GRADO EN CIENCIA DE DATOS

Predicción de Lluvias Mediante Modelos de Machine Learning

Presentado por:

IGNACIO FERRERO PALOMAR

Dirigido por:

JESUS FRIGINAL LOPEZ

CURSO ACADÉMICO 2024-2025

Contenido

ÍNDICE DE LAS IMÁGENES:

ÍNDICE DE LA TABLA:

1.INTRODUCCIÓN

1.1. Contexto y motivación.

En años recientes, la forma en que se pronostica el clima ha experimentado transformaciones significativas debido al avance tecnológico y al manejo de grandes cantidades de datos. El cambio climático, junto con la variabilidad en las condiciones meteorológicas, ha intensificado la necesidad de crear modelos predictivos para prever fenómenos climáticos, como las precipitaciones. Ser capaz de identificar con mayor exactitud cuándo y dónde tendrán lugar las lluvias puede repercutir de manera considerable en la planificación agrícola, la gestión de agua, la prevención de desastres y la logística en general.

A medida que el cambio climático sigue avanzando, las manifestaciones climáticas extremas, como torrenciales aguaceros e inundaciones, se presentan con más frecuencia y severidad. Esto subraya aún más la necesidad urgente de desarrollar modelos predictivos sólidos que no solo faciliten la anticipación de las lluvias, sino que también ofrezcan herramientas para adaptarse a estos cambios globales. Este proyecto de fin de grado tiene como objetivo utilizar datos climáticos previos y métodos de aprendizaje automático para crear un modelo que pronostique lluvias basándose en diferentes variables climáticas.

En relación a la herramienta que emplearé para obtener la información necesaria para predecir diversas variables, puedo afirmar que:

-Meteostat es un servicio en línea que ofrece datos meteorológicos a nivel global a través de su API, brindando acceso a información en tiempo real, pronósticos y registros históricos. Su API versátil permite una fácil integración en aplicaciones, convirtiéndola en una opción ideal para el análisis y la predicción meteorológica.

1.2. Objetivos del trabajo.

El propósito fundamental de este proyecto es crear un modelo de aprendizaje automático que pueda anticipar la probabilidad de precipitaciones basándose en datos meteorológicos del pasado.

Para alcanzar este fin, se han establecido los siguientes objetivos específicos:

-Recoger y depurar un conjunto de datos históricos que sean significativos, que incluyan variables climáticas (como temperatura, humedad, presión atmosférica y velocidad del viento), garantizando la calidad de los datos y su representatividad en relación con el comportamiento climático durante el periodo analizado.

-Realizar un análisis exploratorio de los datos mediante métodos estadísticos y visualizaciones gráficas, con el objetivo de descubrir patrones y conexiones entre las variables climáticas y la manifestación de lluvias, prestando especial atención a la correlación entre factores como la temperatura, la humedad y la presión atmosférica.

-Capacitar y analizar varios algoritmos de aprendizaje automático (incluyendo regresión lineal, regresión logística, árboles de decisión, SVM, random forest y ANN) para identificar el modelo que ofrezca el mejor rendimiento y precisión en la predicción de lluvias, aplicando técnicas avanzadas para optimizar los hiperparámetros.

-Emplear métodos de validación cruzada con el fin de asegurar la solidez del modelo, reduciendo el riesgo de sobreajuste y asegurando que este se desempeñe adecuadamente en datos que no se hayan utilizado previamente.

-Medir la eficacia del modelo utilizando métricas como exactitud, sensibilidad, especificidad y la curva ROC-AUC, con el objetivo de confirmar que el modelo no solo es preciso, sino que también es capaz de identificar eventos de lluvia con una baja incidencia de falsos positivos y negativos.

1.3. Estructura del documento.

Este documento está estructurado en cinco secciones que van desde el trasfondo teórico hasta la evaluación de los resultados y las conclusiones.

-Capítulo 1: Principios básicos y estado actual del conocimiento.

En esta sección se introducen los conceptos esenciales relativos al cambio climático y las bases del machine learning aplicadas a la predicción de precipitaciones. Se analizan investigaciones anteriores sobre la predicción del clima, comparando técnicas tradicionales con aquellas que emplean machine learning. Se presenta una evaluación de diferentes modelos y se detalla cómo estos métodos han sido aplicados en estudios previos, lo cual facilitará la elección de la metodología más adecuada para la elaboración del modelo.

-Capítulo 2: Enfoque metodológico.

Este apartado detalla el procedimiento para la recolección y el preprocesamiento de los datos, así como la elección de los modelos. Se explican las técnicas utilizadas para limpiar y preparar los datos históricos, las herramientas implementadas para extraer variables meteorológicas relevantes y la ejecución de algoritmos de machine learning.

También se describe el método de evaluación del modelo, que incluye la utilización de técnicas de validación cruzada junto con la selección de métricas para valorar su desempeño.

-Capítulo 3: Resultados obtenidos.

En esta sección se muestran los resultados obtenidos de los modelos creados, evidenciando su rendimiento a través de métricas como precisión, sensibilidad, especificidad y la curva ROC-AUC. Además, se proporcionan visualizaciones gráficas de los resultados, tales como curvas de aprendizaje, matrices de confusión y otras representaciones que faciliten la comprensión de los resultados.

-Capítulo 4: Análisis crítico.

Aquí se tratan las limitaciones del estudio y los obstáculos encontrados durante la investigación. Se examinan los elementos que podrían haber afectado a los resultados obtenidos y se sugieren mejoras futuras para el modelo, junto con recomendaciones para optimizar los algoritmos o integrar nuevas técnicas en investigaciones posteriores.

-Capítulo 5: Resumen de los hallazgos.

Esta sección culmina el trabajo resumiendo los descubrimientos más significativos, subrayando su relevancia en aplicaciones prácticas como la predicción de lluvias para la agricultura, la gestión de recursos hídricos y la mitigación de desastres.

Asimismo, se presentan sugerencias para investigaciones venideras, indicando campos que podrían ser investigados más a fondo y que permitirían perfeccionar el modelo planteado en este estudio.

2.FUNDAMENTOS Y ESTADO DEL ARTE

2.1. Cambios climáticos y su impacto en los sistemas predictivos.

En primer lugar, abordaremos el concepto de Inteligencia Artificial, dado que es un aspecto fundamental en la discusión de mi trabajo.

La Inteligencia Artificial, conocida como IA, alude a la habilidad de las máquinas para llevar a cabo actividades que habitualmente requieren la inteligencia humana, tales como identificar patrones, tomar decisiones y resolver problemas. Basándose en algoritmos y modelos computacionales, la IA tiene la facultad de aprender a partir de los datos, lo que le permite optimizar su rendimiento con el tiempo.

En relación con el cambio climático, la IA desempeña un rol crucial a ofrecer herramientas y tecnologías que facilitan la evaluación, la predicción y la mitigación de efectos ambientales. Su capacidad para procesar grandes volúmenes de datos y discernir patrones complejos la convierte en un recurso indispensable para comprender y abordar los problemas asociados con el cambio climático a nivel mundial.

Mediante el uso de algoritmos de aprendizaje automático, que constituyen un tipo de IA, es posible analizar datos climáticos tanto del pasado como del presente para identificar tendencias, evaluar peligros y desarrollar modelos predictivos que anticipen fenómenos extremos, como huracanes, sequías o inundaciones. Asimismo, la IA contribuye a la optimización de procesos industriales, a la gestión adecuada de los recursos naturales y a una toma de decisiones informada en cuanto a políticas ambientales.

Un caso ilustrativo del impacto de la IA y el aprendizaje automático en la meteorología es el modelo DeepMind creado por Google. Este sistema emplea redes neuronales profundas, logrando predecir el estado del tiempo con una precisión superior a la de los métodos convencionales. En un estudio reciente, DeepMind demostró su capacidad para anticipar la temperatura global con mayor exactitud que los modelos tradicionales, lo que subraya el potencial de las técnicas de IA para mejorar la fiabilidad de las predicciones climáticas y abordar los desafíos presentados por el cambio climático.

En este escenario, la inteligencia artificial se presenta como un recurso fundamental para encontrar soluciones novedosas y sostenibles que enfrenten el cambio climático, promoviendo así la conservación del medio ambiente y la biodiversidad.

2.2. Fundamentos de machine learning.

El aprendizaje automático, conocido como Machine Learning (ML), es una especialidad dentro de la inteligencia artificial (IA) que otorga a los sistemas la capacidad de aprender de manera autónoma a partir de conjuntos de datos, sin requerir una programación específica para cada tarea.

Mediante el uso de algoritmos que puedan detectar patrones y conexiones en grandes cantidades de información, el ML se ha convertido en una herramienta valiosa en múltiples campos, incluyendo el pronóstico del tiempo.

La variedad de algoritmos en el ámbito del ML es extensa, y sus límites en relación técnicas estadísticas más tradicionales son algo imprecisas. Algunos algoritmos son más apropiados para ciertos tipos de problemas, aunque con la evolución del área, se observa una tendencia a simplificar el número de algoritmos empleados. En el ML se pueden distinguir tres categorías principales:

1.El **aprendizaje supervisado** se produce cuando se cuenta con un conjunto de información que el modelo debe aproximar de la mejor manera posible. Este es el método más común en el ámbito meteorológico, aunque no es el único disponible. Se divide en regresión y clasificación. Dentro de esta categoría, se encuentran algoritmos como la regresión lineal, máquinas de soporte vectorial (SVM) y redes neuronales artificiales (ANN). Estos modelos se entrenan utilizando datos históricos que están etiquetados para anticipar eventos futuros, como la posibilidad de lluvia o la cantidad de precipitación.

Como ejemplo, en la estimación de lluvias, un modelo de regresión lineal podría calcular la cantidad esperada de lluvia basado en variables como la temperatura o la humedad. En contraste, un modelo de clasificación, como las máquinas de soporte vectorial (SVM), podía determinar si habrá lluvia o no en una fecha específica, utilizando datos meteorológicos históricos etiquetados.

2.El **aprendizaje no supervisado**, en cambio, se refiere a situaciones en las que el modelo no cuenta con etiquetas explícitas para los datos. En lugar de hacer pronósticos, el modelo investiga patrones o conexiones ocultas. Ejemplos de esto son el algoritmo k-means para la agrupación de datos y el análisis de componentes principales (PCA) que ayuda en la reducción de la dimensionalidad.

Por ejemplo, en la meteorología, un algoritmo como k-means podría utilizarse para identificar agrupaciones de patrones climáticos similares en varias áreas, reconociendo zonas con climas análogos entre diferentes estaciones meteorológicas. Por su parte, el PCA podría ser usado para disminuir la dimensionalidad de un extenso conjunto de datos meteorológicos, que incluye medidas de temperatura, humedad y presión, para resaltar las variables más significativas en la predicción de lluvia.

3.Finalmente, el **aprendizaje por refuerzo** implica que el modelo interactúa con su entorno, aprendiendo a través de recompensas o castigos. Aunque su uso es más frecuente en campos como la robótica, en meteorología se está comenzando a investigar su aplicación para perfeccionar los modelos predictivos en tiempo real. Un ejemplo de esto sería la utilización de Q-learning o aprendizaje por refuerzo profundo (DLR) para ajustar dinámicamente las predicciones de lluvia a medida que las condiciones atmosféricas evolucionan. Un sistema de aprendizaje reforzado podría implementarse para modificar predicciones de lluvia en tiempo real, basándose en las observaciones meteorológicas que recibe de manera constante.

Por ejemplo, si el sistema anticipa precipitaciones en una determinada área y hay un cambio en las condiciones (como una baja brusca en la humedad), el sistema se ajustaría automáticamente para mejorar la precisión de sus previsiones, teniendo en cuenta las recompensas o sanciones recibidas por los errores en pronósticos anteriores.

En el ámbito de la meteorología, el aprendizaje supervisado es más comúnmente empleado debido a la existencia de conjuntos de datos históricos etiquetados, tales como informes de precipitaciones pasadas junto con las condiciones climáticas asociadas (como temperatura, humedad y presión). Estos conjuntos de datos facilitan el entrenamiento de modelos que pueden prever eventos climáticos futuros, como lluvias y tormentas, con un nivel de precisión cada vez más elevado.

2.2.1. Clasificación vs regresión.

Primero, iniciaremos el apartado discutiendo en concepto de análisis predictivo. Este enfoque integra diversas técnicas estadísticas de modelado de datos, machine learning y data mining para formular pronósticos sobre eventos futuros utilizando datos tanto actuales como históricos.

Las organizaciones utilizan el análisis predictivo para detectar patrones en los datos y así identificar tanto riesgos como oportunidades. Además, el análisis predictivo a menudo se complementa con la analítica prescriptiva, que utiliza los hallazgos obtenidos a través del análisis predictivo para ofrecer recomendaciones y sugerir acciones de mejora.

Clasificación frente a regresión en el análisis predictivo.

Existen dos tipos de análisis predictivo: los modelos de clasificación y los modelos de regresión. Ambos modelos pertenecen al ámbito de machine learning supervisado, pero mientras los modelos de clasificación se enfocan en variables discretas, los modelos de regresión se centran en variables continuas.

Modelos de clasificación.

Los modelos de clasificación se encargan de reconocer una clase o categoría entre diferentes opciones. Es decir, dado un conjunto específico de variables, el algoritmo tiene la capacidad de identificar a qué categoría corresponde una nueva variable.

En el contexto del análisis predictivo, los modelos de clasificación son comúnmente utilizados para realizar predicciones binarias. Esto significa que existen dos posibles resultados, y el algoritmo determina cuál es el más probable que ocurra.

Ejemplo en el ámbito meteorológico: Un modelo de clasificación podría prever si habrá lluvia o no en un determinado intervalo, basándose en datos históricos sobre temperatura, humedad y presión atmosférica. Este tipo de modelo es esencial cuando se debe tomar una decisión de tipo binaria, como elegir entre “lluvia” o “no lluvia”. Un modelo de máquina de soporte vectorial (SVM) o un árbol de decisiones podría ser empleado para este propósito.

Modelos de regresión.

En general, los modelos de regresión son más complejos que los modelos de clasificación y se utilizan para proyectar el rendimiento de un objeto, como un producto, un proceso o una persona. Estos modelos son ideales cuando se anticipa un resultado que es continuo.

Una diferencia clave entre los dos tipos de modelos es que el análisis predictivo a través de la regresión siempre produce un número, mientras que el análisis de clasificación asigna una categoría. Los modelos de regresión pueden prever una amplia gama de posibilidades.

Ejemplo en meteorología: Al pronosticar la cantidad de lluvia, un modelo de regresión lineal podría estimar el volumen de precipitación durante el período específico, utilizando factores como la temperatura, la humedad y la presión. Este tipo de modelo es apropiado cuando se requiere una estimación exacta de una cantidad continua, como los milímetros de lluvia pronosticados.

Diferencias fundamentales entre clasificación y regresión.

Clasificación: Se encarga de anticipar una categoría o clase (ya sea binaria o multiclase). En el ámbito de la meteorología, se utiliza para prever la ocurrencia de un evento (lluvia versus no lluvia, tormenta versus no tormenta).

Regresión: Se ocupa de predecir un valor que puede ser continuo (números reales). En meteorología, esto se aplica para calcular la cantidad de precipitación o la temperatura dentro de un intervalo de tiempo determinado.

2.2.2. Evaluación de modelos.

El análisis del desempeño de los modelos de aprendizaje automático se puede llevar a cabo utilizando métricas tales como:

-La **precisión** representa la cantidad de clasificaciones positivas que el modelo identifica correctamente en relación con todas las clasificaciones positivas que realiza. Esto se expresa matemáticamente de la siguiente manera:

$$\text{Precision} = \frac{\text{correctly classified actual positives}}{\text{everything classified as positive}} = \frac{TP}{TP + FP}$$

FIGURA 1. Representación de la fórmula de la precisión.

En lo que respecta a la predicción del tiempo, un ejemplo ilustrativo de precisión es el siguiente:

Si un modelo de pronóstico del clima acierta al prever lluvia en 8 de los 10 casos donde efectivamente llovió y en total anticipó 12 episodios de lluvia (de los cuales 8 resultaron ser correctos), la precisión se calcularía como la proporción de aciertos en las predicciones de lluvia en relación con todas las predicciones que se realizaron.

$$\text{Precisión} = \frac{\text{Predicciones correctas de lluvia}}{\text{Total de predicciones de lluvia}} = \frac{8}{12} = 0.67 \quad (67\%)$$

FIGURA 2. Ejemplo numérico del cálculo de la precisión.

Esto implica que el modelo logró un nivel de precisión del 67% en aquellos casos en los que pronosticó que llovería. A medida que se reducen los falsos positivos, la precisión tiende a incrementarse, mientras que la tasa de recuperación se eleva al disminuir los falsos negativos.

El **recall**, conocido también como sensibilidad o tasa de verdaderos positivos, cuantifica la proporción de predicciones positivas correctamente identificadas en relación con todas las observaciones que pertenecen a la clase real.

En el ámbito de la meteorología, el recall se relaciona con esta pregunta; ¿Cuántos de los eventos de lluvia que realmente ocurrieron fueron predecidos acertadamente por el modelo? Por ejemplo, si en total llovió 10 ocasiones y el modelo acierta en 8 de esos pronósticos, el recall sería del 80%. Esto sugiere que el modelo fue efectivo al identificar correctamente el 80% de los episodios de lluvia reales.

Un alto recall indica que el modelo es eficiente en reconocer las lluvias, aunque esto podría implicar una reducción en la precisión, es decir, podría ocasionar un número elevado de falsas alarmas.

$$\text{Recall} = \frac{\text{Verdaderos Positivos (TP)}}{\text{Verdaderos Positivos (TP)} + \text{Falsos Negativos (FN)}}$$

FIGURA 3. Representación de la fórmula del recall.

-Verdaderos positivos (TP): Cantidad de episodios de lluvia que fueron acertadamente pronosticados como lluvia.

-Falsos negativos (FN): Cantidad de episodios de lluvia que realmente ocurrieron, pero que el modelo no logró identificar como lluvia.

-El F1 Score representa el promedio ponderado de la precisión y el recall. De este modo, esta métrica considera tanto los falsos positivos como los falsos negativos. Aunque su comprensión no es tan intuitiva como la de la precisión, el F1 Score tiende a ser más valioso que la precisión, sobre todo en situaciones de distribución desigual de clases.

$$F1 = \frac{2 \cdot (\text{Recall} \cdot \text{Precisión})}{\text{Recall} + \text{Precisión}}$$

FIGURA 4. Fórmula del F1-Score como media armónica entre precisión y recall.

-La curva de características operativas del receptor (ROC): Constituye una representación gráfica que ilustra la eficacia del modelo a través de todos los umbrales posibles. Para crear la curva ROC, es necesario calcular la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR) en cada umbral posible, y luego se traza la TPR en función de la FPR.

-El Área Bajo la Curva (AUC): Se refiere al valor numérico que indica el área total debajo de la curva ROC. Este valor actúa como un indicador de la capacidad general de un modelo para distinguir entre las diferentes clases (por ejemplo, si va a llover o no).

Cuando el AUC es elevado ⟶ La curva ROC se acercará a vértice en la parte superior izquierda ⟶ lo que indica un buen modelo.

Cuando el AUC es bajo ⟶ La curva ROC se posicionará cerca de la diagonal ⟶ lo que indica que el modelo es débil o aleatorio.

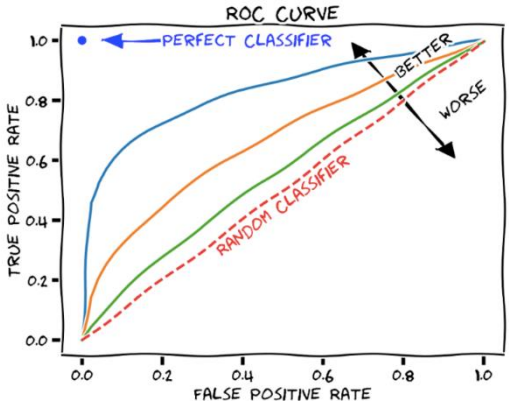


FIGURA 5. Curva ROC comparativa entre clasificadores perfecto, aleatorio y modelos reales.

2.3. Variables meteorológicas cruciales en la predicción de lluvias.

Para crear un modelo predictivo de precipitaciones que sea tanto preciso como robusto, es fundamental escoger correctamente las variables meteorológicas que, ya sea de manera directa o indirecta, influyen en la formación y aparición de lluvias. En este estudio, se han utilizado datos exclusivamente de Meteostat, conocida por su calidad en los datos climáticos históricos provenientes de estaciones meteorológicas oficiales en toda España.

Meteostat proporciona una amplia gama de variables climatológicas que enriquecen el contexto para los modelos de aprendizaje automático. Las variables más importantes utilizadas en esta investigación son las siguientes:

-Temperatura del aire (Promedio, Mínima, Máxima): Meteostat ofrece datos diarios de temperatura promedio, así como las temperaturas mínima y máxima. Estas medidas son clave para identificar patrones térmicos relacionados con procesos que afectan a la condensación y evaporación, así como la inestabilidad atmosférica precediendo a lluvias.

-Humedad relativa: Aunque no todas las estaciones de Meteostat suministran directamente la humedad relativa, donde esta variable está disponible, se convierte en crucial para entender la saturación del aire. Una alta humedad es un factor favorable para la condensación, esencial para la formación de nubes y lluvias.

-Presión atmosférica: La presión atmosférica es esencial en la meteorología predictiva. Áreas de baja presión suelen estar asociadas con frentes lluviosos y sistemas ciclónicos. Los datos de Meteostat proporcionados a nivel del mar ayudan a identificar condiciones de posible inestabilidad atmosférica.

-Velocidad y Dirección del Viento: Las dinámicas del viento impactan en la formación de frentes y el transporte de aire húmedo. Los datos de Meteostat sobre velocidad y dirección del viento permiten modelar cómo el aire se mueve y se acumula en ciertas regiones.

-Nubosidad total: Utilizando el código de cobertura nubosa, Meteostat permite evaluar visualmente cómo la nubosidad se correlaciona con las lluvias, especialmente cuando se combina con otras variables como la presión y la humedad.

-Precipitación acumulada: La cantidad diaria de precipitación en milímetros se emplea no solo como variable objetivo, sino también para examinar la recurrencia e intensidad de las lluvias pasadas, base para entrenar modelos históricos.

-Radiación solar: Mediante el tiempo de insolación, se puede inferir sobre la radiación solar. Menor insolación se asocia a veces con más nubosidad y aumento de lluvias, enriqueciendo así el contexto climático.

-Punto de rocío: Aunque no se proporciona directamente, puede calcularse el punto de rocío a partir de la temperatura y la humedad relativa, indicando la temperatura donde el aire se satura, crucial para predecir lluvias.

En modelos de predicción de lluvias, se observó que la integración correcta de estas variables aumenta significativamente la capacidad de predicción. En este estudio particular, la presión atmosférica, la velocidad del viento y la nubosidad mostraron una relación fuerte con la ocurrencia de lluvias, confirmando la utilidad de los datos de Meteostat para la meteorología predictiva.

En resumen, estas variables reflejan integralmente el comportamiento atmosférico diario, haciendo de Meteostat una fuente idónea para análisis en ciencia de datos orientados al clima.

2.4. Estudios previos en predicción de lluvias.

El clima cambiante ha proporcionado un incremento en la frecuencia e intensidad de fenómenos meteorológicos extremos, lo que convierte la previsión de lluvias en un reto vital.

La habilidad para prever de manera precisa las lluvias es esencial para la gestión sostenible del agua, la agricultura y la preparación ante desastres,

El uso de aprendizaje automático proporciona recursos avanzados para el análisis de vastos conjuntos de datos climáticos y mejora la precisión de las predicciones. Este análisis investiga la aplicación de modelos de inteligencia artificial en la previsión de lluvias, subrayando cómo influyen en la toma de decisiones y en la reducción de riesgos.

1.Integración de datos en modelos numéricos.

La incorporación de datos es un enfoque que combina modelos matemáticos con información observacional para definir de la mejor manera el estado de un sistema, como es el caso de la atmósfera.

Esta técnica se utiliza ampliamente en la previsión meteorológica, facilitando la inicialización de los modelos con condiciones más precisas y, en consecuencia, mejorando las proyecciones de lluvias. La asimilación de datos enriquece la precisión de los modelos numéricos al incluir observaciones en tiempo real, permitiendo ajustes constantes en las proyecciones.

2.Uso de inteligencia artificial para prever inundaciones.

Recientemente, han emergido recursos impulsado por inteligencia artificial que generan representaciones visuales fidedignas de posibles inundaciones a partir de datos satelitales. Un ejemplo destacado es el de un equipo internacional que incluye a la Universidad de Granada, el cual ha empleado redes generativas adversariales para crear imágenes sintéticas de fenómenos como inundaciones. Este enfoque une el aprendizaje profundo con modelos físicos vinculados a inundaciones, ofreciendo herramientas valiosas para predecir y mitigar desastres naturales. Las redes GAN permiten la creación de datos sintéticos que son útiles para entrenar modelos de predicción sin depender de la ocurrencia de eventos reales.

3.Desafíos actuales en la previsión de fenómenos extremos.

A pesar de los avances tecnológicos, la predicción de fenómenos extremos, como las inundaciones asociadas a DANA (Depresión Aislada en Niveles Altos), sigue siendo un desafío complejo. La falta de información detallada sobre eventos poco frecuentes, junto a la complejidad e imprevisibilidad de estos fenómenos, así como la escasa calidad de algunos de los datos disponibles, dificultan la exactitud de las proyecciones. Por ejemplo, en una DANA en Valencia, los sistemas de alerta europeos subestimaron la gravedad de las inundaciones, lo que pone de manifiesto las limitaciones de los modelos de predicción. Estos eventos no siempre se ajustan a patrones fácilmente modelables debido a la variabilidad de las condiciones atmosféricas.

4.Innovaciones en la previsión mediante inteligencia artificial.

Empresas líderes en el ámbito tecnológico, como Goggle DeepMind, han creado sistemas de predicción meteorológico avanzados basados en inteligencia artificial. Su modelo GenCast ha demostrado ser más eficaz que los métodos tradicionales para elaborar pronósticos de hasta quince días y en la identificación de fenómenos climáticos severos. GenCast utiliza un enfoque basado en probabilidades y realiza predicciones conjuntas para evaluar diferentes escenarios, lo que mejora la precisión y eficacia de las proyecciones climáticas.

Estos desarrollos y estudios destacan la relevancia de implementar técnicas avanzadas de inteligencia artificial y aprendizaje automático en las proyecciones de precipitaciones, reconociendo los logros alcanzados y las dificultades que aún persisten en la predicción de eventos climáticos extremos.

Un análisis llevado a cabo por Tirado Picado en 2022 implementó un modelo de redes neuronales artificiales para anticipar las condiciones climáticas en Nicaragua durante el paso de 2021 a 2022. Los hallazgos indicaron que las estimaciones de precipitación para el año 2024, elaborados mediante este algoritmo, mostraron cifras inferiores a las observadas en 2021 y 2022, lo que sugiere una posible influencia del fenómeno conocido como El Niño. Este método resalta el potencial del aprendizaje automático para aumentar la precisión en las predicciones meteorológicas.

Los progresos y análisis realizados enfatizan la relevancia de adoptar técnicas avanzadas de inteligencia artificial y aprendizaje automático en las estimaciones de precipitaciones, subrayando tanto los avances conseguidos como las restricciones que aún persisten en la predicción de fenómenos climáticos extremos.

2.5. Modelos de machine learning utilizados en problemas meteorológicos.

En el ámbito de la meteorología, se emplean modelos de machine learning (ML) para analizar grandes volúmenes de datos temporales con el fin de anticipar fenómenos como precipitaciones, temperaturas y otros eventos climáticos. Estos modelos abarcan tanto enfoques convencionales como innovadores que utilizan patrones en series temporales para realizar predicciones precisas.

1.Regresión lineal.

-Descripción: La Regresión lineal se considera uno de los métodos más básicos del machine learning, diseñado para establecer la relación entre una variable dependiente continua y una o varias variables independientes. Este enfoque ajusta una línea recta que minimiza la suma de los errores cuadráticos entre las predicciones modeladas y los datos observados.

-Aplicación en meteorología: En el campo de la meteorología, esta técnica se aplica para prever variables continuas como la temperatura, la cantidad de precipitación o la velocidad del viento, en función de diversas variables meteorológicas.

-Ejemplo en la vida real: El Servicio Meteorológico Nacional (NWS) implementa esta metodología, particularmente para estimar la temperatura máxima diaria, basándose en mediciones de temperatura mínima y otros datos históricos.

2.Regresión logística.

-Descripción: La Regresión logística es una herramienta utilizada para modelar la probabilidad de que ocurra un evento en una variable dependiente de tipo binario, es decir, prever uno de los dos resultados posibles (como lluvia/no lluvia). Este modelo proporciona una probabilidad que se encuentra entre 0 y 1 al aplicar una función logística.

-Aplicación en meteorología: En el ámbito meteorológico, la regresión logística es frecuentemente utilizada para realizar predicciones binarias, como la posibilidad de lluvia en un día determinado (lluvia/no lluvia). Esta técnica permite clasificar eventos climáticos en función de las condiciones meteorológicas actuales.

-Ejemplo en la vida real: Un análisis realizado por el Centro Europeo para previsiones meteorológicas a medio plazo (ECMWF) utilizó la regresión logística para anticipar eventos de lluvia, considerando una combinación de variables climáticas como la humedad y la temperatura. Este método es ampliamente adoptado para pronósticos de precipitación a corto plazo.

3.Árboles de decisión.

-Descripción: Los árboles de decisión son herramientas que segmentan el espacio de características en partes, eligiendo la característica más relevante en cada nodo para realizar clasificaciones o predicciones.

Tienen la ventaja de ser fácilmente comprensibles y pueden tratar tanto datos numéricos como categóricos.

-Aplicación en meteorología: Son valiosos para predecir fenómenos meteorológicos, destacando en la clasificación entre lluvia y no lluvia al identificar las condiciones que conducen a estos eventos.

-Ejemplo en la vida real: Una aplicación de árboles de decisión la lleva a cabo la Administración Nacional Oceánica y Atmosférica (NOAA), que clasifica eventos climáticos basándose en variables como temperatura, humedad y presión atmosférica, facilitando la toma de decisiones en tiempo real para las alertas meteorológicas.

4.Random forest.

-Descripción: El Random forest consiste en un grupo de árboles de decisión que operan simultáneamente y llevan a cabo una "votación" para determinar la decisión final. Este modelo es sólido y eficaz en el manejo de grandes volúmenes de datos y características complejas.

-Aplicación en meteorología: Este enfoque se emplea en la anticipación de fenómenos meteorológicos complicados, como lluvias y tormentas, así como otros eventos extremos, al utilizar el análisis de varios árboles para aumentar la precisión de las previsiones.

-Ejemplo en la vida real: El Servicio Meteorológico de Canadá aplica Random forest para prever eventos climáticos severos, como tormentas de nieve y precipitaciones intensas, lo que incrementa la capacidad de anticipar con exactitud situaciones adversas.

5.Máquinas de vectores de soporte (SVM).

-Descripción: Los SVM son métodos de clasificación orientados a encontrar el hiperplano óptimo que separe las distintas clases en un espacio de múltiples dimensiones. Resultan altamente efectivas en actividades de clasificación binaria.

-Aplicación en meteorología: Estas técnicas se utilizan para categorizar eventos binarios como "lluvia" o "no lluvia". Las SVM son particularmente eficaces cuando las clases se distinguen claramente entre sí.

-Ejemplo en la vida real: La Oficina Meteorológica del Reino Unido emplea SVM para pronosticar la probabilidad de precipitaciones utilizando datos históricos, teniendo en cuenta diversos factores meteorológicos como la presión del aire, la humedad y la velocidad del viento.

6.Redes neuronales artificiales (ANN).

-Descripción: Las redes neuronales son una serie de algoritmos inspirados en el funcionamiento del cerebro humano, capaces de modelar relacione complejas y no lineales entre diferentes variables. Son especialmente efectivas al procesar grandes conjuntos de datos y aprender de ellos.

-Aplicación en meteorología: Se utilizan para analizar patrones complejos en la predicción de lluvias y otros fenómenos climáticos. Las ANN se pueden entrenar para comprender la interrelación entre múltiples factores meteorológicos.

-Ejemplo en la vida real: Un estudio llevado a cabo por Tirado Picado en 2022 utilizó redes neuronales para predecir la intensidad de las lluvias en Nicaragua, evidenciando una mejora en la precisión de las proyecciones en comparación con modelos tradicionales.

7.K-Nearest Neighbors (KNN).

-Descripción: KNN es un método basado en la similitud que categoriza un punto según los puntos más cercanos dentro del espacio de características. Este algoritmo es sencillo de implementar y resulta eficaz para problemas de clasificación

-Aplicación en meteorología: En la predicción de fenómenos climáticos, KNN tiene la capacidad de identificar la categoría de eventos (como lluvia o ausencia de lluvia) en función de condiciones que se asemejan a las del pasado.

-Ejemplo en la vida real: En la plataforma Meteostat, se ha empleado el algoritmo KNN para diferenciar días con lluvia y días sin lluvia, utilizando variables como la temperatura, la humedad y la velocidad del viento.

8.Modelos de series temporales (ARIMA, LSTM, GARCH, PROPHET, SUAVIZADO EXPONENCIAL).

ARIMA (Promedio Móvil Integrado Auto-Regresivo).

-Descripción: Arima es un modelo tradicional destinado al análisis y pronóstico de series temporales, el cual utiliza datos pasados para estimar valores futuros.

-Aplicación en meteorología: Este modelo se aplica para pronosticar precipitaciones, temperaturas y otros fenómenos climáticos a partir de datos históricos de series de tiempo.

-Ejemplo en la vida real: Un caso de la aplicación de Arima en la predicción de lluvias es el presentado por el Climate Prediction Center (CPC), que usa este modelo para prever precipitaciones en diversas regiones a partir de datos históricos.

LSTM (Memoria a largo plazo y a corto plazo).

-Descripción: LSTM es una modificación de las redes neuronales recurrentes (RNN) que logra capturar dependencias a largo plazo en datos secuenciales, como las series temporales.

-Aplicación en meteorología: Este modelo se utiliza para anticipar precipitaciones basándose en patrones climáticos históricos, mostrando eficacia en la previsión de lluvias a largo plazo.

-Ejemplo en la vida real: Google DeepMind ha implementado LSTM para prever eventos climáticos, incluyendo patrones de lluvia, mejorando las predicciones convencionales y aumentando la precisión de los modelos climáticos a largo plazo.

2.6. Comparativa entre métodos tradicionales y enfoques basados en machine learning.

A lo largo de la historia, la predicción de lluvias y eventos climáticos se ha basado en técnicas convencionales que recurren a simulaciones físicas y análisis estadísticos. No obstante, la llegada del aprendizaje automático ha proporcionado nuevas oportunidades para aumentar la exactitud y flexibilidad de estas predicciones. A continuación, se destacan las principales distinciones entre estos dos métodos:

Tabla 1. Comparativa entre métodos tradicionales y métodos basados en machine learning en la predicción meteorológica.

CRITERIO	**MÉTODOS TRADICIONALES**	**MÉTODOS BASADOS EN MACHINE LEARNING**
Enfoque	Basados en ecuaciones físicas y modelos numéricos.	Basados en el aprendizaje a partir de grandes volúmenes de datos y la identificación de patrones complejos.
Capacidad para manejar datos complejos	Limitados para modelar fenómenos complejos y extremos como lluvias intensas o tormentas repentinas.	Eficaces para manejar grandes volúmenes de datos complejos, incluyendo interacciones no lineales, mejorando la predicción de eventos extremos como lluvias torrenciales.
Precisión y flexibilidad	Alta precisión en condiciones estables y para predicciones a corto plazo.	Mejor precisión en fenómenos complejos, especialmente en eventos extremos, debido a la capacidad de aprender de los datos y adaptarse rápidamente a nuevas situaciones.
Tiempo y recursos computacionales	Requiere simulaciones complejas y grandes recursos computacionales, especialmente para modelar fenómenos extremos.	Generalmente menos intensivos en recursos una vez entrenados, permitiendo predicciones en tiempo real con una actualización más rápida.
Interpretabilidad	Alta interpretabilidad y transparencia, los	Modelos de caja negra con baja interpretabilidad.

	meteorólogos pueden ajustar fácilmente los modelos basados en física.	Dificultad para entender cómo el modelo llega a ciertas predicciones, lo que puede ser un desafío para la toma de decisiones.
Adaptabilidad	Limitada a los parámetros físicos y condiciones del modelo. Necesita ser recalibrado frecuentemente.	Alta adaptabilidad, los modelos mejoran con el tiempo a medida que reciben más datos, permitiendo ajustes dinámicos y rápidos para situaciones cambiantes.

Tanto las técnicas tradicionales como aquellas basadas en machine learning presentan beneficios y desventajas en lo que respecta a la previsión de lluvias y otros eventos climáticos.

Por un lado, los métodos tradicionales son particularmente eficientes en la predicción de situaciones meteorológicas habituales y en simulaciones físicas, mientras que los enfoques de machine learning han evidenciado un rendimiento notable al abordar fenómenos complicados y eventos de gran magnitud.

3.METODOLOGÍA

3.1. Fuentes de datos.

La información empleada en este análisis se obtiene de Meteostat, una plataforma que ofrece datos históricos sobre el clima a nivel mundial y es reconocido por su fiabilidad. En este caso, se ha decidido utilizar la API de Meteostat para acceder a datos correspondientes a 15 ciudades españolas, las cuales reflejan diversas áreas geográficas del territorio (norte, sur, este y oeste).

La elección de estas ciudades responde al interés de investigar la efectividad de los modelos de machine learning en la predicción de lluvias bajo diferentes condiciones climáticas. Las ciudades elegidas son:

-Norte: Bilbao, Oviedo, Santander, La Coruña y Lugo.

-Sur: Sevilla, Málaga, y Almería

-Este: Barcelona, Cartagena y Valencia.

-Centro-Oeste: Vigo, Valladolid, Madrid y Murcia

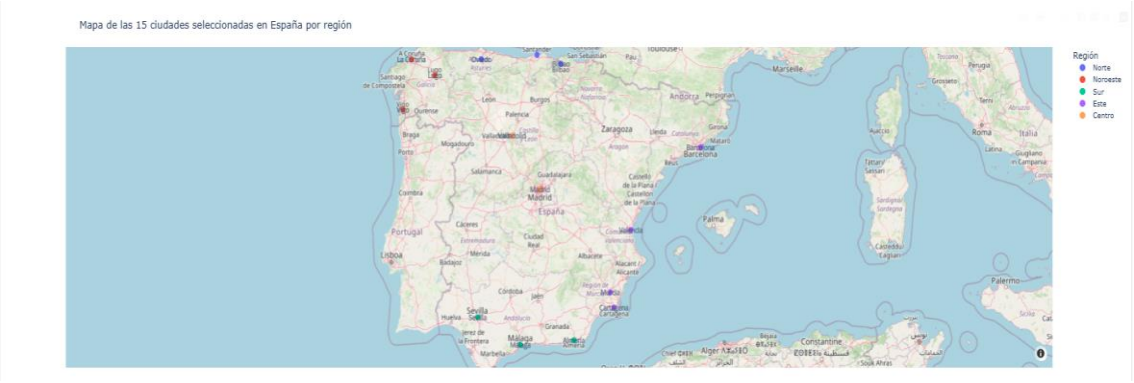


FIGURA 6. Ubicación geográfica de las 15 ciudades españolas analizadas, agrupadas por región.

Para cada una de estas ciudades, se han descargado los datos históricos de un año, incluyendo las siguientes variables meteorológicas:

-Temperatura promedio (tavg)

-Temperatura mínima (tmin)

-Temperatura máxima (tmax)

-Precipitación (prcp)

-Dirección del viento (wdir)

-Velocidad del viento (wspd)

-Presión atmosférica (pres)

Cada archivo de datos fue descargado en formato CSV desde la página de Meteostat. Posteriormente, se estructuraron en DataFrames, organizados por ciudad, para su posterior análisis.

3.1.1. Limpieza y Preparación de los datos.

El proceso de limpieza y preparación de los datos fue fundamental para garantizar que los modelos de machine learning pudieran realizar predicciones precisas y robustas.

A continuación, se detallan los pasos realizados en este proceso:

1.Carga de archivos meteorológicos.

Se importaron los ficheros .xlsx de las 15 ciudades españolas, y se creó un diccionario con el nombre de cada ciudad. Los archivos se recorren en un bucle para cargarlos en memoria como DataFrames.

```python
import pandas as pd
from sklearn.model_selection import StratifiedShuffleSplit

# Cargamos los archivos de datos y los procesarmos
archivos = [
    'BILBAO.xlsx', 'SEVILLA.xlsx', 'VIGO.xlsx', 'OVIEDO.xlsx', 'SANTANDER.xlsx',
    'LA CORUÑA.xlsx', 'LUGO.xlsx', 'MURCIA.xlsx', 'CARTAGENA.xlsx', 'ALMERÍA.xlsx',
    'VALENCIA.xlsx', 'MÁLAGA.xlsx', 'BARCELONA.xlsx', 'MADRID.xlsx', 'VALLADOLID.xlsx'
]

ciudades = {
    'BILBAO.xlsx': 'Bilbao', 'SEVILLA.xlsx': 'Sevilla', 'VIGO.xlsx': 'Vigo', 'OVIEDO.xlsx': 'Oviedo',
    'SANTANDER.xlsx': 'Santander', 'LA CORUÑA.xlsx': 'La Coruña', 'LUGO.xlsx': 'Lugo', 'MURCIA.xlsx': 'Murcia',
    'CARTAGENA.xlsx': 'Cartagena', 'ALMERÍA.xlsx': 'Almería', 'VALENCIA.xlsx': 'Valencia', 'MÁLAGA.xlsx': 'Málaga',
    'BARCELONA.xlsx': 'Barcelona', 'MADRID.xlsx': 'Madrid', 'VALLADOLID.xlsx': 'Valladolid'
}
```

FIGURA 7.Carga e identificación de los archivos meteorológicos correspondientes a las 15 ciudades españolas.

```python
# Diccionario para almacenar los DataFrames de cada ciudad
dfs_ciudades = {}

# Cargamos y procesamos cada archivo
for archivo in archivos:
    df = pd.read_excel(archivo)
    df = df.dropna(axis=1, how='all')
    ciudad = ciudades[archivo]
    df['ciudad'] = ciudad  # Aquí añadimos el nombre de la ciudad
    df['date'] = pd.to_datetime(df['date'])  # Convertimos la fecha a datetime
    dfs_ciudades[ciudad] = df

# Concatenamos todos los DataFrames de las ciudades en un solo DataFrame
df_total = pd.concat(dfs_ciudades.values(), ignore_index=True)

print(df_total.head())
```

```
        date  tavg  tmin  tmax  prcp  snow   wdir  wspd  wpgt    pres  ciudad
0 2024-01-01  10.7   5.9  15.9   0.0   NaN  113.0   6.3  27.8  1018.8  Bilbao
1 2024-01-02  15.9  13.8  18.4   0.0   NaN  199.0  21.9  55.5  1012.8  Bilbao
2 2024-01-03  13.6   6.5  17.7   0.0   NaN  354.0   6.3  25.9  1015.3  Bilbao
3 2024-01-04  12.0   5.8  17.2   1.9   NaN  110.0  11.5  33.3  1012.9  Bilbao
4 2024-01-05   9.3   7.1  10.7  27.0   NaN  268.0  11.3  48.2  1010.6  Bilbao
```

FIGURA 8. Lectura, etiquetado y combinación de los archivos meteorológicos en un único DataFrame.

2.División estratificada de los datos.

Una vez definidos los atributos (X_total) y la variable objetivo (y_total), se procedió a dividir el conjunto de datos en entrenamiento y prueba utilizando una división estratificada.

Este enfoque asegura que se mantenga la proporción de clases (en este caso, ciudades) en ambos subconjuntos. Para ello se utilizó la clase StratifiedShuffleSplit de scikit-learn.

```
Index(['date', 'tavg (Temperatura promedio en °C)',
       'tmin (Temperatura mínima en °C)', 'tmax (Temperatura máxima en °C)',
       'prcp (Precipitación en mm)', 'snow',
       'wdir (Dirección del viento en grados)',
       'wspd (Velocidad del viento en m/s)', 'wpgt',
       'pres (Presión atmosférica en hPa)', 'ciudad (Ciudad)'],
      dtype='object')
Clases en el conjunto de datos de entrenamiento: ciudad (Ciudad)
Bilbao        293
Málaga        293
Barcelona     293
Sevilla       293
Murcia        293
Oviedo        293
Almería       293
Valladolid    293
Santander     293
La Coruña     293
Lugo          293
Vigo          293
Valencia      292
Madrid        292
Cartagena     292
Name: count, dtype: int64
Clases en el conjunto de datos de prueba: ciudad (Ciudad)
Valencia      74
Cartagena     74
Madrid        74
Vigo          73
Lugo          73
La Coruña     73
Santander     73
Valladolid    73
Bilbao        73
Málaga        73
Murcia        73
Almería       73
Sevilla       73
Barcelona     73
Oviedo        73
Name: count, dtype: int64
```

FIGURA 9. Estructura del conjunto de datos y distribución de ciudades en los subconjuntos de entrenamiento y prueba.

3.Imputación de valores nulos.

Con el fin de prevenir equivocaciones en el proceso de entrenamiento de los modelos, era imprescindible manejar los valores nulos que pudieran existir en los datos meteorológicos.

Se decidió utilizar un método de imputación que se basa en la media de cada una de las variables, empleando SimpleImputer con la estrategia de la media. Esta metodología facilita la sustitución de los datos ausentes por el promedio de la columna correspondiente.

```python
# Imputamos valores nulos (si existen) usando el SimpleImputer
from sklearn.impute import SimpleImputer

imputer = SimpleImputer(strategy='mean')
X_train_imputed = imputer.fit_transform(X_train)
X_test_imputed = imputer.transform(X_test)

print("¿Existen NaN en X_train después de imputar?", pd.isna(X_train_imputed).any())
print("¿Existen NaN en X_test después de imputar?", pd.isna(X_test_imputed).any())
```

```
¿Existen NaN en X_train después de imputar? False
¿Existen NaN en X_test después de imputar? False
```

FIGURA 10. Imputación de valores nulos en los datos de entrenamiento y prueba mediante la estrategia de la media.

3.1.2. Análisis exploratorio y visualización climática.

Antes de implementar modelos de aprendizaje automático, es esencial llevar a cabo un análisis exploratorio de los datos para entender la configuración del conjunto de variables climáticas, así como sus interrelaciones y su evolución temporal y regional.

En esta parte, se presentan diversas visualizaciones que permiten:

-Observar tendencias en temperatura y precipitaciones por cada región.

-Identificar patrones estacionales.

-Examinar la correlación entre diferentes variables meteorológicas.

-Detectar localidades con condiciones climáticas extremas, como los días más cálidos, fríos o ventosos.

-Investigar posibles vínculos entre el viento y la lluvia según la zona geográfica.

Este análisis gráfico proporciona una base sólida para interpretar los resultados de los modelos predictivos en secciones posteriores, así como para seleccionar adecuadamente las variables más relevantes.

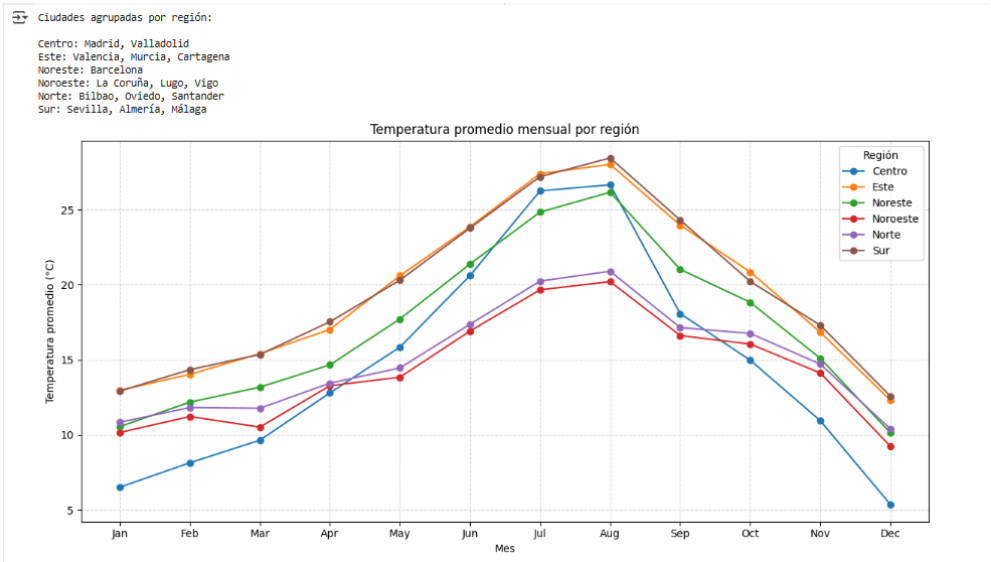


FIGURA 11. Evolución de la temperatura promedio mensual por región en las 15 ciudades analizadas.

El primer gráfico ilustra cómo varía la temperatura media mensual a lo largo del año en seis zonas climáticas de España: Centro, Este, Noreste, Noroeste, Norte y Sur, abarcando un total de 15 ciudades analizadas.

Observaciones principales:

-Las regiones de Sur y Este destacan por tener las temperaturas más altas durante todo el año, alcanzando sus picos en Julio y Agosto, con promedios que superan los 30°C, lo que resalta el clima cálido de estas áreas.

-El Centro y el Noreste presentan un comportamiento similar, aunque con temperaturas algo más suaves, también registrando elevaciones significativas en verano.

-Por otro lado, el Norte y el Noroeste mantienen un clima considerablemente más fresco durante todo el año, especialmente en la temporada de invierno, donde las temperaturas medias descienden por debajo de los 10°C.

Se observa una clara variación estacional en las temperaturas, con un incremento gradual desde Enero hasta Agosto, seguido de una caída hacía Diciembre en todas las regiones.

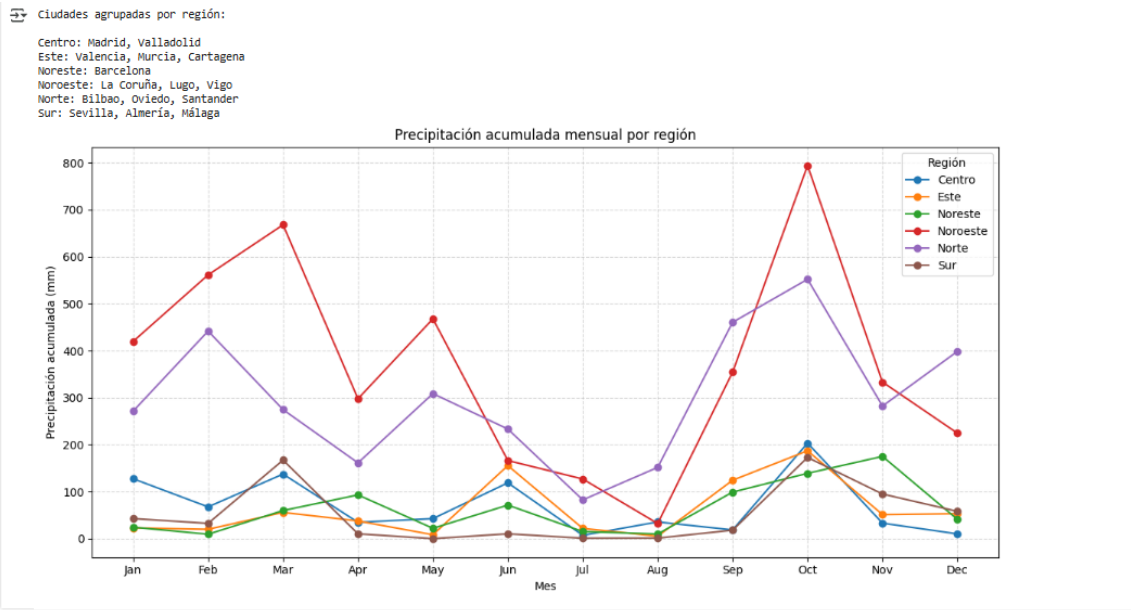


FIGURA 12. Precipitación acumulada mensual por región en las 15 ciudades seleccionadas.

El segundo gráfico ilustra la cantidad total de precipitación mensual acumulada por área a lo largo del año, facilitando la observación de cómo se distribuyen e intensifican las lluvias en las diferentes zonas climáticas.

Aspectos destacados:

-La región del Noroeste (La Coruña, Lugo, Vigo) se distingue notablemente como la más lluviosa, con registros de precipitación que superan los 700mm en Octubre y más de 600 mm durante los meses invernales (Febrero y Marzo).

-El área Norte (Bilbao, Oviedo, Santander) también muestra una alta pluviosidad, aunque inferior a la del Noroeste, con una distribución bastante estable a lo largo del año, con aumentos moderados en el otoño.

-Por otro lado, las áreas del Sur, Este y Centro revelan niveles de precipitación muy bajos, sobre todo en la temporada estival, lo cual es típico del clima mediterráneo y continental.

Se nota una mayor acumulación de precipitaciones en otoño e invierno, especialmente en las regiones atlánticas (Norte y Noroeste, mientras que el verano resulta seco en todas las áreas.

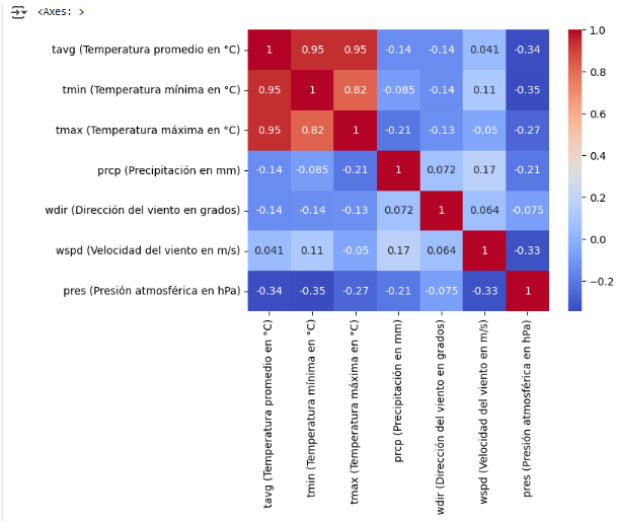


FIGURA 13. Matriz de correlación de Pearson entre las variables meteorológicas utilizadas.

Para una mejor comprensión de las interacciones entre las variables meteorológicas en el conjunto de datos, se ha obtenido la matriz de correlación de Pearson que relaciona las características más relevantes.

El gráfico revela una fuerte relación positiva entre las temperaturas mínimas, medias y máximas, sugiriendo que estas variables proporcionan información muy similar. En contraste, la presión atmosférica muestra correlaciones negativas con la mayoría de las variables, especialmente con las diferentes temperaturas, lo que puede ser significativo para la previsión.

Las variables vinculadas al viento (wdir, wspd) y a la precipitación (prcp) presentan correlaciones que son bajas o moderadas, indicando que podrían ofrecer información variada que resulta valiosa para los modelos.

```python
# Aseguramos que las columnas necesarias estén
df_total['Región'] = df_total['ciudad (Ciudad)'].map(regiones)
df_total['date'] = pd.to_datetime(df_total['date'])

# Obtenemos el día más caluroso por región (según tmax)
dias_mas_calurosos = df_total.loc[
    df_total.groupby('Región')['tmax (Temperatura máxima en °C)'].idxmax()
]

dias_mas_calurosos = dias_mas_calurosos[['Región', 'ciudad (Ciudad)', 'date', 'tmax (Temperatura máxima en °C)']].sort_values('Región')

print("Día más caluroso por región:")
print(dias_mas_calurosos)
```

```
Día más caluroso por región:
      Región ciudad (Ciudad)       date  tmax (Temperatura máxima en °C)
4963    Centro          Madrid 2024-07-24                             39.9
2763      Este          Murcia 2024-07-20                             42.5
4603   Noreste       Barcelona 2024-07-30                             34.5
2419  Noroeste           Lugo 2024-08-11                             36.2
223      Norte          Bilbao 2024-08-11                             42.9
4240       Sur          Málaga 2024-08-02                             42.2
```

FIGURA 14. Día más caluroso registrado por región según la temperatura máxima (tmax).

Se ha determinado también cuál fue el día más cálido registrado en cada área, que se define como aquel que alcanzó la temperatura máxima (tmax) más alta del año. Según los datos presentados en la tabla, los picos de temperatura más significativos ocurrieron durante los meses de Julio y Agosto, superando los 42°C en zonas como el este (Murcia), el sur (Málaga) e incluso, de manera inesperada, en el norte (Bilbao). Este fenómeno puede estar vinculado a eventos extremos ocasionales, como las olas de calor. Esta información proporciona un contexto valioso sobre la variabilidad de las temperaturas entre distintos lugares y permite establecer relaciones con las proyecciones y el diseño del modelo.

```python
df_total['Región'] = df_total['ciudad (Ciudad)'].map(regiones)
df_total['date'] = pd.to_datetime(df_total['date'])

# Obtenmosr el día más friolero por región (según tmin)
dias_mas_friolos = df_total.loc[
    df_total.groupby('Región')['tmin (Temperatura mínima en °C)'].idxmin()
]

dias_mas_frioleros = dias_mas_friolos[['Región', 'ciudad (Ciudad)', 'date', 'tmin (Temperatura mínima en °C)']].sort_values('Región')

print("Día más friolero por región:")
print(dias_mas_frioleros)
```

```
Día más friolero por región:
      Región ciudad (Ciudad)       date  tmin (Temperatura mínima en °C)
5143    Centro      Valladolid 2024-01-20                            -4.2
2918      Este          Murcia 2024-12-22                             2.3
4400   Noreste       Barcelona 2024-01-09                             2.5
2215  Noroeste           Lugo 2024-01-20                            -5.0
1117     Norte         Oviedo 2024-01-20                            -1.9
373        Sur        Sevilla 2024-01-08                            -0.2
```

FIGURA 15. Día más frío registrado por región según la temperatura mínima (tmin).

Del mismo modo, se ha identificado el día que más ha hecho frío por región en base a la temperatura mínima (tmin). Como se observa en la tabla, las mínimas más extremas ocurrieron en el mes de Enero, destacando Lugo con -5,0°C y Valladolid con -4,2°C.

En contra, ciudades del sur y este como Sevilla o Murcia presentaron temperaturas mínimas mucho más suaves, lo que refleja el efecto atenuante del clima mediterráneo y atlántico.

Estos extremos térmicos aportan contexto climático útil para entender la distribución y comportamiento de las variables en los modelos de predicción.

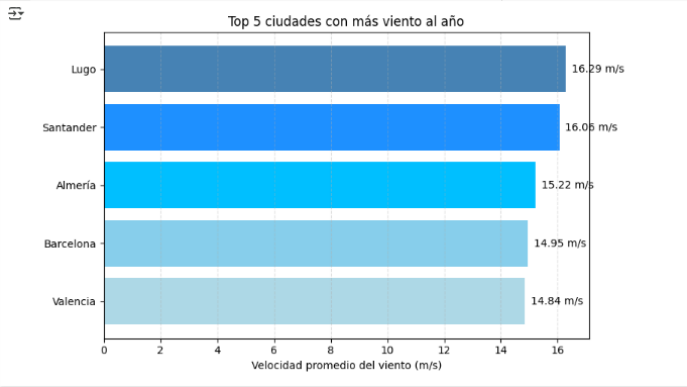


FIGURA 16. Top 5 ciudades con mayor velocidad promedio del viento al año (m/s).

El gráfico muestra las cinco ciudades con más viento del conjunto analizado, según la velocidad promedio anual del viento. Encabezan la lista Lugo(58,6km/h) y Santander (57,8km/h), ambas situadas en la zona norte, seguida por Almería (54,8 km/h), una ciudad del sur.

Esta información resulta útil no solo a nivel descriptivo, sino también para interpretar los resultados de los modelos de clasificación, especialmente si el viento es una variable predictiva relevante. Además, permite analizar relaciones con otros fenómenos meteorológicos como la precipitación.

```
# Filtramos las columnas necesarias y eliminamos filas con NaN
df_corr = df_total[['Región', 'prcp (Precipitación en mm)', 'wspd (Velocidad del viento en m/s)']].dropna()

# Calculamos la correlación entre precipitación y viento para cada región
correlaciones = df_corr.groupby('Región').apply(
    lambda g: g['prcp (Precipitación en mm)'].corr(g['wspd (Velocidad del viento en m/s)'])
)

print("📊 Correlación entre lluvia y viento por región:")
print(correlaciones.sort_values(ascending=False).round(3))
```

```
📊 Correlación entre lluvia y viento por región:
Región
Noroeste    0.316
Norte       0.233
Centro      0.159
Noreste     0.091
Este        0.032
Sur        -0.050
```

FIGURA 17. Correlación entre la precipitación y la velocidad del viento por región.

Se ha calculado el coeficiente de correlación de Pearson entre la velocidad del viento (wspd) y la precipitación (prcp) para cada región, con el objetivo de explorar posibles relaciones entre estos fenómenos meteorológicos.

Como se observa en la tabla, las regiones del noroeste (p =0.316) y norte (p=0.233) presentan una correlación positiva moderada, lo que sugiere que, en estas zonas, la llegada de frentes lluviosos suele ir acompañada de un aumento en la velocidad del viento.

En contra, regiones como el sur (p=-0.050) o el este (p=0.032) muestran una relación casi nula o incluso ligeramente negativa, lo que podría deberse a la presencia de vientos secos y cálido no asociados a la precipitación.

Este análisis refuerza la idea de que el comportamiento meteorológico varía significativamente entre regiones.

3.2. Selección de modelos:

Para este estudio se han seleccionado los 6 modelos de machine learning más comunes que hay actualmente. Cada uno de estos modelos tienen sus pros y sus contras, y se han elegido por su capacidad para manejar problemas de clasificación binaria con diferentes características de los datos. Los modelos seleccionados son:

1.Regresión lineal (baseline).

Aunque la regresión lineal no es un algoritmo específico para clasificación, se ha incluido como modelo de referencia para observar su rendimiento frente a técnicas mucho más avanzadas. Este modelo predice un valor continuo a partir de una combinación lineal de variables independientes, y en este caso se ha adaptado para clasificar ciudades mediante redondeo y control de clase.

```python
# REGRESIÓN LINEAL

from sklearn.linear_model import LinearRegression
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix
from sklearn.preprocessing import LabelEncoder
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

# Codificamos las ciudades como números
encoder = LabelEncoder()
y_train_encoded = encoder.fit_transform(y_train)
y_test_encoded = encoder.transform(y_test)

# Aquí entrenamos el modelo de regresión lineal
model_lr = LinearRegression()
model_lr.fit(X_train_imputed, y_train_encoded)

# Predecimos y redondeamos
y_pred_lr = model_lr.predict(X_test_imputed)
y_pred_lr_class = np.round(y_pred_lr).astype(int)
```

FIGURA 18. Implementación del modelo de Regresión Lineal para la predicción de ciudades a partir de variables meteorológicas.

```
Accuracy (Regresión Lineal - baseline): 0.067
Precision: 0.134
Recall: 0.067
F1-Score: 0.046
Matriz de Confusión:
[[ 0  0  0  1  3 15 39 14  1  0  0  0  0  0  0]
 [ 0  0  0  0  1  5 32 26  7  1  1  0  0  0  0]
 [ 0  0  0  0  1  1 15 23 24  7  2  0  0  0  0]
 [ 0  0  0  5  7  8 28 19  7  0  0  0  0  0  0]
 [ 0  0  0  0  0  4 23 34  9  3  0  0  0  0  0]
 [ 0  0  0  0  2 12 19 13 18  7  2  0  0  0  0]
 [ 0  1  0  0  0  0 11 27 21 13  1  0  0  0  0]
 [ 0  0  0  0  0  1 10 25 26 10  1  0  0  0  0]
 [ 0  0  0  0  0  2 30 28  9  3  0  1  0  0  0]
 [ 0  0  0  0  0  0  1 23 35 11  3  0  0  0  0]
 [ 0  1  0  1  5 13 30 18  4  1  0  0  0  0  0]
 [ 0  0  0  0  0  2  8 27 29  7  0  0  0  0  0]
 [ 0  0  0  2  0  6 23 20 14  6  1  1  1  0  0]
 [ 0  0  0  0  0  0  0  7 38 20  6  2  0  0  0]
 [ 0  0  0  0  0  2 11 22 28  6  3  1  0  0  0]]
```

FIGURA 19. Métricas de evaluación (Accuracy, Precision, Recall y F1-Score) y matriz de confusión del modelo de Regresión Lineal aplicado a la predicción de ciudades.

Estas cifras quieren indicar que la regresión lineal no es adecuada para este problema multiclase. Esto es esperable, ya que este modelo no esta diseñado para clasificación, y menos aún para clasificación multiclase con variables categóricas codificadas.

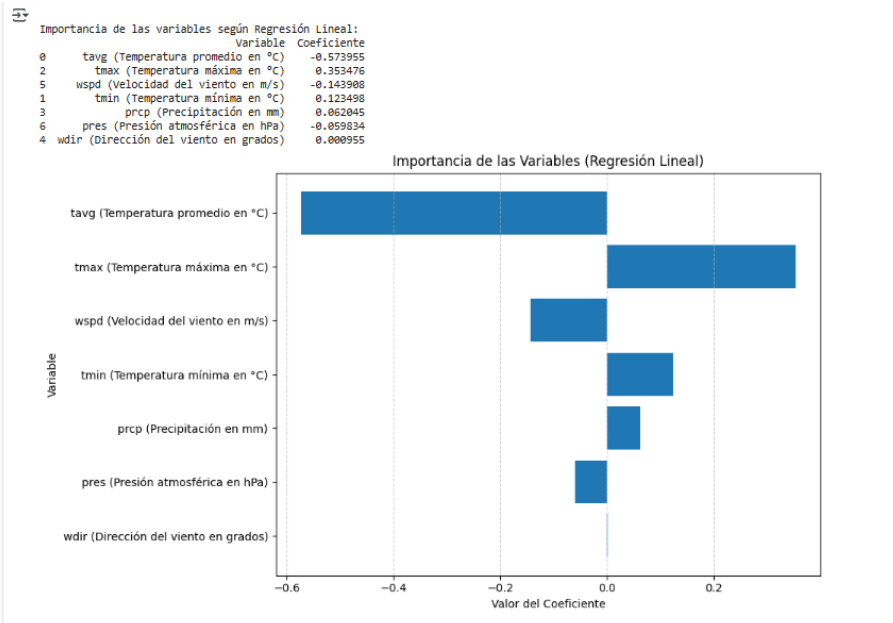


FIGURA 20. Importancia de las variables predictoras según el modelo de Regresión Lineal.

El análisis de coeficientes del modelo de regresión lineal desvela que las variables más importantes en la predicción de la ciudad son, la temperatura promedio y la temperatura máxima. La velocidad del viento también tiene un peso relevante.

En cambio, otras variables como la presión atmosférica o la dirección del viento tienen un impacto limitado. Esta información resulta esencial para entender qué factores meteorológicos tienen mayor capacidad discriminante en el conjunto de datos.

2.Regresión logística.

En estadística, la regresión logística es un tipo de modelo que se utiliza para anticipar el resultado de una variable que tiene categorías. En otras palabras, este método se aplica para calcular la probabilidad de que una variable categórica asuma un valor específico según las variables independientes.

```python
# REGRESIÓN LOGÍSTICA

from sklearn.linear_model import LogisticRegression
from sklearn.multiclass import OneVsRestClassifier
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix

# Usamos OneVsRestClassifier para manejar clasificación multiclase
model_log_reg = OneVsRestClassifier(LogisticRegression(max_iter=1000))

# Entrenamos el modelo
model_log_reg.fit(X_train_imputed, y_train_encoded)

# Realizamos las predicciones
y_pred_log_reg = model_log_reg.predict(X_test_imputed)
```

FIGURA 21. Implementación del modelo de Regresión Logística con OneVsRestClassifier.

```
Accuracy (Regresión Logística): 0.3105646630236794
Precision (Regresión Logística): 0.2950523710441377
Recall (Regresión Logística): 0.3105646630236794
F1-Score (Regresión Logística): 0.28929610549565027
```

FIGURA 22. Métricas de evaluación del modelo de Regresión Logística.

El modelo de regresión logística, más adecuado que la regresión lineal para tareas de clasificación, consiguió mejorar el rendimiento con una precisión del 31,1% y una F1-Score del 28,9%. Aun así, estos resultados muestran que el modelo tiene dificultades para distinguir adecuadamente entre las diferentes ciudades, cometiendo errores frecuentes tanto en la predicción como en la recuperación de clases reales. Esto sugiere que, aunque es un avance respecto al baseline, se necesitan modelos más complejos para capturar mejor los patrones en los datos meteorológicos.

El tipo más común de regresión logística es la regresión logística binaria, que solo presenta dos posibles resultados: “fracaso” o “éxito” (distribución de Bernoulli). El “fracaso” se indica con el valor 0, mientras que el “éxito” se señala con el valor 1.

```
Importancia de las variables según Regresión Logística Multiclase:
                                Variable  Importancia promedio (|coef|)
0      tavg (Temperatura promedio en °C)                       0.361159
1        tmin (Temperatura mínima en °C)                       0.262266
2        tmax (Temperatura máxima en °C)                       0.260081
5     wspd (Velocidad del viento en m/s)                       0.128686
3            prcp (Precipitación en mm)                       0.078086
6      pres (Presión atmosférica en hPa)                       0.019285
4  wdir (Dirección del viento en grados)                       0.001818
```

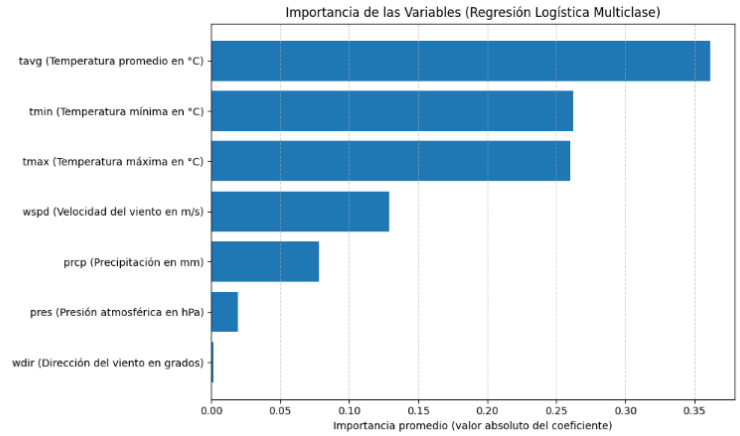


FIGURA 23. Importancia de las variables según Regresión Logística Multiclase.

En el modelo de Regresión logística multiclase, la variable más influyente fue tavg (temperatura promedio en °C), seguida de tmin y tmax. Estas variables presentan valores promedio absolutos de coeficientes significativamente mayores que el resto, lo que indica que las temperaturas son el principal factor que permite diferenciar entre ciudades en función de sus características meteorológicas. Por el contrario, variables como la dirección del viento (wdir) y la presión atmosférica (pres) apenas contribuyen a la clasificación, mostrando coeficientes cercanos a 0 en la mayoría de las clases.

3.Arboles de decisión.

Un árbol de decisión es un método de aprendizaje supervisado que no necesita parámetros, utilizado para actividades de clasificación y regresión. Su forma se asemeja a un árbol, que incluye un nodo principal, ramas, nodos intermedios y nodos finales.

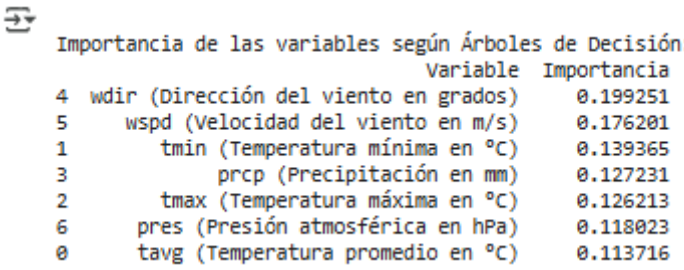
```
Accuracy (Árboles de Decisión): 0.4089253187613843
Precision (Árboles de Decisión): 0.41245165915204285
Recall (Árboles de Decisión): 0.4089253187613843
F1-Score (Árboles de Decisión): 0.4100712577023214
```

FIGURA 24. Métricas del modelo de Árboles de Decisión.

El modelo de árboles de decisión alcanzó un rendimiento notablemente mejor que los modelos anteriores, con una precisión del 40,9% y un F1-Score del 41%. Aunque aún no sigue siendo óptimo, este modelo muestra una mejor capacidad para distinguir entre las distintas ciudades utilizando únicamente variables meteorológicas.

Además, el equilibrio entre precisión y recall indica que el modelo no tiende a favorecer clases específicas, lo cual es positivo en contextos multiclase.

Estos resultados posicionan al árbol de decisión como un modelo más adecuado para el problema planteado, aunque aún se espera mejorar su rendimiento con técnicas de ensamblado como Random Forest.



Importancia de las variables según Árboles de Decisión:

	Variable	Importancia
4	wdir (Dirección del viento en grados)	0.199251
5	wspd (Velocidad del viento en m/s)	0.176201
1	tmin (Temperatura mínima en °C)	0.139365
3	prcp (Precipitación en mm)	0.127231
2	tmax (Temperatura máxima en °C)	0.126213
6	pres (Presión atmosférica en hPa)	0.118023
0	tavg (Temperatura promedio en °C)	0.113716

FIGURA 25. Importancia de las variables según el modelo de Árboles de Decisión.

En el modelo de Árboles de decisión, la variable con mayor importancia fue la dirección del viento (wdir), seguida de la velocidad del viento (wspd) y la temperatura mínima (tmin). Esto contrasta con los resultados obtenidos en regresión lineal y regresión logística, donde predominaban las variables relacionas con la temperatura promedio. El modelo de árbol parece encontrar patrones útiles en variables menos influyentes que en otros modelos, como wdir, lo cual puede deberse a su capacidad para captura relaciones no lineales y divisiones específicas del espacio de características.

4.Random forest.

Es un método de aprendizaje automático que, al ser un modelo de conjunto, une varios árboles de decisión para formar una predicción más sólida y exacta.

Este enfoque emplea una estrategia llamada “bagging” (agregación bootstrap) que produce distintas muestras de entrenamiento a partir del conjunto de datos original, utilizando cada muestra para entrenar un árbol de decisión. Posteriormente, los árboles de decisión se combinan en un modelo global mediante un promedio ponderado.

Cada árbol de decisión en el Random forest se entrena utilizando una selección aleatoria de características y ejemplos de entrenamiento, lo que le permite captar diferentes patrones en varios subconjuntos de los datos. Al hacer una predicción para un nuevo caso, cada árbol de decisión del Random forest emite su propia predicción, y el resultado final del modelo es la media ponderada de todas las predicciones de los árboles individuales.

```
Accuracy (Random Forest): 0.5191256830601093
Precision (Random Forest): 0.5189706383117207
Recall (Random Forest): 0.5191256830601093
F1-Score (Random Forest): 0.5164218876752436
```

FIGURA 26. Métricas de evaluación para el modelo Random Forest.

El modelo Random forest fue con el que obtuve mejores resultados generales, alcanzando un 51,9% de precisión y una F1-Score de 51,6%, Esta mejora significativa respecto a modelos anteriores confirma que la combinación de múltiples árboles de decisión aporta una mayor robustez y generalización al clasificador. Random forest logra un buen equilibrio entre precisión y cobertura, mostrando que es capaz de diferenciar eficazmente entre múltiples ciudades en base a variables meteorológicas. Por tanto, se consolida como el modelo más prometedor dentro del conjunto evaluado.

Importancia de las variables según Random Forest:

	Variable	Importancia
5	wspd (Velocidad del viento en m/s)	0.174338
4	wdir (Dirección del viento en grados)	0.173837
2	tmax (Temperatura máxima en °C)	0.146885
1	tmin (Temperatura mínima en °C)	0.145297
0	tavg (Temperatura promedio en °C)	0.133373
6	pres (Presión atmosférica en hPa)	0.124084
3	prcp (Precipitación en mm)	0.102186

FIGURA 27. Importancia de las variables según el modelo Random Forest.

En el modelo de Random forest, las variables con mayor importancia fueron la velocidad (wspd) y dirección del viento (wdir), con valores de 0.174 y 0.173 respectivamente. Estas variables superaron incluso a la temperatura máxima y mínima, lo cual indica que los patrones del viento pueden ser determinantes en la clasificación de ciudades a partir de datos meteorológicos. Las variables térmicas (tmax, tmin y tavg) mantuvieron una relevancia consistente, mientras que la precipitación (prcp) tuvo una contribución menor pero no nula.

La robustez del modelo de Random forest, al combinar múltiples árboles, permite capturar relaciones complejas y no lineales, proporcionando una visión global más fiable de la importancia de las variables.

5.Support vector machine.

Es un método de aprendizaje automático empleado en tareas de clasificación y regresión. Su objetivo principal es encontrar el hiperplano óptimo que separe correctamente las distintas clases dentro de un conjunto de datos.

Este tipo de modelo es especialmente útil para problemas de clasificación binaria, como por ejemplo distinguir entre spam y no spam, o entre gato y perro. Lo que hace único al SVM es que no se conforma con separar las clases, sino que busca maximizar el margen entre ellas. Cuanto mayor sea ese margen, mejor será la capacidad del modelo para generalizar a datos nuevos.

Gracias a esa característica, los SVM son adecuados para conjuntos de datos complejos, especialmente aquellos con muchas variables o donde la separación entre clases no es evidente.

```python
# SUPPORT VECTOR MACHINE (SVM)

from sklearn.svm import SVC
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix

# Entrenamos el modelo SVM con kernel lineal
model_svm = SVC(kernel='linear', random_state=42)
model_svm.fit(X_train_imputed, y_train_encoded)
```

FIGURA 28. Entrenamiento del modelo Support Vector Machine (SVM) con kernel lineal.

```
Accuracy (SVM): 0.38160291438979965
Precision (SVM): 0.39711457104573905
Recall (SVM): 0.38160291438979965
F1-Score (SVM): 0.3797298964032191
```

FIGURA 29. Métricas de rendimiento del modelo Support Vector Machine (SVM).

El modelo de máquina de vectores soporte (SVM) obtuvo resultados intermedios con una precisión del 38,2% y un F1-Score del 38,0%. Aunque supera a modelos básicos como la regresión logística, queda por debajo de otros como los árboles de decisión o Random forest. Estos resultados sugieren que, si bien SVM puede ser útil para ciertos tipos de datos, no resulta tan eficaz en problemas multiclase con variables meteorológicas como es en este caso, probablemente debido a la dificultad para encontrar márgenes óptimos entre tantas clases diferentes.

```
Importancia de las variables según SVM Lineal Multiclase:
                                Variable  Importancia promedio (|coef|)
0      tavg (Temperatura promedio en °C)                       0.845776
1        tmin (Temperatura mínima en °C)                       0.551417
2        tmax (Temperatura máxima en °C)                       0.488010
5     wspd (Velocidad del viento en m/s)                       0.232045
3            prcp (Precipitación en mm)                       0.137778
6      pres (Presión atmosférica en hPa)                       0.060205
4  wdir (Dirección del viento en grados)                       0.002822
```

FIGURA 30. Importancia de las variables según el modelo SVM lineal multiclase.

En el modelo de SVM con kernel lineal, las variables con mayor relevancia fueron claramente las temperaturas (tavg, tmin y tmax), que dominaron en términos de magnitud promedio de los coeficientes entre pares de clases.

Esto indica que el modelo SVM considera las diferencias térmicas como el principal criterio separador entre ciudades. Las variables relacionadas con el viento y la presión fueron menos determinantes, especialmente la dirección del viento (wdir), cuya influencia fue casi nula. Este modelo refuerza la relevancia de las variables térmicas observada en otros clasificadores lineales como la Regresión logística.

6.Redes neuronales artificiales (ANN).

Una red neuronal artificial (ANN) es un componente de un sistema informático creado para imitar cómo el cerebro humano interpreta y gestiona la información.

Sirve como fundamento de la inteligencia artificial (IA) y resuelve problemas que serían muy difíciles o incluso imposibles de abordar según los criterios humanos o estadísticos.

Las ANN tienen la capacidad de aprender por sí mismas, lo que les permite mejorar sus resultados a medida que reciben más datos. Tras la fase de análisis exploratorio y la selección de modelos candidatos, se procede a implementar los algoritmos de aprendizaje automático que han demostrado un mayor rendimiento en la clasificación binaria. En este caso, el objetivo consiste en predecir si lloverá o no en un determinado día, a partir de variables meteorológicas como temperatura, viento y presión atmosférica.

La retropropagación es el conjunto de pautas de aprendizaje que se utilizan para dirigir las redes neuronales artificiales.

```python
# REDES NEURONALES ARTIFICIALES (ANN)

from sklearn.neural_network import MLPClassifier
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix

# Entrenamos el modelo de Redes Neuronales (ANN)
model_ann = MLPClassifier(hidden_layer_sizes=(100,), max_iter=1000, random_state=42)
model_ann.fit(X_train_imputed, y_train_encoded)
```

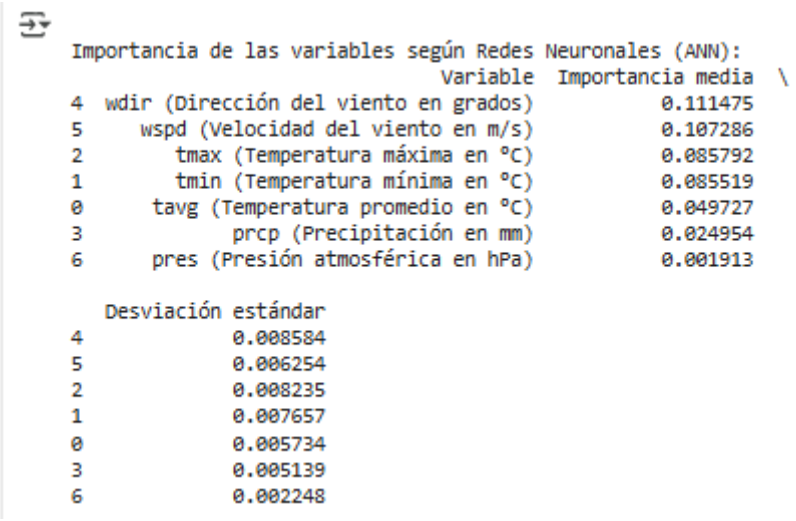
FIGURA 31. Implementación del modelo de Redes Neuronales Artificiales (ANN).

```
Accuracy (ANN): 0.28415300546448086
Precision (ANN): 0.36836053872531727
Recall (ANN): 0.28415300546448086
F1-Score (ANN): 0.22495585854239944
```

FIGURA 32. Evaluación del modelo de Redes Neuronales Artificiales (ANN).

El modelo de redes neuronales artificiales (ANN) obtuvo los peores resultados del conjunto, con una precisión del 28,4% y un F1-Score del 22,5%.

A pesar de ser un modelo potente y flexible, su rendimiento fue limitado posiblemente debido a una mala configuración de hiperparámetros, un número insuficiente de capas o neuronas, o la necesidad de mayor cantidad de datos para aprender correctamente. Esto pone de manifiesto que las ANN requieren una optimización más cuidadosa y un ajuste fino para problemas de clasificación multiclase como este, y que su uso no garantiza buenos resultados sin un diseño adecuado.



Importancia de las variables según Redes Neuronales (ANN):

	Variable	Importancia media	Desviación estándar
4	wdir (Dirección del viento en grados)	0.111475	0.008584
5	wspd (Velocidad del viento en m/s)	0.107286	0.006254
2	tmax (Temperatura máxima en °C)	0.085792	0.008235
1	tmin (Temperatura mínima en °C)	0.085519	0.007657
0	tavg (Temperatura promedio en °C)	0.049727	0.005734
3	prcp (Precipitación en mm)	0.024954	0.005139
6	pres (Presión atmosférica en hPa)	0.001913	0.002248

FIGURA 33. Importancia de las Variables según Redes Neuronales Artificiales (ANN).

El análisis de importancia de variables mediante permutación revela que las características relacionadas con el viento (wdir y wspd) son las que más influyen en las predicciones del modelo de red neuronal, seguidas por la temperaturas máxima y mínima. Por el contrario, la presión atmosférica (pres) tiene un impacto prácticamente nulo en el rendimiento del modelo.

La desviación estándar asociada a cada variable indica que las más importantes no solo tienen un mayor impacto en el accuracy, sino también una mayor consistencia entre repeticiones, lo que aumenta la confianza en su relevancia. Esta información es clave para entender cómo la red neuronal “prioriza” ciertas variables meteorológicas al clasificar las ciudades.

A pesar de que hay varios métodos para abordar los desafíos de la predicción del clima, en este documento se ha elegido una selección de modelos tradicionales de aprendizaje automático supervisado.

Esta elección se hace considerando tantos aspectos educativos como técnicos. Modelos como la Regresión logística, SVM y Random forest proporcionan una buena mezcla de claridad, efectividad y facilidad de uso. Sin embargo, se admite que hay otros algoritmos que podrían haber sido evaluados, como ARIMA o ARIMAX para datos en series temporales, o métodos más sofisticados como XGBoost y LightGBM, que a menudo superan a Random Forest en muchas situaciones competitivas. Estos modelos no se han incorporado debido a razones de simplicidad y el enfoque educativo del trabajo, aunque adopción podría ser una forma evidente de mejorar en estudios futuros.

3.3. Implementación de los modelos.

Tras la fase de análisis exploratorio y la selección de modelos candidatos, se procede a implementar los algoritmos de aprendizaje automático que han demostrado un mayor rendimiento en la clasificación binaria.

En este caso, el objetivo consiste en predecir si lloverá o no en un determinado día, a partir de variables meteorológicas como temperatura, viento y presión atmosférica.

Para ello, se construyó una nueva variable binaria denominada llueve, que toma valor 1 si se registró alguna precipitación (prcp>0) y 0 en caso contrario. El conjunto de datos fue dividido en entrenamiento y prueba utilizando una partición estratificada, a fin de preservar la proporción entre clases y evitar sesgos durante la validación.

De entre los modelos evaluados, se seleccionaron los dos con mayores métricas globales (precisión, F1-Score y AUC).

-Random forest, por su alta capacidad de generalización.

-Árbol de decisión, por su simplicidad y facilidad de interpretación.

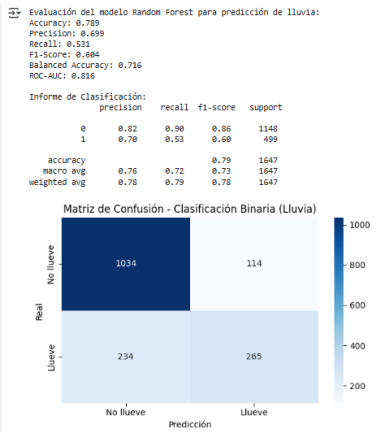


FIGURA 34. Evaluación del Modelo Random Forest para la Clasificación Binaria de Lluvia.

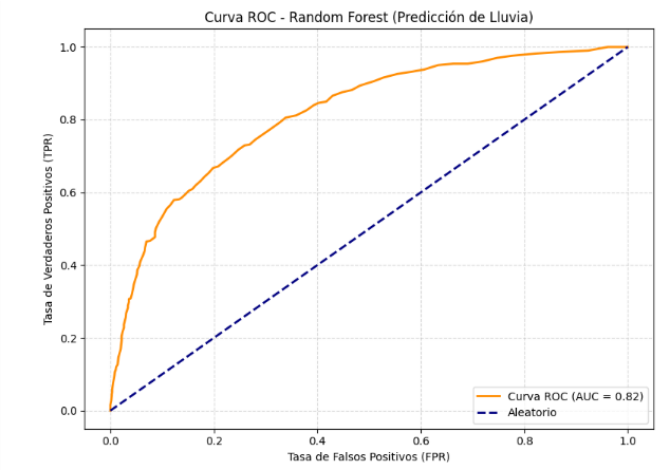


FIGURA 35. Curva ROC del Modelo Random Forest para la Predicción de Lluvia.

Resumen general del rendimiento del modelo: Métricas globales:

-Accuracy 0.789: El modelo acierta en casi el 79% de los casos totales. Buena tasa de acierto.

-Precisión 0.699: Cuando el modelo predice que va a llover, acierta el 70% de las veces.

-Recall (Sensibilidad) 0.531: El modelo detecta el 53% de los días que realmente llueve. Mejora posible.

-F1-Score 0.604: Media armónica entre precisión y recall. Buen balance.

-Balanced Accuracy 0.716: Corrige el desbalance de clases (lluvia vs no lluvia). Aceptable.

-ROC-AUC 0.816: Excelente capacidad para discriminar entre clases (mejor cuanto más cerca de 1).

Informe de Clasificación (por clase):

-El modelo identifica bien los días sin lluvia (recall 0.90).

-Le cuesta más detectar los días de lluvia (solo detecta el 53%).

-Equilibrio aceptable gracias a que la precisión de lluvia es buena (0.70).

Matriz de Confusión:

-Aciertos (diagonales):

1034 días correctamente predichos como sin lluvia.

265 días correctamente predichos como sin lluvia.

-Errores (fuera de la diagonal):

114 días falsamente etiquetados como con lluvia (falsos positivos).

234 días de lluvia no detectados (falsos negativos, más críticos).

El modelo es conservador: tiende a predecir más días sin lluvia, por eso se escapan bastantes días de lluvia reales (bajo recall en clase 1).

Curva ROC:

El modelo de Random forest es capaz de predecir lluvia significativamente mejor que el azar. La curva ROC obtenida para el modelo Random forest muestra un AUC de 0.82, lo que indica una alta capacidad de discriminación entre días con y sin lluvia.

Es especialmente útil cuando se quiere maximizar el recall (por ejemplo, para alertas de lluvia), ya que se puede mover el umbral hacia la izquierda.

El área bajo la curva confirma que hay patrones meteorológicos detectables que correlacionan con la lluvia.

Random Forest no solo da buenas métricas como accuracy o F1-Score, sino que también se comporta de forma robusta a distintos umbrales de decisión.

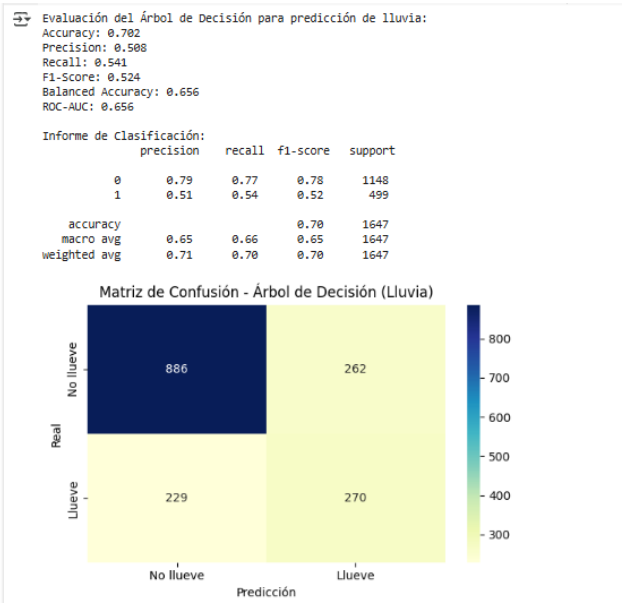


FIGURA 36. Evaluación del Árbol de Decisión para la Predicción de Lluvia.

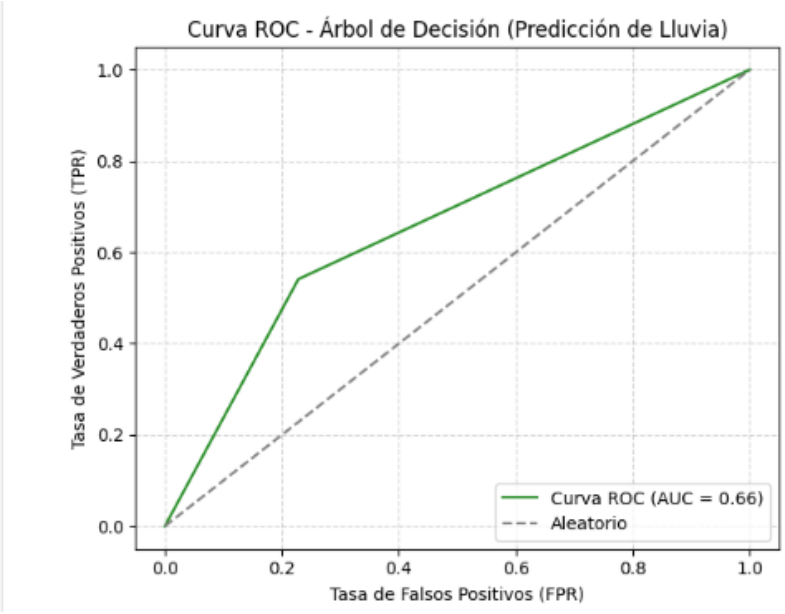


FIGURA 37. Curva ROC - Árbol de Decisión (Predicción de Lluvia).

1.Métricas generales del modelo:

-Accuracy 0.702: El modelo acierta aproximadamente el 70% de las veces.

-Precisión 0.588: De todas las veces que predijo “llueve”, solo el 58.8% erran correctas.

-F1-Score 0.524: Este valor representa un equilibrio entre precisión y recall, y no es alto, lo que indica que el modelo tiene limitaciones para detectar lluvia con fiabilidad.

-ROC-AUC 0.656: La capacidad de discriminar entre días con y sin lluvia es moderada, mejor que aleatoria (0.5) pero inferior a Random Forest (0.816).

2.Matriz de Confusión:

El modelo tiende a confundir muchos días lluviosos como días secos (229 falsos negativos).

También hay un número considerable de falsos positivos (262 veces predice lluvia cuando no llueve).

Es un modelo más conservador que el Random Forest (más cauteloso al predecir lluvia), lo cual puede ser útil en algunas aplicaciones, pero tiene más errores en ambas direcciones.

3.Curva ROC:

El área bajo la curva (AUC=0.66) indica que el modelo tiene cierta capacidad discriminativa, pero no es fuerte.

La curva se acerca ligeramente al borde superior izquierda (ideal), pero está lejos de ser óptima.

Comparado con el Random Forest, este modelo tiene una curva ROC menos curvada y por tanto menos eficaz.

4.RESULTADOS

4.1. Rendimiento de los modelos (métricas y visualización de los resultados).

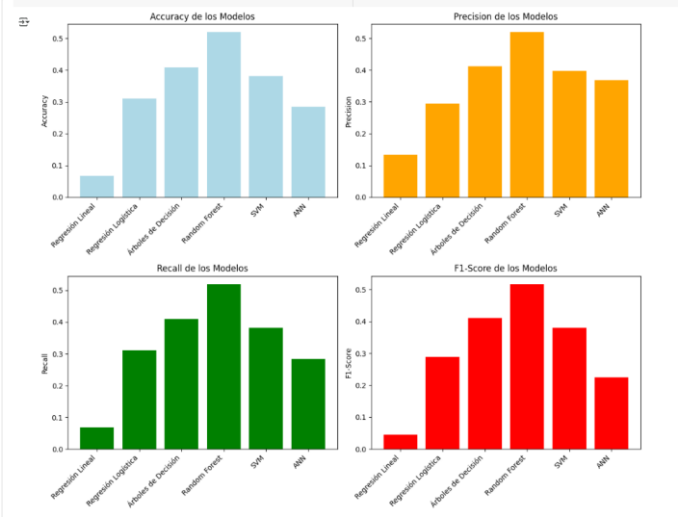


FIGURA 38. Comparativa de Métricas de Desempeño entre los distintos Modelos.

1.Accuracy de los modelos:

-Random Forest destaca con una accuracy de ~0.52, el valor más alto.

-Árboles de decisión le siguen con ~0.41, y luego SVM (~0.38).

-Regresión lineal queda muy por debajo (~0.06), lo que confirma que no es adecuada para este tipo de clasificación.

Este gráfico refuerza la elección de Random Forest y Árboles de decisión como los modelos más prometedores.

2.Precisión:

-Random Forest lidera con ~0.52, mostrando que cuando predice “llueve”, suele acertar.

-Árboles de decisión (~0.41) y SVM (~0.40) también tienen buenos valores.

-Regresión logística (~0.29) es mejor que la lineal, pero queda por debajo de los árboles.

Esto sugiere que los modelos basados en árboles son más seguros al etiquetar días de lluvia.

3.Recall (Sensibilidad):

-Nuevamente, Random Forest logra el mejor valor (~0.52), lo que indica que identifica correctamente la mayoría de los días que llueve. Le sigue el Árbol de Decisión (~0.41), y SVM (~0.38).

-Modelos como Regresión Lineal o Logística tienen dificultades para capturar correctamente los positivos (días de lluvia).

Esto es especialmente importante en meteorología, donde fallar un día lluvioso tiene más impacto que fallar un día seco.

4.F1-Score:

-El F1-Score combina precisión y recall, por lo que es el mejor indicador global del equilibrio del modelo.

-Random Forest nuevamente encabeza (~0.51), seguido por Árboles de decisión (~0.41) y SVM (~0.38).

-ANN no rinde tan bien como se esperaba (~0.22), probablemente por configuración subóptima o necesidad de más datos.

Aunque los resultados muestran que Random forest tuvo el mejor rendimiento, es importante destacar que los valores absolutos de precisión y recall, alrededor de 0.52, se consideran moderados en términos generales. En investigaciones similares sobre predicción de lluvias, las métricas superiores a 0.6 se consideran aceptables, mientras que cifras cercanas a 0.5 sugieren una leve mejora en comparación con un clasificador aleatorio. Así que, a pesar de que el modelo se desempeña mejor que el azar y un baseline simple, todavía hay espacio para mejorar para ser visto como sólido en situaciones reales.

4.2. Análisis de las variables más relevantes en la predicción.

El análisis de la importancia de variables realizado a través del modelo Random Forest reveló que la presión atmosférica y la velocidad del viento fueron las variables con mayor peso a la hora de predecir la ocurrencia de las precipitaciones. Este hallazgo resulta coherente con los patrones identificados en el análisis exploratorio (sección 3.2), donde

ambas variables mostraban correlaciones destacadas con los episodios de lluvia, especialmente en regiones del norte y centro peninsular.

Por el contrario, variables como la temperatura media o la dirección del viento tuvieron un impacto predictivo menor, lo cual sugiere que su relación con los eventos de lluvia, al menos en este conjunto de datos y configuración, no es tan determinante. Esta observación puede orientar futuras implicaciones del modelo, eliminando o sustituyendo variables con bajo aporte informativo, y reforzando así la eficacia del proceso predictivo.

4.3. Comparativa entre los enfoques iniciales y los avanzados.

Durante la fase de experimentación se evaluaron distintos enfoques predictivos, desde modelos lineales como la regresión logística, hasta algoritmos más avanzados como SVM, redes neuronales artificiales (ANN) y modelos basados en árboles. Si bien los modelos lineales ofrecían una implementación más sencilla y rápida, su rendimiento en métricas como accuracy y recall fue significativamente inferior.

Modelos como SVM y ANN mostraron un comportamiento intermedio, pero prestaban una mayor complejidad computacional y menor interpretabilidad. En cambio, los modelos basados en árboles, en especial Random Forest, ofrecieron un rendimiento consistentemente superior en todas las métricas evaluadas, manteniendo además una estructura que permite interpretar las decisiones del modelo de forma más clara.

En consecuencia, se decidió centrar el desarrollo y análisis final únicamente en los dos enfoques más robustos y explicables: Árbol de Decisión y Random Forest, lo cual permite un equilibrio adecuado entre rendimiento predictivo, interpretabilidad y eficiencia práctica.

5.DISCUSIÓN

5.1. Limitaciones del trabajo.

Aunque los resultados obtenidos han sido satisfactorios en varios aspectos, este trabajo presenta ciertas limitaciones que deben ser consideradas. En primer lugar, el dataset utilizado, aunque extenso y multiciudad, abarca únicamente un año de datos meteorológicos, lo que puede restringir la generalización de los modelos a patrones climáticos más amplios o atípicos. Además, la resolución temporal es diaria, lo que impide capturar eventos de lluvia de corta duración o asociados a microclimas locales.

Por otro lado, algunos modelos evaluados, como las redes neuronales, pueden no haber alcanzado su máximo rendimiento debido a la limitada optimización de hiperparámetros o a la falta de mayor volumen de los datos.

Asimismo, se ha asumido una homogeneidad en la calidad de los datos de todas las ciudades, sin realizar una evaluación exhaustiva de la posible presencia de errores de medición o valores atípicos.

5.2. Desafíos encontrados durante el desarrollo.

Durante el proceso de desarrollo se enfrentaron diversos desafíos tanto técnicos como metodológicos. Uno de los principales fue la recolección y preprocesamiento de datos meteorológicos de forma homogénea para todas las ciudades seleccionadas, lo que implicó trabajar con múltiples archivos, formatos y variables faltantes.

A nivel de modelado, se identificaron dificultades para ajustar correctamente algunos algoritmos complejos (como SVM o ANN), que requerían mucho tiempo de entrenamiento o devolvían resultados inestables sin una cuidadosa parametrización. También resultó desafiante establecer una métrica única de comparación entre modelos, ya que algunas métricas favorecen la precisión general mientras que otras priorizan la detección de días lluviosos (recall), lo cual es especialmente relevante en meteorología.

Finalmente, uno de los retos más importantes fue mantener un equilibrio entre la interpretabilidad del modelo y su rendimiento, ya que los algoritmos más precisos (como Random Forest) suelen ser menos transparentes que los modelos lineales.

5.3. Potencial de mejora y optimización futura.

Este trabajo abre múltiples posibilidades de mejora y extensión. En primer lugar, se podría ampliar el período de análisis a varios años para capturar mejor la estacionalidad, la variabilidad interanual y fenómenos extremos poco frecuentes. Asimismo, incorporar variables adicionales como humedad, nubosidad, radiación solar o datos horarios podría enriquecer significativamente la calidad de las predicciones.

Además, es esencial tener en cuenta el gasto computacional que cada modelo implica. Aunque métodos avanzados como Random forest o las redes neuronales proporcionan un mejor rendimiento, su entrenamiento y funcionamiento necesitan más recursos de computación. Esto podría ser un inconveniente en situaciones donde se requiere rapidez, como en aplicaciones en tiempo real o con recursos limitados.

En casos donde la rapidez de las predicciones es crucial, podría ser mejor optar por modelos más sencillos, aunque sean un poco menos precisos, como los árboles de decisión individuales o las regresiones logísticas.

Otra mejora potencial radica en el uso de técnicas de selección de características automáticas, ingeniería de atributos o incluso el uso de técnicas de series temporales como modelos LSTM, que pueden capturar mejora la dependencia temporal entre días consecutivos.

También sería recomendable implementar una búsqueda de hiperparámetros más exhaustiva mediante métodos como Grid Search o Bayesian Optimization, para optimizar el rendimiento de modelos como SVM o ANN.

Otro factor que podría afectar directamente el aumento del rendimiento es la adición de nuevas variables que puedan predecir resultados.

En este estudio, se han utilizado variables climáticas simples, pero se podrían incluir elementos temporales (como la estación, el mes o el día de la semana), geográficos (como la altitud o la proximidad al océano) o incluso información de satélites. Aunque esto podría resultar en un mayor coste de computación y hacer el modelo más complicado, también podría mejorar notablemente la habilidad del sistema para hacer predicciones.

Por último, una futura línea de mejora podría centrarse en construir modelos específicos para cada región climática o estación del año, aumentando así la precisión local del sistema predictivo.

6.CONCLUSIONES

6.1. Resumen de hallazgos.

El presente trabajo ha demostrado la viabilidad de predecir eventos de lluvia utilizando modelos de machine learning aplicados a datos meteorológicos diarios de 15 ciudades españolas. Entre los distintos enfoques evaluados, los modelos basados en árboles de decisión, en particular Random Forest, han mostrado los mejores resultados en términos de precisión, recall y F1-Score.

Asimismo, se ha identificado que variables como la presión atmosférica y la velocidad del viento tienen un papel clave en la predicción de lluvias, lo cual es coherente con el conocimiento meteorológico previo. La comparación entre modelos simples y avanzados permitió justificar la elección final de los algoritmos utilizados.

6.2. Impacto de los resultados en aplicaciones prácticas.

Los resultados obtenidos tienen un claro potencial de aplicación práctica.

Por ejemplo, modelos como los desarrollados pueden integrarse en sistemas de predicción meteorológica a corto plazo para apoyar la toma de decisiones en sectores como la agricultura, el turismo o la gestión de infraestructuras. La posibilidad de contar con modelos explicables y entrenados con datos locales permite crear soluciones adaptadas a necesidades regionales, sin depender únicamente de sistemas globales complejos.

Además, el enfoque utilizado podría escalarse a otras regiones geográficas o aplicarse en plataformas de alerta temprana, mejorando así la resiliencia frente a eventos climáticos adversos.

6.3. Recomendaciones para futuros trabajos.

De cara a futuras investigaciones, se recomienda ampliar la base de datos temporal, así como incorporar nuevas variables y fuentes de datos.

También sería beneficioso evaluar el rendimiento de modelos más complejos, como redes neuronales recurrentes, especialmente si se dispone de datos por horas o en secuencia temporal.

Igualmente, se aconseja trabajar en la mejora de la visualización e interpretación de resultados para que los modelos puedan ser utilizados fácilmente por usuarios no expertos. Finalmente, sería interesante explorar la integración de estos modelos en aplicaciones interactivas, dashboards o APIs para su uso en tiempo real.

7.BIBLIOGRAFÍA

Daniel. (2023, 30 octubre). Machine Learning: definición, funcionamiento, usos. Formación En Ciencia de Datos | DataScientest.com. https://datascientest.com/es/machine-learning-definicion-funcionamiento-usos

El guardián del tiempo | Meteostat. (s. f.-b). https://meteostat.net/es/

Emilio, N. (2022, 30 diciembre). Tipos de análisis predictivo: clasificación vs. regresión | Bismart. info@bismart.com. https://blog.bismart.com/tipos-analisis-predictivo-clasificacion-regresion

González, A. G. (2024, 13 noviembre). ¿Cuál es la diferencia entre regresión y clasificación? Panama Hitek. https://panamahitek.com/cual-es-la-diferencia-entre-regresion-y-clasificacion-en-machine-learning/

Clasificación: Exactitud, recuperación, precisión y métricas relacionadas. (n.d.). Google for Developers. https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall?hl=es-419

Admin. (2020, October 22). *Accuracy, Precision, recall & F1 Score: Interpretation of Performance Measures – Exsilio Blog (Español)*. Hippocrates Guild. https://hippocratesguild.com/es/accuracy-precision-recall-f1-score-interpretation-of-performance-measures-exsilio-blog-2/

colaboradores de Wikipedia. (2025, January 7). *Asimilación de datos*. Wikipedia, La Enciclopedia Libre. https://es.wikipedia.org/wiki/Asimilaci%C3%B3n_de_datos

Efe, A. (2024, November 28). Cadena SER. *Cadena SER*. https://cadenaser.com/nacional/2024/11/28/desarrollan-una-ia-que-genera-imagenes-satelitales-de-futuras-inundaciones-y-puede-servir-para-evitar-catastrofes-cadena-ser/

El sistema europeo avisó del riesgo de inundación en Valencia días antes de la dana, pero subestimó su magnitud | Clima y Medio Ambiente | EL PAÍS - Bing. (n.d.). Bing. https://www.bing.com/search?pglt=299&q=El+sistema+europeo+avis%C3%B3+del+riesgo+de+inundaci%C3%B3n+en+Valencia+d%C3%ADas+antes+de+la+dana%2C+pero+subestim%C3%B3+su+magnitud+%7C+Clima+y+Medio+Ambiente+%7C+EL+PA%C3%8DS&cvid=afbd7604a9af42799a876d1101767627&gs_lcrp=EgRlZGdlKgYIAB

BFGDkyBggAEEUYOTIGCAEQRRg80gEHMTYxajBqMagCALACAA&FORM=ANNTA1&PC=EE19

Google DeepMind hits new milestone in AI weather forecasting - Bing. (n.d.). Bing. https://www.bing.com/search?pglt=299&q=Google+DeepMind+hits+new+milestone+in+AI+weather+forecasting&cvid=e257f0a4f0de4568b77911f3a194a4af&gs_lcrp=EgRlZGdlKgYIABBFGDkyBggAEEUYOTIGCAEQRRhAMgYIAhBFGD3SAQcxOTlqMGoxqAIAsAIA&FORM=ANNTA1&PC=EE19

Tirado Picado, V. R. (2022). *Redes Neuronales Artificiales como Modelo de Predicción de los Factores Climáticos en Nicaragua en el Periodo 2021-2022*.

Sanz, F. (2024, 6 julio). *Regresión Lineal. Qué es, para qué se utiliza y ejemplo práctico*. The Machine Learners. https://www.themachinelearners.com/regresion-lineal/

Estadística, P. Y. (2023, 6 marzo). *Regresión logística*. Probabilidad y Estadística. https://www.probabilidadyestadistica.net/regresion-logistica/

Ibm. (2025, 30 enero). Arboles de decisión. *¿Qué es un árbol de decisión?* https://www.ibm.com/es-es/think/topics/decision-trees

Qué es Random Forest Concepto y definición. Glosario. (s. f.). GAMCO, SL. https://gamco.es/glosario/random-forest/

GeeksforGeeks. (2025, 28 mayo). *Support Vector Machine (SVM) algorithm*. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/support-vector-machine-algorithm/

Duckerman, W. (2021, 3 junio). *¿Qué es una red neuronal artificial (ANN)?* Brita Inteligencia Artificial. https://brita.mx/que-es-una-red-neuronal-artificial-ann/

ENLACE A MI GITHUB:

https://github.com/Ignacioferrero/TFG-Lluvia