Predicción de niveles de polen en la ciudad de Murcia mediante inteligencia artificial

Modelos de Machine Learning basados en datos medioambientales y climáticos

Autor: Javier Rodríguez Zamora

Tutores: Mª Trinidad Cámara Meseguer, Stella Moreno Grau, Juana Re González y Ramón Rodríguez Iborra

Universidad Politécnica de Cartagena

INTRODUCCIÓN

Los granos de polen son partículas que conviven con nosotros a diario, produciendo en ocasiones alergias que pueden llegar a ocasionar problemas para la salud. Existen estudios que indican que la presencia de ciertas tasas de polen viene determinada por factores climáticos y por la presencia de ciertos contaminantes.

Por eso me pregunté si, utilizando los avances de las nuevas tecnologías, se podría predecir la cantidad de estos alérgenos presentes en nuestro entorno a partir de datos climáticos y medioambientales.

Este trabajo trata de resolver esta cuestión construyendo dos modelos de aprendizaje automático que predigan los niveles de polen en la ciudad de Murcia a partir de datos climáticos, medioambientales y de fecha.

OBJETIVO

Construir modelos de aprendizaje automático para la predicción de niveles ambientales de algunos pólenes en la ciudad de Murcia a partir de datos climáticos, medioambientales y fecha.

En concreto, dos modelos de aprendizaje: uno de regresión y otro de clasificación, aplicados a la predicción de cuatro pólenes: Amarantácea, Oleácea, Poácea y Quercus.

PARTES DEL TRABAJO, METODOLOGÍA UTILIZADA Y RESULTADOS OBTENIDOS

1. Recogida de datos

Datos aerobiológicos: proporcionados por el grupo de investigación de Aerobiología y Toxicología Ambiental de la Universidad Politécnica de Cartagena, son datos de los niveles de los pólenes Amarantáceas, Poáceas, Oleáceas y Quercus desde 2009 hasta 2019 en la ciudad de Murcia.

Datos medioambientales: extraídos de la red Sinqlair (Vigilancia y Control de la Calidad del Aire de la Región de Murcia). Se han seleccionado diez contaminantes y de cada uno el mayor de los valores aportados por las estaciones de San Basilio y Alcantarilla.

Datos climáticos: obtenidos de la página MeteoMurcia (estación situada en Puente Tocinos) . Se han utilizado los datos diarios de 15 variables climáticas como Temperatura máx. media y mín. etc

RESULTADO: 3652 datos de cada una de las 29 variables

2. Análisis de datos

Variables aerobiológicas: se han realizado las gráficas tanto de forma global como de cada año concreto a fin de conocer su comportamiento.

Variables medioambientales: para cada contaminante se ha creado una hoja Excel con el valor más alto diario de las dos estaciones consultadas; se han realizado gráficos de estos datos tanto de forma global como por años y se han buscado los límites diarios o anuales permitidos en el Real Decreto 102/2011. A partir de esta información se han seleccionado las variables a utilizar en el estudio informático.

Variables climáticas: como algunas de ellas podrían estar relacionadas entre sí, por ejemplo, temperatura máxima, mínima y media, analizamos tanto el coeficiente de correlación como las nubes de puntos a fin de estudiar su relación y poder eliminar alguna.

RESULTADO: Reducción para nuestro estudio a 9 variables medioambientales de las 10 que teníamos y a 7 climáticas de las 15 iniciales.

NO_x	NO_2	SO_2	O_3	C_6H_6	PM_{10}	CO	C_7H_8	XIL

Temp. Máxima	Humedad relativa media	Presión máxima	Velocidad máxima	Velocidad media	Velocidad de ráfaga	Precipitación

3. Creación de dos modelos de Machine Learning: uno de regresión y otro de clasificación

Los modelos se han desarrollado en la plataforma Google Colab, empleando como lenguaje de programación Python, como entorno de desarrollo los cuadernos Jupyter Notebook y como datos los anteriormente citados junto con los datos de fecha. Estos datos se han dividido en dos grupos (tabla 1): Características y etiquetas.

Características			Etiquetas			
Calendario	Clima	Ambientales	Amarantácea	Oleácea	Poácea	Quercus
Nº día de la semana (1..7) Nº semana en el año (1..54) Nº mes en el año (1..12) Año (2009..2019) Laborable/Festivo	Temp. máx Humedad relativa med Presión máx Velocidad del viento máx Velocidad del viento med Ráfaga de viento máx Precipitación total	NO_2 SO_2 O_3 NO_x C_6H_6 C_7H_8 Xileno PM_{10} CO	Concentración medida en granos/m³ Concentración expresada en nivel de alerta (bajo<10 / 10<medio<20 / alto>20)	Concentración medida en granos/m³ Concentración expresada en nivel de alerta (bajo<100 / 100<medio<200 / alto>200)	Concentración medida en granos/m³ Concentración expresada en nivel de alerta (bajo<10 / 10<medio<50 / alto>50)	Concentración medida en granos/m³ Concentración expresada en nivel de alerta (bajo<50 / 50<medio<150 / alto>150)

Tabla 1. División de datos en: Categorías y Etiquetas

Implementación de los modelos:

Paso 1. Eliminación de muestras con valores no válidos, categorización, normalización y estandarización de las características.

Paso 2. División de forma aleatoria del conjunto de datos en dos subconjuntos (tabla 2): Aprendizaje (80%), que el sistema usa para aprender y Test (20%), que emplea para hallar la fiabilidad de la predicción y que no conocerá hasta el final del proceso.

Paso 3. En el modelo de regresión, la red neuronal definida consta de 2 capas ocultas de 128 nodos cada una, densamente conectadas (fig. 1). A la salida de cada nodo se aplica la función sigmoide. La última capa (de salida) consta de un único nodo a la salida del cual no se aplica ninguna función de activación. En el modelo de clasificación la red es igual que en el modelo de regresión salvo que la capa de salida tiene tres nodos a los que se aplica la función softmax.

Paso 4. Los datos de Aprendizaje se dividen en dos subconjuntos (tabla 2): 75% para entrenamiento y 25% para validación. Con los datos de entrenamiento se obtiene una predicción, se calcula el error cometido (utilizando la función de pérdida huber para regresión y sparse_categorical_crossentropy para clasificación) y se comunica este error a cada nodo de la red neuronal para que reajusten los pesos de sus entradas (usando el optimizador Adam). Con la configuración obtenida tras el aprendizaje se procesan las muestras del conjunto de datos de Validación y se calcula el error cometido en sus predicciones.

Paso 5. Se vuelven a dividir los datos de aprendizaje y se repite el proceso para que el sistema vaya reduciendo el error. El proceso se repite 150 veces.

Paso 6. Para evitar el sobreaprendizaje del modelo se ha empleado la función EarlyStopping con patience=7.

Paso 7. Con los datos de TEST se predice el valor de la etiqueta, se compara con la etiqueta real de la muestra y se obtiene una medida del error absoluto medio (mae) para regresión (tabla 3) o de la exactitud (accuracy) para clasificación.

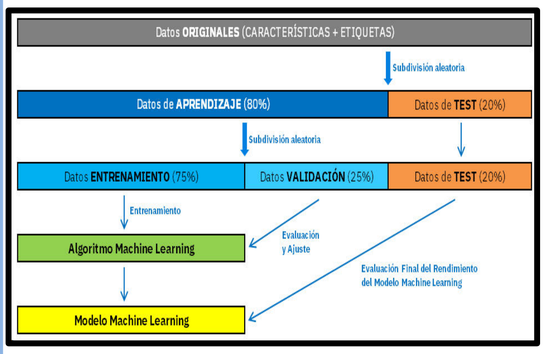

Tabla 2. Esquema de división de los datos originales

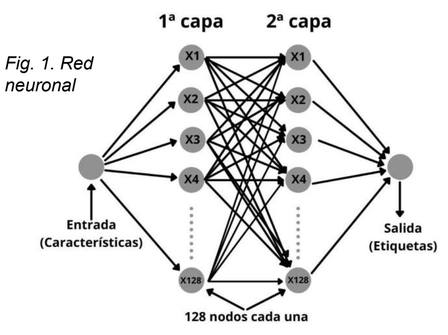

Fig. 1. Red neuronal

PREDICCIÓN DEL MODELO DE REGRESIÓN			
VARIEDAD POLÍNICA	MÉTRICA Error Absoluto Medio	RANGO MUESTRAL	
		Valor Mínimo	Valor Máximo
AMARANTÁCEA	2,49 granos/m³	0	128
OLEÁCEA	3,98 granos/m³	0	909
POÁCEA	1,74 granos/m³	0	99
QUERCUS	3,43 granos/m³	0	320

Tabla 3. Ejemplo de resultados de una ejecución del modelo de regresión

CONCLUSIONES

- Se han construido dos modelos de aprendizaje automático para predecir los niveles de polen en la ciudad de Murcia a partir de datos climáticos, medioambientales y calendario, un modelo de regresión y otro de clasificación.
- El modelo de regresión predice la concentración de polen en granos/m^3 con una gran precisión (ver tabla 3).
- El modelo de clasificación, al tratarse de un escenario de clases severamente no balanceadas, no ofrece resultados buenos por lo que si se desea facilitar una predicción del nivel de alerta en bajo, medio o alto la mejor estrategia es emplear el Modelo de Regresión para obtener el valor de la predicción y convertir esta predicción en el nivel de alerta correspondiente.

BIBLIOGRAFÍA