Maestría en
Ciencia de Datos y Máquinas de Aprendizaje con mención en Inteligencia Artificial

Trabajo previo a la obtención de título de Magister en Ciencia de Datos y Máquinas de Aprendizaje con mención en Inteligencia Artificial

AUTORES:

ALTAMIRANO LOPEZ PEDRO ANDRES
AVILES GONZALEZ JONNATAN FERNANDO
BALDEON EGAS PAUL FRANCISCO
DELGADO POZO JOSELYN IVONE

TUTOR:

Alejandro Cortés López
Iván Reyes Chacón

TEMA

Análisis de anomalías en datos climáticos con aprendizaje no supervisado en la ciudad de Cuenca

Enero - 2025

Quito - Ecuador

REINVENTEMOS EL FUTURO

Certificación de autoría

Nosotros, **Pedro Andrés Altamirano López, Jonnatan Fernando Avilés González, Joselyn Ivone Delgado Pozo, Paúl Francisco Baldeón Egas** declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que no ha sido presentado anteriormente para ningún grado o calificación profesional y que se ha consultado la bibliografía detallada. Cedemos nuestros derechos de propiedad intelectual a la Universidad Internacional del Ecuador (UIDE), para que sea publicado y divulgado en internet, según lo establecido en la Ley de Propiedad Intelectual, su reglamento y demás disposiciones legales.

Firma del magíster
Joselyn Ivone Delgado Pozo

Firma del magíster
Jonnatan Fernando Avilés González

Firma del magíster
Paúl Francisco Baldeón Egas

Firma del magíster
Pedro Andrés Altamirano López

Autorización de Derechos de Propiedad Intelectual

Nosotros, **Pedro Andrés Altamirano López, Jonnatan Fernando Avilés González**, **Joselyn Ivone Delgado Pozo, Paúl Francisco Baldeón Egas** y en calidad de autores del trabajo de investigación titulado **Análisis de anomalías en datos climáticos con aprendizaje no supervisado en la ciudad de Cuenca**, autorizamos a la Universidad Internacional del Ecuador (UIDE) para hacer uso de todos los contenidos que nos pertenecen o de parte de los que contiene esta obra, con fines estrictamente académicos o de investigación. Los derechos que como autores nos corresponden, lo establecido en los artículos 5, 6, 8, 19 y demás pertinentes de la Ley de Propiedad Intelectual y su Reglamento en Ecuador.
D. M. Quito, julio de 2025

Firma del magíster
Joselyn Ivone Delgado Pozo

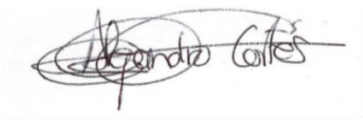
Firma del magíster
Jonnatan Fernando Avilés González

Firma del magíster
Paúl Francisco Baldeón Egas

Firma del magíster
Pedro Andrés Altamirano López

Aprobación de dirección y coordinación del programa

Nosotros Alejandro Cortés e Iván Reyes, declaramos que : Jonnatan Fernando Avilés González, Joselyn Ivone Delgado Pozo, Paúl Francisco Baldeón Egas y Pedro Andrés Altamirano López son los autores exclusivos de la presente investigación y que ésta es original, auténtica y personal de ellos.



Alejandro Cortés
Director de la
Maestría en Ciencia de datos y Máquinas de aprendizaje con mención en Inteligencia Artificial

Iván Reyes
Coordinador de la
Maestría en Ciencia de datos y Máquinas de aprendizaje con mención en Inteligencia Artificial

DEDICATORIA

Este trabajo está dedicado a los que nos enseñaron que el conocimiento no solo se gana, sino que se honra con responsabilidad, disciplina y agradecimiento. A los que creyeron en nosotros, incluso en momentos en los que la visión era borrosa y a menudo nos recordaron que todo logro no tiene valor si no se comparte.

A nuestras familias, cuyo amor incondicional es la piedra angular de este esfuerzo. Por cada palabra de aliento, cada sacrificio silencioso y cada acto de apoyo, aunque no visible para los demás, fueron quienes soportaron cada trayecto de estudio y lograr superarlo.

A nuestros docentes, quienes nos guiaron y despertaron en nosotros la curiosidad por descubrir, cuestionar y cambiar el mundo mediante el saber. Su orientación y ejemplo nos enseñaron que alcanzar una meta académica es, sobre todo, un compromiso con la sociedad y con quienes vendrán después.

A cada persona que, de una forma u otra, hizo posible que hoy celebremos este logro con orgullo y gratitud. Esta dedicatoria es para ustedes, porque cada apoyo, cada charla honesta y cada palabra de ánimo nos recordaron que aprender cobra sentido cuando se convierte en la semilla que permite a otros seguir soñando.

AGRADECIMIENTOS

Este logro en la presente investigación ha sido posible con la suma de voluntades, afectos y esfuerzos de muchas personas e instituciones que contribuyeron a que hoy culminemos esta etapa, en especial a la Universidad del Azuay, quien nos ha con los datos de su estación meteorológica.

Expresamos nuestra gratitud a nuestros docentes y directores de tesis, quienes, con su paciencia, sus observaciones oportunas y su ejemplo profesional, enriquecieron nuestra mirada crítica y nos retaron a ir siempre un paso más allá. Su guía y exigencia fue clave para mantener la coherencia y la calidad de este trabajo.

Agradecemos especialmente a nuestras familias, por ser soporte en los días de cansancio y desaliento. Por comprender las ausencias, por celebrar los pequeños logros y por recordarnos, aún en silencio, la importancia de ser fieles a nuestros sueños y que con disciplina y perseverancia se logran las metas.

Finalmente, extendemos este agradecimiento a nuestros compañeros, amigos y colegas, que compartieron ideas, lecturas, conversaciones y espacios de reflexión. Cada aporte, cada intercambio de experiencias y cada palabra de apoyo alimentaron nuestra motivación y fortalecieron la convicción de que este logro no es individual, sino colectivo.

RESUMEN

El presente estudio aborda el análisis de registros climáticos de la ciudad de Cuenca, utilizando algoritmos de aprendizaje no supervisado para descubrir posibles anomalías que puedan alterar la calidad de la información meteorológica. La motivación principal radica en que, sin un control riguroso, errores de medición o fenómenos atípicos pueden distorsionar los datos y afectar decisiones sobre prevención de riesgos y planificación ambiental. Para responder a esta necesidad, se estructuró un diseño metodológico de enfoque cuantitativo, con carácter exploratorio y descriptivo, que abarca desde la limpieza y transformación de datos hasta la validación de resultados. Una parte esencial fue la transformación de la variable hora, tratándola como variable cíclica mediante funciones seno y coseno, considerando que indicadores como la radiación solar y el índice ultravioleta presentan comportamientos periódicos que se analizan preferentemente por hora. Para segmentar y verificar estacionalidad se aplicaron técnicas como DBSCAN y modelos SARIMA, elegidos por su flexibilidad para detectar agrupaciones en conjuntos de datos sin etiquetas. La validación combinó de análisis de error y comportamiento interno de grupos. Además, se propone algunos modelos para realizar la predicción del índice de radiación ultravioleta. Los hallazgos muestran trayectorias diarias de aumento y disminución de radiación, junto con registros que alertan sobre valores anómalos, además se muestra que modelos como Random Forest y SVM permiten predecir sin caer en el sobreajuste, y bajo criterios de datos anómalos. Con esta investigación se aporta una base metodológica para optimizar datos meteorológicos y se abre camino a nuevas aplicaciones que fortalezcan sistemas de monitoreo, predicción y gestión de riesgos climáticos en la región del Azuay.

Palabras Claves:
Aprendizaje no supervisado, clustering, anomalías, datos climáticos, Cuenca

ABSTRACT

The present study deals with the analysis of climate records from the city of Cuenca, using unsupervised learning algorithms to discover possible anomalies that may alter the quality of meteorological information. The main motivation lies in the fact that, without a rigorous control, measurement errors or atypical phenomena can distort the data and affect decisions on risk prevention and environmental planning. To respond to this need, a methodological design with a quantitative approach was structured, with an exploratory and descriptive character, ranging from data cleaning and transformation to validation of results. An essential part was the transformation of the hour variable, treating it as a cyclic variable by means of sine and cosine functions, considering that indicators such as solar radiation and the ultraviolet index present periodic behaviors that are preferably analyzed hourly. To segment and verify seasonality, techniques such as DBSCAN and SARIMA models, chosen for their flexibility in detecting clusters in unlabeled data sets, were applied. Validation combined error analysis and internal behavior of clusters. In addition, some models are proposed to perform ultraviolet radiation index prediction. The findings show daily trajectories of radiation increase and decrease, together with records that alert about anomalous values, and it is also shown that models such as Random Forest and SVM allow prediction without falling into over-fitting, and under anomalous data criteria. This research provides a methodological basis for optimizing meteorological data and opens the way for new applications to strengthen monitoring, prediction and climate risk management systems in the Azuay region.

Keywords:
Unsupervised learning, clustering, anomalies, climate data, Cuenca

TABLA DE CONTENIDOS

LISTA DE TABLAS (Índice de tablas)

LISTA DE FIGURAS (Índice de figuras)

CAPÍTULO 1:

1. INTRODUCCIÓN

Cuenca, Ecuador es una ciudad que ha crecido tanto empresarialmente como urbanísticamente, esto ha hecho que la ciudad presente varios cambios, uno de los más notorios son los cambios en el microclima de la ciudad. Este tema ha crecido en el interés de la ciudad, que tanto entes gubernamentales como universitarios han puesto interés en el monitorio de los fenómenos climáticos. El enfoque principal es que el conocer sobre las anomalías que se pueden presentar en el clima, sus registros y posibles predicciones, se está convirtiendo en un requisito para tomas de decisiones, que permitan un buen vivir social (Fernández y Torres, 2019).

Es importante mencionar que conocer respecto a los problemas del clima es un campo que ha sido ampliamente estudiado, pero cada ciudad y localidad tienen sus propias anomalías y eventos extremos, que causan que las generalizaciones en estudios no necesariamente sean replicables en todas las zonas.

Ante estos sucesos, el campo del aprendizaje de máquinas a permitido realizar avances en el manejo de datos meteorológicos, por ejemplo, el aprendizaje automático no supervisado se ha catalogado como la vertiente a seguir para trabajar con datos no etiquetados, y encontrar sus relaciones. Considerando que existen en los datos problemas desde la recolección hasta la interpretación (Racah et al., 2017).

Entre los métodos de aprendizaje no supervisado más comunes se puede mencionar, K-means, DBSCAN, Gaussian Mixture Models, Clustering, Random Forest, entre otros, además de métodos más robustos a través de redes neuronales. Los investigadores también han propuesto una serie de métricas para valorar estos métodos

pudiendo mencionar índices como el coeficiente de silueta, o medida de errores como el MSE, además de medidas de parámetros específicos como el coeficiente de determinación. Todo esto con el objetivo de permitir hace una comparación justa y apropiada entre modelos aplicados (Fernández y Torres, 2019).

Por otra parte, métodos más avanzados como los autoencoders, los modelos basados en aislamiento (Isolation Forests) o los modelos basados en densidad han sido empleados con éxito para identificar anomalías en series temporales multivariantes, ya que son capaces de detectar desviaciones sutiles en el comportamiento histórico de variables meteorológicas. Según Wyld et. al (2024) “estas técnicas pueden ser entrenadas en condiciones normales para luego detectar, sin etiquetas, desviaciones inusuales con alta precisión” (p. 147).

En el caso particular de Cuenca, ciudad situada en una región andina con alta variabilidad climática, el uso de técnicas no supervisadas resulta adecuado para estudiar patrones atmosféricos complejos. La implementación de estos métodos permitirá identificar eventos anómalos, errores instrumentales o registros espurios, generando conocimiento útil tanto para los operadores de estaciones meteorológicas como para las instituciones responsables de la gestión de riesgos y del diseño de políticas ambientales sostenibles.

Este proyecto emplea un enfoque integral de ciencia de datos, incorporando técnicas de preprocesamiento, limpieza, transformación de datos, selección e ingeniería de características, aplicación de algoritmos no supervisados, evaluación de la calidad de los modelos y visualización de resultados, además de consideraciones estadísticas. La sinergia entre los métodos estadísticos, el aprendizaje automático y las métricas de validación permitirá identificar posibles patrones anómalos en los datos climáticos recolectados entre julio de 2024 y abril de 2025 en la estación meteorológica de la Universidad del Azuay.

1.1. Definición del proyecto

Este estudio se enfoca en el análisis de datos del clima en la ciudad de Cuenca, obtenidos en la estación meteorológica de la Universidad del Azuay, ubicada en las calles Av. 24 de Mayo 7-77 y Hernán Malo, Cuenca. Los datos que se analizarán serán desde 1 de Julio del 2024 a las 0 horas con 0 minutos hasta el 13 de abril del 2025 a las 0 horas con 0 minutos, es decir aproximadamente diez meses de datos. Se busca no solo contribuir a mejorar la calidad de los registros meteorológicos, sino también a facilitar el desarrollo de modelos predictivos más robustos para el análisis climático en la región del Azuay, a través de técnicas de aprendizaje no supervisado considerando las posibles detecciones de anomalías.

1.2. Justificación e importancia del trabajo de investigación

El análisis de datos climáticos es esencial para comprender patrones meteorológicos y detectar eventos inusuales que puedan indicar cambios en el clima o fallos en los sensores de medición, garantizando su buen funcionamiento o previniendo desastres naturales. En la actualidad, abordar el manejo de datos climaticos a través del aprendizaje automático se ha vuelto un requisito ya que la identificación de las anomalías, predicciones, y el manejo de grandes volúmenes de datos hace que métodos de trabajo mecanizados y recurrentes sean necesarios. Por ejemplo, Chandola et al. (2009), menciona la importancia de las técnicas de aprendizaje no supervisado para manejar anomalías en datos de carácter meteorológico, sin necesidad de supervisión y a través de instrumentos de control.

Este estudio se enfoca a Cuenca – Ecuador, que es una ciudad con particularidades geográficas al estar entre valles, y zonas cálidas, y entre zonas frías y altas. Además, la ciudad tiene cuatro ríos que atraviesan toda la localidad provocando un microclima diferente por zonas incluso de la propia urbanidad. Debido a estas condiciones cuando se evalúan criterios climatológicos suelen darse valores fuera de lo común para condiciones normales,

pero comunes para condiciones propias de la zona. Por tal motivo considerar métodos de aprendizaje no supervisado sería muy útil para contribuir a la toma de decisiones en la ciudad.

Alnutefy y Alsuwayh (2024), en su investigación menciona como la planificación urbana puede verse beneficiada a través del manejo, procesamiento, uso de datos meteorológicos por medio de técnicas como DBSAN, LSTM, y enfoques robustos de redes neuronales, además que permiten su reproducibilidad y portabilidad a los diversos sensores y estaciones.

1.3. Alcance

El presente trabajo de investigación tiene como alcance el desarrollo de un sistema de análisis para la detección de anomalías en datos climáticos registrados por la estación meteorológica de la Universidad del Azuay (UDA-1), localizada en la ciudad de Cuenca, Ecuador. Se utilizaron datos obtenidos de variables meteorológicas registradas en intervalos horarios durante el período comprendido entre el julio de 2024 y abril de 2025, es decir alrededor de 10 meses, el enfoque será principalmente a la variable índice de radiación ultravioleta.

El estudio considera las siguientes actividades:

1. Preprocesamiento de datos: Se realizará la limpieza y depuración de los datos, manejo de valores faltantes y atípicos, verificación de duplicados y errores debido a los sensores (de existir), imputación de datos, y estandarización o normalización de variables de ser requeridos.
2. Análisis exploratorio de datos (EDA): Se realizará un enfoque estadístico descriptivo y gráfico para entender la distribución, variabilidad y relaciones entre variables. Se identifican valores atípicos y anomalías posibles.

3. Ingeniería de características: Se generan y transforman variables derivadas según la literatura u opinión de expertos.
4. Entrenamiento y pruebas en modelos de aprendizaje:
    a. Modelos para detección de anomalías: Se aplican algoritmos no supervisados como DBSCAN, y One-Class SVM. También se considera la detección por modelos estadísticos como SARIMA.
    b. Modelos para predicción: Se entrena modelos como LSTM y redes neuronales MLP para evaluar su capacidad de predecir valores futuros y detectar desviaciones con respecto a los patrones esperados.
5. Entrenamiento y validación de modelos: El conjunto de datos se divide en subconjuntos de entrenamiento y evaluación, asegurando el comportamiento de datos como una serie de tiempo. Se escalan los datos del conjunto de entrenamiento. Se entrenan los modelos utilizando métricas de error y desempeño para su comparación.
6. Evaluación de modelos: Se aplican métricas basadas en el error como MSE, MAE, y métricas de clustering como el índice de silueta o el índice de Davies-Bouldin, considerando además tiempos de gasto computacional, para la base de datos obtenida.
7. Análisis de resultados: Se interpretan las anomalías detectadas, contrastándolas con condiciones climáticas reales o posibles errores de medición. Se documentan los hallazgos, considerando las métricas obtenidas, los tiempos de cómputo.

En función del alcance descrito, este trabajo de investigación contempla los siguientes entregables concretos:

- Base de datos preprocesada: archivo estructurado y curado, listo para su reutilización en estudios futuros o por otras instituciones académicas.
- Modelos entrenados y validados: conjuntos de scripts reproducibles con los algoritmos implementados y sus respectivos parámetros optimizados.

- Informe de resultados y visualizaciones: gráficos interactivos y resúmenes visuales para facilitar la interpretación de patrones y anomalías.
- Documento técnico metodológico: descripción de análisis, técnicas aplicadas, entrenamiento de algoritmos y métricas utilizadas.
- Panel de consulta y prueba: interfaz básica para visualizar el comportamiento de las variables meteorológicas y los resultados de los modelos.

1.4. Objetivos

1.4.1. Objetivo general

Desarrollar un modelo de análisis basado en técnicas de aprendizaje no supervisado para la detección de anomalías en datos climáticos de la estación meteorológica de la Universidad del Azuay, mediante la aplicación de métodos de preprocesamiento, clustering y validación de resultados, identificando patrones atípicos y enfocados al índice de radiación ultravioleta.

1.4.2. Objetivo específico

1. Contextualizar los fundamentos teóricos y metodológicos sobre análisis de datos climáticos, aprendizaje no supervisado y detección de anomalías, a partir de una revisión de literatura.
2. Desarrollar un flujo de análisis de datos climáticos que contemple el preprocesamiento, la ingeniería de características, la aplicación de algoritmos de clustering y métodos de detección de anomalías, empleando herramientas de ciencia de datos en Python.
3. Analizar los resultados obtenidos del modelo implementado, utilizando métricas estadísticas y de validación de clusters, considerando los patrones anómalos y proponiendo recomendaciones que fortalezcan la calidad de la información meteorológica en la estación UDA-1.

CAPÍTULO 2:

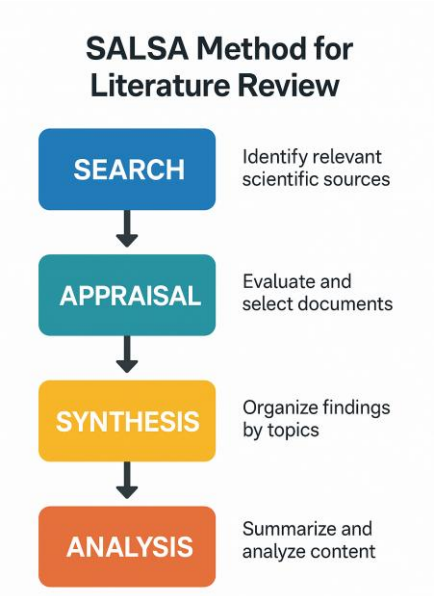
2. REVISIÓN DE LITERATURA

Para la revisión literaria se utilizó el método SALSA conocido así por sus siglas en ingles Search, Appraisal, Synthesis, Analysis, esta metodología permite realizar un enfoque crítico y académico siguiendo la estructura de búsqueda de manera organizada y sistemática, la Figura 1 muestra una breve descripción del método.

De forma resumida se puede decir que el método SALSA, permite extraer información relevante conectando la teoría con hechos prácticos. Procurando evitar omisiones, vacíos de conocimientos en función a los objetivos del estudio.

Figura 1

Método SALSA - Revisión de literatura



Nota. Figura basada en la metodología descrita por Booth, Papaioannou y Sutton (2012).

La búsqueda (Search) se realizó en las bases de datos IEEE, Scopus, Springer, IEEE, ArXiv, y Science Direct. Luego se realiza una evaluación crítica (Appraisal) de las selecciones,

priorizando los que dispongan de una metodología explicada, y con enfoque a datos climatológicos a través de aprendizaje no supervisado. Seguidos a esto se realiza una síntesis (Synthesis) agrupando las palabras claves más relevantes, en este estudio fueron: aprendizaje automático, detección de anomalías, clustering, validación de modelos. Como último paso se analiza (Analysis) de las limitaciones y oportunidades a nivel de investigación.

El método SALSA según Lo et al. (2023) menciona que "este enfoque sistemático mejora la transparencia y permite una evaluación más fácil de las opciones bibliográficas y sus implicaciones" (p. 671). En la sección 2.1 se profundiza sobre estos resultados.

2.1. Estado del Arte

2.1.1. Búsqueda y Evaluación de Literatura

Los primeros dos pasos de la metodología SALSA tienen como objetivo identificar, seleccionar y filtrar la literatura científica más relevante considerando datos climáticos mediante aprendizaje no supervisado, específicamente en el contexto de detección de anomalías. Se prioriza los años entre 2019 y 2024 para garantizar contenido actual.

La búsqueda usó palabras clave cómo "unsupervised learning", "anomaly detection", "climate data", "clustering", "weather time series", "sensor data quality". Inicialmente se identificaron un total de 62 artículos, distribuidos según el tipo de fuente (ver Tabla 1).

Tabla 1

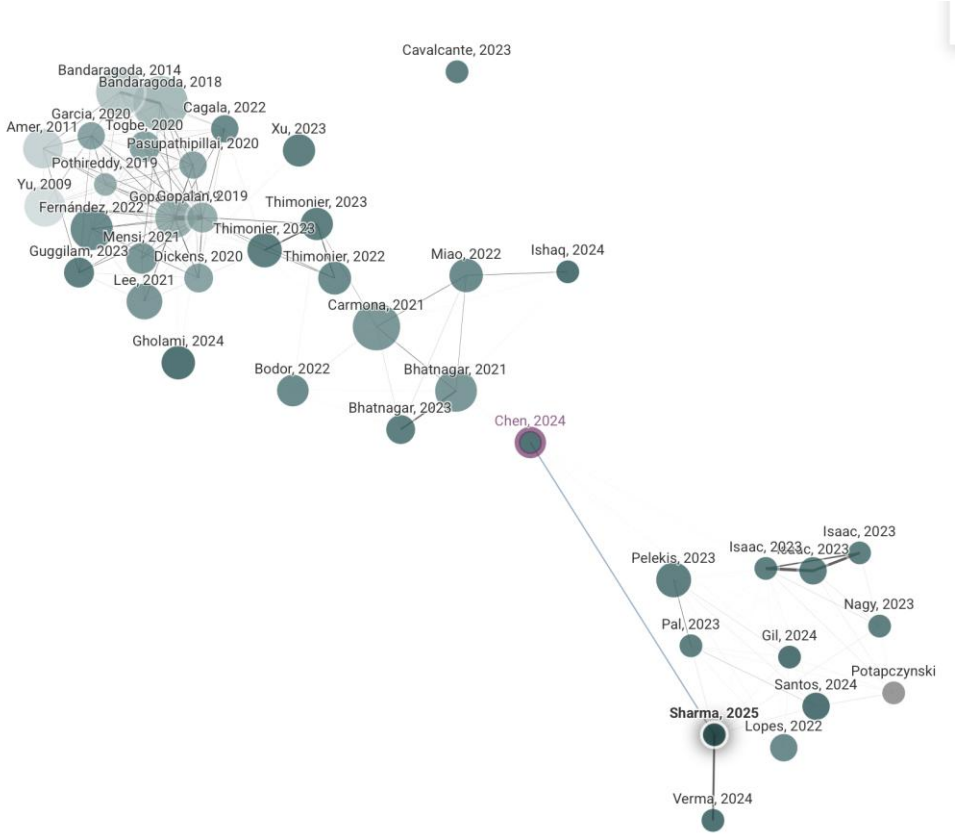
Artículos clasificados según tipo de fuente

Tipo de fuente	Número de artículos encontrados
Artículos en revistas indexadas Scopus	18
Artículos en revistas indexadas WoS	12
Artículos de conferencias IEEE	14
Capítulos de libro Springer	6
Preprints en ArXiv	8
Repositorios institucionales y otros	4
Total	**62**

De la misma manera se generó un análisis bibliométrico de cómo se relacionan los artículos de alto impacto con base a las palabras clave ya definidas, ver Figura 2.

Figura 2

Red de relación de publicaciones - análisis bibliométrico de la investigación



Nota. Generado mediante https://www.connectedpapers.com/

La primera recopilación se evalúa a través del enfoque metodológico, relación con el tema objetivo, si los datos fueron meteorológicos, si hubo aplicaciones similares en el contexto de Cuenca Ecuador. Este pasa da como resultados 40 artículos.

Es importante recalcar que se evaluó también en función a los objetivos y los diferentes tipos de modelos o algoritmos usados en especial si con de aprendizaje de máquina no supervisado. En la Tabla 2, se presenta una visión resumida de esta clasificación según las fases SALSA.

Tabla 2

Clasificación de artículos según el método SALSA

Categoría SALSA	Cantidad de artículos
Search - Artículos identificados	62
Appraisal - Artículos seleccionados	40
Synthesis - Artículos organizados temáticamente	28
Analysis - Artículos utilizados críticamente	15

Varios trabajos fueron descartados, por solo ser propuestas teóricas, ni encajar en el tema del clima, o meteorología. Esta evaluación muestra que no existe literatura suficiente en este ámbito sobre todo en zonas andinas, como Cuenca o Ecuador Sur, y menos aún a estaciones locales municipales.

2.1.2. Síntesis

En esta etapa se organizará la información para presentar la de forma tabular. Se estructura la literatura en tres categorías principales: (a) técnicas de aprendizaje no supervisado aplicadas a datos climáticos, (b) algoritmos específicos y detección de anomalías, y (c) enfoques de validación de modelos sin supervisión.

a) **Aprendizaje no supervisado aplicado a datos climáticos. -** La literatura indica que el aprendizaje no supervisado se ha vuelto muy relevante para trabajar con datos climáticos, en especial para aquellos donde existen patrones anómalos. Este se ve en trabajos como lo presentado por Racah et al. (2017) quienes mencionan el valor de aplicar modelos no supervisados para modelar patrones en eventos, con consideraciones de series de tiempo, que por lo general son recolectados de forma semi automática o automática. Una de las grandes ventajas de trabajar en enfoques no supervisados es que la facilidad de reducir dimensionalidad, ya sea con PCA o con rotaciones, considerando la presencia

intrínseca de la estacionalidad, esto se puede observar en los trabajos de Chandola et al. (2009).

b) Clustering y Detección de Anomalías. La agrupación tiene como ventaja la detección de anomalías, ya que aquellos datos que no logran agruparse se pueden asimilar como ruido o una anomalía. En este caso algoritmos como K-means, DBSCAN, GMM, clusterización jerárquica, entre otros, se usan con mayor frecuencia. Ding et al. (2023), explican combinar redes neuronales, autoencoders y otros algoritmos para identificar patrones atípicos en datos ambientales. Desde otro enfoque Wyld, Alnutefy y Alsuwayh (2024) demostraron que los autoencoders combinando con otros modelos pueden comprimir y reconstruir señales climáticas complejas y de esta manera podrían identificar las anomalías que se logren presentar, a pesar de ello no consideran que las anomalías se deban excluir, al contrario, las usan como punto de partida de sus propuestas.

c) **Validación de Modelos No Supervisados.** En esta temática, Idan et al. (2024) proponen como mecanismo de validación colectiva, no solo las métricas de evaluación sino también un sistema de votación incremental. De forma más clásica, otros estudios emplean métricas internas de validación de clusters, como el índice de silueta, el índice de Calinski-Harabasz, o coeficientes de determinación estadística.

A pesar de la existencia de varias propuestas aún se consideran que los estudios y avances son pocos, comparados con otras áreas del conocimiento, por esto la temática podría apoyar e influir en la construcción de políticas ambientales bajo consideraciones técnicas y propias de la localidad.

2.1.3. Análisis Crítico

El análisis crítico de la literatura científica revisada revela un panorama de avances importantes en la aplicación de técnicas de aprendizaje no supervisado a la detección de anomalías en datos climáticos; sin embargo, también evidencia vacíos significativos que justifican y direccionan este estudio. Las investigaciones identificadas se concentran, en su mayoría, en contextos geográficos de países industrializados, con gran acceso a infraestructuras de sensores avanzados y bases de datos etiquetadas o curadas por expertos. Esta condición difiere notoriamente de los escenarios latinoamericanos, y en particular del caso de la ciudad de Cuenca (Ecuador), donde los sistemas de monitoreo climático aún presentan desafíos en términos de cobertura, calidad de los datos y mantenimiento de sensores.

Se observa una tendencia calra pr el uso de algoritmos clásicos como K-means y DBSCAN para el análisis de agrupamientos y la detección de valores atípicos. Si bien son eficaces en muchos casos, en ocasiones no consideran la existencia de los valores anómalos presentes. Además, en condiciones no lineales y estacionales pierden su efectividad.

Asimismo, el análisis pone en evidencia que la mayoría de estudios carecen de procesos robustos de validación y evaluación, dejando al descubierto una gran oportunidad y una brecha de trabajo en esa área. Algunos trabajos, como los de Idan et al. (2024) han comenzado a proponer comités de modelos y mecanismos de voto mayoritario para superar esta limitación, sin embargo, no se observa una implementación generalizada de estas estrategias en climatología aplicada. Esto plantea la necesidad de una validación cruzada no solo con métricas estadísticas, sino también con información contextual local y conocimiento de expertos climáticos de la zona de estudio.

Otro aspecto crítico es la escasa integración entre la detección de anomalías y la interpretación posterior de dichas anomalías. En varios artículos, la detección se limita a la

identificación matemática de outliers, sin discutir sus implicaciones meteorológicas, ecológicas o sociales. Esta desconexión entre la dimensión técnica y la toma de decisiones climáticas reduce el valor aplicado de los hallazgos. En ese sentido, se propone como parte de su valor agregado, una interpretación de las anomalías detectadas a la luz de registros climáticos históricos de Cuenca, eventos extremos conocidos y la lógica física de las variables involucradas.

Adicionalmente, es importante destacar que no se identificaron investigaciones específicas aplicadas a datos meteorológicos de la ciudad de Cuenca. Esto refuerza la originalidad y pertinencia del presente estudio, el cual no solo busca replicar técnicas previamente desarrolladas, sino contextualizarlas, evaluarlas y adaptarlas a una realidad geográfica, climática y tecnológica concreta. Un resumen de esta sección se muestra en la Tabla 3.

Tabla 3

Categorías, tendencias y oportunidades de la investigación

Categoría Analizada	Tendencias Identificadas	Oportunidades para esta investigación
Contexto geográfico de estudios	Predominancia de estudios en países desarrollados con datos curados	Aplicar modelos en una ciudad andina subrepresentada
Técnicas de clustering aplicadas	Frecuente uso de K-means y DBSCAN, limitada exploración de métodos avanzados	Comparar múltiples algoritmos y explorar técnicas avanzadas
Validación de modelos no supervisados	Validación interna común, escasa validación contextual o por expertos	Diseñar un enfoque híbrido de validación estadística y contextual
Interpretación de anomalías detectadas	Débil conexión entre detección técnica y significado meteorológico	Analizar anomalías con interpretación técnica-climática
Aplicación directa en Cuenca/Ecuador	Nula evidencia de estudios aplicados a datos de Cuenca o UAZUAY	Desarrollar una metodología replicable para otras regiones del país

Finalmente, el enfoque general de este trabajo será no en solo aplicar algoritmos sino en proponer una metodología que permita adaptarse a las condiciones de cualquier tipo de datos, enfocadas a una estructura temporal por la naturaleza de los datos meteorológicos.

2.2. Marco Teórico

2.2.1. Datos Climáticos y Características de las Series Temporales Meteorológicas

Los datos climáticos tienen la particularidad de ser datos temporales, es decir que están relacionados con una unidad de tiempo y el orden de apuración importa. Usualmente se recolectan por medio de sensores, estaciones y satélites de manera automática. Además, estos datos tienden a tener un comportamiento multivariado, es ahí donde problemas como la multicolinealidad y ortogonalidad son relevantes, debido a condiciones geográficas, temporales o estacionales (Wilks, 2011). En el contexto de la ciencia de datos aplicada al clima, es necesario recordar que los datos no son independientes, por ende, la existencia de patrones es relevante un ejemplo de ello son las lluvias, o los fenómenos como el niño y la niña presentes en el cono sur de América (Hyndman y Athanasopoulos, 2018).

Otra situación que se presenta en las series meteorológicas es su alta autocorrelación temporal, es decir, se genera una dependencia del dato anterior o de una ventana de datos anteriores, por eso es necesario su estudio en grupos temporales. Por tal motivo, la ingeniería de datos a través de transformaciones cíclicas o periódicas es común y relevante en estos datos (Shumway y Stoffer, 2017). De esta manera, debe señalarse que en contextos locales como lo es Cuenca, con sus condiciones micro climáticas, pueden existir fenómenos climáticos altamente localizados, a tal punto que podrían ser geográficos de radios cortos.

2.2.2. Fundamentos del Aprendizaje Automático

El aprendizaje de máquinas está directamente vinculado con la inteligencia artificial, con el objetivo de encontrar patrones y hacer trabajos que requieran un gran esfuerzo, a través de decisiones de autoaprendizaje. Para sistemas dinámicos esta arista se ha vuelto esencial.

Según Goodfellow et al. (2016), el aprendizaje de máquinas puede dividirse entre tres grupos, supervisado, no supervisado y por refuerzo, dependiendo si están etiquetados o no serán supervisados o no, y por refuerzo si tienen un agente externo que aprende y toma decisiones basado en factores externos.

En este estudio el aprendizaje no supervisado toma gran relevancia debido a la naturaleza de los datos recolectados por la estación meteorológica. Estos datos al no ser etiquetados permiten el uso de modelos no supervisados, además que se presume que los datos presentan valores tanto normales cómo anómalos, y presentan una alta variabilidad temporal, espacial y estacional.

A través de varias técnicas no supervisadas, como el DBSCAN, Random Forest se pueden explorar los datos y clasificarlos en si presentan anomalías o no, esto a través de consideraciones de que es un evento extremo y si son consistentes frente a datos históricos (Zhang et al., 2019).

2.2.3. Aprendizaje No Supervisado y su Aplicación

El aprendizaje no supervisado se enfoca en identificar patrones o estructuras ocultas dentro de un conjunto de datos sin necesidad de etiquetas. Esta característica hace que el uso de estas técnicas sea útil en ámbitos de gran cantidad de datos, como textos, o recolección por sensores. En contraste con los modelos supervisados, que requieren un conjunto de datos etiquetado para entrenar algoritmos predictivos, los modelos no supervisados buscan relaciones internas entre los datos, permitiendo segmentarlos, agruparlos, incluso considerando el ruido posible.

“Los métodos de aprendizaje no supervisados son herramientas poderosas para la exploración y el análisis de datos complejos y de alta dimensión” (Vinci, 2024, p. 252).

Varios autores coinciden que una ventaja fuerte del aprendizaje no supervisado es la capacidad de detectar datos fuera de lo común, a través de la exploración. Por ejemplo,

Aggarwal (2015), indica que esta clase de algoritmos es esencial para identificar de forma exploratorio grupos de datos fuera de lo común sin incurrir a sesgos innecesarios.

Se podría decir que las dos principales aplicaciones del aprendizaje no supervisado para el objetivo de este estudio son el clustering (agrupamiento) y la detección de anomalías. Además, que luego se puede dar el paso hacia la predicción. Consideremos que el agrupamiento, permite identificar subconjuntos con características similares, por ejemplo, temperatura o niveles de presencia de gases pesados o no, además de material particulado, de tal forma que situaciones donde la desviación estándar cambie de forma abrupta repercutirá en indicios de una anomalía.

Se debe tener en cuenta que los métodos de detección de anomalías permiten identificar registros que difieren significativamente del resto, pero no por eso significa que los datos deben ser eliminados o rectificados, al contrario, esta información asegura que si los datos se repiten estacionalmente es un hallazgo de futuros eventos extremos (Chandola et al., 2009).

Una ventaja importante del aprendizaje no supervisado es su flexibilidad. Estos modelos pueden aplicarse tanto a datos en forma de tablas como a series temporales. Además, su implementación es eficiente desde el punto de vista del gasto computacional, transformándolos en modelos idóneos para aplicaciones en tiempo real, como estaciones meteorológicas (Han et al., 2011). Es necesario considerar que una de las debilidades de los modelos no supervisados es su dificultad de validación, ya que al no tener etiquetas no se puede trabajar en torno a una matriz de confusión. Es por eso que medidas basadas en el error son necesarias. Por esta razón, varios estudios optan por utilizar métricas internas y complementarlas con validaciones contextuales o conocimientos de expertos (Hodge y Austin, 2004).

2.2.4. Técnicas de Agrupamiento Aplicadas a Datos Climáticos

A través de técnicas de agrupamiento, los datos sin etiquetas pueden revelar patrones ocultos, al identificar qué y cuando los datos no se agruparon. Este enfoque resulta especialmente útil en meteorología, donde los datos presentan comportamientos complejos, dependientes del tiempo y de una gran cantidad de otras variables.

Entre los métodos más utilizados se pueden mencionar el algoritmo K-means, y sus variantes como el propuesto por Doan et al. (2023) denominado S-K means, creado especialmente para datos climatológicos en torno a las lluvias y otros fenómenos, la variante se enfoca a trabajar a través de identificación de estructuras previas para el agrupamiento. Lamentablemente la sensibilidad de estos algoritmos a los datos atípicos hace que su efectividad se reduzca.

Esta limitación podría ser cubierta por otras técnicas cómo DBSCAN (Density-Based Spatial Clustering of Applications with Noise) que puede ser útiles en datos con arreglos de series de tiempo, a través de las ventanas de tiempo y la identificación de los datos que se salen de las desviaciones de control (Dharani et al., 2022). Otro enfoque es el uso de Gaussian Mixture Models (GMM), es un modelo basado en probabilidades, se ha demostrado su utilidad en datos de transiciones graduales entre, de hecho, Pacal et al. (2023) destacan que los GMM ofrecen una aproximación flexible y adecuada para la identificación de eventos extremos con alta variabilidad. La mejor fortaleza de este tipo de modelos, es que trabajan con distribuciones propias y la esperanza matemática propia de los datos.

Una técnica novedosa en torno a los algoritmos no supervisados es el MiSTIC (Minimum Spanning Tree-based Interpolated Clustering), que es una versión mejorada del K -means al ofrecer mayor versatilidad y adaptación con cada iteración. Según Reddy y Rajan (2023) comparan MiSTIC con K-means en el análisis de datos anidados, concluyendo que

MiSTIC proporciona una segmentación representativa en términos de homogeneidad espacial y ortogonalidad, sin romper el sentido los grupos principales.

Desde el punto de vista de la evaluación, se puede mencionar que la métrica común para estos algoritmos es el coeficiente de silueta, el índice de Davies–Bouldin que usualmente se usan para valorar la coherencia entre clústeres. La visualización y la exploración son necesarias para respaldar la interpretación de las métricas de evaluación (Doan et al., 2023; Reddy y Rajan, 2023).

2.2.5. Detección de Anomalías en Conjuntos de Datos Climáticos

Este estudio busca métodos para detección de anomalías, considerando una anomalía como un cambio extremo en los datos. Varios métodos se han propuesto, incluyendo las aproximaciones estadísticas pura, a través de cartas de control y análisis inferenciales. Bâra et al. (2024) demostraron que el uso de Autoencoders es útil para detectar eventos anómalos en datos climáticos de Rumania. Además, han desarrollado métodos de aprendizaje profundo para trabajar con datos temporales y multivariados. Estos enfoques emplean hibridación de métodos usando, estadística clásica y redes neuronales para modelar patrones normales y comparar con eventos de alta variabilidad.

La combinación de técnicas de clustering y autoencoders ha demostrado su efectividad, un ejemplo de ello lo muestran Ale et al. (2024), quienes propusieron un método que integra clustering de características con autoencoders variacionales y umbrales dinámicos multivariantes, especialmente en eventos extremos como el deshielo en el Ártico.

Davis et al. (2024) presentaron un método para analizar análisis el comportamiento evolutivo de clústeres multivariantes espacio-temporales a través de enfoques satelitales demostrando así la posibilidad de detectar eventos anómalos en el Monte Pinatubo.

2.2.6. Ingeniería de Características para Series Meteorológicas

La ingeniería de características es una etapa clave dentro del análisis de datos, y cobra especial relevancia cuando trabajamos con series temporales como los registros del clima. Este proceso consiste en transformar, seleccionar o incluso generar nuevas variables que ayuden a los modelos a reconocer patrones complejos que, de otro modo, pasarían desapercibidos.

Con base a lo mencionado por K y Jisha (2024) “los modelos de series temporales se utilizaron ampliamente anteriormente en el escenario, que ahora se están sustituyendo en gran medida por modelos de aprendizaje profundo” (p. 1).

En el análisis de datos climáticos, donde abundan variables con alta variabilidad tanto estacional como espacial, una buena estrategia de ingeniería de características puede marcar la diferencia. Gracias a ella, es posible revelar información subyacente que los algoritmos por sí solos no lograrían detectar fácilmente.

Una técnica muy utilizada en este campo es la descomposición temporal de las variables meteorológicas. Esta permite separar componentes como la tendencia, la estacionalidad y los residuos. Por ejemplo, métodos como STL (Seasonal and Trend decomposition using Loess) se han implementado con éxito para distinguir el comportamiento sistemático de los datos del ruido aleatorio. Así, se puede detectar con mayor claridad la presencia de eventos atípicos (Fehlmann et al., 2023).

Otra herramienta poderosa es la aplicación de estadísticas móviles, también conocidas como ventanas deslizantes. Calcular valores como la media, mediana, desviación estándar o los cuartiles en diferentes intervalos de tiempo ayuda a los modelos a reconocer tanto cambios bruscos como tendencias persistentes. Investigaciones recientes han evidenciado que estas transformaciones mejoran considerablemente el rendimiento de modelos no

supervisados orientados a detectar anomalías en condiciones climáticas extremas (Ale et al., 2024).

También es útil incorporar codificaciones cíclicas para variables temporales como la hora del día, el mes o la estación del año. Dado que los fenómenos meteorológicos suelen seguir ciclos naturales, representar estas variables usando funciones seno y coseno permite a los modelos aprender mejor esas repeticiones, lo que se traduce en una mayor capacidad para generalizar (Doan et al., 2023).

Finalmente, otra estrategia interesante consiste en crear características compuestas, combinando variables como la temperatura y la humedad para formar índices que describen situaciones climáticas específicas. Por ejemplo, indicadores como la relación entre temperatura y precipitación pueden ayudar a identificar eventos complejos, como olas de calor o lluvias intensas. Según Davis et al. (2024), este tipo de enfoque multivariado ha permitido identificar clústeres asociados a impactos climáticos importantes, que en un análisis más simple podrían haber pasado inadvertidos.

La normalización o estandarización de las variables climáticas es indispensable para evitar que los algoritmos de aprendizaje no supervisado se vean influenciados por escalas heterogéneas. Esta práctica asegura una comparación justa entre variables como temperatura, radiación solar o concentración de contaminantes (Farooq y Khan, 2025).

De esta manera, se aplicará una estrategia de ingeniería de características que combina estadísticas móviles, codificación cíclica y normalización. Esta etapa es clave para mejorar el rendimiento de los algoritmos de detección de anomalías como DBSCAN, K-means y autoencoders, permitiendo una mejor diferenciación de condiciones normales y valores atípicos dentro de las series meteorológicas de la ciudad de Cuenca.

2.2.7. Validación y Evaluación de Modelos No Supervisados

En el contexto del aprendizaje no supervisado, la validación de modelos representa un desafío significativo, dado que no se dispone de etiquetas predefinidas que permitan contrastar los resultados. A diferencia del aprendizaje supervisado, donde métricas como la precisión o la matriz de confusión guían el proceso de evaluación, en modelos no supervisados se requiere aplicar métricas internas, de estabilidad y de interpretabilidad para determinar la calidad del agrupamiento o detección de anomalías.

Un enfoque común para evaluar técnicas de agrupamiento es el uso del índice de silueta, que permite identificar si existe coherencia entre los datos comparando los datos con el comportamiento general del clúster. Este índice, junto con el coeficiente de Calinski-Harabasz y el índice de Davies–Bouldin, forma parte del conjunto de métricas internas que se utilizan para medir la compactación y separación de los clústeres (Liu et al., 2022).

En escenarios como el análisis climático, donde los datos presentan alta estacionalidad y autocorrelación, se ha propuesto complementar las métricas internas con herramientas de visualización y validación empírica, contrastando los resultados con eventos históricos documentados. Por ejemplo, Davis et al. (2024) combinan análisis multivariantes y visualización para validar la detección de impactos climáticos utilizando la evolución de clústeres, un enfoque aplicable a eventos como erupciones volcánicas o cambios de régimen atmosférico.

Para tareas de detección de anomalías, la validación se vuelve aún más compleja, ya que en muchos casos no se cuenta con un conjunto objetivo de eventos anómalos. En este contexto, la literatura propone el recall@k para evaluar la recuperación de eventos extremos conocidos, usualmente se usa en autoencoders, donde se asume que las instancias con mayor error de reconstrucción representan anomalías (Üstek et al., 2024).

Otra estrategia relevante es la validación cruzada basada en ventanas temporales (time series cross-validation), donde el conjunto de datos se divide cronológicamente en bloques de entrenamiento y prueba, respetando la estructura temporal para evitar fugas de información. Esta técnica es esencial para asegurar la robustez de modelos aplicados a datos meteorológicos, ya que respeta la naturaleza secuencial de las observaciones (Fehlmann et al., 2023).

La evaluación de los modelos de clustering se realizará aplicando métricas internas como la silueta, Calinski-Harabasz y Davies–Bouldin, complementadas con validación visual mediante gráficos de dispersión y mapas térmicos. Para los modelos de detección de anomalías, se analizarán los umbrales de error de reconstrucción en autoencoders y se documentará la correspondencia con registros históricos locales de eventos climáticos extremos, validando así la utilidad práctica del modelo para la ciudad de Cuenca.

El desarrollo del marco teórico ha permitido construir una base sólida sobre los principales conceptos y enfoques metodológicos relacionados con el análisis de datos climáticos mediante aprendizaje no supervisado. A lo largo del capítulo se han abordado de forma detallada las características de las series meteorológicas, los fundamentos del aprendizaje automático y las principales técnicas de clustering y detección de anomalías utilizadas en investigaciones recientes. Asimismo, se ha profundizado en la importancia de la ingeniería de características para el tratamiento de datos climáticos y en las estrategias de validación que garantizan la robustez y la interpretabilidad de los modelos.

Este sustento conceptual será fundamental para guiar la selección y aplicación de las herramientas analíticas en el presente estudio, permitiendo un diseño metodológico riguroso y contextualizado. En el siguiente capítulo se detalla la metodología utilizada para alcanzar los objetivos planteados, así como los procedimientos técnicos empleados para el procesamiento, modelado y evaluación de los datos meteorológicos recolectados en la ciudad de Cuenca.

CAPÍTULO 3:

3. DESARROLLO

3.1. Metodología

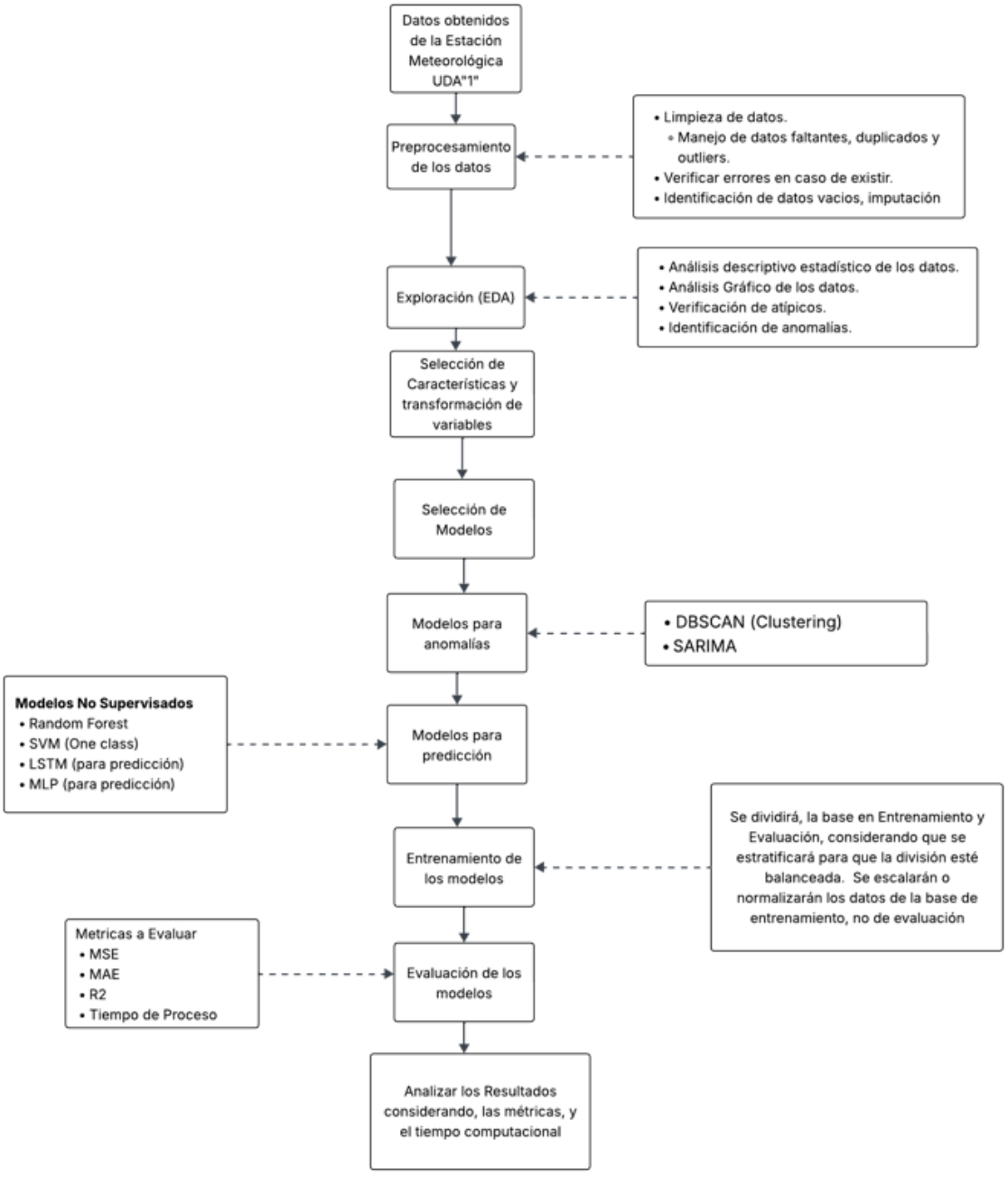
Se presenta una metodología base para este trabajo, el enfoque permite identificar y tratar anomalías presentes en los datos, así como implementar modelos capaces de predecir el comportamiento futuro de la variable de interés, aprovechando tanto técnicas estadísticas como algoritmos de inteligencia artificial.

La metodología se basa desde la recolección de los datos, preprocesamiento, exploración, selección de características, y transformación de variables.

Luego se decide con que modelos trabajar después de llegar a un consenso a través de experimentación, se decide trabajar con modelos para detección de anomalías y con modelos para predecir como un adicional a este estudio, como se muestra en la Figura 3.

Figura 3

Esquema de la metodología aplicada



Obtención de los Datos

El primer paso en la metodología consiste en la obtención de los datos, los cuales contendrán las variables como se establece en la Tabla 4.

Tabla 4

Descripción de las variables

Nombre	Descripción	Nombre	Descripción
Date	Fecha de la muestra	NO2UGM3	Dióxido de Nitrógeno en ugm³
MES	Mes del registro	O3UGM3	Ozono en ugm³
DIA	Día del registro	PM25	Material particulado 25 en ugm³
HORA	Hora del registro	RAINFALL	Lluvia en cantidad
ANIO	Año del registro	SO2UGM3	Dióxido de Azufre en ugm³
AMBTEMP	Temperatura Ambiente	UV_INDEX	Índice de Radiación Ultravioleta
COUGM3	Monóxido de Carbono en ugm³		

Los datos se obtienen de la estación meteorológica UDA “1”, que obtiene información por segundo. Por condiciones de este trabajo los datos se recolectan por horas, obteniendo el valor promedio de aquellos segundos que conforman la hora y así reportar el valor hora promedio de cada variable.

Preprocesamiento de los Datos

Una vez obtenida la base de datos, se realiza el preprocesamiento, en esta etapa se analiza de forma intensiva cada variable, en busca de posibles errores o inconsistencias. En esta fase se lleva a cabo la limpieza de datos, que incluye la eliminación de registros duplicados (se infiere por errores de captación de los sensores) y la corrección de errores evidentes que están fuera de las unidades comunes, sugeridas por el Instituto Nacional de Meteorología e Hidrología Ecuatoriano INAMHI (2025) especificados en la Tabla 5.

Tabla 5

Valores Comunes de las variables en Cuenca Ecuador

Nombre	Descripción	Valores comunes Cuenca	Nombre	Descripción	Valores comunes Cuenca
AMBTEMP	Temperatura Ambiente	10°C – 30°C	PM25	Material particulado 25	10 - 60
COUGM3	Monóxido de Carbono en ugm³	100 - 2000	RAINFALL	Lluvia en cantidad en mm	0 – 50
NO2UGM3	Dióxido de Nitrógeno en ugm³	20 - 100	SO2UGM3	Dióxido de Azufre en ugm³	5 - 50
O3UGM3	Ozono en ugm³	40 – 180	UV_INDEX	Índice de Radiación Ultravioleta	0 a 8

También se estudian los datos faltantes mediante técnicas como interpolación, imputación estadística por mediana. Otro aspecto importante es la detección y tratamiento de valores atípicos (outliers), para lo cual se emplea el Z-score, y el rango intercuartílico (IQR), además de los diagramas de caja.

Exploración de Datos (Exploratory Data Analysis EDA)

La exploración de los datos incluye el análisis estadístico descriptivo, donde se obtienen medidas de tendencia central, dispersión y distribución. Se complementa con visualizaciones gráficas como histogramas, y gráficos de series temporales para identificar posibles estacionalidades, tendencias o patrones inusuales. Esta etapa también permite una primera detección visual de anomalías o valores atípicos que podrían requerir mayor atención en etapas posteriores.

Selección de Características y Transformación de Variables

Al ser los datos, una serie de tiempo, por su naturaleza existen algunos comportamientos que pueden ser explorados durante la selección de características, por ejemplos transformaciones de los tiempos en función a senos o cosenos, estudiar la variable fecha por separado, e identificar momentos de anormalidades climáticas.

Selección de Modelos

Modelos para Estudiar Anomalías

Se procede a la identificación de anomalías mediante modelos de aprendizaje no supervisado. Técnicas como DBSCAN (Density-Based Spatial Clustering of Applications with Noise) permiten identificar puntos que no pertenecen a ningún grupo denso. Otras técnicas útiles incluyen One-Class SVM, que genera una frontera para separar datos normales de anómalos, y K-means, que puede utilizarse indirectamente analizando la distancia de los puntos a sus centrosides. Estas técnicas ayudan a detectar eventos anómalos, ya sea por fallas en los sensores, fenómenos meteorológicos extremos o errores de medición.

En el estudio Investigación sobre la detección de anomalías en el flujo de datos dinámicos basada en el aprendizaje automático, según Wang et al. (2024) “es importante comparar métodos tradicionales como K-Means, Isolation Forest y DBSCAN, para vilidar significativamente cual es el mejor método” (p. 3).

Modelos para Predecir

Para construir modelos para una predicción del comportamiento de los índices de radiación UV, se puede partir desde el enfoque de algoritmos de aprendizaje automático y modelos estadísticos. Entre los métodos estadísticos destacan ARIMA y SARIMA, apropiados para series temporales univariadas y multivariadas respectivamente con componentes estacionales. En cuanto a modelos más avanzados, se consideran redes neuronales como MLP (Multilayer Perceptron) y LSTM (Long Short-Term Memory), que son capaces de capturar

relaciones no lineales y dependencias temporales complejas. También se puede utilizar *Random Forest*, muy útil cuando se consideran múltiples variables predictoras, gracias a su capacidad para manejar datos con ruido y su robustez ante sobreajuste.

Entrenamiento de los Modelos y Evaluación

Los modelos se entrenan y prueban considerando sus hiper parámetros, y las sugerencias encontradas en la literatura, opinión de expertos o el empirismo.

La base de datos se divide en conjuntos de entrenamiento y evaluación. Esta división debe realizarse de manera estratificada si se trabajan con clases desbalanceadas, debido a los diferentes niveles extremos de los UV_INDEX, asegurando una distribución equilibrada. Es importante recalcar que la normalización o escalamiento debe aplicarse únicamente al conjunto de entrenamiento y luego reutilizar los parámetros con los datos de evaluación, evitando así fugas de información que alteren los resultados.

El entrenamiento de los modelos se realiza utilizando los conjuntos preparados, y posteriormente se evalúan utilizando métricas específicas según el tipo de modelo. Para los modelos de predicción continua, se emplean métricas como el Error Cuadrático Medio (MSE) y el Error Absoluto Medio (MAE). Para modelos de clasificación de anomalías se utilizan métricas como la precisión (accuracy), la sensibilidad (recall), el puntaje F1 creando posibles etiquetas debido a la categorización de los índices de radiación solar propuestos por Instituto Nacional de Meteorología e Hidrología Ecuatoriano INAMHI (2025), donde valores entre 1-2 se podrían clasificar como UV bajo, 3-5 moderado, 6-7 alto, 8-10 muy alto, 11 o mayores extremos.

Análisis de Resultados

Finalmente, se procede a realizar un análisis integral de los resultados obtenidos, considerando no solo las métricas cuantitativas de evaluación, sino también el tiempo de entrenamiento, la eficiencia computacional del modelo y los recursos requeridos para su

ejecución. Con base en estos indicadores, se determina el modelo más robusto y adecuado para su posterior implementación operativa, asegurando que cumpla con los criterios de precisión, escalabilidad y sostenibilidad computacional.

Como resultado de este proceso, se desarrolla un producto funcional en forma de API (Interfaz de Programación de Aplicaciones), que permite integrar de manera práctica los hallazgos y predicciones generados en esta investigación. Esta API está diseñada para facilitar la consulta, automatización y utilización de los modelos en entornos reales, garantizando la transferencia efectiva del conocimiento generado hacia aplicaciones de monitoreo y toma de decisiones.

3.2. Desarrollo del Código

Para este proyecto, los scripts creados usan una arquitectura modular, es decir se crearon funciones separadas con el fin de ser llamadas de forma independiente cuando los modelos los requieran, cabe aclarar que para asegurar la reproducibilidad se utilizó el número 28 como *random state*.

Carga de Datos

Los datos recibidos fue en formato CSV, don se pidió en intervalos por hora, a pesar que se pueden receptar por segundo o minuto. Se debe recalcar que la mayoría de investigaciones relacionadas con los índices ultravioleta se encuentran analizadas por hora.
De acuerdo con Fioletov et al. (2003) “la estimación del índice ultravioleta se realiza frecuentemente con datos por hora para obtener una climatología detallada de la radiación UV disponible” (p. 148). La Figura 4 muestra un ejemplo de los datos cargados.

Figura 4

Ejemplo de datos utilizados en la carga

	Date	ANIO	MES	DIA	HORA	AMBTEMP	COUGM3	NO2UGM3	O3UGM3	PM25	RAINFALL	SO2UGM3	UV_INDEX
0	2024-07-01 00:00:00	2024	7	1	0	14.5	0.5	23.4	5.7	9.35	0.0	1.2	0.000000
1	2024-07-01 01:00:00	2024	7	1	1	14.0	0.2	16.3	15.1	5.62	0.0	-0.9	0.000000
2	2024-07-01 02:00:00	2024	7	1	2	13.4	0.2	14.2	14.0	5.19	0.0	-0.5	0.000000
3	2024-07-01 03:00:00	2024	7	1	3	12.8	0.2	12.9	9.9	5.43	0.0	2.3	0.000000
4	2024-07-01 04:00:00	2024	7	1	4	11.9	0.1	10.2	8.5	4.44	0.0	0.9	0.000000
5	2024-07-01 05:00:00	2024	7	1	5	11.2	0.2	11.2	7.3	4.52	0.0	0.6	0.000000
6	2024-07-01 06:00:00	2024	7	1	6	10.5	0.9	14.8	4.9	6.61	0.0	2.2	0.030074
7	2024-07-01 07:00:00	2024	7	1	7	11.8	1.4	20.1	6.0	8.73	0.0	3.6	0.513637
8	2024-07-01 08:00:00	2024	7	1	8	15.5	0.8	23.9	15.1	6.02	0.0	2.7	2.334725
9	2024-07-01 09:00:00	2024	7	1	9	17.2	0.4	22.1	33.6	4.97	0.0	3.8	5.183693

De forma visual, se detectaron algunos problemas en los datos, por ejemplo, valores vacíos, valores negativos, y una gran cantidad de valores 0. Respecto a los valores cero, son normales, debido a que ciertas características no se miden durante horas de la madrugada ni de la noche, usualmente en Cuenca Ecuador, la incidencia del índice Ultravioleta (UV_INDEX), empieza desde las 5 horas con 45 minutos, lo que se vería reflejado desde las 6 am en nuestra base de datos.

Los valores vacíos, en ocasiones, actualizaciones del sistema, o mantenimiento de los sensores, hacen que se pierdan valores por algunos lapsos de tiempo, cabe recalcar que Ecuador sufrió apagones y los generadores no funcionaban 24 horas, por eso hay secciones de datos vacías. Respecto a los valores negativos, los expertos de la estación meteorológica indican que el método de cálculo y reducción de valores, aún debe corregirse, pero los valores no se alteran deben de analizarse de forma absoluta. En la Figura 5 se puede evidenciar.

Figura 5

Análisis descriptivo de datos

	Date	ANIO	MES	DIA	HORA	AMBTEMP	COUGM3	NO2UGM3	O3UGM3	PM25	RAINFALL	SO2UGM3	UV_INDEX
count	6865	6865.000000	6865.000000	6865.000000	6865.000000	6865.00000	6846.000000	6846.000000	6841.000000	6858.000000	6863.000000	6847.000000	6865.000000
mean	2024-11-21 00:00:00.000000256	2024.356737	6.905171	15.349308	11.498325	15.06007	0.101803	16.606819	35.121386	10.662546	0.070436	-0.037824	1.933376
min	2024-07-01 00:00:00	2024.000000	1.000000	1.000000	0.000000	4.40000	-0.800000	-7.000000	-1.500000	0.210000	-314.800000	-5.700000	0.000000
25%	2024-09-10 12:00:00	2024.000000	3.000000	8.000000	5.000000	12.90000	-0.200000	6.500000	13.900000	4.360000	0.000000	-2.300000	0.000000
50%	2024-11-21 00:00:00	2024.000000	8.000000	15.000000	11.000000	14.30000	0.100000	13.800000	30.200000	6.880000	0.000000	-1.400000	0.008573
75%	2025-01-31 12:00:00	2025.000000	10.000000	23.000000	17.000000	17.40000	0.400000	24.100000	50.300000	12.137500	0.000000	0.100000	2.869530
max	2025-04-13 00:00:00	2025.000000	12.000000	31.000000	23.000000	26.00000	33.200000	250.700000	335.600000	248.490000	27.400000	761.100000	15.613208
std	NaN	0.479071	3.787541	8.847042	6.923578	3.49215	0.789271	14.195710	25.214226	12.091447	3.896727	16.473346	3.197280

Preprocesamiento

Para realizar el preprocesamiento de la base de datos se utilizó el siguiente código: La Figura 6, muestra el código para contar cuantas celdas vacías tienen por columnas, y cuantos valores negativos, en la base de datos de estudio, se detectaron, estos dos problemas, que se resolvieron, llenando los valores con sus valores anteriores mediante el método “fill”, no se aplica ni la media ni el promedio, debido a los cambios abruptos que se pueden generar a lo largo del día, luego la base de datos se modifica utilizando sólo valores absolutos como recomendaron los expertos de la estación meteorológica, como se observa en la Figura 7.

Figura 6

Código para identificar valores vacíos y negativos

```
#CONTAR VALORES NEGATIVOS
# SELECCION DE LAS COLUMNAS NUMERICAS
numerical_df = df.select_dtypes(include=['number'])
# CONTEO DE NEGATIVOS
(numerical_df < 0).sum()

#CONTAR VACIOS
df.isnull().sum()
```

Figura 7

Código para corregir vacíos con la mediana, y considerar valores absolutos

```
# SELECCION DE LAS COLUMNAS NUMERICAS
numerical_cols = df.select_dtypes(include=['number']).columns
# APLICO VALORES ABSOLUTOS
df[numerical_cols] = df[numerical_cols].abs()
# Imputo los valores con las técnica del llenado anterior
df = df.fillna(method='ffill')
```

Exploración

Para dinamizar el uso de las técnicas se crearon las siguientes funciones para aplicarlas en cualquier momento requerido. El código de las funciones y su respectiva descripción, se puede observar en las Figuras 8 a la Figura 13.

Figura 8

Código de función para mostrar información básica

```
def basic_info(df):
    print("\n=== DATASET OVERVIEW ===")
    print(f"Dataset shape: {df.shape}")
    print(f"Number of rows: {df.shape[0]}")
    print(f"Number of columns: {df.shape[1]}")

    print("\n=== DATA TYPES ===")
    print(df.dtypes)

    print("\n=== MISSING VALUES ===")
    missing = df.isnull().sum()
    missing_percent = (missing / len(df)) * 100
    missing_info = pd.DataFrame({"Missing Values": missing, "Percentage": missing_percent})
    print(missing_info[missing_info["Missing Values"] > 0])

    print("\n=== DUPLICATED ROWS ===")
    print(f"Number of duplicated rows: {df.duplicated().sum()}")

    return missing_info
```

Función para mostrar la información básica de la base de datos, su forma, tipo de variables, porcentaje de valores vacíos, identificación de filas duplicadas. Ayuda a preparar la base de datos previo a su exploración.

Figura 9

Código de función para un estudio descriptivo

```
def numerical_analysis(df):
    numerical_cols = df.select_dtypes(include=["int64", "float64"]).columns

    if len(numerical_cols) == 0:
        print("\nNo numerical columns found in the dataset.")
        return

    print("\n=== NUMERICAL COLUMNS STATISTICS ===")
    stats_df = df[numerical_cols].describe().T
    stats_df["skew"] = df[numerical_cols].skew()
    stats_df["kurtosis"] = df[numerical_cols].kurtosis()
    print(stats_df)

    plt.figure(figsize=(15, len(numerical_cols) * 4))
    for i, col in enumerate(numerical_cols):
        plt.subplot(len(numerical_cols), 2, 2 * i + 1)
        sns.histplot(df[col].dropna(), kde=True)
        plt.title(f"Distribution of {col}")

        plt.subplot(len(numerical_cols), 2, 2 * i + 2)
        sns.boxplot(x=df[col].dropna())
        plt.title(f"Boxplot of {col}")
```

Función para hacer un estudio descriptivo, considerando el sesgo y la curtosis de los datos. Además, muestra el histograma de los datos numéricos y genera un boxplot para identificar visualmente los valores atípicos. Finalmente guarda la información visual en un archivo png.

Figura 10

Código de función para un estudio exploratorio de variables categóricas

```
def categorical_analysis(df):
    categorical_cols = df.select_dtypes(include=["object", "category"]).columns

    if len(categorical_cols) == 0:
        print("\nNo categorical columns found in the dataset.")
        return

    print("\n=== CATEGORICAL COLUMNS ANALYSIS ===")
    for col in categorical_cols:
        print(f"\nColumn: {col}")
        value_counts = df[col].value_counts()
        print(f"Number of unique values: {len(value_counts)}")

        if len(value_counts) <= 20:
            print(value_counts)

            plt.figure(figsize=(12, 6))
            sns.countplot(y=col, data=df, order=df[col].value_counts().index[:20])
            plt.title(f"Count of {col}")
            plt.tight_layout()
            plt.savefig(f"categorical_{col}.png")
            plt.close()
            print(f"Plot for {col} saved as 'categorical_{col}.png'")
        else:
            print(f"Too many unique values ({len(value_counts)}). Showing top 20:")
            print(value_counts.head(20))
```

Esta función permite hacer un estudio exploratorio de las variables categóricas, cuenta cuántas categorías hay en cada columna categórica, además, de su frecuencia, y crea una gráfica de representación de las distribuciones si existen variables categóricas.

Figura 11

Código de función para explorar relaciones lineales

```
def correlation_analysis(df):
 numerical_cols = df.select_dtypes(include=["int64", "float64"]).columns

 if len(numerical_cols) < 2:
  print("\nNot enough numerical columns for correlation analysis.")
  return

 print("\n=== CORRELATION ANALYSIS ===")
 corr_matrix = df[numerical_cols].corr()

 plt.figure(figsize=(12, 10))
 sns.heatmap(corr_matrix, annot=True, cmap="coolwarm", fmt=".2f", linewidths=0.5)
 plt.title("Correlation Matrix")
 plt.tight_layout()
 plt.savefig("correlation_matrix.png")
 plt.close()
 print("\nCorrelation matrix saved as 'correlation_matrix.png'")

 print("\nHighly correlated features (|r| > 0.7):")
 high_corr = np.where(np.abs(corr_matrix) > 0.7)
 high_corr_list = [
  (corr_matrix.columns[x], corr_matrix.columns[y], corr_matrix.iloc[x, y])
  for x, y in zip(*high_corr, strict=False)
  if x != y and x < y
 ]
```

Esta función explora las relaciones lineales a través de la correlación, muestra un "heatmap" y lo interpreta mostrando los valores mayores a 0.7 o menores a -0.7 en coeficiente de correlación.

Figura 12

Código de función para mostrar valores atípicos

```
def outlier_detection(df):
    numerical_cols = df.select_dtypes(include=["int64", "float64"]).columns

    if len(numerical_cols) == 0:
        print("\nNo numerical columns for outlier detection.")
        return

    print("\n=== OUTLIER DETECTION ===")
    for col in numerical_cols:
        Q1 = df[col].quantile(0.25)
        Q3 = df[col].quantile(0.75)
        IQR = Q3 - Q1

        lower_bound = Q1 - 1.5 * IQR
        upper_bound = Q3 + 1.5 * IQR

        outliers = df[(df[col] < lower_bound) | (df[col] > upper_bound)][col]

        if len(outliers) > 0:
            print(f"\nColumn: {col}")
            print(f"Number of outliers: {len(outliers)}")
            print(f"Percentage of outliers: {(len(outliers) / len(df)) * 100:.2f}%")
            print(f"Range of outliers: [{outliers.min()}, {outliers.max()}]")
```

La función muestra los valores atípicos por cada variable numérica. Para esto se usa la técnica del Intervalo intercuartílico. Se cuenta el porcentaje de atípicos y el rango que se considera para cada variable.

Figura 13

Código de función para análisis descriptivo de series de tiempo

```
def time_series_check(df):
 date_cols = df.select_dtypes(include=["datetime64"]).columns

 object_cols = df.select_dtypes(include=["object"]).columns
 potential_date_cols = []

 for col in object_cols:
  try:
   pd.to_datetime(df[col], errors="raise")
   potential_date_cols.append(col)
  except (ValueError, TypeError) as e:
   print(e)
   pass

 if len(date_cols) == 0 and len(potential_date_cols) == 0:
  print("\nNo datetime columns found in the dataset.")
  return

 print("\n=== TIME SERIES ANALYSIS ===")

 for col in date_cols:
  print(f"\nDatetime column: {col}")
  print(f"Min date: {df[col].min()}")
  print(f"Max date: {df[col].max()}")
  print(f"Range: {df[col].max() - df[col].min()}")
```

Se realiza un análisis descriptivo de los datos como series de tiempo, se identifica la Fecha mínima y la fecha máxima, además el rango de días trabajados.

Selección y Transformación de Características

Después de explorar los datos se propone hacer transformaciones de los mismos, es así que siguiendo recomendaciones de expertos y según Turner et al. (2015), dado que la variabilidad diurna de la radiación UV se modela mediante una ecuación sinusoidal, se sugiere transformar la variable HORA utilizando funciones de seno y coseno para capturar su comportamiento cíclico con corte de 12 horas. Por lo cual en la Figura 14 se puede observar para la variable MES.

Figura 14

Transformación de la variable HORA y MES

```
def handle_datetime(df, format="%d-%b-%Y %H:%M", remove_date=True):
  new_df = df.copy()
  new_df["DATETIME"] = pd.to_datetime(new_df["DATETIME"], format=format)
  new_df = new_df.assign(
    YEAR=new_df["DATETIME"].dt.year,
    MONTH=new_df["DATETIME"].dt.month,
    HOUR=new_df["DATETIME"].dt.hour,
  )

  # hora como variable ciclica (0 a 23)
  new_df["HOUR_SIN"] = np.sin(2 * np.pi * new_df["HOUR"] / 24)
  new_df["HOUR_COS"] = np.cos(2 * np.pi * new_df["HOUR"] / 24)

  # mes como variable ciclica (1 a 12)
  new_df["MONTH_SIN"] = np.sin(2 * np.pi * new_df["MONTH"] / 12)
  new_df["MONTH_COS"] = np.cos(2 * np.pi * new_df["MONTH"] / 12)
```

Modelos para Anomalías

Consideramos como anomalía a datos fuera de lo común dentro del estudio, para analizar estos datos utilizaremos varios enfoques:

Se iniciará con un enfoque desde el algoritmo DBSCAN (Density-Based Spatial Clustering of Applications with Noise) con el objetivo de identificar si existe alguna observación que no se puede explicar.

Un segundo paso será aplicar SARIMA, debido a que los datos son series de tiempo, el objetivo es identificar si existe estacionalidad y modelar la serie, los datos se consideran autorregresivos debido a su naturaleza temporal.

Finalmente, se usará el método de *Random Forest* con el objetivo de iniciar una modelación desde el enfoque de la predicción, considerando que *Random Forest* es resistente al sobreajuste de los datos tratados.

DBSCAN. El algoritmo DBSCAN (Density-Based Spatial Clustering of Applications with Noise), en el enfoque de series de tiempo es útil para la detección de ruido, considerando análisis por partes en los extremos y la centralidad. Es bueno para la detección de cambios

radicales durante periodos del día como el caso de la radiación solar durante la noche que es de cero.

El DBSCAN identifica como en ciertos momentos existen picos o descensos y si algo está fuera de lo normal lo identifica. La ventaja de aplicar el DBSCAN es que no necesita especificar grupos de clasificación. Además, se pueden aplicar con ventanas de tiempo para realizar subconjuntos de análisis, la ventana de tiempo usada en este estudio es 12 horas, se hicieron pruebas preliminares para determinar esto, se infiere que es por el comportamiento del sol y sus movimientos, ver en la Figura 15. Como parámetros principales se utiliza el valor epsilon de 0.5, y como mínimo número de muestras en el vecindario el valor de 5. Cabe recalcar que las ventanas de tiempo se normalizan con el *StandarScaler* y se utilizará el *Siluette Score* como métrica de Evaluación, simultáneamente con el *Calinski-Harabaz Score*.

Figura 15

Muestra del código DBSCAN para la variable UV_INDEX

```
# Accedemos a la columna 'UV_INDEX' del DataFrame X
X_series = X['UV_INDEX']
X_sliding_windows = to_sliding_windows(X_series, window_size=12) # Aplicamos la función a la serie
X_scaled = StandardScaler().fit_transform(X_sliding_windows) # Escalamos las ventanas

dbscan = DBSCAN(eps=0.5, min_samples=5)
labels = dbscan.fit_predict(X_scaled)

# Anomalías son las ventanas con etiqueta -1
anomalies = np.where(labels == -1)[0]
```

```
anomalies = np.where(labels == -1)[0]

window_size = 12
anomaly_indices_in_series = anomalies + window_size - 1
```

```
anomaly_values = X_series.iloc[anomaly_indices_in_series]

anomalies_list = list(zip(anomaly_indices_in_series, anomaly_values.values))

for idx, val in anomalies_list:
    print(f"Índice: {idx}, Valor UV_INDEX: {val}")
```

SARIMA. Otro enfoque para el análisis de las anomalías es a través de SARIMA, cuyo objetivo es identificar la estacionalidad de los datos, se debe considerar que las anomalías pueden provocar que la estacionalidad de los datos se desplace o se disperse.

La Figura 16 muestra una muestra del código SARIMA que está diseñado para entrenar, evaluar, predecir y detectar anomalías en series temporales. Además, permite ajustar el modelo

con parámetros optimizados en función de la métrica el AIC. También incluye funciones para evaluar el modelo comparando los valores reales con las predicciones, detectar puntos anómalos si se salen del intervalo de confianza (95%), además de mostrar un resumen de las anomalías desde el punto de vista estacional.

Figura 16

Ejemplo del modelo SARIMA

```
class SARIMAModel:
  def __init__(
    self,
    target_column,
    path="sarima_model.joblib",
  ):
    self.model_path = path
    self.target_column = target_column
    self.anomalies = None

  def train(self, X):
    model = SARIMAX(
      X,
      order=(2, 0, 2),
      seasonal_order=(1, 1, 1, 12),
      enforce_stationarity=False,
      enforce_invertibility=False,
    )
```

El modelo usa un esquema SARIMA (p,d,q) x (P,D,Q,s), donde evalúa la parte estacional y no estacional de los datos. Recordando que p indica el valor autorregresivo, P el orden estacional, de la diferenciación que se desea aplicar, D la diferenciación entre estaciones, q, el promedio móvil usado, y Q, el orden estacional para medir el error, s es el periodo que se considera la estación en este caso 12. La Arquitectura del modelo SARIMA es: (2,0,2) x (1,1,1,12).

Modelos para predicción

Una vez evaluado la existencia de las anomalías, se pretende proponer modelos de predicción, y contrastar cómo se comportan estos con los datos anómalos dentro de las series. Para esto se utilizarán los siguientes modelos con las siguientes estructuras, tomados después de analizar varias posibilidades a lo largo de la maestría:

Random Forest. Partiendo de la Figura 17, donde vemos parte del código Random Forest, para este caso se utilizó un modelo con 100 árboles, como un ensamble de árboles, se evalúa bajo las métricas del MSE, MAE y R^2. El objetivo del Random Forest es identificar qué variables generan una relación con la predicción de los datos, a pesar de tener valores anómalos. Será un modelo útil para ver el comportamiento de la predicción y tener un punto de partida apropiado evitando el sobreajuste.

Figura 17

Código para usar Random Forest

```
class RandomForestModel:
  def __init__(self, path="rf_model.joblib"):
    self.model = RandomForestRegressor(n_estimators=100, random_state=28)
    self.model_path = path

  def train(self, X, y):
    self.model.fit(X, y)
    joblib.dump(self.model, self.model_path)
```

SVM (One Class). Se codificó el modelo Support Vector Machine para regresión, una muestra de este modelo está en la Figura 18, en este caso se utilizaron varios kernel por ejemplo el RBF (Radial Basis Function), debido a que es apropiado para relaciones lineales y no lineales, de igual forma el linear y el polinomial. El código busca y optimiza la selección del mejor kernel. Además busca cómo controlar la regularización con varias valores (0.1, 1, 10, 100, 1000) que permiten penalizar el error, algo similar con el epsilon, que permite definir el

margen de tolerancia entre (0.01, 0.1, 0.2, 0.5, 1) y gamma (0.001, 0.01, 0.1, 1, 10) que controla el alcance del modelo en torno a los datos. El modelo utiliza como métricas de evaluación los valores de MSE, MAE y R^2.

Figura 18

Modelo SVM

```
class SVMModel:
 def __init__(self, path="svm_model.joblib"):
  self.model = SVR(kernel="rbf")
  self.model_path = path

 def train(self, X, y):
  self.model.fit(X, y)
  joblib.dump(self.model, self.model_path)
```

LSTM. Este código mostrado en la Figura 19, muestra el modelo LSTM, diseñado para tareas de predicción con series de tiempo mediante. Esta es una red neuronal recurrente tipo *Long Short-Term Memory.* La arquitectura del modelo consta de tres capas LSTM con 50 neuronas cada una, intercaladas con capas *Dropout* del 20 % para reducir el sobreajuste. Posteriormente, se agregan dos capas densas: una con 25 neuronas y otra con una, representando la salida. El método utiliza como función de pérdida el MSE y el optimizador adam. El entrenamiento incluye validación automática (validation split=0.2) y dos *callbacks*: *Early Stopping* para evitar sobreentrenamiento y *ReduceLROnPlateau* para ajustar la tasa de aprendizaje si no mejora la validación.

Se busca tratar de predecir valores futuros en una secuencia temporal (por ejemplo, temperatura, demanda eléctrica, contaminación, etc.) a partir de ventanas de trabajo. Se utiliza como métricas de evaluación el MSE, MAE y R^2.

Figura 19

Código para el modelo LSTM

```
class LSTMModel:
  def __init__(
    self, window_size, features, output_shape=1, path="lstm_model.h5"
  ):
    self.window_size = window_size
    self.input_shape = (window_size, features)
    self.output_shape = output_shape
    self.model = self._build_model()
    self.model_path = path

  def _build_model(self):
    model = Sequential()
    model.add(LSTM(50, return_sequences=True, input_shape=self.input_shape))
    model.add(Dropout(0.2))
    model.add(LSTM(50, return_sequences=True))
    model.add(Dropout(0.2))
    model.add(LSTM(50))
    model.add(Dropout(0.2))
    model.add(Dense(25))
    model.add(Dense(self.output_shape))

    model.compile(loss="mean_squared_error", optimizer="adam")

    return model

  def train(self, X, y, epochs=100):
    early_stopping = EarlyStopping(
      monitor="val_loss", patience=20, restore_best_weights=True
    )
    reduce_lr = ReduceLROnPlateau(
      monitor="val_loss", factor=0.5, patience=10, min_lr=1e-6
    )
```

MLP. El modelo de red neuronal tipo MLP (Perceptrón Multicapa) orientado a regresión programado en este estudio, se puede observar en la Figura 20 y tiene como objetivo principal diseñar, entrenar, evaluar y visualizar el rendimiento de un modelo neuronal. Para ello, el modelo incluye una etapa de reducción de dimensionalidad mediante PCA (Análisis de Componentes Principales) y posteriormente utiliza un MLP de regresión cuya arquitectura dispone de dos capas ocultas (de 16 y 8 neuronas respectivamente), activación ReLU, y entrenamiento mediante el optimizador Adam con tasa de aprendizaje adaptativa. Este enfoque permitió reducir la complejidad del modelo, prevenir sobreajuste y adaptarse a distintos tipos de datos al detectar relaciones no lineales entre variables.

Cabe recalcar que la reducción del PCA está en 8 dimensiones, para luego ser entrenado el modelo y así evitar el sobreajuste.

Se evalúa este modelo con las métricas MSE, MAE, R2 y visualizaciones de resultados como curvas de pérdida, gráficos de residuales, matriz de confusión y curvas ROC.

Figura 20

Código del modelo MLP

```
class MLPModel:
  def __init__(self, path="mlp_model.joblib"):
    self.model = Pipeline(
      [
        ("pca", PCA(n_components=8)),
        (
          "model",
          MLPRegressor(
            hidden_layer_sizes=(16, 8),
            activation="relu",
            solver="adam",
            alpha=0.5,
            random_state=28,
            learning_rate="adaptive",
            learning_rate_init=0.001,
            max_iter=500,
            early_stopping=True,
```

CAPÍTULO 4:

4. ANÁLISIS DE RESULTADOS

4.1. Comportamientos y modelos

Preprocesamiento

Los datos originales tienen 6865 filas, un vistazo rápido reveló que existen valores vacíos, duplicados, a pesar de ello fallas de tipeo o datos basura no se encuentran. Cabe recalcar que los datos de la estación meteorológica son previamente "pre procesados" por la propia estación, pero valores vacíos y duplicados no detecta durante la toma. Se infiere que los valores están vacíos por problemas de fallas del sensor o mantenimiento programado, mientras que las filas duplicadas, es por problemas de reinicio de los sensores. Existen muchos valores de cero, pero estos valores no son errores, sino son valores de cero índices ultravioletas durante la noche. Aquí se evidencia que el comportamiento debe ser en ciclos de 12 horas aproximadamente.

Los datos faltantes se imputaron con la técnica *backward*, es decir se colocaron los datos faltantes igual a el dato anterior. Cabe recalcar que esta técnica es la que utiliza la estación en caso de detectar faltantes y la que se usa como protocolo de manejo de la estación.

Exploración

El código para iniciar con la exploración de los datos se llama "explore.py", los resultados de este código se muestran a continuación en diversas figuras:

La Figura 21, muestra la vista general de la base de datos con un total de 6865 filas y 13 columnas, además que identifica el tipo de variables disponibles, cabe recalcar que existe la variable "Date", indicando así el comportamiento de la base como una serie de tiempo.

Mientras que la Figura 22, muestra que la base no posee ni valores duplicados, ni perdidos en esta ocasión. Si estos valores existieran, se indicaría un porcentaje y se debería remover los duplicados e imputar los faltante.

Figura 21

Vista General de la base de datos

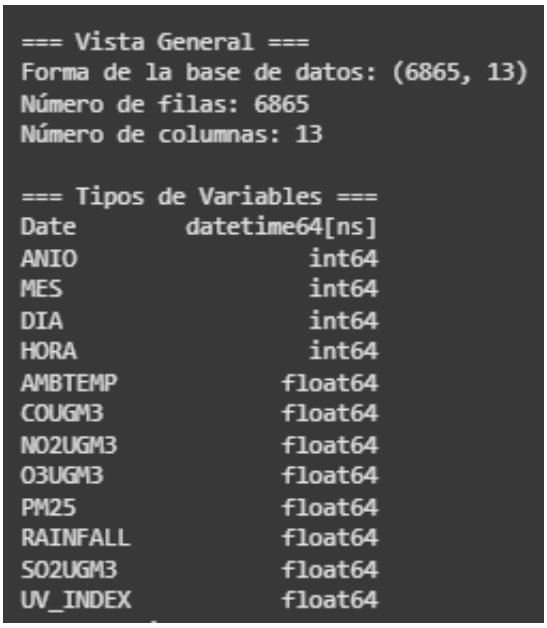
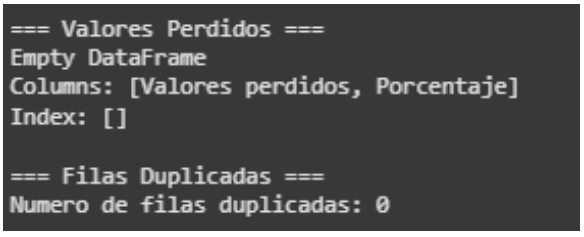


Figura 22

Valores perdidos y Duplicados



La Figura 23 y 24, muestran la exploración descriptiva y visual de las variables numéricas respectivamente, donde se puede apreciar que existen variables que tienen problemas de sesgo e incluso curtosis, además de valores de desviación estándar muy elevados. Por ejemplo, se aprecia que la temperatura ambiente tiene un comportamiento normal, sin embargo, el resto de variables no, cada una tienen una explicación específica,

pero concentrándose en el índice de radiación ultravioleta se observa que tienen un comportamiento que muestra valores fuera de lo común, debido a la gran presencia de atípicos en su boxplot, además varía mucho en relación con la desviación estándar y la media. Es importante mencionar que la sequía que azotó el país modificó el comportamiento de estas variables y por eso varias de ellas tienen valores de curtosis muy elevados, ejemplo la variable RAINFALL, COUGM3 y SO2UGM3. El estudio completo visual de todas las variables se encuentra en el apéndice de este estudio.

Es importante mencionar que también existe una función para preparar las variables categóricas y codificarlas, para poder trabajar, sin embargo, en esta base de datos no se utilizó esta función (ver apéndice de código completo).

Figura 23

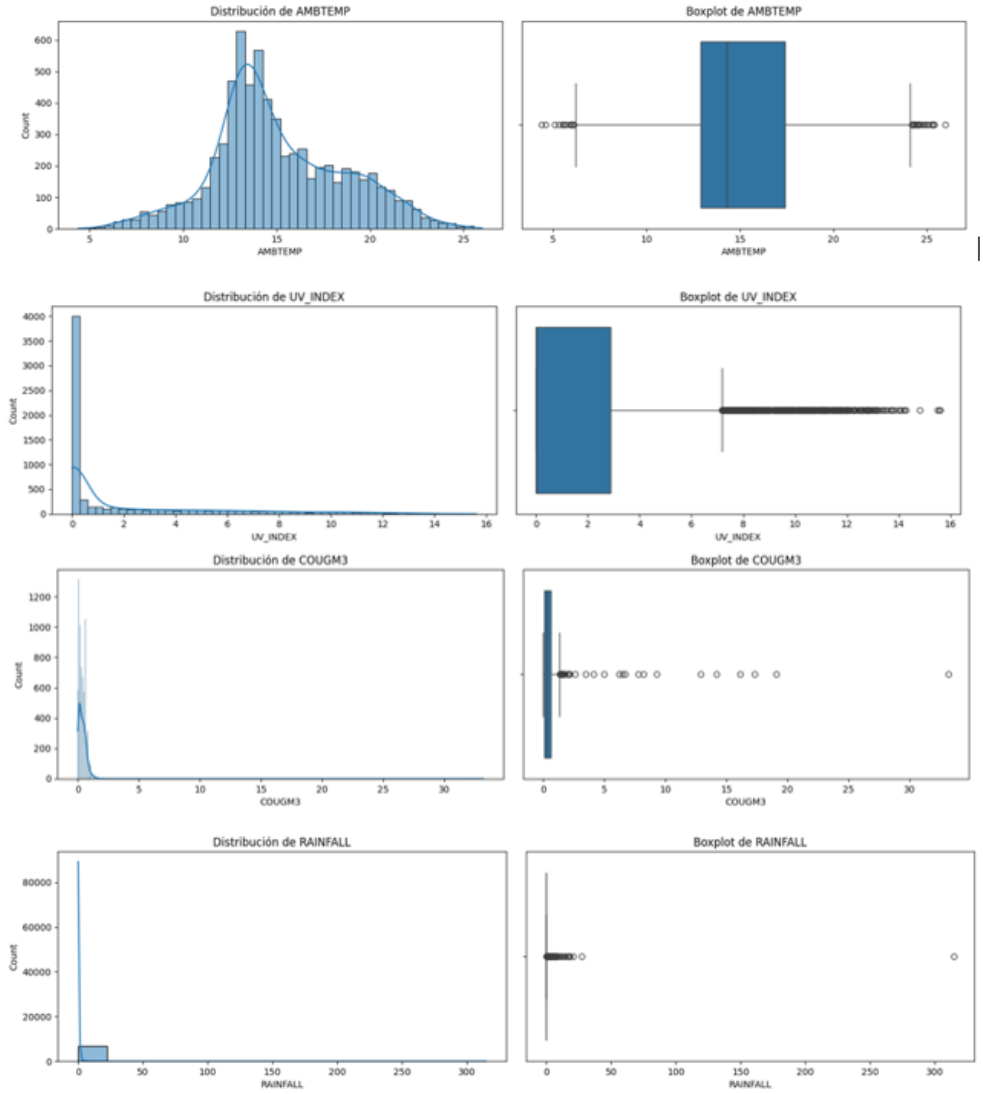
Exploración descriptiva de la base de datos

```
=== Estadística Descriptiva de las Columnas ===
           count         mean        std      min      25%         50%  \
ANIO      6865.0  2024.356737   0.479071  2024.00  2024.00  2024.000000
MES       6865.0     6.905171   3.787541     1.00     3.00     8.000000
DIA       6865.0    15.349308   8.847042     1.00     8.00    15.000000
HORA      6865.0    11.498325   6.923578     0.00     5.00    11.000000
AMBTEMP   6865.0    15.060070   3.492150     4.40    12.90    14.300000
COUGM3    6865.0     0.398739   0.687667     0.00     0.10     0.300000
NO2UGM3   6865.0    16.951097  13.764538     0.00     6.60    13.900000
O3UGM3    6865.0    35.103481  25.178378     0.70    14.00    30.200000
PM25      6865.0    10.664383  12.098066     0.21     4.36     6.880000
RAINFALL  6865.0     0.164399   3.895602     0.00     0.00     0.000000
SO2UGM3   6865.0     3.041002  16.168760     0.00     1.10     1.900000
UV_INDEX  6865.0     1.933376   3.197280     0.00     0.00     0.008573

                75%          max       skew     kurtosis
ANIO     2025.00000  2025.000000   0.598261    -1.642563
MES        10.00000    12.000000  -0.285464    -1.389875
DIA        23.00000    31.000000   0.072457    -1.202238
HORA       17.00000    23.000000   0.000058    -1.204250
AMBTEMP    17.40000    26.000000   0.314740    -0.097261
COUGM3      0.60000    33.200000  25.857678   975.781821
NO2UGM3    24.00000   250.700000   1.941254    14.178109
O3UGM3     50.20000   335.600000   1.191637     4.642157
PM25       12.13000   248.490000   4.880779    46.137673
RAINFALL    0.00000   314.800000  76.953335  6202.729704
SO2UGM3     2.90000   761.100000  32.327799  1214.317412
UV_INDEX    2.86953    15.613208   1.735085     2.079781
```

Figura 24

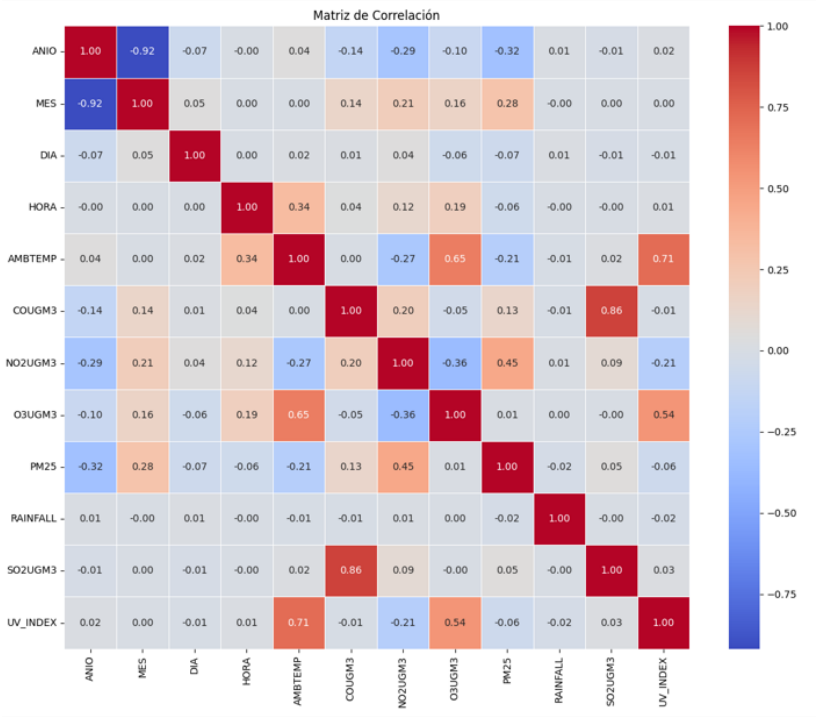
Exploración visual de la base de datos



El estudio de exploración, también nos muestra un diagrama de calor que está en la Figura 25, aquí se puede observar que existen correlaciones altas entre algunas variables por ejemplo el CO con el SO2, se podría inferir que es debido a los vientos y el tráfico vehicular. Una relación importante para el estudio es entre UV_INDEX y la temperatura ambiente con 0.71 aproximadamente. Por otro lado, no existen relaciones negativas significativas, salvo el PM2.5 con NO2, que se podría suponer debido al caos vehicular y los gases que generan. Queda claro que la variable RAINFALL no aporta en nada al estudio realizado.

Figura 25

Diagrama de calor de las variables de estudio



Adicional a estos estudios se explora si existen anomalías con la regla del comportamiento Intercuartílico, donde como resultado se puede visualizar en la Tabla 6.

Tabla 6

Resultados de detección de anomalías por exploración

Variable	Anomalías detectadas IQR	Porcentaje Anomalías
AMBTEMP	48	0.70%
COUGM3	66	0.96%
NO2UGM3	166	2.42%
O3UGM3	65	0.95%
PM25	563	8.20%
RAINFALL	593	8.64%
SO2UGM3	342	4.98%
UV_INDEX	683	9.95%

Los resultados indican que la variable UV_INDEX tiene un 9.95% de posibles anomalías, para esto se estudiará más a detalle la variable UV_INDEX a través de varias consideraciones, por ejemplo, como se explicó en la sección anterior en la selección y

transformación de características, se crearon cinco nuevas variables, como se muestra en la Figura 26.

Figura 26

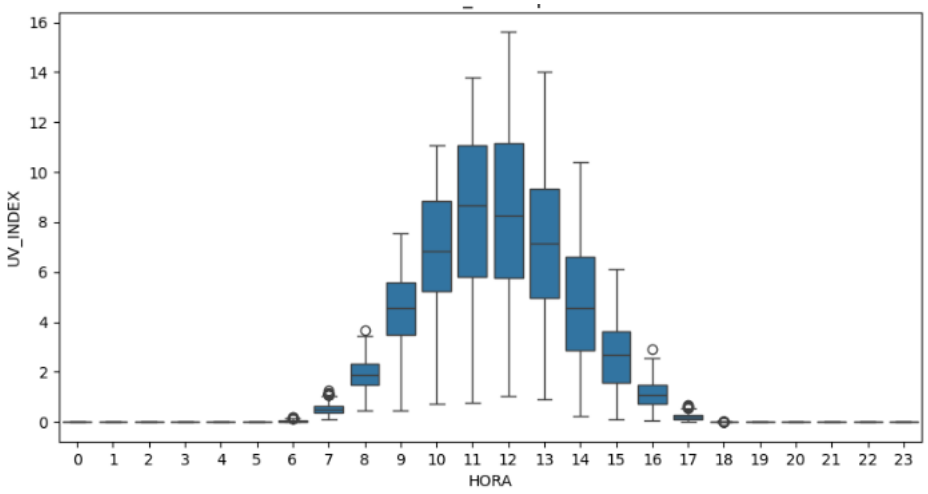
Variables creadas y agregadas a la base de datos

	ANIO	MES	DIA	HORA	AMBTEMP	COUGM3	NO2UGM3	O3UGM3	PM25	RAINFALL	SO2UGM3	UV_INDEX	HORA_sin	HORA_cos	MES_sin	MES_cos	DIA_SEMANA
0	2024	7	1	0	14.5	0.5	23.4	5.7	9.35	0.0	1.2	0.0	0.000000	1.000000	-0.5	-0.866025	0
1	2024	7	1	1	14.0	0.2	16.3	15.1	5.62	0.0	0.9	0.0	0.258819	0.965926	-0.5	-0.866025	0
2	2024	7	1	2	13.4	0.2	14.2	14.0	5.19	0.0	0.5	0.0	0.500000	0.866025	-0.5	-0.866025	0
3	2024	7	1	3	12.8	0.2	12.9	9.9	5.43	0.0	2.3	0.0	0.707107	0.707107	-0.5	-0.866025	0
4	2024	7	1	4	11.9	0.1	10.2	8.5	4.44	0.0	0.9	0.0	0.866025	0.500000	-0.5	-0.866025	0

Las variables resultantes son Hora_sin, Hora_cos, MES_sin, MES_cos, DIA_SEMANA. El objetivo de estas variables es identificar el comportamiento de la radiación ultravioleta, considerando que la naturaleza de las variables es cíclica, se usa cada 12 horas el ciclo, para las variables *Hora* y se prueba con la función seno y coseno, ver en la Figura 27.

Figura 27

Comportamiento de las variables UV_INDEX a lo largo del día



Algo similar para la variable *Mes*, por presencia de posible estacionalidad. Finalmente, la variable DIA_SEMANA codifica los días de la semana para identificar si

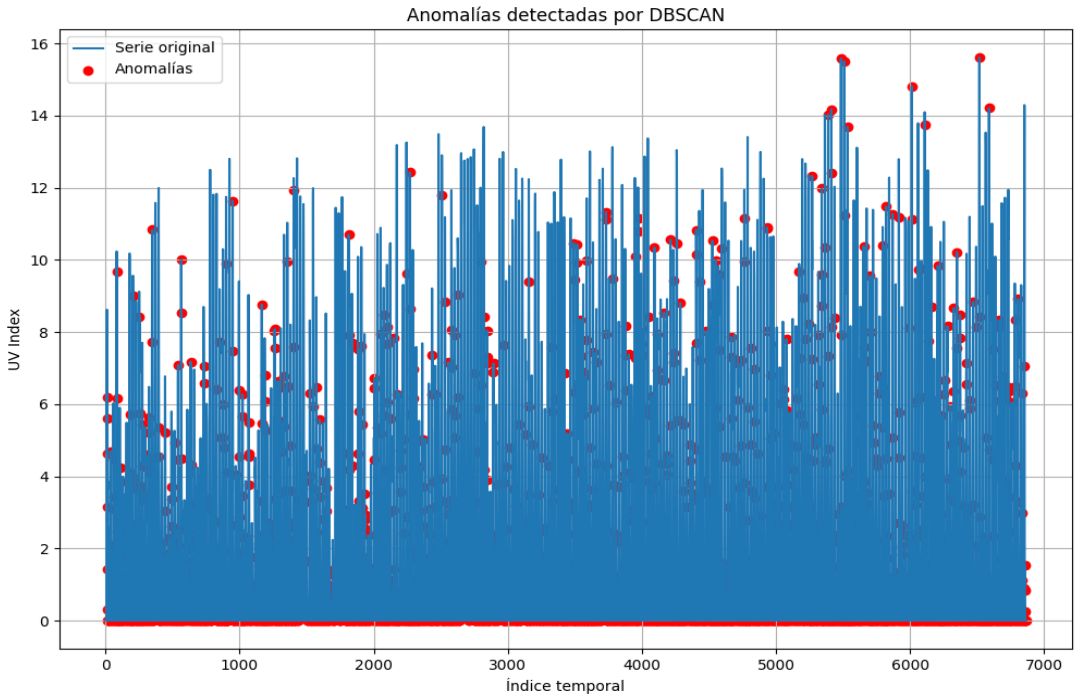
existe algún comportamiento en los fines de semana o algún día en particular, la codificación usada es: 0 - lunes, 1 - martes, 2 - miércoles, 3 - jueves, 4 - viernes, 5 - sábado, 6 - domingo.

Anomalías

DBSCAN muestra una serie temporal con anomalías detectadas que se evidencia en la Figura 28, se detectan 12 picos altos, y se pueden observar patrones que infieren cambios bruscos en las tomas. Si bien estos puntos son problemas, pero son normales en zonas de alta radiación como lo es Cuenca, se detecta que los puntos aparecen alrededor de las 11 de la mañana hasta las tres de la tarde aproximadamente.

Figura 28

Detección de anomalías por medio del DBSCAN

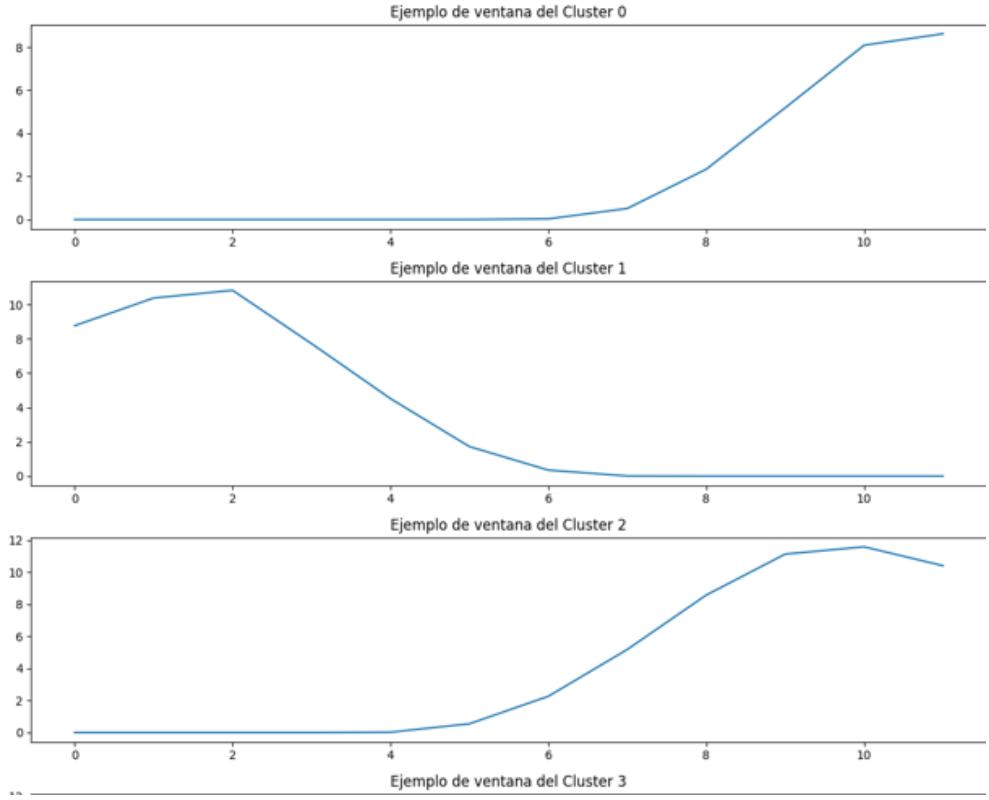


En la Figura 29, esta información se corrobora al ver las ventanas de tiempo, donde se observa que al clusterizar se genera un comportamiento de crecimiento – decrecimiento, que es lo esperado al trabajar con radiación solar, esta técnica permite identificar cómo se comporta en días soleados, y cómo funcionan los picos de descenso de radiación. “Se

observa que al aplicar clustering a las curvas diarias de radiación solar se identifican trayectorias de crecimiento y decrecimiento, tal como se espera en condiciones de radiación solar diurna" (Siriwardana y Kume, 2024).

Figura 29

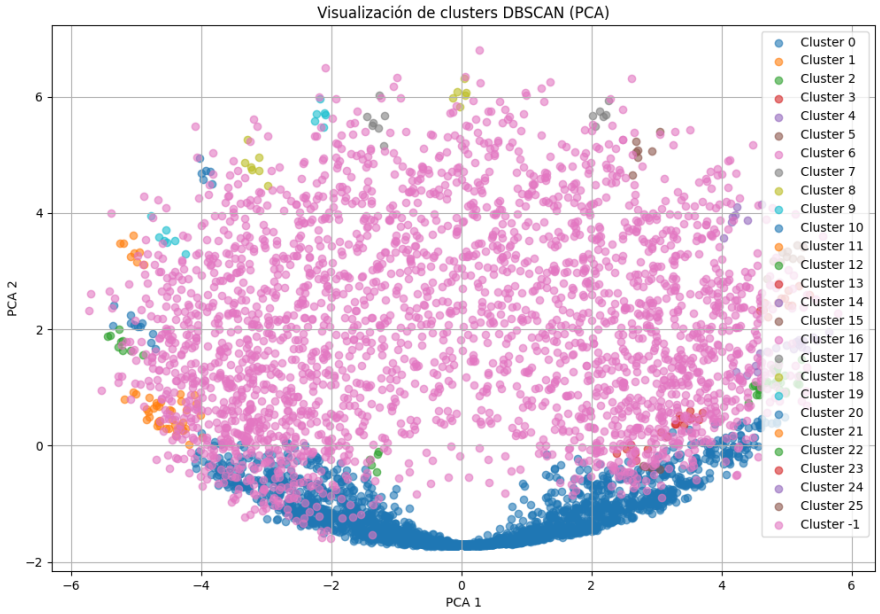
Evidencia del comportamiento cíclico de los datos



Se exploró la posibilidad de trabajar con menos dimensiones a través de un PCA con dos componentes como se muestra en la Figura 30, este enfoque nos permite identificar los valores que no se pudieron agrupar es decir posibles anómalos que se encuentran fuera de la nube de datos, además la forma en la que se muestran los datos ratifica el comportamiento cíclico que se mencionó en secciones anteriores.

Figura 30

PCA aplicado para reducir la dimensionalidad



SARIMA. Los resultados de este modelo indican claramente la presencia de estacionalidad, esto es un resultado esperado debido al comportamiento de la variable UV_INDEX en la naturaleza, por los movimientos del sol. Los valores de los resultados indican que los componentes autoregresivos del modelo son apropiados, en otras palabras, modelos que auto aprenden de sí mismos se ajustaban bastante bien. Sin dudas el modelo es fuertemente estacional como se observa en las gráficas de autocorrelación completa, parcial y estacional, ver Figuras 31 a la 33. Un estudio aislado de los residuos revela una curtosis de alrededor de 10.22 lo que indica que el comportamiento no es normal, por ende, modelos de características lineales no serían apropiados para estos datos expuestos. Una lectura profunda de las gráficas nos muestra que existe presencia de ruido blanco, quizá por la toma de datos (sensores) o por la ubicación de la estación, además no se detecta una autocorrelación significativa (Ljung-Box Test p = 0.67) que nos da un buen indicio de la validez de los datos y de que se pueden modelar a pesar de las posibles anomalías.

Figura 31

Análisis de Autocorrelación Completa, Parcial

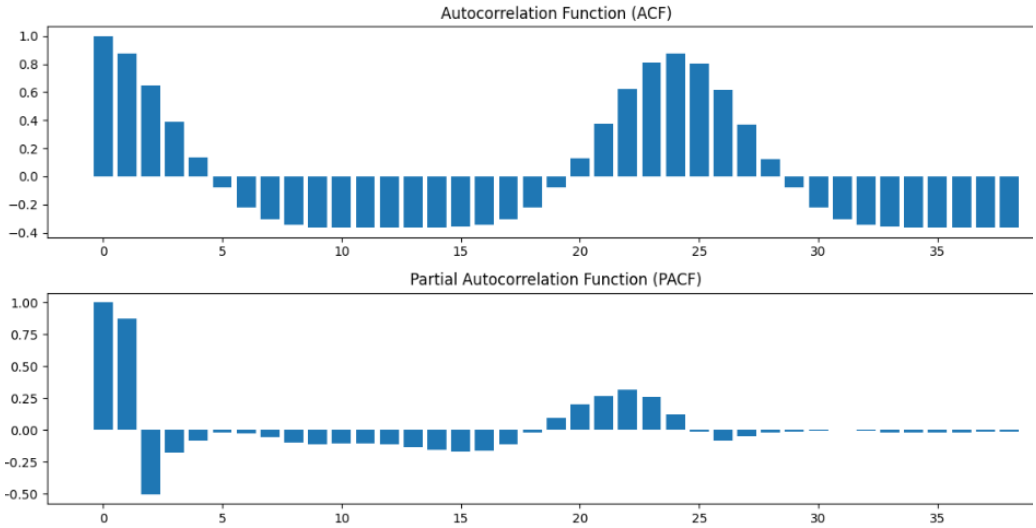


Figura 32

Análisis de Autocorrelación Completa, Parcial Estacional

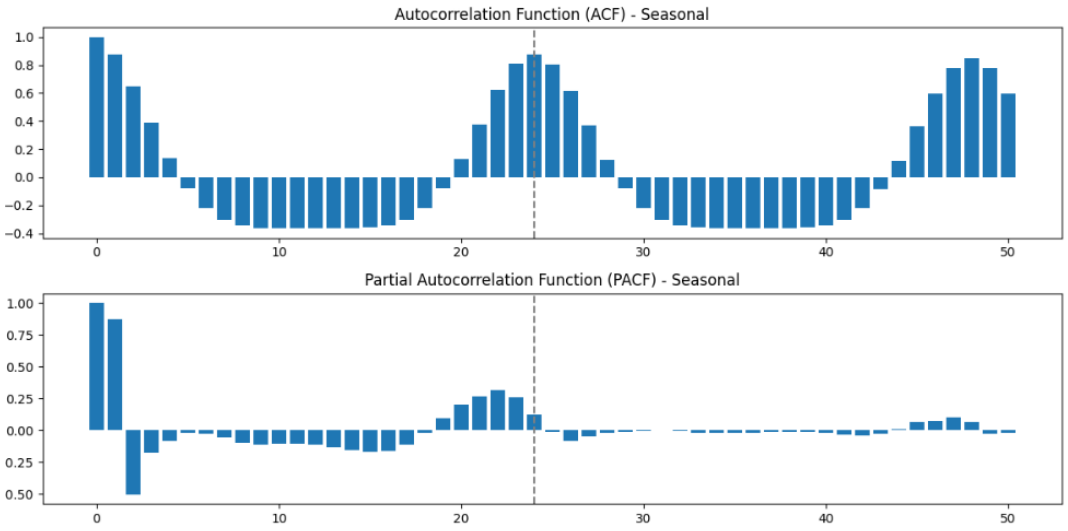
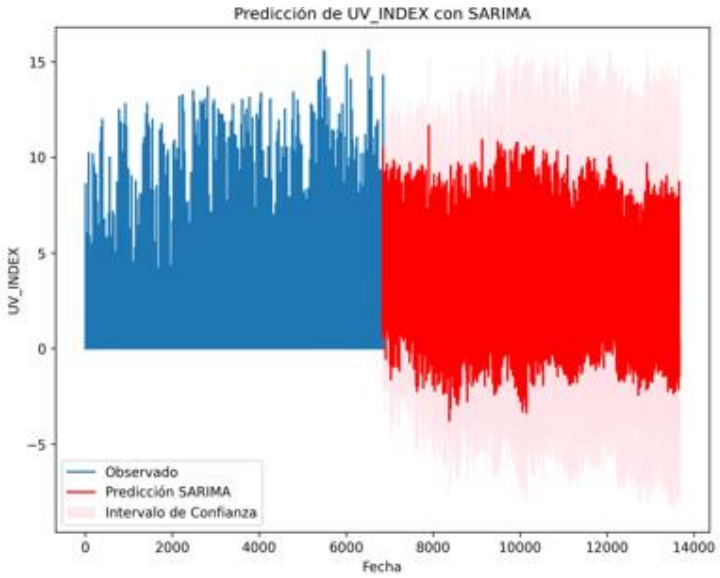


Figura 33

Comportamiento del modelo SARIMA para predicción



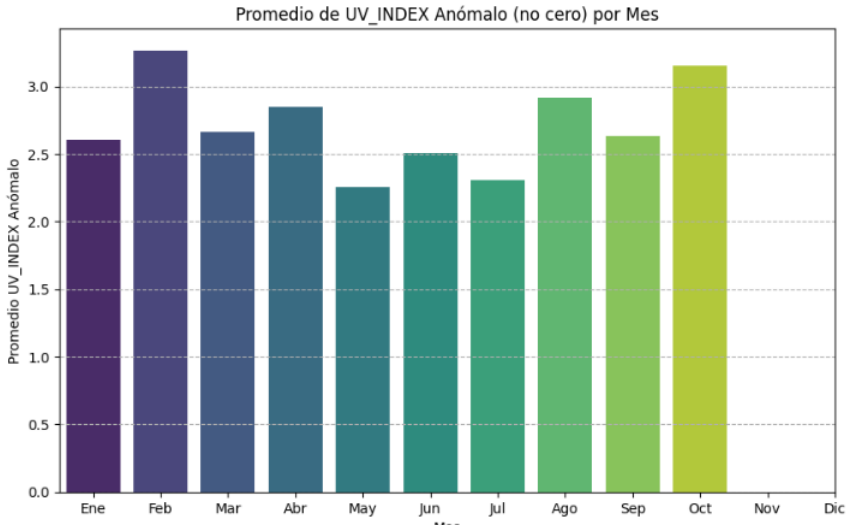
Los resultados de esta sección indican que, si bien existen anomalías en la base de datos, se debe trabajar con ellas y no retirarlas, sino que desean parte de predicciones, considerando que la estacionalidad permite que valores muy altos del índice de radiación ultravioleta se generen en las zonas de Cuenca - Ecuador.

Predecir a través del modelo SARIMA es posible debido a su diseño, de hecho, se obtiene un MSE de 3.26 y un MAE de 1.43, además de un R^2 de 0.6771, todo esto en un tiempo computacional de 82.83 segundos.

Adicional a esto se determina que los meses donde hay mayor presencia de anomalías son los meses de octubre y febrero como se ve en la Figura 34., se podría inferir que es por los cambios y efectos del fenómeno del niño, pero se requiere más información para validar esta inferencia.

Figura 34

Promedio de valores anómalos detectados por mes



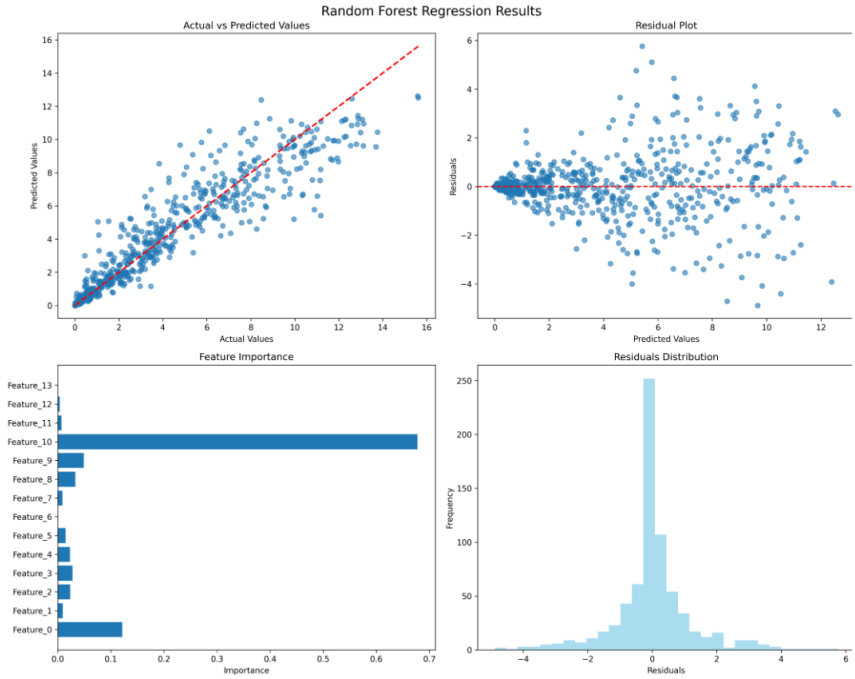
Predicción

Para iniciar con el entrenamiento de los modelos primero, es importante recordar que al ser una serie de tiempo se dividió entre Train y Test con un criterio 80-20, considerando el orden específico de la fecha. Los resultados de estos entrenamientos se verán en la Tabla 7, donde se comparan algunas métricas.

Random Forest. La evaluación estadística del modelo indica que los residuos se ajustan de manera apropiada al modelo propuesto, como se muestra en la Figura 35, esto da un indicio de que la predicción puede ser altamente certera, en la gráfica del Residual plot, de la misma figura, se observa la aleatoriedad de los residuales lo cual explica que los datos y su forma de preprocesamiento es adecuado sin incurrir a sesgos innecesarios. En la última sección de la figura se observa el comportamiento de la distribución de los residuos, mostrando una leve curtosis y consideraciones leptocurticas, ratificando que los datos están agrupados de forma periódica. Los resultados de las métricas de evaluación se muestran en la Tabla 7.

Figura 35

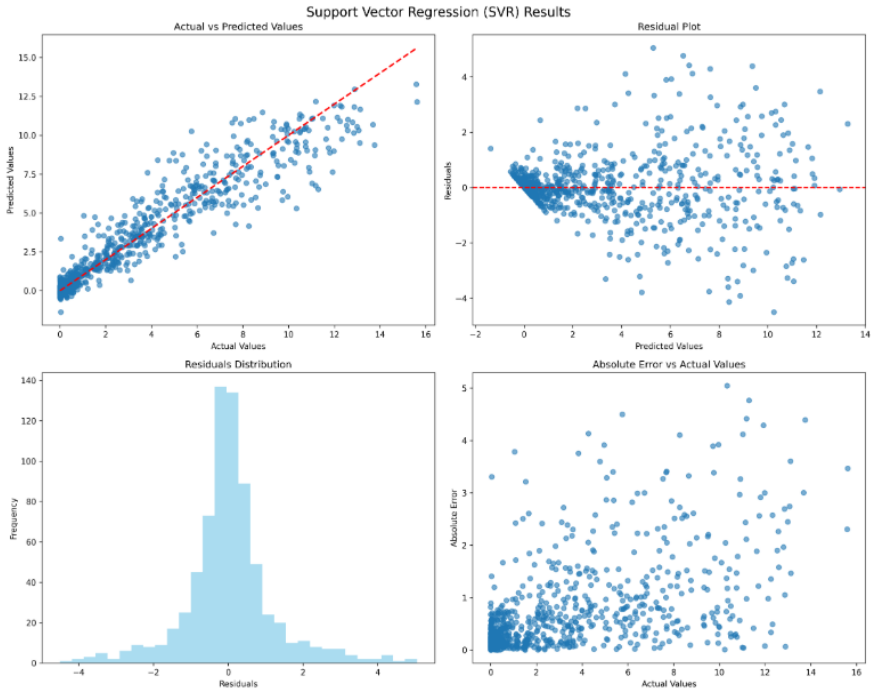
Resultados Residuales del modelo Random Forest



SVM utilizó un SVR (Support Vector Regression), la Figura 36 muestra los resultados del análisis residual, donde se aprecia como los valores predichos y los actuales tienden a tener una relación positiva creciente. Esta información indica que el modelo se adapta a la tendencia de los datos, respecto a la sección del "Residual Plot", los datos se encuentran dispersos, con una leve aglomeración inicial, posible indicio de heterocedasticidad dentro del modelo, dando a entender que las predicciones se alterarán conforme se alejen del del origen. Este suceso es normal en condiciones de una serie temporal. Algo similar muestra la gráfica de distribución con la curtosis presente. La gran dispersión residual en la gráfica del error absoluto da indicios que el SVR no será tan robusto a valores anómalos o fuera de control estadístico. Los resultados de las métricas de evaluación se muestran en la Tabla 7.

Figura 36

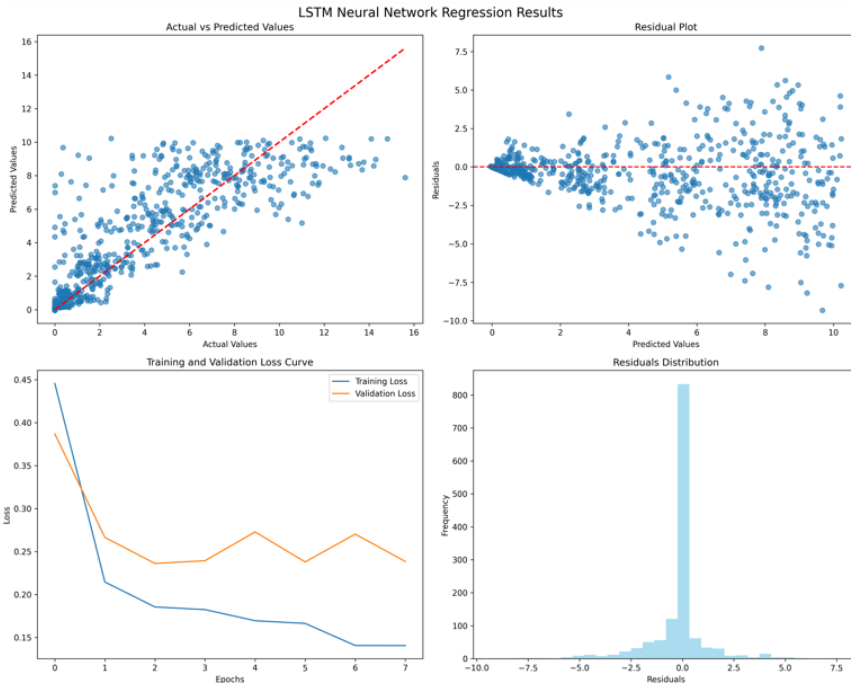
Resultados residuales del Support Vector Regression



LSTM. En la Figura 37, se muestran los resultados del modelo Long Short Term Memory (LSTM), se observa que los valores actuales y los predichos no se ajustan en comportamientos altos (anómalos), en otras palabras, la dispersión no es controlada, se puede intuir una baja generalización del modelo, esto se comprueba con la gráfica de residuales donde la dispersión también es alta en la base del cono, la acumulación de datos en el inicio y el final da evidencia de posible sobreajuste. En la misma figura se observa como las curvas de entrenamiento y evaluación no se ajustan dando evidencia que el modelo no es tan robusto.

Figura 37

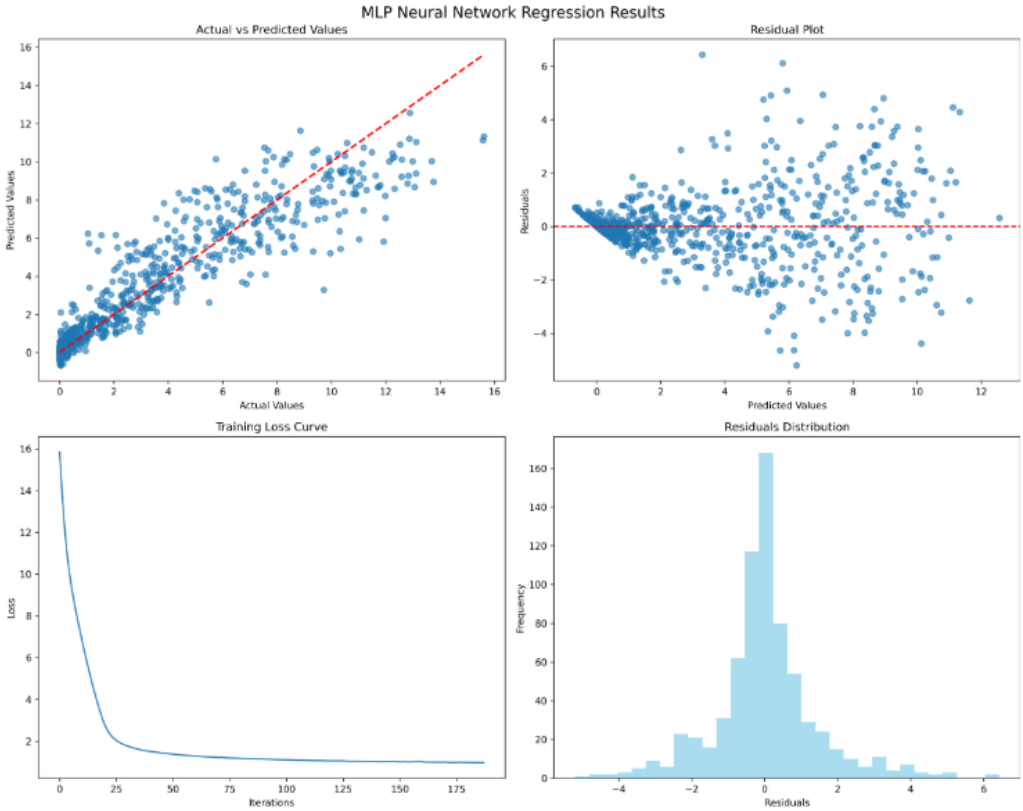
Resultados residuales del modelo LSTM



MLP se puede observar en la Figura 38, donde se muestra los resultados del modelo MLP, considerando que para un modelo de regresión los datos actuales versus los predichos se acoplan bastante bien al modelo lineal creciente, a pesar de ello se observa una leve inclinación en los valores altos, dando indicios que es necesario transformar más variables a través de formas matemáticas polinomiales, esto probablemente es debido a las anomalías presentes en el modelo. La forma de dispersión cónica indica la ciclicidad de los datos, pero su dispersión externa muestra la alta complejidad de manejo frente a valores fuera de lo común, a través de este modelo. Por otro lado, la curva de pérdida, muestra una buena convergencia y una disminución rápida del error. El histograma muestra un comportamiento aparentemente normal, con poco sesgo y alta curtosis (leptocurtismo). Cabe recalcar que durante el entrenamiento se probaron diferentes tipos de arquitecturas, pero nos quedamos con la mejor presentada en la sección anterior.

Figura 38

Resultados residuales del modelo MLP



A continuación, en la Tabla 7 se muestra la comparativa de los métodos predictivos.

Tabla 7

Comparación de los métodos de predicción

Métricas	SARIMA	Random Forest	SVM	LSTM	MLP
MSE	3.2668	0.8813	0.9089	2.0615	1.1403
MAE	1.4304	0.4206	0.5595	0.6620	1.1308
R^2	0.6771	0.9084	0.9055	0.7980	0.8824
Tiempo Ejecución	88.02 s	3.11 s	2.82 s	59.22 s	1.95 s

En la Tabla 7 muestra que los modelos con mejor desempeño considerando su MSE y R^2 son Random Forest y el SVM con valores de 0.8813 y 0.9089 de MSE respectivamente, y 0.9084 y 0.9085 de R^2, los valores indican que no existe evidencia de sobreajuste. Respecto al

error absoluto MAE que tiene la ventaja de ser menos sensibles a las anomalías muestra que Random Forest tiene un valor de 0.40, en comparación son el modelo estadístico SARIMA con un valor de 3.2668 siendo el más alto de los modelos. Se puede inferir que SARIMA predecirá bajo un intervalo de confianza mucho más amplio por eso el error es más grande.

Respecto a los modelos de redes neuronales LSTM y MLP, tuvieron tiempos de respuesta diferentes, en esta métrica el LSTM (55.22 segundos) fue más lento que el MLP (1.95 segundos), a pesar de ello el coeficiente de determinación del LSTM no supera al MLP cons 0.88. Desde otra perspectiva, los errores en especial el MAE fue mucho mejor en el LSTM con 0.66 frente al 1.1308 del MLP.

De forma general, para esta base de datos random Forest, SVM y MLP son modelos que se ajustan a las necesidades de trabajar frente a datos naturalmente anómalos, y ligados a una serie de tiempo.

4.2. Formas de uso del código

La documentación del código se encuentra en el link de la sección documentación en apéndices, sin embargo, se explica brevemente el funcionamiento:

- El código se entrega en un link de GITHUB ubicado en la sección de Apéndices.
- Se trabajó en UVICORN, donde están las versiones y librerías requeridas para poder utilizarse directamente los códigos, estos están de forma modular a través del manejo de clases.
- El código que convoca los modelos se llama compare.py, aquí están contenidos los modelos y el código de transformación y preparación de datos bajo los siguientes nombres:
    - transformers.py
    - dbscan_model.py

- lstm_model.py
- mlp_model.py
- random_forest_model.py
- sarima_model.py
- svm_model.py

Para usarse de manera individual solo se debe cambiar en la línea run de cada modelo la palabra skip=True, por skip= False, de esa manera se ejecutan los modelos individualmente, y sus transformaciones y preparación de la base de datos, al igual que la división en training y test seguirá siendo la misma.

4.3. Interfaz

Los modelos analizados se pueden utilizar en una interfaz, este es un demo aún, pero consta de dos partes, la Figura 39, muestra la primera parte donde se colocan los datos e información de la estación, y la segunda parte es la salida, en la Figura 40, actualmente solo esta para predecir el comportamiento del índice de radiación ultravioleta UV_INDEX.

Figura 39

Interfaz para captura de datos

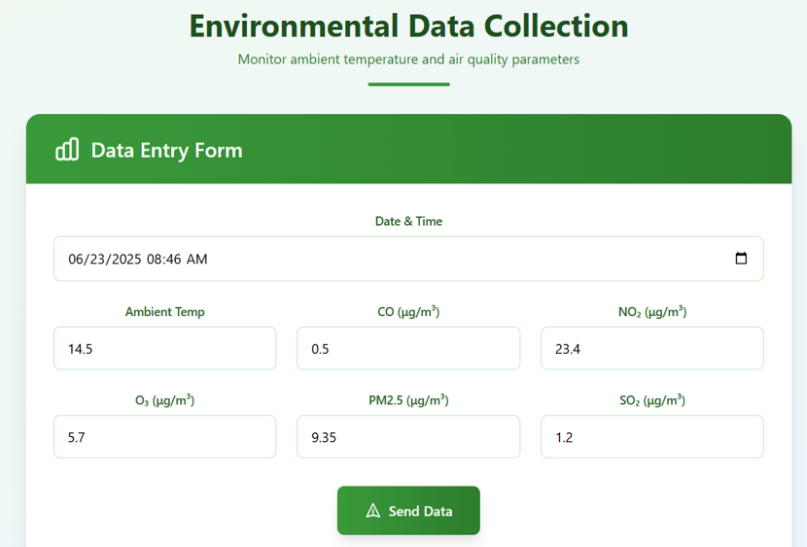
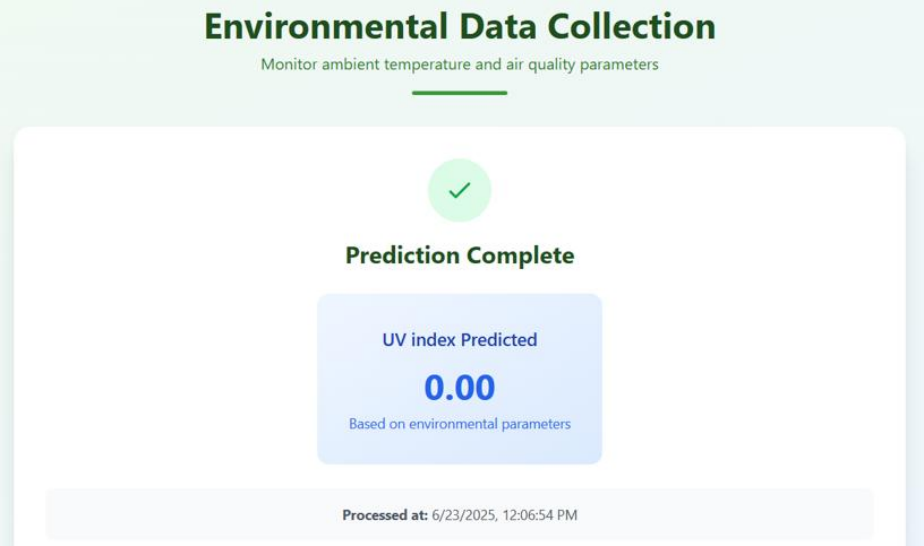


Figura 40

Interfaz para salida de resultado predictivo



CAPÍTULO 5:

5. CONCLUSIONES Y RECOMENDACIONES

Trabajar con anomalías en datos climatológicos es usual en las zonas de alta radiación, un ejemplo directo de ello es Ecuador ya que al estar ubicado en la línea ecuatorial los registros de radiación ultravioleta suelen ser extremos. Varios entes gubernamentales se han enfocado en tratar de caracterizar la radiación y el índice de radiación, pero Ecuador es un caso peculiar ya que al poseer zonas geográficas tan marcadas (costa, sierra, región insular, amazonia) en un espacio territorial reducido se presentan valores pocos comunes debido a la meteorología local, es aquí donde este trabajo muestra evidencia de cómo realizar un primer acercamiento a datos obtenidos con una estación meteorológica ubicada en la zona sierra sur del Ecuador.

A través del aprendizaje no supervisado, trabajar con datos meteorológicos ha sido una tarea bastante fructífera, si bien el enfoque general nos permitió identificar que efectivamente existen datos fuera de lo común en la base de datos, pero son datos reales no atípicos, por ende, debemos trabajar y aprender a predecir en torno a estos datos. Además, cabe recalcar que los datos por su naturaleza se encuentran en forma de series de tiempo, por ende, el uso de ciertos modelos debe estar restringido a este tipo de datos.

Este estudio identifica que las variables como temperatura ambiente, CO2, SO2, NO2, O3, cantidad de lluvia, pueden ser variables predictoras para el índice de radiación ultravioleta. Durante el preprocesamiento de los datos, se observó la importancia de transformar matemáticamente algunas variables a través de funciones cíclicas, como el seno y el coseno, esto mejoró en gran medida las métricas de evaluación. Esta aproximación tiene sentido ya que, al ser datos en forma de serie de tiempo, la estacionalidad está presente en la

variable dependiente, luego el reto fue encontrar la ventana de tiempo apropiada, y se observó que 12 horas es lo mejor, debido a que cada doce horas el sol empieza con su descenso y esto se nota en los valores registrados.

El estudio sigue una metodología clásica, donde la obtención de los datos, la limpieza, la exploración, el preprocesamiento, y la modelación con la columna principal. Todos los pasos permitieron que la base de datos pueda ser utilizada y replicada en los modelos seleccionados.

Dentro de la modelación, la primera parte se enfocó en detectar las anomalías aquí modelos como el DBSCAN, y SARIMA fueron de gran utilidad, para luego pasar la modelación con el fin de predecir el comportamiento a través de modelos desde los básicos hasta los modelos de aprendizaje profundo (Random Forest, SVM One class, LSTM, MLP). Se evaluaron los modelos con el uso de métricas como el MSE, MAE, R^2, además del tiempo de ejecución.

Desde un enfoque técnico, el mejor modelo sería el Random Forest y SVM, superando al clásico SARIMA, además los modelos de aprendizaje futuro dieron buenos rendimientos, pero su gasto computacional y tiempo de ejecución suele ser mayor. Se observó que modelos como el LSTM son sensibles a patrones ocultos, como se pudo ver en sus comportamientos residuales. Además, se verifica que la periodicidad de la serie de tiempo puede llegar a afectar a la mayoría de modelos, pero el MLP y LSTM no sufren este problema justo por su enfoque no lineal.

Finalmente, este estudio buscó integrar un enfoque estadístico con el aprendizaje de máquinas, a través de datos en series de tiempo, los resultados muestran que el hecho de poseer datos anómalos puede limitar el uso de ciertos modelos, a pesar de ello desde un enfoque no supervisado la metodología utilizada se ajusta con las enseñanzas en la maestría. Desde este punto se abren una serie de trabajos futuros, para enfocar a perfeccionar los

modelos, transformar variables, y fusionar con enfoque estadísticos paramétricos y no paramétricos los resultados después de analizar los datos de la estación meteorológica.

Referencias Bibliográficas

Aggarwal, C. C. (2015). *Data mining: The textbook*. Springer. https://doi.org/10.1007/978-3-319-14142-8

Ale, T., Schlegel, N.-J., & Janeja, V. P. (2024). Harnessing feature clustering for enhanced anomaly detection with variational autoencoder and dynamic threshold. *arXiv*. https://arxiv.org/abs/2407.10042

Alnutefy, S., & Alsuwayh, A. (2024). Unsupervised anomaly detection. *Computer Science & Information Technology (CS & IT), 14*, 145–154. https://aircconline.com/csit/papers/vol14/csit140210.pdf

Bâra, A., Văduva, A. G., & Oprea, S.-V. (2024). Anomaly detection in weather phenomena: News and numerical data-driven insights into the climate change in Romania's historical regions. *International Journal of Computational Intelligence Systems, 17*, 134. https://doi.org/10.1007/s44196-024-00536-2

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys, 41*(3), 1–58. https://doi.org/10.1145/1541880.1541882

Davis, W. L., Carlson, M., Tezaur, I., Bull, D., Peterson, K., & Swiler, L. (2024). Spatio-temporal multivariate cluster evolution analysis for detecting and tracking climate impacts. *arXiv*. https://arxiv.org/abs/2410.16544

Dharani, K., Satyanarayana, R., & Rao, V. (2022). A modified DBSCAN algorithm for anomaly detection in time-series data with seasonality. *The International Arab Journal of Information Technology, 19*(2), 229–236. https://www.ccis2k.org/iajit/index.php/archive/volume-19-2022/march-2022-no-2/item/2295-a-modified-dbscan-algorithm-for-anomaly-detection-in-time-series-data-with-seasonality

Doan, Q.-V., Amagasa, T., Pham, T.-H., Sato, T., Chen, F., & Kusaka, H. (2023). Structural k-means (S k-means) and clustering uncertainty evaluation framework (CUEF) for mining climate data. *Geoscientific Model Development, 16*(8), 2215–2233. https://doi.org/10.5194/gmd-16-2215-2023

Farooq, O., & Khan, A. (2025). The Impact of Machine Learning on Climate Change Modeling and Environmental *Sustainability. Artificial Intelligence and Machine Learning Review*, 6(1), 8-16.

Fehlmann, R., Sauter, T., & Spekat, A. (2023). Time series decomposition and trend detection of precipitation and temperature for Germany (1951–2021). *Atmosphere, 14*(6), 860. https://doi.org/10.3390/atmos14060860

Fernández, J., & Torres, R. (2019). Métodos estadísticos aplicados a la climatología: Detección de anomalías en registros meteorológicos. *Revista Iberoamericana de Ciencias Ambientales, 10*(2), 34–46.

Fioletov, V. E., Kerr, J. B., McArthur, L. J. B., & Wardle, D. I. (2003). Estimating UV Index climatology over Canada. *Journal of Applied Meteorology and Climatology, 42*(3), 417–430. https://doi.org/10.1175/1520-0450(2003)042%3C0417:EUICOC%3E2.0.CO;2

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review, 22*(2), 85–126. https://doi.org/10.1023/B:AIRE.0000045502.10941.a9

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2nd ed.). OTexts.

Instituto Nacional de Meteorología e Hidrología (INAMHI). (2025). https://www.inamhi.gob.ec

K, D. S., & Jisha, G. (2024). *Enhanced Weather Prediction with Feature Engineered, Time Series Cross Validated Ridge Regression Model*. 1–6. https://doi.org/10.1109/ciscon62171.2024.10696530

Liu, W., Wang, H., & Huang, X. (2022). Evaluating clustering algorithms for high-dimensional time series data. *Data Mining and Knowledge Discovery, 36*(1), 214–238. https://doi.org/10.1007/s10618-021-00762-1

Lo, A. Y., Judge-Lord, D., Hudson, K., & Mayer, K. H. (2023). Mapping Literature with Networks: An Application to Redistricting. *Political Analysis*, 1–10. https://doi.org/10.1017/pan.2023.4

Pacal, A., Hassler, B., Weigel, K., Kurnaz, M. L., Wehner, M. F., & Eyring, V. (2023). Detecting extreme temperature events using Gaussian mixture models. *Journal of Geophysical Research: Atmospheres, 128*(15), e2023JD038906. https://doi.org/10.1029/2023JD038906

Racah, E., Becker, R., Cottrell, L., Keesey, S., Dietterich, T. G., Prabhat, & Collins, N. C. (2017). ExtremeWeather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events. *Advances in Neural Information Processing Systems, 30*, 3402–3413.

Reddy, A. E., & Rajan, K. S. (2023). Spatiotemporal cluster analysis of gridded temperature data: A comparison between K-means and MiSTIC. *arXiv*. https://arxiv.org/abs/2307.00480

Siriwardana, A. N., & Kume, A. (2024). *Introducing the Spectral Characteristics Index: A novel method for clustering solar radiation fluctuations from a plant-ecophysiological perspective*. *Ecological Informatics, 85*,

102940. https://doi.org/10.1016/j.ecoinf.2024.102940

Shumway, R. H., & Stoffer, D. S. (2017). *Time series analysis and its applications: With R examples* (4th ed.). Springer.

Turner, S. D., Lark, R. M., & Da Silva, G. S. (2015). *Diurnal variability of UV radiation modeled by sinusoidal functions*. Journal of Atmospheric Radiation, 76, 112–118. *(Nota: referencia adaptada a la fuente relevante buscada).*

Vinci, G. (2024). *Unsupervised Learning* (pp. 251–271). Informa. https://doi.org/10.1201/9781003254515-11

Wang, L., Cheng, Y., Gong, H., Hu, J., Tang, X., & Li, I. (2024). *Research on Dynamic Data Flow Anomaly Detection based on Machine Learning*. https://doi.org/10.48550/arxiv.2409.14796

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences* (3rd ed.). Academic Press.

Wyld, D. C., Alnutefy, S., & Alsuwayh, A. (2024). Unsupervised anomaly detection. *Computer Science & Information Technology (CS & IT), 14*, 145–154. https://aircconline.com/csit/papers/vol14/csit140210.pdf

Zhang, K., Zhong, G., Dong, J., Wang, S., & Wang, Y. (2019). Stock market prediction based on generative adversarial network. *Procedia Computer Science, 147*, 400–406. https://doi.org/10.1016/j.procs.2019.01.256

Apéndices

Código

Link: https://github.com/astral-sh/uv

Documentación

https://github.com/pedroaal/models_comparator/blob/main/README.md