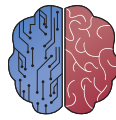




UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería de la Salud



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

19 de junio de 2025

Tutores: Antonio Canepa Oneto – Esther cubo



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



D. Tutor 1, profesor del departamento de departamento, área de área.

Expone:

Que el alumno D. Pepe Pérez, con DNI 123456A, ha realizado el Trabajo final de Grado en Ingeniería de la Salud titulado título del trabajo.

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, 19 de junio de 2025

Vº. Bº. del Tutor:

Vº. Bº. del Tutor:

D. Tutor 1

D. Tutor 2

Resumen

Como herramienta para el análisis y la creación de modelos predictivos para estimar la prevalencia de la enfermedad de Parkinson a partir de ciertas variables ambientales, se han utilizado técnicas de aprendizaje automático y minería de datos. Este enfoque permite identificar los factores ambientales que más influyen en la aparición de la enfermedad y desarrollar modelos capaces de predecir su prevalencia a nivel mundial. Los resultados contribuyen a una mejor comprensión de la relación entre el entorno y la salud, 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

As a tool for the analysis and creation of predictive models to estimate the prevalence of Parkinson's disease based on certain environmental variables, machine learning and data mining techniques have been used. This approach allows for the identification of the environmental factors that most influence the onset of the disease and the development of models capable of predicting its prevalence worldwide. The results contribute to a better understanding of the relationship between the environment and health, 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 general	iii
Índice de figuras	v
Índice de tablas	vi
Introducción	1
Objetivos	3
Conceptos teóricos	5
3.1. Enfermedad de Parkinson	5
3.2. La Biometeorología y su Relación con la Salud	7
3.3. Minería de datos	8
3.4. Estado del arte y trabajos relacionados.	9
Metodología	15
4.1. Descripción de los datos.	15
4.2. Técnicas y herramientas	18
Resultados	21
5.1. Resumen de resultados.	21
5.2. Discusión	27
Conclusiones	31
6.1. Aspectos relevantes.	31
Lineas de trabajo futuras	35

Índice de figuras

3.1. Sustancia negra del cerebro	6
5.1. <i>Ranking</i> promedio final de la importancia de las variables	21
5.2. Predicción de la prevalencia de la enfermedad de Parkinson . . .	22
5.3. Incertidumbre del modelo de predicción	24
5.4. Mapa de anomalías de la enfermedad de Parkinson	25
5.5. Pantalla de inicio de la aplicación	26
5.6. Comparación entre la prevalencia real y la predicha por el modelo combinado.	28

Índice de tablas

5.1. Resumen de métricas de evaluación del modelo combinado . . .	27
---	----

Introducción

El párkinson es una enfermedad neurodegenerativa que afecta a alrededor de 10 millones de personas en el mundo, siendo esta la segunda enfermedad neurodegenerativa más común tras el 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 envejecimiento 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 enfermedad de Parkinson 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 enfermedad de Parkinson [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 el párkinson [Royal Meteorological Society, 2022].

Por todo ello, en este trabajo se ha utilizado la minería de datos y el aprendizaje automático para analizar qué variables ambientales pueden influir en el desarrollo de la enfermedad a través de la elaboración de modelos predictivos.

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 a partir de variables biometeorológicas, con el fin de identificar factores ambientales que puedan influir en su desarrollo.

Objetivos Específicos

■ Objetivos funcionales y técnicos:

- Obtener y extraer 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 párkinson.
- Implementar un modelo final basado en el promedio ponderado de los modelos individuales para mejorar la predicción.
- Desarrollar 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.

■ **Objetivos de aprendizaje:**

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

Conceptos teóricos

3.1. Enfermedad de Parkinson

La enfermedad de Parkinson 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 enfermedad de Parkinson 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 una mayor incidencia en hombres que en mujeres[[Armstrong and Okun, 2020](#)].

La enfermedad de Parkinson 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](#)]). El párkinson 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 de párkinson es la acumulación de una proteína llamada α -sinucleína. Esta proteína, en condiciones normales, tiene la función de liberar neurotransmisores en el cerebro, pero en pacientes

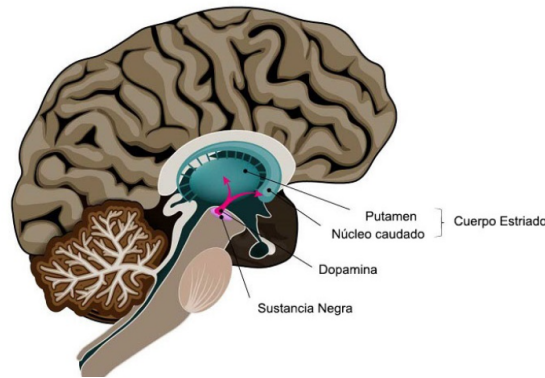


Figura 3.1: Sustancia negra del cerebro

con párkinson, se pliega de manera anormal, formando agregados que se acumulan en las neuronas, lo que contribuye al daño celular. Estos depósitos de α -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].

El diagnóstico de la enfermedad de Parkinson (EP) se basa principalmente en la evaluación clínica. Los criterios más reconocidos fueron establecidos por la *UK Parkinson Disease Society - Brain Bank* e incluyen cuatro signos clave: Bradicinesia ¹, temblor en reposo, rigidez e inestabilidad postural.[Marín et al., 2018]

El tratamiento de los síntomas motores de la enfermedad de Parkinson 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 combinación de

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.

otros), así como no farmacológicas como el ejercicio físico, fisioterapia, terapia de lenguaje entre otras.

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] Además, 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 número 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 transmitidas 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].

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 párkinson, 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, esta revisión bibliográfica puede ser de ayuda a la elaboración del mismo.

Algunos estudios y proyectos sobre la detección genética del párkinson 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 *machine 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 el párkinson.

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 autoencoder 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), logrando un valor de 0.7289, superando significativamente a métodos tradicionales como la caminata aleatoria con reinicio (RWR). Además, identificaron nuevos genes candidatos relacionados con párkinson, 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 predicti-

vos 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 una 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 párkinson. Tras procesar y reducir el número 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 diagnóstico temprano de párkinson.

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 párkinson 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 párkinson.

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 párkinson, 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 párkinson 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 learning* 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.

Metodología

4.1. Descripción de los datos.

Los datos utilizados en este trabajo provienen de la plataforma ***Our World in Data (OWID)***, 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 endpoints³ [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.

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).

³Un endpoint de API es una dirección URL específica donde una aplicación puede interactuar con un servidor para solicitar o enviar datos mediante la API.

4.1.2. Gestión y adaptación de los datos para los modelos

Una vez realizado este procesamiento, se obtuvieron seis *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 párkinson. 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 ⁴, 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 %.

A partir del *dataframe* final, se obtuvieron dos nuevos conjuntos de datos, uno para el entrenamiento de modelos y otro para la predicción. El *dataframe* de entrenamiento contenía tanto las variables independientes como la dependiente, así como los países. Este conjunto abarcaba todos los años disponibles a excepción del último, ya que, es el año que se reservó para la predicción. En cuanto al *dataframe* restante, contenía únicamente las variables independientes y los países utilizando los datos correspondientes al último año disponible en el *dataframe* final.

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.

⁴**Imputación de datos:** técnica estadística que se utiliza para reemplazar valores faltantes o nulos en un conjunto de datos

Para la predicción de la prevalencia de párkinson 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: Modelo GLM con distribución binomial negativa, modelos basados en árboles (*Random Forest Regressor*, *XGBoosting Regressor*), *SVR Regressor*, *KNN Regressor* y *MLP Regressor*. 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 párkinson 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 párkinson, con el objetivo de poder combinar las características de cada uno de los modelos y obtener un resultado lo más 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 párkinson era sobreestimado o subestimado respecto a la prevalencia real.

Para la integración visual de todos los resultados obtenidos se elaboró un aplicación en *Shiny* con *python*.

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](#),].
- **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.
- **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](#),].
- **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](#),].
- **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](#), [a](#)].

- **Microsoft Excel:** Herramienta utilizada para la elaboración y gestión de cronogramas del proyecto, permitiendo la planificación y seguimiento temporal de las tareas.
- **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.

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.
- **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 [PBC, b].

Resultados

5.1. Resumen de resultados.

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 5.1) de la importancia de las variables, calculado como la media de la posición ocupada por cada una de ellas en los distintos modelos considerados.

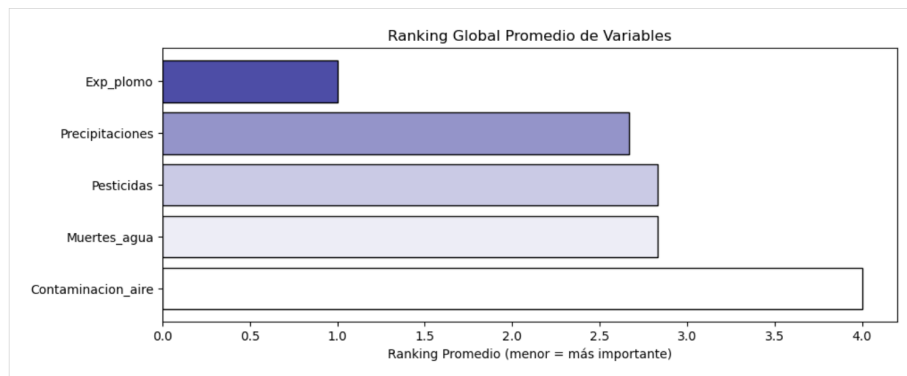


Figura 5.1: *Ranking* promedio final de la importancia de las variables

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 5.1, 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 5.2).

Esta estimación se construyó a partir de las predicciones generadas por seis modelos distintos: GLM (Modelo Lineal Generalizado), Random Forest, XGBoost, Support Vector Regression, K-Nearest Neighbors y Perceptrón Multicapa.



Figura 5.2: 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.

Esta forma de representar los datos facilita la comparación relativa entre países, mostrando claramente cómo se distribuyen las prevalencias estimadas en una escala global.

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.

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 5.3).

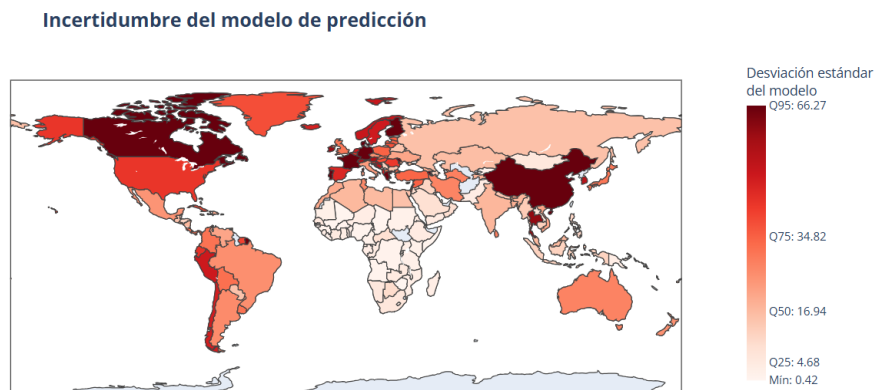


Figura 5.3: Incertidumbre del modelo de predicción

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 5.4) 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.

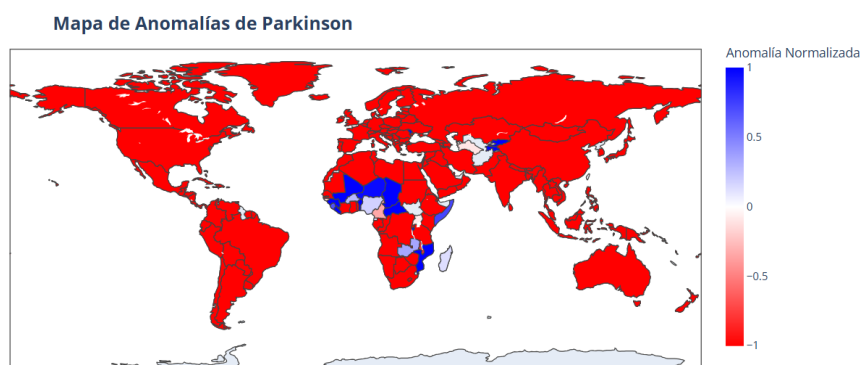


Figura 5.4: Mapa de anomalías de la enfermedad de Parkinson

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.

Desarrollo de Aplicación

Por último, se creó una aplicación interactiva que reúne todo el estudio y sus resultados, pensada para que el usuario pueda explorarlos de forma visual y sencilla. En el Anexo B se explica con más detalle cómo funciona así como el link directo a la misma (Ver figura 5.5).



Figura 5.5: Pantalla de inicio de la aplicación

5.2. Discusión

En este trabajo se probó una estrategia que consiste en combinar las predicciones de varios modelos distintos: GLM con distribución binomial negativa, Random Forest, XGBoost, SVR, MLP y KNN. La idea fue aprovechar las ventajas que ofrece cada uno para mejorar la predicción de la prevalencia de párkinson por país. Usar varios modelos ayuda a captar diferentes patrones en los datos y a evitar que un solo modelo se ajuste demasiado (lo que se llama sobreajuste).

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 (R^2) 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.

Tabla 5.1: Resumen de métricas de evaluación del modelo combinado

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

En la Figura 5.6 se puede ver una comparación entre la prevalencia real de párkinson y la predicha por el modelo en cada país. Cada punto azul representa un país. La línea roja punteada marca el sitio donde las predicciones serían exactas. La mayoría de los puntos están bastante cerca de esa línea, lo que muestra que el modelo funciona bien. Aun así, hay algunos casos, sobre todo en países con valores muy altos, donde el modelo tiende a predecir por debajo del valor real. Esto también ayuda a entender por qué el RMSE no es más bajo.

En resumen, combinar predicciones de varios modelos es una buena estrategia para obtener estimaciones más precisas y consistentes. Esta técnica permite reflejar mejor la complejidad de los datos y es útil cuando se trata de fenómenos como la prevalencia de párkinson, que pueden estar influenciados por muchos factores distintos entre países.

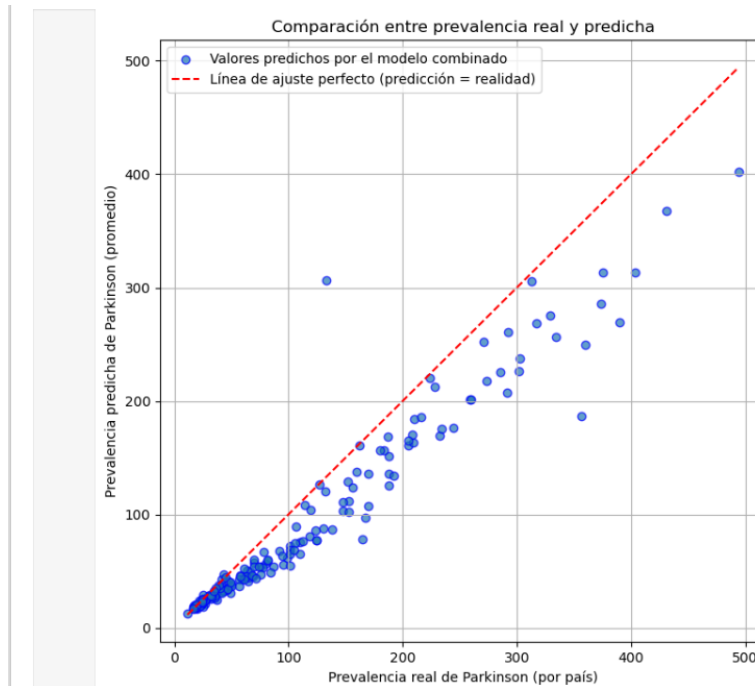


Figura 5.6: Comparación entre la prevalencia real y la predicha por el modelo combinado.

Comparación con el estado del arte

Al comparar este trabajo con el conjunto de investigaciones recogidas en el estado del arte, se identifican diferencias significativas tanto en el objetivo como en el enfoque metodológico y el tipo de datos empleados.

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 párkinson, 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.

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 el párkinson. 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 el párkinson 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 párkinson, 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 recomendable ampliar la búsqueda y el acceso a fuentes de datos, incorporando nuevas variables ambientales o incluso factores socioeconómicos. Esto permitiría contextualizar mejor los resultados y mejorar 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

- [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?utm_source=chatgpt.com. 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.
- [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.

- [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., Haite, 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.
- [GitHub,] GitHub, I. 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.
- [Marín et al., 2018] Marín, D. S., Carmona, H., Ibarra, M., and Gámez, M. (2018). Enfermedad de parkinson: fisiopatología, diagnóstico y tratamiento. *Revista de La Universidad Industrial de Santander. Salud*, 50(1):79–92.
- [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.
- [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,] OpenAI. Chatgpt, modelo de lenguaje desarrollado por openai. <https://chat.openai.com/>. Accedido: 2025-06-13.
- [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.
- [(PAHO), 2025] (PAHO), P. A. H. O. (2025). Cambio climático y salud. https://www.paho.org/es/temas/cambio-climatico-salud?utm_source=chatgpt.com. 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, a] PBC, P. Posit cloud. <https://posit.cloud/>. Accedido: 2025-06-13.
- [PBC, b] PBC, P. Shiny for python. <https://shiny.posit.co/py/>. 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,] PlantUML. Plantuml. <https://plantuml.com/>. Accedido: 2025-06-13.
- [Poewe et al., 2017] Poewe, W., Seppi, K., Tanner, C. M., Halliday, G. M., Brundin, P., Volkman, J., Schrag, A.-E., and Lang, A. E. (2017). Parkinson disease. *Nature reviews Disease primers*, 3(1):1–21.

- [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, 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.
- [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.