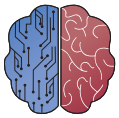
UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería de la Salud



INGENIERÍA
DE LA SALUD

**TFG del Grado en Ingeniería de la Salud**

# Minería de datos y aprendizaje automático aplicado a la predicción de incidencia de párkinson basado en la biometereología.

Presentado por Lorena Calvo Pérez
en Universidad de Burgos

30 de junio de 2025

Tutores: Antonio Jesús Canepa Oneto – Esther Cubo Delgado

UNIVERSIDAD DE BURGOS

ESCUELA POLITÉCNICA SUPERIOR

Grado en Ingeniería de la Salud

D. Antonio Canepa Oneto, profesor del departamento de Ingeniería Informática, área de Lenguajes y Sistemas Informáticos. D. Esther Cubo, profesora del departamento de Ciencias de la Salud, área de Anatomía y Embriología Humana.

Exponen:

Que el alumno D. Lorena Calvo Pérez, con DNI 71705417L, ha realizado el Trabajo final de Grado en Ingeniería de la Salud titulado Minería de datos y aprendizaje automático aplicado a la predicción de incidencia de párkinson basado en la biometereología.

Y que dicho trabajo ha sido realizado por el alumno bajo la dirección del que suscribe, en virtud de lo cual se autoriza su presentación y defensa.

En Burgos, 30 de junio de 2025

Vº. Bº. del Tutor:

D. Antonio Canepa Oneto

Vº. Bº. del Tutor:

D. Esther Cubo

## Resumen

La enfermedad de Parkinson es un trastorno neurodegenerativo cuya prevalencia sigue en aumento a nivel mundial, lo que plantea la necesidad de avanzar en la identificación de factores de riesgo y en el desarrollo de herramientas que apoyen la investigación y la toma de decisiones en salud pública. En este trabajo se aborda esta necesidad mediante el uso de técnicas de aprendizaje automático y minería de datos para analizar la relación entre ciertas variables ambientales y la aparición de la enfermedad. Además, como resultado práctico, se ha desarrollado una aplicación informativa que no solo presenta los resultados de los modelos predictivos, sino que también sirve como una plataforma que podría evolucionar en el futuro hacia una herramienta de apoyo para profesionales y entidades interesadas en el estudio y la prevención de la enfermedad de Parkinson, facilitando el diseño de estrategias preventivas basadas en evidencia.



## Descriptores

machine learning, regresión, Parkinson,variables ambientales, predicción, análisis exploratorio, combinación de modelos, aplicación web

## Abstract

Parkinson's disease is a neurodegenerative disorder whose prevalence continues to rise worldwide, posing the need to advance in identifying risk factors and developing tools that support research and decision-making in public health. This work addresses this need by using machine learning and data mining techniques to analyze the relationship between certain environmental variables and the onset of the disease. Additionally, as a practical result, an informative application has been developed that not only presents the results of predictive models but also serves as a platform that could evolve in the future into a support tool for professionals and entities interested in the study and prevention of Parkinson's disease, facilitating the design of evidence-based preventive strategies.



## Keywords

machine learning, regression, Parkinson's disease, environmental variables, prediction, exploratory analysis, model combination, web application

# Índice general

# Índice de figuras

# Índice de tablas

# Introducción

La enfermedad de Parkinson (EP) es una enfermedad neurodegenerativa que afecta a alrededor de 10 millones de personas en el mundo, siendo esta la segunda enfermedad neurodegenrativa más común tras la enfermedad del Alzheimer. En Europa, su prevalencia oscila entre 1000 y 2000 casos por cada 100.000 habitantes [Mayeux et al., 1997], afectando con más frecuencia a personas mayores de 60 años. La prevalencia de la enfermedad ha aumentado exponencialmente en los últimos años debido al envejecimientos progresivo de la población, por lo que se estima que en 2050 se duplicará el número de casos [World Health Organization, 2022].

Aunque la causa exacta de la EP es desconocida, es considerada una patología multifactorial, es decir, influenciada por diversos factores como pueden ser genéticos, ambientales entre otros [Ball et al., 2019].

Según diversos estudios científicos, se ha demostrado que ciertas variables ambientales pueden influir en el desarrollo y la progresión de la EP [Cao et al., 2024]. Teniendo esto en cuenta, el estudio de la biometeorología está ganando cada vez más relevancia, ya que investiga cómo el clima y las condiciones del ambiente pueden afectar a nuestra salud, especialmente en enfermedades como la EP [Royal Meteorological Society, 2022]. La biometeorología es la ciencia que estudia cómo los factores meteorológicos y climáticos influyen en la salud humana, lo que resulta especialmente relevante en este tipo de patologías [Royal Meteorological Society, 2017].

Por todo ello, en este trabajo no solo se han empleado técnicas de minería de datos y aprendizaje automático para construir modelos predictivos sobre la relación entre variables ambientales y la EP, sino que además se ha desarrollado una plataforma que permite explorar y visualizar estas variables,

así como los resultados obtenidos, proporcionando un enfoque más completo para el análisis de esta enfermedad.

# Objetivos

### Objetivo General

Desarrollar un sistema basado en minería de datos y aprendizaje automático para analizar y predecir la prevalencia de la enfermedad de Parkinson (EP) a partir de variables biometeorológicas, con el fin de identificar factores ambientales que puedan influir en su desarrollo.

### Objetivos Específicos

- **Objetivos funcionales y técnicos:**
  - Automatizar la extracción de datos biometeorológicos de una o varias fuentes para su análisis.
  - Procesar y preparar los datos en formatos adecuados para el entrenamiento y validación de modelos de aprendizaje automático.
  - Desarrollar y entrenar distintos modelos de *machine learning* para la predicción de la prevalencia de la EP.
  - Implementar un modelo final basado en el promedio ponderado de los modelos individuales para mejorar la predicción.
  - Desarrollar y desplegar una aplicación interactiva en *Shiny* con *python* para la visualización de resultados.
- **Objetivos de calidad y fiabilidad:**
  - Evaluar la calidad, precisión y fiabilidad de los modelos mediante análisis estadísticos, incluyendo desviación estándar y detección de anomalías.

- Optimizar la eficiencia y velocidad de ejecución de los modelos para su aplicación práctica.
- Evaluar la importancia de las variables ambientales para interpretar su impacto en las predicciones de los modelos.

- **Objetivos de aprendizaje:**
  - Adquirir habilidades en minería de datos y aprendizaje automático aplicados a datos biometeorológicos.
  - Aprendizaje de la utilizacion del *framework Shiny* con *python* para el desarrollo de la aplicación.

# Conceptos teóricos

## 3.1. Enfermedad de Parkinson

La enfermedad de Parkinson (EP) fue descrita por primera vez por James Parkinson en 1817 en su obra *An Eassy on the shaking palasy* (Un ensayo sobre la parálisis temblorosa) [Parkinson, 2002], en el que explicó las características clínicas de la patología. Aunque él no fue quien dio el nombre a la enfermedad, posteriormente fue reconocida y nombrada en su honor como Enfermedad de Parkinson.

La EP es un trastorno neurodegenerativo del SNC (sistema nervioso central). La prevalencia de la enfermedad es significativa, afectando a un 1-2 % de la población mayor de 65 años. La edad en la que suele manifestarse la enfermedad es entre los 65 y los 70 años, pero puede aparecer en mayores de 50 e incluso en los adolescentes. Cabe destacar que hay mayor incidencia en hombres que en mujeres [Armstrong and Okun, 2020].

La EP se origina cuando ciertas neuronas en el cerebro dejan de funcionar adecuadamente. Estas células son responsables de producir dopamina, una sustancia que transmite señales a la parte del cerebro encargada de controlar el movimiento y la coordinación del cuerpo. Las neuronas afectadas se encuentran en una región llamada sustancia negra [González and Pérez, 2021] (Ver Figura 3.1 [PsicoActiva.com, 2024]). La EP se desarrolla cuando estas células empiezan a morir o deteriorarse, y esto es debido a alteraciones en su metabolismo.
Un aspecto clave de la enfermedad es la acumulación de una proteína llamada $\alpha$-sinucleína. Esta proteína, en condiciones normales, tiene la función de liberar neurotransmisores en el cerebro, pero en pacientes con EP, se pliega de manera anormal, formando agregados que se acumulan en las neuronas,

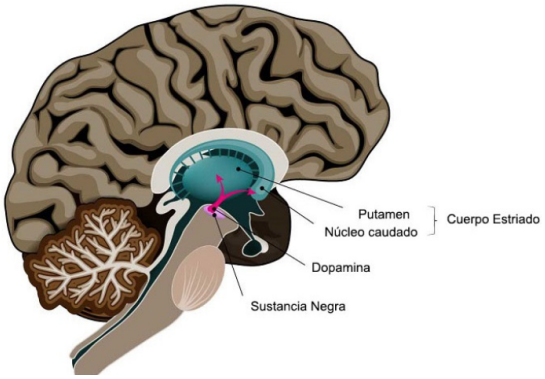



Figura 3.1: Sustancia negra del cerebro

lo que contribuye al daño celular. Estos depósitos de $\alpha$-sinucleína forman estructuras conocidas como cuerpos de *Lewy*, que son característicos de la enfermedad [Zhang et al., 2018]. La producción insuficiente de dopamina desencadena los principales síntomas de la enfermedad, como temblores, lentitud de movimiento, rigidez y problemas de equilibrio [Poewe et al., 2017].

Los criterios diagnósticos de la enfermedad de Parkinson, según la *Movement Disorder Society (MDS)*, se basan en la presencia de bradiscinesia [1] y al menos otro signo motor (temblor en reposo, rigidez e inestabilidad postural) y la exclusión de otras causas [Postuma et al., 2015].

El tratamiento de los síntomas motores de la EP se basa en una terapia de reemplazo de dopamina o en la utilización de agonistas dopaminérgicos [2]. No obstante, la dopamina no puede atravesar la barrera hematoencefálica, por lo que el tratamiento de referencia es la administración de levodopa (L-Dopa), su precursor, que se convierte en dopamina en el cerebro por la acción de la enzima di-hidroxi-fenilalanina descarboxilasa [Hurtado et al., 2016].

En resumen, el tratamiento de la enfermedad combina opciones farmacológicas (principalmente levodopa en monoterapia o en combinacion de otros), así como no farmacológicas como el ejercicio físico, fisioterapia, terapia de lenguaje entre otras.

[1] **Bradicinesia**: Disminución en la velocidad y la amplitud de los movimientos.
[2] **Agonistas dopaminérgicos**: compuestos que activan los receptores de dopamina en el cerebro.

## 3.2. La Biometeorología y su Relación con la Salud

La Biometeorología es la rama de la ciencia que trata las relaciones entre los procesos atmosféricos y los seres vivos [Ramos, 2014]. Por otro lado, la biometeorología médica, estudia cómo los fenómenos meteorológicos repercuten en el cuerpo humano y cómo los cambios del clima a lo largo de un año provocan variaciones importantes en la salud [RTVE, 2017]. Ademas, diferentes investigaciones han demostrado que las condiciones meteorológicas y climáticas, como la temperatura y la humedad entre otras, pueden influir en la aparición de diversas enfermedades por lo que estas variables actúan como factores de riesgo para la salud humana [Rodríguez et al., 2015].

Entre los fenómenos meteorológicos de mayor impacto en la salud podemos encontrar las olas de calor. La frecuencia y la intensidad de estas ha aumentado exponencialmente en los últimos años como consecuencia del cambio climático [Organización Mundial de la Salud, 2021]. Esto ha provocado un incremento en el numero de enfermedades así como de la mortalidad, destacando significativamente en los periodos de tiempo en los que se prolonga una elevada temperatura [Fortoul van der Goes, 2022].

De hecho se ha demostrado a través de investigaciones que las olas de calor tienen un impacto directo en la salud pública, especialmente en poblaciones vulnerables como personas mayores. Un ejemplo significativo es el verano de 2003, cuando España experimentó tres olas de calor que causaron un incremento de la mortalidad del 8 %, concentrado en mayores de 75 años, con aumentos entre el 15 % y el 29 % en los grupos de edad de 75-84 años y mayores de 85, respectivamente [Simón et al., 2005].

Es importante mencionar que existen otros componentes que influyen en la salud de las personas. Entre ellos se encuentran:

- **Calidad del aire:** Entre los contaminantes que representan un grave riesgo para la salud pública se encuentran las partículas en suspensión, el monóxido de carbono, el ozono, el dióxido de nitrógeno y el dióxido de azufre. La contaminación del aire, tanto en interiores como en exteriores, causa enfermedades respiratorias y de otro tipo, y es una fuente importante de morbilidad y mortalidad [World Health Organization, 2025].
- **Agua y saneamiento**: El acceso a agua limpia y un saneamiento adecuado son esenciales para prevenir enfermedades transmiti-

das por el agua. La contaminación de fuentes de agua, por desechos industriales y urbanos, puede generar graves problemas de salud [Organization, 2023].

- **Alimentos y seguridad alimentaria:** Los contaminantes ambientales como pesticidas y metales pesados afectan la cadena alimenticia y pueden causar enfermedades tanto agudas como crónicas [Thompson and Darwish, 2019].
- **Suelos y contaminación:** La contaminación del suelo por productos químicos y desechos peligrosos afecta a la salud humana a través del contacto directo o por ingestión de alimentos contaminados [Biswas et al., 2025].
- **Cambio climático:** Las alteraciones en el clima aumentan los riesgos para la salud, como la propagación de enfermedades, el estrés térmico y los desastres naturales, además de causar desplazamientos poblacionales [Díaz Cordero, 2012] [Ambientum, 2025].

La Organización Panamericana de la Salud (OPS) resalta que el cambio climático representa un riesgo significativo para la salud y el bienestar [(PAHO), 2025].

## 3.3. Minería de datos

La minería de datos es una disciplina que se centra en el desarrollo de métodos y algoritmos diseñados para extraer automáticamente información relevante, lo que facilita la identificación de patrones ocultos en grandes volúmenes de datos. Además, uno de los objetivos de la minería de datos es garantizar que la información obtenida tenga capacidad predictiva, optimizando así, el proceso de análisis de estos datos [Martinez, 2001] (Figura 3.2).


Figura 3.2: Relación entre IA, ML, Minería de Datos y Estadística

Tanto la minería de datos como el aprendizaje automático se han vuelto fundamentales en el campo de la salud, ya que permiten procesar y analizar grandes cantidades de datos clínicos y biométricos, lo que facilita la detección de patrones, la predicción de enfermedades y el apoyo en la toma de decisiones médicas [Raul et al., 2016].

En conclusión, la combinación de la minería de datos y el aprendizaje automático en la predicción de la incidencia de la EP, teniendo en cuenta la biometeorología, permite identificar relaciones entre las variables ambientales y la salud de los pacientes, contribuyendo a una mejor comprensión de la interacción del entorno y la salud.

## 3.4. Estado del arte y trabajos relacionados.

En esta sección se nombrarán algunos proyectos o investigaciones relacionados con el trabajo.

Algunos estudios y proyectos sobre la detección genética de la EP mediante *Machine Learning* son:

1. **Predicción de variantes patogénicas de la enfermedad de Parkinson utilizando sistemas híbridos de aprendizaje automático y características radiómicas** [Hajianfar et al., 2023].

   Realizaron un estudio en el que aplicaron sistemas híbridos de m*achine learning* (HMLS) para predecir variantes patogénicas en los genes LRRK2 y GBA, dos de los principales genes asociados con la enfermedad de Parkinson. El estudio incluyó características clínicas, imágenes convencionales y características radiómicas extraídas de imágenes DAT-SPECT (tomografía por emisión de fotón único), que aportan información detallada sobre la actividad dopaminérgica en el cerebro.

   Para la clasificación y predicción, los autores combinaron diferentes modelos de *machine learning*, incluyendo algoritmos de ensamblado como *Random Forest* y métodos basados en redes neuronales, para aprovechar las fortalezas de cada técnica. Esta combinación permitió mejorar la precisión y robustez del modelo predictivo.

   Los resultados mostraron que el sistema híbrido alcanzó una alta precisión en la identificación de mutaciones patogénicas y en la predicción de la progresión a enfermedad. Además, el uso de características radiómicas mejoró significativamente la capacidad predictiva en comparación con el uso exclusivo de datos clínicos o imágenes convencionales. Este trabajo destaca la eficacia del *machine learning* combinado con análisis de imágenes avanzadas para conseguir la identificación genética y la predicción clínica en la enfermedad de Parkinson.

2. **Predicción de genes de la enfermedad de Parkinson basados en Node2vec y Autoencoder**[Peng et al., 2019]

   En este estudio propusieron un enfoque híbrido de aprendizaje automático para predecir genes relacionados con la enfermedad de Parkinson, combinando algoritmos de representación de grafos y técnicas de deep learning. En su estudio, desarrollaron el modelo N2A-SVM, que emplea

el método node2vec para generar vectores de características a partir de redes de interacción proteína-proteína (PPI), seguido de un auto-encoder para reducir la dimensionalidad de los datos. Finalmente, se utilizó un clasificador SVM (Support Vector Machine) para distinguir entre genes asociados y no asociados con la enfermedad.

El modelo fue entrenado y evaluado utilizando métricas como el área bajo la curva (AUC), alcanzando valores en un rango aproximado de 0.65 a 0.85, con un valor medio cercano a 0.73. Este desempeño superó significativamente a métodos tradicionales como la caminata aleatoria con reinicio (RWR). Además, identificaron nuevos genes candidatos relacionados con la enfermedad de Parkinson, varios de los cuales fueron validados mediante literatura científica.

Este estudio demuestra cómo la combinación de técnicas avanzadas de machine learning y análisis de redes puede ser eficaz para identificar variantes genéticas relevantes, justificando enfoques computacionales similares como los utilizados en el presente trabajo.

3. **Monitorización y predicción eficaz de la enfermedad de Parkinson en ciudades inteligentes mediante un sistema de atención sanitaria inteligente** [Jatoth et al., 2022]

En este estudio, Armananzas y col. (2022) realizan una revisión del uso de técnicas de *Machine Learning* (ML) para identificar biomarcadores genéticos y transcriptómicos relevantes en la enfermedad de Parkinson. El trabajo destaca cómo los algoritmos de ML se aplican al análisis de datos ómicos como SNPs (polimorfismos de un solo nucleótido), perfiles de expresión génica y otras fuentes de información molecular, con el fin de mejorar la detección precoz y caracterización de la enfermedad.

Se evalúan distintos enfoques supervisados y no supervisados, con algoritmos como SVM, *Random Forest* o redes neuronales, aplicados sobre bases de datos públicas y experimentales. Los autores enfatizan el valor del aprendizaje automático para integrar datos genéticos con información clínica e imagenológica, permitiendo modelos predictivos más precisos. Además, se discuten los retos comunes, como el sobreajuste o la falta de interpretabilidad.

Este trabajo destaca el potencial del ML para avanzar en la medicina personalizada y apoya el desarrollo de herramientas que combinen distintas fuentes de datos para la predicción del riesgo genético de la enfermedad de Parkinson.

En cuanto a proyectos relacionados con la aplicación del aprendizaje automático en la detección y predicción de la enfermedad de Parkinson se encuentran:

1. **Predicción de la enfermedad de Parkinson mediante análisis acústico** [Requena Sánchez, 2024]

   En este trabajo se aborda la predicción de la enfermedad de Parkinson a partir de un análisis obtenido de la voz de los pacientes. Para ello, utilizaron grabaciones de voz recogidas de un aplicación móvil, sin supervisión profesional. El objetivo principal era observar si analizando diferentes aspectos de la voz (variaciones en el tono o la intensidad), se podían encontrar patrones comunes en personas con la enfermedad de Parkinson. Tras procesar y reducir el numero de variables, mediante técnicas estadísticas, se entrenó un moldeo SVM (*Support Vector Machine*) para predecir la presencia de la enfermedad solamente a partir de la voz. En cuanto a los resultados obtenidos, el modelo mostró limitaciones a la hora de identificar de forma correcta a los pacientes enfermos (baja sensibilidad), pero a pesar de ello estos estudios sugiere que este tipo de herramienta es útil como sistema complementario, accesible y no invasivo para poder apoyar el diagnostico temprano de la enfermedad de Parkinson.

2. **Estudio longitudinal del declive cognitivo en pacientes con párkinson de novo mediante modelos predictivos y modelos de progresión de la enfermedad** [Dick et al., 2007]

   Este trabajo se centra en estudiar cómo evoluciona el deterioro cognitivo en personas con enfermedad de Parkinson de reciente diagnóstico, utilizando datos recogidos a lo largo del tiempo (estudio longitudinal). Para ello, se emplearon tanto modelos predictivos como modelos de progresión de la enfermedad.

   Se utilizaron múltiples fuentes de datos del estudio PPMI, incluyendo imágenes de resonancia magnética (MRI), biomarcadores del líquido cefalorraquídeo (CSF), estudios del transportador de dopamina (DAT), y pruebas clínicas y neuropsicológicas.

   En cuanto al análisis, se aplicaron modelos de *machine learning* con validación cruzada para predecir qué pacientes podían desarrollar deterioro cognitivo leve (MCI), seleccionando las variables más relevantes mediante el método mRMR. Además, se emplearon modelos de efectos

mixtos lineales (LME) para analizar cambios estructurales en el cerebro y modelos de progresión tipo GRACE para estimar trayectorias cognitivas individuales. También se usaron modelos de supervivencia de Cox para estudiar los tiempos de conversión a MCI.

Los resultados mostraron que ciertos marcadores, como el grosor cortical y el volumen del hipocampo, junto con algunos datos clínicos, son útiles para anticipar el deterioro cognitivo. En general, el trabajo demuestra que los modelos predictivos pueden ser una herramienta valiosa para mejorar el diagnóstico temprano y personalizar el seguimiento de pacientes con enfermedad de Parkinson.

3. **Aprendizaje profundo para la clasificación de la enfermedad de Parkinson mediante imágenes PET/MR multimodales y multisecuencias** [Chang et al., 2025]

   En este estudio se propuso un enfoque basado en *Deep Learning* para la clasificación de la enfermedad de Parkinson utilizando imágenes multimodales obtenidas mediante PET y resonancia magnética (MRI). El modelo, basado en la arquitectura ResNet18, fue entrenado sobre datos de pacientes con enfermedad de Parkinson, atrofia multisistémica (MSA) y controles sanos.

   Gracias a la combinación de datos funcionales (PET) y estructurales (MRI), el sistema logró una precisión del 97 %, una sensibilidad del 98 % y un área bajo la curva (AUC) de 0.96 en la tarea de clasificación. Los resultados muestran el alto potencial de las técnicas de *Deep Learning* para integrar múltiples modalidades de imagen médica y mejorar significativamente la capacidad diagnóstica.

   Además, el estudio destaca la utilidad del aprendizaje profundo para la diferenciación entre enfermedad de Parkinson y otras enfermedades neurodegenerativas con síntomas similares, lo cual representa un avance importante respecto a estudios previos basados en modalidades únicas o técnicas de ML más tradicionales.

4. **Una revisión del aprendizaje automático y el aprendizaje profundo para la detección de la enfermedad de Parkinson** [Rabie and Akhloufi, 2025]

   En este estudio, Rabie y Akhloufi realizaron una revisión sistemática sobre el uso de técnicas de *machine learning* y *deep learning* para la detección de la enfermedad de Parkinson (EP). El análisis incluyó

múltiples fuentes de datos como voz, análisis de la marcha e imágenes médicas. Los autores compararon distintos algoritmos, entre ellos SVM, *Random Forest*, KNN, CNN y LSTM, evaluando su rendimiento, precisión y adecuación clínica.

Los resultados indican que, especialmente en modelos basados en voz y análisis de marcha, se pueden alcanzar precisiones superiores al 99 % en tareas de clasificación. Asimismo, se observa que los enfoques que combinan distintas modalidades de datos (multimodales) suelen superar en rendimiento a los modelos unimodales.

El estudio también destaca varios desafíos actuales: la escasez de conjuntos de datos grandes y balanceados, la falta de interpretabilidad de muchos modelos *deep leearning* y la necesidad de adaptar los modelos al entorno clínico. Esta revisión proporciona un marco de referencia valioso para el desarrollo de modelos predictivos más robustos, justificando enfoques como el propuesto en este trabajo de fin de grado.

A continuación, se presentan algunas aplicaciones web y herramientas informáticas desarrolladas en relación con la detección y predicción de la enfermedad de Parkinson.

1. **Parkinson-Detection-App (Streamlit)** [M., 2023]
   Esta aplicación es un sistema de detección de la enfermedad de Parkinson basado en deep learning, que emplea una red neuronal (TensorFlow/Keras) entrenada con medidas vocales del conjunto de datos de Parkinson del repositorio UCI. Incorpora SHAP para interpretar las predicciones y ofrece una interfaz web local mediante Streamlit, junto con una opción de predicción por línea de comandos. La solución no se encuentra desplegada en un servidor público: requiere que el usuario descargue el repositorio y lo ejecute en su propio entorno. Su código está diseñado para ser modular y escalable, con posibilidad de futuras ampliaciones, como soporte de entrada por voz o despliegue en la nube.

2. **Parkinson's Disease Detection from Voice Data** [Mohit, 2023]
   Esta aplicación consiste en un sistema de detección de la enfermedad de Parkinson mediante el análisis de grabaciones de voz. El proyecto emplea técnicas de machine learning para predecir la probabilidad de la enfermedad a partir de características vocales extraídas de los

datos de audio, como frecuencia fundamental, jitter, shimmer, relación armónicos-ruido, RPDE y PPE, entre otras. Se evaluaron varios modelos de clasificación (árbol de decisión, Random forest, regresión logística, SVM, KNN, naive Bayes y XGBoost), siendo este último el seleccionado por su mayor precisión y capacidad para manejar datos desbalanceados. La solución incluye una aplicación web basada en Streamlit, que permite al usuario subir o grabar su voz, visualizar las características extraídas y obtener una predicción en tiempo real.

La aplicación no está desplegada en línea: requiere ser descargada y ejecutada localmente por el usuario. Está pensada como herramienta demostrativa y educativa, y no sustituye el diagnóstico médico profesional.

3. **Parkinson Disease Web App (Streamlit)** [CoderNitu, 2023]
Este proyecto implementa un modelo de aprendizaje automático para la detección de la enfermedad de Parkinson, integrado en una aplicación web desarrollada con Streamlit. El sistema permite al usuario interactuar con el modelo a través de una interfaz sencilla para realizar predicciones a partir de datos obtenidos del conjunto disponible en Kaggle. La aplicación fue desarrollada utilizando Python, y puede ejecutarse en entornos como Spyder, Jupyter Notebook o Google Colab. Para su uso, el usuario debe descargar el repositorio y ejecutar el código localmente; no se encuentra desplegada en un servidor web accesible de forma pública. El proyecto incluye un flujo de trabajo básico documentado mediante diagramas y capturas de pantalla, y está orientado a fines educativos y de demostración.

# Metodología

## 4.1. Descripción de los datos.

Los datos utilizados en este trabajo provienen de la plataforma ***Our World in Data (OWID)***[3], una fuente de datos abiertos sobre diversas temáticas globales. Para obtener los datos necesarios, primero se exploraron los conjuntos de datos disponibles en su web, seleccionando aquellos que eran más relevantes para el trabajo.

### 4.1.1. Obtención de datos

Una vez identificados los conjuntos de datos relevantes, se procedió a localizar los endpoints[4] [Nosowitz and Goodwin, 2024] de la API de *Our World in Data (OWID)*, lo cual permitió automatizar la descarga de los datos directamente desde la fuente original y asegurar su actualización periódica.

Los endpoints identificados proporcionan archivos en formato JSON, que incluyen tanto datos numéricos como información de metadatos. En particular, se accedió a dos tipos de archivos: `metadata.json`, con información estructural (nombres de países, años disponibles), y `data.json`, que contiene los valores de los indicadores organizados por país y año. Este enfoque facilitó una integración dinámica de los datos en el entorno de análisis.

[3] **URL OWID:** https://ourworldindata.org/

[4] Un endpoint de API es una dirección URL específica donde una aplicación puede interactuar con un servidor para solicitar o enviar datos mediante la API.

El proceso completo de obtención, carga, tratamiento y estructuración de los datos se encuentra descrito con mayor detalle en el Anexo C (manual del programador).

### 4.1.2. Gestión y adaptación de los datos para los modelos

Una vez realizado este procesamiento, se obtuvieron seis tablas de datos (*dataframes*), uno de ellos correspondiente a la variable dependiente y el resto a las diferentes variables independientes. Como no todas las variables cuentan con datos para el mismo conjunto de años y países, se procedió a la unión de los distintos *dataframes*, tomando como referencia la variable objetivo, es decir, la prevalencia de la EP. De esta manera, las demás variables se alinearon únicamente con los años y países disponibles en el conjunto de datos de la variable dependiente. En los casos donde una variable carecía de datos para un país y año determinados, dichos valores se registraron como nulos para mantener la integridad del *dataset*.

Con el objetivo de determinar si era necesario aplicar una imputación de datos[5], se calculó el porcentaje de valores nulos que presentaba el *dataframe* unificado, cuyo resultado fue de un 5.39 %. Estos valores ausentes se encontraban distribuidos en diferentes columnas del *dataframe*, por lo que al realizar una eliminación de los nulos pasamos de tener 7264 filas a 5414, conservando un 75 % de los datos aproximadamente. Por todo ello, no se llevó a cabo la imputación de datos ya que se mantenían más del 50 % de los registros originales.

A partir del *dataframe* final, se obtuvieron dos nuevos conjuntos de datos: uno para el entrenamiento de modelos y otro para la predicción. El conjunto de entrenamiento contenía tanto las variables independientes como la dependiente, así como los países, e incluía todos los años disponibles excepto el último. El conjunto de predicción se componía únicamente de las variables independientes y los países correspondientes al último año disponible.

La división entre entrenamiento y predicción no se realizó mediante un porcentaje fijo predefinido, sino que se optó por una estrategia temporal: los datos del último año presentes en el conjunto de datos (2021) , representa el 3 % de los registros, se reservaron como conjunto de predicción, mientras que el resto de los años (97 %) se emplearon para el entrenamiento.

[5]**Imputación de datos:** técnica estadística que se utiliza para reemplazar valores faltantes o nulos en un conjunto de datos

Para la búsqueda de hiperparámetros y el cálculo de la importancia de variables, las columnas de País y Año se excluyeron como variables predictoras. Sin embargo, estas columnas se conservaron durante el entrenamiento y el análisis final de predicciones para mantener el contexto geográfico y temporal de los resultados.

### 4.1.3. Modelos predictivos

Antes de la etapa de modelado, se realizó un análisis exploratorio preliminar con el objetivo de evaluar la multicolinealidad entre las variables independientes y en estudiar el tipo de relación que cada una mantiene con la variable objetivo. Estos resultados y análisis preliminares se encuentran detallados en el Anexo G.

Para la predicción de la prevalencia de la EP en el año 2021, se utilizaron seis modelos de aprendizaje supervisado con el objetivo de abordar el trabajo desde diferentes metodologías y perspectivas. Los modelos seleccionados para la predicción fueron: Se emplearon distintos modelos: un GLM con distribución binomial negativa, un Random Forest Regressor, un XGBoosting Regressor, un SVR Regressor, un KNN Regressor y un MLP Regressor .

El GLM con distribución binomial negativa es un modelo lineal generalizado que permite trabajar con variables de conteo que presentan sobredispersión, ampliando la regresión de Poisson [University of California, sf]. El Random Forest Regressor es un método de ensamble que combina predicciones de múltiples árboles de decisión para mejorar la precisión y reducir el sobreajuste [Breiman, 2001]. El XGBoosting Regressor implementa un algoritmo de gradiente boosting optimizado para obtener mayor velocidad y rendimiento, creando modelos aditivos de árboles [Chen and Guestrin, 2016]. El SVR Regressor adapta las máquinas de soporte vectorial a tareas de regresión, buscando funciones que se ajusten a los datos dentro de un margen de tolerancia [Altman, 1992]. El KNN Regressor es un método no paramétrico que predice valores en función del promedio de los $k$ vecinos más cercanos [GeeksforGeeks, 2024]. Finalmente, el MLP Regressor es un perceptrón multicapa, una red neuronal capaz de aprender relaciones no lineales complejas mediante capas ocultas y funciones de activación [Kumar, 2023]. Cada modelo fue entrenado utilizando una división del 80 % de los datos para entrenamiento y el 20 % restante para prueba. Se aplicaron técnicas de validación cruzada con cinco particiones y se emplearon métodos de búsqueda de hiperparámetros como GridSearchCV , en función del modelo. Las métricas utilizadas para evaluar el rendimiento fueron el RMSE, el MAE y el coeficiente de determinación $R^2$. La configuración específica de

hiperparámetros y el razonamiento detrás de su selección se encuentran detallados en el Anexo G.

Una vez seleccionados y ajustados los modelos, se procede a determinar la importancia de las variables según cada uno de ellos. El proceso a través del cual se han determinado que variables eran significativas se encuentra explicado en el Anexo G, en el apartado de Selección de modelos.

Posteriormente, se llevó a cabo el entrenamiento individual de cada modelo para cada uno de los países disponibles. Una vez entrenados, se realizó la predicción de la prevalencia de la EP para cada país y una posterior representación visual en forma de mapa para su interpretación e integración en la aplicación.

Finalmente, a partir de los resultados de la predicción de los seis modelos, se construyó un nuevo conjunto de datos que contenía el promedio de las predicciones individuales. Esta medida fue la utilizada como predicción final de la prevalencia de la EP, con el objetivo de poder combinar las características de cada uno de los modelos y obtener un resultado lo mas representativo posible. Además, se calculó la desviación estándar entre las predicciones de los distintos modelos para cada país, con el fin de estimar el grado de incertidumbre asociado a la predicción final. Esta medida permitió identificar regiones donde existía mayor desacuerdo entre modelos, y se representó visualmente mediante un mapa. Asimismo, se elaboró un mapa de anomalías que mostraba la diferencia entre la predicción promedio y la prevalencia observada, con el objetivo de determinar si el resultado de la predicción de la EP era sobreestimado o subestimado respecto a la prevalencia real.

Además, se diseñó una aplicación web desarrollada en *Shiny* para *Python*, con el objetivo de permitir la exploración de la biometorología y los factores ambientales asociados a la EP, así como la consulta y el análisis interactivo de los resultados obtenidos a partir de los modelos predictivos. La aplicación se conecta mediante APIs, lo que posibilita el acceso automatizado y en tiempo real a los datos, ofreciendo herramientas para su visualización, comparación y descarga.

## 4.2. Técnicas y herramientas

Este apartado resume las herramientas y metodologías aplicadas durante la realización del proyecto.

### 4.2.1. Herramientas software

Para el desarrollo del trabajo ha sido necesaria la utilización de diferentes herramientas de desarrollo, tanto para el análisis de datos como para la elaboración de visualizaciones, modelado y documentación. Todas ellas han sido conocidas y utilizadas previamente durante el grado.

- **Anaconda:** Plataforma utilizada para la gestión de entornos virtuales y la instalación de paquetes en *Python*, facilitando la organización y reproducibilidad del entorno de trabajo. Se puede obtener a través de `https://www.anaconda.com/download`.
- **Jupyter Notebook:** Entorno interactivo basado en web que permitió la ejecución de código *Python*, documentación y visualización de resultados en un mismo documento, facilitando la exploración y presentación del análisis.
- **GitHub:** Plataforma de control de versiones y colaboración que se empleó para gestionar el código fuente, mantener el historial de cambios y facilitar el trabajo organizado y seguro en el proyecto [GitHub, sf].
- **Desktop GitHub:** GitHub Desktop es una herramienta de escritorio que ofrece una interfaz gráfica para trabajar con proyectos versionados mediante Git. Permite realizar acciones como guardar cambios, actualizar repositorios y colaborar con otros usuarios en GitHub de manera más accesible y visual.
- **Shiny:** Es un *framework* en R o en Python utilizado para la creación de aplicaciones web interactivas, que permitió la visualización dinámica de los resultados y mapas generados a partir de las predicciones. La elección fue la utilización de Python, ya que me encuentro más familiarizada con ese lenguaje de programación [Posit, 2022].
- **PlantUML:** Herramienta utilizada para la creación de diagramas UML, como diagramas de casos de uso y diagramas de despliegue, facilitando la documentación visual del diseño del proyecto [PlantUML, sf].
- **ChatGPT:** Herramienta de inteligencia artificial utilizada para la generación de ideas, asistencia en redacción y apoyo en la resolución de dudas durante el desarrollo del proyecto [OpenAI, sf].
- **Posit Cloud:** Plataforma en la nube utilizada para desplegar la aplicación web de manera sencilla y accesible, facilitando el acceso remoto y la publicación de resultados interactivos [PBC, 2022].

- **Microsoft Excel:** Herramienta utilizada para la elaboración y gestión de cronogramas del proyecto, permitiendo la planificación y seguimiento temporal de las tareas [Microsoft, sf].
- **Overleaf:** Plataforma online colaborativa para la escritura y edición de documentos en LaTeX, que facilitó la elaboración, revisión y organización de la memoria del proyecto [Overleaf, sf].

### 4.2.2. Lenguajes de programación

En esta sección se detallan los lenguajes de programación usados para el desarrollo del proyecto.

- **Python:** Lenguaje de programación principal empleado para el procesamiento, análisis y modelado de los datos. Se empleó por ser el lenguaje de programación más utilizado durante el grado y por tanto del que mayor conocimiento tengo.

# Resultados

## 5.1. Resumen de resultados.

### Automatización y actualización de datos mediante APIs y desarrollo de la aplicación web

Uno de los primeros resultados obtenidos fue la implementación de un sistema automatizado para la extracción y actualización de datos mediante conexión a APIs externas. Este sistema mantiene actualizadas las variables independientes y dependientes utilizadas en los modelos predictivos, asegurando que la base de datos refleje la información más reciente disponible sin necesidad de intervención manual.

Gracias a esta automatización, se desarrolló una aplicación web interactiva usando Shiny en Python, que facilita la exploración dinámica, visualización y descarga de los datos actualizados, junto con los resultados de las predicciones. La aplicación no solo permite consultar los datos, sino que también integra herramientas para la interpretación de los modelos y la visualización geoespacial de la prevalencia estimada.

Cabe destacar que, aunque los datos se actualizan automáticamente mediante APIs, las predicciones no se generan en tiempo real ni se actualizan automáticamente con cada cambio en los datos. Esta decisión garantiza un control riguroso del proceso de modelado, asegurando la calidad y fiabilidad de los resultados presentados en la aplicación.

En el Anexo B se explica con más detalle cómo funciona asi como el link directo a la misma (Ver figura 5.1).


Figura 5.1: Pantalla de inicio de la aplicación

## Importancia de Variables

Tras la obtención y procesamiento de los datos, así como la selección de los modelos más adecuados según el objetivo de la predicción y la búsqueda de los mejores hiperparámetros para cada uno de ellos, se procedió a la evaluación de la importancia de las variables en cada modelo. Los resultados individuales de esta evaluación se encuentran recogidos en el Anexo G. A partir de estos resultados, se elaboró un *ranking* promedio (Figura 5.2) de la importancia de las variables, calculado como la media de la posición ocupada por cada una de ellas en los distintos modelos considerados.

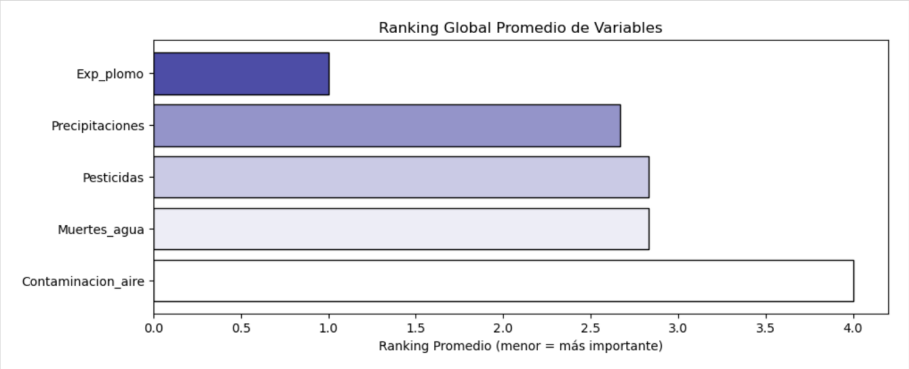



Figura 5.2: *Ranking* promedio final de la importancia de las variables

De esta manera, fue posible determinar de forma global cuáles son las variables con mayor influencia en el conjunto de datos.

Como se observa en la Figura 5.2, la variable `EXP_plomo`, aparece como la más relevante del análisis. Esto sugiere que los efectos del plomo en la salud tienen un peso considerable dentro del fenómeno estudiado. En segundo lugar, se encuentra la variable `Precipitaciones`, lo cual podría relacionarse con su influencia en la propagación de contaminantes o en la disponibilidad de agua segura.

En un nivel intermedio se sitúan `Pesticidas` y `Muertes_agua`. La relevancia de estas variables pone de manifiesto el impacto potencial de los productos químicos agrícolas y de la calidad del agua sobre la salud pública. Finalmente, la variable `Contaminación del aire`,ocupa la última posición en el ranking promedio.

## Prevalencia estimada de la enfermedad de Parkinson

Con el objetivo de obtener una predicción lo más precisa y generalizada posible de la prevalencia de la enfermedad de Parkinson, se elaboró un mapa mundial que muestra los valores promedio predichos para cada país (Figura 5.3).

Esta estimación se construyó a partir de las predicciones generadas por seis modelos distintos: GLM (Modelo Lineal Generalizado con distribución binomial negativa), Random Forest, XGBoost, Support Vector Regression (SVR), K-Nearest Neighbors (KNN) y Perceptrón Multicapa (MLP). La combinación de estas técnicas permite aprovechar las ventajas de cada una, captando distintos patrones presentes en los datos y evitando el sobreajuste asociado al uso de un único modelo.

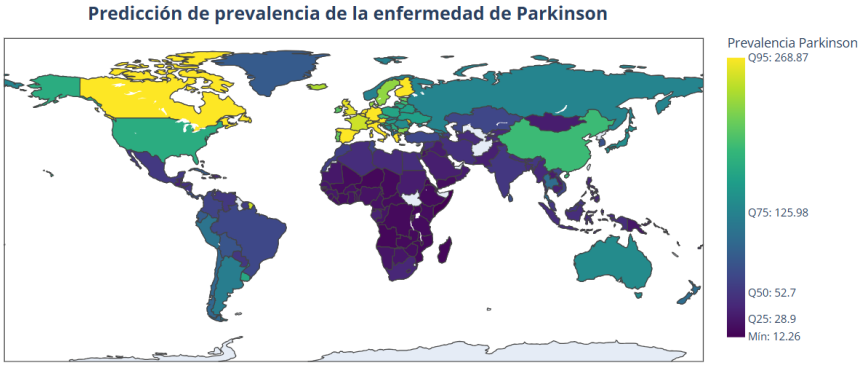



Figura 5.3: Predicción de la prevalencia de la enfermedad de Parkinson.

Para cada país, se calculó un valor promedio a partir de las predicciones de los distintos modelos, con el fin de integrar sus aportaciones individuales y reducir posibles sesgos o variaciones propias de cada técnica. De esta forma, se obtuvo una representación más estable y coherente de la prevalencia estimada de la enfermedad a escala global, al combinar distintas predicciones en lugar de basarse únicamente en una.

El análisis se representa en el mapa mediante una escala de colores basada en los cuantiles de la distribución de los valores predichos. Esta clasificación permite contextualizar los niveles relativos de prevalencia entre países:

- **Q25**: El 25 % de los países presentan valores inferiores a este cuantil, indicando prevalencias bajas.
- **Q50**: Mediana de la distribución, donde la mitad de los países se ubican por debajo y la otra mitad por encima.
- **Q75**: Tres cuartas partes de los países tienen valores menores, señalando prevalencias medias-altas.
- **Q95**: Solo el 5 % de los países superan este umbral, representando las prevalencias más altas estimadas.

A nivel geográfico, se observa una mayor prevalencia estimada en países de Europa Occidental, América del Norte y algunos países nórdicos, mientras que las estimaciones son notablemente más bajas en gran parte del continente africano y del sudeste asiático. Estas diferencias podrían estar relacionadas con factores demográficos, socioeconómicos o de disponibilidad de datos, aunque también pueden reflejar patrones reales en la distribución de la enfermedad.

Además, se evaluó el rendimiento del modelo combinado, lo que permitió cuantificar su capacidad de ajuste:

| Métrica | Valor |
|---|---|
| Error Absoluto Medio (MAE) | 25.98 |
| Error Cuadrático Medio (RMSE) | 39.19 |
| Coeficiente de Determinación ($R^2$) | 0.8547 |

Tabla 5.1: Resumen de métricas de evaluación del modelo combinado

Los resultados obtenidos muestran un error absoluto medio (MAE) de unos 26 casos, lo que significa que, en promedio, las predicciones se desvían en esa cantidad con respecto a los valores reales. El error cuadrático medio (RMSE) fue de 39.2, lo que indica que, aunque hay algunas diferencias grandes, no son muy comunes. Por último, el coeficiente de determinación (R2) fue de 0.85, lo que quiere decir que el modelo combinado es capaz de explicar el 85 % de la variación observada en los datos reales.

La Figura 5.4 permite visualizar estas métricas al comparar la prevalencia real y la predicha por el modelo para cada país. Cada punto azul representa un país, mientras que la línea roja punteada indica el lugar donde las predicciones serían exactas. La mayoría de los puntos se encuentran cerca de esta línea, lo que confirma el buen ajuste, aunque en algunos países con valores muy altos se observa una ligera tendencia a infraestimar la prevalencia, lo que contribuye al valor del RMSE.

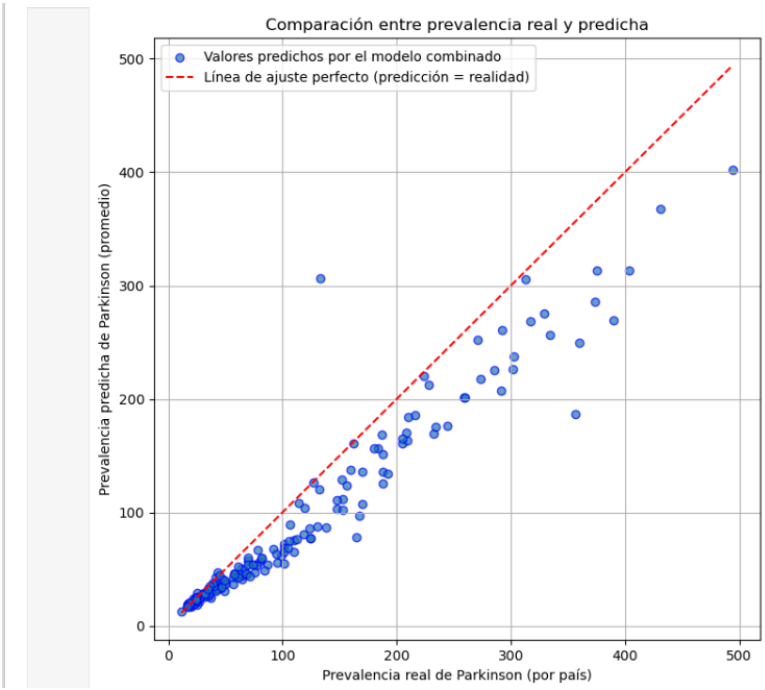



Figura 5.4: Comparación entre la prevalencia real y la predicha por el modelo combinado.

En conjunto, estos resultados reflejan que la estrategia de combinar predicciones de varios modelos es eficaz para obtener estimaciones robustas y consistentes de la prevalencia de la EP a nivel global.

## Evaluación de la calidad de las predicciones

Para valorar la fiabilidad de las estimaciones generadas, se evaluó la incertidumbre asociada a las predicciones de los distintos modelos. Esta se midió a través de la desviación estándar de los valores predichos para cada país: cuanto mayor es esta desviación, mayor es la discrepancia entre modelos y, en consecuencia, menor la confianza en la predicción agregada.(Ver Figura 5.5).

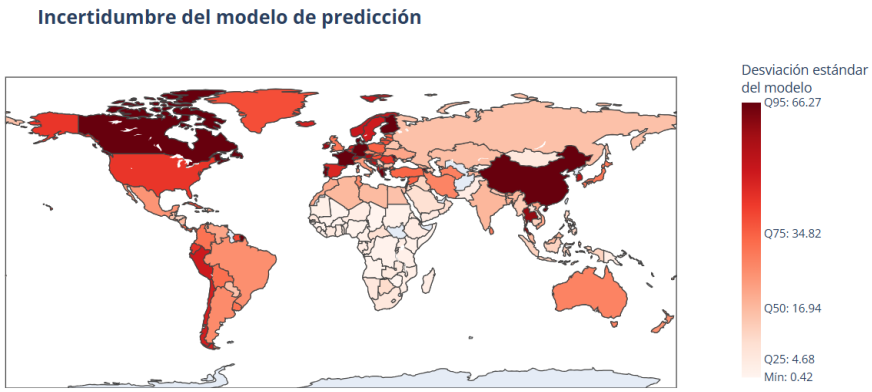



Figura 5.5: Incertidumbre del modelo de predicción

El análisis se representa en el mapa mediante una escala de colores basada en los cuantiles de la distribución de incertidumbre. Esta clasificación permite contextualizar los niveles relativos de variabilidad entre países:

- **Q25 (4.68)**: El 25 % de los países presentan una incertidumbre baja, inferior a este valor.
- **Q50 (16.94)**: Mediana de la distribución; la mitad de los países están por debajo.
- **Q75 (34.82)**: Tres cuartas partes de los países presentan valores inferiores.
- **Q95 (66.27)**: Solo el 5 % de los países superan este umbral, representando los niveles más altos de incertidumbre.

Visualmente, los colores más claros indican mayor concordancia entre modelos (baja incertidumbre), mientras que los tonos oscuros reflejan una alta variabilidad y, por tanto, menor robustez en la estimación.

Geográficamente, se observa una mayor incertidumbre en países como **China, Canadá, Brasil y algunas regiones del norte de Europa**. Esta variabilidad podría deberse tanto a factores relacionados con los datos, como heterogeneidad interna, información incompleta o presencia de ruido, como a diferencias en la forma en que los modelos procesan dichos datos en contextos complejos.

Por el contrario, regiones como **África, Europa del Este, el sudeste asiático y Oceanía** presentan menor variabilidad entre modelos. No obstante, esta aparente consistencia no implica necesariamente una mayor precisión: también podría estar influida por una menor complejidad estructural, baja diversidad en los datos disponibles o incluso limitaciones comunes en todos los modelos al abordar ciertas regiones.

En conjunto, este análisis permite identificar tanto las áreas con predicciones más sólidas como aquellas donde sería necesario mejorar la calidad y cobertura de los datos utilizados por los modelos.

Además de analizar la incertidumbre entre modelos, se evaluó el grado de ajuste de las predicciones frente a los valores reales disponibles. Para ello, se construyó un mapa de anomalías (Figura 5.6) que representa, para cada país, la diferencia entre la prevalencia real y la estimada por el promedio de los modelos. Esta visualización permite identificar de manera clara los países en los que se ha producido una sobreestimación o subestimación significativa, así como aquellos en los que el modelo presenta un buen ajuste.

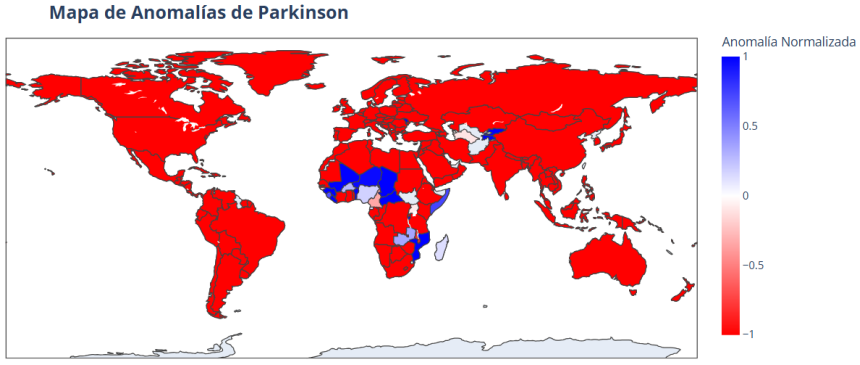



Figura 5.6: Mapa de anomalías de la enfermedad de Parkinson

A nivel global, se observa que en la mayoría de los países las predicciones tienden a subestimar la prevalencia real de la enfermedad de Parkinson, como lo refleja la amplia extensión del color rojo en el mapa. Sin embargo, se identifican anomalías positivas ,es decir, sobreestimaciones significativas.

Estas diferencias pueden deberse a varios factores, como la calidad o cantidad de datos disponibles en cada país, o a que los modelos no se adaptan igual en todas las regiones. Por eso, el análisis de anomalías no solo permite ver tan bien funciona el modelo en cada país, sino que también ayuda a detectar posibles errores o zonas donde las predicciones podrían mejorarse ajustando el enfoque utilizado.

## 5.2. Discusión

Los resultados obtenidos en este trabajo permiten conocer las diferencias respecto a estudios previos y reflexionar sobre las diferencias y aportaciones principales. En este sentido, se observan diferencias claras tanto en el objetivo como en el tipo de datos y el enfoque utilizado.

En general, los estudios revisados se centran en la detección o predicción de la enfermedad de Parkinson a nivel individual. Utilizan modelos de clasificación aplicados sobre datos clínicos, genéticos, de imagen médica (como PET, MRI, DAT-SPECT), o biomarcadores extraídos de la voz y el movimiento. En estos contextos, las métricas comúnmente reportadas son la precisión, la sensibilidad, la especificidad o el área bajo la curva (AUC), alcanzando en muchos casos valores elevados.

Sin embargo, estos resultados suelen depender de entornos controlados, bases de datos limitadas o la disponibilidad de tecnologías especializadas. En contraste, el enfoque propuesto en este trabajo se orienta al ámbito poblacional, en lugar de predecir si una persona padece la EP, se busca estimar la prevalencia de la enfermedad a nivel nacional, a partir de datos abiertos y agregados como indicadores demográficos y ambientales. Esto convierte el problema en una tarea de regresión más amplia, que implica mayor heterogeneidad y ruido en los datos, pero que al mismo tiempo aporta valor desde una perspectiva epidemiológica y de salud pública.

Metodológicamente, mientras que muchos trabajos del estado del arte aplican modelos complejos de *deep learning* sobre datos biomédicos, aquí se propone un enfoque más simple y transparente basado en el promedio de varios modelos de *machine learning* tradicionales. Esta estrategia permitió obtener un buen ajuste general sin necesidad de datos clínicos especializados.

Actualmente, la mayoría de las aplicaciones web disponibles para la detección de la enfermedad de Parkinson se centran en el análisis de un único tipo de dato, como grabaciones vocales, y requieren que el usuario ejecute localmente el código, lo que limita significativamente la interacción, accesibilidad y escalabilidad. Estas herramientas ofrecen funcionalidades básicas de predicción con modelos preentrenados, sin capacidad para incorporar dinámicamente nuevos datos o integrar diferentes fuentes de información, lo que restringe su utilidad para un análisis más amplio o adaptativo.

En contraste, la plataforma desarrollada en este trabajo supera estas limitaciones al facilitar el acceso remoto a datos ambientales y demográficos a través de APIs, integrando funcionalidades avanzadas para la visualización y exploración dinámica de la información mediante mapas interactivos a

nivel mundial y europeo. Además, esta herramienta incorpora resultados predictivos de varios modelos, junto con indicadores de incertidumbre y detección de anomalías, lo que aporta una visión más completa, transparente y confiable de la prevalencia de la enfermedad.

En resumen, aunque los estudios clínicos del estado del arte pueden ofrecer mayor precisión en contextos específicos, este trabajo demuestra que también es posible construir modelos predictivos útiles a partir de datos accesibles y abiertos, con una buena capacidad explicativa. Esto abre la puerta a nuevas formas de aplicar técnicas de *machine learning* a problemas de salud a gran escala, con menor coste y mayor escalabilidad.

# Conclusiones

Realizar este proyecto ha permitido comprobar el potencial que tienen los datos ambientales como fuente de información para el análisis y predicción de enfermedades como la EP. A lo largo del trabajo se han desarrollado modelos de *machine learning* con distintos enfoques, y se ha evaluado su rendimiento, combinando sus predicciones para obtener resultados más robustos.

Uno de los aspectos más satisfactorios ha sido integrar todos los resultados en una aplicación web interactiva, que no solo facilita el acceso a la información, sino que también convierte el análisis en una herramienta práctica y útil.

El desarrollo del proyecto no estuvo exento de dificultades, especialmente durante el tratamiento inicial de los datos y la selección adecuada de modelos. Sin embargo, cada uno de estos desafíos representó una oportunidad de aprendizaje significativa. En conjunto, el trabajo ha demostrado que, utilizando datos abiertos y herramientas accesibles, es posible construir soluciones predictivas fiables con aplicabilidad real, especialmente en el ámbito de la salud pública.

## 6.1. Aspectos relevantes.

Uno de los primeros retos significativos del proyecto fue la elección y obtención de los datos adecuados. Dado que el objetivo era analizar la relación entre variables ambientales y la prevalencia de la enfermedad de Parkinson, se buscó acceder a fuentes de datos abiertas que proporcionaran ambos tipos de información.

El enfoque inicial fue utilizar APIs públicas para automatizar la descarga y actualización de los datos. Sin embargo, este proceso resultó más complejo

de lo esperado: muchas bases de datos requerían pagos para acceder a los datos, otras no ofrecían API o tenían limitaciones técnicas importantes, y en algunos casos, los datos simplemente no estaban disponibles en formatos accesibles.

Tras una búsqueda exhaustiva, se encontró que la plataforma *Our World in Data (OWID)* ofrecía una API abierta con una gran cantidad de datos ambientales y, de forma especialmente útil, también datos de prevalencia relacionados con la EP a nivel mundial. Esta fuente resultó ser la más adecuada para los fines del proyecto, y permitió avanzar con garantías en la fase de recolección y tratamiento de datos.

Después de obtener y estructurar los datos, se realizó un análisis exploratorio preliminar con el objetivo de identificar posibles correlaciones entre las variables. Para ello, se construyó una matriz de correlación que permitió evaluar la relación entre variables independientes y también entre estas y la variable dependiente. Sin embargo, los resultados indicaron que no existían correlaciones significativas que justificaran eliminar alguna variable, por lo que se decidió mantener todas para el modelado.

En cuanto a la selección de modelos, se tuvo en cuenta tanto el tipo de datos disponibles como el objetivo principal del proyecto, que consistía en predecir el número de casos de la EP, por lo que se optó por modelos de regresión en lugar de clasificación. Se seleccionaron varios algoritmos con diferentes características y formas de abordar el problema, con la intención de obtener un modelo combinado que aprovechara las fortalezas de cada uno y redujera sus debilidades individuales.

Durante la fase de optimización, se llevó a cabo una búsqueda de hiperparámetros para mejorar el rendimiento de los modelos. No obstante, en algunos casos esta búsqueda no fue tan exhaustiva como se hubiera deseado, debido principalmente a limitaciones en recursos computacionales y tiempo. A pesar de ello, los parámetros ajustados permitieron obtener modelos suficientemente robustos para avanzar en el proyecto.

Finalmente, se evaluó la significancia de las variables en cada modelo y se procedió a entrenar y realizar predicciones. La combinación de las predicciones de los diferentes modelos permitió obtener resultados más estables y fiables, consolidando así el enfoque adoptado.

Para facilitar la visualización y exploración interactiva de los resultados, se desarrolló una aplicación web utilizando *Shiny* para Python. Esta decisión se basó en la familiaridad con el lenguaje Python, lo que permitió aprovechar mejor los conocimientos previos y acelerar el desarrollo de la herramienta.

La aplicación integra todos los resultados obtenidos durante el proyecto, ofreciendo una interfaz intuitiva que facilita el acceso a las predicciones, análisis de incertidumbre y detección de anomalías. Esto convierte el trabajo en una solución práctica y reutilizable, que puede servir como base para futuros estudios o aplicaciones.

# Lineas de trabajo futuras

Aunque el proyecto ha alcanzado sus objetivos principales, existen diversas áreas que podrían mejorarse o ampliarse en desarrollos futuros.

En primer lugar, sería interesante incluir en la aplicación desarrollada funcionalidades que permitieran al propio investigador seleccionar y cargar nuevos datos según sus necesidades, con la posibilidad de cambiar de proveedor o fuente de información de forma flexible. Esto facilitaría la integración de nuevas variables ambientales, socioeconómicas u otros factores relevantes, enriqueciendo el análisis y potenciando la capacidad predictiva de los modelos construidos.

Del mismo modo, la fase de optimización podría beneficiarse de una búsqueda más exhaustiva de hiperparámetros, utilizando técnicas de ajuste más avanzadas y evaluaciones más robustas, con el fin de evitar sobreajustes y mejorar la generalización del modelo. Aunque ya se ha explorado un modelo no lineal como el *MLP Regressor*, se podrían estudiar variantes más profundas de redes neuronales, así como arquitecturas más complejas que capten relaciones más sutiles entre variables.

En lo que respecta a la aplicación desarrollada, una de las mejoras más relevantes sería optimizar su visualización en dispositivos móviles. Actualmente, ciertos elementos como los mapas no se muestran correctamente en pantallas reducidas, por lo que adaptar la interfaz para *tablets* y *smartphones* mejoraría considerablemente la experiencia de usuario.

Finalmente, una línea de trabajo prometedora sería extender el alcance del proyecto a otras enfermedades neurodegenerativas o a nuevas variables ambientales. Esto permitiría evaluar la utilidad del enfoque propuesto en contextos distintos y comprobar su capacidad de generalización más allá del caso de la enfermedad de Parkinson.

# Bibliografía


[Altman, 1992] Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.

[Ambientum, 2025] Ambientum (2025). Enfermedades emergentes: El rol crucial de la salud ambiental. https://www.ambientum.com/ambientum/cambio-climatico/enfermedades-emergentes-el-rol-crucial-de-la-salud-ambiental.asp? Accedido: 2025-03-26.

[Armstrong and Okun, 2020] Armstrong, M. J. and Okun, M. S. (2020). Diagnosis and treatment of parkinson disease: A review. *JAMA*, 323(6):548–560.

[Ball et al., 2019] Ball, N., Teo, W.-P., Chandra, S., and Chapman, J. (2019). Parkinson's disease and the environment. *Frontiers in Neurology*, 10:218. eCollection 2019.

[Biswas et al., 2025] Biswas, B., Joseph, A., Parveen, N., Ranjan, V. P., Goel, S., Mandal, J., and Srivastava, P. (2025). Contamination of per- and poly-fluoroalkyl substances in agricultural soils: A review. *Journal of Environmental Management*, 380:124993.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

[Cao et al., 2024] Cao, Z., Yuan, Y., White, A. J., Li, C., Luo, Z., D'Aloisio, A. A., Huang, X., Kaufman, J. D., Sandler, D. P., and Chen, H. (2024). Air pollutants and risk of parkinson's disease among women in the sister study. *Environmental Health Perspectives*, 132(1):17001. Epub 2024 Jan 4.

[Chang et al., 2025] Chang, Y., Liu, J., Sun, S., Chen, T., and Wang, R. (2025). Deep learning for parkinson's disease classification using multimodal and multi-sequences pet/mr images. *EJNMMI Research*, 15(1):55.

[Chen and Guestrin, 2016] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

[CoderNitu, 2023] CoderNitu (2023). Parkinson disease web app. Aplicación web desarrollada con la biblioteca Streamlit en Python para la detección de la enfermedad de Parkinson mediante un modelo de aprendizaje automático. Requiere ejecución local y conocimientos básicos de programación.

[Díaz Cordero, 2012] Díaz Cordero, G. (2012). El cambio climático. *Ciencia y sociedad*.

[Dick et al., 2007] Dick, F. D., De Palma, G., Ahmadi, A., Scott, N. W., Prescott, G. J., Bennett, J., Semple, S., Dick, S., Counsell, C., Mozzoni, P., Haites, N., Wettinger, S. B., Mutti, A., Otelea, M., Seaton, A., Söderkvist, P., Felice, A., and study group, G. (2007). Environmental risk factors for parkinson's disease and parkinsonism: the geoparkinson study. *Occupational and environmental medicine*, 64(10):666–672.

[Fortoul van der Goes, 2022] Fortoul van der Goes, T. I. (2022). Cambio climático, la onda de calor y sus efectos en la salud. *Revista de la Facultad de Medicina (México)*, 65(5):3–5.

[GeeksforGeeks, 2024] GeeksforGeeks (2024). K-nearest neighbors (knn) regression with scikit-learn. Accedido: 2025-06-28.

[GitHub, sf] GitHub, I. (s.f.). Github. https://github.com/github. Accedido: 2025-06-13.

[González and Pérez, 2021] González, J. C. and Pérez, M. L. (2021). James parkinson y la parálisis agitante: 200 años después. *Revista de Neurología*, 72(12):501–506.

[Hajianfar et al., 2023] Hajianfar, G., Kalayinia, S., Hosseinzadeh, M., Samanian, S., Maleki, M., Sossi, V., Rahmim, A., and Salmanpour, M. R. (2023). Prediction of parkinson's disease pathogenic variants using hybrid machine learning systems and radiomic features. *Physica Medica*, 113:102647. Epub 2023 Aug 12.

[Hurtado et al., 2016] Hurtado, F., N Cárdenas, M. A., Cardenas, F., and León, L. A. (2016). La enfermedad de parkinson: Etiología, tratamientos y factores preventivos. *Universitas Psychologica*, 15(SPE5):1–26.

[Jatoth et al., 2022] Jatoth, C., E., N., A.V.R., M., and Annaluri, S. R. (2022). Effective monitoring and prediction of parkinson disease in smart cities using intelligent health care system. *Microprocessors and Microsystems*, 92:104547.

[Kumar, 2023] Kumar, A. (2023). Sklearn neural network example – mlpregressor. Accedido: 2025-06-28.

[M., 2023] M., S. (2023). Parkinson detector app. https://github.com/skjeeva/parkinson_detector_app. Aplicación en Streamlit para la predicción de la enfermedad de Parkinson basada en análisis vocal.

[Martinez, 2001] Martinez, B. B. (2001). Minería de datos. *Cómo hallar una aguja en un pajar. Ingenierías*, 14(53):53–66.

[Mayeux et al., 1997] Mayeux, R., Marder, K., Cote, L. J., and et al. (1997). Prevalence of parkinson's disease in europe: the europarkinson collaborative study. *Neurology*, 48(1):9–13.

[Microsoft, sf] Microsoft (s.f.). Microsoft excel en la nube. https://excel.cloud.microsoft/es-es/. consulted June 2025.

[Mohit, 2023] Mohit (2023). Parkinson's disease detection. https://github.com/Mohit6304/Parkinsons-Disease-Detection. Aplicación en Python para detección de la enfermedad de Parkinson usando aprendizaje automático.

[Nosowitz and Goodwin, 2024] Nosowitz, D. and Goodwin, M. (2024). What is an api endpoint? https://www.ibm.com/think/topics/api-endpoint. Accedido el 13 de junio de 2025.

[OpenAI, sf] OpenAI (s.f.). Chatgpt overview. https://openai.com/es-ES/chatgpt/overview/. consulted June 2025.

[Organización Mundial de la Salud, 2021] Organización Mundial de la Salud (2021). Cambio climático y salud. https://www.who.int/es/news-room/fact-sheets/detail/climate-change-and-health. Hoja informativa. Consultado el 28 de mayo de 2025.

[Organization, 2023] Organization, W. H. (2023). Drinking-water. *WHO Fact Sheets*. Consultado en junio de 2025.

[Overleaf, sf] Overleaf (s.f.). Overleaf: Features overview. https://es.overleaf.com/about/features-overview. consulted June 2025.

[(PAHO), 2025] (PAHO), P. A. H. O. (2025). Cambio climático y salud. https://www.paho.org/es/temas/cambio-climatico-salud? Accedido: 2025-03-26.

[Parkinson, 2002] Parkinson, J. (2002). An essay on the shaking palsy. *The Journal of Neuropsychiatry and Clinical Neurosciences*, 14(2):223–236. PMID: 11983801.

[PBC, 2022] PBC, P. (2022). Posit cloud. https://posit.cloud/. Accedido: 2025-06-13.

[Peng et al., 2019] Peng, J., Guan, J., and Shang, X. (2019). Predicting parkinson's disease genes based on node2vec and autoencoder. *Frontiers in Genetics*, 10:226.

[PlantUML, sf] PlantUML (s.f.). Plantuml. https://plantuml.com/. Accedido: 2025-06-13.

[Poewe et al., 2017] Poewe, W., Seppi, K., Tanner, C. M., Halliday, G. M., Brundin, P., Volkmann, J., Schrag, A.-E., and Lang, A. E. (2017). Parkinson disease. *Nature reviews Disease primers*, 3(1):1–21.

[Posit, 2022] Posit (2022). Shiny for python. https://shiny.posit.co/py/. Accedido: 2025-06-13.

[Postuma et al., 2015] Postuma, R. B., Berg, D., Stern, M., Poewe, W., Olanow, C. W., Oertel, W., Obeso, J., Marek, K., Litvan, I., Lang, A. E., Halliday, G., Goetz, C. G., Gasser, T., Dubois, B., Chan, P., Bloem, B. R., Adler, C. H., and Deuschl, G. (2015). Mds clinical diagnostic criteria for parkinson's disease. *Movement Disorders*, 30(12):1591–1601.

[PsicoActiva.com, 2024] PsicoActiva.com (2024). La sustancia negra del cerebro: anatomía, función y relación con el parkinson. https://www.psicoactiva.com/blog/la-sustancia-negra-del-cerebro-anatomia-funcion-relacion-parkinson/. Último acceso: 17 de junio de 2025.

[Rabie and Akhloufi, 2025] Rabie, S. and Akhloufi, M. A. (2025). A review on the application of machine and deep learning for parkinson's disease detection and monitoring. *Discover Artificial Intelligence*, 5(1):41.

[Ramos, 2014] Ramos, M. B. (2014). Biometeorología humana en la ciudad de punta alta.

[Raul et al., 2016] Raul, A., Patil, A., Raheja, P., and Sawant, R. (2016). Knowledge discovery, analysis and prediction in healthcare using data mining and analytics. In *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, pages 475–478. IEEE.

[Requena Sánchez, 2024] Requena Sánchez, C. (2024). *Predicción de la enfermedad de Parkinson mediante análisis acústico*. PhD thesis, UPC, Facultat de Matemàtiques i Estadística.

[Rodríguez et al., 2015] Rodríguez, Y., García, M. d. l. A., Martínez, M. d. C., and Rodríguez, M. d. C. (2015). Variabilidad y cambio climáticos: su repercusión en la salud. *Revista Cubana de Salud Pública*, 41(7):1–12.

[Royal Meteorological Society, 2017] Royal Meteorological Society (2017). Biometeorology: Weather and health. MetMatters. "Biometeorology studies the impact weather has on the natural world, including animals, plants and humans.".

[Royal Meteorological Society, 2022] Royal Meteorological Society (2022). Biometeorology: Weather and health. Accessed: 2025-06-17.

[RTVE, 2017] RTVE (2017). ¿cómo afecta el clima a nuestra salud?

[Simón et al., 2005] Simón, F., López-Abente, G., Ballester, E., and Martínez, F. (2005). Mortality in spain during the heat waves of summer 2003. *Euro Surveillance*, 10(7):156–161.

[Thompson and Darwish, 2019] Thompson, L. A. and Darwish, W. S. (2019). Environmental chemical contaminants in food: Review of a global problem. *Journal of Toxicology*, 2019:2345283. eCollection 2019; acceso vía PMC :contentReference[oaicite:1]index=1.

[University of California, sf] University of California (s.f.). Negative binomial regression. https://stats.oarc.ucla.edu/stata/dae/negative-binomial-regression/. Accedido: 2025-06-28.

[World Health Organization, 2022] World Health Organization (2022). Ageing and health. Accessed: 2025-06-17.

[World Health Organization, 2025] World Health Organization (2025). Air pollution. https://www.who.int/health-topics/air-pollution#tab=tab_1. Accedido: 2025-03-26.

[Zhang et al., 2018] Zhang, G., Xia, Y., Wan, F., Ma, K., Guo, X., Kou, L., Yin, S., Han, C., Liu, L., Huang, J., et al. (2018). New perspectives on roles of alpha-synuclein in parkinson's disease. *Frontiers in aging neuroscience*, 10:370.