“Análisis de series temporales para la predicción de lluvia en estaciones meteorológicas de la ciudad de Manizales - Colombia”

Camilo Andres Pulzara Mora
DNI: 1053827301

Director:
Santiago Torres Jaramillo
Magister en Ciencia - Física
Universidad Nacional de Colombia

Primera Convocatoria

Universidad Internacional de Valencia
Facultad de Ciencia y Tecnología
Valencia, España
2022

Índice

Índice de Ilustraciones

Índice de Tablas

Resumen

El estudio de series de tiempo para realizar predicciones utilizando técnicas de Machine Learning e Inteligencia Artificial, ha despertado el interés de los investigadores por la diversidad de su aplicación en diferentes áreas, como meteorología, finanzas, turismo y medicina. Adicionalmente, se considera una parte esencial para el pronóstico de datos y podría ayudar a encontrar nuevos patrones que aporten nueva información. De igual manera, la tendencia de lluvia es importante para la predicción de eventos de desastres naturales, los cuales no son fáciles de pronosticar ya que dependen de varios factores atmosféricos.

Existen diferentes modelos estadísticos para estudiar las series temporales, tales como ARIMA, ARIMAX, SARIMA, SARIMAX, que han sido usados en una gran variedad de aplicaciones, como la predicción de lluvias, temperatura, contaminantes, enfermedades (COVID -19), viajes, mercado financiero, pues presentan diferentes indicadores para su evaluación y han mostrado gran eficiencia y buenos resultados.

Los datos obtenidos por el Sistema de Monitoreo Automatización y Control (SIMAC) fueron tratados y transformados, según las necesidades del problema. Por otro lado, se encontraron datos faltantes que se imputaron por el método *MissForest* y por imputación múltiple por ecuaciones encadenadas *(MICE)*. Además, para analizar el comportamiento temporal de los datos, se implementó la librería statsmodels, bajo la metodología de *Box Jenkins* y diferentes pruebas estadísticas fueron implementadas (*Augmented Dickey Fuller* - ADF) para determinar el tipo de serie de tiempo. El resultado obtenido evidencia un comportamiento estacionario con un valor de ADF < 0.05.

Se implementó auto ARIMA para buscar mejores resultados en los modelos estocásticos utilizados (ARIMA SARIMA SARIMAX), basados en la predicción por día de la precipitación para las 13 estaciones meteorológicas ubicadas en el municipio de Manizales, Colombia. Adicionalmente, se utilizó la técnica multivariante en la librería de Prophet y Neural Prophet, agregando las fechas de los eventos por deslizamientos de tierra que afectaron algunos sectores, debido a las fuertes lluvias, dentro de la ciudad de Manizales.

Los resultados obtenidos por los modelos ARIMA, SARIMA, SARIMAX y Neural Prophet, para las predicciones de los 7 días, muestran valores del error cuadrático medio (RMSE) y del error absoluto medio (MAE) mayores a 20. Por otro lado, el modelo Prophet alcanzó un valor de RMSE igual a 19.06313 y un MAE de 16.24064, donde se evidenciaron errores más bajos que los modelos estocásticos implementados en este trabajo. Finalmente, los valores predichos por la librería Prophet pueden ayudar como una herramienta para el desarrollo de mejores prácticas en la gestión de análisis y riesgo de deslizamientos de tierra en el área.

Palabras clave: ARIMA, SARIMA, SARIMAX, Prophet, Neural Prophet, Precipitación, Imputación de datos.

Summary

The study of time series for making predictions using Machine Learning and Artificial Intelligence techniques has aroused the interest of researchers due to the diversity of its application in different areas, such as meteorology, finance, tourism and medicine. Additionally, it is considered an essential part of data forecasting and could help to find new patterns that provide new information. Similarly, the rainfall trend is important for the prediction of natural disaster events, which are not easy to forecast as they depend on several atmospheric factors.

There are different statistical models to study time series, such as ARIMA, ARIMAX, SARIMA, SARIMAX, which have been used in a great variety of applications, such as the prediction of rainfall, temperature, pollutants, diseases (COVID -19), traveling, financial market, since they present different indicators for their evaluation and have shown great efficiency and good results.

The data obtained by the Monitoring Automation and Control System (SIMAC) were treated and transformed, according to the needs of the problem. On the other hand, missing data were found and imputed by the MissForest method and by multiple imputation by chained equations (MICE). Moreover, to analyze the temporal behavior of the data, the statsmodels library was implemented under the Box Jenkins methodology and different statistical tests were implemented (Augmented Dickey Fuller - ADF) for computing the type of time series. The result obtained evidence a stationary behavior with an ADF value < 0.05.

Auto ARIMA was implemented to search for better results in the stochastic models used (ARIMA, SARIMA, SARIMA and SARIMAX), based on the daily prediction of precipitation for the 13 meteorological stations located in the city of Manizales, Colombia. Additionally, the multivariate technique was used in the Prophet and Neural Prophet library, adding the dates of the landslide events that affected some sectors, due to heavy rains, within the city of Manizales.

The results obtained by the ARIMA, SARIMA, SARIMAX and Neural Prophet models, for the 7-day predictions, show root mean square error (RMSE) and mean absolute error (MAE) values greater than 20. On the other hand, the Prophet model reached an RMSE value equal to 19.06313 and an MAE of 16.24064, which showed lower errors than the stochastic models implemented in this work. Finally, the values predicted by the Prophet library can help as a tool for the development of best practices in landslide analysis and risk management in the area.

Keywords: ARIMA, SARIMA, SARIMA, SARIMAX, Prophet, Neural Prophet, Precipitation, Data Imputation.

1. Introducción

El cambio climático ha tenido una gran influencia a nivel mundial y tanto las entidades del gobierno como las no gubernamentales, se han encargado de realizar investigaciones con el fin de detectar anomalías o altos riesgos que puedan afectar directa e indirectamente, tanto la económica como en la sociedad. Además, las abundantes lluvias, así como las altas y bajas temperaturas, están causando daños en la variabilidad natural y se están presentando con mayor frecuencia los desastres naturales. Se prevé que el calentamiento en la tierra tendrá un mayor impacto en pocos años. Algunos trabajos realizados pronostican una redistribución de las lluvias a causa de los cambios en la circulación de la atmósfera provocando un aumento en la precipitación en diferentes zonas (Costa Posada, 2007).

En la actualidad la precipitación ha sido considerada como un indicador de productividad importante tanto para el ser humano como para la naturaleza. Además, de ser parte fundamental en países que impulsan la agricultura, como Colombia, México, Argentina, España, entre otros (FAO, 2022). Los fuertes cambios de temperatura en la tierra, a causa del efecto invernadero y a la emisión de CO_2 que producen las fábricas a nivel mundial, ha ocasionado que el nivel de lluvia necesario para abastecer las necesidades del ser humano aumente cada año. A consecuencia de esto se presentan diferentes fenómenos de desastres naturales como deslizamientos, inundaciones, y fuertes vientos. Por lo tanto, científicos y expertos de diferentes áreas han realizado importantes investigaciones para analizar y detectar patrones en la precipitación, con el fin de hacer predicciones para generar alertas sobre desastres naturales que causan las fuertes lluvias (Rudolf et al., 2005), (González-Hidalgo et al., 2011), (Mazzoglio et al., 2019), (Min et al., 2019), (Mazzoglio, 2022).

En Colombia, existen diferentes ciudades que atraviesan la cordillera de los Andes, donde se presentan tormentas y diluvios, asociadas al ascenso de masa del aire producto del choque entre ellas. La ciudad de Manizales, que se encuentra ubicada en la parte central de la cordillera de los andes, es influenciada por el fenómeno del niño y el frente intertropical (CIOH, 2010). Adicionalmente, las condiciones geológicas y topográficas han tenido un impacto negativo en algunos sectores en específico; debido a que, la población ha crecido en zonas de alto riesgo que tienen problemas con el uso del suelo. Igualmente, la falta de políticas de planificación territorial, ha ocasionado inestabilidad en la tierra, causado por sucesos naturales como deslizamientos, generando un impacto social y económico devastador. Por este motivo, Van Westen dedica un primer estudio sobre la probabilidad de ocurrencia de un deslizamiento a razón de la lluvia acumulada de los 25 días en esta ciudad (van Westen & Erlien, 1995).

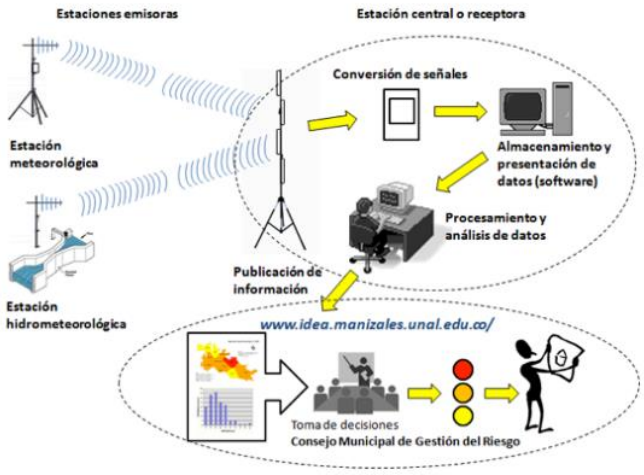
En la ciudad de Manizales, se han desarrollado proyectos en temas de prevención de desastres naturales, tales como inundaciones y deslizamientos con el fin de conocer los factores de causalidad (Grajales García, 2021), (Hardoy & Velásquez Barrero,

2014). Además, por ley en el Decreto 919 de 1989 y la ley 99 de 1993, se hace referencia sobre la parte de prevención y atención de desastres para garantizar la seguridad de los ciudadanos (*Decreto Ley 919 de 1989*, 1997). Por esta misma razón, en la última década, se han implementado en las entidades locales y académicas nuevos proyectos para realizar investigaciones en áreas de alto riesgo. Adicionalmente, pueden presentarse catástrofes que afecten directamente a las personas de las comunidades. De igual manera, se han establecido indicadores y áreas críticas que deben ser analizadas y monitoreadas constantemente.

Dentro del municipio de Manizales se encuentran alrededor de 14 estaciones meteorológicas situadas en diferentes zonas de la ciudad. Este proyecto, empezó con la ayuda de estudiantes de la Universidad Nacional de Colombia con el fin de, optimizar la información recolectada, actualizando los equipos, y haciendo uso de datos en tiempo real. Parte del desafío, era desarrollar un software junto a sistemas integrados que permitieran comunicar la red y las estaciones para obtener información actualizada y de manera inmediata para la toma de decisiones y prevención de desastres. El sistema es capaz de monitorear las siguientes variables: Precipitación, Intensidad Solar, Humedad Relativa, Temperatura, Presión, Dirección del Viento, y la evapotranspiración.

En la ilustración 1, se observa el funcionamiento de la red que se basa en mandar señales cada 5 minutos a la estación centralizada ubicada en las instalaciones del Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM) en la Universidad Nacional de Colombia, a través de VHF (*Very High Frecuency*), donde se almacenan y procesan los datos en tiempo real. Al final estos son analizados y evaluados para darle un uso debido a la información (Pachón Gómez et al., 2018a).

Ilustración 1. Funcionamiento de la Red (Pachón Gómez et al., 2018a)



Los fenómenos naturales son muy difíciles de pronosticar debido a que los sucesos están relacionados con otros factores atmosféricos, tales como cambios en la presión, en la temperatura, la humedad del lugar, la dirección del viento y la intensidad de la

radiación. Por este motivo, tener un conocimiento sobre los cambios del clima se vuelve un requisito principal a la hora de evaluar los riesgos en la infraestructura, medio ambiente, y en la sociedad misma.

Las series de tiempo, toman un papel importante para la predicción del modelamiento de información meteorológica y en diversas áreas donde se presenten fenómenos que varíen en términos de intervalos de tiempo (Collischonn et al., 2005), (Hung et al., 2009), (Mahsin et al., 2012). Existen diferentes objetivos que incluyen: la comprensión y la descripción de un mecanismo de generación, la predicción de valores futuros y el control óptimo de un sistema. La naturaleza de una serie de tiempo se basa en que sus observaciones pueden estar correlacionadas o ser dependientes, donde, y su orden es significativo (Wei, 1991).

La predicción de la lluvia o inundaciones como eventos probabilísticos es un tema esencial para la planificación del recurso hídrico. Este tipo de variables se miden de manera longitudinal en el tiempo. De esta manera, el análisis de series de tiempo de eventos con valores discretos se vuelve apropiado para monitorear el comportamiento hidrológico (Ansari, 2013). De igual manera, la precipitación hace parte de los componentes complejos y desafiantes del ciclo hidrológico para modelar y pronosticar, debido a los fenómenos ambientales y variaciones aleatorias en el espacio - tiempo (Htike & Khalifa, 2010).

Diferentes métodos de *Machine Learning* e inteligencia artificial tales como Redes Neuronales Artificiales, ARIMA, SARIMA y SARIMAX son algunos de los sistemas más utilizados para la predicción de lluvias, y de los cuales muchos analistas en el tema u operarios se han beneficiado debido a sus resultados (Abhishek et al., 2012), (Gorlapalli et al., 2022). Sumado a esto, se han llevado a cabo diferentes estudios sobre la evolución espacial y temporal de la precipitación (Lu et al., 2019). Como ejemplo, algunos de los modelos más importantes es el modelo ARIMA combinando redes neuronales para predecir la lluvia por mes, además, se realizó el análisis de lluvia utilizando la estacionalidad en la serie de tiempo con el modelo SARIMA y el test de Dickey Fuller para determinar la estacionariedad (Jibril et al., 2017) y el modelo SARIMAX para predecir subseries utilizando el método de wavelet para obtener información sobre el tiempo y la frecuencia de la señal (Farajzadeh & Alizadeh, 2018).

Este trabajo contribuye al pronóstico de la precipitación en la ciudad de Manizales, cuyos resultados, se utilizan para construir un sistema de detección temprana de desastres naturales, específicamente, para deslizamientos de tierra. Además, es importante evitar que los individuos que viven en zonas alto riesgo, se vean muy afectados por las fuertes lluvias que se presentan durante todo el año. Por esta razón, se realiza un aporte en el estudio de series de tiempo utilizando los modelos ARIMA, SARIMA, SARIMAX, Prophet y Neural Prophet, con los datos climáticos proporcionados por el SIMAC. Finalmente, se destaca que no se encuentran muchas investigaciones en esta región en particular, lo cual abre las puertas a nuevos resultados que pueden ser la base para mejorar las predicciones de la precipitación.

2. Objetivos

En el contexto de este trabajo, para la predicción de lluvia en la ciudad de Manizales, Colombia, se evaluaron diferentes modelos estadísticos y probabilísticos. En concreto los objetivos establecidos para llevar a cabo el trabajo de fin de master se exponen a continuación.

Objetivo General

Analizar series temporales para la predicción de lluvias en 13 estaciones meteorológicas de la ciudad de Manizales - Colombia utilizando ARIMA, SARIMA, SARIMAX, Prophet y Neural Prophet.

- ***Objetivos específicos***

Para llevar a cabo el objetivo general, se proponen los siguientes objetivos específicos que lo complementan:

1. Extraer, transformar y cargar los datos meteorológicos.
2. Definir una estrategia de imputación de los datos faltantes y atípicos.
3. Implementar los modelos: ARIMA, SARIMA, SARIMAX, Prophet, Neural Prophet.
4. Obtener métricas de evaluación para definir el mejor modelo de serie de tiempo.
5. Llevar a cabo el funcionamiento del análisis de prevención de riesgo usando el modelo óptimo de acuerdo con los resultados.

3. Revisión de Literatura

Predecir es una manera de realizar una estimación sobre los posibles valores futuros. Por otro lado, una serie de tiempo es un conjunto de observaciones medidas secuencialmente en el tiempo. Por lo general, la información recolectada sobre series de tiempo es usada para hacer predicciones. Adicionalmente, diferentes espacios de tiempo pueden ser predichos tanto para corto, medio y largo plazo y dependiendo de la cantidad de información que se adquiera, se escogerá la opción más adecuada de acuerdo con las necesidades del investigador. En la literatura se encuentran investigaciones sobre series de tiempo que se han enfocado en temas sobre el clima y los datos meteorológicos (Talwar, 1990).

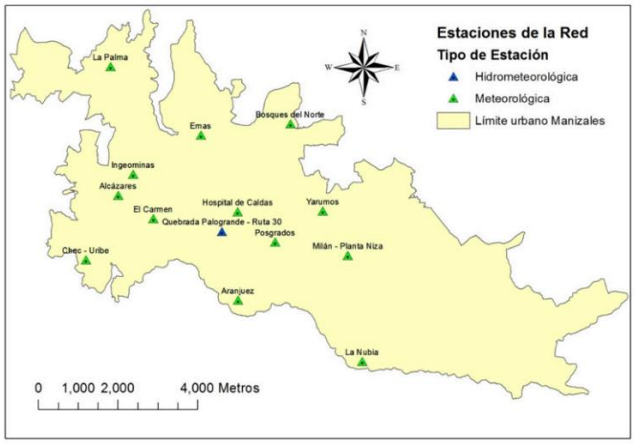
Con el fin de identificar los patrones que se encuentran en los datos recolectados para el proyecto, es conveniente considerar series temporales en sus 4 componentes: tendencia, ciclo, variaciones estacionales y las fluctuaciones. Por consiguiente, la tendencia es un movimiento de manera ascendente y descendente que caracteriza una serie de tiempo durante un periodo de tiempo. Por otro lado, el ciclo se refiere a las fluctuaciones recurrentes alrededor de niveles de tendencia medidos de pico a pico. Agregando a lo anterior, la variación estacional es la repetición de patrones dentro de cierto periodo, por ende, un ejemplo de series de tiempo que presentan variaciones estacionales, son las lecturas de la precipitación y la temperatura (Fox et al., 1973). Para finalizar, la precipitación en la ciudad de Manizales posee un comportamiento estacional, debido a la naturaleza del clima, que depende a su vez de otras variables como la temperatura, la presión, la humedad, entre otras (Vélez Upegui et al., 2015).

La predicción de lluvia no es un trabajo fácil, en especial cuando se espera el valor exacto y preciso de la cantidad estimada, ya que un valor fuera de rango podría cambiar toda la interpretación de análisis de riesgo. Por otra parte, estas predicciones se utilizan para proteger la agricultura, la producción de frutas, prevenir inundaciones, y deslizamientos de tierra, en relación con la cantidad de lluvia por horas, días o semanas (De Lima & Guedes, 2015). Adicionalmente, se utilizan varios algoritmos y resultados finales para ser entregados a los operarios o analistas para completar la labor de riesgo o la toma de decisiones. Por esta razón, lo mencionado anteriormente se vuelve un trabajo necesario en lugares donde las áreas de lluvia son muy frecuentes y se esperan con mayor periodicidad (Biswas et al., 2016). De la misma manera, la predicción de la precipitación acumulada de los 25 días es esencial para la toma de decisiones, con el fin de emitir las alarmas en la ciudad de Manizales, y con ayuda de una red de monitoreo de las variables climáticas en tiempo real, se trata de reducir el impacto en la población más vulnerable que se vean afectadas a los riesgos por deslizamientos. Por consiguiente, la lluvia durante los últimos 25 días (A25) hace referencia a el nivel de peligro a deslizamientos profundos, donde si se supera el umbral se lanza una alarma a la oficina Municipal de Prevención y Atención de Desastres (van Westen & Erlien, 1995).

Las fuertes precipitaciones pueden llegar a causar daño a los seres humanos y la cantidad excesiva de lluvia puede llegar a ocasionar desastres como inundaciones, que a su vez conlleva a pérdidas económicas. Además, es importante mejorar el conocimiento sobre la capacidad de pronóstico de la lluvia y las alertas tempranas, para reducir los impactos (Suhaila et al., 2010).

La alta intensidad de las precipitaciones ha sido identificada como el principal factor relacionado con la ocurrencia de deslizamientos, avalanchas e inundaciones en algunas regiones de Colombia (Caldas - Manizales) (Betancourt Mesa, 2009). Por otro lado, los eventos de lluvia se consideran fuertes o extremos cuando la cantidad de precipitación cae por encima de cierto umbral o percentil. Adicionalmente, el SIMAC y el IDEAM describen los casos de precipitación extrema diaria como la cantidad de al menos 80 milímetros (mm) en los reportes anuales oficiales, para las 14 estaciones meteorológicas (ver Ilustración 2) (IDEAM, 2020). Además, en los meses de lluvia, la precipitación alcanza valores entre 100 y 500 mm durante 21 de marzo y el 22 de septiembre; Mientras que en los meses de poca lluvia, la precipitación oscila entre 50 y 300 mm, durante el 22 de Junio y 21 de Diciembre (Delgado et al., 2020).

Ilustración 2. Localizaciones estaciones de la Red Meteorológica de Manizales – Colombia (Pachón Gómez et al., 2018b).



Los deslizamientos de tierra son un peligro natural que es causado posiblemente por las fuertes lluvias. Así mismo, el predecir la precipitación podría ayudar a combatir la devastación por el deslizamiento de tierra en las regiones más vulnerables. Por último, las series de tiempo, el *Machine Learning* y la inteligencia artificial, se utilizan como las técnicas que permiten a los científicos realizar modelos para aprender de los datos meteorológicos históricos, con el fin de generar predicciones (Schmidt et al., 2008).

Los centros hidrometeorológicos y de investigación de las universidades, han luchado fuertemente para producir la predicción de lluvia más competitiva y precisa para superar los problemas que la lluvia puede causar (Castillo Ruales et al., 2020), (Frame et al., 2022). Por otro lado, los esfuerzos han marcado una mejora considerable en la predicción de lluvia y los datos de pronóstico para muchos algoritmos que utilizan los modelos de series de tiempo y redes neuronales (Hernández et al., 2016).

Las series de tiempo son unas de las herramientas más importantes en el campo de la meteorología, varios métodos han sido implementados para predecir la precipitación, o la temperatura. También, las pequeñas variaciones que se presentan en la serie de tiempo pueden ser estudiadas por la aproximación autorregresiva (AR) y el promedio móvil (MA). Además, el modelo más significativo se conoce como la aproximación de Box – Jenkins, o también conocida como el modelo autorregresivo integrado de media móvil (ARIMA), el cual ha sido utilizado para simular variables hidrológicas y meteorológicas en todo el mundo. Por otro lado, si el modelo presenta una componente estacional, ARIMA puede expandirse e incluir una componente conocida como *Seasonal* ARIMA, SARIMA. Adicionalmente, si el modelo tiene variables exógenas, es posible agregar una nueva componente conocida como *Seasonal* ARIMA *Exogenous,* SARIMAX (Sawsan M, 2013) (Dimri et al., 2020). De modo que, este modelo de probabilidad hace parte de un proceso estocástico que describe el fenómeno estadístico que evoluciona en el tiempo. De esta manera, este proceso involucra elementos aleatorios que son secuenciales en el tiempo y se definen en un conjunto de puntos de tiempo que pueden ser continuos o discretos (Chatfield, 1999).

En las últimas décadas se ha realizado un amplio uso del modelo ARIMA, SARIMA, SARIMAX para comprender las variables climáticas (Precipitación, Temperatura) (Etuk & Mohamed, 2014), (Tularam & Ilahee, 2010), (Sarraf et al., 2011), (Bari et al., 2015), (Murat et al., 2018). Primero, Zakaria utilizó un modelo ARIMA utilizando datos sobre precipitación semanal, donde encontró una tendencia decreciente para el distrito semiárido de Sinjar en Irak (Zakaria et al., 2012). Segundo, Wang realizó una investigación basada en el modelo SARIMA para el riego agrícola y encontró que tiene un buen grado de adaptación en la toma de decisiones (Wang et al., 2013). Finalmente, Dimri y Ahmad, desarrollaron un modelo ARIMA, SARIMA para estudiar la precipitación y la temperatura haciendo los datos estacionarios y removiendo la estacionalidad (Dimri et al., 2020).

Prophet es un modelo OPEN-SOURCE de series de tiempo creado por los científicos de datos de Facebook, el cual ha sido implementado para la investigación sobre datos meteorológicos, enfocado a la predicción y el análisis de la temperatura, la radiación solar y la precipitación. Agregando a lo anterior, Hossain (2021), basa una investigación en predicción de lluvia mensual, sobre la media de conjunto multi-modelo de ensemble (MMEM) como regresor adicional, y utilizando los datos observados con todos los modelos individuales como regresor adicional (Hossain et al., 2021). También, redacta un artículo sobre la predicción de la precipitación utilizando los datos observados (recolectados de Australia) para ser combinados con los datos descargados del portal CMIP5, utilizando la librería de Prophet y algoritmos de Machine Learning (Hossain et al., 2021). Por último, Asha y Rishidas (2020), implementaron algoritmos de Prophet y Random Forest, empleando el parámetro estacionalidad anual, el periodo y un orden de Fourier, para la predicción de la

temperatura en Kerala – India. Por este motivo, su rendimiento basado en cinco estaciones diferentes, se compara en función de la precisión y el error absoluto medio, donde los resultados obtenidos fueron similares (Asha et al., 2020).

Una red neuronal artificial (ANN), se considera como una imitación del sistema neuronal humano. Por otro lado, la capacidad de aprender de los humanos para aprender de los ejemplos, motivó a McCulloch-Pitts a proponer la primera estructura de una red neuronal en 1943. Desde entonces, las ANN han sido exploradas progresivamente en el tiempo por diferentes investigadores del mundo, con nuevas mejoras en su arquitectura, algoritmos de aprendizaje y otras mejoras (G. Zhang et al., 1998) (Hung et al., 2009), (Asha et al., 2020).

En la actualidad, existen muchas variantes de ANN con un rendimiento mejorado en diferentes áreas de aplicación. El tipo más popular de las redes neuronales es el perceptrón multicapa de retropropagación (BPMLP). Así mismo, el científico Zhang resumió una ANN como un algoritmo que puede adaptarse a sí mismo una vez que los datos están disponibles, generalizar de manera efectiva a partir de datos de muestra y realizar modelos no lineales. Por otra parte, la capacidad de una ANN para modelar patrones en los datos la hace adecuada para datos de estacionalidad como la precipitación (G. Zhang et al., 1998).

En el mismo orden de ideas, el autor Singh, presenta un modelo para pronosticar las precipitaciones en la India basado en la estacionalidad y en escala mensual. De modo que, el artículo recomienda usar tres técnicas diferentes: Conjunto *Fuzzy*, la Entropía y las redes neuronales. El modelo de Fuzzy se usa para manejar las incertidumbres que se heredan en el conjunto de datos. Por otro lado, el concepto de entropía computacional se modificó en el modelo y se utilizó para proporcionar la entrada como el grado de pertenencia en la función de entropía. Por último, se utilizan las redes neuronales artificiales para desfuzzificar la función de entropía (Singh, 2018).

Los autores Cramer y kampouridis, comparan el rendimiento predictivo del método “Cadena de Márkov” extendida con la predicción de lluvia (en 20 ciudades de Europa y 22 de EEUU), con otras técnicas de aprendizaje como redes neuronales de base radial, regresión de vectores de soporte y k-vecinos. Así mismo, se reportó una relación entre la precisión y los atributos climáticos como la naturaleza volátil de la lluvia, la cantidad de lluvia máxima y el rango intercuartílico de lluvia (Cramer et al., 2017).

Otros artículos, reportan estudios sobre la predicción y detección de anomalías de lluvia usando la red neuronal envolvente, la selección de características por red auto encoder y la asignación de tareas de clasificación y predicción para la red de perceptrón multicapa, utilizando los datos del IDEAM – Manizales (Universidad Nacional de Colombia) (Gunawansyah et al., 2017), (Hernández et al., 2016). Haidar y Verma, presentan un enfoque basado en algoritmo genético para identificar la mejor combinación de características de entrada y parámetros de red neuronal para obtener el resultado más preciso (Haidar & Verma, 2018).

En la literatura, se encuentran escasos artículos de predicción de precipitación y temperatura para el departamento de Caldas y la ciudad de Manizales, utilizando diferentes métodos como redes neuronales, el método de descomposición wavelet, análisis de Fourier, entre otros. Adicionalmente, los investigadores han desarrollado una arquitectura basada en Deep Learning utilizando autoencoders y ANN, para predecir la lluvia acumulada diaria en una estación meteorológica. Se han realizado predicciones empleando la aproximación Naive, con ANN y la red neurodifusa (ANFIS) (Hatim et al., 2020), (Correa Ortiz et al., 2021), (Hernández et al., 2016), (Moreno Cadavid et al., 2016). A diferencia de estos estudios mencionados anteriormente, este trabajo se basa en modelos estadísticos y la implementación de la librería Prophet para predecir la precipitación.

4. Marco teórico

En este capítulo, se exponen algunos conceptos básicos de los modelos estadísticos y probabilísticos, en especial de los modelos ARIMA, SARIMA, SARIMAX, Prophet y Neural Prophet, para la predicción de la precipitación (mm) en las 13 estaciones de la ciudad de Manizales, Colombia, el cual es el tema central de este trabajo de investigación. Adicionalmente, se realizó una descripción de algunos métodos de imputación de datos y el software implementado de interés para cumplir con los objetivos planteados.

4.1. Introducción al Machine Learning

El aprendizaje automático (ML), ha buscado desarrollar e implementar sistemas informáticos que mejoren automáticamente su rendimiento a través de la experiencia de manera continua. Así mismo, el objetivo final se basa en desarrollar aplicaciones con amplias capacidades de aprendizaje, utilizando algoritmos, la estadística y matemática de manera más compleja y sólida. Adicionalmente, si la investigación tiene éxito puede estar contribuyendo en sistemas más sofisticados como robots que aprenden a operar en nuevos entornos, inclusive sistemas de comprensión del habla que se adaptarían a la dificultad de las condiciones ambientales. También, se crean nuevos sistemas que ayudan al ser humano a resolver y contribuir en la solución de nuevos problemas del medio ambiente, de física o de cálculo, inclusive encontrar la cura para diferentes enfermedades y evitar grandes pandemias que afectan al mundo (Mitchell et al., 2017).

No obstante, las aplicaciones del ML en el mundo real son más complicadas dado que demandan una alta cantidad de información de calidad, que requieren validación y la generación de características a partir de múltiples fuentes de entrada. Adicionalmente, se pueden crear conjuntos de diferentes modelos y con frecuencia apuntan a métricas comerciales que son difíciles de optimizar. Por otro lado, un problema importante es el comportamiento de los modelos de ML que dependen de los datos obtenidos por el investigador, que pueden cambiar debido al diferente comportamiento de cada

usuario, además de encontrarse errores en los pipelines. Finalmente, muchas decisiones dependen de la comprensión profunda de cada algoritmo de ML, y las consecuencias del sistema correspondiente (Schelter et al., 2018).

4.2. Predicción de Series de tiempo

Las series de tiempo son observaciones x_t ordenadas cronológicamente en un tiempo especifico t. Por otro lado, si el conjunto de pasos de tiempo es T donde $t \in T$ es discreto, a esto se le conoce como serie de tiempo discreta, mientras que, si las observaciones se registran continuamente durante algún intervalo de tiempo, la serie de tiempo es continua (Brockwell & Davis, 2016). De modo que, los objetivos principales del análisis de serie de tiempo suelen ser la construcción de un modelo y el poder ajustarlo a las observaciones para estudiar la dependencia en los datos. Por esta razón, se requiere comprender el mecanismo de cómo se generan las observaciones, encontrar patrones y predecir el desarrollo de las variables observadas.

Las series de tiempo se dividen en componentes que representan el tipo de patrón de tendencia, estacionalidad, ciclo y componente restante.

T_t: *Tendencia – el incremento o decremento del valor.*

S_t: *Estacionalidad - Cíclica repetitiva a corto plazo con frecuencia conocida.*

C_t: *Los ciclos también se repiten, su frecuencia no es precisa y su duración es mayor a dos años.*

R_t: *La parte del residuo captura todo lo demás.*

Si se asume una descomposición aditiva, la serie de tiempo obtenida se muestra en la Ecuación 1:

$$y_t = T_t + S_t + C_t + R_t \qquad \textit{Ecuación 1}$$

Si la variación alrededor de la tendencia o magnitud de las fluctuaciones estacionales no difieren desde el nivel de la serie de tiempo, es recomendable utilizar la descomposición aditiva. De lo contrario, la descomposición multiplicativa, que se muestra en la Ecuación 2, es más apropiada.

$$y_t = T_t * S_t * C_t * R_t \qquad \textit{Ecuación 2}$$

En varios métodos de descomposición, los ciclos pueden ser combinados con una tendencia (Hyndman & Athanasopoulos, 2018).

4.2.1. Proceso Autorregresivo (AR)

El modelo autorregresivo (AR) es una variable dependiente del retraso del tiempo (*lagged*), el cual contiene un término autorregresivo. Así mismo, AR forma parte de una serie de tiempo y_t, que contiene un valor que depende de una agrupación lineal del valor anterior, que define retrasos máximos (p), además, tiene un término de error arbitrario ϵ_t. Finalmente, la expresión se muestra en la Ecuación (3) a (5) de la siguiente manera:

AR 1er orden: $\widehat{y_t} = \alpha + b_1 Y_{t-1}$ *Ecuación 3*

AR 2do orden: $\widehat{y_t} = \alpha + b_1 Y_{t-1} + b_2 Y_{t-2}$ *Ecuación 4*

AR 3er orden: $\widehat{y_t} = \alpha + b_1 Y_{t-1} + b_2 Y_{t-2} + b_3 Y_{t-3}$ *Ecuación 5*

De manera general se obtiene:

$$\widehat{y_t} = \alpha + b_1 Y_{t-1} + b_2 Y_{t-2} + b_3 Y_{t-3} + \cdots + b_q Y_{t-p} + \epsilon_t$$ *Ecuación 6*

ϵ_t es un proceso puramente aleatorio, $\widehat{y_t}$ es el proceso autorregresivo de orden p, Y_{t-1} es la regresión de los valores del pasado, α es el término del intercepto y b_q es el coeficiente del retraso que el modelo estima. En este caso, ϵ_t no se explica por los valores pasados, lo que significa que es independiente de Y_{t-1}, Y_{t-2} ....Entonces $\widehat{y_t}$ se llama un proceso autorregresivo (Chatfield, 1999).

4.2.2. Promedio Móvil

El promedio móvil de la serie de tiempo y_t, que se muestra en las ecuaciones (7) a (10), es el valor observado en términos del error aleatorio y de algunas agrupaciones lineales de los términos de los errores arbitrarios previos, hasta un retraso máximo descrito (q):

MA 1er orden: $\widehat{y_t} = \gamma + d_0 u_t + d_1 u_{t-1}$ Ecuación 7

MA 2do orden: $\widehat{y_t} = \gamma + d_0 u_t + d_1 u_{t-1} + d_2 u_{t-2}$ Ecuación 8

MA 3er orden: $\widehat{y_t} = \gamma + d_0 u_t + d_1 u_{t-1} + d_2 u_{t-2} + d_3 u_{t-3}$ Ecuación 9

De manera general se obtiene:

$$\widehat{y_t} = e_t + d_1 u_{t-1} + d_2 u_{t-2} + d_3 u_{t-3} + \cdots + d_q u_{t-q} = \sum_{j=0}^{q} d_j \, u_{t-j}$$ *Ecuación 10*

u_{t-j} es ruido aleatorio, d_j son constantes, e_t se conoce como el ruido blando con media cero y varianza σ_e^2, y y_t, se conoce como el modelo de promedio móvil con orden q (Chatfield, 1999).

4.2.3. Modelo ARIMA

El promedio móvil integrado autorregresivo, también denominado ARIMA, es una Ecuación de pronóstico que puede hacer que las series de tiempo sean estacionarias usando la diferenciación cuando sea necesario. Por esta razón, una serie de tiempo que debe diferenciarse para ser estacionaria es una serie de tiempo integrada (d). Por otro lado, los retrasos de la serie estacionaria se clasifican como autorregresivos (p), el cual designa en el término (AR), y los retrasos de los errores de pronóstico se clasifican como medías móviles (q), que se identifica como (MA). Por último, a un modelo no estacional ARIMA se le conoce comúnmente como ARIMA (p,d,q), en el que:

- p es el número de valores autorregresivos.
- d es el número de diferenciación no estacional necesario para convertir la serie en estacionaria.
- q es el número del error en el pronóstico con retraso en la Ecuación de predicción.

Usualmente, se toma el valor de d = 1 y al menos d = 2.

Si es estacionaria después de diferencias d veces con $W_t = \Delta^d y_t$, donde W_t sigue el proceso ARMA.

Ahora consideremos ARIMA (p,1,q) que se muestra en la Ecuación 11:

$$\widehat{y_t} = b_1 W_{t-1} + b_2 W_{t-2} + b_3 W_{t-3} + \cdots + b_q W_{t-p}$$

$$+ e_t + d_1 u_{t-1} + d_2 u_{t-2} + \cdots + d_q u_{t-q} \qquad \textit{Ecuación 11}$$

Otro detalle importante, toma en consideración los residuos del modelo, los cuales se examinan para determinar si pertenecen a ruido blanco o no. De modo que, si pertenece a ruido blanco, el mejor modelo es probablemente un buen acercamiento al proceso estocástico, es decir que de cierta manera es predecible, pero si no lo son, el proceso se empieza de nuevo. De esta manera se entiende que el método de Box-Jenkins es repetitivo (Corlett & Aigner, 1972). Agregando a lo anterior, el método Box-Jenkins consiste en 4 pasos:

Paso 1. Identificar: Buscar los valores óptimos de p, d y q.

Paso 2. Estimar: Después de encontrar los valores de p, d, y q, se deberá estimar los parámetros de la autoregresión y media móvil del modelo.

Paso 3. Diagnosticar. Una vez elegido el modelo ARIMA en específico y después de estimar sus parámetros, será necesario verificar si el modelo elegido se ajusta razonablemente bien al comportamiento de los datos, ya que es posible que otro modelo ARIMA también pueda hacer su trabajo. Ahora bien, una prueba simple del modelo elegido es ver si los residuos estimados del modelo son ruido blanco. Por

último, en el caso de si pertenecer al ruido blanco se puede aceptar el ajuste determinado, si no, será necesario empezar de nuevo.

Paso 4. Predicción: En varias ocasiones las predicciones que se obtienen por estos métodos son más fiables que las obtenidas por las predicciones a corto plazo. Cada caso debe ser revisado (Corlett & Aigner, 1972).

Uno de los métodos para evaluar el modelo es el de criterio de información Akaike (AIC), el cual puede inscribirse matemáticamente como se muestra en la Ecuación 12:

$$ln(AIC) = \left(\frac{2K}{N}\right) + ln\left(\frac{RSS}{n}\right)$$ *Ecuación 12*

Donde K es el número de regresiones, n es el número de observaciones y RSS es la suma de los cuadrados del residuo, que es una medida de la discrepancia entre los datos y la estimación del modelo. No obstante, la intención de comprar dos o más modelos es considerar el modelo con el valor de AIC más bajo. Además, el método penaliza no solo a los modelos con peores predicciones (subajuste), sino también a los modelos con un gran número de parámetros (sobreajuste). Finalmente, la bondad de los parámetros se calcula mediante el algoritmo de verosimilitud (Zadranska, 2019).

4.2.4. Modelo SARIMA

ARIMA estacional (SARIMA), es una técnica de ARIMA, donde la componente estacional se puede majear en datos de series temporales invariadas. Se adjuntan tres nuevos hiperparámetros para establecer AR(p), I(d), y MA(q) para la componente de estacionalidad. En este caso, se combina tanto la componente no estacional y estacional en un modelo multiplicativo. La notación está definida como (Sulaiman, 2015) (Stitou, 2019):

$$\text{ARIMA (p,d,q) X (PDQ)}_m$$

Donde m es el número de observaciones por año. (P,D,Q) es la componente estacional. Hay 4 componentes estacionales que no hacen parte del modelo ARIMA que se requieren configurar:

P: Orden autorregresivo estacional

D: Orden diferencial estacional

Q: Orden estacional de la media móvil

M: Orden de marca de tiempo para una sola estación.

El modelo SARIMA puede representarse por la Ecuación 13:

$$\Phi_p(B)^s\phi_p(B)(1-B)^d(1-B^s)^D Y_t = \theta_q(B)\Theta_q(B)^s a_t$$ *Ecuación 13*

Donde:

ϕ_p : Operador regular autorregresivo de orden p.

θ_q : operador regular de la media móvil de orden q.

d : Diferenciación de orden d.

Φ_p : Operador estacional autorregresivo de orden P.

Θ_q : Operador regular del promedio móvil de orden Q.

D : Diferenciación de orden D.

s : Periodo estacional.

a_t : Proceso de ruido blanco.

B : Operador de retroceso.

4.2.5. Modelo SARIMAX

El modelo SARIMAX es un modelo SARIMA con variables de influencia externa, denominado SARIMAX (p,d,q) $(P,D,Q)_m(E)$, donde E es el vector de las variables exógenas. Adicionalmente, las variables exógenas modeladas probablemente por la Ecuación 14 de regresión multilineal se obtiene de la siguiente manera:

$$\Phi_p(B)^s\phi_p(B^s)(\nabla)^D\nabla_s^D Y_t = \beta_k\chi_{k,t} + \theta_q(B)\Theta_Q(B^s)\varepsilon_t) \quad \textit{Ecuación 14}$$

Donde $\chi_{k,t}$ es el vector que incluye la K-esima variables de entrada explicativas en el tiempo y β_k es el valor del coeficiente de la k-esima variable exógena de entrada (Vagropoulos et al., 2016).

4.2.6. Facebook Prophet

Prophet es una librería de uso libre creada en Facebook por Sean J. Taylor y Ben Letham, y que es utilizada para pronosticar series temporales. De esta manera, el fin era superar dos problemas que se encuentran a menudo con otras metodologías de predicción: las herramientas automáticas tendían a ser demasiado inflexibles e incapaces de adaptarse a suposiciones adicionales, y el pronóstico más sólido requería de un analista experimentado con habilidades especializadas en ciencia de datos. Por otro lado, debido a la alta demanda de pronósticos comerciales de alta calidad, Facebook lanzo al público Prophet con código abierto.

Prophet fue diseñado para manejar de manera óptima las tareas de pronóstico de negocios, que normalmente presentan cualquier atributo de datos de tiempo capturados a nivel de hora, días o semana; fuertes efectos de estacionalidad que ocurren diaria, semanal o anualmente; días festivos y otros eventos especiales únicos que no necesariamente siguen los patrones de estacionalidad pero que ocurren de manera irregular; y cambios de tendencia significativos que pueden ocurrir con el lanzamiento de nuevas características o productos (Rafferty & Safari, 2021).

Prophet está conformado por tres componentes principales: la tendencia $g(t)$, la estacionalidad $s(t)$, los festivos $h(t)$ y el error ε_t. Se describen por la Ecuación 15 que se muestra a continuación:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \quad \textit{Ecuación 15}$$

Donde $g(t)$ es la tendencia (Lineal o Logística) de la función, $s(t)$ representa la estacionalidad que puede ser semanal o anual, $h(t)$ es el efecto de los festivos el cual ocurren en horarios altamente irregulares durante uno o varios días, y ϵ_t es el error de cualquier cambio que no se adapte al modelo, es típicamente modelado como un ruido con distribución normal (OO & PHYU, 2020).

La estacionalidad se define matemáticamente de la siguiente manera, por la Ecuación 16:

$$s(t) = \sum_{n=1}^{N}(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n sin\,(\frac{2\pi nt}{P})) \quad \textit{Ecuación 16}$$

Donde P es el periodo que tendrá la serie de tiempo, por ejemplo $P = 365$ para datos anuales, o $P = 7$, para la información semanal, o escalamos la variable de tiempo en días, u horas, n es el número de observaciones hasta N (Taylor & Letham, 2017).

Para ajustar la estacionalidad se requieren estimar los parámetros $2N$, siendo $\beta = [a_1, b_1, \dots, a_n, b_n]^T$. Esto se hace mediando la construcción de una matriz de vectores de estacionalidad para cada valor de t en los datos históricos y futuros, por ejemplo, con estacionalidad anual y $N = 10$, se obtiene lo observado en la Ecuación 17:

$$X(t) = \left[\cos\left(\frac{2\pi(1)t}{365.25}\right), \dots, \sin\left(\frac{2\pi(10)t}{365.25}\right)\right] \quad \textit{Ecuación 17}$$

La componente estacional se expresa por la siguiente Ecuación 18:

$$s(t) = X(t)\beta \textit{ Ecuación 18}$$

La incorporación de los días feriados en el modelo se puede simplificar suponiendo que los efectos son independientes. Para cada día festivo i, sea D_i el conjunto de fechas anteriores y próximas de ese festivo, se añade una función que indica si el tiempo t ocurre durante el periodo del festivo i, y se asigna a cada feriado un parámetro k_i, que es el cambio correspondiente en el pronóstico. De esta manera, se genera una matriz de regresores similar a la estacionalidad, que se muestra en la Ecuación 19 y 20 :

$$Z(t) = [1(t\epsilon D_1), \dots, 1(t\epsilon D_L)] \quad \textit{Ecuación 19}$$

Tomando,

$$h(t) = Z(t)\kappa \quad \textit{Ecuación 20}$$

Al igual que la estacionalidad, se usa un $\kappa \sim Normal(0, \sigma^2)$, (Taylor & Letham, 2017).

4.2.7. Neural Prophet

Los científicos de datos de Facebook, desarrollan un nuevo modelo conocido como Neural Prophet para series de tiempo basado en redes neuronales. Por otro lado, algunas diferencias respecto a Prophet son que Neural Prophet permite utilizar el método de gradiente descendente para optimización, modela la autocorrelación de seris de tiempo usando redes autorregresivas (AR-Net), modela regresores usando una red neuronal, y se puede ajustar los horizontes específicos de pronóstico. Sin embargo, una desventaja es que se pierde la ecuación modelo de Prophet, por lo cual se convierte en una caja negra. No obstante, el trabajo que hacen las redes neuronales es aproximar el comportamiento observado de las funciones, pero sin una ecuación base para el modelo. Por último, En el año 2021 Triebe y colaboradores proponen un modelo solo dependiente del tiempo, el cual pude ser expresado matemáticamente de manera equivalente, teniendo en cuenta la predicción de un paso con un horizonte $h = 1$, la cual se muestra en la Ecuación 21:

$$\hat{y}_t = T(t) + S(t) + E(t) + F(t) + A(t) + L(t) \textit{ Ecuación 21}$$

Donde:

$T(t)$ = Tendencia en un tiempo t.

$S(t)$ = Efectos de la estacionalidad en un tiempo t.

$E(t)$ = Efectos de los eventos y festivos en un tiempo t.

$F(t)$ = Efecto de la regresión en un tiempo t para futuras variables conocidas exógenas.

$A(t)$ = Efectos de la autoregresión en un tiempo t basado en observaciones del pasado.

$L(t)$ = Efectos de la regresión en un tiempo t para observaciones con retraso de las variables exógenas.

Todos los módulos de las componentes del modelo se pueden configurar y combinar individualmente para componer el modelo completo. Por defecto, solo están activados los módulos de tendencia y estacionalidad (Triebe et al., 2021).

4.2.8. Métricas de Evaluación

Con el fin de evaluar la eficiencia de los modelos de series de tiempo, es necesario considerar el tipo de métrica para diferenciar el mejor modelo del peor modelo. La función de pérdida son 3:

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

$$MAE = \frac{1}{T}\sum_{i=1}^{T}(y_i - \hat{y}_i)^2$$

$$MSE = \frac{1}{T}\sum_{i=1}^{T}(y_i - \hat{y}_i)^2$$

Donde T es el tamaño de la muestra (Jaramillo Muñoz, 2021).

4.2.9. Imputación de Datos

Los datos faltantes son un problema que se encuentran comúnmente en la recopilación de datos, en especial cuando se obtiene gran volumen de información en el mundo real. Adicionalmente, los datos faltantes impactan en los sistemas de soporte y la realización de métodos de evaluación estadísticos, y modelos de Machine Learning. Por otro lado, los resultados pueden verse afectados mediante el uso de asignaciones arbitrarias o aleatorias a los elementos de datos faltantes (Markey et al., 2006). Además, en las diferentes encuestas que se realizan a la hora de recolectar información, es posible que haya problemas y se obtenga la recopilación incompleta de las variables y la falta de respuesta de las personas, encuestas mal definidas y eliminación de datos por razones como confidencialidad. Además, la selección de datos que son adquiridos por una máquina (Estación Meteorológica) que puede verse afectados debido a las condiciones climáticas que pueden dañar los dispositivos internos e inclusive cambiar el valor del dato en sí (Horton & Kleinman, 2007), (Ssali & Marwala, 2008).

Las técnicas de imputación se pueden dividir en procedimientos de enfoques basados en modelos y no basados en modelos. De esta manera, los procedimientos más comunes no basados en modelos incluyen reemplazar los valores faltantes con la media, e imputación *Hot-Deck* donde el valor que falta se reemplaza con un valor de otro caso similar para el cual ese valor está disponible (Lakshminarayan et al., 1999). Sin embargo, estos métodos tienden a atenuar las estimaciones de la varianza en los procedimientos estadísticos y, además, debido a que la media no es independiente de

otras observaciones en los datos, los analistas que utilizan la imputación de la media tiene menos grados de libertad de lo que se justifica (Donner & Rosner, 1982).

La imputación basada en modelos, es más flexible que los procedimientos anteriores y por lo general incluyen técnicas de regresión que estiman el valor faltante usando un modelo basado en una ecuación de regresión derivado de las observaciones completas observadas previamente (Fogarty, 2006).

MICE

Imputaciones múltiples por ecuaciones de cadena (MICE), consisten en modelos de regresión lineal específicos de variables, que primero se establecen con valores de imputación iniciales. Así mismo, los regresores no lineales basados en aprendizaje automático pueden reemplazar estas ecuaciones de regresión lineal para mejorar la precisión de las estimaciones de valores faltantes (Samad & Yin, 2019). Adicionalmente, se sabe que la parte MI del MICE captura la variabilidad entre modelos, mientras que el aprendizaje en ensemble modela la variabilidad dentro de un modelo (Cevallos Valdiviezo & Van Aelst, 2015). Ahora bien, un ejemplo de algoritmo MICE que utiliza un conjunto de árboles de decisión para la imputación de datos es MICEForest (Stekhoven & Bühlmann, 2012).

La actualización iterativa en MICEForest, se realiza mediante la igualdad de la media predictiva (PMM). En PMM, el valor faltante estimado del modelo se compara con otros valores observados en el conjunto de datos para determinar las muestras vecinas. Además, el valor observado del vecino más cercano se usa luego para determinar las muestras más cercanas. Así, el valor observado del vecino más cercano se usa para imputar el valor faltante. Por lo tanto, PMM realiza la imputación de un solo valor utilizando el valor observado del vecino más cercano en lugar de realizar múltiples imputaciones (Samad et al., 2022).

El proceso de imputación basado en MICE se puede resumir en los siguientes pasos:

Inicialización: Para una variable que contiene valores faltantes, estos se reemplazan con muestras aleatorias de los valores observados de esa variable.

Imputación: El proceso de imputación se realiza secuencialmente para las variables según su orden original en el conjunto de datos. De esta manera, la variable imputada se utiliza como respuesta para la construcción del modelo. Por otro lado, las observaciones en el conjunto de datos se dividen en dos partes según si la variable es observada o faltante en el conjunto de datos original. Finalmente, las observaciones analizadas se utilizan como conjunto de entrenamiento y las observaciones faltantes como conjunto de predicción.

Parar: Cuando se han imputado las variables con datos faltantes, se completa una iteración de imputación.

El proceso de imputación se itera hasta que se alcanza el número máximo de iteraciones (valor predeterminado es 5), y el resultado final es la última imputación (Hong & Lynn, 2020).

MissForest

El árbol de decisión es un algoritmo de Machine Learning de aprendizaje automático que ilustra todos los resultados concebibles y los caminos que conducen a esos resultados en forma de estructura de árbol. Adicionalmente, la imputación de valores perdidos usando este método se hace construyendo arboles de decisión para observar los valores perdidos de cada variable, y luego llena los valores perdidos usando su árbol correspondiente (Twala, 2009). La predicción de valores nulos se muestra luego en el nodo hoja. Además, este algoritmo puede manejar variables numéricas y categóricas. Por otro lado, una desventaja es que los árboles de decisión puede producir un árbol complejo que tiende a consumir mucho tiempo pero con un sesgo bajo (Rokach, 2016).

Otra manera muy conocida usando el enfoque de árbol de decisión es el algoritmo de Random Forest (Bosque Aleatorio), que es un conjunto de árboles de decisión a través de bagging que combina múltiples predictores aleatorios para agregar predicciones. Además, la regla de predicción se basa en el voto mayoritario o el promedio de todos los árboles. Por otro lado, los bosques pueden alcanzar fortalezas de predicción competitivas o incluso superiores en comparación con enfoques bien establecidos, como la regresión y las máquinas de soporte vectorial (Tang & Ishwaran, 2017).

El proceso de imputación, según lo explica Breiman, respecto a los valores faltantes con el Random Forest incluye los siguientes pasos (Breiman, 2001):

1. Seleccionar una muestra aleatoria de las observaciones con reemplazo.
2. Luego se selecciona al azar un con junto de variables.
3. Se elige una variable que proporcione la mejor división.
4. Se repite el paso de elegir una variable que produzca la mejor división hasta alcanzar la profundidad máxima.
5. Se repiten los pasos anteriores hasta alcanzar el número determinado de árboles.
6. Por último, se realiza una predicción del valor faltante por mayoría de votos.

Varios estudios se encuentran en la literatura, donde se utiliza Random Forest para tratar valores faltantes (Pantanowitz & Marwala, 2009), (Stekhoven & Bühlmann, 2012), (Hong & Lynn, 2020), (S. Zhang et al., 2021).

MissForest, es un método de imputación no paramétrico para básicamente cualquier tipo de dato. Además, puede hacer frente a variables de tipo mixto, relaciones no lineales, interacciones complejas y alta dimensionalidad ($p \gg n$). Igualmente, solo requiere de la observación, es decir de las filas del *Dataframe* proporcionadas a la

función y el algoritmo se basa en el modelo de *Random Forest*. Por último, para cada variable, MissForest ajusta un bosque aleatorio en la parte observada y luego predice la parte que falta. De esta manera, el algoritmo continúa repitiéndose los pasos hasta que se cumpla un criterio de parada o se alcanza el máximo de iteraciones especificado por el usuario (Stekhoven, 2012).

4.3. Herramientas de Software

Esta sección está dedicada a una vista general del lenguaje de programación, librerías y herramientas usadas para la implementación.

Python

Python es un lenguaje de programación interoperable y fácil de usar, con una amplia disponibilidad de librerías, módulos e incluso marcos completos de desarrollo de aplicaciones. Esta hace que Python sea una excelente herramienta para usar múltiples campos de estudio, específicamente en ciencia de datos y aprendizaje automático. La biblioteca de Pandas, y NumPy, admite matrices multidimensionales, agrega funciones matemáticas de alto nivel, además del manejo de estructura de datos que permite una fácil modificación, que son útiles para la ciencia de datos (*Welcome to Python.Org*, 2019), (*NumPy*, 2019), (*Pandas*, 2022).

Matplotlib

Es una librería de Python de uso libre usada para la visualización de los datos. Esta librería prevé un amplio rango de diferentes gráficos (*Matplotlib: Python Plotting*, 2022).

Seaborn

Seaborn es una librería basada en Matplotlib, usada para realizar visualizaciones interactivas para gráficos estadísticos (Snoek et al., 2012).

Scikit-learn

Scikit-learn es una librería de software de uso libre que provee herramientas para análisis de datos y diferentes clases para el procesamiento y problemas generales de aprendizaje automático, como clasificación, regresión, agrupamiento o reducción de dimensionalidad. Está construido sobre *NumPy* y *Matplotlib* (*Sci-Kit Learn: Machine Learning in Python*, 2022).

Pandas

Pandas es otra librería de uso libre para Python usada para el análisis de datos. Es bueno para importar la información. Su clase Dataframe es un excelente método para representar los datos de manera tabular, asistiendo en el procesamiento y modificación de los datos (*Python Data Analysis Library - Pandas*, 2022).

StatsModels

StatsModels es un módulo de Python que prevé clases y funciones para la estimación de diferentes modelos estadísticos, tanto para la realización de pruebas y datos (*StatsModels: Statistics in Python*, 2022).

SciPy

Script es otra librería de uso libre basada en Python, destinada para la ciencia y las matemáticas, por ejemplo: Prophet – para series de tiempo (*SciPy.Org*, n.d.).

GitHub

GitHub es una plataforma de alojamiento de código para el control y la colaboración desde cualquier parte del mundo (*GitHub*, 2022).

En este trabajo de fin de master se utilizó GitHub como repositorio para guardar el código y los resultados del proyecto. El enlace es el siguiente:

https://github.com/camilopulzara/TFM

5. Desarrollo del proyecto y resultados

En este capítulo se presenta la metodología, el planteamiento del problema y los resultados obtenidos en la investigación.

5.1. Metodología

El trabajo de fin de master corresponde a un documento de investigación, a partir de estudios que están relacionados con modelos estadísticos y probabilísticos de series de tiempo. Adicionalmente, se pronosticó la precipitación y se realizó un análisis de prevención temprana de riesgo por deslizamientos de tierra.

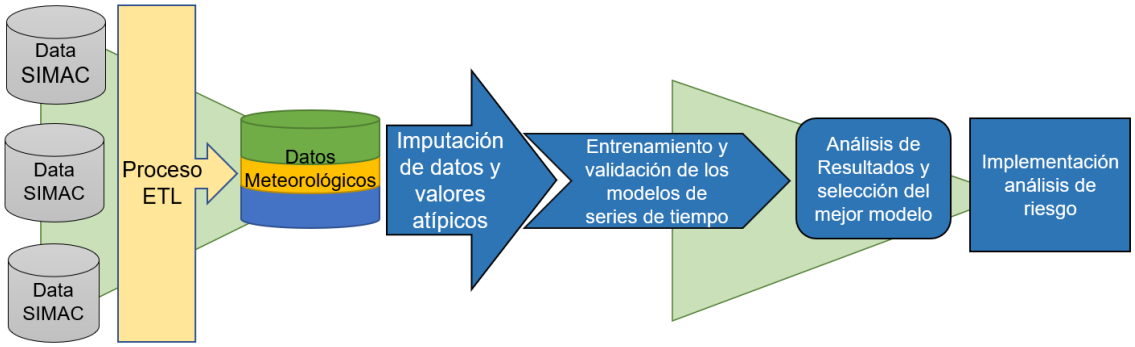
Para la elaboración de este documento fue necesario escribir un cronograma de actividades para cumplir con los objetivos planteados. Primero fue necesario encontrar la base de datos, la cual fue proporcionada por el SIMAC del municipio de Manizales. Segundo, se realizó una extracción, transformación y carga de los datos para ordenar las variables y modificar los valores que eran incorrectos. Tercero, se aplicó una estrategia para la imputación de los valores nulos y otra distinta para los *outliers*. Cuarto, se entrenaron y validaron los modelos ARIMA, SARIMA, SARIMAX, Prophet y Neural Prophet para predecir la precipitación, y de acuerdo con las métricas de evaluación establecidas en el apartado 4.2.8 se escogió el de menor valor. Quinto, se presentaron los resultados obtenidos y se implementó el análisis temprano de riesgo. Sexto, se redacta el documento final realizando una extensa revisión en la literatura sobre modelos de Machine Learning para la predicción de lluvias. Por último, En la Tabla 1 se observa el cronograma de actividades por fechas.

Tabla 1. Cronograma de Actividades

Actividades	Junio	Julio	Agosto
Obtención de base de datos			
Realizar ETL			
Implementar Minería de Datos			
Definir Estrategia de Imputación de Datos			
Modelos de Series de Tiempo			
Resultados			
Implementación análisis de riesgo			
Redacción del TFM			

Para entender mejor el desarrollo del proyecto, en la Ilustración 3 se muestra el diagrama de las distintas fases de la investigación.

Ilustración 3. Diagrama de las fases del proyecto.



5.2. Planteamiento del problema

¿Existen modelos de análisis de riesgo basados en predicción de datos climáticos en la ciudad de Manizales?

El crecimiento de la ciudad de Manizales ha generado un gran cambio en la estructura territorial, en el aumento de la población, migración de habitantes de la zona rural a la ciudad, generando de esta manera, un aumento exponencial en el terreno, lo que da como resultado la suburbanización. Por otro lado, la necesidad de tener un mayor espacio para construir viviendas, locales, centros comerciales trae como consecuencia el establecimiento de asentamientos de las personas inadecuados, debido a que en algunas ocasiones esto se realiza en lugares con condición de riesgo. De esta manera, se generan problemas en el suelo e incrementan la posibilidad de que se presenten desastres que afecten a la población y a la perdida de los hogares, vías, inclusive pérdidas económicas, entre otros.

El municipio de Manizales se encuentra construido sobra montañas que presentan características como pendientes muy altas, suelos susceptibles por las altas precipitaciones, movimientos de tierra, y vulnerabilidad por actividad sísmica, lo cual genera problemas en la construcción debido al poco espacio del terreno. Debido a esto, las organizaciones han decidido construir en zonas de manera vertical, sin tener en cuenta el riesgo de la urbanización, lo cual aumenta la vulnerabilidad a desastres naturales, como el deslizamiento de tierra o inundaciones.

En Manizales, el IDEAM y la Unidad Nacional para la Gestión de Riesgos de Desastres (UNGRD), han implementado un método para disponer de una alarma de prevención en caso de desastres naturales. Sin embargo, han sido pocas las investigaciones realizando predicciones y usando los datos de series de tiempo que se guardan en dichas instituciones. Modelos basados en redes neuronales han sido estudiados, pero no implementados en un proyecto basado en crear un sistema de riesgo.

Este trabajo se enfocó en estudiar los modelos ARIMA, SARIMA, SARIMAX, Prophet y Neural Prophet, para la predicción de la precipitación y el análisis de riesgo de desastres ante un deslizamiento de tierra en Manizales. Es bien sabido que, la naturaleza en cualquier parte del mundo cambia constantemente y se vuelve caótico, más a la hora de estimar la cantidad de lluvia, por lo cual hace que las predicciones sean menos seguras al incrementar el tiempo de pronóstico. De esta manera, se pretende mostrar una nueva perspectiva en el análisis de los datos meteorológicos y aprovechar la información histórica proporciona por el SIMAC, desde el año 2017 hasta el 2020, cada 5 minutos. Por último, es importante mencionar que los resultados se enviarán a una revista científica, a la alcaldía de Manizales y a la oficina de Gestión de Riesgo y Desastres de Manizales, para que puedan aprovechar todo el estudio en sus planes de Riesgo.

Por otro lado, se hace énfasis en que el trabajo se limita solo a estudiar algunas de muchas configuraciones en los modelos presentados y estrategias, tanto para la transformación de variables como en la imputación de datos faltantes.

5.3. Desarrollo del proyecto

5.3.1. Programa

El script principal está definido en Notebook.ipynb, todas las librerías fueron importadas y las más usadas se muestran a continuación:

```
import numpy as np
import pandas as pd
import statsmodels.api as sm
from fbprophet import Prophet
from neuralprophet import NeuralProphet
```

En el script fue necesario utilizar la librería de missingpy para imputar los datos, el cual se implementó de la siguiente manera:

```
import sklearn.neighbors._base
sys.modules['sklearn.neighbors.base'] = sklearn.neighbors._base
from missingpy import MissForest
```

Los datos, se dividen en conjuntos de entrenamiento y de testeo. De esta manera se realizó la predicción de los 7 días siguientes, para cada una de las estaciones meteorológicas.

```
train = dfs[(dfs.index.astype(str) <= fecha)]
test = dfs[(dfs.index.astype(str) >= fecha)]
```

Los hiperparámetros de los 5 modelos estadísticos implementados, fueron optimizados, testeando diferentes combinaciones de entrada tales como el p, d, q, y la estacionalidad de la serie. A continuación, se presenta el código de los mejores resultados para cada modelo.

ARIMA

```
model = auto_arima(train['Precipitacion'], start_p=0, start_q=0,
                   test='adf',
                   max_p=6, max_q=6, m=1,  d=0,
                   seasonal=True,
                   start_P=0,
                   D=None,
                   trace=True,
                   error_action='ignore',
                   suppress_warnings=True,
                   stepwise=True)
```

SARIMA

```
smodel = auto_arima(train['Precipitacion'], start_p=0, start_q=0,
                    test='adf',
                    max_p=5, max_q=5, m=12,
                    start_P=0, seasonal=False,
                    d=None, D=0, trace=True,
                    error_action='ignore',
                    suppress_warnings=True,
                    stepwise=True)
```

SARIMAX

```
sxmodel  = pm.auto_arima(train['Precipitacion'],
                     exog = train[['Temperatura','Presion','Velocidad','Humedad','Radiacion']],
                     start_p=0, start_q=0,
```

```
                test='adf',
                max_p=5, max_q=5, m=6,  start_P=0, seasonal=True,
                d=None, D=0, trace=True,
                error_action='ignore',
                suppress_warnings=True,
                stepwise=True)
```

Prophet

```
m = Prophet(interval_width=0.95, daily_seasonality=True, holidays=holidays)
```

Neural Prophet

```
m = NeuralProphet(yearly_seasonality=True,
                weekly_seasonality=True,
                daily_seasonality=True,
                num_hidden_layers=8,
               d_hidden=16,
               learning_rate=0.5, seasonality_mode="multiplicative", epochs=200)
```

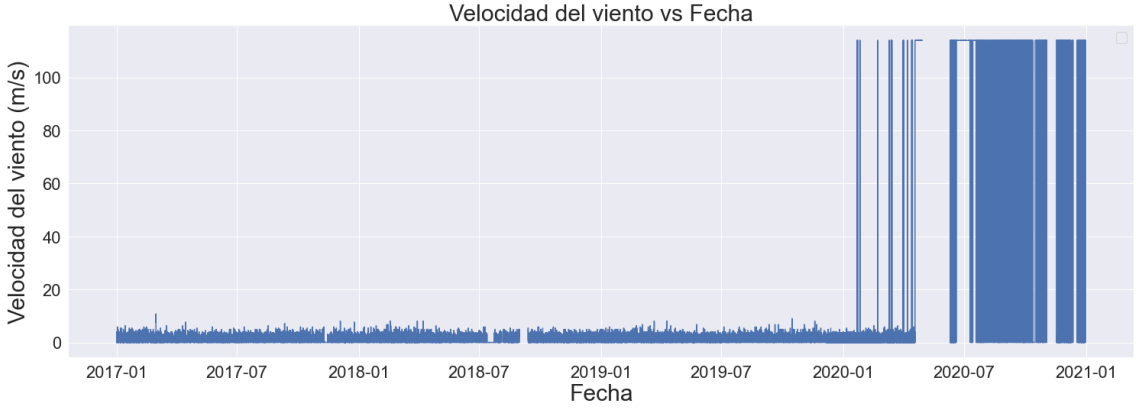
5.3.2. Transformaciones y Análisis de los Datos

Antes de presentar los resultados de la investigación, es necesario explicar la configuración y organización de los datos. En este estudio, se utilizó la base de datos proporcionada por el SIMAC de Manizales, de las 13 estaciones meteorológicas, el cual recoleta información cada 5 minutos de las siguientes variables: la fecha, hora, temperatura (°C), precipitación (mm), velocidad del viento (m/s), presión (mmHg), humedad (%), y radiación solar (W/m^2) (CORPOCALDAS & Universidad Nacional, 2022). Adicionalmente, todo el trabajo fue realizado en Python 3 – Jupyter notebook, utilizando la versión 3.8.12, con un sistema operativo de Windows 10, memoria RAM de 64 Gb y una tarjeta de video dedicada de 8 GB.

Los datos obtenidos estaban en formato csv por año y por estación de manera individual. Como primer paso, se agrupo toda la información de cada estación por año y se creó un nuevo dataset para cada estación completo. Segundo, fue necesario organizar los nombres de todas las columnas de las 13 estaciones y eliminar algunos signos y símbolos innecesarios; además de eliminar palabras en las variables numéricas que se encontraban inmersas dentro de los valores.

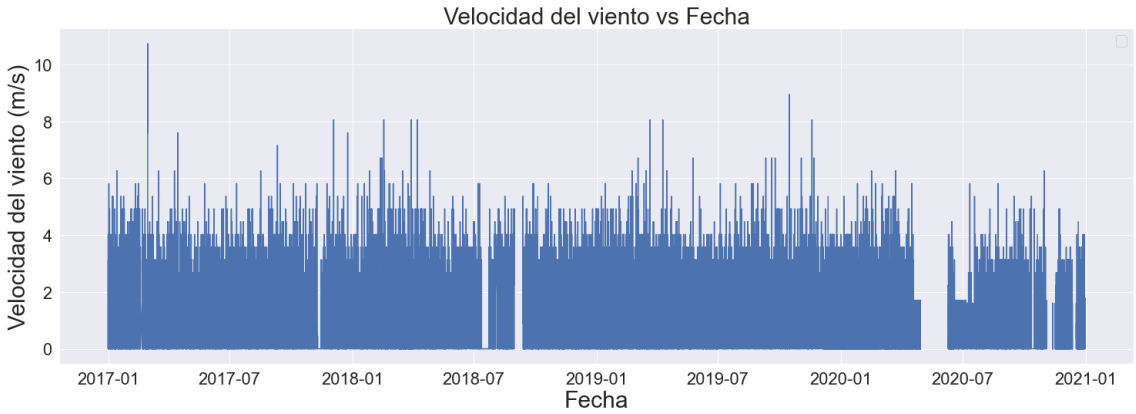
Se identifican diferentes variables meteorológicas con datos atípicos o valores extremos, que no tienen ningún significado más allá del error de medida del equipo. Por esta razón, se propone una estrategia para no perder la información que es importante en la serie de tiempo, ya que, si se eliminan los valores, se pierde la continuidad temporal. Por otro lado, en la Ilustración 4, se muestra un ejemplo para la estación Alcázares, donde se presentan picos muy altos que no coinciden con los valores dentro de los rangos normales:

Ilustración 4. Datos originales de la Velocidad del Viento vs Fecha – Estación Alcázares.



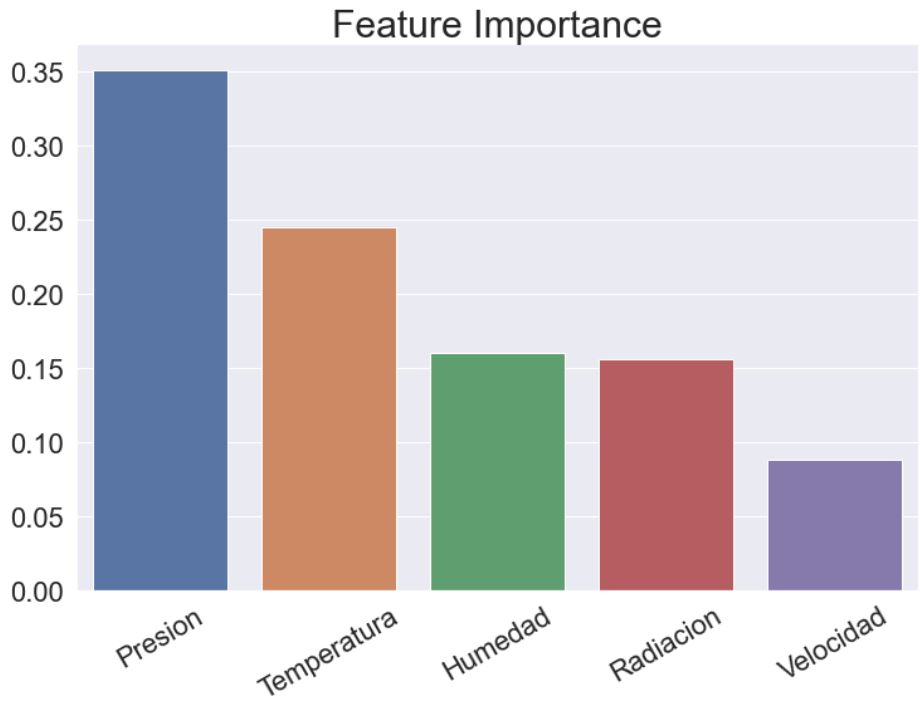
La estrategia propuesta para solucionar el problema se basa en encontrar el valor de la media por periodo de tiempo y añadiendo una desviación en los datos de manera que oscilen dentro del rango de la media. Además, esto se realizó con el fin de simular valores que se asemejen más a la realidad y no una pila de valores de igual magnitud en todo el lapso del tiempo. Igualmente, la estrategia contempla en obtener la media por un periodo de tiempo de cada 3 meses y para cada año en las 13 estaciones meteorológicas. De esta manera, la media no se verá influenciada por el comportamiento de otros valores, sino por periodos que se asemejan a las estaciones de la primavera, verano, otoño e invierno. Dando como resultado lo que se muestra en la Ilustración 5:

Ilustración 5. Datos aplicando la estrategia de la media para la Velocidad del Viento vs Fecha - Estación Alcázares.



Con el fin de identificar las variables que más peso tienen respecto a la variable objetivo (Precipitación), se implementó el modelo Random Forest usando la función *feature_importance*. El resultado se muestra en la Ilustración 6.

Ilustración 6. Feature importance - Random Forest.



Se puede inferir que la variable con mayor peso es la Presión, sin embargo, la magnitud de estas no tiende a valores cercanos a 1, siendo 1 el mayor valor que representa una alta explicación y 0 que no influye de ninguna manera. Por otro lado, este análisis no determina al 100% el aporte o el impacto que podría llegar a tener en un modelo estadístico, pero sí que da indicios sobre que variables utilizar para empezar a implementar una solución. En este caso se seleccionan las primeras 3 más importantes que se presentan en la Ilustración 6, (Presión, Temperatura y Humedad).

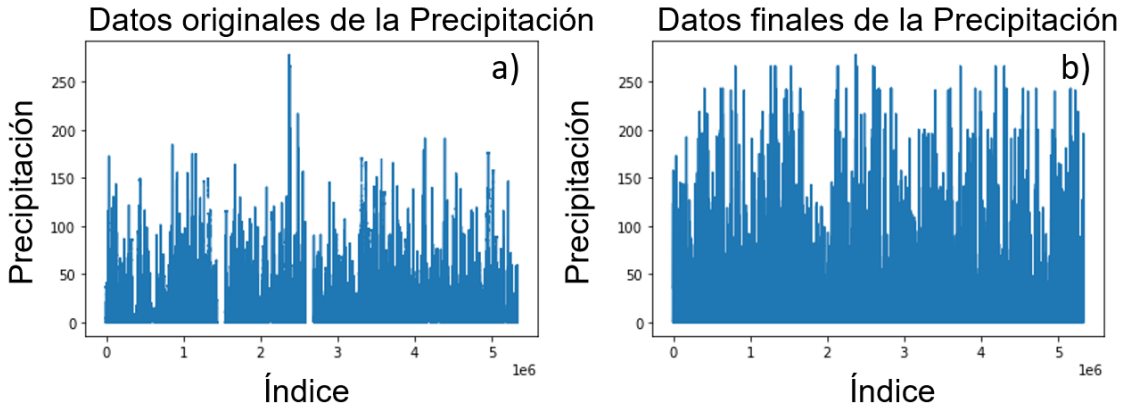
Datos nulos

Se implementaron dos métodos diferentes para la imputación de los valores faltantes de los datos climáticos. Por otro lado, se compara el método de *MissForest* y *MICE*, debido a que son muy utilizados y aceptados en el campo de la investigación. Adicionalmente, se debe tener en cuenta que la gran cantidad de filas hace compleja esta solución. Posteriormente, después de realizar una búsqueda y encontrar un plan eficiente que sirva para determinar los mejores valores, se descartan otros métodos como el reemplazo por la media, moda, y por el *KNN Imputer*. A consecuencia de lo anterior, se investigó un sistema que tome en cuenta las demás variables, diferente a la variable a predecir (Precipitación), ya que la precipitación depende de la temperatura, de la presión, humedad, entre otras. Sin embargo, el *KNN Imputer*, fue descartado después de que llevara mucho tiempo en compilarse y sin éxito alguno de finalizar la imputación; De la misma manera, como se necesita un tiempo de ejecución

bajo para implementarse en un sistema de prevención de riesgo, se proponen los modelos *MICE* y *MissForest*.

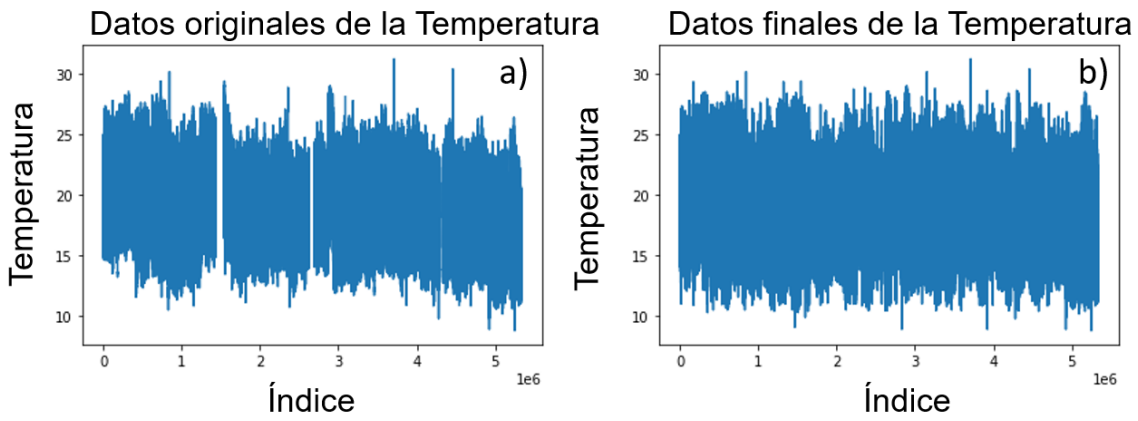
El modelo *MICE*, tardo de 3 a 5 minutos en realizar la imputación de datos faltantes en todo el conjunto de datos. Sin embargo, se encuentran diferencias en los valores predichos por cada uno de los modelos. Después de ajustar algunos hiperparámetros del *MICE*, algunas de las variables toman comportamientos diferentes, el gráfico de la precipitación y la temperatura se muestran en la Ilustración 7 y 8:

Ilustración 7. Precipitación vs Índice – Datos originales y Datos finales aplicando el modelo MICE.



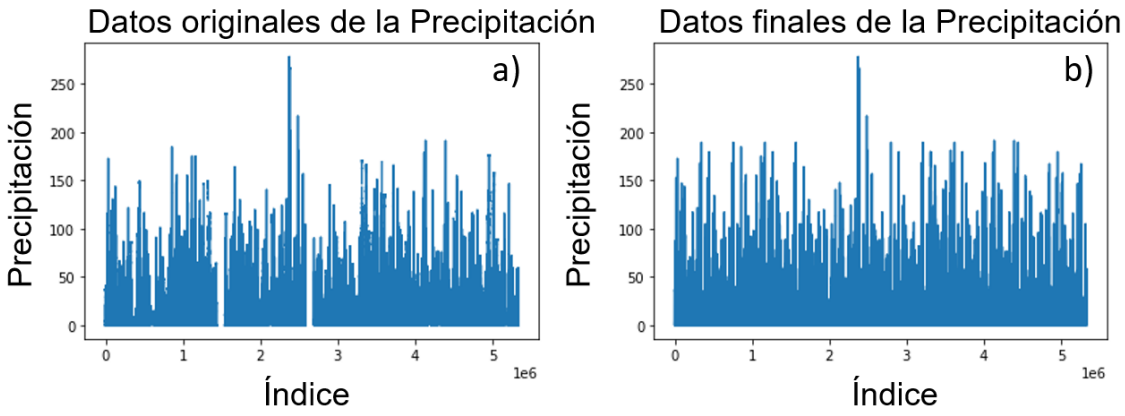
En la Ilustración 7-b, se puede observar que los valores imputados para la precipitación se encuentran cerca del punto más alto, de esta manera hay varios valores que toman esta magnitud. Sin embargo, para las demás variables (Presión, Humedad, Radiación y la Velocidad del viento) como la temperatura que se muestra en la Ilustración 8-b, los valores predichos se encuentran mejor distribuidos y con magnitudes cercanas a 25 °C como en los datos originales (Ilustración 8-a), lo cual evidencia una tendencia a los valores contiguos de la misma variable. Por último, esto se puede atribuir a que el modelo utiliza los vecinos más cercanos para la imputación.

Ilustración 8. Temperatura vs Índice – Datos originales y Datos finales aplicando el modelo MICE.



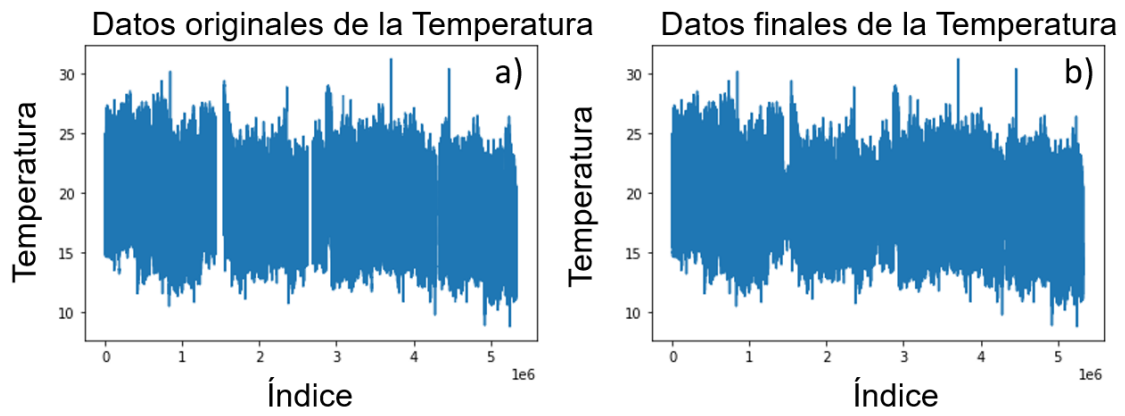
El modelo MissForest, debido a su complejidad tuvo un tiempo de ejecución aproximadamente 1 hora con 31 minutos, bajo las condiciones computacionales mencionadas al inicio de la sección 5.3.2. Por otro lado, se implementaron 4 iteraciones para la imputación de los datos. Finalmente, se identifican diferencias particulares respecto al modelo MICE. El resultado obtenido se muestra en la Ilustración 9 para la precipitación:

Ilustración 9. Precipitación vs Índice – Datos originales y Datos finales aplicando el modelo MissForest.



En la Ilustración 9-b, se pueden observar los valores pronosticados y cómo estos mismos toman magnitudes de precipitación más acorde a las magnitudes por debajo de 150 mm que se observan en los datos original (Ilustración 9-a). Se infiere que, ya que el modelo se basa en árboles de decisión, se podría esperar un comportamiento en los datos más acorde a su propio árbol o a su bosque al realizar la predicción final. Mientras que en el modelo MICE (Ilustración 8-b), los valores imputados se acercan a los valores más altos que superan los 200 milímetros. En este caso, se podría entender que las demás variables meteorológicas están influyendo en la predicción de la precipitación y que el modelo MissForest no se ve tan sesgado por el valor extremo.

Ilustración 10. Temperatura vs Índice – Datos originales y Datos finales (modelo MissForest).



En la Ilustración 10-b, los valores imputados toman magnitudes que oscilan entre el valor de la media de la variable (este comportamiento ocurre para las demás variables restantes). A diferencia del modelo MICE, que la variación de los valores predichos se asemeja más al comportamiento de los valores más cercanos de los datos.

Debido a que el modelo MissForest tiende a ser más robusto, menos sesgado por valores extremos y complejo en cuanto a su desarrollo e implementación, fue el elegido para entrenar los modelos de series de tiempo. Sin embargo, no se descarta la posibilidad de utilizar MICE debido a su rápida ejecución. En la Tabla 2 se muestran los tiempos de ejecución para cada modelo.

Tabla 2. Tiempos de ejecución para modelos de imputación de los datos.

Modelo	***Tiempo de ejecución***
MICE	3 minutos 22 segundos
MissForest	1 hora 22 minutos 46 segundos

5.4. Resultados

En este capítulo, se muestra el estudio de las series de tiempo, la estimación y pronóstico de los modelos descritos en la sección del marco teórico. La investigación incluye diferentes análisis estadísticos y verificación de los modelos ARIMA, SARIMA, SARIMAX, Prophet, Neural Prophet. Adicionalmente, para el desarrollo de los modelos fue necesaria la instalación de las librerías de Pandas, *NumPy*, *MICEForest*, *MissForest*, *Scikit-learn*, *statsmodels*, Prophet, Neural Prophet

El análisis de los modelos de series de tiempo está dividido en 2 secciones. Primero se presentan los modelos ARIMA, SARIMA, Y SARIMAX, donde se realizó la prueba de Dickey-Fuller (ADF) y valor p para determinar la estacionariedad de la serie. Por otro lado, la función de autocorrelación y la autocorrelación parcial fueron obtenidas con el fin de entender el comportamiento de la serie y determinar algunos hiperparámetros del modelo (p,d,q). Segundo, se implementó un modelo baseline y posteriormente se realizó una optimización en los modelos estocásticos, y probabilísticos (Prophet y Neural Prophet), con el fin de encontrar el mejor modelo basado en las métricas de evaluación con el valor más bajo utilizando el RMSE, MSE y MAE.

Hipótesis Nula

La serie de tiempo es no estacionaria.

Con el fin de determinar el comportamiento de la serie de tiempo, fue necesario implementar el método de Augmented Dickey Fuller (ADF). Este método expande la

Ecuación de Dickey Fuller, que se muestra en la Ecuación 22, para incluir un proceso regresivo de alto orden en el modelo.

$$y_t = c + \beta t + \alpha y_{t-1} + \phi_1 \Delta Y_{t-1} + \phi_2 \Delta Y_{t-2} + .. + \phi_p \Delta Y_{t-p} + e_t$$ *Ecuación 22*

$Donde$:

$y_{t-1} = lag\ 1\ de\ la\ serie\ de\ tiempo.$
$\Delta Y_{t-1} = Primera\ derivada\ de\ la\ serie\ de\ tiempo\ (t-1)$

Dado que la hipótesis nula asume la presencia de raíz unitaria, es decir $\alpha = 1$, el valor p obtenido debe ser menor que el nivel de significancia (< .05) y el estadístico de prueba menor que los valores críticos, para rechazar la hipótesis nula. De esta manera, se infiere que la serie es estacionaria (Mushtaq, 2011).

Los valores estadísticos obtenidos de la prueba ADF, el valor – p y valores críticos para cada estación, se muestran en la Tabla 3 y 4.

Tabla 3. Valores críticos

Valores críticos		
1%	5%	10%
-3.430	-2.862	-2.567

En la Tabla 4 se observa que los valores-p de todas las 13 estaciones (*p-valor = 0*) es menor que el nivel de significancia, y además el ADF estadístico es mucho menor a los valores críticos del 1%, 5%, 10% que se muestran en la Tabla 3. Esto implica que las series temporales son estacionarias, rechazando así la hipótesis nula.

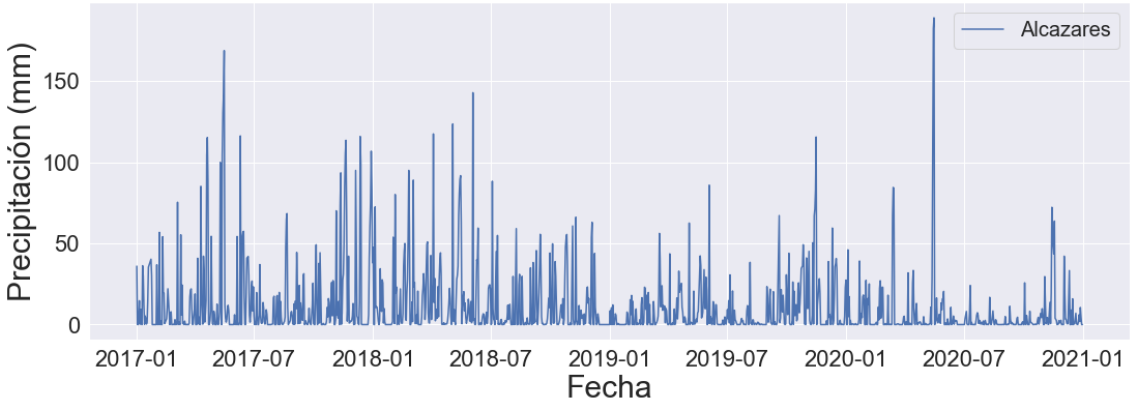
Tabla 4. Valores p y ADF estadístico de las 13 estaciones.

Estación	ADF Estadístico	p-valor
Alcázares	-16.05967	0.000
Aranjuez	-17.53631	0.000
Bosques del Norte	-15.18559	0.000
Chec Uribe	-17.54788	0.000
El Carmen	-16.61556	0.000
EMAS	-15.82624	0.000
Hospital de Caldas	-16.70738	0.000
La Nubia	-15.95407	0.000
La Palma	-16.26608	0.000
Milán	-16.60355	0.000
Obs. Vulcanológico	-15.49576	0.000
Posgrados	-16.75009	0.000
Yarumos	-15.33693	0.000

Autocorrelación (ACF) Y Autocorrelación Parcial (PACF)

El siguiente método se utiliza para determinar la estacionalidad y los parámetros p, q de la serie de tiempo, observando el comportamiento de la gráfica de la función de autocorrelación (ACF) y autocorrelación parcial (PACF). Adicionalmente, se obtuvieron resultados para las 13 estaciones meteorológicas, presentando un comportamiento similar, con sus respectivas variaciones. Finalmente, por razones de presentación en el texto del trabajo, se mostrará la gráfica de la estación Alcázares para el análisis y la explicación de los resultados, la cual se observa en la Ilustración 11 y 12.

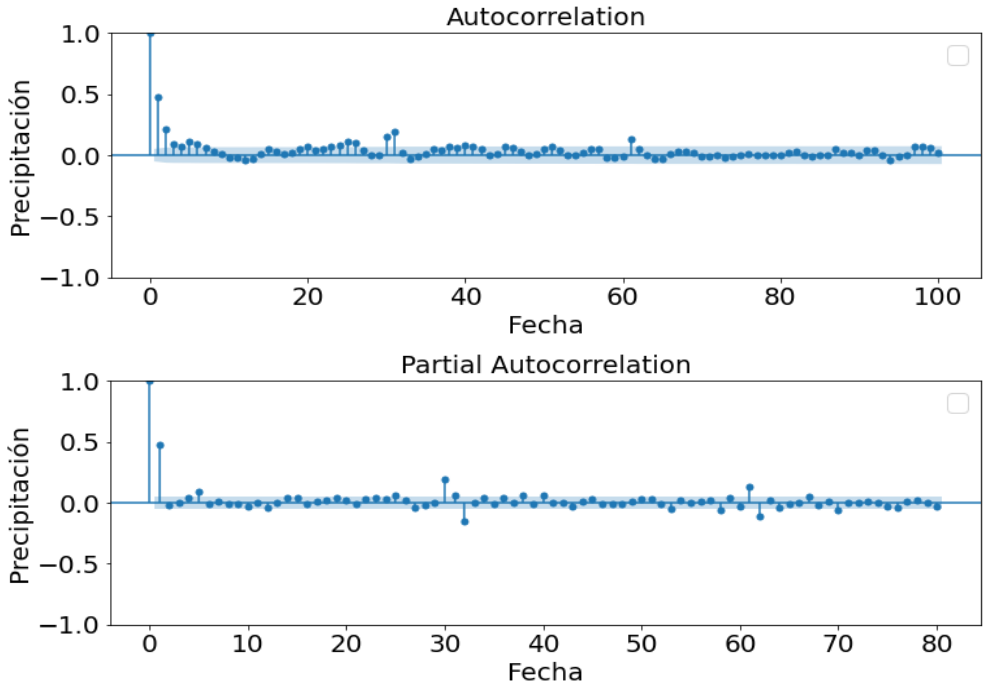
Ilustración 11. Precipitación vs fecha – Estación alcázares.



En la Ilustración 11, se observa como la serie de tiempo no tiene una tendencia, y presenta fluctuaciones que incrementan y decrecen aleatoriamente en el tiempo. Además, el tamaño de las variaciones es constantes en el tiempo, lo cual es otro indicador de la estacionariedad en la gráfica.

En la Ilustración 12, la función de autocorrelación (ACF) cae rápidamente cerca a cero describiendo una serie de tiempo estacionaria. Los demás valores de la ACF no son estadísticamente significativos ya que se encuentran por debajo del límite de la línea azul, lo cual indica que hay presencia de ruido blanco. Es importante resaltar que en los lags de la gráfica se evidencia un comportamiento parecido a una función senoidal, donde los valores oscilan de manera positiva y negativa, dentro del límite de color azul. Esto se podría explicar debido a la presencia de estacionalidad en la serie de tiempo. Adicionalmente, se encuentran picos significativos aproximadamente en el lag 29, 30, 60, que representan estacionalidad (Hyndman & Athanasopoulos, 2018).

Ilustración 12. ACF y PACF - Estación Alcázares.



El comportamiento del ACF y PACF, representa un modelo ARMA (1,1), o en su defecto un modelo ARIMA (101). Adicionalmente, es posible notar que, el proceso ARMA (1,1) contiene el MA (1) como caso especial. Además, el PACF del proceso ARMA (1,1) también disminuye exponencialmente como el ACF, dependiendo de los signos y magnitudes de ϕ_1 y θ_1, de la Ecuación de autocorrelación. Por último, para ARMA (1,1) el modelo sigue la función de autocorrelación (ρ_k) que se presenta en la Ecuación 23:

$$\rho_k \begin{cases} 1 & k = 0 \\ \frac{(\phi_1 - \theta_1)(\phi_1\theta_1)}{1+\theta_1^2-2\phi_1\theta_1} & k = 1 \\ \phi_1\rho_{k-1} & k \geq 2 \end{cases} \quad \textit{Ecuación 23}$$

Donde ϕ_1 es el término de la representación de la media móvil y θ_1 corresponde a la parte autorregresiva. De esta manera, si tanto ACF como PACF disminuyen, indica un modelo ARMA mixto. Adicionalmente, al analizar la Ilustración 12, es posible identificar que debido a efecto combinado de ϕ_1 y θ_1, el PACF del proceso ARMA (1,1), contiene muchas más formas diferentes que el PACF del proceso MA (1), que contiene solo dos posibilidades (Wei, 1991).

Debido a que se presenta un modelo ARMA, de acuerdo con la Ilustración 12 y lo mencionado anteriormente, se implementaron los modelos baseline utilizando la configuración p = 1 y q = 1, que se describen en el siguiente apartado.

Modelo Baseline

Varios hiperparámetros fueron probados inicialmente, realizando algunos experimentos durante la programación para su ajuste final. Adicionalmente, dependiendo de los resultados de los valores estadísticos de ADF, valor p, la función ACF y PACF se seleccionó el modelo baseline.

model = sm.tsa.arima.ARIMA(train['Precipitacion'],order=(1,0,1))
smodel = sm.tsa.statespace.SARIMAX(train['Precipitacion'],order=(1,0,1), seasonal_order = (1,0,1,6))
smodel = sm.tsa.statespace.SARIMAX(train['Precipitacion'],order=(1,0,1), seasonal_order = (1,0,1,6), exog = train[['Temperatura']])

Para encontrar un modelo de referencia, primero se debe tener un conocimiento del comportamiento de la autocorrelación y la autocorrelación parcial. Así mismo, se determinaron los valore de p, d y q para realizar una primera aproximación. De acuerdo con la sección 5.4 sobre el análisis de ACF Y PACF, se encontraron los parámetros de la autoregresión y de media móvil para los modelos estocásticos, siendo p = 1, d = 0 y q = 1.

Por otro lado, la cantidad de dias predichos se estableció de acuerdo con el problema y las necesidades. Adicionalmente, la parte estacional y el uso de las variables exógenas dependerán del tipo de datos y de la tendencia de la serie de tiempo. Por lo general, estos valores se pueden modificar manualmente dentro del modelo. Finalmente, para dividir los datos de entrenamiento y testeo, se decidió realizar la separación de los datos a partir de las siguientes fechas:

- ✓ Entrenamiento: 2017/01/01 – 2020/12/24.
- ✓ Test: 2020/12/25 – 2020/12/31.
- ✓ Predicción: 7 días.

Las métricas establecidas para determinar el menor error de los modelos implementados son el RMSE, MAE, MSE. Por otro lado, para determinar el error total por cada modelo entrenado y validado, se obtuvo la media de los errores sumando el valor de las métricas individuales de las estaciones y dividiendo entre 13, para posteriormente identificar el que presentara el menor valor de los 5 modelos.

En la ilustración 13, se puede observar la gráfica del modelo ARIMA (1,0,1,) de los datos de entrenamiento (línea azul) y los predichos (línea naranja) para la estación Alcázares, en donde se obtuvo el menor resultado de la métrica $RMSE_{train}$ = 19.08935 que se presenta en la Tabla 5.

Ilustración 13. Datos de Entrenamiento para la estación Alcázares, utilizando el Modelo $ARIMA_{101}$.

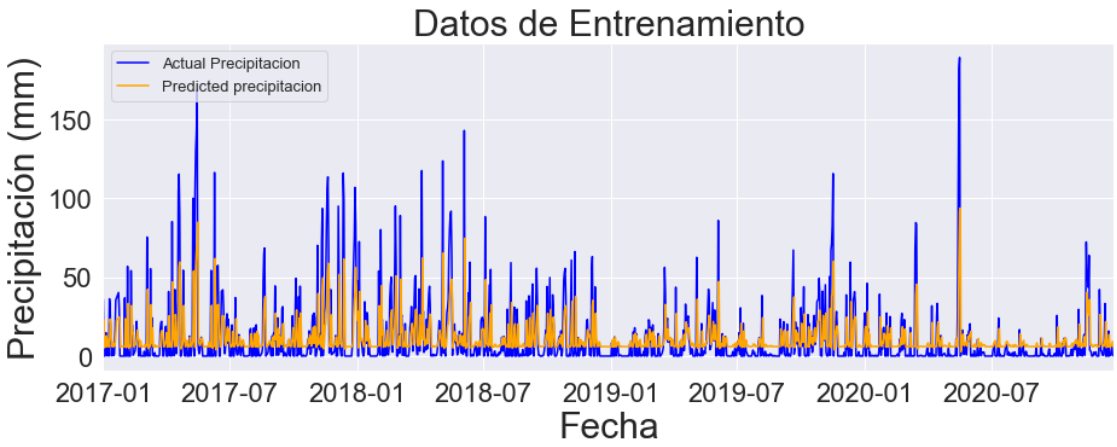


Tabla 5. Métricas de evaluación para los datos de entrenamiento.

Entrenamiento	RMSE	MAE	MSE
$ARIMA_{101}$	19.08935	12.12777	369.58516
$SARIMA_{101}$	19.24187	12.06760	375.48963
$SARIMAX_{101}$	19.29660	11.94998	377.68039
$Prophet_{DÍAS}$	20.83209	13.76298	442.65334
Neural $Prophet_{DÍAS}$	21.08390	13.43836	452.82367

A continuación, se presentan los resultados de los datos de testeo para cada modelo implementado en la Tabla 6:

Tabla 6. Métricas de evaluación para los datos de testeo.

Test	RMSE	MAE	MSE	AIC
$ARIMA_{101}$	27.07459	23.70040	870.11410	12690.11469
$SARIMA_{101}$	27.70922	23.83590	924.10553	12705.52531
$SARIMAX_{101}$	28.34202	24.37122	958.88549	12829.05908
$Prophet_{DÍAS}$	24.91720	21.37104	756.58101	-
Neural $Prophet_{DÍAS}$	30.38424	24.75415	1115.48134	-

Contrario a los datos de entrenamiento, el menor valor del RMSE, MAE y MSE, se encuentra en el modelo Prophet. El RMSE y el MAE están en las mismas unidades de la variable respuesta, por lo que se puede concluir que es necesario realizar una optimización de hiperparametros para disminuir el valor del error en las métricas de evaluación. Por otra parte, el MSE penaliza un pequeño error y es más robusto a los valores atípicos, pero puede llegar a la sobreestimación. De esta manera, en la tabla 6 se observa como el valor más alto se encuentra en el modelo Neural Prophet. De igual

manera, el RMSE se utilizó para penalizar los errores grandes, siendo Neural Prophet el que presenta diferentes variaciones en los datos predichos que no se ajusta muy bien a los datos de entrenamiento (Tabla 5) ni a los de validación.

5.4.1. Optimización de hiperparámetros

ARIMA, SARIMA y SARIMAX

La optimización de hiperparámetros para los modelos ARIMA, SARIMA y SARIMAX, se implementó utilizando la función auto ARIMA, donde se buscan los valores más bajos basados en el valor del AIC, automáticamente. Además, esta opción se implementó como estrategia debido a la gran cantidad de estaciones que se debían entrenar con cada modelo para posteriormente identificar cual presenta un menor error y valor de AIC.

En la Tabla 7, se presentan los resultados de las métricas obtenidas implementando la función auto ARIMA. Adicionalmente, se utilizaron diferentes combinaciones de los parámetros con el fin de encontrar el modelo con el error más bajo para cada estación meteorológica utilizando el criterio AIC. Por último, El tiempo de ejecución en este caso, puede variar entre 1 a 5 minutos. Los valores de las métricas de evaluación de todos los modelos implementados se encuentran en el Apéndice A.

Tabla 7. Resultados de las métricas de evaluación - Modelo ARIMA$_{d=0,\ s=True}$.

ARIMA d = 0 Seasonal = True	Parámetro (pdq)	RMSE	MAE	MSE
Alcázares	(100)	7.47173	6.44317	55.82679
Aranjuez	(200)	18.07889	15.49646	326.84650
Bosques del norte	(101)	16.59625	15.30626	275.43558
CHEC Uribe	(101)	43.47141	36.48689	1889.76429
El Carmen	(101)	16.77893	14.43663	281.53257
EMAS	(100)	11.73780	8.89895	137.77597
Hospital de Caldas	(100)	21.16828	18.71273	448.09638
La Nubia	(400)	18.45964	15.90361	340.75846
La Palma	(201)	41.67653	35.19205	1736.93333
Milán	(303)	29.75427	26.09933	885.31662
Obs. Vulcanológico	(200)	39.50590	31.97134	1560.71687
Posgrados	(100)	24.59690	21.06550	605.00755
Yarumos	(201)	28.92923	25.36730	836.90051
Total		24.47890	20.87540	721.60857

Después de realizar el análisis estadístico de la estacionariedad, utilizando las pruebas ADF y el valor-p donde se rechaza la hipótesis nula, se decidió cambiar algunos hiperparámetros dentro del modelo auto ARIMA. Por consiguiente, el termino de

derivación d, el cual representa la n-esima derivada, fue modifico dependiendo al comportamiento de cada estación meteorológica. Finalmente, con el fin de disminuir el valor de las métricas y el AIC, los valores de autoregresión y de media móvil tomaron valores entre cero y cinco.

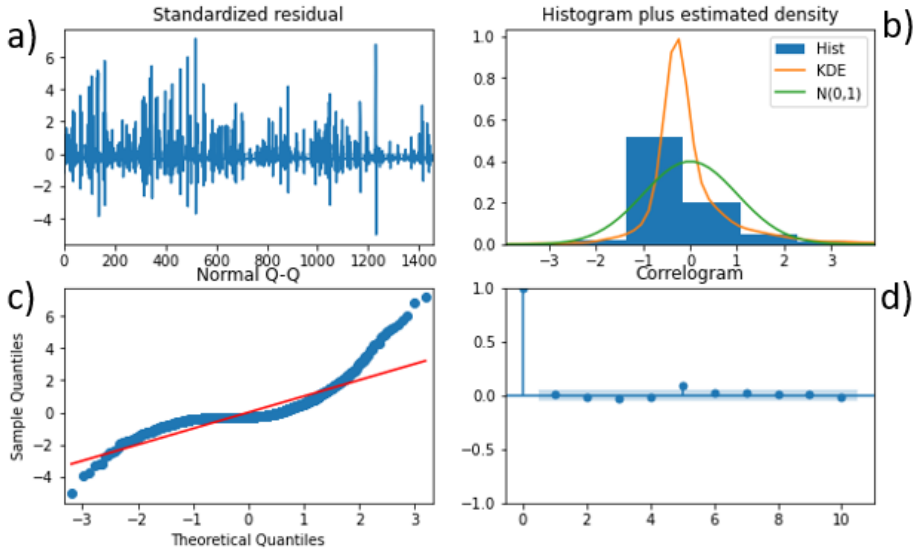
Tabla 8. Resultado del promedio de las métricas de evaluación final – Modelos ARIMA, SARIMA y SARIMAX.

Modelos	RMSE	MAE	MSE	AIC
$ARIMA_{d = 1, Seasonal = False}$	25.31393	21.15138	784.65185	12709.60054
$ARIMA_{d = 0, Seasonal = True}$	24.47890	20.87540	721.60857	12696.14238
$SARIMA_{D=1, m=12}$	27.22153	21.93400	865.01840	13005.91031
$SARIMA_{D=0, m=12}$	24.57647	20.92591	729.32498	12727.42554
$SARIMAX_{D=1, m=6}$	28.32908	23.33140	967.88948	13075.36138
$SARIMAX_{D=0, m=6}$	24.50772	20.88507	707.97779	12693.24569

En la Tabla 8, se observan los resultados de las métricas finales para los modelos ARIMA, SARIMA y SARIMAX. Así mismo, se identificó el mejor modelo cómo ARIMA bajo la configuración de diferenciación cero (d = 0) y presencia de estacionalidad. Por último, los errores de las métricas de evaluación toman valores de RMSE = 24.47890, MAE = 20.87540, MSE = 721.60857 y un AIC = 12696.14238.

Se analizaron los residuos del modelo para las 13 estaciones, los cuales corresponden a la diferencia entre los valores reales y los predichos. Adicionalmente, se mostrará un estudio de la estación Alcázares para explicar los resultados obtenidos. El criterio para seleccionar un buen modelo se basa en obtener ruido blanco aleatorio de los residuos, es decir con media cero. A continuación, en la Ilustración 14 se muestran los resultados residuales usando gráficas de Q-Q, Histograma, análisis cuantil y del residuo estándar, del modelo $ARIMA_{d = 0, Seasonal = True}$, de la estación posgrados.

Ilustración 14. Diagnóstico de residuos del modelo $ARIMA_{d=0, s =True}$ para la estación Alcázares.



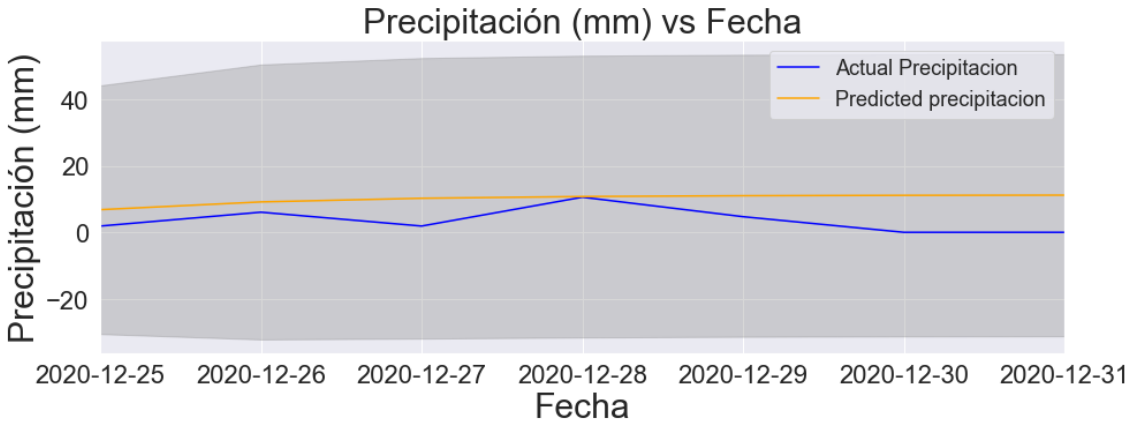
En el análisis del gráfico de diagnóstico que se observa la Ilustración 14, se puede determinar que existe una autocorrelación positiva de los residuos en el correlograma (Ilustración 14-d) y conforme incrementa el eje x toma valores cercanos a cero. Por otra parte, según el gráfico de la estandarización residual (Ilustración 14-a), los valores oscilan cerca a la media. Igualmente, las cantidades de los residuos en el gráfico normal Q-Q (Ilustración 14-c), se muestra que no se alinean perfectamente a la línea recta, debido a los valores extremos. En otras palabras, no se cumple por completo el supuesto de una distribución normal, y esto se corrobora con el histograma. Por consiguiente, el modelo puede mejorarse, hasta satisfacer las propiedades de media cero.

Ilustración 15. Resumen de parámetros del modelo ARIMA$_{d=0,\ s=True,\ (100)}$ - Estación Alcázares.

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	5.9100	0.855	6.914	0.000	4.235	7.585
ar.L1	0.4727	0.010	45.361	0.000	0.452	0.493
sigma2	362.5956	7.257	49.962	0.000	348.371	376.820

El resumen de salida que se obtiene al implementar la función auto ARIMA devuelve diferentes cantidades significativas de información estadística. Sin embargo, se hará énfasis en la Tabla 15 en el apartado de los coeficientes. Adicionalmente, dentro de la Ilustración 15, se observan la columna de coeficientes la cual muestra la importancia de cada característica y cómo influye en la serie temporal. El valor p indica el impacto de cada función de peso. En este caso cada valor es igual a cero, por lo tanto, se dejan los coeficientes dentro del modelo ($p < .05$).

Ilustración 16. Predicción 7 días - Modelo ARIMA$_{d=0,\ s=True,\ (100)}$ para la estación Alcázares.



En la Ilustración 16, se presenta el resultado de la predicción para los 7 días siguientes después de la fecha del 2020/12/24, para la estación Alcázares. Adicionalmente, la línea de tendencia de los valores predichos tiende a ser constantes en el tiempo, con una variación entre 5 y 12 mm de precipitación. De igual manera se observa el umbral en color gris, que representa los valores que podría encontrarse la predicción de

precipitación, fuera del rango se clasifica como un valor atípico. Por otro lado, el valor de la métrica RMSE se muestra en la Tabla 7, donde se puede observar el puntaje más bajo de las 13 estaciones para el modelo $ARIMA_{d=0,\ s=True}$ con una configuración de (1,0,0) y un RMSE = 7.47173.

Prophet y Neural Prophet

Los modelos probabilísticos Prophet y neural Prophet fueron optimizados utilizando diferentes configuraciones en los parámetros de entrada. Además, las variables exógenas (temperatura, presión, radiación, velocidad, humedad) fueron añadidas, con el fin de determinar si podrían influenciar en la variable objetivo y a partir de las métricas de evaluación obtener un valor menor. Por otro lado, el efecto de los días feriados fue reemplazado por los días en donde una catástrofe por deslizamiento de tierra sucedió en la ciudad de Manizales. Finalmente, estos sucesos se modelan de forma análoga a los regresores futuros para el modelo Neural Prophet. EL código implementado es el siguiente:

```
holidays = pd.DataFrame({
            'holiday': 'eventosmanizales',
            'ds': pd.to_datetime(['2017-04-19']),
            'lower_window': 0,
            'upper_window': 1,
        })
    m.add_regressor('Temperatura')
    m.add_regressor('Presion')
    m.add_regressor('Velocidad')
    m.add_regressor('Humedad')
    m.add_regressor('Radiacion')
```

En la Ilustración 17, se puede observar el comportamiento de los valores predichos de entrenamiento, que se ajustan muy bien a los valores de precipitación reales por debajo de 50 mm. Adicionalmente, se identifica un pico cercano a 130 mm en los valores predichos (Línea de color azul) que el modelo entrena y predice con un valor de precipitación alta. Se infiere que, debido al efecto del evento de desastre agregado al modelo Prophet, se presenta como una anomalía que se ve reflejada en el entrenamiento de los datos. Así mismo, se presenta para las demás estaciones meteorológicas.

Ilustración 17. Valores predichos de entrenamiento vs valores reales usando el modelo Prophet multivariado con 5 variables para la estación Alcázares.

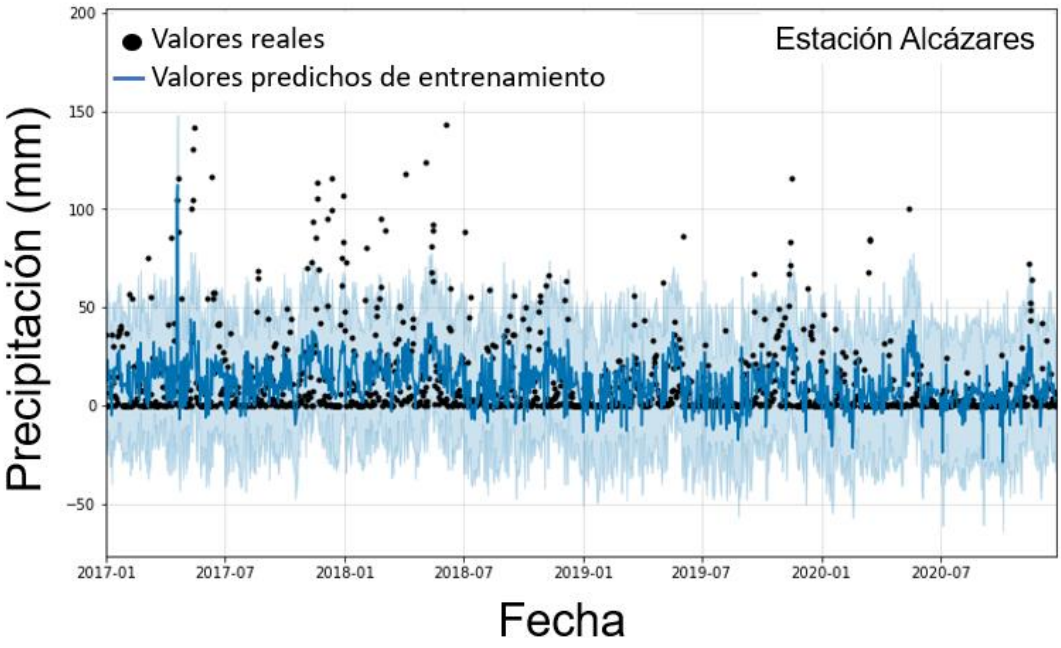


Tabla 9. Resultados del promedio de las métricas de evaluación final para el modelo Prophet.

Modelos	RMSE	MAE	MSE
Prophet Multivariado 3 variables	19.55671	17.07854	460.28991
Prophet Multivariado 5 variables	19.06313	16.24064	430.72827
Prophet Multivariado 5 variables - multiplicative	20.58596	17.63104	504.82400
Neural Prophet Multivariado 3 variables	29.60179	23.35640	1037.28838
Neural Prophet Multivariado 5 variables	28.40272	22.55186	966.35207
Neural Prophet Multivariado 5 variables Multiplicative	27.56973	22.93964	920.99573

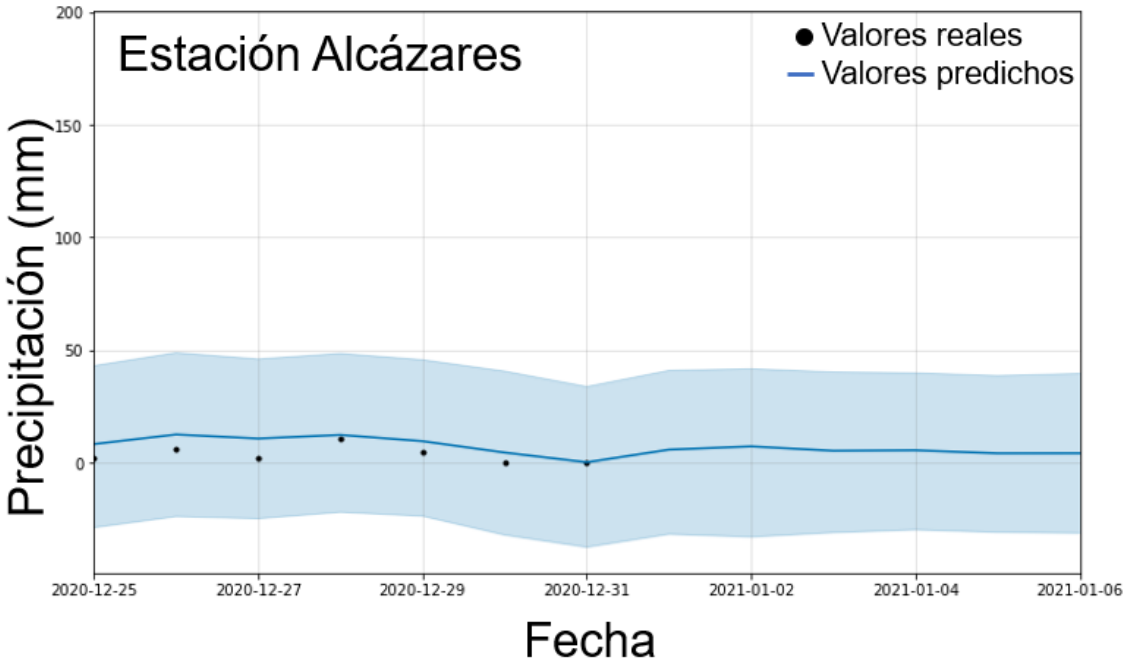
Las métricas de evaluación final se muestran en la Tabla 9. Adicionalmente, los modelos fueron llevados desde 3 variables hasta 5 variables exógenas, para encontrar el mejor ajuste. Por otro lado, se implementaron diferentes hiperparámetros en el modelo Prophet como el tipo de la estacionalidad, que toma los valores de aditivo y multiplicativo. Además, se incorporó las fechas de los desastres por deslizamiento de tierra. Por último, para el modelo Neural Prophet, las redes neuronales se entrenaron con 8 capas ocultas, un epochs de 200 y 16 dimensiones de capas ocultas de AR-Net.

Se determina una disminución en el valor de las métricas de evaluación al utilizar 5 variables exógenas, donde la variable objetivo se ve afectado por la influencia de estas. Finalmente, se puede analizar que el pronóstico de la precipitación con menor valor del RMSE es el modelo Prophet multivariado, con estacionalidad aditiva y con 5

variables exógenas, el valor de RMSE = 19.06313, MAE = 16.24064 y MSE = 430.72827.

En la Ilustración 18, se observa el resultado del modelo con menor error entre los hiperparámetros implementados para la estación Alcázares. De igual manera, se evidencia un buen ajuste y que los datos predichos están muy cerca a los datos reales. Así mismo, la línea azul tiene un comportamiento creciente y decreciente en el tiempo, que varía entre 0 y 10 mm de precipitación. Finalmente, el valor del RMSE para la estación Alcázares es de 5.41035, con un MAE = 4.67015 y un MSE = 29.27193, que se puede observar en el apéndice A.

Ilustración 18. Valores de test predichos vs reales usando el modelo Prophet multivariado con 5 variables para la estación Alcázares.



Aplicación

El sistema de alarma que se implementó en el SIMAC en la sede de la ciudad de Manizales se basa en el artículo escrito por Westen y Terlien (van Westen & Erlien, 1995). Por otra parte, el deslizamiento de tierra puede predecirse a mediano o largo plazo, por medio de la estadística, la probabilidad y con la obtención de los datos históricos recolectados de la red meteorológica. A raíz de esto, es posible crear un boletín con los indicadores de la precipitación acumulada (mm) de los 25 días anteriores, la cual se le conoce como A25.

Westen y Terlien realizaron una investigación sobre la cantidad de lluvia diaria en la estación de agronomía ubicada en Cenicafé (Norte - Manizales), para determinar

alguna correlación con las características del suelo. Posteriormente, como resultado se obtuvo presencia de deslizamiento cuando el valor del A25 alcanzo un límite de 200 mm de precipitación acumulada durante los últimos 25 días (van Westen & Erlien, 1995), (Vélez et al., 2010). Finalmente, como consecuencia de este estudio, las entidades de análisis de riesgo establecieron los niveles de alerta siguiendo los umbrales de la Tabla 10:

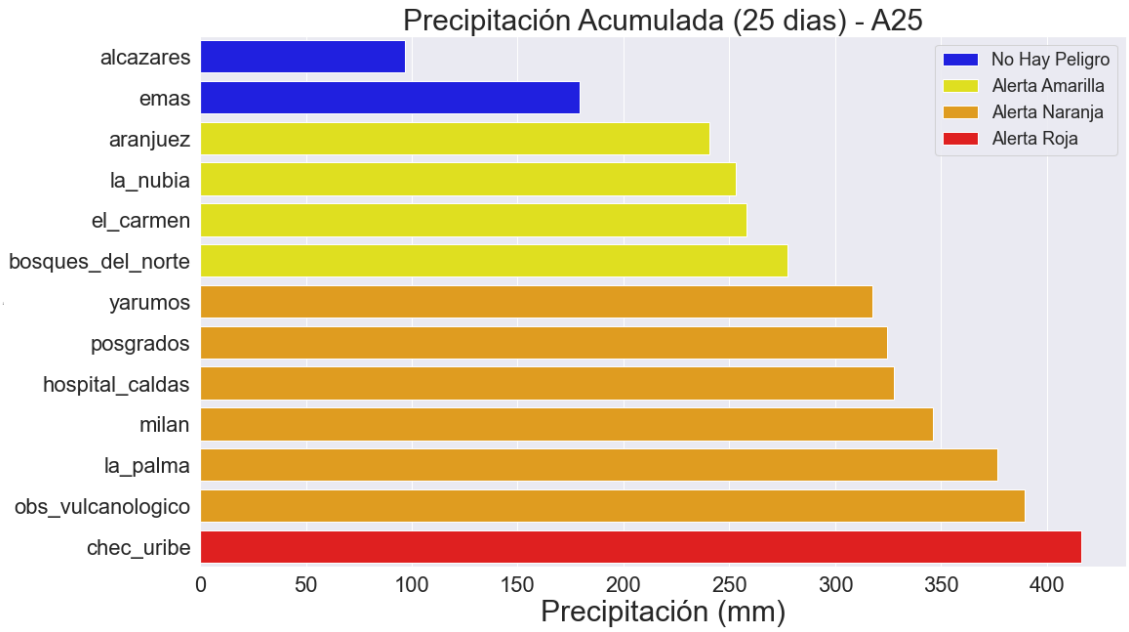
Tabla 10. Estrategia del semáforo de acuerdo con el color.

Color Del Semáforo	Precipitación Acumulada - A25
Alerta Amarilla	$\geq 200\ mm\ y\ < 300\ mm$
Alerta Naranja	$\geq 300\ mm\ y\ < 400\ mm$
Alerta Roja	$\geq 400\ mm$

Para la implementación del sistema de prevención de análisis de riesgo, se decidió predecir la precipitación (mm) de los siguientes 7 días, después de la última fecha de los datos meteorológicos obtenidos. Además, de acuerdo con los resultados el modelo que se seleccionó para predecir la precipitación fue Prophet, donde se obtuvieron los valores más bajos en las métricas de evaluación entre los 5 modelos implementados, para este trabajo.

El objetivo es prevenir y alertar algún tipo de deslizamiento en las diferentes estaciones, con la ayuda del indicador A25, se sumaron los últimos 18 valores de la precipitación de la base de datos y los 7 días de la predicción. Además, el programa imprime el tipo de alerta de acuerdo con su color para cada estación meteorológica, el valor acumulado de los últimos 25 días de la precipitación (mm), tal como se observa en la Ilustración 19.

Ilustración 19. Sistema de alerta para la estación Posgrados.



Implementación del Sistema de anomalías

Se implemento un análisis de anomalías para detectar patrones de la precipitación (mm) que se desvían de su comportamiento, con el fin de alertar sobre tendencias anormales por día. Adicionalmente, este sistema está basado en los valores de la media, la ventana de tiempo y la desviación estándar de los datos de lluvia por cada estación meteorológica. Finalmente, la media se calcula por los datos de cada estación meteorológica, la ventana = 25 y σ toma valores entre 2 y 3. El código implementado para encontrar los valores superiores es el siguiente:

$$data['superior'] = \frac{data[0].rollling(window = wind)}{mean\ ()} - (sigma * data[0].rolling(window = wind).std())$$

A continuación, en la Ilustración 20 a 21 se presentan 2 de los resultados obtenidos por el análisis de las anomalías.

Ilustración 20. Análisis de anomalías para la estación Alcázares.

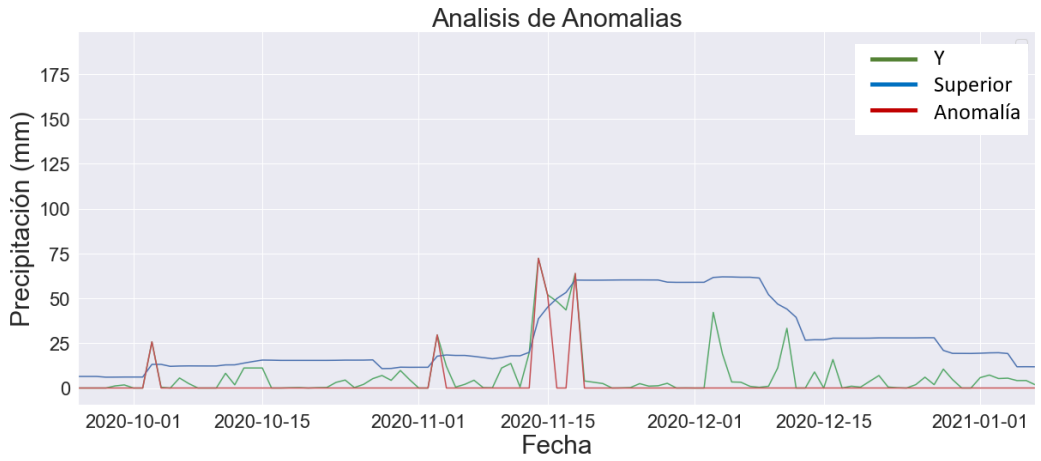
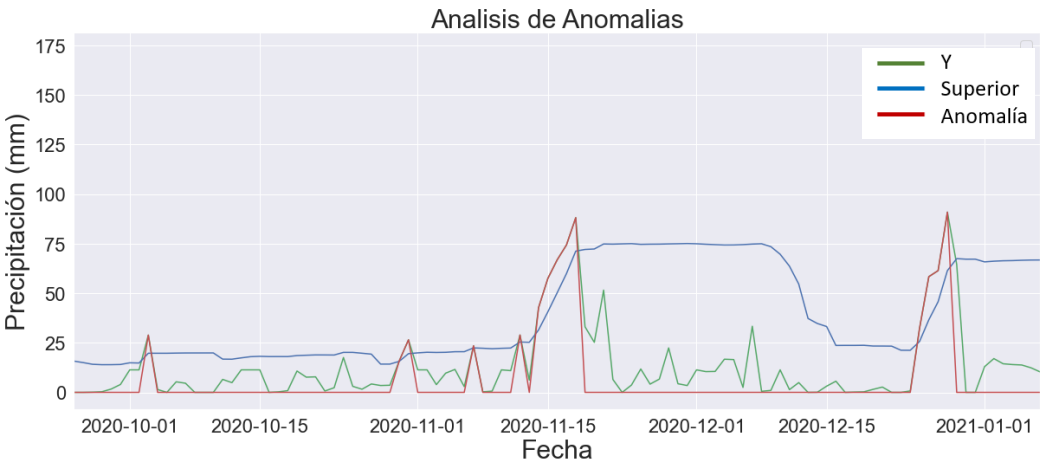


Ilustración 21. Análisis de anomalías para la estación Chec Uribe.



6. Conclusión

- Haciendo uso de las pruebas descritas en el capítulo 5.4 que determinan el tipo de serie para las 13 estaciones meteorológicas, se determinó el valor ADF < 0.05 y un valor p = 0 menor a los valores críticos del 1%, 5% y 10%. Lo anterior permite sugerir que la serie es estacionaria y presenta estacionalidad.

- A partir de los modelos de imputación implementados para los valores nulos, se puede concluir que MissForest presenta valores de imputación menos sesgados a los valores extremos y una mejor tendencia respecto a cada variable climática, del conjunto de datos estudiados. Por otro lado, *MICE* se ve influenciado por los vecinos más cercanos, donde se evidencia la mayor cantidad de datos predichos con altas magnitudes; adicionalmente, en el caso de la precipitación los valores superan los 150 milímetros.

- La función de autocorrelación y autocorrelación parcial permitieron identificar los parámetros iniciales para implementar un modelo base; se puede concluir que los datos siguen un proceso autorregresivo con media móvil - ARMA (1,1). Por último, se reporta presencia de estacionalidad en las series de tiempo con un comportamiento oscilatorio, dentro del umbral estadísticamente no significativo, que toma valores positivos y negativos.

- Con los hiperparámetros optimizados se realizaron las predicciones para obtener el menor error en las métricas de evaluación. Por consiguiente, para los 3 modelos ARIMA, SARIMA y SARIMAX, de acuerdo con el criterio de información de Akaike, se implementó la función auto ARIMA minimizando el valor AIC para las 13 estaciones meteorológicas. Finalmente, ARIMA con derivada cero (*d = 0*) y presencia de estacionalidad (*s = True*), se ajustó más a los datos de precipitación; así mismo, los valores obtenidos en las métricas son RMSE = 24.47890, MAE = 20.87540 y MSE = 721.60857.

- Se evidencia que Prophet funciona mejor que el modelo Neural Prophet y los modelos estocásticos, usando las 5 variables exógenas y las fechas de los desastres por deslizamientos de tierra. Se reporta un valor de RMSE obtenido para Prophet de 19.06313 y para Neural Prophet de 27.56973.

- En comparación con los modelos estocásticos, la librería Prophet prevé parámetros más intuitivos que son fáciles de usar y de mejorar para la predicción de la precipitación. La ventaja de estos modelos radica en que la duración de la implementación es más corta que los modelos ARIMA, SARIMA y SARIMAX.

7. Trabajos Futuros

Esta tesis sirve como base para la implementación de un análisis de prevención de riesgo, utilizando la librería de Prophet. Es posible mejorar el modelo utilizado, obteniendo más información histórica de los datos que sirvan para entrenar el modelo, de igual manera se pueden mejorar más los hiperparámetros, utilizando diferentes combinaciones.

El código realizado en este trabajo, para obtener la imputación de datos, las predicciones y los indicadores A25 de las 13 estaciones meteorológicas se puede mejorar, con el fin de que sea más fácil de entender para el usuario.

Otras técnicas de predicción más complejas quedan pendientes para realizarse, tales como *FUZZY*, redes neuronales, *K-NN*, *Random Forest*, redes neuronales artificiales autorregresivas, combinaciones con ARIMA y la red neuronal de función de base radial, entre otras. Por otro lado, se puede utilizar otro método de imputación diferente, para comparar los resultados obtenidos.

Implementación de un sistema de detección de anomalías para el sistema de alertas.

Implementación del sistema de detección de anomalías de *Amazon Web Service* para recibir alertas vía correo electrones, mensaje de texto, etc.

Por último, está en proceso la redacción de un artículo científico que será publicado en una revista a nivel nacional o internacional, tales como: *Environmental Research, Engineering and Management, Actual Biol, Revista Politecnica, Vitae,* entre otras.

8. Referencias

Abhishek, K., Kumar, A., Ranjan, R., & Kumar, S. (2012). A rainfall prediction model using artificial neural network. *Proceedings - 2012 IEEE Control and System Graduate Research Colloquium, ICSGRC 2012*, *1*, 82–87. https://doi.org/10.1109/ICSGRC.2012.6287140

Ansari, H. (2013). Forecasting Seasonal and Annual Rainfall Based on Nonlinear Modeling with Gamma Test in North of Iran. *International Journal of Engineering Practical Research*, *2*(1), 16–29.

Asha, J. ., Rishidas, S., SanthoshKumar, S., & Reena, P. (2020). Analysis of Temperature Prediction Using Random Forest and Facebook Prophet Algorithms. In *Innovative Data Communication Technologies and Application* (Vol. 46, pp. 432–439). https://doi.org/10.1007/978-3-030-38040-3_94

Bari, S. H., Rahman, M. T., Hussain, M. M., & Ray, S. (2015). Forecasting Monthly Precipitation in Sylhet City Using ARIMA Model. *Civil and Environmental Research*, *7*(1), 69–78. http://www.iiste.org/Journals/index.php/CER/article/view/19069

Betancourt Mesa, J. . (2009). Cambio climatico en colombia. In *Journal of Chemical Information and Modeling* (Vol. 53, Issue 9).

Biswas, S. K., Marbaniang, L., Purkayastha, B., Chakraborty, M., Singh, H. R., & Bordoloi, M. (2016). Rainfall forecasting by relevant attributes using artificial neural networks - a comparative study. *International Journal of Big Data Intelligence*, *3*(2), 111–121. https://doi.org/10.1504/ijbdi.2016.077362

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1007/978-3-030-62008-0_35

Brockwell, P. J., & Davis, R. A. (2016). *Introduction to Time Series and Forecasting, Second Edition* (Springer (Ed.)). http://www.springer.com/series/417

Castillo Ruales, A., Chang, P., Vélez upegui, J. J., Zambrano Nájera, J., & Mejpia Fernandez, F. (2020). Umbrales de precipitación basados en intensidad para crecidas torrenciales en la quebrada Manizales, Colombia. *Revista EIA*, *17*(33), 1–16. https://doi.org/10.24050/reia.v17i33.1302

Cevallos Valdiviezo, H., & Van Aelst, S. (2015). Tree-based prediction on incomplete data using imputation or surrogate decisions. *Information Sciences*, *311*, 163–181. https://doi.org/10.1016/j.ins.2015.03.018

Chatfield, C. (1999). The Analysis of Time Series: an introduction. In *CRC Press* (Vol. 5, Issue 2). https://doi.org/10.2307/2531477

CIOH. (2010). Circulación general de la atmósfera en colombia. In *Bicentenario de la Independencia de Colombia*. www.dimar.mil.co/www.cioh.org.co%0Ahttps://www.cioh.org.co/meteorologia/Climatologia/01-InfoGeneralClimatCaribeCol.pdf

Collischonn, W., Haas, R., Andreolli, I., & Tucci, C. E. M. (2005). Forecasting River

Uruguay flow using rainfall forecasts from a regional weather-prediction model. *Journal of Hydrology*, *305*(1–4), 87–98. https://doi.org/10.1016/j.jhydrol.2004.08.028

Corlett, W. J., & Aigner, D. J. (1972). Basic Econometrics. In *The Economic Journal* (Vol. 82, Issue 326). https://doi.org/10.2307/2230043

CORPOCALDAS, & Universidad Nacional, C. (2022). *CDIAC - Centro de Datos e Indicadores Ambientales de Caldas*. cdiac.manizales.unal.edu.co/

Correa Ortiz, L. C., Ocampo López, O. L., & Alba Castro, M. F. (2021). Análisis de tendencia de temperatura y precipitación para el departamento de Caldas (Colombia), mediante wavelets. *Ciencia e Ingeniería Neogranadina*, *31*(1), 37–52. https://doi.org/10.18359/rcin.4900

Costa Posada, C. (2007). Adaptation to Climate Change in Colombia. *Revista de Ingeniería*, *26*(2), 74–80. https://doi.org/10.1029/TR032i002p00231

Cramer, S., Kampouridis, M., Freitas, A. A., & Alexandridis, A. K. (2017). An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives. *Expert Systems with Applications*, *85*, 169–181. https://doi.org/10.1016/j.eswa.2017.05.029

De Lima, P. M., & Guedes, E. B. (2015). Rainfall prediction for Manaus, Amazonas with artificial neural networks. *2015 Latin-America Congress on Computational Intelligence, LA-CCI 2015*, 1–5. https://doi.org/10.1109/LA-CCI.2015.7435934

Decreto Ley 919 de 1989. (1997). https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=13549

Delgado, V., Zambrano, J., & Vélez, J. (2020). The knowledge of the spatial-temporal rainfall patterns as a tool for storm-design. Case study: Manizales, Colombia. *Authorea Preprints*, *1*, 1–16. https://www.authorea.com/doi/full/10.22541/au.158921470.04015184?commit=6b52eee98e5338f0ce6bf8c58b558d34410ee47e

Dimri, T., Ahmad, S., & Sharif, M. (2020). Time series analysis of climate variables using seasonal ARIMA approach. *Journal of Earth System Science*, *129*(1). https://doi.org/10.1007/s12040-020-01408-x

Donner, A., & Rosner, B. (1982). Missing value problems in multiple linear regression willi two independent Variables. *Communications in Statistics - Theory and Methods*, *11*(2), 127–140. https://doi.org/10.1080/03610928208828222

Etuk, E., & Mohamed, T. M. (2014). A Seasonal ARIMA Model for forecasting Monthly Rainfall in Gezira Scheme. *Journal of Advanced Studies in Agricultural, Biological and Environmental Sciences (JABE)*, *1*.

FAO. (2022). *Organización de las Naciones Unidas para la Alimentación y la Agricultura*.

Farajzadeh, J., & Alizadeh, F. (2018). A hybrid linear–nonlinear approach to predict the monthly rainfall over the Urmia Lake watershed using wavelet-SARIMAX-LSSVM conjugated model. *Journal of Hydroinformatics*, *20*(1), 221–231.

https://doi.org/10.2166/hydro.2017.013

Fogarty, D. J. (2006). Multiple imputation as a missing data approach to reject inference on consumer credit scoring. *Interstat*, *December 2000*, 1–41. http://interstat.statjournals.net/YEAR/2006/articles/0609001.pdf

Fox, R. A., Croxton, F. E., Cowden, D. J., & Bolch, B. W. (1973). Chapter 14 -Time Series: Understanding Changes over Time. In *Practical Business Statistics.* (Sixth Edit, Vol. 22, Issue 4, pp. 430–464). Andrew F. Siegel. https://doi.org/10.2307/2986828

Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., & Gilon, O. (2022). Deep learning rainfall – runoff predictions of extreme events. *Hydrology and Earth System Sciences*, *2019*, 3377–3392.

GitHub. (2022). https://docs.github.com/en/get-started/quickstart/hello-world

González-Hidalgo, J. C., Brunetti, M., & de Luis, M. (2011). A new tool for monthly precipitation analysis in Spain: MOPREDAS database (monthly precipitation trends December 1945-November 2005). *International Journal of Climatology*, *31*(5), 715–731. https://doi.org/10.1002/joc.2115

Gorlapalli, A., Kallakuri, S., Sreekanth, P. D., Patil, R., Bandumula, N., Ondrasek, G., Admala, M., Gireesh, C., Anantha, M. S., Parmar, B., Yadav, B. K., Sundaram, R. M., & Rathod, S. (2022). Characterization and Prediction of Water Stress Using Time Series and Artificial Intelligence Models. *Sustainability*, *14*(11), 6690. https://doi.org/10.3390/su14116690

Grajales García, J. A. (2021). *Landslide hazard assessment by climatic events in the basin of Quebrada El Rosario – Manizales using application ALICE* [Universidad Nacional de Colombia]. https://repositorio.unal.edu.co/handle/unal/80771

Gunawansyah, Liong, T. H., & Adiwijaya. (2017). Prediction and anomaly detection of rainfall using evolving neural network to support planting calender in soreang (Bandung). *2017 5th International Conference on Information and Communication Technology, ICoIC7 2017*, *0*(c). https://doi.org/10.1109/ICoICT.2017.8074671

Haidar, A., & Verma, B. (2018). A novel approach for optimizing climate features and network parameters in rainfall forecasting. *Soft Computing*, *22*(24), 8119–8130. https://doi.org/10.1007/s00500-017-2756-7

Hardoy, J., & Velásquez Barrero, L. S. (2014). Re-thinking “Biomanizales”: Addressing climate change adaptation in Manizales, Colombia. *Environment and Urbanization*, *26*(1), 53–68. https://doi.org/10.1177/0956247813518687

Hatim, M., Siddiqui, F., & Kumar, R. (2020). Addressing challenges and demands of intelligent seasonal rainfall forecasting using artificial intelligence approach. *Proceedings of International Conference on Computation, Automation and Knowledge Management, ICCAKM 2020*, 263–267. https://doi.org/10.1109/ICCAKM46823.2020.9051516

Hernández, E., Sanchez Anguix, V., Vicente, J., Palanca, J., & Nestor, D. (2016). Rainfall prediction: A Deep Learning approach. In *Hybrid Artificial Intelligent Systems* (Vol. 9648, Issue June, pp. 151–162). https://doi.org/10.1007/978-3-319-

32034-2

Hong, S., & Lynn, H. S. (2020). Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. *BMC Medical Research Methodology*, *20*(1), 1–12. https://doi.org/10.1186/s12874-020-01080-1

Horton, N. J., & Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *American Statistician*, *61*(1), 79–90. https://doi.org/10.1198/000313007X172556

Hossain, M. M., Garg, N., Anwar, A. H. M. F., Prakash, M., & Bari, M. (2021). Monthly Rainfall Prediction for Decadal Timescale using Facebook Prophet at a Catchment Level. *Hydrology and Water Resources Symposium (HWRS 2021)*, *September*, 1–14.

Htike, K. K., & Khalifa, O. O. (2010). Rainfall forecasting models using Focused Time-Delay Neural Networks. *International Conference on Computer and Communication Engineering, ICCCE'10*, *May*, 11–13. https://doi.org/10.1109/ICCCE.2010.5556806

Hung, N. Q., Babel, M. S., Weesakul, S., & Tripathi, N. K. (2009). An artificial neural network model for rainfall forecasting in Bangkok, Thailand. *Hydrology and Earth System Sciences*, *13*(8), 1413–1425. https://doi.org/10.5194/hess-13-1413-2009

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice. *Principles of Optimal Design*, 504. https://otexts.com/fpp2/

IDEAM. (2020). *SIMAC*. IDEAM. https://idea.manizales.unal.edu.co/reporte-meteorologico.html

Jaramillo Muñoz, V. D. (2021). *Evaluación de Modelos de Machine Learning para la Predicción de Crímenes en la Ciudad de Medellín*. Nacional de Colombia.

Jibril, Y. K., Abdulkarim, K., & Nathan, S. A. (2017). Time Series Analysis And Forecasting Of Monthly Rainfall Data In Zaria, Nigeria. *3rd YUMSCIC*, *November*, 310–313.

Lakshminarayan, K., Harp, S. A., & Samad, T. (1999). Imputation of missing data in industrial databases. *Applied Intelligence*, *11*(3), 259–275.

Lu, Y., Jiang, S., Ren, L., Zhang, L., Wang, M., Liu, R., & Wei, L. (2019). Spatial and Temporal variability in precipitation concentration over mainland China, 1961-2017. *Water (Switzerland)*, *11*(5). https://doi.org/10.3390/w11050881

Mahsin, M., Akhter, Y., & Begum, M. (2012). Modeling Rainfall in Dhaka Division of Bangladesh Using Time Series Analysis. *Journal of Mathematical Modelling and Application*, *1*(5), 67–73.

Markey, M. K., Tourassi, G. D., Margolis, M., & DeLong, D. M. (2006). Impact of missing data in evaluating artificial neural networks trained on complete data. *Computers in Biology and Medicine*, *36*(5), 516–525. https://doi.org/10.1016/j.compbiomed.2005.02.001

Matplotlib: Python plotting. (2022). https://matplotlib.org/

Mazzoglio, P. (2022). Insights on a global Extreme Rainfall Detection System. *In Precipitation Science*, 135–155.

Mazzoglio, P., Laio, F., Balbo, S., Boccardo, P., & Disabato, F. (2019). Improving an extreme rainfall detection system with GPM IMERG data. *Remote Sensing*, *11*(6), 1–24. https://doi.org/10.3390/rs11060677

Min, M., Bai, C., Guo, J., Sun, F., Liu, C., Wang, F., Xu, H., Tang, S., Li, B., Di, D., Dong, L., & Li, J. (2019). Estimating Summertime Precipitation from Himawari-8 and Global Forecast System Based on Machine Learning. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(5), 2557–2570. https://doi.org/10.1109/TGRS.2018.2874950

Mitchell, T., Buchanan, B., DeJong, G., Dietterich, T., Rosenbloom, P., & Waibel, A. (2017). Machine learning. *Annual Review of Computer Science*, *45*(13), 417–433. https://books.google.ca/books?id=EoYBngEACAAJ&dq=mitchell+machine+learning+1997&hl=en&sa=X&ved=0ahUKEwiomdqfj8TkAhWGslkKHRCbAtoQ6AEIKjAA

Moreno Cadavid, J., Hernández leal, E. J., & Duque Méndez, N. D. (2016). Generación de pronósticos para la precipitación diaria en una serie de tiempo de datos meteorológicos. *Ingenio Magno,* 144–155.

Murat, M., Malinowska, I., Gos, M., & Krzyszczak, J. (2018). Forecasting daily meteorological time series using ARIMA and regression models. *International Agrophysics*, *32*(2), 253–264. https://doi.org/10.1515/intag-2017-0007

Mushtaq, R. (2011). Augmented Dickey Fuller Test. *SSRN Electronic Journal*, 1–19. https://doi.org/10.1177/000331979604700708

NumPy. (2019). https://www.numpy.org/

OO, Z. Z., & PHYU, S. (2020). Time Series Prediction Based on Facebook Prophet: A Case Study, Temperature Forecasting in Myintkyina. *International Journal of Applied Mathematics Electronics and Computers*, *8*(4), 263–267. https://doi.org/10.18100/ijamec.816894

Pachón Gómez, J. A., Mejía Fernández, F., & Zambrano Nájera, J. (2018a). Sistema Integrado de Monitoreo Ambiental de Caldas-SIMAC. In *Boletín Ambiental* (pp. 1–9).

Pachón Gómez, J. A., Mejía Fernández, F., & Zambrano Nájera, J. D. C. (2018b). Sistema Integrado de Monitoreo Ambiental de Caldas-SIMAC. In *Boletín Ambiental* (Vol. 145).

Pandas. (2022). https://pandas.pydata.org/docs/

Pantanowitz, A., & Marwala, T. (2009). Missing data imputation through the use of the random forest algorithm. *Advances in Intelligent and Soft Computing*, *61 AISC*, 53–62. https://doi.org/10.1007/978-3-642-03156-4_6

Python Data Analysis Library - pandas. (2022). https://pandas.pydata.org/

Rafferty, G., & Safari, an O. M. C. (2021). *Forecasting Time Series Data with Facebook Prophet* (S. Editing (Ed.)). Sunith Shetty.

Rokach, L. (2016). Decision forest: Twenty years of research. *Information Fusion*, *27*, 111–125. https://doi.org/10.1016/j.inffus.2015.06.005

Rudolf, B., Beck, C., Grieser, J., & Schneider, U. (2005). Global Precipitation Analysis Products of the GPCC. *Global Precipitation Climatology Centre (GPCC)*, *112*, 1–8. ftp://ftp-anon.dwd.de/pub/data/gpcc/PDF/GPCC_intro_products_2008.pdf

Samad, M. D., Abrar, S., & Diawara, N. (2022). Missing value estimation using clustering and deep learning within multiple imputation framework. *Knowledge-Based Systems*, *249*(Mvi). https://doi.org/10.1016/j.knosys.2022.108968

Samad, M. D., & Yin, L. (2019). Non-linear regression models for imputing longitudinal missing data. *2019 IEEE International Conference on Healthcare Informatics, ICHI 2019*, 1–3. https://doi.org/10.1109/ICHI.2019.8904528

Sarraf, A., Vahdat, S. F., & Behbahaninia, A. (2011). Relative Humidity and Mean Monthly Temperature Forecasts in Ahwaz Station with ARIMA Model in time Series Analysis. *2011 International Conference on Environment and Industrial Innovation IPCBEE*, *12*, 149–153.

Sawsan M, A. (2013). Time Series Analysis of Baghdad Rainfall Using ARIMA Method. *Iraqi Journal of Science*, *54*(4), 1136–1142.

Schelter, S., Biessmann, F., Januschowski, T., Salinas, D., Seufert, S., & Szarvas, G. (2018). On Challenges in Machine Learning Model Management. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 5–13. http://sites.computer.org/debull/A18dec/p5.pdf

Schmidt, J., Turek, G., Clark, M. P., Uddstrom, M., & Dymond, J. R. (2008). Probabilistic forecasting of shallow, rainfall-triggered landslides using real-time numerical weather predictions. *Natural Hazards and Earth System Science*, *8*(2), 349–357. https://doi.org/10.5194/nhess-8-349-2008

Sci-kit learn: machine learning in Python. (2022). https://scikit-learn.org/stable/

SciPy.org. (n.d.). https://www.scipy.org/

Singh, P. (2018). Indian summer monsoon rainfall (ISMR) forecasting using time series data: A fuzzy-entropy-neuro based expert system. *Geoscience Frontiers*, *9*(4), 1243–1257. https://doi.org/10.1016/j.gsf.2017.07.011

Snoek, J., Larochelle, H., & Adams, R. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. *Advances in Neural Information Processing Systems*, *5*, 1–9. https://doi.org/10.1163/15685292-12341254

Ssali, G., & Marwala, T. (2008). Computational intelligence and decision trees for missing data estimation. *Proceedings of the International Joint Conference on Neural Networks*, 201–207. https://doi.org/10.1109/IJCNN.2008.4633790

StatsModels: Statistics in Python. (2022). https://www.statsmodels.org/stable/index.html

Stekhoven, D. J. (2012). *Using the missForest Package*. https://stat.ethz.ch/education/semesters/ss2013/ams/paper/missForest_1.2.pdf

Stekhoven, D. J., & Bühlmann, P. (2012). Missforest-Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Stitou, A. (2019). *SARIMA short to medium-term forecasting and stochastic simulation of streamflow , water levels and sediments time series from the HYDAT database*. Ottawa.

Suhaila, J., Deni, S. M., Zawiah Zin, W. A. N., & Jemain, A. A. (2010). Trends in Peninsular Malaysia rainfall data during the southwest monsoon and northeast monsoon seasons: 1975-2004. *Sains Malaysiana*, *39*(4), 533–542.

Sulaiman, J. (2015). *Heavy Precipitation Forecasting using the Combination of Local and Global Modes with Application to Malaysian Rainfall* (Issue December) [KYUSHU INSTITUTE OF TECHNOLOGY]. http://hdl.handle.net/10228/5457

Talwar, P. P. (1990). The analysis of time series: An introduction with R. In *International Journal of Forecasting* (Vol. 6, Issue 4). https://doi.org/10.1016/0169-2070(90)90041-9

Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining*, *10*(6), 363–377. https://doi.org/10.1002/sam.11348

Taylor, S. J., & Letham, B. (2017). Forecasting at Scale Sean. *PeerJ Preprints 5:E3190v2*, *35*(8), 48–90. https://peerj.com/preprints/3190/%0Ahttp://ezproxy.bangor.ac.uk/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=108935824&site=ehost-live%0Ahttps://peerj.com/preprints/3190/%0Ahttps://peerj.com/preprints/3190.pdf

Triebe, O., Hewamalage, H., Pilyugina, P., Laptev, N., Bergmeir, C., & Rajagopal, R. (2021). *NeuralProphet: Explainable Forecasting at Scale*. 1–40. http://arxiv.org/abs/2111.15397

Tularam, G. A., & Ilahee, M. (2010). Time series analysis of rainfall and temperature interactions in coastal catchments. *Journal of Mathematics and Statistics*, *6*(3), 372–380. https://doi.org/10.3844/jmssp.2010.372.380

Twala, B. (2009). An empirical comparison of techniques for handling incomplete data using decision trees. *Applied Artificial Intelligence*, *23*(5), 373–405. https://doi.org/10.1080/08839510902872223

Vagropoulos, S. I., Chouliaras, G. I., Kardakos, E. G., Simoglou, C. K., & Bakirtzis, A. G. (2016). Comparison of SARIMAX, SARIMA, Modified SARIMA and ANN-based Models for Short-Term PV Generation Forecasting. *2016 IEEE International Energy Conference, ENERGYCON 2016*, 1–6.

van Westen, C. J., & Erlien, M. T. J. (1995). An approach towards deterministic landslide hazard analysis in GIS : a case study from Manizales, Colombia. *Earth Surface Processes and Landforms*, *21*, 853–868.

Vélez, J. J., Mejía, F., Pachón, A., & Vargas, D. (2010). An Operative Warning System of Rainfall-Triggered Landslides at Manizales, Colombia. *Proceedings of World Water Congress and Exhibition IWA*, *1*, 19–24.

Vélez Upegui, J. J., Orozco Alzate, M., Darío, D. M. N., & Aristizabal Zuluaga, B. H. (2015). Entendimiento de fenómenos ambientales mediante análisis. In *Universidad Nacional de Colombia* (Issue 1).

Wang, S., Feng, J., & Liu, G. (2013). Application of seasonal time series model in the precipitation forecast. *Mathematical and Computer Modelling*, *58*(3–4), 677–683. https://doi.org/10.1016/j.mcm.2011.10.034

Wei, W. W. S. (1991). Time Series Analysis: Univariate and Multivariate Methods. In *International Journal of Forecasting* (pearson, Vol. 33, Issue 1). https://doi.org/10.2307/1269015

Welcome to Python.org. (2019). https://www.python.org/

Zadranska, B. L. (2019). *Time Series Forecasting using Deep Neural Networks*. West Bohemia.

Zakaria, S., Al-Ansari, N., Knutsson, S., & Al-Badrany, T. (2012). ARIMA Models for weekly rainfall in the semi-arid Sinjar District at Iraq. *Journal of Earth Sciences and Geotechnical Engineering*, *2*(3), 1792–9660.

Zhang, G., Eddy Patuwo, B., & Y. Hu, M. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, *14*(1), 35–62. https://doi.org/10.1016/S0169-2070(97)00044-7

Zhang, S., Gong, L., Zeng, Q., Li, W., Xiao, F., & Lei, J. (2021). Imputation of GPS coordinate time series using missforest. *Remote Sensing*, *13*(12), 1–17. https://doi.org/10.3390/rs13122312

Apéndice A: Resultados de las Métricas

A continuación, se presentan los resultados de las métricas de todos los modelos respecto a cada estación meteorológica de manera individual y el promedio de estos. De estas Tablas se obtuvieron todos los resultados presentados en la investigación.

BASELINE

$ARIMA_{101}$	RMSE		MAE		MSE	
	Train	Test	Train	Test	Train	Test
Alcázares	19.04523	7.38988	11.64470	6.29880	362.72091	54.61047
Aranjuez	16.52149	20.74056	10.47960	17.68507	272.95995	430.17110
Bosques del norte	22.12685	19.44708	14.62620	18.60872	489.59761	378.18903
CHEC Uribe	17.34438	47.26288	11.03696	40.79465	300.82778	2233.77994
El Carmen	18.18975	18.97653	11.84429	16.98898	330.86732	360.10902
EMAS	23.15731	14.45541	13.62234	11.01883	536.26101	208.95907
Hospital de Caldas	18.77143	23.01478	11.75196	21.19977	352.36664	529.68042
La Nubia	15.51603	20.89552	9.93065	17.59764	240.74746	436.62298
La Palma	21.34193	44.32053	13.83165	38.62917	455.47807	1964.30970
Milán	17.42475	34.95621	11.49359	31.16182	303.62195	1221.93682
Obs. Vulcanológico	21.52523	42.46000	13.51621	35.81760	463.33591	1802.85186
Posgrados	17.18378	27.39093	11.11624	24.37869	295.28259	750.26339
Yarumos	20.01349	30.65941	12.76668	27.92550	400.53998	939.99957
Total	19.08935	27.07459	12.12777	23.70040	369.58516	870.11410

$ARIMA_{101}$	AIC	BIC
Alcázares	12703.491	12724.619
Aranjuez	12290.453	12311.581
Bosques del norte	13140.124	13161.252
CHEC Uribe	12431.833	12452.962
El Carmen	12570.134	12591.262
EMAS	13272.440	13293.569
Hospital de Caldas	12661.801	12682.929
La Nubia	12107.954	12129.082
La Palma	13035.004	13056.132
Milán	12445.371	12466.499
Obs. Vulcanológico	13059.920	13081.048
Posgrados	12404.884	12426.012
Yarumos	12848.082	12869.210
Total	12690.11469	12711.24285

$SARIMA_{101}$	RMSE		MAE		MSE	
	Train	Test	Train	Test	Train	Test
Alcázares	19.16837	4.10226	11.56153	3.41439	367.42676	16.82860
Aranjuez	16.65654	21.33223	10.49608	17.53755	277.65654	455.06424
Bosques del norte	22.31179	19.72759	14.45770	18.73137	497.81625	389.17788
CHEC Uribe	17.51454	47.90993	10.97008	41.20902	306.75924	2295.36166
El Carmen	18.35492	19.83609	11.81873	17.36917	336.90338	393.47050
EMAS	23.30595	14.52739	13.28614	11.16841	543.16743	211.04524
Hospital de Caldas	18.95055	24.31207	11.81043	21.79387	359.12368	591.07681
La Nubia	15.64865	21.52036	9.89061	17.71302	244.88053	463.12610
La Palma	21.50257	46.03678	13.78845	39.22254	462.36006	2119.38544
Milán	17.54502	36.80475	11.44389	32.12092	307.82803	1354.59031
Obs. Vulcanológico	21.71144	42.80361	13.54312	35.62680	471.38704	1832.14951
Posgrados	17.34018	28.12049	11.11566	24.79143	300.68209	790.76219
Yarumos	20.13390	33.18634	12.69645	29.16827	405.37428	1101.33345
Total	19.24187	27.70922	12.06760	23.83590	375.48963	924.10553

$SARIMA_{101}$	AIC	BIC
Alcázares	12721.500	12747.910
Aranjuez	12311.800	12338.210
Bosques del norte	13166.508	13192.918
CHEC Uribe	12457.830	12484.241
El Carmen	12594.254	12620.664
EMAS	12620.664	13322.970
Hospital de Caldas	12685.915	12712.325
La Nubia	12712.325	12158.983
La Palma	13056.289	13082.700
Milán	12466.177	12466.177
Obs. Vulcanológico	13082.110	13108.520
Posgrados	12429.662	12429.662
Yarumos	12866.795	12893.205
Total	12705.52531	12735.26808

$SARIMAX_{101}$	RMSE		MAE		MSE	
	Train	Test	Train	Test	Train	Test
Alcázares	19.26293	6.65930	11.47276	5.76696	371.06085	44.34628
Aranjuez	16.68659	21.54932	10.35026	17.61971	278.44242	464.37356
Bosques del norte	22.33622	19.77186	14.48176	18.74134	498.90711	390.92668
CHEC Uribe	17.47786	48.22373	10.95048	41.42915	305.47575	2325.52835
El Carmen	18.41131	20.16842	11.54598	17.85512	338.97641	406.76525
EMAS	23.36746	15.15297	13.49530	12.33043	546.03843	229.61252
Hospital de Caldas	18.96286	24.52644	11.61428	22.06734	359.59018	601.54639
La Nubia	15.66570	21.10745	9.74643	17.12321	245.41432	445.52457
La Palma	21.66438	46.97505	13.56353	40.08503	469.34546	2206.65624
Milán	17.61440	37.43592	11.29083	32.42567	310.26710	1401.44856
Obs. Vulcanológico	21.82789	45.25755	13.30253	37.45176	476.45691	2048.24647
Posgrados	17.40981	29.77766	10.93109	25.52601	303.10172	886.70942
Yarumos	20.16850	31.84065	12.60461	28.40423	406.76851	1013.82720
Total	19.29660	28.34202	11.94998	24.37122	377.68039	958.88549

$SARIMAX_{101}$	AIC	BIC
Alcázares	12740.528	12772.220
Aranjuez	12323.410	12355.102
Bosques del norte	13171.595	13171.595
CHEC Uribe	13171.595	12489.824
El Carmen	12609.404	12641.097
EMAS	13302.966	13334.659
Hospital de Caldas	12695.338	12727.030
La Nubia	12139.916	12171.608
La Palma	13082.629	13114.321
Milán	13114.321	13114.321
Obs. Vulcanológico	13104.537	13136.229
Posgrados	12446.904	12478.596
Yarumos	12874.625	12906.318
Total	12829.05908	12800.99385

Prophet$_{Días}$	RMSE		MAE		MSE	
	Train	Test	Train	Test	Train	Test
Alcázares	20.23569	3.93692	12.85363	3.49126	409.48341	15.49935
Aranjuez	17.66646	20.86014	11.65279	16.81111	312.10386	435.14562
Bosques del norte	25.52439	14.87858	17.45894	13.15076	651.49463	221.37227
CHEC Uribe	18.61676	42.24147	12.31722	36.74199	346.58390	1784.34224
El Carmen	19.46447	17.26687	13.18885	15.06247	378.86594	298.14485
EMAS	26.43306	12.50682	16.17547	10.28351	698.70678	156.42067
Hospital de Caldas	19.78608	20.97285	13.19078	18.97635	391.48904	439.86063
La Nubia	16.66537	20.14463	10.97902	17.14227	277.73474	405.80624
La Palma	23.44149	42.68405	15.59386	36.07797	549.50384	1821.92883
Milán	18.87792	33.78936	12.95930	29.53320	356.37589	1141.72107
Obs. Vulcanológico	23.92587	40.25925	15.56529	32.84193	572.44769	1620.80776
Posgrados	18.97215	24.38387	12.85654	21.16530	359.94261	594.57325
Yarumos	21.20757	29.99880	14.12711	26.54546	449.76114	899.93044
Total	20.83209	24.91720	13.76298	21.37104	442.65334	756.58101

Neural Prophet	RMSE		MAE		MSE	
	Train	Test	Train	Test	Train	Test
Alcázares	20.55027	3.92471	12.54678	2.76428	422.31360	15.40339
Aranjuez	18.10319	24.38263	11.41124	18.73758	327.72557	594.51285
Bosques del norte	25.59176	21.36790	17.34358	18.71033	654.93866	456.58715
CHEC Uribe	18.76740	52.39486	11.85391	42.51762	352.21558	2745.21894
El Carmen	19.55821	21.62232	12.81569	18.09104	382.52393	467.52478
EMAS	26.57124	15.81873	15.52743	12.67200	706.03128	250.23222
Hospital de Caldas	20.48638	27.34354	12.98058	22.55857	419.69211	747.66952
La Nubia	16.83647	21.92441	10.66775	17.35643	283.46686	480.67985
La Palma	23.50353	50.90579	15.20491	41.13554	552.41627	2591.39964
Milán	19.22650	40.02001	12.88064	33.39986	369.65852	1601.60163
Obs. Vulcanológico	24.15462	47.26254	14.91209	37.13824	583.44606	2233.74853
Posgrados	19.28529	32.83318	12.76155	27.30391	371.92273	1078.01793
Yarumos	21.45592	35.19461	13.79259	29.41863	460.35656	1238.66111
Total	21.08390	30.38424	13.43836	24.75415	452.82367	1115.48134

Optimización de hiperparámetros

ARIMA d = 1 Seasonal = False	Parámetro (pdq)	RMSE	MAE	MSE
Alcázares	(0,1,2)	4.62209	4.12322	21.36379
Aranjuez	(0,1,3)	18.39719	15.40591	338.45680
Bosques del norte	(1,1,5)	16.54704	15.26389	273.80461
CHEC Uribe	(0,1,3)	44.76434	37.219585	2003.84649
El Carmen	(0,1,4)	18.2231	15.338025	332.08303
EMAS	(0.1.3)	11.60278	8.63148	134.62446
Hospital de Caldas	(0,1,3)	23.24900	19.76783	540.51624
La Nubia	(2,1,1)	18.68633	15.68663	344.78253
La Palma	(0,1,3)	43.69625	35.96120	1909.36287
Milán	(1,1,5)	31.03473	26.62790	963.15466
Obs. Vulcanológico	(0,1,3)	41.46059	32.80208	1718.98132
Posgrados	(1,1,5)	26.59478	21.98038	707.28257
Yarumos	(0,1,4)	30.20289	26.15991	912.21471
Total		25.31393	21.15138	784.65185

ARIMA d = 1 Seasonal = False	AIC	BIC
Alcázares	12741.107	12756.954
Aranjuez	12318.532	12339.660
Bosques del norte	13160.812	13197.786
CHEC Uribe	12463.783	12484.911
El Carmen	12576.501	12602.911
EMAS	13303.231	13303.231
Hospital de Caldas	12678.303	12699.431
La Nubia	12114.761	12135.890
La Palma	13065.126	13086.255
Milán	12453.798	12490.772
Obs. Vulcanológico	13096.747	13117.876
Posgrados	12412.734	12449.708
Yarumos	12839.372	12865.782
Total	12709.60054	12733.16669

ARIMA d = 0 Seasonal = True	Parámetro (pdq)	RMSE	MAE	MSE
Alcázares	(100)	7.47173	6.44317	55.82679
Aranjuez	(200)	18.07889	15.49646	326.84650
Bosques del norte	(101)	16.59625	15.30626	275.43558
CHEC Uribe	(101)	43.47141	36.48689	1889.76429
El Carmen	(101)	16.77893	14.43663	281.53257
EMAS	(100)	11.73780	8.89895	137.77597
Hospital de Caldas	(100)	21.16828	18.71273	448.09638
La Nubia	(400)	18.45964	15.90361	340.75846
La Palma	(201)	41.67653	35.19205	1736.93333
Milán	(303)	29.75427	26.09933	885.31662
Obs. Vulcanológico	(200)	39.50590	31.97134	1560.71687
Posgrados	(100)	24.59690	21.06550	605.00755
Yarumos	(201)	28.92923	25.36730	836.90051
Total		24.47890	20.87540	721.60857

ARIMA d = 0 Seasonal = True	AIC	BIC
Alcázares	12709.733	12725.581
Aranjuez	12298.544	12319.675
Bosques del norte	13148.460	13169.591
CHEC Uribe	12441.722	12462.853
El Carmen	12578.333	12599.464
EMAS	13279.231	13295.080
Hospital de Caldas	12668.201	12684.050
La Nubia	12112.104	12143.801
La Palma	13042.230	13068.644
Milán	12444.899	12487.161
Obs. Vulcanológico	13067.422	13088.553
Posgrados	12412.024	12427.872
Yarumos	12846.948	12873.362
Total	12696.14238	12718.899

$SARIMA_{D=1, m=12}$	Parámetro (pdq)(PDQ)	RMSE	MAE	MSE
Alcázares	(1,0,1) (2,1,0)	6.54005	5.39958	42.77238
Aranjuez	(2,0,0) (2,1,0)	20.89700	17.33935	436.68478
Bosques del norte	(1,0,1) (2,1,0)	20.19421	17.25601	407.80614
CHEC Uribe	(1,0,1) (2,1,0)	46.18829	37.50568	2133.35845
El Carmen	(1,0,1) (2,1,0)	19.35651	16.17150	374.674535
EMAS	(2,0,1) (2,1,0)	23.61415	16.54139	557.62834
Hospital de Caldas	(1,0,0) (2,1,0)	23.13737	18.83558	535.33830
La Nubia	(4,0,1) (2,1,0)	19.25604	16.97051	370.79526
La Palma	(1,0,1) (2,1,0)	44.64164	34.70080	1992.87645
Milán	(1,0,0) (2,1,0)	31.11186	26.50276	967.94811
Obs. Vulcanológico	(3,0,1) (2,1,0)	43.23208	32.63767	1869.01287
Posgrados	(1,0,0) (2,1,0)	26.35558	21.15022	694.61684
Yarumos	(1,0,0) (2,1,0)	29.35518	24.13100	861.72682
Total		27.22153	21.93400	865.01840

$SARIMA_{D=1, m=12}$	AIC	BIC
Alcázares	12989.667	13016.039
Aranjuez	12608.854	12635.227
Bosques del norte	13491.974	13518.346
CHEC Uribe	12761.829	12788.201
El Carmen	12886.771	12913.143
EMAS	13574.926	13611.848
Hospital de Caldas	12965.397	12986.495
La Nubia	12438.279	12480.475
La Palma	13341.321	13367.693
Milán	12787.489	12808.587
Obs. Vulcanológico	13351.892	13394.088
Posgrados	12728.548	12749.646
Yarumos	13149.887	13170.985
Total	13005.91031	13033.90562

$SARIMA_{D=0, m=12}$	Parámetro (pdq)(PDQ)	RMSE	MAE	MSE
Alcázares	(500)	6.77978	5.91105	45.96547
Aranjuez	(200)	18.07889	15.49646	326.84650
Bosques del norte	(101)	16.59625	15.30626	275.43558
CHEC Uribe	(101)	43.47141	36.48689	1889.76429
El Carmen	(101)	16.77893	14.43663	281.53257
EMAS	(100)	11.73780	8.89895	137.77597
Hospital de Caldas	(100)	21.16828	18.71273	448.09638
La Nubia	(400)	18.45964	15.90361	340.75846
La Palma	(201)	41.67653	35.19205	1736.93333
Milán	(501)	30.89478	26.89785	954.48802
Obs. Vulcanológico	(200)	39.50590	31.97134	1560.71687
Posgrados	(104)	25.41674	21.45580	646.01082
Yarumos	(201)	28.92923	25.36730	836.90051
Total		24.57647	20.92591	729.32498

$SARIMA_{D=0, m=12}$	AIC	BIC
Alcázares	12703.305	12703.305
Aranjuez	12703.305	12319.675
Bosques del norte	13148.460	13169.591
CHEC Uribe	12441.722	12462.853
El Carmen	12578.333	12599.464
EMAS	13279.231	13295.080
Hospital de Caldas	12668.201	12684.050
La Nubia	12112.104	12143.801
La Palma	13042.230	13068.644
Milán	12452.166	12494.428
Obs. Vulcanológico	13067.422	13088.553
Posgrados	12413.105	12450.084
Yarumos	12846.948	12873.362
Total	12727.42554	12719.45308

$SARIMAX_{D=1, m=6}$	Parámetro (pdq)(PDQ)	RMSE	MAE	MSE
Alcázares	(2,0,1) (2,1,0)	4.97102	4.10065	24.71112
Aranjuez	(3,0,0) (2,1,0)	21.28770	16.78616	453.16622
Bosques del norte	(1,0,1) (2,1,0)	18.76362	16.64081	352.07368
CHEC Uribe	(2,0,0) (2,1,0)	50.50016	41.94078	2550.26628
El Carmen	(1,0,0) (2,1,0)	22.44337	18.41850	503.70511
EMAS	(2,0,0) (2,1,0)	14.60499	11.83781	213.30583
Hospital de Caldas	(1,0,0) (2,1,0)	25.43311	20.65454	646.84309
La Nubia	(4,0,1) (2,1,0)	21.13192	17.38335	446.55814
La Palma	(1,0,0) (2,1,0)	48.12478	39.67678	2315.99475
Milán	(1,0,0) (2,1,0)	35.88494	30.47442	1287.72893
Obs. Vulcanológico	(3,0,0) (2,1,0)	43.19784	34.30772	1866.05411
Posgrados	(1,0,0) (2,1,0)	29.51777	23.90632	871.29920
Yarumos	(1,0,0) (2,1,0)	32.41692	27.18036	1050.85679
Total		28.32908	23.33140	967.88948

$SARIMAX_{D=1, m=6}$	AIC	BIC
Alcázares	13053.937	13090.887
Aranjuez	12670.784	12670.784
Bosques del norte	13513.071	13539.464
CHEC Uribe	12816.651	12843.044
El Carmen	12952.238	12973.352
EMAS	13657.899	13684.292
Hospital de Caldas	13030.453	13051.568
La Nubia	12433.401	12475.630
La Palma	13394.201	13415.316
Milán	12794.172	12815.286
Obs. Vulcanológico	13433.232	13464.904
Posgrados	13464.904	12764.755
Yarumos	12764.755	13219.212
Total	13075.36138	13077.57646

$SARIMAX_{D=0, m=6}$	Parámetro (pdq)(PDQ)	RMSE	MAE	MSE
Alcázares	(1,0,0) (2,0,2)	7.42228	6.58934	55.09025
Aranjuez	(2,0,0) (0,0,0)	18.07889	15.49646	326.84650
Bosques del norte	(1,0,1) (1,0,1)	15.63422	14.23603	244.42912
CHEC Uribe	(1,0,1) (0,0,0)	43.47141	36.48689	1889.76429
El Carmen	(1,0,1) (1,0,1)	16.66133	14.20656	77.60004
EMAS	(1,0,0) (1,0,1)	12.69878	9.40885	161.25904
Hospital de Caldas	(1,0,0) (2,0,0)	21.08364	18.63134	444.51999
La Nubia	(4,0,0) (1,0,1)	18.26022	15.68513	333.43570
La Palma	(2,0,1) (1,0,1)	41.56887	35.12566	1727.97168
Milán	(1,0,0) (0,0,1)	30.66562	26.86484	940.38079
Obs. Vulcanológico	(2,0,0) (0,0,0)	39.50590	31.97134	1560.71687
Posgrados	(1,0,0) (1,0,0)	24.78278	21.27460	614.18657
Yarumos	(1,0,4) (0,0,2)	28.76648	25.52895	827.51050
Total		24.50772	20.88507	707.97779

$SARIMAX_{D=0, m=6}$	AIC	BIC
Alcázares	12699.873	12736.852
Aranjuez	12298.544	12319.675
Bosques del norte	13142.846	13174.543
CHEC Uribe	12441.722	12462.853
El Carmen	12572.773	12604.469
EMAS	13276.367	13302.781
Hospital de Caldas	12666.531	12692.945
La Nubia	12110.545	12152.807
La Palma	13041.430	13078.409
Milán	12443.803	12464.934
Obs. Vulcanológico	13067.422	13088.553
Posgrados	12409.232	12430.364
Yarumos	12841.106	12888.651
Total	12693.24569	12722.91046

Prophet$_{días}$ Multivariado 3 variables	RMSE	MAE	MSE
Alcázares	4.83548	3.99462	23.38189
Aranjuez	17.91290	16.34192	320.87215
Bosques del norte	11.84958	10.25986	140.41266
CHEC Uribe	37.22164	32.21360	1385.45109
El Carmen	11.47092	10.44819	131.58201
EMAS	12.45705	11.52635	155.17824
Hospital de Caldas	14.14326	12.81083	200.03180
La Nubia	21.54910	19.07234	464.36403
La Palma	30.57094	25.23587	934.58246
Milán	22.59212	20.78564	510.40426
Obs. Vulcanológico	30.49123	23.62436	929.71518
Posgrados	16.27749	14.62904	264.95673
Yarumos	22.86561	21.07850	522.83634
Total	19.55671	17.07854	460.28991

Prophet$_{días}$ Multivariado 5 variables	RMSE	MAE	MSE
Alcázares	5.41035	4.67015	29.27193
Aranjuez	17.90099	15.43663	320.44564
Bosques del norte	13.52470	11.44899	182.91759
CHEC Uribe	34.54073	28.71133	1193.06204
El Carmen	11.05293	10.18656	122.16736
EMAS	12.69245	11.47958	161.09848
Hospital de Caldas	13.62043	11.87247	185.51613
La Nubia	18.62989	16.33273	347.07304
La Palma	30.04948	23.89970	902.97172
Milán	21.68560	19.83185	470.26563
Obs. Vulcanológico	31.24599	24.24208	976.31208
Posgrados	16.93534	14.75917	286.80579
Yarumos	20.53193	18.25710	421.56019
Total	19.06313	16.24064	430.72827

Prophet$_{días}$ Multivariado 5 variables - multiplicative	RMSE	MAE	MSE
Alcázares	5.85429	5.30230	34.27278
Aranjuez	17.38153	15.34642	302.11782
Bosques del norte	14.24503	12.46093	202.92101
CHEC Uribe	32.63353	27.62855	1064.94765
El Carmen	11.61025	10.65575	134.79802
EMAS	11.07048	9.22898	122.55571
Hospital de Caldas	14.87896	13.21341	221.38356
La Nubia	19.47771	16.46931	379.38152
La Palma	35.96843	29.83637	1293.72867
Milán	25.00545	22.66388	625.27300
Obs. Vulcanológico	33.29680	25.78774	1108.67702
Posgrados	21.41530	18.47778	458.61545
Yarumos	24.77982	22.13215	614.03985
Total	20.58596	17.63104	504.82400

Neural Prophet$_{días}$ Multivariado 3 variables	RMSE	MAE	MSE
Alcázares	5.86964	4.03487	34.45278
Aranjuez	24.96135	18.49130	623.06919
Bosques del norte	21.60358	17.18507	466.71485
CHEC Uribe	51.85426	41.81837	2688.86531
El Carmen	18.74971	13.95857	351.55184
EMAS	18.09899	13.64652	327.57373
Hospital de Caldas	24.72311	20.09006	611.23217
La Nubia	25.39538	19.40880	644.92573
La Palma	47.46060	36.92036	2252.50891
Milán	40.60026	34.58264	1648.38138
Obs. Vulcanológico	43.96202	33.30651	1932.65960
Posgrados	28.65661	23.66023	821.20159
Yarumos	32.88786	26.52999	1081.61195
Total	29.60179	23.35640	1037.28838

Neural Prophet$_{días}$ Multivariado 5 variables	RMSE	MAE	MSE
Alcázares	6.01627	4.05771	36.19559
Aranjuez	22.84651	17.22822	521.96343
Bosques del norte	23.88094	19.68647	570.29964
CHEC Uribe	50.26313	40.69282	2526.38261
El Carmen	13.60786	10.93002	185.17387
EMAS	17.61144	12.46849	310.16304
Hospital de Caldas	22.99008	18.64608	528.54407
La Nubia	23.01923	17.34612	529.88504
La Palma	43.39826	34.23455	1883.40961
Milán	40.81324	34.75669	1665.72081
Obs. Vulcanológico	43.70289	33.02665	1909.94308
Posgrados	26.72378	22.18140	714.16083
Yarumos	34.36183	27.91902	1180.73537
Total	28.40272	22.55186	966.35207

Neural Prophet$_{días}$ Multivariado 5 variables multiplicative	RMSE	MAE	MSE
Alcázares	5.42631	3.70028	29.44486
Aranjuez	23.94104	23.94104	573.17352
Bosques del norte	17.25135	15.21367	297.60928
CHEC Uribe	50.00718	40.59123	2500.71900
El Carmen	16.90994	11.92268	285.94614
EMAS	17.96168	14.22271	322.62195
Hospital de Caldas	20.66368	17.36396	426.98778
La Nubia	23.10085	23.10085	533.64945
La Palma	44.33557	35.50367	1965.64281
Milán	40.23606	34.26976	1618.94076
Obs. Vulcanológico	42.84453	32.43004	1835.65414
Posgrados	24.00846	20.32252	576.40631
Yarumos	31.71984	25.63292	1006.14859
Total	27.56973	22.93964	920.99573

Neural Prophet$_{días}$ Multivariado 5 variables multiplicative $_{LR = 0.1}$	RMSE	MAE	MSE
Alcázares	6.16094	4.39562	37.95721
Aranjuez	24.46881	18.08341	598.72309
Bosques del norte	20.714846	16.45665	429.10487
CHEC Uribe	50.52413	40.87402	2552.68838
El Carmen	18.41356	13.70826	339.05934
EMAS	17.825177	17.825177	317.73697
Hospital de Caldas	24.35600	19.85856	593.21488
La Nubia	26.24500	20.15227	688.80043
La Palma	46.21619	35.81849	2135.93623
Milán	39.58140	33.71173	1566.68789
Obs. Vulcanológico	42.79181	32.47854	1831.13942
Posgrados	27.09966	22.42052	734.39185
Yarumos	32.05373	25.92828	1027.44196
Total	28.95778869	23.208579	988.6832708