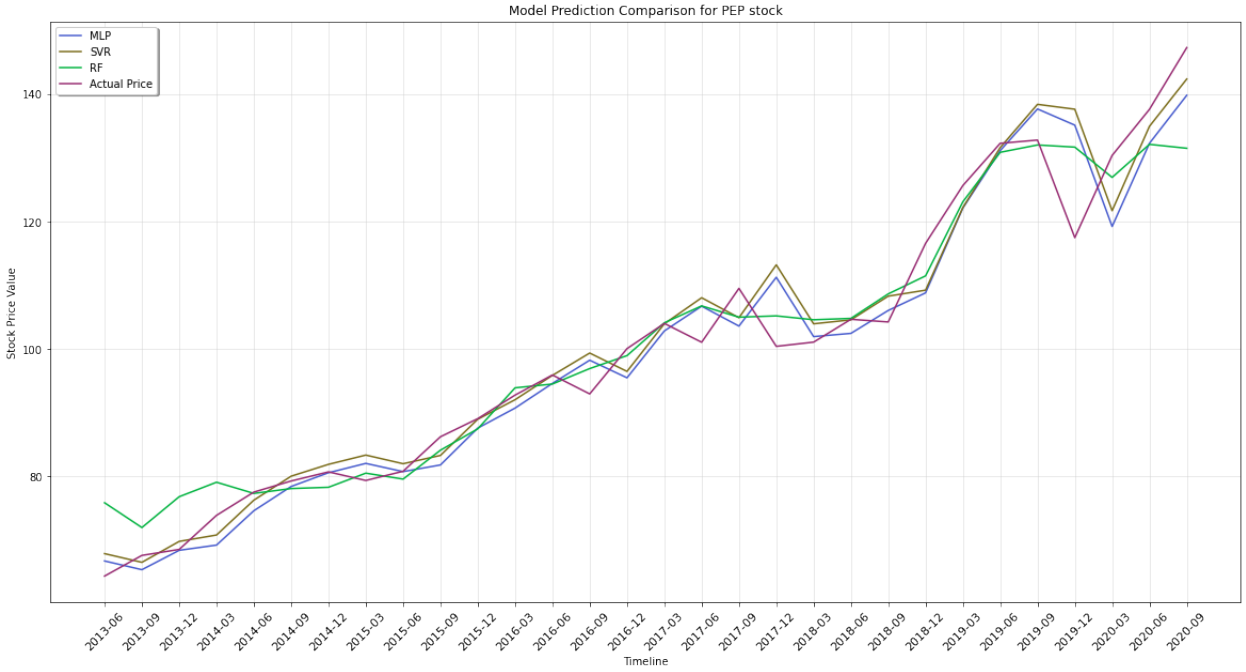

Figura 4.20: Análisis fundamental: Comparación del desempeño del modelo MLP-SXY-T en las distintas acciones frente modelo base respecto a la métrica RMSE

en PEP y \$14,74 (Catorce con setenta y cuatro centavos) en IBM. Siendo, en este caso, SVR el segundo lugar y MLP en tercer lugar generando \$0 dólares en ganancias, puesto que según la predicción que se dio el precio iba en caída, por lo que no sería rentable comprar. Sin embargo, en realidad el precio de estas dos acciones no se comportó así, por lo tanto, la predicción dada por el modelo es errónea. Como se pudo ver gráficamente en los resultados, el modelo de MLP suele quedar levemente por debajo del precio real y esto puede causar que se vea siempre la predicción como pérdidas.

Al igual que observamos con los resultados del backtesting con análisis técnico, también se presenta en estos resultados. El modelo de RF, el cual era el menos confiable debido a sus métricas de error más altas que las de los demás modelos, resulta siendo el modelo que mejor desempeño obtiene en el backtesting por encima de los otros dos modelos. Y la diferencia esta vez está marcada en la acción PEP, la cual es la que mejor desempeño obtuvo en todos los modelos, y por lo tanto, la que mejores resultados debería proporcionar para los modelos que tienen un RMSE más bajo para esta acción.

Figura 4.21: Análisis fundamental: Comparación del desempeño del modelo MLP-SXY-T en las distintas acciones frente modelo base respecto a la métrica RMSE

Tabla 4.8: Análisis fundamental: Resultado del backtesting para las distintas técnicas

Análisis Fundamental	Acción	Inicial	Ganancia	Valor final
RF	PEP	$140	$10,59	$150,59
RF	IBM	$140	$14,74	$154,74
SVR	PEP	$140	$4,92	$144,92
SVR	IBM	$140	$14,74	$154,74
MLP	PEP	$140	$0	140
MLP	IBM	$140	$0	140

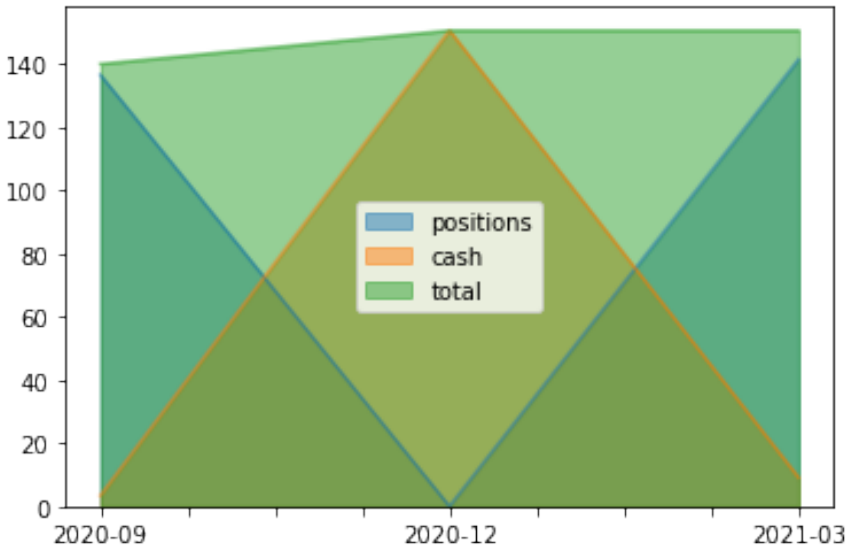

Figura 4.22: Acción PEP: Resultado backtesting técnica RF análisis fundamental

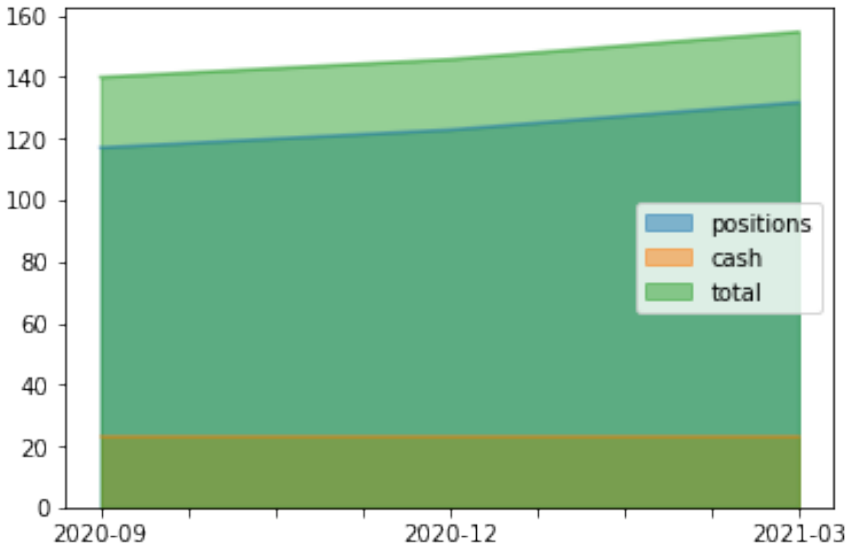

Figura 4.23: Acción IBM: Resultado backtesting técnica RF análisis fundamental

Capítulo 5

Conclusión

Para este estudio se trata el problema de la predicción del precio de acciones del mercado de valores estadounidense haciendo uso de métodos de aprendizaje automático basados en datos de análisis técnico y análisis fundamental. Donde se ponen a prueba 3 técnicas de aprendizaje automático: Regresor de Vectores de Soporte, Bosques Aleatorios y Perceptrón multicapa; haciendo uso de acciones de 5 empresas. Existen numerosos estudios en relación a la predicción del precio de acciones, en particular basados en datos de análisis técnico. Siendo los estudios en este campo basados en datos únicamente de análisis fundamental los que componen una menor parte [11], debido a ciertos obstáculos que complican su desarrollo, los cuales se pudieron comprobar en este proyecto.

Se encontró en la etapa de recolección de datos, un volumen pequeño de datos para análisis fundamental, dada su naturaleza en periodos trimestrales. Lo cual puede terminar generando errores en el resultado, debido a que al contar con un conjunto de entrenamiento tan reducido, se generarían valores sesgados para las métricas utilizadas, en este caso para RMSE y MAE, las cuales utilizan la media, donde es necesario el número total de ejemplos que componen el conjunto para obtener la métrica, afectando claramente este número al valor final.

Con respecto a los resultados de los modelos obtenidos, se muestra que la técnica a la que se le atribuye menor error según la métrica de error MSE para datos de análisis técnico fueron los modelos que implementa la técnica de Regresor de Vectores de Soporte, teniendo resultados en las métricas muy aproximados a las de los modelos de las otras técnicas y contando un un RMSE promedio de $1.07 y un MAE promedio de $0.5 en relación a los resultados de los modelos para cada una de las 5 acciones. En el caso del análisis fundamental, se tiene que fue la técnica de Perceptrón Multicapa la que logró menor error en RMSE con un error promedio de $4.75 con y MAE con un error promedio de $3.1 para los modelos que la implementan. Es posible distinguir que los errores presentados por los modelos de análisis técnico son menores a los de análisis fundamental. Esto es debido a que el cambio diario del precio es mucho menor al del cambio de cada trimestre.

Los resultados de backtesting mostraron que para datos de ambos tipos de análisis el que mejor desempeño obtuvo, según el retorno de inversión fueron los modelos que implementan Bosques Aleatorios. Generando mayor ganancia para la acción de la compañía IBM en análisis técnico y mayor ganancia para la acción PEP en análisis fundamental. Estos resultados dieron a conocer que las métricas de error no representan el desempeño de un un modelo en una situación real como puede ser, en este caso, un backtesting. Por esta razón, los modelos que implementan Bosques Aleatorios, el cual fue el de mayor error en ambos análisis, fue el que obtuvo mayores ganancias a la hora de realizar esta prueba de campo. Esta situación está dada por las diferencias fundamentales entre las métricas de error y el indicador de desempeño utilizado en backtesting, en este caso el retorno de

la inversión o ganancias en un periodo determinado. Estos no se encuentran correlacionados, por lo tanto, tener un error menor no implica necesariamente que se tendrá el mayor retorno.

5.1. Trabajos futuros

Debido a distintas circunstancias como el tiempo, limitación de conocimiento, capacidad de computo y otras variables, no se puede afirmar que se haya construido unos modelos que sean capaces de batir el mercado con grandes rendimientos, a pesar de todo esto se logró aportar una parte en la investigación de este campo. Sin embargo, restan ciertas aproximaciones e ideas que surgieron durante el desarrollo de este proyecto, las cuales podrían ayudar a crear mejores modelos o ahondar más en este campo.

Uno de estos posibles trabajos es la combinación de los dos tipos de análisis fundamental y técnico para así comparar si esto daría lugar a un mejor desempeño. Así como también implementar modelos de ensamble, donde se realicen predicciones sobre ciertos atributos clave, y se tomen estos atributos para realizar una predicción del movimiento del precio de la acción. Luego de este estudio y de haber ahondado más en la literatura, la predicción del valor del precio de una acción es una tarea muy difícil, se podrían conseguir mejores resultados prediciendo el movimiento.

Uno de los retos fue el bajo volumen de datos en análisis fundamental, una solución a este problema podría ser generar artificialmente estos datos para obtener un mayor volumen de estos. De tal manera, también se podría ahondar más en la partición de los datos en entrenamiento para poder determinar si existe *overfitting* o *undefitting* en los modelos.

Sería importante poder explorar modelos con otro tipo análisis como es el análisis de sentimientos, el cual se encuentra en auge actualmente. Este análisis muestra el valor que tiene la influencia de algunas personas o grupos en los mercados a través de redes sociales Esto se puede evidenciar en los 'tweets' de personas relevantes de la plataforma, los cuales son capaces de incrementar el precio de un activo comercial, o los posts que se realizan en lugares como Reddit donde la gente se ha puesto de acuerdo para especular con el precio de una acción como sucedió a inicios de 2021 con la acción de GameStop [30]. Por lo tanto, con este tipo de situaciones se puede observar el posible potencial de este tipo de análisis y es interesante el aporte que puede tener para hacer una mejor predicción.

Otra de las partes donde se podría ahondar en mayor medida, es en el backtesting. Sobretodo para el análisis fundamental, puesto que el número de datos con el que se realizó es demasiado pequeño, lo cual puede llevar a la desconfianza en los resultados. Por otro lado, se podrían probar los modelos con una mejor estrategia que la de vender alto comprar bajo, y llevarla a prueba con un mayor número de acciones formando un portafolio, y de esta manera comprobar el potencial predictivo de los modelos construidos en un ambiente simulado de inversión.

Bibliografía

[1] Investopedia, "Stocks Then And Now: The 1950s And 1970s," 5 2020.

[2] Christina Majaski, "Understanding Fundamental vs. Technical Analysis," 4 2019.

[3] Research and Ranking, "Shocking But True: 90 % People Lose Money In Stocks – Research & Ranking," 11 2019.

[4] YiLi Chien and Paul Morris, "Stock Market Participation Varies Widely by State — St. Louis Fed," 8 2017.

[5] Clay Halton, "Chaos Theory Definition," 7 2019.

[6] Santiagopc, "Diferencia entre fondos de inversión públicos y privados - Rankia," 5 2019.

[7] TranquiFinanzas, "Mejores Fondos de Inversión Colectiva 2019 [actualizado a octubre de 2019]," 2019.

[8] G. A. Wilhelm G., *Technical Analysis in Financial Markets*. Amsterdan: University of Amsterdam, 2003.

[9] C. D. Kirkpatrick and J. R. Dahlquist, *Technical analysis: the complete resource for financial market technicians*. Upper Saddle River, New Jersey: FT Press, second ed., 2007.

[10] M. C. Thomsett, *Getting Started in Fundamental analysis*. Hoboken, New Jersey: John Wiley & Sons, Inc., 2005.

[11] I. K. Nti, A. F. Adekoya, and B. A. Weyori, *A systematic review of fundamental and technical analysis of stock market predictions*, vol. 53. Springer Netherlands, 2020.

[12] Z. Haider Khan, T. Sharmin Alin, and A. Hussain, "Price Prediction of Share Market Using Artificial Neural Network 'ANN'," *International Journal of Computer Applications*, vol. 22, no. 2, pp. 42–47, 2011.

[13] E. F. Fama, "Random Walks in Stock Market Prices," *Financial Analysts Journal*, vol. 21, no. 5, pp. 55–59, 1965.

[14] E. F. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," *The Journal of Finance*, vol. 25, no. 2, pp. 383–417, 1970.

[15] B. G. Malkiel, "A random walk down Wall Street : including a life-cycle guide to personal investing," 1999.

[16] M. Dunne, "Stock Market Prediction Declaration of Originality," *Dept of Computer Science, University College Cork*, vol. 1, no. 1, p. 10, 2017.

[17] L. Downey, "Efficient Market Hypothesis," 2021.

[18] X. Zhang, Y. Hu, K. Xie, S. Wang, E. W. Ngai, and M. Liu, "A causal feature selection algorithm for stock prediction modeling," *Neurocomputing*, vol. 142, pp. 48–59, 2014.

[19] C. Chen, W. Dongxing, H. Chunyan, and Y. Xiaojie, "Exploiting social media for stock market prediction with factorization machine," *Proceedings - 2014 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IAT 2014*, vol. 2, pp. 142–149, 2014.

[20] L. A. Teixeira and A. L. I. De Oliveira, "A method for automatic stock trading combining technical analysis and nearest neighbor classification," *Expert Systems with Applications*, vol. 37, no. 10, pp. 6885–6890, 2010.

[21] A. A. Adebiyi, A. O. Adewumi, and C. K. Ayo, "Comparison of ARIMA and artificial neural networks models for stock price prediction," *Journal of Applied Mathematics*, vol. 2014, pp. 9–11, 2014.

[22] S. Jansen, *Hands-On Machine Learning for Algorithmic Trading*. Birmingham: Packt Publishing, first ed., 2018.

[23] Ashwin Raj, "Unlocking the True Power of Support Vector Regression," 10 2020.

[24] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. Sebastopol, California: O'Reilly Media, Inc., second edi ed., 2019.

[25] S. Trenn, "Multilayer perceptrons: Approximation order and necessary number of hidden units," *IEEE Transactions on Neural Networks*, vol. 19, no. 5, pp. 836–844, 2008.

[26] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature," *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, 2014.

[27] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, "Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques," *Expert Systems with Applications*, vol. 42, no. 1, pp. 259–268, 2015.

[28] E. Beyaz, F. Tekiner, X. J. Zeng, and J. Keane, "Comparing Technical and Fundamental Indicators in Stock Price Forecasting," in *Proceedings - 20th International Conference on High Performance Computing and Communications, 16th International Conference on Smart City and 4th International Conference on Data Science and Systems, HPCC/SmartCity/DSS 2018*, pp. 1607–1613, Institute of Electrical and Electronics Engineers Inc., 1 2019.

[29] Y. Huang, "Machine Learning for Stock Prediction Based on Fundamental Analysis," *Electronic Thesis and Dissertation Repository*, 4 2019.

[30] A. Christoforous, "The GameStop short squeeze was 'the grand awakening': expert," 2021.