PRONÓSTICO DE LA PRECIPITACIÓN PARA LA ZONA DE INFLUENCIA DE LA ESTACIÓN AGROCLIMÁTICA YARIGUIES, UTILIZANDO TÉCNICAS DE MACHINE LEARNING

PRECIPITATION FORECAST FOR THE AREA OF INFLUENCE OF THE YARIGUIES AGROCLIMATIC STATION, USING MACHINE LEARNING TECHNIQUES

Diego Fernando Naranjo Polania, dfnaranjop@libertadores.edu.co, Fundación Universitaria los libertadores, José John Fredy González Veloza, jjgonzalezv02@libertadores.edu.co, Fundación Universitaria Los Libertadores

RESUMEN:

Pronosticar la precipitación es ideal porque ayuda a la planeación de la actividad agrícola y humana, en la actividad agronómica se podría determinar si los requerimientos hídricos de un cultivo se van a presentar y así no perder una cosecha, o conocer cuántos milímetros se necesitan conseguir para mantener el cultivo hidratado, por otro lado, como sociedad nos interesa, porque se puede determinar cuándo se presentarán precipitaciones fuertes o torrenciales que conlleven a una inundación o a deslizamientos del suelo que pongan en peligro la vida. Para el análisis estadístico se tomaron los datos de la estación agroclimática Yariguies ubicada en el municipio de Barrancabermeja, departamento de Santander, país Colombia, la serie de tiempo está comprendida entre el 01/07/1967 a 30/09/2009, la unidad de la variable precipitación es milímetros (mm), en total fueron 19266 datos, de los cuales 15.412 (80%) se utilizaron para entrenamiento y 3.854 (20%) para probar el modelo, los

modelos elaborados fueron Holt Winters, Árboles de decisión y una Red Neuronal secuencial (GRU), las métricas utilizadas fueron el MAE, MSE y RMSE para los modelos, destacándose la red neuronal GRU con 0,05, 0,01 y 0,1 mm respectivamente, sin embargo las lluvias fuertes (20-70 mm), intensas (70-150 mm) y torrenciales (>150 mm) no se observan en la figura porque el error es más alto de lo esperado, con el árbol de decisión se logró predecir lluvias fuertes, intensas y torrenciales pero el ajuste del modelo no es adecuado; a pesar que la predicción realizada por el modelo tiende a tener un comportamiento similar a los datos de reales, posiblemente porque los datos de precipitación no son lineales en la naturaleza ya que la cantidad, la frecuencia y la intensidad son tres características principales de las series de tiempo de lluvia y los valores varían por la ubicación, día, mes y año según Mohini P., Vipul K., & Harshadkumar B., (2015) y a un desbalance en los datos causado porque el 75% de la base de datos corresponde a precipitaciones inferiores a 5,3 mm, y el 50% a precipitaciones inferiores a 0,2 mm.

Palabras clave: Precipitación, Machine Learning, Modelo, Métricas, Predicción.

ABSTRACT:

Predicting precipitation is ideal because it helps the planning of agricultural and human activity, in agronomic activity it could be determined if the water requirements of a crop are going to be present and thus not lose a harvest, or know how many millimeters need to be achieved to keeping the crop hydrated, on the other hand, as a society we are interested in it, because it can be determined when there will be heavy or torrential rains that lead to a flood or landslides that endanger life. For the statistical analysis, data were taken from the Yariguies agroclimatic station located in the municipality of Barrancabermeja, department of Santander, Colombia, the time series is between 07/01/1967 to 09/30/2009, the unit of the precipitation variable is millimeters (mm), in total there were 19,266 data, of which 15,412 (80%) were used for training and 3,854 (20%) to test the model, the models developed were Holt Winters, Decision trees and a Sequential Neural Network (GRU), the metrics used were the MAE, MSE and RMSE for the models, standing out the GRU neural network with 0.05, 0.01 and 0.1 mm respectively, however heavy rains (20- 70 mm), intense (70-150 mm) and

torrential (> 150 mm) are not observed in the figure because the error is higher than expected, with the decision tree, it was possible to predict heavy, intense and torrential rains but the fit of the model is not adequate; despite the fact that the prediction made by the model tends to have a behavior similar to the actual data, possibly because the precipitation data are not linear in nature since the quantity, frequency and intensity are three main characteristics of the rain time series and the values vary by location, day, month and year according to Mohini P., Vipul K., & Harshadkumar B., (2015) and an imbalance in data caused by the fact that 75% of the database corresponds to rainfall less than 5.3 mm, and 50% to rainfall less than 0.2 mm.

Keywords: Precipitation, Machine Learning, Model, Metrics, Prediction.

INTRODUCCIÓN

La precipitación la define el (IDEAM, 2019) como la caída de partículas de agua líquida o sólida que se originan en una nube, atraviesan la atmósfera y llegan al suelo. La cantidad de precipitación es el volumen de agua lluvia que pasa a través de una superficie en un tiempo determinado según el Instituto de Hidrología Meteorología y Estudios Ambientales – IDEAM. La precipitación analizada en este artículo corresponde a la cantidad de partículas de agua líquida que cae sobre el área de influencia de la estación Climática Aeropuerto Yariguies, ubicada en el municipio de Barrancabermeja-Santander en Colombia.

La precipitación es importante en el ciclo hidrológico, por ser la cantidad de agua que cae a la superficie terrestre y provee de agua dulce para el desarrollo de la vida (Priyan, 2015) citado por (Morales Rojas, Diaz Ortiz, Garcia, & Milla Pino, 2021) hace varios años se viene calentando el planeta lo que pone en riesgo la existencia de muchos seres vivos. (Castillo, Montero, Amador, & Duran, 2018) pronosticaron que en la costa Pacífica de Colombia, Ecuador y Perú habrá un aumento en la precipitación especialmente en invierno y con diferencias muy marcadas para el RCP 8.5. (Vías de concentración representativas o Representative Concentration Pathways) Estos resultados se relacionan bastante bien con lo mencionado por Parry et al. (2007) citado por (Castillo, Montero, Amador, & Duran, 2018),

donde se destaca la disminución de precipitación en Centroamérica y el aumento en las regiones de la costa Pacífica del norte de Sudamérica debido a la potencial intensificación de los eventos del ENOS, el aumento del nivel del mar y su temperatura superficial. Lo anterior lo explica (Allan & Soden, 2008; Donat et al., 2016; Papalexiou & Montanari, 2019) citados por (Li, Jin, & Shao, 2021) cuando argumenta que las precipitaciones son muy variables y los cambios climáticos ya han provocado una tendencia creciente en la frecuencia de los eventos de precipitaciones extremas.

En este artículo se busca pronosticar la precipitación mediante el uso de técnicas aprendizaje automático o machine learning ya que las lluvias extremas a menudo provocan desastres como deslizamientos de tierra e inundaciones, que provocan la muerte de miles de personas y afectan a miles de millones cada año (Shrabani S., Subhankar, & Subimal, 2021), en la topografía Colombiana se encuentran terrenos quebrados utilizados para cultivar, como también zonas con riesgo de inundación; para (Jia Yi, y otros, 2021) muchos países en desarrollo se enfrentan a conflictos de datos hidrológicos (escasez, escasos, etc.) y la utilización de los datos satelitales de acceso público podría ser una alternativa útil, el Instituto de Hidrología Meteorología y Estudios Ambientales – IDEAM en Colombia proporciona datos de diferentes variables climáticas y son de uso público, datos utilizados en el presente análisis estadístico. Por otra parte, el sector agrícola es uno de los factores más críticos que determinan la economía del país, y la agricultura depende completamente de las lluvias (Sethupathi M., Sai Ganesh, & Mansoor Ali, 2021) Colombia es un país agrícola y posee un gran potencial.

Para (Lima, Kwon, & Kim, 2021) el uso de modelos empíricos, técnicas estadísticas y modelos basados en aprendizaje automático (Clark & Slater, 2006, Sloughter et al., 2007, Hong, 2008, Chau y Wu, 2010, Daoud et al., 2011, Ortiz-García et al., 2014, Shrestha et al., 2015, Asanjan et al., 2018, Pham et al., 2020, Ni et al., 2020) han surgido como una alternativa o complemento a dichos modelos meteorológicos para las previsiones diarias de precipitaciones. Para el análisis estadístico de la precipitación se han utilizado diferentes métodos como cálculo de las diferencias de los ensambles de los multi-modelos, random forest, SARIMA, ARIMA, Box – Jenkins, redes neuronales artificiales (ANN), SVM

(Máquina de vectores de soporte), Adaptive Neuro Fuzzy Inference System (ANFIS), SLIQ, K-Nearest Neighbour (KNN), Naïve Byes, posprocesamiento basado en cópula (CPP), Holt-Winters (aditivo) Decision Tree, para el desarrollo de esta artículo se aplicaron las técnicas de Holt Winters, árboles de decisión y redes neuronales GRU.

(Traspuesto Abascal, 2019) ejecutó la técnica del Random Forest y concluyó que es en general, una técnica competitiva para la regionalización de precipitación, mostrando resultados similares a GLM (modelos globales) para la mayoría de las métricas de validación consideradas, y pecando de algunas de las desventajas conocidas para las técnicas basadas en modelos de regresión. (Zainundin, Jasim, & Bakar, 2016) encontraron que con datos de entrenamiento pequeños (10%) de 1581 instancias, Random Forest clasificó correctamente 1043 instancias. (Fonseca Garzón, sf) en su trabajo relaciona tres fortalezas al utilizar Random Forest, es un modelo multiuso que funciona en la mayoría de los problemas, puede manejar datos ruidosos o faltantes (características categóricas o continuas) y selecciona solo las características más importantes y como desventajas tiene que a diferencia de un árbol de decisión, el modelo no es fácilmente interpretable y puede requerir algo de trabajo para ajustar el modelo a los datos (Brett Lantz, 2013) citado por (Fonseca Garzón, sf).

(Sinay & Kembauw, 2021) lograron predecir si la precipitación en la ciudad de Ambon presenta componentes estacionales mediante el método de series de tiempo utilizando ARIMA/SARIMA y Holt-Winter Exponential Smoothing.

(Cowpertwait et al., 1996) citados por (Diez Sierra & Del Jesus, sf) argumenta que ha quedado demostrado que estás técnicas dependen del tipo de clima, y de las zonas de estudio concretas donde se quieran aplicar, y por lo tanto son difíciles de generalizar a otras zonas. Debido a la compleja relación no lineal de lluvia entre las diferentes estaciones en una cuenca (Tongal & Sivakumar, 2021) por esta razón en esta investigación se decidió tomar una estación climática y analizar la variable precipitación de tal forma que se pueda generar un modelo y replicar la metodología a una estación objetivo.

Es un reto poder pronosticar la precipitación, ya que según (Mohini P., Vipul K., & Harshadkumar B., 2015) las técnicas estadísticas para el pronóstico de lluvia no pueden funcionar bien para el pronóstico de lluvia a largo plazo debido a la naturaleza dinámica de

los fenómenos climáticos, porque los datos de precipitación no son lineales en la naturaleza. La cantidad, la frecuencia y la intensidad son tres características principales de las series de tiempo de lluvia. Los valores varían por la ubicación, día, mes y año; sin embargo, en este análisis estadístico se busca pronosticar la precipitación hasta el año 2024.

METODOLOGÍA

Para el análisis estadístico de la precipitación se tomaron los datos de la estación agroclimática Yariguies, los datos fueron proporcionados por el Instituto de Hidrología Meteorología y Estudios Ambientales – IDEAM., en donde se generó una serie de tiempo entre el 01/07/1967 y 30/09/2009, la imputación de los datos faltantes se realizó mediante la interpolación (method =”time”, librería pandas), la variable “Fecha” se tomó como índice y con la precipitación se generó la serie de tiempo.

Respecto a la base de datos se dividió en entrenamiento (80%) y testeo (20%) quedando con 15412 datos para entrenamiento y 3854 datos para prueba, los datos fueron escalados con la herramienta MinMaxScaler.

Para la predicción de la precipitación de los datos se generaron modelos utilizando las técnicas de machine learning como Holt Winters, Decision Tree Regressor y una Red Neuronal secuencial GRU la cual tiene una arquitectura dos gatillos (reinicio y actualización), para predecir un día el hiperparámetro fue de 365 días, se utilizó un Dropout de 0,2, y una capa de salida, el método de optimización utilizado es adam y la función de costo es el error cuadrático medio (MSE), para la red neuronal se utilizaron 100 epocas, 16 batch, y el 20% para validación. El método utilizado para la predicción con la serie de tiempo se denomina Recursive multi-step forecasting.

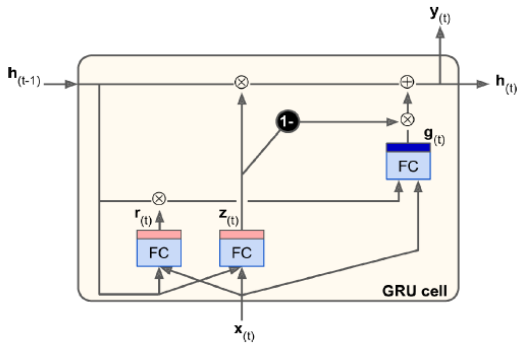

Figura 1 celda GRU tomado de Aurélien Géron (2019)

Las métricas utilizadas para evaluar el modelo son el Mae (mean_absolute_error), Mse (mean_squared_error), RMSE (Root Mean Square Error)

$$error_{MAE} = \frac{1}{m}\sum_{j=i}^{m} \left|y^{(j)} - \hat{y}^{(j)}\right|$$

$$error_{MSE} = \frac{1}{m}\sum_{j=i}^{m} \left(y^{(j)} - \hat{y}^{(j)}\right)^2$$

$$error_{RMSE} = \sqrt{\frac{1}{m}\sum_{j=i}^{m} (y^{(j)} - \hat{y}^{(j)})^2}$$

RESULTADOS

Los datos de precipitación de la estación agroclimática Yariguies presenta una media de 7,57 mm/día, el 75% de la base de datos corresponde a precipitaciones inferiores a 5,3 mm, según Brown Manrique, Diaz Ruiz, Gallardo Ballat, & Valero Freyre, (2017) estas precipitaciones se consideran lluvias ligeras (ver tabla 1), el registro más alto de precipitación fue de 188,4 mm siendo una lluvia torrencial.

Tabla 1 Clasificación de las precipitaciones diarias

Tabla 1. Clasificación de las precipitaciones diarias

Clasificación	**Rango (mm)**	**Estación Yariguies No de días entre el rango**	**Rango adaptado variable cualitativa**
Lluvia nula	0	9191	0
Lluvias ligeras	0-5	5159	1
Lluvias moderadas	5-20	2539	2
Lluvias fuertes	20-70	2019	3
Lluvias intensas	70-150	351	4
Lluvias torrenciales	>150	7	5

(Brown Manrique, Diaz Ruiz, Gallardo Ballat, & Valero Freyre, 2017)

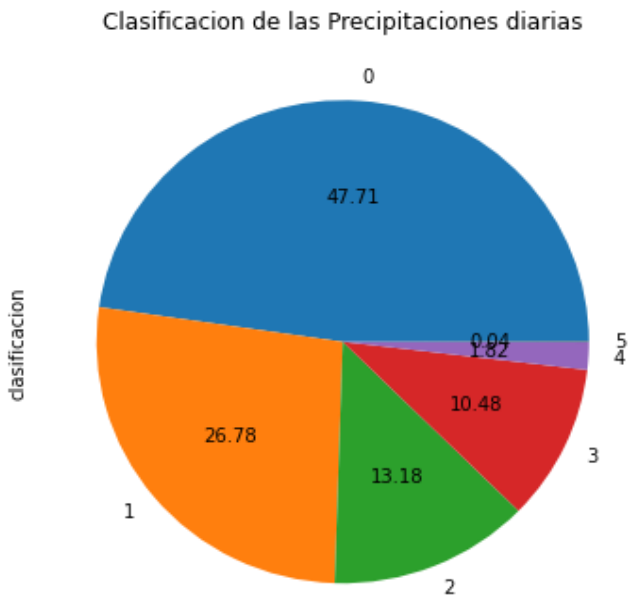

Figura 2. No de eventos (días) de la base de datos que se encuentran dentro de la clasificación realizada por (Brown Manrique, Diaz Ruiz, Gallardo Ballat, & Valero Freyre, 2017)

Respecto a la autocorrelación se observa que durante el periodo analizado anualmente se encuentran 4 ciclos, comprendidos en 2 periodos secos y 2 lluviosos, lo anterior se puede evidenciar en la figura 3.

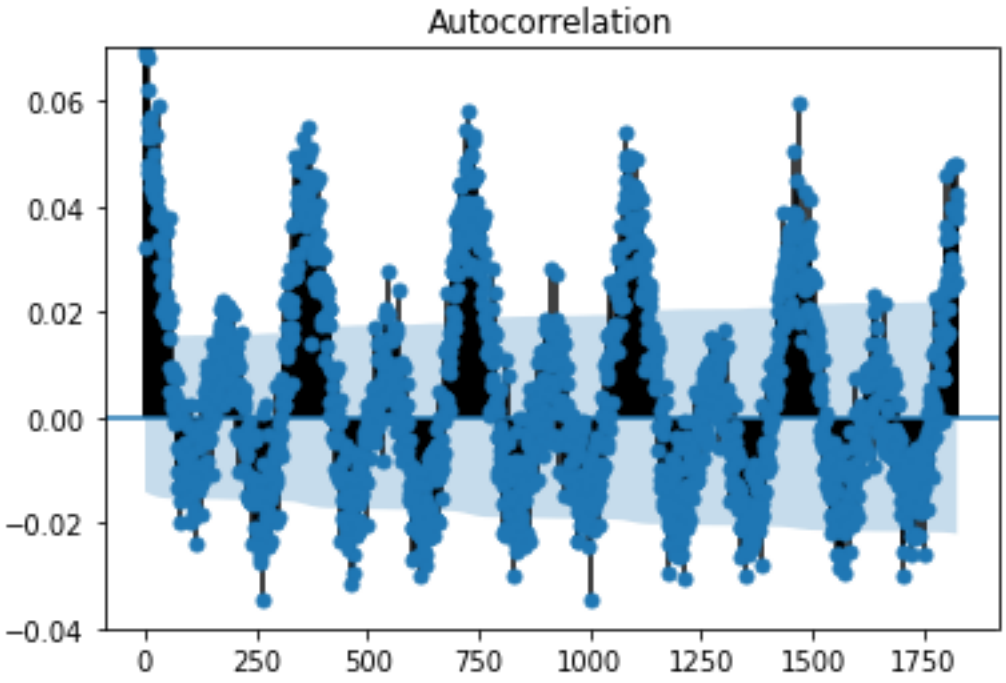

Figura 3. Autocorrelación de la precipitación en la estación Yariguies -Barrancabermeja

Se ejecutaron tres modelos de Machine Learning, los cuales presentaron las siguientes métricas, siendo el modelo de red neuronal secuencial GRU (NNR-GRU) el que obtuvo mejores métricas, (ver tabla 2).

Tabla 2. Métricas de los modelos

Modelo	***MAE***	***MSE***	***RMSE***
Holt-Winters	0,06	0,09	0,3
DecisionTreeRegressor	0,11	0,19	0,43
NNR-GRU	0,05	0,01	0,1

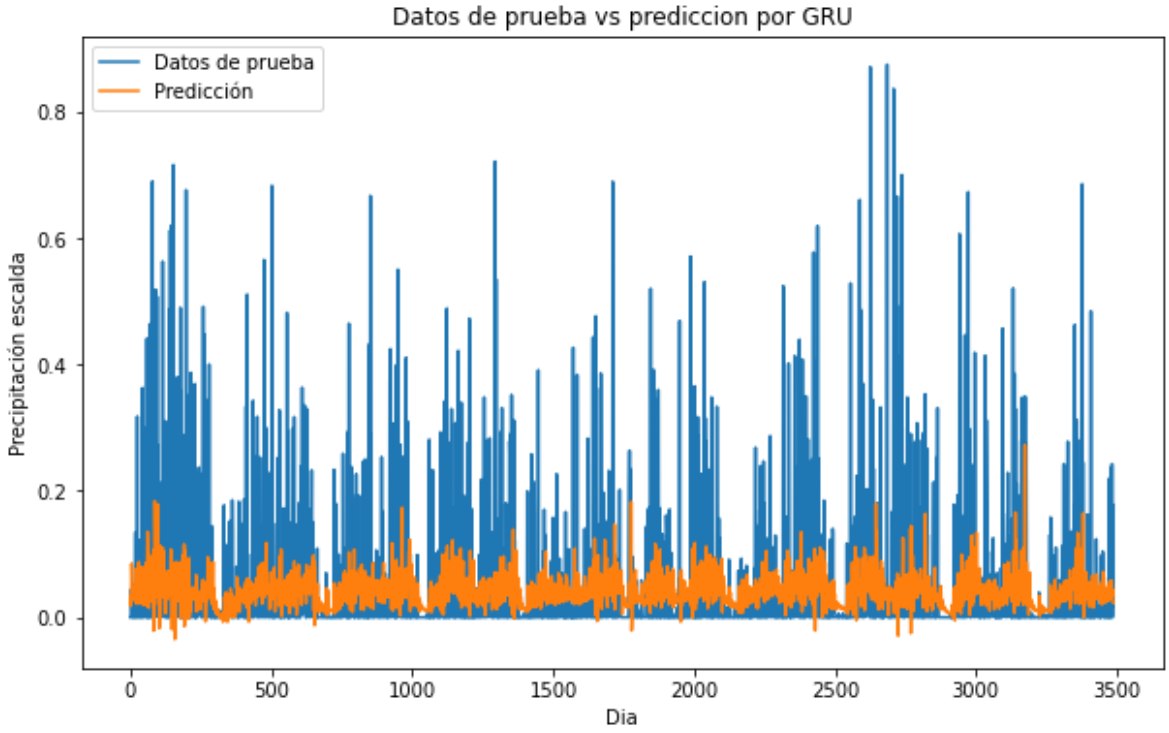

Figura 4. Predicción de la NNR-GRU vs los datos de prueba de la estación Yagires-Barrancabermeja

en la figura 4 se observa la NNR_GRU donde están los datos de prueba versus las predicciones y se puede observar el ajuste que presenta el modelo, las precipitaciones consideradas por Brown Manrique, Diaz Ruiz, Gallardo Ballat, & Valero Freyre, (2017) como fuertes, intensas y torrenciales no se logran predecir siendo una limitante del modelo ya que estás lluvias son las de mayor interés debido que se pueden tomar medidas preventivas para minimizar los riesgos a los que está expuesto el hombre (inundaciones y deslizamientos del suelo), sin embargo el modelo predice con mayor facilidad las eventos secos.

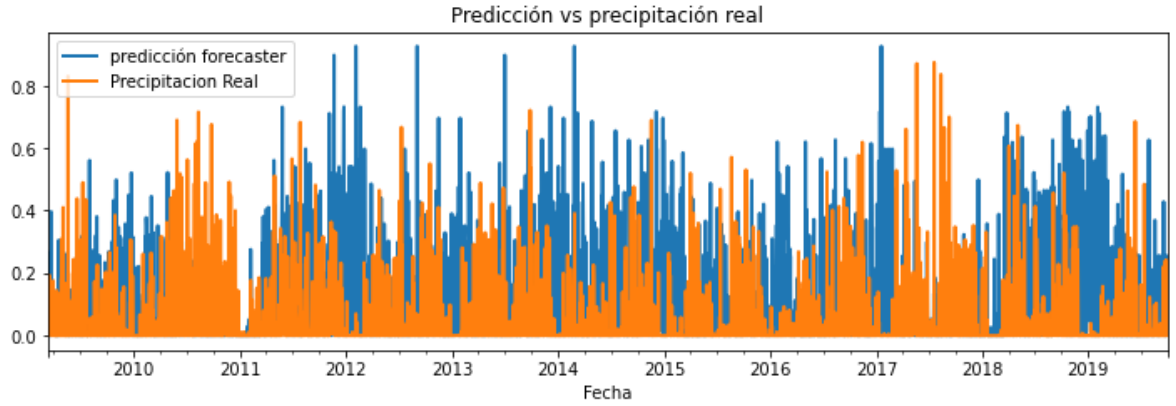

Figura 5. Predicción árbol de decisión vs los datos de precipitación real de la estación Yagires-Barrancabermeja

En la figura 5 se observa la predicción realizada con el modelo de árbol de decisión donde se logran predecir lluvias fuertes, intensas y torrenciales pero el modelo sobredimensiona la precipitación.

DISCUSIÓN DE RESULTADOS

Tongal & Sivakumar, (2021) pronóstico la precipitación mensual para unas estaciones de una cuenca hidrográfica y encontró que los índices de rendimiento obtenidos para este método son 38,39 mm, 0,76, 0,86 y 0,61 para RMSE, NSE, KGE y VE, respectivamente, para la estación agroclimática Yariguies con la red neuronal (GRU) se tuvo un RMSE de 0,1 mm. Lima, Kwon , & Kim, (2021) realizaron un modelo para predecir la precipitación diaria incluyendo una variable exógenas (viento) encontrando influencia estadísticamente significativa hasta 4 días para la ocurrencia de lluvia y 5 días para la cantidad de lluvia, sin embargo concluyeron que el modelo propuesto tiene limitaciones, en cuanto a que los pronósticos de la precipitación diaria siguen siendo un desafío, inclusive pronosticar datos por día; en el mismo sentido Sethupathi M., Sai Ganesh, & Mansoor Ali, (2021) realizaron un pronóstico utilizando metodologías de clasificación obteniendo un accuracy del 95% con random forest, pero concluyen que los resultados de los procedimientos de caracterización utilizados funcionaron bien para la clase sin aguacero, pero no tan bien para la clase de aguacero, Geetha & Selvaraj, (2011) realizaron un pronóstico de la precipitación mensual incluyendo variables como velocidad del viento, temperatura media, humedad relativa y obtuvieron resultados similares a los encontrados en este artículo, posiblemente causado por el desbalance que presentan los datos de precipitación. Geetha & Selvaraj, (2011) sugieren

incluir para futuras investigaciones otras variable como por ejemplo temperatura máxima. En este artículo solo se analizó una variable que fue precipitación, la precipitación hace parte del ciclo del agua y según Peña & Melgarejo, (2016) Cuando el agua lluvia (la principal fuente de agua) cae al territorio puede tomar varios caminos. Puede infiltrarse en el suelo; desplazarse por la superficie (y desembocar en el mar); evaporarse o transpirar (cuando es captada por la vegetación) y llegar de nuevo a la atmósfera, en forma de vapor de agua, y cerrar así el ciclo, en este recorrido existen muchos factores que influyen sobre el ciclo del agua por lo tanto hay que continuar midiendo e incluyendo nuevas variables, como la evaporación por dar un ejemplo, y así posiblemente se puedan mejorar estos pronósticos.

CONCLUSIONES

Los resultados obtenidos en este trabajo no son los esperados ya que las precipitaciones intensas, fuertes y torrenciales no se lograron predecir, siendo estas las de mayor interés.

La red neuronal GRU es la que presentó mejores métricas para la predicción de la precipitación.

Las predicciones realizadas con el árbol de decisión lograron pronosticar lluvias fuertes, intensas y torrenciales, pero no se ajusta a los datos de testeo.

La escasez de datos o datos incompletos no permiten construir una adecuada base de datos, por lo tanto, es necesario que las entidades independientes y públicas que monitoreen o se encuentren interesados en el pronóstico del clima empiecen a tomar datos de nuevas variables y a continuar llevando los registros de las mediciones actuales.

REFERENCIAS BIBLIOGRÁFICAS

Brown Manrique, O., Díaz Ruiz, R., Gallardo Ballat, Y., & Valero Freyre, J. (2017). Caracterización de precipitaciones diarias en el municipio de Ciego de Ávila, Cuba. *Ingeniería Hidráulica y Ambiental*, 44-58.

Castillo, R., Montero, R., Amador, J., & Durán, A. M. (2018). Cambios futuros de precipitación y temperatura sobre América Central y el Caribe utilizando proyecciones climáticas de reducción de escala estadística). *Revista de Climatología*, 1-12.

Diez Sierra, J., & Del Jesus, M. (sf). Generación sintética de series de precipitación horarios, mediante modelos puntuales en zonas sin información sub-diaria. *Universidad de Cantabria*, 223-232.

Dimas, L. (2006). *Agua: recurso estrategico para nuestro crecimiento economico y progreso social. Situacion y desafios.* San Salvador: FUSADES.

Fonseca Garzón, J. A. (sf). Series de tiempo y Random Forest Regression en el periodo 2015-2019, modelamiento de la temperatura en Bogotá D.C. *Los Libertadores*, 1-27.

Geetha, G., & SELVARAJ, R. (2011). Predicción of monthly rainfall in Chennai using back propagation neural network model. *International Journal of Engineering Science and Technology*, 211-213.

IDEAM. (2019). *www.ideam.gov.co.* Obtenido de Glosario Meteorologico: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjHsLiHr5j0AhViRTABHWYxCpIQFnoECC8QAQ&url=http%3A%2F%2Fwww.ideam.gov.co%2Fdocuments%2F11769%2F72085840%2FAnexo%2B10.%2BGlosario%2Bmeteorol%25C3%25B3gico.pdf%2F6a90e554-6607-43cf-8845

Jia Yi, T., Rizaludin Mahmud, M., Nadzri Md Reba, M., Hashim, M., Norman, M., Shafrina, W., & Jaafar, W. (2021). Forecasting the near future of rainfall in humid tropics using the high-resolution satellite precipitation data. *Physics and Chemistry of the Earth*, 1-9.

Li, M., Jin, H., & Shao, Q. (2021). Improvements in subseasonal forecast of rainfall extremes by statical postprocessing methods. *Weather and Climate Extremes*, 1-17.

Lima, C. H., Kwon, H.-H., & Kim, Y.-T. (2021). A Bernoulli-Gamma hierarchical Bayesian model for daily rainfall forecast. *Journal of Hydrology*, 113-118.

Mohini P., D., Vipul K., D., & Harshadkumar B., P. (2015). Rainfall forecasting using neural network: a survey. *ResearchGate*, 1-10.

Morales Rojas, E., Diaz Ortiz, E. A., Garcia, L., & Milla Pino, M. (2021). Forecasting monthly rainfall: a case study in Peruvian native communities. *Pakamuros*, 71-85.

Sethupathi M., G., Sai Ganesh, Y., & Mansoor Ali, M. (2021). Efficient rainfall prediction and analysis using machine learning techniques. *Turkish Journal of Computer and Mathematics Education*, 3467-3474.

Shrabani S., T., Subhankar, K., & Subimal, G. (2021). Hazard at weater scale for extreme rainfall forecast reduces uncertainty. *water security*, 1-8.

Sinay, L. J., & Kembauw, E. (2021). Monthly rainfall components in Ambon City: evidence from the serious time analysis. *The Electrochemical Society*, 1-8.

Tongal, H., & Sivakumar, B. (2021). Forecasting rainfall using transfer entropy coupled directed-weighted. *Atmospheric Research*, 1-13.

Traspuesto Abascal, M. (2019). Estudio de idoneidad de la técnica Random Forest para la regionalización estadística de proyecciones de cambio climático. *Tesis de maestría*, 1-55. España.

Zainundin, S., Jasim, D. S., & Bakar, A. A. (2016). Comparative analysis of data mining techniques for Malaysian rainfall prediction. *International Journal on Advanced Science, Engineering and Information Technology*, 1148-1153.