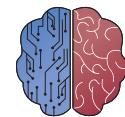




UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería de la Salud



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 parkinson
basado en la biometereología.
Documentación Técnica**

Presentado por Lorena Calvo Pérez
en Universidad de Burgos

29 de junio de 2025

Tutores: Antonio Canepa Oneto – Esther Cubo

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	4
Apéndice B Documentación de usuario	7
B.1. Requisitos software y hardware para ejecutar el proyecto.	7
B.2. Instalación / Puesta en marcha	7
B.3. Manuales y/o Demostraciones prácticas	8
Apéndice C Manual del desarrollador / programador / investigador.	21
C.1. Estructura de directorios	21
C.2. Compilación, instalación y ejecución del proyecto	25
C.3. Pruebas del sistema	29
C.4. Instrucciones para la modificación o mejora del proyecto.	29
Apéndice D Descripción de adquisición y tratamiento de datos	31
D.1. Descripción formal de los datos	31
D.2. Descripción clínica de los datos.	33

Apéndice E Manual de especificación de diseño	37
E.1. Diseño arquitectónico	37
Apéndice F Especificación de Requisitos	41
F.1. Diagrama de casos de uso	41
F.2. Explicación casos de uso.	43
F.3. Prototipos de interfaz o interacción con el proyecto	48
Apéndice G Estudio experimental	49
G.1. Cuaderno de trabajo.	49
G.2. Configuración y parametrización de las técnicas.	60
G.3. Detalle de resultados	69
Apéndice H Anexo de sostenibilización curricular	75
H.1. Introducción	75
H.2. Reflexión sobre salud,tecnología y sostenibilidad	75
H.3. Competencias desarrolladas	76
H.4. Aplicación práctica en el TFG	76
H.5. Conclusión	77
Bibliografía	79

Índice de figuras

A.1. Cronograma Inicial del Trabajo	2
A.2. Cronograma Final del Trabajo	2
B.1. Página de inicio de la aplicación (Pestaña Home)	9
B.2. Contenido de la Sección de Enfermedad de Parkinson	9
B.3. Contenido de la Sección de Mapa Mundial de la enfermedad de Parkinson	10
B.4. Contenido del Ver Mapa europeo en la sección de Mapa Mundial de la enfermedad de Parkinson	11
B.5. Descarga de datos en la sección de Mapa Mundial de párkinson	11
B.6. Contenido de la sección de Variables ambientales	12
B.7. Contenido botón Contaminacion de aire	12
B.8. Contenido botón Ver Mapa Europeo en sección de contaminacion de aire	12
B.9. Contenido botón uso de Pesticidas	13
B.10. Contenido botón Ver Mapa Europeo dentro de la sección de uso de pesticidas	13
B.11. Contenido de la sección de uso de pesticidas (Descarga de datos)	13
B.12. Contenido de la sección de Contaminación del aire Descarga de datos	14
B.13. Contenido de la sección de Predicciones	14
B.14. Contenido de la sección de Predicciones (mapas)	15
B.15. Contenido de la sección de Importancia de Variables	16
B.16. Contenido inferior de la sección de Importancia de variables	16
B.17. Contenido botón Modelo lineal	17
B.18. Contenido botón Modelo lineal al pulsar Ver Mapa Europeo	17
B.19. Contenido botón Modelos basados en árboles	17
B.20. Contenido botón Random Forest	18

B.21. Contenido botón Random Forest al pulsar Ver Mapa Europeo	18
B.22. Contenido botón Otros modelos de regresión	18
B.23. Contenido de la sección de Mas Información	18
E.1. Diagrama de despliegue de la aplicación web.	38
E.2. Diagrama de flujo del proceso de modelado y predicción.	39
E.3. Diagrama de paquetes del sistema.	40
F.1. Diagrama de casos de uso de la aplicación.	42
G.1. Matriz de correlación entre variables independientes.	50
G.2. Relación entre el párkinson y las Muertes atribuidas a fuentes de agua inseguras.	50
G.3. Relación entre el párkinson y la Tasa de carga de enfermedad por exposición al plomo.	51
G.4. Relación entre el párkinson y la Tasa de mortalidad por contaminación del aire.	51
G.5. Relación entre el párkinson y el uso de pesticidas.	52
G.6. Relación entre el párkinson y las precipitaciones.	52
G.7. Modelo Cuassi-poisson.	55
G.8. Modelo-Binomial Negativo	55
G.9. Importancia de variables Modleo GLM.	55
G.10. Importancia de variables Modleo RF.	56
G.11. Importancia de variables Modleo XG.	57
G.12. Importancia de variables Modleo SVR.	58
G.13. Importancia de variables Modleo KNN.	59
G.14. Importancia de variables Modleo MLP.	60

Índice de tablas

A.1. Costes de hardware	3
A.2. Costes de luz	3
A.3. Costes de personal simulados para el proyecto	4
A.4. Coste total del proyecto	4
G.1. Configuración aplicada al modelo GLM (Binomial Negativa) . .	61
G.2. Configuración aplicada al modelo Random Forest	62
G.3. Configuración aplicada al modelo XGBoost	64
G.4. Parámetros utilizados en el modelo SVR con variables transformadas	65
G.5. Parámetros utilizados en el entrenamiento del modelo KNN . .	66
G.6. Parámetros utilizados en el entrenamiento del modelo MLP . .	68
G.7. Resultados del modelo GLM	70
G.8. Resultados del modelo Random Forest	71
G.9. Resultados del modelo XGBoost	71
G.10.Resultados del modelo SVR	72
G.11.Resultados del modelo kNN	73
G.12.Resultados del modelo MLP	73
G.13.Comparativa global entre modelos	74

Apéndice A

Plan de Proyecto Software

A.1. Introducción

En este anexo se presenta el Plan de Proyecto Software realizado 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, para garantizar que el proyecto cumple con la normativa aplicable.

A.2. Planificación temporal

La planificación temporal del proyecto ha sido clave para estructurar adecuadamente el desarrollo del proyecto. El trabajo se ha dividido en distintos hitos o *milestones*, cada uno compuesto por un conjunto de tareas o *issues*. A cada *issue* se le ha asignado una estimación temporal.

Para ello, inicialmente se llevó a cabo la elaboración de un cronograma con la duración considerada para cada una de las tareas según su complejidad, representado con un diagrama de Gantt¹, elaborado a través de excel (Ver Figura A.1).

¹herramienta gráfica cuyo objetivo es exponer el tiempo de dedicación previsto para diferentes tareas

MILESTONES	ISSUES	Febrero	Marzo	Abril	Mayo	Junio
		07-14 07-14	15-21 01-07	22-28 08-14 15-21	01-04 05-11 12-18 19-25	02-09 10-16 17-23 24-31 02-06
RECOPILACIÓN DE DATOS	Obtención de datos necesarios					
ANÁLISIS DE LOS DATOS	Extracción, limpieza y análisis de datos					
	Construcción de mapas predesarrollo					
DESARROLLO DEL ALGORITMO	Definir los modelos a desarrollar					
	Diseñar los modelos de predicción					
	Creación de mapas resultantes					
	Realizar pruebas y corregir errores					
DESARROLLO DE LA APLICACIÓN SHINY	Aprendizaje de la aplicación SHINY					
	Diseñar la estructura y funcionalidad de la APP					
	Integración de los resultados obtenidos					
	Corrección de errores					
ELABORACIÓN DE LA MEMORIA	Escribir el documento del TFG					
	Revisar y corregir el documento					
ELABORACION DE LA PRESNTACIÓN	Preparar la presentación final					

Figura A.1: Cronograma Inicial del Trabajo

A su vez, durante la ejecución de las tareas, se ha llevado a cabo reuniones semanales con el tutor del proyecto para el seguimiento y evaluación del cumplimiento de los objetivos y, en caso necesario, reajustar el cronograma. Debido a que el tiempo necesario para la finalización de cada actividad no se ajustaba a lo establecido inicialmente, ha sido necesario reajustar temporalmente el cronograma, lo que ha llevado a la representación mediante Gantt de un cronograma final (Ver Figura A.2).

MILESTONES	ISSUES	Febrero	Marzo	Abril	Mayo	Junio	Julio
		07-14 07-14	15-21 01-07	22-28 08-14 15-21	01-04 05-11 12-18 19-25	02-09 10-16 17-23 24-31	02-08 09-15 16-22 23-30 01-08
RECOPILACIÓN DE DATOS	Obtención de datos necesarios						
ANÁLISIS DE LOS DATOS	Extraer, limpiar y analizar los datos						
	Construcción de mapas predesarrollo						
DESARROLLO DEL ALGORITMO	Definir los modelos a desarrollar						
	Diseñar los modelos de predicción						
	Creación de mapas resultantes						
	Realizar pruebas y corregir errores						
DESARROLLO DE LA APLICACIÓN SHINY	Aprendizaje de la aplicación SHINY						
	Diseñar la estructura y funcionalidad de la APP						
	Integración de los resultados obtenidos						
	Corrección de errores						
ELABORACIÓN DE LA MEMORIA	Escribir el documento del TFG						
	Revisar y corregir el documento						
ELABORACION DE LA PRESNTACIÓN	Preparar la presentación final						

Figura A.2: Cronograma Final del Trabajo

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 han generado 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 (5 meses).

Concepto	Coste en €	Coste amortizado (5 meses)
Ordenador portátil	1.300	90,28

Tabla A.1: Costes de hardware

A su vez, se debe considerar el gasto aproximado de luz (Iberdrola) necesario para la elaboración y ejecución de todo el proyecto (5 meses). Considerando un gasto de 60 kw/hora al mes, el gasto de la luz sería el siguiente;

Concepto	Coste mensual en €	Coste total (5 meses)
Luz	14	70

Tabla A.2: Costes de luz

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 durante 5 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, 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 [Pastoriza, 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 5 meses	12.500,00

Tabla A.3: 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	90.28
Costes de luz	70
Costes de personal	12.500
Coste total estimado	12.660,28

Tabla A.4: Coste total del proyecto

A.4. Viabilidad legal

En este proyecto se han utilizado datos publicados en línea y se debe considerar la viabilidad legal de los mismos, así como el cumplimiento de regulaciones vigentes.

Leyes vigentes y derechos de autor

Los datos publicados en línea se rigen sobre el **Reglamento General de Protección de Datos (RGPD)** de la Unión Europea que garantiza la protección de datos personales. Éstas normas se aplican tanto a empresas como a organizaciones (públicas o privadas), con sede en la unión europea o fuera de ella.

Además, el uso de datos públicos está respaldado por la Ley 19/2013, de 9 de diciembre, de transparencia, acceso a la información pública y buen gobierno, así como por la Ley 18/2015, que regula la reutilización de la información del sector público. Estas leyes fomentan el acceso libre a los

datos generados por las administraciones públicas y respaldan su utilización en proyectos de investigación, siempre que se respeten las condiciones de uso.

En cuanto a los derechos de autor, se rigen por el Real Decreto Legislativo 1/1996 de la Ley de Propiedad Intelectual en el que define las bases de la propiedad intelectual y explica cómo funcionan los derechos de autor, incluyendo las condiciones y características que los regulan.

Licencias del software

La plataforma de la cual se han adquirido los datos es *Our World In Data* (OWID), una organización sin fines de lucro que publica datos de libre acceso relacionados con temáticas sociales, económicas y medioambientales. Los datos escogidos para el desarrollo del trabajo ofrecen información estadística de carácter agregado, sin incluir datos personales identificables.

Una licencia de software es un contrato legal que define cómo un usuario puede usar un programa, indicando qué acciones están permitidas y cuáles no, en cuanto a su uso, copia, distribución o modificación [Eguía, 2024].

OWID publica sus datos bajo la licencia **Creative Commons BY 4.0 (CC BY 4.0)**, lo que implica que los datos pueden ser copiados, modificados, distribuidos y utilizados incluso con fines comerciales, siempre que se refiera adecuadamente a la fuente original. Además, las herramientas y el software desarrollados por OWID son de código abierto y están disponibles bajo la **licencia MIT**, lo que permite su uso, modificación y redistribución sin restricciones significativas.

Es importante mencionar que, aunque la mayoría del contenido de OWID es de libre acceso, algunos gráficos o conjuntos de datos específicos pueden estar sujetos a licencias de terceros, como es el caso de ciertos datos utilizados en este trabajo que provienen del **Institute for Health Metrics and Evaluation (IHME)**. Este organismo establece sus propias políticas de uso y condiciones legales. En particular, los datos del IHME están disponibles para su uso no comercial, pero no permite su redistribución, bajo una licencia personalizada, y requieren el reconocimiento explícito de la fuente y el cumplimiento de sus términos y condiciones de uso.

Finalmente en cuanto al 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 (Shiny), es de

código abierto bajo licencias compatibles con el desarrollo y distribución de proyectos académicos y científicos.

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.



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 de la enfermedad de 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 de 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 de parkinson a lo largo del tiempo.(Véase la Figura ??)



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

Tras ello 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.

Inmediatamente debajo del control deslizante y del botón "Ver mapa europeo", se encuentra una sección titulada "Descarga de datos", en la que



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

aparece un mensaje comunicándole al usuario que estos datos no se encuentra disponible para la descarga porque no se permite su redistribución, ya que se encuentran bajo licencia.

Seguido de esto, aparece una sección titulada "Citas", que incluye la citación de las webs a través de las cuales se han obtenido dichos datos y que son necesarias realizarlas para el cumplimiento de su política.(vease Figura B.5)



Figura B.5: Descarga de datos en la sección de Mapa Mundial de párkinson

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 de la Enfermedad de 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)



Figura B.7: Contenido botón Contaminación de aire



Figura B.8: Contenido botón Ver Mapa Europeo en sección de contaminación de aire

- Herramientas para filtrar los datos por país y año, y descargar los resultados en formato CSV o JSON en aquellas variables cuyos conjunto se datos tengan el premiso de redistribución, como el uso de pesticidas y precipitaciones (Véanse las figuras B.9,B.10, B.11)

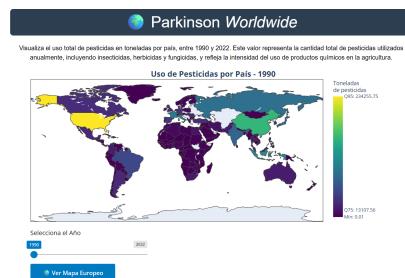


Figura B.9: Contenido botón uso de Pesticidas



Figura B.10: Contenido botón Ver Mapa Europeo dentro de la sección de uso de pesticidas



Figura B.11: Contenido de la sección de uso de pesticidas (Descarga de datos)

En el caso de el resto de variables aparece un mensaje como el descrito en la sección Mapa Global de la enfermedad de Parkinson (Vease Figura B.12).

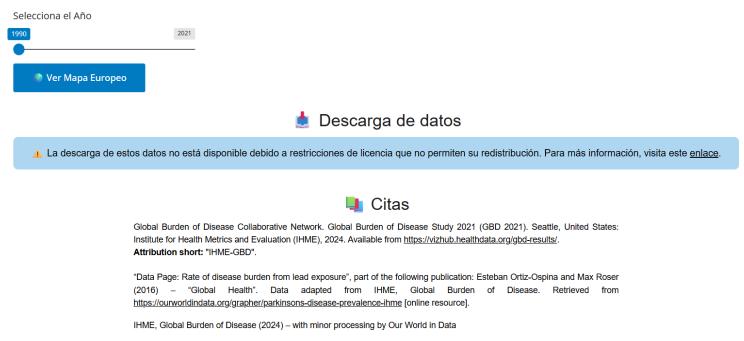


Figura B.12: Contenido de la sección de Contaminación del aire Descarga de datos

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.13)



Figura B.13: Contenido de la sección de Predicciones

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 de parkinson predicho por seis modelos diferentes

de aprendizaje automático. La visualización permite observar cómo se distribuyen estas estimaciones por país y por año.

- **Mapa de incertidumbre:**Este mapa 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.14

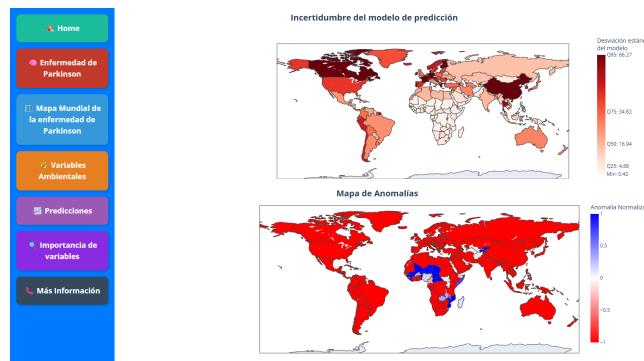


Figura B.14: Contenido de la sección de Predicciones (mapas)

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.15)



Figura B.15: Contenido de la sección de Importancia de Variables

Debajo del gráfico de ranking global, el usuario encontrará tres botones interactivos (Figura B.16), cada uno de los cuales da acceso a secciones específicas dedicadas a los modelos de predicción utilizados.

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.



Figura B.16: Contenido inferior de la sección de Importancia de variables

- 1. 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 de parkinson por país(Figura B.17), 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 B.18)



Figura B.17: Contenido botón Modelo lineal



Figura B.18: Contenido botón Modelo lineal al pulsar Ver Mapa Europeo

2. **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.19)

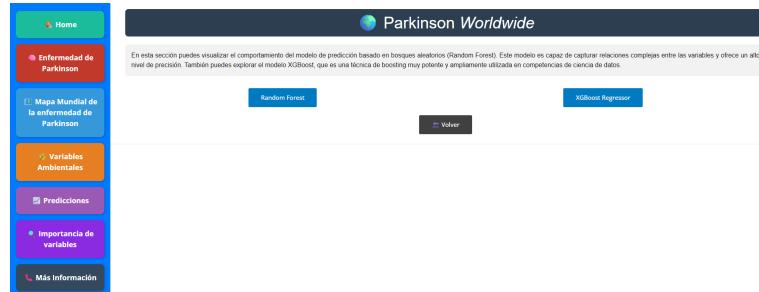


Figura B.19: Contenido botón Modelos basados en árboles

Ejemplo de contenido del botón Random Forest (mismo contenido para el segundo botón) puede verse en Figura G.2 Y B.21.

3. **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 entre la vista global y la vista europea, siguiendo la misma dinámica



Figura B.20: Contenido botón Random Forest



Figura B.21: Contenido botón Random Forest al pulsar Ver Mapa Europeo

que en otras secciones.(véase Figura B.22) el contenido de cada botón es similar al ya explicado en los anteriores

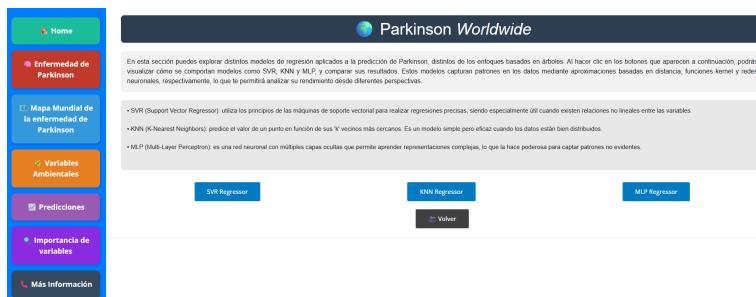


Figura B.22: Contenido botón Otros modelos de regresión

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



Figura B.23: Contenido de la sección de Mas Información

con el desarrollo de esta aplicación, lo que permite ampliar información o conocer otros proyectos vinculados (Ver Figura B.23).

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

La estructura que presentan los directorios del proyecto es la se encuentra en [mi GitHub](#) y consta de:

Carpeta de Referencias

La carpeta **Referencias** contiene:

- **APIS.txt:**

Archivo que contiene las URLs de las APIs utilizadas para obtener los datos asociados a cada una de las variables del estudio.

- **link_datos.txt:**

Contiene los enlaces directos a los sitios web desde los cuales se obtuvieron los conjuntos de datos utilizados en el análisis.

Carpeta notebooks

En la carpeta notebooks se encuentra:

1. Obtencion_procesamiento_datos.ipynb:

En él podemos encontrar la carga de los datos a través de peticiones a la API, junto con su posterior procesamiento y estructurara para obtener el Dataframe final. A su vez, contiene la construcción de las tablas necesarias para la realización del resto del proyecto.

2. Analisis_exploratorio_preliminar.ipynb:

Se realiza un análisis exploratorio de los datos. Se estudia la relación entre las variables independientes y la variable objetivo. Además, se incluye una matriz de correlación para detectar posibles problemas de multicolinealidad entre variables.

3. Búsqueda_hiperparametros.ipynb:

Se prueban diferentes modelos y combinaciones de hiperparámetros con el fin de identificar los más adecuados para el conjunto de datos.

4. Significancia.ipynb:

Utilizando los modelos con los mejores hiperparámetros identificados previamente, se evalúa la significancia de cada variable en los distintos modelos. También se genera un ranking de importancia para cada uno de ellos así como un *ranking* promedio final de todos ellos.

5. Entrenar_predecir.ipynb:

Este notebook contiene el entrenamiento final de los modelos y la predicción de la prevalencia de parkinson. Se generan:

- Un mapa promedio final que resume los resultados de todos los modelos.
- Un mapa de incertidumbre entre modelos.
- Un mapa de anomalías para identificar sobreestimación o subestimación.

6. Mapas.ipynb:

Contiene todos los mapas generados a partir de los datos crudos justo después de su obtención, antes de aplicar los modelos o análisis posteriores.

Carpeta src

En la carpeta src se encuentra:

- **Entrenamiento_Modelos.py:**

Contiene todas las funciones necesarias para el entrenamiento de modelos. Este script es importado y utilizado en el notebook `Entrenar_predecir.ipynb` para realizar el entrenamiento de forma estructurada y reproducible.

- **Prediccion_Modelos.py:**

Incluye las funciones encargadas de realizar las predicciones con los modelos previamente entrenados. También es importado dentro de `Entrenar_predecir.ipynb`, facilitando la ejecución de predicciones sobre los datos y la generación de los mapas correspondientes.

Carpeta de Resultados

La carpeta Resultados contiene:

- **Archivos_csv:**

Dentro de esta carpeta se encuentran dos subcarpetas:

- **Datos:**

Contiene los archivos `.csv` correspondientes a los datos inmediatamente después de ser obtenidos, antes de ser procesados o utilizados en el análisis. También incluye las tablas base utilizadas para el análisis posterior.

- **Predicciones:**

Almacena los archivos `.csv` que contienen las predicciones generadas por cada uno de los modelos implementados y los correspondientes a los mapas de incertidumbre, anomalías y promedio de la predicción.

- **Importancia_variables:**

Esta carpeta contiene:

- **CSV:**

Contiene los archivos `.csv` que representan los *rankings* de importancia de las variables generados por cada modelo de forma individual, así como un *ranking* promedio final que resume la importancia global de las variables a través de todos los modelos evaluados.

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

■ Mapas:

Esta carpeta almacena todos los mapas generados durante el proyecto, organizados en dos subcarpetas:

- **Mapas_datos_originales:**

Esta carpeta contiene los mapas correspondientes a los datos crudos, es decir, los mapas generados justo después de la obtención de los datos, antes de aplicar modelos o transformaciones.

- **Predicciones:**

Incluye los mapas generados a partir de las predicciones de los modelos. Aquí se encuentran:

- Mapas individuales por modelo.
- Mapa promedio de predicción de todos los modelos.
- Mapa de incertidumbre entre modelos.
- Mapa de anomalías.

Carpeta de App

La carpeta App contiene los elementos necesarios para la ejecución de una aplicación desarrollada como parte del proyecto. Su contenido es el siguiente:

■ **diseño.py:**

Contiene el código fuente de la aplicación. En este *script* se define la lógica de visualización, interacción y funcionalidad general de la app, incluyendo la presentación de resultados y/o visualizaciones dinámicas.

■ **requirements.txt:**

Archivo que especifica las dependencias y bibliotecas necesarias para ejecutar correctamente la aplicación.

Carpeta Documentacion

Incluye los documentos formales del proyecto:

- Documento de la memoria.
- Documento del anexo.

Carpeta *Images*

Contiene imágenes utilizadas principalmente para la elaboración del archivo README.txt.

Otros archvivos

En el directorio raíz del proyecto se encuentran los siguientes archivos:

- **README.txt**: Breve descripción del proyecto, instrucciones de uso e información general.
- **LICENSE**: Licencia bajo la cual se distribuye el proyecto.

C.2. Compilación, instalación y ejecución del proyecto

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.

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
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. Acceso a datos mediante API

Para ampliar el conjunto de datos con nuevas variables y mantener la aplicación actualizada automáticamente, es necesario acceder a la fuente original de datos a través de la API o los archivos JSON disponibles públicamente.

En el caso de *Our World In Data (OWID)*, el acceso se realiza de la siguiente forma:

- A partir de la URL base del conjunto de datos, se accede al archivo metadata.json cambiando la extensión de la URL correspondiente.
- El archivo metadata.json contiene información sobre la estructura del conjunto de datos, así como un enlace a metadatos más detallados.
- Para acceder a los datos restantes (data.json) que no contiene el metadata.json , se utilizaron herramientas de desarrollo web (como el panel "Network" del inspector del navegador) que permiten visualizar las peticiones que realiza la página al cargar los gráficos. De esta forma, se identificó la URL desde la que se cargan los datos completos en formato JSON, accesible directamente y actualizada periódicamente.
- Estos archivos JSON (metadata.json y data.json) pueden ser consultados directamente para obtener tanto la descripción de las variables como los datos actualizados, permitiendo construir o actualizar dinámicamente el dataset en la aplicación.

6. Procesamiento de datos

Una vez identificadas las URLs de acceso a los archivos ya mencionados de la plataforma *Our World In Data*, el siguiente paso consistió en obtener y estructurar los datos para su análisis.

Para ello, se implementaron funciones que realizan peticiones HTTP (requests) a estas direcciones. La información fue descargada en formato JSON, lo que permitió su tratamiento directo en lenguajes como Python.

Cabe destacar que los datos estaban distribuidos en dos archivos con estructuras distintas:

- **data.json:** Contiene los valores de los indicadores a lo largo del tiempo, asociados a identificadores numéricos de países y años.
- **metadata.json:** Proporciona información auxiliar como la correspondencia entre identificadores numéricos y nombres de países, así como los años disponibles en el conjunto de datos.

Dada la diferencia en la estructura de estos archivos, fue necesario implementar un proceso de integración para combinar ambos y construir un dataset coherente y estructurado.

7. Estructuración de Datos

Posteriormente, se procedió a extraer los valores relevantes del archivo data.json, y a cruzarlos con la información presente en metadata.json. Esto permitió:

- Asociar correctamente cada valor de indicador con su país correspondiente.
- Filtrar los años válidos para el análisis, definidos en la metadata.
- Organizar los datos en una estructura clara y homogénea.

En esta etapa también se implementaron controles para manejar posibles valores faltantes o inconsistencias entre los archivos.

8. Conversión a DataFrame

Finalmente, los datos estructurados fueron convertidos en un *Data-Frame*, una estructura tabular ampliamente utilizada para el análisis de datos. Cada fila representa una combinación única de país y año, con su correspondiente valor del indicador. En los casos donde un

valor no estaba disponible, se utilizó un valor nulo o se dejó en blanco, manteniendo así la integridad del conjunto de datos.

Gracias a este procesamiento, fue posible construir un conjunto de datos listo para análisis, visualización y uso dinámico dentro de la aplicación.

9. Selección y entrenamiento de modelos

Dado que el objetivo de la predicción es un valor numérico continuo, se optó por abordar el problema mediante modelos de regresión. Por ello, dependiendo del objetivo de la predicción, los modelos a considerar serán distintos.

Para cada modelo, se realizó una búsqueda exhaustiva de hiperparámetros óptimos utilizando `GridSearchCV` con validación cruzada de 5 particiones (`cv=5`), garantizando así una estimación robusta del rendimiento y evitando el sobreajuste.

La división de los datos se realizó en un 80 % para entrenamiento y un 20 % para prueba, asegurando la evaluación en datos no vistos durante el entrenamiento.

10. Evaluación de la importancia de variables

Para interpretar la relevancia de las variables predictoras en los modelos ajustados, se utilizaron métodos específicos según el modelo:

- En modelos basados en árboles (Random Forest, XGBoost), así como SVR, MLP Y KNN se analizó la *feature importance* para identificar las variables con mayor impacto en la predicción.
- En el caso del modelo GLM, se examinaron los coeficientes y sus correspondientes p-valores para determinar la significancia estadística de cada variable.

11. Despliegue de la aplicación

La aplicación interactiva desarrollada con *Shiny* fue desplegada en la plataforma *Posit Cloud* (anteriormente RStudio Cloud), facilitando el acceso web sin necesidad de instalación local.

Este entorno gestionado permite actualizar y mantener la aplicación fácilmente, además de compartirla con usuarios y colaboradores.

El código fuente, junto con toda la documentación necesaria, está disponible en el repositorio GitHub vinculado al proyecto.

C.3. Pruebas del sistema

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

- **Pruebas de despliegue y accesibilidad:** Se realizaron pruebas para verificar el correcto funcionamiento de la aplicación una vez desplegada. Para ello, se compartió el enlace público de acceso con distintos usuarios, permitiendo comprobar que era posible acceder a la aplicación desde dispositivos y ubicaciones diferentes a la del desarrollador.

Asimismo, se verificó que múltiples usuarios pudieran acceder simultáneamente sin que se presentaran errores de funcionamiento o caídas del servicio.

Por último, se comprobó que la aplicación estuviera disponible de forma continua, es decir, que no entrara en estado de suspensión tras un período de inactividad, garantizando así su accesibilidad en cualquier momento.

- **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.

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

C.4. Instrucciones para la modificación o mejora del proyecto.

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.

30 *Apéndice C. Manual del desarrollador / programador / investigador.*

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.
- 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

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, la tasa de carga de morbilidad por exposición al plomo, las muertes atribuidas a fuentes de agua insalubres, el uso de pesticidas y la precipitación anual.

D.1.1. Prevalencia de la enfermedad de Parkinson

- **Definición y unidad de medida:** Esta variable se define como el numero estimado de personas con enfermedad de 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:** La prevalencia de la EP 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 de 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 es contextualizar de que manera estas variables pueden influir en la prevalencia de la enfermedad de Parkinson.

D.2.1. Prevalencia de la enfermedad de 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 párkinson [Kirrane 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 párkinson 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 párkinson 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 tetrachloroetileno (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)**.

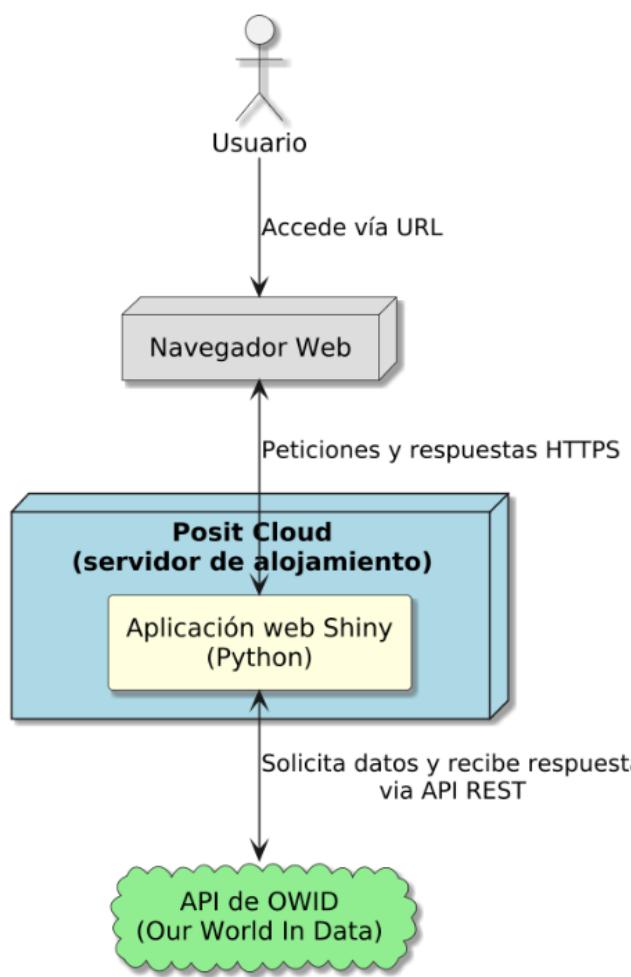


Figura E.1: Diagrama de despliegue de la aplicación web.

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.

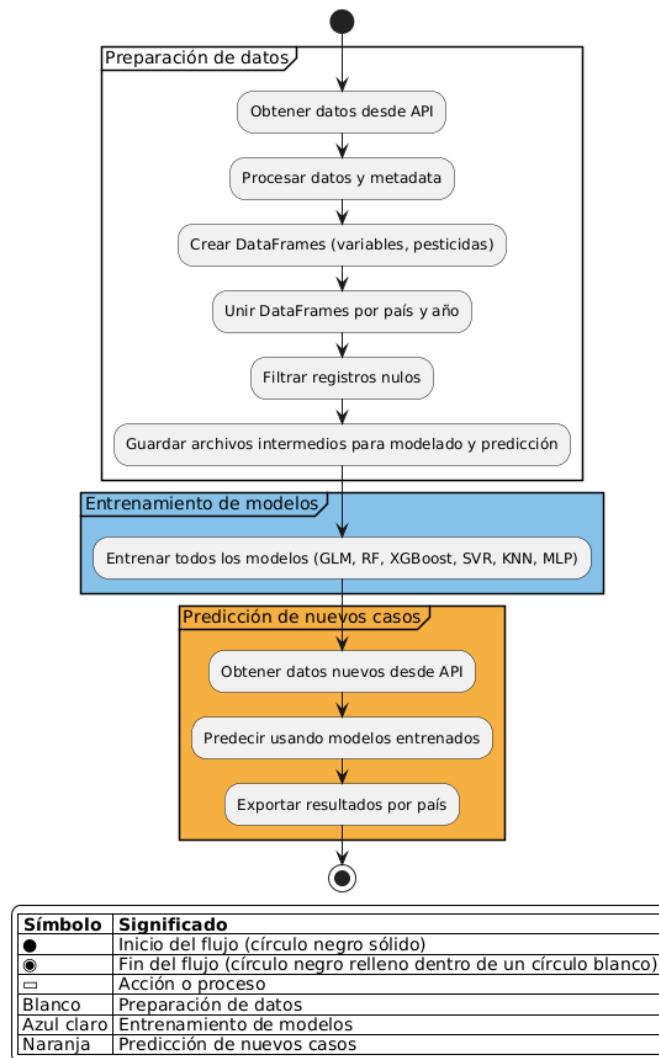


Figura E.2: Diagrama de flujo del proceso de modelado y predicción.

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 los componentes relacionados, facilitando la comprensión de la estructura general.

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.

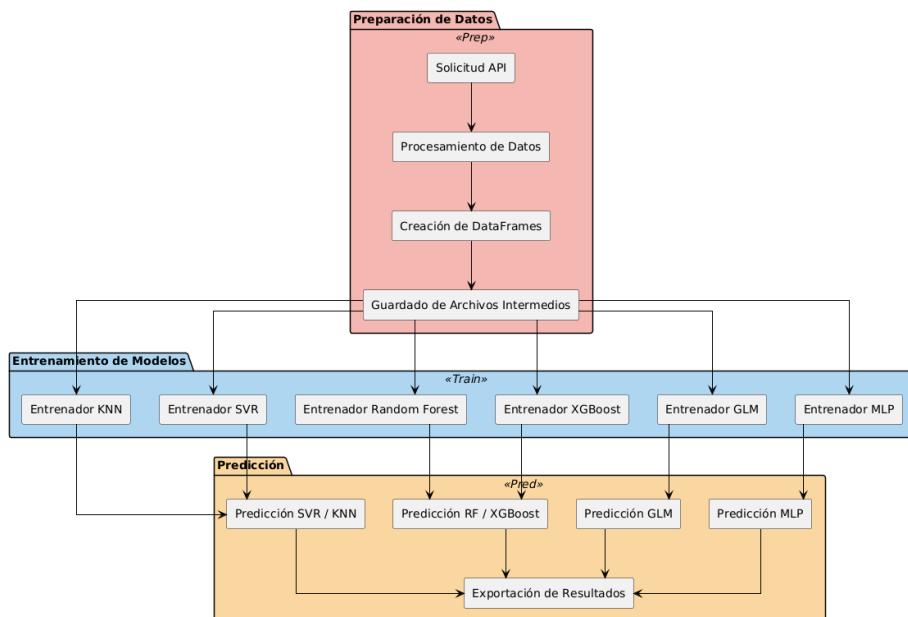


Figura E.3: Diagrama de paquetes del sistema.

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.

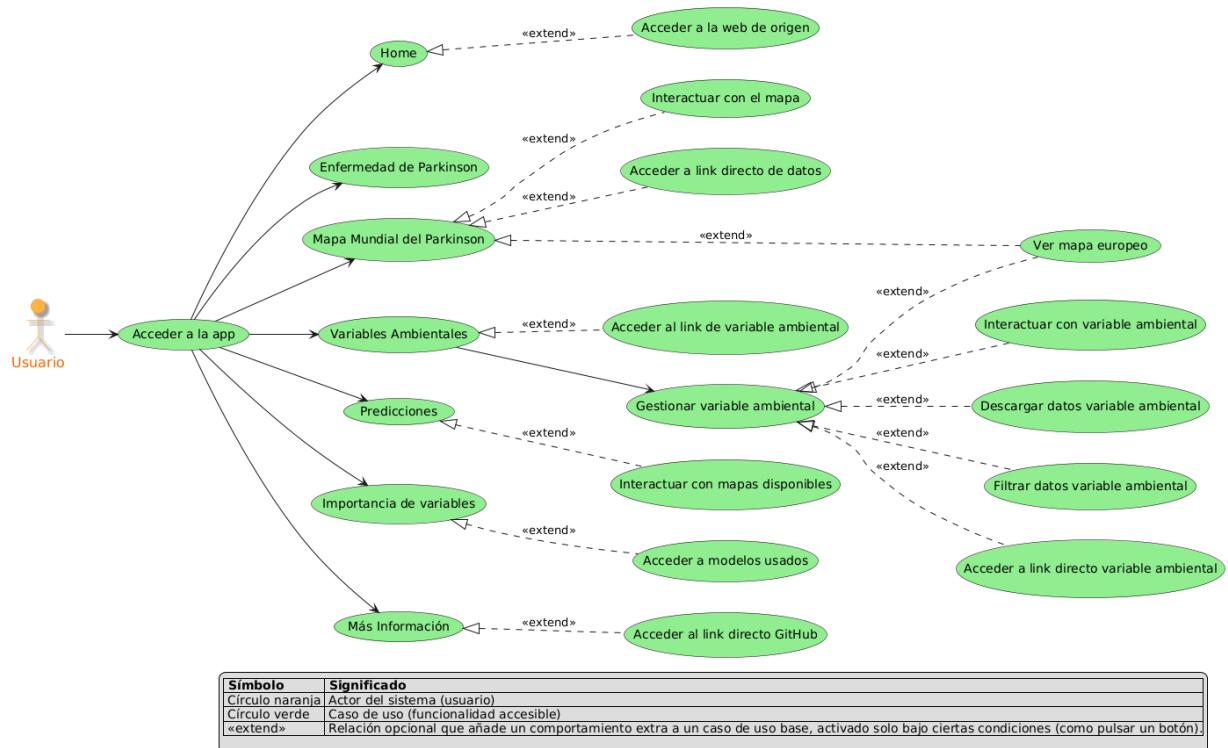


Figura F.1: Diagrama de casos de uso de la aplicación.

F.2. Explicación casos de uso.

CU-1: Acceder a la app

Campo	Descripción
Versión	1.0
Autor	Lorena Calvo Pérez
Descripción	Permite al usuario acceder a la aplicación.
Precondición	Usuario tiene dispositivo con conexión a internet.
Acciones	<ol style="list-style-type: none"> 1. Usuario abre la app. 2. Sistema muestra pantalla de inicio.
Postcondición	Usuario está dentro de la aplicación.
Excepciones	Error de conexión a internet.
Importancia	Alta

CU-2: Home

Campo	Descripción
Versión	1.0
Autor	Lorena Clavo Pérez
Requisitos asociados	RF-01
Descripción	Pantalla principal donde se muestra la navegación general. Desde aquí, el usuario puede navegar a otras funcionalidades principales o, de forma opcional, acceder a la web de origen de los datos (extensión).
Precondición	Usuario ha accedido a la app.
Acciones	<ol style="list-style-type: none"> 1. Mostrar información sobre la aplicación. 2. Opcional: Acceder a la web de origen de los datos (extensión).
Postcondición	El usuario ve la pantalla principal con la información.
Importancia	Alta

CU-3: Enfermedad de Parkinson

Campo	Descripción
Versión	1.0
Autor	Lorena Calvo Pérez
Requisitos asociados	RF-01
Descripción	Proporciona información sobre la enfermedad de Parkinson.
Precondición	Usuario ha accedido a la app.
Acciones	<ol style="list-style-type: none"> 1. Usuario selecciona la opción. 2. Sistema muestra información.
Postcondición	Usuario obtiene información relevante.
Importancia	Media

CU-4: Mapa Mundial de la enfermedad de Parkinson

Campo	Descripción
Versión	1.0
Autor	Lorena Calvo Pérez
Requisitos asociados	RF-01
Descripción	Muestra un mapa mundial con datos relacionados de la enfermedad de Parkinson.
Precondición	Usuario ha accedido a la app.
Acciones	<ol style="list-style-type: none"> 1. Mostrar mapa. 2. Permitir visualización de la enfermedad de Parkinson a través de los años disponibles. 3. Permite acceder al mapa europeo en el caso de que el usuario quiera visualizarlo.
Postcondición	Usuario visualiza datos del mapa.
Excepciones	Error al cargar datos del mapa.
Importancia	Alta

CU-5: Variables Ambientales

Campo	Descripción
Versión	1.0
Autor	Lorena Calvo Pérez
Requisitos asociados	RF-01
Descripción	Acceso a información sobre variables ambientales relacionadas.
Precondición	Usuario ha accedido a la app.
Acciones	<ol style="list-style-type: none"> 1. Mostrar variables ambientales. 2. Permitir selección y análisis. 3. Permite acceso a los datos vía enlace
Postcondición	Visualiza información sobre variables
Excepciones	Error en la carga de datos.
Importancia	Media

CU-6: Gestionar variable ambiental

Campo	Descripción
Versión	1.0
Autor	Lorena Calvo Pérez
Requisitos asociados	RF-01, RF-05
Descripción	Permite al usuario gestionar las variables ambientales.
Precondición	Usuario ha accedido a la app y a variables ambientales.
Acciones	<ol style="list-style-type: none"> 1. Seleccionar variable. 2. Visualizar datos a nivel mundial europeo 3. Descarga y filtrado de datos para los conjuntos disponibles para redistribución.
Postcondición	Visualización de variables.
Excepciones	Error en la carga de datos.
Importancia	Alta

CU-7: Predicciones

Campo	Descripción
Versión	1.0
Autor	Alumno
Requisitos asociados	RF-01
Descripción	Permite interactuar con mapas y datos de predicciones.
Precondición	Usuario ha accedido a la app.
Acciones	<ol style="list-style-type: none"> 1. Mostrar mapas predicción. 2. Permitir interacción.
Postcondición	Usuario obtiene datos predictivos.
Excepciones	Error en conexión.
Importancia	Media

CU-8: Importancia de variables

Campo	Descripción
Versión	1.0
Autor	Alumno
Requisitos asociados	RF-01
Descripción	Muestra la importancia o peso de variables usadas en modelos.
Precondición	Usuario ha accedido a la app.
Acciones	<ol style="list-style-type: none"> 1. Mostrar variables y su importancia. 2. Permitir consulta de los modelos utilizados 3. Al seleccionar un modelo se podrá visualizar el mapa asociado a la predicción con es modelo.
Postcondición	Usuario comprende relevancia de variables.
Excepciones	Error de conexión.
Importancia	Baja

CU-9: Más Información

Campo	Descripción
Versión	1.0
Autor	Alumno
Requisitos asociados	RF-01
Descripción	Proporciona enlaces adicionales, como GitHub.
Precondición	Usuario ha accedido a la app.
Acciones	<ol style="list-style-type: none">1. Permitir acceso a enlaces para ponerse en contacto.
Postcondición	Usuario navega a recursos externos.
Excepciones	Enlace inaccesible.
Importancia	Baja

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.

Apéndice G

Estudio experimental

G.1. Cuaderno de trabajo.

Con el fin de evaluar la relación entre las variables ambientales y la prevalencia de párkinson 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 párkinson). 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 párkinson, donde se observa una curvatura inicial que luego se estabiliza, lo que sugiere una relación no lineal que podría capturarse adecuadamente mediante modelos polinómicos.

De manera similar, las variables Exposición al plomo y Contaminación del aire presentan una estructura curvada en sus relaciones con los casos de párkinson, lo que también apunta hacia la necesidad de modelos que consideren términos no lineales (ver Figura G.3 y Figura G.4).

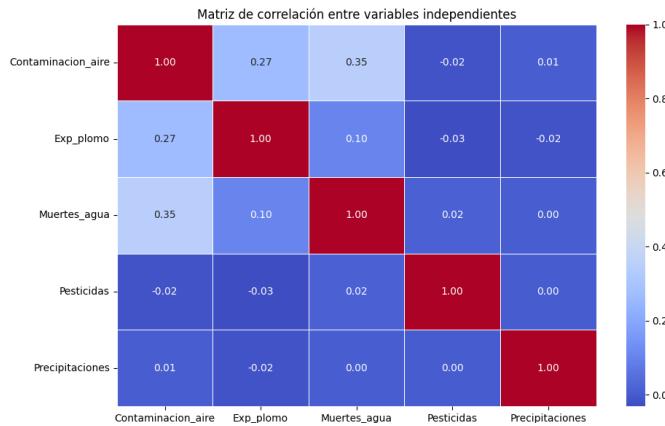


Figura G.1: Matriz de correlación entre variables independientes.

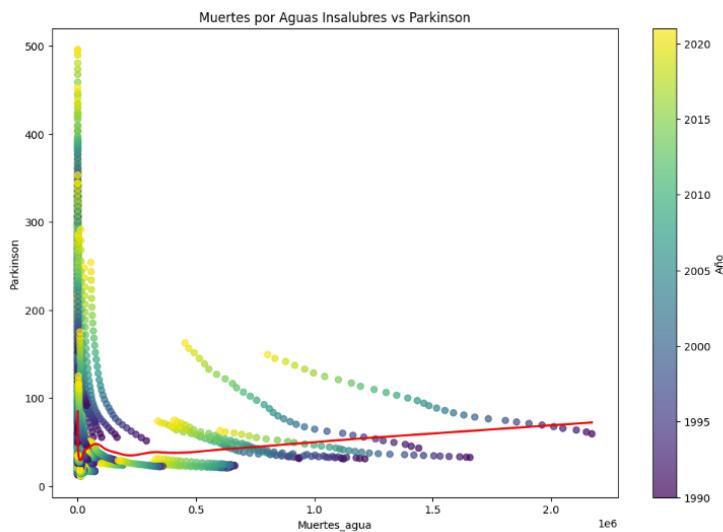


Figura G.2: Relación entre el párkinson y las Muertes atribuidas a fuentes de agua inseguras.

En el caso de los Pesticidas, la relación muestra varios picos, cambios de tendencia y una clara estructura no lineal. Este patrón respalda el uso de modelos polinómicos de mayor orden, que permitan capturar adecuadamente la complejidad observada (Figura G.5).

Por su parte, la variable Precipitaciones también presenta un patrón curvado, con un aumento inicial que se estabiliza posteriormente. Este comportamiento no es consistente con una relación lineal simple, por lo que

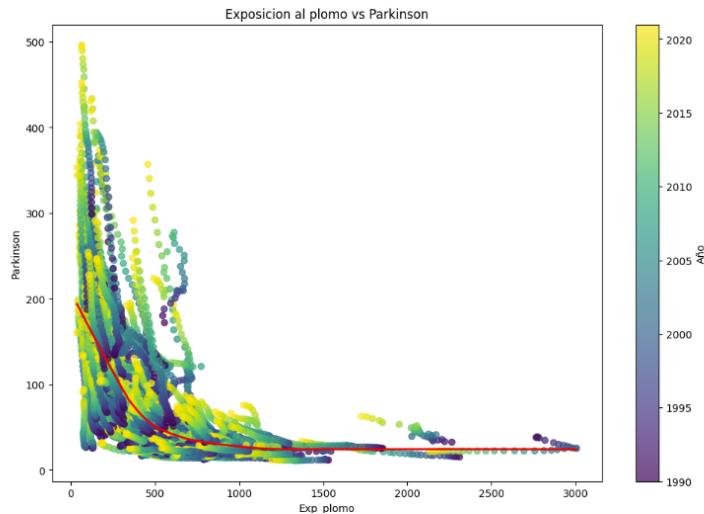


Figura G.3: Relación entre el párkinson y la Tasa de carga de enfermedad por exposición al plomo.

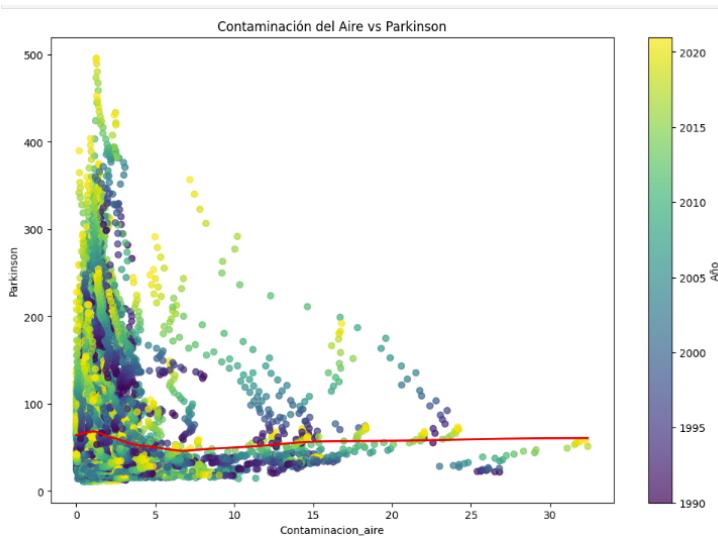


Figura G.4: Relación entre el párkinson y la Tasa de mortalidad por contaminación del aire.

nuevamente se opta por modelos polinómicos para reflejar adecuadamente esta forma (Figura G.6).

Este análisis exploratorio se centró en identificar la forma de las relaciones entre las variables predictoras y los casos de párkinson, con el objetivo de seleccionar modelos que se ajusten a los patrones observados. Así, se garantiza

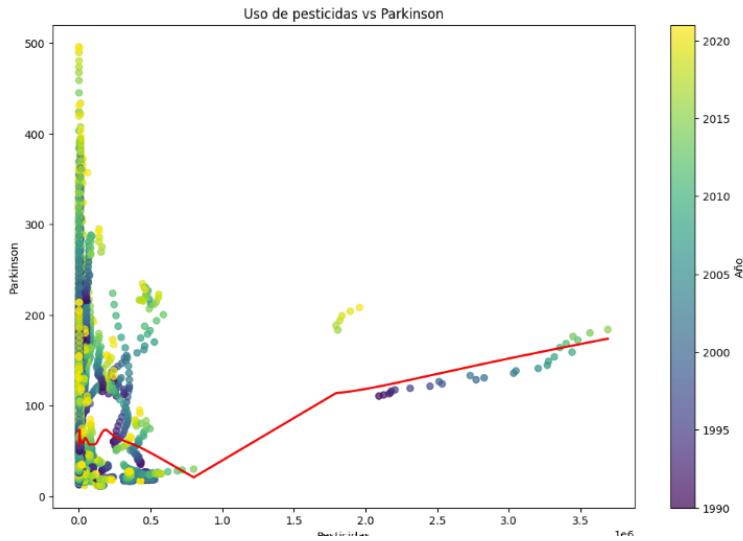


Figura G.5: Relación entre el párkinson y el uso de pesticidas.

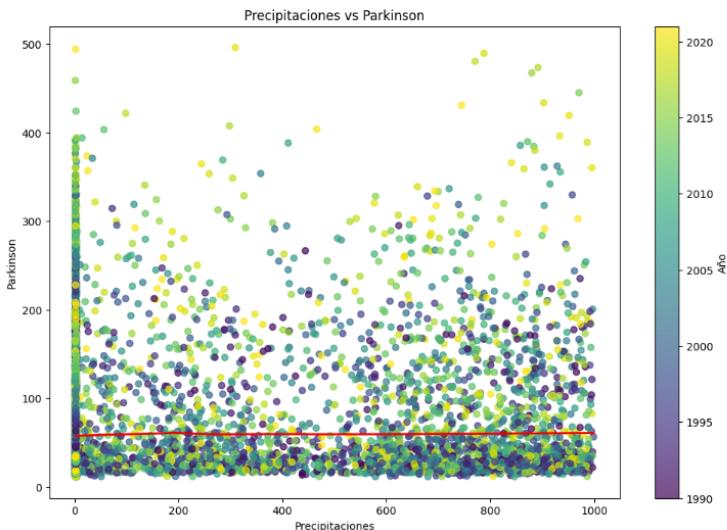


Figura G.6: Relación entre el párkinson y las precipitaciones.

una mejor capacidad explicativa y predictiva al respetar la estructura propia de los datos. 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

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 qué 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 párkinson, 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 párkinson y X_i son las variables ambientales consideradas. Esta fórmula 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) ya que estos permiten capturar la curvatura observada sin perder sentido práctico ni interpretabilidad. 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.

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.

- **Selección e importancia de variables:** Para determinar qué variables ambientales debían formar parte del modelo final, se utilizó el procedimiento de selección hacia atrás (***backward elimination***). Este consiste en comenzar con todas las variables candidatas, incluyendo sus términos polinómicos (hasta grado 3), evaluando su significancia estadística (p-valor) y eliminando iterativamente aquellas que no alcanzaban el umbral de significancia ($p > 0.05$).

En cada iteración se reentrenó el modelo GLM con las variables restantes, hasta obtener una combinación en la que todos los términos incluidos fueran estadísticamente significativos. Este

Modelo Cuasi-Poisson:		Generalized Linear Model Regression Results					
Dep. Variable:	Parkinson	No. Observations:	4323	Model:	GLM	Df Residuals:	4313
Model Family:	Poisson	Df Model:	9	Link Function:	NegativeBinomial	Df Model:	1.0000
Method:	IRLS	Log Scale:	1.0000	Date:	Thu, 10 Apr 2025	Deviance:	57750.
Time:	18:05:10	Log-Likelihood:	-57750.	No. Iterations:	13	Pearson chi2:	1150.3
Date:	Thu, 10 Apr 2025	Deviance:	90197.	No. Iterations:	13	Pseudo R-squ. (CS):	0.4104
Time:	18:05:10	Pearson chi2:	1.00e+05	Covariance Type:	nonrobust		
No. Iterations:	6	Pseudo R-squ. (CS):	1.000				
Covariance Type:	nonrobust						
const	4.1322	0.002	1913.827	0.000	4.128	4.136	
Contaminacion_aire	0.2118	0.005	44.254	0.000	0.202	0.221	
Exp_plomo	-1.5512	0.005	-310.109	0.000	-1.561	-1.541	
Muertes_agua	-0.3351	0.009	-36.244	0.000	-0.353	-0.317	
Pesticidas	-0.0001	0.000	-3.000	0.000	-0.000	-0.000	
Precipitaciones	0.0024	0.002	1.448	0.148	-0.001	0.206	
Contaminacion_aire_2	-0.0715	0.005	-15.303	0.000	-0.081	-0.062	
Muertes_agua_2	0.2830	0.008	34.030	0.000	0.267	0.299	
Exp_plomo_2	0.9234	0.005	175.366	0.000	0.913	0.934	
Pesticidas_log	0.0535	0.000	29.559	0.000	0.050	0.057	

Modelo Binomial Negativo:							
Generalized Linear Model Regression Results							
Dep. Variable:	Parkinson	No. Observations:	4323	Model:	GLM	Df Residuals:	4313
Model Family:	NegativeBinomial	Df Model:	9	Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-22292.	Date:	Thu, 10 Apr 2025	Deviance:	1150.3
Time:	18:05:01	Pearson chi2:	1.56e+03	No. Iterations:	13	Pseudo R-squ. (CS):	0.4104
Covariance Type:	nonrobust						
coef	std err	z	P> z	[0.025	0.975]		
const	4.1466	0.015	269.926	0.000	4.116	4.177	
Contaminacion_aire	0.1897	0.039	4.896	0.000	0.114	0.266	
Exp_plomo	-1.4937	0.043	-34.496	0.000	-1.579	-1.409	
Muertes_agua	-0.2855	0.050	-5.722	0.000	-0.383	-0.188	
Pesticidas	-0.0057	0.017	-0.341	0.733	-0.039	0.027	
Precipitaciones	0.0027	0.013	0.205	0.837	-0.007	0.013	
Contaminacion_aire_2	-0.0577	0.037	-1.546	0.122	-0.131	0.015	
Muertes_agua_2	0.2609	0.049	5.354	0.000	0.165	0.356	
Exp_plomo_2	0.9095	0.043	21.178	0.000	0.817	0.984	
Pesticidas_log	0.0653	0.017	3.841	0.000	0.032	0.059	

Error Cuadrático Medio (RMSE): 47.63369176286732
Error Absoluto Medio (MAE): 32.00714819047614

Figura G.7: Modelo Cuassi-poisson.

Figura G.8: Modelo-Binomial Negativo

enfoque garantiza que cada predictor aportara una contribución relevante a la explicación de la variable objetivo, evitando la inclusión de términos innecesarios que puedan introducir ruido o colinealidad.

A continuación, se presenta el ranking final, Figura G.9, de variables según sus p-valores.

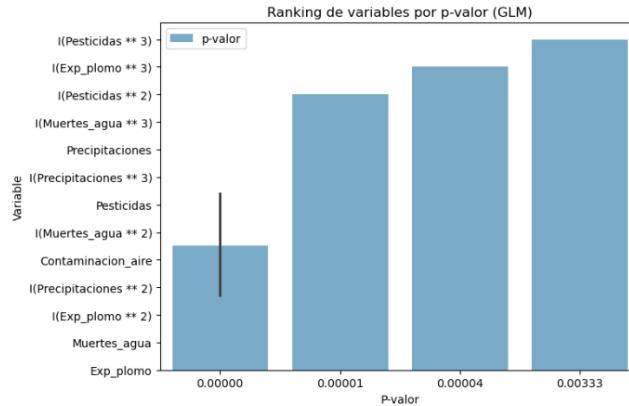


Figura G.9: Importancia de variables Modleo GLM.

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 párkinson). Este modelo permite obtener predicciones precisas sin necesidad de asumir una forma funcional específica entre las variables independientes y la variable objetivo.

- **Importancia de variables:** Para evaluar la contribución de cada variable al modelo, se utilizó la importancia basada en permutación, que mide la disminución en la precisión del modelo al permutar aleatoriamente los valores de cada variable.

La Figura G.10 presenta la importancia de las variables según el modelo.

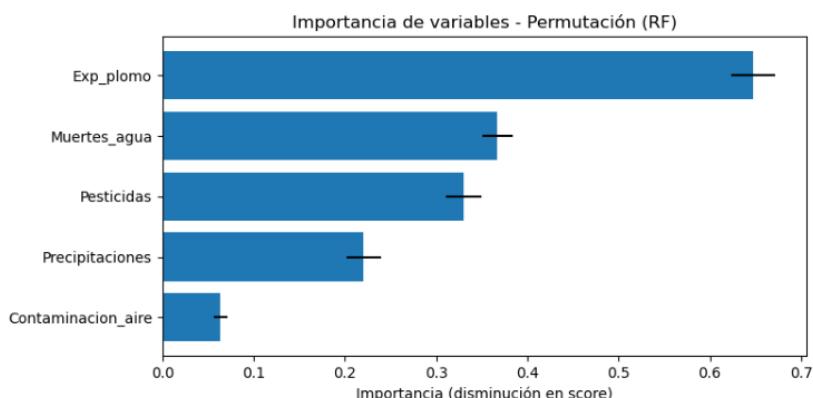


Figura G.10: Importancia de variables Modleo RF.

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 párkinson, 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 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.

- **Importancia de variables:** Para evaluar la importancia de las variables en el modelo XGBoost, se empleó el mismo método de permutación descrito en el apartado de Random Forest. Esto permitió comparar directamente la contribución relativa de cada predictor en ambos modelos. La Figura G.11 muestra los resultados obtenidos.

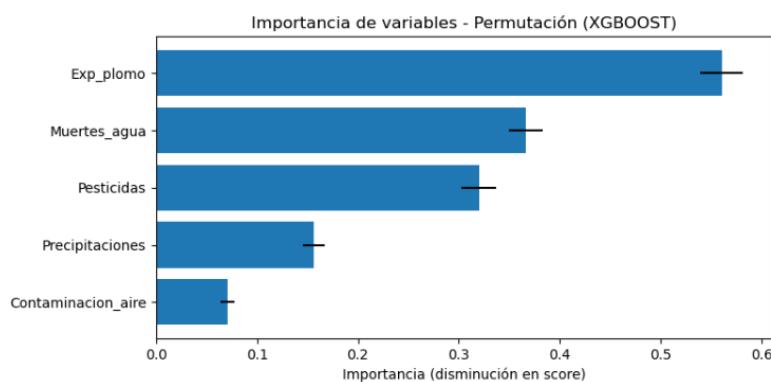


Figura G.11: Importancia de variables Modleo XG.

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 párkinson con buenos resultados en cuanto a precisión y generalización, empleando la implementación del modelo disponible en `scikit-learn`.

- **Importancia de variables:** Al igual que en los modelos Random Forest y XGBoost, la importancia de las variables para el SVR se evaluó mediante el método de permutación descrito anteriormente. Esto permite comparar de manera homogénea la relevancia de cada predictor en los distintos modelos utilizados. Los resultados se presentan en la Figura G.12.

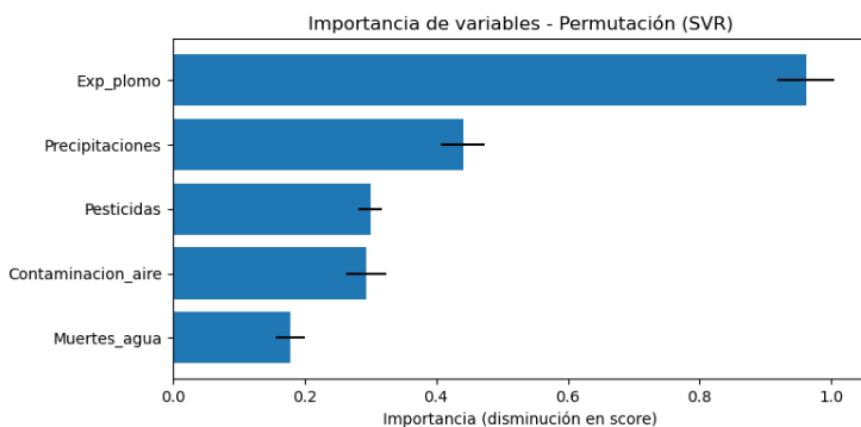


Figura G.12: Importancia de variables Modleo SVR.

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.

- **Importancia de variables:** Al igual que en los modelos Random Forest, XGBoost y SVR, la importancia de las variables(Figura G.13) en KNN se evaluó mediante el método de permutación, siguiendo el mismo criterio explicado anteriormente. Esto permite una comparación homogénea entre modelos.

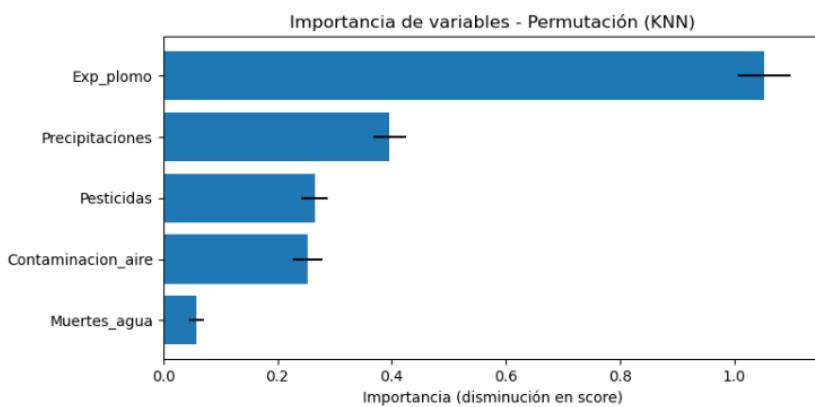


Figura G.13: Importancia de variables Modleo KNN.

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.

- **Importancia de variables:** De la misma manera que hemos determinado la importancia de las variables en modelos anteriores, se ha empleado la técnica de la importancia por permutación. Los resultados se pueden visualizar en la Figura G.14

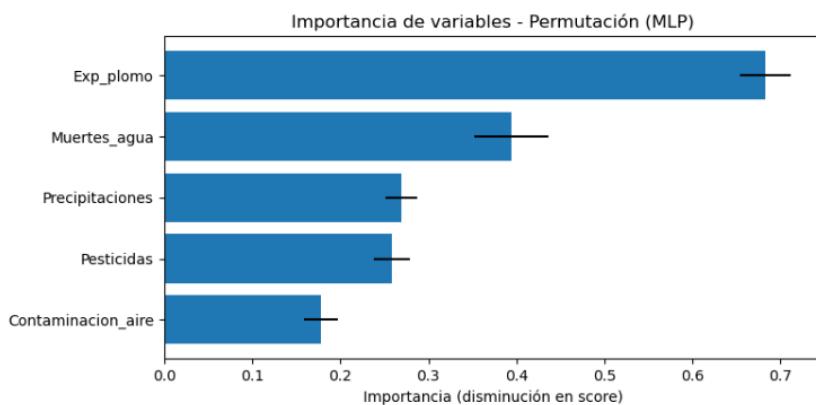


Figura G.14: Importancia de variables Modlelo MLP.

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.

Para la selección de los hiperparámetros más adecuados se empleó la técnica de *búsqueda aleatoria* (`RandomizedSearchCV`) con validación cruzada de 5 (*5-fold cross-validation*). Este enfoque permite explorar de manera eficiente un amplio espacio de combinaciones de parámetros, evaluando únicamente un subconjunto aleatorio (en este caso, 50 combinaciones) para reducir el coste computacional. La métrica utilizada para evaluar el rendimiento durante esta búsqueda fue el *error cuadrático medio negativo* (`neg_mean_squared_error`), y se seleccionó el conjunto de hiperparámetros que obtuvo el mejor desempeño promedio durante la validación cruzada.

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.

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

- **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 caracte-

rísticas 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. Para garantizar la comparabilidad con el modelo Random Forest, se aplicaron las mismas condiciones experimentales, incluyendo el uso de `RandomizedSearchCV` con 50 iteraciones, validación cruzada de 5 pliegues y la métrica `neg_mean_squared_error`.

La división de los datos se realizó en un 80 % para entrenamiento y un 20 % para prueba. Además, 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 del modelo y reducir la varianza.

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

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

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 se encuentran en la [G.5](#)

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

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

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:

- **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 de 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 error lo máximo posible.

En conclusión, 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
hidden_layer_sizes	(256, 128)
activation	relu
max_iter	10000
alpha	0.01
random_state	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

En esta sección se evalúa el rendimiento de cada uno de los modelos predictivos empleados mediante tres métricas comunes en regresión: el error absoluto medio (MAE), la raíz del error cuadrático medio (RMSE), y el coeficiente de determinación (R^2).

El **MAE (Error Absoluto Medio)** mide, en promedio, la magnitud de los errores en las predicciones, expresándolos en las mismas unidades que la variable objetivo. Por ejemplo, si la variable a predecir es una cantidad de casos o una concentración, el MAE indica cuánto se desvía en promedio la predicción respecto al valor real. Un MAE bajo implica que las predicciones están generalmente muy cercanas a los valores observados, mientras que un MAE alto indica que, en promedio, las diferencias son mayores.

Estas métricas se han elegido por su utilidad práctica: el MAE proporciona una medida intuitiva del error medio, el RMSE penaliza más los errores grandes, y el R^2 permite estimar la proporción de varianza explicada por el modelo. No obstante, valores de R^2 cercanos a 1 pueden indicar sobreajuste, especialmente si no se valida en datos independientes.

A continuación, se presentan los resultados para cada modelo de forma individual.

G.3.1. GLM (Modelo Lineal Generalizado)

El modelo GLM presenta un MAE relativamente bajo de 15.45, lo que indica que, en promedio, las predicciones se desvían de los valores reales en unas 15 unidades, mostrando un buen desempeño en términos de error absoluto medio.

Sin embargo, el RMSE es considerablemente alto (86.37), lo que evidencia que existen errores puntuales importantes; es decir, algunos errores de predicción son mucho más grandes y están penalizados de forma más fuerte por el RMSE, lo que sugiere que el modelo puede tener dificultades para manejar valores extremos o casos atípicos.

El bajo coeficiente de determinación R^2 (0.2942) indica que el modelo solo explica alrededor del 29 % de la variabilidad total en los datos, reflejando una capacidad limitada para capturar las relaciones complejas del fenómeno estudiado.

En conjunto, el GLM ofrece una aproximación básica con errores medios razonables, pero muestra limitaciones claras para ajustarse adecuadamente a los datos más complejos y dispersos (ver Tabla G.7).

Métrica	Valor
MAE	15.45
RMSE	86.37
R^2	0.2942

Tabla G.7: Resultados del modelo GLM

G.3.2. Random Forest

El modelo Random Forest mejora notablemente la capacidad predictiva en comparación con el GLM, mostrando un coeficiente de determinación R^2 de 0.6699, lo que significa que explica cerca del 67 % de la variabilidad observada en los datos. Este valor indica un buen ajuste general del modelo.

El RMSE, con un valor de 59.06, refleja una reducción significativa en comparación con modelos más simples, sugiriendo que los errores grandes son menos frecuentes o menos extremos.

Sin embargo, el MAE es relativamente alto, en 40.93 unidades, lo que implica que, en promedio, las predicciones se desvían de los valores reales en aproximadamente 41 unidades. Esta discrepancia puede estar relacionada con la capacidad del modelo para manejar errores dispersos, posiblemente

debido a la naturaleza de los datos o a la estructura del modelo, que puede suavizar las predicciones en algunas regiones.

En conjunto, Random Forest ofrece un equilibrio sólido entre precisión y generalización, haciéndolo adecuado para esta tarea de predicción. (ver Tabla G.8).

Métrica	Valor
MAE	40.93
RMSE	59.06
R^2	0.6699

Tabla G.8: Resultados del modelo Random Forest

G.3.3. XGBoost

El modelo XGBoost presenta un MAE de 41.84, lo que indica que sus predicciones se desvían en promedio unas 42 unidades de los valores reales. Este error absoluto medio es comparable al de Random Forest, mostrando un rendimiento consistente en términos de precisión promedio.

El RMSE es de 61.41, ligeramente mayor que el de Random Forest, lo que señala que existen algunos errores grandes que afectan el ajuste general del modelo, aunque no de manera excesiva.

Con un coeficiente de determinación R^2 de 0.6431, XGBoost logra explicar aproximadamente el 64 % de la variabilidad en los datos, demostrando una capacidad explicativa sólida aunque algo inferior a la de modelos como SVR o Random Forest.

En resumen, XGBoost mantiene un buen equilibrio entre precisión y capacidad explicativa, posicionándose como un modelo robusto y confiable para esta tarea. (ver Tabla G.9).

Métrica	Valor
MAE	41.84
RMSE	61.41
R^2	0.6431

Tabla G.9: Resultados del modelo XGBoost

G.3.4. SVR (Máquinas de Vectores de Soporte)

El modelo SVR presenta un rendimiento superior en comparación con modelos previos, reflejado en un MAE de 32.71, lo que indica que en promedio sus predicciones se desvían por alrededor de 33 unidades respecto a los valores reales. Este menor error absoluto respecto a modelos como kNN o Random Forest implica una mayor precisión.

El RMSE es de 48.04, lo que sugiere que los errores grandes son menos frecuentes o menos severos que en otros modelos, contribuyendo a un mejor ajuste global.

El valor del coeficiente de determinación R^2 es 0.7816, lo que significa que aproximadamente el 78 % de la variabilidad en los datos es explicada por el modelo. Este valor es lo suficientemente alto para indicar un buen ajuste, pero sin alcanzar niveles tan elevados que puedan sugerir sobreajuste.

En conjunto, SVR ofrece un equilibrio efectivo entre la reducción de errores y la capacidad explicativa, posicionándose como uno de los modelos más sólidos en este análisis. (ver Tabla G.10).

Métrica	Valor
MAE	32.71
RMSE	48.04
R^2	0.7816

Tabla G.10: Resultados del modelo SVR

G.3.5. kNN (K-Vecinos más Cercanos)

El modelo kNN presenta un desempeño sólido y consistente, con un MAE de aproximadamente 40.72, lo que indica que, en promedio, las predicciones se desvían en unas 40 unidades de los valores reales. Este error absoluto medio es comparable al de otros modelos como Random Forest y XGBoost.

El RMSE, que penaliza más los errores grandes, se sitúa en 58.94, mostrando que aunque la mayoría de las predicciones están razonablemente cerca, existen algunos errores más significativos que afectan el ajuste general.

El coeficiente de determinación R^2 de 0.6713 indica que el modelo explica cerca del 67 % de la variabilidad observada, lo que señala una capacidad de predicción decente pero no sobresaliente.

En conjunto, kNN ofrece un rendimiento estable que es competitivo frente a otros métodos de aprendizaje automático, sin destacar especialmente

en ninguna métrica, pero manteniendo un buen equilibrio entre error y capacidad explicativa. (ver Tabla G.11).

Métrica	Valor
MAE	40.72
RMSE	58.94
R^2	0.6713

Tabla G.11: Resultados del modelo kNN

G.3.6. MLP (Perceptrón Multicapa)

El modelo MLP muestra un desempeño excepcional, con un MAE muy bajo, lo que indica que en promedio se equivoca por menos de 6 unidades en las predicciones, lo cual es un error muy pequeño respecto a la escala de los datos. Además, el RMSE también es bajo, lo que sugiere que no solo los errores promedio son bajos, sino que además no existen errores grandes o muy puntuales que puedan afectar el rendimiento general del modelo.

El valor de R^2 cercano a 1 indica que el modelo explica prácticamente toda la variabilidad observada en los datos de entrenamiento. Esto significa que las predicciones se ajustan muy bien a los valores reales.

Sin embargo, un desempeño tan elevado también puede indicar que el modelo está aprendiendo demasiado bien los datos de entrenamiento, un fenómeno conocido como sobreajuste (overfitting). Esto implica que, aunque el modelo funciona muy bien con los datos usados para entrenarlo, podría no generalizar igual de bien cuando se enfrente a datos nuevos o no vistos anteriormente.(Ver Tabla G.12).

Métrica	Valor
MAE	5.96
RMSE	10.15
R^2	0.9902

Tabla G.12: Resultados del modelo MLP

G.3.7. Resumen comparativo

En conclusión, el MLP presenta el mejor rendimiento en términos de error y explicación de la variabilidad, aunque su elevado R^2 podría reflejar sobreajuste. Por otra parte, SVR ofrece un buen equilibrio entre precisión

Modelo	MAE	RMSE	R^2
GLM	15.45	86.37	0.2942
Random Forest	40.93	59.06	0.6699
XGBoost	41.84	61.41	0.6431
SVR	32.71	48.04	0.7816
kNN	40.72	58.94	0.6713
MLP	5.96	10.15	0.9902

Tabla G.13: Comparativa global entre modelos

y capacidad explicativa sin indicios claros de sobreajuste. Modelos como Random Forest, XGBoost y kNN muestran rendimientos similares, adecuados pero con errores promedio mayores. El GLM, aunque básico, ayuda a entender las limitaciones de un modelo lineal simple para este problema.

Para obtener predicciones más robustas y generalizables, se considerará la integración o combinación de modelos, aprovechando las fortalezas individuales y minimizando las debilidades, especialmente para evitar los riesgos de sobreajuste del MLP.

Apéndice H

Anexo de sostenibilización curricular

H.1. Introducción

En este anexo se incluye una reflexión sobre la sostenibilidad desde el punto de vista de Ingeniería de la Salud, con un enfoque en la utilización de herramientas de aprendizaje automático (*machine learning*), para poder llevar a cabo un análisis y predicción de factores ambientales que puede llegar a ser influyentes en el desarrollo de la enfermedad de parkinson. Para el desarrollo de este trabajo he aplicado competencias que se encuentran relacionadas con la integración de la tecnología, así como de la salud, a través de datos ambientales y de la elaboración de modelos predictivos con el objetivo de una mejor gestión preventiva.

H.2. Reflexión sobre salud, tecnología y sostenibilidad

La sostenibilidad en el ámbito sanitario conlleva tanto a la atención de forma efectiva de las enfermedades como a la prevención y el uso eficiente de los recursos tecnológicos y científicos para reducir impactos negativos en la sociedad y en el medio ambiente. El empleo de aprendizaje automático en la Ingeniería de la Salud permite analizar patrones y factores de riesgo que nos ayuden a anticipar problemas de salud.

Haber trabajado con datos ambientales para poder estudiar qué relación tienen con enfermedades como el parkinson, me ha hecho reflexionar sobre el

impacto que tiene el entorno en la salud. La utilización de modelos predictivos nos permite tomar decisiones mucho más informadas y precisas tanto para la gestión sanitaria como para el diseño de políticas públicas. Por otro lado, el acceso a información en tiempo real y su análisis nos va a permitir una respuesta mucho más rápida frente a cambios ambientales.

H.3. Competencias desarrolladas

Durante el proyecto, he desarrollado habilidades clave relacionadas con el análisis tecnológico y sanitario, entre ellas:

- **Obtención y análisis de datos ambientales y sanitarios:** Uso de APIs públicas para obtener datos y preparación de conjuntos de datos limpios y estructurados, necesario para los análisis posteriores.
- **Diseño e implementación de modelos de aprendizaje automático:** Selección y entrenamiento de modelos como regresiones, árboles de decisión y redes neuronales para anticipar la influencia de variables ambientales en el desarrollo de parkinson, convirtiendo datos complejos en información útil.
- **Interpretación crítica y comunicación de resultados:** Análisis de los resultados con una perspectiva clínica y social, valorando su aplicabilidad en la toma de decisiones y en la planificación de intervenciones sanitarias.
- **Ética y responsabilidad en el manejo de datos:** Conciencia sobre el uso adecuado y respetuoso de datos sensibles, promoviendo la confianza y la integridad en las tecnologías de salud.

H.4. Aplicación práctica en el TFG

El proyecto ha consistido en la recolección de datos ambientales mediante APIs, seguida del análisis y creación de un dataset para entrenar varios modelos de aprendizaje automático. El objetivo principal fue predecir la influencia de factores ambientales en la aparición de parkinson, ofreciendo una herramienta que puede apoyar la vigilancia epidemiológica y la formulación de políticas preventivas.

La integración de tecnología avanzada con conocimiento sanitario permite una gestión innovadora y eficiente en salud pública, optimizando recursos

y contribuyendo a mitigar el impacto social y económico de enfermedades neurodegenerativas.

H.5. Conclusión

En definitiva, el TFG ha permitido combinar competencias técnicas con una visión sanitaria orientada a mejorar la prevención y gestión de enfermedades relacionadas con el entorno. El uso innovador de herramientas digitales promueve sistemas de salud más eficientes, proactivos y sensibles al bienestar humano y ambiental, fomentando el trabajo interdisciplinar y el manejo ético de los datos como bases para un futuro saludable y equilibrado.

En este contexto, la sostenibilidad no solo implica cuidar el entorno físico, sino también garantizar que las soluciones tecnológicas desarrolladas sean escalables, éticas y centradas en las personas. Desde mi formación en Ingeniería de la Salud, considero fundamental impulsar una cultura profesional orientada a la innovación con impacto social, en la que la prevención de enfermedades se aborde desde múltiples dimensiones: tecnológica, ambiental, sanitaria y ética. Esta experiencia ha reforzado mi convicción de que la colaboración entre disciplinas es clave para construir un sistema de salud más resiliente, justo y preparado para los desafíos del futuro.

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.
- [Eguía, 2024] Eguía (2024). ¿qué es una licencia de software? tipos y un ejemplo práctico. Accedido: 2025-06-11.
- [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.
- [Pastoriza, 2024] Pastoriza, D. R. (2024). ¿cuánto gana un data scientist en españa en 2024? Hack a Boss Blog. Consultado el 29 de junio de 2025.

- [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.