UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE MEDELLÍN
FACULTAD DE MINAS

DYNA

http://dyna.medellin.unal.edu.co/

Estimación de anomalías de temperatura del aire en el departamento de Antioquia

Estimation of air temperature anomalies in the department of Antioquia

Valentina Sánchez-Castaño [b]

[a] *Facultad de Minas, Universidad Nacional de Colombia, Medellín, Colombia. vasanchezca@unal.edu.co*

Diciembre 9 de 2022

Resumen

En el presente estudio se aplicaron y evaluaron diferentes modelos de machine learning para la estimación de anomalías mensuales positivas de temperatura superficial del aire en el departamento de Antioquia. El conjunto de variables usado se conformó por ochos variables de medición satelital y una variable medida en tierra. Tras la aplicación de los modelos, los resultados mostraron que los mejores algoritmos para la estimación de las anomalías son las redes neuronales por medio del modelo MLPClassifier y los modelos ensamblados de Ada Boost Classifier.

Palabras clave: Temperatura del aire; anomalías, bases de datos; machine learning, redes neuronales, modelos ensamblados.

Abstract

In the present study, different machine learning models were applied and evaluated for the estimation of positive monthly anomalies of surface air temperature in the department of Antioquia. The set of variables used was made up of eight satellite measurement variables and one variable measured on land. After the application of the models, the results showed that the best algorithms for the estimation of the anomalies are the neural networks through the MLPClassifier model and the assembled models of the Ada Boost Classifier.

Keywords: Air temperature; anomalies, databases; machine learning, neural networks, assembled models**.**

1 Introducción

La temperatura del aire cerca de la superficie del suelo (*Tas*) es uno de los parámetros meteorológicos más críticos e importantes en los estudios ambientales y climáticos, además, esta se encarga de describir las condiciones ambientales terrestres, haciéndola una de las variables climáticas más utilizadas en los estudios de cambio global [1,2]. *Tas* desempeña un papel fundamental en múltiples procesos biológicos y físicos entre la atmósfera, hidrósfera y biósfera [3]. La estimación precisa de *Tas* y el mapeo de su distribución espacial son útiles para predecir las consecuencias del cambio climático, tales como el desencadenamiento de fenómenos meteorológicos extremos, que están asociados con cambios en el régimen de incendios forestales [4,5,6], la distribución de la biomasa forestal [7] y el rendimiento de los cultivos [8,9]. Por otro lado, la revisión bibliográfica sugiere que la transferencia de calor convectiva entre *Tas* y el suelo se ha acelerado durante el último siglo [10] induciendo en muchas partes del mundo el calentamiento del subsuelo a profundidades de aproximadamente 100-200 m [11].

Por las razones expuestas anteriormente se hace necesaria la obtención de datos *Tas* espaciales precisos a gran escala, sin embargo, los datos globales de observaciones de esta variable no son realistas. De igual forma, la falta de homogeneidad en la distribución de estaciones en tierra ha provocado que las observaciones sean menos representativas horizontalmente, permitiendo la introducción de sesgos cuando se utiliza esta información para caracterizar la variación de temperatura en grandes áreas, especialmente en regiones subdesarrolladas y montañosas [1].

La temperatura del aire a menudo se mide con termómetros a 1,5 a 2 m sobre el suelo en estaciones

meteorológicas, y se suele obtener observaciones con alta precisión y resolución temporal [12]. Sin embargo, los registros están limitados por la escasa distribución de las estaciones meteorológicas. Por lo tanto, la resolución espacial de los datos de es baja [13]. Varios métodos de interpolación, como los interpoladores globales, Kriging y ponderación de densidad inversa (IDW), a menudo se llevan a cabo para compensar esta deficiencia [12]. Se ha demostrado que estos métodos tienen éxito en la estimación de la temperatura cerca de las estaciones meteorológicas, pero podrían generar incertidumbres considerables causadas por la distribución no homogénea de las estaciones meteorológicas, especialmente en las regiones con topografía complicada como las zonas montañosas [14]. No obstante, existen otras técnicas con las cuales es posible obtener mejores resultados en cuanto a la estimación de *Tas*, como los métodos de aprendizaje automático (machine learning), pues, la temperatura se estima con modelos de aprendizaje automático no lineales, como redes neuronales, bosques aleatorios y máquinas de vectores de soporte [15,16]. En comparación con métodos tradicionales de regresión, los métodos de machine learning tienen la ventaja de modelar fácilmente relaciones no lineales y altamente interactivas [17].

Los métodos de machine learning son empleados en muchos campos de las ciencias ambientales principalmente por la capacidad de los diferentes modelos propuestos para el procesamiento de datos, predicción de eventos, pronóstico de la calidad del aire, análisis y modelado de datos ambientales, pronóstico oceanográfico e hidrológico, modelado ecológico, y seguimiento de la nieve, el hielo y los bosques, entre otros; esto es necesario para conocer, por ejemplo los efectos que se tendrán sobre el clima del planeta ante perturbaciones del comportamiento normal de variables ambientales [18].

Este documento tiene como objetivo estimar las anomalías positivas mensuales de temperatura superficial del aire a nivel del suelo (*Tas)* de resolución de 1 km a partir de un conjunto de datos con información desde el 2003 hasta el 2014, basado en variables ambientales de origen satelital derivadas de imágenes de resolución moderada (MODIS) y del reanálisis ERA-5 del ECMWF y modelo de elevación digital (DEM) de Alos Palsar e información medida en tierra de la temperatura del aire por medio de diferentes algoritmos de aprendizaje automático de redes neuronales, específicamente el Multi-layer Perceptron classifier (MLPClassifier) y modelos ensamblados, principalmente el algoritmo Ada Boost Classifier.

2 Área de estudio y bases de datos

La zona de estudio del presente documento es el departamento de Antioquia ubicado al noroeste de Colombia con coordenadas 6°13′00″N 75°34′00″O, cubre un área de 60000 km² aproximadamente y tiene una elevación promedio de 2099 m.s.n.m, la figura 1 muestra la zona de estudios con las 36 estaciones meteorológicas de las cuales se obtuvo la información de temperatura superficial del aire.

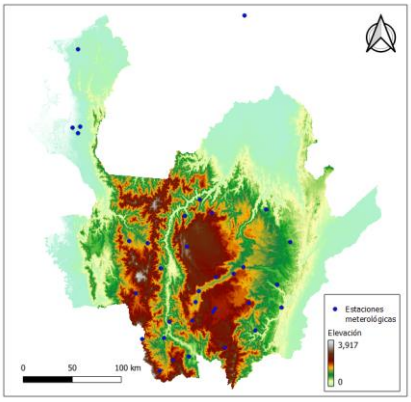

Figura 1. Localización de estaciones en el departamento de Antioquia. Fuente: Elaboración propia.

Para desarrollar el presente estudio en total se emplearon ocho bases de datos, de los cuales siete son productos de origen satelital, es decir, las mediciones se realizaron por medio de satélites, estas variables fueron la evaporación total, temperatura del suelo, cobertura de nubes, velocidad del viento y tipo de cobertura vegetal, estos datos fueron proveídos por Copernicus, El Centro Europeo de Previsiones Meteorológicas a Plazo Medio (ECMWF) y el Servicio de Cambio Climático [19], otra de las variables usadas es el índice NDVI obtenido a partir del satélite MODIS perteneciente a las NASA [20], también se requiere el modelo de elevación digital (DEM) obtenido de Alos Palsar [21], por último, la variable medida en tierra (temperatura del aire cerca de la superficie del suelo - *Tas*) se obtuvo a partir de 36 estaciones meteorológicas descargada del portal Dhime del Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM) distribuidas en el departamento de Antioquia. En la Tabla 1 se observa el resumen de la información básica de las bases de datos.

Tabla 1.
Bases de datos.

Variable	Fuente	Resolución espacial	Resolución temporal
Evaporación total [m de agua equivalente]	ERA-5	0.1° x 0.1°	Mensual
Temperatura del suelo [K]	ERA-5	0.1° x 0.1°	Mensual
Cobertura de nubes [%]	ERA-5	0.1° x 0.1	Mensual
Velocidad del viento [m/s]	ERA-5	0.1° x 0.1°	Mensual
Tipo de cobertura vegetal [-]	ERA-5	0.1° x 0.1°	Mensual
NDVI [-]	MODIS	5600 m	Diaria
Modelo digital de elevación [m]	Also Palsar	0.0003° x 0.0003°	-
Temperatura del aire [K]	IDEAM	-	Diaria

Fuente: Elaboración propia.

3 Metodología

3.1. Variables

Para la estimación de *Tas* se utilizó la temperatura del suelo *Ts* de nivel 1 debido a que ambas variables tienen una alta correlación según la revisión bibliográfica. *Ts* se tomó del reanálisis ERA-5 con resolución mensual y resolución espacial de 0.1° x 0.1°. Además, se tuvieron en consideración variables como la altitud, la cubierta vegetal, la evaporación, la velocidad del viento, la cobertura de nubes y la radiación solar puesto que estas influyen en la relación *Tas* – Ts. La altitud se extrajo del portal de The Alaska Satellite Facility, la cubierta vegetal, la evaporación y la velocidad del viento en la componente x se obtuvieron del mismo dataset empleado para *Tas* (ERA-5).

El problema que se abordará a continuación es de tipo categórico o de clasificación, ya que lo que se busca es estimar si se tienen anomalías positivas o no de temperatura del aire cerca al suelo en el departamento de Antioquia. Para la solución del problema se requiere del uso de datos en tierra proveniente de estaciones meteorológicas distribuidas en el departamento, la cual en el desarrollo de los modelos será la variable dependiente *y*, en este caso se emplearon datos de 36 estaciones descargados del portal Dhime del Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM).

Luego de realizar el tratamiento de datos correspondiente la base de datos final tiene en total 5134 datos correspondientes a datos desde el 2003 hasta el año 2014 de las coordenadas de las 36 estaciones del IDEAM.

3.2. Selección de variables

Para la selección de variables se implementaron técnicas de análisis de componentes principales para conocer cuales representan mejor el problema. Sin embargo, tras usar esta técnica se concluye que cada variable tiene un peso importante en la descripción del modelo dependiendo de la dimensión, esto se muestra en la Figura 3, por lo cual se opta por seguir con el mismo conjunto de variables. Además, cada variable se encarga de darle sentido físico al modelo, pues todas influyen en la variación de *Tas.*

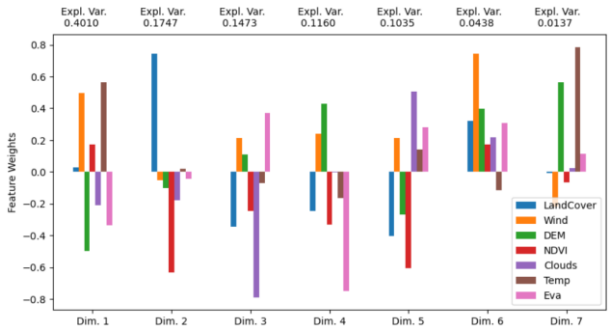

Figura 3Analisis de componentes principales.
Fuente: Elaboración propia.

Es importante mencionar, que, si bien no se realizó una elección de variables significativas, se debió tener en cuenta el uso de técnicas de preprocesamiento debido al desbalance presente en el conjunto de datos, la técnica empleada fue el *Random Under Sampler*, cuyo propósito es reducir el peso de una variable de clasificación para que otra pueda ser más representativa en el modelo, en este caso, se tiene dos categorías o clasificaciones, la 0 y la 1, la categoría 0 corresponde a temperaturas que están por debajo del umbral para el cual se considera que una anomalía es positiva (este valor de temperatura se encuentra dos veces por debajo de la desviación estándar del conjunto de datos), las anomalías positivas se categorizaron con el valor 1, es decir, se sobrepasa el umbral, para este estudio, 1 es la variable categórica de interés, sin embargo, 0 es la que más frecuencia tiene en el conjunto de datos.

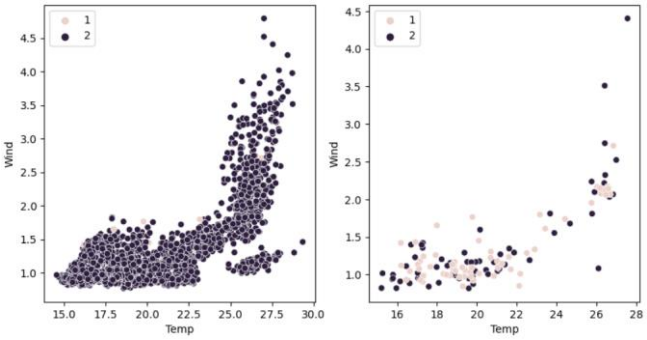

Figura 4. Remuestreo con Random Under Sampler..
Fuente: Elaboración propia

3.3. Modelos de validación

El rendimiento de los modelos de aprendizaje automático se evaluó mediante una validación cruzada de seis conjuntos. Todas las muestras se dividieron aleatoriamente en seis subconjuntos. De los cuales el 70% de los datos se destinaron para el entrenamiento del modelo y el 30% restante se usó en su validación. Por otro lado, se construyeron curvas ROC para conocer y evaluar el grado de precisión de los modelos, y se estimó la *precisión* y el *recall* para cada modelo por medio del reporte de clasificación que se le realiza a los datos de evaluación de la variable dependiente y los datos predichos.

3.4. Modelos supervisados

Para la predicción de eventos de anomalías de temperaturas positivas en el departamento de Antioquia se emplearon varios algoritmos con el propósito de saber con cual se tenía un mejor desempeño. Entre estos, los modelos cuyos resultados cumplieron con el objetivo de la identificación de anomalías son los modelos de redes neuronales y los modelos ensamblados, *Multi-layer Perceptron classifier* y *Ada Boosting* respectivamente.

- **Redes neuronales - Multi-layer Perceptron classifier (MLPClassifier).**

El perceptrón multicapa (MLPClassifier) es un tipo de red neuronal artificial (RNA) formada por múltiples capas de neuronas interconectadas de tal manera que tiene la capacidad de resolver problemas que no son linealmente separables y realizar predicciones precisas.

Las redes neuronales MLPClassifier pueden ser utilizadas para predecir *Tas* en una ubicación determinada. Para ello, se necesita un conjunto de datos históricos que incluya información sobre la temperatura en diferentes momentos y se entrena la red neuronal (Fig. 4) para que aprenda a reconocer patrones en los datos y a realizar predicciones precisas. Esto puede ser útil en aplicaciones como la elaboración de pronósticos meteorológicos. Además, para entrenar el modelo también se requiere emplear variables independientes que sirvan para la identificación de dicho patrón.

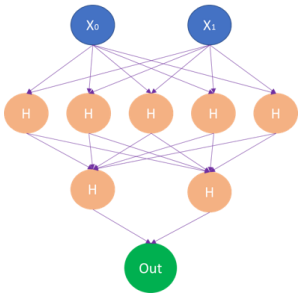

Figura 5. Estructura de una red neuronal.
Fuente: Michael Fuchs.

- **Modelos ensamblados - Ada Boost Classifier:**

Los modelos ensamblados o combinados son un conjunto de técnicas que permiten combinar varios modelos a la hora de realizar predicciones en vez de utilizar uno solo. Desde las más sencillas hasta las más sofisticadas (Random Forest, Boosting y sus variantes). El modelo Ada Boost Classifier fue el primer algoritmo de boosting adaptativo, creado con el objetivo de mejorar la capacidad predictiva de clasificadores binario, este se encuentra diseñado para que en cada paso del entreno la distribución de los datos se adapte a los resultados del clasificador actual, para así poner más peso en los puntos $x \in S$ que el modelo clasifica de forma incorrecta. Finalmente, se usa un promedio ponderado de todos estos modelos secuenciales para dar lugar a la predicción final ensamblada [22].

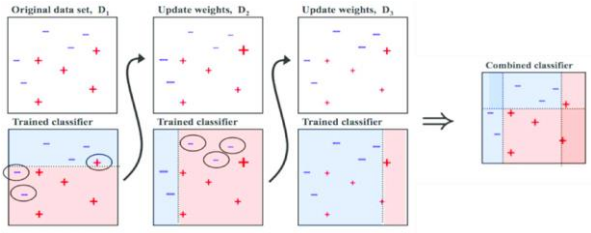

Figura 6. Estructura de modelo ensamblado Ada Boosting.
Fuente: Valentina Alto.

4 Resultados y discusión

Tras aplicar diferentes modelos de machine learning y realizar las respectivas evaluaciones se obtuvieron dos modelos que presentaban el mejor desempeño en cuanto a la identificación de eventos anómalos de temperatura del aire, estos modelos fueron *Multi-layer Perceptron Classifier* y *Ada Boosting Classifier*.

Los resultados obtenidos con el modelo *Multi-layer Perceptron Classifier* se pueden observar en la Figura 5, de esta se puede apreciar que para la variable categórica 1 se tiene una *precisión* del 0.02 y un *recall* de 0.86, esto significa que el modelo está realizando la identificación de anomalías positivas, pero está teniendo problemas en cuanto a la calidad del modelo en tareas de clasificación, es decir, que el modelo se equivocara al pronosticar datos positivos o de categoría cero, sin embargo, se debe tener en cuenta, que el interés del modelo es identificar eventos de anomalías positivas, no que sea preciso al pronosticar datos que no corresponden a la categoría de interés.

	precision	recall	f1-score	support
0	0.96	0.07	0.14	1765
1	0.02	0.86	0.04	35
accuracy			0.09	1800
macro avg	0.49	0.47	0.09	1800
weighted avg	0.94	0.09	0.14	1800

Figura 7. Reporte clasificación para el modelo Multi-layer Perceptron Classifier.
Fuente: Elaboración propia.

De igual forma, la curva ROC (Fig. 6) no muestra un buen desempeño del modelo en cuento precisión, lo que significa, que es necesario tener en cuenta los hiperparámetros del modelo para llegar a resultados más congruentes y precisos.

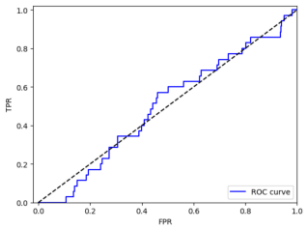

Figura 8. Curva ROC para el modelo Multi-layer Perceptron Classifier.
Fuente: Elaboración propia.

Los resultados al aplicar el modelo ensamblado *Ada Boost Classifier* reduce el *recall*, pero la precisión del modelo mejora (Fig. 7), además la curva ROC, para este caso tiene una mejora en comparación con la curva ROC del modelo *Multi-layer Perceptron Classifier*.

	precision	recall	f1-score	support
0	0.99	0.47	0.64	1761
1	0.03	0.72	0.06	39
accuracy			0.47	1800
macro avg	0.51	0.59	0.35	1800
weighted avg	0.97	0.47	0.62	1800

Figura 9. Reporte clasificación para el modelo Ada Boost Classifier.
Fuente: Elaboración propia.

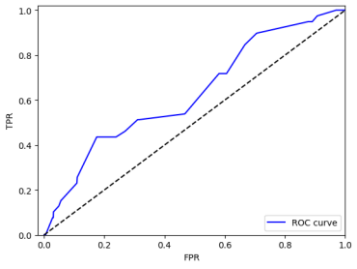

Figura 10. Curva ROC para el modelo Ada Boost Classifier.
Fuente: Elaboración propia.

5 Conclusiones

Para obtener mejores resultados en la aplicación de modelos de machine learning se requiere tener una base de datos con más variabilidad y mejor balanceada, para evitar problemas de underfitting u overfitting, además es necesario conocer en profundidad la física del problema para evitar llegar a resultados incongruentes.

Si bien se obtienen buenos resultados para el *recall* en los dos modelos empleados, se deben varias los hiperparametros para poder mejorar el desempeño de los modelos, puesto que se están sacrificando la predicción de los ceros para poder predecir la categoría 1. Si se mejoran estos modelos, es posible tener una buena aproximación de la *Tas* en todo el departamento de Antioquia.

6 Referencias

[1] WANG, Aihui; ZENG, Xubin. Development of global hourly 0.5 land surface air temperature datasets. Journal of Climate, 2013, vol. 26, no 19, p. 7676-7691.
[2] PRIHODKO, Lara; GOWARD, Samuel N. Estimation of air temperature from remotely sensed surface observations. Remote Sensing of Environment, 1997, vol. 60, no 3, p. 335-346.
[3] STISEN, Simon, et al. Estimation of diurnal air temperature using MSG SEVIRI data in West Africa. Remote sensing of Environment, 2007, vol. 110, no 2, p. 262-274.
[4] WESTERLING, Anthony L., et al. Warming and earlier spring increase western US forest wildfire activity. science, 2006, vol. 313, no 5789, p. 940-943.
[5] CHEN, Yang, et al. Forecasting fire season severity in South America using sea surface temperature anomalies. Science, 2011, vol. 334, no 6057, p. 787-791.
[6] MANZO-DELGADO, L.; SÁNCHEZ-COLÓN, S.; ÁLVAREZ, R. Assessment of seasonal forest fire risk using NOAA-AVHRR: a case study in central Mexico. International Journal of Remote Sensing, 2009, vol. 30, no 19, p. 4991-5013.
[7] REICH, Peter B., et al. Temperature drives global patterns in forest biomass distribution in leaves, stems, and roots. Proceedings of the National Academy of Sciences, 2014, vol. 111, no 38, p. 13721-13726.
[8] RUANE, Alex C., et al. Carbon–Temperature–Water change analysis for peanut production under climate change: a prototype for the AgMIP Coordinated Climate-Crop Modeling Project (C3 MP). Global change biology, 2014, vol. 20, no 2, p. 394-407.
[9] ROSENZWEIG, Cynthia, et al. Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison. Proceedings of the national academy of sciences, 2014, vol. 111, no 9, p. 3268-3273.
[10] BROHAN, Phillip, et al. Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. Journal of Geophysical Research: Atmospheres, 2006, vol. 111, no D12.
[11] KOOI, Henk. Spatial variability in subsurface warming over the last three decades; insight from repeated borehole temperature measurements in The Netherlands. Earth and Planetary Science Letters, 2008, vol. 270, no 1-2, p. 86-94.
[12] VANCUTSEM, Christelle, et al. Evaluation of MODIS land surface temperature data to estimate air temperature in different ecosystems over Africa. Remote Sensing of Environment, 2010, vol. 114, no 2, p. 449-465.
[13] ZHU, Wenbin; LŰ, Aifeng; JIA, Shaofeng. Estimation of daily maximum and minimum air temperature using MODIS land surface temperature products. Remote Sensing of Environment, 2013, vol. 130, p. 62-73.
[14] CARLSON, Toby. An overview of the “triangle method” for estimating surface evapotranspiration and soil moisture from satellite imagery. Sensors, 2007, vol. 7, no 8, p. 1612-1629.
[15] APPELHANS, Tim, et al. Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania. Spatial Statistics, 2015, vol. 14, p. 91-113.
[16] ZHANG, Hongbo, et al. Estimating daily air temperatures over the Tibetan Plateau by dynamically integrating MODIS LST data. Journal of Geophysical Research: Atmospheres, 2016, vol. 121, no 19, p. 11,425-11,441.
[17] OLDEN, Julian D.; LAWLER, Joshua J.; POFF, N. LeRoy. Machine learning methods without tears: a primer for ecologists. The Quarterly review of biology, 2008, vol. 83, no 2, p. 171-193.
[18] HSIEH, William W. Machine learning methods in the environmental sciences: Neural networks and kernels. Cambridge university press, 2009.].
[19] Copernicus Climate Data Store |. Copernicus Climate Data Store | [en línea]. [sin fecha] [consultado el 8 de diciembre de 2022]. Disponible en: https://cds.climate.copernicus.eu/
[20] LP DAAC - MOD13C1. LP DAAC - Homepage [en línea]. [sin fecha] [consultado el 8 de diciembre de 2022]. Disponible en: https://lpdaac.usgs.gov/products/mod13c1v061/
[21] ASF Home Page. ASF [en línea]. [sin fecha] [consultado el 8 de diciembre de 2022]. Disponible en: https://asf.alaska.edu/
[22] Métodos de ensamblado en Machine Learning [en línea]. 2020 [consultado el 8 de diciembre de 2022]. Santiago de Compostela: Proyecto_1686.pdf. Disponible en: http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_1686.pdf

V. Sánchez-Castaño, estudiante de la especialización de recursos hidráulicos de la Universidad Nacional de Colombia, sede Medellín.