

TECNOLÓGICO NACIONAL DE MÉXICO/ INSTITUTO TECNOLÓGICO DE CIUDAD MADERO

DEPARTAMENTO DE DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN



INFORME FINAL DE AÑO SABÁTICO

USO DE TÉCNICAS PARA LA MEJORA DEL PRONÓSTICO EN SERIES DE TIEMPO

Que presenta

Dr. Juan Javier González Barbosa

Realizado en

Instituto Tecnológico de Ciudad Madero

Tamaulipas, México

Duración del periodo

01/SEP/2024 al 31/AGO/2025

Dictamen No. AS-2-205/2024

Ciudad Madero, Tamaulipas, septiembre de 2025



Ciudad de México, **22 de agosto de 2024**

AUTORIZACIÓN No AS-2-205/2024

**JUAN DIONISIO CRUZ GUERRERO
DIRECTOR DEL INSTITUTO TECNOLÓGICO DE
CIUDAD MADERO
PRESENTE**

Con fundamento en el Reglamento Interior de Trabajo del Personal Docente de los Institutos Tecnológicos, en las Políticas Académicas Generales del Periodo Sabático y las bases de la Convocatoria correspondiente, el Tecnológico Nacional de México a través de la Comisión Central, ha revisado la solicitud del **Año Sabático** del **C. JUAN JAVIER GONZÁLEZ BARBOSA** en el Programa de Investigación Científica, Tecnológica o Educativa, bajo la modalidad: **A.1. Proyecto de Investigación Científica: Uso de Técnicas para la mejora del Pronóstico en Series de Tiempo**, teniendo a bien emitir el siguiente dictamen de:

AUTORIZACIÓN

Del programa presentado, a partir del **02 de septiembre de 2024** hasta el **01 de septiembre de 2025**, quedando su seguimiento bajo la responsabilidad de la Dirección a su cargo, debiendo recibir: reporte intermedio y final (evaluaciones por parte de la Academia correspondiente y de la Comisión Dictaminadora de las actividades establecidas en la calendarización que se encuentra en el expediente de la solicitud del docente), ejemplar en físico y en electrónico (en formato PDF) del producto final comprometido.

Así mismo, se deberá remitir a la Dirección de Docencia e Innovación Educativa del Tecnológico Nacional de México las evaluaciones antes mencionadas.



M00-PR-04-R08



Av. Universidad 1200, Col Xoco, Alcaldía Benito Juárez, C.P. 06630, Ciudad de México.
Tel. (55) 3601-7500, Ext. 65050, e-mail: direccion@tecnm.mx tecnm.mx




SECRETARÍA DE EDUCACIÓN PÚBLICA
TECNOLÓGICO NACIONAL
DE MÉXICO
DIRECCIÓN GENERAL

Versión 4



2024
Felipe Carrillo
PUERTO
GOBIERNO DEL ESTADO DE YUCATÁN
SECRETARÍA DE EDUCACIÓN Y DEPORTES

Índice	Pag
I.- Identificación del Proyecto	3
II.- Resultados.....	3
1.- Resumen del Proyecto.....	3
2.- Objetivo del Proyecto.....	4
3.- Metas. Cumplimiento de metas.....	4
4.- Metodología.....	8
5.- Difusión.....	14
6.- Beneficios y Problemas.....	15
7.- Información adicional.....	16
Anexos.....	16
Referencias.....	17

I.- Identificación del Proyecto

Instituto o Centro: **Tecnológico Nacional de México/ Instituto Tecnológico de Ciudad Madero**

Nombre del docente: **Juan Javier González Barbosa**

Título del proyecto: **Uso de Técnicas para la mejora del Pronostico en Series de Tiempo**

Tipo de investigación: **PROYECTO DE INVESTIGACIÓN CIENTÍFICA**

Duración del proyecto: **Un año**

Fecha de inicio del proyecto: **01 de septiembre 2024**

Fecha de término del proyecto: **31 de agosto de 2025**

II.- Resultados

1.- Resumen del Proyecto

Este proyecto busca realizar un análisis y descripción sobre algunas técnicas que ayudan a mejorar el pronóstico de diferentes tipos de series de tiempo, particularmente series financieras y de cambio climático.

El problema de pronóstico es vital en una gran cantidad de áreas de investigación, como son problemas asociados a energía, industria, series financieras y cambio climático. En la actualidad, este tipo de problemas usa una combinación de métodos de pronóstico para encontrar la solución. Es decir, el método a desarrollar utiliza diferentes métodos individuales para lograr un buen pronóstico. Sin embargo, se requiere el uso de otras técnicas, antes, durante y al final del proceso del pronóstico para obtener el mejor desempeño o el menor error.

Dado lo anterior es necesario hacer una descripción y análisis teórico y en algunos casos practico para su uso. Para lo cual tendremos dos casos de estudio.

2.- Objetivo del Proyecto

Analizar el uso de Técnicas de mejora para el pronóstico de Series de Tiempo.

Objetivos Particulares

Objetivo particular 1 (OP1): Analizar el estado del arte de los métodos de pronóstico que formarán parte de la investigación.

Objetivo particular 2 (OP2). Analizar y evaluar los mejores métodos de pronóstico reportados en el estado del arte.

Objetivo particular 3 (OP). Desarrollar los métodos de solución de los problemas considerados.

Objetivo particular 4 (OP4). Evaluar experimentalmente el desempeño de los métodos de solución considerados.

Objetivo particular 5 (OP5). Difundir resultados en diversos foros.

3.- Metas. Cumplimiento de metas

Las metas que se alcanzarán como resultado de este proyecto de investigación son:

- Consolidar la línea de investigación de Simulación y Computo Aplicado en el Instituto Tecnológico de Cd. Madero.
- Generar nuevo conocimiento relativo a la Ciencia de Datos, validado por la comunidad científica especializada en esta área.
- Buscar la permanencia en el SNP de los programas Ciencias de la Ingeniería y Ciencias de la Computación.
- Incrementar la Producción Científica con el fin de continuar permaneciendo en el SNII.
- Formar estudiantes de doctorado, maestría y licenciatura en el campo de la Ciencia de Datos.

Metas cuantificables	Cantidad programada	Cantidad lograda	Observaciones
Incorporación de estudiantes de licenciatura al proyecto, créditos complementarios:			
Incorporación de estudiantes de licenciatura al proyecto, servicio social:			
Estudiantes con residencia concluida			
Tesis en desarrollo de Licenciatura			
Tesis concluida de Licenciatura			
Tesis en desarrollo de Maestría	1	1	
Tesis concluida de Maestría			
Tesis en desarrollo de Doctorado	1	1	
Tesis concluida de Doctorado			
Artículos científicos publicados en revistas indizadas enviados			
Artículos científicos en revistas arbitradas enviados			
Artículos de divulgación enviados			
Artículos científicos en revistas indizadas publicados		2	
Artículos científicos en revistas arbitradas publicados			
Artículos de divulgación publicados	2	2	
Memorias en extenso en congresos			
Capítulos de libros enviados para revisión			
Libros enviados para revisión			
Libros editados y publicados			
Registro de Patente (IMPI)			
Registro de Modelo de Utilidad (IMPI)			
Registro de Signos Distintivos (IMPI)			
Registro de Diseño Industrial (IMPI)			
Registro de Esquemas de Trazado de Circuitos Integrados (IMPI):			
Registro de Obra (INDAUTOR):			
Carta de Usuario (Empresa):			

Metas cumplidas

Incorporación de estudiantes de licenciatura al proyecto, créditos complementarios:	
Fecha	Nombre de los estudiantes incorporados y su Programa Educativo

Incorporación de estudiantes de licenciatura al proyecto, servicio social:	
Fecha	Nombre de los estudiantes incorporados y su Programa Educativo
Estudiantes residentes participantes en el proyecto	
Fecha	Nombre de los estudiantes residentes, nombre del proyecto de residencia y su Programa Educativo

Tesis en desarrollo de Licenciatura	
Fecha	Nombre de la tesis

Tesis concluidas de Licenciatura	
Fecha	Nombre de la tesis

Tesis en desarrollo de Maestría	
Fecha	Nombre de la tesis
Agosto 2025	ESTIMACIÓN DE LA TEMPERATURA USANDO ANÁLISIS MULTIVARIABLE

Tesis concluidas de Maestría	
Fecha	Nombre de la tesis

Tesis en desarrollo de Doctorado	
----------------------------------	--

Fecha	Nombre de la tesis
Agosto 2025	Portafolios de Inversión Utilizando Métodos

	Heurísticos y de Aprendizaje Profundo en Mercados Estandarizados
Tesis concluidas de Doctorado	
Fecha	Nombre de la tesis
Artículos científicos publicados o enviados en revistas indizadas	
Fecha	Nombre del artículo y acuse de envío
Septiembre 2024	A Comparative Analysis of the Classical and Machine learning Forecasting Methods for the Mexican Stock Exchange
Marzo 2025	MASIP: A Methodology for Assets Selection in Investment Portfolios
Artículos científicos publicados o enviados en revistas arbitradas	
Fecha	Nombre del artículo y acuse de envío
Artículos de divulgación publicados o enviados	
Fecha	Nombre del artículo y acuse de envío
Mayo 2025	Multivariate Temperature Prediction using deep learning
Mayo 2025	Forecasting and Optimization of stock Market assets using metaheuristics and deep learning
Memorias en extenso en congresos	
Fecha	Nombre y memoria del Congreso
Capítulo(s) de libros enviados para revisión	
Fecha	Nombre del Libro y del/de los capítulo(s) y acuse de enviado
Libro(s) enviados para revisión	
Fecha	Nombre del/de los Libro(s) y acuse de enviado
Libro(s) editado(s) y publicado(s)	
Fecha	Nombre del/de los Libro(s) y Editorial(es)
Registro de Patente (IMPI):	
Fecha	Acuse de solicitud de registro de patente
Registro de Modelo de Utilidad (IMPI)	

Fecha	Acuse de solicitud de registro

Registro de Signos Distintivos (IMPI):	
Fecha	Acuse de solicitud de registro

Registro de Diseño Industrial (IMPI)	
--------------------------------------	--

Fecha	Acuse de solicitud de registro

Registro de Esquemas de Trazado de Circuitos Integrados (IMPI):	
Fecha	Acuse de solicitud de registro

Registro de Obra (INDAUTOR):	
Fecha	Acuse de solicitud de registro

Carta de Usuario (Empresa):	
Fecha	Acuse de solicitud de registro

4.- Metodología

Explique el o los procedimientos científico-metodológicos que se siguieron para cumplir los objetivos y metas del proyecto, indicando las pruebas estadísticas, diseño experimental y técnicas a utilizar (máximo dos cuartillas).

El problema de pronóstico es vital en una gran cantidad de áreas de investigación, como son problemas asociados a energía, industria, series financieras y cambio climático. En la actualidad, este tipo de problemas usa una combinación de métodos de pronóstico para encontrar la solución. Es decir, el método a desarrollar utiliza diferentes métodos individuales para lograr

un buen pronóstico. Sin embargo, se requiere el uso de otras técnicas, antes, durante y al final del proceso del pronóstico para obtener el mejor desempeño o el menor error.

Dado lo anterior es necesario hacer una descripción y análisis teórico y en algunos casos práctico para su uso. Para lo cual tendremos dos casos de estudio.

- Caso 1.- Para Series de Tiempo Financieras, se realizará el pronóstico para diferentes activos.
- Caso 2.- Para series de tiempo de cambio climático, se realizará el pronóstico para al menos cuatro variables (Ciudades).

La metodología consta de las siguientes etapas.

Etapas 1. El primer paso es establecer los conjuntos de datos (Series financieras y series de cambio climático). En esta etapa es importante analizar de forma precisa las series. Posteriormente se hará una preselección para formas conjuntos con diferentes problemáticas.

Después de haber concluido con la preselección, se procede a la limpieza de los datos recolectados, lo que facilitará el procesado al brindar datos confiables para poder realizar el análisis. Para este punto es de gran relevancia el revisar las técnicas de limpieza de datos clásicas para seleccionar la más adecuada al problema.

Etapas 2. Una vez que se tengan los datos listos, se sintonizan los hiperparametros de los métodos individuales con el fin de obtener el mejor resultado de estos, posteriormente se seleccionan los métodos individuales que se utilizaran para la construcción del ensamble de métodos.

Etapas 3. Dado el pronóstico inicial se procede a su optimización, para lo cual se varían los pesos de cada método participante. En este punto es necesario utilizar alguna métrica de desempeño como: MAPE (Mean absolute percentage error), SMAPE (symmetric Mean Absolute

Percentage Error) o MASE (Mean Absolute Scaled Error), con el fin de encontrar la mejor combinación de pesos asignada a cada método.

Etapas 4. Se comparan resultados con respecto a los mejores métodos participantes y se realiza el reporte con los detalles del método para su difusión.

Caso 1

Para Series de Tiempo Financieras, se realizará el pronóstico para diferentes activos. Los pronósticos son estimaciones futuras basadas en patrones y tendencias identificados en datos históricos. Estas estimaciones se utilizan en una variedad de industrias, desde finanzas y economía hasta la gestión de inventarios y la toma de decisiones empresariales (Makridakis, S., 1998).

Etapas 1. Para este caso se utilizó la Bolsa de Valores de Bovespa, que actualmente cuenta con más de 1,500 empresas y una capitalización bursátil de un billón de dólares (Yahoo Finance, s. f.). Se utilizaron los registros del valor de los activos del mercado de valores de Bovespa, que constan de 22 series de tiempo financieras que contienen información desde el 2 de enero de 2020 hasta el 28 de julio de 2022.

Después de haber seleccionado el mercado y las series se procede a realizar el preprocesamiento lo cual implica ejecutar una serie de análisis y para que las series puedan ser utilizadas por los algoritmos que harán el pronóstico.

Dado que los datos generalmente provienen de bases de datos reales, implica limpiar, transformar y organizar los datos para su análisis (Kelleher, J. D., & Tierney, B., 2018; Provost, F., & Fawcett, T., 2013; Kuhn, M., & Johnson, K., 2019). Este paso crucial asegura que tus datos estén libres de errores, inconsistencias y elementos innecesarios, a esto se le conoce como preprocesamiento.

Etapas 2. Selección de los métodos individuales:

El conjunto de datos utilizado para desarrollar esta etapa forma parte de la Bolsa de Valores de Bovespa, que actualmente cuenta con más de 1,500 empresas y una capitalización bursátil de un billón de dólares (Yahoo Finance, s. f.). Se utilizaron los registros del valor de los activos del mercado de valores de Bovespa, que constan de 22 series de tiempo financieras que contienen información desde el 2 de enero de 2020 hasta el 28 de julio de 2022. El conjunto de datos incluye 641 precios de cierre diarios para cada serie temporal, los cuales fueron divididos en un 80% para el bloque de entrenamiento, un 10% para la validación y un 10% para prueba. Las series se clasificaron en cuatro sectores principales: Energía, Financiero, Industrial y Transporte.

Con la finalidad de mejorar el desempeño de los modelos de pronóstico es fundamental configurar los hiperparámetros, ya que esto define su capacidad para generalizar y evitar problemas como el sobreajuste o el desajuste (Victoria, A.H., & Maragatham, G., 2021; Wu, J. et al. 2019). Un modelo sobreajustado presenta un error bajo en el conjunto de entrenamiento, pero un rendimiento deficiente con datos desconocidos, mientras que un modelo desajustado carece de la flexibilidad para capturar patrones relevantes.

Algunas técnicas pueden ser el ajuste bayesiano, el cual prioriza las configuraciones de buen rendimiento utilizando un modelo probabilístico iterativo, lo que reduce el espacio de búsqueda y aumenta la eficiencia u otro más tradicional es el GridSearch, el cual se adapta dinámicamente a resultados anteriores, maximizando la probabilidad de encontrar configuraciones óptimas con menos evaluaciones (Wu, J. et al. 2019). Una vez ajustado los hiperparámetros del modelo se procede a su utilización.

Etapas 3 y 4. Dado el pronóstico inicial se procede a medir su desempeño utilizando métricas de error como: MAPE (Mean absolute percentage error) o SMAPE (symmetric Mean

Absolute Percentage Error), con el fin de comparar los resultados y realizar el reporte con los detalles del método para su difusión.

Los resultados en esta etapa mostraron que el algoritmo para pronóstico CNN, tiene mejor desempeño para el sector Industrial, con el menor porcentaje de error registrado para las métricas. Los resultados se muestran en la Tabla 2.

Tabla 2.- Desempeño en sectores			
Sectores	MAPE	MSE	SMAPE
Energía	3.79	0.01	3.89
Financiero	4.27	0.01	4.46
Industrial	2.70	0.002	2.69
Transporte	4.45	0.002	4.36
Error promedio	3.74	0.005	3.80

Conclusiones

El algoritmo CNN muestra un desempeño consistente al procesar las 22 series de forma general para obtener el pronóstico de los activos, alcanzando un MAPE de 3.74%. Durante las experimentaciones realizadas se pudo observar que el algoritmo CNN tiene una mayor fortaleza para pronosticar las series del sector Industrial al obtener un MAPE de 2.70%.

Caso 2

Para series de tiempo de cambio climático, se realizará el pronóstico para al menos cuatro variables (Ciudades). Siguiendo la metodología del presente trabajo tenemos las siguientes etapas.

Etapas 1. El conjunto de datos utilizado para este estudio proviene de fuentes de libre acceso correspondiente a mediciones de temperatura de los aeropuertos en las ciudades objeto de estudio,

complementadas con registros de las estaciones meteorológicas locales, (tomados de tutiempo.net). Estos datos representan los promedios mensuales de temperatura para cuatro ciudades costeras del Golfo de México: Tampico, Matamoros, Houston y Miami, abarcando el periodo de enero de 1997 hasta diciembre de 2023, proporcionando una base amplia para su análisis.

Etapas 2. Una vez que se tengan los datos listos, se inicia con la selección de los métodos individuales que se utilizarán para hacer el pronóstico. Para este caso se utilizará el método Holt and Winters, el cual es un procedimiento de suavizado exponencial para series temporales que exhiben tendencias y patrones estacionales.

El método de Holt-Winters aplica suavizado exponencial a cada uno de los tres componentes mencionados. Utiliza tres parámetros de suavizado: α para el nivel, β para la tendencia, y γ para la estacionalidad. Estos parámetros controlan la ponderación relativa de las observaciones más recientes frente a las más antiguas, y son cruciales para el ajuste del modelo; su selección puede basarse en criterios de optimización como la minimización del error cuadrático medio (Sheng, J. et al. 2023). Una de las principales ventajas de Holt-Winters es su capacidad para adaptarse a cambios en la tendencia y la estacionalidad sin necesidad de modelos complejos (Domínguez, D. et al. 2023).

Etapas 3 y 4. Dado el pronóstico inicial se procede a medir su desempeño utilizando métricas de error como: MAPE (Mean absolute percentage error) o SMAPE (symmetric Mean Absolute Percentage Error), con el fin de comparar los resultados y realizar el reporte con los detalles del método para su difusión. Los resultados se muestran a continuación.

Tabla 3. Desempeño de Holt and Winters.

	MAPE	DA	RMSE
Tampico	3.6254	74.4681	1.1132
Matamoros	4.89	85.11	1.57
Miami	2.68	93.62	0.83
Houston	6.5292	85.1064	1.6479

Conclusiones

El algoritmo Holt and Winters muestra un buen desempeño para las ciudades de Miami, Tampico y Matamoros al tener un error menor al 5%.

5.- Difusión

Se participo en el 4º Congreso Internacional de tecnología y ciencia Aplicada (CITCA), celebrado en el Centro Nacional de Investigación y desarrollo Tecnológico, con dos ponencias:

Multivariate Temperature Prediction using deep learning, presentada por Allan Mizael Jiménez Fernández, estudiante de maestría.

Forecasting and Optimization of stock Market assets using metaheuristics and deep learning, presentada por Francisco Rivera Vargas, estudiante de Doctorado.

Ambos estudiantes están desarrollando su tesis como parte de este proyecto.

La primera ponencia fue elegida para contribuir como **ARTICULO PARA CONGRESO**, en el área de CIENCIAS DE LA COMPUTACION.

La segunda fue elegida para contribuir como **ARTICULO PARA REVISTA**.

Otro logro de este Proyecto fue la publicación de **DOS ARTICULOS CIENTIFICOS EN REVISTAS INDIZADAS**

1. A Comparative Analysis of the Classical and Machine learning Forecasting Methods for the Mexican Stock Exchange
2. MASIP: A Methodology for Assets Selection in Investment Portfolios

6.- Beneficios y Problemas

Los beneficios que aporta el presente proyecto se verán reflejados principalmente en:

- a) fortalecimiento del cuerpo académico.
- b) Incremento en la productividad científica y tecnológica.
- c) Formación de recursos humanos.

De manera particular este proyecto beneficia a los programas de Ingeniería en Sistemas y Computación, al Posgrado en Ciencias de la Computación y al Posgrado en Ciencias de la Ingeniería, del Tecnológico Nacional de México/ Instituto Tecnológico de Ciudad Madero.

7.- Información adicional

ANEXO 1.- Boletas y actas del comité tutorial que dan cuenta de la formación de recursos humanos en el desarrollo de sus tesis.

ANEXO 2.- Reconocimientos y artículos de divulgación.

ANEXO 3.- Artículos científicos en revistas indizadas.

 Subdirector Académico Jose Aaron Melo Banda Nombre y Firma	 Docente Juan Javier Gonzalez Barbosa Nombre y Firma
---	---

Referencias

- Domínguez, D., Barriuso Pastor, J., Pantoja-Díaz, O., González-Rodríguez, M., 2023, Forecasting Worldwide Temperature from Amazon Rainforest Deforestation Using a Long-Short Term Memory Model. *Sustainability*, 15(20), 15152.
- Estrada-Patiño, E., Castilla-Valdez, G., Frausto-Solis, J., González-Barbosa, J., Sánchez-Hernández, J.P., 2024, A Novel Approach for Temperature Forecasting in Climate Change Using Ensemble Decomposition of Time Series. *Int. J. Comput. Intell. Syst.*, 17, 253, doi:10.1007/s44196-024-00667-6.
- Frausto-Solis, J., Hernandez-Gonzalez, L.J., Gonzalez-Barbosa, J.J., Sanchez-Hernandez, J.P., Roman-Rangel, E., 2021, Convolutional Neural Network–Component Transformation (CNN–CT) for Confirmed COVID-19 Cases. *Math. Comput. Appl.*, 26, 29. <https://doi.org/10.3390/mca26020029>.
- Johnstone, I.M., Titterington, D.M., 2009, Statistical Challenges of High-Dimensional Data. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.*, 367, 4237–4253, doi:10.1098/rsta.2009.0159.
- Kelleher, J. D., Tierney, B., 2018, *Data Science*. The MIT Press.
- Kingma, D.P., Ba, J., 2015, Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR, San Diego, CA, USA, 7–9 May*.
- Kuhn, M., Johnson, K., 2019, *Feature Engineering and Selection: A Practical Approach for Predictive Models*, 1st ed. Chapman and Hall/CRC.

- Lu, Y., Cohen, I., Zhou, X.S., Tian, Q., 2007, Feature Selection Using Principal Feature Analysis. In Proceedings of the 15th ACM international conference on Multimedia; ACM: Augsburg Germany, September 29; pp. 301–304.
- Makridakis, S., 1998, Forecasting: methods and applications (Vol. 2). John Wiley & Sons.
- Malhi, A., Gao, R.X., 2004, PCA-Based Feature Selection Scheme for Machine Defect Classification. IEEE Trans. Instrum. Meas., 53, 1517–1525.
- Provost, F., Fawcett, T., 2013, Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking, 1st ed. O'Reilly Media.
- Sheng, J., Miao, Y., Wang, L., 2023, An Assessment of the Impact of Climate Change on Asphalt Binder Selection in East China Based on the ARIMA Model. Sustainability, 15(21), 15667.
- Song, F., Guo, Z., Mei, D., 2010, Feature Selection Using Principal Component Analysis. In Proceedings of the 2010 international conference on system science, engineering design and manufacturing informatization; IEEE, Vol. 1, pp. 27–30.
- Victoria, A.H., Maragatham, G., 2021, Automatic Tuning of Hyperparameters Using Bayesian Optimization. Evol. Syst., 12, 217–223, doi:10.1007/s12530-020-09345-2.
- Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H., Deng, S.-H., 2019, Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization. J. Electron. Sci. Technol., 17, 26–40.
- Yahoo Finance, s. f., Historial de precios de IBOVESPA (^BVSP). Yahoo Finance. Recuperado de <https://finance.yahoo.com/quote/%5EBVSP/history/>.

ANEXO 1

NÚMERO DE CONTROL: G15070895

FECHA: 05 de Febrero de 2025

ESTUDIANTE: JIMENEZ FERNANDEZ ALLAN MIZAE

SEMESTRE: 3

CARRERA: M.C. DE LA INGENIERIA

RETÍCULA: 174(MCING-2011-45)

ESPECIALIDAD:

CLAVE	MATERIA	CR	CALIFICACIÓN	TIPO DE EVALUACION	OBSERVACIONES
31741	SEM DE INV III	4	90	Ev.Ord.1ra	
41743	Temas selectos II	6	94	Ev.Ord.1ra	

2475a22f32318c8cbcd2e82a75a7a862

CERTIFICADO DIGITAL

Verifica su autenticidad en

<http://sii.cdmadero.tecnm.mx>



IMPRESO EL DIA 2025-06-20 A LAS 17:26:25 HRS



Instituto Tecnológico de Ciudad Madero
Subdirección Académica
División de Estudios de Posgrado e Investigación

**EVALUACIÓN DEL COMITÉ TUTORIAL
POSGRADO EN CIENCIAS DE LA INGENIERÍA**

Ciudad Madero, Tam., a 14 de mayo de 2025

Nombre	Allan Mizael Jiménez Fernández		
Número de Control	G15070895	Semestre Inscrito	4to
Programa	Maestría	en Ciencias de la Ingeniería	

PROYECTO DE INVESTIGACIÓN			
Título (Dato obligatorio)	Estimación de la temperatura usando análisis multivariable		
Director(es) de Tesis	Dr. Juan Javier González Barbosa		
Porcentaje de avance anterior		Calificación anterior	

COMENTARIOS GENERALES			
TERMINAR INVESTIGACION			
Porcentaje de avance actual	85	Calificación del profesor	90

Juan Francisco Solís
NOMBRE Y FIRMA DEL MIEMBRO DEL
COMITÉ TUTORIAL



2025
Año de
La Mujer
Indígena

Av. 1° de Mayo y Sor Juana I. de la Cruz S/N Col. Los Mangos
C.P. 89440 Cd. Madero, Tam. Tel. (833) 3574820 ext. 3110
e-mail: depi_cdmadero@tecnm.mx www.tecnm.mx
www.cdmadero.tecnm.mx





**EVALUACIÓN DEL COMITÉ TUTORIAL
POSGRADO EN CIENCIAS DE LA INGENIERÍA**

Ciudad Madero, Tam., a 14 de mayo de 2025

Nombre	Allan Mizael Jiménez Fernández		
Número de Control	G15070895	Semestre Inscrito	4to
Programa	Maestría	en Ciencias de la Ingeniería	

PROYECTO DE INVESTIGACIÓN			
Titulo (Dato obligatorio)	Estimación de la temperatura usando análisis multivariable		
Director(es) de Tesis	Dr. Juan Javier González Barbosa		
Porcentaje de avance anterior		Calificación anterior	

COMENTARIOS GENERALES			
Porcentaje de avance actual	85%	Calificación del profesor	70

NOMBRE Y FIRMA DEL MIEMBRO DEL
COMITÉ TUTORIAL

Juan Javier González B.

Av. 1° de Mayo y Sor Juana I. de la Cruz S/N Col. Los Mangos
C.P. 89440 Cd. Madero, Tam. Tel. (833) 3574820 ext. 3110
e-mail: depi_cdmadero@tecnm.mx www.tecnm.mx
www.cdmadero.tecnm.mx



2025
Año de
La Mujer
Indígena



**EVALUACIÓN DEL COMITÉ TUTORIAL
POSGRADO EN CIENCIAS DE LA INGENIERÍA**

Ciudad Madero, Tam., a 14 de mayo de 2025

Nombre	Allan Mizael Jiménez Fernández		
Número de Control	G15070895	Semestre Inscrito	4to
Programa	Maestría	en Ciencias de la Ingeniería	

PROYECTO DE INVESTIGACIÓN			
Titulo (Dato obligatorio)	Estimación de la temperatura usando análisis multivariable		
Director(es) de Tesis	Dr. Juan Javier González Barbosa		
Porcentaje de avance anterior		Calificación anterior	

COMENTARIOS GENERALES			
* Recomiendo actualizar los periodos utilizados en tu serie, considerar hasta 2024 al menos. * Considerar incorporar la sintonización de parámetros. * Normalización $\Rightarrow 0$ a 1 puede provocar errores con algunas métricas de evaluación. * Mejorar la comparación con los trabajos del estado del arte. * Terminar la investigación.			
Porcentaje de avance actual	85%	Calificación del profesor	90

Ju. Castella V.
NOMBRE Y FIRMA DEL MIEMBRO DEL
COMITÉ TUTORIAL



2025
Año de
La Mujer
Indígena

Av. 1° de Mayo y Sor Juana I. de la Cruz S/N Col. Los Mangos
C.P. 89440 Cd. Madero, Tam. Tel. (833) 3574820 ext. 3110
e-mail: depi_cdmadero@tecnm.mx www.tecnm.mx
www.cdmadero.tecnm.mx





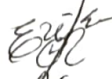
**EVALUACIÓN DEL COMITÉ TUTORIAL
POSGRADO EN CIENCIAS DE LA INGENIERÍA**

Ciudad Madero, Tam., a 14 de mayo de 2025

Nombre	Allan Mizael Jiménez Fernández		
Número de Control	G15070895	Semestre Inscrito	4to
Programa	Maestría	en Ciencias de la Ingeniería	

PROYECTO DE INVESTIGACIÓN			
Titulo (Dato obligatorio)	Estimación de la temperatura usando análisis multivariable		
Director(es) de Tesis	Dr. Juan Javier González Barbosa		
Porcentaje de avance anterior		Calificación anterior	

COMENTARIOS GENERALES			
* Actualizar los datos conforme a los resultados obtenidos en experimentaciones anteriores * En cuanto a los hiperparámetros no presenta la experimentación * En cuanto a la selección de variables no presenta los contenidos que se siguieron * Los resultados obtenidos con respecto al estado del arte no considerados sean susceptibles a poder compararse conforme a lo presentado, hacer énfasis en las variables, técnicas, parámetros de error. * Describir de manera técnica la metodología y alinearla con la presentación de resultados.			
Porcentaje de avance actual	85%	Calificación del profesor	92


Enika Alarcón Ruiz

NOMBRE Y FIRMA DEL MIEMBRO DEL
COMITÉ TUTORIAL



2025
Año de
La Mujer
Indígena

Av. 1° de Mayo y Sor Juana I. de la Cruz S/N Col. Los Mangos
C.P. 89440 Cd. Madero, Tam. Tel. (833) 3574820 ext. 3110
e-mail: depl_cdmadero@tecnm.mx www.tecnm.mx
www.cdmadero.tecnm.mx



NÚMERO DE CONTROL: D13071963

FECHA: 05 de Febrero de 2025

ESTUDIANTE: RIVERA VARGAS FRANCISCO

SEMESTRE: 2

CARRERA: DCI

RETÍCULA: 166(DING-2010-13)

ESPECIALIDAD:

CLAVE	MATERIA	CR	CALIFICACIÓN	TIPO DE EVALUACION	OBSERVACIONES
21661	PROY. DE INV. I	16	90	Ev.Ord.1ra	
21662	SEM PREDOCTORAL	8	90	Ev.Ord.1ra	

f87788049bce960f946ea201d3002f61

CERTIFICADO DIGITAL

Verifica su autenticidad en

<http://sii.cdmadero.tecnm.mx>



IMPRESO EL DIA 2025-08-18 A LAS 18:54:46 HRS

NÚMERO DE CONTROL: D13071963

FECHA: 10 de Junio de 2025

ESTUDIANTE: RIVERA VARGAS FRANCISCO

SEMESTRE: 3

CARRERA: DCI

RETÍCULA: 166(DING-2010-13)

ESPECIALIDAD:

CLAVE	MATERIA	CR	CALIFICACIÓN	TIPO DE EVALUACION	OBSERVACIONES
31661	PROY. DE INV. II	16	95	Ev.Ord.1ra	

dc8b806019cc45433be5a247e98315da

CERTIFICADO DIGITAL

Verifica su autenticidad en

<http://sii.cdmadero.tecnm.mx>



IMPRESO EL DIA 2025-08-14 A LAS 05:50:18 Hrs

ANEXO 2



EL TECNOLÓGICO NACIONAL DE MÉXICO
A TRAVÉS DEL CENTRO NACIONAL DE INVESTIGACIÓN
Y DESARROLLO TECNOLÓGICO

OTORGA EL PRESENTE

RECONOCIMIENTO

A

ALLAN MIZAE L JIMÉNEZ FERNÁNDEZ | JUAN JAVIER GONZÁLEZ
BARBOSA | ERIKA ALARCÓN RUIZ | JUAN FRAUSTO SOLÍS | MIRNA
PATRICIA PONCE FLORES

Por su destacada participación con la Contribución: ARTÍCULO PARA CONGRESO, en al
área: CIENCIAS DE LA COMPUTACIÓN, con el Título:

PREDICCIÓN MULTIVARIABLE DE LA TEMPERATURA USANDO APRENDIZAJE
PROFUNDO

presentada en el marco del 4.º Congreso Internacional de Tecnología y Ciencia
Aplicada (CITCA), celebrado en el Centro Nacional de Investigación y Desarrollo
Tecnológico del 21 al 23 de mayo de 2025.

Cuernavaca, Morelos, 23 de mayo de 2025

LORENZO O. HERNÁNDEZ
Director del CENIDET

MANUEL ADAM MEDINA
Coordinador General
4.º CITCA



Sello Digital:

F2OMO/KBoZnJ8OhpDMFKHm6Rd91+vmUm819RKrCmG72ayTa31r3EJIHMOSfNyrHth+xsVGd
2AC04UK+ckkHlwTxV+VoQn1YTmPMig1wO6sK5+U6Yva+fnDZUKjwz6vK5nFfPbBkRdYggw1V
fAUphrfkl+tg13gqpEwv5fXDfiy7BYT4YRcf06tqU3W81K3S96X5DGEW3x5+8xXn6ZUYvOMb
3AUs5dGVGgK3tIffLSlefziLayYM+phWMeQYnXGE37Pq2GQikfewTJN3GUuDv/vuWuDrvwB
UBFAoLz1fFXLxyUd7/ROIQqGIYFovoJJuBAQRQunziiEeqQNdqNzUA==

CITCA277A25 <http://constancias.cenidet.tecnm.mx>



Educación
Secretaría de Educación Pública



TECNOLÓGICO
NACIONAL DE MÉXICO®

EL TECNOLÓGICO NACIONAL DE MÉXICO
A TRAVÉS DEL CENTRO NACIONAL DE INVESTIGACIÓN
Y DESARROLLO TECNOLÓGICO

OTORGA EL PRESENTE

RECONOCIMIENTO

A

FRANCISCO RIVERA VARGAS | JUAN JAVIER GONZÁLEZ BARBOSA |
JUAN FRAUSTO SOLÍS | ERIKA ALARCÓN RUIZ | JOSE LUIS PURATA
ALDAZ

Por su destacada participación con la Contribución: ARTÍCULO PARA REVISTA, en al
área: CIENCIAS DE LA COMPUTACIÓN, con el Título:
FORECASTING AND OPTIMIZATION OF STOCK MARKET ASSETS USING
METAHEURISTICS AND DEEP LEARNING
presentada en el marco del 4.º Congreso Internacional de Tecnología y Ciencia
Aplicada (CITCA), celebrado en el Centro Nacional de Investigación y Desarrollo
Tecnológico del 21 al 23 de mayo de 2025.

Cuernavaca, Morelos, 23 de mayo de 2025

LORENZO O. HERNÁNDEZ
Director del CENIDET

MANUEL ADAM MEDINA
Coordinador General
4.º CITCA



Sello Digital:

M2O9pha5mMaeC6CsSNuwKrbzyqeCtERrVxGYX3cqZUhRwcgnAFsRTcrVpnyOUUp7gBrFk25E8
mXKORT406gPfJ07GiA/YPdgxD2czkmvYWbK5ftUxDS8Yy5adRUD1XdDnzw07ThXyegFH1KCL
pC1YxFLqjDor1eJpV+JDiHYFP1KVJAY2BKavyfC97A/VHcuG9sgzFKGHiEe5yAbUiHHdojPX
SWbKbUuZYUW5CGAALSeV5BBPa5ti5CKiGbs/GdvFv0zN7WJpmpde5t3W2ExedJBzq8bZKue
3bH11d0z4b+9UZpcxE7OZBEFgAbNkDaInzkNYRKgDAE+uFpe700aXQ==

CITCA254A25 <http://constancias.cenidet.tecnm.mx>



Multivariate temperature prediction using deep learning

Allan Mizael Jiménez Fernández*, Juan Javier González Barbosa*, Erika Alarcón Ruiz*, Juan Frausto Solís*, Mirna Patricia Ponce Flores*

**Tecnológico Nacional de México/Instituto Tecnológico de Ciudad Madero, Ciudad Madero, Tamaulipas, México.*

**G15070895@cdmadero.tecnm.mx, jjgonzalezbarbosa@hotmail.com, erika.ar@cdmadero.tecnm.mx, juan.frausto@gmail.com, mirna.pf@cdmadero.tecnm.mx*

Abstract: Climate change is a phenomenon that affects planet Earth and living beings, due to this, different types of challenges arise when facing its mitigation. One of these challenges is predicting temperature. To perform temperature prediction, it is necessary to select a machine learning technique such as neural networks, multiple linear regression, or support vector machines, among others. This work focuses on neural networks for use in different climatological time series from four cities in Mexico (Tampico, Tijuana, Monterrey, and Guadalajara). The time series contain data on the behavior of variables such as humidity, dew point, wind speed, and pressure over 11 years (2012-2022). These variables affect the temperature, which will be taken as the dependent variable to be forecasted. The result is a multivariate model that normalizes time series data using min-max scaling. A temperature prediction is then performed using a recurrent neural network (RNN) such as a long-short-term memory (LSTM) based on multivariate analysis. The purpose of this work is to contribute to climate change mitigation and provide a multivariate model for various problems encountered in different research areas. The results show satisfactory model performance, achieving mean absolute percentage error (MAPE) of less than 5.4% in all cities analyzed, validating the multivariate approach and the use of LSTM for climatological applications.

Keywords: Climate change, Temperature, Prediction, Multivariate, Deep learning, RNN, LSTM, Normalization.

1. INTRODUCCIÓN

El cambio climático es uno de los fenómenos más críticos del siglo XXI, afectando tanto a los ecosistemas como a las actividades humanas en muchos ámbitos y que provoca cambios de temperatura extremos, estos cambios tan significativos iniciaron desde que los seres humanos lograron industrializar los distintos procesos que hoy en día movilizan al mundo. En particular, la temperatura es una variable clave en los estudios climáticos, dado que su comportamiento influye directamente en sectores como la agricultura, la energía, el transporte y la salud pública.

La predicción de la temperatura y de las variables que afectan el cambio climático es de suma importancia ya que son utilizadas en la vida del ser humano, tiene aplicación en el desarrollo de áreas como agricultura, transporte, energía, turismo, etc. Así como también permite generar programas de mitigación para enfrentar el problema del cambio climático y tener una idea clara de lo que serán las temperaturas en un futuro.

El desarrollo de modelos predictivos basados en inteligencia artificial ha mostrado resultados prometedores, especialmente en lo que se refiere a series temporales multivariantes. Este trabajo presenta una propuesta de modelo basado en redes neuronales LSTM, entrenado con datos multivariantes de cuatro ciudades mexicanas, con el objetivo de generar pronósticos precisos de temperatura.

1.1 Estado del arte y antecedentes

A través del tiempo, el estudio de la predicción de la temperatura ha avanzado a pasos agigantados en búsqueda de obtener mejores resultados y un menor error en su pronóstico debido a la importancia que esta predicción tiene en las diferentes áreas donde se utiliza esta información diariamente.

En los trabajos previos se han realizado diversas predicciones de datos climatológicos, cada una de estas predicciones con diferentes configuraciones. Se han utilizado distintos métodos tanto de aprendizaje automático como de aprendizaje profundo. También en la selección de variables, pueden cambiar la importancia de las variables independientes debido a la posición geográfica o a la estación del año. Además, se puede incluir un modelo multivariable donde se utilizan distintas variables independientes para predecir la variable dependiente o un modelo univariable el cual trabaja con una variable.

El trabajo que fue hecho por Roy (2020) en la ciudad de Nueva York es un ejemplo de como se puede utilizar el aprendizaje profundo en la predicción de la temperatura del aire. En el trabajo obtuvieron un MAPE que oscila entre 10.85% - 2.58%. Utilizó las siguientes técnicas: Multi-Layer Perceptron (MLP), Long Short Term Memory Network (LSTM) y una combinación de Convolutional Neural Network (CNN) y LSTM. Las variables utilizadas fueron el promedio de la velocidad del viento, la lluvia, la caída de nieve, profundidad



de la nieve, el promedio de la temperatura, la temperatura máxima y la temperatura mínima.

Oshodi (2022) realizó un trabajo en la ciudad de Seattle donde obtuvo un MAPE aproximado del 15%. El proyecto desarrolla una investigación sobre algoritmos basados en aprendizaje automático para la predicción meteorológica donde propuso un modelo de pronóstico meteorológico basado en aprendizaje automático. El modelo sugiere implementar cuatro algoritmos de clasificación el bosque aleatorio, el árbol de decisión, el modelo bayesiano naive gaussiano y el de aumento de gradiente.

El trabajo realizado por Appau. (2023) en 5 ciudades de China (Beijing, Shanghai, Guangzhou, Chengdu, y Shenyang) sirve como antecedente en el uso del aprendizaje profundo, donde obtuvieron un MSE de 8.62 hasta 24.20 dentro de las observaciones de las 5 ciudades antes mencionadas. En la experimentación utilizaron redes neuronales recurrentes con un regularizador Ridge durante la etapa de entrenamiento y un optimizador de hiperparámetros como lo es la optimización bayesiana para predecir la temperatura en 5 ciudades de China. Las variables utilizadas fueron la temperatura, el punto de rocío, la humedad relativa, la presión y la velocidad del viento.

1.2 Justificación y contribución del trabajo

Dado que los modelos para predecir la temperatura pueden ser diversos, en este trabajo se propone un modelo multivariable utilizando una red neuronal recurrente como lo es Long Short-term Memory (LSTM) donde las variables independientes a utilizar sean las que tengan mejor desempeño en la etapa de selección de variable y buscando una configuración adecuada para los hiperparámetros de la red neuronal.

Todo esto con la finalidad de obtener el mínimo de error de porcentaje absoluto medio llamado (MAPE) en la predicción de la temperatura.

2. METODOLOGÍA

En el presente trabajo se hace uso del aprendizaje profundo mediante una red neuronal recurrente como lo es LSTM aplicando un modelo multivariable para predecir la temperatura. A continuación, se describe la metodología utilizada para desarrollar el modelo antes mencionado.

2.1 Series de tiempo

La base de datos utilizada en esta investigación son series de tiempo con datos climatológicos obtenidas en los sitios de internet weather underground (2023) y SINAICA (2023). Las series de tiempo describen el comportamiento de 5 variables climatológicas durante el periodo de enero del 2012 a diciembre del 2022 en las ciudades de Guadalajara, Monterrey, Tampico y Tijuana.

Tabla 1. Variables climatológicas

Variables
Temperatura
Punto de rocío
Humedad
Velocidad del viento
Presión

Todos los datos de las variables antes mencionadas fueron normalizados para facilitar el análisis e interpretación de la base de datos. El método de normalización utilizado es el de normalización Min-Max; mediante este método obtenemos un valor dentro del rango de 0 a 1 y su fórmula se describe como:

$$T \text{ normalizada} = \frac{T \text{ promedio} - T \text{ mínima}}{T \text{ máxima} - T \text{ mínima}} \quad (1)$$

Donde: T es temperatura, T promedio es la temperatura promedio obtenida de los valores de temperatura máxima y mínima de la base de datos, T mínima es la temperatura mínima absoluta observada en todo el conjunto de datos y T máxima es la temperatura máxima absoluta observada en todo el conjunto de datos.

Después de normalizar los datos, se divide la base de datos para la fase de experimentación, dichas series de tiempo fueron divididas en 3 etapas: entrenamiento, validación y prueba. Los porcentajes fueron distribuidos de la siguiente manera:

Tabla 2. Valores de cada fase en la experimentación

Etapas	Porcentaje
Entrenamiento	80%
Validación	10%
Prueba	10%

2.1 Modelo multivariable usando aprendizaje profundo

En el modelo propuesto se realiza una predicción de la temperatura utilizando todas las variables restantes de la base de datos, es decir, un modelo multivariable donde el punto de rocío, la humedad, la velocidad del viento y la presión son utilizadas para pronosticar la temperatura.

Considerar dos o más variables independientes beneficia al modelo ya que se realiza un análisis multivariable que es una técnica estadística que se utiliza para examinar y comprender las relaciones entre múltiples variables simultáneamente. A diferencia del análisis univariable, que se centra en una sola variable, el análisis multivariable considera la interacción entre dos o más variables

El objetivo principal del modelo multivariable es identificar patrones, tendencias y asociaciones complejas que no pueden ser descubiertas mediante el análisis de una sola variable.



Para obtener la predicción de la temperatura usando un modelo multivariable se utiliza una red neuronal recurrente (RNN) como lo es una LSTM (Long Short-Term Memory) que es utilizada para predecir series de datos. Gracias a que la LSTM puede aprender dependencias a largo plazo gracias a su arquitectura la hace especialmente útil para predecir series de datos temporales, y en este caso, para predecir series de datos climatológicos que fueron recopilados a través del tiempo.

En la metodología de este trabajo se entrena y se valida la red neuronal para pronosticar la etapa de prueba, tal como lo muestra la tabla 2, con los porcentajes correspondientes a cada etapa.

El pseudocódigo de la red neuronal es el siguiente:

```
model = Sequential()
model.add(LSTM(128,return_sequences=True,
input_shape=(X_combined_train_val.shape[1],
X_combined_train_val.shape[2])))

model.add(Dropout(0.2))
model.add(LSTM(128))
model.add(Dropout(0.2))
model.add(Dense(64, activation='relu'))
model.add(Dense(1))

model.compile(optimizer=Adam(learning_rate=0.001),
loss='mean_squared_error')

model.fit(X_combined_train_val, y_combined_train_val,
epochs=20, batch_size=32, verbose=1)
```

Donde los hiperparámetros utilizados se muestran a continuación:

Tabla 3. Valores de los hiperparámetros

Hiperparámetro	Valor
Número de capas	5
Units primera capa	128
Units segunda capa	128
Units capa densa	64
Return sequences 1ra capa	True
Return sequences 2da capa	False
Número de características por entrada	3
Dropout ambas capas	0.2
Función de activación	relu
Optimizador	Adam (0.001)

Épocas	20
Batch size	32
Units capa densa de salida	1
Time step	10

Para evaluar el modelo anteriormente descrito se utiliza la métrica de error MAPE (Mean Absolute Percentage Error) que nos ayuda a comprobar la precisión de modelos de predicción, como en análisis de series temporales y regresión. Esta métrica nos permite observar el error promedio entre los valores reales y los valores pronosticados como un porcentaje del valor real.

Una vez configurada la red neuronal y sus hiperparámetros, se introducen las series de tiempo de los datos de las variables climatológicas para obtener los resultados de un pronóstico de la temperatura con un análisis multivariable de las ciudades de Monterrey, Guadalajara, Tampico y Tijuana.

3.RESULTADOS Y/O DISCUSIÓN

Para evaluar el modelo se llevaron a cabo experimentaciones en cada ciudad donde se obtuvieron bases de datos con suficiente información de las variables independientes para pronosticar la temperatura.

Los resultados obtenidos son los siguientes:

3.1 Guadalajara

En la ciudad de Guadalajara se realizó la predicción de la temperatura utilizando la métrica de error MAPE para evaluar el modelo. Los resultados de la experimentación son los siguientes:

Tabla 4. Resultado del MAPE en Guadalajara.

Pronóstico LSTM Guadalajara		
Variable dependiente	Variables independientes	MAPE
Temperatura	- Punto de rocío - Humedad - Velocidad del viento - Presión	2.3133 %

Con estos resultados obtenemos la siguiente gráfica:

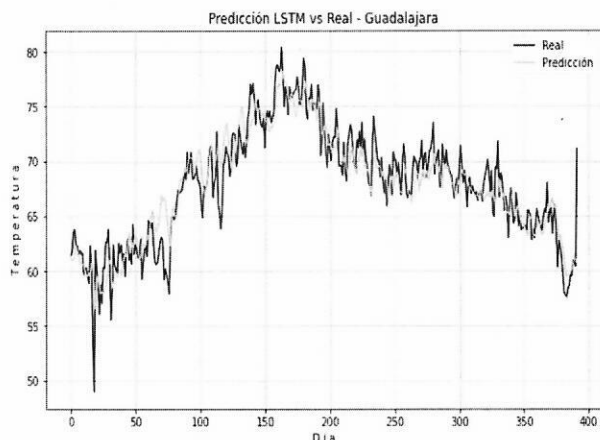


Fig. 1. Tendencia de los datos reales y datos pronosticados. En la ciudad de Guadalajara

La figura 1 muestra en el eje X los días y en el eje Y la temperatura. La gráfica indica la tendencia de las dos series de tiempo en la ciudad de Guadalajara, la serie de tiempo de los datos reales en color azul y la serie de tiempo de los datos pronosticados el color naranja.

3.2 Monterrey

En la ciudad de Monterrey se realizó la predicción de la temperatura utilizando la métrica de error MAPE para evaluar el modelo. Los resultados de la experimentación son los siguientes:

Tabla 5. Resultado del MAPE en Monterrey.

Pronóstico LSTM Monterrey		
Variable dependiente	Variables independientes	MAPE
Temperatura	• Punto de rocío • Humedad • Velocidad del viento • Presión	5.3553 %

Con los resultados de la tabla 5 podemos visualizar la gráfica siguiente:

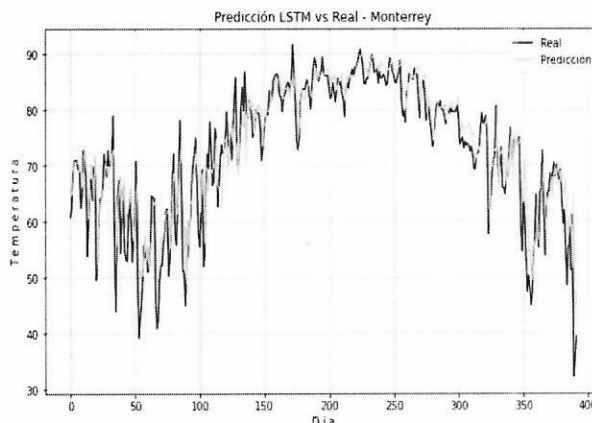


Fig. 2. Tendencia de los datos reales y datos pronosticados en la ciudad de Monterrey.

La figura 2 expresa la tendencia temporal de los datos reales y los datos pronosticados de la ciudad de Monterrey.

3.3 Tampico

En la ciudad de Tampico se realizó la predicción de la temperatura utilizando la métrica de error MAPE para evaluar el modelo. Los resultados de la experimentación son los siguientes:

Tabla 6. Resultado del MAPE en Tampico.

Pronóstico LSTM Tampico		
Variable dependiente	Variables independientes	MAPE
Temperatura	• Punto de rocío • Humedad • Velocidad del viento • Presión	3.1500 %

Con los resultados de la tabla 6 obtenemos la siguiente gráfica:

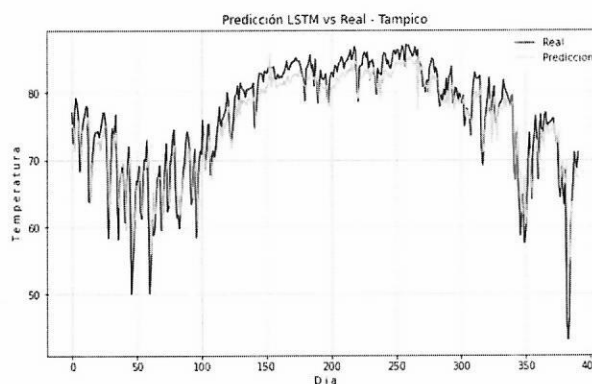


Fig. 3. Tendencia de los datos reales y datos pronosticados en la ciudad de Tampico.



La figura 3 ayuda a visualizar la tendencia de las series temporales reales y las series pronosticadas por la red neuronal en la ciudad de Tampico.

3.4 Tijuana

En la ciudad de Tijuana se realizó la predicción de la temperatura utilizando la métrica de error MAPE para evaluar el modelo. Los resultados de la experimentación son los siguientes:

Tabla 7. Resultado del MAPE en Tijuana.

Pronóstico LSTM Tijuana		
Variable dependiente	Variables independientes	MAPE
Temperatura	- Punto de rocío - Humedad - Velocidad del viento - Presión	3.5047 %

Al graficar la serie temporal real y la serie temporal pronosticada el resultado es el siguiente:

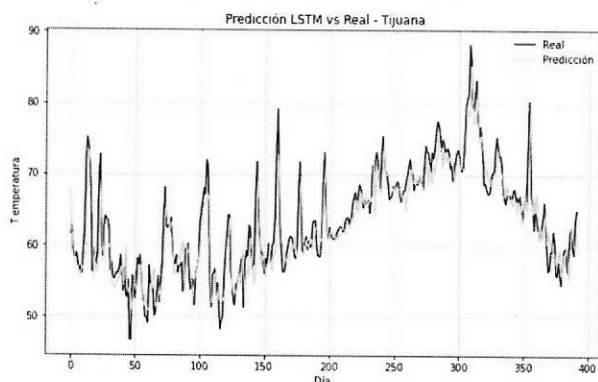


Fig. 4. Tendencia de los datos reales y datos pronosticados en la ciudad de Tijuana.

En la figura 4 se aprecia cómo se comportan las series temporales, tanto la serie real como la serie de datos pronosticada en la ciudad de Tijuana.

La tabla a continuación hace una comparativa del desempeño del modelo multivariable usando aprendizaje profundo en cada ciudad. El error fue medido con MAPE.

Tabla 8. Resultado del MAPE por ciudades.

Pronostico LSTM	
Ciudad	MAPE
Guadalajara	2.3133 %
Tampico	3.1500 %
Tijuana	3.5047 %
Monterrey	5.3553 %

El análisis muestra que el modelo LSTM multivariable puede adaptarse eficazmente a diferentes condiciones climáticas regionales. El bajo margen de error obtenido en todas las ciudades resalta la capacidad del modelo para capturar patrones complejos de temperatura a partir de variables meteorológicas asociadas.

Asimismo, los resultados sugieren que la inclusión de múltiples variables independientes mejora significativamente la capacidad predictiva respecto a modelos univariados tradicionales, al permitir una caracterización más rica del sistema atmosférico.

A continuación, una comparación de los resultados de las investigaciones encontradas en el estado del arte con el trabajo actual:

Análisis comparativo				
	Roy, D. S.	Oshodi, I.	Appau Nketiah et. al	Trabajo actual
MAPE	2.58 - 10.85%	15%	-	2.31 - 5.35%
MSE	-	-	8.62 - 24.20	4.09 - 22.96

Podemos observar que el desempeño del modelo en el trabajo actual es tan bueno como el desempeño de los modelos que se encuentran en el estado del arte.

6. CONCLUSIONES

El cambio climático plantea importantes desafíos para la gestión ambiental y la planificación territorial. Una de las herramientas más prometedoras en la mitigación y adaptación al cambio climático es la predicción precisa de variables atmosféricas, como la temperatura. En este estudio se presenta un modelo multivariable de predicción de temperatura basado en redes neuronales recurrentes del tipo LSTM (Long Short-Term Memory), aplicado a series de tiempo climatológicas de las ciudades mexicanas de Guadalajara, Monterrey, Tampico y Tijuana. Las variables predictoras incluyen punto de rocío, humedad relativa, velocidad del viento y presión atmosférica, recopiladas en un periodo de 11 años (2012–2022). Se empleó una estrategia de normalización min-max y se dividieron los datos en conjuntos de entrenamiento (80%), validación (10%) y prueba (10%). Los resultados muestran un rendimiento satisfactorio del modelo, alcanzando errores porcentuales absolutos medios (MAPE) inferiores al 5.4% en todas las ciudades analizadas, lo que valida el enfoque multivariable y el uso de LSTM para aplicaciones climatológicas.

REFERENCIAS

Appau Nketiah, E., Chenlong, L., Yingchuan, J. and Appah Aram, S. (2023) *Recurrent neural network modeling of multivariate time series and its application in temperature*



forecasting.

PLOS

ONE.

<https://doi.org/10.1371/journal.pone.0285713>

Oshodi, I. (2022) *Machine Learning-Based Algorithms for Weather Forecasting. International Journal of Artificial Intelligence and Machine Learning*, 2(2), pp. 12–20.
<https://doi.org/10.51483/ijaiml.2.2.2022.12-20>

Roy, D. S. (2020) *Forecasting the air temperature at a weather station using deep neural networks. Procedia Computer Science*, 170, pp. 517–522.
<https://doi.org/10.1016/j.procs.2020.11.005>

Weather Underground (n.d.) *Weather history and forecasting*. Available at: <https://www.wunderground.com/>

SINAICA (n.d.) *Sistema Nacional de Información de la Calidad del Aire*. Available at: <https://sinaica.inecc.gob.mx/>

ANEXO 3



www.editada.org

A Comparative Analysis of the Classical and Machine learning Forecasting Methods for the Mexican Stock Exchange

Juan Frausto-Solis¹, Javier Alberto Rangel-González,^{2,*} Erick Estrada-Patiño¹, Juan Javier González-Barbosa¹, Erika Alarcón Ruiz¹, Guadalupe Castilla-Valdez¹, Ocotlán Díaz-Parra³

¹Tecnológico Nacional de México/Instituto Tecnológico de Ciudad Madero, Ciudad Madero, Tamaulipas, México

²Universidad Autónoma de Tamaulipas, Morelos, México.

³Universidad Politécnica de Pachuca, Pachuca, México

E-mails: juan.frausto@gmail.com, javieralberto64@hotmail.com

Abstract. There is no recent comparison in the literature on the application of classical and machine learning methods for forecasting financial assets on the Mexican Stock Exchange (BMV). These methods divide the time series into three sections: training, validation, and testing. They predict future values using the training data and are evaluated in the validation phase using error metrics; once the lowest error is obtained, the best parameters and algorithms are used to predict values using the test section. This paper aims to find the most accurate regression algorithm to make predictions in the financial time series of the BMV. The regression methods compared include linear regression, neural networks, decision trees, and support vector regression. The study uses historical BMV asset price data to compare the accuracy of each of these algorithms.

Keywords: Statistical forecasting methods, Regression Algorithms, Mexican Stock Exchange, Machine learning forecasting.

Article Info

Received Jan 26, 2024

Accepted July 11, 2024

1 Introduction

Financial markets have been consolidated as elements of great importance for national economies, where companies and traders find business opportunities to increase their capital through stock trading (Roca et al, 2009) (Misra & Chaurasia, 2020). However, those who wish to perform their investment portfolios or purchase and sell assets require unavailable future stock prices; they should estimate them, and forecasting machine learning and statistical methods represent the best alternative to estimate them. The development of advanced technologies and the increased accessibility of financial data have boosted interest in trading-oriented forecasting techniques to reduce the uncertainty inherent in the sector due to fluctuating asset values, market volatility, and macroeconomic and environmental factors (Wasserbacher & Spindler, 2022) (Rouf et al, 2021). Not all assets exhibit the same level of volatility; some maintain specific stability, while others are highly volatile (Idrees et al, 2019). This variability highlights the importance of predicting their future behavior to make informed decisions with high confidence (Granger & Poon, 2024) (Xiao & Aydemir, 2007).

Historically, the behavior of assets has been represented by time series, which capture the asset's value at regular intervals (Weigend, 1994). The graphical representation of the time series allows efficient visualization of the asset's behavior. Furthermore, the numerical data can be used in time series forecasting models to estimate the future values of the stocks, which is of great importance for creating investment portfolios with favorable rates of return (Hyndman & Athanasopoulos, 2021). The most commonly used forecasting models are the classical or statistical models, such as linear regression, exponential smoothing, and autoregressive strategies; besides modern machine learning techniques, such as support vector machines, decision trees, nearest neighbor techniques, and neural networks are common (Dougherty, 2011) (Gerlein et al, 2016) (Makridakis et al, 2008). The later models generally take advantage of the possible data patterns present in the time series, such as the trend, which refers to the behavior of the series in the medium or long term and can be increasing, decreasing, or constant; its seasonality, which refers to constant repetitive patterns within the series; its cyclicity, which refers to very long-term behaviors attributable to

macroeconomic factors; and the residual component, which encompasses behaviors over which there is no control (Hyndman & Athanasopoulos, 2021). Understanding these components facilitates understanding the asset's behavior and enables more accurate predictions (Werner, 2016).

The Mexican Stock Exchange (BMV) is one of Mexico's leading financial institutions and is crucial in the domestic financial market. The BMV was established in 1894 to facilitate securities trading. It provides a transparent and efficient environment for purchasing and selling stocks, bonds, and other financial instruments. Its importance lies in channeling national savings into productive investment, boosting the country's economic growth. In addition, the BMV offers a wide range of products and services, including stock market indexes that reflect market performance and advanced tools for financial analysis. As a result, it is an essential point of reference for investors and analysts (Bolsa Mexicana de Valores, n.d.).

In the context of this paper, the selection of assets from the Mexican Stock Exchange enables us to assess the suitability and efficacy of diverse time series forecasting techniques in an emerging and dynamic market (San Martin-Reyna & Duran-Encalada, 2012). The BMV, with its rich data history and relevance in the Mexican economy, provides an ideal scenario to assess forecasting models and gain deeper insights into market dynamics that encompass both stability in specific sectors and high volatility in others (Herrera & Lockwood, 1995). These facts enhance the comparative analysis and provide valuable insights for investors seeking to optimize their trading strategies. This paper presents a comparative study of various time series forecasting techniques applied to predict the value of financial assets. Several assets of the Mexican Stock Exchange (BMV) were selected to analyze each method's capabilities, strengths, and limitations. The objective is to provide an overview of the performance of these methods using this market's data.

This paper is organized as follows: Section 2 introduces the theoretical framework of the related strategies; Section 3 provides a detailed description of the experimental process with the selected strategies; and Section 4 presents the conclusions obtained and a discussion of the work.

2 Related Methods

The following is an overview of the methods and strategies used for comparison in this work. We remark that there are many variants of each method in the literature, each successful for the type of problem addressed. However, it is not possible to establish a definitive superiority for all problems. Therefore, the approaches discussed are described in general terms. Although they have been optimized to produce the best possible results, other variants could improve or worsen the predictive value.

2.1 Time Series

In the forecasting area, following the traditional definition, a time series is defined as a set of one or more variables recorded at regular intervals (Hyndman & Athanasopoulos, 2021). The value that an asset has taken in a specific period of the time series was previously influenced by events that occurred before the said period and whose number of periods and variables affecting it is usually impossible to establish. Thus, all we can say is that the value that has taken an asset in a specific time series has been influenced by the event that occurred in a specific period. In forecasting financial assets, such as those traded on the Mexican Stock Exchange (BMV), these time series permit modeling the evolution of the value of these assets as a function of time (Osisanwo & Atanda, 2012).

In general, each variable in a time series can be decomposed into three main components: the trend, which represents the general direction in which the series moves over the long term; the seasonality, which reflects repetitive patterns over shorter periods; and the residual component, which captures the variability not explained by the trend and seasonality (Zarnowitz & Ozyildirim, 2006). Furthermore, in the case of long-term time series, a cyclical component may emerge that is not associated with seasonality (Harvey & Trimbur, 2003); instead, it may be linked to factors such as economic policies or long-term phenomena (Alexander, 2017).

The objective of a forecasting method is to fit the predicted values to the observed actual values in the given time series. A regression function typically represents the predictions to ensure that the fitted curve is as close as possible to the original curve of the time series; this enables the capture and learning of patterns and behaviors over time with the known data, which in turn makes it possible to produce accurate estimates of the future over a forecast horizon (De Gooijer & Hyndman, 2006). The similarity between regression functions is evaluated using error metrics, which quantify the discrepancy between predicted and

actual values (Naser & Alavi, 2023). In the context of this paper, the error metrics used to evaluate the quality of predictions are discussed in the paper.

2.2 Components in time series

Time series are a set of successive observations measured over time, with the characteristic that current observations directly depend on past observations (Araghinejad, 2014). Thus, time series do not present randomized behaviors but follow specific patterns that tend to oscillate over time. These patterns are trend, seasonality, and randomness handled here with residual techniques (Hyndman & Athanasopoulos, 2021). Graphically, we can observe the typical behavior of the components in Figure 1, where we also observe that these components are the product of the differentiation or decomposition of the original time series.

Trend refers to the long-term behavior or movement of the data, i.e., the general direction of the data, which can be increasing, decreasing, or in a direction with a slope close to zero. The trend allows us to understand the general pattern of the data over time (Hamilton, 1994).

Seasonality refers to movements or fluctuations that occur over a fixed period and are repeated within the series (Hyndman & Athanasopoulos, 2021). These variations can increase, decrease, or remain stable, and close prices are commonly used for predicting asset prices. In multivariable forecasting, several variables are considered to predict a single variable (Hamilton, 1994). Nevertheless, in the case of asset price prediction, the Market Efficient Hypothesis establishes that for predicting any asset price, all the information required is contained only in the time series of the asset. Thus, in this work, we only consider this variable (Fama, 1970). The component of the times series called residual is the value that does not adjust to seasonality or trend and is generally a product of the random nature of the data. This component captures irregularities and unforeseen variations that the other two components are unable to explain why these changes happen (Araghinejad, 2014).

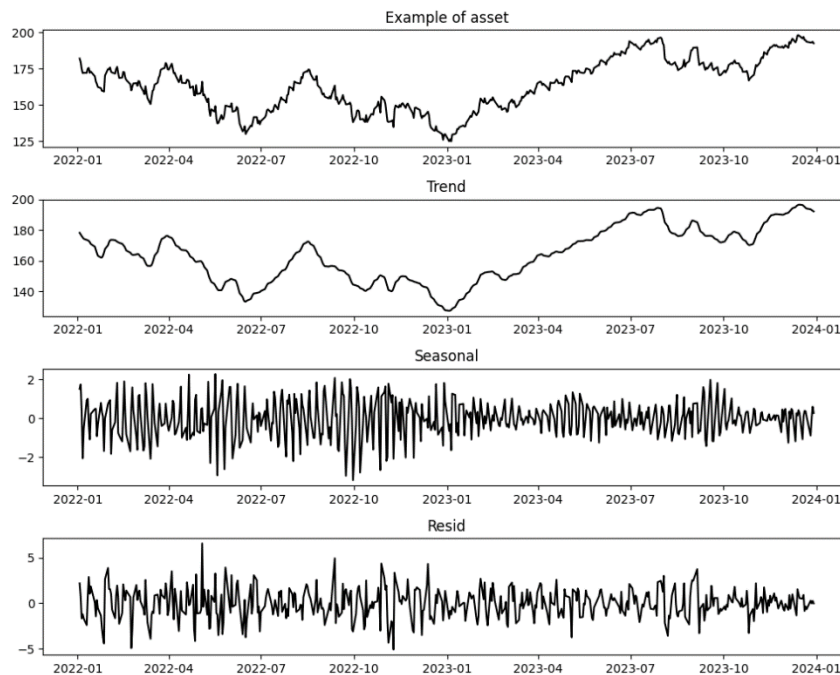


Fig. 1. Example of an asset taken randomly from Yahoo Finance

2.3 Representation of assets as a time series

Financial assets have characteristics that allow their representation through time series, which is applicable not only in the Mexican Stock Exchange (BMV), but in any asset exchange platform. The presentation of these data in a time format is fundamental for their analysis and for applying techniques to forecast their value. A time series is a sequence of values measured in continuous and uniform time intervals; for example, an asset can be represented through its price, trading volume, and derivative indexes in regular intervals such as every hour, every market close, and every week (Rapach & Zhou, 2022).

This temporal representation allows visualizing patterns and trends that are crucial for the development of predictive methods. Asset time series data includes vital information such as price movements, volatility, and cyclical behavior (Idrees et al, 2019). By analyzing these patterns, predictive models can identify trends and make inferences about future market movements. This analysis is essential for making informed investment and risk management decisions (Pincus & Kalman, 2004). In addition, the use of advanced time series analysis techniques and machine learning algorithms, such as autoregressive models, moving averages, recurrent neural networks, and regression methods, enhances the ability to transform historical data into powerful tools for predicting and anticipating the behavior of financial assets (Kumar, 2021).

2.4 Asset Forecasting Models

As mentioned above, there are different approaches to time series forecasting. Depending on the data set used, the methods can have varying performance; no universally superior method exists (Wolpert & Macready, 1997). A brief review of the forecasting methods selected for this work is presented below; these methods were chosen based on the strategies that showed the best average results and are widely studied in the literature.

2.4.1 Linear Regression

Linear regression is one of the most straightforward statistical strategies for modeling the relationship between input and output variables (Altay & Satman, 2005). This method allows an approximation of a linear function that is adjusted to minimize the distance between observed values and model predictions by reducing an error metric (Su et al, 2012). Generally speaking, simple linear regression is expressed by the equation $y = \beta_0 + \beta_1 x + \epsilon$ where y is the dependent variable, x is the independent variable, β_0 is the intercept or bias term, β_1 is the regression slope, and ϵ is the error or residual term that captures variations not explained by the model.

One of the main advantages of linear regression is its high interpretability. The β coefficients provide a direct measure of the expected change in the dependent variable per unit change in the independent variable, holding all other variables constant. However, the effectiveness of this model is compromised if the underlying relationships between the variables are not linear or if there is a high presence of outliers, which can lead to an incorrect fit of the data and inaccurate predictions.

2.4.2 Support Vector Regression

Support Vector regression machines (SVR) are a variant of support vector machines (SVM), which are widely used machine learning strategies (Drucker et al, 1996). The model was initially developed by Vapnik, and has been successfully tested in numerous state-of-the-art papers (Cortes & Vapnik, 1995). The SVR strategy generally attempts to fit a kernel to the regression line established by a known data set. As expected, this task does not always result in an exact fit, and it is necessary to determine confidence intervals or boundary areas bounded by an ϵ value that allows capturing observations outside the regression curve (Balabin & Lomakina, 2011). Values within the interval are accepted, and those outside are considered errors. The idea is to adjust the confidence intervals to small values to reduce the number of values outside the interval, thus reducing the error (Bathla, 2020).

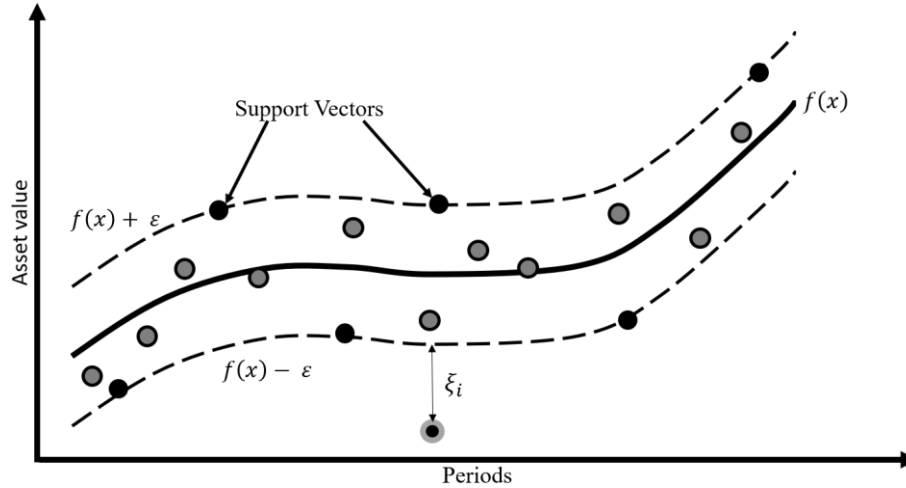


Fig. 2. Adjustment of SVR

In the context of machine learning, a hyperplane is the representation of a multidimensional space, which in SVM is used to separate sets or classes, seeking to maximize the distance between them and thus improve the accuracy of the model. SVR relies on constructing a hyperplane in which all values within a margin ε are accepted and, therefore, are not considered errors, thus allowing to approximate the best value within that margin, as seen in Figure 2.

$$y = \sum_{i=1}^N (a_i - a_i^*) \cdot K(x_i, x) + b \quad (1)$$

The standard function representing the prediction of an SVR model is given by equation 1, where y is the model prediction, N is the number of cases in the training set, while $(a_i - a_i^*)$ represent the Lagrange multipliers, which quantify the importance of each data point, looking for a simple and generalizable model; $K(x_i, x)$ are the kernel function where x_i is the feature vector for the i -th element, and x is the vector to forecast; finally, b is the bias value (Drucker et al, 1996).

Within the family of Support Vector Machine algorithms for regression, we find variants such as NuSVR (Nu Support Vector Regression), which incorporates a parameter ν . This parameter controls the fraction of data points used as support vectors, replacing the traditional ε value. This strategy allows for an easier-to-interpret model and, to some extent, greater flexibility. However, it can be susceptible to outliers or very noisy data, which could negatively affect the performance (Bathla, 2020).

Linear SVR is another variant of the traditional strategy, designed exclusively to model linear relationships between model variables. It uses a linear hyperplane that minimizes the sum of absolute errors between model predictions and actual output values. This model is usually efficient in terms of processing and interpretation. However, modeling exclusively linear relationships may not be suitable if the variables have nonlinear relationships or high dimensionality, limiting its ability to capture the complexity of the data set (Ho & Lin, 2012).

2.4.3 Decisions trees

Decision trees are a widely used strategy in machine learning, proving effective in classification and regression problems (De Ville, 2013). This model is based on constructing one or more trees that divide a training data set, using the variable that provides the greatest information gain at each step (Ali et al, 2012).

The structure of decision trees is highly interpretable, as it can be visualized in a clear and understandable way. In addition, each tree can be interpreted as a set of decision rules based on the input variables.

Generally, decision trees start with a root node containing the entire data set. The model evaluates all input variables from this node to determine which best divides the data set. This iterative process continues until a predetermined stopping condition is met, which can be a minimum number of samples per node or the maximum depth of the tree. This approach allows the problem

to be broken down into more straightforward and manageable decisions, thus facilitating the interpretation and implementation of the model (Apté & Weiss, 1997).

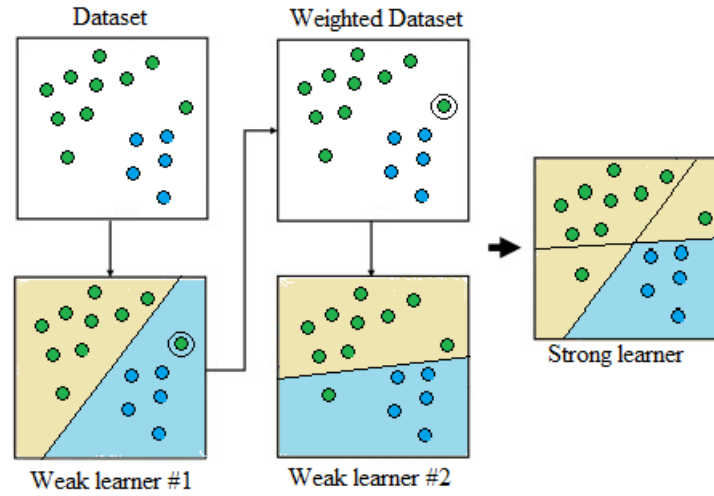


Fig. 3. Principle of AdaBoost algorithm (Imtiaz et al, 2020).

Despite the success of this model, decision trees can be prone to overfitting, mainly when applied to non-representative data sets or when the structures are too complex (Mazumder et al, 2022). To mitigate these problems, it is essential to thoroughly analyze the data and employ pruning techniques that trim unnecessary branches from the tree. In addition, ensemble approaches, such as random forests or boosting, can be used, which combine multiple decision trees to improve the accuracy and robustness of the model. These strategies help to avoid overfitting and ensure that the model generalizes well to new data. In Figure 3, the generic process for AdaBoost is presented. Different strategies, both individual and ensemble or hybridization techniques, are derived from this approach. In this work, the boosting strategies of AdaBoost, an adaptive ensemble technique that improves the efficiency of individual decision trees, are used. AdaBoost applies a sample-weighting approach, focusing on the predictions with the highest error. This approach allows the model to generate a new, improved binary tree, sequentially building trees that correct the mistakes of the previous ones to obtain a better forecast result (Ying et al, 2013).

AdaBoost generally starts by assigning equal weights to all samples in the training dataset. The total number of samples dictates the type of AdaBoost strategy used and how the decision trees are trained; samples that are incorrectly classified are given greater weight in subsequent iterations. In this way, AdaBoost adjusts its approach to focus on the most difficult-to-predict cases, gradually improving the model's accuracy. This iterative process continues until a predefined number of trees is reached or until the prediction error reaches an acceptable threshold. Thus, AdaBoost generates a robust and accurate model and maintains the high interpretability characteristic of decision trees since each tree in the sequence can be visualized as a set of decision rules based on the input variables (Schapire, 2013).

2.4.5 *K-nearest neighbors*

The k-nearest neighbors (KNN) algorithm is a widely used supervised classification algorithm applied in tasks such as data preprocessing, recommendation engines, finance, healthcare, and pattern recognition (Cover & Hart, 1967). Unlike many machine learning algorithms, KNN does not have an explicit learning phase; instead, it performs its prediction mechanism using the entire available dataset (Guo et al, 2003). Unlike many machine learning algorithms, KNN does not have an explicit learning phase; instead, it performs its prediction mechanism using the entire available dataset (Guo et al, 2003). Its strategy is based on finding similarities between variables in the hyperplane, considering that close values belong to the same neighborhood and, therefore, to the same class or value. Thus, the number of neighbors for searching must be determined, and the neighbors with the shortest distance to the target value should be found. Equations 3, 4, and 5 show the calculation functions for Euclidean, Manhattan, and Minkowski distances, respectively.

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^k (p_i - q_i)^2} \quad (3)$$

$$d(p, q) = \sum_{i=1}^k |p_i - q_i| \quad (4)$$

$$d(p, q) = \left(\sum_{i=1}^k (|p_i - q_i|)^p \right)^{1/p} \quad (5)$$

where $d(p, q)$ is the distance between the n -dimensional points p and q , k is the number of coordinates, p_i and q_i represent the i -th coordinates, and the parameter p defines the order of the distance. Once these K neighbors are obtained, the class is chosen through a weighting function. This technique is known for its simple and intuitive implementation. However, predictions can be computationally expensive, especially when the dataset is large. Moreover, KNN can be highly sensitive to outliers and noise, affecting the model's accuracy if the data has not been adequately cleaned (Zhang, 2021).

There are improvements and variants to address some KNN limitations. A common strategy is weighted KNN, which assigns weights to neighbors based on their distance to the data point, giving more importance to closer neighbors. Despite its limitations, KNN is a powerful and versatile algorithm with applications in various fields. Its simplicity and effectiveness, in many cases, make it a valuable tool. Using enhancements and optimizations, KNN can be adapted to handle a wide range of classification and regression problems, providing accurate and valuable results in diverse contexts (Pérez-Ortega et al, 2022) (Pérez-Ortega et al, 2024).

2.4.6 Multilayer Perceptron

The MLP (Multi-Layer Perceptron) regression algorithm is a learning model that employs several hidden layers to process the input data and generate a continuous output (Haykin, 2000) (Stathakis, 2009). The structure of the MLP model consists of multiple layers of interconnected neurons, including an input layer, one or more hidden layers, and an output layer. In the input layer, data are received and distributed to the neurons of the first hidden layer. This initial layer does not perform significant transformations on the data. Still, it gives the foundation for further processing (Popescu, 2009). The hidden layers are crucial in the MLP's ability to learn complex patterns. An activation function is applied on each neuron in these layers, which introduces nonlinearity into the model and allows it to capture more complex relationships (Windeatt, 2006). Among the most common activation functions are the sigmoid function, the hyperbolic tangent (tanh), and the rectified linear unit (ReLU). These functions transform the inputs' weighted sum into an output passed to the next layer. The sigmoid function, for example, maps the inputs to a range between 0 and 1, which is helpful for modeling probabilities. The tanh function maps the inputs to a range $[-1, 1]$, centering the data and thus facilitating learning in deep networks. The ReLU function, meanwhile, introduces computational simplicity by keeping only positive values and setting negative values to zero, enabling faster and more efficient training of deep models (Rasamoelina et al, 2020).

The MLP output layer produces the final model output, which is a continuous value in the case of regression. Depending on the problem, this layer may contain a single neuron (for single-valued predictions) or multiple neurons (for multivariate predictions). MLP training aims to adjust the weights of all neural connections by minimizing the error between the model predictions and the actual output values. This optimization process is performed using algorithms such as gradient descent, which adjusts the weights based on the error calculated at each iteration. In addition, more advanced variants of gradient descent, such as stochastic gradient descent (SGD) and Adam, can be employed, which improves the efficiency and speed of training.

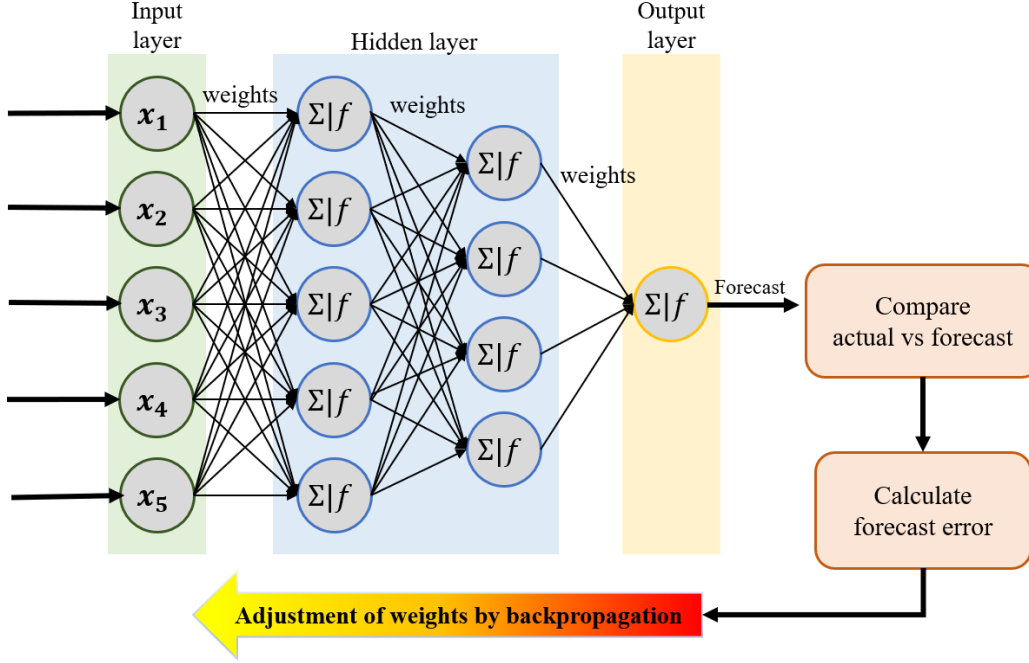


Fig. 4. MLP neural network structure for regression

In Figure 4, we represent how the input data is propagated through the hidden layers and how the activation functions and the adjusted weights influence the generation of the final output. This visualization can also highlight the complete connectivity between neurons in adjacent layers, showing how each neuron in one layer is connected to all neurons in the next layer, enabling feature interdependence during the learning process.

During training, the MLP adjusts the weights through an iterative process of error backpropagation. This process involves calculating the error at the output layer and propagating it backward through the network, adjusting the weights accordingly to reduce the error. Backpropagation is combined with the optimization algorithm to adjust the weights efficiently (Hecht-Nielsen, 1992). This method allows the MLP to learn from the training data and improve its generalization ability to make accurate predictions on new data. The MLP's capabilities to capture complex, nonlinear relationships make it especially useful in many applications, including forecasting, where prediction accuracy is essential. Its flexible structure and the ability to adjust multiple hyperparameters, such as the number of hidden layers, the number of neurons in each layer, and activation functions, allow researchers and practitioners to tailor the model to their specific needs and continuously improve its performance.

2.4.7 Partial least squares

The least squares technique is a statistical strategy widely applied in economics, engineering, data analysis, and many others. Its objective is to find the best approximation of a linear model to a set of observed data (Pirouz, 2006).

$$E(a, b) = \sum_{i=1}^n (y_i - (a \cdot x_i + b))^2 \quad (6)$$

where $E(a, b)$ is the sum of squared errors, y_i represents the observed values of the dependent variable, x_i contains the observed values of the independent variable, a is the slope, and b is the y-intercept of the regression line. This is achieved by minimizing the sum of the squares of the differences between the observed values and the values predicted by the model. Thus, given the set (x_i, y_i) , we seek to find the values a and b on the straight line that minimizes Equation 6 (Cha, 1994). A variant used in this work is partial least squares (PLS), which is used to find linear relationships when the data are highly collinear, finding a data projection that maximizes the covariance between the set of independent and dependent variables. On the other hand, PLS-Canonical seeks to find linear relationships between two sets of multivariate variables, unlike the standard strategy, which maximizes the covariance between two blocks of independent and dependent variables simultaneously (Esposito Vinzi & Russolillo, 2013).

2.4.8 Passive-aggressive algorithm

The Passive-aggressive (PA) algorithm is a powerful machine learning technique designed to minimize model loss by leveraging a unique loss function tailored to the specific learning problem. This approach is particularly effective when the model needs to adapt quickly to new data. When the model makes an accurate prediction, its parameters remain largely unchanged, preserving the knowledge it has already acquired. Conversely, when the model errs, it adjusts its parameters in a direction that compensates for the mistake, thereby refining its accuracy over time. This dual response mechanism—staying passive for correct predictions and becoming aggressive for incorrect ones—enables the PA algorithm to maintain a balance between stability and adaptability (Chang et al, 2010). One of the key strengths of the PA algorithm lies in its simplicity and efficiency. Unlike other machine learning methods that may require extensive computational resources or complex parameter tuning, the PA algorithm operates with straightforward update rules. This makes it particularly suitable for real-time applications where rapid decision-making is crucial. Additionally, the algorithm's ability to handle binary and multi-class classification tasks broadens its applicability across various domains, from spam detection to financial forecasting. By continuously adjusting its parameters in response to errors, the PA algorithm ensures that it remains relevant and accurate, even as the underlying data distribution evolves.

2.4.9 Lasso

Lasso is a regularization technique that effectively selects a subset of essential predictor variables by penalizing the regression coefficients based on their magnitude. By applying this penalty, Lasso encourages the model to shrink some coefficients to zero, thereby excluding less essential variables from the model. This reduction in model complexity helps to avoid overfitting, ensuring that the model generalizes well to new, unseen data. The simplicity and interpretability of the resulting models make Lasso particularly valuable in fields where understanding the underlying relationships between variables is crucial (Ranstam & Cook, 2018). Moreover, Lasso's ability to perform feature selection within the modeling process sets it apart from other regularization methods. Lasso simplifies the model and enhances its predictive performance by automatically identifying and retaining only the most relevant predictors. This is especially beneficial in high-dimensional datasets where the number of predictors exceeds the number of observations. The penalization mechanism not only improves model accuracy but also aids in uncovering the most significant factors influencing the response variable. Consequently, Lasso remains a popular choice for regression analysis, offering a balanced approach to managing model complexity and interpretability (Alhamzawi & Ali, 2018).

3. Experimental Process

The conditions and results of the experimental process are presented below, detailing the obtaining of the instances, the tests with algorithm implementations, the error metrics used and the interpretation of the results.

3.1 Data collection

The assets used in this study belong to the Mexican Stock Exchange (BMV). A sample of 15 assets was taken to evaluate the performance and results of each of the algorithms against the time series of these companies. The selected assets were: ACTINVRB.MX, ALPEKA.MX, ALSEA.MX, AMXL.MX, ASURB.MX, AZTECACPO.MX, FIBRAMQ12.MX, FSHOP13.MX, GBMO.MX, GIGANTE.MX, HOMEX.MX, LABB.MX, PINFRA.MX, POCHTECB.MX and URBI.MX.

The values of these assets were downloaded from Yahoo Finance, covering a 24-month period, from 08/09/2020 to 07/09/2022. In this way, we have 505 data items for each asset in their respective time series. The training phase was performed using the first 485 elements of each series, while the time window for prediction corresponds to the remaining 20 values of each series. This approach allowed the evaluation of the predictive capability of the algorithms on a data set not seen during training, providing a more robust assessment of their performance. The selection of the 15 assets was based on their diversity in terms of industry and market capitalization, aiming to capture a broad spectrum of market behavior. The chosen period from 08/09/2020 to 07/09/2022 included significant market events, such as the recovery phase following the initial impact of the COVID-19 pandemic, which added an extra layer of complexity to the prediction task. Overall, this study contributes to understanding how different predictive models perform in the context of financial time series data, offering valuable insights for investors and analysts in developing more effective forecasting strategies.

3.2 Implementation of the forecasting models

The algorithms used to generate the predictions of each time series were diverse, including machine learning and advanced statistical techniques. These include linear regression, SVR with its variants NuSVR and linear SVR, MLP Regressor, decision tree, and a combination of decision tree with the AdaBoost method using two different configurations for the estimator parameter, with values of 50 and 300. Additionally, Bayesian regression, nearest neighbors, PLS regression, PLS canonical, passive-aggressive regression, and Lasso were included, totaling 15 different algorithms.

The experimentation was conducted using the Python language and the scikit-learn library, which provides a wide range of tools and functions for implementing and evaluating machine learning models. This approach allowed for efficient and reproducible implementation of the algorithms, allowing us to compare their performance in predicting the time series of the selected assets. These algorithms were selected based on their ability to handle time series data and their prevalence in previous financial prediction studies. Multiple experiments were conducted to tune and validate the models, ensuring each was optimized to deliver its best performance. Furthermore, cross-validation techniques were used to assess the models' robustness by splitting the data into multiple subsets to repeatedly train and test the algorithms. This helped mitigate the risk of overfitting and provided a more accurate evaluation of their predictive capabilities. The results obtained from these experiments provide valuable insights into the relative performance of each algorithm but also highlight the importance of selecting the appropriate model based on the specific characteristics of financial data. Therefore, this study offers a valuable resource for researchers and professionals interested in time series prediction in financial markets.

3.3 Error Metric Algorithms in Time Series Forecasting

MAPE measures the prediction errors in percentage, providing an easy interpretation of the model's accuracy.

$$MAPE = \frac{100}{N} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (7)$$

Equation 7 shows the MAPE formula, where y_i represents the actual values, $\hat{y}_i$ the predicted values, and N is the total number of observations. The advantage of MAPE is its ability to be interpreted intuitively, allowing the performance of different models to be compared in relative terms. Moreover, being a standardized metric, it facilitates the comparison of errors between series of different magnitudes. However, MAPE can be influenced by real values close to zero, which can distort its measurement. Despite this limitation, MAPE is widely used in the literature and practical applications due to its simplicity and ease of interpretation.

3.4 Experimentation

The comparison metric between the algorithms was MAPE. In the first phase of the experimentation, 15 assets were selected out of the 76 corresponding to the Mexican Stock Exchange, and the data were cleaned, resulting in a time series of 505 items. The regression algorithms were implemented, and the MAPE was calculated to measure the performance of each algorithm. Tables 1 to 3 show the results of the algorithms, with first and second places shaded in blue and orange, respectively.

Table 1. MAPE results for different regression algorithms

Assets	Lineal Regression	SVR	NuSVR	Linear SVR	MLP
ACTINVRB.MX	0.00536	0.0119232	0.0081861	0.1326803	0.0450910
ALPEKA.MX	0.0257636	0.0613773	0.0704218	0.0185813	0.0192532
ALSEA.MX	0.2954848	0.0185516	0.0176108	0.0198921	0.3018038
AMXL.MX	0.1760139	0.0236742	0.0257789	0.0625258	0.2180682
ASURB.MX	0.0222357	0.1152942	0.1072539	0.0690743	0.0402737

AZTECACPO.MX	0.3766623	0.0673360	0.1709530	0.4937773	0.3411057
FIBRAMQ12.MX	0.0571168	0.0445952	0.0355279	0.1928604	0.0502488
FSHOP13.MX	0.0150617	0.0085153	0.0087321	0.0969803	0.0643222
GBMO.MX	0.3822901	0.0930643	0.0607561	0.7910832	0.3003851
GIGANTE.MX	0.0862386	0.0446353	0.0491523	0.2464847	0.0802588
HOMEX.MX	0.4170411	0.6820765	0.0924926	0.4358983	0.6844230
LABB.MX	0.1758850	0.1544537	0.1405102	0.6620405	0.1864919
PINFRA.MX	0.0174443	0.0225572	0.0191010	0.0176238	0.0408576
POCHTECB.MX	0.0888154	0.0177750	0.0157200	0.9597231	0.0123351
URBL.MX	0.2033036	0.1774260	0.1618387	0.5173878	0.0625209

Table 2. MAPE results for decision trees and Bayesian regression

	Decision Tree	Decision Tree AB300	Decision Tree AB50	Decision Tree AB	Bayesian Regression
ACTINVRB.MX	0.0043461	0.0047185	0.0047185	0.0056505	0.0053580
ALPEKA.MX	0.0306404	0.0325426	0.0325426	0.0352677	0.0256860
ALSEA.MX	0.0373365	0.0453149	0.0361077	0.0454955	0.2952312
AMXL.MX	0.0334045	0.0333927	0.0334257	0.0344670	0.1759125
ASURB.MX	0.0536812	0.0563681	0.0563681	0.0524089	0.0221853
AZTECACPO.MX	0.1132440	0.1086410	0.1000998	0.1106137	0.3763246
FIBRAMQ12.MX	0.0355879	0.0330402	0.0321266	0.0335353	0.0574206
FSHOP13.MX	0.0075302	0.0081633	0.0077545	0.0181335	0.0149623
GBMO.MX	0.0037096	0.0053010	0.0053010	0.0058177	0.3817450
GIGANTE.MX	0.0185030	0.0156415	0.0156415	0.0156415	0.0859883
HOMEX.MX	0.0438634	0.1104971	0.1104971	0.0739705	0.4170404
LABB.MX	0.0317674	0.0315030	0.0311961	0.0312173	0.1753393
PINFRA.MX	0.0204846	0.0200824	0.0200824	0.0217340	0.0176023
POCHTECB.MX	0.0199644	0.0310731	0.0310731	0.0327981	0.0891546
URBL.MX	0.0000000	0.0000000	0.0000000	0.0000000	0.2017392

Table 3. MAPE results for other regression algorithms

	KNN	PLS Regression	PLS Canonical	Passive Aggressive	Lasso
ACTINVRB.MX	0.0062669	0.0053634	0.0093543	0.0055041	0.0051448
ALPEKA.MX	0.0319678	0.0257636	0.0474303	0.2550770	0.0253920
ALSEA.MX	0.0659569	0.2954848	0.3523334	0.4080364	0.2951424
AMXL.MX	0.0324556	0.1760139	0.1996637	0.1148840	0.1753052
ASURB.MX	0.0536812	0.0222357	0.0390828	0.0168124	0.0222203

AZTECACPO.MX	0.1237246	0.3766623	0.5035424	0.4782619	0.3611923
FIBRAMQ12.MX	0.0355879	0.0571168	0.0224625	0.0552952	0.0576214
FSHOP13.MX	0.0099524	0.0150617	0.1076158	0.0321609	0.0130257
GBMO.MX	0.0035219	0.3822901	0.4963385	0.1854387	0.3812948
GIGANTE.MX	0.0160628	0.0862386	0.1358957	0.2116499	0.0857482
HOMEX.MX	0.0360502	0.4170411	1.0165530	1.0000000	0.3229325
LABB.MX	0.0432522	0.1758850	0.2604541	0.1309772	0.1751085
PINFRA.MX	0.0193715	0.0174443	0.0576212	0.0308392	0.0174646
POCHTECB.MX	0.0114438	0.0888154	0.0181548	0.3966782	0.0902525
URBL.MX	0.0711590	0.2033036	0.4803909	1.8495231	0.2015713

When executing the algorithms for different time series, we rank the algorithms by the number of times they finish in first and second place; the results are shown in Table 4, where the five best algorithms found can be seen: first, the decision tree with AdaBoost followed by KNN. The performance of the decision tree algorithm with AdaBoost for the assets GBM Corporative (GBMO.MX) and Fibra Shop (FSHOP13.MX) is shown in Figures 5 and 6, respectively.

Table 4. Best Performing Algorithms

Algorithm	Ranked first
Decision Tree AB50	6
KNN	3
PLS Regression	2
SVR	2
Decision Tree	2

The performance of the decision tree algorithm with AdaBoost for the assets GBM Corporative (GBMO.MX) and Fibra Shop (FSHOP13.MX) is shown in Figures 5 and 6, respectively. In Figure 5, we can observe that the algorithm adapts quite well when there is a stationary trend. However, in Figure 6, where volatility is high, the algorithm tends to fail at the steepest peaks, which increases the calculated error.

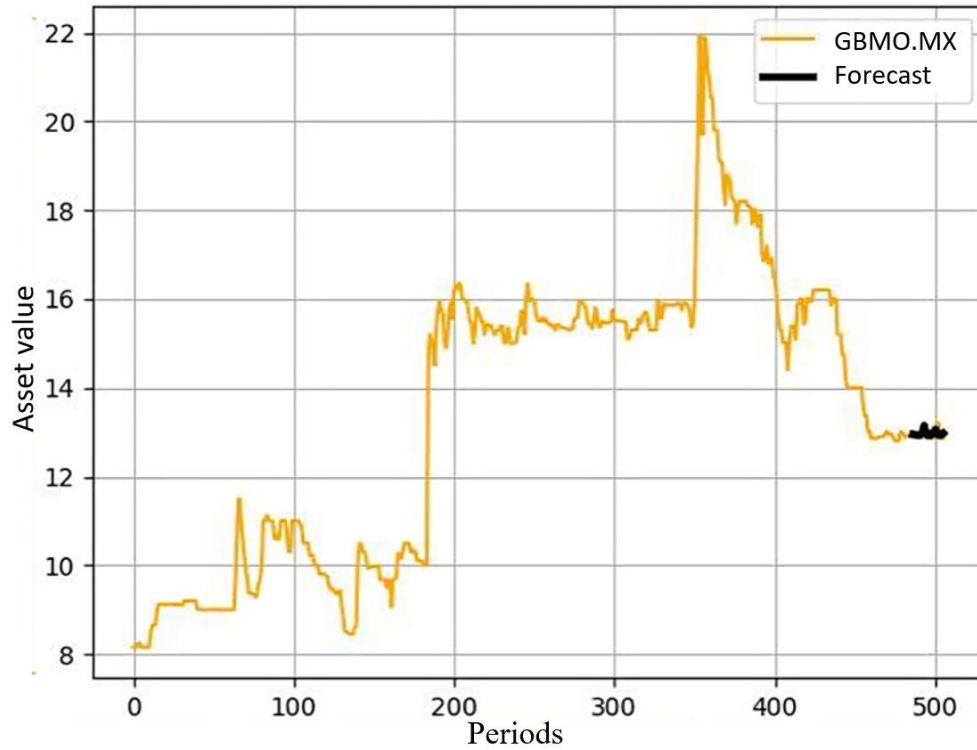


Fig. 5. Performance of GBM Corporative (GBMO.MX)

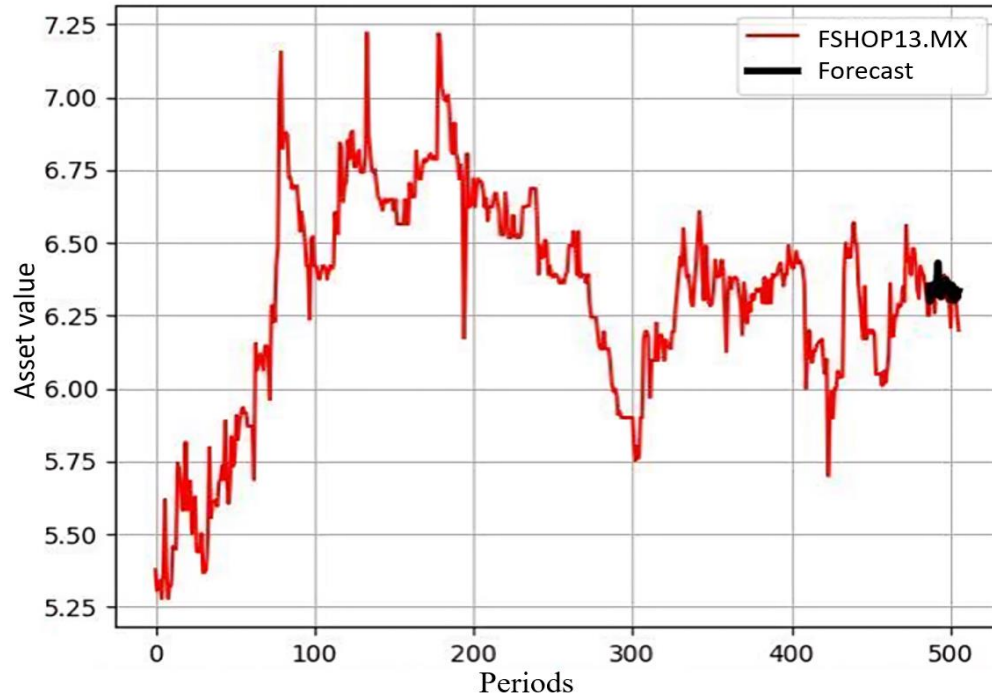


Fig. 6. Performance of Fibra Shop (FSHOP13.MX)

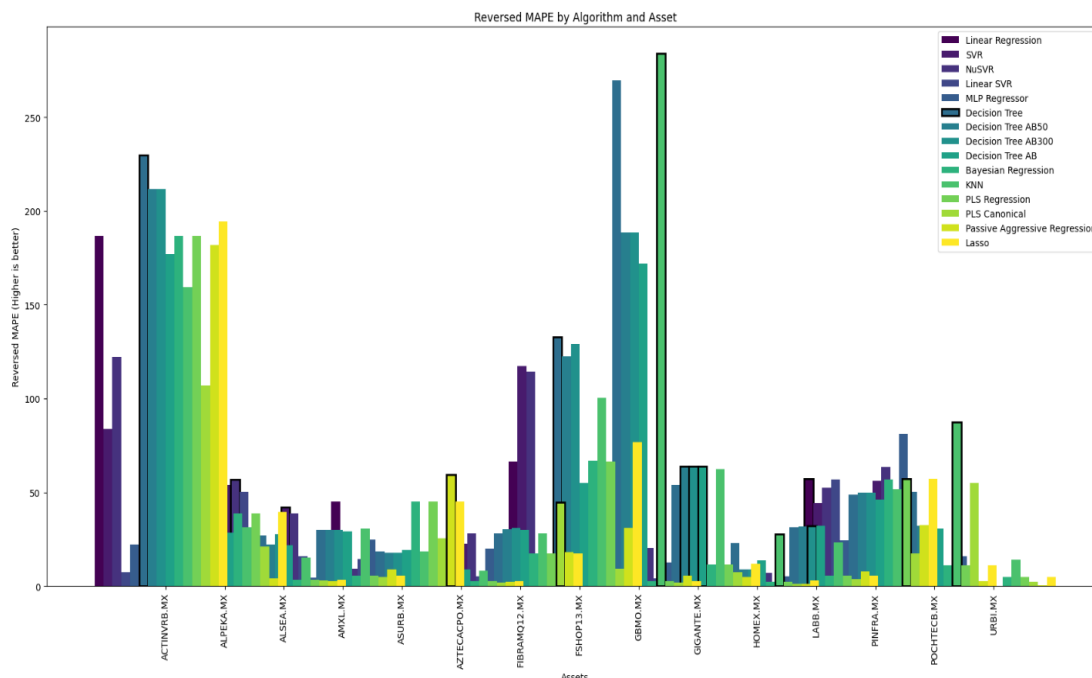


Fig. 7. Reversed MAPE by each algorithm and asset

Figure 7 shows the performance of various machine learning algorithms in terms of the inverse MAPE for different financial assets. The algorithms evaluated include linear regression, support vector machines (SVR), Bayesian regression, decision trees, and partial least squares regression. In the graph, those algorithms with the lowest MAPE for each asset are highlighted with black edges, indicating the most efficient methods for each case. The inversion of the MAPE values allows the most efficient algorithms to be visually highlighted since those with a lower MAPE are shown at the top of the graph. This facilitates the identification of the most accurate methods, providing a clear comparison between algorithms for each asset. Overall, this approach allows a more intuitive and quicker interpretation of the results, highlighting the best options for prediction in the financial domain; with these results, we can conclude that the decision tree algorithm with AdaBoost showed the best performance in terms of MAPE compared to other algorithms, especially in time series with lower volatility.

4 Conclusions

This work compares the classic and machine learning forecasting methods applied to the Mexican Stock Exchange or BMV). The characteristics, advantages, and limitations of each approach were analyzed. While classical forecasting or statistical methods have been used for decades to model and predict time series in financial markets, machine learning methods are more recent and used to capture complex patterns in large volumes of data. The experimentation presented the obtaining of the instances, the tests with the implementations of the algorithms, the error metrics used, and the interpretation of the results. The data sets used correspond to stocks of the Mexican Stock Exchange (BMV) and consist of time series with a temporal format that is essential for their analysis and for applying techniques that allow forecasting their value. Fifteen regression models were selected and implemented to predict the assets of the selected data sets. It was determined which algorithms have a better performance for the set of evaluated instances, highlighting the decision tree with AdaBoost in the first place. In future work, the implementation of an automatic parameterization algorithm using self-generated fuzzy logic will be proposed.

References

- Alexander, N. (2017). Factors affecting earnings management in the Indonesian Stock Exchange. *Journal of Finance and Banking Review*, 2(2), 8-14.
- Alhamzawi, R., & Ali, H. T. M. (2018). The Bayesian adaptive LASSO regression. *Mathematical Biosciences*, 303, 75-82.
- Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, 9(5), 272.

- Altay, E., & Satman, M. H. (2005). Stock market forecasting: Artificial neural network and linear regression comparison in an emerging market. *Journal of Financial Management and Analysis*, 18(2), 18.
- Apté, C., & Weiss, S. (1997). Data mining with decision trees and decision rules. *Future Generation Computer Systems*, 13(2-3), 197-210.
- Araghinejad, S. (2014). Time series modeling. En *Data-driven modeling: Using MATLAB® in water resources and environmental engineering* (pp. 85-137). Springer Netherlands. https://doi.org/10.1007/978-94-007-7506-0_4.
- Balabin, R. M., & Lomakina, E. I. (2011). Support vector machine regression (SVR/LS-SVM)—an alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near-infrared (NIR) spectroscopy data. *Analyst*, 136(8), 1703-1712. <https://doi.org/10.1039/C0AN00387E>.
- Bathla, G. (2020). Stock price prediction using LSTM and SVR. En *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)* (pp. 211-214). IEEE.
- Bolsa Mexicana de Valores. (n.d.). Retrieved July 12, 2024, from https://camaradecomercioempresarial.org/wp-content/uploads/wpcfto_files/12a0586a2cbf54b32fd7f21c7bf404f0Material%20de%20apoyo%206%20-%20Johan%20Romero%20Azocar.pdf
- Cha, J. (1994). Partial least squares. *Advances in Methods for Marketing Research* (Vol. 407, pp. 52-78).
- Chang, C. C., Lee, Y. J., & Pao, H. K. (2010). A passive-aggressive algorithm for semi-supervised learning. En *2010 International Conference on Technologies and Applications of Artificial Intelligence* (pp. 335-341). IEEE.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297. <https://doi.org/10.1007/BF00994018>.
- Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21-27.
- De Gooijer, J. G., & Hyndman, R. J. (2006). 25 years of time series forecasting. *International Journal of Forecasting*, 22(3), 443-473.
- De Ville, B. (2013). Decision trees. *WIREs Computational Statistics*, 5(6), 448-455. <https://doi.org/10.1002/wics.1278>.
- Dougherty, C. (2011). *Introduction to econometrics*. Oxford University Press, USA.
- Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1996). Support vector regression machines. En *Advances in Neural Information Processing Systems* (Vol. 9). MIT Press.
- Esposito Vinzi, V., & Russolillo, G. (2013). Partial least squares algorithms and methods. *WIREs Computational Statistics*, 5(1), 1-19. <https://doi.org/10.1002/wics.1239>.
- Fama, E. F. (1970). Efficient capital markets. *The Journal of Finance*, 25(2), 383-417.
- Gerle, E. A., McGinnity, M., Belatreche, A., & Coleman, S. (2016). Evaluating machine learning classification for financial trading: An empirical approach. *Expert Systems with Applications*, 54, 193-207.
- Granger, C. W., & Poon, S. H. (2024). Forecasting financial market volatility: A review. *SSRN*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=268866.
- Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN model-based approach in classification. En *On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE* (Vol. 2888, pp. 986-997). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-39964-3_62.
- Hamilton, J. D. (1994). *Time series analysis*. Princeton University Press. ISBN: 9780691042893.
- Harvey, A. C., & Trimbur, T. M. (2003). General model-based filters for extracting cycles and trends in economic time series. *The Review of Economic Statistics*, 85(2), 244-255.
- Haykin, S. (2000). Neural networks: A guided tour. *Nonlinear Biomedical Signal Processing*, 1, 53-68.
- Hecht-Nielsen, R. (1992). Theory of the backpropagation neural network. En *Neural networks for perception* (pp. 65-93). Academic Press.
- Herrera, M. J., & Lockwood, L. J. (1995). The size effect in the Mexican stock market. *Journal of Bank & Finance*, 18(4), 621-632.
- Ho, C. H., & Lin, C. J. (2012). Large-scale linear support vector regression. *Journal of Machine Learning Research*, 13(1), 3323-3348.
- Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice*. OTexts.
- Idrees, S. M., Alam, M. A., & Agarwal, P. (2019). A prediction approach for stock market volatility based on time series data. *IEEE Access*, 7, 17287-17298.
- Imtiaz Khan, N., Mahmud, T., Islam, M. N., & Mustafina, S. N. (2020). Prediction of cesarean childbirth using ensemble machine learning methods. <https://doi.org/10.1145/3428757.3429138>.
- Kumar Dubey, A., Kumar, A., García-Díaz, V., Kumar Sharma, A., & Kanhaiya, K. (2021). Study and analysis of SARIMA and LSTM in forecasting time series data. *Sustainable Energy Technologies and Assessment*, 47, Article 101474. <https://doi.org/10.1016/j.seta.2021.101474>
- Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (2008). *Forecasting methods and applications*. John Wiley & Sons.

- Mazumder, R., Meng, X., & Wang, H. (2022). Quant-BnB: A scalable branch-and-bound method for optimal decision trees with continuous features. En *International Conference on Machine Learning* (pp. 15255-15277). PMLR.
- Misra, P., & Chaurasia, S. (2020). Data-driven trend forecasting in stock market using machine learning techniques. *Journal of Information Technology Research (JITR)*, 13(1), 130-149.
- Naser, M. Z., & Alavi, A. H. (2023). Error metrics and performance fitness indicators for artificial intelligence and machine learning in engineering and sciences. *Applications of Artificial Intelligence and Machine Learning in Architecture, Structure, and Construction*, 3(4), 499-517. <https://doi.org/10.1007/s44150-021-00015-8>.
- Osisanwo, B. G., & Atanda, A. A. (2012). Determinants of stock market returns in Nigeria: A time series analysis. *African Journal of Scientific Research*, 9(1).
- Pérez-Ortega, J., Moreno-Calderón, C. F., Roblero-Aguilar, S. S., Almanza-Ortega, N. N., Frausto-Solis, J., Pazos-Rangel, R., & Rodríguez-Lelis, J. M. (2024). A new criterion for improving convergence of fuzzy C-means clustering. *Axioms*, 13(1), 35. <https://doi.org/10.3390/axioms13010035>.
- Pérez-Ortega, J., Roblero-Aguilar, S. S., Almanza-Ortega, N. N., Frausto-Solis, J., Zavala-Díaz, C., Hernández, Y., & Landero-Nájera, V. (2022). Hybrid fuzzy C-means clustering algorithm oriented to big data realms. *Axioms*, 11(8), 377. <https://doi.org/10.3390/axioms11080377>.
- Pincus, S., & Kalman, R. E. (2004). Irregularity, volatility, risk, and financial market time series. *PNAS*, 101(38), 13709-13714. <https://doi.org/10.1073/pnas.0405168101>.
- Pirouz, D. M. (2006). An overview of partial least squares. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.1631359>.
- Popescu, M. C., Balas, V. E., Perescu-Popescu, L., & Mastorakis, N. (2009). Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, 8(7), 579-588.
- Ranstam, J., & Cook, J. A. (2018). LASSO regression. *Journal of the British Surgery*, 105(10), 1348.
- Rapach, D., & Zhou, G. (2022). Asset pricing: Time-series predictability. *Oxford Research Encyclopedia of Economics and Finance*. <https://doi.org/10.2139/ssrn.3941499>.
- Rasamoelina, A. D., Adjailia, F., & Sinčák, P. (2020). A review of activation function for artificial neural networks. En *2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI)* (pp. 281-286). IEEE.
- Roca, J. C., García, J. J., & De La Vega, J. J. (2009). The importance of perceived trust, security and privacy in online trading systems. *Information Management & Computer Security*, 17(2), 96-113.
- Rouf, N., Malik, M. B., Arif, T., Sharma, S., Singh, S., Aich, S., & Kim, H. C. (2021). Stock market prediction using machine learning techniques: A decade survey on methodologies, recent developments, and future directions. *Electronics*, 10(21), 2717. <https://doi.org/10.3390/electronics10212717>.
- San Martin-Reyna, J. M., & Duran-Encalada, J. A. (2012). The relationship among family business, corporate governance and firm performance: Evidence from the Mexican stock exchange. *Journal of Family Business Strategy*, 3(2), 106-117.
- Schapire, R. E. (2013). Explaining AdaBoost. En *Empirical Inference* (pp. 37-52). Springer. https://doi.org/10.1007/978-3-642-41136-6_5.
- Stathakis, D. (2009). How many hidden layers and nodes? *International Journal of Remote Sensing*, 30(8), 2133-2147. <https://doi.org/10.1080/01431160802549278>.
- Su, X., Yan, X., & Tsai, C. (2012). Linear regression. *WIREs Computational Statistics*, 4(3), 275-294. <https://doi.org/10.1002/wics.1198>.
- Wasserbacher, H., & Spindler, M. (2022). Machine learning for financial forecasting, planning and analysis: Recent developments and pitfalls. *Digital Finance*, 4(1), 63-88. <https://doi.org/10.1007/s42521-021-00046-2>.
- Weigend, A. S. (1994). *Time series prediction: Forecasting the future and understanding the past* (1ra ed.). Routledge. <https://doi.org/10.4324/9780429492648>.
- Werner, R. A. (2016). A lost century in economics: Three theories of banking and the conclusive evidence. *International Review of Financial Analysis*, 46, 361-379. <https://doi.org/10.1016/j.irfa.2015.08.014>.
- Windeatt, T. (2006). Accuracy/diversity and ensemble MLP classifier design. *IEEE Transactions on Neural Networks*, 17(5), 1194-1211.
- Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1), 67-82.
- Xiao, L., & Aydemir, A. (2007). Volatility modelling and forecasting in finance. En *Quantitative finance: Forecasting volatility in the financial markets* (3ra ed.). <https://doi.org/10.1016/B978-075066942-9.50003-0>.
- Ying, C., Qi-Guang, M., Jia-Chen, L., & Lin, G. (2013). Advance and prospects of AdaBoost algorithm. *Acta Automatica Sinica*, 39(6), 745-758.
- Zarnowitz, V., & Ozylidirim, A. (2006). Time series decomposition and measurement of business cycles, trends and growth cycles. *Journal of Monetary Economics*, 53(7), 1717-1739.
- Zhang, S. (2021). Challenges in KNN classification. *IEEE Transactions on Knowledge and Data Engineering*, 34(10), 4663-4675.

Article

MASIP: A Methodology for Assets Selection in Investment Portfolios

José Purata-Aldaz ¹, Juan Frausto-Solís ^{1,*}, Guadalupe Castilla-Valdez ¹, Javier González-Barbosa ¹
and Juan Paulo Sánchez Hernández ²

¹ Graduate Program Division, Tecnológico Nacional de México/Instituto Tecnológico de Ciudad Madero, Ciudad Madero 89440, Mexico; luis.pa@cdmadero.tecnm.mx (J.P.-A.); gpe_cas@yahoo.com.mx (G.C.-V.); jjgonzalezbarbosa@hotmail.com (J.G.-B.)

² Dirección de Informática, Electrónica y Telecomunicaciones, Universidad Politécnica del Estado de Morelos, Boulevard Cuauhnáhuac 566, Jiutepec 62574, Mexico; juan.paulosh@upemor.edu.mx

* Correspondence: juan.frausto@itcm.edu.mx

Abstract: This paper proposes a Methodology for Assets Selection in Investment Portfolios (MASIP) focused on creating investment portfolios using heuristic algorithms based on the Markowitz and Sharpe models. MASIP selects and allocates financial assets by applying heuristic methods to accomplish three assignments: (a) Select the stock candidates in an initial portfolio; (b) Forecast the asset values for the short and medium term; and (c) Optimize the investment portfolio by using the Sharpe metric. Once MASIP creates the initial portfolio and forecasts its assets, an optimization process is started in which a set with the best weights determines the participation of each asset. Moreover, a rebalancing process is carried out to enhance the portfolio value. We show that the improvement achieved by MASIP can reach 147% above the SP500 benchmark. We use a dataset of SP500 to compare MASIP with state-of-the-art methods, obtaining superior performance and an outstanding Sharpe Ratio and returns compared to traditional investment approaches. The heuristic algorithms proved effective in asset selection and allocation, and the forecasting process and rebalancing contributed to further improved results.



Academic Editor: Oliver Schütze

Received: 31 January 2025

Revised: 20 March 2025

Accepted: 21 March 2025

Published: 24 March 2025

Citation: Purata-Aldaz, J.; Frausto-Solís, J.; Castilla-Valdez, G.; González-Barbosa, J.; Sánchez Hernández, J.P. MASIP: A Methodology for Assets Selection in Investment Portfolios. *Math. Comput. Appl.* **2025**, *30*, 34. <https://doi.org/10.3390/mca30020034>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: investment portfolios; optimization; forecast; heuristics

1. Introduction

In financial asset management and investment decision-making, portfolio creation and optimization represent a fundamental paradigm and essential study area. Building investment portfolios addresses the complex task of allocating limited resources to a diverse range of financial assets, as well as forecasting them [1]. This process requires careful analysis of the risks and returns associated with each asset and understanding their interaction in the context of a diversified portfolio. Optimization involves the search for asset combinations that maximize expected returns or minimize risk, all subject to specific constraints and objectives. Similarly, there are diversification and asset selection methodologies, such as the Markowitz portfolio theory and its extensions, value at risk (VaR) models, and asset allocation strategies [2]. The Markowitz Mean-Variance model, also known as the Modern Portfolio Theory (MPT), is presented in Equations (1)–(4), which is a theoretical framework developed by Harry Markowitz [3]. This study is relevant because it allows portfolio optimization and considers risk-taking based on performance

and diversification. This quantitative approach provides an objective framework for making investment decisions and lays the foundations for investment management.

$$\text{Max } E(X_i) = \sum_{i=1}^N X_i \mu_i \quad (1)$$

$$\text{Min } V(X_i, X_j) = \sum_{i=1}^n \sum_{j=1}^n X_i X_j \sigma_{ij} \quad (2)$$

$$\sum_{i=1}^n X_i = 1 \quad (3)$$

$$0 < X_i \leq 1 \quad (4)$$

where μ_i is the expected return of the i th asset, X_i is the relative value invested in the asset i , σ_{ij} represents the covariance between assets, and $E(X_i)$ is the expected return of the portfolio. Notice that the invested amount is normalized to one in the constraint defined by Equation (3), and no negative investments X_i s are permitted according to Equation (4). The decision variables are the expected value $E(X_i)$ and the variance V , and the objectives are to achieve maximum profit and minimum risk.

In the model defined by Equations (1)–(4), the variance V of the portfolio is introduced as a risk measure, which is essential for constructing diversified portfolios. Covariance and correlation are also key elements for good diversification [3]. From the mean-variance model, two objectives must be met: firstly, to maximize the portfolio return and secondly, to minimize the risk. These two objectives become the main difficulty for solving the problem. The Sharpe Ratio (ζ) is a popular and straightforward function [4]; ζ is commonly used in modern portfolio methods [5–7], requiring only the expected value of the portfolio, its volatility, and a reference rate [8]. They are called Sharpe Ratio models and offer several advantages, including the facility to reduce the portfolio evaluation problem from a bi-objective scheme to a mono-objective optimization problem with Equations (5)–(7), thereby focusing solely on maximizing the ζ function.

$$\max_{\zeta}(X_i) = \frac{\sum_{i=1}^N X_i \mu_i - R_B}{\sqrt{\sum_{i=1}^n \sum_{j=1}^n X_i X_j \sigma_{ij}}} \quad (5)$$

$$\sum_{i=1}^n X_i = 1 \quad (6)$$

$$0 < X_i \leq 1 \quad (7)$$

where the $\zeta(X_i)$ function indicates the expected differential return per unit of risk associated with different assets conforming to the portfolio, and R_B is the return on a benchmark or the risk-free rate. This portfolio optimization model has only one decision variable, defined by the Sharpe Ratio $\zeta(X_i)$; this variable involves the relation between $E(X_i)$ and V . Thus, Equations (5)–(7) represent a mono-objective problem.

The composition of an investment portfolio entails more than merely selecting assets with the most significant potential to achieve stated objectives and satisfy constraints. Moreover, the optimal weights of the assets that satisfy the constraints and optimize the objective function must be determined. Also, the future value of the assets and their weighting in the portfolio for the periods during which the investment is planned must be known. These future values lead to a realistic portfolio being obtained for the investment plans. Therefore, it is necessary to forecast every asset as accurately as possible for the investment period. There are many successful forecasting methods, such as classical and modern

machine learning, and combining forecast methods has proven valuable in enhancing accuracy and reducing the variance of forecasting errors [9]; they can generate more accurate forecasts than those from individual methods [10]. Optimization methods can obtain the best weights of the individual forecasting methods participating in an ensemble. One of these methods is Threshold Accepting (TA). This iterative algorithm controls the number of iterations by a temperature parameter and accepts some poor solutions when they fulfill an acceptance criterion defined in terms of a threshold parameter [11]. TA is an efficient tool in many applications [12].

The associated literature about how to conform an initial portfolio is scarce. However, most of the existing literature focuses on a single aspect of portfolio construction: (a) The identification of the candidate assets to be integrated in the portfolio [13,14]; (b) The optimization of the portfolio [15–17]; or (c) The application of a forecast strategy to predict the portfolio return [18–20]. This observation underscores a promising area of opportunity, which we seek to address in this study by implementing a novel approach. MASIP integrates three fundamental components: the preselection of assets, considering a minimum acceptable return; the forecasting of the selected assets based on an assembly method optimized by an enhanced threshold acceptance algorithm; and the rebalancing of assets to create the final portfolio. The above-mentioned points are performed by MASIP, an innovative methodology incorporating an asset pre-selection stage, an objective function based on the median-variance and Sharpe Ratio model, and asset diversification. This methodology also implements an acceptance scheme for a minimum return and a maximum permissible risk rate. The modern combining forecast method and the well-tested portfolio optimization heuristic algorithm accomplish MASIP.

The rest of this paper is organized as follows. Section 2 reviews the background of portfolios, including financial time series forecast modeling and portfolio integration. Section 3 introduces the details of MASIP methodology for portfolio integration models and describes the construction process of an ensemble forecasting model. Section 4 presents the experimentation and results from the studies using financial time series data and the application of MASIP. Detailed analysis and result interpretations are provided. Section 5 presents the conclusions.

2. Background

The construction of investment portfolios has been a central topic in the financial literature for decades, driven by the search for effective strategies that maximize return and minimize risk. From Markowitz's pioneering work on portfolio theory to the most recent advances in financial forecasting, research has generated a vast body of knowledge that underscores the importance of a meticulous approach to portfolio construction and management. This section considers the modern aspects related to portfolio construction: asset selection, forecasting stock methods, and portfolio optimization.

2.1. Asset Selection and Portfolio Construction

Cesarone et al. (2020) provided theoretical results for the risk-parity approach to general risk measures. They proposed an efficient and accurate Multi-Greedy heuristic to solve the portfolio selection problem [21]. Empirical results on real-world data show that diversified optimal portfolios are only slightly suboptimal in sample and generally exhibit better out-of-sample performance than purely diversified or optimized portfolios [22]. The authors analyzed NASDAQ100, FTSE100, and SP500 assets, finding good Sharpe values returns ratios for the SP500 index. The combined diversification–optimization approach is better than pure diversification or optimization.

A framework proposed by Puerto et al. (2023) adopts a bi-criteria mean-risk optimization² approach to model the portfolio optimization problem, where the average portfolio rate of return is maximized while minimizing a given risk measure. These authors propose the use of mixed-integer linear programming (MILP) formulation to solve portfolio clustering and selection problems in a unified phase [23]. The authors focus on using Conditional Value at Risk (CVaR) as the objective function for portfolio optimization and evaluate the performance of the proposed integer linear programming formulation for the SP500 market. In another work performed by Kaczmarek and Perez, they propose the mean-variance model and hierarchical risk parity (HRP) for SP500, achieving a Sharpe Ratio ζ in the range of 0.76 to 0.97 [24]. Unfortunately, they consider assets from an ancient dataset. In 2023, Zhang used datasets from the Chinese market based on the mean-variance model and Sharpe Ratio; therefore, a Monte Carlo process is proposed for asset selection to adjust the portfolio within the efficient frontier [25].

Ma et al. (2020) analyze the efficacy of a combination of machine learning and deep learning algorithms in predicting returns [26]. The analysis utilizes two machine learning models: Random Forest (RF) and Support Vector Regression (SVR), as well as three deep learning models: Long Short-Term Memory (LSTM), Deep Multilayer Perceptron (DMLP), and Convolutional Neural Network (CNN). These models were utilized for the purposes of stock preselection and portfolio optimization through the mean-variance (MV) and omega models for the Chinese market. The findings of this study indicated that RF + MV demonstrated superior performance.

For the European markets, LSTM neural networks were recently applied to predict asset returns and incorporated into the mean-variance optimization model to determine the optimal asset allocation. The average return prediction accuracy in this case was 95.8%, and the integrated portfolio outperformed the benchmark [27]. Table 1 presents a summary of related studies. The present study is focused on SP500; thus, we will compare our results with the references [21,23].

Table 1. Related studies.

Approach	Asset Preselection	Asset Forecasting	Portfolio Optimization
Cesarone [21]			X
Puerto [23]			X
Kaczmarek [24]			X
Zhang [25]			X
Ma [26]		X	X
Martinez [27]		X	X
MASIP	X	X	X

2.2. Asset Portfolio Prediction Methods

Nowadays, it is well known that combining several methods has advantages over individual forecasting models, and several ensemble methods are available [28]. For portfolio design and asset selection, an efficient way to analyze and predict risky assets is required, as well as a comparative analysis of the structure of a complete portfolio, including risky and risk-free assets, depending on the risk aversion coefficient [29,30]. Combinational forecasting has several approaches, such as the Simple Average, which is the sum of the forecasts divided by the number of forecasts. The Weighted Average model assigns weight to each forecast. The combined forecast is calculated as the sum of the forecasts multiplied by their respective weights, divided by the sum of the weights [31].

The Aggregation Model employs a machine learning algorithm to integrate forecasts derived from multiple models. It utilizes a range of algorithms, including neural networks, support vector machines, and decision trees, among others [32–34].

An ensemble model integrates various forecasting methodologies, employing a criterion to determine their contribution to the ensemble. These methodologies may encompass machine learning algorithms, autoregressive integrated moving average (ARIMA) model variants, exponential smoothing techniques, and other approaches. The mathematical formulation for this integration is dependent on the specific models being combined [30,35–38]. Metaheuristic methods, such as simulated annealing and genetic algorithms, can be applied to analyze the datasets while integrating the assets to determine their participation in a portfolio. Neural networks and support vector machines have also been used [39]. In other words, it is crucial to acknowledge that the models employed in forecast combination methodologies may differ depending on the specific approach utilized and the characteristics of the forecasting issue at hand. When there are a limited number of forecasting methods, assigning weights or balances to them is relatively straightforward. However, as the number of models increases, the task becomes significantly more complex. Thus, it is necessary to rely on optimization tools.

The integration of machine learning, deep learning, and probabilistic techniques into stock market analysis has the potential to enhance the precision of predictions and provide novel insights into portfolio management. However, it is imperative to recognize the limitations and challenges inherent in these approaches, such as data quality and model complexity, to ensure optimal effectiveness in practical applications [40,41].

2.3. Portfolio Optimization Methods

The potential for employing diverse forecasting techniques and integrating them into a portfolio introduces a novel challenge: identifying the optimal approach or discerning the most effective combination of forecasting methods that minimizes error and yields superior outcomes compared to individual forecasts. To address this, combinatorial and optimization strategies must be employed.

Optimization algorithms, such as TA, have become important in several areas [11]. The last reference highlights the superior performance of TA over other methods, such as Simulated Annealing, for combinatorial challenges. A new version of TA is used to improve portfolio asset weights by exploiting its convergence property [42], its applicability to econometrics applications [43], and in financial sector problems [44]. Gilli and Kellezi showed the effectiveness of TA in selecting a subset of assets that closely replicate the performance of the benchmark index, achieving high performance in investment portfolio design. Liu and Tadesse used the LSTM machine learning method to forecast assets from the SP500 market [45]; they applied the Sharpe Ratio and Equations (3)–(5) to integrate the investing portfolio. However, they considered only four assets of this market without a method for selecting the assets to incorporate into the portfolio. In addition, no risk-free reference rate was considered. When this value is considered, the yield and the Sharpe Ratio performance would be different. As shown in another section, this modification is performed in the model we propose in the present study.

These approaches permit investors to anticipate forthcoming market changes, optimize resource allocation, and mitigate potential risks. By employing classical, advanced, or combined models, it is possible to make more accurate predictions of market trends and asset behavior. This work, through the combination of various methods and their optimization, is intended to be a powerful tool for decision-making.

3. MASIP Methodology

This section presents the MASIP methodology for asset selection in investment portfolios and its related algorithms. A step-by-step approach addressing the issues relating to the following complex tasks is presented: asset selection, initial portfolio creation, asset

forecasting, and final rebalancing or optimization. Furthermore, MASIP involves acquiring, analyzing, segmenting, and evaluating extensive time series data to achieve investment goals. Figure 1 presents the elements of MASIP and their interactions.

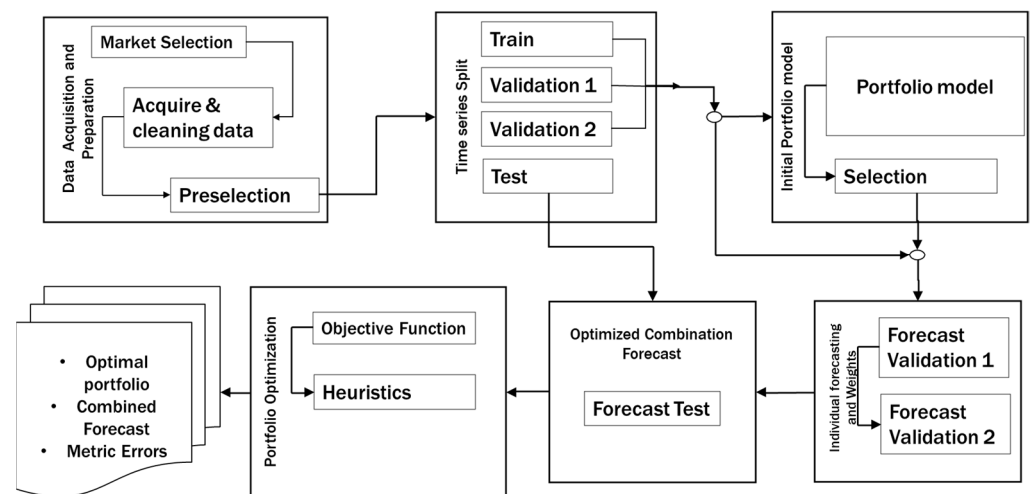


Figure 1. The MASIP methodology.

3.1. Data Acquisition and Preparation

Market selection is the first step to integrate the investment portfolio. At this point, the target market to be analyzed is selected; among the candidate markets to be evaluated, we could mention those listed in the New York Stock Exchange, the European region, the Asian region, or those belonging to developing countries such as BRICS or Latin America. In this case, the analyst must perform a previous analysis of the target market and establish the period to be analyzed. In acquisition and cleaning of data, the information on each of the candidate assets is obtained. We consider the daily closing prices to conform to the time series in a determined time. As we know, it is important to have quality, cleanliness, and order of information since, if erroneous or missing information is entered, we will have inaccurate conclusions that can lead to severe losses of resources. Likewise, the data must be complete; thus, the necessary techniques must be applied to enable the data series cleaning process. Some of the methods used may be normalization, other data transformation, inter- or extrapolation, application of simple averages, moving averages, or any other variation that allows us to have complete information without altering the quality of the data [46–48].

The selection of assets should be made before integrating the portfolio. At this point, the analyst knows which assets of the selected marketplace are already the candidates for integrating the optimal portfolio. Thus, only the corresponding time series would be used to design it, and the effort for this task can be reduced. For example, if we analyze SP500, one of the largest financial exchanges in the world, we will have to analyze more than 500 time series, which consumes a lot of time and resources. To streamline the evaluation process, we apply the discrimination of assets that meet a percentage of profit or return by the Minimum Acceptable Rate of Return or MARR [49], which could be used to select an investment that assures us a fixed return with minimal utility and no risk. Thanks to this scheme, we can significantly reduce the number of candidate assets that need to be evaluated [8].

3.2. Time Series Split

The segmentation of time series is an important practice since it allows for evaluating the performance of models using known data and comparisons with previously unseen data. When a model is trained with a time series, it is expected to be able to make accurate

predictions. However, if the model is evaluated with the same training data, there is a risk of overfitting, and only training data are stored.

When the time series data are divided into training, validation, and test sets, it is possible to adjust the model parameters to minimize the error in the validation set and then evaluate the model performance in the test set, which contains data that are “new” to the model and not previously analyzed.

Segmentation also allows for the identification of problems with the data, such as outliers or anomalous values, which allows for treating and correcting the input data before training the models and obtaining a final model [35,50,51].

In this implementation, the training portion of each time series is initially employed to forecast the validation section, applying all available individual forecast methods. The outcomes of this process are used to determine the most effective methods for forecasting the validation section and to establish the distribution of weights for the combination of these optimal methods. Subsequently, the forecast is evaluated using the validation data, and the most effective methods are identified, along with the previously calculated weight distribution. The resultant forecasts are stored, and the optimization process with TA is initiated using Validation 1 to forecast Validation 2. The weights are adjusted until no further improvement is observed [30].

3.3. Initial Portfolio Model

Building a portfolio is not a simple task today since several factors can influence the selection, integration, and weighting of the assets of the investment portfolio. That is why it requires robust and reliable tools which allow us to analyze many candidate assets and the data that integrate them. As we present in Section 1, the MPT model (Equations (1)–(4)), proposed by Markowitz, is the pioneering work in portfolio management which considers two main objectives: maximizing the expected return and minimizing the risk [3,52]. The Sharpe model represented by Equations (5)–(7) is an MPT extension considering the two objective functions in only one. In the present work, we propose a like-Markowitz Model of MASIP defined by Equations (8)–(12).

$$\max SR = \frac{E(R_{pf}) - \text{MARR}}{\sqrt{\sum_{i=1}^n \sigma_i^2 x_i^2 + 2 \sum_{i=1}^n \sum_{j \neq i}^n \sigma_i \sigma_j x_i x_j \rho_{ij}}} \quad (8)$$

subject to

$$\sum_{i=1}^n x_i = 1 \quad (9)$$

$$\sum_{i=1}^n E(R_{pf}) > \text{MARR} \quad (10)$$

$$0 < x_i \leq 1 \quad (11)$$

$$\sum_{i=1}^n \sigma_i^2 x_i^2 + 2 \sum_{i=1}^n \sum_{j \neq i}^n \sigma_i \sigma_j x_i x_j \rho_{ij} \leq \Phi_{\max}^2 \quad (12)$$

where $E(R_{pf})$ is the expected return, i and j are the assets, σ_{pf} is the portfolio risk, x_i, x_j are the contribution of i and j assets, σ_i, σ_j are the standard deviation of the assets, ρ_{ij} are the correlations between assets i and j , and $\Phi_{\max}^2$ is the Maximum Risk Rate Allowable.

The MASIP model defined by Equations (8)–(12), as we mentioned above, is an MPT extension with the following differences:

- In the objective function, we used the parameter MARR to represent the Minimum Acceptable Rate of Return that the investor could accept; instead, the Markowitz model uses a risk-free rate;
- Constraint (10) uses the MARR parameter for weighting assets into the portfolio;
- Constraint (12) uses the Φ_{max}^2 risk parameter to define the assets integrating the portfolio. This parameter binds the risk associated with the asset.

We use the parameter MARR, representing a common aspect used by decision-makers to accept an investment; however, the MPT and other models use a risk-free rate. Moreover, the average is sensitive to outliers and can be affected by extreme fluctuations. Conversely, the Expected Value (mathematical expectation) is calculated by multiplying each possible value by its probability of occurrence and adding up the results. In short, while the average is a basic measure of central tendency, the expected value more accurately represents the inherent uncertainty in the data [53].

In the state-of-the-art methods, a genetic algorithm is used for initial portfolio selection (GENPO), as well as algorithms based on threshold acceptance (TAIPO) and simulated annealing (SAIPO), which were tested and showed superior performance to other models, mainly focused on risk management (Yu, Gilli, and Masese) [16,17,54]. The last models are cited in the present study as YGM models. The individual YGM models versus MASIP are analyzed. The results of this analysis demonstrate the performance of the 12 combinations obtained between the YGM models and the hybridized algorithms. These results show that the YGM models using the maximum SR objective function applying the TAIPO algorithm obtain superior performance [8].

3.4. Individual Forecasting and Weights

Forecasting the performance of an investment portfolio is crucial for various reasons, such as identifying and quantifying the risks associated with the portfolio, financial planning, and evaluating strategy. Forecasting enables financial analysts to anticipate and adjust to changing market conditions, allowing them to make more timely and effective decisions.

The forecasting section of the Methodology is supported by FCTA (Forecasting Combined Method with Threshold Accepting) [30], which consists of assembling 14 forecasting methods and an enhanced threshold-accepting algorithm for weighting individual predictions. The forecast methods involved are Autoregressive integrated moving average ARIMA, exponential smoothing state space (ETS), Neural network forecasting method NNetar, Exponential smoothing state space-Box-Cox-ARMA (ETS), STLM with AR method, Naive and random walk forecasts with drift activated, Decomposing Time Series with Smoothing Methods (Theta), Naive method, Autoregressive fractionally integrated moving average (ARFIMA), Bootstrapped method (Bagged), SplinesCubic, Holt, Feature-based forecast model averaging (FFORMA), and Jaganathan.

Notably, at least two of the models incorporated, FFORMA [36] and Jaganathan [55], obtained second and fourth place, respectively, in the competition, as both FFORMA and Jaganathan are ensemble and hybridization methods in certain cases [56]. Within this framework, it is essential to underscore the more than 24,500 financial series within the dataset, which suggests that the application of FCTA for the analysis of financial series can be considered reliable.

The FCTA employs a set of criteria to evaluate the efficacy of forecasting methods, including the uniform distribution, the normalized weights criterion, the Bayesian information criterion (BIC), and the Akaike information criterion (AIC). These criteria have been demonstrated to yield superior performance in the analysis and forecasting of the M4 competition series, which encompasses financial series.

The forecasting process is presented in Figure 2. In this stage, individual forecasts are made for each asset and its time series, using each forecasting technique considered in the algorithm, that is, the n assets selected in the initial portfolio are evaluated and forecasted by the M forecasting models. This is performed for each split of the time series, that is, each of the forecasting models is fed with the training section of each of the asset's data and compared with the validation section. The error is measured for all the best methods, and the ones with the lowest error evaluation are selected and compared against the test section.

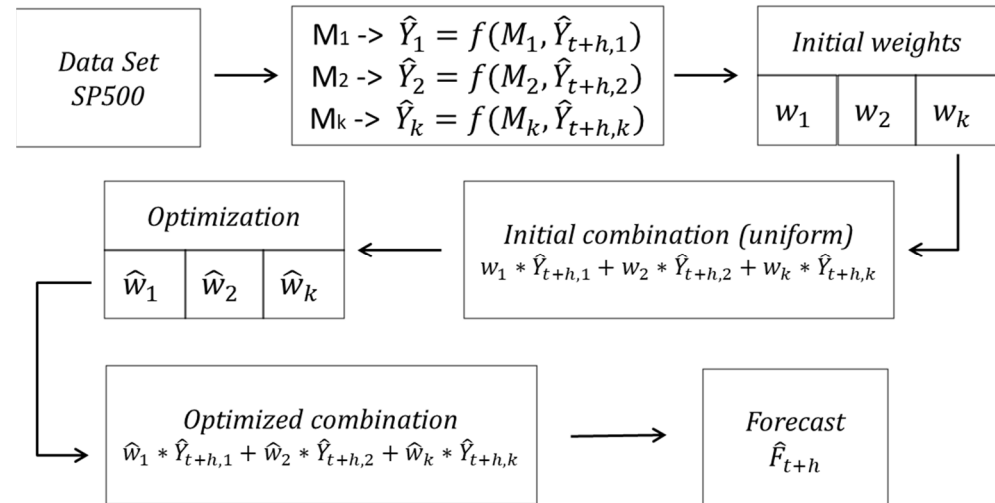


Figure 2. Combination of forecasting methods.

The initial combination of forecasting methods is established through the selection of the best-evaluated ones.

Through experiments involving enhanced weights, it was observed that a collection of methods yielded superior results compared to individual methods. They identified which features were crucial for selecting methods for the combination. Notice that the most effective combination is typically achieved when the best individual methods are employed.

Since the mid-20th century, ensemble methods for forecasting have been applied [57]; specifically, Bates [9] proposed using several forecasting methods in a weighted combination with weights w to be determined in terms of the error of each of the individual methods. This method is commonly known as the Bates method and was initially tested using Brown's exponential smoothing methods. The way to compare $n \geq 2$ methods in the Bates method is to determine the Mean Squared Error (MSE) for all methods for the series considered in an evaluation period; then, the methods are combined so that each participates in the forecast in terms of the accuracy corresponding to its MSE. A final forecast ensemble or combination can be expressed as Equation (13):

$$\hat{Y}_f = \sum_{i=1}^n w_i \hat{y}_i \quad (13)$$

It can be said that the relationship between the combined forecast $\hat{Y}_f$ and the individual forecast $\hat{y}_i$ are linear, where we have n individual forecast models, and for each model $\hat{y}_i$ we have a weight w_i .

Moreover, the objective function is measured using the symmetric mean absolute percentage error (sMAPE), a variation of the classic mean absolute percentage error (MAPE)

metric [58]. However, other convenient error metrics could be used. The metric used in this paper is expressed in Equation (14).

$$sMAPE = \left[\frac{2}{h} * \sum_{t=n+1}^{n+h} \frac{|Y_i - \hat{Y}_i|}{|Y_i| + |\hat{Y}_i|} \right] * 100 (\%) \quad (14)$$

where Y_i is the current observation of the time series, $\hat{Y}_i$ represents the predicted value, n is the amount of data in the time series, and h represents the length of the forecast horizon. This measure has been the subject of analysis since the late 1980s [59], but it gained popularity with the M3 competitions [60,61] as a metric to compare various forecasting methods. sMAPE is simple to calculate and understand, and unlike other errors, sMAPE considers both overestimates and underestimates equally.

Other error metrics considered for evaluation are MSE (15), Root Mean Squared Error (RMSE) (16), and Mean Absolute Percentage Error (MAPE) in Equation (17).

$$MSE = \frac{1}{n} \sum_{t=1}^n (Y_i - \hat{Y}_i)^2 \quad (15)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (Y_i - \hat{Y}_i)^2} \quad (16)$$

$$MAPE = \frac{100}{n} \sum_{t=1}^n \left| \frac{(Y_i - \hat{Y}_i)}{Y_i} \right| \quad (17)$$

$$Rsquared = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y}_i)^2} \quad (18)$$

In the context of machine learning forecasting, the evaluation of model performance is subject to selecting appropriate metrics. This study includes MSE, MAPE, and RMSE to evaluate forecasting models. MSE is used as an objective function because of its ability to penalize large errors, although its sensitivity to outliers can be a disadvantage. MAPE can overestimate small errors and generate problems when actual values are close to zero. sMAPE, on the other hand, is used to present results which are easy to understand, as it provides a more symmetric error penalty compared to MAPE [62]. While MSE is useful for optimizing models, its interpretation is not intuitive for users. Therefore, sMAPE is used for internal model fitting, while MSE is used for algorithm comparison. This approach is designed to ensure a comprehensive and equitable evaluation of forecasting models, aligning with practices employed in time series forecast competitions [63].

The RMSE is useful in cases where large errors are unacceptable because it takes the square root of the errors. This makes the RMSE the same size as the original variable, which makes it easier to understand the average size of the error. However, it can also be too harsh on big errors, which may be a problem if they are rare or not important in the analysis. Outliers can also have a big effect on the RMSE, which can change how we judge the model when there are outliers [35]. The coefficient of the determination Rsquared (R^2) measures how well a forecasting model explains the variability of the actual data [35].

3.5. Optimized Combination Forecast

Once an initial combination of forecasting methods is obtained using the best-evaluated models, it is necessary to establish the optimal combination of the forecast models, i.e., to weigh them in such a way that this combination delivers the lowest possible error. For this, it is necessary to use again the heuristics that allow solutions, which, in this

case, try to minimize the error of the combined forecast against real test data. As mentioned above, it is possible to use any of the heuristic methods.

Figure 3 shows the general process of the FCTA methodology. In the first step, the training data of the n time series are used in the m forecasting methods, $\hat{Y}_{t+h}(M1)$. These methods generate a weighted combination of methods for each time-series. This combination is then compared to validation segment 1. The process is repeated with the data of Validation 1. These data are used to forecast into the periods of Validation 2. The results are then compared to Validation 2. The test split is applied to evaluate the optimized combined forecasting. The error analysis in stages 1 and 2 selects the best individual forecasting methods for each of the selected assets in the initial portfolio, thereby establishing an initial mix of forecasting methods. This combination is uniform when equal weights are assigned to each forecasting method.

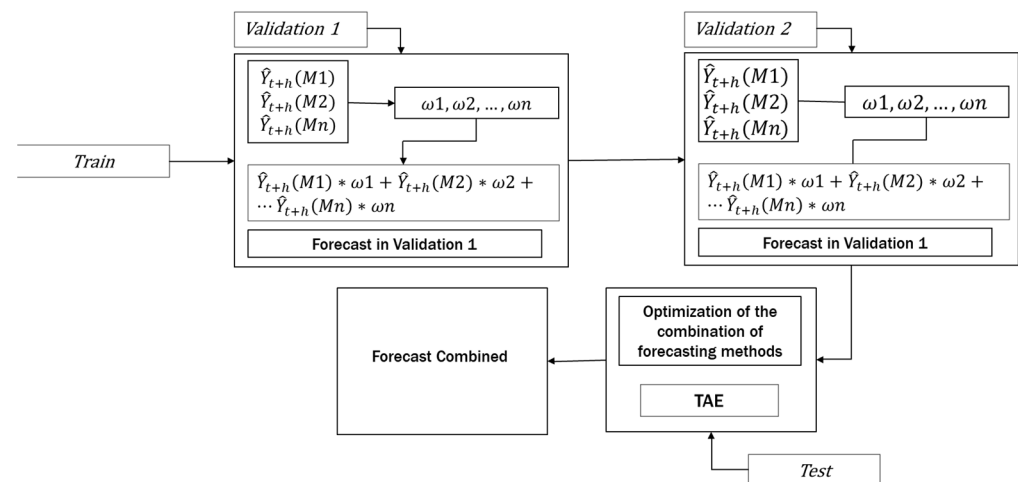


Figure 3. General outline of FCTA.

The initial combination of the best forecasting methods enters the optimization heuristic based on TA, which has been modified to obtain a superior performance to the classical TA algorithm. The algorithm Threshold Accepting Enhanced (TAE) is an algorithm including a modified cooling scheme, a reheating scheme, and thermal equilibrium. This algorithm was tested and showed a high performance in M4 time series [30].

The output of the TAE algorithm is the weighted combination of the best forecasting methods that obtain the minimum value of the sMAPE error metric. The weighted methods are then combined to generate forecasts, which are subsequently compared against the test section to calculate the error metric.

3.6. Portfolio Optimization

Following the conclusion of the forecasting process, it is imperative to perform a portfolio rebalancing or optimization. In this stage, the candidate assets' past and forecasted data are evaluated using the methodologies or algorithms outlined in Section 3.3. The anticipated outcome of this process is a portfolio that exhibits enhanced performance in comparison to its initial state. The logical sequence and steps of MASIP are shown in Algorithm 1.

Algorithm 1. Pseudo code for MASIP methodology

```

1:  Market: Time series for portfolio forecasting (assets)
   Result: Optimal Portfolio Forecast
2:  Initialization
3:  Market <- Get Market data
   Market <- Clear Market data
4:  For i in each asset in the Market
5:    If profit_asset >= MARR #Minimum Acceptable Rate Return
6:      Data <- asset
7:      Train, validation 1, validation 2, Test <- splitData()
8:      InitialPortfolio <- PortfolioSelection(Data[Train, validation 1, validation 2],
      ObjectiveFunction)
9:      IndividualForecast <- IndividualMethods(Data[InitialPortfolio])
10:     InitialCombinationMethod <- IncludeProcess(IndividualForecast)
11:     ForecastCombination <- FCTA(InitialCombinationMethod, Test)
12:     OptimalPortfolio <- PortfolioSelection(Data[Train, validation 1, validation 2, Test],
      ForecastCombination)
13:  Return (Optimal portfolio, Combined Forecast, Metric Errors)

```

4. Experimentation and Results

This section reviews the equipment and material resources used for developing this study. Since this is complex work, several computer parts equipment were required: Intel(R) Core(TM) i7-13620H, CPU @ 2.40 GHz, RAM: 16 GB, Intel(R) Xeon(R) Gold 6230 CPU @ 2.10 GHz, and 2.10 GHz, RAM: 12 GB (National Laboratory of Information Technology, Tamaulipas, Mexico).

The languages used for this experiment are Python 3.7 for initial and final portfolios and R 4.3.3 for forecasting.

A description of the time series utilized and an account of the experimentation conducted with this series through the portfolio integration and portfolio forecasting algorithms are also provided. A concise overview of both is included. The results and discussion of these experiments are subsequently presented.

4.1. Methodology Application

In this section the methodology application is described in detail and each step of our methodology is explained from the data acquisition and preparation to the final portfolio optimization.

4.1.1. Data Set Acquisition and Preparation

To evaluate the algorithms and the proposed solution methodology, a set of assets and their data series of the SP500 were used, consisting of 479 companies listed between January 2020 and the end of February 2025, giving a total of 262 observations (weekly prices). This period was chosen because there was significant variability in the index and the stocks due to the health emergency that affected the assets of the United States of America and the whole financial world.

Justifications for using the SP500 are the market representativeness and diversification (wide variety of companies of different sectors and sizes), access to historical data, liquidity which allows for easy and efficient trading, it being commonly used as a benchmark to compare the performance of an investment portfolio, and, finally, the U.S. equity market enjoys strong regulation and institutional stability.

A computer program was developed to obtain information on the SP500 market assets corresponding to the daily closing prices of the selected period from the Yahoo Finance portal. Once the initial set was formed, the amount of empty or non-numeric data (NAN) was checked, and those with missing data representing more than 50% of the total were discarded.

In the first phase of the proposed model, data validation and cleaning were performed, which consisted of verifying that all the assets had the complete and correct information for the period to be analyzed. The main tasks performed in this phase were detecting atypical data and data supplementation in case of missing data at the beginning, end, or within the series. The complementation process was performed by applying the extrapolation (missing data at the beginning or end of the series) and interpolation methods.

As a final step, the expected value of each asset was compared to the return of the SP500 index, as shown in Figure 4, where the assets with a return lower than this value were discarded, since the assets that do not exhibit a return higher than this limit would be of no interest to financial investors. As mentioned above, this helps to reduce computation time since the final set of data to be analyzed is reduced.

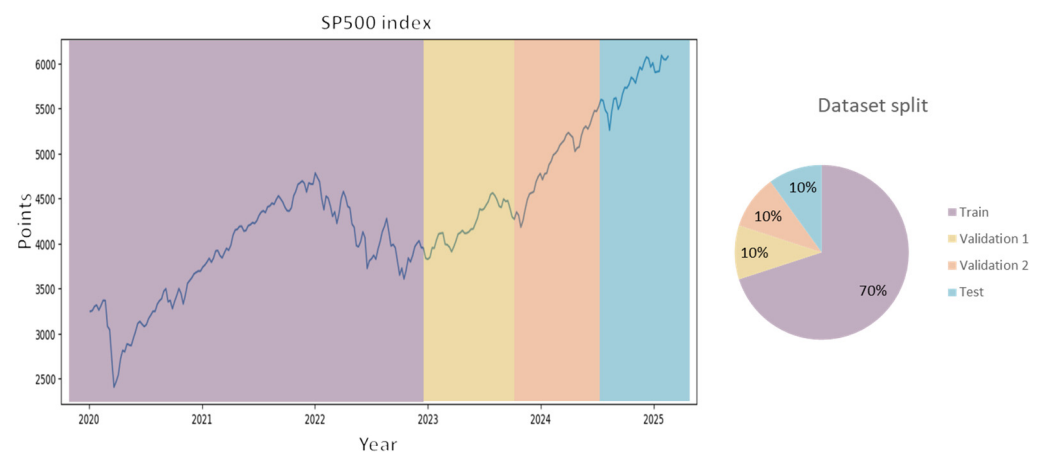


Figure 4. SP500 performance and split.

4.1.2. Segmentation of the Data Set

At this point, it was decided to segment the data series into four sections: the first segment is the training set, which was established as 70% of the data, i.e., 183 weekly observations, two validation segments, which consist of 10% of the observations each, i.e., 26 observations each, and finally a test segment equivalent to the remaining 10%. We decided to segment in this way to make the dataset compatible with the techniques applied in the FCTA methodology for data forecasting. Similarly, this same segmentation is applicable in constructing the initial portfolio where the training and validation sets are taken as input data. To summarize, the segmentation written above is in Figure 4.

4.1.3. Initial Portfolio Hyperparameters

For the construction of the initial portfolio, the TAIPO algorithm was used. This had already been tested and compared previously with other alternatives, demonstrating its advantages, such as obtaining high values in the Sharpe Ratio as well as a good return and a low correlation of assets.

The hyperparameters of the TAIPO metaheuristic were tuned by applying the analytical tuning method described in [8,64]. The values obtained for the parameters are: Initial Temperature: 273.8164, Final Temperature: 0.0000169, Cooling ratio: 0.96, while the equilibrium cycle length L , was dynamically calculated as a function of the number of assets and the temperature scheme.

Given that TAIPO is an approximate algorithm, 30 executions were carried out to obtain the average of the performance indicators, thereby enabling the selection of assets that maximize returns. The average values obtained were: Sharp Ratio (SR): 0.0221, Return: 0.0114, Risk: 0.2980, Correlation between assets: 0.2044, and the number of selected assets: 57, in weekly terms.

Compared to the performance of the SP500 during 2024, equivalent to 0.24 per year, the portfolio integrated in this stage had a performance of 147% above the index.

Figure 5 shows a sample of the graphs generated by the results of the algorithm. It can be observed that the Sharpe Ratio behaves as expected regarding the TAIPO algorithm; in the initial interactions, the system tolerates suboptimal solutions. However, as the iterations progress under the cooling scheme, the system becomes less tolerant of accepting bad solutions, ultimately identifying a higher SR value.

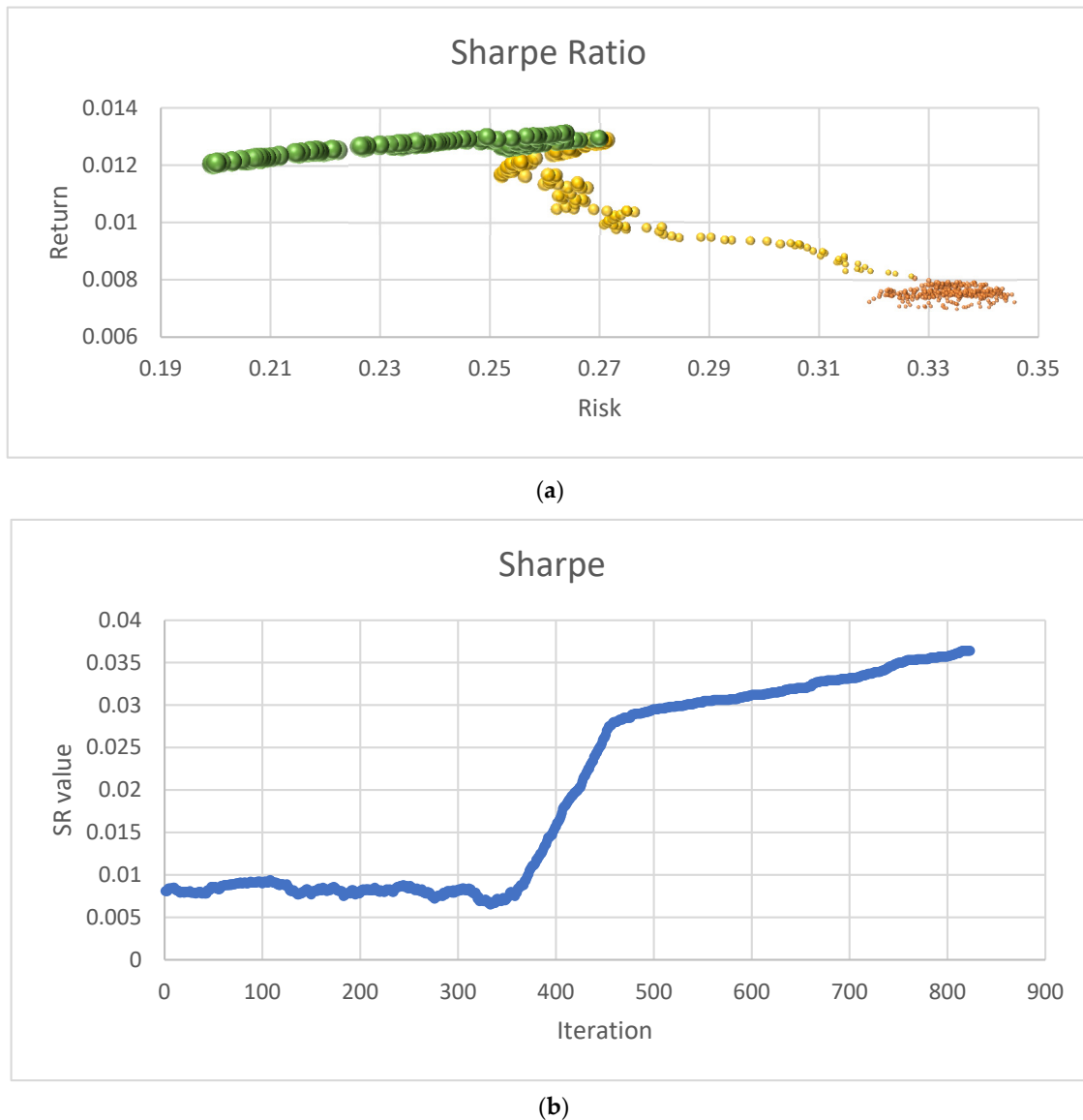


Figure 5. Algorithm performance in optimizing the Sharpe-based objective function value. (a) Efficient Frontier Risk vs. Return; (b) Sharpe Ratio values through algorithm execution.

Figure 5a presents the growth of the Sharpe Ratio value, which provides a representative example of the algorithm's performance. The size of the bubbles represents the SR, with smaller bubbles in the lower part corresponding to SR values between 0.0 and 0.01, indicating lower returns (below 0.008). In the intermediate section, the SR values range from 0.01 to 0.0299, with returns varying between 0.008 and 0.012. The larger bubbles at the top illustrate the efficient frontier, which is defined as a set of portfolios with high SR values above 0.03 and returns greater than 0.012.

The behavior of the SR value, which functions as the evaluation criterion for the TAIPO algorithm, is highlighted and shown in Figure 5b. Up to iteration 350, the values exhibit the anticipated behavior, accepting suboptimal solutions. However, after this initial phase, there is a decline in the acceptance of suboptimal solutions, leading to enhanced SR values. Notice that from iteration 500 to the conclusion of the process, the highest SR values correspond to the efficient frontier.

4.1.4. Forecasting: Weights and Optimization

For the initial portfolio forecasting, we applied the FCTA predicting tools [30]. Figure 6 shows the average result of the fourteen forecasting methods; we executed thirty executions according to the FCTA process. The sMAPE metric is presented for the initial combination using a uniform (sMAPE U) distribution for the weights. A 35.28% value was obtained for this initial solution, while the optimized combination (sMAPE O) yielded a sMAPE value of 23.60%. Notice that a large improvement of 49% of sMAPE was obtained. The forecast horizon for this case was set at 7 weeks ahead; nevertheless, MASIP works for any other short and medium forecast horizon.

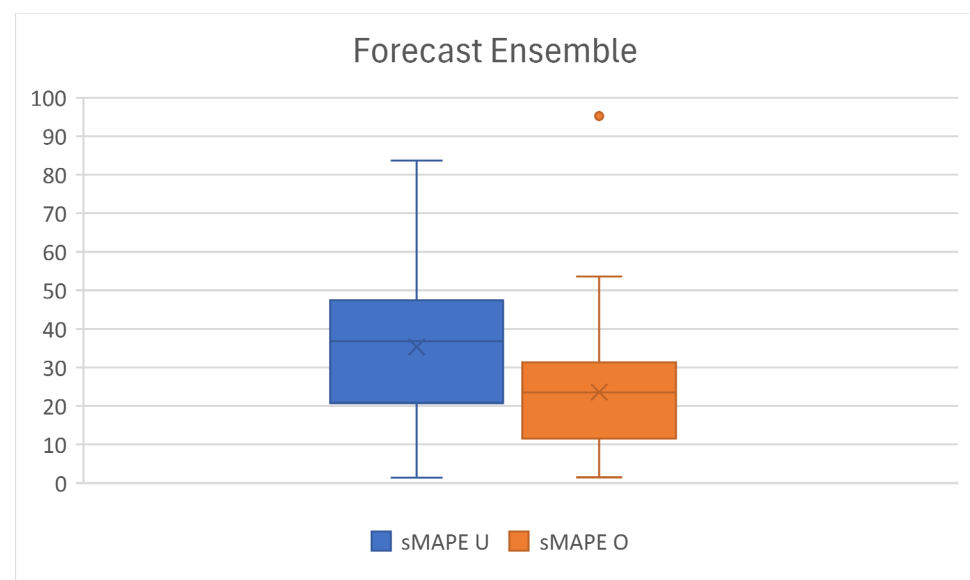


Figure 6. Forecast ensemble metric.

As illustrated in Table 2, a comparison is presented that integrates four performance error metrics.

Table 2. Forecasting performance of FCTA from 2020 to 2025.

	sMAPE	MSE	RMSE	MAPE	R ²
Mean	23.60	24.46	4.94	54.99	0.143
Std	16.5	16.22	9.03	83.20	0.178

4.1.5. Final Portfolio Optimization

In the final stage of the methodology, the combined forecasts of the assets were obtained. These were then refined using one of the integration algorithms and objective functions that have been previously mentioned. For this phase, the TAIPO algorithm was employed once more with the following metrics: Sharp Ratio: 0.0359, Return: 0.0126, Risk: 0.2170, Correlation between assets: −0.1061, and number of selected assets: 6 incorporating the complete time series as well as the forecasts calculated in the preceding stage. In terms of performance, the final portfolio presented an increase of 10.5% in returns and a reduction

in correlation value compared to the initial portfolio. The graph of the evolution of the Sharpe-based objective function is shown in Figure 7.

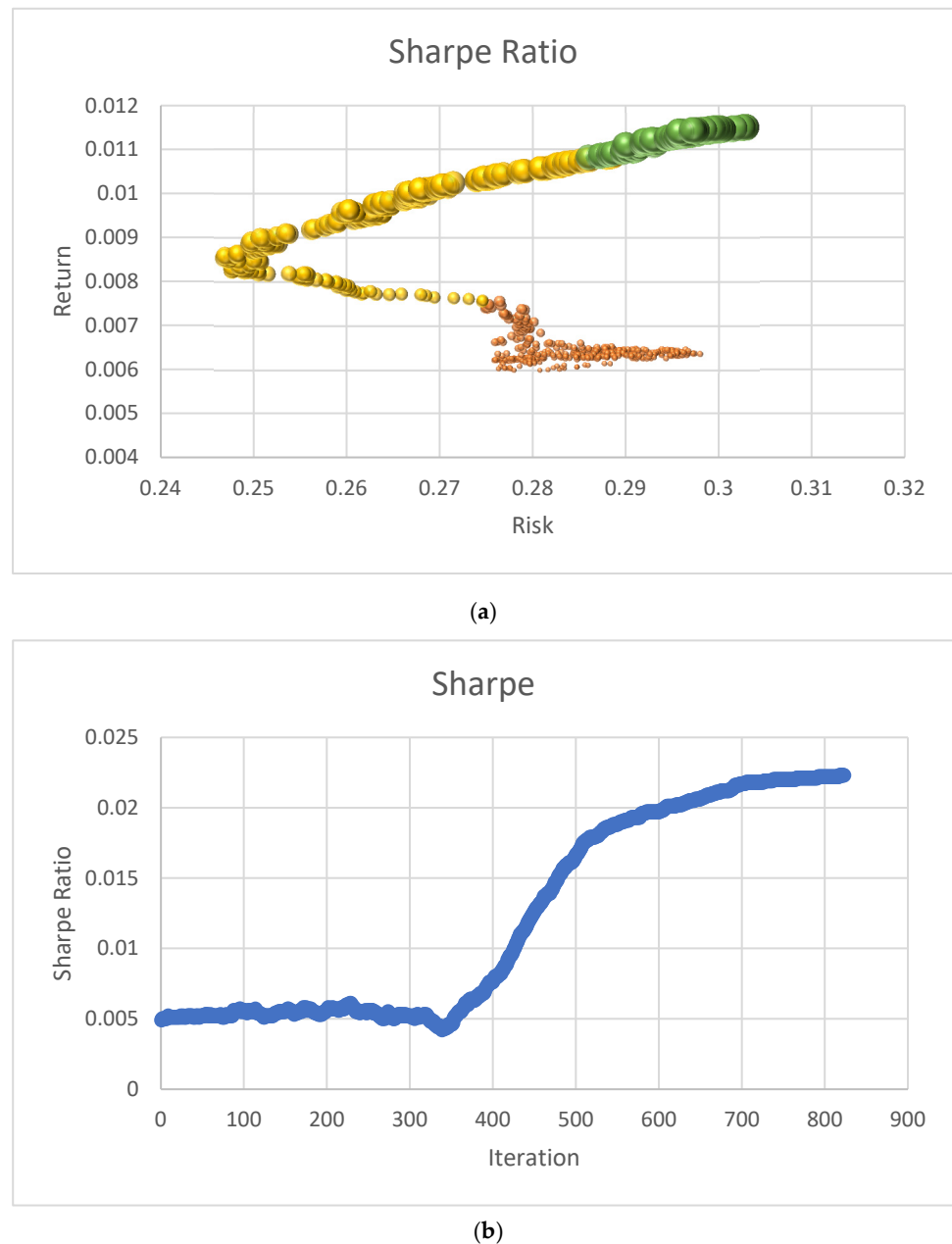


Figure 7. The Sharpe Ratio during final portfolio optimization: (a) Sample of Sharpe Ratio portfolios; (b) Sharpe Ratio through the TAIPO algorithm.

Figure 7a presents the growth of the Sharpe Ratio value, which provides a representative example of the algorithm's performance. The size of the bubbles represents the Sharpe Ratio (SR), with smaller bubbles in the lower part corresponding to SR values between 0.0 and 0.0099, indicating lower returns (below 0.00756). In the intermediate section, the SR values range from 0.01 to 0.02, with returns varying between 0.0075 and 0.0108. The larger bubbles represent the efficient frontier, which is defined as a set of portfolios with high SR values above 0.022 and returns greater than 0.011. The behavior of the SR value is presented in Figure 7b, which serves as the evaluation function for the TAIPO algorithm; up to iteration 250, the values behave as expected, accepting suboptimal solutions. However, in some periods ahead, the acceptance of suboptimal solutions decreases, resulting in better SR values. From iteration 310 to completion, the highest SR values form the efficient

frontier. We can, therefore, conclude that the algorithm finds the best solutions and begins a stagnation process. This means that after 20 repetitions of the algorithm without finding an improvement in the solution, the algorithm stops [8].

By way of comparison and observing the results of the performance of the GENPO algorithm, an experiment was carried out for the optimization of the final portfolio, producing the following results: SR: 0.0308, Return: 0.0132, Risk: 0.2399, Correlation between assets: -0.0674 , and number of selected assets: 10. In the results, we notice that, in practically all the items, there are similar values except for the assets where there are 10, as well as a slight increase in both return value and risk.

4.2. MASIP Comparison with State-of-the-Art Methods

To validate the proposed methodology, we compared MASIP with Puerto and Cesarone's model [21,23], and we used datasets from the SP500 market from 2005 to 2016. This period comprises the widest range of time common in both studies. For these datasets, we compared the average MASIP results with those reported by these authors in Table 3. In the last references, a portfolio for 52 weeks considering a zero risk-free rate was performed. The metrics used in the three alternatives were: the expected return, the risk value, the Sharpe Ratio, and the number of assets. These results are discussed in subsequent sections. We developed the programming codes necessary to emulate the former authors' works.

Table 3. MASIP results comparison with state-of-the-art methods for SP500.

Avg Results	Puerto [21]	Cesarone [23]	MASIP
Expected Return	0.002168	0.0026	0.0720
Risk	0.0231	0.0279	0.2988
Sharpe Ratio	0.094	0.0931	0.2218
Number of Assets	15	10	11

4.2.1. Data Acquisition, Preparation, and Split

We gathered datasets between 2005 and 2016 using Yahoo Finance, comprising, after data cleaning, a preparation dataset of 445 assets and 572 weekly observations. Observations were split into 70% for training, 10% each for Validation 1 and 2, and 10% for testing.

4.2.2. Initial Portfolio MARR Results

The TAIPO algorithm was used to build the initial portfolio, with the maxSharpe model. Table 4 shows a sensitivity analysis of the average results of the portfolios with different minimum acceptable rates of return (MARR). The value of the SR is directly related to the value of the MARR, since as the return value increases, the algorithm will look for a set of assets that delivers a higher return. We can also observe that as the minimum acceptable rate of return increases, the number of assets decreases; this is because, internally, the algorithm selects the assets with the highest return and that is close to the risk-free value. On the other hand, the results show a slight reduction in the correlation value.

Table 4. Initial portfolio applying MARR.

MARR %	Sharpe	Return	Risk	Correlation	Assets
15	0.1924	0.0383	0.1842	0.0305	45
20	0.2162	0.0572	0.2467	0.0142	17
25	0.2181	0.0636	0.2697	0.0091	10
30	0.2230	0.0723	0.2983	0.0135	11

The improvement in SR values achieved by the TAIPO algorithm and the target function can be seen in Figure 8.

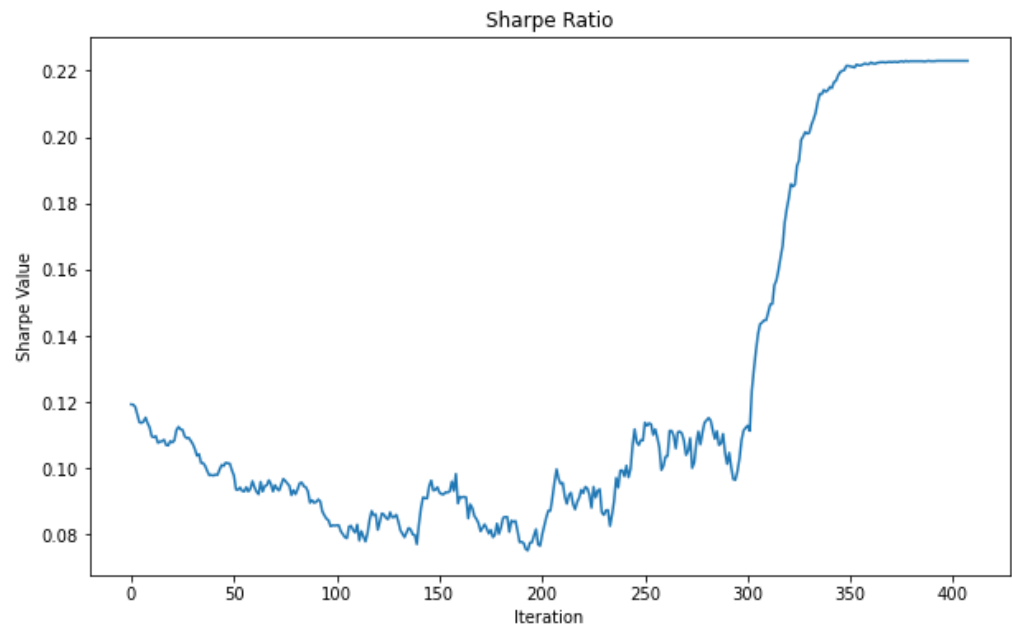


Figure 8. Sample of Sharpe Ratio in the initial portfolio.

4.2.3. Portfolio Forecasting

The initial portfolio was generated with the heuristic algorithm, and individual forecasts were obtained for each selected asset through the 14 forecasting methods within the FCTA methodology. A series of experiments were conducted, varying the forecast horizon within the range of 1 to 24 weeks.

Table 5 shows the differences in the error metric values of the forecast method combination models in a uniform distribution scheme, sMAPE U, and the contrast with the sMAPE O.

Table 5. Average sMAPE values at different forecasting horizons.

Horizon Weeks	sMAPE U	sMAPE O	Improvement
1	35.08	15.67	55%
4	10.64	7.74	27%
8	11.6	9.71	16%
24	12.59	11.68	7%

In conclusion, the application of the FCTA algorithm, once more, provides results that, at least in the case of the 1-week forecast horizon, demonstrate an improvement of up to 50%.

As is shown in Table 6, a comparison is presented that integrates four performance error metrics.

Table 6. Forecasting performance of FCTA from 2005 to 2016.

	sMAPE	MSE	RMSE	MAPE	R ²
Mean	13.15	27.47	4.62	74.93	0.241
Std	12.63	27.36	2.51	64.91	0.220

4.2.4. Final Optimized Portfolio

The subsequent section entails implementing rebalancing or optimization procedures. For this purpose, the TAIPO algorithm was used to obtain comparative results for each forecast horizon, as well as to contrast the advantages of each of the algorithms and the case study. It becomes evident that as the forecast horizon increases, there is a notable decline in the Sharpe Ratio and return values. Concurrently, the risk factor experiences a reduction,

and the number of assets exhibits a decline. The performance shown in Figure 9 illustrates the evolution of the SR within the algorithm for the integration of the final portfolio. The final portfolios generated through MASIP methodology exhibit Sharpe values that are higher than those of the state-of-the-art methods in the short term, and this advantage is further enhanced once these portfolios have been forecasted. As we observe, they maintain or even increase the SR in the presented instances.

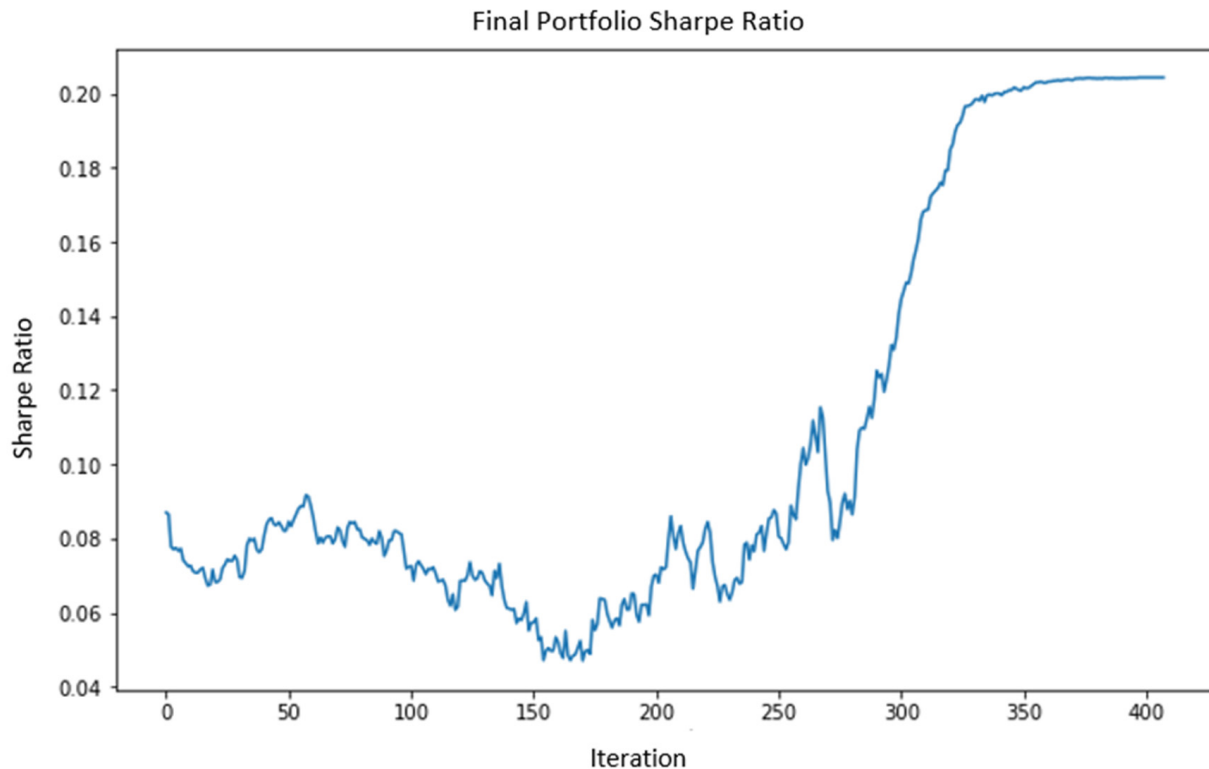


Figure 9. Sample of Sharpe Ratio in the final portfolio.

Table 7 shows the results obtained from these experiments, prioritizing the average of the highest Sharpe values found, identifying the algorithms that achieved superior performance in return terms. As shown in Table 7, the integrated portfolios with brief time horizons and the model MASIP yield outcomes surpass the contemporary standard. Conversely, in the medium-term horizons, between 8 and 24 weeks, Cesarone's model attains superior results in terms of SR. Conversely, when the focus is directed toward the variable yield, the MASIP algorithm demonstrates a markedly superior performance in comparison to the other two algorithms. This distinction can be attributed to the inherent characteristics of the MASIP algorithm and the mathematical models employed. The MASIP model is designed to optimize the SR by prioritizing the maximization of yield.

Conversely, the implementation of the MASIP methodology in the models under comparison yields analogous or superior outcomes. This is evident in the case of the Puerto approach, where predictions for the medium term (i.e., 8 to 24 weeks) indicate an SR considerably higher than those reported in the reference studied, reaching up to 415%. In contrast, Cesarone's model consistently shows superior performance, particularly in medium-term predictions, where significant improvements of up to 348% in terms of the SR are observed. As mentioned above, the MASIP methodology can be applied with various models, yielding favorable outcomes.

Table 7. Final comparison of forecasted values.

		Horizon Weeks					
		1	2	4	8	12	24
MASIP	Expected Return	0.072	0.0721	0.0721	0.06	0.0598	0.0609
	Risk	0.2991	0.2989	0.2986	0.2	0.1993	0.2091
	SR	0.2215	0.2218	0.2222	0.1752	0.1746	0.1717
	Num. Assets	11	11	11	7	7	7
Puerto [23]	Expected Return	0.0061	0.0061	0.0061	0.0339	0.0338	0.033
	Risk	0.0081	0.0081	0.008	0.0066	0.0065	0.0064
	SR	0.0829	0.0832	0.0839	0.482	0.4845	0.477
	Num. Assets	21	22	22	22	22	20
Cesarone [21]	Expected Return	0.0057	0.0056	0.0062	0.0068	0.0068	0.0067
	Risk	0.0282	0.0282	0.0321	0.0163	0.0163	0.0161
	SR	0.204	0.2018	0.195	0.4167	0.4171	0.4148
	Num. Assets	15	15	15	15	15	15

5. Conclusions

This paper proposes MASIP, an efficient methodology for investing portfolios. This methodology presents for the first time all the elements required for designing a portfolio: the selection of the best asset candidates for integrating the portfolio, the forecasting of the asset values in the investing horizon, including the best method selection for this task, the integration the best assets in this horizon and their participation, as well as the optimization model considering new constraints related to the maximum risk rate allowable and the minimum acceptable rates of return. MASIP was applied to SP500 and compared with state-of-the-art methods.

Regarding the initial phase of the portfolio construction process, it is imperative to identify the most suitable candidates. To this end, MASIP employs the TAIPO algorithm, a proposed mathematical model, and a constraint related to the lowest acceptable rate of return. This approach facilitates the identification of effective portfolios. In a subsequent stage, MASIP employs the FCTA to generate individual forecasts for each candidate asset and each of the prediction methods, selecting those methods with optimal performance and assembling them through the TAE algorithm to identify the optimal combination that minimizes the sMAPE and generates short- and medium-term forecast horizons. A rebalancing of the portfolio is subsequently implemented in the final stage, utilizing the TAIPO algorithm. This results in a portfolio comprising forecasted data and optimized weightings.

Experimental evaluations of the algorithms and data used yielded outstanding asset selection and evaluation results with SR, as well as expected return. The MASIP methodology was complemented by the forecast of the assets participating in the portfolio, thus collaborating in the final optimization, resulting in a final weighted, forecasted, and optimized portfolio.

A comparative analysis was conducted between MASIP and two alternative investment models, the Puerto and Cesarone models. The outcomes of this analysis were prioritized based on the meaning of the highest Sharpe values identified. The results of this study demonstrate that the MASIP algorithm exhibited robust performance in terms of returns. As the forecast horizon grows, the SR and return values drop significantly, concurrently with a drop in the risk factor and the asset number. The SR values from the created portfolios exceed those of the state-of-the-art methods, and forecasting enhances this advantage. The superiority of MASIP performance relative to the other two algorithms lies in its design, which prioritizes the maximization of short-term yield to optimize SR. Furthermore, the efficacy of the MASIP methodology in enhancing the performance of

other methodologies is evident from the observed improvement in SR performance. This observation supports the potential for further experimentation with the same approach.

For future research, the application of other optimization methods is proposed, including other objective functions and metrics, both for asset preselection and portfolio rebalancing, the experimentation of novel methods or forecasting ensembles in financial time series, as well as the application of the MASIP methodology in other markets.

Author Contributions: Conceptualization, J.F.-S. and J.P.-A.; methodology and software, J.P.-A.; validation, G.C.-V., J.G.-B. and J.F.-S.; formal analysis, J.F.-S. and J.G.-B.; investigation, J.P.-A. and J.F.-S.; writing—original draft preparation, J.P.-A.; writing—review and editing, J.G.-B., J.F.-S., J.P.S.H. and G.C.-V.; supervision, J.F.-S. and J.G.-B.; project administration, J.F.-S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: <https://github.com/DrJuanFraustoSolis/TAIPO.git> (accessed on 30 January 2025).

Acknowledgments: The authors would like to acknowledge SECIHTI (Secretariat of Science, Humanities, Technology and Innovation), TecNM (National Technology of Mexico), Madero City Technological Institute, and the National Laboratory of Information Technologies (LaNTI) for access to the cluster.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Guerard, J.B. *Introduction to Financial Forecasting in Investment Analysis*, 1st ed.; Springer: New York, NY, USA, 2013. [CrossRef]
2. Andersen, J.V. Investment Decision Making in Finance, Models of. In *Encyclopedia of Complexity and Systems Science*; Meyers, R.A., Ed.; Springer: New York, NY, USA, 2009; pp. 4971–4983. [CrossRef]
3. Markowitz, H. Portfolio Selection. *J. Financ.* **1952**, *7*, 77–91. [CrossRef]
4. Sharpe, W.F. Mutual Fund Performance. *J. Bus.* **1966**, *39*, 119–138. [CrossRef]
5. Scholz, H. Refinements to the Sharpe ratio: Comparing alternatives for bear markets. *J. Asset Manag.* **2007**, *7*, 347–357. [CrossRef]
6. Zakamouline, V.; Koekebakker, S. Portfolio performance evaluation with generalized Sharpe ratios: Beyond the mean and variance. *J. Bank. Financ.* **2009**, *33*, 1242–1254. [CrossRef]
7. Landete, M.; Monge, J.F.; Ruiz, J.L.; Segura, J.V. *Sharpe Portfolio Using a Cross-Efficiency Evaluation BT—Data Science and Productivity Analytics*; Charles, V., Aparicio, J., Zhu, J., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 415–439. [CrossRef]
8. Solis, J.F.; Aldaz, J.L.P.; Del Angel, M.G.; Barbosa, J.G.; Valdez, G.C. SAIPO-TAIPPO and Genetic Algorithms for Investment Portfolios. *Axioms* **2022**, *11*, 42. [CrossRef]
9. Bates, J.M.; Granger, A.W.J. The Combination of Forecasts. *Oper. Res. Soc.* **1969**, *20*, 451–468. [CrossRef]
10. Winkler, R.L.; Makridakis, S. The Combination of Forecasts. *J. R. Stat. Soc. Ser. A* **1983**, *146*, 150–157.
11. Dueck, G.; Scheuer, T. Threshold accepting: A general purpose optimization algorithm appearing superior to simulated annealing. *J. Comput. Phys.* **1990**, *90*, 161–175. [CrossRef]
12. Winker, P.; Maringer, D. The Threshold Accepting Optimisation Algorithm in Economics and Statistics. In *Optimisation, Econometric and Financial Analysis*; Kontoghiorghes, E.J., Gatu, C., Eds.; Springer: Berlin/Heidelberg, Germany, 2007; pp. 107–125.
13. Choueifaty, Y.; Coignard, Y. Toward maximum diversification. *J. Portf. Manag.* **2008**, *35*, 40–51. [CrossRef]
14. Tella, R.; Rogel-Salazar, J. Portfolio Construction Based on Implied Correlation Information and VAR. *SSRN Electron. J.* **2013**, *12*, 125–144. [CrossRef]
15. Wang, S.; Xia, Y. *Portfolio Selection and Asset Pricing*, 1st ed.; Springer: Berlin/Heidelberg, Germany, 2002. [CrossRef]
16. Gilli, M.; Këlezi, E. A Heuristic Approach to Portfolio Optimization. Available online: <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=f6500f280e2c91a2ce11d0e46f90424aac749f4> (accessed on 31 January 2025).
17. Masese, J.M.; Othieno, F.; Njenga, C. Portfolio Optimization under Threshold Accepting: Further Evidence from a Frontier Market. *J. Math. Financ.* **2017**, *7*, 941–957. [CrossRef]
18. Rangel-González, J.A.; Frausto-Solis, J.; González-Barbosa, J.J.; Pazos-Rangel, R.A.; Fraire-Huacuja, H.J. Comparative study of ARIMA methods for forecasting time series of the mexican stock exchange. *Stud. Comput. Intell.* **2018**, *749*, 475–485. [CrossRef]

19. Vijh, M.; Chandola, D.; Tikkiwal, V.A.; Kumar, A. Stock Closing Price Prediction using Machine Learning Techniques. *Procedia Comput. Sci.* **2020**, *167*, 599–606. [\[CrossRef\]](#)
20. Singh, P.; Jha, M. Portfolio Optimization Using Novel EW-MV Method in Conjunction with Asset Preselection. *Comput. Econ.* **2024**, *64*, 3683–3712. [\[CrossRef\]](#)
21. Cesarone, F.; Scozzari, A.; Tardella, F. An optimization–diversification approach to portfolio selection. *J. Glob. Optim.* **2020**, *76*, 245–265. [\[CrossRef\]](#)
22. Cesarone, F.; Mottura, C.; Ricci, J.M.; Tardella, F. On the Stability of Portfolio Selection Models. 2018, pp. 1–27. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3420081 (accessed on 31 January 2025).
23. Puerto, J.; Rodríguez-Madrena, M.; Scozzari, A. Clustering and portfolio selection problems: A unified framework. *Comput. Oper. Res.* **2020**, *117*, 104891. [\[CrossRef\]](#)
24. Kaczmarek, T.; Perez, K. Building portfolios based on machine learning predictions. *Econ. Res. Istraz.* **2021**, *35*, 1–19. [\[CrossRef\]](#)
25. Zhang, A. Portfolio Optimization of Stocks—Python-Based Stock Analysis. *Int. J. Educ. Humanit.* **2023**, *9*, 32–38. [\[CrossRef\]](#)
26. Ma, Y.; Han, R.; Wang, W. Portfolio optimization with return prediction using deep learning and machine learning. *Expert Syst. Appl.* **2021**, *165*, 113973. [\[CrossRef\]](#)
27. Martínez-Barbero, X.; Cervelló-Royo, R.; Ribal, J. Portfolio Optimization with Prediction-Based Return Using Long Short-Term Memory Neural Networks: Testing on Upward and Downward European Markets. *Comput. Econ.* **2024**, *65*, 1479–1504. [\[CrossRef\]](#)
28. Elliott, G.; Timmermann, A. Economic Forecasting. *J. Econ. Lit.* **2008**, *46*, 3–56. [\[CrossRef\]](#)
29. Centeno, V.; Georgiev, I.R.; Mihova, V.; Pavlov, V. Price forecasting and risk portfolio optimization. *AIP Conf. Proc.* **2019**, *2164*, 060006. [\[CrossRef\]](#)
30. Frausto-Solis, J.; Rodriguez-Moya, L.; González-Barbosa, J.; Castilla-Valdez, G.; Ponce-Flores, M. FCTA: A Forecasting Combined Methodology with a Threshold Accepting Approach. *Math. Probl. Eng.* **2022**, *2022*, 6206037. [\[CrossRef\]](#)
31. Zhang, Y.; Qu, H.; Wang, W.; Zhao, J. A Novel Fuzzy Time Series Forecasting Model Based on Multiple Linear Regression and Time Series Clustering. *Math. Probl. Eng.* **2020**, *2020*, 9546792. [\[CrossRef\]](#)
32. Rao, P.S.; Srinivas, K.; Mohan, A.K. *A Survey on Stock Market Prediction Using Machine Learning Techniques BT—ICDSMLA 2019*; Kumar, A., Paprzycki, M., Gunjan, V.K., Eds.; Springer: Singapore, 2020; pp. 923–931.
33. Becha, M.; Dridi, O.; Riabi, O.; Benmessaoud, Y. Use of Machine Learning Techniques in Financial Forecasting. In Proceedings of the 2020 International Multi-Conference on: “Organization of Knowledge and Advanced Technologies” (OCTA), Tunis, Tunisia, 6–8 February 2020; pp. 1–6. [\[CrossRef\]](#)
34. He, K.; Yang, Q.; Ji, L.; Pan, J.; Zou, Y. Financial Time Series Forecasting with the Deep Learning Ensemble Model. *Mathematics* **2023**, *11*, 1054. [\[CrossRef\]](#)
35. Hyndman, R.J.; Athanasopoulos, G. Forecasting: Principles and Practice. In *Forecasting: Principles and Practice*; Econometrics & Business Statistics: Monash, Australia, 2018; p. 504. Available online: <https://otexts.com/fpp2/> (accessed on 31 January 2025).
36. Montero-Manso, P.; Athanasopoulos, G.; Hyndman, R.J.; Talagala, T.S. FFORMA: Feature-based forecast model averaging. *Int. J. Forecast.* **2020**, *36*, 86–92. [\[CrossRef\]](#)
37. Petropoulos, F.; Apiletti, D.; Assimakopoulos, V.; Babai, M.Z.; Barrow, D.K.; Ben Taieb, S.; Bergmeir, C.; Bessa, R.J.; Bijak, J.; Boylan, J.E.; et al. Forecasting: Theory and practice. *Int. J. Forecast.* **2022**, *38*, 705–871. [\[CrossRef\]](#)
38. Estrada-Patiño, E.; Castilla-Valdez, G.; Frausto-Solis, J.; González-Barbosa, J.; Sánchez-Hernández, J.P. A Novel Approach for Temperature Forecasting in Climate Change Using Ensemble Decomposition of Time Series. *Int. J. Comput. Intell. Syst.* **2024**, *17*, 253. [\[CrossRef\]](#)
39. Wasserbacher, H.; Spindler, M. Machine learning for financial forecasting, planning and analysis: Recent developments and pitfalls. *Digit. Financ.* **2022**, *4*, 63–88. [\[CrossRef\]](#)
40. Dharrao, D.S.; Bongale, A.M.; Deokate, S.T.; Doreswamy, D.; Bhat, S.K. Forecasting Stock Market Prices Using Machine Learning and Deep Learning Models: A Systematic Review, Performance Analysis and Discussion of Implications. *Int. J. Financial Stud.* **2023**, *11*, 94. [\[CrossRef\]](#)
41. Cheng, L.; Shadabfar, M.; Khoojine, A.S. A State-of-the-Art Review of Probabilistic Portfolio Management for Future Stock Markets. *Mathematics* **2023**, *11*, 1148. [\[CrossRef\]](#)
42. Althöfer, I.; Koschnick, K.-U. On the convergence of ‘Threshold Accepting’. *Appl. Math. Optim.* **1991**, *24*, 183–195. [\[CrossRef\]](#)
43. Winker, P. *The Stochastics of Threshold Accepting: ANALYSIS of an Application to the Uniform Design Problem BT—Compstat 2006—Proceedings in Computational Statistics*; Rizzi, A., Vichi, M., Eds.; Physica: Heidelberg, HD, USA, 2006; pp. 495–503.
44. Gilli, M.; Kellezi, E. The Threshold Accepting Heuristic for Index Tracking. In *Financial Engineering, E-commerce and Supply Chain*; Springer: Boston, MA, USA, 2011; pp. 1–18. [\[CrossRef\]](#)
45. Ta, V.D.; Liu, C.M.; Tadesse, D.A. Portfolio optimization-based stock prediction using long-short term memory network in quantitative trading. *Appl. Sci.* **2020**, *10*, 437. [\[CrossRef\]](#)
46. Kelleher, J.D.; Tierney, B. *Data Science*; The MIT Press: Cambridge, MA, USA, 2018.

47. Provost, F.; Fawcett, T. *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking*, 1st ed.; O'Reilly Media: Sebastopol, CA, USA, 2013.
48. Kuhn, M.; Johnson, K. *Feature Engineering and Selection: A Practical Approach for Predictive Models*, 1st ed.; CRC Chapman and Hall: Boca Raton, FL, USA, 2019.
49. S Ross, A.; Westerfield, R.W.; Jordan, B.D. *Fundamentals of Corporate Finance*, 13th ed.; McGraw Hill: New York, NY, USA, 2021.
50. Box, G.E.P.; Jenkins, G.M.; Reinsel, G.C. *Time Series Analysis: Forecasting and Control*; Prentice Hall: Englewood Cliffs, NJ, USA, 1994; Volume SFB 373, Chapter 5; pp. 837–900.
51. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016; Available online: <http://www.deeplearningbook.org> (accessed on 31 January 2025).
52. Markowitz, H. Foundations of portfolio theory. In *Harry Markowitz: Selected Works*; World Scientific Publishing Company: Singapore, 2009; Volume 46, pp. 481–490. [CrossRef]
53. Markowitz, H.M. *Portfolio Selection*; Yale University Press: New Haven, CT, USA, 1959; Available online: <http://www.jstor.org/stable/j.ctt1bh4c8h> (accessed on 31 January 2025).
54. Yu, L.; Wang, S.; Lai, K.K. Multi-attribute portfolio selection with genetic optimization algorithms. *INFOR* **2009**, *47*, 23–30. [CrossRef]
55. Jaganathan, S.; Prakash, P.K.S. A combination-based forecasting method for the M4-competition. *Int. J. Forecast.* **2020**, *36*, 98–104. [CrossRef]
56. Cawood, P.; Van Zyl, T. Evaluating State-of-the-Art, Forecasting Ensembles and Meta-Learning Strategies for Model Fusion. *Forecasting* **2022**, *4*, 732–751. [CrossRef]
57. Barnard, G.A. New Methods of Quality Control New Methods of Quality Control. *J. R. Stat. Soc.* **1963**, *126*, 255–258.
58. Flores, B.E. A Pragmatic View of Accuracy Measurement in Forecasting. *Omega* **1986**, *14*, 93–98. [CrossRef]
59. Armstrong, J.S. Long-range Forecasting: From Crystal Ball to Computer. In *Wiley-Interscience Publication*; Wiley: Hoboken, NJ, USA, 1985; Available online: <https://books.google.com.mx/books?id=t7V8AAAAIAAJ> (accessed on 31 January 2025).
60. Makridakis, S. Accuracy measures: Theoretical and practical concerns. *Int. J. Forecast.* **1993**, *9*, 527–529. [CrossRef]
61. Makridakis, S.; Hibon, M. The M3-Competition: Results, conclusions and implications. *Int. J. Forecast.* **2000**, *16*, 451–476. [CrossRef]
62. Armstrong, J.S.; Franke, G. *Principles of Forecasting: A Handbook for Researchers and Practitioners*, 1st ed.; Springer: Boston, MA, USA, 2001.
63. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. The M4 Competition: 100,000 time series and 61 forecasting methods. *Int. J. Forecast.* **2020**, *36*, 54–74. [CrossRef]
64. Frausto-Solis, J.; Román, E.F.; Romero, D.; Soberon, X.; Liñán-García, E. *Analytically Tuned Simulated Annealing Applied to the Protein Folding Problem BT—Computational Science—ICCS 2007*; Shi, Y., van Albada, G.D., Dongarra, J., Sloot, P.M.A., Eds.; Springer: Berlin/Heidelberg, Germany, 2007; pp. 370–377.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.