Revista Cubana de Meteorología

Revista Cubana de Meteorología, Vol. 30, No. 2, abril-junio 2024, ISSN: 2664-0880

Artículo Original

Predicción de las precipitaciones para la temporada lluviosa en Cuba

Rainfall prediction for the rainy season in Cuba

https://cu-id.com/2377/v30n2e02

Julio Enrique Rojas Cantero[1]*, Beatriz Bello García[1], Aldo Saturnino Moya Álvarez[2]

[1]Universidad Central Marta Abreu de Las Villas, Cuba.
[2]Instituto de Geofísica de Perú, Cuba.

RESUMEN: Los pronósticos de las precipitaciones tienen gran importancia para prevenir desastres, pérdidas económicas en cultivos y varios aspectos importantes para la hidroeconomía del país, de ahí la necesidad de desarrollar modelos de predicción de precipitaciones. En el Centro Meteorológico Provincial de Villa Clara (CMPVC) se emplean modelos matemáticos con un nivel de precisión alto, pero requieren mucho tiempo de procesamiento. De ahí la necesidad de buscar maneras más actuales y precisas para realizar pronósticos de lluvia acertados, pero de manera más rápida. El progreso de la inteligencia artificial y en especial, las redes neuronales artificiales han permitido desarrollar modelos para dar solución al problema de predicción de precipitaciones; mediante las métricas proporcionadas por la Organización Meteorológica Mundial (OMM) comprendidas entre los años 1975 hasta el 2016. En este trabajo se realiza un análisis de los diferentes modelos de regresión con múltiples salidas para predecir las precipitaciones en los meses mayo, junio, julio y agosto. Los modelos de regresión que más se ajustaron a esta problemática fueron la regresión lineal, el árbol de regresión, el k-vecinos más cercanos, la regresión directa, la regresión encadenada y el perceptrón multicapa. Luego de realizar las corridas de los modelos resultó el perceptrón multicapa, el modelo de regresión con mejores resultados, con un alto grado de eficacia para el pronóstico de las precipitaciones.

Palabras claves: modelos de regresión con múltiples salidas, aprendizaje automatizado.

ABSTRACT: Rainfall forecasts are of great importance to prevent disasters, economic losses in crops and several important aspects for the country's hydroeconomy, hence the need to develop rainfall prediction models. At the Villa Clara Provincial Meteorological Center (CMPVC), mathematical models with a high level of accuracy are used, but they require a lot of processing time. Hence, the need to look for more current and accurate, ways to make accurate rainfall forecasts but in a faster way. The progress of artificial intelligence and especially artificial neural networks have allowed the development of models to solve the problem of rainfall prediction; using the metrics provided by the World Meteorological Organization (WMO) from 1975 to 2016. In this work, an analysis of the different regression models with multiple outputs to predict rainfall is carry out in the months of May, June, July and August. The regression models that best fit this problem were linear regression, regression tree, k-nearest neighbors, direct regression, chained regression and multilayer perceptron. After running the models, the multilayer perceptron was the regression model with the best results, with a high degree of efficiency for rainfall forecasting.

Key words: regression models with multiple outputs, machine learning.

INTRODUCCIÓN

La predicción de las precipitaciones es un asunto de suma importancia en Cuba, solamente considerando en el ámbito económico dirigido a la agricultura es simplemente esencial. Los cultivos de alto rendimiento ya sean de ciclo corto o largo requieren riego o planificaciones estrictas de las fechas de siembra o

*Autor para correspondencia: Julio Enrique Rojas Cantero. E-mail: julio.rojas@vcl.insmet.cu
Recibido: 12/03/2024
Aceptado: 06/06/2024
Julio Enrique Rojas Cantero. Universidad Central Marta Abreu de Las Villas. E-mail: julio.rojas@vcl.insmet.cu
Beatriz Bello García. Universidad Central Marta Abreu de Las Villas. E-mail: beatrizbellogarcia@gmail.com
Aldo Saturnino Moya Álvarez. Instituto de Geofísica de Perú. E-mail: aldomoya00@gmail.com
Conflicto de intereses: Los autores de este artículo científico, declaran que no existe ningún conflicto de intereses.
Contribución de autores: Julio Enrique Rojas Cantero: Diseño de la investigación, búsqueda bibliográfica y síntesis de los antecedentes. Recolección de los datos, análisis de los resultados y en la revisión crítica de su contenido, así como en la redacción y aprobación del informe final. Presentación de los resultados. **Aldo Saturnino Moya Álvarez:** Asesoría metodológica para el diseño de la investigación. Recolección de datos, análisis de los resultados y revisión de su contenido. **Beatriz Bello García:** Asesoría metodológica y selección de algoritmos a utilizar. Consultoría para la redacción del informe.

La predicción de las precipitaciones es un asunto de suma importancia en Cuba, solamente considerando en el ámbito económico dirigido a la agricultura es simplemente esencial. Los cultivos de alto rendimiento ya sean de ciclo corto o largo requieren riego o planificaciones estrictas de las fechas de siembra o recolección, teniendo en cuenta que el regadío implica sustanciales inversiones en equipos y portadores energéticos. Hay cultivos que necesitan estar con un mayor nivel de regadío o estar en un ambiente muy húmedo como el caso de arroz. Otro aspecto dependiente de las precipitaciones con alto impacto económico es el tipo y cantidad de fitofármacos a planificar, los cuáles varían en tipo y cantidad a utilizar en dependencia del régimen de lluvias. También es importante en el sector energético, especialmente en las instalaciones hidroeléctricas, para la generación de energía, así como en el sector ambiental debido a que "el agua" constituye un factor fundamental e imprescindible para mantener el medio ambiente en equilibrio y además una predicción a tiempo de las precipitaciones que pueden ocasionar grandes inundaciones en zonas rurales y ciudades, provocando derrumbes, y afectaciones en el tráfico aéreo y en la organización de eventos al aire libre que representen peligro para la vida de las personas.

En Cuba, el Ministerio de Ciencia, Tecnología y Medio Ambiente, fundamentalmente mediante el trabajo del Instituto de Meteorología realiza una amplia investigación para mitigar los posibles problemas que puedan ocasionar las fuertes precipitaciones o la falta de ellas. El pronóstico de las precipitaciones es una de las tareas que tiene el servicio meteorológico provincial, para ello se han reunido una serie de datos a partir de varios estudios a lo largo de los últimos años en las diferentes sedes del Instituto de Meteorología de Cuba y de otros miembros de la Organización Meteorológica Mundial (OMM), que luego se usan en la predicción a corto, mediano y largo plazo, mediante diferentes modelos matemáticos, en la actualidad, casi todos basados en análisis de series cronológicas de varias variables meteorológicas para obtener la predicción de la lluvia a esperar en una zona y período determinados. Un sistema de pronóstico de las precipitaciones mediante el uso de inteligencia artificial en Cuba es inédito, no así en el Centro Meteorológico de Villa Clara porque ya se ha utilizado en el pronóstico de las temperaturas mínimas del país (Roque, 2015). El uso de las redes neuronales artificiales en el pronóstico de las precipitaciones se ha utilizado en otros países desarrollados por compañías de desarrollo de software (García, 2020). La regresión es un conjunto de técnicas que son usadas para establecer una relación entre una variable cuantitativa llamada variable dependiente y una o más independientes, llamadas predictoras. Estas deben ser por lo general cuantitativas, sin embargo, usar cualitativas es permisible. Posee grandes aplicaciones en los campos de predicción, descripción, control y selección de variables (Jiménez *et al*., 2020). La regresión de múltiples salidas se define por un modelo de regresión que aplica ecuaciones que representan la relación entre las variables para determinar más de una salida (S. M. S. Ahmed y Guneyli, 2023). Para estimar la ecuación del modelo se debe tener una muestra de entrenamiento la cual le permita al modelo predecir las salidas. Ejemplo de estos modelos son la regresión lineal (Starbuck, 2023), árbol de regresión (Singh Kushwah *et al*., 2022), k-NN (Kigo *et al*., 2023), regresión directa (S. M. S. Ahmed y Guneyli, 2023), regresión encadenada (M. I. Ahmed *et al*., 2023) y el perceptrón multicapa (Nair y Vijaya, 2022).

Como antecedente se puede destacar luego de un estudio en la bibliografía adecuada que no existe uso de los modelos de regresión de salidas múltiples para el uso de la predicción de las precipitaciones en Cuba en el período lluvioso. Sin embargo, fuera de Cuba haciendo uso de la regresión lineal y las redes neuronales se fue capaz de realizar las predicciones de las precipitaciones (Pardo Navarro, 2018) (Muñoz Herrera *et al*., 2020). Con la investigación realizada en la bibliografía adecuada se pude afirmar que en Cuba aplicado a las precipitaciones no hay ningún trabajo precedente. El propósito de este trabajo es desarrollar modelos de predicción de las precipitaciones para el servicio meteorológico provincial, para los meses de mayo, junio, julio y agosto en las diferentes regiones del país, utilizando técnicas de aprendizaje automático (del inglés, *machine learning*), específicamente, los métodos de regresión con múltiples salidas y las redes neuronales artificiales.

MATERIALES Y MÉTODOS

Modelos de regresión con múltiples salidas

Algunos de los problemas de regresión involucran la predicción de dos o más valores numéricos dado un ejemplo de entrada como es el caso del problema planteado en este trabajo. Un ejemplo podría ser predecir una coordenada dada una entrada "p", predecir los valores de "x" e "y". Otro ejemplo sería el pronóstico de series de tiempo de varios pasos que implica predecir múltiples series de tiempo futuras de una variable dada (Géron, 2017).

Varios algoritmos de aprendizaje automático están diseñados para predecir un solo valor numérico, denominado simplemente como regresión. Algunos algoritmos admiten la regresión de múltiples salidas, como la regresión lineal y árboles de decisión. También hay modelos de soluciones especiales que se pueden usar para envolver y usar esos algoritmos que no admiten de forma nativa la predicción de múltiples salidas (Mahesh R N y Nelleri, 2023). La regresión se refiere a un problema de modelado predictivo que implica predecir un valor numérico. Por ejemplo, predecir un tamaño, peso, cantidad, número de ventas y número de clics es un problema de regresión.

Los problemas de regresión con múltiples salidas se conocen como: regresión de salida múltiples:

- Regresión: predice una sola salida numérica dada una entrada.
- Regresión de múltiples salidas: predice dos o más salidas numéricas dada una entrada.

En la regresión de múltiples salidas, normalmente las salidas dependen de la entrada. Esto significa que a menudo los resultados no son independientes entre sí y pueden requerir un modelo que prediga ambas salidas juntas (Mu, 2016).

El pronóstico de series de tiempo de múltiples pasos puede considerarse un tipo de regresión de resultados múltiples donde se predice una secuencia de valores futuros y cada valor predicho depende de los valores anteriores en la secuencia.

El análisis de Regresión Lineal de Múltiples Salidas nos permite establecer la relación que se produce entre una variable dependiente "Y" y un conjunto de variables independientes (X1, X2, ... XK) (Finlay, 2020). El análisis de regresión lineal múltiple, a diferencia del simple, se aproxima más a situaciones de análisis real puesto que los fenómenos, hechos y procesos sociales, por definición, son complejos y, en consecuencia, deben ser explicados en la medida de lo posible por la serie de variables que directa e indirectamente, participan en su concreción. Al aplicar el análisis de regresión múltiple lo más frecuente es que tanto la variable dependiente como las independientes sean variables continúas medidas en escala de intervalo o razón (Bonaccorso, 2020).

No obstante, caben otras posibilidades: también se puede aplicar este análisis cuando relacionemos una variable dependiente continua con un conjunto de variables categóricas; o bien, también aplicaremos el análisis de regresión lineal múltiple en el caso de que relacionemos una variable dependiente nominal con un conjunto de variables continuas (Negnevitsky, 2005).

Se utilizaron como objeto de estudio para esta investigación los Modelos de regresión de múltiples salidas independientes:

a. Regresión lineal (del inglés, *linear regression*): ajusta un modelo lineal con coeficientes w = (w1, …, wp) para minimizar la suma de cuadrados residual entre objetivos observados en el conjunto de datos y los objetivos predichos por la aproximación lineal (González et al., 2016).

b. K vecinos más cercanos (del inglés, *k-NN regressor*): predice mediante la interpolación local de los objetivos asociados a los vecinos más cercanos del conjunto de entrenamiento (Tkatek et al., 2023).

c. Árbol de regresión (*del inglés, decision tree regressor*): construye modelos de regresión en forma de árbol, descompone un conjunto de datos en subconjuntos cada vez más pequeños, al tiempo que desarrolla un árbol de decisión asociado y el resultado final es un árbol con nodos decisión y nodos hojas (Singh Kushwah et al., 2022).

d. Regresión directa (del inglés, *direct regression*): implica dividir el problema de regresión para cada variable objetivo a predecir, las salidas son independientes entre sí (Brownlee, 2021).

e. Regresión encadenada (del inglés, *chained regression*): es una forma de convertir un modelo de regresión de salida única para regresión de múltiples salidas, modelo multietiqueta que ordena las regresiones en una cadena (Brownlee, 2021).

f. Perceptrón multicapa (MLP) (del inglés, *multilayer perceptron*): es una red neuronal que se puede utilizar para realizar regresión. En el aprendizaje automático, la regresión se puede considerar como un mapeo de un espacio a otro donde cada espacio puede tener cualquier número de dimensiones y consta de tres tipos de capas: la capa de entrada, la capa oculta y la capa de salida (Corona *et al*., 2020).

Los datos contienen las cuatro diferentes regiones en las que divide el océano Pacífico, en donde influyen los eventos meteorológicos conocido como El Niño y La Niña, estos sectores se conocen como regiones N12, N3, N4 y N34, las cuáles afectan de forma diferente. La primera y segunda fase de la Madden-Julian Oscillation (MJO) en las tres decenas de días de las que está conformado el mes, la media de la MJO y la amplitud media del mes para cada uno de los meses que componen la temporada lluviosa del país, así como las anomalías de la temperatura del mar y temperatura del mar al principio de la temporada lluviosa y las precipitaciones de los meses de lluvia.

El conjunto de datos posee 36 atributos, de los cuales 32 atributos constituyeron el conjunto de datos predictores y cuatro salidas que correspondieron a las precipitaciones de los meses lluviosos del año en Cuba, los atributos predictantes.

Los predictores (variables de las que depende la precipitación) para la realización del pronóstico fueron los valores de los índices climáticos conocidos durante los seis meses previos al comienzo de la temporada de lluvias, mientras que el predictando (variable a pronosticar) resultaron las anomalías mensuales de la precipitación para cada uno de los meses de la temporada de lluvias.

Se utilizaron los datos grillados de precipitación con resolución de 0.5 grados, tomados del Laboratorio de Ciencias Físicas de la NOAA, Estados Unidos. Se empleó la precipitación mensual acumulada entre 1975 y 2016 para el período lluvioso mayo - octubre. Los índices climáticos fueron tomados del mismo sitio, igualmente para el período 1975 - 2016, de los

seis meses previos al comienzo de la temporada de lluvias. Las últimas correspondieron a las variables objetivos o las precipitaciones a predecir. El dominio del rasgo objetivo se consideró numérico. En este caso los valores que toman los atributos fueron números de un universo finito comprendido entre + 4, por tanto, se trató de un problema de regresión de predicción de precipitaciones con datos obtenidos durante todo el año.

A causa de la diferencia entre los datos ofrecidos por el Centro, se procedió a normalizar para un mejor resultado en el entrenamiento de los modelos y pérdida de precisión. Este estimador escala y traduce cada característica individualmente de manera que se encuentre en el rango dado en el conjunto de entrenamiento, en este caso se realiza una normalización de todos los atributos en un rango de entre + 1.

La transformación viene dada por:

$$X_{std} = (X - X.min(axis = 0))/(X.\max(axis = 0)) - X.\min(axis = 0) \quad (1)$$

$$X_{scale} = X_{std} * (\max - min) + min \quad (2)$$

Donde:

X_std: Valor de la variable a normalizar
X: Valor de la variable original
X.min: Valor mínimo de X
X.max: Valor máximo de X
X_scaled: Valor de la variable a transformar a su escala original
max: máximo valor de los valores tomados
min: mínimo valor de los valores tomados (Unpingco, 2016)

Validación de los modelos

El procedimiento *dejando uno afuera* (del inglés, *Leave-One-Out Cross-Validation* LOOCV) se utiliza para estimar el rendimiento de los algoritmos de aprendizaje automático cuando se utiliza para hacer predicciones sobre datos no utilizados para entrenar el modelo (Brownlee, 2020a). Este procedimiento es una configuración de la validación cruzada *k-fold* en la que se fija en el número de ejemplos del conjunto de datos y a su vez, es una versión extrema de la validación cruzada k-fold que tiene el máximo coste computacional. Requiere la creación y evaluación de un modelo para cada ejemplo del conjunto de datos de entrenamiento (Michalski *et al*., 2013).

El beneficio de tantos modelos ajustados y evaluados es una estimación más sólida del rendimiento del modelo, ya que cada fila datos tiene la oportunidad de representar la totalidad del conjunto de datos de prueba. Dado el coste computacional, LOOCV no apropiado para conjuntos de datos muy grandes, como más de decenas o cientos de miles de ejemplos, o para modelos costosos de ajustar, como las redes neuronales.

Una vez que evaluados los modelos mediante LOOCV y se ha elegido un modelo y una configuración definitivos, se ajustó un modelo final a todos los datos disponibles y se utiliza para hacer predicciones sobre nuevos datos.

Selección de rasgos

Los rasgos se refieren a las características particulares de cada individuo como el temperamento, la adaptación, la habilidad emocional y los valores que le permiten al individuo girar en torno a una característica en particular (Cunningham *et al*., 2021). Los conjuntos de datos, en ocasiones pueden ser pequeños mientras que otros son tremendamente grandes en tamaño, en especial cuando cuentan con un gran número de características, ocasionando que sean muy difíciles de procesar. Cuando se tiene este de tipo de conjuntos de datos de alta dimensión y se utilizan todos para la creación de modelos de aprendizaje automático puede ocasionar que:

1. Las características adicionales actúan como un ruido para el cual el modelo de aprendizaje automático puede tener un rendimiento extremadamente bajo.
2. El modelo tarda más tiempo en entrenarse.
3. Asignación de recursos innecesarios para estas características.

La selección de características o rasgos es el proceso de seleccionar las más importante y/o relevantes características de un conjunto de datos, con el objetivo de mejorar el rendimiento de predicción de los predictores, proporcionar predictores más rápidos y más rentables y proporcionar una mejor comprensión del proceso subyacente que generó los datos.

La importancia futura (del inglés, *Future importance*) de las características (Brownlee, 2020b) se refiere a una clase de técnicas para asignar puntuaciones a las características de entrada a un modelo predictivo que indica la importancia relativa de cada característica al hacer una predicción. Las puntuaciones de importancia de las características pueden calcularse para los problemas que implican la predicción de un valor numérico, llamados de regresión, y para los problemas que implican la predicción de una etiqueta de clase, llamados de clasificación. Las puntuaciones son útiles y pueden utilizarse en una serie de situaciones en un problema de modelado predictivo, como, por ejemplo:

a. Entender mejor los datos.
b. Entender mejor un modelo.
c. Reducir el número de características de entrada.

La importancia de las características puede utilizarse para mejorar un modelo de predicción. Esto puede lograrse utilizando las puntuaciones de importancia para seleccionar aquellas características que se deben eliminar (puntuaciones más bajas) o aquellas características que se deben mantener (puntuaciones más altas). Esto es un tipo de selección de características y

puede simplificar el problema que se está modelando, acelerar el proceso de modelado (la eliminación de características se llama reducción de la dimensionalidad) y, en algunos casos, mejorar el rendimiento del modelo (Pudjihartono *et al*., 2022).

CART es un algoritmo de predicción utilizado en el aprendizaje automático y explica cómo se pueden predecir los valores de la variable objetivo basándose en otras cuestiones. Es un árbol de decisión en el que cada bifurcación se divide en una variable de predicción y cada nodo tiene una predicción para la variable objetivo al final (Cunningham *et al*., 2021).

En el árbol de decisión, los nodos se dividen en subnodos en función de un valor umbral de un atributo. El nodo raíz se toma como conjunto de entrenamiento y se divide en dos teniendo en cuenta el mejor atributo y el valor umbral. Además, los subconjuntos también se dividen utilizando la misma lógica. Esto continúa hasta que se encuentra el último subconjunto puro en el árbol o el máximo número de hojas posible en ese árbol en crecimiento

RESULTADOS Y DISCUSIÓN

La evaluación de un modelo de regresión es fundamental en cualquier aplicación, ya sea para encontrar parámetros óptimos para una regresión, para decidir cual tiene mejores resultados ante un problema específico o para comparar con otro nuevo modelo d regresión. A continuación, se caracterizan los procesos realizados a los datos antes de su uso, los modelos implementados y se explican los experimentos y las pruebas estadificas realizadas a los algoritmos, así como los resultados obtenidos.

Se realizó una normalización de los datos debido a su diversidad de valores, en un rango de + 1 con el objetivo de mantener en un rango cercano los atributos seleccionados al asignarle los pesos a la red cuando realiza el entrenamiento de la misma.

Para darle solución a este problema se recurre al aprendizaje automático, haciendo uso de modelos de regresión capaces de realizar la predicción de los meses lluviosos a partir de los atributos facilitados por la empresa. Entre los modelos de regresión se encuentran la regresión lineal, árbol de regresión, k-vecinos más cercanos, regresión directa, regresión encadenada y el MLP.

Se realizó después una validación cruzada para con el objetivo de estimar el rendimiento de los algoritmos de aprendizaje automático cuando se utiliza para hacer predicciones sobre datos no utilizados para entrenar el modelo, dentro de las muchas que existen se escogió LOOCV.

Se efectuó una selección de rasgos para disminuir la dimensionalidad de datos debido a la poca cantidad de datos brindados; para esto se utilizó el algoritmo CART, método de selección de rasgos de los árboles de decisión. Por último, se procedió a realizar una corrida de todos los modelos haciendo uso de los atributos más importantes resultantes de algoritmo de selección de rasgos y haciendo uso de la validación cruzada.

En la evaluación para comprobar la eficacia de los algoritmos durante su ejecución se procedió a calcular el error medio absoluto (del inglés, *Mean Absolute Error* MAE), el cual representa la diferencia entre los valores originales y los predichos, se extrae promediando la diferencia absoluta sobre el conjunto de datos.

En el caso específico del modelo de regresión del MLP se utilizó un parámetro de ajuste un GridSearchCV, con el objetivo de obtener los parámetros que más ajustaran el algoritmo con los datos; se obtuvo una capa de activación logística, un parámetro de penalización 0.0005, dos capas ocultas de 10 cada una, un optimizador de pesaje Adam y cuatro salidas, para cada uno de los meses lluviosos.

Como se puede observar en la Tabla 1, se puede seleccionar como el mejor modelo de regresión el MLP con un sobresaliente 0.188 MAE, también desatancan los modelos de regresión del *Direct Regression* y el *Decision Tree Regressor*.

Tabla 1. Comparación de los distintos modelos de regresión.

Modelo	MAE
Linear Regression	0.554
Decision Tree Regressor	0.247
k-NN Regressor	0.234
Direct Regression	0.216
Chained Regression	0.303
MLP Regressor	**0.188**

Fuente: Elaboración propia.

Debido a que los resultados arrojados por los modelos, se procedió a realizar una validación cruzada, usando en específico LOOCV con el objetivo de mejorar las predicciones.

Luego de realizar las corridas de los distintos modelos de regresión haciendo uso de la LOOCV como se puede observar en la Tabla 2; todos presentan mejoría excepto la regresión lineal, la cual empeora considerablemente debido a que acumula el error medio a medida que realizada la validación cruzada provocando un MAE muy alto, lo cual hace no sea factible una posible elección.

Tabla 2. Comparación entre los modelos de regresión aplicando una LOOCV.

Modelo	MAE
Linear Regression	16380865535.629
Decision Tree Regressor	0.229
k-NN Regressor	0.185
Direct Regression	0.190
Chained Regression	0.198
MLP Regressor	**0.178**

Fuente: Elaboración propia.

Selección de rasgos

El algoritmo de clasificación de rasgos por su importancia de los árboles de decisión CART fue la forma usada para realizar una selección de atributos, con el objetivo de disminuir la dimensionalidad de los datos y mejorar los resultados de los distintos modelos. En la Figura 1 aparecen los rasgos con mayor nivel de importancia:

A partir del resultado anterior, se seleccionaron los diez mejores con mejor valor, trabajando con una importancia mayor a 0.02290 (Tabla 3), donde destaca RMM2_3D_Julio_AnoAnterior como el rasgo de mayor importancia con 0.21800 y el rasgo amplitude_1D_Julio_AnoAnterior como el de menor importancia de los seleccionados con 0.02314:

A partir de los resultados que arrojó la selección de rasgos, se procede a realizar una nueva corrida de los modelos de regresión, pero esta vez haciendo uso de los rasgos más importantes y la validación LOOCV con el objetivo de tener mejores resultados a la hora de la predicción de las precipitaciones en los meses lluviosos.

La Tabla 4 muestra la comparación de las corridas de los modelos de regresión; resaltando nuevamente el MLP como el mejor modelo con un MAE de 0.172.

Como conclusión parcial, primero se pudo afirmar que la selección de rasgos y la validación cruzada no fue todo lo efectivo que se pensó, debido a que solo mejoró el rendimiento de solo tres de los seis modelos en cuestión y como segundo que el modelo MLP con la topología siguiente: capa de activación logística, un parámetro de penalización de 5e-05, tres capas ocultas de 30, 30 y 20 neuronas respectivamente y el optimizador de pesaje Adam resulta el mejor modelo para la predicción de las precipitaciones (Figura 2).

Luego de haber realizado todas las corridas de los modelos de regresión propuestos, se procede a seleccionar el mejor modelo; el MLP con la topología capa de activación logística, un parámetro de penalización de 0.0005, con tres capas ocultas de 30, 30 y 10 neuronas respectivamente y un optimizador de pesaje Adam usando la validación cruzada y la selección de rasgos, con un MAE de 0.171, lo convierte en el mejor modelo de regresión para la predicción de las precipitaciones en Cuba en los meses lluviosos del año (Tabla 5).

Fuente: Elaboración propia.

Figura 1. Gráfico de los niveles de importancia de los rasgos.

Tabla 3. Rasgos ordenas descendentemente por su importancia.

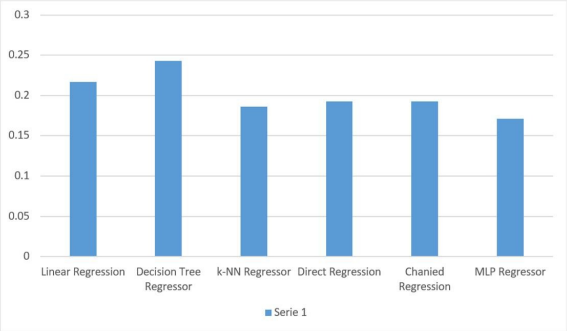
Número de la característica	Nombre de la característica	Importancia del rasgo
22	RMM2_3D_Julio_AnoAnterior	0.21800
8	RMM2_2D_Marzo	0.13376
30	SST_G de Mexico_Marzo	0.10093
1	Anom_SST_N3_Marzo	0.09231
4	RMM1_1D_Marzo	0.07701
5	RMM2_1D_Marzo	0.07308
2	Anom_SST_N4_Marzo	0.06689
21	amplitude_2D_Julio_AnoAnterior	0.03782
17	RMM2_1D_Julio_AnoAnterior	0.02914
19	amplitude_1D_Julio_AnoAnterior	0.02314

Fuente: Elaboración propia.

Tabla 4. Comparación de los modelos con uso de la selección de rasgos por su MAE.

Modelo	MAE
Linear Regression	0.218
Decision Tree Regressor	0.254
k-NN Regressor	0.194
Direct Regression	0.193
Chained Regression	0.193
MLP Regressor	0.171

Fuente: Elaboración propia.



Fuente: Elaboración propia.

Figura 2. Gráfico de comparación de los modelos con selección de rasgos y LOOCV.

Tabla 5. Comparación de modelos con LOOCV y selección de rasgos junto con LOOCV.

Modelo	MAE(LOOCV)	MAE(Selección de rasgos y LOOCV)
Linear Regression	16380865535.629	0.218
Decision Tree Regressor	0.229	0.254
k-NN Regressor	0.185	0.194
Direct Regression	0.190	0.193
Chained Regression	0.198	0.193
MLP Regressor	0.178	0.171

Fuente: Elaboración propia.

Luego de realizar la validación pertinente de los distintos modelos de regresión para múltiples salidas podemos concluir de manera parcial que el *MLP Regressor* es que modelo de todos los presentados; ya el mismo obtuvo el menor MAE con 0,171.

CONCLUSIONES

1. Del estudio del estado del arte de los modelos de regresión múltiples y atendiendo al problema de investigación resultan factibles aplicar los métodos regresión lineal, árbol de regresión, k-NN, regresión directa, regresión encadenada y el MLP.
2. El conjunto de datos es pequeño, se cuenta solamente con 41 años, resultó necesario la escalarización de los datos entre +1, y la selección de rasgos como método de preprocesamiento con el algoritmo CART basado en árbol para mejorar los resultados de la regresión de múltiples salidas.
3. Quedó implementado los métodos recomendados los cuales fueron evaluados utilizando el error medio absoluto (MAE) y como método de evaluación LOOCV, manejando tanto el conjunto de datos descrito por todos los atributos y analizando estos con el conjunto de datos reducido según la selección de rasgos.
4. Atiendo a la métrica MAE se seleccionó como mejor modelo de predicción el MLP como el más óptimo de los modelos que se utilizaron en la investigación.

REFERENCIAS

Ahmed, M. I., -, N., Mahbub-Or-Rashid, Md., y Islam, F. (2023). Prediction of Death Counts Based on Short-term Mortality Fluctuations Data Series using Multi-output Regression Models. *International Journal of Advanced Computer Science and Applications*, *14*(5), Article 5. https://doi.org/10.14569/IJACSA.2023.01405120

Ahmed, S. M. S., y Guneyli, H. (2023). Robust Multi-Output Machine Learning Regression for Seismic Hazard Model Using Peak Crust Acceleration Case Study, Turkey, Iraq and Iran. *Journal of Earth Science*, *34*(5), Article 5. https://doi.org/10.1007/s12583-022-1616-2

Bonaccorso, G. (2020). Mastering machine learning algorithms: Expert techniques for implementing popular machine learning algorithms, fine-tuning your models, and understanding how they work (Segunda edición). PACKT.

Brownlee, J. (18 de Agosto de 2020). Machine Learning Mastery. Recuperado el 18 de Diciembre de 2023, de How to Perform Feature Selection for Regression Data: https://machinelearningmastery.com/feature-selection-for-regression-data/

Brownlee, J. (26 de Agosto de 2020). Machine Learning Mastery. Recuperado el 23 de Diciembre de 2023, de LOOCV for Evaluating Machine Learning Algorithms: https://machinelearningmastery.com/loocv-for-evaluating-machine-learning-algorithms/

Brownlee, J. (27 de Abril de 2021). Machine Learning Mastery. Recuperado el 18 de Diciembre de 2023, de How to Develop Multi-Output Regression Models with Python: https://machinelearningmastery.com/multi-output-regression-models-with-python/

Corona, J. C., Diez, H. R. G., y Morell, C. (2020). Un estudio empírico del modelo de red neuronal MLP para problemas de predicción con salidas múltiples. (6). 13(6), Article 6.

Cunningham, P., Kathirgamanathan, B., y Delany, S. J. (2021). *Feature Selection Tutorial with Python Examples* (arXiv:2106.06437; Número arXiv:2106.06437). arXiv. http://arxiv.org/abs/2106.06437

Finlay, S. (2020). *Artificial Intelligence for Everyone*. Relativistic.

Géron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow*.

González, H. R., Morell, C., y Blanco, A. (2016). *Regresión lineal local con reducción de rango para problemas de predicción con salidas compuestas* (4). *10*(4), Article 4.

Jiménez, P. M. E., López, P. J. N., y Avilés, R. R. (2020). *Aplicación de la regresión de múltiples objetivos en la estimación de componentes fitoquímicos*.

Kigo, S. N., Omondi, E. O., y Omolo, B. O. (2023). Assessing predictive performance of supervised machine learning algorithms for a diamond pricing model. *Scientific Reports*, *13*(1), Article 1. https://doi.org/10.1038/s41598-023-44326-w

Michalski, R. S., Carbonell, J. G., y Mitchell, T. M. (2013). *Machine Learning: An Artificial Intelligence Approach*. Springer Science y Business Media.

Mu, A. C. (2016). *Introduction to Machine Learning with Python*.

Muñoz Herrera, W., Bedoya, O. F., y Rincón, M. E. (2020). Aplicación de redes neuronales para la reconstrucción de series de tiempo de precipitación y temperatura utilizando información satelital. *Revista EIA*, *17*(34), Article 34. https://doi.org/10.24050/reia.v17i34.1292

Nair, J. P., y Vijaya, M. S. (2022). River Water Quality Prediction and index classification using Machine Learning. *Journal of Physics: Conference Series*, *2325*(1), Article 1. https://doi.org/10.1088/1742-6596/2325/1/012011

Negnevitsky, M. (2005). *Artificial intelligence: A guide to intelligent systems* (Segunda edición). Addison-Wesley.

Pardo Navarro, F. (2018). Aplicación del modelo de regresión múltiple para la interpolación de las temperaturas y precipitaciones de la península ibérica y las Islas Baleares. *Estudios Geográficos*, *78*(283), Article 283. https://doi.org/10.3989/estgeogr.201717

Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., y O'Sullivan, J. M. (2022). A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Frontiers in Bioinformatics*, *2*, 927312. https://doi.org/10.3389/fbinf.2022.927312

Roque Rodríguez, J. C. (2015). *Pronóstico de temperaturas mínimas en todas las estaciones meteorológicas cubanas utilizando redes neuronales artificiales*. Universidad Central "Marta Abreu" de Las Villas, Facultad de Matemática, Física y Computación. Santa Clara: Universidad Central "Marta Abreu" de Las Villas.

Singh Kushwah, J., Kumar, A., Patel, S., Soni, R., Gawande, A., y Gupta, S. (2022). Comparative study of regressor and classifier with decision tree using modern tools. *Materials Today: Proceedings*, *56*, 3571-3576. https://doi.org/10.1016/j.matpr.2021.11.635

Starbuck, C. (2023). Linear Regression. En C. Starbuck, *The Fundamentals of People Analytics* (pp. 181-206). Springer International Publishing. https://doi.org/10.1007/978-3-031-28674-2_10

Tkatek, S., Amassmir, S., Belmzoukia, A., y Abouchabaka, J. (2023). Predictive fertilization models for potato crops using machine learning techniques in Moroccan Gharb region. *International Journal of Electrical and Computer Engineering (IJECE)*, *13*(5), Article 5. https://doi.org/10.11591/ijece.v13i5.pp5942-5950

Unpingco, J. (2016). *Python for Probability, Statistics, and Machine Learning*. Springer International Publishing. https://doi.org/10.1007/978-3-319-30717-6

Mahesh R N, U., y Nelleri, A. (2023). Multi-Class Classification and Multi-Output Regression of Three-Dimensional Objects Using Artificial Intelligence Applied to Digital Holographic Information. *Sensors*, *23*(3), Art. 3. https://doi.org/10.3390/s23031095

García, J. (14 de Enero de 2020). Xataca. Recuperado el 18 de Diciembre de 2023, de Google desarrolla una inteligencia artificial capaz de predecir las lluvias de las próximas seis horas en menos de 10 minutos: https://www.xataka.com/inteligencia-artificial/google-desarrolla-inteligencia-artificial-capaz-predecir-lluvias-proximas-seis-horas-10-minutos