# FACULTAD DE ESTUDIOS ESTADÍSTICOS

## GRADO EN ESTADISTICA APLICADA

**Curso 2024 / 2025**

---

## Trabajo de Fin de Grado

***TÍTULO:***
***Modelización predictiva en situación de alarma sanitara***


***Alumno*: Santiago García Morante**

***Tutor*: Fernando Pérez Contreras**


Junio de 2025


UNIVERSIDAD COMPLUTENSE
MADRID

*"Todos los modelos son incorrectos,*
*pero algunos son útiles"*
(Box & Draper, 1987, p.424)

*"En teoría, hay poca diferencia entre teoría*
*y práctica. Pero, en la práctica, hay una*
*gran diferencia"*
(Brewster, 1882)

*"La gente puede elaborar estadísticas para*
*demostrar cualquier cosa; el cuarenta por*
*ciento de la gente lo sabe"*
(Homer Simpson.
Swartzwelder & Moore, 1994)

*"Si no puedes convencerlos, confúndelos"*
(Les Luthiers, 2008)

## A modo de introducción no paramétrica

Estas cuatro citas ilustran, desde diferentes perspectivas, la complejidad de la relación que existe entre los modelos estadísticos, la realidad y nuestra capacidad para su interpretación. Es en ese punto en el que pretende situarse este trabajo: el uso de modelos estadísticos sencillos pero robustos, construidos a partir de datos accesibles, que permitan el acercamiento a una determinada realidad. Sin duda, existirán limitaciones, pero el objetivo no es otro que construir herramientas útiles que, aunque imperfectas, permitan mejorar la toma de decisiones.

**Índice**

## Índice de figuras



## Índice de tablas

# RESUMEN

Con el fin de poder realizar una asignación eficiente de los recursos sanitarios disponibles, y a partir de datos publicados por el Ayuntamiento de Madrid, se ha construido un modelo predictivo capaz de permitir descartar con confianza a los individuos con menor vulnerabilidad en un contexto de emergencia sanitaria similar a la pandemia de COVID-19.

A nivel metodológico, los resultados obtenidos en el uso de técnicas como la regresión logística binaria o el bagging han permitido alcanzar un modelo parsimonioso basado en la edad, otras patologías previas y el nivel de exposición a determinados contaminantes atmosféricos, con una especificidad superior al 89%, un AUC aceptable y sin signos de sobreajuste.

La variable edad presentó un efecto protector contraintuitivo, por lo que se amplió el estudio con ANOVA y diferentes pruebas no paramétricas como el coeficiente de correlación de Spearman, el test U de Mann-Whitney o el de la Chi-cuadrado. Esta exploración complementaria confirmó la existencia de una relación no lineal que podría ser consecuencia de sesgos muestrales.

Por supuesto, el estudio no pretende ofrecer un diagnóstico clínico sino proponer una herramienta complementaria que permita mejorar la toma de decisiones en políticas de salud pública en situaciones de escasez de recursos sanitarios. El trabajo muestra la integración entre técnicas clásicas y modernas de la estadística aplicada lo que, además de proponer un modelo con utilidad práctica, ilustra la evolución de la estadística en la era del aprendizaje automático.



Palabras clave: modelo predictivo, especificidad, enfermedades respiratorias, emergencia sanitaria, salud pública, regresión logística, bagging.

## ABSTRACT

With the purpose of being able to accomplish an efficient assignment of the available health resources, and parting from the data published by the Ayuntamiento de Madrid, a predictive model has been designed to be able to confidently discard the individuals with a lesser vulnerability in a sanitary emergency context similar to the COVID-19 pandemic.

On a methodological level, the gathered results in the use of techniques such as the logistic binary regression or bagging, have allowed the accomplishment of a frugal model based on the age, other previous pathologies and the level of exposure to certain atmospheric pollutes, with a specificity over 89%, an acceptable AUC and with no signs of over adjustment.

The variable age presented a protective anti-intuitive effect, thus the study was amplified with ANOVA and different non parametric tests such as the Spearman correlation coefficient, Mann-Whitney's U test or Chi-square. This complementary exploration confirmed the existence of a non linear relation that could be a consequence of sampling biases.

Of course, the study does not pretend to offer a clinical diagnosis but suggests a complementary tool that can improve the political decision making in public health care in the light of limited health resources. The study shows the integration of classic and modern techniques of applied statistics which, in addition to suggesting a model with a practical utility, illustrates the evolution of statistics in the automatic learning era.



Key words: Predictive model, specificity, respiratory diseases, sanitary emergency, public health, logistic regression, bagging.

# 1 Introducción

La optimización en la asignación de los recursos sanitarios ha sido un asunto muy tratado en las últimas décadas en el ámbito de los organismos con competencias en materias de salud. La crisis sanitaria derivada de la pandemia de COVID-19 ha hecho crecer la relevancia de esta cuestión en las políticas públicas. La imagen de los hospitales colapsados y del personal sanitaria completamente desbordado permanecerá para siempre en la memoria colectiva. La pandemia no solo fue una emergencia sanitaria mundial de máximo nivel, también se convirtió en una importante crisis de gestión de la información y de los recursos disponibles.

En aquellos momentos, a cualquiera con cierta inquietud por los datos y su tratamiento le surgía una pregunta que, en el fondo, es bastante elemental: ¿cómo podemos saber, rápidamente y con datos simples y de fácil acceso, con quién se deben priorizar los recursos existentes? Encontrar una respuesta a esa pregunta encontrando un modelo predictivo sencillo, basado en datos fácilmente obtenibles que, en plena emergencia, permitiese descartar a las personas cuyo riesgo de enfermedad respiratoria grave fuese mínimo, hubiese supuesto una distribución más eficiente los escasos recursos disponibles.

Por otro lado, y volviendo a la actualidad, cabe señalar que prestigiosos organismos internacionales han puesto de manifiesto que algunos factores como la calidad del aire, las condiciones de vida o el nivel socioeconómico, pueden tener un impacto especialmente significativo en la prevalencia de determinadas enfermedades respiratorias crónicas (World Health Organization, 2021). De manera que existen personas con mayor riesgo de contraer una enfermedad de tipo respiratorio que otras y que esta situación podría estar determinada por un puñado de variables.

En España, en el año 2021, las enfermedades respiratorias supusieron la tercera causa de muerte y hospitalización, con un incremento en el número de fallecimientos del 20,9% respecto al año anterior (SEFAC, 2023). Algunos autores como Marmot (2005) y Wilkinson & Pickett (2009) sostienen que las desigualdades sociales y económicas pueden llegar a resultar muy influyentes en la incidencia de determinadas enfermedades respiratorias. Concretamente, diversos estudios epidemiológicos han

concluido que existe una relación directa entre la exposición a contaminantes atmosféricos, como los denominados conjuntamente Óxidos de Nitrógeno (Monóxido y Dióxido de Nitrógeno), y el desarrollo de asma, bronquitis crónica o EPOC. (Brunekreef & Holgate, 2002; Dominici et al. 2006).

En este contexto, parece bastante evidente la utilidad de encontrar modelos predictivos que, con el uso de variables fácilmente observables y, antes de la aparición de síntomas y de la realización de costosas pruebas diagnósticas, ayuden a identificar poblaciones de riesgo. Especialmente en el marco de una crisis sanitaria que pudiera volver a desbordar el uso de servicios sanitarios.

Además, se dan dos circunstancias especialmente relevantes que podemos enlazar con esta necesidad: por un lado, el marco legal obliga a las administraciones públicas a ofrecer datos abiertos de todo tipo, incluidos de salud y socioeconómicos para su reutilización (Parlamento Europeo y Consejo, 2019). Por otro lado, contamos con modelos estadísticos como la regresión logística binaria o el bagging que, aplicados sobre datos de salud o socioeconómicos, permiten la construcción de herramientas predictivas de bajo coste que son capaces de complementar la práctica clínica. Trabajos como los de Sun et al. (2021) y Zou et al. (2019) han probado que estos modelos pueden alcanzar niveles óptimos de precisión utilizando este tipo de variables antes de recurrir a pruebas médicas invasivas, costosas y limitadas.

En las fases iniciales de una emergencia sanitara como la que vivimos como consecuencia del COVID-19, identificar rápidamente a las personas que no presentan riesgo de evolución grave es, probablemente, el enfoque más eficaz. Como resultado de esta estrategia sería posible liberar recursos sanitarios que se pueden dedicar a aquellos individuos que realmente necesitan de ellos por la evolución inminente de la enfermedad. Tal y como señala la OMS (2020), *“la identificación rápida de pacientes que no tienen riesgo de enfermedad grave permite a los sistemas sanitarios concentrar recursos en quienes tienen mayor probabilidad de deterioro”.* De la misma manera, Rosenbaum (2020) destaca que un triaje eficiente *“ha de permitir descartar con confianza a aquellos que no requieren atención urgente”*.

Con estos antecedentes, el presente estudio toma datos de la Encuesta de Salud de Madrid 2021 y Datos de calidad del aire de 2021 (Ayuntamiento de Madrid, 2021) para

elaborar modelos estadísticos de predicción que permitan, a partir de variables sanitarias, sociales, económicas y ambientales, identificar aquellos individuos que tengan una menor probabilidad de verse afectados por enfermedades respiratorias crónicas para poder descartarlos en un triaje preliminar y focalizar el uso de los recursos sanitarios disponibles en el resto de la población, evitando el colapso de los recursos sanitarios.

En ningún caso este trabajo contiene información con valor médico ni de diagnóstico, simplemente plantea una herramienta complementaria de apoyo estadístico orientada a la prevención y a la mejora en la gestión de las políticas de salud pública vinculadas a los casos de emergencia sanitaria.

# 2 Fuentes y preparación de los datos

La principal fuente utilizada en este estudio para la obtención de datos ha sido el Portal de Datos Abiertos del Ayuntamiento de Madrid. De él se han obtenido los microdatos anonimizados de la Encuesta de salud de la ciudad de Madrid 2021, que recoge información sobre hábitos de vida, condiciones socioeconómicas, estado de salud en diferentes áreas y variables demográficas. También se utilizaron los datos horarios de la calidad del aire del año 2021, que contienen mediciones de varios contaminantes atmosféricos recogidas por las estaciones distribuidas por los diferentes distritos de la ciudad.

Inicialmente, se seleccionaron de la Encuesta de salud las variables relacionadas con las enfermedades respiratorias y algunas otras que, en base a la literatura previa, podrían resultar influyentes en su aparición, como la edad, el sexo, la situación laboral o el consumo habitual de sustancias potencialmente adictivas. A partir de los datos de calidad del aire se calculó la exposición media anual por estación de medición a los contaminantes más peligrosos para, posteriormente, asignarle el valor correspondiente a cada individuo en función a la cercanía de la misma con su lugar de residencia.

Para facilitar la interpretación posterior de los resultados y conseguir robustez en los modelos estadísticos, se ha realizado un proceso de agrupación y recodificación de algunas de las variables originales. Este tipo de procesos responde a criterios metodológicos ampliamente aceptados tanto en el análisis estadístico aplicado a las ciencias sociales como en la investigación epidemiológica.

Se ha buscado evitar el sobreajuste en los modelos multivariantes a través de la reducción del número de categorías de las variables cualitativas. Además, esta reducción facilita la comparación entre grupos con muestras suficientemente grandes. Como señalan Hosmer, Lemeshow y Sturdivant (2013), “la categorización de variables cualitativas debe atender al equilibrio entre la fidelidad a la información original y la estabilidad de las estimaciones del modelo” (p. 33).

El detalle de las variables seleccionadas de la Encuesta de salud de la ciudad de Madrid 2021 se muestra en el Anexo I. Algunas variables tuvieron que ser transformadas y recodificadas para adaptarlas a las necesidades del análisis estadístico, tal y como se detalla en el apartado correspondiente. El proceso incluyó la recodificación de las variables categóricas, la creación de variables binarias a partir de otras ya existentes y la reagrupación de categorías que permitiesen facilitar la interpretación posterior y mejorar la robustez de los modelos.

Por otro lado, en casos como la variable “vida saludable” o “consuno de sustancias adictivas” se han creado variables binarias a partir de indicadores múltiples. Esto está justificado cuando es necesario construir indicadores compuestos que sean capaces de captar dimensiones relevantes de una manera sintética y coherente con la literatura previa. Este tipo de transformación está recomendada en estudios observacionales donde el objetivo es identificar factores de riesgo mediante técnicas de clasificación o regresión (Tabachnick & Fidell, 2019).

Asimismo, la reagrupación de los diferentes niveles educativos o tramos de ingresos en categorías de tipo ordinal más amplias no solo permite cumplir con algunos requisitos estadísticos, como el tamaño muestral por celda en tablas de contingencia, sino que, además, facilita la interpretación en términos de desigualdad socioeconómica, como es habitual en estudios sobre determinantes sociales de la salud (Marmot, 2005, Solar & Irwin, 2010).

Además, el uso del valor medio anual por distrito en el caso de las variables relacionadas con la exposición a contaminantes atmosféricos responde a criterios estandarizados en estudios de salud ambiental, con el fin de asegurar la comparabilidad espacial y temporal de los datos (WHO, 2021; EEA, 2023).

De los datos de la calidad del aire recogidos se seleccionaron el monóxido de nitrógeno (NO) y el dióxido de nitrógeno ($NO_2$), a los que nos referimos conjuntamente como los óxidos de nitrógeno ($NO_x$), ya que forman el grupo de contaminantes atmosféricos de mayor importancia en los entornos urbanos. Ambos son consecuencia, especialmente, del resultado de los procesos de combustión del tráfico rodado. La Organización Mundial de la Salud (2021) y la Agencia Europea de Medio Ambiente (2023) señalan al $NO_2$ como responsable de efectos adversos en la función

pulmonar, sobre todo en poblaciones vulnerables. Por su parte, el NO es el precursor del $NO_2$, ya que surge de su oxidación. Además, diversos estudios epidemiológicos han demostrado la asociación entre una exposición prolongada a $NO_x$ y un aumento en la incidencia de enfermedades respiratorias crónicas (Khomenko, et al. 2021).

A continuación, se describen las variables finales utilizadas para el análisis estadístico.

- *Edad*: Variable cuantitativa continua. Representa la edad del individuo expresada en años cumplidos en el momento de contestar a la encuesta. Se han considerado valores iguales o superiores a 15.
- *Peso*: Variable cuantitativa continua. Representa el peso corporal del individuo en kilos en el momento de contestar a la encuesta.
- *Altura*: Variable cuantitativa continua. Representa la altura del individuo, medida en centímetros, en el momento de contestar a la encuesta.
- *Sexo*: Variable cualitativa nominal. Indica el sexo con el que el individuo se identifica en el momento de contestar a la encuesta.
- *Distrito*: Variable cualitativa nominal. Indica el distrito municipal de residencia del individuo en el momento de contestar a la encuesta. Son considerados los 21 distritos establecidos por el Ayuntamiento de Madrid. Esta variable nos permite relacionar al individuo encuestado con el nivel de exposición a contaminantes.
- *Nivel de estudios*: Variable cualitativa ordinal. Indica el nivel educativo más alto finalizado por el individuo en el momento de responder a la encuesta. Los posibles valores de han reagrupado desde la variable original resultando: “Sin estudios”, “Estudios primarios”, “Estudios secundarios”, “Bachillerato superior o equivalente” y “Estudios universitarios”.
- *Metros de vivienda*: Variable cuantitativa continua. Indica la superficie total de la vivienda habitual del individuo. Se expresa en metros cuadrados.
- *Ingresos netos*: Variable cualitativa ordinal. Indica la suma de ingresos mensuales del hogar al que pertenece el individuo en el momento de realizar la encuesta después de impuestos y cotizaciones a la seguridad social. Se expresa en euros. Se ha reagrupado desde la variable original, resultando los siguientes posibles valores: “Menos de 1.100”, “Entre 1.100 y 1.650 euros”,

"Entre 1.650 y 2.300 euros", "Entre 2.300 y 3.800 euros" y "Más de 3.800 euros".

- *Vida saludable*: Variable dicotómica. Indica si el individuo presenta, en el momento de la encuesta, una vida saludable. Los criterios para considerar una vida saludable o no son la combinación de diferentes variables relacionadas con el tipo de alimentación (consumo habitual de frutas, verduras, refrescos azucarados o comida rápida, por ejemplo), los hábitos de descanso y de actividad física. Los posibles valores que toma la variable son "Sí" o "No".
- *Drogas*: Variable dicotómica. Indica si el individuo, en el momento de la encuesta, era consumidor de sustancias potencialmente adictivas. Se han considerado a las personas que han reportado tabaquismo o un uso continuado de cigarrillos electrónicos o vapeadores, consumo de alcohol excesivo, uso regular de ansiolíticos, antidepresivos u otros fármacos con potencial adictivo o consumo de marihuana u otras sustancias con carácter recreativo.
- $ON_x$: Variable cuantitativa continua. Representa el nivel de exposición ambiental a óxidos de nitrógeno (NO y $NO_2$) al que está sometido el individuo en el momento de la encuesta. Se expresa en microgramos por metro cúbico de aire. Se ha considerado el valor medio del año de referencia a partir de los datos diarios de cada una de las estaciones medidoras. En los distritos en los que existen diferentes estaciones de medida, se ha considerado el valor medio de todas ellas. En los distritos de Salamanca, Usera, Vicálvaro y San Blas-Canillejas, que no cuentan con estación medidora propia, se han utilizado las mediciones de Retiro, Carabanchel, Villa de Vallecas y Ciudad Lineal, respectivamente. En todos los casos la distancia respecto a la estación medidora más cercana ha sido inferior a 4 kilómetros.
- *Enfermedad cardio*: Variable dicotómica. Indica si el individuo, en el momento de contestar a la encuesta, ha sido diagnosticado de alguna enfermedad del aparato cardiovascular. Las enfermedades consideradas para esta variable han sido: Infarto, tensión alta crónica, colesterol alto crónico, varices y diabetes. Toma valores "Sí" o "No"

- *Enfermedad muscular*: Variable dicotómica. Indica si el individuo, en el momento de contestar a la encuesta, ha sido diagnosticado de alguna enfermedad o trastorno crónico de origen musculoesquelético. Las enfermedades consideradas han sido: artrosis, dolor muscular crónico, dolor cervical crónico, dolor lumbar crónico. Toma valores “Sí” o “No”
- *Enfermedad neuro*: Variable dicotómica. Indica si el individuo, en el momento de contestar a la encuesta, ha sido diagnosticado de alguna enfermedad de origen neuropsiquiátrica. Las enfermedades consideradas en esta variable han sido: depresión, ansiedad, trastorno bipolar, esquizofrenia, epilepsia, demencia, Parkinson u otras afecciones propias del sistema nervioso central o de la salud mental. Toma valores “Sí” o “No”
- *Enfermedad respi*: Variable dicotómica. Indica si el individuo, en el momento de contestar a la encuesta, ha sido diagnosticado de alguna enfermedad del aparato respiratorio. Las enfermedades consideradas han sido: Asma, bronquitis crónica, enfisema pulmonar, EPOC, covid persistente y otras patologías respiratorias. Toma valores “Sí” o “No”

# 3 Objetivos y metodología

Establecer unos objetivos claros y una metodología adecuada y funcional resulta fundamental a la hora de poner en marcha un estudio como este. Solo se puede alcanzar eficientemente el lugar al que se quiere llegar si se establece correctamente el punto de inicio y se definen claramente cada uno de los pasos que se han de ir dando. En este caso, el lugar de partida es la Encuesta de salud 2021 y los datos de calidad del aire del mismo año referidos a la ciudad de Madrid. Cada uno de los pasos son las herramientas estadísticas con las que alcanzaremos las pretendidas conclusiones. Por supuesto, no sin antes detenernos en solventar los problemas que la falta de respuesta nos pueda causar, intentando minimizar la pérdida de información.

El objetivo principal de este trabajo es desarrollar herramientas que sean útiles en la toma de decisiones respecto a las políticas de salud pública, especialmente en el ámbito de una emergencia sanitaria, a través de la construcción de modelos predictivos que permitan descartar a aquellos individuos que no están en riesgo de padecer una enfermedad respiratoria basándonos, únicamente, en información accesible.

Para alcanzar el objetivo principal, ha sido necesario ir superando otros objetivos parciales previos que han permitido dar una sólida estructura a cada una de las fases del análisis y dotar de un soporte robusto a los modelos finales. Algunos de estos objetivos parciales son los siguientes:

- Comprensión integral de la base de datos de partida a través de un amplio análisis exploratorio de variables y datos.
- Transformación de las variables iniciales y las posibles respuestas para adaptarlas a las necesidades de los análisis y que, de esta manera, capturen de una forma más clara las relaciones menos explícitas en los datos originales.
- Depuración de los datos transformados para actuar ante la falta de respuesta de determinadas variables y así conseguir eliminar cuando ha sido posible o, al menos, minimizar la pérdida de información.

- División de la base de datos definitiva y depurada en grupos de entrenamiento y prueba para evaluar los modelos obtenidos a través de métodos de validación cruzada.
- Uso de diferentes técnicas estadísticas para el análisis de los datos y, a partir de ellas, la construcción de los modelos predictivos.
- Comparación del rendimiento de los diferentes modelos predictivos aplicados al mismo conjunto de datos.
- Interpretación de los resultados obtenidos.

Las técnicas estadísticas que han sido utilizadas en todo este proceso han sido las que a continuación se enumeran:

- Estadística descriptiva
- KNN
- Validación cruzada
- Regresión logística binaria
- ANOVA
- Transformación Box-Cox
- Coeficiente de correlación de Spearman
- Test de la Chi-cuadrado
- Test U de Mann-Whitney
- Árboles de decisión
- Bagging

# 4 Imputación

A menudo, los estudios empíricos presentan datos perdidos. Dado que los datos que se han utilizado para este trabajo han sido recogidos de un estudio riguroso realizado por una administración pública local, estos son muy pocos en porcentaje y se deben, casi exclusivamente, a la falta de respuesta en variables especialmente sensibles sobre asuntos económicos y/o sociales. Reducir el tamaño muestral eliminando los casos en los que no existe respuesta hubiese aumentado significativamente la posibilidad de introducir sesgos, por lo que se ha optado por la imputación (Little y Rubin, 2019). Por otro lado, la reducción significativa del tamaño de la muestra podría suponer una reducción de la potencia estadística y un aumento del error estándar.

La imputación de datos faltantes es una técnica que se utiliza en estadística para sustituir los valores ausentes por estimaciones que se consideran razonables. La aplicación de técnicas de imputación está respaldada por numerosos autores que justifican su uso en el análisis estadístico y en el desarrollo de modelos predictivos. Básicamente, permite no perder información sin alterar en modo alguno la estructura subyacente de los datos además de, en muchos casos, reducir el sesgo. Todo ello, podría también traducirse en una mejora considerable de la interpretabilidad de los modelos, especialmente en los casos en los que la pérdida de información no es aceptable porque podría implicar cambios significativos en los modelos (Rubin, D.B., 1987).

Las variables que han precisado de imputación son las siguientes:

- *Edad*: Esta variable solo tenía un dato perdido. La imputación por la mediana en este caso permite no desvirtuar la tendencia central de los datos ya que estamos ante un valor robusto frente a valores extremos, como sucedería en el caso de individuos demasiado mayores. El impacto en los análisis posteriores, al tratarse de un único caso, será mínimo.
- *Peso, Altura, Nivel de estudios, Metros de vivienda, Ingresos netos*: La técnica de imputación utilizada en el caso de estas variables es la conocida como k-Nearest Neigbors (KNN). Se trata de una técnica compleja que utiliza la información contenida en múltiples variables relacionadas con la variable objeto de imputación para estimar los valores perdidos, basándose, en este

caso, en las 5 observaciones completas más parecidas. De esta manera, no se pierden los patrones y correlaciones existentes en los datos originales. Esto último resulta especialmente importante cuando, como es el caso, los análisis posteriores dependerán, precisamente, de estas relaciones entre variables.

Se ha utilizado tanto para las variables numéricas (salvo edad) y las categóricas y resulta especialmente útil cuando existe relación entre las variables, pero esta no es lineal. Este tipo de imputación se realiza a partir de casos similares, por lo que tiene a determinar valores más representativos y consistentes, lo que evita el sesgo que se introduce a partir de otros tipos de imputación. Por último, destacar que la imputación KNN es especialmente adecuada para los análisis del estudio dada la necesidad de preservar los posibles patrones de los datos para la modelización multivariante (Troyanskaya et al., 2001).

# 5 Marco teórico

## 5.1 K-Nearest Neighbors (KNN)

La técnica KNN es un método de aprendizaje supervisado no paramétrico que se utiliza en el campo de la imputación y consiste en la asignación del valor más común entre las k observaciones más parecidas del conjunto de datos de entrenamiento. Estamos ante un método de los conocidos como “lazy learning” que no construye un modelo general durante la fase de entrenamiento, es decir, ni ajusta parámetros ni genera reglas o fórmulas. Por el contrario, almacena los datos de entrenamiento para realizar una predicción en el momento en que llega una nueva instancia buscando entre los ejemplos almacenados más parecidos.

La manera más usual de calcular los vecinos más cercanos para variables numéricas es a través de la distancia euclídea, que responde a la siguiente expresión:

$$d(p,q\ ) = \sqrt{\sum_{i=1}^{n}(p_i - q_j\ )^2}$$

donde $p, q$ son dos puntos en el espacio n euclidiano, $p_i\ y\ q_j$ son vectores euclidianos a partir del origen del espacio y $n$ es el espacio $n$.

Su aplicación en la imputación de datos faltantes consiste en calcular la distancia de la observación con dato faltante a otras observaciones completas utilizando, únicamente, las variables comunes con datos. Después se seleccionan los k vecinos completos con menor distancia y, finalmente, se imputa el valor faltante por la media o la mediana, en el caso de variables numéricas, y por la moda si las variables son categóricas.

Las principales ventajas de esta técnica son que tiene en cuenta la estructura multivariante de los datos y que, al tratarse de un método no paramétrico, no precisa de presunciones sobre distribuciones. En cambio, el principal problema es las necesidades computacionales que requiere que, en conjuntos de datos muy grandes, puede llegar a hacerla prácticamente inviable. (Jastie et al., 2009)

## 5.2 Validación cruzada

La evaluación de la capacidad predictiva de un modelo es algo que resulta fundamental para que se pueda garantizar el uso del mismo en datos no utilizados para su elaboración. Evaluar un modelo con los datos que se han utilizado para su construcción puede llevar al sobreajuste, es decir, un modelo que funciona muy bien con los datos que se han utilizado para su desarrollo, pero no con otros diferentes. La validación cruzada es una de las técnicas estadísticas que permite evitar este problema.

Consiste en la división aleatoria del conjunto de datos con el que se trabaja en dos subconjuntos. Uno de los subconjuntos será el destinado al entrenamiento del modelo. El otro subconjunto se utilizará para la evaluación posterior del modelo. La proporción de datos de los subconjuntos suelen estar en torno al 80 y 20%. De esta manera, el los datos del subconjunto de entrenamiento serán únicamente utilizados para la estimación de los parámetros del modelo, mientras que los datos del subconjunto de prueba se utilizarán, exclusivamente, para su evaluación.

Esta técnica permite que el modelo se evalúe con datos no utilizados para su construcción, de manera que se podrá estimar el rendimiento del modelo de una forma más objetiva, ya que los datos no conocidos resultan ser una simulación de datos futuros no observados. Esta técnica resulta especialmente útil por su simplicidad, bajo coste computacional y rápida implementación (Kuhn & Johnson, 2013).

## 5.3 Regresión logística binaria: Fundamentos teóricos

El objetivo de una regresión logística binara es estimar la probabilidad con la que ocurre un evento a partir de una o más variables independientes. A la variable dicotómica a predecir se le asignará un valor 0 o 1 en base a su probabilidad. Generalmente se comienza estableciendo la probabilidad de ocurrencia o no en 0,5 pero, habitualmente, se suele cambiar este valor en base a diferentes criterios. A diferencia de la regresión lineal que establece relaciones lineales entre la variable dependiente y las predictoras, en este caso, las relaciones son logísticas. La probabilidad se modela a través de la siguiente expresión:

$$P(Y = 1|X_1, \ldots, X_m) = \frac{e^{(\beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m)}}{1 + e^{(\beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m)}}$$

La función logística permite transformar la combinación lineal de las variables explicativas en una probabilidad. La función sigmoide muestra como las probabilidades tienen a acumularse asintóticamente en 0 y 1, lo que resulta más que adecuado para poder clasificar eventos de carácter binario (Hosmer et al., 2013).

La regresión logística binaria modela el logaritmo de las odds ratio como una función lineal de las variables regresoras, tal y como se ha visto con anterioridad. Por lo tanto, otra manera similar de expresar el modelo es a través de la función logit que realiza una transformación. La expresión es la siguiente:

$$logit(p) = \frac{p}{1 - p} = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m$$

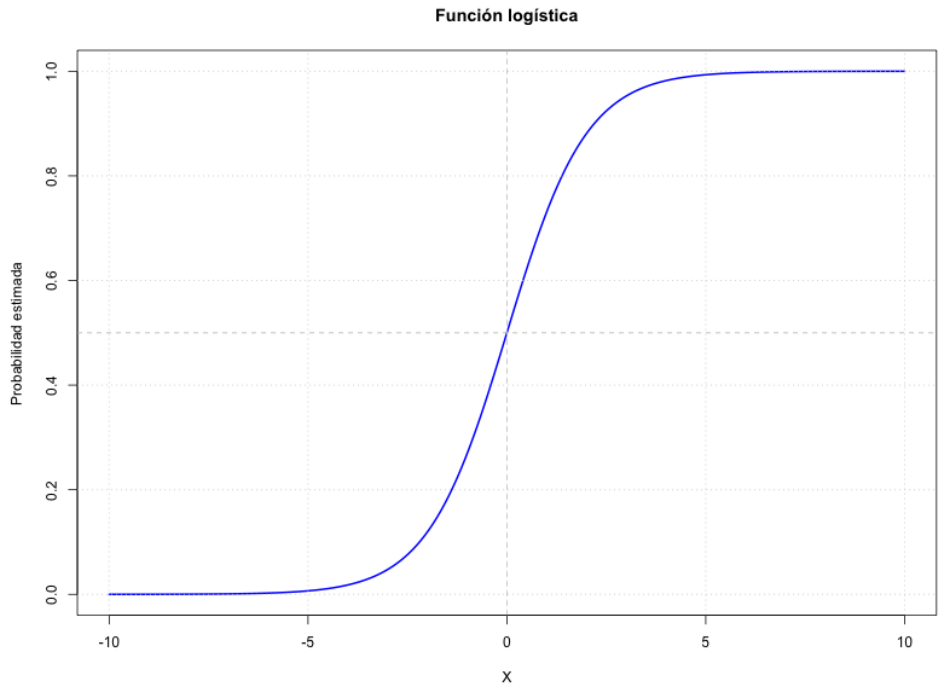


*Figura 1 Representación gráfica de la función logit*

Una de las ventajas más destacadas de este modelo es que permite la interpretación de los coeficientes cuantificando la relación entre la variable respuesta y las variables explicativas. Como se puede observar en la Figura 1, el salto de probabilidad que se produciría en el centro de la función sería mucho mayor que si este se produjese en alguno de los extremos. Esto es debido a que estamos ante un modelo no lineal (Landis & Koch, 1977).

Las estimaciones de los parámetros del modelo se calculan considerando que la variable respuesta Y se distribuye según una Bernoulli. A partir de aquí, se utiliza el método de máxima verosimilitud, que busca los valores de $\beta_j$ que maximizan la función de verosimilitud. La expresión es la siguiente:

$$L(\beta) = \prod_{i=1}^{n} p_{li}^{yi} \ (1 - p_{li})^{1-yi} \ = \prod_{i/yi=1}^{n} p_{li} \prod_{j/yi=0}^{n} (1 - p_{li})$$

donde $p_{li} \ = \ P(Y = 1|X_1, \ldots, X_m)$

El análisis de la significancia individual de cada uno de los parámetros del modelo se realiza a través del estadístico de Wald. El contraste consiste en determinar si cada uno de los parámetros $\beta$ del modelo son significativamente distintos de cero. La hipótesis nula del contraste y el estadístico de Wald tienen las siguientes expresiones:

$$H_0: \beta_k = 0$$

$$W = \left( \frac{\hat{\beta}_k}{SE(\hat{\beta}_k)} \right)^2$$

donde $\hat{\beta}_k$ es el estimador del coeficiente $\beta_k$ y $SE(\hat{\beta}_k)$ error estándar del estimador $\hat{\beta}_k$.

Un p-valor pequeño, inferior al nivel de significación que se elija, sugiere que existe evidencia estadísticamente significativa para rechazar la hipótesis nula, lo que indicaría que el coeficiente es significativamente distinto de cero, de manera que la variable en cuestión tendría un impacto significativo en la predicción (Hosmer et al., 2013).

Los tres métodos se selección de variables más utilizados son forward, backward y step. En los tres casos se busca la mejor selección de variables, lo que cambia es el punto de partida que utiliza cada uno de ellos. El método forward parte del modelo vacío y va incluyendo en cada paso la variable que mayor mejora aporte al modelo. El método backward, por el contrario, parte del modelo saturado y, en cada paso, excluye a la variable que menos aportación realiza al modelo. Ambos métodos se detienen en el momento que detectan que incluir o excluir una variable más hace que empeore el modelo alcanzado. En el caso del método stepwise se utiliza una combinación de las técnicas anteriores, de manera que, de forma iterativa, se van añadiendo y/o eliminando variables en cada paso, según resulte más conveniente. De la misma manera que en los métodos anteriores, el método deja de excluir o incluir variables cuando hacerlo empeora el modelo alcanzado (Alonso Revenga & Calviño Martínez, 2025).

Los criterios AIC (Criterio de información de Akaike) y BIC (Criterio de información bayesano) son criterios generales de selección de modelos que se pueden aplicar a una amplia gama de contextos. A menudo se utilizan para comparar modelos construidos a partir de la regresión logística binaria.

Una de las ventajas del uso de estos criterios para la selección de modelos es que permiten comparar modelos de regresión logística incluso cuando el número de variables es diferente. En ambos casos el objetivo es encontrar un equilibrio entre buen ajuste y simplicidad. Sin embargo, la penalización de los modelos más complejos es mayor en el criterio BIC que en el AIC.

$$AIC = -2ln(L) + 2p$$
$$BIC = -2ln(L) + p\ ln(n)$$

donde $ln(L)$ es la función log-verosimilitud, *p* el número de parámetros y $n$ el número de variables independientes.

La odds ratio asociada a un coeficiente estimado es el cambio multiplicativo en las probabilidades relativas de que ocurra el evento de interés por cada unidad adicional en la variable explicativa, manteniendo constantes las demás variables del modelo. (Alonso Revenga & Calviño Martínez, 2025)

En definitiva, las odds ratio representan cuántas veces resulta más probable que se dé el éxito que el fracaso bajo determinadas circunstancias. La expresión de la odds ratio si $X_1$ cambia de 0 a 1, cuando el resto de variables predictoras permanecen constantes es la siguiente:

$$OR = \frac{\frac{P(Y=1|X_1=1,X_2\ldots,X_m)}{1\ -\ P(Y=1|X_1=1,X_2\ldots,X_m)}}{\frac{P(Y=1|X_1=0,X_2\ldots,X_m)}{1\ -\ P(Y=1|X_1=0,X_2\ldots,X_m)}} = e^{\beta_1}$$

En conjunto, la regresión logística binara es una herramienta poderosa para modelizar fenómenos dicotómicos. Además de su capacidad predictiva, es preciso valorar la posibilidad que ofrece de interpretar, a través de las odds ratios, la influencia de cada variable en el modelo. (Hosmer et al., 2013).

Existen diferentes métricas que permiten evaluar la bondad del ajuste en una regresión logística binaria. Algunas de las más destacadas son las siguientes:

- Prueba de Hosmer-Lemeshow: Es uno de los métodos de evaluación de la bondad del ajuste más utilizados en regresión logística binaria. Compara a través del estadístico $\chi^2$ las frecuencias observadas con las esperadas de la variable dependiente en subgrupos ordenados por deciles de riesgo predicho.

$$\chi^2 = \sum_{g=1}^{G} \frac{(O_g - E_g)^2}{E_g(1 - \widehat{p_g}}$$

  Donde G es el número de grupos y $O_g$ y $E_g$ son los valores observados y esperados en el grupo *g.*

  Un pvalor alto indica que el modelo se ajusta bien a los datos. En cualquier caso, la validez del modelo puede verse afectada por el tamaño de la muestra y por el número de grupos, por lo que resulta muy conveniente utilizarlo como complemento a otras medidas (Hosmer et al., 2013).

- $R^2$ de Cox & Snell: Está basado en la comparación de la verosimilitud del modelo con la del modelo nulo. Al no alcanzar el valor de 1, su interpretabilidad es bastante limitada.

$$R^2_{Cox\_Snell} = 1 - \left(\frac{L_0}{L_1}\right)^{\frac{2}{n}}$$

  Donde $L_0$ es la verosimilitud del modelo nulo, $L_1$ es la verosimilitud del modelo ajustado y *n* es el tamaño de la muestra.

- $R^2$ de Nagelkerke: Es una versión del anterior en la que se realiza un ajuste para poder escalar el resultado para que su máximo valor sea 1. Ofrece una medida de la proporción de variabilidad explicada ajustada a escala completa.

$$R^2_{Nagelkerke} = \frac{R^2_{Cox\_Snell}}{1 - (L_0)^{\frac{2}{n}}}$$

Sin embargo, la bondad del ajuste no nos proporciona toda la información relevante sobre la calidad del modelo. Generalmente, se obtiene información valiosa sobre el comportamiento del modelo a través de las métricas que sirven para cuantificar el desempeño predictivo del modelo. Las principales son las siguientes:

- Curva ROC y AUC: La Curva ROC es la representación cartesiana de la tasa de falsos positivos (1-Especificidad) frente a la Sensibilidad del modelo (Tasa de verdaderos positivos). El AUC es el área bajo la curva ROC y se utiliza para evaluar la capacidad del modelo de diferenciar entre los dos grupos de la variable objetivo. Un AUC cercano a 1 indica una excelente capacidad del modelo, en cambio un valor de 0,5 sugiere una capacidad del modelo similar al azar (Fawcett, ,2006).

- Accuracy: Mide el porcentaje total de aciertos del modelo, es decir, el total de los verdaderos positivos y verdaderos negativos sobre el total de las observaciones. Es de fácil interpretación, aunque puede dar una sensación de rendimiento equivocada cuando hay desequilibrio entre las clases (Landis & Koch, 1977).

- Índice Kappa de Cohen: Cuantifica el nivel de acuerdo entre las predicciones realizadas por el modelo y los valores reales, realizando una corrección por el azar. Un calor superior a 0,20 comienza a ser aceptable (Viera & Garret, 2005).

- Sensibilidad: Explica la capacidad del modelo para predecir positivos reales. Un valor alto indica buen desempeño del modelo en la predicción de casos positivos (Altman & Bland, 1994).

- Especificidad: Es la capacidad del modelo de predecir negativos reales. Un valor alto indica buen desempeño del modelo en la predicción de casos negativos (Altman & Bland, 1994).

## 5.4 ANOVA: Fundamentos teóricos

El objetivo de un análisis de la varianza (ANOVA por sus siglas en inglés: Analysis of Variance) es comparar las medias de más de dos grupos de poblaciones. Esta comparación se realiza a partir de la descomposición de la variabilidad total observada en partes atribuibles a diferentes fuentes de variación.

Esta técnica estadística inferencial permite analizar, de forma simultánea, si se dan diferencias estadísticamente significativas entre las medias de más de dos grupos, evaluando si una de ellas, al menos, difiere significativamente del resto. Esta técnica se basa en que resulta poco probable que, si las diferencias observadas entre medias resultan muy grandes en relación con la variabilidad dentro de los grupos, se deban al azar.

El caso unifactorial es la forma más sencilla de aquellas en las que se puede llevar a cabo un análisis de la varianza. Se utiliza cuando existe un único factor explicativo con más de dos niveles y se busca evaluar si el cambio de un nivel a otro produce un efecto significativo sobre una variable respuesta cuantitativa. Para ello, se realiza un contraste de hipótesis en el que se contrasta la hipótesis nula de igualdad de medias frente a la alternativa de que al menos una es diferente.

El modelo matemático con el que se pretenden explicar los datos $Y_{ij}$ obtenidos por la variable respuesta es el siguiente:

$$Y_{ij} = \mu + \alpha_{ij} + \epsilon_{ij}$$

para $i$ =1,…,a; $j$ = 1,…,n

con $\epsilon_{ij} \sim N(0, \sigma^2)$ independientes

donde $Y_{ij}$ es el valor que toma la variable respuesta para el nivel i-ésimo del factor y la réplica j-ésima y $\epsilon_{ij}$ es el término de error aleatorio.

Es importante mencionar que para que los resultados observados en un ANOVA se puedan considerar válidos deben cumplirse las hipótesis de normalidad, homocedasticidad e independencia.

Se define la suma total de cuadrados como:

$$SCT = SCTrat + SCE$$

con $SCTrat = n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2$

y $SCE = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2$

$$SCT = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 = n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2$$

con: $\frac{SCTrat}{\sigma^2} \sim \chi^2_{a-1}$ y $\frac{SCE}{\sigma^2} \sim \chi^2_{N-a}$, independientes y $\frac{SCT}{\sigma^2} \sim \chi^2_{N-1}$

Se define a las medias de cuadrados como las sumas de cuadrados divididas por sus grados de libertad, de manera que:

$MCTrat = \frac{SCTrat}{a-1}$ y $MCE = \frac{SCE}{N-a}$, por lo que $\frac{MCTrat}{MCE} \sim F_{a-1,N-a}$

La hipótesis nula es $H_0 : \mu_1 = \ldots = \mu_a = \mu$. Valores grandes de $\frac{MCTrat}{MCE}$ aportan evidencia en contra de *H0* y la región crítica del contraste de hipótesis con un nivel de significación $\alpha$ es:

$$C = \left\{\frac{MCTrat}{MCE} > F_{a-1;\, N-a,\alpha}\right\}$$

Para poder considerar válidos los resultados del test *F,* basado en el estadístico $\frac{MCTrat}{MCE}$, es necesario que los residuos sean variables aleatorias independientes, con distribución normal y con una varianza constante (Montgomery, 2013).

## 5.5 Transformación Box-Cox: Fundamentos teóricos

La transformación Box-Cox es una técnica estadística que se utiliza mejorar un modelo estadístico cuyos residuos no cumplen con las hipótesis de homocedasticidad y normalidad. En ocasiones, en técnicas como el ANOVA o la regresión lineal, los supuestos necesarios para que el modelo sea interpretable, no se cumplen. Esta técnica puede corregir las deficiencias del modelo con la aplicación de una transformación paramétrica a la variable respuesta. Se define a partir de la siguiente expresión:

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ log(y), & \lambda = 0 \end{cases}$$

donde $y^{(\lambda)}$ es la variable transformada y $\lambda \in \mathbb{R}$ es el parámetro de transformación, estimado con los datos de la muestra.

El valor óptimo de $\lambda$ se estima mediante el método de máxima verosimilitud, buscando el valor de $\lambda$ que, bajo los supuestos de normalidad y homocedasticidad, maximiza la verosimilitud de los datos transformados.

La utilidad de la transformación Box-Cox radica en la capacidad que tienen para estabilizar la varianza y aproximar la normalidad de los errores, lo que resulta fundamental si se quiere garantizar la validez de los contrastes estadísticos en los modelos lineales. A nivel práctica, su uso está especialmente recomendado cuando los residuos no presentan homocedasticidad o en casos de asimetría muy marcada. (Box & Cox, 1964).

## 5.6 Test U de Mann-Whitney: Fundamentos teóricos

El test U de Mann-Whitney, también se le conoce como la prueba de Wilcoxon, es una prueba no paramétrica que se utiliza para determinar si existen diferencias significativas entre las distribuciones de las que provienen dos muestras independientes. Al tratarse de una prueba no paramétrica, no requiere que las variables se ajusten a ninguna distribución conocida ni que las varianzas resulten ser homogéneas. Esto lo convierte en una herramienta muy robusta, pero muy flexible, en aquellos casos en los que no se cumplen los supuestos de normalidad y homocedasticidad y, por lo tanto, no es posible utilizar otras pruebas paramétricas comunes como la t de Student para muestras independientes.

La prueba evalúa la hipótesis nula de que ambas muestras provienen de poblaciones con la misma distribución. El estadístico de contraste responde a la siguiente expresión:

$$U = min(U_1, U_2)$$

con $U_1 = R_1 - \frac{n_1(n_1+1)}{2}$ y $U_2 = R_2 - \frac{n_2(n_2+1)}{2}$ ; donde $U_1 + U_2 = n_1 n_2$

En muestras suficientemente grandes, es común aplicar una aproximación normal mediante la corrección de continuidad de acuerdo a la siguiente expresión:

$$Z = \frac{U - \mu_U \pm 0.5}{\sigma_U}$$

donde $\mu_U = \frac{n_1 n_2}{2}$ y $\sigma_U = \sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}$

Para poder ser considerados válidos, los resultados del tets U de Mann-Whitney deben cumplirse los siguientes supuestos:

- Muestras independientes
- Los datos deben presentar un orden jerárquico
- Las distribuciones deben tener la misma forma

Generalmente, se interpreta como una prueba de comparación de medianas, pero, en realidad, lo que hace es comparar la tendencia central de las distribuciones, asumiendo que las formas son iguales (Mann & Whitney, 1947).

## 5.7 Coeficiente de correlación de Spearman: Fundamentos teóricos

El coeficiente de Spearman, también conocido como rho de Spearman, es una técnica no paramétrica capaz de evaluar si existe una asociación monótona entre dos variables ordinales o numéricas continuas. Es la alternativa no paramétrica al coeficiente de correlación de Pearson en los casos en los que los datos no cumplen las condiciones de normalidad y/o linealidad.

El coeficiente de correlación de Spearman no se basa en los valores originales sino en los rangos, lo que le dota de robustez en los casos en los que la distribución no es normal y permite detectar relaciones monótonas, aunque no sean lineales.

Si existen dos variables *X* e *Y*, con *n* observaciones, y se asignan a cada una sus respectivos rangos $R(X_i)$ y $R(Y_i)$, se puede definir el coeficiente de Spearman como:

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(R(X_i) - R(Y_i))^2}{n(n^2-1)}$$

El coeficiente de correlación de Spearman toma valores entre -1 y 1, indicando 1 una relación completa monótona positiva, -1 relación completa monótona negativa y 0 ausencia de relación. Además, es invariante a las transformaciones monótonas de los datos y no se ve afectado por la escala de las variables ni por los outliers (Spearman,1904).

## 5.8 Test de la Chi-cuadrado: Fundamentos teóricos

El test de la Chi-cuadrado es una prueba que permite evaluar la existencia o no de relación entre dos variables categóricas. Resulta especialmente relevante en casos en los que se buscan relaciones a partir de datos agrupados en tablas de contingencia. Bajo la hipótesis de independencia, se espera que la frecuencia observada en cada una de las celdas sea proporcional al producto de los totales marginales de la fila y la columna correspondiente, ajustado al tamaño total de la muestra.

El estadístico de contraste resulta de la suma de cuadrados de las diferencias entre las frecuencias observadas y las esperadas, ponderadas por las frecuencias esperadas. Un valor alto del estadístico de contraste advierte de la discrepancia entre lo que se ha observado y lo que se esperaría en caso independencia, lo que permitiría rechazar la hipótesis nula. La expresión es la siguiente:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

donde $O_{ij}$ es la frecuencia observada en la celda $ij$ y $E_{ij}$ es la frecuencia observada en la celda $ij$.

Para considerar válidos los resultados del contraste, deben darse las siguientes condiciones:

- Observaciones independientes
- La frecuencia esperada de más del 80% de las observaciones no debe ser inferior a 5.
- No debe haber celdas con frecuencias esperadas inferiores a 1.

Cuando no se dan las circunstancias anteriores, resulta recomendable aplicar la corrección de Yates o, en su defecto, utilizar otro tipo de pruebas como el test exacto de Fisher. (Siegel & Castellan, 1988)

## 5.9 Árboles de decisión y Bagging: Fundamentos teóricos

Para explicar a nivel teórico en qué consiste el bagging, es necesario acercarnos antes a los árboles de decisión. Un árbol de decisión es un sencillo modelo predictivo que representa decisiones a través de una estructura jerárquica que se ramifica según avanza. Este tipo de modelo va dividiendo la muestra con la que trabaja en regiones homogéneas en base a decisiones de tipo binario sobre las variables explicativas. En los árboles de clasificación, la variable objetivo es categórica.

En esta modalidad de árboles de decisión, el modelo se encarga de seleccionar en cada nodo tanto la variable más importante como el punto de corte óptimo. Para decidir cuál es la variable más importante en cada nodo, el modelo evalúa todas ellas. Para cada combinación, calcula cuánto reduciría el error esa selección y elige aquella que produzca una mayor reducción de este. Una vez seleccionada la variable de un nodo, para decidir cuál es el punto de división utiliza el Índice de Gini que, básicamente, mide la probabilidad de que dos elementos que han sido seleccionados al azar pertenezcan a clases distintas (James et al., 2013).

$$IG(nodo) = 1 - \sum_{k=1}^{K} (p_k^j)^2$$

Donde $p_k^j$ es la proporción de observaciones de la categoría *k* en el nodo *j*. Por lo tanto, cuanto menor sea el *IG* mayor es la capacidad predictiva de la variable observada y, por lo tanto, mejor será el modelo.

Se trata de un modelo predictivo sencillo y que resulta muy fácil de interpretar. Sin embargo, suele presentar una alta variabilidad, lo que provoca que cambios mínimos en los datos supongan un gran cambio en el árbol resultante y, por lo tanto, en las predicciones.

El Bagging, acrónimo de Bootstrap Aggregating, nace, precisamente, para tratar de minimizar la elevada variabilidad asociada a los árboles de decisión. Dado que los árboles de decisión tienden al sobreajuste cuando se entrenan sin ningún tipo de

restricción, el hecho de promediar múltiples modelos con varianza alta podría estabilizar la predicción final. Para conseguirlo, utiliza una técnica llamada Bootstrap, que no es más que realizar sucesivos muestreos con reemplazamiento sobre el conjunto de datos con el que se trabaja. De esta manera, se construye un árbol de decisión sobre cada muestra Bootstrap para luego promediar las predicciones, en el caso de que la variable a predecir fuese numérica, o seleccionar la más probable, en el caso de que la variable objetivo fuese categórica. (Breiman, 1996)

Tal y como se ha señalado, la reducción de la variabilidad es el objetivo principal del bagging y viene determinada por el hecho de que la varianza de la media de una muestra aleatoria simple siempre será menor que la varianza de cada uno de los elementos que componen la muestra, como queda demostrado con la siguiente expresión:

$$Var[\bar{y}] = Var\left[\frac{1}{B}\sum_{i=1}^{B}\hat{y}_i\right] = \frac{1}{B^2}\sum_{i=1}^{B}Var(\hat{y}_i) = \frac{Var(\hat{y}_i)}{B}$$

La construcción de cada uno de los árboles que utiliza el modelo deja fuera un número de observaciones. A estas observaciones se les conoce como OOB (de sus siglas en inglés Out Of Bag) y supone en torno a un tercio del total de los datos. Estas observaciones OOB son utilizadas posteriormente, como una partición de prueba, para probar el modelo de manera de que para cada observación se busca la predicción solamente con los árboles en los que no fue utilizada en la creación del modelo, pudiendo así estimar el error del modelo.

El bagging es una técnica robusta, de interpretación sencilla, con una variabilidad baja y con mejor rendimiento predictivo que otras técnicas dentro del machine learning. Sin embargo, la interpretabilidad del modelo es limitada, aunque, a pesar de esta limitación, es posible generar un ranking de importancia de las variables predictoras que mayor información aportan a la predicción de la variable objetivo. Esta información se obtiene a partir de la contribución de cada variable a la mejor del índice de Gini

calculado en cada nodo de cada árbol del conjunto. De esta manera es posible identificar las variables más influyentes en la predicción, lo que aporta al modelo cierta capacidad explicativa.

Una vez que el modelo ha detectado la importante una relación importante entre una variable predictora y la variable objetivo, puede resultar interesante ver cómo es el tipo de relación existente. El PDP o Gráfico de Dependencia Parcial es una utilidad gráfica que permite interpretar modelos como el bagging, profundizando en la relación entre una variable explicativa y la variable respuesta, manteniendo constantes los efectos del resto de las variables. Resulta especialmente útil para detectar relaciones no lineales ya que permite localizar tendencias, umbrales o no linealidades (Molnar, C., 2022).

# 6 Modelización

## 6.1 Regresión logística binaria

Después de todo el trabajo de agrupación de variables, reasignación de clases e imputación, en la base de datos que servirá de base al trabajo se observa que, de un total de 8.625 observaciones, el 37,15% corresponden a individuos que sufren una enfermedad respiratoria.

Con objeto de evaluar, a través de validación cruzada, el rendimiento real del modelo predictivo que se obtenga, el conjunto total de las observaciones se ha dividido, aleatoriamente, en dos partes: un 80% de las observaciones entrenarán al modelo mientras que el 20% restante se utilizarán para su evaluación. Con el uso de esta técnica, se consigue que la evaluación del modelo se realice con datos desconocidos, lo que ofrece una evaluación sin posibilidades de sobreajuste.

Para la selección automática de variables se han utilizado son los métodos forward, backward y stepwise y, como criterios de selección, se han utilizado AIC y BIC. La combinación de los tres métodos de selección con los dos criterios genera seis modelos. Algunos resultan ser coincidentes alcanzando, finalmente, tres modelos diferentes. Son los siguientes:

- Modelo 1: Método stepwise y criterio BIC

  respi ~ neuro + edad + otras + muscu + $ON_x$

- Modelo 2: Método stepwise y criterio AIC

  respi ~ neuro + edad + otras + muscu + $ON_x$ + nivel_estudios + peso + sexo + ingresos_netos

- Modelo 3: Método backward y criterio BIC

  respi ~ sexo + edad + muscu + neuro + otras + peso + $ON_x$

Con la finalidad de evaluar la capacidad explicativa de los modelos seleccionados se ha utilizado el coeficiente de determinación de Nagelkerke. De esta manera se podrá comparar la variabilidad explicada por los modelos.

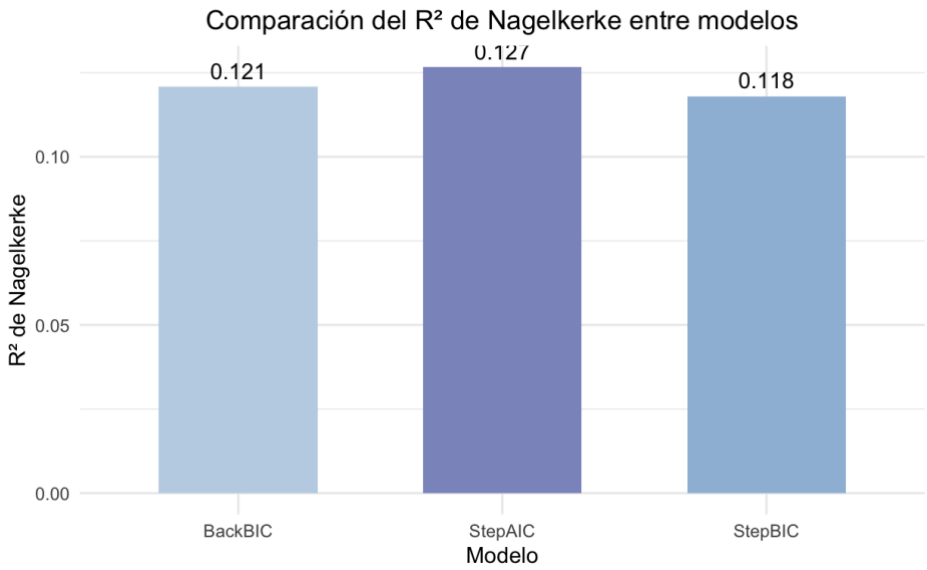


*Figura 2 Comparativa Coeficiente de determinación de Nagelkerke*

La *Figura 2* muestra que el modelo con un coeficiente de determinación de Nagelkerke más alto es el correspondiente al modelo StepAIC, lo que indica que es el que mayor capacidad predictiva ofrece. Sin embargo, las diferencias entre los 3 modelos son tan pequeñas que no resultan relevantes. Es necesario destacar que todos los valores se sitúan por debajo de 0,13, lo que resulta habitual en estudios observacionales con datos reales en el ámbito de la salud pública. En ningún caso debe interpretarse como una falta de utilizad del modelo, sino como una derivada del carácter multifactorial del fenómeno objeto de estudio (Hosmer, Lemeshow & Sturdivant, 2013; Menard, 2002).

Una vez evaluada la capacidad explicativa global del modelo, se avanzará hacia la comparación de los modelos en base al desempeño predictivo. Este tipo de métricas nos permite conocer la eficacia del modelo a la hora de realizar la clasificación de nuevas observaciones.

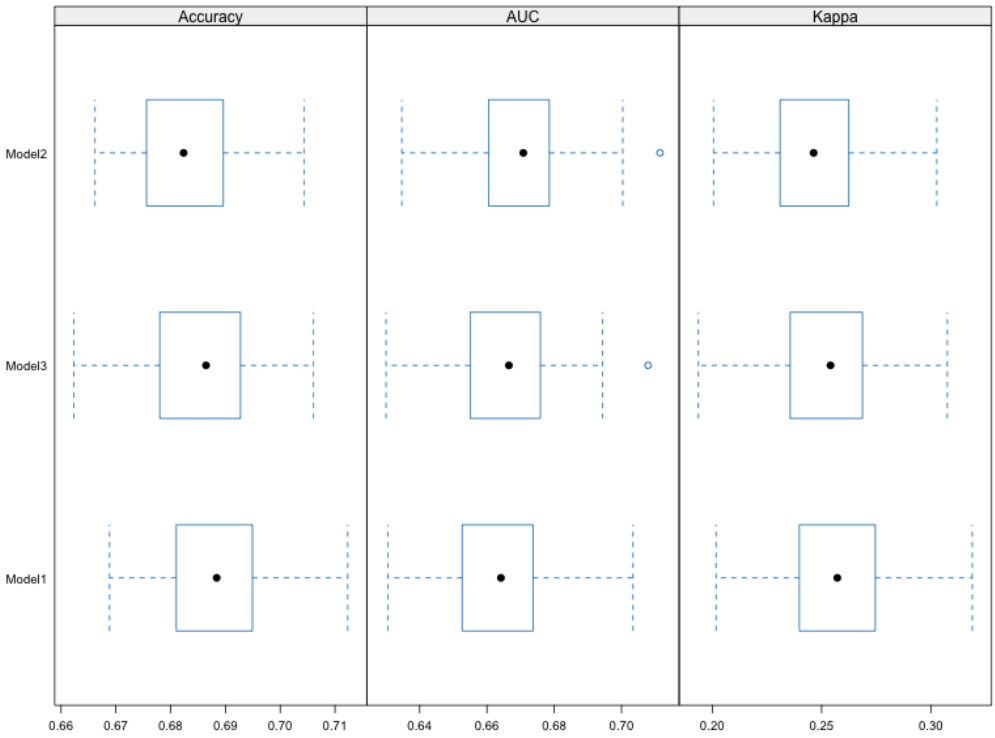


*Figura 3 Comparativa desempeños predictivos*

Las métricas de desempeño predictivo, como muestra la *Figura 3,* resultan ser muy similares entre los tres modelos considerados. Esta similitud ya se había observado en las métricas de bondad del ajuste, lo que refuerza la idea de que no hay diferencias sustanciales ni en la capacidad explicativa ni en el poder predictivo de los modelos evaluados. Dado que el rendimiento que los tres modelos ofrecen es bastante parecido se utiliza el principio de parsimonia como criterio de decisión.

El principio de parsimonia es ampliamente aceptado en el ámbito de la modelización estadística y sugiere que ante modelos de rendimientos similares es preferible seleccionar el que presenta una estructura más simple (Burnhan & Anderson). Por lo tanto, se selecciona el Modelo 1 como el más adecuado, ya que es el más sencillo de los tres propuestos, lo que facilita su interpretación, implementación y generalización a nuevos datos.

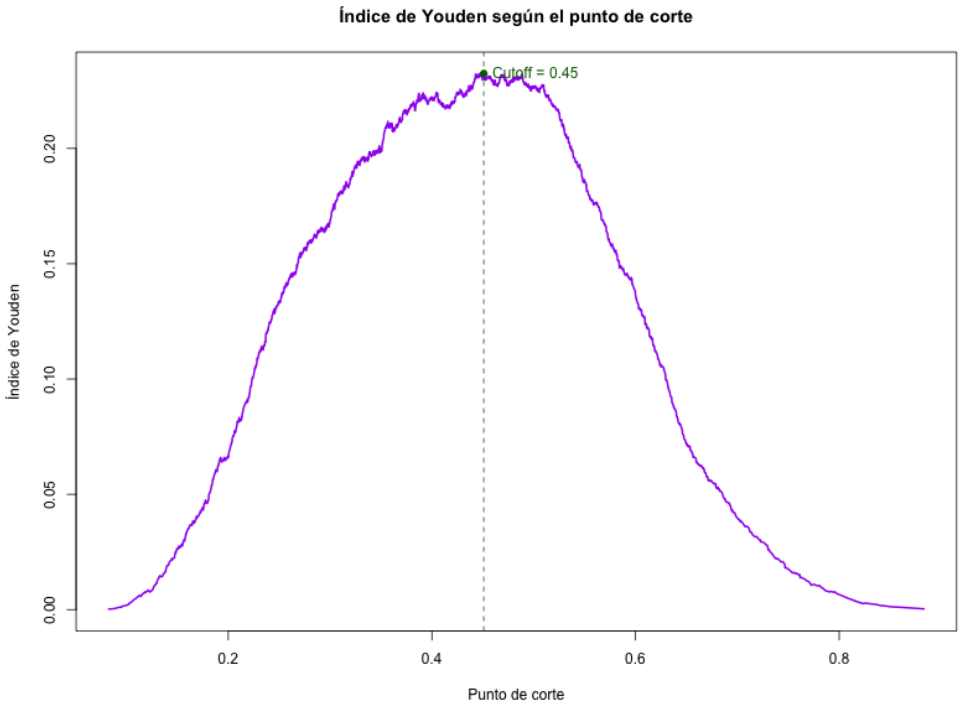


*Figura 4 Índice de Youden*

En la *Figura 4* se muestra el índice de Youden, que se maximiza en el punto en el que existe mayor equilibrio entre las curvas de sensibilidad y 1-especificidad. En este caso, el máximo se alcanza en 0,45. Se evaluará de nuevo el modelo utilizando ese valor como umbral de decisión.

| prob_threshold | Accuracy | Kappa | Sensivity | Specificity |
|---|---|---|---|---|
| 0.45 | 0.6720765 | 0.2557354 | 0.4204368 | 0.8208439 |
| 0.50 | 0.6888857 | 0.2584198 | 0.3326833 | 0.8994697 |

*Tabla 1 Comparación umbrales de clasificación*

Se observa que la única métrica que mejora cambiando el umbral de clasificación es la sensibilidad. Sin embargo, obtener una especificidad alta supone tener mucha precisión a la hora de descartar casos positivos, probablemente el objetivo más importante del estudio. Así se decide seleccionar como umbral de decisión el que proporciona una mayor especificidad al modelo.

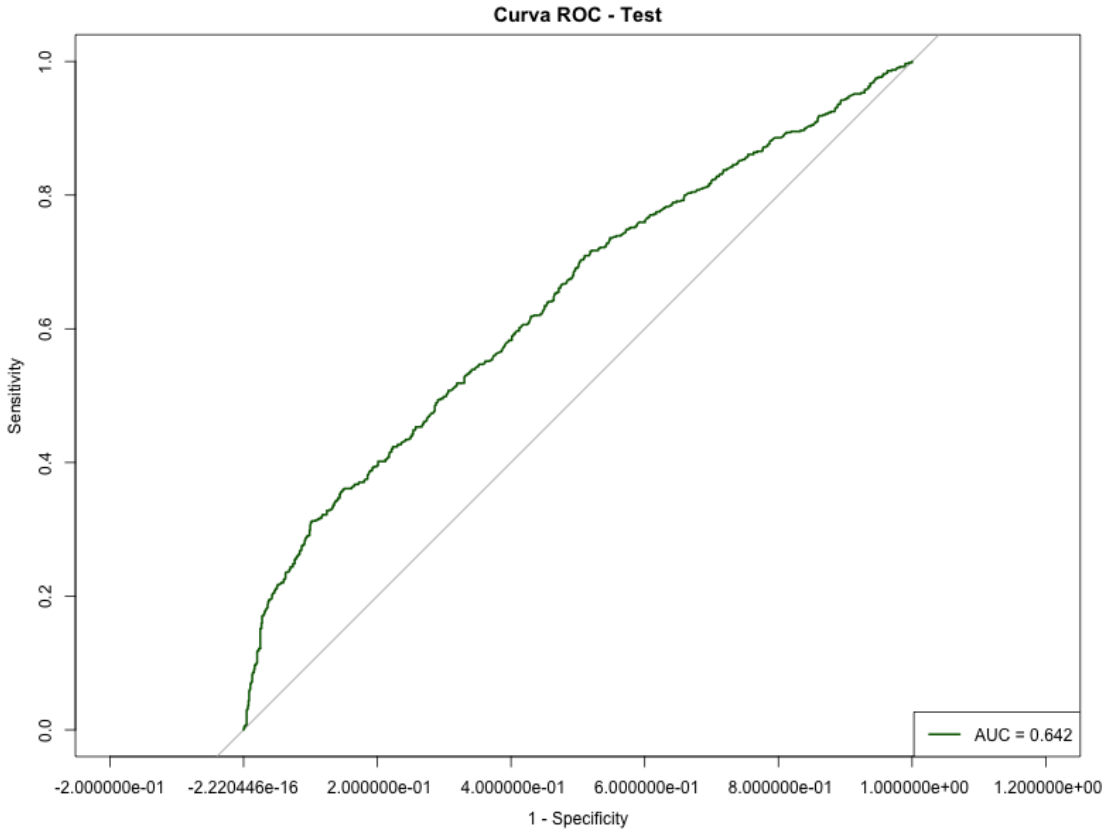


*Figura 5 Curva ROC con datos de prueba*

El análisis de la curva ROC permite concluir que el AUC está por encima del azar, pero se encuentra distante de valores altos que indiquen una alta capacidad predictiva general del modelo. Tanto Accuracy como Kappa alcanzan valores aceptables, pero no excesivamente buenos. Sin embargo, la especificidad del modelo resulta ser muy alta, lo que implica que clasificará correctamente el 89,95% de los negativos, lo que resulta determinante para cumplir con el objetivo del estudio.

| parameter | AUC | Accuracy | Kappa | Sensivity | Specificity |
|---|---|---|---|---|---|
| train | 0.664 | 0.6888857 | 0.2584198 | 0.3326833 | 0.8994697 |
| test | 0.642 | 0.6774942 | 0.2298080 | 0.3140625 | 0.8920664 |

*Tabla 2 Desempeño predictivo del modelo con datos de entrenamiento y prueba*

Las diferencias que se han observado en las métricas obtenidas con el conjunto de datos de entrenamiento y de prueba son mínimas. Esto evidencia claramente que estamos ante un modelo robusto que no ha caído en sobreajuste, manteniendo su capacidad predictiva también con los datos no utilizados durante el proceso de ajuste del modelo de regresión logística binaria.

La siguiente tabla presenta los coeficientes estimados del modelo de regresión logística binaria. También incluye los errores estándar, valores Z y p-valores. El modelo tiene como objetivo de predecir la presencia de enfermedades respiratorias crónicas en un individuo en función de una serie de variables explicativas. La tabla solo muestra aquellas que resultaron ser significativas.

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -0.3378919 | 0.13074946 | -2.5842696 | 0.00975855 |
| neuro1 | 0.58531292 | 0.06170163 | 9.48618192 | 2.3965E-21 |
| edad | -0.0262249 | 0.00163509 | -16.038822 | 6.844E-58 |
| otras1 | 0.84707275 | 0.0659347 | 12.8471474 | 8.9241E-38 |
| muscu1 | 0.57011018 | 0.06209186 | 9.18172211 | 4.2425E-20 |
| OxN | 0.01219364 | 0.00226505 | 5.38337908 | 7.31E-08 |

*Tabla 3 Coeficientes del modelo de regresión logística*

A continuación, se presentan las odds ratios asociadas a las variables explicativas que han resultado seleccionadas para el modelo. Se consiguen a partir de la exponenciación de los coeficientes estimados de la regresión logística y permiten realizar una realizar una explicación más intuitiva del efecto de cada variable predictora.

| (Intercept) | neuro1 | edad | otras1 | muscu1 | OxN |
|---|---|---|---|---|---|
| 0.7132724 | 1.7955528 | 0.9741160 | 2.3328081 | 1.7684619 | 1.0122683 |

*Tabla 4 Odds ratios estimadas para el modelo de regresión logística*

La interpretación que se puede realizar de las odds ratios del modelo, *caeteris paribus en media,* es decir, considerando el efecto medio de cada variable sobre la variable respuesta, manteniendo constantes el resto de variables incluidas en el modelo, son las siguientes:

- Las personas que tienen alguna enfermedad neurológica tienen casi el doble de probabilidad de tener una enfermedad respiratoria que las que no la tienen.
- Las personas con algún tipo de enfermedad crónica diferente a las mencionadas tienen más del doble de probabilidad de tener una enfermedad respiratoria que las que no la tienen.
- Las personas con enfermedades musculoesqueléticas tienen casi el doble de probabilidad de tener una enfermedad respiratoria que las que no la tienen.
- Por cada microgramo por metro cúbico de óxidos de nitrógeno (NO y $NO_2$) adicional de exposición a la que se somete a una persona, la probabilidad de tener una enfermedad respiratoria aumenta en un 1,2%.
- Por cada año adicional de edad, la probabilidad de contraer una enfermedad respiratoria se ve reducida en un 2,6%.

En resumen, siempre manteniendo constantes el resto de variables del modelo, el hecho de tener una enfermedad crónica de cualquier tipo, salvo si esta es de origen cardiaca, prácticamente duplica la probabilidad de llegar a desarrollar una enfermedad respiratoria si se comparan con aquellos que no manifiestan ninguna enfermedad crónica. La exposición a mayores cantidades de óxidos de nitrógeno también aumenta significativamente las probabilidades de desarrollar una enfermedad respiratoria.

Llama la atención que el modelo presente, de forma contraintuitiva, la edad como un factor protector respecto al diagnóstico de enfermedades respiratorias crónicas. Sin embargo, hay varias explicaciones científicas plausibles documentadas. Una de ellas es el sesgo de supervivencia. Ocurre cuando la muestra seleccionada no incluye en el análisis más que a los individuos que “sobrevivieron” al proceso y no a aquellos que o no lo hicieron o no están disponibles para la encuesta (Delgado-Rodríguez & Llorca, 2004). Otra posible explicación es el subregistro o la omisión de reporte de ciertos eventos, lo que puede llevar a una representación inexacta de la realidad. En el caso de la encuesta en que se basa este trabajo podría deberse a la normalización de los

síntomas por la edad (Geriatrics Editorial Office, 2024) o, simplemente, a menor acceso a un diagnóstico o al seguimiento (National Advisory Comittee on Rural Health and Human Services, 2018).

Aunque lo cierto es que existen explicaciones tanto a nivel teórico como empírico que podrían justificar un aparente efecto protector de la edad frente a las enfermedades respiratorias crónicas, como los expuestos anteriormente, este hallazgo ha de interpretarse con mucha cautela. Resulta necesario profundizar más en el análisis estadístico de esta variable, evaluando su comportamiento dentro del modelo predictivo de una forma más profunda. De esta manera se podrá obtener una comprensión más precisa del papel que desempeña la edad en el desarrollo del tipo de patologías objeto de estudio, lo que serviría no solo para enriquecer el análisis objeto de este trabajo sino, también, para abrir nuevas líneas para trabajos futuros.

## 6.2 ANOVA

Dado que la interpretación que se deduce de la odd ratio relacionada con la variable edad va en contra de lo que intuitivamente pudiera parecer lógico, ya que se sugiere que, según aumenta la edad, disminuye la probabilidad de padecer una enfermedad respiratoria, se va a proceder a realizar un análisis adicional para tratar de entender mejor cómo se comporta esta variable edad en relación con la probabilidad de contraer una enfermedad respiratoria.

Se han utilizado las probabilidades predichas por el modelo de regresión logística binaria seleccionado como la variable dependiente en un análisis de la varianza (ANOVA) en el que se ha considerado la edad como factor único agrupado en cuatro niveles diferentes. Los niveles establecidos para los grupos edad son los compuestos por los individuos menores de 20 años, los individuos entre 20 y 40, entre 40 y 60 y el grupo de personas mayores de 60 años.

Para proceder a este análisis, en lugar de la variable dicotómica que indica presencia o ausencia de enfermedad respiratoria, se han utilizado las probabilidades estimadas por el modelo de regresión logística binaria para la aparición de la misma. De esta manera, se podrá estudiar si existe o no una variación en la probabilidad de aparición de una enfermedad respiratoria crónica relacionada con la pertenencia a determinados grupos de edad.

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| grupos_edad | 3 | 25.492 | 8.4973 | 500.45 | <2.2e-16 |
| Residuals | 6897 | 117.107 | 0.0170 | | |

*Tabla 5 Análisis de la Varianza*

El análisis, a priori, parece confirmar que existe una diferencia significativa en la probabilidad media predicha entre los diferentes grupos de edad establecidos. El valor del estadístico de contraste sugiere que la diferencia entre los grupos es muy marcada. Este resultado refuerza la idea de que la edad está fuertemente asociada a tener una enfermedad respiratoria.

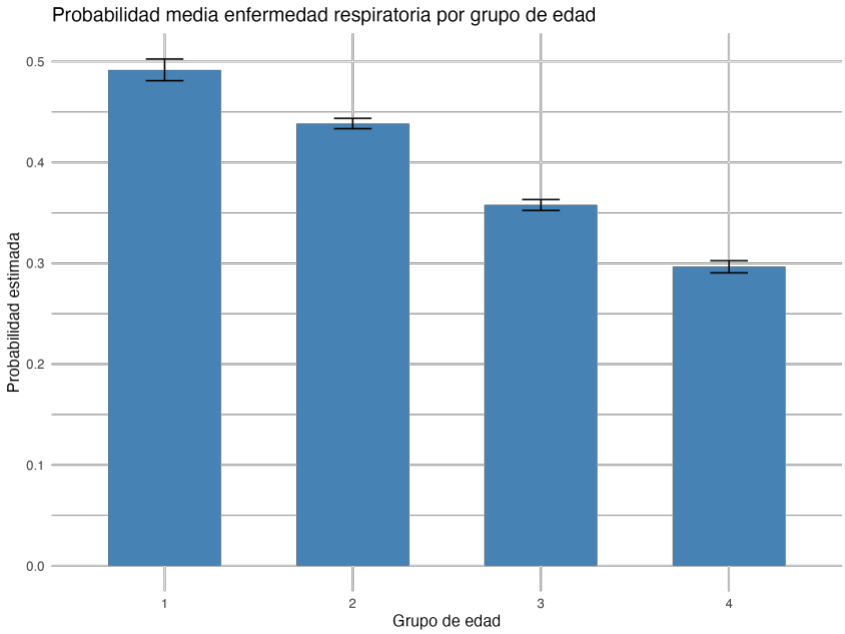


*Figura 6 Gráfico probabilidad media enfermedad respiratoria por grupo de edad*

Sin embargo, antes de tomar las conclusiones del ANOVA como válidas es necesario validar el modelo a través del análisis de los residuos. Se realizan los tests de Kolmogorov Smirnov, Levene y Durbin Watson para contrastar normalidad, homocedasticidad y autocorrelación, respectivamente.

| Test | p-value |
|---|---|
| Kolmogorov-Smirnov test | < 2.2e-16 |
| Levene's Test for Homogeneity of Variance | < 2.2e-16 |
| Durbin-Watson test | 5.69e-06 |

*Tabla 6 Análisis de los residuos*

En todos los casos el valor es muy pequeño, por lo que se puede afirmar que no se cumplen ninguna de las tres hipótesis del modelo (normalidad, homocedasticidad y autocorrelación de los residuos). Por lo tanto, los resultados del ANOVA no pueden ser tenidos en cuenta. Sin embargo, antes de descartarlos, se ha de buscar si una transformación de los datos podría cambiar esta situación. La transformación Box-Cox se utiliza para, en general, mejorar las hipótesis clásicas asociadas a los modelos lineales. De esta manera, a menudo reduce la variabilidad no constante entre los

niveles de un factor, aproxima la normalidad de los residuos y ayuda a linealizar relaciones no aditivas entre los factores del modelo y la variable respuesta.

| Test | p-value |
|---|---|
| Kolmogorov-Smirnov test | < 2.2e-16 |
| Levene's Test for Homogeneity of Variance | < 2.2e-16 |
| Durbin-Watson test | 5.587e-05 |

*Tabla 7 Análisis de los residuos después de la transformación Box-Cox*

A pesar de la transformación, las hipótesis del modelo siguen sin cumplirse. La explicación podría estar en que el ANOVA clásico, basado en la regresión lineal, supone que el efecto de cada nivel es aditivo respecto del nivel general. Si la relación entre esas variables no fuese lineal explicaría esta situación, ya que el ANOVA no es capaz de capturar relaciones que no sean lineales (Montgomery, 2017). Por lo tanto, se recurrirá a otro tipo de técnicas estadísticas no paramétricas que sean capaces de detectar relaciones de carácter no lineal.

## 6.3 Pruebas no paramétricas

Para profundizar en el tipo de relación existente entre la variable edad y la presencia de una enfermedad respiratoria se han aplicado tres pruebas no paramétricas complementarias. Cada una de ellas está dirigida a capturar aspectos diferentes en la relación que se está estudiando. De esta manera es posible reforzar la solidez de los hallazgos a través de métodos suficientemente robustos que, además, enfocan el análisis desde diferentes niveles de medida y supuestos distribucionales.

- La prueba de U de Mann-Whitney compara al grupo de individuos con enfermedad respiratoria con el grupo de individuos que no la tienen para ver si las distribuciones de la variable edad son diferentes en ambos grupos. De esta manera se puede saber si existen diferencias significativas de edad entre los dos grupos. Los resultados muestran un estadístico de prueba, basado en la suma de los rangos de la variable edad agrupadas según la variable dependiente, de W= 9997208, con un pvalor < 2.2e-16, por lo que la diferencia entre los grupos es estadísticamente significativa.

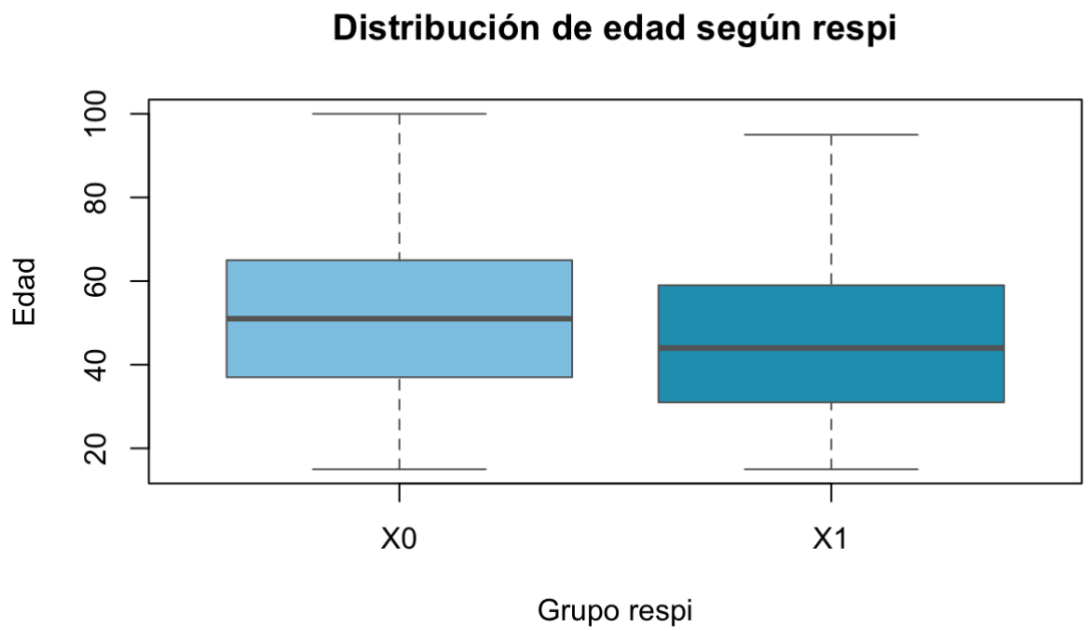


*Figura 7 Comparación de edad según presencia de enfermedad respiratoria*

- El coeficiente de correlación de Spearman permite cuantificar la fuerza y dirección de una relación monótona entre la edad y la probabilidad de sufrir una enfermedad de carácter respiratorio. El resultado obtenido fue $\rho$= - 0.127, con un pvalor <2.2e-16. El valor negativo del coeficiente de correlación de rangos de Spearman indica que existe una asociación monótona, débil y negativa entre las variables, por lo que, a mayor edad existe una ligera tendencia a presentar menos probabilidad de contraer una enfermedad respiratoria de carácter crónico. El pvalor tan pequeño permite concluir que esta asociación es estadísticamente significativa.

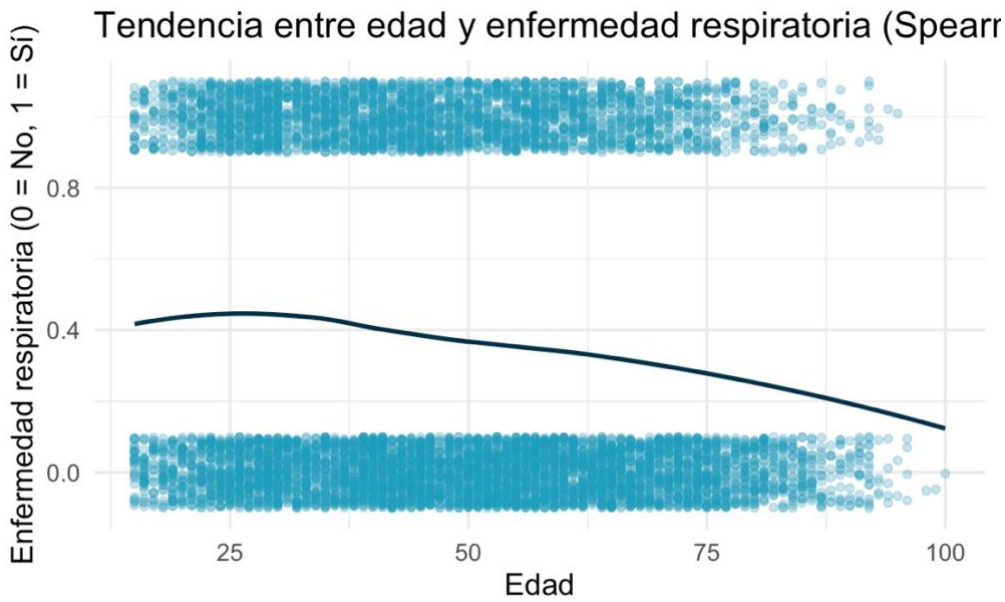


*Figura 8 Edad vs Enfermedad respiratoria (Correlación de Spearman)*

- El test de Chi cuadrado de independencia de Pearson permite explorar si la frecuencia de aparición de una enfermedad respiratoria varía significativamente entre diferentes grupos de edad. El estadístico de contraste toma un valor 107.23, con 3 grados de libertad y un pvalor<2.2e-16. Estos resultados indican que existe una diferencia significativa en la distribución entre personas con enfermedad respiratoria crónica según los diferentes grupos de edad. Por lo tanto, existen evidencias estadísticamente significativas de que la aparición de enfermedades respiratorias crónicas varía según la edad.

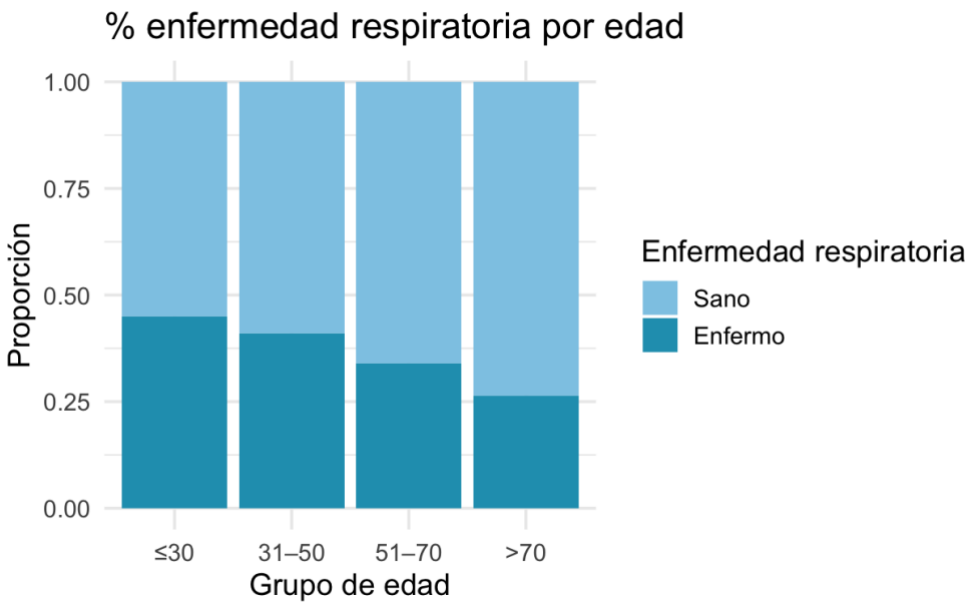


*Figura 9 Enfermedad respiratoria por edad (Test Chi cuadrado)*

Los tres enfoques no paramétricos, comparación de distribuciones, asociación monótona y dependencia categórica, muestrean resultados convergentes. Estos resultados refuerzan la idea de que existe una relación entre la edad y la aparición de una enfermedad respiratoria de carácter crónico. Sin embargo, la dirección negativa de esta relación resulta ser negativa lo que, unido a la baja magnitud del coeficiente de correlación de Spearman, podría sugerir una relación no lineal entre las variables. La diferencia significativa entre los distintos grupos de edad que revela el test de la Chi-cuadrado, refuerza esta hipótesis. Todo ello, apunta a una estructura compleja que no puede ser adecuadamente explicada mediante una relación lineal.

## 6.4 Bagging

El modelo de regresión logística binaria ha encontrado una relación contraintuitiva entre la edad y la presencia de enfermedad respiratoria. Tanto del ANOVA como del resto de técnicas no paramétricas realizadas se ha podido concluir que, el sentido de esta relación, podría responder a la falta de linealidad. Por ello, se ha considerado oportuno recurrir a métodos de aprendizaje automático más flexibles.

El bagging es una alternativa robusta a los problemas de falta de linealidad y permite la reducción de la varianza del modelo al combinar múltiples árboles de decisión generados a partir de Bootstrap, lo que es muy útil cuando las relaciones entre las variables no se pueden captar haciendo uso de los modelos paramétricos tradicionales. Además de capturar patrones no lineales, es capaz de detectar interacciones complejas entre variables. Por todo ello, resulta una opción adecuada para aportar un enfoque complementario al análisis que permita contrastar la robustez de los hallazgos anteriores y dotar de una mayor flexibilidad la exploración de posibles relaciones entre las variables.

Para encontrar el mejor modelo, se comienza el proceso con la construcción del modelo saturado y eligiendo 500 como el número de árboles, lo que produce un modelo con un error OOB estimado del 32,65%, lo que quiere decir que ese porcentaje de las observaciones totales fueron mal clasificadas utilizando datos no vistos por cada árbol. La tasa de error de negativos es del 11,48% y la de positivos del 68,45%. El modelo saturado ofrece una buena capacidad predictiva para los negativos, pero pobre para los positivos.

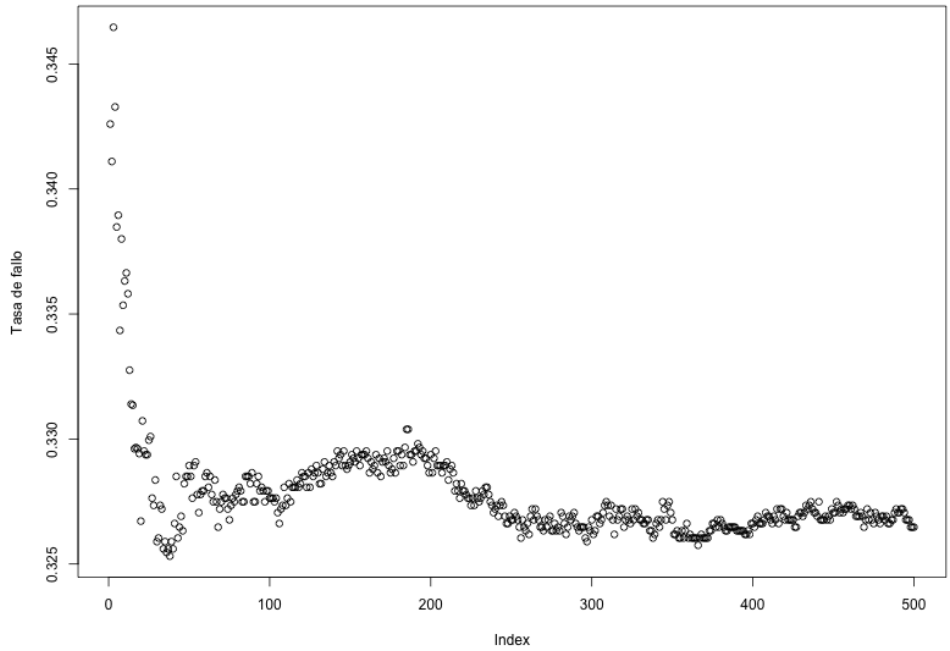


*Figura 10 Tasa de error según número de árboles*

En la *Figura 10* se puede observar con claridad cómo desciende la tasa de error de forma bastante pronunciada en cuanto aumenta la cantidad de árboles del modelo. Este comportamiento es una de las características principales de este tipo de métodos de ensamblado, donde la agregación de un gran número de clasificadores individuales es capaz de reducir de una forma muy significativa la varianza del modelo global. Se observa pues que, con un número pequeño de árboles, la variabilidad es mayor, poniendo de manifiesto la inestabilidad propia de los modelos individuales. La tasa de error desciende súbitamente, aunque muestra cierta variabilidad hasta los 250 árboles, aproximadamente, donde comienza a mantenerse con una variabilidad mínima.

Uno de los parámetros más determinantes para la obtención del mejor modelo es el tamaño de la hoja, que es el mínimo número de observaciones que ha de haber en un nodo para que no continúe dividiéndose. A menor tamaño de hoja, mayor es el tamaño y la complejidad del árbol resultante.

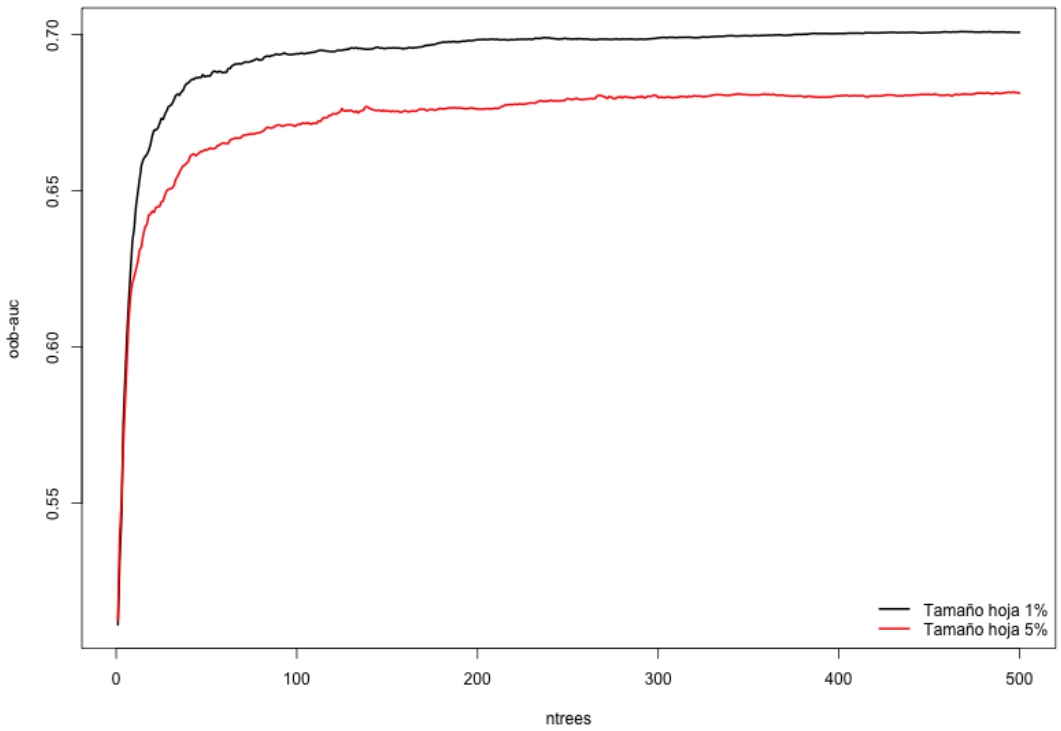


*Figura 11 AUC para hojas 1% y 5%*

La *Figura 11* muestra cómo el AUC resulta ser mejor para un tamaño de hoja del 1%. Por lo tanto, se construye un nuevo modelo de 250 árboles y un tamaño de hoja del 1%. Las métricas son las siguientes con los dos posibles umbrales de corte.

| Threshold | Accuracy | Kappa | AUC | Sensitivity | Specificity |
|---|---|---|---|---|---|
| 0.50 | 0.6682135 | 0.2176915 | 0.6784 | 0.3296875 | 0.8680812 |
| 0.45 | 0.6629930 | 0.2231681 | 0.6784 | 0.3734375 | 0.8339483 |

*Tabla 8 Métricas de evaluación del modelo Bagging según umbral de clasificación*

Dado que las métricas con ambos umbrales resultan bastante parecidas, se elige el que mayor especificidad ofrece, dado que es el objetivo prioritario del estudio, de manera que se selecciona el punto 0,5 como umbral de selección.

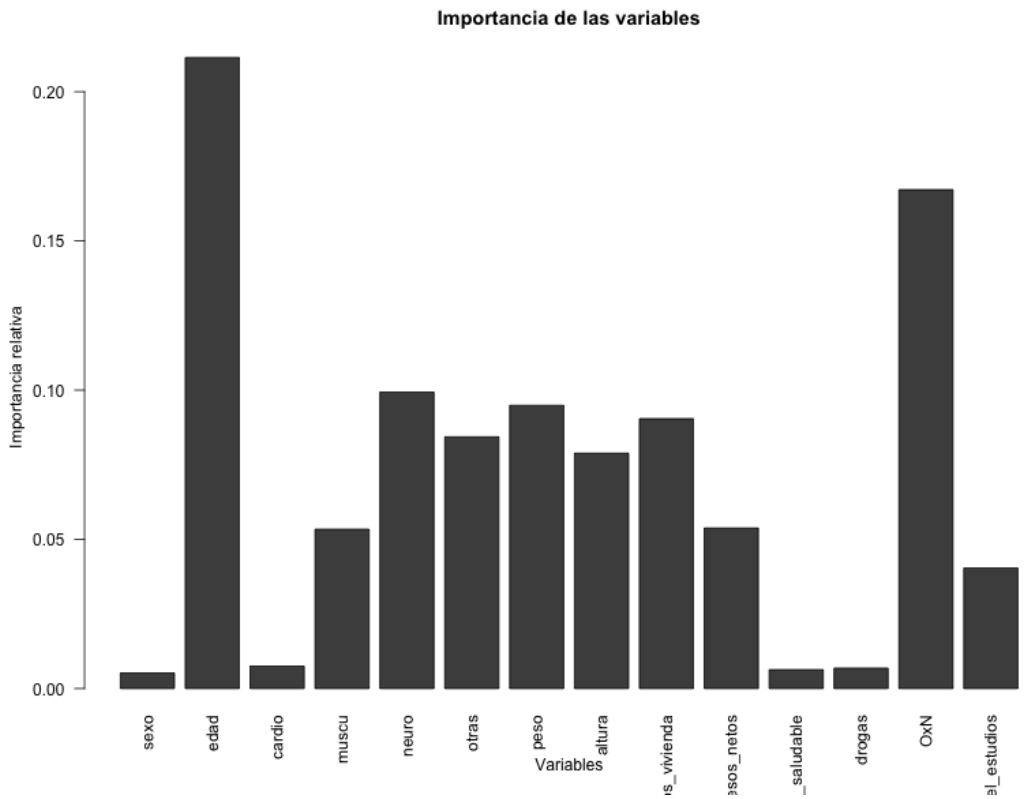


*Figura 12 Importancia de variables*

La *Figura 12* muestra que la variable edad resulta ser la más importante del modelo bagging, seguida de cerca del nivel de exposición a los óxidos de nitrógeno. Algo más lejos están un grupo de variables entre las que se incluyen el resto de enfermedades crónicas, a excepción de las cardiacas, lo que coincide con los resultados obtenidos previamente en en el modelo de regresión logística. Aunque aparecen otras variables, como son peso, altura, metros de vivienda e ingresos netos, su importancia relativa es muy baja. En conjunto, se puede afirmar que las variables más relevantes en el modelo bagging coinciden en gran medida con las seleccionadas por la regresión logística binaria, lo que refuerza la coherencia y robustez del análisis.

A continuación, se continuará profundizando en la naturaleza de la relación entre la variable edad y la variable objetivo, a partir de un gráfico PDP (Partial Dependence Plot), con el objetivo de visualizar de una forma más clara al efecto marginal de la variable regresora en cestión sobre la probabilidad de desarrollar una enfermedad respiratoria.

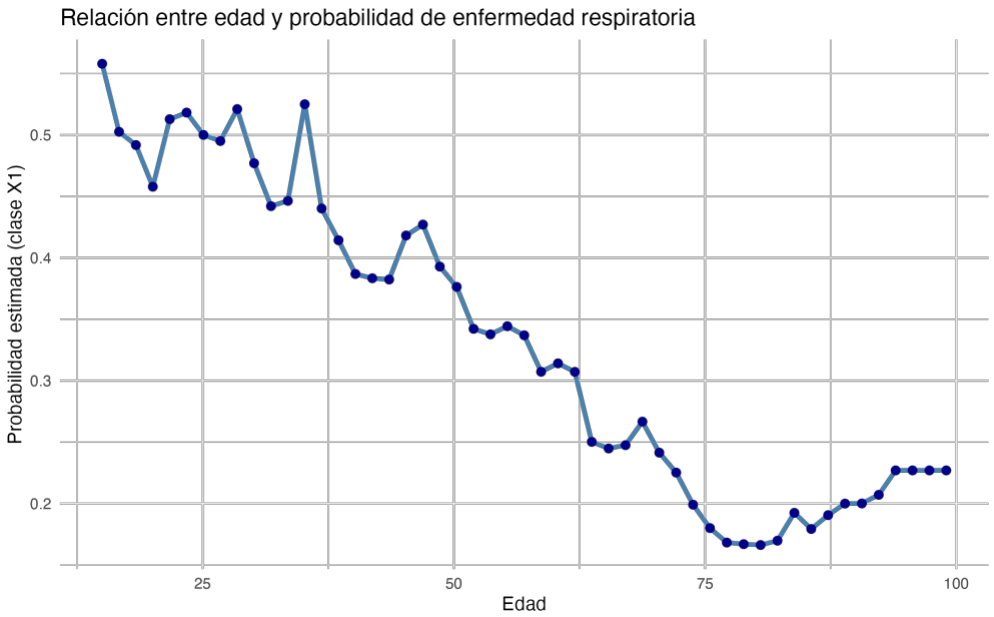


*Figura 13 Gráfico PDP entre edad y probabilidad de enfermedad respiratoria*

Observando la *Figura 13* podemos llegar a varias conclusiones:

- Grupo de edad entre 15 y 40 años: No existe un patrón muy definido, pero sí se aprecia cierta inestabilidad. En este grupo de edad, la probabilidad de contraer una enfermedad respiratoria es moderadamente alta, pero con bastante variabilidad.
- Grupo de edad entre 40 y 70 años: La probabilidad de contraer una enfermedad respiratoria disminuye progresivamente.
- Grupo de edad de mayores de 75 años: Comienza un repunte en la probabilidad de contraer una enfermedad respiratoria que crece hasta la edad de 90 años en la que se estabiliza.

El modelo, claramente, refleja una relación no lineal entre la edad y la probabilidad de contraer una enfermedad respiratoria. El gráfico muestra un riesgo moderado en las personas más jóvenes, menor en las personas de mediana edad y un aumento en las edades más avanzadas.

Este modelo ha sido capaz de profundizar en una relación no lineal entre las variables edad y probabilidad de contraer una enfermedad respiratoria que el ANOVA unifactorial no ha sido capaz detectar al incumplirse claramente, incluso a pesar de la transformación de los datos, las hipótesis de independencia, normalidad y homocedasticidad de los residuos. Además, los resultados son coherentes con los obtenidos con las pruebas no paramétricas.

## 6.5 Regresión logística binaria por estratos de edad

Se contempla la posibilidad de que la relación entre la edad y los problemas respiratorios no guarden una relación uniforme en toda la población. Ante esta posibilidad se decide realizar un análisis por cada uno de los grupos de edad. De esta manera se busca evaluar si el efecto de la edad sobre la variable respuesta se comporta de forma similar en todos los grupos. Este enfoque está basado en las recomendaciones metodológicas de Hosmer, Lemeshw & Sturdivant (Applied Logistic Regressioin, 2013), quienes sugieren la exploración de interacciones no homogéneas mediante el análisis por grupos cuando se tiene la sospecha, como es el caso, de una posible no linealidad.

Sin embargo, tras realizar el estudio con los modelos separados por grupo se pudo observar que las odds ratio apenas diferían entre sí, por lo que se concluye que la estratificación de la muestra por edad no aporta ninguna información relevante respecto al modelo global. En la tabla siguiente se muestran tanto los modelos ajustados por cada uno de los grupos de edad como las odds ratios de esta variable.

| Grupo | Modelo | OR_Edad |
|---|---|---|
| [15,40) | respi ~ sexo + edad + muscu + neuro + otras + altura + ingresos_netos + OxN | .972 |
| [40,60) | respi ~ sexo + edad + muscu + neuro + otras + peso + OxN | .976 |
| > 60 años | respi ~ edad + cardio + muscu + neuro + otras + altura + OxN + nivel_estudios | .962 |

*Tabla 9 Modelos por grupos de edad y OR de la variable Edad*

En definitiva, aunque el análisis por estratos de edad permitió la exploración de posibles diferencias en el efecto de esta variable sobre la variable respuesta, los resultados obtenidos muestran una consistencia notable en las odds ratios entre los diferentes grupos que sugiere que el comportamiento de la variable edad como predictor es muy estable en todos ellos, por lo que, tanto a nivel metodológico como de sencillez del modelo, resulta más adecuado utilizar el modelo global.

# 7 Conclusiones

Este trabajo ha desarrollado un modelo predictivo capaz de identificar el riesgo de contraer una enfermedad respiratoria a partir de variables fácilmente observables. Para alcanzar el objetivo propuesto, se ha buscado intencionadamente conseguir un modelo con alta especificidad, incluso a costa de sacrificar otras métricas. De esta manera, se consigue minimizar los falsos positivos, es decir, evitar clasificar como en riesgo a personas que no lo están. Este enfoque podría resultar especialmente útil en contextos de emergencia sanitaria, donde los recursos asistenciales son limitados y es imprescindible priorizar adecuadamente la atención médica.

Después de un proceso iterativo de selección de predictores, evaluación de métricas y validación de supuestos, se ha conseguido un modelo sorprendentemente sencillo y, sin embargo, robusto. La edad, el nivel de exposición a óxidos de nitrógeno en función del lugar de residencia, la observación de los antecedentes en enfermedades neurológicas, esquelético-musculares y otras comorbilidades resultan suficientes para conseguir una especificidad muy elevada, lo que convierte al modelo en una herramienta complementaria extremadamente útil en funciones de cribado inicial.

El modelo global mostró una relación contraintuitiva entre la edad y el riesgo de contraer enfermedades respiratorias. Se ha explorado esta relación a través de diferentes técnicas paramétricas y no paramétricas que, finalmente, han puesto de manifiesto la falta de linealidad en la relación.

Este trabajo, además de un ejercicio de modelización estadística, ha explorado una vía sobre cómo tomar decisiones cuando los recursos no son suficientes, el tiempo apremia y los datos son imperfectos. En una situación de emergencia sanitaria, la posibilidad de poder localizar con cierto grado de seguridad a quienes no están en riesgo, con el único uso de información accesible, puede marcar la diferencia entre el colapso y la eficiencia.

El desarrollo del modelo predictivo ha permitido comprobar que técnicas como la regresión logística binaria o el bagging son capaces de aportar una precisión razonable con el uso de variables básicas. Sin embargo, es preciso destacar que la

utilidad del modelo no debe ser explicada solo en términos de exactitud sino también en su capacidad para ser entendido, implementado y usado con responsabilidad.

Este análisis, tal como se ha expuesto ampliamente a lo largo del trabajo, puso de manifiesto un comportamiento inesperado en una de las variables más determinantes: la edad. A pesar de lo que intuitivamente cabría esperar, la edad no mostró una relación directamente proporcional al riesgo de contraer la enfermedad respiratoria. Este resultado sugiere la posibilidad de sesgos en los datos o efectos confusores que no se han podido determinar. Esta anomalía pone de manifiesto la necesidad de, en futuros trabajos, profundizar en la calidad de la selección de la muestra, la segmentación por grupos de edad y la interacción entre variables.

Desde una perspectiva personal, este trabajo cuenta con un valor simbólico añadido. Inicié mis estudios de estadística en la última década del siglo pasado. Entonces, la disciplina estaba muy centrada en los enfoques analíticos más tradicionales. Precisamente, en esos mismos años, Leo Breiman introdujo el bagging a nivel teórico. Esta técnica no formaba parte del currículo académico ni de la práctica estadística a nivel profesional, principalmente, por las limitaciones computacionales de la época. Con la tecnología del siglo XXI, ha pasado a convertirse en uno de los pilares sobre los que se fundamenta el machine learning. Incluir el bagging en este trabajo no ha sido solamente una elección metodológica sino una forma de cerrar un ciclo personal conectando el inicio y el fin, por el momento, de esta etapa académica, utilizando para ello la propia evolución de la Estadística.

# 8 Bibliografía y referencias


Agencia Europea de Medio Ambiente (AEMA / EEA). (2023). Air quality in Europe — 2023 report. Publications Office of the European Union. https://www.eea.europa.eu/publications/air-quality-in-europe-2023

Alonso Revenga, J. M., & Calviño Martínez, A. (2025). Introducción a la ciencia de datos con R (1.ª ed.). García Maroto Editores.

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests 1: Sensitivity and specificity. BMJ, 308(6943), 1552. https://doi.org/10.1136/bmj.308.6943.1552

Ayuntamiento de Madrid. (2021). Encuesta de Salud de la Ciudad de Madrid 2021 [conjunto de datos]. Datos Abiertos Madrid. https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=77e22cbf3ee07510VgnVCM1000001d4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default

Ayuntamiento de Madrid. (2021). Calidad del aire 2021 [conjunto de datos]. Datos Abiertos Madrid. https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=d51453c73ab16910VgnVCM1000001d4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default

Brewster, B. (1882). The Yale Literary Magazine, 47(8), 188–190.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. Journal of the Royal Statistical Society: Series B (Methodological), 26(2), 211–252.

Box, G. E. P., & Draper, N. R. (1987). Empirical model-building and response surfaces. Wiley.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140. https://doi.org/10.1007/BF00058655

Brunekreef, B., & Holgate, S. T. (2002). Air pollution and health. The Lancet, 360(9341), 1233–1242. https://doi.org/10.1016/S0140-6736(02)11274-8

Burnham, K. P., & Anderson, D. R. (2002). Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach (2nd ed.). Springer.

Delgado-Rodríguez, M., & Llorca, J. (2004). Sesgos en estudios epidemiológicos. Salud Pública de México, 46(2), 165–170. https://doi.org/10.1590/S0036-36342004000200013

Dominici, F., Peng, R. D., Bell, M. L., Pham, L., McDermott, A., Zeger, S. L., & Samet, J. M. (2006). Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. JAMA, 295(10), 1127–113. https://doi.org/10.1001/jama.295.10.1127
Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Geriatrics Editorial Office. (2024). Pulmonary diseases in older patients: Understanding and improving outcomes. Geriatrics, 9(2), Article 34. https://doi.org/10.3390/geriatrics9020034

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning (2nd ed.). Springer.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). Wiley.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: With applications in R. Springer.

Khomenko, S., Cirach, M., Pereira-Barboza, E., et al. (2021). Premature mortality due to air pollution in European cities: A health impact assessment. The Lancet Planetary Health, 5(3), e121–e134. https://doi.org/10.1016/S2542-5196(20)30272-2

Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174. https://doi.org/10.2307/2529310

Les Luthiers. (2008). Lutherapia [Espectáculo en vivo]. Buenos Aires: Les Luthiers Producciones.

Little, R. J. A., & Rubin, D. B. (2019). Statistical analysis with missing data (3rd ed.). Wiley.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. The Annals of Mathematical Statistics, 18(1), 50–60. https://doi.org/10.1214/aoms/1177730491

Marmot, M. (2005). Social determinants of health inequalities. The Lancet, 365(9464), 1099–1104. https://doi.org/10.1016/S0140-6736(05)71146-6

Menard, S. (2002). Applied logistic regression analysis (2nd ed.). Sage Publications.

Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable (2nd ed.). Leanpub.
https://christophm.github.io/interpretable-ml-book/

Montgomery, D. C. (2017). Design and analysis of experiments (9th ed.). John Wiley & Sons.

National Advisory Committee on Rural Health and Human Services. (2018). Addressing the burden of chronic obstructive pulmonary disease (COPD) in rural America: Policy brief. U.S. Department of Health and Human Services.
https://www.hrsa.gov/sites/default/files/hrsa/advisory-committees/rural/ruralcopd.pdf

Organización Mundial de la Salud. (2020). Operational considerations for case management of COVID-19 in health facility and community. World Health Organization.
https://www.who.int/publications/i/item/10665-331492

Organización Mundial de la Salud (OMS). (2021). Directrices mundiales sobre la calidad del aire: Actualización global de 2021. Ginebra: OMS.
https://www.who.int/publications/i/item/9789240034228

Parlamento Europeo y Consejo. (2019). Directiva (UE) 2019/1024 relativa a los datos abiertos y la reutilización de la información del sector público. Diario Oficial de la Unión Europea, L172/56.
https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32019L1024

Rosenbaum, L. (2020). The untold toll — The pandemic's effects on patients without Covid-19. The New England Journal of Medicine, 382(24), 2368–2371.
https://doi.org/10.1056/NEJMms2009984

Rubin, D.B. (1987) Multiple imputation for nonresponse in surveys. John Wiley & Sons.

SEFAC. (2023). Documento de posicionamiento sobre enfermedades respiratorias crónicas. El papel del farmacéutico comunitario en el abordaje integral del paciente respiratorio. Sociedad Española de Farmacia Clínica, Familiar y Comunitaria.
https://www.sefac.org/system/files/2024-02/Documento%20Respirar.pdf

Siegel, S., & Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences* (2nd ed.). McGraw-Hill.

Spearman, C. (1904). The proof and measurement of association between two things. The American Journal of Psychology, 15(1), 72–101.
https://doi.org/10.2307/1412159

Sun, Y., Song, X., Han, Y., & Zhang, Z. (2021). Predicting respiratory disease risk based on socioeconomic and environmental factors: A machine learning approach. Environmental Research, 201, 111548.
https://doi.org/10.1016/j.envres.2021.111548

Swartzwelder, J. (Guionista), & Moore, J. (Director). (1994, 6 de enero). Homer the Vigilante (Temporada 5, Episodio 11) [Episodio de serie de televisión]. En M. Groening (Productor ejecutivo), The Simpsons. 20th Century Fox.
Tabachnick, B. G., & Fidell, L. S. (2019). Using Multivariate Statistics (7th ed.). Pearson.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics,17(6), 520–525. https://doi.org/10.1093/bioinformatics/17.6.520

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. Family Medicine, 37(5), 360–363.

Wilkinson, R., & Pickett, K. (2009). The spirit level: Why more equal societies almost always do better. Allen Lane.

World Health Organization (2021). WHO global air quality guidelines: Particulate matter ($PM_{2.5}$ and $PM_{10}$), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. https://www.who.int/publications/i/item/9789240034228

**ANEXO I**. Ficha técnica de la Encuesta de Salud de la Ciudad de Madrid 2021

| *Elemento* | *Descripción* |
|---|---|
| ***Promueve*** | Organismo Autónomo Madrid Salud |
| ***Objetivo*** | Conocer los principales problemas de salud de la población madrileña y evaluar las desigualdades sociales y territoriales en salud. |
| ***Universo*** | Hogares de la ciudad de Madrid con al menos una persona residente habitual de 15 o más años. En caso de no poder responder, se utilizó un proxy. |
| ***Diseño muestral*** | Muestreo aleatorio estratificado |
| ***Criterios de estratificación*** | Distrito, sexo, edad y país de origen |
| ***Tamaño muestral*** | 8.625 entrevistas válidas (400 por distrito) |
| ***Error muestral*** | ±1,5% para el conjunto de la ciudad y ±5% por distrito (nivel de confianza del 95,5%, p = q = 50%) |
| ***Trabajo de campo*** | Del 7 de octubre al 14 de diciembre de 2021 |
| ***Método de recogida*** | Entrevistas telefónicas asistidas por ordenador (CATI) |
| ***Cuestionario*** | Dos versiones con un bloque común y bloques diferenciados |
| ***Temáticas abordadas*** | Salud (morbilidad, salud mental, COVID-19); cuidados (medicación, vacunación, sistema sanitario); hábitos de vida; aspectos sociales |
| ***Entidad ejecutora*** | Demométrica S.L. |
| ***Presupuesto*** | 101.502,33 € |

**ANEXO II**. Variables originales seleccionadas

A continuación, se muestran las variables originales utilizadas en el estudio.

| Variable |
|---|
| DISTRITO |
| EDAD |
| A1. ¿Me podría decir con qué género se identifica? |
| B1. En los últimos doce meses, ¿diría que su estado de salud ha sido... |
| C1. Su médico le ha dicho que la padece o la ha padecido en los últimos doce meses - 6. Alergia crónica, como rinitis, conjuntivitis o dermatitis alérgica, alergia alimentaria o de otro tipo (asma alérgica excluida) |
| C1. Su médico le ha dicho que la padece o la ha padecido en los últimos doce meses - 7. Asma |
| C1. Su médico le ha dicho que la padece o la ha padecido en los últimos doce meses - 10. Depresión |
| C1. Su médico le ha dicho que la padece o la ha padecido en los últimos doce meses - 11. Ansiedad crónica |
| C1. Su médico le ha dicho que la padece o la ha padecido en los últimos doce meses - 12. Migraña o dolor de cabeza frecuente |
| C1. Su médico le ha dicho que la padece o la ha padecido en los últimos doce meses - 17. Síndrome post COVID/COVID persistente |
| C2. ¿Ha sido usted diagnosticado/a de infección por coronavirus? |
| G1. ¿Cuánto tiempo lleva residiendo en el municipio de Madrid? AÑOS |
| G3. ¿Cuál es el mayor nivel de estudios que ha completado? |
| G4. ¿Qué número de años ha pasado usted en el sistema educativo? |
| G5. ¿En cuál de las siguientes situaciones respecto al empleo se encuentra Ud. actualmente? |
| G14. ¿Cuál es la categoría profesional que tiene o tenía en la empresa donde trabaja o trabajaba? |
| G16. ¿Trabaja actualmente? |
| G26. ¿Puede permitirse mantener la vivienda con una temperatura adecuada durante todo el año? |
| E3. Dígame si consume... - Fruta fresca (excluyendo zumos) |
| E3. Dígame si consume... - Verduras |
| E3. Dígame si consume... - Legumbres |
| E3. Dígame si consume... - Carne, Pescado |
| E3. Dígame si consume... - Huevos |
| E3. Dígame si consume... - Leche y/o derivados lácteos |
| E3. Dígame si consume... - Dulces y/o bollería |
| E3. Dígame si consume... - Refrescos y/o zumos azucarados |
| E3. Dígame si consume... - Comida rápida (pizzas, hamburguesas…) |
| E4. ¿Es usted Vegano o Vegetariano? |
| E5. ¿Podría decirme si fuma tabaco actualmente? |
| E6. ¿Utiliza usted cigarrillos electrónicos u otros dispositivos de administración de nicotina? |
| E7. A continuación, valora de 0 a 3 las siguientes actividades recientes de la agencia SINC. - CANNABIS / MARIHUANA / HACHÍS |
| E8. ¿Con qué frecuencia toma alguna bebida alcohólica? |
| E9. ¿Cuántas bebidas alcohólicas consume normalmente cuando bebe? |
| E10. ¿Con qué frecuencia toma 6 o más bebidas alcohólicas en un solo día? |
| C6 - LO HA TOMADO... Tranquilizantes, ansiolíticos o medicación para dormir |
| C6 - RECETADO... Tranquilizantes, ansiolíticos o medicación para dormir |
| C6 - LO HA TOMADO... Antidepresivos |
| C6 - RECETADO... Antidepresivos |
| C6 - LO HA TOMADO... Medicamentos fuertes para el dolor |
| C6 - RECETADO... Medicamentos fuertes para el dolor |
| D1. Forma Física: Durante las dos últimas semanas, ¿cuál ha sido la máxima actividad física que pudo realizar durante al menos dos minutos? |
| D2. Sentimientos: Durante las dos últimas semanas, ¿en qué medida le han molestado los problemas emocionales tales como sentimientos de ansiedad, depresión, irritabilidad o tristeza y desánimo? |
| E.3. ¿Podría indicarme, aproximadamente, cuántas horas duerme habitualmente al día? |
| G29. ¿Podría decirme cuál de los intervalos siguientes representa mejor el ingreso mensual neto de todo su hogar, tras las deducciones por los impuestos, Seguridad Social, etc.? |

*Tabla 10 Variables originales seleccionadas*

**ANEXO III**. Datos medios anuales de calidad del aire de Madrid referidos a 2021

A continuación, se muestran los datos medios anuales de óxidos de Nitrógeno de la ciudad de Madrid, referidos a 2021, recogidos por la técnica de quimioluminiscencia, expresados en µg/m3, desglosados por distritos.

| | DISTRITO | Monóxido de Nitrógeno | Dióxido de Nitrógeno | Óxidos de Nitrógeno |
|---|---|---|---|---|
| 1 | Centro | 22,85981 | 29,82854 | 59,13216 |
| 2 | Arganzuela | 10,942 | 28,03803 | 44,73697 |
| 3 | Retiro | 8,473931 | 28,11495 | 41,09039 |
| 4 | Salamanca | 8,473931 | 28,11495 | 41,09039 |
| 5 | Chamartín | 10,86208 | 31,85439 | 48,46859 |
| 6 | Tetuán | 12,61514 | 31,72964 | 51,02161 |
| 7 | Chamberí | 6,964192 | 28,51481 | 39,20808 |
| 8 | Fuencarral-El Pardo | 6,957572 | 23,60307 | 34,27268 |
| 9 | Moncloa-Aravaca | 4,234553 | 16,96243 | 23,46099 |
| 10 | Latina | 9,545702 | 27,07796 | 41,74066 |
| 11 | Carabanchel | 20,26517 | 39,37102 | 70,45688 |
| 12 | Usera | 20,26517 | 39,37102 | 70,45688 |
| 13 | Puente de Vallecas | 9,432935 | 28,99705 | 43,51171 |
| 14 | Moratalaz | 9,045248 | 30,14523 | 43,98538 |
| 15 | Ciudad Lineal | 9,093499 | 27,119 | 40,99422 |
| 16 | Hortaleza | 7,869945 | 25,02517 | 37,14424 |
| 17 | Villaverde | 19,77697 | 34,31459 | 64,52134 |
| 18 | Villa de Vallecas | 11,40845 | 26,2518 | 43,72461 |
| 19 | Vicálvaro | 11,40845 | 26,2518 | 43,72461 |
| 20 | San Blas-Canillejas | 9,093499 | 27,119 | 40,99422 |
| 21 | Barajas | 8,530303 | 26,51194 | 39,61869 |

*Tabla 11 Datos medios anuales de óxidos de nitrógeno en µg/m3*