UNIVERSIDAD DISTRITAL
"FRANCISCO JOSE DE CALDAS"

TRABAJO FINAL
ESPECIALIZACION EN PROYECTOS INFORMATICOS

MOC-TEML:
CASO DE ESTUDIO: PREDICCIÓN DE TENDENCIA EN LOS ÍNDICES BURSÁTILES DE LA BOLSA DE VALORES DE COLOMBIA

Autores
Diego Esteban Moreno Morales
Maira Alejandra Quintana Duarte

Director
Sandro Bolaños

Bogotá 2020

RESUMEN

Este proyecto investiga técnicas de aprendizaje supervisado de una rama de la inteligencia artificial (IA) conocida como Machine Learning, para poner en práctica la aplicación de dos algoritmos, Maquinas de Vectores de Soporte (SVM) y redes neuronales artificiales (ANN), que permitan generar un comparativo de resultados de precisión en los pronósticos de acierto en la trayectoria que siguen los precios de los índices bursátiles de la Bolsa de Valores de Colombia (BVC), lo que comúnmente se conoce como tendencia.
Como resultado se entrega un documento con los resultados del análisis comparativo de los modelos aplicados, donde es posible identificar los niveles de desviación y proximidad en la certeza de los resultados de cada modelo. También se mencionan las ventajas representativas de cada modelo vs los demás aplicados. De manera que sea posible orientar a quien interese, sobre cuál de los dos algoritmos presenta el mayor grado de certidumbre.

PALABRAS CLAVE: *Machine Learning, Modelo supervisado, Maquinas de soporte vectorial, Redes Neurales Artificiales, Índice bursátil, Predicción.*

ABSTRACT

This project investigates supervised learning techniques from a branch of artificial intelligence (AI) known as Machine Learning, to put into practice the application of two algorithms, Support Vector Machines (SVM) and artificial neural networks (ANN), which allow generating a comparison of precision results in the predictions of success on the path followed by the prices of the stock market indices of the Colombian Stock Exchange (BVC), what is commonly known as a trend.

As a result, a document is delivered with the results of the comparative analysis of the models applied, where it is possible to identify the levels of deviation and proximity in the certainty of the results of each model. The representative advantages of each model vs. the others applied are also mentioned. So that it is possible to guide whoever is interested, on which of the two algorithms has the highest degree of certainty.

KEYWORDS: *Machine Learning, Supervised Model, Vector Support Machines, Artificial Neural Networks, Stock Market Index, Prediction*

Tabla de contenido

Tabla de Figuras:

Índice de Tablas:

INTRODUCCIÓN

Los cambios tecnológicos y la innovación sustentada en la ciencia computacional, ha dado lugar a la inteligencia artificial (IA), que propone una serie de algoritmos con el fin de crear máquinas con capacidad de interpretar, aprender, adaptarse y tomar decisiones. Estos algoritmos se pueden soportar en datos externos que generen conocimiento o por imitación del comportamiento humano [1] [2].

Para efectos de esta investigación, se denota sobre modelos supervisados, puntualmente Maquinas de Vectores de Soporte (SVM) y redes neuronales artificiales (ANN); Para los que se aplica un modelo comparativo con un caso de estudio que permite evaluar los resultados de la predicción de la tendencia en los índices bursátiles de la Bolsa de Valores de Colombia BVC, se toma como ejemplo por el tipo de datos que se almacena de forma histórica en la página oficial de la bolsa.

Esto permitirá evidenciar las ventajas relevantes y el poder de predicción de cada modelo, medibles en el nivel de certeza con respecto al comportamiento de la tendencia en el mundo real. El resultado de la práctica y comparación de los modelos ofrece a los interesados, orientación en términos de aplicación y viabilidad de cada modelo permitiendo orientar la toma de decisiones con respecto al uso de un algoritmo de vectores de soporte o redes neuronales artificiales.

PARTE I FUNDAMENTACIÓN DE LA INVESTIGACIÓN

El objeto de investigación se ubica en la tecnología de Machine Learning y corresponde a un modelo de comparación de técnicas de Machine Learning (MOC-TEML) - caso de estudio: predicción de tendencia en los índices bursátiles de la Bolsa de Valores de Colombia.

1. DESCRIPCIÓN DE LA INVESTIGACIÓN

El estudio del problema a desarrollar consta de tres pasos que se detallan a continuación:

1.1 Planteamiento del problema

Las técnicas de Machine Learning son variadas y la aplicación de las mismas también apunta a infinidad de requerimientos de distintas industrias, por ello, en la revolución tecnológica, han nacido nuevas profesiones y una de las más deseadas a nivel de las organizaciones, es la de científico de datos, profesionales con formación analítica, estadística y lógica de programación, que sean capaces de crear los sistemas de aprendizaje automatizado, que resuelvan problemas de predicción con mayor efectividad y eficiencia.

En la actualidad la demanda de análisis de datos es muy alta y los profesionales requieren buscar alternativas que suplan sus necesidades, sin embargo, en su búsqueda encuentran múltiples ofertas de algoritmos aplicables y es difícil elegir la opción más apropiada para la investigación deseada.

Se hace visible la necesidad de orientar el trabajo de estos nuevos profesionales, a partir de ejemplos prácticos y aplicados a casos reales, tomando como ejemplo la predicción de la tendencia de los índices bursátiles de la Bolsa de Valores de Colombia, donde se validará el comportamiento de los modelos supervisados, Maquinas de Vectores de Soporte (SVM) y redes neuronales artificiales (ANN), mostrando los resultados comparativos que permiten medir el grado de certeza de cada modelo.

1.1.1 Pregunta de investigación

De acuerdo con el planteamiento, se puede establecer la siguiente pregunta de investigación: ¿Cómo comparar el grado de certeza en la predicción de los modelos de aprendizaje supervisados SVM y ANN, para elegir el algoritmo con menos probabilidad de desviación, teniendo en cuenta que el conjunto de datos objeto de estudio sea útil para aplicar regresión logística?

1.1.2 Sistematización del problema

- ¿Tiene sentido comparar diferentes modelos supervisados?
- ¿Es posible predecir la tendencia en índices bursátiles con series de tiempo?

- ¿En qué se beneficiarán los científicos de datos, que apenas inician su desarrollo profesional, con el modelo de comparación de nivel de predicción en algoritmos supervisados SVM y ANN?

- ¿Qué impacto tendrá este modelo comparativo en la aplicación de técnicas para predecir comportamientos con datos de otras industrias?

1.2 Objetivos

Los objetivos por cumplir en este proyecto se presentan a continuación:

1.2.1 Objetivo general

Diseñar un modelo de comparación de técnicas de Machine Learning, aplicado a la predicción de la tendencia de los índices bursátiles de la Bolsa de Valores de Colombia (BVC), con el propósito de apoyar la elección del algoritmo con mayor grado de certeza y mejores ventajas comparativas

1.2.2 Objetivos específicos

- Recopilar información de algoritmos de Machine Learning que resuelvan problemas de regresión lineal basados en una serie temporal.

- Desarrollar dos modelos predictivos, usando los algoritmos SVM y ANN para la predicción de índices bursátiles de la BVC.

- Realizar un análisis comparativo de las ventajas y desventajas de los modelos SVM y ANN, a partir de su comportamiento en la predicción de índices bursátiles.

1.3 Justificación

Hoy en día la información es comprendida como un activo estratégico, además actualmente los volúmenes de información crecen exponencialmente; por esta razón es relevante poder tomar decisiones a partir del análisis avanzado de datos. Es en este punto donde los científicos y analistas de datos deben encontrar lineamientos que les permitan obtener los mejores resultados cuando desarrollan soluciones de inteligencia de negocios.
El análisis de datos avanzado, mediante la implementación de modelos predictivos, se enfrenta a una gran variedad de algoritmos de Machine Learning que, dependiendo del caso de implementación, pueden obtener distintos resultados en cuanto al grado de certeza de dichas predicciones. Sin embargo, frente a la implementación de varios algoritmos para resolver un mismo problema, existe un marco teórico limitado en cuanto a la definición de un modelo de validación que permita escoger cuál de estos está presentando un mejor desempeño.
Esta es una justificación de tipo práctico, debido a que propone el diseño de un modelo estructurado de validación de los resultados obtenidos por diversos algoritmos de Machine Learning, el cual será aplicado al problema de predicción de la tendencia de los índices bursátiles de la Bolsa de Valores de Colombia, comprendido como un pronóstico de series de tiempo. En conclusión, se demostrará cómo el modelo propuesto ayuda a tomar la decisión de elección del algoritmo de aprendizaje automático que da una solución más acertada al caso expuesto.

1.4 Hipótesis

La aplicación de técnicas de Machine Learning se ha vuelto una necesidad para resolver problemas de predicción, por ello se han definido una serie de algoritmos que permiten generar información de valor para apoyar la toma de decisiones de forma eficiente, rápida, automática y escalable en todos los campos de acción.

Teniendo en cuenta el enunciado anterior y el planteamiento del problema, se define la siguiente Hipótesis:

“Un modelo comparativo de técnicas de Machine Learning, facilita la elección de un modelo supervisado, basado en ventajas comparativas y resultados con mayor precisión en la predicción, minimizando el tiempo aplicado por científicos o analistas de datos en ejercicios de ensayo y error.”

1.5 Metodología a aplicar

A continuación, se describen los aspectos metodológicos que se utilizan en el desarrollo del proyecto.

1.5.1 Tipo de investigación

Para el proyecto que se va a desarrollar, se identifica un tipo de investigación de carácter "**Descriptiva**", porque puede servir de base para investigaciones que requieran un mayor nivel de profundidad y además responder una necesidad de los interesados en ciencia de datos. Esta será complementada con una investigación de tipo "**Aplicada**", haciendo uso de los conocimientos en técnicas de Machine Learning y utilizando las lecciones aprendidas durante las fases de la investigación.

1.5.2 Metodología de la investigación

Se toma como referencia una metodología de trabajo ágil, teniendo en cuenta entregar el mínimo producto viable en el menor tiempo posible, por ello se aplicara el marco de trabajo Scrum, que dicta un trabajo colaborativo, donde no existe la jerarquía tradicional de jefes y se tiene mayor confianza en el conocimiento y capacidades del grupo, además se define una lista de tareas alcanzable y de valor en un tiempo definido (sprint) a la cual es posible hacer un seguimiento diario que ayudara a gestionar impedimentos en caso de que existan.

En el marco de trabajo de scrum se aplica un método **inductivo** porque se desagregan las tareas para poder entregar valor en cada sprint y en un determinado punto llegar al producto final, además se complementa con el método **empírico** porque crece el conocimiento del producto a través de la experiencia.

1.5.3 Técnicas de recolección de información

Las técnicas de levantamiento de información que serán utilizadas en la ejecución del proyecto serán las siguientes:

- Consulta de documentación: Se investigarán casos de uso similares aplicando técnicas de Machine Learning para identificar estrategias de comparación útiles o apoyar el conocimiento que permita desarrollar satisfactoriamente la prueba de comparación.
- Extracción de datos históricos: Se descargará de la página oficial de la Bolsa de Valores de Colombia el comportamiento histórico de los índices bursátiles: COLCAP, COLEQTY, COLIR, COLSC, desde 2010 a corte de Febrero, porque son la base para el entrenamiento de los modelos.
- Observación: Se realizará observación sobre el comportamiento de los índices bursátiles para identificar factores anormales en la historia del índice y observación sobre los resultados de la predicción de cada modelo aplicado.

1.5.4 Tratamiento de la información

La aplicación de algoritmos de Machine Learning requieren el uso de información histórica para la etapa de aprendizaje o también llamado entrenamiento, por ende, hay dos formas de tratamiento de datos que se deben implementar en el proyecto:

- **Manual**: Para acceder a la página oficial de la Bolsa de Valores de Colombia y descargar uno a uno el histórico de los índices bursátiles de estudio esto en modo consulta.
- **Automática**: Para ordenar, clasificar, agrupar, comparar, analizar y demás operaciones que ejecutaran los modelos de Machine Learning de forma automática una vez estén programados en lenguaje Python.

1.6 Organización del trabajo

El trabajo de grado se encuentra dividido en 3 partes, así:

- Parte I fundamentación de la investigación.

  Se plantea la sistematización del problema, se definen los objetivos y se explica la metodología de investigación necesaria para implementar la solución.

- Parte II desarrollo de la investigación.

  Se aborda el desarrollo de la solución a partir del modelo propuesto para comparación de algoritmos de Machine Learning, se aplica el modelo para el caso de uso de los índices de la Bolsa de Valores de Colombia, se detalla paso a paso el desarrollo, se evalúan métricas de error y predicción para comparar comportamientos de los algoritmos y al final se generan los análisis sobre los resultados.

- Parte III cierre de la investigación.

  En esta fase, se generan las conclusiones y se plantean líneas de investigación futuras que se puedan derivar de la investigación.

PARTE II DESARROLLO DE LA INVESTIGACIÓN

2. DESCRIPCIÓN DE LA SOLUCION

2.1 Marco referencial

Se describe el siguiente marco teórico y marco conceptual, para soportar el caso de estudio aplicado en esta investigación:

2.1.1 Marco Teorico

A continuación, se describe un conjunto de conocimientos que permiten orientar esta investigación, y ofrecen una conceptualización adecuada de los términos que se emplearán para abordar el problema planteado.

2.1.1.1 Aprendizaje automático (Machine Learning)

El Machine Learning, también conocido como aprendizaje automático o de máquinas, hace parte de las ciencias de la computación y en muchos casos es definido como parte de la inteligencia artificial, dado que su función es definir, desarrollar e implementar técnicas y métodos que originen un proceso de aprendizaje en las máquinas [3] [4]. En otras palabras el objetivo del Machine Learning (ML) es programar computadores para que usen un conjunto de datos o eventos del pasado que les permitan resolver un problema dado [5] [6].

2.1.1.2 Algoritmos de aprendizaje supervisado

Las técnicas de Machine Learning comprenden un gran número de algoritmos que, basados en diversos modelos matemáticos, permiten encontrar patrones en los datos; dichos algoritmos pueden ser clasificados en tres grupos principales a partir de la forma en que estos aprenden: aprendizaje supervisado, no supervisado y por refuerzo[7].

El aprendizaje supervisado se caracteriza por el uso de conjuntos de datos donde se representa un resultado a partir de una o más variables de entrada [8]. Dichos algoritmos, además, pueden ser clasificados en dos grupos, de regresión y de clasificación, los cuales dependen del tipo de problema que se busca solucionar. Sin embargo, ambos tipos de algoritmos buscan encontrar el valor de una variable a partir de un conjunto de atributos; la diferencia radica en el hecho de que la variable dependiente, la cual es el resultado de la predicción, puede tomar valores categóricos (clasificación) o numéricos (regresión) [9] [10].

La elección de los algoritmos a implementar y evaluar se realiza basado en el tipo de problema a resolver; como se ha mencionado anteriormente, para el caso de estudio planteado se refiere a un problema de aprendizaje supervisado basado en regresión. En este punto se resalta la importancia de realizar una revisión sistemática del estado del arte que permite identificar qué algoritmos pueden dar mejores predicciones a partir de los datos recolectados.

Sin embargo, no se pretende descartar la posibilidad de elección de los algoritmos a partir del conocimiento empírico y científico del analista de información o científico de datos.

2.1.1.3 Máquinas de vectores de soporte (Support Vector Machine - SVM)

En Machine Learning las máquinas de vectores de soporte, mejor conocidas como SVM por sus siglas en inglés, son un algoritmo de aprendizaje supervisado que puede ser implementado para la solución de problemas de regresión o clasificación [11]. Este algoritmo de entrenamiento es conocido por su capacidad para generar un vector de soporte, el cual se define a partir de la distancia existente entre un conjunto de puntos de un espacio determinado, dicho vector divide las clases en dos espacios mediante un hiperplano de separación [12].

Una de las principales bondades del SVM es que, a diferencia de otros algoritmos de ML, este trabaja bajo el supuesto de definición de una función de pérdida que ignora el error que se sitúa a una cierta distancia del valor real [13].

2.1.1.4 Redes neuronales artificiales ANN

Un algoritmo predictivo de Machine Learning que ha tomado gran relevancia en la actualidad es el de Redes Neuronales Artificial (RNA), este es un modelo de aprendizaje automático que está inspirado en las neuronas del sistema nervioso y el funcionamiento del cerebro para establecer modelos no lineales diseñados específicamente para resolver diversos problemas [14]. Este se compone por un sistema enlazado de neuronas que trabajan en conjunto para producir un estímulo, estos procesos son conocidos por su capacidad de procesar información de forma eficiente a altas velocidades [15].

Las RNA son utilizadas principalmente para predecir resultados a partir de la combinación de un conjunto de parámetros, que se compone por una gran variedad de elementos simples altamente conectados que procesan grandes cantidades de información a grandes velocidades a través de un estado dinámico [16] que responde a entradas externas. Estas redes están organizadas jerárquicamente, interconectadas masivamente en paralelo e interactuando con objetos del mundo real tal cual lo hace el sistema nervioso [17].

Las redes neuronales se constituyen por neuronas conectadas que se agrupan en tres capas principales. La capa por medio de la cual ingresan los datos es conocida como capa de entrada, luego estos pasan por medio de una capa conocida como la capa oculta y salen del sistema por la capa de salida. Sin embargo, la capa oculta muchas veces se constituye por un número mayor de capas. Las funciones de una red neuronal describe la forma en que la red va a recibir, procesar y presentar la información [18].

2.1.1.5 Predicciones de series de tiempo

Las series de tiempo son una forma estructurada para la representación y el análisis de datos, la cual muestra la evolución de una variable en intervalos regulares de tiempo. Por otra parte, se puede considerar como un tipo de análisis que, basado en métodos y técnicas estadísticas, permite identificar tendencias en datos que varían en función del tiempo [19].

2.1.1.6 Métodos de comparación de predicciones: mediciones del error y desempeño

A continuación, se hablará sobre de tipos de error, sobreajuste y subajuste:

2.1.1.6.1 Tipos de errores

Cuando se habla de pronósticos es indispensable cuantificar la precisión de las proyecciones calculadas; para esto se pueden implementar medidas que calculan la desviación existente entre un conjunto de valores proyectados y el valor real. La importancia de esta medición se encuentra la posible elección o validación de un modelo de pronóstico.

2.1.1.6.2 Sobreajuste y subajuste

Con el objetivo de obtener el mejor pronóstico posible, se debe medir el rendimiento de un algoritmo de Machine Learning en cuanto a la validez de su predicción, por eso es importante verificar que los resultados obtenidos a partir del entrenamiento sean objetivos y consistentes. Sin embargo, al tratar de generalizar el aprendizaje de un modelo de ML, puede presentarse un fallo dada la relación encontrada entre los datos de entrada y salida implementados para el entrenamiento.

Por consiguiente, resultado arrojado por los algoritmos de ML puede verse afectado por el proceso de aprendizaje, que en consecuencia se ve reflejado en un sesgo o desviación de los datos de salida. Con base en lo anterior, se pueden presentar dos escenarios posibles; en el primero, conocido como Overfitting (sobreajuste) [20], el algoritmo aprende a solucionar solamente el caso particular con el que fue entrenado y por ende no es capaz de obtener buenas predicciones a partir de datos de entrada diferentes [21]; en el segundo se obtiene el resultado contrario, conocido como Underfitting (subajuste), donde el algoritmo no logra generalizar el aprendizaje, por lo cual no tiene buenas predicciones a partir del entrenamiento [22].

2.1.1.7 Índices bursátiles en la bolsa de valores de Colombia

La predicción del precio de las acciones es una de las tareas más importantes y desafiantes debido a su naturaleza altamente dinámica, puesto que se ve afectada por factores económicos, políticos, condiciones del mercado global, tendencias del mercado, tasas de referencia, que producen movimientos rápidos en los precios. Un método de predicción de índices, se puede apoyar en técnicas de Machine Learning, donde se destacan algoritmos de redes neuronales artificiales (ANN) [23] [24] y las máquinas de soporte vectorial [25].

Información de interés sobre los índices evaluados:

- **COLCAP**

El COLCAP es un índice de capitalización que refleja las variaciones de los precios de las acciones más líquidas de la Bolsa de Valores de Colombia (BVC), donde la participación de cada acción en el índice está determinada por el correspondiente valor de la capitalización bursátil ajustada (flotante de la compañía multiplicado por el último precio). El valor inicial del índice es equivalente a 1.000 puntos y su primer cálculo se realizó el día 15 de enero de 2008 [29].

- **COLEQTY**

El COLEQTY es un índice general que está compuesto por las 40 acciones con mejor función de selección de la Bolsa de Valores de Colombia (BVC), donde la participación de cada acción dentro del índice es determinada por el flotante (capitalización ajustada) de cada una de las especies.

El valor del COLEQTY será igual a la sumatoria del precio de cada acción que conforma la canasta del índice por el peso que tiene dentro de la misma ajustado por un factor de enlace [29].

- **COLSC**

El COLSC es un índice que está compuesto por las 15 acciones de las empresas más pequeñas en capitalización bursátil que hacen parte del COLEQTY, donde la participación de cada acción dentro del índice es determinada por el flotante de cada una de las especies.

El valor del COLSC será igual a la sumatoria del precio de cada acción que conforma la canasta del índice por el peso que tiene dentro de la misma ajustado por un factor de enlace [29].

- **COLIR**

El COLIR es un índice que está compuesto por las acciones de las empresas que cuentan con el Reconocimiento Emisores 013 IR (o 01CReconocimiento IR01D) en los términos de la Circular Única BVC y que hacen parte del COLIR, donde la participación de cada acción dentro del índice es determinada por el flotante de cada una de las especies.

Fórmula: El valor del COLIR será igual a la sumatoria del precio de cada acción que conforma la canasta del índice por el peso que tiene dentro de la misma ajustado por un factor de enlace [29].

A continuación, se muestra la formula general para el cálculo del valor de índices de renta variable:

$$I(t) = E \sum_{i=1}^{n} Wi\ Pi(t) \qquad (1)$$

Donde:

I(t): Valor del índice en el instante t.

E: Factor de enlace mediante el cual se da continuidad al índice cuando se presente un rebalanceo o recomposición de la canasta o en caso de darse eventos corporativos que lleven a variaciones en el índice.

t: Instante en el cual se calcula el valor del índice.

i: 1, 2, ..., n acciones que componen el índice.

n: Número de acciones en el índice en el instante t.

Wi: Ponderador de la acción i en el instante t.

Pi: Precio de la acción i en el instante t.

2.1.2 Marco conceptual

Para apoyar el entendimiento de esta investigación, a continuación, se relacionan los conceptos más relevantes:

- **Índice Bursátil:** Las acciones son consideradas como parte fundamental del mercado de renta variable, ya que el valor de las acciones va cambiando con el tiempo como consecuencia de la oferta y demanda, así como por efecto de la volatilidad de los mercados [25].
- **Inteligencia Artificial:** La inteligencia artificial (IA) es conocida en ciencias de la computación como un término aplicado para definir procesos donde una máquina logra imitar funciones cognitivas, como el aprendizaje y la solución de problemas [26].
- **Predicción:** La predicción en el contexto científico de datos es una declaración precisa de lo que ocurrirá en un momento futuro teniendo en cuenta determinadas condiciones.
- **Lenguaje de programación Scripting:** Un lenguaje de secuencia de comandos o scripting, se refiere al tipo de programación que puede interpretar el código en tiempo de ejecución sin necesidad de ser compilado por el procesador de la máquina [27] esta es la gran diferencia con respecto a los lenguajes de programación que se conocen tradicionalmente, pero no implica que no sea un lenguaje de programación.
- **Python:** Es un lenguaje de programación de propósito general [27], el más completo de la industria del desarrollo y se ajusta perfectamente a las necesidades de programación de técnicas de Machine Learning, también es de los más comunes porque es un lenguaje interpretado, dinámico y multiplataforma [28].

2.2 Diseño de la solución

En la Figura 1, se presenta el diagrama de proceso que expresa las etapas para tratamiento y modelado de datos hasta obtener el resultado esperado aplicando los lineamientos para uso de algoritmos de analítica, se exalta que este diagrama resulta muy útil para que un científico o analista de datos establezca las pautas a aplicar en una solución de Machine Learning.

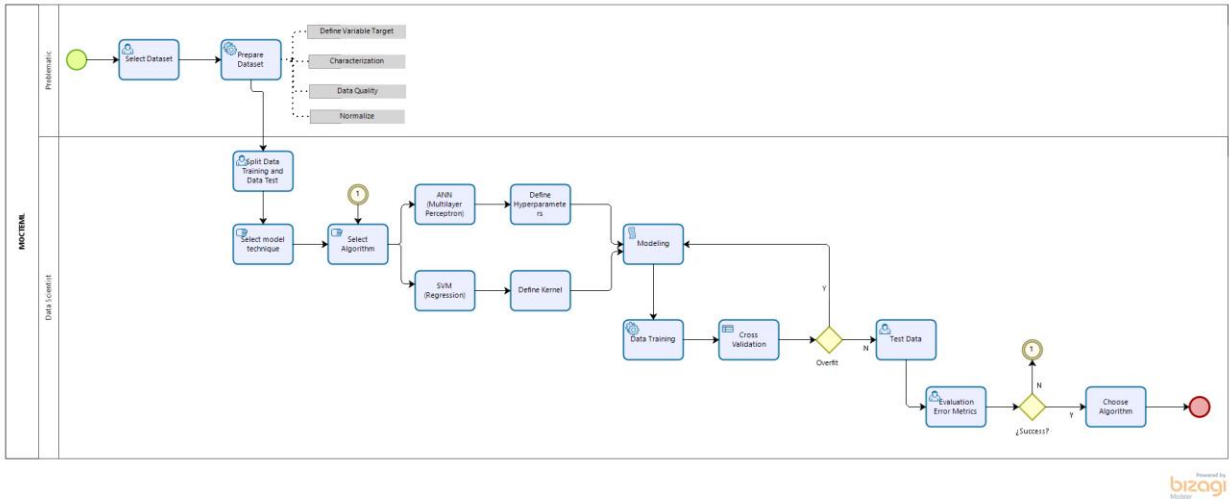

Figura 1 Proceso ML. Los autores.

A partir de este diagrama se establece el punto de partida para explicar cada una de las etapas del proceso:

2.2.1 Gestión de los datos

En esta fase es necesario aplicar una serie de tareas en relación a los datos que se explicarán a continuación:

2.2.1.1 Seleccionar los datos

La información sobre los índices bursátiles es de acceso público y está disponible en la página oficial de la Bolsa de Valores de Colombia; los índices son actualizados varias veces al día y se almacena su comportamiento histórico, lo que permite hacer trazabilidad en una línea de tiempo; la fiabilidad de estos datos esta soportada por la Superintendencia Financiera de Colombia.

En la Figura 2, se relacionan los atributos de los índices de renta variable, que serán la entrada de datos para el modelamiento.

Índices — Mercado Cerrado

Información en línea

Indices	Mercado	Valor Hoy	Valor Ayer	Variación Absoluta	Variación %	
COLCAP	RENTA VARIABLE	919.5900	899.7500	19.8400	2.2051	▲
COLSC	RENTA VARIABLE	815.1500	844.0400	-28.8900	-3.4228	▼
COLEQTY	RENTA VARIABLE	646.3800	636.5900	9.7900	1.5379	▲
COLIR	RENTA VARIABLE	604.0400	594.5800	9.4600	1.5910	▲

Figura 2 Índices BVC [29].

2.2.1.2 Preparar los datos

Se determinó descargar la información histórica de los últimos 10 años de cada índice para contar con suficiente información que permita segmentar los conjuntos de datos de entrenamiento y pruebas.

Una vez se ha descargado toda la información histórica, se consolida en una única fuente con los atributos, como se observa en la Figura 3.

Fecha	Indice	Valor Hoy	Valor Ayer	Variacion %	Variacion Absoluta	Variacion 12 me	Variacion
06/09/2019	COLCAP	1567,53	1567,35	0,01148435	0,18	5,546203776	18,2211731
06/09/2019	COLEQTY	1082	1081,51	0,04530702	0,49	7,699198726	18,5961374
06/09/2019	COLIR	1016,97	1016,51	0,04525287	0,46	7,943702037	18,6510483
06/09/2019	COLSC	929,02	926,03	0,3228837	2,99	16,57046778	27,6950779
06/09/2019	COLTES CP	251,5	251,05	0,18	0,45	7,350179273	5,27417329
06/09/2019	COLTES LP	319,62	318,96	0,21	0,66	12,74869479	9,64665523
06/09/2019	COLTES UVR	301,08	300,88	0,07	0,2	12,35166803	9,07905224
06/09/2019	COLTES	285,88	285,31	0,2	0,57	11,32398754	8,48512447
05/09/2019	COLCAP	1567,35	1565,89	0,09323771	1,46	3,870929262	18,2075977

Figura 3 Set de datos consolidado. Los autores.

Es importante aplicar las siguientes acciones previo al uso de la información:

- **Generales:**

    - Determinar el nivel de análisis de los datos (Hora, Dia, Mes), en este caso la naturaleza de los datos es diaria.

    - Identificar columnas confiables, importantes y fáciles de entender, en este caso, los atributos relevantes son Fecha, índice y valor hoy:

        - **Fecha**: Día de la operación en jornada diurna habitual y únicamente días hábiles.
        - **Índice**: Nombre del índice evaluado: COLCAP, COLEQTY COLIR COLSC
        - **Valor Hoy**: Valor del índice con el que cerró el día.

- **Definir la Variable Objetivo:**

    - Para aplicar un modelo de regresión la variable objetivo debe ser numérica. En este caso la variable objetivo es "***Valor Hoy***" este atributo permitirá hacer la predicción.

    - Analizar la variable de destino en el tiempo, como se observa en la Figura 4, el comportamiento de la variable en el rango de un mes, no presenta variaciones de alto impacto, se mantiene sobre los 1505 y 1559, lo que posibilita la aplicación de algoritmos de Machine Learning de regresión para elaborar predicciones.

Figura 4 Valor Hoy del Índice COLCAP – 2019-08. Los autores.

    - Detectar cualquier efecto de estacionalidad o comportamiento extraño en el tiempo de la variable objetivo. Se evidenció que no existe riesgo de sesgo por estacionalidad.

- **Evaluar la Calidad de los datos:**

  Identificar y tratar las inconsistencias en los datos (valores absurdos, fechas incorrectas, valores nulos) y excluirlos de la base de análisis si su valor no es representativo dentro del conjunto. En el set de datos evaluado desde Enero 2010 a Abril 2020 no hay valores de fecha inválidos, es de aclarar que solo están disponibles los días hábiles de cada año, que es cuando esta activa la operación de la BVC, los valores de los índices, también están de forma correcta.

- **Normalizar los datos:**

  Esta tarea consiste en comprimir o extender los valores de la variable para que estén en un rango definido, ejemplo, que todos los valores estén en una escala entre 0 y 1, para esto se aplicó el método de escalado de variables ***MinMax Scaler***, que consiste en normalizar entre los limites mínimo y máximo definidos por el mismo conjunto de datos, aplicando la siguiente formula [30] :

$$X' = \frac{x - min(x)}{max(X) - \min(X)} \qquad (2)$$

2.2.1.3 Crear conjuntos de Datos

Cuando se pretende predecir el futuro con información del pasado los datos no se deben arrastrar, ya que la secuencia de los datos es una característica esencial y el arrastre podría ocasionar una pérdida temporal de información.

Por ello es muy importante que en este ejercicio se segmenten los datos en tres grupos, datos de entrenamiento, datos de validación y datos de prueba; es difícil saber con exactitud en que porcentaje se deben segmentar los grupos puesto que es un factor codependiente del comportamiento de los datos, el tamaño de la población, la historia disponible, estacionalidades, entre otros, esta tarea queda expuesta a la experiencia del analista de datos.

Como apoyo para una evaluación inicial, se propone catalogar el set de datos original, en set de datos pequeños y set de datos grandes, para ambos escenarios puede aplicar una segmentación de datos de forma porcentual, Figura 5:

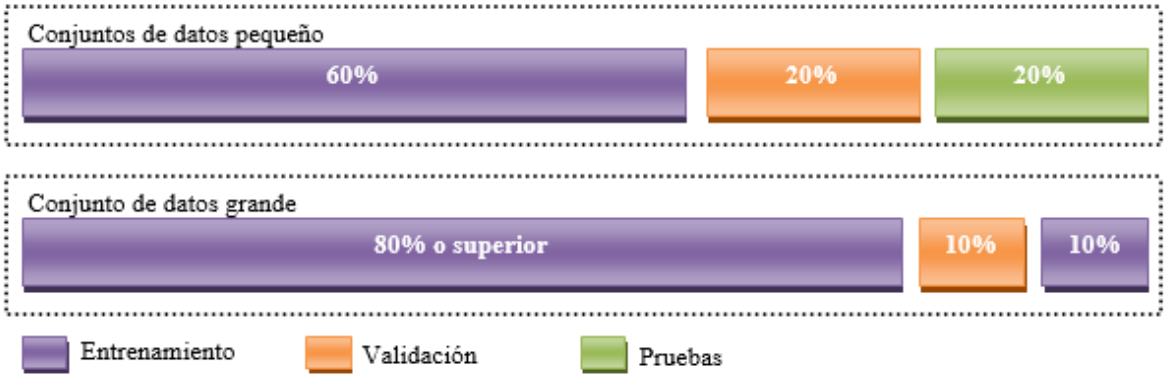

Figura 5 Método recomendado para dividir el conjunto de datos. Los autores.

Sobre el consolidado del conjunto de datos que se creó en la etapa de preparación de datos, se obtuvo un aproximado de 10.000 registros, esta no es una cifra muy representativa, por lo que se catalogó como un conjunto de datos pequeño.

2.2.1.3.1 Datos de Entrenamiento

Es el conjunto de datos más importante para el modelo, porque se utiliza para enseñar los comportamientos más repetitivos en los datos, de esta manera se genera valor a la predicción, el modelo ve y aprende de estos datos, durante el entrenamiento se reproducen las condiciones reales, para que el resultado del modelo sea cercano a la realidad; Con respecto a la historia disponible para los índices de renta variable, se tomara el 75% del conjunto total de datos para entrenamiento.

2.2.1.3.2 Datos de Validación

Con este conjunto se realiza una evaluación imparcial del modelo, mientras se ajustan características como hiperparámetros del modelo. La evaluación se vuelve más sesgada a medida que la habilidad en el conjunto de datos de validación se incorpora a la configuración del modelo, por lo tanto, el modelo ocasionalmente ve estos datos, pero nunca "Aprende" de este grupo.

2.2.1.3.3 Datos de Pruebas

Este último grupo de datos se utilizó para proporcionar una evaluación imparcial de un ajuste final del modelo en el conjunto de datos de entrenamiento, se asignara el restante de registros del conjunto total.

2.2.2 Modelamiento Analítico de los Datos

2.2.2.1 Técnica del modelo

Dado que los índices bursátiles tienen valores que fluctúan a lo largo del tiempo, se ha modelado su predicción como un pronóstico basado en series de tiempo, donde la variable que representa el valor del índice, es una variable dependiente de la forma y=f(x), explicada en función de la variable dependiente x que representa un instante de tiempo determinado. [33]

Con base en lo anterior se puede notar que el caso de estudio responde a un problema de regresión, teniendo en cuenta que lo que se busca es pronosticar el valor de los índices en instantes de tiempo futuros a partir de sus históricos. [34]

En la Figura 6, se puede analizar las series de tiempo, entendidas como la variación de los valores diarios de los índices estudiados a lo largo del tiempo, se observa que en varios puntos se presentan comportamientos tendenciales, estacionales y cíclicos, sin embargo, en otros puntos parecen presentar comportamientos de tipo irregulares: sin embargo, es concluyente, que para todos los casos de estudio las series de tiempo, no presentan linealidad.

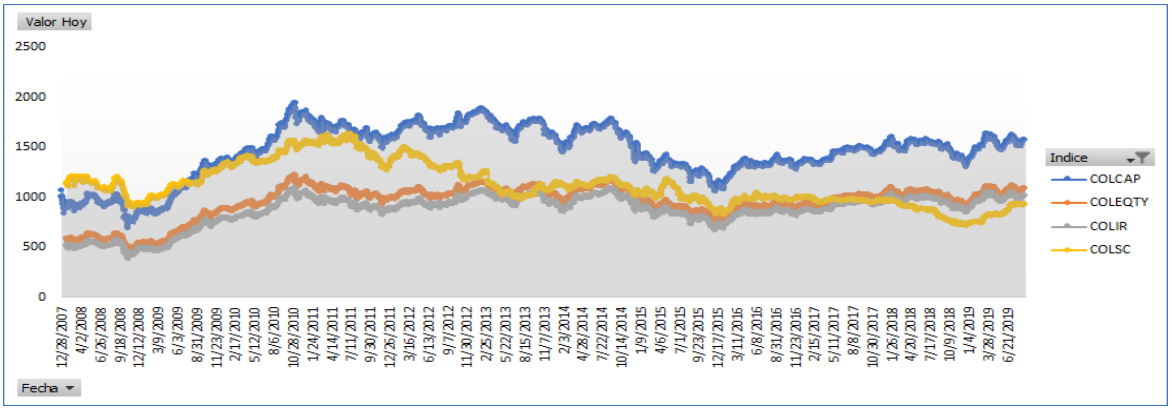

Figura 6 Representación gráfica de los índices como series temporales. Los autores.

2.2.2.2 Algoritmo para el modelo

En esta etapa, para el caso de estudio se han seleccionado dos algoritmos; el primero corresponde a las redes neuronales artificiales de avance (feedforward neural network) del tipo perceptrón multicapa [35]; el segundo corresponde a las máquinas de vectores de soporte (support vector machine) con enfoque de regresión (también conocido como SVR).

2.2.2.3 ANN (Multilayer Perceptron)

El perceptrón multicapa es un tipo de red neuronal Feedforward, en la Figura 7, se puede evidenciar su estructura, conta de tres tipos de capas, entrada, ocultas y de salida; cada una de estas capas están compuestas por un conjunto de nodos.

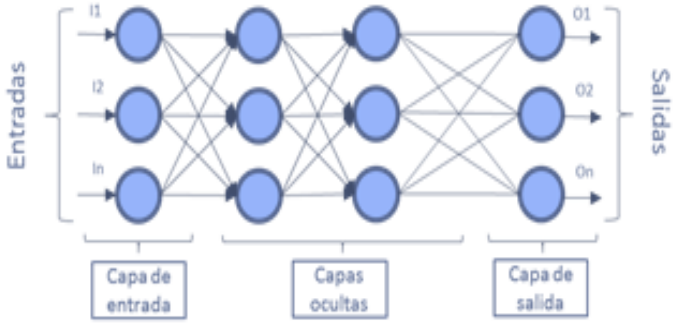

Figura 7 Ejemplo de estructura del perceptrón multicapa. Los autores.

Adicionalmente en este tipo de redes cada nodo de las capas ocultas usa una función de activación no lineal.

- Definición de hiperparámetros

Consiste en seleccionar los parámetros adecuados de forma heurística, evaluando diferentes escenarios de configuración [36], de esta manera se garantiza mayor confianza en el resultado.

Sin embargo y como se sugiere en una etapa previa, este modelo comparativo no descarta la realización de una revisión de literatura o el conocimiento empírico y científico del implementador, como criterios para la definición de dichos hiperparámetros.

La Tabla 1, muestra la parametrización del perceptrón multicapa escogida para el caso de estudio.

Hiperparámetro	Definición
Numero de capas ocultas	1 capa
Nodos de la capa oculta	Igual al número de nodos de la capa de entrada
Función de activación de la capa oculta	Tangente Hiperbólica

Tabla 1 Definición de hiperparámetros del perceptrón multicapa. Los autores.

Debido a que el modelo corresponde a la predicción de series de tiempo, es considerado un problema de predicción no tan complejo, se ha definido solo una capa oculta y el mismo número de nodos de la capa de entrada. Esto debido a que arquitecturas y configuraciones más complejas son usadas para resolver problemas avanzados de Deep Learning y reconocimiento de imagen; adicionalmente, este tipo de configuraciones pueden incurrir en un requerimiento de hardware mayor.

Por otra parte, la función de activación tangente hiperbólica fue determinada en primera instancia debido a que en diversas investigaciones ha sido usada para los problemas de regresión basados en series de tiempo.

2.2.2.4 SVM (Regression)

El algoritmo SVM también puede ser usado como un método de regresión conservando las características principales del algoritmo, esta aproximación también se conoce como regresión de vectores de soporte o SVR por sus siglas en inglés [37]. Para el caso de la regresión, se debe definir un margen de tolerancia (épsilon) el cual representa una aproximación al SVM. Vale la pena aclarar que este algoritmo en comparación al SVM puede requerir una mayor capacidad de computo. A pesar de sus diferencias, conceptualmente ambos algoritmos funcionan de forma similar, ya que buscan minimizar el error, individualizando el hiperplano que maximiza el margen, teniendo en cuenta que se tolera parte del error. [38]

- Definición función Kernel

Los algoritmos basados en vectores de soporte suelen usar un conjunto de funciones matemáticas que se definen como el Kernel, que consiste en tomar datos como entrada y transformarlos en salidas que satisfacen una necesidad. Dado que el objeto de estudio genera una solución a partir de una aproximación no lineal, se debe definir un Kernel de transformación, es posible implementar diferentes algoritmos basados en vectores de soporte parametrizando el uso de las diversas funciones Kernel existentes. Estas funciones pueden ser de diferentes tipos. Por ejemplo, lineal, no lineal, polinomial, función de base radial (RBF) y sigmoide. [39]

Es importante tener en cuenta que las funciones Kernel devuelven el producto interno entre dos puntos en un espacio de características adecuado. Por lo tanto, al definir una noción de similitud, incurre en un recurso computacional alto incluso en espacios de dimensiones relativamente pequeños. Nuevamente se recalca que la elección se este tipo de parametrizaciones como funciones Kernel, se puede realizar de forma heurística sin descartar la revisión de literatura y el conocimiento científico y empírico del implementador.

Como definición para el caso de estudio se selecciona como Kernel la función en base radial (RBF) [40] la cual se define de la siguiente forma:

Dadas dos muestras x y x'

$$K(x, x') = \exp(-\frac{||x - x'||^2}{2\sigma^2}) \qquad (3)$$

Donde

$\|\|x - x'\|\|^2$	Puede entenderse como la distancia euclidiana al cuadrado entre los vectores caracterizados.
σ	Es un parámetro libre.

Al igual que la función tangente hiperbólica en el modelo de redes neuronales, varias investigaciones apuntan a que una abstracción de dicha función puede ser similar a la forma en que trabajan las funciones de base radial. Además, esta función Kernel es demasiado efectiva cuando se trabaja con series de tiempo que no tienen un comportamiento lineal.

2.2.3 Evaluación

2.2.3.1 Validación Cruzada

La validación cruzada se utiliza principalmente en modelos de clasificación y regresión [41] para reducir el sesgo entre todo el conjunto de datos y conjunto de entrenamiento (o conjunto de prueba). La técnica de validación cruzada K-fold, divide el conjunto de datos sin procesar en K subconjuntos. Se elige uno de los subconjuntos como el conjunto de prueba, y los restantes subconjuntos de datos K - 1 se consideran como el conjunto de entrenamiento en cada iteración [42], generando la media aritmética de los resultados de cada iteración para obtener un único resultado.

En el desarrollo del caso de estudio, se aplicó esta técnica de validación creando segmentos sobre la base del conjunto de datos de entrenamiento, donde los datos fueron distribuidos proporcionalmente. Lo anterior permite que, a través de una serie de iteraciones, se incluyan todos los segmentos tanto en el grupo de datos de prueba como en el grupo de entrenamiento. De esta manera se logra mayor precisión en los resultados, sin perder de vista todo el comportamiento de los datos.

En la Figura 8, se puede apreciar el proceso de validación cruzada que se aplicó para el conjunto de datos del proyecto:

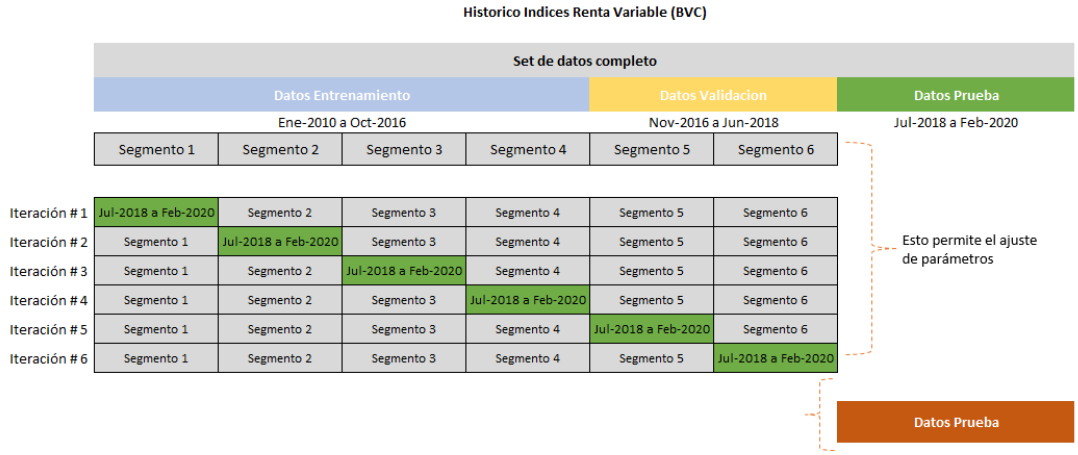

Figura 8 Validación cruzada del set de datos de índices. Los autores.

2.2.3.2 Testeo de los datos

El modelo definido propone evaluar y comparar la precisión de dichos pronósticos con datos de testeo, los cuales corresponden a observaciones reales del problema de regresión que no han sido utilizados ni para el entrenamiento ni para la validación de los algoritmos aplicados. Estos datos deben estar desnormalizados, a partir de esto tomando como input el conjunto de fechas seleccionadas en este conjunto de datos, se aplican las respectivas predicciones con el modelo entrenado las cuales, también desnormalizadas, se compararán con los valores reales del set de testeo.

2.2.3.3 Evaluación de métricas de error

Para validar el nivel de certeza de un pronóstico o predicción es necesario cuantificar la precisión del mismo. Existen diversas medidas que permiten evaluar que tan desviado se encuentra un valor pronosticado del valor real, a continuación, se describen las principales que serán tenidas en cuenta para la evaluación de los modelos aplicados al caso de estudio.

Teniendo en cuenta que el error corresponde a la cuantificación de la diferencia existente entre el valor pronosticado y el real ocurrido, es posible medir esta desviación promedio o total para todo el conjunto de datos de testeo con ciertas variaciones matemáticas. Por esto a continuación se definen algunos de los métodos más comunes para hacer esta medición los cuales serán aplicadas al case de estudio como métricas de medición del error, aclarando la siguiente notación:

Medida de error:

$$e = y - \hat{y} \qquad (4)$$

Donde

e Es el error calculado.

y Es el valor real.

$\hat{y}$ Es el valor pronosticado.

- Error al cuadrado medio: MSE o Error al cuadrado medio es el promedio de la diferencia al cuadrado entre el valor objetivo y el valor predicho por el modelo de regresión.

$$MSE = \frac{1}{n}\sum(y - \hat{y})^2 \qquad (5)$$

- Error cuadrático medio de raíz: RMSE es la raíz cuadrada de la diferencia cuadrática promedio entre el valor objetivo y el valor predicho por el modelo.

$$RMSE = \sqrt{\frac{\sum(y - \hat{y})^2}{n}} \qquad (6)$$

- Error absoluto medio: MAE es la diferencia absoluta entre el valor objetivo y el valor predicho por el modelo.

$$MAE = \frac{1}{n}\sum|y - \hat{y}| \qquad (7)$$

2.2.4 Comparación

Una vez cuantificada la precisión de cada uno de los modelos evaluados es preciso identificar cual modelo ha resultado más acertado en su pronóstico. Teniendo en cuenta las métricas de error existentes y evaluadas, se expresa la necesidad de dar un resultado final ya que posiblemente en algunos casos un modelo obtenga los mejores resultados al evaluarlo puntualmente con una métrica, pero al evaluarlo con otras métricas no genere la misma precisión.

En esta línea se propone escoger el algoritmo con mayor nivel de certeza a partir del cálculo de su desviación porcentual con base en el valor obtenido en cada una de las métricas de error, siendo el algoritmo más preciso aquel que logre obtener una desviación porcentual media menor.

La desviación porcentual media es calculada de la siguiente forma:

$$\overline{D_i} = \sum_{i\,\varepsilon J} {}^{D_{ij}}\!/_{n} \;\nabla\, \mathrm{i}\;\varepsilon\;\mathrm{I} \qquad (8)$$

$$D_{ij} = \frac{e_{ij} - e_{0j}}{e_{0j}} \times 100\ \% \qquad (9)$$

Donde

I	Corresponde al conjunto de algoritmos evaluados.
J	Corresponde al conjunto de métricas de error evaluadas.
n	Es el número de métricas de error evaluadas.
$\overline{D_i}$	Es la desviación porcentual media del algoritmo i.
D_{ij}	Es la desviación porcentual del algoritmo i en la métrica j.
e_{0j}	Es el menor error obtenido en la métrica j.
e_{ij}	Es el error obtenido por el algoritmo i en la métrica j.

La matriz de mediciones de error por algoritmo, que se usa para registrar los valores obtenidos en cada una de las métricas, es ilustrada en la Tabla 2, esto permitirá visualizar de forma tabular el desempeño de cada uno de los algoritmos para posteriormente calcular las denominadas desviaciones porcentuales.

Métricas de error J				
Algoritmo I	**Métrica 1**	**Métrica 2**	**…**	**Métrica n**
Algoritmo 1	e11	e12	…	e1n
Algoritmo 2	e21	e22	…	e2n
…	…	…	…	…
Algoritmo k	ek1	ek2		ekn

Tabla 2 Matriz de mediciones de error por algoritmo. Los autores.

La Tabla 3, ilustra el cálculo de las diferencias porcentuales de cada valor obtenido versus el mínimo de la métrica de error analizada. Esta matriz de desviaciones porcentuales es la base para calcular la desviación porcentual media de cada algoritmo.

Métricas de error J				
Algoritmo I	**Métrica 1**	**Métrica 2**	**…**	**Métrica n**
Algoritmo 1	(e01-e11)/e01	(e02-e12)/e02	…	(e0k-e1n)/e0k
Algoritmo 2	(e01-e21)/e01	(e02-e22)/e02	…	(e0k-e2n)/e0k
…	…	…	…	…
Algoritmo k	(e01-ek1)/e01	(e02-ek2)/e02		(e0k-ekn)/e0k

Tabla 3 Matriz de desviaciones porcentuales por error para cada algoritmo. Los autores.

2.3 Desarrollo de la solución

Teniendo en cuenta que el modelo comparativo propone un marco de trabajo, para la implementación de proyectos de analítica predictiva de datos basados en la implementación de algoritmos Machine Learning supervisados de regresión. La implementación del MOC-TEML permite estandarizar el desarrollo y la implementación de este tipo de soluciones cuyo valor agregado es la posibilidad de identificar el algoritmo más preciso en la predicción de un problema.

En consecuencia y como se ha especificado anteriormente, el propósito del caso de estudio ha sido implementar el modelo MOC-TEML a la predicción del valor de los índices bursátiles a partir del modelado de estos como series temporales.

Con base en lo anterior, el desarrollo de la solución se ha enfocado en aplicar a la codificación del software el marco de trabajo propuesto en el modelo. Dicha aplicación se ha implementado en el lenguaje de programación Python, debido a que en la actualidad es uno de los lenguajes más usados para la implementación de soluciones de analítica y ciencia de datos.

Adicionalmente por facilidad y disponibilidad de acceso a las herramientas de desarrollo, se utilizó la plataforma Google Colaboratory, que es un entorno de máquinas virtuales en la nube y además de permitir el desarrollo del código, también permitió crear el documento técnico de la solución de forma clara y ordenada, esto se conoce como Jupyter Notebook.

El despliegue del desarrollo se realizó de forma independiente para cada índice, por ende, se crearon cuatro Notebooks aplicando la solución y se anexan al final del documento. La Tabla 4, describe la relación de anexos con los índices.

Número	Índice Bursátil	Anexo - Software
1	COLCAP	Anexo 1 COLCAP MOC-TEML ANN-SVR.ipynb
2	COLIR	Anexo 2 COLIR MOC-TEML ANN-SVR.ipynb
3	COLSC	Anexo 3 COLSC MOC-TEML ANN-SVR.ipynb
4	COLEQTY	Anexo 4 COLEQTY MOC-TEML ANN-SVR.ipynb

Tabla 4 Soluciones desarrolladas. Los autores.

Vale la pena poner en contexto que, adicional a las librerías de Python para tratamiento y gráficas de datos, en el desarrollo se usó para el entrenamiento del SVR la librería Scikit-learn también conocida como SKLearn y para el caso de la red neuronal la librería Keras.

PARTE III CIERRE DE LA INVESTIGACIÓN

3. RESULTADOS Y DISCUSIÓN

3.1 Resultados

En primera instancia, el modelo propuesto permitió estandarizar el proceso de comparación de los dos algoritmos de Machine Learning, lo cual facilita el desarrollo del código y el despliegue de la solución. En consecuencia, la implementación del MOC-TEML en el caso de estudio, estableció un esquema de trabajo para hacer que los resultados de los algoritmos fueran realmente comparables y definió la implementación de cada una de las etapas necesarias en el desarrollo de la solución de un problema de analítica de datos, se aplican dos modelos de Machine Learning basados en aprendizaje supervisado del tipo regresión, los cuales son SVR y ANN-MLP.

En la Figura 9, se visualiza gráficamente los resultados independientes de la serie correspondiente al índice COLCAP, el color negro representa el valor real de los datos de testeo, el color rojo representa los resultados de predicción de la red neuronal (ANN-MLP) y el color azul representa los resultados de predicción de las máquinas de vectores de soporte (SVR).

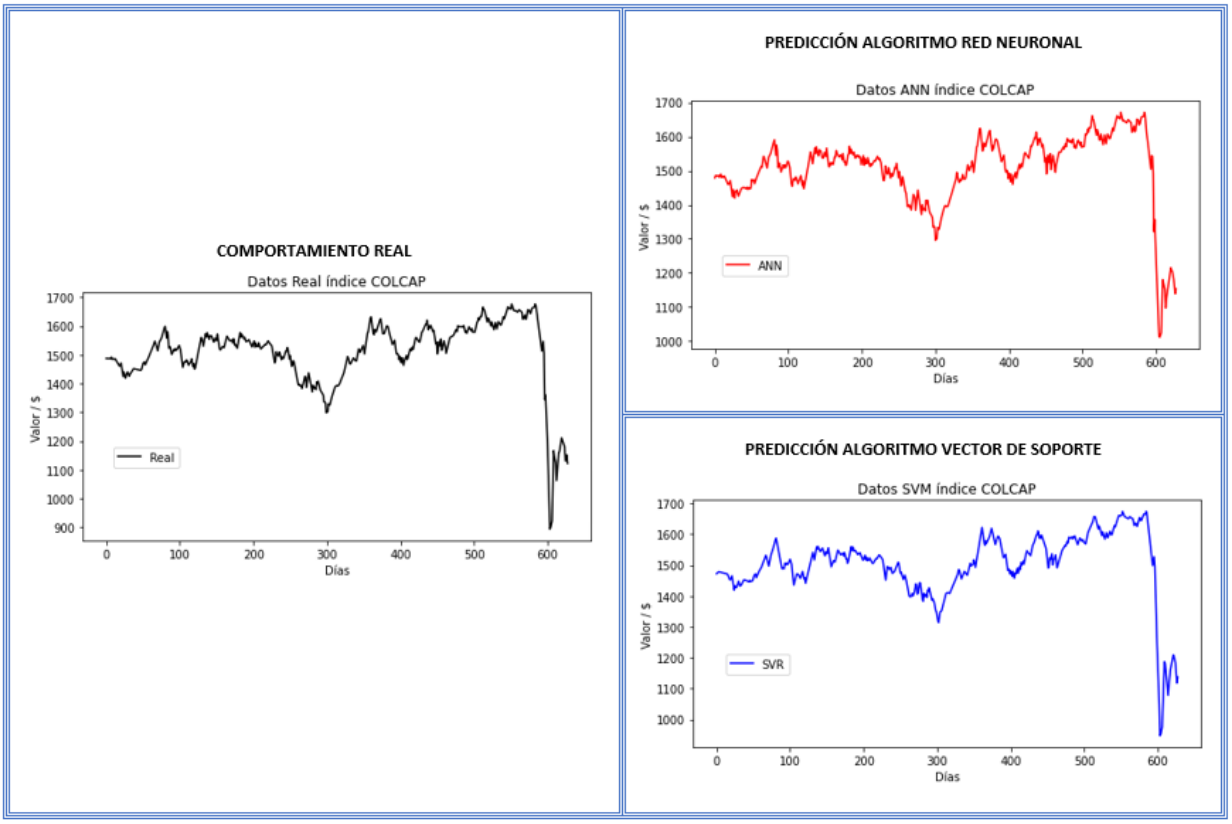

Figura 9 Resultados independientes índice COLCAP. Los autores.

Ahora de forma comparativa, la Figura 10 ilustra el solapamiento de las tres series para el índice COLCAP, el valor real, la predicción de la red neuronal y la predicción del vector de soporte.

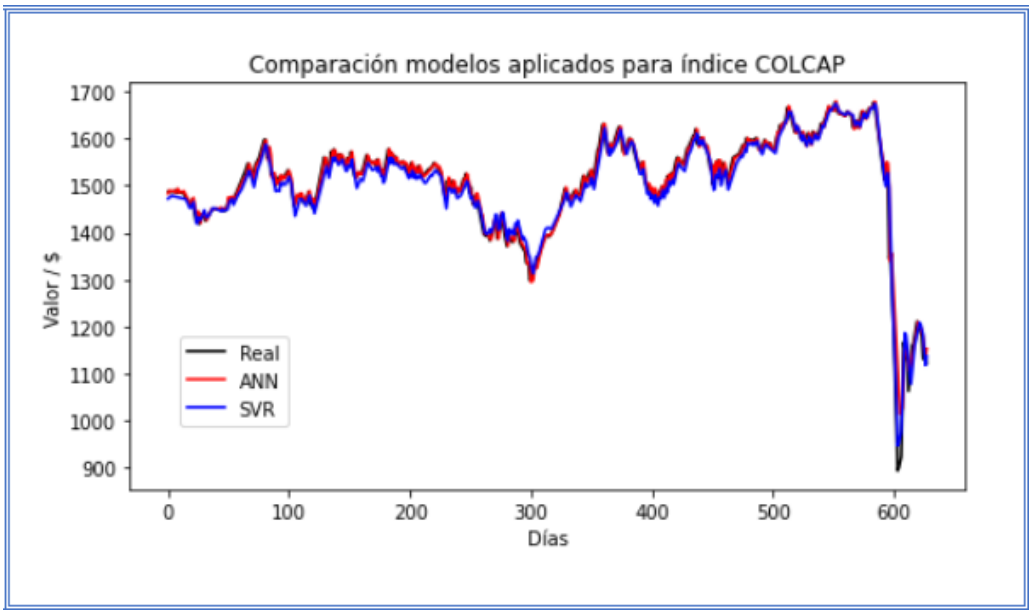

Figura 10 Resultado comparación índice COLCAP. Los autores.

Evidentemente, los resultados demuestran un pronóstico bastante acertado, debido a que el patrón de las tres series es muy similar en cuanto a forma y valores.

La implementación del modelo fue desplegada para cada uno de los 4 índices bursátiles estudiados y el comportamiento lógico de predicción es similar, por ello en la Figura 11 se evidencian los resultados comparativos de todos los índices.

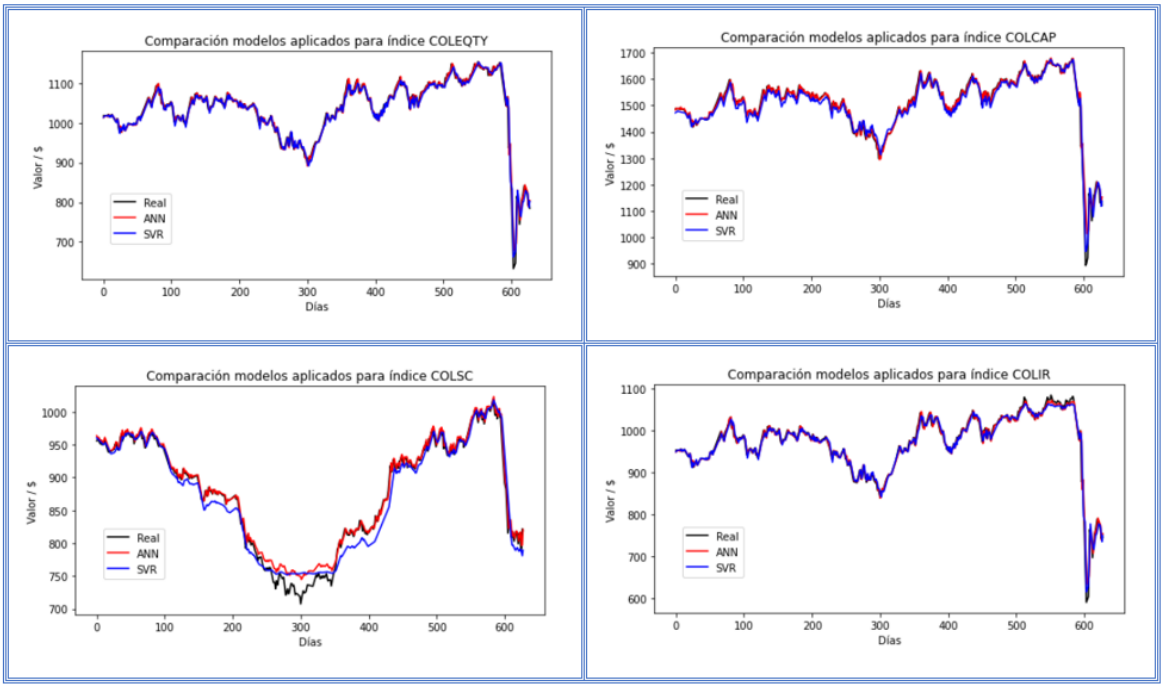

Figura 11 Resultados comparación modelos por índice. Los autores.

El análisis gráfico permite establecer visualmente la precisión de los pronósticos y da un indicio de si estos están siendo realmente acertados. Para el caso de estudio se puede evidenciar que los pronósticos de ambos algoritmos están siendo efectivos en su predicción para cada uno de los cuatro índices, lo cual añade mayor valor al modelo propuesto, ya que este permite cuantificar dicho grado de certeza.

Si bien, para los índices COLCAP, COLIR y COLEQTY los resultados son notablemente acertados, para el COLSC fueron substancialmente menos precisos, esto debido a que, en un periodo de tiempo, aunque la predicción de la serie en su forma, es similar, esta se encuentra desfasada en cuanto al valor real.

3.2 Análisis de resultados

3.2.1 Medidas de error

Una vez aplicado el modelo MOC-TEML en su fase de evaluación de métricas de error, se presenta como resultado para cada índice, el valor obtenido por cada algoritmo en las mediciones establecidas, Figura 12.

Medidas de error - COLEQTY

	Algoritmo	MSE	RMSE	MAE
0	SVR	142.891719	11.953732	7.926489
1	ANN	140.408540	11.849411	6.995393

Medidas de error - COLCAP

	Algoritmo	MSE	RMSE	MAE
0	SVR	415.671962	20.388035	15.079040
1	ANN	374.097698	19.341605	10.887762

Medidas de error - COLSC

	Algoritmo	MSE	RMSE	MAE
0	SVR	230.084902	15.168550	11.893502
1	ANN	117.160075	10.824051	7.664283

Medidas de error - COLIR

	Algoritmo	MSE	RMSE	MAE
0	SVR	126.477150	11.246206	7.467315
1	ANN	129.345728	11.373026	6.780882

Figura 12 Resultados métricas de error por índice. Los autores.

Los resultados de las mediciones de error, dan un primer indicio del grado de acierto de los modelos. Si se analiza el resultado calculado de las métricas de error contra el valor promedio de cada índice en el periodo analizado se puede ver que estos errores son proporcionalmente muy bajos, lo cual indica que el pronóstico tiene un nivel de certeza bastante aceptable.

Adicionalmente, estos resultados ya empiezan a dar nociones de cuál es el algoritmo con mayor rendimiento en cuanto a su predicción, debido a que en la mayor parte de las métricas obtenidas para los Índices COLCAP, COLEQTY y COLSC el menor error lo obtuvo la red neuronal; no obstante, el resultado de la medición de error en el COLIR es menor para el SVR en dos de las tres métricas (MSE y RMSE).

Nuevamente se encuentra un escenario donde se pone en evidencia el valor del modelo y el cálculo de la desviación media propuesta, ya que para este último índice se cuantifica la precisión de un algoritmo con respecto al otro.

3.2.2 Medidas de desviación

Posteriormente se muestra para cada índice, Figura 13, el valor obtenido de la aplicación de la medición de certeza. Esta cuantificación es expresada mediante el cálculo del promedio aritmético de las desviaciones porcentuales obtenidas por los algoritmos en las medidas de error evaluadas.

Medidas de desviación - COLEQTY

	Algoritmo	Desv MSE	Desv RMSE	Desv MAE
0	SVR	0.017685	0.008804	0.133101
1	ANN	0.000000	0.000000	0.000000

Medidas de desviación - COLCAP

	Algoritmo	Desv MSE	Desv RMSE	Desv MAE
0	SVR	0.111132	0.054103	0.384953
1	ANN	0.000000	0.000000	0.000000

Medidas de desviación - COLSC

	Algoritmo	Desv MSE	Desv RMSE	Desv MAE
0	SVR	0.963851	0.401375	0.551809
1	ANN	0.000000	0.000000	0.000000

Medidas de desviación - COLIR

	Algoritmo	Desv MSE	Desv RMSE	Desv MAE
0	SVR	0.000000	0.000000	0.101231
1	ANN	0.022681	0.011277	0.000000

Figura 13 Resultados métricas de desviación por índice. Los autores.

Las desviaciones porcentuales para los índices COLCAP, COLEQTY y COLSC, permiten ver con mayor claridad que en cada una de las métricas de error la red neuronal tuvo resultados más precisos que el SVR, ya que, al obtener resultados de cero, indica el primer algoritmo con mejor resultado en cada una de las mediciones.

Sin embargo, para el caso del índice COLIR, el SVR tuvo una desviación porcentual con valor cero en los errores MSE y RMSE, indicando que, en estas métricas tuvo un mejor resultado; en contraparte, la red neuronal fue la mejor en la métrica MAE, pero al analizar las desviaciones porcentuales en para el MSE y RMSE con respecto a la obtenida por el SVR, estas toman valores de 2% y 1% respectivamente, los cuales son considerados significativamente bajos.

Lo anterior quiere decir que, si bien para este último caso, el SVR tuvo buen desempeño en las dos mediciones analizadas, no fue significativamente muy superior a la red neuronal; lo contrario ocurre si se analizan los resultados de la red neuronal en la medición del MAE donde el SVR tuvo una desviación porcentual media del 10% la cual es bastante superior.

3.2.3 Medidas de desviación media

En la Figura 14, se puede evidenciar el resultado de la comparación de la predicción de los dos algoritmos aplicados:

Desviación media - COLEQTY

	Algoritmo	desv_media
0	SVR	0.053197
1	ANN	0.000000

Desviación media - COLCAP

	Algoritmo	desv_media
0	SVR	0.183396
1	ANN	0.000000

Desviación media - COLSC

	Algoritmo	desv_media
0	SVR	0.639011
1	ANN	0.000000

Desviación media - COLIR

	Algoritmo	desv_media
0	SVR	0.033744
1	ANN	0.011319

Figura 14 Resultados métricas de comparación por índice. Los autores

El promedio aritmético de dichas desviaciones, logra establecer por algoritmo cual fue el que tuvo mejor resultado al ser comparados. Se observa que para los índices COLCAP, COLEQTY y COLSC, la red neuronal obtiene valores de cero, dado que fue el algoritmo con mayor precisión en las tres métricas evaluadas. En el caso del COLIR, dada su complejidad, el algoritmo ANN tuvo un valor de 1%, mientras que, para el SVR fue de 3%, lo cual explica el análisis realizado previamente, donde si bien el SVR tuvo mejor resultado en dos mediciones de error, la diferencia frente al ANN no fue tan notable, mientras que, en el resultado obtenido en el MAE, su bajo desempeño fue bastante mayor.

3.2.4 Algoritmo con mayor precisión

Finalmente, la Tabla 5, muestra los algoritmos con mayor precisión en la predicción de cada uno de los índices bursátiles que fueron elegidos a partir del cálculo de las desviaciones medias, teniendo en cuenta que los más precisos son los que toman un valor menor en la métrica.

Índice Bursátil	Anexo - Software
COLCAP	Red Neuronal (ANN)
COLIR	Red Neuronal (ANN)
COLSC	Red Neuronal (ANN)
COLEQTY	Red Neuronal (ANN)

Tabla 5 Resultados MOC-TEML: Algoritmo más preciso por índice

En consecuencia, la red neuronal fue escogida como el mejor algoritmo para predecir los cuatro índices ya que, en todos los casos logró pronosticar los resultados esperados con mayor nivel de precisión, expresado cuantitativamente mediante el cálculo de la desviación porcentual media, obteniendo siempre el menor valor.

4. CONCLUSIONES

4.1 Verificación, contraste y evaluación de los objetivos

El objetivo general planteó, el diseño de un modelo de comparación de técnicas de Machine Learning, para un caso de uso puntual, que corresponde a la predicción de los índices de la Bolsa de Valores de Colombia (BVC), para llegar a esto, se dio cumplimiento al primer objetivo específico que se refiere a la documentación de técnicas de Machine Learning, donde se seleccionaron los algoritmos SVM y Redes Neuronales Artificiales, por ser de las técnicas más apropiadas para resolver problemas de regresión lineal basadas en series de tiempo, seguidamente se desarrollaron con lenguaje Python dos modelos predictivos uno para cada algoritmo, haciendo uso del historial de datos disponible, desde enero de 2010 hasta abril del presente año, la ejecución de estos modelos dio como resultado, que las Redes Neuronales Artificiales presentan una desviación menor, como se observó en las gráficas de resultados, por ende, al comparar este dos algoritmos para resolver un problema con el mismo modelamiento de datos, este algoritmo tiene más probabilidades de certeza, aunque pueden tener distintos niveles de precisión dependiendo de la serie de tiempo modelada los resultados del caso de estudio favorecieron mayormente a las redes neuronales.

MOC-TEML se establece como un modelo que concuerda con un pipeline adecuado para la aplicación de modelos de aprendizaje automático de tipo supervisados para regresión, por ello el flujo de proceso y marco de trabajo definido permiten estandarizar la implementación de proyectos analíticos, con este modelo se facilita el desarrollo de software de Machine Learning, ya que especifica las pautas y actividades para aplicar satisfactoriamente varios algoritmos, que pretendan solucionar el mismo problema y así elegir el más apropiado para el modelamiento basado en los resultados de sus predicciones.

Modelar la predicción del valor de los índices bursátiles como series de tiempo, que pueden ser pronosticadas a partir de algoritmos Machine Learning supervisados de regresión, ha demostrado ser una técnica bastante efectiva dadas las mediciones de error obtenidas incluso en la predicción de los datos de testeo que no han sido introducidos al aprendizaje del modelo previamente.

En la práctica se identificó que, sin la aplicación de la etapa de validación cruzada, que consolida y apoya potencialmente el entrenamiento del modelo, los niveles de precisión eran poco satisfactorios y además daba a pensar que el algoritmo estaba sobre ajustado, sin embargo, al aplicar el uso de esta técnica e incluso después de evaluar los resultados con el conjunto de datos de pruebas, se descartó la hipótesis y se mejoraron notablemente los resultados.

En cuanto a los recursos de hardware y contrario a lo que se esperaba, dada la revisión de literatura, con el conjunto de datos evaluado y las librerías de Python usadas, el SVR presenta un procesamiento más eficiente en términos de tiempos de cálculo. Para todos los índices evaluados, los tiempos de ejecución del entrenamiento del SVR mediante la librería SKLearn fueron siempre menores a los de las redes neuronales desarrollados con la librería Keras. Incluso con la validación cruzada la diferencia en tiempo de cálculo se hizo más notable teniendo siempre como resultado un tiempo menor para el SVR.

El cálculo de la precisión de los algoritmos a partir de las medidas de comparación definidas como la desviación porcentual media, fueron útiles para la escogencia de los algoritmos con mejor desempeño en su precisión. Se destaca su valor agregado cuando se tienen varios algoritmos cuyos pronósticos son relativamente igual de precisos. Se resalta que esta métrica permite evaluar la certeza a partir de diversas mediciones de error y no solo se enfoca en los mejores valores obtenidos, sino que permite evaluar porcentualmente que tan distantes se encuentran los algoritmos del más preciso en cada métrica de error.

El desarrollo del software logra reafirmar a Python como lenguaje de programación bastante idóneo para la programación de soluciones analíticas basadas en ciencia de datos ya que permite un despliegue e implementación relativamente rápido. Adicionalmente, se lograron evaluar librerías de Machine Learning como SKLearn y Keras las cuales demostraron ser idóneas para el procesamiento y modelamiento de modelos de ciencia de datos. En consecuencia, se resaltan las características de los notebooks como soluciones para documentar desarrollos de analítica avanzada de datos y como entorno de ejecución práctico para el análisis.

La aplicación de este caso de estudio y el software desarrollado pueden ser considerados un aporte de gran valor para el estudio de mercados bursátiles y la toma de decisiones en las bolsas de valores para las operaciones de compra y ventas de valores tales como las acciones. Este caso de estudio permite pronosticar el valor de los índices a futuro, por lo cual presenta una información base para el análisis de diversos conjuntos de acciones.

Finalmente, los resultados del caso de estudio le proporcionan a la Bolsa de Valores de Colombia, economistas e inversores en general, un estudio mediante el cual puede desarrollarse una aplicación que sea capaz de realizar periódicamente un pronóstico del valor de los índices bursátiles a futuro. Teniendo en cuenta que los índices bursátiles están estadísticamente compuestos por una agregación del valor de varias acciones en el mercado reflejando sus variaciones y rentabilidades, dicha aplicación proporciona una base para la toma de decisiones y planeación de acuerdo con el comportamiento del mercado de valores previsto. Adicionalmente los inversores y analistas del mercado de valores, pueden reducir su incertidumbre al momento de realizar transacciones de compra o venta, dado que una proyección a futuro del valor de estos índices permite establecer la percepción del mercado frente al comportamiento de las empresas y de la economía, gestionar portafolios de inversión y realizar un análisis eficiente de riesgos del mercado.

4.2 Síntesis del modelo propuesto

El prototipo propuesto en el presente documento se puede analizar a través del siguiente diagrama. Figura 15.

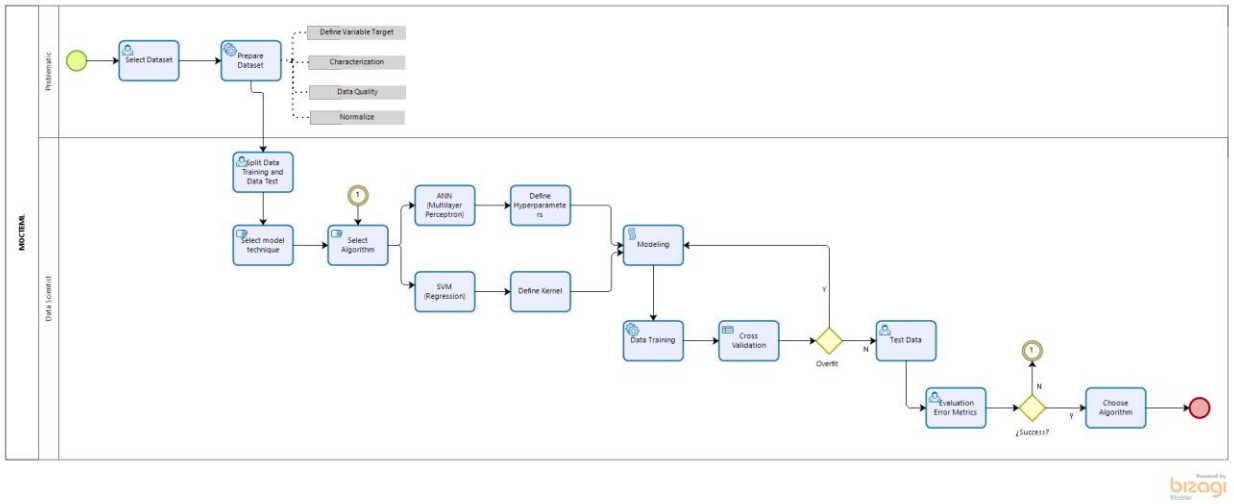

Figura 15 Proceso ML. Los autores.

4.3 Aportes originales

Esta investigación se enfoca en una de las profesiones que están revolucionando el mercado laboral y además se está posicionando como una de las profesiones más deseadas en la industria; el rol de analista o científico de datos permite la toma de decisiones aplicando métodos de estadística, predicciones y procesos automatizados, de esta forma, el modelo de comparación de los algoritmos SVM y Redes Neuronales genera valor, tanto para los profesionales como para las organizaciones, porque les permite acceso al conocimiento aplicado de forma práctica, por ejemplo, en el caso de uso de predicción de índices bursátiles y se complementa con el diagrama del modelo expuesto anteriormente y la fórmula de comparación diseñada para evaluar métricas de error vs resultados.

4.4 Trabajos o Publicaciones derivadas

Académicamente, de esta investigación se derivan dos trabajos, el primero, corresponde a un artículo de revisión, donde a partir del enfoque de técnicas de Machine Learning se hace una evaluación de las tecnológicas de la información en las organizaciones que pueden usar estas herramientas como método de innovación y el estado del arte de las mismas; el segundo con un enfoque gerencial, corresponde a la aplicación de los procesos y procedimiento para la gestión de un proyecto según el marco de trabajo de PMI.

5. PROSPECTIVA DEL TRABAJO DE GRADO

5.1 Líneas de investigación futuras

A futuro es posible aplicar modelos de comparación entre otro tipo de algoritmos, teniendo en cuenta que se clasifican en algoritmos de aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo, se pueden realizar diferentes trabajos comparativos y evaluativos de otro tipo de algoritmos.

Por otra parte, es posible ampliar el caso de estudio aplicando otros algoritmos de regresión como los árboles de decisión, random forest, incluso otros tipos de redes neuronales. En consecuencia, se abre campo a la aplicación de este modelo, incluso con el mismo marco de trabajo, para solucionar diversos casos de estudio como proyecciones de ventas, propensión de clientes, consumos energéticos, predicción de fraude, entre otros.

En cuanto a la tecnología usada para el desarrollo es posible aplicar MOC-TEML para la comparación de precisión entre lenguajes de programación como R o Julia y diversos softwares de advanced analytics como lo son SAS, Alteryx, BigML, entre otros e incluso soluciones en la nube como las propuestas por Microsoft (Azure), Amazon Web Services y Google Cloud.

5.2 Trabajos de investigación futuros

- Se propone crear un modelo con la intervención de más algoritmos de Machine Learning para que la comparación de técnicas sea más enriquecedora para los usuarios finales, además que la integración de estas técnicas sea dinámica y ágil.
- Realizar la integración en línea con dataset de datos abiertos, por ejemplo, los que dispone el programa del MINTIC, Gobierno en Línea, para que la información sea obtenida y analizada en la nube, apoyados en las plataformas online que ofrece Google Colaboratory para el desarrollo de analítica.

BIBLIOGRAFÍA

[1] P. Ongsulee, "Artificial intelligence, machine learning and deep learning," in *2017 15th International Conference on ICT and Knowledge Engineering (ICT&KE)*, 2017, pp. 1–6.

[2] M. G. Pecht and M. Kang, "Machine Learning: Fundamentals," in *Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things*, IEEE, 2019, p. 1.

[3] F. J. Alexander, "Machine Learning," *Comput. Sci. Eng.*, vol. 15, no. 5, pp. 9–11, 2013.

[4] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2012.

[5] E. Alpaydin, *Introduction to machine learning*. MIT press, 2014.

[6] S. D. Essinger and G. L. Rosen, "An introduction to machine learning for students in secondary education," *2011 Digit. Signal Process. Signal Process. Educ. Meet.*, pp. 243–248, 2011.

[7] R. Saravanan and P. Sujatha, "Algorithms : A Perspective of Supervised Learning Approaches in Data Classification," *2018 Second Int. Conf. Intell. Comput. Control Syst.*, no. Iciccs, pp. 945–949, 2018.

[8] I. Gandin and C. Cozza, "Can we predict firms' innovativeness? The identification of innovation performers in an Italian region through a supervised learning approach," *PLoS One*, vol. 14, no. 6, pp. 1–17, 2019.

[9] K. Thirunavukkarasu, A. S. Singh, P. Rai, and S. Gupta, "Classification of IRIS Dataset using Classification Based KNN Algorithm in Supervised Learning," *2018 4th Int. Conf. Comput. Commun. Autom.*, pp. 1–4, 2019.

[10] J. S. Angarita-Zapata, A. D. Masegosa, and I. Triguero, "A Taxonomy of Traffic Forecasting Regression Problems from a Supervised Learning Perspective," *IEEE Access*, vol. 7, pp. 68185–68205, 2019.

[11] A. C. Braun, U. Weidner, and S. Hinz, "Support vector machines, import vector machines and relevance vector machines for hyperspectral classification - A comparison," *Work. Hyperspectral Image Signal Process. Evol. Remote Sens.*, vol. 2, no. 3, pp. 1–4, 2011.

[12] P. Bohra and H. Palivela, "Understanding and formulation of various kernel techniques for suport vector machines," *2015 IEEE Int. Conf. Comput. Intell. Comput. Res. ICCIC 2015*, pp. 1–6, 2016.

[13] D. X. Niu, Q. Wang, and J. C. Li, "Short term load forecasting model using support vector machine based on artificial neural network," *2005 Int. Conf. Mach. Learn. Cybern. ICMLC 2005*, no. August, pp. 4260–4265, 2005.

[14] C. N. Babu and P. Sure, "Partitioning and interpolation based hybrid ARIMA – ANN model for time series forecasting," *Sādhanā*, vol. 41, no. 7, pp. 695–706, 2016.

[15] J. R. Hilera González and V. J. Martínez Hernando, *Redes neuronales artificiales: fundamentos, modelos y aplicaciones*. 2000.

[16] E. Varela and E. Campbells, "Redes Neuronales Artificiales : Una Revisión del Estado del Arte , Aplicaciones Y Tendencias Futuras Artificial Neural Networks : A Brief Review," *Investig. y Desarro. en TIC*, vol. 2, no. 1, pp. 18–27, 2011.

[17] S. J. Kwon, *Artificial neural networks*. New York: Nova Science Publ., 2011.

[18] R. Lackes and D. Mack, "Neural Networks. Basics and Applications." CBT Program on Neural Nets (c) Springer-Verlag Berlin, Heidelberg, 1998.

[19] W. A. Mahmood, "Dynamics Using Locally Weighted Projection," 2017.

[20] B. Imanol and J. Bilbao, "Overfitting problem and the over-training in the era of data: Particularly for Artificial Neural Networks," *Eighth Int. Conf. Intell. Comput. Inf. Syst.*, no. Icicis, pp. 173–177, 2017.

[21] S. Lawrence and C. L. Giles, "Overfitting and neural networks: Conjugate gradient and backpropagation," *Proc. Int. Jt. Conf. Neural Networks*, vol. 1, pp. 114–119, 2000.

[22] A. Ghasemian, H. Hosseinmardi, and A. Clauset, "Evaluating Overfit and Underfit in Models of Network Community Structure," *IEEE Trans. Knowl. Data Eng.*, no. October, pp. 1–1, 2019.

[23] M. M. R. Majumder, M. I. Hossain, and M. K. Hasan, "Indices prediction of Bangladeshi stock by using time series forecasting and performance analysis," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2019, pp. 1–5.

[24] A. P. Palacio, C. P. Orozco, S. R. Arenas, and C. Lochmüller, "Daily Ecopertrol stock performance estimate by using estimation of distribution algorithms (EDAs): Evolutionary computation," in *7th Iberian Conference on Information Systems and Technologies (CISTI 2012)*, 2012, pp. 1–6.

[25] P. A. Peña, F. Gómez, and J. M. Vélez, "Vectorial model for progressive adaptation for purchase and sale of shares using stock market indicators," in *2016 11th Iberian Conference on Information Systems and Technologies (CISTI)*, 2016, pp. 1–7.

[26] A. García, *Inteligencia artificial: fundamentos, práctica y aplicaciones*. Rc Libros, 2012.

[27] J. G. Naragund, C. Sujatha, K. G. Karibasappa, S. Giraddi, and S. Yaligar, "Experimental learning in Scripting Languages laboratory," in *2015 IEEE 3rd International Conference on MOOCs, Innovation and Technology in Education (MITE)*, 2015, pp. 213–218.

[28] R. Filguiera, I. Klampanos, A. Krause, M. David, A. Moreno, and M. Atkinson, "dispel4py: A Python Framework for Data-Intensive Scientific Computing," in *2014 International Workshop on Data Intensive Scalable Computing Systems*, 2014, pp. 9–16.

[29] "Bolsa de Valores de Colombia." .

[30] L. Latha and S. Thangasamy, “Efficient approach to Normalization of Multimodal Biometric Scores,” *Int. J. Comput. Appl.*, vol. 32, no. 10, pp. 975–8887, 2011.

[31] “Dividir los datos en datos de formación y evaluación - Amazon Machine Learning.” .

[32] “Preventing Deep Neural Network from Overfitting – mc.ai.” .

[33] C. Deb, F. Zhang, J. Yang, S. E. Lee, and K. W. Shah, “A review on time series forecasting techniques for building energy consumption,” *Renew. Sustain. Energy Rev.*, vol. 74, pp. 902–924, 2017.

[34] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, “Manipulating machine learning: Poisoning attacks and countermeasures for regression learning,” in *2018 IEEE Symposium on Security and Privacy (SP)*, 2018, pp. 19–35.

[35] A. Tealab, “Time series forecasting using artificial neural networks methodologies: A systematic review,” *Futur. Comput. Informatics J.*, vol. 3, no. 2, pp. 334–340, 2018.

[36] H. Ramchoun, M. A. J. Idrissi, Y. Ghanou, and M. Ettaouil, “Multilayer Perceptron: Architecture Optimization and Training.,” *IJIMAI*, vol. 4, no. 1, pp. 26–30, 2016.

[37] G. Santamaría-Bonfil, A. Reyes-Ballesteros, and C. Gershenson, “Wind speed forecasting for wind farms: A method based on support vector regression,” *Renew. Energy*, vol. 85, pp. 790–809, 2016.

[38] S. Suthaharan, “Support vector machine,” in *Machine learning models and algorithms for big data classification*, Springer, 2016, pp. 207–235.

[39] B. Feizizadeh, M. S. Roodposhti, T. Blaschke, and J. Aryal, “Comparing GIS-based support vector machine kernel functions for landslide susceptibility mapping,” *Arab. J. Geosci.*, vol. 10, no. 5, p. 122, 2017.

[40] H. Shi, H. Xiao, J. Zhou, N. Li, and H. Zhou, “Radial Basis Function Kernel Parameter Optimization Algorithm in Support Vector Machine Based on Segmented Dichotomy,” in *2018 5th International Conference on Systems and Informatics (ICSAI)*, 2018, pp. 383–388.

[41] S. Arlot and A. Celisse, “A survey of cross-validation procedures for model selection,” *Stat. Surv.*, vol. 4, pp. 40–79, 2010.

[42] Q. Ren, M. Li, and S. Han, “Tectonic discrimination of olivine in basalt using data mining techniques based on major elements: a comparative study from multiple perspectives,” *Big Earth Data*, vol. 3, no. 1, pp. 8–25, 2019.

ANEXOS

Anexo 1: COLCAP MOC-TEML ANN-SVR.ipynb
Anexo 2: COLIR MOC-TEML ANN-SVR.ipynb
Anexo 3: COLSC MOC-TEML ANN-SVR.ipynb
Anexo 4: COLEQTY MOC-TEML ANN-SVR.ipynb