|  |  |
| --- | --- |
| **COVID persistente y factores neuropsicológicos.**  Diseño y validación de un instrumento de cribado breve mediante Inteligencia Artificial. | |
|  | |
|  | **ARIEL ERNESTO CARIAGA MARTÍNEZ**  Bioinformática Estadística y Aprendizaje Automático    **Nombre Tutor/a de TF**  Joan Muniesa (UOC), Meritxell Carmona Cervelló; Concepció Violán Fors; Pere Torán Monserrat. (Unitat de Suport a la Recerca Metropolitana Nord, Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), Mataró, Spain).  **Profesor/a responsable de la asignatura**  Agnès Pérez Millan.  **Fecha Entrega** |

© ARIEL ERNESTO CARIAGA MARTÍNEZ

Reservados todos los derechos. Está prohibido la reproducción total o parcial de esta obra por cualquier medio o procedimiento, comprendidos la impresión, la reprografía, el microfilme, el tratamiento informático o cualquier otro sistema, así como la distribución de ejemplares mediante alquiler y préstamo, sin la autorización escrita del autor o de los límites que autorice la Ley de Propiedad Intelectual.

**FICHA DEL TRABAJO FINAL**

|  |  |
| --- | --- |
| **Título del trabajo:** | **COVID persistente y factores neuropsicológicos.** Diseño y validación de un instrumento de cribado breve mediante Inteligencia Artificial. |
| **Nombre del autor:** | *ARIEL ERNESTO CARIAGA MARTÍNEZ* |
| **Nombre del director/a:** | Joan Muniesa (UOC), Meritxell Carmona Cervelló (1); Concepció Violán Fors (1); Pere Torán Monserrat (1). (1. Unitat de Suport a la Recerca Metropolitana Nord). |
| **Nombre del PRA:** | *Agnès Pérez Millan.* |
| **Fecha de entrega (mm/aaaa):** | *MM/AAAA* |
| **Titulación o programa:** | Máster Universitario en Bioinformática y bioestadística UOC-UB |
| **Área del Trabajo Final:** | *Bioinformática estadística y aprendizaje automático* |
| **Idioma del trabajo:** | *Español* |
| Palabras clave | *Long COVID, Neuropsychological assessment, processing speed, verbal memory, Screening Test, XGBoost, Logistic Regression, SHAP values.* |
| **Resumen del Trabajo** | |
| Este trabajo aborda el desarrollo y validación de un instrumento de cribado breve para identificar pacientes con COVID persistente y afectación neuropsicológica, utilizando técnicas de inteligencia artificial. Integrando datos sociodemográficos y resultados de pruebas neuropsicológicas validadas, sobre una población de 241 participantes, se evaluaron modelos predictivos supervisados, destacando XGBoost por su precisión y robustez. El modelo alcanzó una sensibilidad del 78,38% y una especificidad del 90,91%, superando al modelo base de regresión logística. Además, se incorporaron valores SHAP para interpretar la contribución de cada variable a las predicciones, mejorando la transparencia y aplicabilidad clínica del modelo.  Los resultados subrayan la relevancia de variables como la velocidad de procesamiento y la memoria verbal, así como variables sociodemográficas (como el sexo) en la clasificación de las personas diagnosticadas con COVID persistente con afectación cognitiva. Se destaca el potencial uso del modelo para reducir la carga en los sistemas de salud al optimizar el diagnóstico temprano y la planificación de tratamientos, pasando de la necesidad de aplicar 15 pruebas neuropsicológicas a 3. Finalmente, como líneas de interés futuro, se recalca la validación del modelo en cohortes más amplias y la integración de marcadores adicionales. Este trabajo busca contribuir especialmente a la atención personalizada de pacientes con COVID persistente, alineándose con objetivos de sostenibilidad e inclusión. | |
| **Abstract** | |
| This work addresses the development and validation of a brief screening test to identify patients with long COVID and neuropsychological impairment using artificial intelligence techniques. Integrating sociodemographic data and validated neuropsychological test results from a population of 241 participants, supervised predictive models were evaluated, with XGBoost standing out for its accuracy and robustness. The model achieved a sensitivity of 78.38% and a specificity of 90.91%, outperforming the baseline logistic regression model. Additionally, SHAP values were incorporated to interpret the contribution of each variable to predictions, enhancing the model's transparency and clinical applicability.  The results highlight the relevance of variables such as processing speed and verbal memory, as well as sociodemographic variables (e.g., sex) in classifying individuals diagnosed with long COVID and cognitive impairment. The model’s potential to alleviate healthcare system burdens is emphasized, optimizing early diagnosis and treatment planning by reducing the need to administer 15 neuropsychological tests to just 3. Finally, as future research directions, the validation of the model in larger cohorts and the integration of additional markers are underscored. This work significantly contributes to the personalized care of long COVID patients, aligning with sustainability and inclusion objectives. | |

Índice

[1. Introducción 7](#_Toc186628036)

[1.1. Contexto y justificación del Trabajo 7](#_Toc186628037)

[1.2. Objetivos 8](#_Toc186628038)

[1.3. Impacto en sostenibilidad, ético-social y de diversidad. 9](#_Toc186628039)

[1.4. Enfoque y método seguido 11](#_Toc186628040)

[1.5. Planificación del Trabajo 12](#_Toc186628041)

[1.6. Breve sumario de productos obtenidos. 17](#_Toc186628042)

[1.7. Breve descripción de los otros capítulos de la memoria 17](#_Toc186628043)

[2. Estado del arte 18](#_Toc186628044)

[3. Materiales y métodos 22](#_Toc186628045)

[4. Resultados 29](#_Toc186628046)

[5. Conclusiones y trabajos futuros 43](#_Toc186628047)

[6. Glosario 47](#_Toc186628048)

[7. Bibliografía 48](#_Toc186628049)

[8. Anexos 52](#_Toc186628050)

Lista de tablas y figuras

[Figura 1: Imputación de datos faltantes. a) Proporción de datos faltantes por variable antes de la aplicación del algoritmo kNN. b) Proporción de datos faltantes por variable tras la imputación. 29](#_Toc186627600)

[Figura 2: Análisis exploratorio de los datos. Distribución univariada de variables cuantitativas (incluyendo variables binarizadas). En el eje X se representan los valores cuantitativos en tanto que en el eje Y se representan las frecuencias correspondientes. Se observa una distribución aproximadamente normal en variables de interés para el desarrollo de análisis subsiguientes. 30](#_Toc186627601)

[Figura 3: Análisis exploratorio de los datos. Distribución univariada de variables cualitativas. En el eje X se indica la ausencia (cero) o presencia (uno) de la categoría y en el eje Y se indica la frecuencia de observaciones. 31](#_Toc186627602)

[Figura 4: Análisis exploratorio de los datos. Mapa de calor de correlación. La matriz muestra los coeficientes de correlación de Pearson entre variables seleccionades. Los colores que tienden al azul corresponden a coeficientes de correlación positiva en tanto que los colores que tienden al rojo corresponden a coeficientes de correlación negativa. 32](#_Toc186627603)

[Figura 5: Análisis exploratorio de los datos. Análisis bivariado, agrupado por diagnóstico. Las cajas representan el rango intercuartílico (entre el primer y tercer cuartil), con la mediana indicada por una línea dentro de cada caja. Los bigotes se extienden hasta 1.5 veces el rango intercuartílico, y los puntos negros fuera de los bigotes indican posibles valores extremos. 33](#_Toc186627604)

[Figura 6: Análisis exploratorio de los datos. Análisis de Componentes Principales, agrupado por diagnóstico. Representación de las proyecciones de las observaciones sobre los dos primeros componentes principales, agrupadas por diagnóstico. Los ejes muestran el porcentaje de varianza explicada por cada componente. Las observaciones están coloreadas y simbolizadas según su grupo de diagnóstico (0/Círculo = Sin infección por COVID/PCC No Cog; 1/Triángulo = PCC Cog). 34](#_Toc186627605)

[Figura 7: Análisis exploratorio de los datos. *Scree Plot*. El eje Y muestra la proporción de varianza explicada por cada componente principal, mientras que el eje X indica el número de componentes. La selección del número óptimo de componentes puede basarse en el "codo" del gráfico, donde la varianza explicada acumulada deja de aumentar significativamente. En este caso, el "codo" podría hallarse entre el tercer y cuarto componente, pero la varianza acumulada es muy baja ya que subsiguientes componentes explican una proporción baja de la varianza. 34](#_Toc186627606)

[Figura 8: Curva ROC (Modelo logístico). Se generó un modelo logístico para utilizarse como modelo de base sobre el que comparar modelos más complejos. Esta curva representa la la especificidad en el eje X y la sensibilidad en el eje Y. Un modelo con sensibilidad y especificidad máxima generaría una gráfica en forma de L invertida, en tanto que un modelo que generara predicciones al azar generaría una gráfica que se acercaría a una línea diagonal. A medida que los valores de sensibilidad y especificidad obtenidos por el modelo mejoran, el gráfico se acerca a la forma ideal de L invertida. 36](#_Toc186627607)

[Figura 9: Curva ROC (Modelo XGBoost). Se generó un modelo de tipo XGBoost, como ejemplo de modelo complejo, para comparar los resultados obtenidos frente al modelo logístico. La curva ROC representa la sensibilidad en el eje Y y la especificidad en el eje X. Un modelo con sensibilidad y especificidad máxima generaría una gráfica en forma de L invertida, en tanto que un modelo que generara predicciones al azar generaría una gráfica que se acercaría a una línea diagonal. A medida que los valores de sensibilidad y especificidad obtenidos por el modelo mejoran, el gráfico se acerca a la forma ideal de L invertida. 38](#_Toc186627608)

[Figura 10: Comparación de curvas ROC para el modelo de base (logístico, en rojo) en comparación con el modelo complejo (XGBoost, en azul). 39](#_Toc186627609)

[Figura 11: Importancia de las variables (modelo XGBoost). La importancia de las variables se calcula con base en la métrica *gain*, representando la contribución relativa de cada variable a la construcción del modelo. Las barras indican la magnitud de esta importancia, con las variables más influyentes listadas en la parte superior. Este análisis permite identificar las características clave que impactan en las predicciones, facilitando un primer acceso a la interpretación del modelo y a su conexión con el contexto clínico. 40](#_Toc186627610)

[Figura 12: Ejemplo de cálculo de valores SHAP para una observación seleccionada del conjunto de prueba. El gráfico muestra la contribución individual de cada variable a la predicción realizada por el modelo. Las barras representan los valores SHAP, donde un valor positivo indica un aumento en la probabilidad predicha, y un valor negativo indica una disminución. Este análisis permite interpretar cómo las variables específicas (tales como tn12, tn30 y el sexo, en este caso de ejemplo), influyen en la predicción final, promoviendo la transparencia del modelo en aplicaciones clínicas. 41](#_Toc186627611)

[Figura 13: Valores SHAP medios absolutos para el conjunto de prueba. El gráfico presenta la importancia promedio de cada variable en las predicciones realizadas por el modelo, medida a través de los valores SHAP medios absolutos. Las variables con mayores valores, como tn12, tn30, tn6 y el sexo (mujer), destacan por su impacto consistente en la toma de decisiones del modelo, lo que subraya su relevancia en el análisis global. 42](#_Toc186627612)

[Tabla 1: Diagrama de Gantt tentativo. Se desglosan las tareas y fechas propuestas, así como los hitos esperados. En rombos rojos, los hitos correspondientes a PECs y como barra horizontal roja las fechas propuestas de defensa de TFM. 16](#_Toc186627995)

[Tabla 2: Evaluaciones estandarizadas de dominios cognitivos. Una breve descripción de cada prueba puede encontrarse en Anexos del trabajo. 20](#_Toc186627996)

[Tabla 3: Variables sociodemográficas de la base de datos analizada. 23](#_Toc186627997)

[Tabla 4: Variables seleccionadas en la construcción de la base de datos definitiva utilizada en los análisis subsiguientes. La columna de codificación recoge el nombre codificado para cada variable en los análisis posteriores. 25](#_Toc186627998)

[Tabla 5: Paquetes de R, con sus versiones y responsables, utilizados en el desarrollo de este trabajo. 28](#_Toc186627999)

[Tabla 6: Comparativa de métricas entre los modelos desarrollados. Se compara el modelo XGBoost contra el modelo base (regresión logística). Valores positivos indican mejora de la métrica por parte del modelo XGBoost comparado con el logístico. Valores negativos indican empeoramiento de la métrica en la misma comparación. 38](#_Toc186628000)

# Introducción

## Contexto y justificación del Trabajo

Desde la primavera de 2020, tras las olas de contagio provocadas por la pandemia de COVID-19 y con la aparición de nuevas variantes [1], se observó un fenómeno inusual relacionado con las infecciones víricas: personas que habían superado la enfermedad inicial comenzaron a presentar un amplio rango de síntomas persistentes durante semanas posteriores, indicando que la COVID-19 podría no ser una infección autocontenida. Entre los múltiples síntomas indicados se incluyó la fatiga, debilidad, dificultad para concentrarse, dolores de cabeza [2] y más de 50 síntomas adicionales [3]. Para septiembre de 2020, la Organización Mundial de la Salud (OMS) actualizó sus códigos de clasificación internacional de enfermedades para incluir una nueva condición post-COVID [4].

La COVID persistente (en inglés también denominada *Long COVID o post-COVID condition*, en adelante, en este texto, PCC) se define por la OMS como aquella condición que aparece en personas con un historial de infección probable o confirmada por el virus SARS-CoV-2, generalmente tres meses después del inicio, con síntomas que duran al menos dos meses y que no pueden ser explicados por un diagnóstico alternativo [5]. Los síntomas más comunes incluyen fatiga, dificultad para respirar y disfunción cognitiva, afectando la capacidad de pensamiento y memoria entre otras funciones. Además, también pueden presentarse otros síntomas que impactan en la vida diaria, como trastornos relacionados con ansiedad o depresión, entre otros [6]. La PCC afecta aproximadamente al 10-20 % de las personas que han padecido COVID-19, independientemente de la gravedad inicial de la infección [5], ya que puede afectar tanto a pacientes que tuvieron una forma leve o incluso asintomática, como a las personas que presentaron los casos más graves. Además, afecta a todas las franjas etarias, siendo más frecuente en mujeres de mediana edad, impactando significativamente en la calidad de vida familiar, social y laboral [7].

Para reflejar la forma en que la COVID afecta el sistema nervioso, se ha acuñado el término “NeuroCOVID” el cual engloba un conjunto de síntomas neurológicos, neuropsicológicos y neuropsiquiátricos persistentes durante más de 2 meses no justificados por otra causa neuropsicológica [8]. Estos síntomas son particularmente incapacitantes y se ha observado que la disfunción cognitiva post-COVID, a menudo descrita como "niebla mental", incluye déficits en la memoria, la atención, las funciones ejecutivas y la fluidez verbal. Si bien algunos estudios han vinculado estas quejas subjetivas con déficits cognitivos objetivos [9], otros no han encontrado esta asociación [10], lo que resalta la necesidad de investigaciones adicionales [3]. Un aspecto crucial es que la mayoría de los estudios se han centrado en pacientes que tuvieron un ingreso hospitalario, dejando una brecha de conocimiento importante sobre quienes padecieron formas leves de la infección [11].

Por otra parte, también se han descrito otras complicaciones neurológicas y neuropsiquiátricas de la COVID-19, como dolores de cabeza, síndrome de Guillain-Barré, temblores y ataxia, y se ha propuesto que estos síntomas pueden estar relacionados con la invasión directa del virus al sistema nervioso o con la respuesta inflamatoria sistémica exacerbada [12]. Esta posible invasión podría generar la presencia de biomarcadores derivados tanto de cambios funcionales y estructurales en el cerebro y de estructuras relacionadas (como la retina), así como otros indicadores indirectos (como aquellos provenientes de la respuesta inflamatoria o de alteraciones en el sistema inmunológico) [13].

En el desarrollo de este trabajo de fin de máster (TFM), buscamos generar un instrumento de cribado, que incorpore tanto variables sociodemográficas como neuropsicológicas, y que permita una clasificación rápida de pacientes con PCC con o sin déficits cognitivos. Además, buscamos comprender el peso específico de cada una de las variables para facilitar su uso en el diagnóstico temprano o en el seguimiento de pacientes que puedan sufrir esta presentación de la enfermedad. Finalmente, conocer la aportación de cada aspecto neuropsicológico mediante el modelado de las variables, puede permitir su uso para optimizar la atención, el seguimiento y la implementación de estrategias específicas de rehabilitación cognitiva previniendo el deterioro cognitivo en pacientes con PCC, mejorando su bienestar y calidad de vida.

## Objetivos

*a. Objetivo general.*

- Desarrollar un instrumento de cribado breve que, teniendo como base pruebas neuropsicológicas validadas y variables sociodemográficas específicas, permita identificar de manera eficaz a personas con COVID persistente con afectación neurocognitiva de otras formas de COVID persistente.

*b. Objetivos específicos.*

- Realizar una evaluación exploratoria de los datos para caracterizar y definir la calidad de los datos recogidos.

- Evaluar las diferencias en variables neuropsicológicas y sociodemográficas entre personas con COVID persistente con afectación cognitiva y otras poblaciones que actúen como controles.

- Evaluar la utilidad de modelos de aprendizaje automático no supervisado, en particular, modelos de reducción de dimensionalidad.

- Desarrollar modelos de aprendizaje automático supervisados de clasificación, teniendo en cuenta variables sociodemográficas y neuropsicológicas.

- Validar los modelos (mediante validación cruzada) para analizar su capacidad de generalización.

- Seleccionar los mejores modelos considerando las mejoras métricas de cada uno (con un interés en métricas clínicas como sensibilidad, especificidad y área bajo la curva -en adelante, AUC-), así como la importancia de las variables específicas, en particular, las pruebas neuropsicológicos que más aporten a la mejora de las métricas definidas.

- Dotar de explicabilidad a las predicciones generadas por el modelo seleccionado, de cara a su uso como apoyo a la decisión clínica.

## Impacto en sostenibilidad, ético-social y de diversidad.

Este TFM pretende impactar en aspectos fundamentales ético-sociales y relacionados con la sostenibilidad y la diversidad en personas afectadas por PCC. Para ello, hemos diseñado soluciones que mejoren, en la medida de nuestras posibilidades, estos factores. Estos aspectos son cruciales no solo para mejorar la calidad de vida de las personas con la enfermedad bajo estudio, sino también para alinearse con los Objetivos de Desarrollo Sostenible (ODS), que buscan promover un desarrollo inclusivo, equitativo y respetuoso con el medio ambiente y la sociedad. En este sentido, este trabajo impacta y se alinea con los ODS número 3 (Salud y Bienestar) buscando un diagnóstico más rápido y eficiente de la COVID persistente con afectación cognitiva; 5 y 10 (Igualdad de género y Reducción de las desigualdades) al generar modelos respetuosos con la igualdad y permitiendo la mejora constante y escalabilidad de los mismos; 9 (Industria, Innovación e Infraestructura) trabajando hacia modelos accesibles, innovadores y de alta tecnología y 8 (Trabajo decente y crecimiento económico) al permitir una clasificación adecuada, sencilla y aplicable al entorno clínico que permita el acceso al bienestar y al crecimiento económico de las personas que sufren la afectación neurocognitiva de la PCC. De forma más específica, recogemos el impacto previsto en los siguientes apartados.

*Impacto en la sostenibilidad.*

La sostenibilidad se encuentra imbricada en el proyecto desde su diseño. El objetivo fundamental del proyecto es simplificar la evaluación neuropsicológica utilizada en un proyecto de investigación para poder aplicarla y trasladarla a la práctica clínica habitual. El análisis de datos generado en este trabajo al respecto de la PCC se alinea con los objetivos globales de salud pública para un sistema de atención sanitaria sostenible (ODS 3). Al generar evidencia que contribuye a la comprensión de la COVID persistente, los hallazgos de este estudio permitirán desarrollar estrategias para la asignación de recursos, intervenciones tempranas y planificación de cuidados a largo plazo, incrementando la incorporación de personas afectadas por PCC a la sociedad, alineándose también así con los objetivos 8 (Trabajo decente y crecimiento económico) y 9 (Industria, Innovación e Infraestructura). Estas estrategias son cruciales para minimizar los costos económicos y ambientales asociados con las condiciones de salud crónicas. Reducir la carga a largo plazo de la PCC en los sistemas de salud promoverá un uso más eficiente de los recursos sanitarios, disminuyendo las hospitalizaciones y las visitas a los centros sanitarios lo que ayudará a reducir la huella de carbono de los servicios de salud, contribuyendo a los objetivos más amplios de sostenibilidad en la atención sanitaria. Por otra parte, durante el desarrollo de este trabajo, se priorizaron los modelos más eficientes en cuanto al cálculo y a la explicabilidad para que, manteniendo la transparencia requerida para la posible aplicación clínica de la inteligencia artificial (IA), se mantuviera un uso responsable de recursos con objeto de disminuir la huella de carbono debida al coste energético del propio modelo.

Finalmente, desde el punto de vista de la sostenibilidad, consideramos que el instrumento desarrollado, al disminuir la necesidad de realizar múltiples pruebas neuropsicológicas, aliviará el coste humano y social para el sistema de salud, reduciendo indirectamente la huella ecológica asociada a traslados y consumibles. Este impacto es propio para la herramienta desarrollada por lo que la escalabilidad es un punto de optimización clave para su posible adopción clínica. También, el hecho de que pueda transformarse en una herramienta con bajo coste económico y aplicable en herramientas informáticas de baja demanda tecnológica (equipos de escritorio convencionales), la hará adaptable a diversos contextos por lo que se puede promover su implementación en regiones con recursos sanitarios limitados.

*Impacto en la ética social.*

Desde su concepción, esta investigación prioriza los principios éticos de beneficencia, no maleficencia y justicia al abordar la urgente necesidad de obtener conocimientos basados en datos sobre la PCC, una condición que afecta desproporcionadamente a poblaciones vulnerables, incluidas aquellas con condiciones de salud preexistentes y grupos socioeconómicamente desfavorecidos. Este aspecto se alinea específicamente con el objetivo 10 (Reducción de las desigualdades). A lo largo de todo el trabajo se mantendrá estrictamente el consentimiento informado, la privacidad de los datos y la confidencialidad de quienes accedieron a participar, de acuerdo con los estándares éticos internacionales. Por otra parte, el proyecto en el que se engloba este TFM cuenta con la aprobación del comité de ética del IDIAPJGol (Ref: IDIAPJGol, 21/220-P) y buscó una representación amplia de las comunidades afectadas por la PCC, al incluir pacientes de 70 centros de salud y un hospital universitario: de esta manera se garantizó el acceso equitativo y la representatividad. El estudio también busca abordar las inequidades en salud generando conocimientos que beneficiarán a dichas poblaciones, prestando especial atención a aquellas en mayor riesgo de sufrir síntomas prolongados post-COVID. La metodología aplicada en los análisis presentados (y, de forma general, a todo el trabajo presentado) garantiza la transparencia, al agregar el cálculo de índices explicativos y hacerlos inherentes e indivisibles de la solución presentada, lo que permite reforzar la confianza de pacientes y profesionales para comprender las decisiones adoptadas. Indirectamente, estas decisiones aplicadas en el trabajo impactan sobre el respeto por la autonomía de quienes participaron (y puedan hacer uso futuro del modelo), ya que el modelo no pretende sustituir la evaluación clínica, sino que pretende ampliarla, acercarla a todas las poblaciones, con un desarrollo científico sólido y que permita decisiones clínico/médicas basadas en evidencia y datos explicables, reduciendo así el sesgo. La solución presentada por este trabajo busca contribuir al bien común mejorando el acceso a diagnósticos, pronósticos o seguimientos personalizados, mientras se abren nuevas puertas de investigación, siempre de cara a la mejora colectiva. Finalmente, la opción de escalar, modificar y, de forma general, reutilizar el modelo en diferentes zonas geográficas, o considerando poblaciones o aspectos raciales específicos, ayuda a reducir desigualdades sociales contribuyendo a una mayor equidad en la atención médica.

*Impacto en diversidad e inclusión.*

Este trabajo está comprometido con los principios de diversidad e inclusión, asegurando que el conjunto de datos analizado refleje un amplio espectro de identidades raciales, étnicas, socioeconómicas y de género. La PCC afecta desproporcionadamente a diferentes comunidades, y esta investigación tiene como objetivo descubrir las formas matizadas en las que se manifiestan estas disparidades. Este trabajo no pretende quedarse en la esfera técnica, sino que pretende responder a desafíos clave: la PCC tiende a afectar desproporcionadamente a mujeres y a poblaciones condiciones preexistentes por lo que este punto de análisis busca alinearse con el cumplimiento real y efectivo de los ODS 5 y 10 (Igualdad de género y Reducción de desigualdades).

Al desagregar los datos según factores demográficos clave, el estudio resaltará las diferencias en el acceso a la atención, los resultados del tratamiento y los impactos sociales y económicos a largo plazo de la PCC. Los conocimientos obtenidos apoyarán el desarrollo de políticas de salud pública más inclusivas que aborden las necesidades específicas de los grupos marginados, promoviendo la equidad en la atención sanitaria. Este trabajo pone de relieve, de manera objetiva, la afectación desproporcionada que experimentan las mujeres en relación con la COVID persistente (PCC), facilitando el diseño de intervenciones específicas y adaptadas a este grupo. Aunque investigaciones previas ya han señalado esta desigualdad basada en el género, nuestro modelo no solo confirma esta tendencia, sino que la visibiliza de forma inherente a su estructura, al reflejar de manera clara cómo el género influye en los resultados. En este sentido, el modelo permite identificar y analizar la importancia de variables sociodemográficas y de género como elementos clave para comprender la PCC y desarrollar estrategias que aborden esta situación de vulnerabilidad.

En definitiva, consideramos que el modelo no solo mejora aspectos relacionados con las sostenibilidad y el acceso equitativo a los recursos sanitarios, sino que es una herramienta inclusiva y que, de ser adoptado (y adaptado), permitirá dar respuesta a necesidades específicas de las poblaciones más afectadas por la PCC, siempre basado en la ética, la transparencia y el respeto a la autonomía de las personas.

## Enfoque y método seguido

La PCC representa un desafío en salud pública debido a su impacto sobre la calidad de vida de las personas afectadas y, en especial, la de aquellas aquejadas por alteraciones neuropsicológicas (como problemas de memoria, atención y funciones ejecutivas) ya que dificultan especialmente la integración laboral y social [14]. Por ello, es clave el desarrollo de instrumentos de cribado rápido que permitan identificar de manera eficiente y precisa a quienes necesitan una intervención más exhaustiva. En este trabajo, el enfoque metodológico se basó en una estrategia cuantitativa, con la aplicación de pruebas neuropsicológicas específicas (que evalúan funciones ejecutivas, memoria, atención y velocidad de procesamiento, entre otros dominios) tanto en pacientes que no sufrieron la infección, como en las personas que presentaron PCC con y sin afectación cognitiva. Una vez desarrolladas y valoradas las mediciones, la estrategia consistió en evaluar la calidad del dato recogido para proponer los modelos de aprendizaje automático factibles. Entre los modelos barajados se incluyó la regresión logística, como una opción para una primera clasificación, ya que este modelo es reconocido por su simplicidad y facilidad de interpretación [15]. Por otro lado, los modelos de aprendizaje automático más complejos como los derivados de árboles (entre los que se hallan los modelos de tipo *Decision Tree, Random Forest* o *XGBoost*) son particularmente útiles en contextos donde podrían existir relaciones no lineales o interacciones complejas, características comunes en los datos clínicos [16]. Aunque la explicabilidad de los modelos basados en árboles es menor, la mejora en las métricas de aplicación requeridas (como sensibilidad o especificidad) es sustancialmente superior, a la vez que están sujetos a menores restricciones teóricas y/o estadísticas. Por otra parte, y para mejorar el déficit explicativo de los modelos basados en árboles, se hizo uso de los valores denominados *SHapley Additive exPlanations* (en adelante, SHAP) que brindan una forma de cuantificar el peso de cada variable sobre el resultado predicho por el modelo [17].

Se consideraron dos estrategias principales: adaptar productos existentes o desarrollar un nuevo instrumento. Adaptar productos ya validados, como pruebas neuropsicológicas estandarizadas, permitiría aprovechar su base empírica y reducir tiempos de implementación. Sin embargo, dado que no están diseñadas para este reto clínico, estas herramientas pueden no capturar las especificidades de la PCC con afectación cognitiva, limitando su sensibilidad y especificidad. También, la aplicabilidad puede verse reducida por su falta de validez psicométrica o su coste (licencias, etc.), entre otros factores. En contraste, desarrollar un instrumento nuevo, aunque requiere mayor tiempo y recursos, permite un diseño más personalizado y alineado con las necesidades de esta población. En este caso, se cuenta con datos de interés cuya utilidad puede ser elevada para la población clínica de interés. Por ello, tras analizar ambas alternativas, se optó por desarrollar un nuevo producto que combine elementos de pruebas neuropsicológicas tradicionales con un enfoque innovador de análisis basado en aprendizaje automático. Esta estrategia asegura que el instrumento sea altamente sensible a las particularidades de la COVID persistente con afectación cognitiva, permitiendo una identificación más precisa de los casos. Además, al seleccionar el uso de técnicas avanzadas se maximiza la capacidad predictiva del cribado y su aplicabilidad en escenarios clínicos diversos. Esto no solo promueve la detección temprana, la inclusión y la sostenibilidad, sino que también tiene el potencial de mejorar significativamente la asignación de recursos y la planificación de tratamientos.

## Planificación del Trabajo

*Recursos.*

El principal recurso estuvo constituido por los datos adquiridos provenientes de la población clínica de interés que corresponde a pacientes con un seguimiento desde el inicio de la pandemia (en marzo de 2020) por parte de los servicios de sanidad pública de Cataluña (70 centros de atención primaria del área metropolitana norte de Barcelona y del Hospital Universitario Germans Trias I Pujol). El uso de estos datos fue aprobado por el comité ético del Instituto Universitario de Investigación en Atención Primaria (IDIAPJGol, 21/220-P), en tanto que los protocolos seguidos para la inclusión de participantes y la recolección de datos se publicaron tras revisión por pares [11].

*Tareas.*

Las diversas tareas se organizaron en forma de trabajo semanal y se asignaron tiempos estimados tal como recogemos en los siguientes ítems:

- Recopilación de datos: consistió en la recolección de información sociodemográfica y resultados de pruebas neuropsicológicas de la muestra de pacientes, siempre bajo la adhesión a los planteamientos indicados en el punto correspondiente al impacto en sostenibilidad, ético-social y de sostenibilidad y diversidad y según las aprobaciones del comité de ética interviniente. El tiempo empleado fue de 4 a 6 semanas.

- Control de calidad del dato: consistió en el análisis de los parámetros requeridos para el posterior modelado. Entre otras acciones, se revisaron posibles inconsistencias, datos faltantes y desviaciones que puedan afectar la calidad del análisis y del modelo de cribado. El tiempo empleado fue de 4 a 5 semanas.

- Análisis exploratorio de los datos con realización de los análisis descriptivos uni, bi y multivariantes. De forma más específica, se identificaron patrones y relaciones preliminares (tanto de forma analítica como gráfica) entre las variables antes de proceder con análisis más complejos. El tiempo empleado en la realización de estas tareas fue de 3 a 4 semanas.

- Análisis inferencial con comparaciones entre grupos entre otros análisis inferenciales. Este análisis constituyó un punto de referencia para identificar diferencias significativas que pueden ser relevantes para la comprensión de las variables, sus relaciones o el modelado posterior. El tiempo empleado fue de 1 semana.

- Modelado y validación, implementando un primer modelado piloto y un posterior análisis de regresión logística como primeras aproximaciones a la clasificación. Se evaluaron, además, modelos alternativos para asegurar una clasificación precisa y robusta a través de diversas agrupaciones de variables de interés clínico y aplicado para obtener modelos generalizables y que, eventualmente, puedan aplicarse a un entorno de la realidad de la atención primaria u hospitalaria. El tiempo empleado fue de 4 a 5 semanas.

- Evaluación de métricas y gráficas (análisis de la importancia de variables y obtención de valores SHAP): tras la selección de las variables de interés obtenidas en las tareas anteriores, se realizó el cálculo de métricas de interés clínico como sensibilidad, especificidad y AUC para complementar la validación del instrumento. También se analizó el impacto de cada variable seleccionada en los modelos escogidos para mejorar la interpretación y ampliar su posible utilidad en entornos realistas. El tiempo empleado fue de 2 a 3 semanas.

- Elaboración de la memoria y resultados: la redacción de los hallazgos y conclusiones del trabajo fue una tarea transversal que se desarrolló durante todo el proyecto (10-12 semanas).

*Calendario.*

El desglose del calendario se indica en el diagrama de Gantt (presente en la tabla 1). El calendario mensual incluyó:

- Mes 1: recopilación y control de calidad de datos. Análisis exploratorio y primeras pruebas con la muestra de datos. Se evaluaron posibles inconsistencias en los datos y se realizaron ajustes preliminares para asegurar su calidad. En este mes se realizó la entrega de la prueba de evaluación continua número 1 (PEC1).

- Mes 2: modelado de datos (regresión logística, discriminante lineal, XGBoost y otros). Se realizó un modelado piloto para obtener una primera aproximación a las variables de interés que pudieran aportar valor clínico y analítico al cribado y se probaron diferentes enfoques para seleccionar el modelo más adecuado según el desempeño en la predicción. En este mes se realizó la entrega de la PEC2.

- Mes 3: evaluación de métricas y validación del modelo. Se profundizó en el análisis de relaciones subyacentes, inferenciales y en el cálculo de métricas de interés clínico (como sensibilidad, especificidad, AUC y curvas *Receiver Operating Characteristic -*en adelante, curvas ROC*-)* para validar la robustez del modelo, así como la comparativa entre posible modelos diferenciales generados. En este mes se realizó la entrega de la PEC3.

- Mes 4: se llevó adelante la redacción de la memoria final contando con los hitos anteriores como base de escritura. En este mes también se realizará la entrega de la PEC4 y posterior defensa (PEC5).

*Hitos.*

El trabajo se estructuró en hitos de interés que están recogidos en el diagrama de Gantt (ver tabla 1). De forma resumida, los hitos establecidos fueron:

- Hito 1: finalización de la recopilación de datos y primera presentación del análisis exploratorio de los datos (final del primer mes).

- Hito 2: resultados preliminares del modelado piloto y de análisis estadísticos descriptivos e inferenciales básicos, con selección de variables clínicas, neuropsicológicas y sociodemográficas de interés (mitad del segundo mes).

- Hito 3: selección inicial del modelo con mejor rendimiento. Validación final de instrumento (final del tercer mes).

- Hito 4: redacción del documento y defensa pública (final del cuarto mes – enero 2025).

*Análisis de riesgos.*

Como en todo proyecto propuesto, existieron riesgos que podrían haber afectado el correcto desarrollo de este. Entre los detectados se destacaron los siguientes:

- Tiempo limitado: el análisis amplio de variables de interés clínico requiere de suficiente tiempo, considerando la posterior utilización que podría darse a este instrumento, por lo que la validación del modelo en el plazo asignado y cumpliendo los hitos (en particular, considerando el tiempo constreñido por el hecho de que este sea un TFM con entregables específicos) puede ser un factor de riesgo. Para mitigar esta opción, se propusieron objetivos iniciales realistas y alcanzables sin olvidar que existe la opción de escalar el modelo en siguientes fases del proyecto.

- Control de calidad del dato y complejidad de los análisis: la obtención de datos es un punto crucial por lo que el control de calidad de este puede desvelar imponderables que podrían afectar el correcto desarrollo temporal del proyecto (presencia de valores faltantes, presencia de *outliers/inliers*, etc.) lo que además podría impactar en los subsiguientes análisis estadísticos. Como medida de mitigación se propuso una política de comunicación fluida con investigadores e investigadoras con amplia experiencia en el trabajo (bio)estadístico y de aprendizaje automático, tanto desde dentro de UOC como por parte del grupo de supervisión externa que apoyaron el avance del trabajo.

- Tamaño de la muestra: aunque este es un riesgo técnico, es interesante recordar que un tamaño de muestra pequeño podría afectar la generalización de los resultados, en particular en muestras con diversos bloques. Sin embargo, una vez establecidas las bases, los modelos pueden rápidamente escalarse y mejorarse por lo que, si existe un bajo rendimiento, los resultados podrán ser tomados como una prueba de concepto de alto impacto y ampliar la muestra en posteriores fases del proyecto.

- Poder computacional requerido: un aspecto que se observó como un reto posterior de la propuesta inicial es el poder computacional requerido tanto para el entrenamiento como para la aplicación del modelo. Aunque este aspecto no suele considerarse problemático, trabajar con conjuntos de datos complejos, como los utilizados en casos clínicos con cientos de registros y variables, resultó ser un desafío. Para abordarlo, se optó por ampliar el modelo a su máxima capacidad, utilizando más de 150 variables simultáneamente, con el objetivo de evaluar la viabilidad en ordenadores de sobremesa convencionales. A pesar del tiempo requerido, se comprobó que era factible realizar el análisis con equipos comúnmente disponibles, lo que permitió ajustar el calendario y realizar el análisis exploratorio de datos en una segunda fase. Este enfoque fue crucial, ya que eliminar variables sin justificación podría comprometer la utilidad del modelo, mientras que crear modelos excesivamente demandantes para equipos estándar carecería de interés práctico. Finalmente, se logró un equilibrio que permitió seleccionar variables relevantes tanto desde una perspectiva clínica como analítica y constituye el principal logro presentado en este trabajo.

Interfaz de usuario gráfica, Aplicación, Tabla, Excel

Descripción generada automáticamente

Tabla 1: Diagrama de Gantt tentativo. Se desglosan las tareas y fechas propuestas, así como los hitos esperados. En rombos rojos, los hitos correspondientes a PECs y como barra horizontal roja las fechas propuestas de defensa de TFM.

## Breve sumario de productos obtenidos.

Como resultado del proyecto presentado y el trabajo realizado, se desarrolló una herramienta de cribado breve basada en técnicas de aprendizaje automático. Esta herramienta permite clasificar, con altos niveles de sensibilidad y especificidad, a personas con COVID persistente que presenten afectación cognitiva en comparación con aquellas que no presenten esta afectación. La clasificación se basa en tres variables principales, derivadas de puntuaciones de pruebas neuropsicológicas tradicionales, agregando un nivel explicativo gracias a la aplicación de los valores SHAP. Todo el desarrollo de este producto se plasmó en un documento reproducible escrito en R (de tipo RMarkdown), cuyo código permite entrenar el modelo, obtener predicciones y calcular valores SHAP para explicar las predicciones obtenidas.

Una vez validada clínicamente, esta herramienta tendrá el potencial de reducir la necesidad de aplicar pruebas cognitivas exhaustivas, facilitando un seguimiento personalizado de pacientes. Además, su combinación con otras variables de interés clínico, como datos sociodemográficos y otros marcadores, permitirá maximizar la eficacia y eficiencia en el diagnóstico y tratamiento de la PCC.

Finalmente, todo el proceso llevado a cabo servirá como base para generar una memoria y una presentación que permitirán la defensa para optar a la titulación de Máster en Bioestadística y Bioinformática.

## Breve descripción de los otros capítulos de la memoria

Capítulo 2: Estado del arte. Este capítulo ofrece una recopilación de aportes bibliográficos que analiza los avances más recientes en el campo de interés, enfocándose específicamente en el problema que se pretende resolver. Proporciona el contexto científico y técnico necesario para entender la relevancia y los desafíos asociados al proyecto.

Capítulo 3: Materiales y métodos. Se describen en detalle los materiales y métodos utilizados durante el desarrollo del proyecto. Este apartado incluye tanto las herramientas empleadas como los procedimientos aplicados, para contribuir a la reproducibilidad y transparencia del trabajo presentado.

Capítulo 4: Resultados. Este capítulo presenta los principales resultados obtenidos a lo largo del desarrollo del proyecto. Se seleccionan y estructuran los datos más relevantes, acompañados de un análisis que permita interpretar su significado en el marco del problema planteado.

Capítulo 5: Conclusiones y trabajos futuros. En este apartado se exponen las conclusiones más significativas derivadas de los resultados. Además, se identifican posibles líneas de investigación futura que podrían ampliar o complementar los hallazgos obtenidos.

Capítulo 6: Glosario. Sin ánimo de ser exhaustivo, este capítulo proporciona un listado de términos técnicos y/o específicos utilizados en la memoria, con sus respectivas definiciones, para facilitar la comprensión de los conceptos clave del presente trabajo.

Capítulo 7: Bibliografía. Finalmente, este capítulo incluye todas las referencias bibliográficas que han sido utilizadas a lo largo del trabajo.

# Estado del arte

La COVID-19 ha sido uno de los mayores retos sanitarios de la historia moderna, afectando a más de 676 millones de personas y causando cerca de 6.9 millones de muertes a nivel global, considerando datos recopilados hasta el tercer trimestre de 2023 [18]. Aunque las campañas masivas de vacunación han reducido considerablemente la gravedad y mortalidad de la enfermedad [19], una proporción significativa de pacientes continúa experimentando síntomas debilitantes mucho después de la fase aguda de la infección. Este conjunto de síntomas persistentes se denomina COVID persistente (o, en inglés, *Long COVID* o *post-COVID condition* -en adelante, en este texto, PCC-), una condición caracterizada por la presencia de síntomas que perduran al menos dos meses tras la infección inicial y no pueden explicarse por otros diagnósticos [5]. La PCC afecta aproximadamente al 10-20 % de las personas que han padecido COVID-19, independientemente de la gravedad inicial de la infección [5]. Esta condición persistente tiene un impacto profundo en la calidad de vida de quienes la sufren, afectando su capacidad laboral, bienestar psicosocial y relaciones personales. Además, se ha observado que afecta en mayor proporción a mujeres de mediana edad y a personas con condiciones preexistentes, exacerbando las desigualdades en salud [20]. Los síntomas más comunes de la PCC incluyen fatiga, dificultad respiratoria y un amplio conjunto de más de 50 síntomas [3]. También se han documentado complicaciones neurológicas, como síndrome de Guillain-Barré, miopatías y temblores, que podrían estar relacionadas con la invasión directa del virus en el sistema nervioso central [12] o con la inflamación sistémica exacerbada [21]. Estas alteraciones también pueden reflejarse en estructuras periféricas, como la retina, la cual comparte características anatómicas y fisiológicas con los vasos cerebrales, proporcionando un biomarcador potencialmente útil para estudiar los efectos vasculares de la enfermedad [13]. Aunque la fisiopatología de la PCC aún no se comprende completamente, diversas hipótesis apuntan a la disfunción endotelial, la inflamación crónica y la neuroinflamación como mecanismos principales. La activación persistente de citocinas proinflamatorias, como IL-6 y TNF-alfa, junto con la activación glial, puede contribuir al daño neuronal y a la disfunción neurológica y cognitiva observada en las personas diagnosticadas [22], [23].

Entre las múltiples manifestaciones de la COVID persistente, las alteraciones neurológicas y neuropsicológicas, conocidas colectivamente como "NeuroCOVID", representan uno de los aspectos más incapacitantes [8]. Estudios de neuroimagen han identificado adelgazamiento cortical y alteraciones en la conectividad cerebral, lo que refuerza la base biológica de la afectación del sistema nervioso central [24].

Desde el punto de vista más estrictamente neuropsicológico, las deficiencias cognitivas presentes en la PCC incluyen déficits en la memoria a corto y largo plazo, atención, velocidad de procesamiento, fluidez verbal y funciones ejecutivas, entre otras. Este conjunto de síntomas es a menudo descrito por quienes lo padecen como una "niebla mental" y podría estar vinculada tanto a cambios estructurales como funcionales en el cerebro [25]. Considerando la sintomatología autorreportada por quienes padecen PCC, los dominios más comúnmente afectados son:

*A) Funciones ejecutivas*: estas habilidades, que incluyen la planificación, el control inhibitorio y la flexibilidad cognitiva, son fundamentales para la resolución de problemas y la toma de decisiones. Dentro de este dominio se pueden identificar subdominios tales como:

- Memoria de trabajo: fundamental para mantener y manipular información en tiempo real, especialmente útil para tareas complejas como el cálculo mental o la planificación.

- Fluidez verbal: evalúa la capacidad para generar palabras de manera rápida y organizada, reflejando tanto el acceso a la memoria como la habilidad lingüística.

- Inhibición de respuestas automáticas: implica la capacidad de controlar respuestas impulsivas o automáticas, esencial para el autocontrol y la flexibilidad cognitiva.

Las alteraciones en este dominio pueden explicar las dificultades para gestionar tareas complejas reportadas por pacientes con afectación cognitiva bajo el cuadro de COVID persistente.

*B) Atención y velocidad de procesamiento:* la capacidad para mantener la atención y procesar información rápidamente es esencial para la eficacia en actividades cotidianas. La atención se refiere a la capacidad de concentrarse en estímulos relevantes mientras se ignoran los irrelevantes, en tanto que la velocidad de procesamiento mide la rapidez con la que se procesa y se responde a la información. La fatiga mental y la ralentización cognitiva son síntomas frecuentes en el cuadro de COVID persistente y afectan directamente a este dominio.

*C) Memoria:* tanto la memoria verbal (la habilidad de recordar palabras o información auditiva, crítica para la comunicación y el aprendizaje, como la memoria visual (relacionada con la retención de información visual, como figuras o patrones, esencial para la orientación y la percepción espacial) pueden verse comprometidas en la PCC, afectando la capacidad de retener y recuperar información, lo cual impacta significativamente la calidad de vida de las personas diagnosticadas.

*D) Funciones visoespaciales y visoconstructivas:* estas habilidades, necesarias para interpretar y organizar información visual, son críticas para tareas como la orientación y la planificación espacial.

*E) Lenguaje:* la fluidez verbal, la denominación de objetos y la comprensión son aspectos esenciales del lenguaje que pueden estar afectados en la PCC, limitando la comunicación y las interacciones sociales. Además del acceso semántico general, este dominio también incluye habilidades como la fluidez verbal, ya mencionada, y la denominación, que son cruciales para la comunicación efectiva.

La evaluación sistemática de estos dominios cognitivos es clave para comprender el impacto de la COVID persistente en las funciones neuropsicológicas por lo que se aplican pruebas validadas específicas para alcanzar este propósito. Aunque existen diversas metodologías, la tabla 2 recoge algunas de las evaluaciones más comúnmente aplicadas.

|  |  |  |
| --- | --- | --- |
| **Dominios** | **Subdominio** | **Test** |
| Funciones ejecutivas | Memoria de trabajo | Digit span backward (WAIS-III) |
| TMT B - A (temps) |
| Fluidez verbal | Phonetic fluency (PMR) |
| Semantic fluency (animals) |
| Inhibición de respuestas automáticas | Stroop word-colors (interference) |
| Atención y velocidad de procesamiento | Atención | Digit span forward (WAIS-III) |
| Velocidad de procesamiento | SDMT (WAIS-III) |
| TMT A (time) |
| Symbol search (WAIS-III) |
| Memoria | Memoria verbal | RAVLT (summarize) |
| RAVLT (delayed recall) |
| Memoria visual | ROCF (delayed recall) |
| Funciones visoespaciales y visoconstructivas |  | ROCF (copy accuracy) |
| Lenguaje |  | BNT |
| Vocabulary (WAIS-III) |

Tabla 2: Evaluaciones estandarizadas de dominios cognitivos. Una breve descripción de cada prueba puede encontrarse en Anexos del trabajo.

Como se desprende de lo anterior, el diagnóstico de la PCC (con o sin afectación cognitiva) es un desafío clínico significativo debido a la heterogeneidad de los síntomas y la falta de (bio)marcadores específicos. Sin embargo, la aplicación de metodologías de ciencia de datos e IA han demostrado su utilidad para predecir cuadros complejos como la PCC [26].

De forma global, la IA es un área de investigación que comprende diversas subáreas, siendo el aprendizaje automático una de las más prometedoras en el área sanitaria, ya que facilita el análisis más rápido y preciso de grandes cantidades de datos, así como potencia la capacidad de los profesionales para tomar decisiones informadas [27]. Sin ser exhaustivos, y en términos generales, el aprendizaje automático utiliza algoritmos para identificar relaciones en grandes volúmenes de datos y construir modelos predictivos capaces de realizar clasificaciones, regresiones o agrupaciones. En el contexto de la COVID persistente, el aprendizaje automático permite integrar datos para crear modelos que podrían facilitar el diagnóstico y el manejo de la enfermedad [28], [29]. En la práctica clínica, los modelos predictivos son herramientas valiosas para identificar pacientes en riesgo, anticipar resultados clínicos y priorizar intervenciones. Los algoritmos de aprendizaje automático, como los basados en árboles de decisión, redes neuronales y métodos de ensamble, han demostrado ser particularmente efectivos en la clasificación de enfermedades complejas y multifactoriales como la COVID persistente [30]. Entre los algoritmos basados en árboles de decisión, el *Extreme Gradient Boosting* (XGBoost, en adelante) se ha destacado como una herramienta de vanguardia en el área biomédica. Este método utiliza una estrategia de ensamble que combina múltiples árboles de decisión para mejorar la precisión y reducir el riesgo de sobreajuste [31]. Además, es altamente eficiente en términos computacionales, lo que lo hace adecuado para conjuntos de datos grandes y complejos, siendo posible aplicarlo a un amplio abanico de situaciones clínicas [32], [33], incluyendo la enfermedad por COVID-19 [34]. XGBoost ofrece varias ventajas en el análisis de datos clínicos y neuropsicológicos. Primero, incorpora técnicas de regularización, como L1 (Lasso) y L2 (Ridge), para minimizar el impacto de variables irrelevantes y mejorar la interpretabilidad del modelo. Segundo, su capacidad para manejar valores faltantes y tratar relaciones no lineales lo convierte en una opción ideal para los estudios sobre COVID persistente, donde la heterogeneidad y la complejidad de los datos son desafíos comunes [31].

Considerando que nuestra *hipótesis de investigación* radica en que existen síntomas diferenciales en el grupo de pacientes que sufre de PCC con afectación cognitiva cuando son comparados con una población no infectada (o en pacientes sin PCC), consideramos que es posible el desarrollo de un instrumento de cribado, basado en pruebas neurocognitivas e IA, que permita una separación (clasificación) de ambos grupos de pacientes. Al combinar el poder predictivo de modelos de IA con datos neuropsicológicos validados, en este trabajo buscamos ofrecer una solución escalable, precisa y accesible para la detección temprana de esta condición, siendo de particular relevancia para la atención primaria, donde los recursos son limitados. En definitiva, este trabajo se centra en la creación de un instrumento de cribado que no solo facilite el diagnóstico de la COVID persistente con afectación cognitiva, sino que también permita generar un enfoque accesible para la atención clínica permitiendo diagnósticos más tempranos y personalizados, reduciendo la carga sobre los sistemas de salud y mejorando la calidad de vida de quienes sufren de esta condición.

# Materiales y métodos

Este proyecto busca el desarrollo de un instrumento de cribado breve para detectar casos de COVID persistente con afectación neuropsicológica, integrando datos neuropsicológicos y variables sociodemográficas mediante técnicas de aprendizaje automático. Para ello, se han utilizado y aplicado los materiales y métodos que se describen en este capítulo.

*3.1) Diseño del estudio y participantes.*

Se incluyeron personas mayores de 25 años atendidas en centros de atención primaria de Cataluña (centros de atención primaria en el área metropolitana norte de Barcelona y en el Hospital Universitario Germans Trias I Pujol). Los criterios de inclusión fueron los siguientes [11]:

*(a) Para el grupo de pacientes con infección por COVID y afectación neuropsicológica (PCC Cog):* las personas incluidas como pacientes deben cumplir al menos uno de los siguientes síntomas:

1) Presentar cualquier síntoma neurológico persistente relacionado con COVID-19 durante un período mayor a 12 semanas desde el inicio de la enfermedad, incluyendo migrañas recurrentes, anosmia, ageusia, mareos y vértigo.

2) Presentar cualquier síntoma neuropsicológico durante un período mayor a 12 semanas desde el inicio de la enfermedad, incluyendo la sensación subjetiva de deterioro en una o más funciones cognitivas.

3) Presentar cualquier síntoma vestibular o neuropsiquiátrico persistente relacionado con COVID-19 durante un período mayor a 12 semanas desde el inicio de la enfermedad, incluyendo trastornos depresivos, ansiedad o trastorno de estrés postraumático.

4) Presentar dos o más síntomas de los puntos 1 y 3.

*(b) Para el grupo de control sin infección por COVID (NoCOV):* quienes participen de este grupo no deben haber estado infectados con COVID-19, demostrado mediante una prueba Rt-PCR y serología negativas para nucleocápside.

*(c) Para el grupo de control de pacientes diagnosticados con PCC pero sin afectación neurocognitiva (PCC No cog):* quienes participen de este grupo deben presentar algunos síntomas persistentes de COVID-19 pero sin déficits neurocognitivos, durante un período mayor a 12 semanas desde el inicio de la enfermedad.

*(d) Para el grupo de pacientes sin PCC (NoPCC):* quienes participen deben haberse recuperado de la infección por COVID-19, demostrado mediante una prueba Rt-PCR positiva, y no presentar síntomas persistentes de COVID-19 durante un período mayor a 12 semanas desde el inicio de la enfermedad.

Por otra parte, los criterios de exclusión fueron:

1) Deterioro cognitivo, síntomas psiquiátricos como psicosis, esquizofrenia, trastorno por déficit de atención, antecedentes de abuso de drogas o alcohol, o enfermedad neurológica previa a la infección por COVID-19.

2) Presencia de cualquier condición que impida o contraindique la realización de una resonancia magnética (MRI).

3) Pronóstico vital menor a 6 meses.

Tras la aplicación de los criterios de inclusión y exclusión, se obtuvo una base de datos de 268 pacientes cuyas principales características sociodemográficas se resumen en la tabla 3.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Valor** | **n**  **(número de participantes)** | **Porcentaje** |
| **Sexo** | Femenino | 217 | **80,97** |
| Masculino | 51 | **19,03** |
| **Nivel educativo** | Educación primaria | 11 | **4,10** |
| Educación secundaria | 10 | **3,73** |
| Bachillerato | 85 | **31,72** |
| Grado universitario | 101 | **37,69** |
| Especialista/Máster | 43 | **16,04** |
| Doctorado | 18 | **6,72** |
| **Espectro clínico de la primera infección** | Asintomático | 10 | **3,73** |
| Sintomatología media/moderada | 185 | **69,03** |
| Hospitalizado | 35 | **13,06** |
| Unidad de cuidados intensivos | 7 | **2,61** |
| No disponible | 31 | **11,57** |
| **Índice de masa corporal (BMI) según clasificación OMS** | Bajo peso | 9 | **3,36** |
| Normal | 100 | **37,31** |
| Sobrepeso | 78 | **29,48** |
| Obesidad I | 36 | **13,43** |
| Obesidad II | 26 | **9,70** |
| Obesidad III | 12 | **4,48** |
| No disponible | 6 | **2,24** |
| **Diagnóstico** | Sin infección por COVID (No COVID) | 20 | **7,46** |
| Sin COVID persistente (COVID No PCC) | 47 | **17,54** |
| Con COVID persistente sin afectación neurocognitiva (COVID PCC No Cog) | 27 | **10,07** |
| Con COVID persistente con afectación neurocognitiva (COVID PCC Cog) | 174 | **64,93** |

Tabla 3: Variables sociodemográficas de la base de datos analizada.

La edad media de quienes participaron fue de 48,58 años y el BMI medio fue de 27,7 kg/m2.

*3.2) Generación de la base de datos de análisis.*

Con el objeto de balancear las categorías diagnósticas, se generaron 2 grupos principales de análisis: por un lado, la sumatoria de pacientes sin infección por COVID-19 y aquellos sin PCC (n=67; 27,8% de las observaciones), y por el otro, el conjunto de los diagnosticados con PCC con afectación neuropsicológica (n=174; 72,2% de las observaciones). En este trabajo no se hizo uso de los datos provenientes de pacientes con un diagnóstico de PCC pero sin afectación cognitiva. En consecuencia, la base de datos final contó con datos provenientes de 241 pacientes.

**A) Mayores de 25 años.**

**B) Generación del conjunto final.**

**Diagrama 1: A)** Diagrama de flujo que describe los criterios de inclusión aplicados para clasificar a los participantes en cuatro grupos principales: No COVID, COVID No PCC, COVID PCC No Cog y COVID PCC Cog. **B)** Se detalla el proceso de generación del conjunto final, que incluye la eliminación de participantes no elegibles y la creación de grupos de comparación finales: Grupo 1 (No COVID + COVID No PCC) y Grupo 2 (COVID PCC Cog). El total inicial fue de 268 participantes, culminando en 241 participantes distribuidos en los grupos de comparación.

*3.3) Administración, corrección y evaluación del estado neuropsicológico de las personas participantes.*

El equipo del área de psicología clínica sanitaria realizó la administración de las pruebas neuropsicológicas validadas, así como la corrección y registro de estas (centros de atención primaria en el área metropolitana norte de Barcelona y en el Hospital Universitario Germans Trias I Pujol). Tras una generación inicial de un modelado piloto (datos no mostrados) donde se contó con 151 variables iniciales (incluyendo variables binarizadas), se seleccionó el conjunto de variables sociodemográficas, clínicas y neuropsicológicas definitivas de interés que se recogen en la tabla 4.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Subdominio** | **Test/**  **Herramienta** | **Tipo de variable** | **Codificación/**  **Nombre de variable** |
| Análisis de la función ejecutiva | Memoria de trabajo | Digit span backward (WAIS-III) | Cuantitativa | tn46 |
| TMT B - A (temps) | Cuantitativa | tn52 |
| Fluidez verbal | Phonetic fluency (PMR) | Cuantitativa | tn36/tn38/tn40 |
| Semantic fluency (animals) | Cuantitativa | tn42 |
| Inhibición de respuestas automáticas | Stroop word-colors (interference) | Cuantitativa | tn22 |
| Análisis de la atención y velocidad de procesamiento | Atención | Digit span forward (WAIS-III) | Cuantitativa | tn44 |
| Velocidad de procesamiento | SDMT (WAIS-III) | Cuantitativa | tn14 |
| TMT A (time) | Cuantitativa | tn24 |
| Symbol search (WAIS-III) | Cuantitativa | tn12 |
| Memoria | Memoria verbal | RAVLT (summarize) | Cuantitativa | tn6 |
| RAVLT (delayed recall) | Cuantitativa | tn30 |
| Memoria visual | ROCF (delayed recall) | Cuantitativa | tn34 |
| Funciones visoespaciales y visoconstructivas |  | ROCF (copy accuracy) | Cuantitativa | tn8 |
| Lenguaje |  | BNT | Cuantitativa | tn48 |
| Vocabulary (WAIS-III) | Cuantitativa | tn50 |
| Edad |  |  | Cuantitativa | ag |
| BMI |  |  | Cuantitativa | ptg19 |
| Sexo |  |  | Binaria | sex |
| Espectro clínico |  |  | Multinomial | spec1 |
| Nivel educativo |  |  | Multinomial | el |
| Diagnóstico |  |  | Binaria | Target |

Tabla 4: Variables seleccionadas en la construcción de la base de datos definitiva utilizada en los análisis subsiguientes. La columna de codificación recoge el nombre codificado para cada variable en los análisis posteriores.

En total se seleccionaron 23 variables de interés provenientes de datos de 241 participantes. Tras la generación de la base de datos definitiva para realizar el análisis, el desarrollo del trabajo siguió un enfoque cuantitativo, con fases progresivas que abarcaron desde el análisis exploratorio de los datos hasta la validación de modelos predictivos.

*3.4) Análisis exploratorio de los datos.*

Se llevó a cabo un análisis descriptivo inicial de las variables seleccionadas para identificar características relevantes. En cuanto a los valores faltantes, se implementó una estrategia de imputación general mediante el algoritmo de k vecinos más cercanos (k-NN). De manera complementaria, se separaron las variables cualitativas de las cuantitativas y, siempre que fue posible, se priorizó el uso de análisis gráficos (histogramas, mapas de calor, etc.). Se realizaron análisis gráficos univariantes, bivariantes y multivariantes utilizando paquetes específicos de R.

*3.5) Reducción de la dimensionalidad.*

Se llevó a cabo una selección automatizada de variables de baja variabilidad utilizando el paquete caret. Paralelamente, se realizó un análisis de componentes principales (PCA) para evaluar la posibilidad de reducir la dimensionalidad de la base de datos (paquetes FactoMineR y factoextra). Asimismo, se exploraron patrones en los datos mediante agrupamiento no supervisado, empleando el método de agrupamiento jerárquico.

*3.6) Análisis estadístico inferencial.*

Para las variables cualitativas, se realizó un análisis de asociación utilizando la prueba de Chi-cuadrado. Se consideró que existían diferencias estadísticamente significativas cuando el valor p fue inferior a 0,05. En el caso de las variables cuantitativas, para la comparación de medias entre los grupos definidos se utilizó la prueba de la U de Mann-Whitney, considerando diferencias estadísticamente significativas cuando los valores p fueron menores a 0,05.

*3.7) Modelos de análisis predictivo.*

La metodología para generar modelos predictivos se centró en los siguientes pasos:

*Selección de variables y pruebas:* como se indicó previamente, se realizó un modelado piloto que permitió que se seleccionaron datos provenientes de pruebas neuropsicológicas validadas que evaluaron los dominios cognitivos susceptibles de ser afectados por el COVID persistente con afectación neurocognitiva. Adicionalmente, se incluyeron variables sociodemográficas clave (como edad, sexo y nivel educativo) para explorar su impacto en el diagnóstico y mejorar la capacidad predictiva del modelo.

*Preparación de los datos:* los datos se sometieron a un proceso de limpieza y preprocesamiento para garantizar su calidad. Esto incluyó la imputación de valores faltantes, transformaciones de las variables (como binarizaciones y escalados donde fuera necesario) y reconstrucción del conjunto de datos para optimizar su adecuación a los modelos de aprendizaje automático.

*Modelado predictivo:* brevemente, tras lo indicado previamente, se entrenaron dos modelos con el objetivo de realizar una comparación: un modelo de regresión logística, considerado como el modelo base, y un modelo de tipo XGBoost. Los datos se dividieron en dos subconjuntos: un 80% para entrenamiento y un 20% para prueba. Tras realizar validación cruzada, se obtuvieron las matrices de confusión para ambos modelos, obteniéndose métricas clave de cara a la aplicación clínica de los modelos. Más precisamente sea analizó sensibilidad, especificidad, valor F1 y AUC. Además, se generaron las curvas ROC correspondientes para evaluar la capacidad discriminativa de los modelos correspondientes.

*Decisiones metodológicas.*

Dada la naturaleza compleja de la base de datos y el objetivo de desarrollar un instrumento escalable y clínicamente aplicable, se consideraron diferentes enfoques metodológicos:

*Alternativas evaluadas:* inicialmente se exploraron métodos simples, como análisis univariados, multivariados o de estadística inferencial. También se buscaron opciones de reducción de la dimensionalidad como un primer paso analítico. Sin embargo, estos enfoques mostraron limitaciones para manejar la complejidad de los datos y capturar relaciones no lineales.

*Criterios para la selección de metodología:* tras sopesar las opciones previas (y contrastarlas estadística y programáticamente) se optó por modelos avanzados de aprendizaje automático debido a su capacidad para manejar datos complejos, generar predicciones precisas y proporcionar herramientas interpretativas como los valores SHAP, que pueden ser esenciales en la explicabilidad y aplicabilidad de los modelos generados en entornos clínicos.

*3.8) Producto/s obtenido/s*

- Modelos predictivos iniciales: desarrollo de modelos que pueden actuar como un cribado de pacientes que sufren PCC con alteración neurocognitiva. Los resultados, a la espera de la validación clínica, indican que las pruebas neuropsicológicas son, efectivamente, variables altamente relevantes para la obtención de esta clasificación.

- Gráficas y cálculo de valores SHAP: que permitirán una interpretación visual y un análisis de importancia de variables personalizado a cada observación, lo que permite comprender el peso de cada prueba neuropsicológica en las predicciones aumentando la transparencia en el uso de los modelos generados como apoyo a decisiones clínicas.

*3.9) Programas y paquetes utilizados.*

Todo el análisis se realizó utilizando el software R (versión 4.4.2, lanzado en octubre de 2024) dentro del entorno de desarrollo integrado RStudio® (versión 2024.12.0+467, Kousa Dogwood). RStudio® está desarrollado por RStudio Team (2023) como un entorno de programación para R, distribuido por Posit Software, PBC (Boston, MA, USA. Más información sobre el software en [https://posit.co](https://posit.co/)). Las funciones empleadas en este trabajo se obtuvieron de paquetes disponibles bajo licencia GNU, alojados en *The Comprehensive R Archive Network* (CRAN, [https://cran.r-project.org](https://cran.r-project.org/)). La tabla 5 presenta un resumen de los principales paquetes utilizados (no se incluyen las correspondientes dependencias).

|  |  |  |
| --- | --- | --- |
| **Paquete** | **Versión** | **Responsable/Web/Cita** |
| library(knitr) | 1.49 | Xie Y <[https://yihui.org](https://yihui.org/)> |
| library(fastDummies) | 1.7.4 | Kaplan J |
| library(xgboost) | 1.7.8.1 | Tianqi C and Guestrin C [35] |
| library(tidyverse) | 2.0.0 | Wickham H |
| library(naniar) | 1.1.0 | Cook D, McBain M, Fay C. |
| library(corrplot) | 0.95 | Wei T, Simko V, Levy M. |
| library(caret) | 7.0-1 | Kuhn M. y cols. |
| library(DMwR2) | 0.0.2 | Torgo, L. [36] |
| library(reshape2) | 1.4.4 | Wickham H. |
| library(FactoMineR) | 2.11 | Husson F, Josse J, Le S, Mazet J [37] |
| library(factoextra) | 1.0.7 | Kassambara A. |
| library(ggrepel) | 0.9.6 | Slowikowski K. y cols. |
| library(MASS) | 7.3-61 | Ripley B y cols [38] |
| library(car) | 3.1-3 | Fox y cols. [39] |
| library(multcomp) | 1.4-26 | Hothron T y cols [40] |
| library(pROC) | 1.18.5 | Robin X. [41] |
| library(fastshap) | 0.1.1 | Greenwell B. |

Tabla 5: Paquetes de R, con sus versiones y responsables, utilizados en el desarrollo de este trabajo.

# Resultados

**4.1) Análisis exploratorio de los datos.**

*4.1.1) Manejo de datos faltantes (NAs).*

En 18 variables se observaron valores faltantes. El máximo porcentaje de valores faltantes se observó en la variable ptg19 (índice de masa corporal) con un 2,5% del total de las observaciones recogidas. En el caso de las variables que cuantificaron aspectos neuropsicológicos, se observaron datos faltantes que correspondieron a entre un 1,25% hasta un máximo del 1,75% (para el caso de tn22, relacionada con la evaluación de la inhibición de respuestas automáticas) de las observaciones. Considerando el bajo porcentaje, se optó por una imputación aplicando el algoritmo de vecinos más cercanos (kNN). Tras la imputación, la base de datos se completó adecuadamente, careciendo de datos faltantes (Figura 1).

a)



b)



Figura 1: Imputación de datos faltantes. a) Proporción de datos faltantes por variable antes de la aplicación del algoritmo kNN. b) Proporción de datos faltantes por variable tras la imputación.

*4.1.2) Análisis univariado.*

Tras la selección de variables, el análisis exploratorio de los datos se inició con un análisis univariado de las variables cuantitativas (y binarizadas) de interés incluidas en la base de datos (Figura 2). De forma visual, la variable ag (edad) tiene una distribución normal aparente. En el caso del BMI (ptg19), existe un desplazamiento hacia la izquierda concordante en que no existen valores extremos para esta variable. Con respecto a las variables neuropsicológicas, se observa que algunas de ellas pueden no seguir distribuciones normales. Estas variables están asociadas a subdominios como la memoria verbal, las funciones visoespaciales y visoconstructivas, la atención, la memoria de trabajo y el lenguaje. (tales como tn6, tn8, tn30, tn44, tn46 y tn48).



Figura 2: Análisis exploratorio de los datos. Distribución univariada de variables cuantitativas (incluyendo variables binarizadas). En el eje X se representan los valores cuantitativos en tanto que en el eje Y se representan las frecuencias correspondientes. Se observa una distribución aproximadamente normal en variables de interés para el desarrollo de análisis subsiguientes.

También como parte del análisis exploratorio de los datos se realizó un análisis gráfico de variables cualitativas, considerando las frecuencias de datos. Se observa un desbalance en clases correspondientes al diagnóstico clínico (Target). También existe un desbalance en la categoría de sexo (sex) y en diversos niveles educativos (el) (Figura 3).



Figura 3: Análisis exploratorio de los datos. Distribución univariada de variables cualitativas. En el eje X se indica la ausencia (cero) o presencia (uno) de la categoría y en el eje Y se indica la frecuencia de observaciones.

*4.1.3) Análisis de correlaciones.*

Como continuación del análisis exploratorio de los datos, se realizó un análisis de correlación de Pearson para las principales variables cuantitativas. Tal como se esperaba, las variables que evaluaron dominios/subdominios específicos mostraron altos niveles de correlación (tales como tn36, tn38 y tn40). Adicionalmente, algunas variables binarizadas presentaron patrones de correlación esperados, aunque cabe señalar que estos resultados deben interpretarse con cautela debido a las limitaciones matemáticas de aplicar esta medida a variables categóricas o binarias. Por ejemplo, se observó una alta correlación negativa entre personas con sintomatología media o moderada y hospitalización, así como entre distintos niveles educativos. Por último, se identificó que el diagnóstico (Target) mostró correlaciones negativas significativas con ciertas variables neuropsicológicas cuantitativas, tales como tn12, tn14 y tn42, (las 2 primeras relacionadas con la velocidad de procesamiento y la última, con la fluidez verbal) lo que refuerza su posible relación con estos dominios.



Figura 4: Análisis exploratorio de los datos. Mapa de calor de correlación. La matriz muestra los coeficientes de correlación de Pearson entre variables seleccionades. Los colores que tienden al azul corresponden a coeficientes de correlación positiva en tanto que los colores que tienden al rojo corresponden a coeficientes de correlación negativa.

*4.1.4) Análisis bivariado.*

Como continuación del análisis exploratorio de los datos, se llevó a cabo un análisis bivariado de las principales variables incluidas en el estudio, utilizando el diagnóstico como variable de agrupación. Mediante el uso de diagramas de cajas y bigotes, el análisis bivariado permitió observar la presencia de ciertos valores que podrían considerarse extremos en las variables.



Figura 5: Análisis exploratorio de los datos. Análisis bivariado, agrupado por diagnóstico. Las cajas representan el rango intercuartílico (entre el primer y tercer cuartil), con la mediana indicada por una línea dentro de cada caja. Los bigotes se extienden hasta 1.5 veces el rango intercuartílico, y los puntos negros fuera de los bigotes indican posibles valores extremos.

*4.1.5) Reducción de la dimensionalidad.*

Con el objeto de facilitar análisis posteriores, y siguiendo con el análisis exploratorio, se optó por realizar una reducción de la dimensionalidad aplicando un análisis de componentes principales (PCA). La proporción de varianza explicada por los dos componentes principales solo alcanzó el 30,6% sin observarse una clara definición cuando se agruparon las observaciones considerando el diagnóstico (Figura 6).



Figura 6: Análisis exploratorio de los datos. Análisis de Componentes Principales, agrupado por diagnóstico. Representación de las proyecciones de las observaciones sobre los dos primeros componentes principales, agrupadas por diagnóstico. Los ejes muestran el porcentaje de varianza explicada por cada componente. Las observaciones están coloreadas y simbolizadas según su grupo de diagnóstico (0/Círculo = Sin infección por COVID/PCC No Cog; 1/Triángulo = PCC Cog).

Tras el PCA, se observó que se requieren al menos 13 componentes para alcanzar un 80% de varianza explicada, por lo que la reducción de dimensionalidad por PCA no generó resultados optimizados para el problema planteado en este trabajo (Figura 7).



Figura 7: Análisis exploratorio de los datos. *Scree Plot*. El eje Y muestra la proporción de varianza explicada por cada componente principal, mientras que el eje X indica el número de componentes. La selección del número óptimo de componentes puede basarse en el "codo" del gráfico, donde la varianza explicada acumulada deja de aumentar significativamente. En este caso, el "codo" podría hallarse entre el tercer y cuarto componente, pero la varianza acumulada es muy baja ya que subsiguientes componentes explican una proporción baja de la varianza.

Finalmente, el uso de otros modelos no supervisados, como el agrupamiento jerárquico, tampoco permitió la separación de grupos claramente definidos (datos no mostrados).

**4.2) Análisis estadísticos inferenciales.**

*4.2.1) Variables cualitativas.*

Para contrastar la hipótesis de asociación entre variables cualitativas, se aplicó la prueba de Chi cuadrado para cada par de variables. Se observaron asociaciones estadísticamente significativas (p < 0.05) entre las variables spec1 (espectro clínico de la primera infección) y el (nivel educativo) con el diagnóstico emitido (Target). Sin embargo, estos resultados deben interpretarse con cautela debido al desbalance de clases observado en el análisis exploratorio de los datos, ya que este podría haber influido en los valores p obtenidos.

*4.2.2) Variables cuantitativas.*

Considerando el problema planteado y la naturaleza de las variables explicativas cuantitativas, se llevó a cabo un análisis exploratorio para verificar los supuestos necesarios para aplicar pruebas paramétricas. Al constatar que dichas variables no cumplían con los requisitos de normalidad y/o homogeneidad de varianzas, se optó por emplear una prueba no paramétrica, específicamente la U de Mann-Whitney. Los resultados mostraron diferencias estadísticamente significativas (p<0.05) para todas las variables evaluadas, con excepción de las variables tn8, tn34 y tn52 (relacionadas con las funciones visoespaciales y visoconstructivas, memoria visual y memoria de trabajo, respectivamente), las cuales no presentaron diferencias significativas entre los grupos comparados.

**4.3) Análisis predictivos.**

*4.3.1) Generación de un modelo de clasificación utilizando regresión logística.*

Tras el análisis exploratorio de los datos, se procedió a desarrollar un modelo inicial de aprendizaje automático supervisado con el objetivo de clasificar las observaciones. Dado que la variable objetivo presenta dos clases, se optó por construir un modelo de regresión logística como punto de partida. Este tipo de modelo es ampliamente utilizado debido a su balance entre simplicidad, interpretabilidad y capacidad para estimar probabilidades asociadas a cada clase. Sin embargo, también está limitado por una serie de asunciones estadísticas, tales como la linealidad de las relaciones entre las variables predictoras y el logit de la variable objetivo, la independencia de los errores, la ausencia de multicolinealidad significativa entre las variables explicativas, entre otras. Al contrastarse estos supuestos (mediante el cálculo del factor de inflación de la varianza -VIF- y mediante análisis gráficos) se observó que todas las variables cumplen los supuestos, excepto por la variable tn6 (que evalúa memoria verbal) la cual no cumple el supuesto de independencia de los errores (datos no mostrados). A pesar de esta limitación, la regresión logística se seleccionó como modelo base para proporcionar un marco de referencia inicial, que permitiera evaluar inicialmente la capacidad de las variables explicativas para discriminar entre las clases y establecer una línea de rendimiento. Este enfoque será complementado posteriormente con modelos más complejos y flexibles que no dependan de las mismas restricciones, para garantizar un análisis más robusto y mejorar el rendimiento predictivo.

Para evaluar el desempeño del modelo logístico, se dividió la base de datos en dos subconjuntos: un 80% de las observaciones fue utilizado para el entrenamiento del modelo, mientras que el 20% restante se destinó para la prueba, asegurando que este conjunto de datos no se hubiera presentado previamente al modelo durante su entrenamiento. Los resultados del análisis indicaron que los coeficientes asociados a las variables ptg19, tn12 y tn30 (que corresponden a las variables de índice de masa corporal, velocidad de procesamiento y memoria verbal, respectivamente) fueron estadísticamente significativos, sugiriendo una contribución relevante de estas variables en la predicción de la variable objetivo.

En línea con los objetivos clínicos de este trabajo, se priorizaron métricas clave para evaluar el rendimiento del modelo. La sensibilidad, definida como la capacidad del modelo para identificar correctamente los casos positivos, alcanzó un 87,10%. Por otro lado, la especificidad, que mide la proporción de verdaderos negativos correctamente identificados, fue del 33,33%. Además, se obtuvo el valor F1, una métrica que combina la precisión y la sensibilidad en una única media armónica, que alcanzó un valor de 77,14%. En cuanto a la exactitud (*accuracy*, en inglés), que representa la proporción de predicciones correctas en el total de observaciones, fue del 67,35%. Finalmente, el AUC fue 0,719. La curva ROC generada por el modelo es recogida en la figura 8.



Figura 8: Curva ROC (Modelo logístico). Se generó un modelo logístico para utilizarse como modelo de base sobre el que comparar modelos más complejos. Esta curva representa la la especificidad en el eje X y la sensibilidad en el eje Y. Un modelo con sensibilidad y especificidad máxima generaría una gráfica en forma de L invertida, en tanto que un modelo que generara predicciones al azar generaría una gráfica que se acercaría a una línea diagonal. A medida que los valores de sensibilidad y especificidad obtenidos por el modelo mejoran, el gráfico se acerca a la forma ideal de L invertida.

Otro modelo que puede ser utilizado como base puede ser un modelo de tipo discriminante lineal, pero el conjunto de datos no cumple con los postulados requeridos por lo que no es aplicable (datos no mostrados).

*4.3.2) Generación de un modelo de clasificación utilizando XGBoost.*

Dada la naturaleza de los datos clínicos utilizados, se consideró que un modelo basado en árboles sería más adecuado para abordar el problema planteado. Entre las diversas opciones disponibles, XGBoost destaca como una de las técnicas típicamente seleccionadas en el ámbito biomédico debido a su alta escalabilidad, robustez frente a datos desbalanceados y flexibilidad en cuanto a los supuestos matemáticos y estadísticos. Sin embargo, una de las limitaciones de XGBoost es su carácter de “caja negra”, lo que dificulta la interpretación directa de las predicciones. Una forma de abordar esta limitación es mediante el análisis de la importancia de las variables, un parámetro que mide la contribución de cada variable a la generación de las “ramas” de decisión de cada árbol generado durante el entrenamiento del modelo. No obstante, esta métrica solo proporciona una visión general del modelo y no permite interpretar de forma específica el impacto de las variables en cada observación individual. Para superar esta restricción, se pueden emplear los valores SHAP, que permiten cuantificar la contribución de cada variable a cada una de las predicciones realizadas, ofreciendo así una interpretación más detallada y personalizada del modelo, contribuyendo a la transparencia de las decisiones clínicas que puedan ser apoyadas por la aplicación de este.

Para la construcción del modelo, los datos se dividieron en dos subconjuntos: un 80% para entrenamiento y un 20% para prueba. Posteriormente, se aplicó validación cruzada para ajustar los hiperparámetros óptimos, utilizando el área bajo la curva ROC (AUC-ROC) como métrica de optimización. Finalmente, el modelo fue entrenado de forma definitiva utilizando los hiperparámetros seleccionados, con el objeto de maximizar su rendimiento predictivo.

Nuevamente, en línea con los objetivos clínicos de este trabajo, se priorizaron métricas clave para evaluar el rendimiento del modelo. La sensibilidad alcanzó un 78,38% y la especificidad fue del 90,91%. Además, se obtuvo un valor F1 de 86,57%. En cuanto a la exactitud, fue del 81,25%. Finalmente, el valor AUC fue 0,885. La curva ROC generada por el modelo es recogida en la figura 9.



Figura 9: Curva ROC (Modelo XGBoost). Se generó un modelo de tipo XGBoost, como ejemplo de modelo complejo, para comparar los resultados obtenidos frente al modelo logístico. La curva ROC representa la sensibilidad en el eje Y y la especificidad en el eje X. Un modelo con sensibilidad y especificidad máxima generaría una gráfica en forma de L invertida, en tanto que un modelo que generara predicciones al azar generaría una gráfica que se acercaría a una línea diagonal. A medida que los valores de sensibilidad y especificidad obtenidos por el modelo mejoran, el gráfico se acerca a la forma ideal de L invertida.

En la tabla 6 observamos la comparativa de las métricas obtenidas en los modelos y su cambio porcentual.

|  |  |  |  |
| --- | --- | --- | --- |
| **Métrica de interés** | **Modelo Logístico** | **Modelo XGBoost** | **Cambio en porcentaje (Log vs XGBoost)** |
| Sensibilidad (%) | 87,10 | 78,38 | -10,01 |
| Especificidad (%) | 33,33 | 90,91 | +172,75 |
| Valor F1 (%) | 77,14 | 86,57 | +12,22 |
| Exactitud (%) | 67,35 | 81,25 | +20,64 |
| AUC | 0,719 | 0,885 | +23,08 |

Tabla 6: Comparativa de métricas entre los modelos desarrollados. Se compara el modelo XGBoost contra el modelo base (regresión logística). Valores positivos indican mejora de la métrica por parte del modelo XGBoost comparado con el logístico. Valores negativos indican empeoramiento de la métrica en la misma comparación.

En conclusión, a pesar de una pérdida en la sensibilidad, el modelo XGBoost mejora las métricas de interés propuestas para este trabajo, mejorando particularmente en los valores F1 que representan una media armónica entre sensibilidad y especificidad. Este valor es clave para analizar resultados en bases de datos clínicas donde pueden existir desbalances de clases (como nuestro caso) o ante la presencia de relaciones no lineales y de alta complejidad. Por otra parte, el modelo XGBoost mejora en un 23,08% los valores de AUC cuando son comparados con el modelo de base (logístico), mostrando su eficacia y eficiencia en sensibilidad y especificidad (Figura 10).



Figura 10: Comparación de curvas ROC para el modelo de base (logístico, en rojo) en comparación con el modelo complejo (XGBoost, en azul).

*4.3.3) Cálculo de la importancia de las variables en las decisiones del modelo XGBoost.*

El modelo XGBoost generado presenta limitaciones en cuanto a su interpretabilidad. Como primera aproximación, se calculó la importancia de las variables utilizando la métrica de ganancia (*gain*, en inglés), que mide la mejora en la función objetivo atribuida a cada variable en la generación de las ramas de decisión (Figura 11). En el modelo desarrollado en este trabajo, la variable tn12 (que evalúa velocidad de procesamiento) resultó ser la que más contribuye a la creación de nuevas ramas. Le siguen, aunque con una contribución significativamente menor, las variables tn30 (memoria verbal), tn6 (memoria verbal), el sexo (mujer), tn14 (velocidad de procesamiento) y tn40 (fluidez verbal), en ese orden.

Además, la métrica de ganancia mostró concordancia con la métrica de cobertura (*coverage*, en inglés), que evalúa qué variables se utilizan para dividir nodos que impactan un mayor número de instancias en el árbol. Esta concordancia se observó particularmente para las variables tn12, tn30, tn6 y tn40, reforzando la idea de que estas características tienen un papel clave en la clasificación eficaz y eficiente entre los grupos de pacientes analizados.



Figura 11: Importancia de las variables (modelo XGBoost). La importancia de las variables se calcula con base en la métrica *gain*, representando la contribución relativa de cada variable a la construcción del modelo. Las barras indican la magnitud de esta importancia, con las variables más influyentes listadas en la parte superior. Este análisis permite identificar las características clave que impactan en las predicciones, facilitando un primer acceso a la interpretación del modelo y a su conexión con el contexto clínico.

*4.3.4) Cálculo de valores SHAP.*

Aunque las métricas previas permitieron comprender el rol general de las variables en la construcción del modelo, estas no facilitan una evaluación explicativa a nivel individual para cada predicción realizada. Para superar esta limitación, se calcularon los valores SHAP, que permiten interpretar la contribución específica de cada variable en las predicciones individuales. La Figura 12 muestra un ejemplo de los valores SHAP correspondientes a una observación seleccionada del conjunto de prueba. En este ejemplo, la observación seleccionada se incluye en el diagnóstico de PCC con afectación cognitiva y el modelo predice correctamente su integración en este grupo de pacientes, siendo los valores de la variable tn12 (y, en menor proporción, los de las variables tn30, el sexo y tn40) los que más aportan a que dicha observación sea predicha como parte de dicho grupo. De esta forma, los valores SHAP brindan una información de interés para poder realizar una evaluación crítica del resultado de la predicción ejecutada por el modelo y facilitar la comprensión por parte del personal clínico de la adecuación (o no) del resultado observado. En definitiva, este enfoque permite una explicación más transparente del comportamiento del modelo y puede ser aplicado a cada observación para garantizar la explicabilidad, lo cual es esencial en contextos como el apoyo a decisiones clínicas (figura 12).



Figura 12: Ejemplo de cálculo de valores SHAP para una observación seleccionada del conjunto de prueba. El gráfico muestra la contribución individual de cada variable a la predicción realizada por el modelo. Las barras representan los valores SHAP, donde un valor positivo indica un aumento en la probabilidad predicha, y un valor negativo indica una disminución. Este análisis permite interpretar cómo las variables específicas (tales como tn12, tn30 y el sexo, en este caso de ejemplo), influyen en la predicción final, promoviendo la transparencia del modelo en aplicaciones clínicas.

Con el objetivo de obtener una visión global de la importancia de las variables en el conjunto de prueba, se calcularon los valores SHAP medios absolutos para cada variable. Este enfoque permite evaluar la contribución promedio de cada característica a las predicciones realizadas por el modelo. Como se muestra en la Figura 13, las variables tn12 (velocidad de procesamiento), tn30, tn6 (ambas relacionadas con la memoria verbal) y el sexo (mujer) destacan como las más influyentes en el modelo, con valores SHAP medios absolutos más altos. Esto refuerza su relevancia en la clasificación, ya que tienen un impacto consistente y significativo en las predicciones.



Figura 13: Valores SHAP medios absolutos para el conjunto de prueba. El gráfico presenta la importancia promedio de cada variable en las predicciones realizadas por el modelo, medida a través de los valores SHAP medios absolutos. Las variables con mayores valores, como tn12, tn30, tn6 y el sexo (mujer), destacan por su impacto consistente en la toma de decisiones del modelo, lo que subraya su relevancia en el análisis global.

# Conclusiones y trabajos futuros

En el transcurso de este trabajo, hemos logrado desarrollar un instrumento de cribado que responde de manera eficaz al desafío clínico de identificar pacientes con COVID persistente y afectación neurocognitiva. Los resultados obtenidos no solo confirman la validez de la metodología aplicada, sino que también revelan hallazgos interesantes que apuntalan la necesidad de tener en cuenta los factores neuropsicológicos durante esta condición. Si bien la reducción de dimensionalidad a través de PCA no produjo resultados óptimos esperados, el uso de un elevado número de variables (mediante algoritmos avanzados) permitió superar esta limitación, dando lugar a modelos efectivos, a los que se intentó dotar de facilidad de interpretación y transparencia. El instrumento finalmente desarrollado ha demostrado una capacidad elevada para diferenciar pacientes con PCC y afectación neurocognitiva frente a aquellos que no presentan dicha condición. Más precisamente, los modelos predictivos, en particular el modelo XGBoost, alcanzaron elevados niveles de sensibilidad y especificidad, indicando que ciertas variables neuropsicológicas específicas, como tn12 (que está relacionada con la evaluación de la velocidad de procesamiento), tn30 y tn6 (ambas evaluando la memoria verbal) han emergido como variables cruciales en las predicciones. Aunque la riqueza informativa del modelo generado es mayor, podríamos concluir que estos 2 subdominios son claves para poder diferenciar a los grupos. Por otra parte, dichas 3 variables (y, en consecuencia, dominios) mostraron un peso de discriminación que las transforma en elementos de interés para promover avances en el rol que pueden desempeñar en modelos más complejos (o simplificados) en fases posteriores del proyecto. También, considerando el aspecto sociodemográfico, en este trabajo demostramos que el sexo femenino puede representar un factor de riesgo para el desarrollo de PCC con afectación cognitiva, recalcando la utilidad de la aplicación adecuada de una perspectiva de género y diversidad al diseñar los proyectos de investigación. De forma global, los resultados apoyan el rol que tiene la investigación multidisciplinar (neurobiología, neuropsicología, medicina e IA) para responder frente a retos de salud pública. Por otra parte, este trabajo podría representar un caso claro de búsqueda de mejora en la efectivad y eficiencia: el modelo de cribado breve desarrollado podría aplicarse en entornos clínicos para identificar personas con PCC, de manera rápida y con pocas pruebas neuropsicológicas validadas (pasando de las 15 utilizadas inicialmente a tan solo 3), subrayando la importancia del rol de la modelación predictiva para el acompañamiento en la toma de decisiones clínicas. El modelo, complementado con otras variables de interés clínico, neuropsicológico y/o biológico, podría permitir también el seguimiento y pronóstico personalizado para aquellas personas afectadas de PCC.

Yendo hacia la reflexión crítica de los objetivos propuestos, consideramos que el objetivo principal de este trabajo, que consistía en desarrollar un instrumento de cribado breve y efectivo, fue alcanzado con éxito. Los objetivos específicos, como la validación de modelos predictivos y la identificación de variables relevantes, también se lograron en gran medida. Sin embargo, aspectos como la evaluación de métodos no supervisados quedaron limitados debido a restricciones en la efectividad real de su aplicación y en la aplicabilidad de los resultados obtenidos. Algo similar se pudo observar al momento de realizar otros análisis paramétricos. En definitiva, es clave un diseño estadístico realista cuando se debe trabajar con datos de alta complejidad, con patrones no lineales y con un gran número de variables, tal como suele ser frecuente en datos clínicos. Esto también nos lleva a reflexionar críticamente de cara a los resultados que podrían presentarse como parte de publicaciones científicas, ya que es clave que estos representen adecuadamente una parte de la realidad, pero siempre ajustados a sus supuestos matemáticos/estadísticos básicos. Finalmente, a pesar de estas limitaciones sobre los objetivos, los hallazgos obtenidos representan un avance interesante en el área, destacando la utilidad de combinar datos neuropsicológicos validados con herramientas avanzadas de aprendizaje automático.

Posteriormente, centrándonos en el estudio crítico de la planificación y metodología, podríamos indicar que, en general, la planificación inicial se cumplió en los plazos previstos, aunque ciertas tareas tuvieron que reorganizarse. En futuros proyectos, podría ser beneficioso implementar una cantidad de tiempo inicial para realizar estudios piloto, ya que son fases preliminares que representan una inversión inicial pero que se recupera con creces en fases avanzadas. Probablemente este es un punto de mejora y crítica constructiva, siempre teniendo en cuenta que el tiempo se encuentra constreñido para la realización de un TFM. En cuanto a la metodología utilizada demostró ser adecuada, destacando la efectividad de combinar pruebas neuropsicológicas con modelos avanzados como XGBoost. Este modelo resultó particularmente útil debido a su capacidad para manejar datos complejos y desbalanceados, integrar variables de diferentes tipos (como sociodemográficas y neuropsicológicas), y proporcionar interpretaciones con cierto nivel de detalle a través de valores SHAP. De hecho, este último punto se introdujo en la estrategia analítica y apuntaló la mejora en la transparencia del modelo para su uso clínico (el que también era un objetivo clave).

Considerando los impactos del proyecto en las áreas de sostenibilidad, ético-social y de diversidad consideramos que se han logrado. Desde el punto de vista de la sostenibilidad, el instrumento desarrollado contribuirá a optimizar recursos en los sistemas de salud, a la vez que abre las puertas a nuevos desarrollos que sigan maximizando la eficacia y eficiencia con respecto al diagnóstico, pronóstico y seguimiento personalizado de la condición de PCC con afectación cognitiva. Esta condición es altamente incapacitante por lo que también resaltamos la consecución de los impactos positivos desde el punto de vista de la ética social, buscando garantizar transparencia del modelo (en el caso de que se use para apoyo a la decisión clínica) y respetando en todo momento la necesidad de generar conocimiento inclusivo y con perspectiva de género. De hecho, si bien existen trabajos publicados al respecto [42], nuestro modelo objetivamente indica que la población femenina se encuentra en mayor riesgo frente a la PCC, por lo que es necesario validar clínicamente los hallazgos para transformarlos en herramientas de cribado reales, que se apliquen en todos los estamentos sanitarios, empezando por la atención primaria y que tengan en cuenta esta distinción poblacional. En este sentido, este punto refuerza trabajos científicos previos que intentan comprender de qué forma el sexo puede ser un aspecto diferencial en la PCC [43]. Se ha indicado que las mujeres son más susceptibles al desarrollo de esta condición por diversos motivos, entre los que se destacan:

* Diferencias en las características de los sistemas inmunitarios. En este sentido, las mujeres presentan respuestas inmunitarias más contundentes que las de los hombres, siendo estas respuestas protectoras pero, a la vez, problemáticas. Por ejemplo, las mujeres producen niveles más elevados de anticuerpos tipo IgG que puede llevar a mantener un cuadro prolongado de inflamación [44]. También los marcadores inflamatorios (como IL-6) se mantienen elevados en mujeres después de la infección primaria por COVID-19 [45].
* Factores hormonales propios del sexo femenino provocan una mayor susceptibilidad al desarrollo de PCC, en particular en edades reproductivas donde el organismo debe adaptarse a desarrollar el embarazo. Igualmente, los estrógenos pueden brindar una barrera inicial de protección pero pueden contribuir a una activación inmunitaria prolongada [46].
* Factores fisiológicos propios de las mujeres pueden coadyuvar al desarrollo de cuadros más graves considerando las sintomatología de la PCC. Por ejemplo, bajo esta condición, las mujeres sufren una disminución mayor de la capacidad pulmonar total [47].

Y aunque no esté en relación con aspectos de susceptibilidad biológica, los factores sociodemográficos influyen directamente en la alta prevalencia de PCC observada en mujeres. Factores tales como el nivel educativo o las características raciales, han demostrado su asociación en mujeres con dicha condición [48]. Estos aspectos también incluyen situaciones tales como su rol en los sistemas sanitarios, la presión social, la explotación de migrantes y las consecuencias económicas a largo plazo que sufren las mujeres tras la pandemia [49]. También, la recuperación de la actividad económica de las mujeres y la recuperación de sus empleos presenta una elevada dificultad debido a la PCC [50]. En definitiva, nuestro modelo apoya la existencia de patrones específicos relacionados con el sexo femenino y abre nuevas líneas de investigación para entender mejor las diferencias en la afectación neurocognitiva y cómo puede afectar esta situación de forma diferencial a aspectos laborales, sociales y personales a las mujeres. Comprender mejor estos patrones permitirá el diseño de intervenciones clínicas más personalizadas, como terapias adaptadas a factores de riesgo específicos o estrategias de seguimiento dirigidas, mejorando así la eficacia y equidad en el tratamiento de las pacientes.

En cuanto a la diversidad e inclusión, el trabajo ha velado por tener representación sociodemográfica de todos los estratos, comunidades y niveles sociales y educativos permitiendo enriquecer y comprender los matices que pueden revelarse al respetarse estos principios equitativos, siempre teniendo en cuenta los hallazgos previamente indicados con respecto al sexo femenino [42]. Finalmente, el desarrollo de un modelo breve, con una tecnología avanzada, pero fácilmente aplicable en ordenadores convencionales, permite asegurar un abordaje responsable frente a las desigualdades en salud.

En cuanto al trabajo futuro, consideramos que este trabajo puede sentar las bases para desarrollos de interés, que podrían incluir:

- Validación clínica del instrumento: incluyendo nuevas cohortes para evaluar la robustez del desarrollo intentando ampliar tanto como sea posible la representatividad de hombres. Idealmente, este trabajo debería haberse realizado con una muestra estratificada, sin embargo, la PCC afecta desproporcionadamente a mujeres (por las razones previamente indicadas) por lo que el número de participantes de sexo masculino, en este caso, no ha sido suficiente.

- Exploración de métodos más robustos: incluyendo desarrollos más amplios que se enfoquen, por ejemplo, en redes neuronales (o aprendizaje profundo – *Deep Learning* - , en general) que, si bien poseen una explicabilidad menor, pueden mejorar la sensibilidad obtenida por este trabajo. Por otra parte, considerando que se aplique a modo de cribado inicial o como continuación de la investigación para obtener conocimiento del rol de los factores neuropsicológicos en la PCC, la explicabilidad no debería representar un obstáculo insalvable. De hecho, podría brindar nuevas perspectivas, las cuales deberían ser analizadas y aplicadas con transparencia para apoyar la decisión clínica y respetar los derechos de las personas con PCC.

- Integración de otras áreas clínicas de interés, tales como estudios oftalmológicos (centrados, por ejemplo, en retina), posturográficos u otros biomarcadores que puedan ser fácilmente adquiribles en atención primaria para generar modelos más amplios, explicativos y/o personalizados, mientras se mantiene la facilidad de administración.

- Desarrollo de herramientas prácticas, aplicables a la población de interés o accesibles para profesionales de la salud, por ejemplo, mediante aplicaciones digitales autoadministrables, basadas en las variables observadas como clave y en el modelo desarrollado en este trabajo (o en otros modelos multidisciplinares que puedan desarrollarse como consecuencia del punto anterior).

En definitiva, este trabajo representa un paso hacia la mejora en el diagnóstico y comprensión de la COVID persistente cuando existe afectación cognitiva. Esta forma de presentación es particularmente incapacitante, afectando dominios clave para el desarrollo personal tanto a nivel laboral como social por ello, obtener un instrumento sencillo como el proporcionando, con el trasfondo científico, práctico y explicativo puesto de relieve en esta memoria, puede apoyar la implementación de la IA en entornos clínicos (en especial, en atención primaria, que es el primer punto de contacto con el sistema de salud por parte de quienes sufren esta condición) para facilitar decisiones informadas, transparentes y personalizadas. Además, los hallazgos sobre variables clave y patrones específicos ofrecen una guía para optimizar futuras estrategias de atención y seguimiento de pacientes poniendo el énfasis en la comprensión de factores neuropsicológicos. Creemos que estos resultados no solo validan la metodología empleada, sino que también abren nuevas posibilidades para la investigación y la aplicación clínica, ética y transparente, de la IA a un reto para la salud pública como lo es la COVID persistente.

# Glosario

**AUC (*Area Under the Curve*)**: métrica utilizada para evaluar el rendimiento de un modelo clasificatorio, representando la relación entre sensibilidad y especificidad. Un AUC cercano a 1 indica un modelo con alta capacidad discriminativa.

**COVID persistente (*post-COVID condition* -PCC-)**: condición caracterizada por la presencia de síntomas persistentes durante al menos 12 semanas después de una infección inicial por SARS-CoV-2, que no pueden explicarse por otro diagnóstico alternativo.

**Especificidad**: medida de rendimiento de un modelo predictivo que indica la proporción de verdaderos negativos correctamente identificados entre todos los casos negativos reales. Refleja la capacidad del modelo para evitar falsos positivos.

**Exactitud (*Accuracy*)**: proporción de predicciones correctas realizadas por un modelo en relación con el total de observaciones evaluadas.

**F1 (Puntuación F1)**: métrica que combina precisión y sensibilidad en una única media armónica, siendo especialmente útil para evaluar modelos en conjuntos de datos desbalanceados.

**Imputación de datos**: técnica utilizada para sustituir valores faltantes en un conjunto de datos con el objetivo de mantener la integridad analítica.

**Precisión (*Precision*)**: proporción de verdaderos positivos entre todas las observaciones clasificadas como positivas por un modelo.

**Reducción de dimensionalidad**: proceso de reducción del número de variables, optimizando la preservación de la mayor cantidad posible de información relevante. Una técnica típica es, por ejemplo, el Análisis de Componentes Principales (PCA).

**Regresión logística**: modelo estadístico utilizado para predecir la probabilidad de una variable categórica dependiente basada en una o más variables independientes. Mediante la transformación matemática adecuada, a pesar de ser un modelo de regresión, puede ser usado en clasificación binaria.

**ROC (*Receiver Operating Characteristic*)**: curva que representa la relación entre la sensibilidad y el valor (1-especificidad) de un modelo clasificatorio, y es utilizada para evaluar su rendimiento. Un modelo ideal representa gráficamente una “L invertida”, es decir, una curva más cercana al punto superior izquierdo del gráfico indica un mejor desempeño. A medida que el desempeño del modelo se ve afectado, la curva tiende a acercarse a la diagonal (que representaría un modelo que realiza una clasificación al azar).

**Sensibilidad**: proporción de casos positivos correctamente identificados por un modelo.

**Valores SHAP (*Shapley Additive Explanations*)**: técnica que permite la obtención de valores utilizados para interpretar modelos de aprendizaje automático. Estos valores cuantifican la contribución de cada variable a las predicciones específicas realizadas por el modelo.

**XGBoost (*Extreme Gradient Boosting*)**: algoritmo de aprendizaje automático basado en árboles de decisión, conocido por su capacidad de manejo eficiente de grandes conjuntos de datos, flexibilidad y rendimiento en tareas de clasificación y regresión.

# Bibliografía

[1] A. Fontanet, B. Autran, B. Lina, M. P. Kieny, S. S. A. Karim, y D. Sridhar, «SARS-CoV-2 variants and ending the COVID-19 pandemic», *The Lancet*, vol. 397, n.o 10278, pp. 952-954, mar. 2021, doi: 10.1016/S0140-6736(21)00370-6.

[2] «Long-term consequences of COVID-19: research needs - The Lancet Infectious Diseases». Accedido: 31 de diciembre de 2024. [En línea]. Disponible en: https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30701-5/fulltext#app-1

[3] S. Lopez-Leon *et al.*, «More than 50 long-term effects of COVID-19: a systematic review and meta-analysis», *Sci. Rep.*, vol. 11, n.o 1, p. 16144, ago. 2021, doi: 10.1038/s41598-021-95565-8.

[4] «Post COVID-19 condition». Accedido: 31 de diciembre de 2024. [En línea]. Disponible en: https://www.who.int/teams/health-care-readiness/post-covid-19-condition

[5] «Post COVID-19 condition (Long COVID)». Accedido: 26 de diciembre de 2024. [En línea]. Disponible en: https://www.who.int/europe/news-room/fact-sheets/item/post-covid-19-condition

[6] B. Khraisat, A. Toubasi, L. AlZoubi, T. Al-Sayegh, y A. Mansour, «Meta-analysis of prevalence: the psychological sequelae among COVID-19 survivors», *Int. J. Psychiatry Clin. Pract.*, vol. 26, n.o 3, pp. 234-243, sep. 2022, doi: 10.1080/13651501.2021.1993924.

[7] M. Peghin *et al.*, «Post-COVID-19 symptoms 6 months after acute infection among hospitalized and non-hospitalized patients», *Clin. Microbiol. Infect. Off. Publ. Eur. Soc. Clin. Microbiol. Infect. Dis.*, vol. 27, n.o 10, pp. 1507-1513, oct. 2021, doi: 10.1016/j.cmi.2021.05.033.

[8] R. Oliveira, F. D. Sotero, y T. Teodoro, «NeuroCOVID: critical review of neuropsychiatric manifestations of SARS-CoV-2 infection», *Ir. J. Med. Sci.*, vol. 190, n.o 2, pp. 851-852, may 2021, doi: 10.1007/s11845-020-02367-4.

[9] C. García-Sánchez *et al.*, «Neuropsychological deficits in patients with cognitive complaints after COVID-19», *Brain Behav.*, vol. 12, n.o 3, p. e2508, mar. 2022, doi: 10.1002/brb3.2508.

[10] S. Albu, N. R. Zozaya, N. Murillo, A. García-Molina, C. A. F. Chacón, y H. Kumru, «What’s going on following acute covid-19? Clinical characteristics of patients in an out-patient rehabilitation program», *NeuroRehabilitation*, vol. 48, n.o 4, pp. 469-480, 2021, doi: 10.3233/NRE-210025.

[11] R. Dacosta-Aguayo *et al.*, «Neurocognitive Profile of the Post-COVID Condition in Adults in Catalonia-A Mixed Method Prospective Cohort and Nested Case-Control Study: Study Protocol», *Vaccines*, vol. 10, n.o 6, p. 849, may 2022, doi: 10.3390/vaccines10060849.

[12] M. Boldrini, P. D. Canoll, y R. S. Klein, «How COVID-19 Affects the Brain», *JAMA Psychiatry*, vol. 78, n.o 6, pp. 682-683, jun. 2021, doi: 10.1001/jamapsychiatry.2021.0500.

[13] M. Carmona-Cervelló *et al.*, «Long COVID: cognitive, balance, and retina manifestations», *Front. Med.*, vol. 11, p. 1399145, 2024, doi: 10.3389/fmed.2024.1399145.

[14] «Impact of long COVID on health-related quality-of-life: an OpenSAFELY population cohort study using patient-reported outcome measures (OpenPROMPT) - The Lancet Regional Health – Europe». Accedido: 26 de diciembre de 2024. [En línea]. Disponible en: https://www.thelancet.com/journals/lanepe/article/PIIS2666-7762(24)00074-7/fulltext

[15] P. Schober y T. R. Vetter, «Logistic Regression in Medical Research», *Anesth. Analg.*, vol. 132, n.o 2, pp. 365-366, feb. 2021, doi: 10.1213/ANE.0000000000005247.

[16] F. Asadi, M. Rahimi, A. H. Daeechini, y A. Paghe, «The most efficient machine learning algorithms in stroke prediction: A systematic review», *Health Sci. Rep.*, vol. 7, n.o 10, p. e70062, oct. 2024, doi: 10.1002/hsr2.70062.

[17] Y. Nohara, K. Matsumoto, H. Soejima, y N. Nakashima, «Explanation of machine learning models using shapley additive explanation and application for real data in hospital», *Comput. Methods Programs Biomed.*, vol. 214, p. 106584, feb. 2022, doi: 10.1016/j.cmpb.2021.106584.

[18] «COVID-19 Map», Johns Hopkins Coronavirus Resource Center. Accedido: 26 de diciembre de 2024. [En línea]. Disponible en: https://coronavirus.jhu.edu/map.html

[19] M. Català *et al.*, «The effectiveness of COVID-19 vaccines to prevent long COVID symptoms: staggered cohort study of data from the UK, Spain, and Estonia», *Lancet Respir. Med.*, vol. 12, n.o 3, pp. 225-236, mar. 2024, doi: 10.1016/S2213-2600(23)00414-9.

[20] «Expanding our understanding of post COVID-19 condition: report of a WHO webinar - 9 February 2021». Accedido: 26 de diciembre de 2024. [En línea]. Disponible en: https://www.who.int/publications/i/item/9789240025035

[21] P. Poyatos *et al.*, «Endothelial dysfunction and cardiovascular risk in post-COVID-19 patients after 6- and 12-months SARS-CoV-2 infection», *Infection*, vol. 52, n.o 4, pp. 1269-1285, ago. 2024, doi: 10.1007/s15010-024-02173-5.

[22] R. N. Low, R. J. Low, y A. Akrami, «A review of cytokine-based pathophysiology of Long COVID symptoms», *Front. Med.*, vol. 10, p. 1011936, mar. 2023, doi: 10.3389/fmed.2023.1011936.

[23] M. A. F. Queiroz *et al.*, «Severe COVID-19 and long COVID are associated with high expression of STING, cGAS and IFN-α», *Sci. Rep.*, vol. 14, n.o 1, p. 4974, feb. 2024, doi: 10.1038/s41598-024-55696-0.

[24] B. Besteher *et al.*, «Cortical thickness alterations and systemic inflammation define long-COVID patients with cognitive impairment», *Brain. Behav. Immun.*, vol. 116, pp. 175-184, feb. 2024, doi: 10.1016/j.bbi.2023.11.028.

[25] F. Ceban *et al.*, «Fatigue and cognitive impairment in Post-COVID-19 Syndrome: A systematic review and meta-analysis», *Brain. Behav. Immun.*, vol. 101, pp. 93-135, mar. 2022, doi: 10.1016/j.bbi.2021.12.020.

[26] «Machine learning identifies long COVID patterns from electronic health records», *Nat. Med.*, vol. 29, n.o 1, pp. 47-48, 2023, doi: 10.1038/s41591-022-02130-5.

[27] V. A. Binson, S. Thomas, M. Subramoniam, J. Arun, S. Naveen, y S. Madhu, «A Review of Machine Learning Algorithms for Biomedical Applications», *Ann. Biomed. Eng.*, vol. 52, n.o 5, pp. 1159-1183, may 2024, doi: 10.1007/s10439-024-03459-3.

[28] P. P. Ariza-Colpas, M. A. Piñeres-Melo, M. A. Urina-Triana, E. Barceló-Martinez, C. Barceló-Castellanos, y F. Roman, «Machine Learning Applied to the Analysis of Prolonged COVID Symptoms: An Analytical Review», *Informatics*, vol. 11, n.o 3, Art. n.o 3, sep. 2024, doi: 10.3390/informatics11030048.

[29] E. Cordelli *et al.*, «Machine learning predicts pulmonary Long Covid sequelae using clinical data», *BMC Med. Inform. Decis. Mak.*, vol. 24, n.o 1, p. 359, nov. 2024, doi: 10.1186/s12911-024-02745-3.

[30] B. Antony *et al.*, «Predictive models of long COVID», *eBioMedicine*, vol. 96, oct. 2023, doi: 10.1016/j.ebiom.2023.104777.

[31] «XGBoost | Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining». Accedido: 26 de diciembre de 2024. [En línea]. Disponible en: https://dl.acm.org/doi/10.1145/2939672.2939785

[32] X. Guan *et al.*, «Construction of the XGBoost model for early lung cancer prediction based on metabolic indices», *BMC Med. Inform. Decis. Mak.*, vol. 23, n.o 1, p. 107, jun. 2023, doi: 10.1186/s12911-023-02171-x.

[33] A. Moore y M. Bell, «XGBoost, A Novel Explainable AI Technique, in the Prediction of Myocardial Infarction: A UK Biobank Cohort Study», *Clin. Med. Insights Cardiol.*, vol. 16, p. 11795468221133611, nov. 2022, doi: 10.1177/11795468221133611.

[34] W. Hong *et al.*, «A Comparison of XGBoost, Random Forest, and Nomograph for the Prediction of Disease Severity in Patients With COVID-19 Pneumonia: Implications of Cytokine and Immune Cell Profile», *Front. Cell. Infect. Microbiol.*, vol. 12, abr. 2022, doi: 10.3389/fcimb.2022.819267.

[35] T. Chen y C. Guestrin, «XGBoost: A Scalable Tree Boosting System», en *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, en KDD ’16. New York, NY, USA: Association for Computing Machinery, ago. 2016, pp. 785-794. doi: 10.1145/2939672.2939785.

[36] «DMwR 2nd Edtion». Accedido: 27 de diciembre de 2024. [En línea]. Disponible en: https://ltorgo.github.io/DMwR2/

[37] S. Lê, J. Josse, y F. Husson, «FactoMineR: An R Package for Multivariate Analysis», *J. Stat. Softw.*, vol. 25, pp. 1-18, mar. 2008, doi: 10.18637/jss.v025.i01.

[38] «Modern Applied Statistics with S, 4th ed». Accedido: 27 de diciembre de 2024. [En línea]. Disponible en: https://www.stats.ox.ac.uk/pub/MASS4/

[39] «R Companion 3E». Accedido: 27 de diciembre de 2024. [En línea]. Disponible en: https://www.john-fox.ca/Companion/

[40] T. Hothorn, F. Bretz, y P. Westfall, «Simultaneous inference in general parametric models», *Biom. J. Biom. Z.*, vol. 50, n.o 3, pp. 346-363, jun. 2008, doi: 10.1002/bimj.200810425.

[41] X. Robin *et al.*, «pROC: an open-source package for R and S+ to analyze and compare ROC curves», *BMC Bioinformatics*, vol. 12, n.o 1, p. 77, mar. 2011, doi: 10.1186/1471-2105-12-77.

[42] F. Bai *et al.*, «Female gender is associated with long COVID syndrome: a prospective cohort study», *Clin. Microbiol. Infect. Off. Publ. Eur. Soc. Clin. Microbiol. Infect. Dis.*, vol. 28, n.o 4, p. 611.e9-611.e16, abr. 2022, doi: 10.1016/j.cmi.2021.11.002.

[43] A. Rodríguez Onieva, C. A. Soto Castro, V. García Morales, M. Aneri Vacas, y A. Hidalgo Requena, «Long COVID: Factors influencing persistent symptoms and the impact of gender», *Semergen*, vol. 50, n.o 5, p. 102208, 2024, doi: 10.1016/j.semerg.2024.102208.

[44] F. Zeng *et al.*, «A comparison study of SARS‐CoV‐2 IgG antibody between male and female COVID‐19 patients: A possible reason underlying different outcome between sex», *J. Med. Virol.*, vol. 92, n.o 10, pp. 2050-2054, oct. 2020, doi: 10.1002/jmv.25989.

[45] J. Klein *et al.*, «Distinguishing features of Long COVID identified through immune profiling», 10 de agosto de 2022, *medRxiv*. doi: 10.1101/2022.08.09.22278592.

[46] B. Pollack *et al.*, «Female reproductive health impacts of Long COVID and associated illnesses including ME/CFS, POTS, and connective tissue disorders: a literature review», *Front. Rehabil. Sci.*, vol. 4, p. 1122673, abr. 2023, doi: 10.3389/fresc.2023.1122673.

[47] L. Spicuzza *et al.*, «Female Sex Affects Respiratory Function and Exercise Ability in Patients Recovered from COVID-19 Pneumonia», *J. Womens Health 2002*, vol. 32, n.o 1, pp. 18-23, ene. 2023, doi: 10.1089/jwh.2022.0104.

[48] J. Cohen y Y. van der Meulen Rodgers, «An intersectional analysis of long COVID prevalence», *Int. J. Equity Health*, vol. 22, p. 261, dic. 2023, doi: 10.1186/s12939-023-02072-5.

[49] W.-H. Chang, «Understanding the COVID-19 pandemic from a gender perspective», *Taiwan. J. Obstet. Gynecol.*, vol. 59, n.o 6, pp. 801-807, nov. 2020, doi: 10.1016/j.tjog.2020.09.004.

[50] M. Hossain y M. A. Hossain, «COVID-19 and Gender Gap in Labor Market Recovery: Evidence From Nigeria», *J. Afr. Econ.*, p. ejae012, sep. 2024, doi: 10.1093/jae/ejae012.

# Anexos

**ANEXO 1: BREVE DESCRIPCIÓN DE ALGUNOS DE LAS PRUEBAS NEUROPSICOLÓGICAS UTILIZADAS EN ESTE TRABAJO.**

**Digit Span Backward (WAIS-III)**: evalúa la memoria de trabajo al pedirle al sujeto repetir una serie de números en orden inverso al presentado.

**TMT B - A (Trail Making Test, tiempo diferencial)**: mide la capacidad de cambio mental y atención al calcular la diferencia de tiempo entre las partes B (alternar entre números y letras) y A (conectar números en orden).

**Phonetic Fluency (PMR)**: examina la capacidad de generar palabras que comiencen con una letra específica en un tiempo determinado, evaluando funciones ejecutivas y lenguaje.

**Semantic Fluency (Animals)**: mide la capacidad de generar palabras dentro de una categoría semántica específica (por ejemplo, animales), evaluando memoria semántica y fluidez verbal.

**Stroop Word-Colors (Interference)**: evalúa la inhibición cognitiva al pedir que se nombren los colores de palabras que representan nombres de colores impresos en un color diferente.

**Digit Span Forward (WAIS-III)**: mide la memoria inmediata al solicitar al sujeto repetir una serie de números en el orden presentado.

**SDMT (Symbol Digit Modalities Test)**: evalúa la velocidad de procesamiento cognitivo y atención al asociar símbolos con números siguiendo una clave.

**TMT A (Trail Making Test, tiempo)**: mide la velocidad psicomotora y atención al pedir al sujeto conectar números en orden secuencial lo más rápido posible.

**Symbol Search (WAIS-III)**: examina la velocidad de procesamiento y atención al identificar si un símbolo está presente en una lista.

**RAVLT (Rey Auditory Verbal Learning Test, summarize)**: evalúa la capacidad de aprendizaje y memoria verbal al presentar una lista de palabras que el sujeto debe recordar y repetir.

**RAVLT (Delayed Recall)**: mide la memoria a largo plazo pidiendo al sujeto recordar una lista de palabras previamente presentada después de un intervalo de tiempo.

**ROCF (Rey-Osterrieth Complex Figure, delayed recall)**: evalúa la memoria visual y la capacidad de organización al reproducir una figura compleja tras un intervalo de tiempo.

**ROCF (Copy Accuracy)**: mide habilidades visoespaciales y organización al reproducir una figura compleja inmediatamente después de observarla.

**BNT (Boston Naming Test)**: examina habilidades de denominación al pedir al sujeto nombrar objetos representados en imágenes.

**Vocabulary (WAIS-III)**: evalúa habilidades verbales y comprensión del lenguaje al pedir al sujeto definir palabras.