



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



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

**título del TFG
Documentación Técnica**

Presentado por nombre alumno
en Universidad de Burgos

27 de mayo de 2025

Tutores: nombre tutor – nombre tutor 2

Índice general

Índice general	i
Índice de figuras	iii
Índice de tablas	v
Apéndice A Plan de Proyecto Software	1
A.1. Introducción	1
A.2. Planificación temporal	1
A.3. Planificación económica	2
A.4. Viabilidad legal	3
Apéndice B Documentación de usuario	5
B.1. Requisitos software y hardware para ejecutar el proyecto.	5
B.2. Instalación / Puesta en marcha	5
B.3. Manuales y/o Demostraciones prácticas	6
Apéndice C Manual del desarrollador / programador / investigador.	17
C.1. Estructura de directorios	17
C.2. Compilación, instalación y ejecución del proyecto	17
C.3. Pruebas del sistema	19
C.4. Pruebas del sistema	20
C.5. Instrucciones para la modificación o mejora del proyecto.	20
Apéndice D Descripción de adquisición y tratamiento de datos	23
D.1. Descripción formal de los datos	23

D.2. Descripción clínica de los datos.	26
Apéndice E Manual de especificación de diseño	29
E.1. Diseño arquitectónico	29
Apéndice F Especificación de Requisitos	33
F.1. Diagrama de casos de uso	33
F.2. Explicación casos de uso.	34
F.3. Prototipos de interfaz o interacción con el proyecto	35
Apéndice G Estudio experimental	37
G.1. Cuaderno de trabajo.	37
G.2. Configuración y parametrización de las técnicas.	45
G.3. Detalle de resultados.	53
Apéndice H Anexo de sostenibilización curricular	55
H.1. Introducción	55
Bibliografía	57

Índice de figuras

B.1. Página de inicio de la aplicación (Pestaña Home)	7
B.2. Contenido de la Sección de Enfermedad de Parkinson	7
B.3. Contenido de la Sección de Mapa Mundial de Parkinson	8
B.4. Contenido del Ver Mapa europeo en la sección de Mapa Mundial de Parkinson	9
B.5. Desacarga de datos en la sección de Mapa Mundial de Parkinson	9
B.6. Contenido de la sección de Variables ambientales	10
B.7. Contenido botón Contaminación del aire	11
B.8. Contenido botón Ver Mapa Europeo dentro de la sección de Contaminación de aire	11
B.9. Contenido de la sección de Variables ambientales (Descarga de datos)	11
B.10. Contenido de la sección de Predicciones	12
B.11. Contenido de la sección de Predicciones (mapas)	12
B.12. Contenido de la sección de Importancia de Variables	13
B.13. Contenido inferior de la sección de Importancia de variables	13
B.14. Contenido botón Modelo lineal	14
B.15. Contenido botón Modelo lineal al pulsar Ver Mapa Europeo	14
B.16. Contenido botón Modelos basados en árboles	15
B.17. Contenido botón Random Forest	15
B.18. Contenido botón Random Forest al pulsar Ver Mapa Europeo	15
B.19. Contenido botón Otros modelos de regresión	15
E.1. Diagrama de despliegue de la aplicación web.	30
E.2. Diagrama de flujo del proceso de modelado y predicción.	31
E.3. Diagrama de paquetes del sistema.	32
F.1. Diagrama de casos de uso de la aplicación.	33

G.1. Matriz de correlación entre variables independientes	38
G.2. Relación entre el Parkinson y las Muertes atribuidas a fuentes de agua inseguras.	38
G.3. Relación entre el Parkinson y la Tasa de carga de enfermedad por exposición al plomo.	39
G.4. Relación entre el Parkinson y la Tasa de mortalidad por conta- mación de aire.	39
G.5. Relación entre el Parkinson y el uso de pesticidas.	40
G.6. Relación entre el Parkinson y el uso de pesticidas.	40
G.7. Modelo Cuassi-poisson.	42
G.8. Modelo-Binomial Negativo	42

Índice de tablas

A.1. Costes de hardware	2
A.2. Costes de personal simulados para el proyecto	3
A.3. Coste total del proyecto	3
F.1. CU-1 Nombre del caso de uso.	36
G.1. Configuración aplicada al modelo GLM (Binomial Negativa) . .	45
G.2. Configuración aplicada al modelo Random Forest	46
G.3. Configuración aplicada al modelo XGBoost	48
G.4. Parámetros utilizados en el modelo SVR con variables transformadas	49
G.5. Parámetros utilizados en el entrenamiento del modelo KNN . .	51
G.6. Parámetros utilizados en el entrenamiento del modelo MLP . .	52

Apéndice A

Plan de Proyecto Software

A.1. Introducción

Este anexo presenta el Plan de Proyecto Software elaborado para guiar el desarrollo del trabajo. En él se abordan distintos aspectos clave relacionados con la planificación y gestión del proyecto, con el objetivo de asegurar su viabilidad tanto técnica como económica y legal. Entre los elementos que se tratan se encuentra la planificación temporal, que permite organizar las tareas en distintas fases; la estimación de costes, necesaria para valorar el esfuerzo económico aproximado del desarrollo; y un análisis de la viabilidad legal, centrado en garantizar que el proyecto cumple con la normativa aplicable.

Ojo ¹

A.2. Planificación temporal

La planificación temporal del proyecto ha sido clave para estructurar adecuadamente el desarrollo y garantizar el cumplimiento de los objetivos en los plazos establecidos. Para ello, el trabajo se ha dividido en diferentes hitos o *milestones*, cada uno de los cuales agrupa un conjunto de tareas o *issues* específicas.

Cada *issue* ha sido estimada y registrada con su correspondiente tiempo de dedicación, permitiendo así un seguimiento detallado del progreso. Esta

¹Los anexos deben de tener su propia bibliografía, eso es tan fácil como utilizar referencias igual que en la memoria [?]

organización ha facilitado una gestión eficiente del tiempo y ha servido como referencia para evaluar el grado de avance en cada fase del proyecto.

En esta sección se presenta el cronograma completo del desarrollo, incluyendo los distintos hitos alcanzados y el desglose temporal de las tareas realizadas.

A.3. Planificación económica

En esta sección se presenta una estimación económica teórica del proyecto, considerando los costes de hardware, software y personal, con el fin de reflejar el valor aproximado que tendría su desarrollo en un entorno profesional. Cabe destacar que, al tratarse de un Trabajo de Fin de Grado, no se ha incurrido en gastos reales más allá del uso del equipo personal del autor.

A.3.1. Costes de hardware

No se ha adquirido nuevo hardware para el desarrollo del proyecto. No obstante, se ha estimado el coste de amortización del equipo utilizado (ordenador portátil personal) a lo largo de un periodo de 6 años, distribuyendo el coste proporcionalmente al tiempo de duración del proyecto (4 meses).

Concepto	Coste en €	Coste amortizado (4 meses)
Ordenador portátil	1.300	72,22

Tabla A.1: Costes de hardware

A.3.2. Costes de software

Todo el software empleado en el desarrollo del proyecto es de código abierto y gratuito, por lo que no ha supuesto ningún coste económico.

A.3.3. Costes de personal

Para estimar los costes de personal, se ha simulado una contratación ficticia durante 4 meses con una dedicación equivalente a media jornada. Dado que el proyecto involucra tareas propias de un perfil mixto —minería de datos, aprendizaje automático, desarrollo de scripts automatizados y construcción de una aplicación web en R (Shiny)— se ha tomado como referencia un salario bruto mensual estimado de 2.500€, basado en los rangos

salariales habituales en España para perfiles junior con competencias en ciencia de datos y desarrollo web [y ESADE, 2023, Glassdoor, 2024].

Concepto	Coste en €
Salario mensual neto estimado	1.697,22
Retención IRPF (19 %)	322,47
Seguridad Social (28.3 %)	480,31
Salario mensual bruto	2.500,00
Total 4 meses	10.000,00

Tabla A.2: Costes de personal simulados para el proyecto

Coste total estimado

Sumando los costes de hardware y personal, y considerando que el software no ha supuesto gasto, el coste total estimado del proyecto sería el siguiente:

Concepto	Coste en €
Costes de hardware	72,22
Costes de personal	10.000
Coste total estimado	10.072,22

Tabla A.3: Coste total del proyecto

A.4. Viabilidad legal

El proyecto se ha basado en la utilización de bases de datos públicas y abiertas relacionadas con enfermedades neurológicas y contaminación ambiental, las cuales cuentan con licencias que permiten su uso y análisis para fines académicos y de investigación. Se ha verificado que estas fuentes cumplen con las normativas vigentes en materia de protección de datos personales y confidencialidad.

Todo el software utilizado, incluyendo Python, sus bibliotecas de aprendizaje automático (como scikit-learn, pandas, etc.), y herramientas para el análisis y tratamiento de datos como Jupyter Notebook, así como frameworks para la creación de aplicaciones web (por ejemplo, Streamlit o Dash), es de código abierto bajo licencias compatibles con el desarrollo y distribución de proyectos académicos y científicos.

No se ha empleado ninguna información personal identificable, por lo que no se requiere autorización específica para el tratamiento de datos sensibles conforme al Reglamento General de Protección de Datos (RGPD).

En conclusión, el desarrollo del proyecto cumple con los requisitos legales y éticos necesarios para garantizar la viabilidad legal del mismo.

Apéndice B

Documentación de usuario

B.1. Requisitos software y hardware para ejecutar el proyecto.

Para utilizar correctamente la aplicación web, el usuario debe contar con un entorno básico que garantice el acceso y correcto funcionamiento del sistema. A continuación se detallan los requisitos mínimos recomendados:

- **Navegador web:** Google Chrome, Mozilla Firefox, Microsoft Edge o Safari, en sus versiones actualizadas.
- **Conexión a internet:** Se recomienda una conexión estable con una velocidad mínima de 6 Mbps. Esta cifra se basa en pruebas reales del tiempo y volumen de carga de la aplicación (aproximadamente 1.6 MB de recursos en menos de 2 segundos), lo que garantiza una experiencia fluida al visualizar mapas y utilizar funcionalidades clave.
- **Resolución de pantalla:** Mínimo 1280x720 píxeles para una visualización óptima.

B.2. Instalación / Puesta en marcha

La aplicación está desplegada en línea, por lo tanto no requiere instalación local. Para acceder a ella, el usuario simplemente debe abrir un navegador web compatible y dirigirse a la siguiente dirección:

- <https://lorenacalvoperez-parkinson-worldwide.share.connect posit.cloud/>

Una vez en el sitio web, el usuario podrá comenzar a utilizar las funcionalidades de la plataforma.

B.3. Manuales y/o Demostraciones prácticas

Esta sección proporciona una guía práctica para el uso de la aplicación web **Parkinson Worldwide**, una herramienta interactiva de acceso público que no requiere registro ni inicio de sesión. La aplicación está diseñada para ser intuitiva y fácil de explorar, incluso para usuarios sin experiencia técnica previa.

La navegación por la plataforma se realiza a través de una barra lateral situada en la parte izquierda de la pantalla. Desde esta barra, el usuario puede acceder a las distintas secciones de la aplicación simplemente haciendo clic sobre el botón correspondiente. Cada sección muestra un tipo de información específico, ya sea contenido explicativo, visualizaciones interactivas o gráficos analíticos, según el objetivo de cada módulo.

A continuación, se describen de forma detallada las funcionalidades de cada sección, acompañadas de capturas de pantalla que ilustran el contenido visual y la experiencia de uso.

1. Página de Inicio (Home)

Al acceder a la aplicación web, el usuario es recibido en la sección principal llamada **Home**(ver Figura B.1). Esta página ofrece una introducción general sobre el contenido y el propósito de la aplicación. Se explica de forma sencilla qué tipo de información podrá explorar el usuario, así como los análisis que se han realizado en torno a la enfermedad de Parkinson.

También se informa sobre el origen de los datos utilizados en la elaboración de los gráficos y visualizaciones disponibles en la plataforma. Para facilitar al usuario el acceso a la web de la cual se han obtenido los datos, se ha colocado en link al enlace de forma directa.

B.3. Manuales y/o Demostraciones prácticas

7



Figura B.1: Página de inicio de la aplicación (Pestaña Home)

2. Enfermedad de Parkinson

Esta es una sección informativa que tiene como objetivo ofrecer al usuario una introducción general sobre la enfermedad de Parkinson. El contenido está redactado de forma clara y accesible, pensado para personas sin conocimientos médicos previos como puede verse en la Figura B.2



Figura B.2: Contenido de la Sección de Enfermedad de Parkinson

Se explican brevemente los síntomas más comunes de la enfermedad, así como los principales factores de riesgo asociados a su desarrollo. La información se presenta de manera visual, mediante una infografía sencilla, que permite una lectura rápida y comprensible.

No se requiere ninguna interacción por parte del usuario en esta sección; su función es puramente divulgativa y sirve como base contextual para entender mejor el resto de los análisis presentados en la aplicación.

3. Mapa Global del Parkinson

Esta sección permite visualizar, de forma interactiva, la prevalencia estimada de la enfermedad de Parkinson a nivel mundial. Al acceder, el usuario encontrará una breve descripción que introduce los datos representados en el mapa, así como un enlace directo a la fuente original de dichos datos. En concreto, se trata de los datos brutos sobre la prevalencia global del Parkinson, expresados como el número estimado de casos en cada país a lo largo del tiempo.

El mapa está diseñado para ser intuitivo: el usuario puede desplazarse por las regiones del mundo y pasar el cursor sobre un país para consultar los valores específicos correspondientes a cada año. Además, se incorpora un control deslizante (slidebar) en la parte inferior, que permite seleccionar el año deseado dentro del rango disponible en el conjunto de datos. Esto facilita el análisis visual de cómo ha evolucionado la prevalencia del Parkinson a lo largo del tiempo.(Véase la Figura B.3)



Figura B.3: Contenido de la Sección de Mapa Mundial de Parkinson

Debajo del mapa se incluye un botón titulado "Ver mapa europeo", que brinda al usuario la oportunidad de visualizar la zona europea con un mayor nivel de detalle, mediante un enfoque tipo zoom. Esta funcionalidad permite una mejor visualización de los datos específicos de esta región, lo que resulta especialmente útil dado el tamaño reducido de algunos países europeos en la vista global.(Véase la Figura B.4)

Una vez en esta vista detallada de Europa, el usuario dispone de un botón adicional titulado "Volver al mapa global", que permite regresar fácilmente a la visualización completa del mapa mundial. De este modo, se facilita la navegación fluida entre ambas vistas.

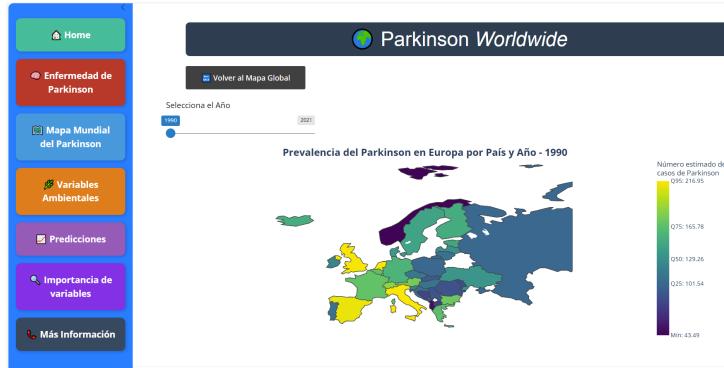


Figura B.4: Contenido del Ver Mapa europeo en la sección de Mapa Mundial de Parkinson

Inmediatamente debajo del control deslizante y del botón "Ver mapa europeo", se encuentra una sección que permite descargar el conjunto completo de datos en formatos JSON y CSV. Esta opción ofrece al usuario la posibilidad de obtener los datos brutos sin filtros, útiles para análisis externos o integración en otros sistemas.

Más abajo, aparece una subsección titulada "Filtrar los datos por año y país", donde el usuario puede seleccionar uno o varios años específicos, así como los países de interés. Esta herramienta de filtrado permite personalizar el conjunto de datos a descargar, facilitando análisis más concretos. Una vez realizada la selección, se habilitan los botones correspondientes para descargar los datos filtrados en formato CSV o JSON, según se prefiera.(vease Figura B.5)

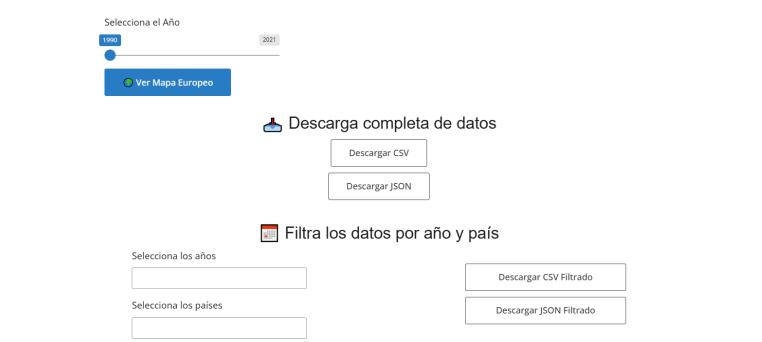


Figura B.5: Desacraza de datos en la sección de Mapa Mundial de Parkinson

4. Variables Ambientales

Esta sección de la aplicación está dedicada a la exploración de distintos factores ambientales que se han estudiado por su posible relación con la enfermedad de Parkinson. El objetivo es ofrecer una herramienta visual e interactiva que permita observar cómo varían estos factores en el tiempo y entre distintas regiones del mundo.

En concreto, se incluye Contaminación del Aire, Exposición al Plomo, Aguas Inseguras, Uso de Pesticidas, Precipitaciones.(Véase la Figura B.6)

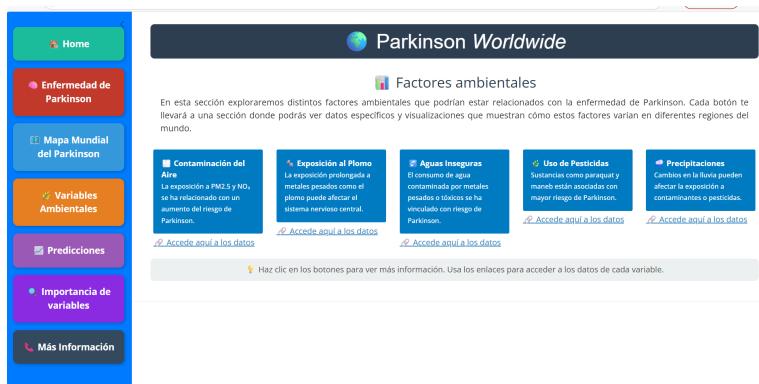


Figura B.6: Contenido de la sección de Variables ambientales

Cada uno de estos factores se presenta mediante un botón interactivo que permite acceder a una sección específica con información detallada. Debajo de cada botón, se incluye un enlace directo a la fuente original de los datos brutos utilizados, lo que permite al usuario consultar o descargar la información directamente desde su origen. La estructura de estas secciones es idéntica a la utilizada en la sección "Mapa Mundial del Parkinson", lo cual ofrece una experiencia de usuario coherente y sencilla. Concretamente, cada página incluye:

- Un mapa mundial interactivo que muestra la distribución del factor ambiental junto con un control deslizante (slider) para explorar los datos en distintos años.

Además de un botón para visualizar un mapa europeo con mayor nivel de detalle así como un botón para volver al mapa global, facilitando la navegación.(vease Figura B.7 y Figura B.8)

- Herramientas para filtrar los datos por país y año, y descargar los resultados en formato CSV o JSON.(vease Figura B.9)



Figura B.7: Contenido botón Contaminación del aire

Figura B.8: Contenido botón Ver Mapa Europeo dentro de la sección de Contaminación de aire



Figura B.9: Contenido de la sección de Variables ambientales (Descarga de datos)

5. Predicciones

Esta sección permite visualizar, de forma interactiva, las estimaciones generadas por modelos de predicción respecto a la prevalencia de la enfermedad de Parkinson en distintos países del mundo. Está diseñada para que el usuario pueda explorar fácilmente tanto los valores estimados como la fiabilidad de dichas predicciones y las posibles desviaciones respecto a los datos reales.(vease Figura B.10)

Al acceder a esta pestaña, el usuario encontrará tres mapas principales, presentados de forma ordenada y con navegación similar a otras secciones de la aplicación:

- **Mapa de prevalencia estimada:** Este mapa muestra el valor promedio de la prevalencia del Parkinson predicho por seis modelos diferentes de aprendizaje automático. La visualización permite observar cómo se distribuyen estas estimaciones por país y por año.



Figura B.10: Contenido de la sección de Predicciones

- **Mapa de incertidumbre:** Justo debajo del mapa anterior, el usuario encontrará un segundo mapa que representa la desviación estándar de las predicciones realizadas por los seis modelos. Este valor indica el grado de acuerdo entre los modelos: una menor desviación implica que los modelos han generado estimaciones similares para ese país y año, mientras que una desviación mayor refleja discrepancias entre los resultados.
- **Mapa de anomalías:** Por último, se presenta un mapa que muestra las diferencias entre los valores predichos y los valores reales conocidos. Las anomalías ayudan a detectar posibles zonas donde el modelo haya sobreestimado (países en azul) o subestimado (países en rojo) la prevalencia de la enfermedad. Véase ambos mapas en la Figura B.11

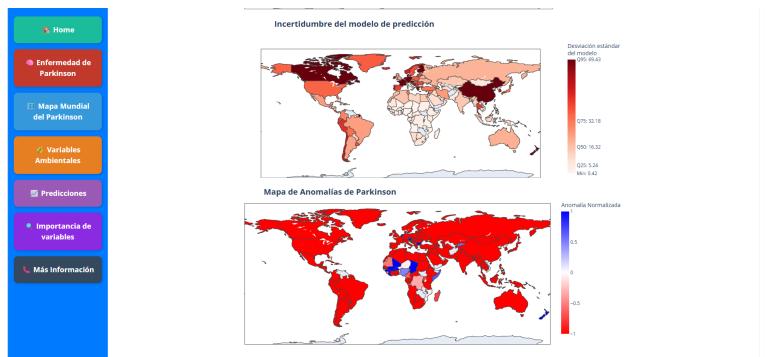


Figura B.11: Contenido de la sección de Predicciones (mapas)

6. Importancia de variables

Esta sección ofrece una visión general sobre la influencia de las distintas variables ambientales en los modelos de predicción de la enfermedad de Parkinson. Su objetivo es ayudar al usuario a identificar qué factores han sido considerados más relevantes por los modelos de aprendizaje automático utilizados.

Al acceder, el usuario encontrará una breve explicación sobre el propósito de esta sección, seguida de un gráfico interactivo que muestra el ranking promedio de importancia de cada variable. Esta importancia se ha calculado a partir de los resultados obtenidos por todos los modelos entrenados, proporcionando así una perspectiva agregada y global. En el gráfico, cuanto más bajo es el valor del ranking, mayor es la relevancia de la variable en las predicciones. Es decir, las variables que ocupan las primeras posiciones han sido las más influyentes y consistentes en los diferentes modelos. (vease Figura B.12)

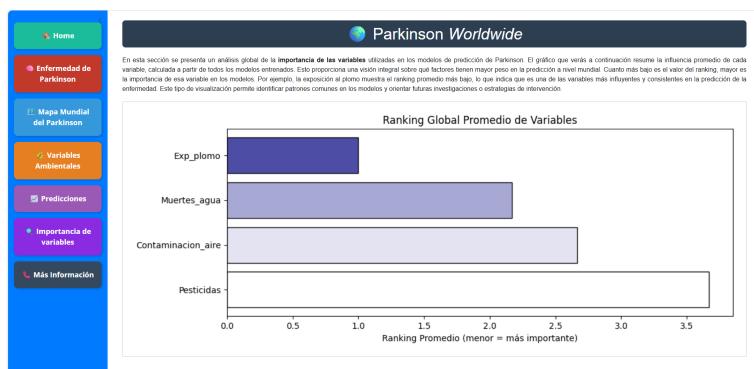


Figura B.12: Contenido de la sección de Importancia de Variables

Debajo del gráfico de ranking global, el usuario encontrará tres botones interactivos (Figura B.13) , cada uno de los cuales da acceso a secciones específicas dedicadas a los modelos de predicción utilizados.



Figura B.13: Contenido inferior de la sección de Importancia de variables

Estos botones permiten explorar en detalle cómo cada tipo de modelo ha predicho la prevalencia de la enfermedad y qué variables han considerado más relevantes.

- Modelo Lineal:** Al pulsar este botón, el usuario accede a la visualización del resultado de la predicción generada por el modelo lineal. Se muestra un mapa interactivo con la prevalencia estimada del Parkinson por país(Figura B.14), junto con el botón "Ver mapa europeo" para una visualización más detallada de esta región. También se incluye un botón de "Volver"para regresar fácilmente a la sección principal de Importancia de las Variables.(Figura ??)

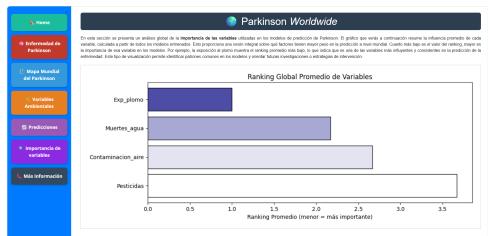


Figura B.14: Contenido botón Modelo lineal



Figura B.15: Contenido botón Modelo lineal al pulsar Ver Mapa Europeo

- Modelos basados en árboles:** Este botón agrupa los resultados de los modelos Random Forest Regressor y XGBoost Regressor, dos algoritmos de predicción basados en estructuras de árbol. Al acceder, el usuario encontrará una breve explicación comparativa entre ambos modelos, seguida de dos botones, uno para cada modelo, que permiten consultar sus resultados de forma separada. En cada uno se incluyen los mapas interactivos, controles de navegación y botones para cambiar entre la vista global y la vista europea, siguiendo la misma dinámica de otras secciones.(Figura B.16)

Ejemplo de contenido del botón Random Forest (mismo contenido para el segundo botón) puede verse en Figura G.2 Y B.18.

- Otros modelos de regresión:** Esta sección agrupa tres modelos adicionales: SVR Regressor, KNN Regressor y MLP Regressor. Al acceder, el usuario verá una breve descripción general de estos enfoques y podrá explorar los resultados de cada uno mediante tres botones individuales. Cada botón lleva a una página específica con el mapa correspondiente a ese modelo, que incluye las opciones de cambiar

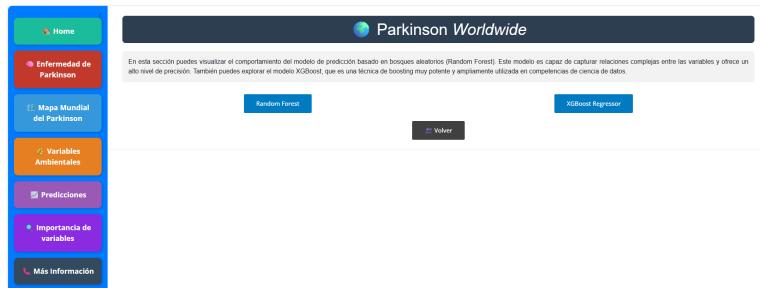


Figura B.16: Contenido botón Modelos basados en árboles



Figura B.17: Contenido botón Random Forest



Figura B.18: Contenido botón Random Forest al pulsar Ver Mapa Europeo

entre la vista global y la vista europea, siguiendo la misma dinámica que en otras secciones.(véase Figura B.19) el contenido de cada botón es similar al ya expilcado en los anteriores

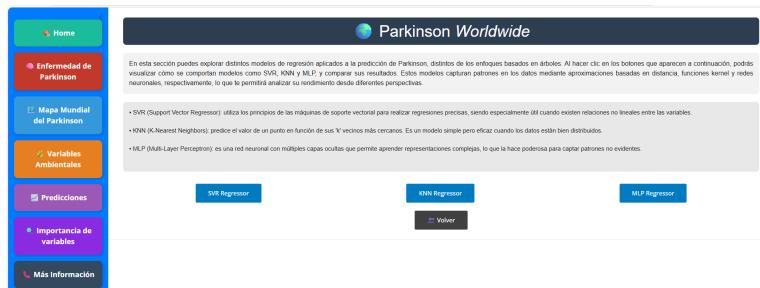


Figura B.19: Contenido botón Otros modelos de regresión

7. Más información

En esta sección, el usuario encontrará distintas formas de contacto directo con la persona responsable del proyecto. Está pensada para quienes

deseen realizar consultas, enviar sugerencias o explorar posibilidades de colaboración.

Además, se proporciona acceso a recursos externos relacionados con el desarrollo de esta aplicación, lo que permite ampliar información o conocer otros proyectos vinculados.

Apéndice C

Manual del desarrollador / programador / investigador.

Este anexo recoge la documentación técnica del proyecto, con el objetivo de facilitar su comprensión, ejecución y posible modificación por parte de otros desarrolladores o investigadores. El proyecto abarca distintas fases: obtención y tratamiento de datos, análisis exploratorio, desarrollo de modelos predictivos y creación de una aplicación web interactiva mediante Shiny para Python.

C.1. Estructura de directorios

Descripción de los directorios y ficheros entregados.

C.2. Compilación, instalación y ejecución del proyecto

En caso de ser necesaria esta sección, porque la compilación o ejecución no sea directa.

La aplicación desarrollada en este proyecto se encuentra desplegada en la web, por lo que no es necesario instalar ni compilar nada para su uso habitual.

18 Apéndice C. Manual del desarrollador / programador / investigador.

No obstante, si se desea replicar el desarrollo completo, modificar el análisis o reutilizar la aplicación con otro conjunto de datos, se deben seguir los siguientes pasos:

1. **Instalar Anaconda (incluye Python 3.12)** Se recomienda instalar Anaconda para gestionar entornos y dependencias de Python fácilmente. Puede descargarse desde: <https://www.anaconda.com/download>
2. **Crear un entorno de trabajo** (opcional, pero recomendable):

```
conda create --name tfg-env python=3.12 # (opcional especificar version)
conda activate tfg-env
```

3. Descargar el proyecto completo

Para clonar el repositorio, se puede usar el siguiente comando en la terminal o consola:

```
git clone <URL-del-repositorio>
```

El enlace al repositorio GitHub está disponible en la propia aplicación web, en la sección *Más Información*, para facilitar el acceso a quien desee continuar o modificar el proyecto.

4. Instalar las librerías necesarias

Se recomienda instalar manualmente las librerías que se usaron en el proyecto para garantizar la compatibilidad. Por ejemplo, ejecutando:

```
pip install pandas==1.5.3 numpy==2.2.6 matplotlib==3.10.1 \
seaborn==0.12.2 scikit-learn==1.6.1 plotly==6.1.1 shiny==1.4.0 \
xgboost==3.0.1 requests
```

Esta lista incluye las librerías principales para análisis, modelado y desarrollo de la app.

5. Desarrollo desde cero

Si se desea iniciar un proyecto similar desde cero, basta con crear un entorno virtual con Python y proceder a instalar las librerías mencionadas para análisis, visualización y desarrollo de la aplicación web.

6. Acceso a datos adicionales mediante API

En caso de querer ampliar el conjunto de datos con nuevas variables (por ejemplo, ambientales) o actualizar los datos automáticamente, es necesario acceder a la fuente original mediante la API proporcionada por la web de origen.

Para ello, debe seguirse el siguiente procedimiento:

- En la página original, elegir el conjunto de datos deseado y descargar el archivo `metadata.json`.
- Dentro del `metadata.json` se encuentra la URL base de la API.
- Utilizando esta URL, se puede acceder a los endpoints que devuelven los datos en formato JSON:
 - Endpoint `metadata` para obtener información sobre las variables disponibles.
 - Endpoint `data` para obtener los valores completos de los datos.
- Con estos datos se puede construir o actualizar el dataset en la aplicación.

Por ejemplo, en Python:

```
import requests

url_metadata = ''
url_data = ''

# Obtener metadata
response_meta = requests.get(url_metadata)
metadata = response_meta.json()

# Obtener datos
response_data = requests.get(url_data)
data = response_data.json()
```

C.3. Pruebas del sistema

Esta sección puede ser opcional.

Puede tratarse de validación de la interfaz por parte de los usuarios, mediante encuestas o similar o validación del funcionamiento mediante pruebas unitarias.

C.4. Pruebas del sistema

Para validar el correcto funcionamiento y la usabilidad de la aplicación, se realizaron las siguientes pruebas:

- **Pruebas de descarga de datos:** Se comprobó que los usuarios pueden filtrar los datos según diferentes criterios y descargar los datos resultantes sin errores, asegurando que los archivos exportados contienen la información correcta.
- **Pruebas de despliegue y accesibilidad:** Comprobación de acceso mediante enlace público y compatibilidad con distintos navegadores y dispositivos.

Los resultados indican que la aplicación es estable, intuitiva y cumple los objetivos planteados.

C.5. Instrucciones para la modificación o mejora del proyecto.

Instrucciones y consejos para que el trabajo pueda ser mejorado en futuras ediciones.

Para garantizar la evolución y mantenimiento del proyecto, se plantean a continuación varias recomendaciones y mejoras que pueden implementarse en futuras versiones.

- Implementar un mecanismo de respaldo para la obtención de datos: la aplicación debe intentar obtener los datos en tiempo real desde la API oficial. En caso de fallo en la conexión o en la respuesta de la API, la aplicación debe cargar automáticamente una copia local de los datos, almacenada previamente en formatos CSV o JSON dentro del proyecto. Esta solución garantiza que la aplicación siga operativa incluso sin acceso a internet o si la API está temporalmente indisponible. Además, es recomendable informar al usuario con un mensaje claro cuando se utilice esta copia local para evitar confusiones.

- Ampliar los formatos de descarga disponibles, incorporando opciones adicionales como Excel o formatos compatibles con software estadístico para mayor versatilidad.
- Incorporar funcionalidades de exportación de reportes automáticos en formatos PDF o HTML para facilitar la comunicación de resultados.
- Implementar pruebas automatizadas adicionales y monitoreo en producción para detectar y corregir rápidamente posibles fallos en el acceso a datos o en la visualización.

Apéndice D

Descripción de adquisición y tratamiento de datos

Tablas, imágenes, señales, secuencias de ADN...

D.1. Descripción formal de los datos

Los datos empleados para la elaboración del trabajo provienen de la plataforma Our World in Data (OWD). Las variables consideradas son la prevalencia de la enfermedad de Parkinson, la tasa de mortalidad por contaminación del aire Número estimado de muertes atribuidas a diferentes tipos de contaminación atmosférica, la tasa de carga de morbilidad por exposición al plomo, las muertes atribuidas a fuentes de agua insalubres, el uso de plaguicidas y la precipitación anual.

D.1.1. Prevalencia de la enfermedad del parkinson (Variable dependiente)

- **Definición y unidad de medida:** Esta variable se define como el numero estimado de personas con enfermedad del Parkinson, cuya unidad de medida se expresa por cada 100.000 habitantes.
- **Estructura de los datos:** Los datos se encuentran organizados por país y año, con un rango temporal que abarca desde 1990 hasta 2021.
- **Descripción:** es la variable dependiente en este trabajo, ya que con el estudio de esta se busca entender como factores como la contaminación,

el uso de pepticidas u otras variables pueden estar relacionadas con la prevalencia de la enfermedad del Parkinson.

D.1.2. Variables independientes

Las variables independientes son aquellas que se consideran factores que pueden influir o tener un impacto sobre la prevalencia de la enfermedad de Parkinson.

1. Tasa de mortalidad por contaminación del aire

- **Definición y unidad de medida:** Representa el numero estimado de muertes atribuidas a diferentes tipos de contaminación del aire por cada 100.000 habitantes.
- **Estructura:** Los datos están disponibles por país y año desde 1990 hasta 2021.
- **Descripción:** Esta variable mide el impacto de la contaminación del aire en la mortalidad. A través de esta variable, se puede evaluar como la exposición a ciertos contaminantes como las partículas PM2.5, podría estar relacionada con la prevalencia de la enfermedad.

2. Tasa de carga de enfermedad por exposición al plomo

- **Definición y unidad de medida:** Numero estimado de años de vida ajustados por discapacidad (AVAD) debido a la exposición al plomo, estandarizados por edad, provenientes de todas las causas, por cada 100.000 habitantes.
- **Estructura:** Los datos se encuentran organizados por pais y año, con un rango temporal que abarca desde 1990 hasta 2021.
- **Descripción:** Los años de vida ajustado por discapacidad (AVAD) miden la carga total sobre la salud de la población, considerando los años de vida perdidos por muertes prematuras y los años vividos con discapacidad. En este caso, la exposición al plomo se asocia con diversos problemas de salud que afectan a la calidad de vida y la mortalidad. La carga total se calcula sumando todos los efectos de salud relacionados con esta exposición, sin especificar las causas exactas de las muertes o discapacidades.

3. Muertes atribuidas a fuentes de agua inseguras

- **Definición y unidad de medida:** Se define como el número total de muertes causadas por fuentes de agua no seguras.
- **Estructura:** Los datos están organizados por país y año, con un rango temporal que abarca desde 1990 hasta 2021.
- **Descripción:** Esta variable mide el impacto del consumo de agua no segura en la mortalidad, sumando todas las muertes que pueden estar relacionadas con el agua insalubre, como enfermedades transmitidas por el agua o infecciones gastrointestinales. Se considera el total de muertes atribuidas a esta causa, sin especificar cada enfermedad o condición que causó la muerte.

4. Uso total de pesticidas

- **Definición y unidad de medida:** Se define como el uso total de pesticidas medido en toneladas.
- **Estructura:** Los datos se encuentran organizados por país y año, con un rango temporal que cubre desde 1990 hasta 2022.
- **Descripción:** Los pesticidas totales, incluyen los insecticidas, fungicidas y bactericidas, herbicidas, reguladores de crecimiento de las plantas, rodenticidas, desifenctantes entre otros.

5. Precipitaciones anuales

- **Definición y unidad de medida:** Se define como las precipitaciones anuales totales (lluvia y nieve), calculada como la suma de los promedios diarios y expresada como la profundidad del agua que cae a la superficie de la Tierra, excluyendo la niebla y el rocío. La variable se mide por milímetros de precipitación.
- **Estructura:** Los datos están organizados por país y por año, con un rango temporal que abarca desde 1940 hasta 2024.
- **Descripción:** Esta variable representa la cantidad total de precipitación que ocurre en un área durante un año, incluyendo tanto la lluvia como la nieve derretida. La medida se expresa en milímetros, indicando la profundidad del agua que caería sobre la superficie terrestre si se recogiera toda la precipitación. Los valores no incluyen fenómenos como la niebla o el rocío, que no aportan agua de manera significativa al suelo.

D.2. Descripción clínica de los datos.

En esta sección se presenta la perspectiva clínica de las variables consideradas para el estudio, el objetivo de esto es contextualizar de que manera estas variables pueden influir en la prevalencia de la enfermedad del Parkinson.

D.2.1. Prevalencia de la enfermedad del Parkinson

La enfermedad de Parkinson es un trastorno neurodegenerativo progresivo que afecta principalmente al sistema motor, causado por la pérdida de neuronas dopaminérgicas en la sustancia negra del cerebro. Clínicamente, se manifiesta con síntomas como temblores en reposo, rigidez muscular, bradicinesia (lentitud de movimientos) y alteraciones posturales. Su prevalencia aumenta con la edad y puede estar influenciada por factores ambientales y genéticos.[Instituto Nacional sobre el Envejecimiento (NIA), 2022] .

D.2.2. Tasa de mortalidad por contaminación del aire

La exposición prolongada a contaminantes del aire como las partículas finas ($PM_{2,5}$), dióxido de nitrógeno (NO_2) y ozono (O_3) se ha asociado con un mayor riesgo de enfermedades cardiovasculares y neurodegenerativas. Estudios recientes sugieren que la contaminación del aire puede inducir estrés oxidativo e inflamación sistémica, lo que podría contribuir a la neurodegeneración observada en enfermedades como el Parkinson.[Kerrane et al., 2015]

D.2.3. Carga de enfermedad por exposición al plomo

El plomo es un neurotóxico conocido que puede acumularse en el cerebro y alterar funciones neurológicas. En adultos, la exposición crónica al plomo ha sido relacionada con una mayor incidencia de deterioro cognitivo y enfermedades neurodegenerativas. Desde una perspectiva clínica, su asociación con el Parkinson se explica por el daño oxidativo y la disfunción mitocondrial inducida por este metal pesado.[Pyatha et al., 2022]

D.2.4. Muertes atribuidas a fuentes de agua inseguras

Aunque las enfermedades derivadas del consumo de agua contaminada no tienen una relación directa con el Parkinson en todos los casos, la exposición a ciertos contaminantes químicos presentes en el agua, como pesticidas y metales pesados, ha sido asociada con efectos neurotóxicos. Varios estudios

indican que la exposición prolongada a contaminantes del agua, como el tetracloroetileno (TCE) y otros productos químicos, puede estar relacionada con un mayor riesgo de desarrollar enfermedades neurodegenerativas, incluida la enfermedad de Parkinson.[Pacheco Moisés et al., 2011, inf, 2023, ken, 2022].

D.2.5. Uso total de pesticidas

El uso de pesticidas, especialmente herbicidas como el paraquat y fungicidas como el maneb, ha sido consistentemente asociado con un mayor riesgo de desarrollar la enfermedad de Parkinson. Estos compuestos pueden inducir estrés oxidativo y afectar la función mitocondrial, contribuyendo al daño neuronal característico de la enfermedad. Varios estudios han encontrado que la exposición prolongada a estos pesticidas aumenta significativamente el riesgo de desarrollar Parkinson, particularmente en áreas agrícolas donde su uso es elevado.[Pearce et al., 2013, Tanner et al., 2011, Starks et al., 2013]

D.2.6. Precipitaciones anuales

Aunque las precipitaciones no influyen directamente en la salud humana, pueden actuar como moduladores del entorno, afectando la dispersión de contaminantes o el uso agrícola de pesticidas. Desde un punto de vista clínico, su relevancia radica en su potencial para modificar la exposición a factores ambientales vinculados con la neurotoxicidad.[America, 2023]

Apéndice E

Manual de especificación de diseño

Este anexo recoge los aspectos clave relacionados con el diseño estructural y funcional de la aplicación desarrollada. El objetivo es proporcionar una visión clara y detallada de cómo se organiza internamente el sistema, tanto a nivel de componentes lógicos como de su interacción y despliegue.

Se incluyen diversos diagramas elaborados con el lenguaje de modelado *PlantUML*[Roques, 2025], los cuales permiten representar gráficamente el flujo de trabajo del sistema, la organización por paquetes funcionales, y la arquitectura de despliegue.

E.1. Diseño arquitectónico

E.1.1. Diagrama de Despliegue

En esta sección se presenta el diseño arquitectónico de la aplicación desarrollada. Dado que el proyecto está implementado en **Python** utilizando el *framework Shiny*, y no sigue una estructura orientada a objetos tradicional, no se incluyen diagramas de clases. En su lugar, se proporciona un **diagrama de despliegue** (Figura E.1), que representa cómo se estructura e interconecta el sistema en tiempo de ejecución.

La aplicación está desplegada en la plataforma *Posit Cloud*, lo que permite el acceso de los usuarios a través de un navegador web y una *URL* pública. Los datos utilizados en la visualización y análisis se obtienen en

tiempo real mediante peticiones a la **API pública de *Our World In Data* (OWID)**.

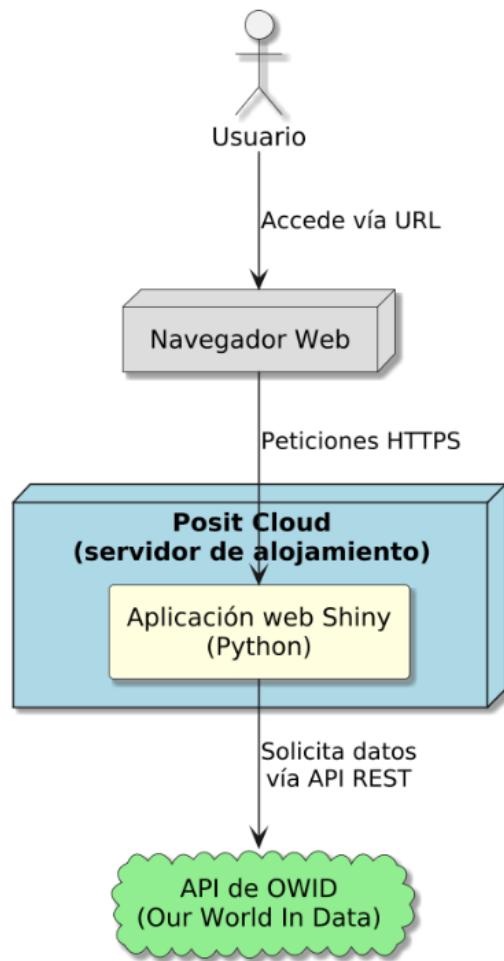


Figura E.1: Diagrama de despliegue de la aplicación web.

E.1.2. Diagrama de Flujo

En esta sección se presenta el diagrama de flujo([E.2](#)) que ilustra el proceso general seguido para la preparación de datos, entrenamiento de modelos y predicción de resultados. Este diagrama facilita la comprensión visual de las etapas principales del flujo de trabajo implementado en la aplicación, permitiendo identificar de forma clara las decisiones clave y las tareas realizadas en cada fase.

El diagrama refleja la estructura lógica del sistema, desde la adquisición y procesamiento inicial de datos, hasta las fases de entrenamiento y predicción, mostrando las posibles rutas según las condiciones definidas en el proceso. Esta representación contribuye a una mejor comunicación técnica del funcionamiento interno del proyecto y sirve como referencia para futuros desarrollos o mantenimiento.

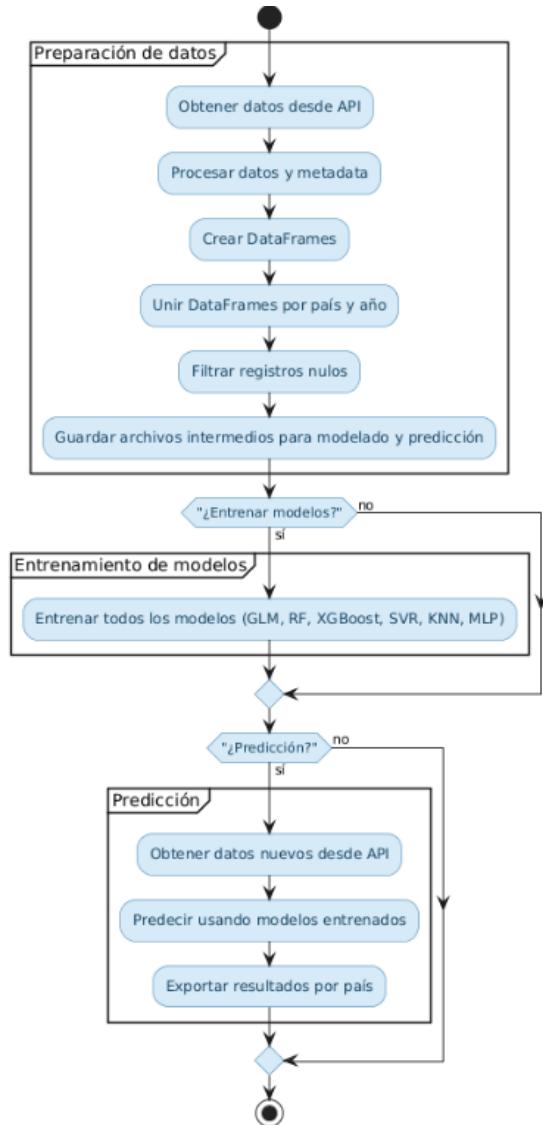


Figura E.2: Diagrama de flujo del proceso de modelado y predicción.

E.1.3. Diagrama de Paquetes

En esta sección se presenta el diagrama de paquetes E.3 que representa la organización modular del sistema desarrollado. Este diagrama muestra la división del proyecto en tres módulos principales: Preparación de Datos, Entrenamiento de Modelos y Predicción. Cada paquete agrupa las responsabilidades y componentes relacionados, facilitando la comprensión de la estructura general y las dependencias entre los distintos subsistemas.

El diagrama sirve como referencia para visualizar cómo se agrupan y conectan los diferentes procesos y componentes, contribuyendo a una mejor organización del código y simplificando futuras tareas de mantenimiento o ampliación del sistema.

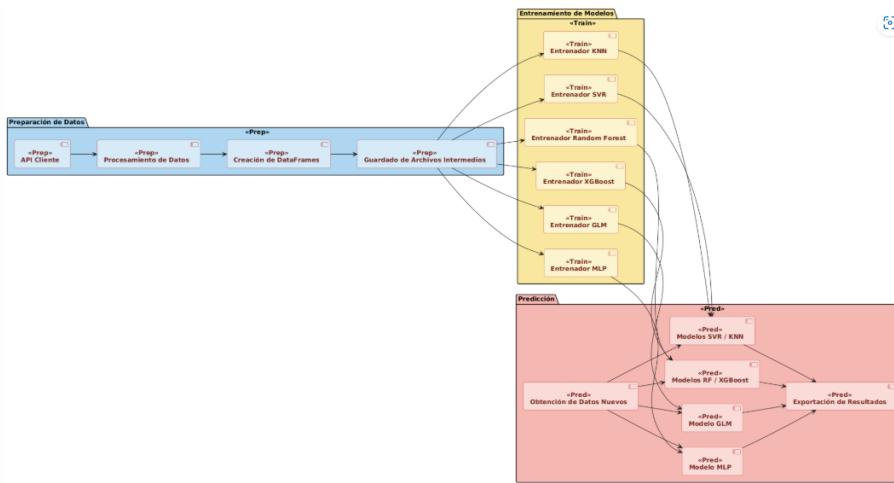


Figura E.3: Diagrama de paquetes del sistema.

Apéndice F

Especificación de Requisitos

F.1. Diagrama de casos de uso

En la Figura F.1 se muestra el diagrama de casos de uso que describe la interacción del usuario con la aplicación.



Figura F.1: Diagrama de casos de uso de la aplicación.

F.2. Explicación casos de uso.

El diagrama de casos de uso representa las diferentes formas en que el usuario puede interactuar con la aplicación. A continuación, se describe cada funcionalidad principal y las interacciones asociadas:

- **Acceder a la aplicación:** El usuario inicia la interacción accediendo a la aplicación, desde donde puede elegir distintas secciones para visualizar.
- **Home:** Al seleccionar la sección Home, el usuario puede acceder a un enlace que lo dirige a la página web original de donde se han obtenido los datos utilizados en la aplicación.
- **Enfermedad de Parkinson:** Esta sección permite al usuario obtener información general sobre la enfermedad.
- **Mapa Mundial del Parkinson:** En esta sección, el usuario puede interactuar con un mapa que muestra la prevalencia de la enfermedad a nivel mundial. Además, puede filtrar los datos por año y país, descargar los datos completos, y acceder a un enlace directo con la fuente de los datos. Adicionalmente, existe la opción de visualizar un mapa europeo, que es una extensión de esta funcionalidad.
- **Variables Ambientales:** El usuario puede explorar diferentes variables ambientales relacionadas con la enfermedad. Al seleccionar cada variable, se ofrecen funcionalidades similares a las del mapa mundial, como interacción con mapas, filtrado de datos, descarga y acceso a enlaces directos. También es posible visualizar un mapa europeo para estas variables.
- **Predicciones:** Esta sección permite al usuario interactuar con los mapas disponibles para observar predicciones relacionadas con la enfermedad. También incluye la opción de ver el mapa europeo.
- **Importancia de Variables:** Aquí, el usuario puede acceder a los modelos utilizados para el análisis y elaboración del trabajo, profundizando en el impacto de las diferentes variables.
- **Más Información:** Finalmente, el usuario puede acceder a un enlace directo al repositorio GitHub donde se encuentra el código y documentación relacionada con la aplicación.

Estas funcionalidades reflejan la interacción completa entre el usuario y la aplicación, mostrando tanto las opciones básicas como las extensiones opcionales que enriquecen la experiencia.

F.3. Prototipos de interfaz o interacción con el proyecto

En este proyecto no se han desarrollado prototipos visuales de la interfaz o de la interacción, dado que la implementación se ha realizado directamente en código funcional.

La estructura y diseño de la aplicación se basa en la programación y la definición directa de las funcionalidades, sin pasar por una fase previa de prototipado gráfico.

CU-1	Ejemplo de caso de uso
Versión	1.0
Autor	Alumno
Requisitos asociados	RF-xx, RF-xx
Descripción	La descripción del CU
Precondición	Precondiciones (podría haber más de una)
Acciones	<ul style="list-style-type: none"> 1. Pasos del CU 2. Pasos del CU (añadir tantos como sean necesarios)
Postcondición	Postcondiciones (podría haber más de una)
Excepciones	Excepciones
Importancia	Alta o Media o Baja...

Tabla F.1: CU-1 Nombre del caso de uso.

Apéndice G

Estudio experimental

G.1. Cuaderno de trabajo.

Enumeración de todos los métodos probados con resultados positivos o no.

Con el fin de evaluar la relación entre las variables ambientales y la prevalencia del Parkinson a escala mundial, se ha llevado a cabo un análisis exploratorio preliminar. Esta etapa tuvo como objetivo detectar posibles problemas de multicolinealidad entre las variables independientes y orientar adecuadamente la fase de modelado.

Para ello, se construyó una matriz de correlación (véase Figura G.1), a partir de la cual se observó que ninguna de las variables presentaba coeficientes de correlación superiores a 0.7. Esto permitió descartar redundancia estadística y justificar el uso conjunto de todas ellas en los modelos predictivos.

Posteriormente, se estudiaron las relaciones individuales entre cada variable independiente y la variable objetivo (casos de Parkinson). Para ello, se realizaron gráficos de dispersión con líneas de tendencia ajustadas, lo que permitió observar que algunas relaciones eran no lineales.

La Figura G.2 muestra la relación entre Muertes por agua contaminada y los casos de Parkinson, donde se puede observar una curvatura inicial que luego se estabiliza.

De manera similar, las variables Exposición al plomo y Contaminación del aire también mostraron un comportamiento curvado. (ver Figura G.3 y Figura G.4).

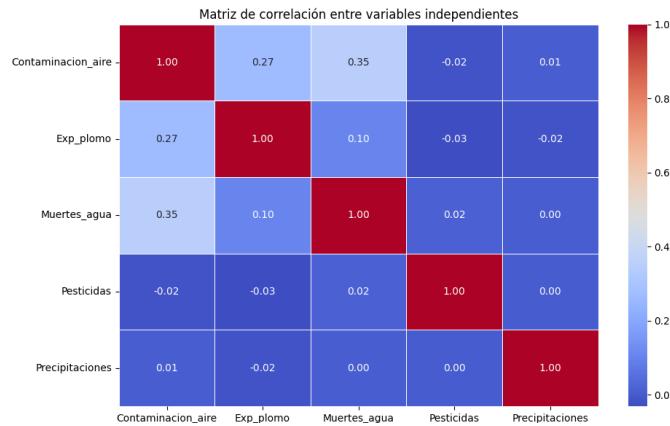


Figura G.1: Matriz de correlación entre variables independientes.

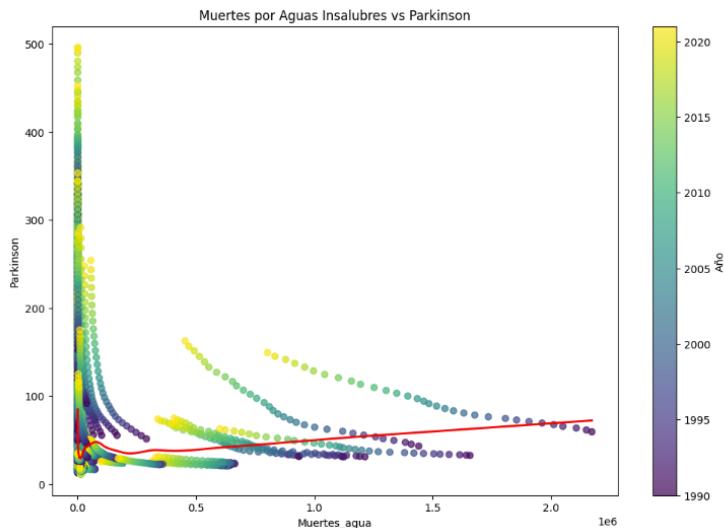


Figura G.2: Relación entre el Parkinson y las Muertes atribuidas a fuentes de agua inseguras.

Por otro lado, la variable Pesticidas mostró un patrón logarítmico con picos, lo que sugiere que enfoques de modelado que contemplen relaciones no lineales o funciones logarítmicas podrían resultar más apropiados para capturar su comportamiento. La Figura G.5 ilustra este comportamiento logarítmico.

En cuanto a Precipitaciones, la relación con los casos de Parkinson fue lineal, lo que sugiere que modelos lineales podrían ser adecuados para este predictor, como se muestra en Figura G.6.

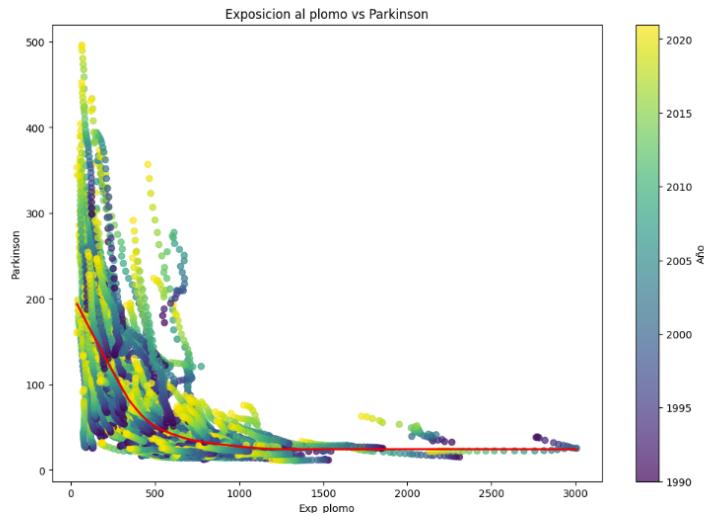


Figura G.3: Relación entre el Parkinson y la Tasa de carga de enfermedad por exposición al plomo.

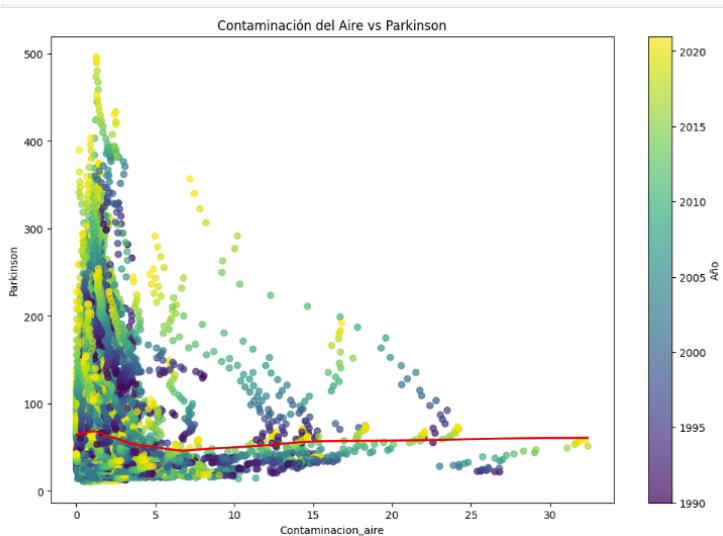


Figura G.4: Relación entre el Parkinson y la Tasa de mortalidad por contaminación de aire.

Este análisis preliminar no implicó la transformación directa de las variables, sino que se enfocó en identificar la naturaleza de sus relaciones con el objetivo de orientar la elección y formulación de modelos adecuados en etapas posteriores. De este modo, los modelos se adaptarán a los datos y

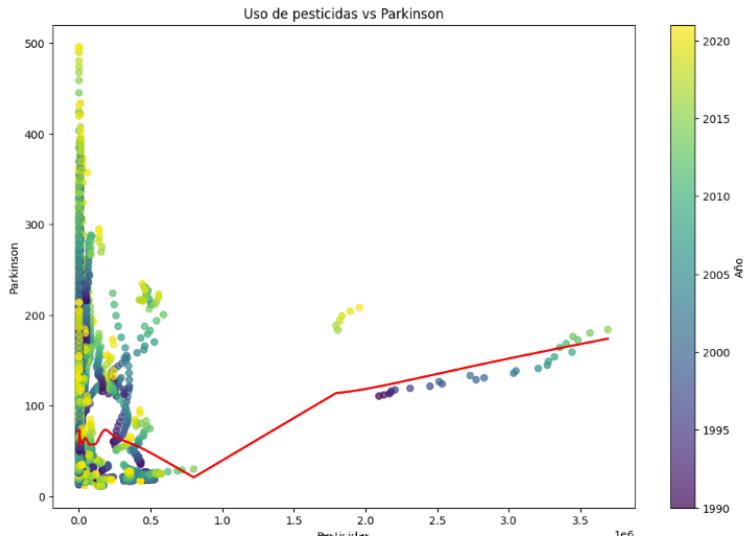


Figura G.5: Relación entre el Parkinson y el uso de pesticidas.

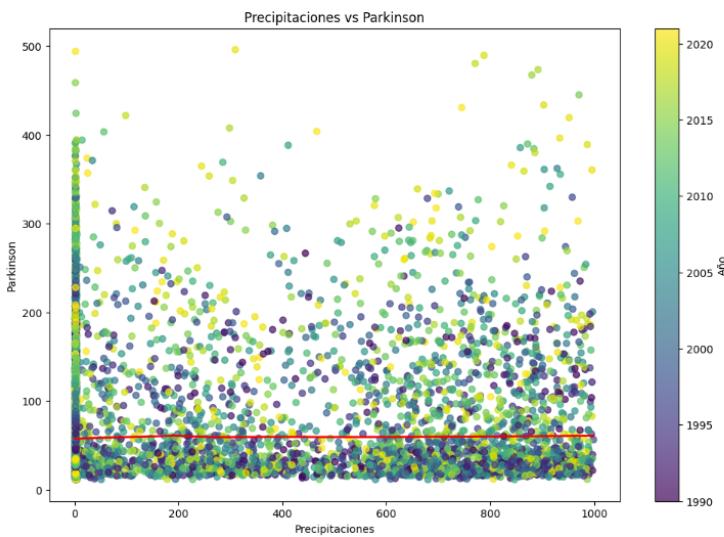


Figura G.6: Relación entre el Parkinson y el uso de pesticidas.

no al revés, respetando sus patrones inherentes y optimizando la capacidad predictiva de cada enfoque.

G.1.1. Selección de modelos

Una vez realizado el análisis preliminar de los datos, se estudian que modelos son los más afines a utilizar para la predicción según las características

que presentan los datos. El objetivo es aplicar una combinación de modelos de diferentes familias ya que ningún modelo por sí solo es capaz de capturar completamente la complejidad de las relaciones presentes en los datos. Cada tipo de modelo tiene características y capacidades particulares que lo hacen más adecuado para ciertos patrones de datos.

Dado que la variable objetivo en este estudio es el número estimado de casos de Parkinson, es decir, un conteo que representa el número de casos en diferentes países, se requiere un enfoque que se adapte específicamente a variables de recuento. Los modelos seleccionados para esta tarea son aquellos que son capaces de manejar correctamente datos con características no lineales, distribuciones sesgadas o complejas y relaciones no evidentes entre las variables predictoras. A continuación, se explica por qué se eligieron los siguientes modelos:

1. Generalized Linear Model (GLM) con distribución Binomial Negativa

- **Motivo y aplicación:** La familia GLM es una opción adecuada para modelar variables de recuento, ya que permite ajustar la distribución de la variable objetivo según la naturaleza de los datos. En este caso, se seleccionó la distribución Binomial Negativa en lugar de Poisson, ya que la varianza de los recuentos era mayor que la media, y este modelo supone que ambas son iguales.

A pesar de probar con el modelo Cuasi-Poisson (Figura G.7), los resultados obtenidos fueron insatisfactorios, con un Pseudo R-squared de 1, lo que indicaba un sobreajuste de los datos. Al comparar la verosimilitud entre ambos modelos y al obtener resultados similares, el modelo seleccionado fue el modelo Binomial Negativo (Figura G.8).

- **Estructura del modelo:** El modelo GLM se rige por la siguiente fórmula:

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

donde $\mathbb{E}[Y]$ representa el numero estimado de casos de Parkinson y X_i son las variables ambientales consideradas. Esta forma permite capturar efectos multiplicativos, lo que resulta adecuado para datos de recuento.

- Adaptación del modelo a los datos:** A partir de los resultados del análisis exploratorio preliminar, se observó que algunas variables presentaban relaciones no lineales con la variable objetivo. Para reflejar estas dinámicas sin transformar directamente los datos, se adaptó la fórmula del modelo GLM incorporando diferentes formas funcionales según la naturaleza observada de cada variable. En particular, para las variables que mostraban curvatura inicial, se evaluaron términos polinómicos hasta grado 3 (X, X^2, X^3), no se utilizaron polinomios de grado superior a 3 porque pueden generar sobreajuste. Además, dado que la variable objetivo representa un número de casos de una enfermedad, no tiene sentido aplicar funciones más complejas que podrían dar lugar a predicciones poco realistas.

El uso de términos hasta grado 3 permite capturar la curvatura observada sin perder sentido práctico ni interpretabilidad. En el caso de *Pesticidas*, se incorporó el término $\log(1 + X)$ al reflejar un comportamiento logarítmico, mientras que *Precipitaciones* se mantuvo en forma lineal, al no requerir ajustes adicionales.

La inclusión final de cada término se basó exclusivamente en su significancia estadística, conservando solo aquellos que aportaban valor explicativo real al modelo. Este enfoque permitió ajustar la forma funcional del modelo a la naturaleza de los datos, mejorando su capacidad predictiva y manteniendo la estructura propia del GLM.

Modelo Binomial Negativo: Generalized Linear Model Regression Results							
<hr/>							
Dep. Variable:	Parkinson	No. Observations:	4323	Parkinson	No. Observations:	4323	
Model:	GLM	Df Residuals:	4313	Model:	GLM	Df Residuals:	4313
Model Family:	Poisson	Df Model:	9	Model Family:	NegativeBinomial	Df Model:	9
Link Function:	log	Scale:	1.0000	Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-57759.	Method:	IRLS	Log-Likelihood:	-22292.
Date:	Thu, 10 Apr 2025	Deviance:	98097.	Date:	Thu, 10 Apr 2025	Deviance:	1150.3
Time:	18:05:00	Pearson chi2:	1.06e+05	Time:	18:05:01	Pearson chi2:	1.56e+03
No. Iterations:	6	Pseudo R-squ. (C5):	1.000	No. Iterations:	13	Pseudo R-squ. (C5):	0.4104
Covariance Type:	nonrobust			Covariance Type:	nonrobust		
<hr/>							
const	4.1322	0.003	1913.827	0.000	4.128	4.138	
Contaminacion_aire	0.2118	0.005	44.254	0.000	0.202	0.221	
Exp_plomo	-1.5512	0.005	-310.105	0.000	-1.561	-1.541	
Muertes_agua	-0.3351	0.009	-36.244	0.000	-0.353	-0.317	
Pesticidas	-0.0024	0.002	-3.703	0.000	-0.009	-0.008	
Precipitaciones	0.0024	0.002	1.248	0.000	0.001	0.000	
Contaminacion_aire_2	-0.0715	0.005	-15.363	0.000	-0.081	-0.062	
Muertes_agua_2	0.2835	0.008	34.030	0.000	0.267	0.299	
Exp_plomo_2	0.9238	0.005	175.366	0.000	0.913	0.934	
Pesticidas_log	0.0535	0.002	29.559	0.000	0.050	0.057	
<hr/>							
Error Cuadrático Medio (RMSE):	47.4063187307246			const	4.1466	0.015	
Error Absoluto Medio (MAE):	31.852376321380544			Contaminacion_aire	0.1897	0.039	
				Exp_plomo	-1.4937	0.043	
				Muertes_agua	-0.2855	0.050	
				Pesticidas	-0.0057	0.017	
				Precipitaciones	0.0027	0.015	
				Contaminacion_aire_2	0.0777	0.037	
				Muertes_agua_2	0.2669	0.049	
				Exp_plomo_2	0.0005	0.043	
				Pesticidas_log	0.0653	0.017	
				Error Cuadrático Medio (RMSE):	47.43369176286732		
				Error Absoluto Medio (MAE):	32.00714819647614		

Figura G.7: Modelo Cuassi-poisson.

Figura G.8: Modelo-Binomial Negativo

2. Random Forest

- **Motivo y aplicación:** El modelo Random Forest es una técnica basada en la combinación de múltiples árboles de decisión, lo que le permite capturar relaciones complejas y no lineales entre las variables. Se adapta bien a datos con ruido y a relaciones no evidentes, lo que lo convierte en una buena opción para el tipo de datos utilizados en este estudio.

Se empleó el algoritmo RandomForestRegressor dado que la variable objetivo es numérica (número estimado de casos de Parkinson). Este modelo permite obtener predicciones precisas sin necesidad de asumir una forma funcional específica entre las variables independientes y la variable objetivo.

3. XGBoost

- **Motivo y aplicación:** XGBoost es un modelo de boosting basado en árboles que destaca por su alta precisión y capacidad para capturar relaciones no lineales y complejas entre variables. Dado que el análisis exploratorio mostró patrones no lineales en las relaciones entre las variables ambientales y los casos de Parkinson, XGBoost resultó adecuado para modelar este tipo de datos.

Aunque es ampliamente utilizado en tareas de clasificación, XGBoost también dispone de una versión para regresión (XGBRegressor), que fue la empleada en este trabajo, ya que la variable objetivo (número estimado de casos de Parkinson) es de tipo continuo y de recuento. Este modelo es eficaz incluso en presencia de ruido y relaciones complejas difíciles de capturar por modelos lineales.

4. Support Vector Regression (SVR)

- **Motivo y aplicación:** El modelo SVR es adecuado para capturar relaciones complejas entre variables, incluso cuando estas no siguen patrones lineales. Esto se logra gracias al uso de funciones núcleo (kernel), que permiten proyectar los datos a espacios de mayor dimensión, donde las relaciones no lineales pueden ser modeladas mediante una función lineal en ese nuevo espacio.

En este trabajo se utilizó el kernel radial (RBF), que es especialmente útil para detectar patrones no lineales suaves. A diferencia de modelos basados en transformaciones explícitas de las variables, como el GLM, el SVR incorpora la no linealidad de forma implícita a través del kernel.

Aunque el SVR se emplea comúnmente en problemas de regresión continua, puede aplicarse también en contextos de datos de recuento si la escala y la naturaleza de la variable objetivo lo permiten. En este caso, se modeló el número estimado de casos de Parkinson con buenos resultados en cuanto a precisión y generalización, empleando la implementación del modelo disponible en `scikit-learn`.

5. K-Nearest Neighbors Regression (KNN):

- **Motivo y aplicación:** El modelo KNN es un algoritmo basado en instancias que realiza predicciones en función de la similitud entre observaciones. Su principal ventaja es que no requiere asumir una forma funcional específica entre las variables predictoras y la variable objetivo, lo que lo hace especialmente útil para modelar patrones locales o relaciones complejas que varían en distintas regiones del espacio de características.

Para este estudio, se utilizó la implementación `KNeighborsRegressor` de `scikit-learn`. Debido a su sensibilidad a la escala y a la dispersión de los datos, las variables fueron estandarizadas previamente. Aunque KNN no realiza ninguna inferencia paramétrica ni captura directamente relaciones no lineales generales, sí puede adaptarse bien a estructuras no lineales locales presentes en los datos.

Este modelo resultó útil para capturar tendencias locales en el número estimado de casos de Parkinson, particularmente en combinaciones de variables donde se observaban patrones heterogéneos o no globalmente lineales.

6. Multi-Layer Perceptron (MLP):

- **Motivo y aplicación:** El MLP es una red neuronal de tipo *feedforward* que permite modelar relaciones altamente complejas y no lineales entre las variables. Gracias a su arquitectura basada en capas ocultas y funciones de activación no lineales, posee una gran capacidad de aprendizaje y es especialmente útil cuando los patrones subyacentes no pueden ser capturados adecuadamente por modelos más simples.

Se utilizó la implementación `MLPRegressor` de `scikit-learn`. Aunque el MLP puede aprender transformaciones internas complejas, se realizó una estandarización previa de las variables

predictoras para mejorar la estabilidad numérica y acelerar la convergencia del modelo durante el entrenamiento.

Este modelo fue capaz de capturar interacciones no evidentes entre variables y mostró un buen rendimiento predictivo. No obstante, su principal desventaja radica en la menor interpretabilidad en comparación con modelos lineales o basados en reglas.

G.2. Configuración y parametrización de las técnicas.

Con el fin de garantizar un ajuste óptimo de los modelos predictivos empleados en el estudio, se llevó a cabo un proceso de configuración y ajuste de hiperparámetros específico para cada técnica. Esta etapa es fundamental, ya que permite optimizar el rendimiento de cada modelo en función de las características de los datos.

G.2.1. Generalized Linear Model (GLM - Binomial Negativo)

Para el ajuste del GLM, se utilizó la librería statsmodels, empleando la familia Binomial Negativa (NegativeBinomial), adecuada para variables de recuento con sobredispersión, es decir, cuando la varianza excede a la media.

La tabla G.1 resume la configuración y parametrización aplicada al modelo Generalized Linear Model (GLM) utilizando la distribución Binomial Negativa.

Tabla G.1: Configuración aplicada al modelo GLM (Binomial Negativa)

Aspecto	Descripción
Distribución elegida	Binomial Negativa
Función de enlace	Logarítmica
Estandarización	Aplicada
División de datos	80 % entrenamiento / 20 % prueba
Evaluación	RMSE, MAE

G.2.2. Modelo Random Forest

A continuación se presenta la tabla G.2 con la configuración y los parámetros aplicados al modelo Random Forest utilizado en este análisis. Los

parámetros descritos incluyen configuraciones clave como el número de árboles, la profundidad de los árboles, y otras opciones relacionadas con el proceso de entrenamiento y la división de los datos. Estos parámetros fueron seleccionados con el objetivo de optimizar el rendimiento del modelo, asegurando que se capture la complejidad de los datos sin alcanzar el sobreajuste.

Tabla G.2: Configuración aplicada al modelo Random Forest

Aspecto	Descripción
Número de estimadores	1000
Profundidad máxima	Ningún límite (None)
Mínimo de muestras para dividir	2
Mínimo de muestras por hoja	1
Características por división	sqrt
Valor de semilla	42
Estandarización	No aplicada (Random Forest no lo requiere)
División de datos	80 % entrenamiento / 20 % prueba
Evaluación	RMSE, MAE, R ²

Descripción de los campos:

- **Número de estimadores (n_estimators):** Este parámetro determina cuántos árboles se van a construir en el modelo. Un número mayor de árboles aumenta la precisión y la estabilidad del modelo, ya que cada árbol contribuye a reducir el error total. Sin embargo, un número muy alto también puede aumentar el tiempo de entrenamiento.
- **Profundidad máxima (max_depth):** Define la profundidad máxima de cada árbol. Es decir, cuántos niveles puede tener el árbol desde la raíz hasta las hojas. Una profundidad mayor permite capturar relaciones más complejas entre las características, pero si es demasiado alta, puede llevar a un sobreajuste (overfitting).
- **Mínimo de muestras para dividir (min_samples_split):** Este parámetro especifica el número mínimo de muestras requeridas para

dividir un nodo. Si se establece un valor alto, el modelo se vuelve más conservador y evita crear divisiones en nodos con pocos datos. Esto puede ayudar a reducir el sobreajuste, aunque a su vez limita la capacidad del modelo para aprender patrones más complejos.

- **Mínimo de muestras por hoja (min_samples_leaf):** Controla el número mínimo de muestras que debe haber en un nodo hoja. Este parámetro es importante porque asegura que las hojas del árbol contengan una cantidad significativa de datos, lo que ayuda a evitar que el modelo aprenda demasiado de los ruidos o las fluctuaciones pequeñas de los datos.
- **Características por división (max_features):** Este parámetro controla cuántas características se consideran para la división en cada nodo. Si se utiliza una fracción más pequeña de las características, se introduce mayor aleatoriedad, lo que puede ayudar a reducir el sobreajuste y hacer el modelo más robusto. Usar todas las características puede llevar a un modelo más específico para los datos de entrenamiento, pero puede ser propenso a sobreajustarse.
- **Valor de semilla (random_state):** Se utiliza para fijar la aleatoriedad del modelo, garantizando que los resultados sean reproducibles. Si no se establece un valor de semilla, los resultados pueden variar en cada ejecución debido a la selección aleatoria de muestras y características.
- **Estandarización:** No es necesaria, ya que Random Forest no depende de las escalas de las variables.
- **División de datos:** Se utiliza un 80 % de los datos para entrenamiento y un 20 % para prueba.
- **Evaluación:** Se emplean métricas como RMSE, MAE y R² para evaluar el rendimiento del modelo.

G.2.3. Modelo XGBoost

En la Tabla G.3 se detallan los principales hiperparámetros utilizados para entrenar el modelo XGBoost. Estos valores fueron seleccionados con el objetivo de optimizar el rendimiento del modelo y evitar el sobreajuste. Cabe destacar que la división de los datos se realizó en un 80 % para entrenamiento y un 20 % para prueba, y que el subsampling corresponde a una técnica interna de XGBoost que selecciona aleatoriamente una fracción del conjunto de entrenamiento en cada iteración para mejorar la generalización.

Tabla G.3: Configuración aplicada al modelo XGBoost

Aspecto	Descripción
Número de estimadores	1000
Tasa de aprendizaje	0.05
Profundidad máxima	7
Peso mínimo por hoja	5
Submuestreo	80 % de los datos por árbol
Proporción de características por árbol	100 % (colsample_bytree = 1.0)
Valor de semilla	42
Estandarización	No aplicada (escalado interno)
División de datos	80 % entrenamiento / 20 % prueba
Evaluación	RMSE, MAE, R ²

Descripción de los campos:

- **Tasa de aprendizaje (learning_rate):** La tasa de aprendizaje controla cuánto cambia el modelo con cada árbol. Un valor bajo (como 0.05) hace que el modelo aprenda más lentamente, lo que ayuda a evitar el sobreajuste y mejora la generalización. Sin embargo, valores bajos también requieren más árboles para alcanzar un buen rendimiento.
- **Peso mínimo por hoja (min_child_weight):** Indica el número mínimo de instancias que deben estar en una hoja del árbol. Un valor de 5 asegura que un nodo hoja contenga una cantidad significativa de datos, lo que previene que el modelo se ajuste a pequeñas fluctuaciones o ruidos en los datos. Si se establece demasiado bajo, el modelo podría aprender patrones no representativos.
- **Submuestreo (subsample):** Especifica el porcentaje de datos que se utilizarán para entrenar cada árbol. Con un valor del 80 %, solo una parte del conjunto de entrenamiento se usa para cada árbol. Este submuestreo introduce variabilidad en el modelo y ayuda a prevenir el sobreajuste, ya que no todos los datos se usan en cada árbol.
- **Proporción de características por árbol (colsample_bytree):** Controla la fracción de las características que se usan para entrenar cada árbol.

Con un valor del 100% , el modelo utiliza todas las características disponibles en cada árbol. Si se reduce este valor, se puede introducir más aleatoriedad y reducir el riesgo de sobreajuste.

- El número de estimadores, la profundidad máxima, el valor de semilla, el mínimo de muestras para dividir, la estandarización, la división de datos y la evaluación tienen la misma definición que la explicada en el modelo Random Forest.

G.2.4. Modelo SVR

El modelo Support Vector Regression (SVR) requiere la configuración de varios hiperparámetros clave que impactan su rendimiento. Para optimizarlos, se utilizó GridSearchCV con validación cruzada de 5 particiones ($cv=5$), lo que mejora la estimación del rendimiento y ayuda a evitar el sobreajuste. La métrica de optimización empleada fue el error cuadrático medio negativo (`neg_mean_squared_error`), ya que esta penaliza los errores grandes, lo que favorece un modelo preciso.

Los parámetros escogidos fueron el resultado de aplicar transformaciones.(Véase La tabla G.4)

Tabla G.4: Parámetros utilizados en el modelo SVR con variables transformadas

Parámetro	Valor
C	1000
ϵ	1
γ	1
<code>kernel</code>	<code>rbf</code>

A continuación, se describen los efectos generales de cada parámetro y cómo influyen en el comportamiento del modelo:

- **C (parámetro de regularización):** Este parámetro controla el equilibrio entre la maximización del margen y la minimización de los errores de predicción. Un valor alto de C penaliza fuertemente los errores, lo que permite al modelo ajustarse bien a los datos de entrenamiento. Sin embargo, un valor muy alto puede llevar al sobreajuste, ya que el modelo se adapta demasiado a los detalles específicos del conjunto de entrenamiento.

- **Epsilon:** Este parámetro define un margen de tolerancia dentro del cual los puntos de datos no afectan el modelo. Un valor pequeño de `epsilon` indica que el modelo tratará de minimizar todos los errores, incluso los pequeños, lo que puede hacer que el modelo sea más sensible y propenso al sobreajuste. Un valor mayor proporciona mayor margen de error y, por tanto, un modelo menos sensible a fluctuaciones pequeñas.
- **Gamma:** Este parámetro controla la influencia de cada punto de datos en el modelo. Un valor bajo de `gamma` implica que los puntos de datos lejanos tienen mayor influencia, mientras que un valor alto hace que solo los puntos cercanos tengan impacto, lo que puede permitir capturar relaciones complejas, pero también puede aumentar el riesgo de sobreajuste si el modelo se ajusta demasiado a las pequeñas variaciones en los datos.
- **Kernel:** El kernel define la función que transforma los datos en un espacio de mayor dimensión, lo que permite al modelo encontrar patrones no lineales. El kernel `rbf` (Radial Basis Function) es una opción común debido a su capacidad para modelar relaciones complejas entre las variables. Este kernel es especialmente útil cuando las relaciones en los datos no son lineales y requiere que el modelo aprenda patrones de forma flexible.

G.2.5. Modelo K-Nearest Neighbors (KNN)

El modelo K-Nearest Neighbors (KNN) fue optimizado mediante una búsqueda de hiperparámetros utilizando lo mismo mencionado en el modelo SVR.

Estas combinaciones de hiperparámetros fueron seleccionadas para obtener el mejor rendimiento del modelo, y los valores de **MAE** y **RMSE** se utilizaron como criterios de evaluación para la calidad de las predicciones realizadas por el modelo.

Finalmente los hiperparámetros escogidos para el modelo fueron que se encuentran en la [G.5](#)

Los siguientes parámetros son esenciales para el funcionamiento del modelo KNN y afectan su rendimiento al determinar cómo se calculan las predicciones. A continuación, se describen los efectos generales de cada parámetro:

Tabla G.5: Parámetros utilizados en el entrenamiento del modelo KNN

Parámetro	Valor
<i>n_neighbors</i>	15
<i>weights</i>	'distance'
<i>algorithm</i>	'auto'
<i>metric</i>	'manhattan'

- ***n_neighbors***: Este parámetro especifica el número de vecinos más cercanos que se considerarán al hacer una predicción. En este caso, el modelo utilizará 3 vecinos. Un valor más bajo de ***n_neighbors*** puede hacer que el modelo sea más sensible al ruido, mientras que un valor más alto puede hacer que el modelo sea más suave y menos sensible a los detalles locales de los datos.
- ***weights***: El parámetro ***weights*** define la forma en que se ponderan los vecinos en la predicción. El valor '***distance***' significa que los vecinos más cercanos tendrán más peso en la predicción, lo que puede mejorar la precisión en áreas donde los puntos de datos son más densos.
- ***algorithm***: Este parámetro especifica el algoritmo utilizado para calcular los vecinos más cercanos. El valor '***auto***' permite que el modelo elija automáticamente el algoritmo más adecuado según el número de muestras y las características del conjunto de datos. Los algoritmos disponibles son '***ball_tree***', '***kd_tree***', '***brute***', y '***auto***'.
- ***metric***: El parámetro ***metric*** define la métrica de distancia utilizada para calcular la proximidad entre los puntos de datos. En este caso, se utiliza '***manhattan***', que mide la distancia entre puntos sumando las diferencias absolutas de sus coordenadas. Esta métrica es útil cuando los datos tienen características discretas o si los patrones de distancia tienen una forma lineal.

G.2.6. Modelo Perceptrón Multicapa (MLPRegressor)

El modelo **Perceptrón Multicapa** (MLPRegressor) fue optimizado mediante una búsqueda de hiperparámetros al igual que en los modelos anteriores.

Aunque la combinación los hiperparámetros resultantes proporcionó los mejores resultados, el **MSE** seguía siendo relativamente alto, lo que

indicaba que, a pesar de los esfuerzos de optimización, el modelo no logró una predicción precisa. En este punto, se exploraron alternativas para la búsqueda de otros hiperparámetros que minimizaran el MSE (*RandomizedSearchCV*).

Debido al tiempo de cómputo necesario para la obtención de hiperparámetros que produjeran mejores resultados se optó por mantener los mejores hiperparámetros obtenidos de la combinación de variables y ajustar manualmente algunos parámetros para observar su impacto en la reducción del error. Este proceso se repitió hasta minimizar el máximo posible el error.

Al final, los parámetros finales seleccionados para el modelo fueron los que proporcionaron el **mejor rendimiento** sin necesidad de un proceso de búsqueda exhaustiva debido al tiempo de cómputo elevado. (Tabla G.6)

Tabla G.6: Parámetros utilizados en el entrenamiento del modelo MLP

Parámetro	Valor
<code>hidden_layer_sizes</code>	(256, 128)
<code>activation</code>	<code>relu</code>
<code>max_iter</code>	10000
<code>alpha</code>	0.01
<code>random_state</code>	42

A continuación, se describen los efectos generales de cada parámetro:

- **hidden_layer_sizes:** Este parámetro define la arquitectura de las capas ocultas de la red neuronal. En este caso, tiene dos capas ocultas, una con 256 neuronas y otra con 128. El número y el tamaño de las capas ocultas afectan directamente la capacidad del modelo para aprender representaciones complejas de los datos. Un mayor número de neuronas o capas permite que el modelo capture patrones más complejos, pero también puede aumentar el riesgo de sobreajuste si no se ajusta adecuadamente.
- **activation:** El parámetro de activación define la función utilizada en las neuronas de las capas ocultas. En este caso, se utiliza ReLU (Rectified Linear Unit), que es una de las funciones de activación más comunes y eficientes. La función ReLU introduce no linealidad en el modelo, permitiendo que aprenda representaciones complejas de los datos. Además, es menos propensa a problemas de desvanecimiento del gradiente en redes profundas, lo que facilita el entrenamiento de redes grandes.

- **max_iter:** Este parámetro establece el número máximo de iteraciones (o épocas) para entrenar el modelo. En este caso, se fijó en 10,000. A mayor número de iteraciones, el modelo tiene más oportunidades para aprender de los datos, lo que puede mejorar el rendimiento. Sin embargo, un número demasiado alto puede llevar a un tiempo de entrenamiento innecesariamente largo, especialmente si el modelo ya ha convergido.
- **alpha:** El parámetro **alpha** controla la regularización L2, que es una técnica para prevenir el sobreajuste penalizando los pesos grandes. Un valor pequeño de **alpha** significa que la regularización tiene menos impacto, permitiendo que el modelo se ajuste más estrechamente a los datos de entrenamiento. Un valor más grande aumenta la regularización, lo que puede ayudar a generalizar mejor el modelo, pero puede reducir su capacidad para ajustarse a los detalles específicos del conjunto de entrenamiento.
- **random_state:** Este parámetro se utiliza para establecer la semilla aleatoria para la inicialización de los pesos y la división de los datos. Fijar un valor para **random_state** asegura que los resultados sean reproducibles. Si no se establece, cada ejecución del modelo puede resultar en diferentes configuraciones, lo que puede afectar la consistencia de los resultados.

G.3. Detalle de resultados.

Apéndice H

Anexo de sostenibilización curricular

H.1. Introducción

Este anexo incluirá una reflexión personal del alumnado sobre los aspectos de la sostenibilidad que se abordan en el trabajo. Se pueden incluir tantas subsecciones como sean necesarias con la intención de explicar las competencias de sostenibilidad adquiridas durante el alumnado y aplicadas al Trabajo de Fin de Grado.

Más información en el documento de la CRUE [https://www.crue.org/
wp-content/uploads/2020/02/Directrices_Sostenibilidad_Crue2012.pdf](https://www.crue.org/wp-content/uploads/2020/02/Directrices_Sostenibilidad_Crue2012.pdf).

Este anexo tendrá una extensión comprendida entre 600 y 800 palabras.

Bibliografía

- [ken, 2022] (2022). Parkinson's disease and camp lejeune contaminated water claims. *Ken Allen Law*.
- [inf, 2023] (2023). Sustancia química que permanece en el agua puede aumentar un 70 *InfoSalus*.
- [America, 2023] America, P. N. (2023). Los pesticidas y el cambio climático: Un círculo vicioso. Accedido: 2025-04-10.
- [Glassdoor, 2024] Glassdoor (2024). Salario medio data scientist junior - españa. Consultado en mayo de 2025.
- [Instituto Nacional sobre el Envejecimiento (NIA), 2022] Instituto Nacional sobre el Envejecimiento (NIA) (2022). La enfermedad de parkinson: causas, síntomas y tratamientos. Consultado el 10 de abril de 2025.
- [Kirrane et al., 2015] Kirrane, E. F., Bowman, C., Davis, J. A., Hoppin, J. A., Blair, A., Chen, H., Patel, M. M., Sandler, D. P., Tanner, C. M., Vinikoor-Imler, L., et al. (2015). Associations of ozone and pm2.5 concentrations with parkinson's disease among participants in the agricultural health study. *Journal of Occupational and Environmental Medicine*, 57(5):509–517.
- [Pacheco Moisés et al., 2011] Pacheco Moisés, F. P. et al. (2011). Toxicidad de plaguicidas y su asociación con la enfermedad de parkinson. *Archivos de neurociencias*, 16(1):33–39.
- [Pearce et al., 2013] Pearce, N. et al. (2013). Paraquat and parkinson's disease: A systematic review and meta-analysis of observational studies. *Environmental Health Perspectives*, 121(5):704–709.

- [Pyatha et al., 2022] Pyatha, S., Kim, H., Lee, D., and Kim, K. (2022). Association between heavy metal exposure and parkinson's disease: A review of the mechanisms related to oxidative stress. *Antioxidants*, 11(12):2467.
- [Roques, 2025] Roques, A. (2025). Plantuml: Herramienta de código abierto para la creación de diagramas uml. Versión consultada en mayo de 2025.
- [Starks et al., 2013] Starks, Z. et al. (2013). Pesticide exposure and parkinson's disease: The potential role of environmental factors. *Journal of Clinical Neuroscience*, 20(6):794–799.
- [Tanner et al., 2011] Tanner, C. M. et al. (2011). Pesticide exposure and parkinson's disease: A review of the literature. *Environmental Health Perspectives*, 119(6):823–827.
- [y ESADE, 2023] y ESADE, I. (2023). Informe del mercado laboral en españa 2023. Consultado en mayo de 2025.