Máster en dirección de sistemas y TIC de la salud y en digitalización sanitaria

Curso académico 2022-2023

Trabajo de Fin de Máster

Estimación del sesgo de género en la frecuencia de ingresos hospitalarios en la red de asistencia sociosanitaria entre los años 2020 y 2022 mediante el uso de Inteligencia Artificial

Autores:

Iria Rodríguez Cobo

Roberto González Novas

Susana Beatriz Nájera Cano

Tutor:

D. Manuel José Fernández Iglesias

Octubre 2023

Trabajo de Fin de Máster

Tipo: Proyecto de investigación

Título: Estimación del sesgo de género en la frecuencia de ingresos hospitalarios en la red de asistencia sociosanitaria entre los años 2020 y 2022 mediante el uso de AI

Autoría: Iria Rodríguez Cobo, Roberto González Novas y Susana Beatriz Nájera Cano

Tutoría: D. Manuel José Fernández Iglesias

fecha y firma de los alumnos

FERNANDEZ IGLESIAS MANUEL JOSE - 36073249A
2023.10.12 12:13:37 +02'00'

Iria Rodríguez Cobo	Roberto González Novas	Susana Beatriz Nájera Cano
Firmado por RODRIGUEZ COBO IRIA - ***9803** el día 10/10/2023 con un certificado emitido por AC FNMT Usuarios	GONZALEZ NOVAS ROBERTO - 53114410G Firmado digitalmente por GONZALEZ NOVAS ROBERTO - 53114410G Fecha: 2023.10.11 01:50:31 +02'00'	Firmado por NAJERA CANO SUSANA BEATRIZ - 12381037E el día 10/10/2023 con un certificado emitido por AC FNMT Usuarios

Índice

1 Resumen

El objetivo de este estudio es la estimación del sesgo de género en la prevalencia de derivaciones hospitalarias en la red de asistencia de residencias sociosanitarias entre los años 2020 y 2022 mediante el uso comparativo de la IA con las técnicas estadísticas convencionales. A partir de la base de datos de residencias de mayores de *DomusVi* se obtienen las variables a analizar en la población de estudio mayor de 59 años. Mediante la elaboración de modelos predictivos se obtendrán resultados ajustados por las variables sociodemográficas, de estado funcional, comorbilidades y tratamientos recibidos. Además de la elaboración de un modelo general, se entrenarán los modelos de forma separada en hombres y en mujeres para comparar los resultados pronosticados con los observados al cruzar ambos modelos, como estimación del sesgo de género en la hospitalización. En una segunda fase se plantea la integración de estas bases de datos sociosanitarias con el Registro de Actividad de Atención Especializada (RAE-CMBD) del Ministerio de Sanidad para enriquecer el estudio con nuevas variables resultado como los ingresos en UCI o los procedimientos y/o tratamientos recibidos.

2 Introducción justificativa de los objetivos

En el ámbito sanitario el sesgo de género es el planteamiento erróneo de igualdad o de diferencia entre hombres y mujeres que puede generar una conducta desigual en la atención sanitaria y un resultado discriminatorio para un sexo con respecto al otro. La investigación realizada en los últimos 30 años ha descrito la existencia de un sesgo en detrimento del sexo femenino, entre otros, en la investigación básica, en los ensayos clínicos y en la intensidad terapéutica (utilización hospitalaria, aplicación de procedimientos terapéuticos, demora de la asistencia sanitaria, prescripción y consumo de fármacos). Una de las áreas más estudiadas en cuanto al denominado sesgo de género es la de la enfermedad cardiovascular.

A partir de la crisis sanitaria provocada por el COVID-19 algunas investigaciones señalan la existencia de un sesgo de género en la asistencia sanitaria con un mayor impacto en las poblaciones vulnerables como son las personas mayores institucionalizadas. Parece ser que también con el COVID-19, al igual que se viene demostrando desde los años 90 con otras enfermedades, la intensidad diagnóstica y terapéutica ha sido menor en mujeres que en hombres ante situaciones clínicas similares. El análisis de las diferencias objetivas entre sexos (en las hospitalizaciones y los ingresos en UCI) y la posibilidad de ajuste por diversas características clínico-epidemiológicas en el contexto de bases de datos clínico-administrativas, nos permite una aproximación a la cuestión de hasta qué punto la variable sexo ha influido en los resultados en salud durante la reciente pandemia.

El planteamiento inicial del estudio fue seleccionar, dentro de todos los usuarios ingresados en los centros sociosanitarios de DomusVi, a los afectados por COVID y, de ellos, medir como variable resultado las hospitalizaciones por COVID. Sin embargo, un análisis preliminar nos mostró que el tamaño muestral obtenido de esta forma (15.936 personas con COVID en los tres años de estudio, de la cuales apenas 929 llegaron a ingresar en el hospital) era insuficiente para poder realizar predicciones, tanto con estadística convencional como con analítica avanzada, puesto que se trataba de un evento de muy baja frecuencia. Fue entonces cuando decidimos ampliar el foco del estudio y analizar el sesgo de género en las hospitalizaciones por todas las causas de todas las personas de la base de datos.

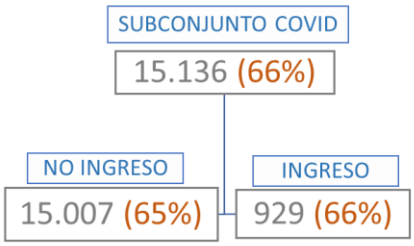


Imagen 1: Tamaño muestral del subconjunto COVID (entre paréntesis, el porcentaje de mujeres)

El grupo DomusVi es una empresa de atención a la dependencia que cuenta con centros de atención a personas mayores, de atención a la discapacidad, a la salud mental y centros de día. En el año 2022 contaba con 148 residencias de personas mayores con un total de 19.561 camas. Aunque cuenta con centros a lo largo de toda la geografía española, las Comunidades Autónomas -CCAA- donde más presencia tiene son la Comunidad Autónoma de Madrid, Galicia, Comunidad Valenciana, Cataluña y Andalucía.

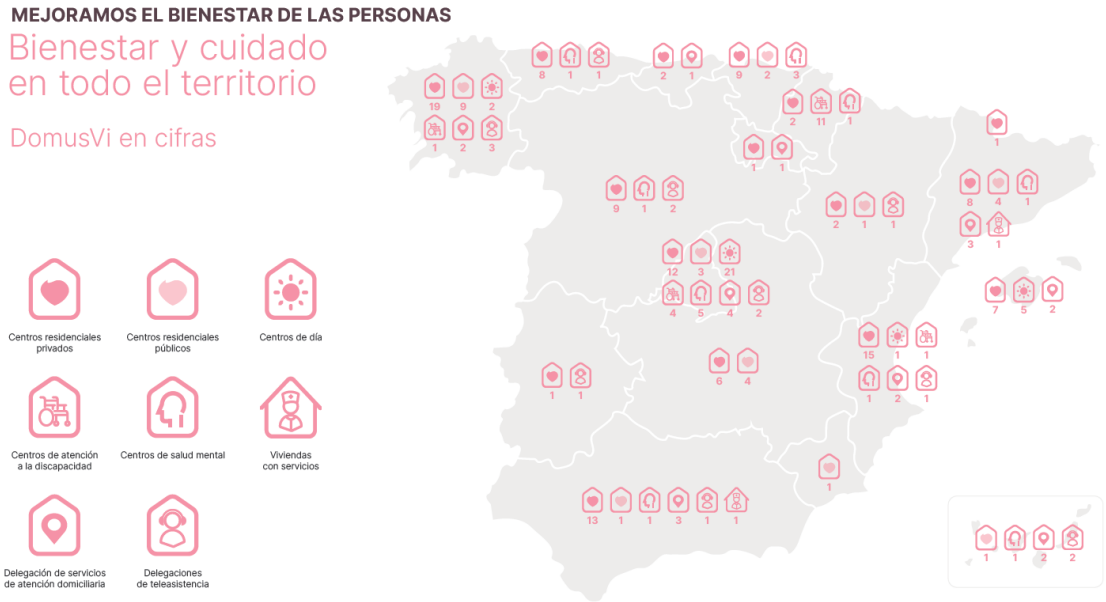


Imagen 2: Distribución de centros DomusVi por CCAA

DomusVi es el mayor grupo residencial en número de centros y plazas en España. Hay que destacar que las residencias de mayores es un sector muy atomizado. Según el último informe del sector geriátrico publicado por Alimarket para los años de este estudio en febrero 2022 había 378.465 camas disponibles en los 5.252 centros geriátricos operativos. Los diez primeros operadores del mercado aglutinan el 20% del total de camas geriátricas, lo que refleja la estructura altamente atomizada del mismo. En la siguiente tabla se muestran los principales grupos geriátricos por número de camas, donde se observa que DomusVi representa el 25,48% de sus plazas.

	Empresa/Grupo	Nº de Centros	Nº Camas	%
1	**DomusVi España**	**148**	**19.561**	**25,94%**
2	Grupo Vitalia Home	58	8.771	11,63%
3	Orpea Ibérica	56	8.732	11,58%
4	Grupo Ballesol	52	7.555	10,02%
5	Grupo Colisée España	56	6.524	8,65%
6	Grupo Amavir	43	6.390	8,48%
7	Asoc. Edad Dorada-Mensajeros de la Paz	101	3.628	4,81%
8	Sanitas Mayores	44	6.112	8,11%
9	Grupo Clece – División de Mayores	64	4.581	6,08%
10	Fundación San Rosendo	22	3.543	4,70%
	T O T A L	**644**	**75.397**	

Tabla 1: Principales grupos geriátricos por número de camas (febrero 2022)

2.1 Objetivo principal

Estudiar y estimar el sesgo de género en la prevalencia de hospitalización en personas ingresadas en la red de asistencia de residencias sociosanitarias, ajustando por las variables sociodemográficas, comorbilidades, estado funcional y tratamientos recibidos.

2.2 Objetivos secundarios

Serían objetivos secundarios de este estudio:

- Evaluar la calidad de los datos sociosanitarios y ahondar en el conocimiento de las características epidemiológicas de la población de las residencias de personas mayores.
- Explorar la factibilidad a corto-medio plazo de enriquecer esta información sociosanitaria con otras bases de datos clínico-administrativas.
- Promover la estandarización de los protocolos de compartición de datos entre los ámbitos sanitario y sociosanitario, explorando el camino de la integración tanto para uso primario como secundario.
- Analizar el desempeño de la Inteligencia Artificial y concretamente el aprendizaje automático (*machine learning*) en bases de datos del mundo real (RWD).
- Comprobar si el análisis mediante técnicas de Inteligencia Artificial mejora los resultados de las regresiones logísticas de la estadística convencional.
- Conocer de primera mano las tareas concretas involucradas en la implementación de proyectos de Inteligencia Artificial para que sirva como

referencia a la hora de evaluar otros proyectos o en la dirección de proyectos de este tipo.

3 Contenidos del temario

Los contenidos del temario que utilizados en la elaboración de este estudio son los siguientes temas de la IX edición del máster en dirección de sistemas y TIC para la salud y en digitalización sanitaria:

Tema 2.5 La seguridad TIC. Legislación aplicable. Aplicación del Reglamento General de Protección de Datos. El papel del Delgado de Protección de Datos. Auditorias. Metodologías / Herramientas de Seguridad MAGERIT. SGSI. PILAR.

Tema 2.8 Cloud Computing. BIG DATA. Inteligencia Artificial. Casos de uso de la Inteligencia Artificial. Infraestructuras de procesamiento. Servicios de apoyo.

Tema 2.9 Entornos / Metodologías / Plataformas de desarrollo. Desarrollos en el ámbito científico. Entornos Bibliográficos. Gestores de contenidos / documentales. Repositorios. Control de versiones. Desarrollos para movilidad. Herramientas de *back end*.

Tema 3.1 Los Sistemas de Información de Salud y Socio Sanitarios. Estrategia y Gobernanza.

Tema 3.2 La aplicación de la normativa de Protección de Datos en el sector Salud.

Tema 3.3 La Interoperabilidad en al ámbito de la Salud.

Tema 3.7 Las TIC y la continuidad asistencial.

Tema 3.8 Sistemas de Información para la Salud Pública.

Tema 3.10 Estrategias, infraestructuras y aplicaciones avanzadas basada en datos para la Investigación en Salud y Biomedicina.

Tema 4.7 Aplicaciones en asistencia sanitaria e investigación.

Tema 4.8 Analítica y modelos predictivos en salud.

4 Metodología

4.1 Justificación del abordaje metodológico

La hipótesis de trabajo, basada en la revisión de la literatura científica, es la existencia de un sesgo de género en cuanto a la intensidad diagnóstica y terapéutica, superior en los hombres que en las mujeres.

El abordaje crudo de la pregunta de investigación, como la comparación directa del porcentaje bruto de ingresos entre hombres y mujeres, no resulta adecuado debido a las características demográficas de este grupo poblacional mayor de 59 años, muy sesgado hacia las mujeres en estas edades de la vida, puesto que representan aproximadamente el 60% de la población de estudio. Además, debido a una mayor esperanza de vida de las mujeres, y al efecto de la supervivencia de los hombres más sanos, el perfil de morbilidad no sería comparable entre ambos sexos.

Por todo esto, se considera necesario realizar un ajuste por aquellas variables sociodemográficas y, sobre todo, por las morbilidades y comorbilidades que presentan cada una de las personas. Este ajuste nos permitirá concluir si, ante la misma combinación de diagnósticos y tratamientos, el resultado de hospitalizaciones es diferente en hombres y en mujeres.

El abordaje utilizado para conseguir ajustar por todas las variables sociodemográficas y clínicas de interés es utilizar un modelo que nos permita obtener el efecto en la variable resultado, hospitalizaciones, ajustando por las variables explicativas que introducimos en el mismo. Entrenando el modelo en la base de datos de hombres y posteriormente obteniendo las predicciones de este modelo de hombres en la base de datos de mujeres podemos comparar si existen diferencias entre los ingresos reales (observados) en mujeres y los esperados (predichos) por el modelo entrenado en hombres. Esto nos dará una estimación del sesgo de género ajustado por morbilidad además de por otras variables sociodemográficas y de estado funcional disponibles en la base de datos.

Este análisis se realizó con dos metodologías de forma paralela, por un lado, con la estadística convencional, mediante regresiones logísticas, y por el otro con inteligencia artificial, con técnicas de aprendizaje automático (*machine learning)*. Esto nos ha servido para comprobar si el análisis con de inteligencia artificial mejora los resultados de las regresiones logísticas y que información adicional nos aporta cada una de las metodologías utilizadas.

4.2 Población de estudio

Personas mayores de 59 años ingresadas en la red de asistencia de residencias sociosanitarias durante los años 2020 a 2022. Se utilizará la base de datos de residencias

de mayores *DomusVi* de la que obtendremos la selección de la población de estudio y sus variables predictoras, así como la variable resultado del ingreso hospitalario.

En una segunda fase, esta base de datos se fusionará con los datos del Registro de Actividad de Atención Especializada (RAE-CMBD), que serán solicitados al Ministerio de Sanidad para obtener otras variables resultado del análisis (ingresos en UCI, tratamientos y procedimientos médicos).

4.3 Bases de datos

La base de datos asistencial de DomusVi almacena los datos sociosanitarios de más de 176.500 personas, de las cuales más de 170.000 corresponden a personas mayores institucionalizadas. De dicha base de datos obtuvimos la selección de la población de estudio (personas mayores de 59 años ingresadas en los años 2020, 2021 y 2022) y sus variables predictoras, así como la variable resultado del ingreso hospitalario.

A continuación, se detallan cada una de las tablas generadas mediante consultas SQL que componen el origen de datos para este estudio:

1. Genérico: Datos genéricos de los usuarios. 64.421 registros

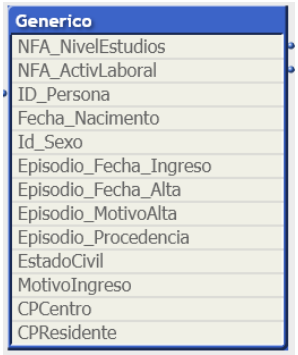


Imagen 3: Tabla con los datos genéricos de los residentes

2. Positivos: Usuarios que fueron positivo COVID en el período del estudio. 15.939 registros

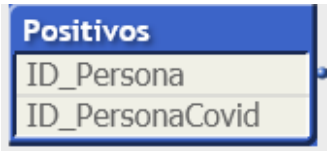


Imagen 4: Tabla con campo de usuarios positivos de COVID

3. Resultado_COVID_Filtrado: registro de los test de COVID. 230.552 registros

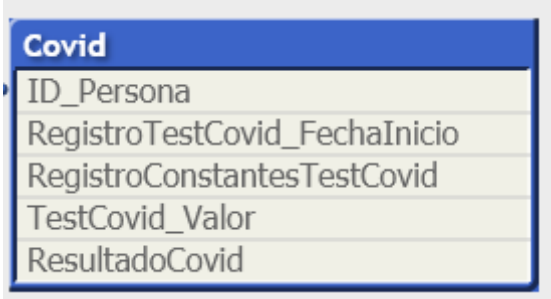


Imagen 5: Tabla con campos de resultados de test COVID

4. Vacunas: Fecha y tipo de vacunas suministradas. 94.114 registros

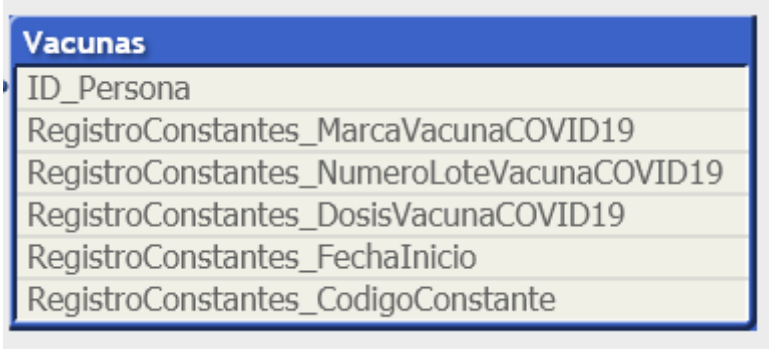


Imagen 6: Tabla con campos relativos a las vacunas COVID recibidas

5. Diagnósticos: Diagnósticos de los usuarios (todos los diagnósticos, incluido COVID). 285.771 registros

Imagen 7: Tabla con campos de Diagnósticos

6. MiniMental: Evolución de la escala MiniMental para valorar el deterioro cognitivo. 35.683 registros

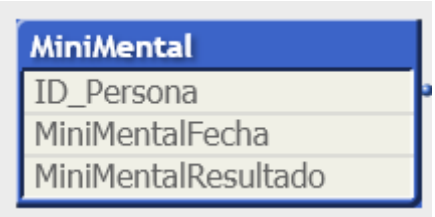


Imagen 8: Tabla con campos de los resultados de la escala MiniMental

7. IMC: Evolución del índice de masa corporal. 307.037 registros

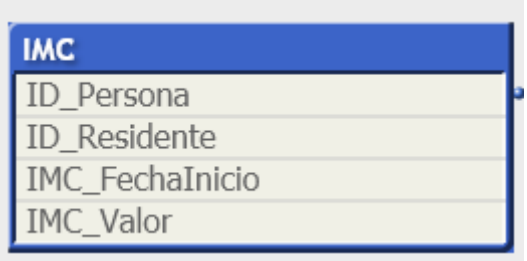


Imagen 9: Tabla con campos del IMC

8. Barthel: Evolución de la escala Barthel para valorar la capacidad funcional. 146.969 registros

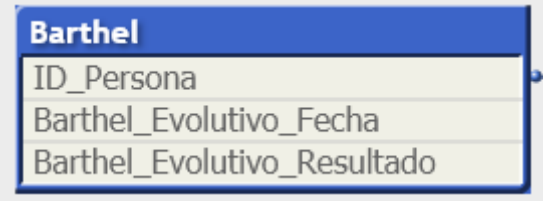


Imagen 10: Tabla con campos de los resultados de la escala Barthel

9. Hospital_COVID: Registro de salidas al hospital por COVID. 1.173 registros

Imagen 11: Tabla con los registros de salida al hospital

10. Salidas_Residentes: Registro de salidas de la residencia 742.178 registros

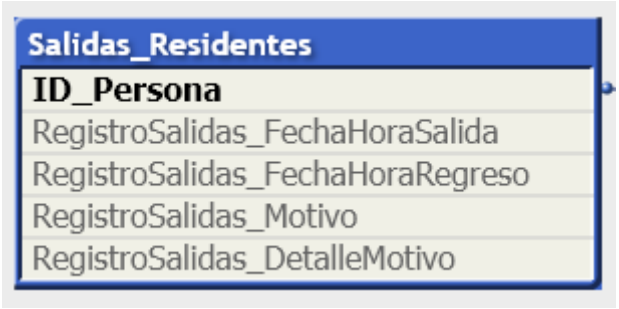


Imagen 12: Tabla con los registros de salidas del residente del Centro

11. Plan Farma: Registro de medicamentos administrados. 1.048.575 registros

Imagen 13: Tabla con los campos del Plan Farmacéutico

12. Medicamentos y Grupos: Equivalencia de medicamentos. 56.845 registros

Imagen 14: Tabla con campos de los Medicamentos pautados

13. Actividad Laboral: equivalencia Actividad Laboral. 18 registros

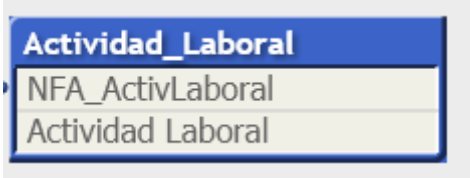


Imagen 15: Tabla con los campos de Actividad Laboral

1

14. Nivel de Estudios: equivalencia Nivel de Estudios. 12 registros

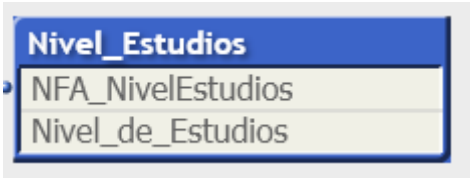


Imagen 16: Tabla con los campos de Nivel de Estudios

En el gráfico siguiente se detallan todas las relaciones entre las tablas extraídas mediante consultas SQL de la base de datos de DomusVi.

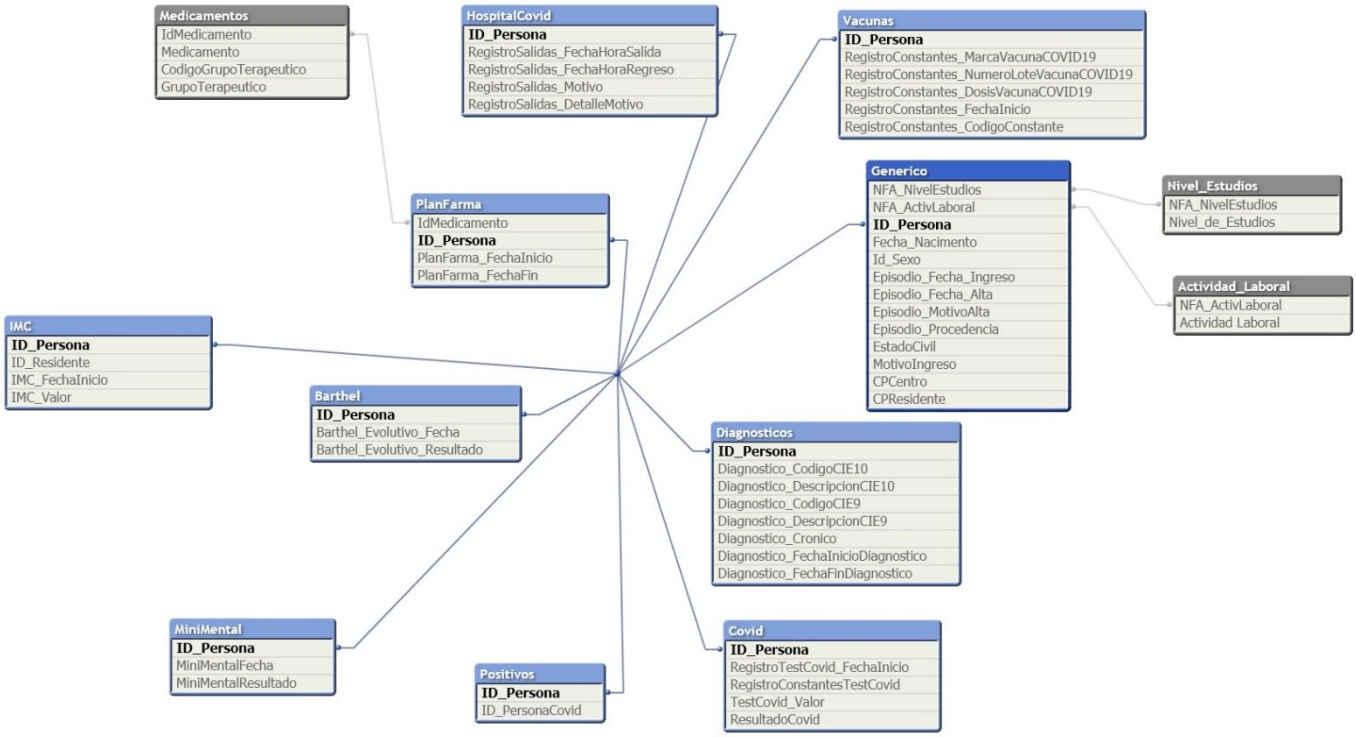


Imagen 17: Modelo relacional de las tablas extraídas de la base de datos DomusVi

4.4 Variables predictoras

Para cada una de las personas incluidas en el estudio se incluyeron como predictoras las siguentes variables:

- Sociodemográficas: edad, sexo, nivel de estudios, situación laboral, estado civil y comunidad autónoma de origen.
- Estado funcional y de salud mental (puntuación en los índices de Barthel y MiniMental).
- Diagnósticos crónicos (codificados en CIE-10-ES).
- Tratamientos farmacológicos: número de administraciones por grupo terapéutico (ATC) y dosis totales recibidas por persona.
- Número de vacunas COVID recibidas durante el período de análisis.
- Número de positivos a COVID para cada una de las personas en el mismo período.

Con la intención de incluir una variable de ajuste que pudiera ser un *proxy* del exposoma o conjunto de exposiciones a sustancias químicas ambientales, se decidió extraer la variable código postal. Sin embargo, por motivos de protección de datos, esta variable se agrupó en la más genérica CCAA que, aunque en cierta forma también puede medir

el exposoma, se trata sobre todo de un indicador que permite comparar los sistemas sociosanitarios de cada una de las CCAA.

4.5 Variables resultado

A partir del campo "Salidas_residentes" se obtuvieron las hospitalizaciones de las personas de la base de datos *DomusVi* que conformaron la variable resultado de interés: número de hospitalizaciones, denominada "Total_hospital". En una segunda fase del estudio se plantea el enriquecimiento de los datos con la información clínico-administrativa del Ministerio de Sanidad del Conjunto Mínimo Básico de Datos de Atención Especializada, denominado RAE-CMBD. En este caso no sólo se utilizarían las variables predictoras sociosanitarias si no que se obtendría información del ingreso hospitalario (RAE-CMBD) como los diagnósticos del episodio en cuestión, así como tratamientos y procedimientos realizados y el ingreso en UCI, que también podrán ser explorados como variables resultado.

4.6 Fases de extracción, transformación, carga (ETL) y ensamblaje

En la fase de extracción se obtienen los datos desde sus fuentes originales, que podrían ser bases de datos, archivos CSV, APIs, etc. En esta parte del proceso las herramientas utilizadas son:

- Motores SQL para extraer datos de bases de datos relacionales.
- Python con bibliotecas como "Pandas" para manejar los datos tabulares.
- R para ensamblaje dentro del propio script de modelado.

En la fase de transformación se aplican diversas transformaciones a los datos para prepararlos y limpiarlos para su posterior análisis. Las tareas comunes incluyen:

- Limpieza de datos: Tratar valores nulos, valores atípicos, etc.
- Transformaciones de datos: Cambio de formatos, cálculos de nuevas columnas, etc. En nuestro trabajo se utiliza la librería de *tidyverse* y asociadas.
- Integración de datos: Unir tablas mediante claves comunes.

En la fase de carga, los datos transformados se cargan en una estructura centralizada, como un único conjunto de datos o un almacén de datos. En general, se pueden utilizar diversos enfoques:

- Bases de datos: PostgreSQL, MySQL, SQLite, etc.
- Almacenes de datos: Apache Parquet, Apache Avro, etc.
- Archivos: CSV, JSON, etc.

De todos ellos, en este trabajo se opta por utilizar un fichero CSV.

Imagen 18: Esquema del proceso de extracción y ensamblaje

4.7 Seguridad y anonimización de los datos

Para proteger la privacidad de los datos personales, es esencial aplicar técnicas de anonimización antes de compartir o analizar los datos. Algunas técnicas incluyen:

- Perturbación de datos: Agregar ruido a los valores numéricos.
- Máscaras: Reemplazar algunos dígitos de números identificativos.
- Generalización: Reducción de detalles (por ejemplo, sustituir edades exactas por rangos de edad).

Las técnicas de anonimización deben asegurar los siguientes aspectos:

- Integridad de los datos: Asegurarse de que los datos sean precisos y coherentes. Esto implica la detección y corrección de valores atípicos y valores nulos.
- Normalización y limpieza: Siguiendo principios de normalización de bases de datos, se trata de organizar los datos de manera eficiente, eliminando redundancias y evitando anomalías.
- Privacidad y anonimización: Se debe garantizar que los datos no contengan información personal identificable. Esto implica comprender las regulaciones de privacidad relevantes, como el RGPD.
- Calidad de los datos: Asegurarse de que los datos sean útiles y confiables para el análisis. Esto incluye la validación de los datos para garantizar su coherencia.
- Escalabilidad: Garantizar que las herramientas y procesos utilizados sean escalables para manejar grandes volúmenes de datos.

Para garantizar el cumplimiento del RGPD, además de la pseudoanonimización de los identificadores de los residentes, tras un análisis de riesgos centrado en los requerimientos, se evitó el tratamiento de las variables más sensibles no necesarias para el análisis, como datos personales o la fecha de nacimiento, utilizando el principio de

minimización de la información compartida a aquella estrictamente necesaria para resolver la pregunta de investigación. La asignación códigos de identificación se aplicó utilizando un algoritmo irreversible. El valor del código postal se agrupó por CCAA y los diagnósticos y tratamientos fueron también agrupados en los valores jerárquicos más generales: a dos caracteres en el caso de los diagnósticos y a tres en el caso de los ATC (*Anatomical, Therapeutic and Chemical classification system*). Estas últimas medidas de anonimización, también nos han servido para simplificar el análisis al reducir la cardinalidad de las variables de diagnósticos y medicamentos. En una segunda fase, para realizar la fusión con los datos del RAE-CMBD, se ha de obtener el campo CIPautonómico_SNS mediante anonimización requerida por el Ministerio de Sanidad para permitir el cruce de las bases de datos.

5 Análisis de los datos

5.1 Análisis exploratorio y tratamiento de la calidad del dato

Se realizó un estudio de la calidad del dato, detallando variable a variable los problemas de calidad y tomando las decisiones de tratamiento de los datos en función del conocimiento funcional y de la pregunta de investigación. Se adoptó un enfoque pragmático y contextual de análisis de la calidad del dato, tal y como sugiere la literatura científica al respecto. En el anexo 1 se detallan los resultados más importantes de este análisis exploratorio de la calidad del dato y en la imagen 20 se muestran los resultados de los tratamientos recibidos con respecto al tamaño de la muestra. En cuanto a la edad, se limitó el análisis a los mayores de 59 años, tal y como viene implícito en la pregunta de investigación, lo que supuso una pérdida de tamaño muestral de 8.750 registros (imagen 20).

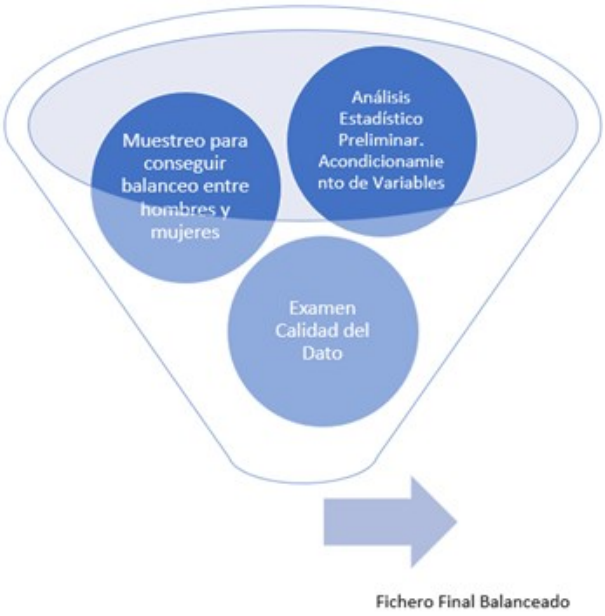


Imagen 19: Esquema del tratamiento de los datos

Para valorar la proporción de personas con valores perdidos en las variables de ajuste más importantes se creó una variable intermedia denominada “DataQ”. Esta variable contabiliza la presencia o ausencia de valores en las variables explicativas funcionalmente más destacadas: escala Barthel y MiniMental, número de administraciones de medicamentos y número total de diagnósticos. No se incluyeron las variables sociodemográficas, también muy importantes, porque se comprobó que estas sí estaban siempre registradas. Así, un valor de DataQ=0 en un registro indica que esa

persona no tiene ninguna información registrada ni en diagnósticos, ni en tratamientos ni en escalas funcionales.

Se excluyeron aquellos registros de muy baja calidad (DataQ<5). El número total de registros eliminados por este criterio más otros problemas de calidad detectados ascendió a 7.435 registros. Esta medida contribuyó a balancear ligeramente la muestra puesto que el número de casos con baja calidad de registro resultó considerablemente superior en mujeres que en hombres. Dado que nuestro análisis se centra en las diferencias por sexos esta variable debía estar suficientemente balanceada. Por eso, a mayores, se eliminó una muestra de conveniencia de 2.089 mujeres con peor calidad (DataQ=5), hasta obtener unas cifras más balanceadas y de paso equilibrar las diferencias de calidad del dato entre hombres y mujeres. Nos quedamos entonces con una N de 37.190 casos, ligeramente más balanceada con un 57% de mujeres, pues partíamos de un 63% de mujeres.

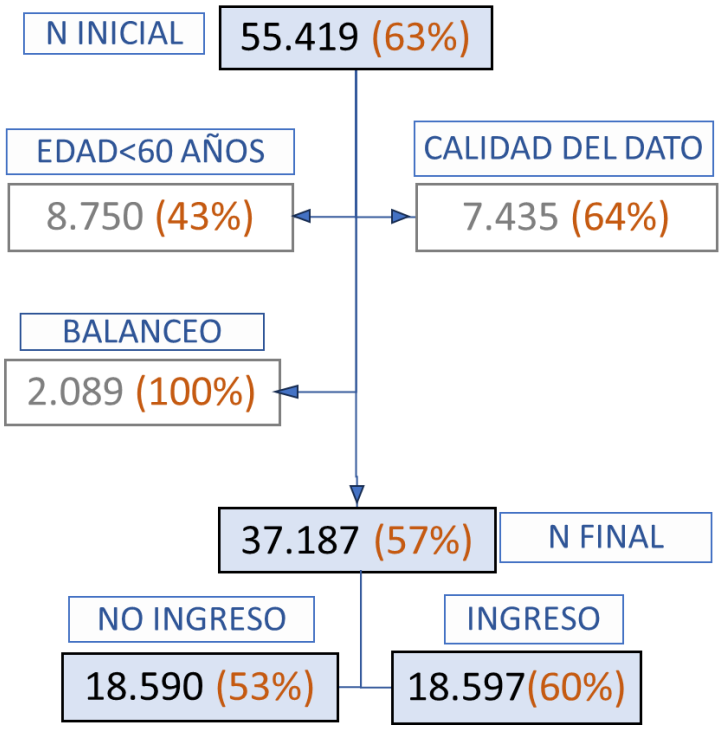


Imagen 20: Número (N) de personas incluidas en el estudio (entre paréntesis, porcentaje de mujeres)

Con respecto al resto de variables fue necesario realizar asunciones para el tratamiento de los valores perdidos, valores alejados y las inconsistencias o errores (explicadas en el anexo 1). Así, por ejemplo, en las escalas Barthel y MiniMental, se completaron los valores perdidos con el valor medio para cada una de estas variables, con la idea de interferir lo menos posible en los resultados y asumiendo que la ausencia de registro no implicaba una situación basal. En aquellos casos como el Índice de Masa Corporal (IMC), en el que había muchos valores alejados (*outliers*), la mediana resultó más indicada para

substituir los valores perdidos pues no es influenciada tanto como la media por los valores alejados.

En cuanto a los diagnósticos, seleccionamos como variables explicativas los diagnósticos crónicos, desechando los diagnósticos agudos (aquellos que tienen una fecha de inicio y una fecha de fin) por ser mucho más complejos de interpretar y requerir un análisis temporal (sólo se podrían utilizar como variable explicativa aquellos que ocurriesen con anterioridad al resultado). Para reducir la alta cardinalidad de los diagnósticos de la CIE-10, sus valores fueron truncados quedando los dos primeros caracteres y después se agruparon en 27 grupos con sentido clínico. En cuanto a las medicaciones, se agruparon los fármacos codificados mediante códigos ATC en el nivel 2 (ATC2), con 3 caracteres. Estas agrupaciones se detallan en el anexo 2.

5. 2 Métodos de análisis de los datos

La estadística convencional y la evaluación de la calidad del dato se realizó mediante el programa estadístico SPSS. También se utilizó la librería de *phyton seaborn* para elaborar el gráfico de correlaciones. El modelo utilizado para predecir los ingresos en hombres y en mujeres ajustando por las variables explicativas seleccionadas fue la regresión logística. En el anexo 3 se encuentra una explicación detallada de la interpretación de los resultados y estadísticos de la regresión logística y los indicadores de precisión de la predicción más importantes.

La analítica avanzada se realizó con técnicas de *machine learning* que es un subcampo de ciencias de la computación y una rama de la IA cuyo objetivo es simular el aprendizaje automático gracias a modelos matemáticos que nos sirven para detectar patrones en los datos y realizar predicciones. Sus algoritmos se pueden clasificar según la salida que produce cada uno de ellos:

- El aprendizaje supervisado se basa en conjuntos de datos previamente etiquetados, produciendo una función que establece correspondencias entre las entradas y las salidas deseadas del sistema. Es decir, se ajusta para poder predecir la respuesta de nuevo material, en función de lo aprendido sobre el conjunto de datos original etiquetado.

- En el aprendizaje no supervisado todo el modelado se basa solo en los datos de entrada que no están etiquetados, es decir, que no se tiene información a priori de las categorías de los ejemplos. Entonces, el sistema se ajusta para buscar patrones que agrupen los datos y esas segmentaciones se utilizan para poder “etiquetar” los diferentes grupos.

- El aprendizaje semi-supervisado es una mezcla de lo anterior. Se utiliza el aprendizaje no supervisado para la etiquetación y después se utilizan esas categorías para etiquetar nuevos datos.

- La información de entrada del aprendizaje por refuerzo es el *feedback* o retroalimentación que obtiene del mundo exterior como respuesta a sus acciones. Es un sistema que aprende a base de ensayo-error.

En este trabajo utilizamos modelos de aprendizaje supervisado ya que los datos con los que entrenaremos al modelo ya vienen etiquetados, puesto que sabemos si se produjo o no el ingreso hospitalario (ésa sería la etiqueta). Las fases naturales de un proyecto de aprendizaje supervisado, y que coinciden con los pasos que llevamos a cabo, son las siguientes:

- Recolección de datos: Acceso a bases de datos, extracción y carga.

- Preparación de los datos: Análisis exploratorio de datos (EDA en inglés). Incluye la transformación y homogeneización de los distintos tipos de datos y el control de los procesos de ETL y rendimiento de carga.

- Elección del algoritmo y realizar ajustes de los hiperparámetros.

- Entrenamiento del modelo: Para evitar problemas de generalización y sobreajuste, distribuimos los datos en dos conjuntos. El 70% fueron datos para entrenamiento, y el 30% restante se utilizaron como conjunto de validación.

- Validación del modelo: En esta fase se utilizan los datos restantes para validar diferentes indicadores del modelo (sensibilidad, especificidad), y como ya se ha señalado en el punto anterior, se utiliza ese 30% de datos restantes.

- Utilización del modelo: En este trabajo se utilizó el modelo para predecir resultados en el dataset de mujeres y valorar así si han existido sesgos.

Para realizar todo lo anterior, y en el aspecto técnico, los dos entornos de programación más populares en el campo del aprendizaje son: Python y R. En este trabajo, para el entrenamiento del modelo y todas las tareas de validación, se utilizó R con las librerías de manipulación de datos y creación de gráficas *dplyr*, *ggplot2* entre otras librerías clásicas. Dentro de los diferentes tipos de algoritmos de aprendizaje supervisado, se ha seleccionó el algoritmo *XGBoost*. A continuación, señalamos los principales puntos fuertes del modelo.

En primer lugar, este algoritmo proporciona un buen mecanismo para para evaluar las características que influyen más en las predicciones. Esta capacidad para seleccionar las

características relevantes que influyen en el modelo final, nos permite una cierta flexibilidad a la hora de explicar su funcionamiento, algo que no es viable con otros modelos que actúan más como cajas negras.

En este proyecto, como fiel reflejo de la realidad en el mundo de los datos sanitarios, se presentan valores perdidos. Pues bien, el propio algoritmo puede manejar estas circunstancias al tratarlos como un caso especial cuando construye los árboles, sin necesidad de que tengamos que realizar imputaciones y asunciones previas. En este caso, este factor está mitigado porque nuestro *dataset* final –como también se analiza por medios clásicos- ya va limpio en este sentido, pero el algoritmo podría utilizarse con un nivel de preparación más bajo que el presente, y eso siempre es una ventaja.

Otra de las características de este algoritmo es que, dentro de sus hiperparámetros, tiene uno que puede compensar el desbalanceo existente en los datos, algo que otros no tienen y que por tanto supone una ventaja a la hora de elegirlo. En nuestro caso, por los mismos motivos anteriores, este problema se soluciona en fases previas, pero sigue siendo de interés.

Además, en la literatura revisada, se expone que los modelos generados por *XGBoost* suelen ser más robustos y generalizables, ya que es capaz de controlar el sobreajuste y la complejidad. Esto hace que nos evite la personalización excesiva, es decir, que sobreajuste en exceso y por tanto se pierda generalización en las predicciones, y además evita que el rendimiento sea mucho peor con datos nuevos.

Otra característica interesante es que es capaz de capturar relaciones no lineales a través de sus árboles de decisión ensamblados, y porque puede realizar divisiones complejas en los datos. Teniendo en cuenta que la naturaleza de los datos sanitarios implica una no-linealidad subyacente, nos parece una característica muy interesante que inclinó la balanza hacia su selección.

Para concluir, y teniendo en cuenta que hasta hace poco tiempo era *Random Forest* uno de los algoritmos de elección, nos gustaría mostrar de una forma intuitiva la diferencia cualitativa entre utilizar un algoritmo más conocido como éste de *Random Forest* y el nuevo *XGBoost*. Para ello, supongamos que contamos con un grupo de expertos (que serán los árboles de decisión) que tienen que decidir cada uno si un paciente va a ingresar dado un conjunto de factores. Cada uno lo estudia y emite un voto (ingresará o no). Tras cada voto, tomaremos en cuenta todas las decisiones y se realizará un conteo para tomar en consideración la decisión más votada. Salvando las distancias y teniendo en cuenta que es un ejemplo intuitivo, en este caso estaríamos utilizando un algoritmo de tipo *Random Forest*.

Ahora bien, supongamos que, para la misma tarea, cambiamos un poco la estrategia y ahora lo que hacemos con el grupo de expertos es obligarles a trabajar en conjunto para

obtener la mejor decisión. Esta colaboración es iterativa, cada uno va haciendo su estimación, pero en la siguiente ronda –para mejorar- toma en cuenta los errores del anterior para mejorar su predicción. Esto se hace reiteradamente en cada caso y se mide el rendimiento de cada uno de los modelos propuestos por cada experto para elegir al que tiene mejor rendimiento. Digamos que, en este caso, los procesos de decisión se van refinando y se toma el mejor de todos ellos.

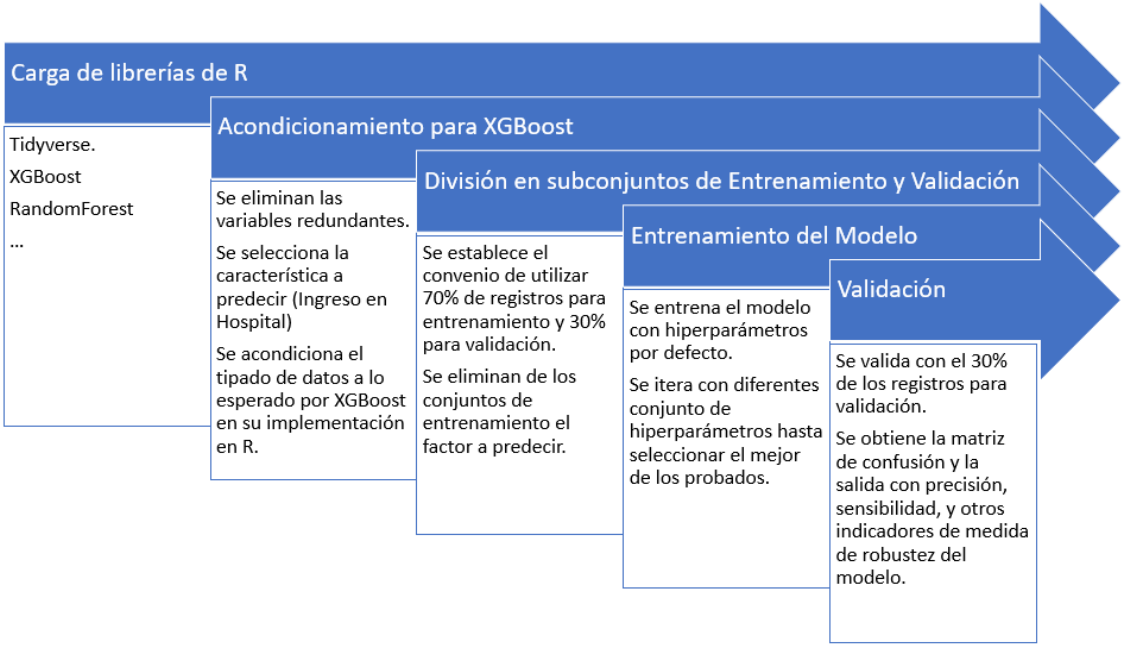


Imagen 21: Proceso específico de entrenamiento del modelo XGBoost

Por último, señalar que en este modelo y en cuanto a la interpretación de los hiperparámetros (anexo 4), la ganancia mide cómo cada característica contribuye a la mejora del rendimiento, la cobertura muestra cuántos datos se vieron afectados por las decisiones basadas en una característica, y la frecuencia indica cuánto se utilizó la característica para tomar decisiones. La importancia es una métrica ponderada que combina estas medidas para evaluar la relevancia general de las características en el modelo entrenado.

5.3 Resultados

5.3.1 Analítica convencional

En cuanto al análisis con estadística convencional, los modelos de regresión logística construidos con el *dataset* de pacientes positivos a COVID, obtuvieron una baja capacidad explicativa de la variabilidad que no llegaba al 10% (R^2 ajustada de 0,082). Esto concordaba también con que los modelos de analítica avanzada (*XGBoost* y *Radom Forest*) no consiguieran realizar predicciones con garantías suficientes ni en sensibilidad ni en precisión.

Debido a este resultado tan bajo en la predicción de las hospitalizaciones por COVID, se reenfocó el estudio a un análisis más amplio que incluyese las hospitalizaciones por todas las causas. Tras la limpieza y el tratamiento de los datos, este nuevo modelo consiguió explicar más de un 30% de la variabilidad (R^2=0,327). Esta razón de prevalencias (1,327) nos indica que los hombres, con respecto a las mujeres, a igualdad de diagnósticos, tratamientos y resto de variables predictoras, tienen un 33% más probabilidades de ser hospitalizados que las mujeres. En el anexo 3 se explica la interpretación de la regresión logística y en el anexo 5 se muestran los resultados detallados de la misma.

Tras ensayar con distintas combinaciones de variables a incluir en el modelo de regresión logística, finalmente se llega al modelo final, que es el que maximiza la capacidad explicativa. Inicialmente se trata de un modelo general, entrenado en conjunto de datos completo, con hombres y mujeres. Esto nos permite obtener un primer estimador del sesgo de género, gracias a la interpretación del coeficiente de la variable sexo que, como hemos visto, es de 1,327. Este primer estimador del sesgo de género, además, resulta consistente en todos los modelos intermedios en los que el coeficiente de la variable sexo siempre se mantiene entre 1,200 y 1,300.

Después, entrenamos este modelo de forma separada con los hombres (sexo=1) y las mujeres (sexo=0) y comparamos los resultados de los parámetros de interés. Tras esta comparación se concluye que el modelo entrenado en hombres explica mayor porcentaje de la variabilidad por lo que será el que utilicemos para realizar predicciones. Así, cuando pasamos este modelo entrenado hombres a cada uno de los dos subconjuntos (hombres y mujeres) nos encontramos con diferentes capacidades predictivas, como es de esperar. En la tabla 2 se muestran los principales indicadores de predicción: área bajo la curva, sensibilidad, especificidad, valor predictivo positivo y valor predictivo negativo, conceptos que son explicados en el anexo 3.

	Área bajo la curva (AUC)	**Sensibilidad**	**Especificidad**	**VPP**	**VPN**
Modelo general	76,3%	68,1%	70,5%	69,8%	68,8%
Modelo de hombres testado en hombres	**77,1%**	61,8%	**78,6%**	**71%**	**70,8%**
Modelo de hombres testado en mujeres	75%	**72,9%**	64,1%	69,7%	67,6%

Tabla 2: Indicadores de la predicción de los 3 modelos analizados

En la tabla 2 se señala en negrita el mejor de los resultados de predicción para cada uno de los indicadores. Se puede observar que cuando el modelo entrenado en hombres predice en el subgrupo de hombres obtiene los mejores valores con excepción de la sensibilidad, que es más alta cuando este mismo modelo predice en el subgrupo de las mujeres.

El modelo entrenado en hombres es capaz de detectar con mayor acierto a las mujeres hospitalizadas que a los hombres hospitalizados, esto es, predice menos hospitalizaciones en hombres y más en mujeres de las que realmente ocurren. Hay que señalar que estos resultados están ajustados por diversas variables, entre ellas edad, diagnósticos crónicos (comorbilidades), tratamientos farmacológicos, estado funcional y estado mental. Por lo tanto, ante las mismas variables de ajuste las mujeres están recibiendo un trato más conservador que los hombres en lo que respecta a la hospitalización, lo que sugiere que si fueran tratadas como ellos tendrían más hospitalizaciones.

El siguiente paso para cuantificar este sesgo es la estimación mediante el cálculo de cuantas mujeres serían hospitalizadas si fueran tratadas como los hombres, o sea, la diferencia entre el número de mujeres hospitalizadas observadas en el conjunto de datos y las esperadas o pronosticadas según el modelo de hombres. Esta diferencia asciende a 508 mujeres que, con los mismos diagnósticos, tratamientos, edad y el resto de las variables explicativas, hubieran sido hospitalizadas si fuesen tratadas como los hombres, o sea, según el modelo entrenado en hombres. Haciendo este mismo cálculo con los hombres la diferencia es de 957 hombres que según sus variables explicativas no serían hospitalizados según el mismo modelo entrenado en hombres. Esta diferencia de predicción entre hombres y mujeres (1.465 hospitalizaciones de diferencia) es una manera de estimar y cuantificar el sesgo de género en la hospitalización. Podría existir, eventualmente, una falta de ajuste por otras variables, que podrían afinar más los resultados, si bien las variables más importantes están presentes en nuestro modelo (edad, diagnósticos y tratamientos). Además, los resultados han resultado consistentes a lo largo de todos nuestros análisis y, como veremos ahora, también con el análisis de *machine learning*. Estos cálculos son explicados en detalle en el anexo 5.

5.3.2 Analítica avanzada

Como ya se ha mencionado en el apartado de metodología, hemos seguido los pasos habituales en este tipo de proyectos:

- En primer lugar, comenzamos con una fase de extracción de los datos de las fuentes de origen y los procesos de acondicionamiento que incluyen la selección de características y subconjuntos.
- Después pasamos a la fase de análisis estadístico y acondicionamiento de variables, submuestreo para conseguir balanceo y examen de calidad del dato.

- Y por último se realiza el entrenamiento y la validación del modelo/s elegidos.

En nuestro caso, como ya comentamos anteriormente, seleccionamos el modelo *XGBoost* aunque también se hizo una prueba con el modelo de *Random Forest*. Se obtuvieron rendimientos parecidos, siendo mejor el *XGBoost* que es el que vamos a examinar en detalle en este trabajo. Adicionalmente a la parte de IA, se realizó también en paralelo un estudio con estadística convencional (EC) que complementa y permite comparar los resultados entre ambas aproximaciones.

Las tres fases mencionadas se articulan secuencialmente, siendo la primera de las fases (la extracción de los datos) común tanto para esta parte de análisis con IA como para la parte de estadística convencional. Es interesante mencionar que para la fase de validación hemos aplicado lo que ya constituye un estándar operativo en el que los datos disponibles se dividen en dos conjuntos: el de entrenamiento, que tiene un 70% de los datos del total, y el 30% restante, que se reserva para realizar con él las validaciones del modelo. Estas validaciones se realizan calculando indicadores como la precisión, el *recall* o retirada, el F1-score y la curva ROC para evaluar la precisión de las predicciones del modelo comparando la sensibilidad frente a la especificidad de la clasificación.

Actualmente ya existen librerías que permiten automatizar todo este tipo de validaciones y combinarlos con la selección de los diferentes hiperparámetros con el objetivo de ahorrar tiempo en la selección del ajuste del mejor modelo posible. En este trabajo esta iteración y selección de los hiperparámetros y la validación han sido manuales, probando varios modelos diferentes hasta encontrar el mejor de todos los probados. Es probable que los resultados hubieran mejorado al utilizar estos métodos más sofisticados, pero en diversas pruebas realizadas las mejoras no eran significativas en la evaluación del coste-beneficio y los resultados obtenidos eran similares para nuestro objetivo principal y los secundarios que nos hemos marcado.

Con los datos ya extraídos y acondicionados, y completada la fase de evaluación previa, comenzamos el modelado, y siguiendo las fases explicadas antes, nos planteamos varios escenarios:

1. Entrenar el modelo sobre mujeres y aplicarlo en hombres.
2. Entrenar el modelo sobre hombres y aplicarlo en mujeres.
3. Entrenar un modelo sobre un conjunto mixto de hombres y mujeres y utilizarlo en dos conjuntos de testeo de hombres y mujeres, separados

En las pruebas realizadas, los tres escenarios condujeron a las mismas conclusiones. Como queríamos poder compararlo con la Estadística Convencional, elegimos finalmente el mismo escenario que en EC, entrenando un modelo en hombres y luego utilizándolo en el conjunto de mujeres.

Una vez obtenido el modelo realizaremos dos validaciones: una del propio modelo respecto al grupo en el que se ha entrenado (hombres) y luego validaremos con el grupo de mujeres; o si se prefiere, podemos decir que aplicaremos el modelo en ese conjunto. Si ambos subgrupos (hombres y mujeres) fueran similares, los resultados de las validaciones deberían ser similares también. Pero si los resultados no son parecidos, será un indicio de la existencia de posibles sesgos.

Gráficamente y de forma esquemática, el procedimiento explicado es el siguiente:

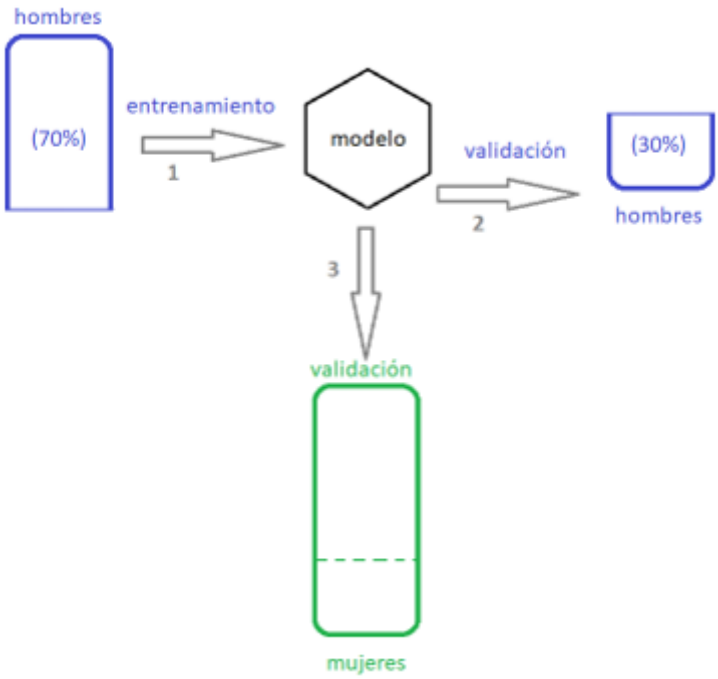


Imagen 22: Figura de elaboración propia. Se entrena el modelo sobre hombres. Luego se valida con el subconjunto de hombres reservado para testeo, y se valida de nuevo con el conjunto de mujeres nunca usado ni para entrenar. Se comparan las validaciones. Si fueran no habría indicios de sesgo.

En relación a los hiperparámetros seleccionados de entre todas las opciones ensayadas fueron un objetivo logístico ya que la variable a predecir es si ingresa o no. El número de rondas seleccionada fueron 200 con una profundidad máxima de los árboles de 3 niveles, un eta de 0,2. Como los modelos fueron entrenados en un ordenador personal de potencia media (suficiente para este tipo de modelados básicos) el número de hilos fueron 2. Para optimizar el tiempo de modelado se seleccionó que tras diez rondas sin cambios significativos en la ganancia se detuviera el proceso.

Una vez generado el modelo, comenzamos con la primera de las validaciones: la que se refiere al propio grupo de hombres. Así pues, se testeó con el subconjunto de hombres reservados para esta tarea (el 30% de los registros que no se usaron para entrenar). La predicción que calcula el modelo es en realidad una probabilidad y se utilizó un umbral del 0,5 sobre 1 para establecer si se le asignaba la etiqueta “ingresa” o “no ingresa”. Es un criterio similar al utilizado en modelos de regresión logística convencional.

Después realizamos la segunda validación: en el conjunto de mujeres. Utilizamos todo el conjunto puesto que lo que nos interesa es comparar los resultados de ambas validaciones sin importar tanto la reserva de casos de entrenamiento puesto que ahora ya no tiene sentido.

Es necesario comparar ambos resultados, y para ello podemos utilizar las métricas que se proporcionan en cada salida. En nuestro caso, los que ingresan están etiquetados como 0 en el *dataset* acondicionado para *XGBoost.* Esto es contraintuitivo respecto a las convenciones habituales pero esta librería funciona así, toma la clase positiva como la 0. Así que re-etiquetamos el dataset para que podamos utilizar los indicadores habituales. En este sentido, usaremos sensibilidad, especificidad, valor predictivo positivo y valor predictivo negativo. Para ello lo que hacemos es calcular la matriz de confusión obteniendo los resultados que exponemos en el anexo 6, y resumimos a continuación.

En nuestro caso, sobre todo nos interesa la sensibilidad, ya que esta métrica indica cómo de bien se pueden identificar correctamente los casos positivos verdaderos (los que ingresan), esto es, nos interesa que los casos positivos sean clasificados correctamente.

Modelo	Sensibilidad	Especificidad	Precisión	Valor Pred. Positivo	Valor Pred. Negativo
1 Validado en Mujeres	0,7605	0,6347	0,7024	0,7080	0,6947
2 Validado en hombres	0,6440	0,7902	0,7215	0,7314	0,7146

Tabla 3: Comparación de los resultados en hombres y mujeres

Al comparar las validaciones entre los grupos de hombres y mujeres (del modelo entrenado en hombres) vemos que la sensibilidad es mayor para las mujeres (0,7658 frente a 0,644). Así pues, el modelo tiene una mejor capacidad para identificar correctamente los casos positivos verdaderos en el conjunto de las mujeres.

Como ya habíamos explicado al principio de esta sección, otro escenario implicaba hacerlo a la inversa, entrenar un modelo en mujeres y aplicarlo en hombres. Vemos que las mujeres tienen mejor validación en su propio grupo y peor al aplicarlo a los hombres. Esto es simétrico frente a lo anterior y lleva a las mismas conclusiones, como veremos más adelante.

Sensiblidad	
Modelo Masculino	Modelo Femenino
0,644	0,7605
Aplicado a Mujeres	Aplicado a Hombres
0,7658	0,6139

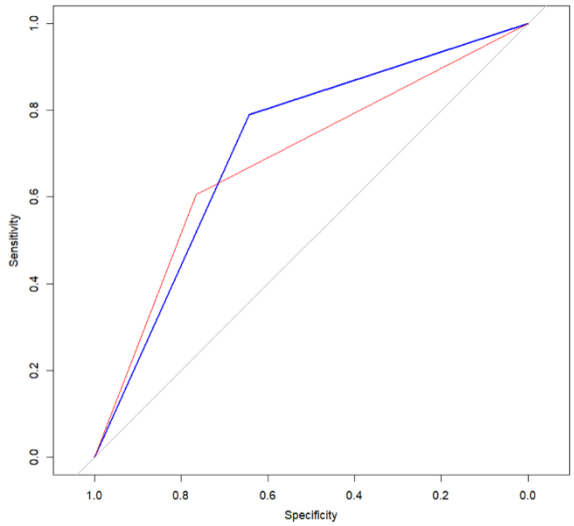
Tabla 4: Comparación de la sensibilidad de los distintos modelos en hombres y mujeres

Además de los indicadores anteriores, también suele ser útil realizar una curva ROC y evaluar su área. A mayor área, mejor sería el modelo en conjunto, considerando la globalidad de indicadores.

Calculando las áreas ROC de la validación en hombres y en mujeres, encontramos que la mayor área es la correspondiente al modelo entrenado en hombres. Esto es porque tiene en cuenta también la especificidad (los que no ingresan) en los que los modelos masculinos son mejores, y de forma global el modelo tiene más área. Sin embargo, en el puntaje de sensibilidad se ve que concuerda con los datos anteriores, mejorando ese indicador en el modelo masculino aplicado a mujeres.

Área ROC modelo masculino	Área ROC modelo aplicado a mujeres
0,7107	0,6896
1,03 (normalizado frente a mujeres)	1

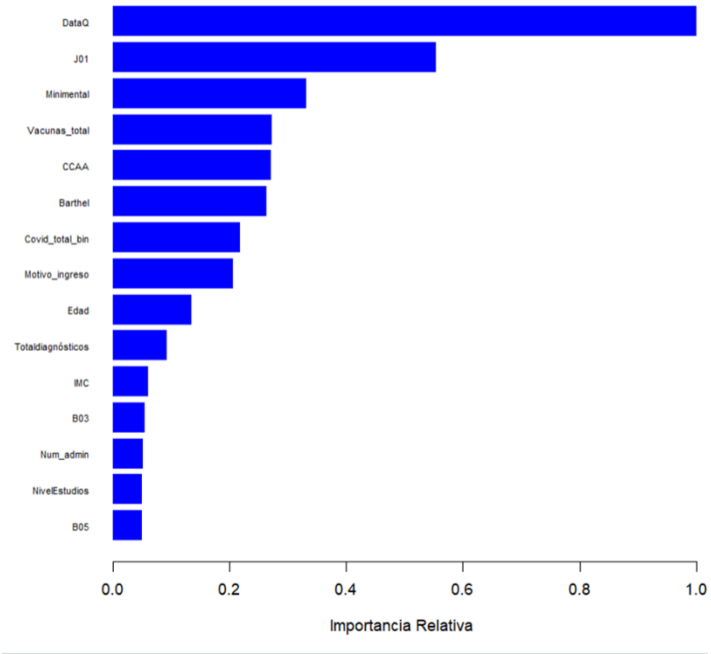
Tabla 5: Comparativa de las áreas ROC entre las dos validaciones



Gráfica 1: Curva ROC del modelo entrenado en hombres pasado sobre el conjunto de mujeres (rojo) frente a hombres (azul).

Además de todo lo anterior, y con el objetivo de hacer más explicativo el modelo, se calcularon las características más relevantes del modelo con una de las funcionalidades de XGBoost, que permite calcular de forma automática las más relevantes.

En el modelo masculino, de todas ellas, la variable que representa la calidad del dato es la principal, junto con el J01 (que se corresponde con los antibióticos sistémicos), el MiniMental, el total de vacunas, la escala Barthel, la comunidad autónoma o el motivo de ingreso entre otros. En el anexo 6 se puede ver la relación completa.



Gráfica 2: Top 15-factores más importante del modelo entrenado con hombres

Como se comentó al principio, otro escenario posible era entrenar con mujeres y aplicarlo a hombres. Esto se hizo para determinar si había diferencias en las características principales. Aquí se muestra la comparativa, en la que se ve que en ambos subconjuntos los factores relevantes son similares, aunque existen algunas diferencias, tanto en el orden como en dos factores que aparecen exclusivamente en los primeros puestos en cada conjunto:

- En mujeres: C03 (del grupo de diuréticos), B01 (antitrombóticos) y la procedencia del episodio.
- En hombres: B05 (perfusiones de sangre) y el B03 (antianémicos).

Entrenado en Mujeres	Entrenado en Hombres
DataQ	DataQ
J01	J01
Vacunas_total	MiniMental
MiniMental	Vacunas_total
Barthel	CCAA
COVID_total_bin	Barthel
Motivo_ingreso	COVID_total_bin
CCAA	Motivo_ingreso
Edad	Edad
Num_admin	Totaldiagnósticos
C03	IMC
IMC	B03
Totaldiagnósticos	Num_admin
B01	NivelEstudios
Episodio_Procedencia	B05

Tabla 6: Factores más importantes obtenidos en hombres y mujeres

Por último, se realizó un análisis con la librería *Fairness* para confirmar los resultados anteriores. En concreto utilizamos el indicador de paridad demográfica que considera todos los positivos (tanto los verdaderos como los falsos). Si ambos grupos fueran equitativos, deberían ser muy similares. Y sin embargo los datos obtenidos por el modelo indican que se asignarían más ingresos a las mujeres que a los hombres. Es decir, que el modelo masculino aplicado a las mujeres predice más ingresos de los observados realmente en el conjunto de datos. Tomando como referencia a las mujeres, los hombres tienen un factor de 0,5691 frente a la unidad. Es decir, no existe equidad entre ambos grupos según estos parámetros. En el siguiente apartado veremos la interpretación de estos resultados y su relación con los cálculos de estadística convencional.

5.3.3 Comparación de resultados

Los resultados de estimación del sesgo de género resultan consistentes en ambas metodologías. En ambos casos las predicciones del modelo entrenado hombres -obtenidas en el subgrupo de hombres-, consiguen los mejores resultados con excepción de la sensibilidad, que es más alta en el modelo de hombres testado en mujeres. Esto significa que el modelo entrenado en hombres es capaz de detectar con mayor acierto a

las mujeres que a los hombres que son hospitalizados. Se observa también que las predicciones del *machine learning* mejoran ligeramente las de la EC, con la excepción de la especificidad en mujeres, aunque en ningún caso se trata de diferencias relevantes, puesto que los modelos de *machine learning* tampoco alcanzan el dintel del 75% (exceptuando la sensibilidad en las mujeres) considerado generalmente como la capacidad predictiva aceptable en estos indicadores.

	Sensibilidad		**Especificidad**		**VPP**		**VPN**	
	EC	**IA**	**EC**	**IA**	**EC**	**IA**	**EC**	**IA**
Modelo de hombres testado en hombres	61,8%	64,4%	78,6%	79%	71%	73,1%	70,8%	71,5%
Modelo de hombres testado en mujeres	72,9%	76,1%	64,1%	63,5%	69,7%	70,8%	67,6%	69,5%

Tabla 7: Comparación de los indicadores de precisión en ambas metodologías

Si comparamos las estimaciones de sesgo de género mediante el cálculo de la diferencia en el número de casos pronosticados (esperados) y el de casos observados para cada uno de los sexos, nos encontramos los resultados de la tabla 8. Nuevamente, aunque no coinciden de forma exacta, los resultados entre los distintos modelos tanto de EC como de *machine learning* sí resultan coherentes y consistentes.

Esta diferencia entre los casos observados y los esperados resulta negativa en ambos casos para las mujeres, con 508 y 652 casos de mujeres respectivamente (EC/IA), que hubieran sido ingresadas si fuesen tratadas como los hombres en función de sus valores registrados en la base de datos. Haciendo este mismo cálculo con los hombres la diferencia es de 957 y 882 (EC/IA) hombres que según sus variables explicativas no serían hospitalizados. También es destacable que ante la mejora de la capacidad predictiva que aporta el *machine learning,* las diferencias de casos observados y esperados (que es nuestro estimador final del sesgo de género) se suavizan en los hombres y sin embargo se acentúan en las mujeres, lo que también resulta coherente con nuestras conclusiones, pues al mejorar las predicciones con la IA se acentúa el sesgo (diferencia de 1.534 hospitalizaciones con la IA frente a las 1.465 de la EC).

	EC modelo final				IA modelo final			
	Obsv	Esp	Obsv-Esp	Diferencia H-M	Obsv	Esp	Obsv-Esp	Diferencia H-M
Hombres	7.399	6.442	**957**	**1.465 ingresos**	7.399	6.517	**882**	**1.534 ingresos**
Mujeres	11.194	11.702	**-508**		8.725	9.377	**-652**	

Tabla 8: Comparación de las estimaciones del sesgo de género (casos observados menos casos esperados) mediante los distintos métodos utilizados. Obsv: casos observados; Esp: casos esperados; Obsv-esp: diferencia entre casos esperados y observados; Diferencia H-M: diferencia de ingresos hospitalarios predichos en hombre y mujeres por el modelo ajustado en hombres.

6 Limitaciones

Una de las principales limitaciones de este estudio, como todos aquellos realizados en bases de datos del mundo real en el ámbito sociosanitario, es la **calidad del dato**. Esta siempre es un reto en estas bases de datos, que son tanto clínicas como administrativas, sobre todo porque es necesario mantener unos niveles de calidad que nos permitan desarrollar los usos asistenciales, primarios, de la información. Los usos secundarios, como es el caso de este estudio, suelen quedar en un segundo plano a la hora de diseñar la recogida y tratamiento de los datos, lo que convierte la calidad del dato en la piedra angular de estos usos secundarios, como puede ser la investigación, la gestión o la salud pública. Además, con el auge de las tecnologías de analítica avanzada la calidad del dato se convierte en un asunto crítico para poder explotar todas las potencialidades de la ciencia guiada por datos. En nuestro estudio un 13% de los registros (7.435) fueron eliminados del estudio por baja calidad del dato, lo que normalmente puede acarrear sesgos y disminuir la significación estadística de los resultados. Este efecto no es tan acusado en este estudio porque se centra en la variable sexo que al ser binaria es muy robusta y permite obtener resultados con menor tamaño muestral.

El tipo de **análisis transversal**, que recoge tanto las variables predictoras como las variables resultado en un mismo periodo temporal (2020-2022) también limita las conclusiones y dificulta la inferencia de causalidad. Lo ideal hubiese sido tener en cuenta la secuencia temporal, esto es, que aquello utilizado para predecir sea anterior a lo que vamos a predecir o resultado. Esto no ha sido posible en este estudio y por ello la interpretación de los modelos predictivos no puede incluir asunciones de causalidad y los coeficientes (ExpB) del modelo han de interpretarse más como una razón de prevalencias –descriptiva- que como una *odds ratio* -OR- o razón de riesgos, que tienen connotación causal.

El objetivo de nuestro estudio es la comparación de resultados entre ambos sexos y no la inferencia causal de los modelos predictivos, por lo que esta limitación tampoco inhabilita los hallazgos. Por el mismo motivo, la limitación que puede suponer la baja **capacidad predictiva** obtenida de los modelos también es relativa. Aunque utilizamos métodos predictivos para estimar las hospitalizaciones, nuestro objetivo real es señalar los diferentes patrones con respecto a la derivación hospitalaria en hombres y mujeres, sin pretensiones de predecirlas o explicar sus causas. Por tanto, los modelos serían considerados poco aptos para un estudio predictivo en sí mismo, por el bajo porcentaje de la variabilidad explicada, pero en nuestro caso sí nos permiten extraer conclusiones con respecto a una variable binaria como el sexo.

Integrar en el análisis los diagnósticos agudos, que requieren una interpretación temporal, hubiese probablemente aumentado la capacidad predictiva de los modelos. Y tampoco podemos descartar la existencia de otros factores independientes de los

analizados que influyan en el resultado diferencial entre sexos, más allá del propio sesgo de género que es lo que se pretende medir y que incluir estos factores mejoraría la estimación del sesgo. Sin embargo, hay que tener en cuenta que las variables más importantes que pueden influir en la hospitalización, según se desprende de la literatura científica, están presentes en el estudio (edad, diagnósticos y tratamientos).

E**l tratamiento de las clasificaciones** clínicas como la CIE que por su alta cardinalidad no pueden entrar en el modelo al completo y las estrategias para abordar este problema, puede ser también discutible. En esta ocasión nos ha bastado con truncar el código CIE quedándonos con un prefijo de 2 caracteres, para evitar tener demasiadas características que dificultarían a los algoritmos encontrar buenos modelos, y que tendrán un coste computacional alto. Esta estrategia es útil cuando se sabe que la pérdida de información puede no ser muy importante en el resultado global, o cuando se prioriza el coste del entrenamiento. También hemos utilizado en parte otras formas de abordaje como la denominada *"feature agglomeration*" reduciendo el número de variables al agrupar jerárquicamente aquellas con patrones similares y con sentido clínico. Mucho menos tratamiento recibió la clasificación ATC de los medicamentos que entró directamente en su segundo nivel (ATC2). Es posible que el tratamiento de esta variable clínica de forma similar a los diagnósticos hubiera mejorado los resultados o al menos los hubiera hecho mucho más interpretables.

Otra posible limitación sería la **representatividad** de la base de datos de *DomusVi* y hasta qué punto se pueden extrapolar estos resultados al conjunto de la población geriátrica española. *DomusVi* representa el 25,48% de las plazas de centros geriátricos del país, por lo que podemos decir que se trata de una amplia muestra del sector geriátrico en España. Aunque es posible que la representatividad no sea la misma en todas las comunidades autónomas, nuevamente, al tratarse de una variable de segmentación tan robusta como el sexo, los resultados son muy probablemente extrapolables al conjunto de población de mayores de nuestro país.

7 Discusión

Un detalle interesante de los resultados del estudio es que estos no se aprecian en los **porcentajes brutos** mostrados en la imagen 20 (porcentaje de mujeres de cada uno de los grupos). En este análisis crudo el porcentaje de mujeres hospitalizadas es superior al de hombres, al contrario de lo que ocurre en los análisis ajustados. Se deduce por tanto la existencia de confusión, al menos por la variable edad, siendo esta una de las más importantes de las variables del ajuste. El distinto perfil demográfico en estas edades entre hombres y mujeres es una característica muy estudiada que se refleja en la pirámide poblacional de nuestro país (imagen 23). El ajuste que nos proporcionan tanto los modelos de regresión logística como los de *machine learning*, por características como la edad, los diagnósticos y tratamientos, el estado funciona y el mental, son fundamentales en este análisis puesto que si no fuera por el ajuste el sesgo de género no saldría a la luz.

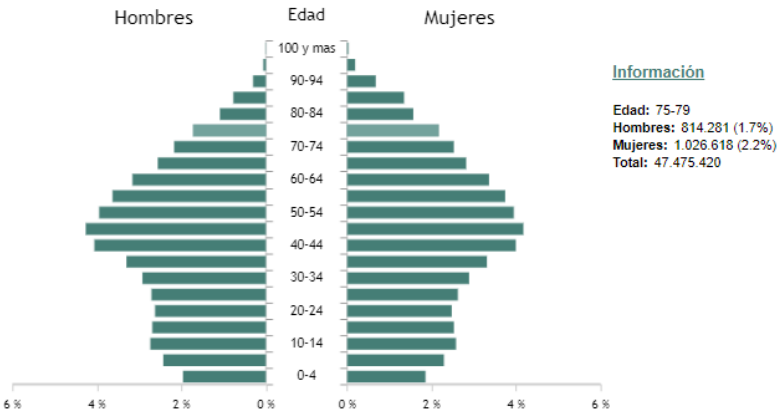


Imagen 23: Pirámide poblacional de España a 1 de enero de 2022. Fuente: INE

Otra característica destacable del análisis realizado ha sido la **coherencia** y persistencia del sesgo de género en todos los análisis intermedios y en ambas metodologías (EC e IA). Por ejemplo, el análisis inicial en pacientes COVID resultó insuficiente para obtener predicciones. Sin embargo, los modelos de regresión logística sí eran estadísticamente significativos, esto es, el conjunto de las variables incluidas explicaba los resultados mejor que el azar, y el efecto de la variable sexo sobre el ingreso por COVID también resultó estadísticamente significativo siendo la prevalencia ajustada de ser hospitalizado para los hombres con COVID casi un 30% mayor con respecto a las mujeres con COVID (razón de prevalencias –RP- de 0,296). Se trata de resultados siempre en el mismo sentido y de una magnitud entre el 20% y el 30%.

El sesgo de género en los resultados sociosanitarios se puede descomponer a su vez en varias causas convergentes, que han sido descritas por la literatura científica en los

últimos 30 años. Tenemos, por una parte, el **sesgo de información** o de registro, que se refiere a una diferente exhaustividad en el registro en hombres y en mujeres. Por otro lado, está el **sesgo diagnóstico**, que implica diferencias en los diagnósticos atribuidos a hombres y a mujeres ante los mismos signos, síntomas y/o parámetros. Este sesgo también se ha relacionado con la diferente intensidad de pruebas diagnósticas aplicada sobre mujeres y hombres ante las mismas señales clínicas. Otro sesgo es el **terapéutico**, que normalmente comporta una menor intensidad terapéutica en mujeres, antes problemas clínicos similares, excepto en aquellos medicamentos de la esfera psicológica, como ansiolíticos o antidepresivos, mucho más utilizados en mujeres. Este sesgo a su vez se puede evidenciar tanto en actuaciones médicas y tratamientos farmacológicos como en intervenciones quirúrgicas u hospitalizaciones, como sería nuestro caso.

En las bases de datos basadas en registros clínico-administrativos, como las utilizadas en nuestro estudio, es muy difícil valorar el **sesgo de información** pues este es introducido de forma inconsciente por los profesionales, y no tenemos normalmente una verdad base o *"Gold estándar*" con el que comparar los diagnósticos. Sin embargo, en este estudio sí hemos constatado una menor calidad del dato en mujeres, puesto que el número de registros con calidad nula (DataQ=0) fue más elevado entre las mujeres que entre los hombres pues el 63,3% de estos registros pertenecían a mujeres. Además, los modelos predictivos nos han permitido cuantificar la situación registrada de hombres y mujeres en cuanto a resultados objetivos (hospitalizaciones), ajustando por comorbilidades, tratamientos y otros factores de riesgo registrados, lo que sería un **sesgo terapéutico**, considerando la hospitalización como una modalidad de tratamiento, que de por sí podría también llevar implícito un **sesgo diagnóstico**.

8 Conclusiones

Con respecto al objetivo principal del estudio, que era la valoración y cuantificación del sesgo de género utilizando el ajuste aportado por los modelos predictivos, podemos decir que se confirma tanto por la estadística convencional como por la analítica avanzada la existencia de un sesgo en la hospitalización a favor de los hombres, siendo un 30% más prevalente la hospitalización de hombres que de mujeres ante los mismos perfiles de edad, diagnósticos, tratamientos y el resto de variables de ajuste. Este sesgo se ha cuantificado de dos maneras, por un lado, utilizando los coeficientes de la variable sexo en el modelo general y por el otro calculando las diferencias entre las hospitalizaciones observadas y las predichas por los modelos en hombres y mujeres. Los resultados han sido consistentes a lo largo de todos nuestros análisis y también entre las distintas metodologías (regresión logística/m*achine learning*).

Hemos observado en este estudio que las metodologías de IA como el *machine learning* son válidas para valorar el sesgo de género, tanto al menos como la estadística convencional. Nuestro estudio no es más que un ejemplo de cómo podemos abordar esta pregunta de investigación sobre las diferencias entre sexos en la atención sociosanitaria recibida, aunque pueden existir otras formas de abordar el problema del ajuste por las variables clínicas y sociodemográficas.

Uno de los objetivos secundarios del estudio es evaluar la calidad de los datos sociosanitarios. A este respecto, hemos tenido que realizar limpieza y tratamiento de los mismos para poder obtener resultados predictivos, tal y como se describe en los apartados correspondientes, de lo que se deduce que existe recorrido de mejora en este ámbito. Sin embargo, también sabemos que la calidad del dato es el punto débil en cualquier contexto sociosanitario, y sanitario en general, sobre todo cuando lo que se pretende es el uso de analítica avanzada puesto que la finalidad con la que son recogidos los datos por parte de los profesionales difiere de estos usos secundarios. La necesidad de **mejorar la calidad del dato** ocurre en todas las bases de datos del mundo real. Los problemas de calidad del dato detectados en este estudio resultan muy comparables a los de otras bases de datos clínico-administrativas como pudieran ser las del Ministerio de Sanidad.

Cuando hablamos de un uso secundario de los datos, se ha demostrado que la existencia de sistemas de retroalimentación hacia aquellos profesionales que los registran supone normalmente una mejora progresiva de la calidad. Se trata del “retorno de inversión” del esfuerzo del profesional sanitario a la hora del registro. Si no ven para qué sirve (más allá de lo evidente) pierden motivación por mantener su calidad o ir un poco más allá de lo mínimo. En este sentido podría ser positivo para la calidad del dato ofrecer resultados y análisis de los datos a los profesionales de los centros sociosanitarios puesto que de

alguna manera les hace sentirse partícipes de otro uso de los datos diferente del asistencial y les hace más conscientes de la importancia del correcto registro.

También hemos visto en este estudio, no sólo la factibilidad técnica de interoperar estas bases de datos sociosanitarias con otras bases de datos clínico-administrativas, sino también el potencial beneficio de enriquecer estos datos con datos sanitarios para mejorar la capacidad explicativa de los modelos al añadir variables contextuales muy relacionadas con los eventos de interés. El cruce de ambas bases de datos, sanitarias y sociosanitarias, a nivel funcional no sólo nos permitirá afinar los ajustes sino también ampliar el estudio a otros eventos de interés, como por ejemplo los ingresos en UCI. También hemos de tener en cuenta que estamos hablando de un sector poblacional en crecimiento y que genera un importante porcentaje del gasto sanitario, por lo que nos interesa conocer sus características epidemiológicas y sociodemográficas y poder asociarlas a su información sanitaria. Sería muy interesante por lo tanto promover la estandarización de los protocolos de compartición de datos entre los ámbitos sanitario y sociosanitario, explorando el camino de la integración tanto para uso primario como secundario.

Además, aunque no era el objetivo principal del estudio, hemos podido de forma secundaria ahondar en el conocimiento de las características epidemiológicas de los mayores ingresados. Hemos comprobado la baja tasa de ingresos por COVID durante la pandemia, la desbalanceada distribución por sexos y también hemos podido analizar algunas variables socioeconómicas como el nivel de estudios o el estado civil (ver anexo 1). En relación con el desempeño de la Inteligencia Artificial aplicada en las bases de datos del mundo real (RWD), en este trabajo hemos constatado y trabajado con algunos de los puntos de fricción cuando se ha de acometer un proyecto de IA en un entorno real. Soslayando los aspectos jurídicos y de protección de datos -que son el paso previo e imprescindible para iniciar siquiera la recolección de los datos-, nos encontramos con el principal escollo que es la calidad del dato ya mencionado anteriormente. Es probable que muchos neófitos en el tema no sean conscientes de que los sistemas de IA también necesitan procesos de limpieza del dato y acondicionamiento, y que ante datos incompletos esta tecnología también fallará como pasa en las técnicas más clásicas.

Otro de los puntos que son tratados sistemáticamente al revisar las oportunidades de estos sistemas nuevos de IA se circunscriben al sesgo y la equidad. Nuestro trabajo hace de este punto su objetivo central y nos parece adecuado incluir el apunte de que la propia IA, si bien puede verse sesgada -porque aprende de datos que lo están- también puede servir para complementar y ayudar a los estudios que pretenden revisar estos aspectos, como el presente, lo que siempre es un punto muy positivo ya que cuantas más herramientas tengamos a nuestra disposición, mejor podremos acomodar la tecnología sanitaria a su objetivo principal que es ayudar a los profesionales y los pacientes a mejorar la calidad de la atención prestada y recibida. En este sentido,

consideramos que nuestro trabajo expone de una forma clara y fácil de entender a los no expertos, las limitaciones, las necesidades de las diferentes etapas y el esfuerzo que hay que invertir en este tipo de proyectos para que el gestor o directivo tenga una visión más objetiva de lo que suponen este tipo de proyectos y el reto que implica.

Aunque los resultados en EC resultan menos precisos que los obtenidos con el *XGBoost,* hay que destacar la riqueza de información que aporta la estadística tradicional frente a los modelos de *machine learning* por todo lo que nos permite interpretar, y no sólo las variables que más explican el modelo, si no que con los coeficientes podemos cuantificar la relación de cada variable con la variable resultado. Además, en un contexto de tamaño muestral limitado como en el primer análisis realizado con las hospitalizaciones por COVID, la estadística convencional aporta bastante información, aunque el modelo predictivo no resulte suficiente.

Al hilo de lo anterior, aunque al contar con relativamente pocos datos la IA " parece no funcionar en su faceta predictora" y la EC es mejor porque aporta al menos explicaciones. Sin embargo, la IA permite también ayudar a la EC como un complemento que permita cuestionar los resultados de la EC o incluso una herramienta para refinar los datos obtenidos de forma clásica. Quizás donde menos se nota la diferencia es en estudios como éste donde en núcleo final tenemos una teoría matemática muy parecida: regresión logística "pura y clásica" frente a un modelo de *machine learning* que también se basa en algo muy similar, aunque con un método de cálculo diferente. La IA, actualmente brilla con intensidad deslumbrante en campos menos accesibles (o directamente inaccesibles para los métodos de estadística clásica como el de visión o examen de imágenes médicas) y alcanza logros tremendos. La IA también está sorprendiendo desde hace unos años con el despegue de los LLM (*Large Language Model*) que están suponiendo un salto cualitativo en muchos aspectos, que exceden las pretensiones de este trabajo. Lo interesante es integrar las nuevas tecnologías sin despreciar las anteriores por dos razones: las nuevas se basan en las antiguas añadiendo una potencia de cálculo junto con capacidades heurísticas que hasta ahora no se había desarrollado, y conocer por tanto ambas y ponerlas al servicio de objetivos comunes puede potenciar nuestra forma de abordar los problemas, y por tanto de alcanzar nuevas y satisfactorias soluciones.

9 Gráficas, tablas, imágenes y abreviaturas

9.1 Índice de Gráficas

9.2 Índice de Tablas

9.3 Índice de Imágenes

9.4 Índice de abreviaturas

AI / IA: Inteligencia Artificial

ATC: Anatomical, Therapeutic and Chemical classification: clasificación de medicamentos por grupos anatómico, terapéutico y químico,

AUC: Área Bajo la Curvaevaluar la precisión de las predicciones de modelo altrazar la sensibilidad frente a la especificidad de una prueba de clasificación

BDCAP: Base de Datos Clínicos de Atención Primaria

CCAA: Comunicades Autónomas

CIE-10: Clasificación Internacional de Enfermedades, 10ª edición

CIP: Código de Identificación Personal

EC: Estadística Convencional

EDA: Análisis Exploratorio de Datos

ETL: Extraction, Transformation, Load

ExpB: representa la razón-cambio en las probabilidades del evento de interés para un cambio de una unidad en el predictor.

IMC: Índice de Masa Corporal

OR: Odds Ratio o Razón de de probabilidad

RAE-CMBD: Registro de Actividad de Atención Especializada – Conjunto Mínimo Básico de Datos

RGPD: Reglamento General de Protección de Datos

RP: Razón de Prevalencias

ROC: Característica Operativa del Receptor

RWD: Datos del Mundo Real

SNS: Sistema Nacional de Salud

SPSS: Paquete Estadístico para las Ciencias Sociales

TIC: Tecnologías de la Información y la Comunicación

UCI: Unidad de Cuidados Intensivos

VPN: Valor Predictivo Negativo

VPP: Valor Predictivo Positivo

10 Referencias

Bergdahl, M., Ehling, M., Elvers, E., Földesi, E., Körner, T., Kron, A., Lohauß, P., Mag, K., Morais, V., Nimmergut, A., Viggo-Sæbø, H., Timm, U. & Zilhão, M.J. (2007). Handbook on data quality assessment methods and tools. Ehling, Manfred Körner, Thomas.

Boada, L. D., Henríquez Hernández, L.A., Pérez Luzardo, O., Álvarez-León, E. E. & Zumbado Peña, M. (2023). Contaminación química, exposoma y salud en la población de las islas Canarias: una revisión sistemática de los estudios realizados y análisis de la situación. Revista Española de Salud Pública, 97(e202304033), e1-e19.

Burke, M.A. & Eichler, M. (2006). The bias free framework. A practical tool for identifying and eliminating social biases in health research. Global Forum for Health.

Cajachahua Espinoza, L. A. (2015). Predicción de Fuga de Clientes µ Una Aplicación de Técnicas de Data Mining En Telefonía Móvil. Universidad Complutense de Madrid.

Chang, Y. C., Chang, K. H. & Wu, G.J. (2018) Application of EXtreme Gradient Boosting Trees in the Construction of Credit Risk Assessment Models for Financial Institutions. Applied Soft Computing Journal, 73, 914–20. doi: 10.1016/j.asoc.2018.09.029.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, R., Mitchell, I. & Zhou, T. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4), 1-4.

Deng, Y., & Lumley, T. (2023). Multiple imputation through xgboost. Journal of Computational and Graphical Statistics, (*just accepted)*, 1-18.

De Vito, L. (2017). LinXGBoost: extension of XGBoost to generalized local linear models. arXiv preprint arXiv:1710.03634.

Dheeru, D. & Casey, G. (2017). UCI Machine Learning Repository. Retrieved September 6, 2021. http://archive.ics.uci.edu/ml.

Doménech Massons, J.M. & Pastor Navarro, J.B. (2007). Regresión logística binaria, multinomial, de Poisson y binomial negativa. Signo.

Doménech Massons, J.M. & Pastor Navarro, J.B (2007). Regresión múltiple con predictores cuantitativos y categóricos. Signo.

García Dauder, S. & Pérez Sedeño, E. (2018). Las "mentiras" científicas sobre las mujeres, 2ªed.: Los libros de la catarata.

García-Espinosa, E., Rojas-Concepción, A. A., Vitón-Castillo, A. A., & López, C. O. (2022). Principios FAIR de gestión de datos de investigación en ciencias de la salud. Revista Información Científica, 101(6), 13.

Gordon, B., Barrett, J., Fennessy, C., Cake, C., Milward, A., Irwin, C., Jones, M. & Sebire, N. (2021). Development of a data utility framework to support effective health data curation. BMJ health & care informatics, 28(1), pp. 1-8. DOI: 10.1136/bmjhci-2020-100303

Henley-Smith, S., Boyle, D., & Gray, K. (2019) Improving a Secondary Use Health Data Warehouse: Proposing a Multi-Level Data Quality Framework. eGEMs, 7(1). DOI: 10.5334/egems.298

Heras, M. (2006). Cardiopatía isquémica en la mujer: presentación clínica, pruebas diagnósticas y tratamiento de los síndromes coronarios agudos. Revista Española de Cardiología, 59(4): 371-81.

Hunter, D. J. & Holmes, C. (2023). Where Medical Statistics Meets Artificial Inteligence. The New England journal of medicine, 389(13), 1211-9. DOI: 10.1056/NEJMra2212850

Khan, M. G., Raebel, M. A., Glanz, J. M., Riedlinger, K. & Steiner, J. F. (2012). A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Med Care, 50(0). DOI: 10.1097/MRL.0b013e318257dd67.

Liaw, S. T., Guo, J. G. N., Ansari, S., Jonnagaddala, J., Godinho, M. A., Borelli Jr, A. J., de Lusignan, S., Capurro, D., Liyanage, H., Bhattal, N., Bennett, V., Chan, J. & Kahn, M. G. (2021). Quality assessment of real-world data repositories across the data life cycle: a literature review. Journal of the American Medical Informatics Association, 28(7), 1591-1599. DOI: https://doi.org/10.1093/jamia/ocaa340.

López-Sendón J.L., Alonso-Rodríguez, D., Barón-Esquivias, G., Cosin-Sales, J., Marín, F, Galera-Llorca, J., Jiménez, N., Marler, S., Huisman, M.V. & Lip, G.Y.H. (2021). Gender differences in antithrombotic treatment in patients with atrial fibrillation form Spain versus de rest of Western Europe. GLORIA-AF Program. Medicina Clinica. DOI: 10.1016/j.medcli.2021.09.016

Ottenbacher, K. J., Smith, P. M., Illig, S. B., Linn, R. T., Fiedler, R.C., Granger, C.V. (2001). Comparison of logistic regression and neural networks to predict rehospitalization in patients with stroke. Journal of Clinical Epidemiology, 54(11), 1159-1165. https://doi.org/10.1016/s0895-4356(01)00395-x.

Ruiz-Cantero, M.T. (2009). Sesgos de género en la atención sanitaria. Escuela Andaluza de Salud Pública.

Ruiz-Cantero, M.T. & Verdú-Delgado, M. (2004). Sesgo de género en el esfuerzo terapéutico. Gaceta Sanitaria, 18(4).

Sociedad Española de Cardiología (2016). Enfermedad Cardiovascular en la mujer. Estudio de la situación en España. Observatorio de Salud de la mujer. Ministerio de Sanidad y Consumo.

Valls-Llobet, C. (2020). La mujer invisible en la medicina. Capitán Swing.

Vogel, B. A., Appelman, Y., Bairey Merz, C. N., Chieffo, A. & Figtree, G. A. (2021). The Lancet women and cardiovascular disease Commission: reducing the global burden by 2030. The Lancet.

Wade, C. & Glynn, K. (2020). Hands-On Gradient Boosting with XGBoost and scikit-learn: Perform accessible machine learning and extreme gradient boosting with Python. Packt Publishing Ltd.

Zamora, A., Masana, L., Plana, N., Comas-Cufi, M., Gil, M., Rodríguez-Borjabad, C. & Ramos, R. (2019) ¿Existen desigualdades de género en la hipercolesterolemia familiar? La visión desde el manejo masivo de datos. En Ruiz-Cantero, M.T. (Eds), Perspectiva de género en medicina. Monografía 39 (64-80). Fundación Dr. Antoni Esteve

Zolotareva, E. (2021). Aiding long-term investment decisions with XGBoost machine learning model. In Artificial Intelligence and Soft Computing: 20th International Conference, ICAISC 2021, Virtual Event, June 21–23, 2021, Proceedings, Part II 20 (pp. 414-427). Springer International Publishing.

11 Webgrafía

https://biblus.us.es/bibing/proyectos/abreproy/72362/fichero/TFM-2362+Carmona+Pardo.pdf

https://cybertesis.unmsm.edu.pe/bitstream/handle/20.500.12672/18207/Huaquipaco_es.pdf?sequence=1&isAllowed=y

https://www.alimarket.es/sanidad/informe/346291/informe-2022-del-sector-geriatrico-en-espana

https://www.domusvi.es/memoria-anual-domusvi/

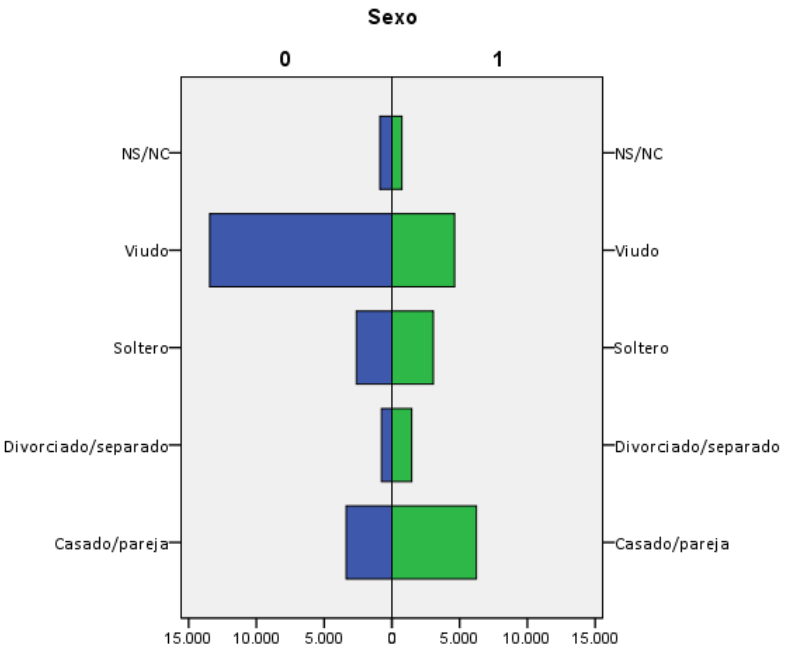
12 Anexos

Anexo 1: Evaluación de la calidad del dato

La evaluación de la calidad del dato se realizó de forma sistemática siguiendo las pautas habituales de un análisis exploratorio de datos. Las variables continuas como la edad, el índice de Barthel, el MiniMental y el índice de masa corporal (IMC) se estudiaron con los estadísticos básicos (media, mediana, cuartiles, desviación estándar, rango y valores perdidos) y gráficos tipo *box-plot* o diagramas de caja; Para las variables dicotómicas, como el sexo y la variable resultado (hospitalización sí/no) resultó interesante, además de la distribución de frecuencias y la existencia de valores perdidos, estudiar la distribución de los estadísticos descriptivos de algunas variables continuas de interés, como la edad, como veremos algún ejemplo. Los *box-plot* también son útiles para estos casos.

Sexo y edad son las variables de desagregación más utilizadas por ser las más relevantes en cualquier estudio epidemiológico. Todas aquellas variables categóricas nominales, como el motivo de ingreso en el centro, el estado civil, el nivel de ingresos, la CCAA (calculada a partir del código postal), situación laboral, procedencia… se analizan mediante distribuciones de frecuencias e histogramas.

Gráfica 1: Distribución por sexo de la variable estado civil (0=Mujeres; 1=Hombres)



Este análisis exploratorio nos permite evaluar variable a variable si existen valores erróneos, anómalos y perdidos y tomar las decisiones más adecuadas para el tratamiento de los datos. Esta limpieza o tratamiento de los datos implica en ocasiones realizar asunciones que han de basarse en el conocimiento funcional de cómo son recogidos los datos y en las implicaciones que pueden tener en los resultados, intentando optar siempre por las opciones más conservadoras de menor

impacto potencial en los resultados. A continuación, presentaremos los artefactos o transformaciones más importantes que fueron realizados.

En muchas ocasiones las clases de las variables categóricas nominales fueron **agrupadas**, manteniendo el sentido funcional, para que entrasen en el modelo con el menor número de categorías posibles. Así, por ejemplo, la variable "motivo de ingreso", que se refiere al motivo de internamiento en el centro sociosanitario, pasó de tener 17 categorías diferentes a tener siete (tabla 1): 1. Problemas familiares y sociales; 2. Dificultades en las actividades de la vida diaria (AVD); 3. Ingreso por orden judicial, del ayuntamiento o de la diputación; 4. Ingreso voluntario; 5. Ingreso para rehabilitación o terapéutico; 6. Ingreso temporal o por convalecencia; 9. Otros. Estas agrupaciones también ayudan en la interpretación de los resultados.

Tabla 1: Distribución de frecuencias de variable motivo de ingreso antes y después de ser agrupada

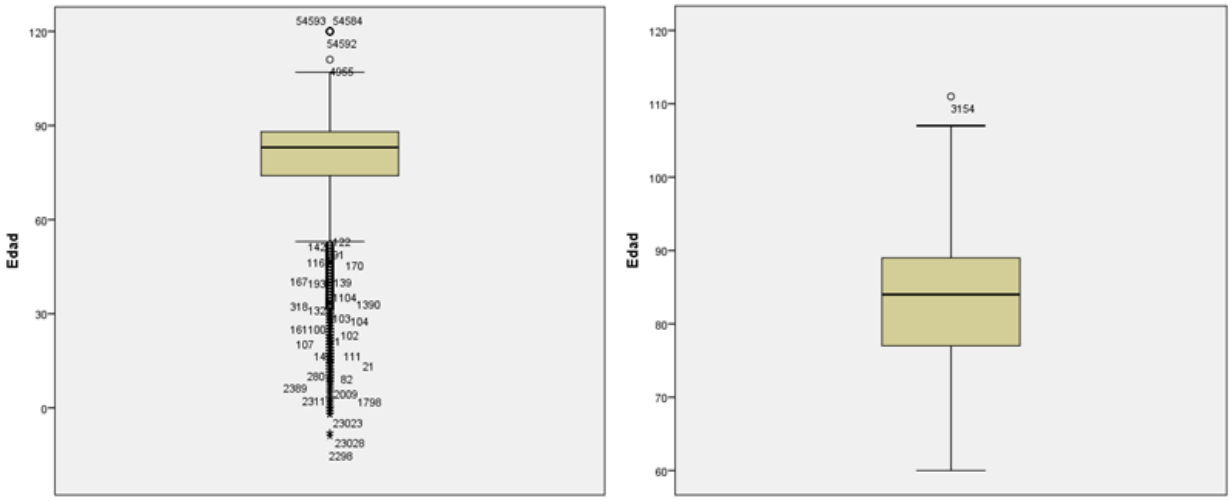
	Frecuencia	Porcentaje
Ausencia de Familia	197	1,2
Convalecencia	110	0,7
Descarga de Familia	2155	13,5
Dificultad para cubrir las AVD	7993	50,2
Dificultad/Seguimiento Tratamiento en Domicilio	1693	10,6
Emergencia Social	605	3,8
Falta de Soporte Familiar	271	1,7
Ingreso Temporal	247	1,5
Internamiento Terapéutico	388	2,4
Orden de Ayuntamiento	90	0,6
Orden de Diputación	144	0,9
Orden Judicial (Auto de internamiento)	240	1,5
Otros	576	3,6
Rechazo de Familia	20	0,1
Rehabilitación	124	0,8
Vivienda Inadecuada	21	0,1
Voluntario	1062	6,7
Total	15936	100,0

	Frecuencia	Porcentaje
1	7355	19,8
2	22182	59,6
3	1102	3,0
4	2022	5,4
5	1668	4,5
6	1467	3,9
9	1394	3,7
Total	37190	100,0

En cuanto al tratamiento de la **edad,** se trata de una variable fundamental en cualquier estudio epidemiológico que, además, aporta un gran valor explicativo cuando analizamos resultados en salud. En primer lugar, se seleccionaron aquellos residentes con edad mayor o igual a 60 años, pues son los que cumplen los criterios de inclusión. Después, se abordó el problema de los valores anómalos, que en este caso eran *outliers* -valores alejados-*,* considerándose "erróneos" los casos por encima de los 111 años, después de un cuidadoso estudio de su distribución. Se trataba de 16 casos que, junto con los 24 valores perdidos, en total no superaban el 0,1% de la muestra. Sus valores se sustituyeron por la mediana (83 años) por ser en este caso una medida mucho más robusta que la media. Las gráficas 2 y 3 presentan la distribución de la variable edad antes y después del tratamiento de los datos.

La variable **sexo** se encontraba muy desbalanceada en el *dataset* inicial, pues partíamos de un 63% de mujeres (imagen 20). Al ser precisamente la variable de estudio y de desagregación, se utilizó la limpieza de los registros con menor calidad del dato para compensar ligeramente su desbalanceo, obteniéndose finalmente un 57% de mujeres en la muestra final. Esto fue posible porque los registros con menor calidad de la información eran más frecuentes en mujeres (63,3% de los registros con DataQ=0 eran de mujeres).

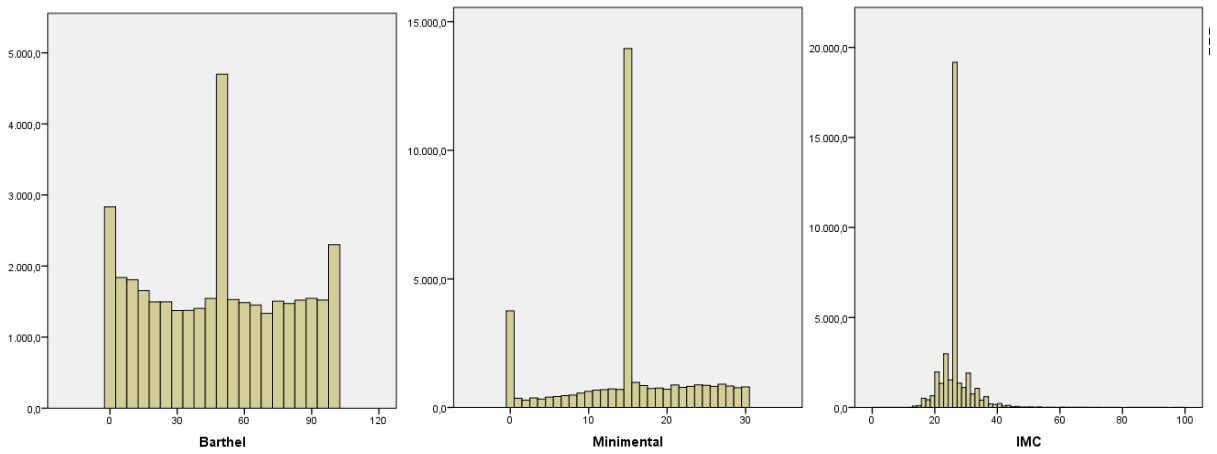
Gráficas 2 y 3: *Box-plot* de la variable edad antes y después del tratamiento de los datos



El caso de las variables Barthel, IMC y MiniMental, fue más complicado de tratar porque los valores perdidos/erróneos se situaban entre el 5% y el 10% de la muestra. Se analizaron estos registros de forma detallada, observando sus valores en el resto de variables y finalmente se decidió sustituirlos por la mediana en el caso del IMC (mediana 26) y por la media en el caso del índice de Barthel y del MiniMental (media 50 y 15 respectivamente) por tratarse de escalas con un rango de valores ya fijado de antemano (entre 0-100 en Barthel y entre 0-30 en Minimental). Estas asunciones son artefactos que nos permiten mantener a todos estos sujetos en el estudio sin alterar de forma significativa los valores medios de la variable, basándonos en el principio de regresión a la media. Se trata de todas formas de una asunción que debemos de tener en cuenta

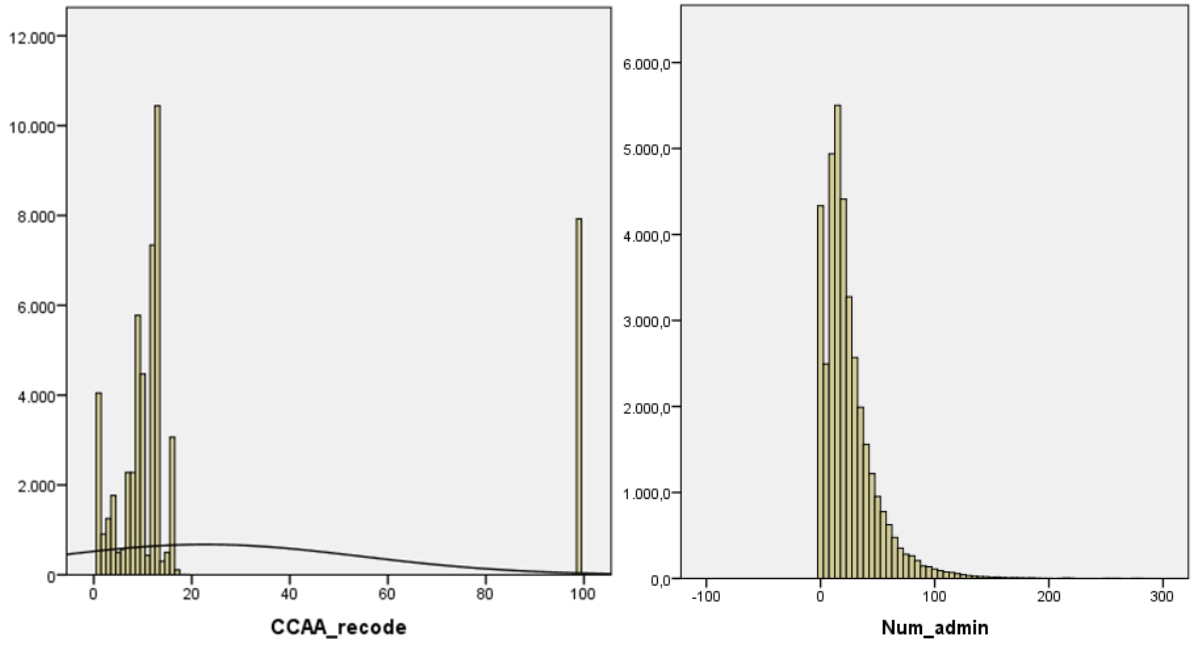
a la hora de interpretar los resultados. Las gráficas 4, 5 y 6 muestran claramente los efectos de este tratamiento.

Gráficas 4, 5 y 6: Histogramas de las variables Barthel, Minimental e IMC después del tratamiento de los datos



En otras ocasiones, los valores perdidos se asignaron con un código sin significado, como el código 99 en el caso de la variable CCAA, que fue extraída de dataset inicial a partir del campo "código postal". Como vemos en el histograma de la gráfica 6 los valores perdidos en este campo estaban por encima del 10% de los casos (alrededor de un 22%) y podría ser por este motivo por el que esta variable no siempre resulta estadísticamente significativa en los modelos estudiados. En otras variables como el número de administraciones de cualquier medicamento (Num_admin) se obtuvo una distribución "semi-normal" (gráfica 7), que nos sugiere fiabilidad del registro.

Gráficas 6 y 7: Histogramas de las variables CCAA y número de administraciones, después del tratamiento de los datos



En cuanto al tratamiento de la información clínica como diagnósticos crónicos registrados con códigos CIE-10 y medicamentos mediante los códigos ATC, el reto fue afrontar la elevada cardinalidad de estos códigos puesto que ambas clasificaciones tienen una gran granularidad (la CIE-10 consta de unos 100.000 códigos, por ejemplo). La solución que adoptamos fue por un lado truncar los códigos, quedándonos así con los 2 primeros caracteres en el caso de los diagnósticos CIE-10 y los 3 primeros caracteres en el caso de los ATC, lo que se corresponde con los grupos terapéuticos predefinidos ATC2. En el anexo 2 se detallan ambas codificaciones terminológicas (CIE-10 y ATC). En cuanto a los diagnósticos, además, estos fueron agrupados posteriormente en los siguientes grupos con sentido clínico, para facilitar la interpretación funcional de los resultados:

Enfermedades infecciosas=A0 + A1 + A2 + A3 + A4 + A5 + A6 + A8 + B0 + B1 + B2 + B3 + B4 + B5 + B6 + B7 + B8 + B9

Neoplasias malignas=C0 + C1 + C2 + C3 + C4 + C5 + C6 + C7 + C8 + C9

Neoplasias benignas=D0 + D1 + D2 + D3 + D4

Enfermedades de la sangre=D5 + D6 + D7 + D8

Diabetes Mellitus=E0 + E1

Desnutrición=E4 + E5

Obesidad y Sobrepeso=E6

Otras enfermedades endocrinas=E2 + E3 + E4 + E5 + E6 + E7 + E8

Enfermedades endocrinas=E0 + E1 + E2 + E3 + E4 + E5 + E6 + E7 + E8

Trastornos mentales=F0 + F1 + F2 + F3 + F4 + F5 + F6 + F7 + F8 + F9

Sistema Nervioso=G0 + G1 + G2 + G3 + G4 + G5 + G6 + G7 + G8 + G9

Enfermedades de los ojos y los oidos=H0 + H1 + H2 + H3 + H4 + H5 + H6 + H7 + H8 + H9

Enfermedad cardiovascular=I0 + I1 + I2 + I3 + I4 + I5 + I6 + I7 + I8 + I9

Enfermedad cerebrovascular=I6

Hipertensión arterial y asimilados=I1

Isquemia y enfermedad cardíaca pulmonar=I2

Enfermedades del sistema respiratorio=J0 + J1 + J2 + J3 + J4 + J6 + J7 + J8 + J9

Enfermedades del sistema digestivo=K0 + K1 + K2 + K3 + K4 + K5 + K6 + K7 + K8 + K9

Enfermedades de la piel=L0 + L1 + L2 + L3 + L4 + L5 + L6 + L7 + L8 + L9

Enfermedades del sistema músculo-esquelético=M0 + M1 + M2 + M3 + M4 + M5 + M6 + M7 + M8 + M9

Artritis reumatoide y asimiladas=M0 + M1 + M2 + M3 + M4 + M5

Enfermedades tratcto genito-urinario=N0 + N1 + N2 + N3 + N4 + N5 + N6 + N7 + N8 + N9

Insuficiencia renal=N1;

Malformaciones=Q0 + Q1 + Q2 + Q3 + Q4 + Q5 + Q6 + Q7 + Q8 + Q9

Signos y síntomas=R0 + R1 + R2 + R3 + R4 + R5 + R6 + R7 + R8 + R9

Traumatismos=S0 + S1 + S2 + S3 + S4 + S5 + S6 + S7 + S8 + S9 +T0 + T1

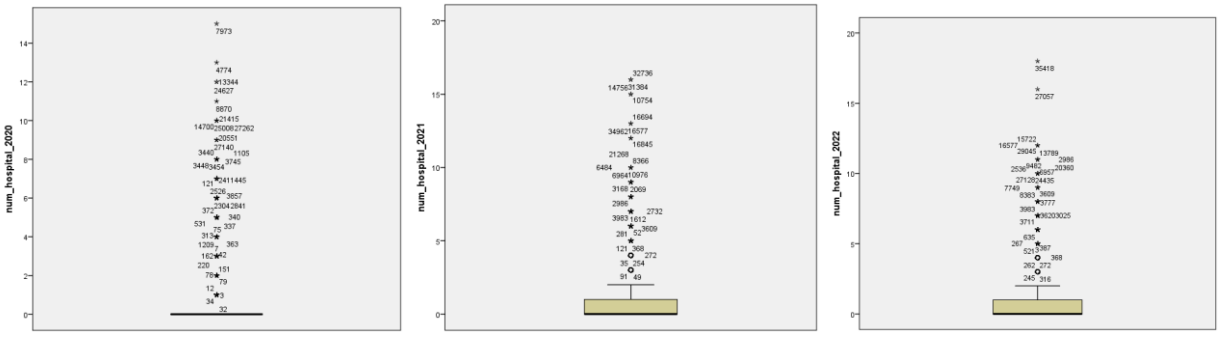
Accidentes=T2 + T3 + T4 + T5 + T6 + T7 + T8

Códigos Z=Z0 + Z1 + Z2 + Z3 + Z4 + Z5 + Z6 + Z7 + Z8 + Z9

Causas externas=V0 + V1 + W0 + W1 + X1 + X5 + X8 + Y0 + Y7 + Y8 + Y9

En cuanto a la variable resultado (número de hospitalizaciones) denominada “Total_hospital”, realizamos, además del análisis exploratorio, un análisis temporal para comprobar que estas hospitalizaciones tenían sentido contextual, puesto que estamos analizando los años de la pandemia (2020-2022). Como se ve en las gráficas 8, 9 y 10, el año 2020 fue el de menos hospitalizaciones que progresivamente se van recuperando en los dos años siguientes.

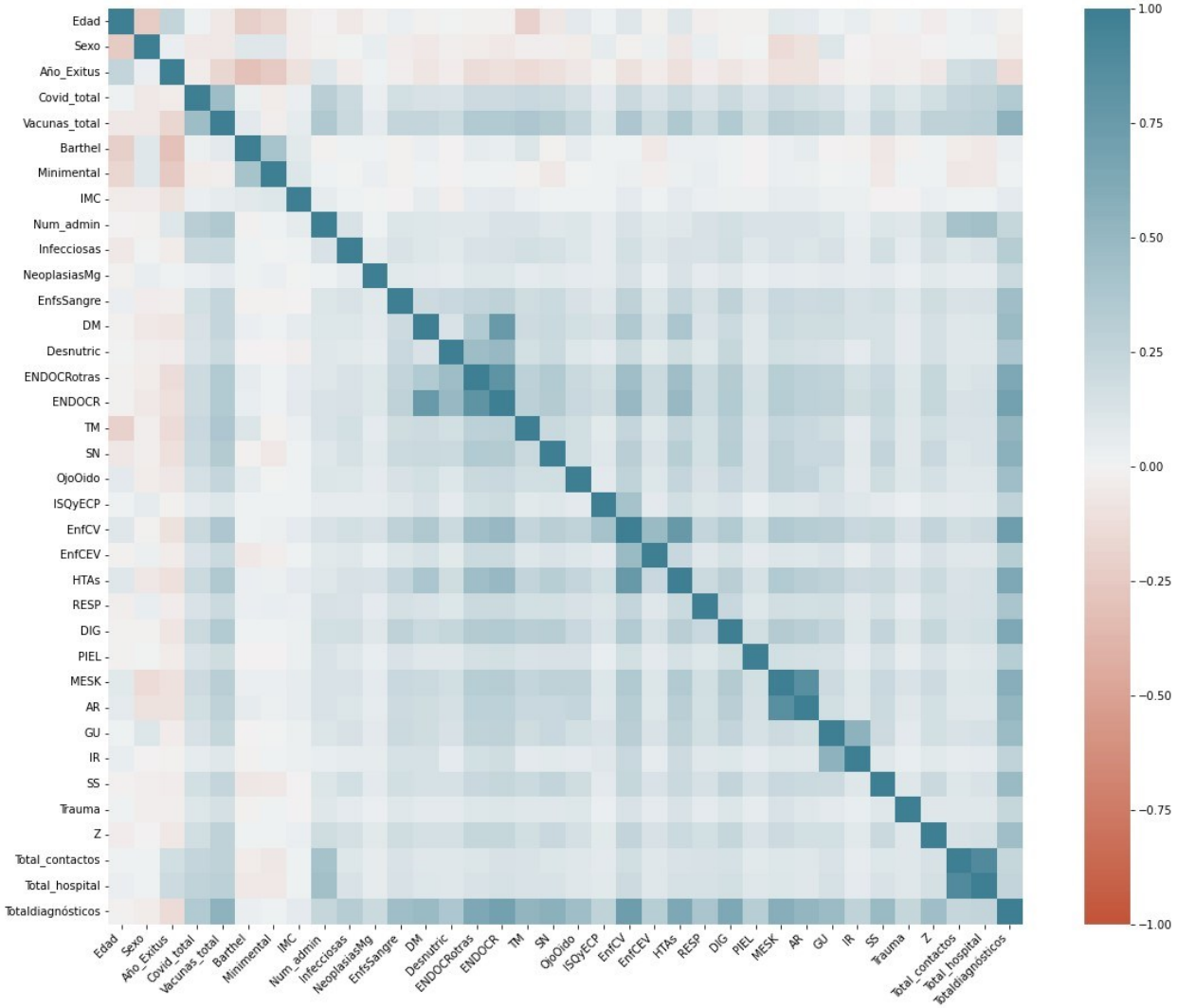
Gráficas 8, 9 y 10: Distribución del número de hospitalizaciones en los tres años del análisis (2020, 2021 y 2022 de izquierda a derecha).



Además, como parte del análisis exploratorio de los datos, se analizaron las correlaciones de todas las variables entre sí, como una primera aproximación analítica a los datos y para estudiar la presencia de colinealidad entre ellas. Como podemos observar en la gráfica 10 mediante un mapa de calor, las correlaciones más intensas ocurren entre las entidades clínicas más cercanas

(artritis reumatoide y enfermedades musculoesqueléticas) o en aquellas entidades incluidas unas dentro de otras (endocrino-diabetes; HTA y enfermedad cardiovascular), lo que se corresponde con lo esperable. También se detecta correlación entre la edad y el sexo y entre ésta y el año de defunción, así como entre ambos índices funcionales (Barthel y MiniMental).

Gráfica 10: mapa de calor que analiza la correlación entre variables



Anexo 2: Tablas de correspondencias códigos CIE-10 y ATC

A continuación, se presenta la información básica para entender las clasificaciones utilizadas en nuestro conjunto de datos para codificar, por un lado, los tratamientos farmacológicos (ATC: Anatomical, Therapeutical and Chemical classification) y, por otro lado, los diagnósticos crónicos (CIE-10: Clasificación internacional de enfermedades, versión 10). En la tabla 1 se indican los descriptivos del primer nivel de los ATC que es el nivel ATC1 y que se codifica con una letra. Cada letra se corresponde con un nivel anatómico sobre el que actúan los medicamentos codificados con esa letra, que es el primer carácter de los códigos ATC. En nuestro caso utilizamos para incorporar a los modelos los medicamentos agrupados en el segundo nivel, ATC2, que se corresponde con códigos de 3 caracteres (una letra y dos números) asignados a grupos terapéuticos. Así, por ejemplo, el código A02 se refiere a aquellos medicamentos para los trastornos de la secreción gástrica, el A10 son antidiabéticos y el C10 son fármacos modificadores de los lípidos. La tabla 2 presenta los descriptivos de las agrupaciones de dos caracteres de la CIE-10 y su correspondencia con los capítulos correspondientes de la misma.

Tabla 1: Descripciones terminológicas del primer nivel de los ATC (ATC1)

Nivel ATC	COD_ ATC	Descripción
1	A	APARATO DIGESTIVO Y METABOLISMO
1	B	SANGRE Y ORGANOS HEMATOPOYETICOS
1	C	APARATO CARDIOVASCULAR
1	D	DERMATOLOGICOS
1	G	APARATO GENITO-URINARIO Y HORMONAS SEXUALES
1	H	HORMONAS SISTEMICAS, EXCL. HORMONAS SEXUALES
1	J	ANTIINFECCIOSOS DE USO SISTEMICO
1	L	ANTINEOPLASICOS E INMUNOMODULADORES
1	M	SISTEMA MUSCULO-ESQUELETICO
1	N	SISTEMA NERVIOSO
1	P	ANTIPARASITARIOS, INSECTICIDAS Y REPELENTES
1	R	SISTEMA RESPIRATORIO
1	S	ORGANOS DE LOS SENTIDOS
1	V	VARIOS
1	Y	EFECTOS Y ACCESORIOS
1	Z	DESCONOCIDO

Tabla 2: Descripciones terminológicas del primer nivel de los ATC (ATC1)

Capítulos CIE 10	Códigos CIE-10 truncados	Agrupaciones
1 Enfermedadess infecciosas	A0	Infecciones GI
	A1	**Tuberculosis**
	A2	zoonozis bacterianas
	A3	Otras enfermedades infecciosas (meningitis, neumonía, tos ferina)
	A4	Sepsis y otras infecciones
	A5	**Sífilis, gonorrea y otras ETS**
	A6	Herpes virus y enf Lyme
	A8	Enfermedades virales y por priones del SNC
	B0	Infecciones virales caracterizadas por lesiones piel y mucosas
	B1	Otros herpesvirus humanos y hepatitis viral
	B2	VIH y otras infecciones virales
	B3	Micosis
	B4	Micosis
	B5	protozoos y helmintos
	B6	
	B7	
	B8	helmintos, pediculosis y ácaros
	B9	secuelas de infecciones/agentes infecciosos/ otras enfs infecciosas
2 Neoplasias	C0	***Neoplasias malignas***
	C1	
	C2	
	C3	
	C4	
	C5	
	C6	
	C7	
	C8	
	C9	
	D0	Neoplasias in situ
	D1	Neoplasias benignas
	D2	Otras neoplasias beningnas
	D3	Otras neoplasias de comportamiento incierto
	D4	Neoplasias de comportamiento no especificado
3 enfermedades de la sangre y órganos hematopoyéticos	D5	*Enfs sangre y órganos hematopoyéticos*
	D6	
	D7	
	D8	
4 Enfermedades endocrinas	E0	**Trastornos del tiroides y DM por afección subyacente y DM inducida**
	E1	**DM y trastornos regulación glucosa**
	E2	otras glándulas
	E3	complicadiones intraoperatorias
	E4	**desnutrición**
	E5	**otras deficiencias nutricionales**
	E6	**Sobrepeso y obesidad**
	E7	Trastornos metabólicos
	E8	otras complicaciones endocrinas
5 Trastornos mentales	F0	TM por afecciones fisiológicas conocidas
	F1	TM debidos al consumo de sustancias psicoactivas
	F2	Esquizofrenia y otros trastornos psicóticos
	F3	Trastornos afectivos
	F4	Trastorno de ansiedad y otros TM no psicóticos
	F5	Síndromes de comportamiento asociados a trastornos fisiológicos
	F6	Trastornos de personalidad y de comportamiento del adulto
	F7	Discapacidad intelectual
	F8	Trastornos del desarrollo
	F9	Trastornos de personalidad y de comportamiento de comienzo en la infancia
6 Enfermedades sistema nervioso	G0	Enf inflamatorias SNC
	G1	Atrofias sistémicas
	G2	Trastornos extrapiramidales
	G3	Enfs degenerativas y desmielinizantes
	G4	Trastornos episódicos y paroxísticos
	G5	Trastornos de los nervios, raíces y plexos nerviosos
	G6	Polineuropatías y SNP
	G7	Enf musculares y de la unión neuromuscular
	G8	Parálisis cerebral y otros síndromes paralíticos
	G9	Otros trastornos del SN
7 Enf del ojo, oído y sus anexos	H0	*Enf del ojo y sus anexos*
	H1	
	H2	
	H3	
	H4	
	H5	
	H6	*Enf del ojo, oído y sus anexos*
	H7	
	H8	
	H9	
	I0	Enfermedades reumáticas
	I1	**Enfermedades hipertensivas**
	I2	**Enf isquémica cardíaca y enf cardíaca pulmonar**
	I3	

Enf aparato respiratorio	J0	Infecciones agudas tracto superior, gripe y neumonía
	J1	
	J2	Otras infecciones agudas tracto inferior
	J3	Otras enf tracto superior
	J4	Enf crónicas tracto inferior
	J6	*Enf pulmonares por agentes externos*
	J7	
	J8	Otras enf respiratorias intersticiales y trastornos supurativos y necróticos vías inferiores
	J9	Otras enf aparato respiratorio
Enf aparato digestivo	K0	*Enf cavidad oral y glandulas salivales*
	K1	
	K2	*Enfs estómago, esófago y duodeno*
	K3	
	K4	Hernia
	K5	**Enteritis y colitis no infecciosas**
	K6	**Otras enf intestinales**
	K7	Enf del hígado
	K8	Enf vesícula biliar, vías biliares y páncreas
	K9	Otras enf aparato digestivo
Enf de la piel y del tejido subcutáneo	L0	Infecciones piel y tej subcutáneo
	L1	Trastornos ampollosos
	L2	*Dermatitis y eccema*
	L3	
	L4	*Enf de la piel y del tejido subcutáneo*
	L5	
	L6	
	L7	
	L8	
	L9	
Enf aparato musculo esquelético	M0	*Artropatías*
	M1	
	M2	
	M3	Trastornos del tejido conectivo
	M4	Dorsopatías
	M5	idem
	M6	*otras enf aparato musculo esquelético*
	M7	
	M8	Osteopatías y condropatías
	M9	idem
Enf aparato genitourinario	N0	Enf glomerulares
	N1	**IRA e IRC**
	N2	Litiasis
	N3	*Enf aparato genitourinario*
	N4	
	N5	
	N6	
	N7	
	N8	
	N9	
Malformaciones congénitas y anomalías cromosómicas	Q0	Malf congénitas SN
	Q1	*Malformaciones congénitas y anomalías cromosómicas*
	Q2	
	Q3	
	Q4	
	Q5	
	Q6	
	Q7	
	Q8	Otras malf congénitas
	Q9	Otras anomalías cromosómicas
Lesiones traumáticas, envenenamientos y consecuencias de causas externas	S0	*Traumatismos*
	S1	
	S2	
	S3	
	S4	
	S5	
	S6	
	S7	
	S8	
	S9	
	T0	Traumatismo múltiple
	T1	Traumatismo NE
	T2	Quemaduras
	T3	Quemaduras, congelación, envenenamiento
	T4	Envenenamiento
	T5	Efectos tóxicos de sustancias
	T6	otros
	T7	otros
	T8	complicaciones de atención médico quirúrgica
Causas externas	V0	Peatón lesionado en accidente de transporte
	V1	Ciclista lesionado en accidente de transporte
	W0	Caídas
	W1	Caídas
	X1	Contacto con calor y sustancias calientes
	X5	sobreesfuerzo y otras exposiciones
	X8	lesiones autoinfligidas
	Y0	agresiones
	Y7	efectos adeversos de dispositivos

Anexo 3. Interpretación de la regresión logística

3.1. Estadísticos de la regresión logística

Los modelos de regresión logística permiten modelar la relación entre una variable de respuesta (variable dependiente) binaria con variables independientes o explicativas cuantitativas y categóricas. Así como en una regresión lineal los coeficientes obtenidos representan estimaciones de los efectos cuantitativos de las variables explicativas sobre la variable respuesta, en la regresión logística se interpretan las exponenciales de los coeficientes como los efectos expresados en razones de *odds*, o razones de probabilidad.

Los modelos de regresión pueden tener una utilidad explicativa, cuando interesa saber cuáles son las variables que más influyen en la variable respuesta, o bien, una utilidad predictiva, cuando se trata de predecir la variable resultado a partir de las variables independientes. En el primer caso, se recomienda utilizar lo que se denominan "modelos parsimoniosos", con pocas variables, sólo aquellas que realmente aportan explicabilidad al modelo. En el caso del uso predictivo, que es el nuestro, lo que se intenta es maximizar el porcentaje de la variabilidad observada que es explicado por el modelo, no importando tanto el número de variables incluidas. Esto no excluye que podamos conocer cuáles son aquellas variables independientes más significativas a nivel estadístico, como veremos.

A continuación, explicamos aquellas métricas de un modelo de regresión logística más importantes para entender sus resultados. Utilizamos como ejemplo la salida de SPSS del modelo final de nuestro estudio, seleccionado con el criterio de optimización del porcentaje de variabilidad explicado, mediante la R^2 ajustada (*R^2 de Nagelkerke*).

Tabla 4: Pruebas ómnibus de coeficientes de modelo

	X2 (Chi-cuadrado)	gl (grados libertad)	Sig. (significación)
Modelo	8447,342	180	0,000

La prueba de *χ2* o "prueba de la razón de verosimilitud" se utiliza para valorar si el modelo estimado es estadísticamente significativo, esto es, si el conjunto de términos que incluye es predictivo de la muestra. Valora la significación del conjunto de variables predictoras incluidas. En este caso (tabla 4: "Pruebas ómnibus de coeficientes de modelo") observamos que sí es estadísticamente significativo pues la p (significación) es menor de 0,05 (p=0,000).

La tabla 5 ("Resumen del modelo") nos muestra los estadísticos que describen el ajuste global del modelo derivados de la verosimilitud del modelo, puesto que en este caso el modelo está estimado por el método de máxima verosimilitud. El coeficiente *R^2 de Nagelkerke* estima la proporción de la variabilidad de la variable respuesta explicada por la regresión logística. Se trata de un índice corregido que vale 1 si el modelo explica el 100% de la incertidumbre de los datos. En nuestro caso el modelo explica casi un 33% de la variabilidad (0,327).

Tabla 5: Estadísticos resumen del modelo

Escalón	**Logaritmo de la verosimilitud -2**	**R^2 de Cox y Snell**	**R^2de Nagelkerke**
1	18135,097	0,225	0,327

Tabla 6: Prueba de Hosmer y Lemeshow

Escalón	**χ2**	**gl**	**Sig.**
1	10,850	8	0,210

Otra forma de valorar la calibración o "bondad de ajuste" del modelo es la "*prueba de Hosmer y Lemeshow*"(tabla 6) que valora la concordancia entre la probabilidad observada en la muestra y las probabilidades predichas por el modelo. Esta prueba indica un buen ajuste cuando la *χ2* no es significativa (p > 0,05). En nuestro modelo final comprobamos que el ajuste es bueno puesto que esta prueba no es significativa (p=0,210).

Una de las tablas de salida de SPSS que más información nos aporta sobre la regresión logística son las "Variables en la ecuación". A continuación, se muestra una selección de las variables más relevantes en el modelo final para explicar la interpretación de los estadísticos. Para ello, hemos seleccionado tres de los estadísticos más importantes: el error estándar, la significación estadística y el exponencial de Beta (coeficiente que el modelo atribuye a cada variable).

En la regresión logística el coeficiente de la ecuación de regresión beta (B) ha de transformarse en exponencial y se interpreta como una razón de probabilidades *(odds ratio*) y en el caso de un estudio transversal, como el nuestro, como una razón de prevalencias. Se interpreta como el factor por el que se multiplica la prevalencia, o presencia de la variable resultado, en el paso de un valor a otro de la variable explicativa, por ejemplo, en este caso del valor 0 al valor 1. Lo veremos con varios ejemplos extraídos del análisis.

En la tabla 7 vemos como la variable estado civil no resulta estadísticamente significativa en su conjunto (p > 0,05) pero, sin embargo, vemos que el valor "2" de esta variable (que se corresponde con divorciado/separado) es significativo bajo el criterio de p<0,05 con una RP estimada (Exp(B)) de 1,231. Esto significa que, con respecto a los otros valores de estado civil, el hecho de tener registrado un valor "2" (separado/divorciado) en esta variable implica un 23% más prevalencia de hospitalización (variable resultado), o sea, que estar separado o divorciado sería un "factor de riesgo" de ingresar en un hospital.

La misma interpretación se puede hacer del motivo de ingreso “3” (Ingreso por orden judicial, de la diputación o del ayuntamiento) pero en este caso al tener un coeficiente menor de 1 (0,660) significa que este motivo hace que disminuya la prevalencia de ingreso en el hospital un 40%. Se trataría de un “factor protector” de hospitalización, por decirlo de una manera más fácil de entender.

Por tanto, la razón de prevalencias se interpreta como una OR o un Riesgo relativo (RR): si es mayor de 1 es un factor de “riesgo” y si es menor de 1 es un factor “protector”. Si la variable es continua, como la edad, se interpreta como el factor por el que se multiplica la prevalencia de hospitalización para cada unidad de incremento en esa variable continua, esto es, por cada año de edad. En las variables dicotómicas como el sexo (0=mujeres y 1=hombres) interpretamos de la misma forma el cambio en una unidad, de 0 a 1, o sea, el hecho de ser hombre significa en este caso un riesgo de ingreso casi un 33% superior al de ser mujer (RP=1,327).

Tabla 7: Variables en la ecuación

	Error estándar	Sig.	Exp(B)
Est_Civil		0,078	
Est_Civil(1)	0,067	0,846	1,013
Est_Civil(2)	0,081	0,040	1,231
Est_Civil(3)	0,071	0,350	1,068
Est_Civil(4)	0,066	0,298	1,071
Motivo_ingreso		0,000	
Motivo_ingreso(1)	0,067	0,484	0,954
Motivo_ingreso(2)	0,063	0,018	1,161
Motivo_ingreso(3)	0,094	0,000	0,660
Motivo_ingreso(4)	0,080	0,023	1,198
Motivo_ingreso(5)	0,094	0,000	0,283
Motivo_ingreso(6)	0,088	0,000	0,721
Edad	0,002	0,000	1,010
Sexo	0,032	0,000	1,327
CCAA	0,000	0,059	1,001
Barthel	0,000	0,000	0,997
Minimental	0,002	0,000	0,982
NeoplasiasMg	0,060	0,001	1,226
Desnutric	0,050	0,180	0,935
TM	0,035	0,028	0,927
SN	0,035	0,000	0,873
OjoOido	0,041	0,000	0,858
PIEL	0,054	0,016	0,879
RESP	0,043	0,038	1,093
EnfCEV	0,046	0,076	0,922
A01	0,086	0,096	1,154

A02	0,028	0,002	0,915
A07	0,040	0,001	1,136
A15	0,108	0,000	1,537
B03	0,028	0,000	1,155
B04	0,147	0,300	0,859
B05	0,055	0,000	0,607
C01	0,047	0,000	1,287
C02	0,062	0,279	1,070
Num_admin	0,013	0,000	0,942
DataQ	0,013	0,000	1,074
Covid	0,073	0,000	1,585
Vacunas	0,015	0,000	0,946

Destacamos como curiosidad los resultados de las variables “Covid” (si ha pasado el COVID) y “Vacunas”, que se refiere a si el individuo ha recibido o no alguna vacuna frente al COVID-19. Este modelo nos dice que de forma estadísticamente significativa la prevalencia de hospitalizaciones es un 58% superior (RP=1,585) en aquellos con algún diagnóstico de COVID y también que el haber recibido alguna vacuna frente al COVID resulta “protector” con respecto a las hospitalizaciones (aunque el resultado es bajo, quizás por el diseño transversal de este estudio). Otras variables observadas en la tabla de abajo son el número de administraciones de cualquier medicamento (“Num_admin”) que es una variable continua que vemos que disminuye ligeramente la prevalencia de ingresos hospitalarios o la presencia de algún diagnóstico de neoplasia maligna (“NeoplasiasMg”) que aumenta un 22% la prevalencia de ingresos de forma estadísticamente significativa.

3.2 Indicadores de precisión de la predicción

Aplicando a cada uno de los sujetos del conjunto de datos la ecuación de regresión estimada por el modelo, obtenemos la probabilidad estimada de hospitalización. Por defecto el modelo asigna el punto de corte de 0,5 a cada uno de los sujetos para calcular el valor pronosticado, así, el sujeto se asigna al grupo pronosticado 1 (hospitalización) cuando la probabilidad estimada supera el umbral mayor o igual a 0,5 y grupo pronosticado 0 (no hospitalización) si éste es menor de 0,5. Se puede considerar en este caso la regresión logística como una prueba diagnóstica del evento a predecir y calcular, a partir de la tabla de clasificación, los indicadores propios de las pruebas diagnósticas como sensibilidad, especificidad, valor predictivo positivo , valor predictivo negativo y la curva ROC.

Dado un estimador para una variable estadística discreta binaria que clasifica la población de estudio en sujetos que son hospitalizados (1) y sujetos que no lo son (0), la **sensibilidad** nos indica la capacidad del modelo para predecir como hospitalizados aquellos casos que realmente han sido hospitalizados, esto es, la proporción de hospitalizados correctamente identificados por el

modelo. La sensibilidad caracteriza la capacidad de la prueba para detectar el evento en aquellos sujetos que lo han sufrido.

La **especificidad**, por el contrario, nos indica la capacidad de nuestro estimador para predecir como no hospitalizados los casos que realmente no han sido hospitalizados, esto es, la proporción de no hospitalizados correctamente identificados. La especificidad caracteriza la capacidad de predecir la ausencia de hospitalización en aquellos sujetos que no la han tenido. De este modo, si un test tiene una sensibilidad del 98% y una especificidad del 92% quiere decir que su tasa de falsos negativos (aquellos que el modelo predice que no han sido hospitalizados, pero sí lo han sido), es de un 2% y la de falsos positivos (el modelo predice que han sido hospitalizados, pero en realidad no lo han sido) es de un 8%.

Otros indicadores de predicción interesantes son la probabilidad de que un individuo para el que se haya obtenido un resultado (1) -hospitalización-, haya sido efectivamente hospitalizado; y lo contrario, la probabilidad de que un individuo con un resultado (0) -no hospitalización- no haya sido hospitalizado. Estos indicadores se conocen como valores predictivos de una prueba positiva (**valor predictivo positivo, VPP**) y de una prueba negativa (**valor predictivo negativo, VPN**) respectivamente. Todos estos indicadores se calculan a partir de la tabla de clasificación que compara los valores observados con los esperados o predichos por el modelo. Además, las curvas ROC son útiles para comparar el poder de clasificación de diferentes modelos ajustados. En general se puede afirmar que el mejor modelo es al que le corresponde una mayor área bajo la curva. Esta área bajo la curva (AUC) es otra medida de la validez predictiva del modelo. Todos estos indicadores se presentan en el anexo 5.

El poder de clasificación generalmente se considera aceptable si la especificidad y la sensibilidad superan el 75%, pero este umbral puede ser modificado según el objetivo de la clasificación. En nuestro caso, no tratamos de predecir ingresos sino de detectar las diferencias de comportamiento en hombres y mujeres en lo que respecta a hospitalizaciones, por lo que obtener un alto poder de clasificación no es nuestro principal objetivo, aunque sí hubiese sido deseable.

Anexo 4. Interpretación modelos XGBoost

4.1 Hiperparámetros

Los modelos de IA implementados en los distintos leguajes cuentan con una serie de variables de configuración que permiten ajustar el entrenamiento y los modos de funcionamiento interno a la hora de ejecutar el entrenamiento del modelo. A estas variables se las suele denominar hiperparámetros. A continuación, se señalan los más importantes en XGBoost:

Tasa de Aprendizaje (*learning rate*): El *learning rate* controla la contribución de cada árbol en el ensamble al resultado final. Un valor menor reduce la contribución de cada árbol y puede ayudar a evitar el sobreajuste. Por ejemplo, si el learning rate es 0.1, cada árbol contribuirá con el 10% de su predicción al resultado final.

Profundidad máxima de los árboles (*Max Depth*): Establece la profundidad máxima de cada árbol del ensamblaje. La limitación en la profundidad puede prevenir el sobreajuste, ya que los árboles más profundos pueden capturar relaciones ruidosas en los datos. Por ejemplo: Si *Max Depth* es 5, los árboles en el ensamble tendrán como máximo cinco niveles.

Números de árboles para ensamblaje (*Number of trees*): Este hiperparámetro establece cuántos árboles se construirán en el ensamble. Un número mayor de árboles puede mejorar el rendimiento, pero también aumenta el tiempo de entrenamiento. Ejemplo: Si el número de árboles es 100, se construirán 100 árboles en el ensamble.

El objetivo define la función de pérdida a optimizar durante el entrenamiento. Dependerá de la naturaleza del problema, como regresión o clasificación binaria/multiclase. Por ejemplo, para una clasificación binaria, el objetivo podría ser "*binary:logistic*" para optimizar la función de pérdida logística.

***Scale Pos Weight* (Peso de Clases Positivas):** Este hiperparámetro ajusta los pesos de las clases en un problema de clasificación desbalanceada. Puede ser útil para dar más importancia a la clase minoritaria. Por ejemplo, si el peso de la clase positiva es 5, las instancias de la clase positiva tendrán cinco veces más peso durante el entrenamiento.

En nuestro caso se ha hecho una búsqueda manual del mejor conjunto, pero el método teórico implicaría realizar una búsqueda exhaustiva en el espacio de hiperparámetros para encontrar la combinación que maximice el rendimiento, siempre sobre un conjunto de datos específicos.

Es decir, para la búsqueda del mejor conjunto de hiperparámetros posible sería necesario seleccionar siempre el mismo conjunto de datos para ir probando en él las diferentes alternativas y realizar cada entrenamiento. Una forma efectiva de hacer lo anterior es dividir el conjunto total de datos disponibles en tantos conjuntos de entrenamiento, validación y prueba como escenarios diferentes se vayan a probar. Esto permite ajustar y evaluar diferentes combinaciones de hiperparámetros de una manera efectiva. Adicionalmente a la preparación del conjunto global de

datos hay que definir qué hiperparámetros son los que se busca ajustar y definir un rango de posibles valores para cada uno de ellos.

Se pueden utilizar diferentes estrategias para buscar los mejores hiperparámetros:

- **Búsqueda en Cuadrícula**: Define una cuadrícula de combinaciones de hiperparámetros y prueba todas las combinaciones posibles. Puede ser intensivo en tiempo computacional, pero es exhaustivo.
- **Búsqueda Aleatoria:** Se seleccionan combinaciones de hiperparámetros al azar dentro de los rangos definidos. Es más eficiente en términos de tiempo que la búsqueda en cuadrícula.
- **Optimización Bayesiana**: Esta estrategia utiliza algoritmos de optimización bayesiana para encontrar las combinaciones de hiperparámetros más prometedoras. Puede ser más eficiente que las búsquedas anteriores.

El resto de las fases siguen la misma estructura que mostramos en nuestro ejemplo: para cada combinación de hiperparámetros, se entrena el modelo con esos valores y se evalúa su rendimiento. Después se utiliza la estrategia elegida para iterar sobre las diferentes combinaciones y se va registrando el rendimiento de cada modelo en el conjunto de validación. Se comparan los resultados de los diferentes modelos, y una vez seleccionado el que aparece mejor, se evalúa este modelo en el conjunto de prueba para obtener una estimación realista de su rendimiento.

4.2 Importancia de las características

La matriz de importancia en XGBoost proporciona información cómo de influyentes son las características en el modelo final. Los siguientes conceptos explican la salida de XGBoost con respecto a la importancia de las características:

***Feature* (Característica):** Esta columna muestra el nombre de la característica (variable) para la cual se está calculando la importancia.

***Gain* (Ganancia):** La ganancia representa la contribución acumulativa de una característica a través de la creación de árboles en el ensamble. Cuanto mayor sea la ganancia, más influyente es la característica en términos de mejorar la predicción del modelo. La ganancia se calcula sumando las mejoras (ganancias) en la función de pérdida que resultan de las divisiones basadas en la característica en todos los árboles del modelo.

***Cover* (Cobertura):** La cobertura indica la cantidad de datos que se beneficiaron de las divisiones basadas en una característica en el proceso de construcción de árboles. Es una medida de cuánto afecta una característica en las decisiones del modelo. La cobertura se calcula sumando la cantidad de ejemplos de entrenamiento que pasaron por las divisiones basadas en la característica en todos los árboles.

***Frequency* (Frecuencia):** La frecuencia es simplemente el número de veces que se utilizó la característica para hacer divisiones en el conjunto de árboles del modelo. Indica cuán a menudo la característica se utiliza para tomar decisiones durante el proceso de predicción.

***Importance* (Importancia):** La importancia se calcula combinando la ganancia y la cobertura de una característica de manera ponderada. Puede verse como una medida general de la importancia relativa de la característica en el modelo. Algunas implementaciones también normalizan la importancia para que la suma de todas las importancias sea igual a 1, lo que facilita la comparación entre características.

4.3 Explicación de la salida de los modelos de XGBoost

Confusion Matrix: La matriz de confusión muestra la clasificación real (verdadero o falso) versus la clasificación predicha (positivo o negativo) para el conjunto de datos de prueba.

- La celda superior izquierda (0, 0) representa los verdaderos negativos (TN) - ejemplos que fueron clasificados correctamente como negativos.
- La celda superior derecha (0, 1) representa los falsos positivos (FP) - ejemplos que fueron clasificados incorrectamente como positivos.
- La celda inferior izquierda (1, 0) representa los falsos negativos (FN) - ejemplos que fueron clasificados incorrectamente como negativos.
- La celda inferior derecha (1, 1) representa los verdaderos positivos (TP) - ejemplos que fueron clasificados correctamente como positivos.

Accuracy: La precisión general del modelo en predecir correctamente ambas clases (positiva y negativa).

95% CI: Intervalo de confianza del 95% para la precisión del modelo.

No Information Rate: La tasa de aciertos si el modelo siempre predijera la clase mayoritaria.

Kappa: Coeficiente de Kappa, una medida de la concordancia entre las predicciones del modelo y las observaciones reales, que ajusta la precisión observada por la probabilidad de coincidencia al azar.

Mcnemar's Test P-Value: El valor p del test de McNemar, que evalúa si existe una diferencia significativa entre las tasas de error de los dos clasificadores comparados.

***Sensitivity**:* También conocida como tasa de verdaderos positivos o *recall*, mide la capacidad del modelo para identificar correctamente los casos positivos.

Specificity: También conocida como tasa de verdaderos negativos, mide la capacidad del modelo para identificar correctamente los casos negativos.

Pos Pred Value: Valor predictivo positivo, representa la proporción de predicciones positivas que son correctas.

Neg Pred Value: Valor predictivo negativo, representa la proporción de predicciones negativas que son correctas.

Prevalence: La prevalencia real de la clase positiva en los datos de prueba.

Detection Rate: Tasa de detección, mide la proporción de casos positivos que el modelo ha detectado correctamente.

Detection Prevalence: La prevalencia estimada de casos positivos detectados por el modelo.

Balanced Accuracy: Promedio de sensibilidad y especificidad, proporciona una medida equilibrada del rendimiento en ambas clases.

'Positive' Class: Indica cuál es la clase positiva en la evaluación.

El **valor del coeficiente Kappa** es una medida que evalúa la concordancia entre las predicciones de un modelo y las observaciones reales, ajustando la precisión observada por la probabilidad de coincidencia al azar. El rango de valores de Kappa va desde -1 hasta 1, donde:

- Un valor de Kappa de -1 indica una completa discordancia entre las predicciones y las observaciones reales.
- Un valor de Kappa de 0 indica que las predicciones son iguales a las que se esperaría al azar.
- Un valor de Kappa de 1 indica una concordancia perfecta entre las predicciones y las observaciones reales.

En términos generales, se interpreta el valor de Kappa de la siguiente manera:

- Kappa < 0: Discordancia peor que al azar.
- 0 < Kappa < 0.2: Discordancia ligera.
- 0.2 < Kappa < 0.4: Discordancia moderada.
- 0.4 < Kappa < 0.6: Discordancia sustancial.
- 0.6 < Kappa < 0.8: Concordancia moderada a sustancial.
- 0.8 < Kappa < 1: Concordancia casi perfecta.

Dado que el valor de Kappa se basa en una comparación con la probabilidad de coincidencia al azar, es especialmente útil cuando las clases están desbalanceadas y se espera cierta cantidad de coincidencias al azar. Sin embargo, también es importante tener en cuenta el contexto específico del problema al interpretar el valor de Kappa. Por ejemplo, en algunas aplicaciones sanitarias, se puede considerar que un valor de Kappa en el rango de 0.6 a 0.8 es bastante aceptable, mientras que, en otras áreas, se podría requerir un nivel de concordancia más alto.

En resumen, no existe un "mejor" valor de Kappa universal, ya que su interpretación depende del dominio del problema y de las expectativas en cuanto a la concordancia entre las predicciones y las observaciones reales. Es importante considerar el contexto y la magnitud de Kappa en relación con la naturaleza y la importancia del problema en cuestión.

El **test de McNemar** es una prueba estadística utilizada para evaluar si existe una diferencia significativa en las tasas de error entre dos clasificadores o modelos en un problema de clasificación binaria. El valor p del test de McNemar se utiliza para determinar si la diferencia entre las tasas de error es estadísticamente significativa.

El valor p del test de McNemar puede variar entre 0 y 1. Sin embargo, el rango de valores que se considera "normativo" o "significativo" depende del nivel de significación (alfa) que se elija. El nivel de significación es un umbral predefinido que se utiliza para determinar si se rechaza o no la hipótesis nula.

En general, en la mayoría de las disciplinas científicas, se utiliza un nivel de significación común de 0.05 (5%). Esto significa que si el valor p calculado es menor que 0.05, se considera que hay evidencia suficiente para rechazar la hipótesis nula y concluir que existe una diferencia significativa entre las tasas de error de los dos clasificadores.

Por lo tanto, el rango de valores "normativos" o "significativos" para el valor p del test de McNemar es generalmente:

- Valor $p < 0.05$: Se considera que hay una diferencia estadísticamente significativa entre las tasas de error.
- Valor $p \geq 0.05$: No hay suficiente evidencia para concluir que hay una diferencia estadísticamente significativa entre las tasas de error.

Adicionalmente a lo anterior, solo comentar de forma general:

- **Especificidad**: El modelo con el valor más alto es mejor para identificar casos negativos verdaderos.
- **Sensibilidad**: El modelo con este valor más alto es mejor para identificar casos positivos verdaderos.
- **Precisión**: El mayor valor indica que predice correctamente más casos en general.
- Valor Predictivo Positivo: Cuanto más alto indica que cuando predice una clase positiva, es más probable que sea correcta.
- **Kappa**: Ambos modelos tienen valores de Kappa relativamente bajos (sobre 0,400), lo que sugiere un acuerdo moderado entre las predicciones y las etiquetas reales.

Anexo 5: Resultados estadística convencional

5.1 Modelo general

El mejor de los modelos obtenidos fue el usado para los cálculos de sensibilidad y especificidad y para la comparación con los resultados de *machine learning.* Este modelo final se seleccionó, tras realizar diversas pruebas, con el criterio de mejor variabilidad explicada (R^2 de Nagelkerke), junto con la valoración del test de Hosmer-Lemeshow -con menor significación estadística-. Estos estadísticos son explicados en el anexo 3.

El modelo final utiliza como variable resultado el total de hospitalizaciones, sin incluir las visitas a urgencias. Las variables explicativas incluidas en el modelo se muestran a continuación en la sintaxis de la regresión logística (SPSS). Estas variables son aquellas por las que estamos ajustando el modelo, esto es, los resultados diferenciales entre hombres y mujeres están ajustados, entre otras variables, por los grandes grupos diagnósticos y de medicamentos recibidos, por la edad, el estado civil, el índice de dependencia de Barthel y el de estado mental Minimental, el estado de vacunación de COVID y el haber pasado el COVID, entre otras.

Sintaxis del modelo de regresión logística seleccionado:

```
LOGISTIC REGRESSION VARIABLES Total_hospital_bin

/METHOD=ENTER NivelEstudios ActivLaboral Est_Civil MotivoIngreso Edad Sexo CCAA Covid Vacunas Barthel
Minimental IMC Num_admin DataQ Total_hospital Acc Trauma Malf GU AR MESK  DIG Infecciosas NeoplasiasMg
NeoplasiasBg EnfsSangre DM Desnutric OBySP TM SN OjoOido PIEL ENDOCR RESP EnfCV EnfCEV ISQ  HTA A01
A02 A03 A04 A05 A06 A07 A09 A10 A11 A12 A13 A15 A16 B01 B02 B03 B04 B05 B06 C01 C02 C03 C04 C05 C06
C07 C08 C09 C10 D01 D02 D03 D04 D05 D06 D07 D08 D09 D10 D11 F00 F40 F62 G01 G02 G03 G04 H01 H02 H03
H04 H05 H30 J01 J02 J03 J04 J05 J07 J08 L01 L02 L03 L04 M01 M02 M03 M04 M05 M09 N01 N02 N03 N04 N05
N06 N07 P01 P02 P03 R01 R02 R03 R05 R06 S01 S02 S03 V02 V03 V04 V06 V07 V08 V09 X00 Y10 Y20

/CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

Estas variables son, por orden de aparición: nivel de estudios, actividad laboral, estado civil, motivo de ingreso en el centro o residencia, edad, sexo, comunidad autónoma, haber pasado el COVID y estado de vacunación para el mismo, índice de Barthel, resultado MiniMental, índice de masa corporal, número bruto de administraciones de cualquier fármaco (N_admin), indicador de calidad del dato (DataQ), hospitalizaciones (Total_hospital). El resto de los códigos son las agrupaciones de los códigos diagnósticos CIE-10 en grupos con sentido clínico, además de las agrupaciones ATC2 de los medicamentos recibidos. La gran mayoría de estas variables son dicotómicas (0/1) con excepción de las categóricas multinominales (nivel de estudios, actividad laboral, estado civil, motivo de ingreso y CCAA) y las continuas (edad, Barthel, Minimental, IMC, DataQ y N_admin). En la tabla 8 se presentan las variables que resultaron estadísticamente significativas en el modelo final general, construido con todo el conjunto de datos, hombres y

mujeres. De entre ellas nos interesa especialmente la variable sexo por la interpretación del exponente del coeficiente B (Exp(B)) como una razón de prevalencias (RP=1,327).

Tabla 8: Variables que resultan estadísticamente significativas en el modelo final (p>0,005)

	Error estándar	Sig.	Exp(B)		Error estándar	Sig.	Exp(B)
MotivoIngreso		0,000		B01	0,028	0,000	1,310
Motivo(1)	0,067	0,484	0,954	B03	0,028	0,000	1,155
Motivo (2)	0,063	0,018	1,161	B05	0,055	0,000	0,607
Motivo(3)	0,094	0,000	0,660	C01	0,047	0,000	1,287
Motivo(4)	0,080	0,023	1,198	C03	0,028	0,000	1,250
Motivo(5)	0,094	0,000	0,283	C09	0,029	0,000	0,895
Motivo(6)	0,088	0,000	0,721	C10	0,032	0,000	0,788
Edad	0,002	0,000	1,006	D08	0,136	0,002	0,661
Sexo	0,032	0,000	1,327	H02	0,034	0,000	1,221
Vacunas	0,015	0,000	0,946	J01	0,029	0,000	2,008
Covid	0,073	0,000	1,585	J03	0,094	0,000	1,563
Barthel	0,000	0,000	0,997	L01	0,083	0,000	0,724
Minimental	0,002	0,000	0,982	L02	0,057	0,007	1,165
NeoplasiasMg	0,060	0,001	1,226	M02	0,056	0,001	1,203
SN	0,035	0,000	0,873	M04	0,056	0,007	1,164
OjoOido	0,041	0,000	0,858	N03	0,033	0,000	1,184
A02	0,028	0,002	0,915	P01	0,102	0,000	1,711
A07	0,040	0,001	1,136	Num_admin	0,013	0,000	0,942
A15	0,108	0,000	1,537	DataQ	0,013	0,000	1,074

A partir de la tabla de clasificación (Tabla 9) con los valores observados y los valores pronosticados por el modelo podemos obtener los indicadores de sensibilidad, especificidad, valor predictivo positivo, valor predictivo negativo, curvas ROC y área bajo la curva y compararlos con los obtenidos con los modelos de *machine learning.* Los indicadores predictivos resultan por debajo del 75%, con excepción del área bajo la curva que es del 76,3 %. Como ya se ha comentado se trata de unos resultados predictivos pobres pero nuestro objetivo no es la capacidad predictiva en sí mima si no el poder comparar los resultados por sexos.

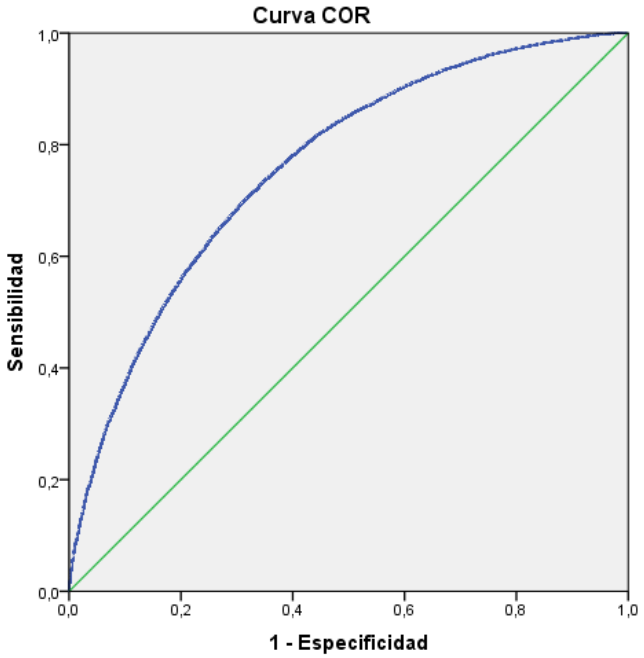
Tabla 9: Tabla de clasificación del modelo general final

		Valores pronosticados Total_hospital		Total
		0	1	
Valores observados Total_hospital	0	13111	5483	18.594
	1	5938	12655	18.593
Total		19.049	18.138	37.187

Tabla 10: Indicadores de precisión del modelo general

	Área bajo la curva (AUC)	Sensibilidad	Especificidad	VPP	VPN
Modelo general	76,3%	68,1%	70,5%	69,8%	68,8%

Gráfico 10: Curva ROC del modelo general



5.2 Análisis por sexos

Construimos primero el modelo general para poder obtener una interpretación del coeficiente de la variable sexo, que en este caso es de 1,327. Después, entrenamos este modelo de forma separada con los hombres (sexo=1) y las mujeres (sexo=0) y comparamos los resultados de los parámetros de interés (R^2 de Nagelkerke y test de Hosmer-Lemeshow).

Tablas 10 -13: Estadísticos de los modelos entrenados en hombres y en mujeres

	Significación del modelo	**R^2 Nagelkerke**	**X^2 Hosmer-Lemeshow**	**Grados de libertad**	**Significación X^2 Hosmer-Lemeshow**
Hombres	0,000	0,308	10,850	8	0,210
Mujeres	0,000	0,255	6,820	8	0,556

Tras la comparación de los estadísticos se concluye que el modelo entrenado en hombres explica un mayor porcentaje de la variabilidad (R^2) por lo que será el que utilicemos para validar tanto en hombres como en mujeres. Así, cuando validamos este modelo de hombres en cada uno de los sexos nos encontramos con diferentes capacidades predictivas, tal y como se registra en la tabla 14. A partir de estos valores podemos obtener los indicadores de sensibilidad, especificidad, valor predictivo positivo y valor predictivo negativo. También podemos construir las curvas ROC a partir de los valores pronosticados.

Tabla 14: Tabla de clasificación del modelo final de hombres validado en hombres y en mujeres

Modelo de hombres validado en HOMBRES		**Valores pronosticados Total_hospital**		**Total**
		0	**1**	
Valores observados Total_hospital	**0**	6.855	1.867	8.722
	1	2.824	4.575	7.399
Total		9.679	6.442	16.121

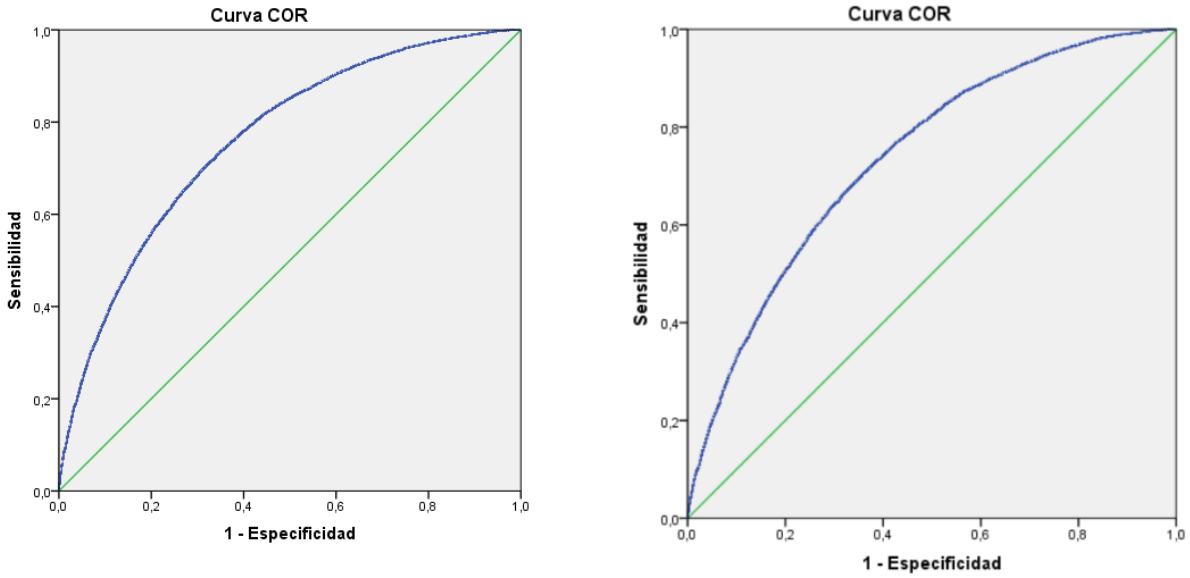
Modelo de hombres validado en MUJERES		**Valores pronosticados Total_hospital**		**Total**
		0	**1**	
Valores observados Total_hospital	**0**	6.332	3.540	9.872
	1	3.032	8.162	11.194
Total		9.364	11.702	21.066

Tabla 15: Indicadores de la predicción de las validaciones del modelo final de hombres. Se señalan en negrita los valores más elevados para cada indicador.

	Área bajo la curva (AUC)	Sensibilidad	Especificidad	VPP	VPN
Modelo de hombres validado en hombres	**77,1%**	61,8%	**78,6%**	**71%**	**70,8%**
Modelo de hombres validado en mujeres	75%	**72,9%**	64,1%	69,7%	67,6%

Se puede observar que el modelo de hombres validado en hombres obtiene los mejores valores predictivos con excepción de la sensibilidad, que es más alta en el modelo de hombres validado en mujeres. Esto indica que el modelo de hombres es capaz de detectar con mayor sensibilidad a las mujeres que van a ingresar que a los hombres.

Gráfico 11: Curva ROC del modelo de hombres testado en hombres y del modelo de hombres testado en mujeres



Anexo 6: Resultados de la IA

6.1 Entrenamiento en Mujeres y aplicado en Hombres

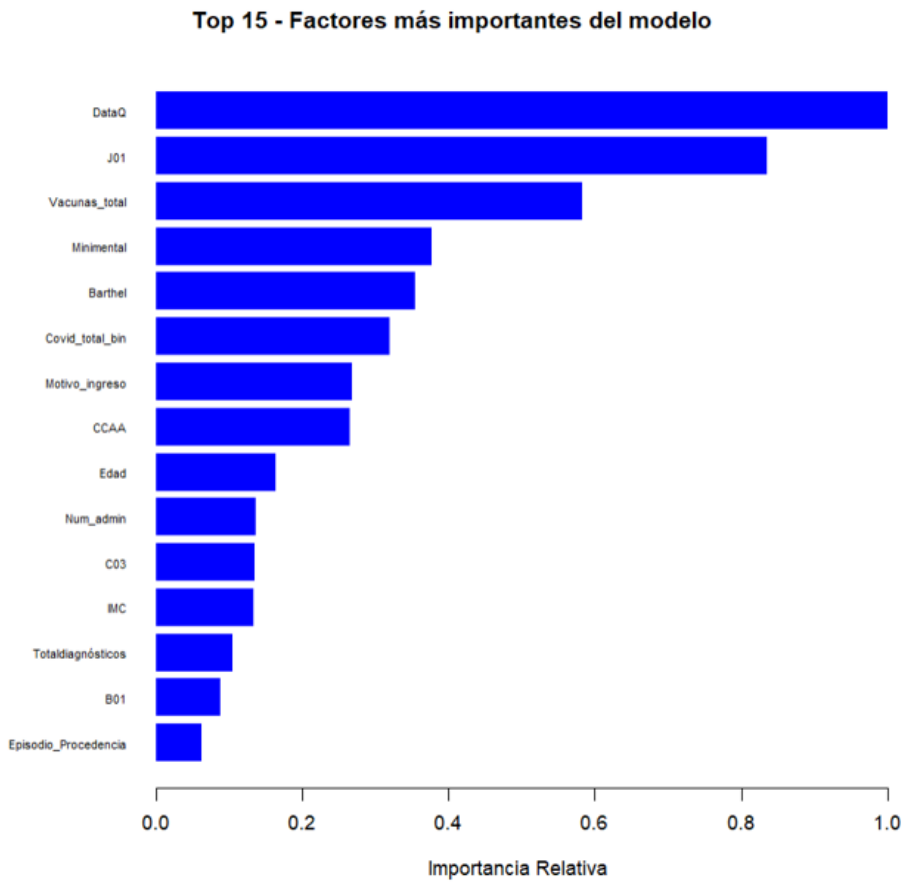
En primer lugar, entrenamos el modelo en el conjunto de las mujeres, en el *dataset* de 70%.

Probamos diferentes conjuntos de hiperparámetros y finalmente seleccionamos el siguiente:

```
xgboost(
                                data = almacen$train_mat,
                                objective = "binary:logistic",
                                nrounds = 200, max.depth = 3, eta = 0.2, nthread = 2,
                                early_stopping_rounds = 10
)
```

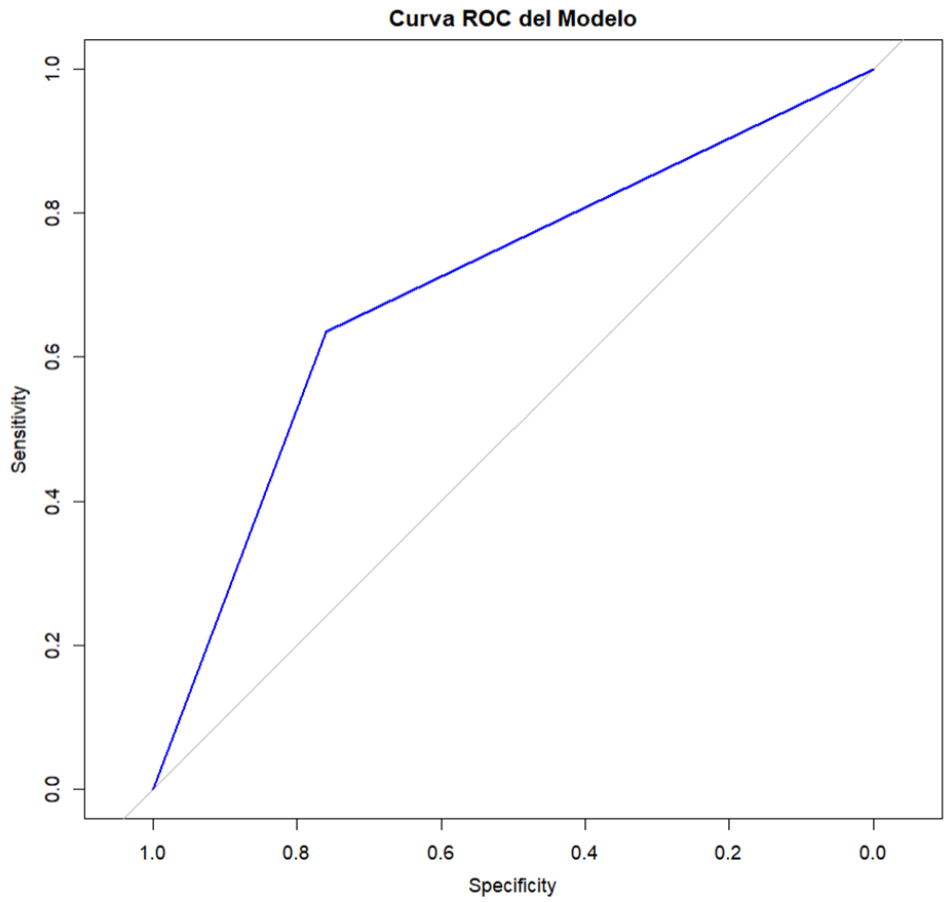
Confusion Matrix and Statistics
X2
X1 0 1
0 2687 1108
1 846 1925
Accuracy : 0.7024
95% CI : (0.6912, 0.7134)

```
    No Information Rate : 0.5381
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3977

 Mcnemar's Test P-Value : 3.539e-09

            Sensitivity : 0.7605
            Specificity : 0.6347
         Pos Pred Value : 0.7080
         Neg Pred Value : 0.6947
             Prevalence : 0.5381
         Detection Rate : 0.4092
   Detection Prevalence : 0.5780
      Balanced Accuracy : 0.6976

       'Positive' Class : 0
```



Feature	Gain	Cover	Frequency	Importance
DataQ	0,15990312	0,0281766	0,02859351	0,159903119
J01	0,13369451	0,03103093	0,0324575	0,133694509
Vacunas_total	0,09325561	0,0225389	0,02550232	0,093255613
Minimental	0,06036654	0,03953966	0,04868624	0,060366543
Barthel	0,05650007	0,02811158	0,05023184	0,056500070

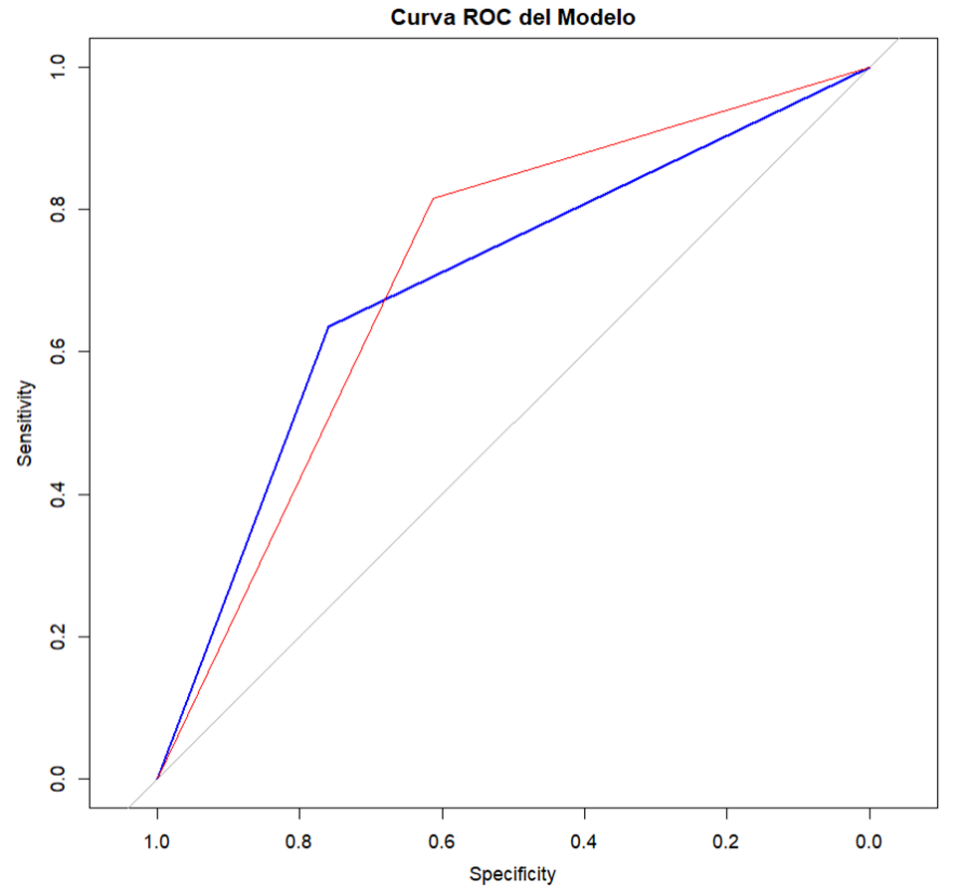
Covid_total_bin	0,05116591	0,00974261	0,01159196	0,051165905
Motivo_ingreso	0,04278753	0,01283067	0,02395672	0,042787530
CCAA	0,04231008	0,05619967	0,05641422	0,042310080
Edad	0,0259992	0,03124554	0,04327666	0,025999199
Num_admin	0,02171398	0,01050506	0,02163833	0,021713978
C03	0,02152747	0,03499988	0,02395672	0,021527466
IMC	0,02127023	0,03026402	0,03709428	0,021270230
Totaldiagnósticos	0,01676895	0,01349402	0,02163833	0,016768945
B01	0,01390851	0,01214902	0,01159196	0,013908513
Episodio_Procedencia	0,00995875	0,00874499	0,01159196	0,009958748



Confusion Matrix and Statistics
X2
X1 0 1
0 5212 1812
1 3278 7987

Accuracy : 0.7217
95% CI : (0.7151, 0.7282)
No Information Rate : 0.5358
P-Value [Acc > NIR] : < 2.2e-16
Kappa : 0.434
Mcnemar's Test P-Value : < 2.2e-16
Sensitivity : 0.6139
Specificity : 0.8151
Pos Pred Value : 0.7420
Neg Pred Value : 0.7090
Prevalence : 0.4642
Detection Rate : 0.2850
Detection Prevalence : 0.3841
Balanced Accuracy : 0.7145

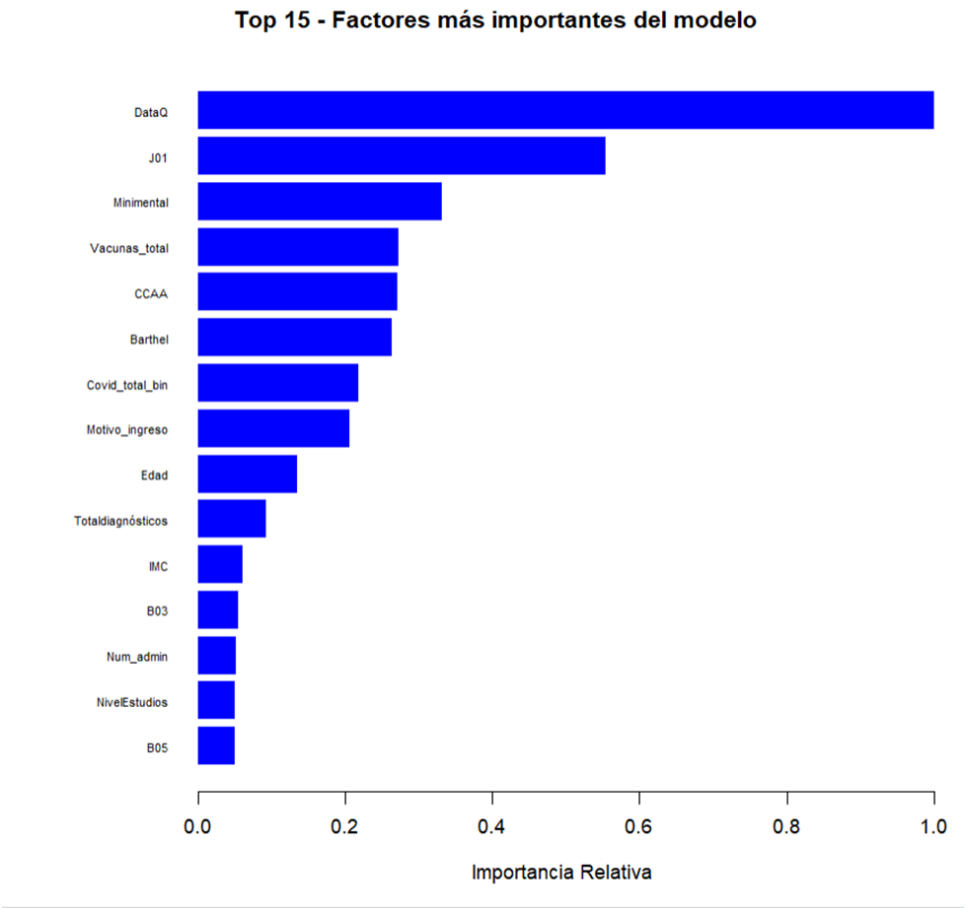
```
'Positive' Class : 0
```



6.2 Entrenamiento en Hombres y aplicado en Mujeres

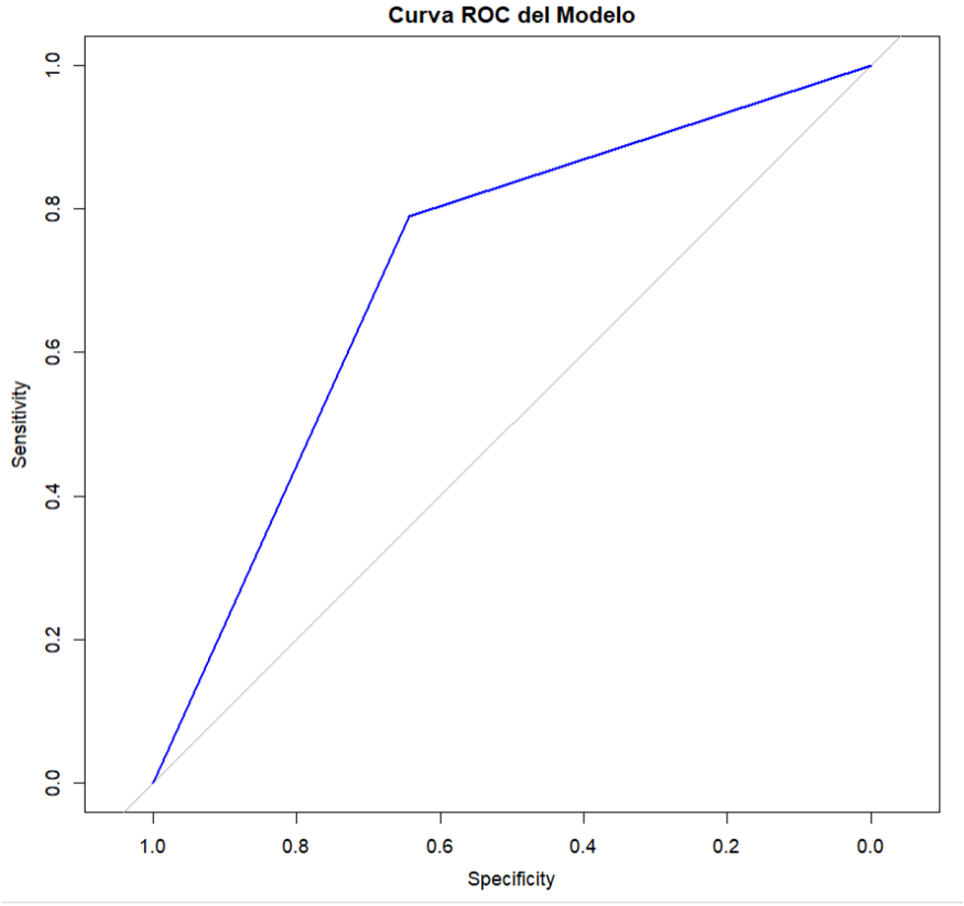
Confusion Matrix and Statistics
X2
X1 0 1
0 1661 610
1 918 2298
Accuracy : 0.7215
95% CI : (0.7095, 0.7334)
No Information Rate : 0.53
P-Value [Acc > NIR] : < 2.2e-16
Kappa : 0.4372
Mcnemar's Test P-Value : 4.038e-15
Sensitivity : 0.6440

Specificity : 0.7902
Pos Pred Value : 0.7314
Neg Pred Value : 0.7146
Prevalence : 0.4700
Detection Rate : 0.3027
Detection Prevalence : 0.4139
Balanced Accuracy : 0.7171
'Positive' Class : 0



Feature	Gain	Cover	Frequency	Importance
DataQ	0,22578922	0,03520391	0,04212168	0,22578922
J01	0,12513827	0,04384465	0,02886115	0,12513827
Minimental	0,07483074	0,04930585	0,0475819	0,07483074
Vacunas_total	0,06136165	0,03209247	0,03042122	0,06136165

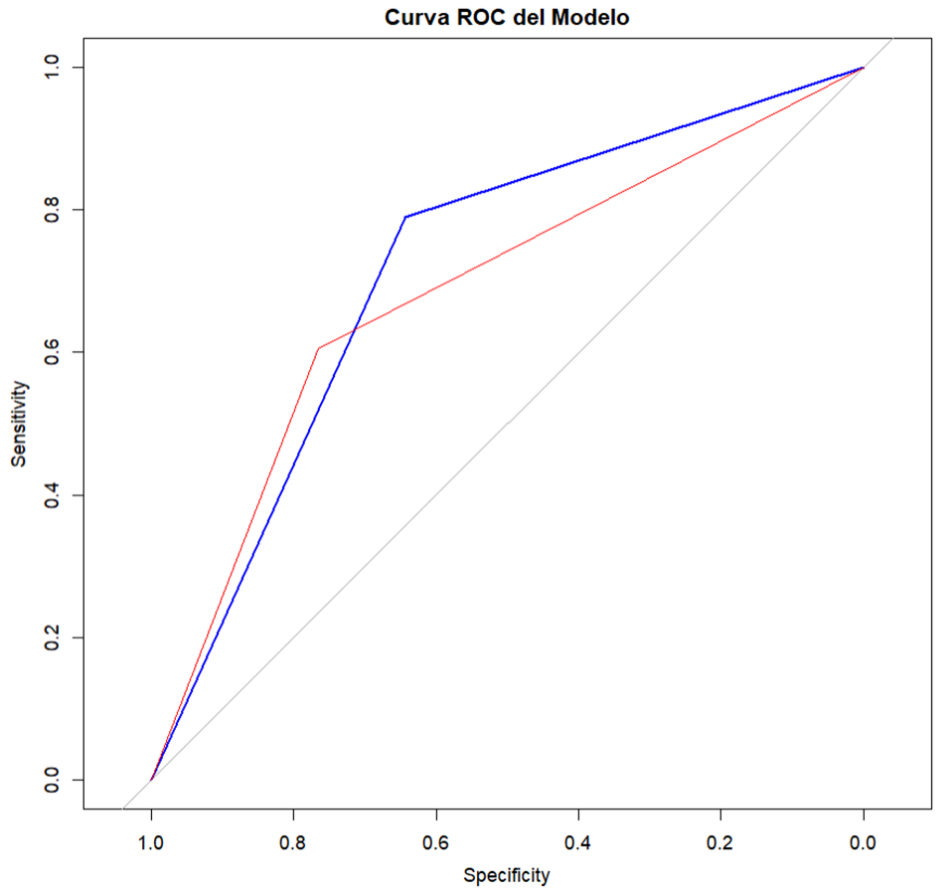
CCAA	0,06105174	0,06810131	0,06864275	0,06105174
Barthel	0,05936862	0,03699694	0,049922	0,05936862
Covid_total_bin	0,04908775	0,01460871	0,01716069	0,04908775
Motivo_ingreso	0,04629859	0,03380459	0,03900156	0,04629859
Edad	0,03041557	0,04020901	0,05226209	0,03041557
Totaldiagnósticos	0,02067005	0,01627491	0,02028081	0,02067005
IMC	0,0136138	0,02351475	0,03042122	0,01361380
B03	0,01227008	0,01231591	0,01560062	0,01227008
Num_admin	0,01147745	0,01072419	0,02574103	0,01147745
NivelEstudios	0,01127658	0,01713002	0,02574103	0,01127658
B05	0,01124463	0,0151326	0,01404056	0,01124463



Confusion Matrix and Statistics
X2
X1 0 1
0 9171 3913
1 2805 5996

```
               Accuracy : 0.693
                 95% CI : (0.6869, 0.6991)
    No Information Rate : 0.5472
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3745

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7658
            Specificity : 0.6051
         Pos Pred Value : 0.7009
         Neg Pred Value : 0.6813
             Prevalence : 0.5472
         Detection Rate : 0.4191
   Detection Prevalence : 0.5979
      Balanced Accuracy : 0.6854
```

```
'Positive' Class : 0
```



Comparativas

Entrenado en Mujeres	Entrenado en Hombres
DataQ	DataQ
J01	J01
Vacunas_total	Minimental
Minimental	Vacunas_total
Barthel	CCAA
Covid_total_bin	Barthel
Motivo_ingreso	Covid_total_bin
CCAA	Motivo_ingreso
Edad	Edad
Num_admin	Totaldiagnósticos
C03	IMC
IMC	B03
Totaldiagnósticos	Num_admin
B01	NivelEstudios
Episodio_Procedencia	B05

Modelo	Sensibilidad	Especificidad	Precisión	Valor Pred. Positivo	Valor Pred. Negativo
1 Entrenado en Mujeres	0,7605	0,6347	0,7024	0,7080	0,6947
2 Entrenado en hombres	0,6440	0,7902	0,7215	0,7314	0,7146

Sensiblidad	
Modelo Masculino	Modelo Femenino
0,644	0,7605
Aplicado a Mujeres	Aplicado a Hombres
0,7658	0,6139

Anexo 7. Domusvi: Acuerdo cesión de datos para desarrollo de proyectos de Investigación. Compromiso de Confidencialidad.

Acuerdo de cesión de datos para desarrollo de proyectos de Investigación. Compromiso de confidencialidad

En Vigo a 21 de Abril de 2023

REUNIDOS

De una parte, Dña. María Taboada Fernández, con DNI nº 76926486B, como Directora Corporativa de Excelencia de DomusVi (Geriavi, SAU) con CIF A87332292, actuando en nombre y representación de DomusVi.

De otra parte, D. Roberto González Novas, con DNI nº 53114410G, como Profesional que va a realizar el Proyecto de Investigación como parte de su TFM del *Máster en Dirección de Sistemas y TIC para la Salud y en Digitalización Sanitaria*, organizado por el Instituto de Salud Carlos III y la Sociedad Española de Informática de la Salud (SEIS) bajo el título de: ***"Estimación del sesgo de género en la hospitalización e ingreso en UCI debido al COVID en la red de asistencia de residencias de la tercera edad entre los años 2020 y 2022 mediante el uso de AI"***, actuando en su propio nombre.

Y D. Manuel José Fernández Iglesias como Director del proyecto llevado a cabo por Roberto González Novas, Susana Beatriz Nájera e Iria Rodríguez Cobo.

MANIFIESTAN

I. Que la estrategia de Investigación e Innovación de DomusVi está dirigida a la realización de proyectos científico-técnicos que generen valor añadido para los centros y servicios en las áreas de salud y bienestar, promoción de la autonomía, prevención de la dependencia, aplicación de las TIC e integración comunitaria.

II. Que un objetivo prioritario en el marco de esta estrategia es facilitar a los profesionales de DomusVi la cesión de datos asistenciales agregados (GCR) para la realización de proyectos de fin de máster o tesis doctorales. De esta forma, se materializa el compromiso de la Organización con la generación de conocimiento útil y aplicable y se colabora en la formación científica de los profesionales, potenciando su competencia investigadora.

III. Que habiendo revisado la documentación aportada por Roberto González Novas acerca del proyecto *"Estimación del sesgo de género en la hospitalización e ingreso en UCI debido al COVID en la red de asistencia de residencias de la tercera edad entre los años 2020 y 2022 mediante el uso de AI"* coordinado por D. Manuel José Fernández Iglesias Catedrático de Universidad – Área de Ingeniería Telemática *de la Universidad de Vigo, se establecen las siguientes*

ESTIPULACIONES

Primera.- Que DomusVi autoriza a D. Roberto González Novas el acceso a los datos para la realización del proyecto *"Estimación del sesgo de género en la hospitalización e ingreso en UCI debido al COVID en la red de asistencia de residencias de la tercera edad entre los años 2020 y 2022 mediante el uso de AI"*, coordinado por D. Manuel José Fernández Iglesias *de la Universidad de Vigo.*
Segunda.- Que esta autorización implica que ambas partes adquieran una serie de compromisos:

Por un lado, DomusVi se compromete a:

1. Facilitar el archivo de datos agregados con las variables solicitadas por D. Roberto González Novas para la realización de su proyecto de investigación. Se dará acceso a los datos anonimizados de todos los centros o servicios de la compañía.
2. Velar por la correcta utilización de los datos y su explotación según los objetivos marcados en el proyecto de investigación.
3. Revisar las presentaciones, material técnico, publicaciones o cualquier otro material que se derive de la realización del proyecto con el objetivo de comprobar que se reconoce adecuadamente la colaboración de DomusVi.
4. Por parte de la organización se garantiza que los datos entregados al Profesional para su uso en el Proyecto siempre serán anónimos, sin que pueda individualizarse al residente o usuario del Centro.

Por otro lado, D. Roberto González Novas (en adelante "profesional") en calidad de alumno del Máster en Dirección de Sistemas y TIC para la Salud y en Digitalización Sanitaria del Instituto de Salud Carlos III y la SEIS se compromete a:

1. Incluir una referencia expresa a la realización del trabajo en DomusVi en cualquiera de los materiales que se produzcan (documentación final del proyecto, ponencias, posters, etc.), independientemente de que el profesional siga o no en la compañía. Se debe mencionar la colaboración de la Organización en el marco de su estrategia de investigación e innovación y su apoyo a la formación científica de los profesionales.

2. Permanecer en su mismo puesto por un período no inferior a un año desde la aprobación del acceso a los datos. Esta condición podrá ser modificada por parte de la organización. El incumplimiento de este compromiso conllevará la retirada de la autorización y, por tanto, la imposibilidad de elaborar el trabajo o de presentarlo con los datos facilitados.

3. Hacer una presentación o ponencia sobre los resultados obtenidos en el foro que DomusVi considere conveniente.

Tercera.- DomusVi, de acuerdo con lo que establece la Ley orgánica 3/2018 de 5 de diciembre de Protección de Datos Personales y garantía de los derechos digitales y con el fin de darle cumplimiento, quiere recordarle que, en el supuesto de que en el desarrollo de la relación de investigación, su empresa u organización intervenga en

cualquier fase del tratamiento de datos de carácter personal existentes en nuestros ficheros, están obligados al secreto profesional y a no hacer un uso diferente de la información al de las finalidades por las cuales se establece la colaboración. Esta obligación subsistirá incluso en caso de que dicha relación se extinga

La presente Autorización y compromiso de confidencialidad se firma por triplicado y a todos los efectos en el lugar y fecha indicados.

Vigo a 21 de abril de 2023,

76926846B MARIA TABOADA FERNANDEZ (R: A87332292) Firmado digitalmente por 76926846B MARIA TABOADA FERNANDEZ (R: A87332292) Fecha: 2023.09.12 18:54:02 +02'00'

FERNANDEZ IGLESIAS MANUEL JOSE - 36073249A 2023.09.12 17:19:07 +02'00'

GONZALEZ NOVAS ROBERTO - 53114410G Firmado digitalmente por GONZALEZ NOVAS ROBERTO - 53114410G Fecha: 2023.09.12 17:00:20 +02'00'

Dña. María Taboada Fernández	D. Manuel José Fernández Iglesias	D. Roberto González Novas
Representante DomusVi	Director del proyecto	Trabajador/a DomusVi

Anexo 8. Dictamen del Comité de Ética Asistencial de Domusvi

DICTAMEN DEL COMITÉ DE ÉTICA ASISTENCIAL DE DOMUSVI

El Comité de Ética Asistencial de DomusVi, en su reunión del día 14/09/2023, con relación al Proyecto denominado "TFM" que lidera Roberto González Novas, teniendo en cuenta la documentación presentada y los aspectos que se detallan a continuación:

	Adecuada	Dudosa	Incorrecta	No procede
Justificación del estudio	x			
Definición del objeto de estudio	x			
Implicaciones éticas en el diseño, metodología y financiación	x			
Obtención del consentimiento informado y otros informes necesarios	x			
Información, adecuación de las instalaciones e instrumentos requeridos	x			
Competencia del investigador y/o de su grupo	x			
Compromiso de confidencialidad	x			

Observaciones/ Comentarios:

Se concluye el visto bueno del CEA como "favorable", pero condicionado a:

- Visto bueno del DPO del Grupo DomusVi.

Se resuelve emitir el siguiente dictamen como[1]:

Favorable X Favorable condicionado ___ Desfavorable ___

(Ver observaciones)

Fecha: 14/09/2023

Firma del presidente del CEA DomusVi:

Dr. Francesc Torralba Roselló

TORRALBA ROSELLO FRANCESC - 46231424K

Signat digitalment per TORRALBA ROSELLO FRANCESC - 46231424K
Data: 2023.09.17 09:00:41 +02'00'

[1] Cualquier modificación o incidencia que afecte al desarrollo del proyecto (finalidad, personas del equipo, etc.), se deberá notificar al CEA DomusVi para proceder a una nueva valoración del proyecto.

Todos los miembros del CEA DomusVi se comprometen a garantizar la confidencialidad de la información a la que tienen acceso en el desarrollo de sus funciones. Se garantiza así el tratamiento adecuado de la documentación recibida para la evaluación de protocolos y de la identidad de los sujetos que participan en las propuestas que se evalúan.