**DiagnosIA**

**Sistema de diagnóstico médico adaptativo mediante inteligencia artificial**

Imagen en blanco y negro de un reloj con números romanos

El contenido generado por IA puede ser incorrecto.



Autor: Andrés Lara Romero de Ávila

Tutor: Pedro Ignacio Dorado Diaz

Julio 2025

**DiagnosIA**

**Sistema de diagnóstico médico adaptativo mediante inteligencia artificial**

Imagen en blanco y negro de un reloj con números romanos

El contenido generado por IA puede ser incorrecto.



Autor: Andrés Lara Romero de Ávila

Tutor: Pedro Ignacio Dorado Diaz

Junio 2025

Logotipo

El contenido generado por IA puede ser incorrecto. Texto

El contenido generado por IA puede ser incorrecto.

Certificado del/los tutor/es TFG

D./Dña. Pedro Ignacio Dorado Diaz, profesor/a del Departamento de Estadística de la Universidad de Salamanca,

HACE/N CONSTAR:

Que el trabajo titulado “DiagnosIA: Sistema de diagnóstico médico adaptativo mediante inteligencia artificial “que se presenta, ha sido realizado por Andrés Lara romero de Ávila, con DNI 20616689H y

constituye la memoria del trabajo realizado para la superación de la asignatura Trabajo de Fin de Grado en Estadística en esta Universidad.

Salamanca,02 de Julio de 2025

Fdo.:

# **Agradecimientos**

A mi padre Andrés, sin el que no sería la persona que soy ahora mismo, y que gracias a su insistencia en mi carrera académica, en la mayoría de ocasiones he sido capaz de superar los retos académicos y personales.

A mi madre Goyi, la estrella que irradia felicidad en la casa y a la que le debemos todo. Gracias por estar siempre ahí aunque en muchas ocasiones no te valoremos como te mereces.

A mi hermana Ana, que aunque tenga un carácter un poco enérgico, es la mujer más trabajadora que conozco y que estoy seguro de que podrá conseguir todos los retos que se proponga.

También acordarme de toda mi familia, mis abuelos, tíos y primos, los cuales me han guiado en mi camino hasta la actualidad y espero que lo sigan haciendo mucho tiempo más. Gracias por tratarme con tanto cariño.

Concluyendo con una de las célebres frases de mi padre de entre las cientos que le habré escuchado:

***“Haz lo que esté en tu mano y deja de darle vueltas a lo demás”***

# **Resumen**

Este Trabajo Fin de Grado presenta **DiagnosIA**, un prototipo de sistema de ayuda al diagnóstico médico que, a partir de un breve cuestionario de síntomas binarios, estima la enfermedad más probable para aliviar la presión asistencial.

Se analizaron 4 962 historias clínicas anonimizadas con 132 síntomas y 41 enfermedades equilibradas. Tras la exploración y depuración inicial, se aplicaron técnicas de selección de variables (χ², información mutua, regresión logística, árboles y Random Forest) para extraer un subconjunto mínimo de síntomas informativos, reduciendo el cuestionario a unas veinte preguntas sin pérdida apreciable de rendimiento.

La estructura latente se visualizó con MCA, UMAP y Logistic PCA, revelando clústeres clínicamente coherentes.

En la fase predictiva se entrenaron Árboles de Decisión, Random Forest, XGBoost y un modelo bayesiano secuencial, evaluados con exactitud, sensibilidad, especificidad y el área bajo la curva ROC.

El modelo secuencial bayesiano fue el que mejores soluciones aportó, resultando un modelo adaptativo, que manejaba cada paciente de forma individualizada según el resultado de los síntomas.

Con un máximo de 20 preguntas, el modelo bayesiano secuencial alcanzó una precisión del 97,78 % formulando solo 6,7 preguntas de media, lo que avala su capacidad para priorizar dinámicamente los síntomas más informativos y reducir la carga para el paciente.

En conjunto, DiagnosIA demuestra que un cuestionario optimizado y modelos interpretables pueden proporcionar diagnósticos rápidos y fiables, constituyendo una base sólida para futuras herramientas de triaje digital en atención primaria.

# **Abstract**

This Final Year Project presents **DiagnosIA**, a prototype clinical decision-support system that predicts the most likely disease from a brief binary-symptom questionnaire to ease primary-care workload.

A corpus of 4 962 anonymised clinical records with 132 symptoms and 41 balanced diseases was explored and cleaned. Variable-selection techniques (χ² test, mutual information, logistic regression, decision trees and random forest) distilled an informative subset of about twenty symptoms without performance loss.

Latent structure visualisation through MCA, UMAP and Logistic PCA revealed clinically coherent clusters. For prediction, Decision Trees, Random Forest, XGBoost and a sequential Bayesian model were trained and assessed with accuracy, sensitivity, specificity and area under ROC curve.

The sequential Bayesian model delivered the best performance, adaptively tailoring questions to each patient. With a maximum of twenty questions, it achieved 97.78 % accuracy while asking only 6.7 questions on average, demonstrating its ability to prioritise the most informative symptoms and reduce patient burden.

Overall, DiagnosIA shows that an optimised questionnaire combined with interpretable models can provide fast, reliable diagnoses and forms a solid foundation for future digital triage tools in primary care.

**Índice de ilustraciones**

[Ilustración 1: Dendograma del clustering jerárquico de los 20 síntomas más significativos 15](#_Toc202302985)

[Ilustración 2: Dendograma del clúster jerárquico de enfermedades usando solo los 20 síntomas más significativos 16](#_Toc202302986)

[Ilustración 3: Biplot de enfermedades agrupadas 19](#_Toc202302987)

[Ilustración 4: UMAP usando los 20 síntomas más significativos 22](#_Toc202302988)

[Ilustración 5: Resultado Análisis de componentes principales logístico para los 20 síntomas más significativos 24](#_Toc202302989)

[Ilustración 6: Proceso de construcción de un árbol de decisión. Navlani, Avinash. (2024) Tutorial de clasificación de árbol de decisión en Python. recuperado de https://www.datacamp.com/es/tutorial/decision-tree-classification-python 29](#_Toc202302990)

[Ilustración 7: Grafico con la relación entre el valor de la precisión, el número de síntomas y la profundidad del árbol 32](#_Toc202302991)

[Ilustración 8: Relación entre la profundidad máxima del gráfico y el valor de precisión 33](#_Toc202302992)

[Ilustración 9: Gráfico comparando el número máximo de preguntas, precisión y número medio de preguntas obtenido 41](#_Toc202302993)

[Ilustración 10: Gráfico comparando el umbral de confianza con el nivel de precisión y la proporción de preguntas utilizadas 42](#_Toc202302994)

[Ilustración 11: Resultado de la simulación de diagnóstico para un paciente 45](#_Toc202302995)

[Ilustración 12: Dendograma del clúster jerárquico de enfermedades usando los 132 síntomas 63](#_Toc202302996)

[Ilustración 13: MCA por enfermedades usando 20 síntomas sin agrupar 65](#_Toc202302997)

[Ilustración 14: MCA por enfermedades usando todos los síntomas sin agrupar 66](#_Toc202302998)

[Ilustración 15: Biplot conjunto de enfermedades agrupadas y síntomas resultado del MCA 67](#_Toc202302999)

[Ilustración 16: MCA de enfermedades agrupadas con círculos delimitando el área de cada grupo 68](#_Toc202303000)

[Ilustración 17: UMAP usando todos los síntomas del conjunto de datos original 69](#_Toc202303001)

[Ilustración 18: Resultado logistic PCA usando todos los síntomas del dataset 69](#_Toc202303002)

[Ilustración 19: Precisión del modelo según la profundidad escogida usando solo 20 síntomas 70](#_Toc202303003)

[Ilustración 20: Resultados del árbol de decisión usando todos los síntomas sin profundidad máxima 70](#_Toc202303004)

[Ilustración 21: Resultado de la aplicación del Random Forest usando los 20 síntomas más significativos y con 200 estimadores 71](#_Toc202303005)

[Ilustración 22: Resultado de Random Forest usando todos los síntomas 72](#_Toc202303006)

[Ilustración 23: Resultado del algoritmo XGBoost para la ejecución con los 20 síntomas más representativos 73](#_Toc202303007)

**Índice de ecuaciones**

[Ecuación 1:Ecuación de la información mutua entre dos sucesos X e Y 10](#_Toc202303010)

[Ecuación 2: Función de la regresión lineal 11](#_Toc202303011)

[Ecuación 3: Función sigmoide 11](#_Toc202303012)

[Ecuación 4: Fórmula de la regresión logística 11](#_Toc202303013)

[Ecuación 5: Fórmula regresión logística multinomial 11](#_Toc202303014)

[Ecuación 6: Fórmula del coeficiente de correlación phi 14](#_Toc202303015)

[Ecuación 7: Fórmula alternativa para calcular el coeficiente de correlación phi 14](#_Toc202303016)

[Ecuación 8: Coeficiente de similitud de Jaccard 15](#_Toc202303017)

[Ecuación 9: Distancia de Jaccard 15](#_Toc202303018)

[Ecuación 10: Fórmula del método de la media para clusterización (average linkage) 16](#_Toc202303019)

[Ecuación 11: Probabilidad de conexión entre dos puntos usando la distancia de Jaccard 20](#_Toc202303020)

[Ecuación 12: Proyección en el espacio de baja dimensión UMAP 21](#_Toc202303021)

[Ecuación 13: Función a minimizar entre los dos grafos producidos en la técnica UMAP 21](#_Toc202303022)

[Ecuación 14: Modelo Latente de Logs-Odds 22](#_Toc202303023)

[Ecuación 15: Función logística 23](#_Toc202303024)

[Ecuación 16: Entropía del dataset (D) 29](#_Toc202303025)

[Ecuación 17: Entropía condicionada tras la partición del atributo A 30](#_Toc202303026)

[Ecuación 18: Ganancia de información de A 30](#_Toc202303027)

[Ecuación 19: Índice de impurezas de Gini 30](#_Toc202303028)

[Ecuación 20:Impureza media ponderada tras la partición 30](#_Toc202303029)

[Ecuación 21: Ganancia de Gini 30](#_Toc202303030)

[Ecuación 22: Fórmula de la probabilidad condicionada 39](#_Toc202303031)

[Ecuación 23: Probabilidad de la intersección 39](#_Toc202303032)

[Ecuación 24: Probabilidad de tener el conjunto de síntomas x 39](#_Toc202303033)

[Ecuación 25: Fórmula final del algoritmo de clasificación de bayes 39](#_Toc202303034)

**Índice de tablas**

[Tabla 1: Tabla de síntomas 56](#_Toc202303035)

[Tabla 2: Tabla de enfermedades 57](#_Toc202303036)

[Tabla 3: Relación entre el número de grupos de individuos y los individuos repetidos por enfermedad 59](#_Toc202303037)

[Tabla 4: Síntomas más frecuentes 59](#_Toc202303038)

[Tabla 5: Síntomas menos frecuentes 59](#_Toc202303039)

[Tabla 6: Tabla con los 20 síntomas más representativos frente a las enfermedades 60](#_Toc202303040)

[Tabla 7: Resultado de importancia de síntomas aplicando la regresión logística de nivel 1 61](#_Toc202303041)

[Tabla 8: Síntomas más relevantes según el árbol de decisión 61](#_Toc202303042)

[Tabla 9: Tabla con los 20 síntomas más significativos según la técnica del Random Forest 62](#_Toc202303043)

[Tabla 10: Tabla conjunta de todas las técnicas de selección de variables con los respectivos síntomas por orden de importancia 62](#_Toc202303044)

[Tabla 11: Tabla generada para calcular el coeficiente de correlación phi 63](#_Toc202303045)

[Tabla 12: Coeficiente de correlación phi entre pares de síntomas 64](#_Toc202303046)

[Tabla 13: Síntomas con mayor contribución a la componente 1 del MCA 64](#_Toc202303047)

[Tabla 14: Síntomas con mayor contribución a la componente 2 del MCA 65](#_Toc202303048)

# **Introducción**

La profusión de información médica disponible en la actualidad es, a la vez, un privilegio y un quebradero de cabeza. Cada jornada se suman nuevas bases de datos, artículos científicos, guías clínicas y foros de pacientes que inundan la red con descripciones de patologías, fármacos y procedimientos. Ese caudal nutre a los profesionales cuando necesitan evidencias frescas, pero también los expone a una sobrecarga cognitiva que entorpece la consulta diaria. En este contexto se sitúa **DiagnosIA**, un proyecto concebido para transformar el exceso de datos en valor clínico tangible mediante un modelo predictivo adaptativo que actúa como primer filtro diagnóstico.

Desde los primeros sistemas expertos en medicina, como ***MYCIN***, se ha reconocido que los datos clínicos son incompletos y a menudo ambiguos. *Shortliffe* y *Buchanan* (Shortliffe & Buchanan, 1975) fueron pioneros en formalizar un esquema de “razonamiento impreciso” mediante reglas de producción con factores de certeza, demostrando cómo propagar y combinar grados de confianza para apoyar el diagnóstico cuando la información es parcial o contradictoria. Esta línea temprana de trabajo subraya la importancia de incorporar explícitamente la incertidumbre en los sistemas de ayuda a la decisión clínica.

El punto de partida es sencillo: un paciente se enfrenta a un malestar y, antes de ocupar una cita presencial, escasa y muy demanda, responde a un breve cuestionario digital. A diferencia de los sistemas tradicionales, que proponen siempre la misma batería de preguntas, **DiagnosIA** emplea algoritmos de aprendizaje secuencial para elegir, en tiempo real, cuál es la siguiente cuestión más informativa. De ese modo evita los interrogatorios interminables y centra el diálogo en síntomas relevantes, logrando que la interacción sea ligera para el usuario y, al mismo tiempo, altamente discriminante para el sistema.

La flexibilidad del modelo es su primer rasgo distintivo. Cada paciente posee un historial, unos antecedentes y un contexto vital que condicionan la expresión de su dolencia; **DiagnosIA** integra esa variabilidad adaptando el itinerario de preguntas sobre la marcha. Si las respuestas iniciales apuntan a un cuadro respiratorio, la ruta se especializa en signos neumológicos; si sugieren un proceso digestivo, la secuencia pivota hacia ese dominio. Así, el sistema imita el razonamiento clínico dinámico que un médico ejercería en su consulta, pero lo hace apoyándose en una base estadística rigurosa que optimiza la información recopilada.

El segundo pilar es la eficacia. Al seleccionar únicamente los síntomas con mayor poder de discriminación, e iniciar la búsqueda por los de prevalencia e impacto combinados, el esquema secuencial reduce la necesidad de pruebas complementarias y descarta desde el principio diagnósticos muy improbables. Incluso cuando el cuadro resulta complejo, el modelo recurre a su estructura probabilística para ordenar las hipótesis de modo que las menos plausibles se descarten pronto, minimizando la duración total del cuestionario. Esta estrategia repercute en una reducción directa de la carga asistencial: menos tiempo de consulta para los facultativos, menos horas de espera para los pacientes y, en consecuencia, un sistema de salud más ágil.

En tercer lugar se encuentra la transparencia. Frente al recelo que a veces suscitan los algoritmos de caja negra, **DiagnosIA** incorpora un mecanismo de interpretación paso a paso basado en probabilidades bayesianas actualizadas con cada respuesta. El profesional puede visualizar por qué una pregunta fue considerada necesaria, cómo se ajustó la probabilidad de cada patología y qué factores inclinaron la balanza hacia el diagnóstico propuesto. Esta claridad favorece la confianza del clínico, facilita la validación y abre la puerta a la auditoría externa y a la mejora continua del modelo.

La génesis del proyecto pasó por un filtrado previo de la información, imprescindible cuando se dispone de un repositorio masivo y heterogéneo. Se recabaron historias clínicas anonimizadas, se depuraron registros inconsistentes y se aplicaron técnicas de selección de variables, pruebas de asociación, métodos de importancia en árboles de decisión y aproximaciones de información mutua, para aislar el conjunto mínimo de síntomas necesarios sin merma en la capacidad predictiva. Ese núcleo compacto aligera el modelo, rebaja la demanda computacional y reduce la susceptibilidad al sobreajuste, todo ello sin sacrificar fiabilidad.

Una vez destilado el vocabulario de síntomas, se optó por un diseño adaptativo de “inspiración bayesiana”. En lugar de entrenar un único clasificador determinista, se construyó un sistema que actualiza distribuciones de probabilidad en tiempo real. Cada nueva respuesta modula las creencias sobre la enfermedad subyacente y recalcula la utilidad esperada de las preguntas pendientes; la que mayor valor informativo aporta se formula a continuación. Este bucle se repite hasta que la incertidumbre cae por debajo de un umbral de confianza predefinido o se alcanza un límite máximo de ítems, tras lo cual el sistema emite su diagnóstico. El resultado es un engranaje flexible que aprende de cada interacción y ajusta sus rutas de interrogación a los matices de cada caso.

En la actualidad, la medicina de alto rendimiento exige la convergencia entre la experiencia clínica y las capacidades de la inteligencia artificial ***(Topol, 2019)***. Sin embargo, este enfoque adaptativo no pretende desbancar al médico, sino complementarlo. El algoritmo ofrece una primera hipótesis respaldada por estadísticas y justificable mediante métricas de certeza; a continuación, el facultativo revisa esa propuesta, la valida o la corrige y, en los escenarios más complejos, decide qué pruebas adicionales son pertinentes. De este modo, la inteligencia humana y la artificial cooperan: la máquina filtra el ruido y entrega un resumen probabilístico, mientras que el profesional aplica su juicio clínico y su experiencia contextual para la decisión definitiva.

En el largo plazo, la implantación de herramientas como **DiagnosIA** promete múltiples beneficios. Al derivar sólo los casos necesarios a especialistas y filtrar consultas de baja complejidad, libera recursos para patologías graves y contribuye a un sistema sanitario más sostenible. Para el paciente, implica obtener una orientación fiable en minutos, sin desplazamientos innecesarios y con un lenguaje adaptado a su nivel de comprensión. Para el profesional, supone disponer de un compañero digital que preclasifica la información y le permite concentrarse en el juicio clínico y la relación humana.

En suma, **DiagnosIA** no es únicamente un algoritmo que asigna diagnósticos; es un marco adaptativo, flexible y eficientemente diseñado para gestionar la sobreabundancia de datos médicos y potenciar la toma de decisiones en salud. Al aunar bioestadística rigurosa, aprendizaje secuencial y transparencia operativa, demuestra que la inteligencia artificial puede integrarse de forma ética y productiva en la práctica clínica, ofreciendo una herramienta de triaje capaz de evolucionar con la medicina y centrada en la dignidad del paciente.

# **Estructura del proyecto**

El trabajo avanza en tres etapas enlazadas que conducen, paso a paso, a la construcción de un “motor diagnóstico adaptativo”. Todo comienza con una exploración preliminar de los datos: se limpia, se detectan valores faltantes y se ofrecen al lector solo las visualizaciones indispensables (frecuencias de síntomas, balance de enfermedades, mapa de ausencias) que justifican las decisiones de depuración; los cuadros numéricos extensos y las tablas auxiliares permanecen en el Anexo (***ANEXO***) para no entorpecer la narrativa. Sobre esa base se inicia la segunda fase, en la que convergen la selección de variables y una reducción dimensional puramente ilustrativa. El objetivo es identificar el subconjunto mínimo de síntomas que conserva la capacidad diagnóstica sin transformar las variables originales; por ello, el texto principal resume la comparativa entre métodos (**χ²**, **información mutua, regresión logística**, **árboles**, **Random Forest**) y expone la lista final de síntomas.

La tercera y más importante etapa concentra el esfuerzo en el modelo predictivo. Aquí se describe, con la profundidad justa, el funcionamiento de un algoritmo bayesiano secuencial que actualiza probabilidades tras cada respuesta y elige la pregunta siguiente según el valor informativo esperado. La exposición destaca tres fortalezas: la flexibilidad para adaptar la entrevista a cada paciente, la eficiencia lograda al reducir el número de preguntas y la transparencia que habilita la supervisión clínica. Se añade un breve contraste con modelos no adaptativos para evidenciar la ganancia en interacción y se ilustran dos rutas de preguntas divergentes a modo de ejemplo. Las **curvas ROC** por clase, los ajustes finos de umbrales y el resto de métricas detalladas se consignan en el Anexo (***ANEXO***), de modo que el lector interesado pueda rastrear la evidencia sin que la sección principal pierda fluidez.

Este itinerario, de la limpieza esencial al filtrado inteligente de síntomas y, finalmente, a un cuestionario que se reconfigura en tiempo real, muestra cómo la estadística y el aprendizaje secuencial se combinan para ofrecer un diagnóstico preliminar ágil, verificable y, sobre todo, adaptativo.

# **Herramientas utilizadas**

Para este proyecto se ha optado por la dupla **Python** y **Jupyter Notebooks**, una elección que equilibra simplicidad de uso, potencia técnica y amplia aceptación tanto en la industria como en la academia.

**Python** destaca por una sintaxis intuitiva y una comunidad masiva que mantiene un ecosistema de bibliotecas decisivo para ciencia de datos y aprendizaje automático: Pandas y NumPy para manipulación eficiente de datos, Matplotlib y Seaborn para visualización, y marcos de ML como Scikit-Learn, TensorFlow y PyTorch. Esa combinación cubre todo el flujo de trabajo, desde la depuración inicial de los datos hasta la puesta en producción de modelos complejos. Frente a ***R***, muy valorado en estadística tradicional pero menos extendido en ingeniería de sistemas o despliegues de IA, Python ofrece un lenguaje único para tareas exploratorias, backend y automatización. En comparación con **MATLAB**, sobresale por ser gratuito y de código abierto, lo que facilita la colaboración y evita restricciones de licencia, algo crucial cuando el proyecto puede crecer o migrar a infraestructuras en la nube.

Para el entorno de trabajo se escogió **Jupyter Notebooks** porque integra código ejecutable, texto explicativo y gráficos en un único documento interactivo. Esta característica permite iterar rápidamente: se prueba un bloque de código, se analizan los resultados y se añade la interpretación justo debajo. Herramientas como **VS Code** o **PyCharm** son excelentes entornos de desarrollo generales, pero requieren más configuración (entornos virtuales, extensiones, depuración paso a paso) y no ofrecen, de forma nativa, la narrativa computacional que hace de Jupyter el estándar de facto en notebooks científicos. Además, la compatibilidad con plataformas en la nube (p. ej., Google Colab) permite ejecutar experimentos sin una estación de trabajo potente ni instalaciones locales complejas, algo que reduce la fricción cuando se colabora con otros investigadores o se comparte el cuaderno para revisión.

En conjunto, Python aporta la versatilidad y el ecosistema; Jupyter, la interactividad y la documentación viva. Esta combinación facilita un flujo de trabajo reproducible, colaborativo y escalable, perfectamente alineado con las exigencias de un proyecto que abarca desde la exploración de datos hasta la implementación de modelos de aprendizaje automático avanzados.

Para facilitar el entendimiento con el nombre de las variables, el cual está en inglés, encontramos un apartado del anexo donde aparece un diccionario explicando cada síntoma (***Tabla 4: Tabla de síntomas***). Lo mismo ocurre con las enfermedades que vamos a estudiar y que aparecen en la última variable de la base de datos denominada *“prognosis”* (**pronóstico**)(**Tabla 5: Tabla de enfermedades**)*.*

# **Análisis Exploratorio de Datos (EDA)**

## **Introducción**

El **Análisis Exploratorio de Datos** (EDA, por sus siglas en inglés) constituye una fase esencial en cualquier estudio estadístico, ya que permite examinar la estructura de los datos, detectar patrones y relaciones, identificar valores atípicos y corregir inconsistencias. Como su nombre indica, el **EDA** se centra en la exploración inicial del conjunto de datos para comprender sus propiedades antes de aplicar técnicas de modelado predictivo. Este análisis proporciona una visión general del dataset y facilita la toma de decisiones sobre la limpieza y transformación de los datos, asegurando su calidad para las etapas posteriores del estudio.

La sección de análisis exploratorio de datos se inspira en la filosofía de ***(«Exploratory Data Analysis», 2008***), para quien “mirar antes de modelar” es fundamental. Siguiendo sus recomendaciones, empleamos diagramas de dispersión, histogramas y otras visualizaciones junto con estudios de distribución, para revelar estructuras, detectar valores atípicos y comprender las relaciones entre variables antes de proceder a cualquier técnica de selección o modelado.

En este estudio se analiza un conjunto de datos compuesto por **132 variables explicativas** dicotómicas y **una variable respuesta** categórica con **41 categorías** distintas, cada una representando una enfermedad diagnosticada. Las variables explicativas representan síntomas presentes en individuos afectados por diferentes enfermedades, mientras que la variable respuesta identifica la patología diagnosticada. El dataset contiene un total de **4.962 observaciones**, donde cada fila representa a un paciente y cada columna indica la presencia o ausencia de un síntoma mediante valores binarios (0: ausencia del síntoma, 1: presencia del síntoma). La variable respuesta, en cambio, es de tipo categórico y abarca diversas enfermedades, incluyendo infecciones fúngicas, acné y psoriasis, entre otras. Cabe destacar que la mayoría de patologías no están relacionadas entre sí, ya que afectan a distintas partes del cuerpo y pertenecen a diversas especialidades médicas.

Dado el tamaño y la estructura de este conjunto de datos, el **EDA** se enfrenta a varios retos, entre los que destacan:

* **Identificación de la distribución de los síntomas:** Determinar cuáles son los síntomas más frecuentes y su posible relación con grupos específicos de enfermedades.
* **Estudio de correlaciones:** Evaluar la posible redundancia entre variables, lo que puede impactar la eficacia del modelo predictivo.
* **Visualización de datos:** Implementar herramientas gráficas para una mejor interpretación de la información disponible.
* **Evaluación del desbalance de clases:** Analizar si algunas enfermedades están sobrerrepresentadas en comparación con otras, lo que podría afectar el rendimiento de los modelos predictivos.

Para abordar estos desafíos, el análisis exploratorio se dividirá en dos fases principales. En primer lugar, se llevará a cabo un proceso de limpieza de datos, en el cual se identificará la existencia de valores faltantes, se eliminarán duplicados y se verificarán inconsistencias en las variables. Posteriormente, se procederá a la exploración visual y estadística del dataset, empleando diversas técnicas que permitan extraer información clave sobre la distribución de los datos y las relaciones entre las variables.

## **Primeros pasos**

El análisis comenzó con la carga e inspección del conjunto de datos, compuesto por 4962 observaciones (pacientes) y 132 variables explicativas (síntomas), junto a una variable respuesta categórica que representa la enfermedad diagnosticada, con un total de 41 clases distintas. Todas las variables explicativas son de tipo dicotómico, codificadas como 0 (síntoma ausente) y 1 (síntoma presente), y se ha verificado que no existen valores nulos ni inconsistencias en el dataset. Cada fila del dataset representa a un paciente, identificado por el conjunto de síntomas que presenta. Dado que muchos pacientes comparten una misma enfermedad y presentan síntomas similares o idénticos, es esperable encontrar registros duplicados. En concreto, se identificaron **4657 filas duplicadas**, lo que representa el 93,39 % del total. Esta repetición de tuplas no se considera un problema, sino una característica natural del conjunto, que refleja la presencia de perfiles clínicos recurrentes. Lejos de eliminarse, estos duplicados son útiles para el modelo, ya que aportan robustez al proceso de aprendizaje.

Agrupando pacientes que comparten los mismos síntomas y la misma enfermedad, se identificaron 304 combinaciones únicas, distribuidas entre las 41 enfermedades. Esto implica que cada patología puede expresarse a través de varios perfiles sintomáticos distintos. Esta diversidad intraclase es especialmente valiosa, ya que permite construir modelos que no dependan de combinaciones rígidas de síntomas y que puedan generalizar mejor ante nuevas observaciones.

## **Cálculo de frecuencias**

Una parte clave del EDA es el estudio de la frecuencia de aparición de los síntomas (columnas) y la distribución de las enfermedades (categorías de la variable respuesta). En particular, se analizaron:

* **Ratio de pacientes por grupo sintomático único**

Se calculó el ratio entre el número total de pacientes duplicados por enfermedad y el número de combinaciones únicas de síntomas dentro de esa enfermedad o grupos de duplicados de cada enfermedad (***Tabla 3***). Este valor osciló entre 11,10 y 23,20, lo que significa que cada grupo sintomático se repite, como mínimo, en 11 pacientes dentro de cada enfermedad. Este resultado es muy prometedor, ya que indica que existen perfiles sintomáticos recurrentes y representativos dentro de cada patología. Un ratio alto implica menor variabilidad dentro de la clase, es decir, que la mayoría de los pacientes duplicados pertenecen al mismo grupo de síntomas. Un ratio bajo implica mayor diversidad, dicho de otro modo, que existe un mayor número de grupos de individuos que presentan una enfermedad, lo que aumenta la flexibilidad del modelo sin comprometer su capacidad para aprender.

* **Número de grupos únicos por enfermedad**

El número de grupos únicos de pacientes por enfermedad varía entre 5 y 10 por enfermedad. Las enfermedades con mayor número de perfiles distintos son el dengue, la hepatitis D, la migraña y la varicela. Esta variabilidad garantiza que el modelo pueda captar distintas manifestaciones clínicas de una misma enfermedad, aumentando su capacidad de generalización. Véase ***Tabla 3***.

* **Frecuencia de aparición de síntomas (por columnas)**

Estudiando la frecuencia de aparición de los síntomas en la base de datos, se observó que algunas variables tienen una presencia muy elevada, como fatiga, vómitos, fiebre alta, pérdida de apetito, náuseas y dolor de cabeza (***Tabla 4***). Estos síntomas, al ser comunes a múltiples enfermedades, teóricamente tienen un bajo poder discriminativo. Por el contrario, los síntomas menos frecuentes, como hipervolemia, mal olor en la orina, debilidad unilateral, sensación de vértigo o lagrimeo excesivo (***Tabla 5***), tienden a estar asociados a algunas enfermedades, pero de cara al modelo final, si lo que queremos es reducir el número de síntomas, no serán la mejor opción para incluir en el modelo.

Una prueba concreta ilustra este fenómeno: El síntoma “*foul\_smell\_of\_urine*” (mal olor en la orina), el menos frecuente en el dataset, aparece únicamente en pacientes diagnosticados con infección del tracto urinario, lo que indica una asociación directa. En este punto del proyecto no podemos conocer si las variables menos frecuentes son las que nos aportarán mayor información a la hora de clasificar a los pacientes. A pesar de las hipótesis o conclusiones que podamos sacar para fases futuras, saber qué variables están relacionada de manera unívoca con una enfermedad nos ayuda a conocer mejor la base de datos.

* **Frecuencia de aparición de enfermedades**

El número de ocurrencias para todas las enfermedades es el mismo. Cada enfermedad la presentan 121 individuos, menos la infección fúngica que la presentan 122 individuos. Esto puede ser normal si el encargado de recopilar la información quería obtener el mismo número de pacientes para cada enfermedad.

## **Selección de variables**

El conjunto de datos que se utiliza en este proyecto es demasiado grande para trabajar y visualizar datos de forma ordenada, sobre todo dado el gran número de variables que se tienen. Es por ello, que se hace necesaria la labor de recopilar las variables que más información aportan.

Para seleccionar las variables más importantes del conjunto se emplearán una serie de técnicas, que retornarán las variables por orden de importancia. Finalmente se estudiarán los resultados de todas ellas y se seleccionarán aquellas variables que aparecen como significativas en la mayoría de las técnicas.

Para asegurar la validez de los resultados y evitar el posible filtrado de información que se produce al realizar pruebas usando el conjunto de datos completo, el dataset se ha dividido en:

* **Conjunto de entrenamiento** (70% de las observaciones), sobre el que se realiza todo el EDA, la selección de variables (Chi-Cuadrado, información mutua, coeficiente phi, etc.) y el ajuste de los modelos.
* **Conjunto de pruebas** (30%restante**)**, que permanece completamente aislado hasta la fase final de evaluación.

Esta estrategia garantiza que la elección de los síntomas más informativos (y el posterior cálculo del **coeficiente φ**) se base únicamente en información disponible en el entrenamiento, sin interferir con la información de la prueba para la posterior construcción de los modelos de predicción.

### **Prueba chi-cuadrado + Información Mutua**

En primer lugar se evalúa la asociación entre síntomas y patologías. Para cada síntoma del conjunto de datos se construye una tabla de contingencia 2×41, ausencia del síntoma frente a las 41 enfermedades, que recoge las frecuencias observadas.

Sobre cada tabla se aplica la pruebade independencia, contrastando las frecuencias empíricas con las esperadas bajo la hipótesis nula de no asociación. Un valor de insuficientemente pequeño indica que la distribución de enfermedades varía en función de la presencia del síntoma, de modo que el síntoma resulta informativo para el diagnóstico. Para garantizar la validez del contraste cuando existen numerosas celdas con frecuencia esperada inferior a cinco, se emplea la corrección por simulación Monte Carlo, que proporciona estimaciones más fiables del **-valor**.

Complementariamente, se calcula la I**nformación Mutua (MI)** de cada síntoma respecto al diagnóstico. (Peng et al., 2005) este trabajo sustenta la fase de selección de síntomas mediante información mutua, pues demuestra cómo combinar de forma óptima relevancia diagnóstica y diversidad sintomática para reducir la dimensionalidad.

Esta medida cuantifica la reducción de la incertidumbre sobre la enfermedad al conocer el estado del síntoma: **MI** igual a cero implica independencia absoluta, mientras que valores positivos crecientes reflejan un aporte informativo cada vez mayor. En términos probabilísticos, la **MI** capta cualquier desviación entre la distribución conjunta de síntoma y enfermedad y el producto de sus marginales, señalando por tanto la dependencia entre ambos eventos.

La combinación del con la información mutua permite jerarquizar los síntomas según su relevancia diagnóstica y fundamenta los pasos posteriores de selección de variables.

Ecuación 1:Ecuación de la información mutua entre dos sucesos X e Y

En la ***Ecuación 1*** (***Capítulo+3.pdf, s. f***.) podemos comprobar la fórmula de la información mutua, la cual se basa en la probabilidad conjunta de ambos sucesos multiplicado por el logaritmo entre la probabilidad del suceso x condicionado por y (), entre la probabilidad del suceso x ().

En la ***Tabla 6*,** podemos comprobar el resultado del estudio del análisis entre síntomas y enfermedades, mostrándose solamente los 20 más representativos. Vemos que entre los síntomas mostrados destacan la fatiga, los vómitos y la fiebre alta. Esto nos indica que, conociendo la respuesta a la pregunta de si el paciente presenta alguno de estos síntomas, seremos capaces de reducir el número de enfermedades posibles en el diagnóstico. Estos resultados son altamente esperanzadores para la construcción del modelo de predicción, puesto que, usando estos síntomas, la predicción de la enfermedad será de mayor exactitud.

### **Regresión Logística**

La regresión logística constituye uno de los métodos de referencia en los modelos de clasificación de **Machine Learning**, especialmente cuando la variable respuesta es categórica. Para problemas con únicamente dos categorías se emplea la versión binaria, mientras que la regresión logística multinomial extiende el procedimiento a variables dependientes con más de dos clases. Durante la fase de selección de variables, el diseño e implementación del modelo de regresión logística se llevaron a cabo siguiendo los lineamientos metodológicos establecidos por (Jr et al., 2013), aprovechando sus recomendaciones sobre codificación de covariables, criterios de inclusión basados en razón de verosimilitud y estrategias de validación.

Las variables explicativas, también llamadas covariables, pueden ser cuantitativas o cualitativas, pero si estas son cualitativas, deben tomar valores 0 para su ausencia y 1 para su presencia. Si alguna de estas covariables tuviera más de dos categorías, debería hacerse una transformación en varias variables ficticias (***variables dummy***).

Los modelos de regresión logística persiguen dos finalidades:

1. **Estimación de efectos** : cuantificar, a través de los coeficientes y sus errores estándar, la contribución individual de cada covariable a la probabilidad de pertenecer a una categoría determinada.
2. **Clasificación de individuos** : asignar cada observación a la clase cuya probabilidad condicional es máxima dado su vector de características.

La regresión logística es un caso especial de una regresión lineal donde la variable objetivo es categórica. Trabaja con predicciones de la probabilidad de ocurrencia de casos binarios utilizando una función logit.

Ecuación 2: Función de la regresión lineal

En la ***Ecuación 2*** encontramos la fórmula de la regresión lineal donde la variable dependiente es “y”, y todas las demás variables son las variables explicativas.

Ecuación 3: Función sigmoide

La ***Ecuación 3*** muestra la forma de la función sigmoide, que si aplicamos a la ***Ecuación 2*** obtenemos la fórmula de la regresión logística.

Ecuación 4: Fórmula de la regresión logística

Cuando la variable dependiente tiene más de dos categorías existen dos enfoques principales:

* **One-vs-Rest:** Se ajustan K (siendo k el número de categorías de la variable dependiente) regresiones binarias, cada una distinguiendo clase k frente al resto.
* **Multinomial (softmax):** Se modela directamente con una verosimilitud basada en la categoría correcta para cada ejemplo.

Ecuación 5: Fórmula regresión logística multinomial

Esta ***Ecuación 5*** convierte un conjunto de puntuaciones arbitrarias en una distribución de probabilidad sobre K clases. A diferencia de la función sigmoide que solo sirve para dos clases, el ***Softmax*** normaliza un vector de dimensión K.

En la ***Tabla 7*** encontramos el resultado de importancia de los síntomas resultado de la regresión logística. Para la ejecución del modelo se ha utilizado una penalización de nivel 1 (penalización Lasso) que fuerza muchos coeficientes a cero, lo cual ayuda a seleccionar las variables más importantes evitando el sobreajuste. El optimizador que soporta el nivel 1 de penalización es el *“saga”*, consiguiendo un buen escalado en datasets binarios dispersos.

En cuanto a los síntomas cuya importancia supera las 10 unidades encontramos los picores, la fatiga, el dolor muscular, el dolor detrás de los ojos, flema rojizo, sangre en el flema, vómitos, orina de color oscuro, fiebre alta y nauseas.

### **Árbol de decisión**

El árbol de decisión se empleó como instrumento de selección de variables porque, durante su construcción, evalúa recursivamente cada síntoma según su capacidad para reducir la impureza del nodo (**ganancia de información**). En cada partición se escoge el síntoma que maximiza dicha ganancia, calculada aquí con el **criterio de entropía**, y se establece como regla de división. Concluido el entrenamiento, la importancia global de un síntoma se obtiene sumando las reducciones de impureza que genera en todos los nodos en los que interviene. La ordenación descendente de estas importancias proporciona un ranking objetivo de relevancia clínica, útil para definir un subconjunto óptimo de variables antes de entrenar modelos de mayor complejidad.

Para el ajuste del árbol se impusieron las siguientes restricciones:

* **Criterio de división**: entropía.
* **Profundidad**: sin límite explícito; el crecimiento se controla mediante parámetros de tamaño de nodo.
* **Muestras mínimas para dividir un nodo**: 2.
* **Muestras mínimas en nodo hoja**: 1, lo que permite obtener hojas altamente específicas.

Los resultados de este procedimiento se presentan en la ***Tabla 8***, que recoge, en orden descendente, la importancia relativa de cada síntoma en el proceso de clasificación.

### **Random Forest**

En el contexto de la selección de variables, un **Random Forest** actúa como un ensamblado de muchos árboles de decisión entrenados sobre muestras ***Bootstrap*** y considerando, en cada división, solo un subconjunto aleatorio de características. Gracias a esta diversidad, además de proporcionar predicciones robustas, ofrece una medida de importancia de las variables (por ejemplo, la reducción media de impureza asociada a cada síntoma). Al ordenar esas importancias, podemos identificar qué síntomas contribuyen más al poder predictivo del modelo y usarlos para filtrar o priorizar variables en etapas posteriores.

La explicación teórica del árbol se realiza en el apartado **Random Forest (RF)**, pero en esta sección, se usarán la técnica del Random Forest para obtener información sobre las variables más importantes del conjunto de datos.

La ***Tabla 9*** nos muestra los síntomas que más información aportan de entre todos los del dataset según la técnica del Random Forest. Para la construcción del bosque se han utilizado un número de 100 árboles de decisión, un uso de la técnica de boostraping y un criterio de impureza para elegir la mejor división de cada nodo con el índice de Gini.

### **Comparación de las técnicas para selección de variables**

Las cuatro técnicas han provisto resultados muy similares, sobre todo porque han escogido los mismos síntomas, aunque el orden de importancia varie en algunos de ellos. Las variables más importantes coinciden en la mayoría de técnicas.

Todas estas técnicas son bastante fiables para la selección de variables, aunque el **Random Forest** es uno de los más fiables. Esto se debe a que construye modelos de predicción reales, en los que trata de conseguir la precisión máxima. Además, a diferencia del **árbol de decisión** tradicional, la técnica de Random Forest construye muchos árboles de decisión (en nuestro caso 100) y esto aporta una confianza extra a los resultados que nos brinda.

La ***Tabla 10*** muestra los síntomas ordenados por importancia según cada técnica utilizada. La parte alta de la tabla es similar en la mayoría de técnicas y casi todos los síntomas situados entre los 20 más significativos son los mismos en todas ellas.

Finalmente, los 20 síntomas que más se han repetido por orden de importancia son: Fatiga, vómitos, fiebre alta, dolor abdominal, dolor en el pecho, picores, dolor muscular, escalofríos, coma, dolor de cabeza, pérdida de apetito, malestar, eflorescencia (lesión cutánea), ojos amarillentos, orina de color oscuro, diarrea, irritabilidad, irritación en el ano, dolor en las articulaciones y pérdida del equilibrio.

## **Relación entre variables y categorías**

### **Estudio de la relación entre síntomas**

Una vez descubiertos cuales son las variables más importantes a nivel explicativo en el conjunto de datos, el análisis se centra en estudiar la relación entre síntomas, utilizando herramientas específicas para variables binarias que permiten detectar patrones, redundancias y asociaciones relevantes entre los síntomas.

Para averiguar relaciones se utiliza el coeficiente de correlación entre variables. Para variables dicotómicas, el coeficiente de correlación que hay que usar es el **coeficiente phi**. El coeficiente phi es un derivado del coeficiente de correlación de Pearson aplicado a variables binarias. Su fórmula es la que aparece en la ***Ecuación 6***.

Ecuación 6: Fórmula del coeficiente de correlación phi

donde los parámetros A, B, C y D provienen de una tabla de contingencia entre las dos variables a analizar (***Tabla 11***). También podemos calcular el coeficiente de correlación phi de la manera que nos muestra la Ecuación 7.

Ecuación 7: Fórmula alternativa para calcular el coeficiente de correlación phi

En ***Tabla 12*** podemos encontrar el valor del coeficiente de correlación phi para los pares de síntomas que mayor relación poseen según el estadístico phi. Los síntomas cuya relación tiene mayor fuerza son la pérdida de apetito y los ojos amarillentos con un coeficiente phi del 0.77. Este valor nos indica que los síntomas presentan una relación positiva muy fuerte, y que cuando encontremos el síntoma de pérdida de apetito, lo normal será encontrar también el de ojos amarillentos.

### **Análisis de clúster de síntomas**

Con el objetivo de identificar patrones de coocurrencia entre síntomas y estudiar su agrupamiento natural, se llevará a cabo un análisis de clustering jerárquico sobre las variable dicotómicas (síntomas) más relevantes obtenidas en los pasos previos del análisis exploratorio.

Ecuación 8: Coeficiente de similitud de Jaccard

La **distancia de Jaccard** mide la disimilitud entre dos conjuntos de muestras y se define como el complemento del coeficiente de Jaccard (Lorenzetti, 2018).

=

Ecuación 9: Distancia de Jaccard

Gráfico

El contenido generado por IA puede ser incorrecto.

Ilustración 1: Dendograma del clustering jerárquico de los 20 síntomas más significativos

El dendrograma refleja la estructura de similitud en la muestra analizada, agrupando los síntomas que comparten patrones similares de aparición. En este caso, se pueden distinguir claramente cuatro clústeres principales: **Clúster 1 (amarillo)**, **Clúster 2 (verde)** agrupando síntomas compatibles con infecciones respiratorias leves, **Clúster 3 (rojo)** afecciones dermatológicas, **Clúster 4 (violeta)** con síntomas gastrointestinales y **Clúster 5 (marrón)** que representa el resto de síntomas.

Desde el punto de vista clínico, estas agrupaciones tienen coherencia, lo que refuerza la utilidad del análisis de clustering como herramienta exploratoria. No obstante, aunque estos grupos permiten identificar síntomas con alto grado de coocurrencia, no debe asumirse que basta con un único síntoma por clúster para construir un modelo predictivo robusto. Cabe mencionar que este clúster ha sido realizado utilizando solo los 20 síntomas más significativos del conjunto de datos, lo cual permite reducir el ruido generado por otras variables menos importantes del dataset.

### **Análisis de clúster de las enfermedades**

El análisis de clústeres se incluyó únicamente para comprobar la coherencia médica de los síntomas seleccionados y descartar redundancias antes del modelado.

La distancia utilizada para crear el clúster es la distancia de Jaccard (explicada en **Análisis de clúster de los síntomas**) y el método utilizado es *average* (media).

Ecuación 10: Fórmula del método de la media para clusterización (average linkage)

La ***Ecuación 10*** muestra la fórmula para calcular esta distancia media (Ana Belén Nieto Librero & Nerea Gonzalez García, s. f.), siendo P y Q dos objetos del mismo grupo, y R otro objeto de un grupo distinto.

Gráfico

El contenido generado por IA puede ser incorrecto.

Ilustración 2: Dendograma del clúster jerárquico de enfermedades usando solo los 20 síntomas más significativos

Tras haber realizado la selección de variables en **Comparación de las técnicas para selección de variables** y habiendo obtenido estos síntomas más significativos, los usamos para el análisis de clúster. El análisis de clúster, que se muestra en Ilustración 2, representa clara y ordenadamente todas las enfermedades, y solo hay 4 que no consigue incluir en ningún grupo. Además, la mayoría de las uniones se realizan a distancias bajas, lo cual indica que consigue detectar similitudes entre enfermedades con poca información.

El dendrograma agrupa las patologías en siete clústeres clínicamente coherentes:

1. **Articulares**: artritis y osteoartritis se sitúan juntas porque comparten inflamación crónica del aparato locomotor.
2. **Dermatológico-inmune**: afecciones de piel y reactividad inmunitaria; acné, psoriasis, impétigo, reacción a fármacos, infección fúngica, VIH, alergia, hemorroides e infección urinaria, aparecen muy próximas por su patrón cutáneo-mucoso y la participación del sistema inmune.
3. **Hepático**: se fusionan las enfermedades con afectación directa del hígado o cuadro de ictericia (ictericia, hepatitis alcohólica, úlcera péptica) junto con las hepatitis virales A-E y la colestasis crónica; todas comparten manifestaciones hepatobiliares.
4. **Metabólico-circulatorio**: hipotiroidismo, hipertiroidismo, diabetes, hipoglucemia y varices convergen porque derivan de alteraciones endocrinas o vasculares que cursan con trastornos metabólicos sistémicos.
5. **Respiratorio infeccioso**: resfriado común, neumonía y tuberculosis forman un bloque definido por la afectación aguda del tracto respiratorio.
6. **Infecciones tropicales**: malaria, dengue y fiebre tifoidea quedan juntas debido a su origen endémico, curso febril y sintomatología sistémica superpuesta.
7. **Mixto neuro-cardio-digestivo**: infarto de miocardio, reflujo gastroesofágico, gastroenteritis, parálisis cerebral, vértigo, espondilosis cervical e hipertensión comparten un núcleo de síntomas neurovegetativos (dolor torácico, náuseas, inestabilidad) que el algoritmo detecta como similitud suficiente para agruparlos.

Estas siete agrupaciones reproducen de forma compacta la proximidad sintomática observada y ofrecen una referencia clínica valiosa para las fases posteriores de modelado, donde se aprovechará la redundancia interna de cada grupo para reducir preguntas sin perder capacidad diagnóstica.

# **Reducción de la dimensionalidad**

## **Introducción**

En conjuntos con cientos de variables es habitual que sólo una fracción aporte auténtica información; el resto introduce ruido o repite contenido por colinealidad. Para comprobarlo se recurrió a técnicas de reducción de dimensionalidad (**MCA**, **UMAP**, **Logistic PCA**), cuyo propósito aquí fue puramente exploratorio: proyectar los datos a dos o tres dimensiones y visualizar si las enfermedades se agrupan de forma coherente cuando se emplean los síntomas más relevantes.

Estos métodos no intervienen en el modelo diagnóstico adaptativo por dos motivos. Primero, el objetivo del TFG es conservar los síntomas originales, no generar combinaciones abstractas que luego habría que recalcular en producción. Segundo, el modelo ya incorpora un proceso de selección de variables que elimina redundancias y baja varianza sin necesidad de transformaciones adicionales.

En suma, la reducción de dimensionalidad se mantuvo como herramienta gráfica para respaldar la coherencia de los síntomas escogidos; el aprendizaje adaptativo se entrena y opera directamente con esas variables originales

## **Análisis de correspondencias múltiples (MCA)**

### **Desarrollo teórico**

El **análisis de correspondencias múltiples (MCA)** es la versión para variables categóricas del PCA. Proyecta una gran tabla de frecuencias a dos o tres ejes que concentran la mayor parte de la inercia y permiten visualizar, en un ***biplot***, la proximidad entre síntomas y entre enfermedades (Greenacre, 2017). En este TFG se emplea solo con fines exploratorios: verificar si las categorías se agrupan lógicamente y corroborar la relevancia de los síntomas ya seleccionados. Para evitar el ruido que generan variables poco informativas, el MCA se ejecuta únicamente sobre el subconjunto de síntomas significativos (**Selección de variables**); así la proyección resulta más estable y refleja con mayor claridad las relaciones clínicamente coherentes entre filas (enfermedades) y columnas (síntomas).

### **Estudio práctico**

Una vez definidos los principios fundamentales del **Análisis de Correspondencias Múltiples**, en este proyecto se aplicará esta técnica para representar gráficamente la estructura de las relaciones existentes entre los síntomas seleccionados y entre las enfermedades.

Para mejorar la legibilidad del ***biplot*** de MCA se procedió, previamente, a agrupar las enfermedades en conjuntos clínicamente homogéneos. Al representar decenas de entidades individuales, la superposición de puntos dificultaba la interpretación; por ello, y apoyándonos en el dendrograma de similitud, se fusionaron las patologías que comparten mecanismos fisiopatológicos y sintomatología dominante. Esta agregación, basada en coherencia clínica y validada por el análisis de clúster, permite reducir el número de puntos trazados, enfatizar las distancias relevantes entre grupos y, en definitiva, obtener una proyección gráfica más nítida y discriminativa.

Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.

Ilustración 3: Biplot de enfermedades agrupadas

La proyección MCA de la ***Ilustración 3*** muestra a los pacientes en el plano definido por los dos primeros ejes y ya agrupados en los siete grandes clústeres de enfermedades. Cada etiqueta aparece en el centroide de su grupo, de un vistazo se aprecia que la mayor parte de los conglomerados, hepáticos, respiratorios, tropicales, metabólicos, etc. quedan bien diferenciados, mientras que otros se solapan parcialmente, reflejando la superposición clínica real de sus síntomas. Esa mezcla localizada confirma que, aunque la reducción a 20 síntomas conserva gran parte del poder discriminante, ningún grupo resulta completamente monolítico: el modelo adaptativo deberá seguir interrogando al paciente para desambiguar cuadros limítrofes.

La ***Ilustración 16***, que añade circunferencias envolventes, refuerza esta lectura visual y pone de relieve dos pares especialmente próximos: las enfermedades hepáticas virales junto con las ictericias no virales y las respiratorias frente a las infecciones tropicales, ambas pareadas por compartir síntomas sistémicos (fiebre, escalofríos, tos, malestar). La tabla de contribuciones confirma el patrón: la primera componente está dominada por manifestaciones hepáticas (orina oscura, dolor abdominal, ictericia), mientras que la segunda viene definida por síntomas respiratorio-sistémicos (dolor muscular, escalofríos, tos, malestar).

Cuando se prescinde de la agregación y se analiza cada enfermedad por separado (***Ilustración 12***), la disposición global permanece, lo que valida la coherencia clínica del agrupamiento previo. En contraste, repetir el MCA con los 131 síntomas originales (***Ilustración 14***) dispersa los puntos y difumina las fronteras: la información relevante queda diluida entre variables irrelevantes. Este contraste subraya la razón práctica de limitarse al subconjunto de síntomas significativos y confirma que la fase exploratoria, sin intervenir en el motor diagnóstico, aporta evidencia visual de que el (diagnosticar con el menor número de preguntas) es alcanzable sin sacrificar discriminación clínica.

## **Aproximación y proyección de variedad uniforme (UMAP)**

### **Desarrollo teórico**

**UMAP** es un método no lineal de reducción de dimensionalidad diseñado para proyectar datos de alta dimensión en pocos ejes (habitualmente 2) preservando, tanto como sea posible, la vecindad local y la geometría global. Su popularidad radica en la rapidez, la escalabilidad y la capacidad de mantener estructuras complejas que otros algoritmos (p. ej. **t-SNE**) tienden a distorsionar en conjuntos extensos.

Ecuación 11: Probabilidad de conexión entre dos puntos usando la distancia de Jaccard

El proceso consta de tres etapas (véanse las Ecuaciones **Ecuación 11**-**Ecuación 13** incluidas en la memoria):

1. **Grafo de alta dimensión**: Se construye un grafo ponderado donde cada nodo es un paciente y las aristas reflejan la probabilidad de conexión ​, calculada con la distancia de Jaccard y ajustada por el parámetro para equilibrar la densidad local (**Ecuación 11**).
2. **Grafo proyectado:** Los mismos nodos se ubican en el plano reducido, allí se definen pesos ​​, mediante una función paramétrica controlada por y (**Ecuación 12**).
3. **Alineación de grafos**: Se minimiza la divergencia cruzada entre ambas distribuciones de pesos (**Ecuación 13**), de modo que la topología del grafo original se refleje, en la medida de lo posible, en la representación de baja dimensión.

Ecuación 12: Proyección en el espacio de baja dimensión UMAP

En la práctica, el mapa resultante ilustra cuán aislada o solapada queda cada enfermedad: cuanto más separado aparece un grupo, mejor distinguen sus síntomas a los pacientes que lo componen. Sin embargo, la técnica se emplea aquí solo con fines exploratorios y visuales; el modelo diagnóstico adaptativo opera directamente con los síntomas originales, por lo que UMAP no interviene en la fase predictiva. Su utilidad reside en confirmar, de un vistazo, que el subconjunto de variables seleccionadas contiene información suficiente para diferenciar clínicamente los cuadros más relevantes.

Ecuación 13: Función a minimizar entre los dos grafos producidos en la técnica UMAP

### **Estudio práctico**

Calendario

El contenido generado por IA puede ser incorrecto.

Ilustración 4: UMAP usando los 20 síntomas más significativos

La **proyección UMAP** basada únicamente en los 20 síntomas seleccionados genera un mapa con micronúcleos densos que representan enfermedades fácilmente distinguibles (acné, alergia, tuberculosis, varicela, fiebre tifoidea, dengue). Sin embargo, una fracción considerable de diagnósticos aparece aglomerada en la zona central del plano (aprox. –10 ≤ Comp1 ≤ 10), consecuencia de compartir síntomas genéricos, fatiga, fiebre o dolor muscular, que el subconjunto reducido no logra desambiguar.

Este patrón indica lo siguiente:

* **Alta especificidad** : UMAP separa con nitidez las patologías cuya sintomatología es muy característica.
* **Riesgo de solapamiento** : los agrupamientos centrales revelan pares o tríos de enfermedades clínicamente próximas que, con solo 20 síntomas, podrían inducir confusiones en el modelo predictivo.
* **Valor exploratorio** : la técnica confirma qué cuadros requieren preguntas adicionales o variables más discriminantes antes de desplegar el clasificador adaptativo.

Al volver a ejecutar UMAP con los 131 síntomas originales, la distribución de las enfermedades en el plano resulta mucho más coherente desde el punto de vista clínico: los grupos se separan mejor y el solapamiento disminuye. Esta comparación subraya que UMAP cumple aquí un papel meramente exploratorio (validar cómo se organizan las patologías en el espacio de variables) y no formará parte del motor predictivo. Asimismo, confirma que el subconjunto de 20 síntomas basta para diferenciar muchos diagnósticos, pero que los casos con sintomatología común requerirán las preguntas adicionales del modelo adaptativo para evitar confusiones.

## **Análisis de componentes principales logístico (Logistic PCA)**

### **Desarrollo teórico**

En presencia de un gran número de síntomas binarios altamente colineales, la combinación de regresión logística con análisis de componentes principales logístico **(logistic PCA**) ofrece una alternativa estable a las técnicas lineales convencionales.

El **PCA clásico** genera componentes ortogonales que concentran la mayor varianza, pero presupone variables continuas gaussianas; aplicado a datos binarios, pierde sentido estadístico. El logistic PCA resuelve esta limitación al modelar, para cada paciente y síntoma , los ***log-odds*** ​ de presencia del síntoma:

Ecuación 14: Modelo Latente de Logs-Odds

donde ​ es un efecto fijo del síntoma, ​ representa las coordenadas latentes del paciente y ​ las cargas del síntoma en dichos ejes. La probabilidad estimada se obtiene mediante la función logística.

Ecuación 15: Función logística

**Ventajas.**

* **Reducción drástica de dimensión**: los primeros componentes capturan la mayor covariación entre síntomas, mitigando la inestabilidad por colinealidad.
* **Compatibilidad binaria**: al operar sobre log-odds, mantiene la interpretación probabilística inherente a datos 0/1.
* **Base para modelos supervisados**: los componentes resultantes pueden alimentarse a una regresión logística multinomial o a un clasificador adaptativo con menos riesgo de sobreajuste.

En síntesis, el **logistic PCA** es el análogo no lineal y probabilístico del PCA clásico para variables binarias, proporcionando factores latentes independientes que conservan la información diagnóstica esencial con mayor solidez estadística.

### **Estudio práctico**

La Ilustración 5 muestra el resultado del **logistic PCA**, usando los 20 síntomas más significativos. Cada punto representa a un paciente, proyectado en el plano definido por los dos primeros componentes principales, y el nombre de la enfermedad aparece en el centroide del conjunto de pacientes diagnosticados con dicha patología.

La proyección obtenida con logistic PCA sobre los veinte síntomas seleccionados muestra que los pacientes se ordenan en el plano de los dos primeros componentes de acuerdo con patrones clínicos bien reconocibles: las hepatitis virales y la hepatitis alcohólica se concentran en la región superior izquierda, mientras que neumonía, resfriado común, tuberculosis y las infecciones tropicales (dengue, malaria, fiebre tifoidea) comparten el cuadrante inferior derecho conforme a su sintomatología febril y respiratoria. En otras zonas, especialmente el sector inferior izquierdo, varias patologías se solapan, lo que indica que, con solo dos componentes, la varianza capturada no basta para desambiguarlas por completo. Esta representación confirma, no obstante, que el núcleo de veinte síntomas retiene suficiente información para diferenciar los grandes bloques clínicos, y justifica su empleo como base del cuestionario adaptativo.

Cuando el logistic PCA se repite con los 131 síntomas originales la separación visual se degrada: la varianza se reparte entre más componentes y los dos primeros dejan de discriminar categorías con claridad. El contraste subraya el carácter exclusivamente exploratorio de estas técnicas de reducción; el modelo predictivo final se entrena con los síntomas originales y relevantes, manteniendo interpretabilidad y capacidad de preguntar al paciente por variables concretas.

Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.

Ilustración 5: Resultado Análisis de componentes principales logístico para los 20 síntomas más significativos

## **Comparación de métodos**

El examen comparativo de **MCA, UMAP y logistic PCA** ha cumplido una función estrictamente exploratoria, orientada a verificar la coherencia clínica de los datos antes de implantar el motor diagnóstico adaptativo.

* **MCA** ofrece la visión más explícita de las relaciones entre categorías. Al situar síntomas y enfermedades en un mismo biplot permite identificar qué síntomas impulsan cada eje y qué grupos patológicos se separan con ellos.
* **UMAP**, en cambio, sobresale cuando se busca la individualización de diagnósticos. Con los veinte síntomas seleccionados aísla con nitidez enfermedades de perfil muy específico, mientras que, al incorporar los 131 síntomas originales, reproduce con mayor fidelidad las vecindades clínicas entre entidades próximas. Esta decisión concuerda con la del artículo (Becht et al., 2019), donde al comparar UMAP con otros métodos como t-SNE y PCA, se demuestra que UMAP mantiene con mayor fidelidad tanto la estructura global como en la vecindad local de muestras.
* **Logistic PCA** facilita una lectura de vecindad semejante a MCA, pero revela solapamientos relevantes que alertan sobre diagnósticos difícilmente separables con sólo dos componentes.

En conjunto, estos tres métodos han servido para confirmar que el subconjunto de veinte síntomas conserva la estructura clínica esencial, señalar los focos de solapamiento donde el cuestionario adaptativo deberá profundizar y reforzar la decisión de trabajar operativamente con los síntomas originales, no con variables transformadas. De este modo, la fase exploratoria respalda el objetivo último del proyecto: construir un modelo predictivo flexible, interpretable y capaz de diagnosticar con el menor número posible de preguntas concretas al paciente.

# **Modelo de predicción**

## **Introducción**

Los modelos de predicción son formulaciones matemático-estadísticas que, a partir de datos históricos, aprenden patrones para estimar resultados futuros. En este trabajo se pretende anticipar el diagnóstico de una entre 42 patologías posibles utilizando vectores de síntomas codificados como presentes o ausentes; en esencia, se aborda un problema de **clasificación multiclase**. La experimentación con distintos algoritmos de minería de datos persigue un doble objetivo: medir su capacidad discriminante y determinar cuántos síntomas bastan para mantener una precisión clínicamente aceptable, con vistas a una implantación operativa.

Los métodos evaluados son:

* **Árbol de decisión**. Estructura jerárquica donde cada nodo interno formula una regla sobre un síntoma y cada hoja asigna la enfermedad correspondiente.
* **Random Forest**. Conjunto de árboles entrenados sobre subconjuntos aleatorios de datos y variables; la votación agregada aumenta la robustez y mitiga el sobreajuste.
* **XGBoost (Extreme Gradient Boosting)**. Algoritmo de ***boosting*** basado en árboles que corrige iterativamente los errores residuales y suele superar en rendimiento a los bosques aleatorios.
* **Modelo bayesiano secuencial**. Clasificador probabilístico que actualiza, tras cada respuesta, las distribuciones a priori mediante la regla de Bayes, lo que habilita un cuestionario adaptativo.

La comparación sistemática de estos enfoques permitirá identificar el más idóneo para el propósito del proyecto y asentar un prototipo eficiente, interpretable y aplicable en la práctica clínica.

## **Separación de datos en train y test**

En un problema de clasificación, el objetivo del algoritmo es inducir un modelo a partir de un conjunto de ejemplos para que, ante observaciones nunca vistas, devuelva la categoría correcta. Para evaluar esa capacidad de generalización resulta imprescindible particionar el conjunto de datos en dos subconjuntos mutuamente exclusivos: uno de **entrenamiento** (alrededor del 70 %) y otro de **prueba** (alrededor del 30 %). El primero sirve para ajustar los parámetros; el segundo, para estimar de forma objetiva el rendimiento fuera de muestra.

Debe advertirse, sin embargo, que la selección de variables constituye en sí misma una etapa de entrenamiento, pues se apoya en la información de las etiquetas (o en la estructura intrínseca) para decidir qué síntomas retener. Si este filtrado se realiza antes de la partición, se produce una fuga de información, lo que se conoce como “***data leakage****”*, que provoca:

1. **Un sesgo optimista en las métricas de validación,**
2. **Una estimación irreal de la capacidad predictiva**
3. **Un modelo aparentemente preciso pero no extrapolable al entorno clínico real, donde solo dispone de datos genuinamente inéditos.**

En este trabajo la división en entrenamiento y prueba se efectuó previa a cualquier procedimiento de selección de variables, garantizando que el modelo final se evalúe únicamente con información no utilizada durante su construcción. Con esta salvaguarda metodológica se procede, a continuación, al análisis comparado de los distintos algoritmos de clasificación.

## **Árbol de decisión**

### **Desarrollo teórico**

Un **árbol de decisión** es un clasificador jerárquico organizado como un grafo acíclico en forma de árbol: los nodos internos formulan pruebas sobre un atributo, las ramas representan los posibles resultados de la prueba y los nodos hoja asignan la clase final . La construcción se basa en una partición recursiva del espacio de datos que emula el razonamiento humano paso a paso: a cada división, las instancias se encaminan por una rama distinta hasta converger en una hoja que contiene el diagnóstico. Por su transparencia, la ruta de decisión es explícita, el árbol se clasifica como algoritmo de ***caja blanca***, idóneo cuando se requiere interpretabilidad.

El esquema de trabajo que presenta el algoritmo es el siguiente:

1. **Selección del atributo óptimo:** En cada nodo se escoge la variable que mejor separa las instancias mediante una ***Medida de Selección de Atributos* (MSA)**, como la ganancia de información, el índice de Gini o el *gain ratio*.
2. **Generación del nodo de decisión:** El atributo elegido se establece como nodo; las particiones resultantes forman las ramas y originan nuevos subárboles.
3. **Recursión y parada:** El proceso se repite sobre cada subconjunto hasta que se cumple alguno de los siguientes criterios:
   * Todas las instancias del nodo pertenecen a la misma clase
   * No quedan atributos por evaluar
   * El nodo queda sin observaciones.

El procedimiento finaliza con un modelo interpretable que clasifica nuevas instancias siguiendo la ruta definida por las pruebas sucesivas.

Diagrama

El contenido generado por IA puede ser incorrecto.

Ilustración 6: Proceso de construcción de un árbol de decisión. (*Tutorial de clasificación de árboles de decisión en Python*, s. f.)

Las medidas de selección de atributos (**MSA**) son heurísticas que permiten conocer el atributo que divide las tuplas de la mejor manera posible. En el caso de atributos categóricos, los criterios de división se construyen usando cada una de las categorías del atributo estudiado. Para datos continuos se pueden usar rangos o intervalos de valores. Las medidas de selección más populares son la **Ganancia de Información**, y el Í**ndice de Gini**. El método de medidas de selección utilizado determinará el resultado, y es posible que dependiendo de los datos con los que trabajamos haya un método más preciso que otro.

#### **Ganancia de información**

La ganancia de información mide la impureza del conjunto de entrada y la compara con el conjunto de salida. Para ello, usa el término de entropía (aleatoriedad o impureza en el sistema), tratando de buscar la disminución de esta en el sistema de salida (Navlani, 2024).

Ecuación 16: Entropía del dataset (D)

En la ***Ecuación 16*** encontramos la entropía del conjunto de datos. Para calcular esta entropía se usan las proporciones de cada categoría de la variable respuesta en el conjunto de datos, representadas por . Si todas las tuplas del dataset pertenecen a una misma categoría, es decir, que la variable respuesta solo tuviera un posible valor, la incertidumbre sería 0. El caso contrario se consigue cuando todas las categorías de la variable respuesta se encuentran uniformemente distribuidas, y por lo tanto H(D) alcanza su valor máximo.

Ecuación 17: Entropía condicionada tras la partición del atributo A

En la ***Ecuación 17*** encontramos la segunda parte de la fórmula de la ganancia de información. Al dividir D según el atributo A, obtenemos la entropía media ponderada tras el corte. representa el peso de las tuplas que contienen el valor v dentro del atributo A.

Ecuación 18: Ganancia de información de A

La ***Ecuación 17*** muestra que la ganancia de información al usar el atributo A para dividir el dataset (D), es la disminución de entropía que se consigue.

#### **Índice de Gini**

El índice de Gini utiliza el método Gini para crear los puntos de división. Usa las probabilidades de que una tupla de D pertenezca a una categoría del atributo en estudio. Este índice considera una división binaria para cada atributo. Se realizan cálculos de la suma de las impurezas para las particiones realizadas en un atributo. Finalmente, se selecciona como atributo de división el subconjunto que da el índice de Gini mínimo. Esta heurística es parecida en su base a la ganancia de información.

Ecuación 19: Índice de impurezas de Gini

En la Ecuación 19 denotamos como la proporción de ocurrencias de la categoría i en la variable explicativa. Si solo hay una categoría en la variable respuesta el índice de Gini es 0, indicando una impureza igual a 0.

Ecuación 20:Impureza media ponderada tras la partición

Ecuación 21: Ganancia de Gini

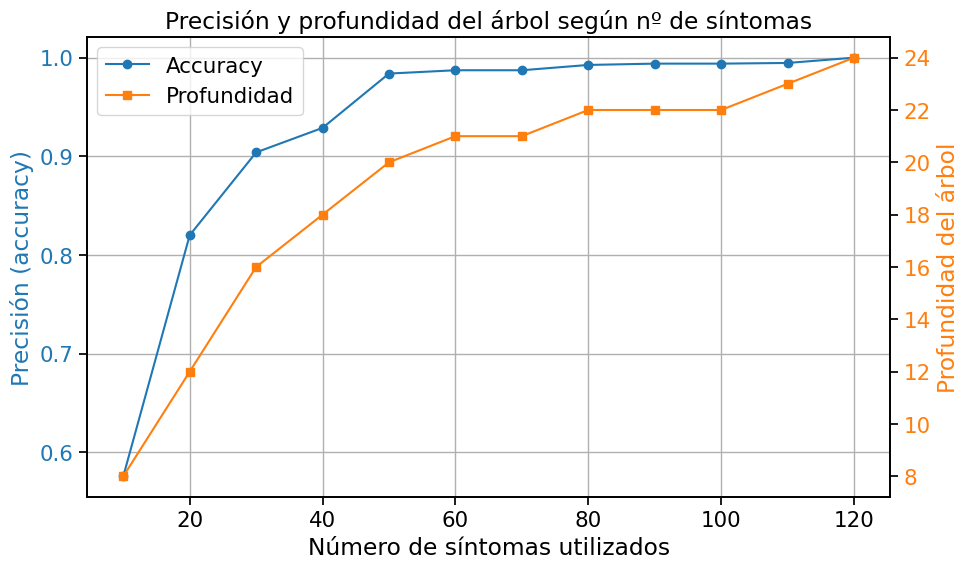
En la Ecuación 21 podemos ver la fórmula de la ganancia de Gini, la cual se trata de maximizar para escoger en cada nodo el atributo que sea capaz de explicar mayor información.

Como criterio de partición, el árbol de decisión individual utiliza ***entropy***(ganancia de información), mientras que el Random Forest opera con el índice de Gini. Estos son los dos métodos de elección de nodo empleados en el proyecto.

### **Estudio práctico**

En el modelo inicial se entrenó un árbol sin restricción de profundidad utilizando la totalidad de los síntomas disponibles. Tal como recoge la ***Ilustración 19***, el clasificador alcanza una precisión del 100% un resultado previsible dado que cada patología posee un patrón sintomático exclusivo: con el repertorio de síntomas completo, el algoritmo puede distinguirlas de forma unívoca. No obstante, esta capacidad se logra a costa de un árbol de 24 niveles, equivalente en términos prácticos a interrogar al paciente sobre 24 síntomas diferentes antes de emitir un diagnóstico. Esta extensión del cuestionario se aleja del objetivo principal del proyecto, que persigue reducir de manera drástica el número de preguntas manteniendo una precisión clínicamente aceptable.

La Ilustración 7 muestra el resultado del estudio sobre la relación entre el valor de la precisión, el número de síntomas y la profundidad del árbol. La profundidad mínima se obtiene cuando se usan 10 síntomas pero la precisión del árbol está por debajo del 60% de precisión para ese caso. Usando 20 síntomas, la precisión consigue una mejora importante aunque la profundidad también aumenta llegando hasta los 12 niveles. La estabilización de la precisión se consigue con alrededor de 50 síntomas, pero la profundidad sigue aumentando hasta alcanzar su valor máximo (24 niveles de profundidad).

Ilustración 7: Grafico con la relación entre el valor de la precisión, el número de síntomas y la profundidad del árbol

La ***Ilustración 7*** condensa, en un único gráfico, los resultados de múltiples árboles de decisión; superpone la precisión(eje izquierdo) y la profundidad del árbol(eje derecho) frente al número de variables utilizadas. Esta representación permite valorar de un vistazo el compromiso entre rendimiento diagnóstico y número de preguntas. La figura, por tanto, evidencia que alcanzar un acierto clínicamente satisfactorio exige todavía un cuestionario más largo de lo deseado y subraya la necesidad de métodos adaptativos que mantengan la precisión con menos preguntas.

El incremento simultáneo del número de síntomas empleados en el entrenamiento y de la profundidad del árbol mantiene una relación directa: cada nuevo síntoma introduce un eje adicional a lo largo del cual el algoritmo puede segmentar el espacio de características, incrementando el repertorio de divisiones posibles y, en consecuencia, la complejidad de la frontera de decisión. Una mayor cantidad de variables también amplía el abanico de interacciones sintomáticas que el modelo puede explotar, lo que se traduce en trayectorias más extensas desde la raíz hasta las hojas y, por tanto, en árboles más profundos. Esta profundidad adicional acostumbra a mejorar la precisión sobre los datos de entrenamiento, si bien incrementa a la vez el riesgo de sobreajuste.

La ***Ilustración 18***, elaborada con los veinte síntomas más relevantes, confirma que la precisión se estabiliza alrededor del 85 % cuando la profundidad máxima alcanza los nueve niveles. Tal exactitud resulta insuficiente para un escenario clínico, ya que implicaría un error de diagnóstico en aproximadamente quince de cada cien pacientes. Estos resultados conducen a priorizar la restricción de la profundidad del árbol, idealmente a no más de diez niveles, como parámetro crítico para cumplir el objetivo del proyecto: reducir el número de preguntas sin comprometer la fiabilidad diagnóstica. En consecuencia, el siguiente paso consistirá en estudiar la relación entre precisión y profundidad del árbol, manteniendo fijo el número de variables (usando todos los síntomas).

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

Ilustración 8: Relación entre la profundidad máxima del gráfico y el valor de precisión

En la ***Ilustración 8*** encontramos el valor de precisión obtenido para cada modelo según la profundidad utilizada. Se puede comprobar como la precisión se comienza a estabilizar alrededor de los 8 niveles de profundidad lo cual es muy buena señal, ya que es posible conseguir un modelo con una precisión aceptable (alrededor del 90% de éxito para una profundidad de 8 niveles) y un pocos niveles de profundidad (pocas preguntas a realizar al paciente).

En las fases iniciales se supuso que entrenar el árbol de decisión únicamente con los 20 síntomas más relevantes constituiría la opción óptima; sin embargo, dicha premisa pasaba por alto el verdadero factor que condiciona la experiencia del paciente: la profundidad del árbol, esto es, el número máximo de nodos (preguntas) en la trayectoria más larga hasta una hoja. Si se fija un límite a esa profundidad, el algoritmo formulará, en el peor de los casos, solo *X* preguntas, con independencia de que el espacio de variables contenga 20 o 131 síntomas.

Los experimentos lo confirman: con las 131 variables completas y una profundidad máxima de 10 niveles, el modelo alcanza una precisión del 94,35 %. Al restringirlo a los 20 síntomas seleccionados, la exactitud desciende aproximadamente diez puntos porcentuales, pese a plantear idéntico número de preguntas al paciente; la reducción de atributos elimina información discriminante sin acortar el cuestionario. Elevar la profundidad a **20 niveles** eleva la precisión hasta el **98,8 %**, pero duplica el diálogo clínico y aumenta el riesgo de sobreajuste, sin aportar una mejora proporcional al esfuerzo adicional.

Conviene subrayar, además, que el árbol de decisión es un clasificador no adaptativo, esto es, que una vez entrenado, no incorpora conocimiento procedente de nuevos casos ni reconfigura la jerarquía de atributos. Este resultado marca el paso a la siguiente fase del estudio, en la que se evaluarán modelos de mayor complejidad, Random Forest y XGBoost**,** con el objetivo de comprobar si los ensamblados y las técnicas de *boosting* pueden mejorar la precisión sin incrementar el número efectivo de preguntas al paciente.

## **Random Forest (RF)**

### **Desarrollo teórico**

El **Random Forest (RF)** es un método de aprendizaje automático basado en el ensamblado (***bagging***) de múltiples árboles de decisión. Los árboles individuales son modelos altamente interpretables, pero, si se les permite crecer sin restricciones, tienden al sobreajuste; a la inversa, un control excesivo de su profundidad conduce al subajuste. Frente a estas limitaciones existen dos estrategias: el ***boosting***, empleado por algoritmos como Gradient Boosting o XGBoost, y el ***bagging***, que es la base del RF(Jauregui, 2024). Estos bosques de árboles producen estimaciones de clasificación y regresión con **alto poder predictivo** y **resistencia al sobreajuste**, incluso en problemas de alta dimensionalidad ***(Breiman, 2001).***

El procedimiento de RF genera un número de árboles (normalmente más de 50 o 100) sobre subconjuntos de entrenamiento obtenidos mediante **bootstrapping**; cada árbol, además, recibe en cada división una muestra aleatoria de variables candidatas. Esta doble aleatorización (muestras y atributos) garantiza la diversidad entre árboles y, al promediar sus predicciones, reduce la varianza global del modelo. No obstante, la arquitectura resultante no es compatible con una entrevista secuencial (la entrevista que se le realiza al paciente para conocer el padecimiento de ciertos síntomas): para inferir la enfermedad en un paciente nuevo es necesario proporcionar valores para todos los síntomas empleados en el entrenamiento. En consecuencia, el RF se utiliza aquí principalmente como herramienta de análisis, por ejemplo, para estimar la importancia de las variables, más que como componente operativo del sistema de diagnóstico, cuyo objetivo es minimizar el número de preguntas planteadas al paciente.

### **Estudio práctico**

En esta fase se ha configurado un Random Forest de 100 árboles, número a partir del cual la curva de error se estabiliza y las mejoras adicionales resultan marginales. Cada árbol se construye sobre una muestra ***bootstrap*** extraída con reemplazo del conjunto de entrenamiento, lo que significa que algunas observaciones pueden repetirse mientras otras quedan fuera; estas últimas reciben el nombre de **casos *out-of-bag*** y permiten estimar de forma interna y fiable el error de generalización sin recurrir a una partición adicional de prueba.

Durante el crecimiento de cada árbol, en cada nodo se evalúa únicamente un subconjunto aleatorio de síntomas, de modo que distintos árboles exploran combinaciones distintas de variables y fronteras de decisión. Esta doble aleatorización, de instancias y de atributos, genera la diversidad necesaria para que el promedio de las predicciones individuales reduzca la varianza global del modelo y mitigue el sobreajuste inherente a un árbol profundo único.

Con la estructura del bosque entrenada, el algoritmo proporciona además una medida robusta de importancia de variables, útil para corroborar qué síntomas son realmente discriminantes en el diagnóstico. Sin embargo, hay una limitación intrínseca: el Random Forest exige disponer de todos los síntomas antes de emitir la clasificación, por lo que no se adapta al enfoque de **entrevista secuencial** que persigue el proyecto, esto es, formular al paciente sólo las preguntas imprescindibles. En consecuencia, su papel aquí es principalmente analítico y comparativo, sirviendo de referencia de rendimiento y de fuente de información sobre relevancia sintomática, antes de avanzar hacia modelos que combinen precisión con un diálogo clínico más breve.

## **Extreme Gradient Boosting (XGBoost)**

### **Desarrollo teórico**

**Extreme Gradient Boosting** (**XGBoost**) es un método de aprendizaje automático que combina la potencia de los árboles de decisión con la estrategia de ensamblaje **boosting**. Partiendo del esquema de **Gradient Boosting**, cada árbol sucesivo se ajusta sobre los residuos generados por el conjunto acumulado de árboles anteriores; así, las predicciones se corrigen de forma iterativa hasta que la función de pérdida deja de mejorar (« Todo lo que necesitas es XGBoost », s. f.).

XGBoost incorpora un término de regularización (penalización de la complejidad del árbol y de los pesos de las hojas) en la función objetivo, lo que atenúa el sobreajuste característico de los modelos de boosting clásicos y favorece soluciones con mejor capacidad de generalización.

El algoritmo se diseñó desde su origen para el cómputo paralelo. Divide la matriz de características en bloques comprimidos, descomprimidos únicamente cuando se requieren, y distribuye esos bloques entre núcleos o hilos, maximizando el rendimiento de CPU y la utilización de memoria caché. Esta arquitectura permite entrenamientos significativamente más rápidos que los de otros métodos de boosting.

Para determinar los puntos de división en los nodos, XGBoost recurre a un esquema de cuantiles que discretiza las variables continuas en intervalos equiprobables, reduciendo la búsqueda exhaustiva y manteniendo una alta precisión en el corte del espacio de características. Los valores faltantes se asignan al lado de la partición que maximiza la ganancia, preservando la información potencial sin introducir sesgos sistémicos.

Estas características, regularización explícita, paralelización eficiente, manejo inteligente de cuantiles y valores ausentes, hacen de XGBoost un referente cuando se busca un compromiso óptimo entre velocidad de entrenamiento, capacidad predictiva y control del sobreajuste.

### **Desarrollo práctico**

Cuando el modelo XGBoost se entrena con la totalidad de los síntomas disponibles, reproduce el comportamiento observado en el árbol de decisión y en el Random Forest: alcanza una precisión del 100 %, pues la combinación completa de variables permite discriminar de forma unívoca cada patología. Sin embargo, al restringir el entrenamiento a los 20 síntomas de mayor varianza explicada, la precisión desciende al 86,03 % (véase ***Ilustración 23***), superando ligeramente el rendimiento del Random Forest y del árbol único bajo la misma reducción de variables. Esta convergencia confirma que la configuración de hiperparámetros elegida para XGBoost es adecuada y se alinea con los resultados de los modelos basados en árbol.

Cabe subrayar un matiz operativo que comparten **Random Forest** y **XGBoost**. Aunque en ambos casos se limite la profundidad máxima, de modo que, en la práctica, un árbol concreto solo “consulte” como mucho X síntomas en la ruta hasta la hoja, el motor de inferencia exige que el vector de entrada contenga todas las variables empleadas en el entrenamiento. En Random Forest ello es estrictamente obligatorio; en XGBoost puede pasarse *“np.nan”*, pero la columna ha de estar presente y la decisión sobre la rama para valores faltantes quedó fijada durante el ajuste. En cualquier caso, para diagnosticar a un paciente se necesita rellenar los 131 síntomas del esquema original, aunque la predicción termine utilizando solo una fracción. Esta exigencia invalida a ambos modelos como herramienta de triaje secuencial: restringir la profundidad reduce la longitud lógica del árbol, pero no elimina la necesidad de suministrar el conjunto completo de variables, ni dota al sistema de la flexibilidad de preguntar al paciente únicamente lo imprescindible.

## **Modelo bayesiano secuencial**

### **Introducción**

Tras evaluar los enfoques estáticos basados en árboles, Árbol de Decisión, Random Forest y XGBoost, se revela una limitación común: aun restringiendo la profundidad, exigen disponer de todos los síntomas antes de emitir un diagnóstico. Para cumplir el objetivo principal del proyecto, predecir la enfermedad con el número mínimo de preguntas en tiempo real, resulta imprescindible un modelo generativo y adaptativo.

El **enfoque bayesiano secuencial** satisface precisamente esta necesidad. Durante la fase de entrenamiento puede aprovechar la matriz completa de síntomas, pero, en la fase de inferencia, actualiza de forma iterativa las probabilidades de cada patología conforme recibe nuevas respuestas y detiene el interrogatorio tan pronto como alcanza un umbral de certeza predefinido. De este modo, el sistema formula únicamente aquellas preguntas que aportan información marginal relevante, eliminando la carga de solicitar el valor de todos los atributos al paciente.

P. Domingos y M. Pazzani exponen en su libro (Domingos & Pazzani, 1997) porqué el clasificador bayesiano es óptimo en términos de minimizar la tarea de error, cuando se manejan variables cuyo valor es acierto o error (1 o 0), asignando siempre la clase de máxima probabilidad posterior. Este enfoque es justo el que presenta este proyecto.

En síntesis, el modelo bayesiano representa la solución idónea para el problema planteado: combina la riqueza informativa del conjunto completo de variables con un mecanismo de consulta adaptativa que reduce drásticamente el número de síntomas necesarios para llegar a un diagnóstico fiable.

### **Desarrollo teórico**

El teorema de bayes es uno de los métodos más conocidos y utilizados en ámbito probabilístico y estadístico, especialmente en el estudio de la probabilidad condicionada. Este teorema permite inferir la probabilidad de un sucesión mediante probabilidades conocidas previamente. Thomas Bayes, en su primer texto sobre el cálculo de probabilidades comenta: *“Si conocemos cuantas veces ha ocurrido un suceso en una serie de ensayos, ¿Cómo podemos estimar la probabilidad de la causa subyacente que lo genera?”* (Dale, 2005)

La técnica de clasificación de bayes está basada en el teorema de bayes y es uno de los algoritmos de aprendizaje supervisado más simples. El algoritmo asume que el efecto de una característica particular en una categoría es independiente de otras características. Para conseguir esta ingenuidad en la que los atributos se consideran independientes, se usa la fórmula de probabilidad condicionada.

Ecuación 22: Fórmula de la probabilidad condicionada

La ***Ecuación 22*** muestra la fórmula de la probabilidad condicionada, base en el algoritmo de clasificación de bayes. En ella se puede encontrar P(B), que es la probabilidad de que ocurra B; y que es la probabilidad de que ocurra A y B al mismo tiempo.

Esta definición la podemos aplicar a nuestro ámbito de estudio. Convertimos A en el paciente que tiene la categoría y B en el vector de síntomas x, referentes a ese paciente. Lo siguiente es obtener la probabilidad de la intersección:

Ecuación 23: Probabilidad de la intersección

La ***Ecuación 23*** muestra la fórmula de la probabilidad de que el paciente tenga esa categoría concreta de la variable respuesta y que se observen los síntomas x. En este caso hace referencia a la frecuencia con la que aparece esa enfermedad en el conjunto de datos, y es la verosimilitud de tener x (los síntomas) si realmente se tiene esa categoría ().

Ecuación 24: Probabilidad de tener el conjunto de síntomas x

Ahora se calcula la probabilidad de tener el conjunto de síntomas (***Ecuación 24***). En esta ecuación se calcula para el conjunto de categorías de la variable respuesta, la probabilidad de tener una categoría y además tener el conjunto de síntomas concreto.

Ecuación 25: Fórmula final del algoritmo de clasificación de bayes

La ***Ecuación 25*** muestra la fórmula utilizada en el clasificador de bayes. En el numerador encontramos la probabilidad de aparición de una categoría de la variable respuesta, multiplicada por la probabilidad condicionada de que si aparece esa categoría también aparezcan los valores para las variables “x”. El numerador comprende el sumatorio de la probabilidad de cada una de las categorías de la variable respuesta multiplicada por la probabilidad condicionada de que teniendo los valores para las variables del vector “x” se tenga como resultado esa categoría.

En la fase de entrenamiento, el algoritmo de clasificación de Bayes se divide en varias etapas. En la primera realiza un recuento de clases, contando cuantos ejemplares en el conjunto de entrenamiento pertenecen a una determinada categoría. Con esto consigue las probabilidades por categorías. Posteriormente realiza un recuento condicional de atributos de modo que para cada atributo y cada valor posible , cuenta cuantas tuplas con tienen . Una vez tiene estos datos estima la probabilidad condicional para cada uno de los posibles valores y los almacena en una tabla para su posterior uso.

En la fase de prueba o de clasificación real, el algoritmo realiza el cálculo de , de modo que obtengamos la con mayor probabilidad. Para decidir esto basta con comparar los numeradores de la ***Ecuación 25***, porque los denominadores son comunes a todas las k.

El **clasificador bayesiano secuencial** arranca con una fase de aprendizaje en la que se calcula, a partir del conjunto de entrenamiento, la frecuencia con la que aparece cada enfermedad y la probabilidad de que cada síntoma se presente cuando esa enfermedad está realmente presente. Con esos parámetros, el sistema diagnostica a nuevos pacientes de forma dinámica y gradual: inicia con una estimación inicial de la probabilidad de cada patología y formula, en cada paso, la pregunta cuyo posible resultado más reduciría la incertidumbre sobre el diagnóstico. Tras conocer la respuesta, actualiza inmediatamente las probabilidades y decide si ya existe una enfermedad con fiabilidad suficiente para concluir la entrevista, o bien continúa interrogando hasta alcanzar el límite de preguntas prefijado.

Para evitar que la entrevista se prolongue indefinidamente, y siguiendo con los consejos de la función de parada de (van der Linden & Ren, 2020), el sistema incorpora dos hiperparámetros de control: un **límite máximo de preguntas** y un **umbral de confianza** () en la probabilidad posterior. El mecanismo de parada funciona así: en cada paso, si la probabilidad de la enfermedad más probable alcanza o supera , la consulta finaliza, independientemente de cuántas preguntas queden disponibles; si no se cumple y todavía no se ha alcanzado ​, el algoritmo formula la siguiente cuestión.

Este umbral permite emitir diagnósticos con un número medio de preguntas reducido, ya que no es necesario agotar el cupo si con menos ítems se ha obtenido suficiente certeza. No obstante, a medida que disminuye, la entrevista se acorta, pero la probabilidad de error aumenta, dado que un valor posterior moderado no garantiza la veracidad de la predicción. En sentido contrario, elevar exige más preguntas para alcanzar la confianza deseada, mejorando la fiabilidad diagnóstica a costa de un interrogatorio más extenso.

### **Estudio práctico**

En la práctica, los dos parámetros utilizados proporcionan un equilibrio dinámico entre la velocidad de la consulta y su exactitud: ​ acota la longitud máxima del cuestionario, mientras que regula la certeza requerida para finalizar la entrevista, configurando así el compromiso óptimo entre eficiencia y precisión clínica.

En un primer experimento, fijamos el máximo en diez preguntas y el umbral en un 85 %. Los resultados mostraron una tasa de acierto del 93,3 %, con una media de apenas 6,35 preguntas por paciente: un 60 % de las veces el diagnóstico se emitía antes del décimo ítem, gracias a que el algoritmo prioriza los síntomas más informativos. Al relajar el límite a veinte preguntas, la precisión subió al 97,78 %, pero el promedio de ítems apenas creció (6,67), lo cual indica que raramente hace falta agotar el plazo para alcanzar la certeza requerida.

Este comportamiento pone de manifiesto la eficacia del enfoque: la entrevista se adapta al caso particular, explotando al máximo cada respuesta y deteniéndose tan pronto como las probabilidades alcanzan un nivel clínicamente aceptable. Además, al comparar estos resultados con los obtenidos en modelos estáticos (árboles de decisión, Random Forest, XGBoost), se evidencia que solo el método bayesiano secuencial logra compaginar un cuestionario breve con una fiabilidad cercana al 98 %.

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

Ilustración 9: Gráfico comparando el número máximo de preguntas, precisión y número medio de preguntas obtenido

La ***Ilustración 9*** muestra, para un umbral de confianza del 99 %, la evolución de la **precisión** (eje izquierdo) y del **número medio de preguntas** realizadas (eje derecho) al variar el límite máximo de preguntas de 5 a 30.

Con un máximo de solo 5 preguntas, la precisión se sitúa en torno al 67 % y el sistema agota siempre el cupo (media = 5). Al aumentar el límite a 10, la precisión da un salto hasta aproximadamente el 94 %, mientras que el promedio de preguntas desciende a 7,3, lo que indica que en un número apreciable de casos el diagnóstico se emite antes de alcanzar el tope. Con 15 preguntas el modelo roza el 98 % de acierto con una media de 7,5 ítems, y a partir de 20 preguntas la precisión supera el 99 % sin que el promedio de preguntas exceda 7,8. A 25 y 30 preguntas la curva se aplana en la cota máxima: el sistema raramente necesita más de ocho cuestiones para alcanzar el umbral fijado.

En conjunto, el gráfico evidencia que, bajo un umbral muy exigente, fijar ​ por encima de 15 aporta rendimientos marginales, ya que el diagnóstico suele alcanzarse con menos de ocho preguntas; al mismo tiempo, muestra el clásico compromiso entre número de interacciones y fiabilidad: cuanto mayor es el límite, más cerca se encamina la precisión hacia el 100 %. Esto confirma la idoneidad de un punto de operación intermedio (por ejemplo, 15 preguntas) para equilibrar brevedad y exactitud en un entorno clínico real.

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

Ilustración 10: Gráfico comparando el umbral de confianza con el nivel de precisión y la proporción de preguntas utilizadas

Los resultados que se visualizan en la ***Ilustración 10***, son muy alentadores consiguiendo una precisión del 97,98% usando un umbral de confianza de 0,95 (***Ilustración 10***). El número medio de preguntas para este caso es de 6.90, es decir, que estableciendo un límite máximo de 15 preguntas (algún caso clínico necesitará de 15 síntomas para ser diagnosticado), una media de preguntas por paciente menor a 7 y un umbral de confianza de 0,95 obtenemos un modelo con una precisión casi completa.

## **Comparación de resultados**

En resumen, los métodos basados en árboles binarios, **Random Forest** y **XGBoost,** han mostrado su capacidad para procesar grandes volúmenes de variables y alcanzar precisiones comprendidas entre el 84,35 % y el 86,03 % con solo veinte síntomas. No obstante, su rigidez inherente impide un diálogo interactivo: requieren conocer de antemano el valor de todos los atributos entrenados, lo que los hace poco adecuados para un escenario clínico en el que el paciente debe responder únicamente las preguntas estrictamente necesarias.

El árbol de decisión simple, por su parte, sí puede aplicarse de forma secuencial; al fijar su profundidad máxima en diez niveles, equivalente a diez preguntas, se consigue un 95,10 % de acierto. A pesar de esta ventaja, su estructura permanece invariable una vez entrenada y no incorpora ningún mecanismo de adaptación en tiempo real, por lo que la jerarquía de síntomas en cada nodo no puede modificarse al enfrentarse a nuevos patrones emergentes.

Frente a estas limitaciones, el clasificador bayesiano secuencial se revela como la solución óptima para nuestro objetivo. Entrenado con el conjunto completo de síntomas, ofrece un proceso de triaje puramente adaptativo: en cada paso calcula qué pregunta aportará la máxima reducción de incertidumbre, actualiza instantáneamente las probabilidades de cada enfermedad y detiene la entrevista tan pronto como alcanza un umbral de confianza predefinido o un límite de preguntas. Gracias a este enfoque, con un máximo de quince preguntas logra un 97,78 % de precisión, pero solo necesita en promedio siete ítems por paciente, reduciendo drásticamente la carga interrogatoria sin sacrificar fiabilidad.

Sus ventajas son múltiples: **eficiencia en el diálogo clínico**, pues solo interroga sobre las variables realmente informativas; **flexibilidad**, al adaptar la ruta diagnóstica a cada caso individual; **transparencia**, al permitir rastrear cómo cada respuesta modifica las probabilidades; y **escalabilidad**, ya que emplea todo el conocimiento estadístico de los síntomas sin abrumar al usuario. Su única desventaja, un tiempo de procesamiento mayor, queda fácilmente mitigado mediante el uso de hardware más potente o el uso de computación en la nube.

Dado este balance de prestaciones, el modelo bayesiano secuencial se impone como el componente central de nuestro sistema de diagnóstico: no solo ofrece la precisión necesaria para el entorno sanitario, sino que además encarna el principio de “preguntar solo lo imprescindible” que conforma el corazón de este Trabajo Fin de Grado. A modo de demostración, se ha implementado una función interactiva que simula la conversación con un paciente real: tras cada respuesta, el modelo muestra las tres patologías más probables y, al cabo de tan solo cinco preguntas en un escenario de prueba (***Ilustración 11***), arroja un diagnóstico con un 94,14 % de fiabilidad. Este prototipo ilustra con claridad cómo la adaptabilidad bayesiana puede transformar la sobrecarga de datos médicos en una experiencia de triaje ágil, precisa y centrada en el paciente.

Tabla

El contenido generado por IA puede ser incorrecto.

Ilustración 11: Resultado de la simulación de diagnóstico para un paciente

# **Conclusión**

El propósito de este Trabajo Fin de Grado era diseñar un sistema de triaje médico que minimizase el número de preguntas necesarias para alcanzar un diagnóstico fiable. Tras explorar enfoques clásicos de Machine Learning, **árboles de decisión**, **Random Forest** y **XGBoost,**  se constató que, aunque todos ellos pueden lograr precisiones cercanas al 95 % o superiores, requieren conocer de antemano el valor de la totalidad de los síntomas y carecen de la flexibilidad necesaria para detener el cuestionario tan pronto como la información recopilada sea suficiente.

En contraste, el **clasificador Bayesiano** **secuencial** cumple con los requisitos fundamentales del proyecto:

1. **Adaptabilidad en tiempo real:** A medida que el paciente responde, el modelo recalcula las probabilidades de cada enfermedad y selecciona siempre la pregunta más informativa para la siguiente iteración.
2. **Control estricto de la carga diagnóstica:** Con un límite de 10 preguntas y un umbral de confianza del 85 %, el sistema alcanza una precisión del 93,3 % con una media de 6,35 preguntas. Extendiendo el tope a 15 preguntas, la precisión sube al 97,8 % con 6,67 preguntas de media, mientras que con 20 preguntas roza el 99 % sin superar una media de ocho ítems.
3. **Eficiencia sin renunciar a la fiabilidad:** Incluso con apenas 5 preguntas y manteniendo un umbral alto (99 %), el diagnóstico acierta en un 67 % de los casos, lo que demuestra que los primeros ítems aportan la mayor parte de la información discriminante.

Estos resultados confirman que preguntar sólo lo imprescindible no sacrifica la exactitud diagnóstica y que existe un punto de equilibrio entre el número de preguntas y la precisión (ver curva «rendimiento vs. preguntas» ***Ilustración 9***), esencial para la adopción clínica.

En el trayecto de este proyecto también se abordaron retos personales y técnicos relevantes:

* **Límites de los modelos estáticos: S**e aprendió que, sin un mecanismo interactivo, los clasificadores convencionales no pueden traducirse en entrevistas vivas con pacientes.
* **Desarrollo como profesional estadístico:** El autor, a través del desarrollo de este proyecto, ha conseguido unir todos los conocimientos adquiridos en las diferentes asignaturas del grado en estadística, mezclándolos con estudio personal para desarrollar un plan real. Este proceso construye el puente entre la etapa puramente académica del estudiante y su acceso al mercado laboral.
* **Equilibrio precisión–carga:** La parametrización de umbrales y topes de preguntas resultó clave para ajustar el modelo a las necesidades reales del clínico y del paciente.

## **Límites del proyecto**

El desarrollo de este trabajo ha supuesto un desafío tanto académico como personal, marcado por la imprevisibilidad inherente al análisis de datos reales. En particular, la construcción de un plan de trabajo partiendo de un conjunto inicial de 4962 historias clínicas y 131 síntomas expuso varias dificultades fundamentales:

1. **Calidad y complejidad de los datos.** Las historias clínicas anonimizadas contenían valores erróneos, duplicidades y frecuencias muy desiguales entre síntomas, lo que exigió dedicar más tiempo del previsto a la depuración y validación de la información antes de diseñar cualquier modelo.
2. **Selección y reducción de variables.** El elevado número de síntomas, unido a la alta dimensionalidad, llevó inicialmente a plantear técnicas de reducción como MCA, UMAP y Logistic PCA. Sin embargo, pronto apareció el obstáculo de la **inoperatividad** de esas variables transformadas para un sistema interactivo: no se traducen en preguntas concretas. Este hallazgo supuso un giro metodológico que obligó a priorizar la selección de un subconjunto de síntomas originales, reequilibrando las fases de estudio exploratorio y la fase de modelado.
3. **Equilibrio precisión–simplicidad.** El objetivo de minimizar el número medio de preguntas de diagnóstico chocó con la tendencia natural de los algoritmos de maximizar la exactitud introduciendo más variables o aumentando la profundidad. Encontrar el punto óptimo, diez a quince preguntas con un umbral de confianza adecuado, fue un proceso iterativo de parámetros y experimentos que en ocasiones ofrecía resultados contradictorios, lo que requirió replantear estrategias y revisar supuestos.
4. **Limitaciones computacionales.** El clasificador bayesiano secuencial, a pesar de su idoneidad teórica, demandó tiempos de entrenamiento y validación notoriamente superiores a los de modelos estáticos. La necesidad de iterar numerosas simulaciones para ajustar umbrales y topes de preguntas puso de manifiesto la dependencia del hardware y sugirió la conveniencia de soluciones en la nube o técnicas de aprendizaje activo para acelerar ciclos de prueba.

Estos retos reflejan que el camino hacia un diagnóstico adaptativo no ha sido “de rosas”, sino un proceso de aprendizaje sobre cómo traducir un problema clínico real en un flujo de trabajo estadístico robusto. Cada obstáculo validó la importancia de la flexibilidad metodológica y subrayó que, en estadística aplicada, la capacidad de adaptación y la priorización de resultados realistas pesan tanto como el rigor técnico en la consecución de objetivos.

## **Proyección de futuro**

El prototipo desarrollado ha demostrado un alto grado de precisión en ambientes controlados, lo que abre la puerta a su extrapolación a entornos clínicos reales. En una primera fase, resulta prioritario **ampliar la cohorte de pacientes** incorporando nuevos registros, especialmente de las patologías más prevalentes y responsables de la saturación de la atención primaria. Este aumento en la diversidad de casos no solo enriquecerá las distribuciones a priori de cada enfermedad, sino que permitirá ajustar mejor las verosimilitudes condicionales y mejorar la capacidad del modelo para distinguir procesos clínicos similares en poblaciones heterogéneas.

Una segunda línea de trabajo consistiría en **integrar el sistema en las salas de espera** de urgencias mediante terminales de autoevaluación. Los pacientes interesados podrían responder de forma voluntaria al cuestionario secuencial mientras aguardan consulta, obteniendo una orientación preliminar sobre su posible diagnóstico. Lejos de sustituir el juicio médico, esta aplicación serviría para familiarizar al usuario con las entrevistas asistidas por IA, reducir su ansiedad y, simultáneamente, proporcionar al personal sanitario datos agregados sobre la tipología de casos que llegan al servicio, contribuyendo así a optimizar la organización de turnos y recursos.

Por último, conviene extender el modelo **incorporando variables no binarias** de naturaleza demográfica y antropométrica, edad, sexo, índice de masa corporal, porcentaje de grasa corporal, antecedentes familiares, que podrían aportar información predictiva adicional. Para ello, se exploraría la integración de técnicas bayesianas mixtas o modelos de regresión jerárquica que combinen datos continuos y categóricos, permitiendo al sistema plantear preguntas cuantitativas (por ejemplo, “¿Cuál es su edad?” o “¿Cuál es su peso?”) junto a las binarias clásicas. Esta evolución conducirá a un asistente más versátil y personalizado, capaz de adaptar el cuestionario a la configuración clínica y sociodemográfica de cada paciente y, por tanto, de mejorar su rendimiento diagnóstico real.

# **Bibliografía**

Todo lo que necesitas es XGBoost . Recuperado 1 de julio de 2025, de https://datos.ninja/tutorial/xgboost/

Ana Belén Nieto Librero & Nerea Gonzalez García. (s. f.). *Análisis de Clúster*. 2022.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., & Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, *37*(1), 38-44. https://doi.org/10.1038/nbt.4314

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

*Capítulo+3.pdf*. (s. f.). Recuperado 1 de julio de 2025, de https://biblus.us.es/bibing/proyectos/use/abreproy/10506/fichero/Cap%C3%ADtulo+3.pdf

Dale, A. I. (2005). Thomas Bayes, *An essay towards solving a problem in the doctrine of chances* (1764). En I. Grattan-Guinness, R. Cooke, L. Corry, P. Crépel, & N. Guicciardini (Eds.), *Landmark Writings in Western Mathematics 1640-1940* (pp. 199-207). Elsevier Science. https://doi.org/10.1016/B978-044450871-3/50096-6

Domingos, P., & Pazzani, M. (1997). On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning*, *29*(2), 103-130. https://doi.org/10.1023/A:1007413511361

Exploratory Data Analysis. (2008). En *The Concise Encyclopedia of Statistics* (pp. 192-194). Springer, New York, NY. https://doi.org/10.1007/978-0-387-32833-1\_136

Greenacre, M. (2017). *Correspondence Analysis in Practice* (3.a ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315369983

Jauregui, A. F. (2024, noviembre 24). Guía completa sobre Random Forest. *Ander Fernández*. https://anderfernandez.com/blog/guia-completa-random-forest/

Jr, D. W. H., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. John Wiley & Sons.

Lorenzetti, C. M. (2018). Caracterización Formal y Análisis Empírico de Mecanismos Incrementales de Búsqueda basados en Contexto. *arXiv.Org*. https://www.proquest.com/publiccontent/docview/2118097610?pq-origsite=primo

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1226-1238. https://doi.org/10.1109/TPAMI.2005.159

Shortliffe, E. H., & Buchanan, B. G. (1975). A model of inexact reasoning in medicine. *Mathematical Biosciences*, *23*(3), 351-379. https://doi.org/10.1016/0025-5564(75)90047-4

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44-56. https://doi.org/10.1038/s41591-018-0300-7

*Tutorial de clasificación de árboles de decisión en Python: Scikit-Learn DecisionTreeClassifier*. (s. f.). Recuperado 1 de julio de 2025, de https://www.datacamp.com/tutorial/decision-tree-classification-python

van der Linden, W. J., & Ren, H. (2020). A Fast and Simple Algorithm for Bayesian Adaptive Testing. *Journal of Educational and Behavioral Statistics*, *45*(1), 58-85. https://doi.org/10.3102/1076998619858970

# **Lista de abreviaciones**

* EDA – Análisis exploratorio de datos
* ML – Machine Learning
* LDA – Análisis discriminante lineal
* PCA – Análisis de componentes principales
* RF - Random Forest
* Logistic PCA - PCA logístico
* MCA - Análisis de correspondencias múltiples
* UMAP – Uniform Manifold Aproximation and Projection
* MSA – Medidad de selección de atributos

# **ANEXO**

|  |  |  |
| --- | --- | --- |
| **Síntoma (EN)** | **Síntoma (ES)** | **Descripción (ES)** |
| Abdominal pain | Dolor abdominal | Dolor en el área del abdomen |
| Abnormal menstruation | Menstruación irregular | Irregularidades en el ciclo menstrual |
| Acidity | Acidez | Sensación ardiente por reflujo gástrico |
| Acute liver failure | Insuficiencia hepática aguda | Pérdida rápida de la función hepática |
| Altered sensorium | Confusión mental | Alteración del estado de conciencia |
| Anxiety | Ansiedad | Nerviosismo o sensación constante de preocupación |
| Back pain | Dolor de espalda | Dolor en la parte baja o alta de la espalda |
| Belly pain | Dolor bajo abdominal | Dolor en la parte inferior del abdomen |
| Blackheads | Puntos negros | Obstrucción del poro con sebo y células muertas |
| Bladder discomfort | Malestar vesical | Sensación incómoda en la vejiga |
| Blister | Ampolla | Lesión cutánea con líquido interno |
| Blood in sputum | Sangre en la flema | Presencia de sangre en el esputo respiratorio |
| Bloody stool | Sangre en heces | Presencia de sangre en las evacuaciones |
| Blurred and distorted vision | Visión borrosa o distorsionada | Dificultad para ver con claridad |
| Brittle nails | Uñas quebradizas | Fragilidad de las uñas |
| Bruising | Hematomas | Aparición de moretones sin golpe aparente |
| Burning micturition | Ardor al orinar | Sensación de ardor al orinar |
| Chest pain | Dolor torácico | Dolor en el pecho, de origen respiratorio o cardíaco |
| Chills | Escalofríos | Sensación de frío extremo sin causa aparente |
| Coma | Coma | Estado de inconsciencia prolongado |
| Congestion | Congestión | Bloqueo nasal por exceso de mucosidad |
| Constipation | Estreñimiento | Dificultad o retraso en la evacuación intestinal |
| Continuous feel of urine | Ganas constantes de orinar | Urgencia urinaria frecuente |
| Continuous sneezing | Estornudos constantes | Estornudos repetitivos sin pausa |
| Cough | Tos | Expulsión brusca de aire por irritación respiratoria |
| Cramps | Calambres | Espasmos musculares involuntarios y dolorosos |
| Dark urine | Orina oscura | Tono más oscuro de lo normal, asociado a problemas hepáticos o deshidratación |
| Dehydratation | Deshidratación | Pérdida excesiva de líquidos corporales |
| Depression | Depresión | Tristeza profunda y prolongada |
| Diarrhoea | Diarrea | Evacuaciones frecuentes y líquidas |
| Dischromic patches | Manchas con cambio de color | Áreas cutáneas con alteración de pigmentación |
| Distention of abdomen | Distensión abdominal | Hinchazón anormal del abdomen |
| Dizziness | Mareo | Sensación de inestabilidad o vértigo |
| Drying and tingling lips | Labios secos y hormigueo | Sensación de sequedad y cosquilleo en los labios |
| Enlarged thyroid | Bocio | Agrandamiento de la glándula tiroides |
| Excessive hunger | Hambre excesiva | Sensación constante de necesidad de comer |
| Extra marital contacts | Contactos sexuales extramatrimoniales | Relaciones sexuales fuera del matrimonio |
| Family history | Antecedentes familiares | Historia clínica de enfermedades hereditarias |
| Fast heart rate | Taquicardia | Ritmo cardíaco acelerado |
| Fatigue | Fatiga | Cansancio extremo |
| Fluid overload | Retención de líquidos | Acumulación anormal de fluidos corporales |
| Foul smell of urine | Orina con mal olor | Orina con olor fuerte o desagradable |
| Headache | Dolor de cabeza | Dolor que puede variar en intensidad y localización |
| High fever | Fiebre alta | Temperatura corporal mayor a 38 °C |
| Hip joint pain | Dolor de cadera | Dolor en la articulación de la cadera |
| History of alcohol consumption | Consumo habitual de alcohol | Antecedente de ingesta frecuente de bebidas alcohólicas |
| Increased appetite | Aumento del apetito | Deseo excesivo de comer |
| Indigestion | Indigestión | Malestar estomacal después de comer |
| Inflammatory nails | Uñas inflamadas | Enrojecimiento o hinchazón en la base de las uñas |
| Internal itching | Picazón interna | Sensación de picor sin causa aparente |
| Irritability | Irritabilidad | Tendencia a enojarse fácilmente |
| Irritation in anus | Irritación anal | Picazón o molestia en la zona anal |
| Irregular sugar level | Glucosa irregular | Niveles anormales de azúcar en sangre |
| Itching | Picazón | Picazón o sensación de irritación en la piel |
| Joint pain | Dolor articular | Dolor en una o varias articulaciones |
| Knee pain | Dolor de rodilla | Dolor localizado en la articulación de la rodilla |
| Lack of concentration | Falta de concentración | Dificultad para mantener la atención sostenida |
| Lethargy | Letargo | Sensación de debilidad y somnolencia constante |
| Loss of apetite | Falta de apetito | Disminución o pérdida del deseo de comer |
| Loss of balance | Pérdida de equilibrio | Dificultad para mantenerse en pie o caminar |
| Loss of smell | Anosmia | Incapacidad para percibir olores |
| Malaise | Malestar general | Sensación de incomodidad o indisposición inespecífica |
| Mild fever | Fiebre leve | Temperatura moderadamente elevada |
| Mood swings | Cambios de humor | Alteraciones bruscas en el estado emocional |
| Movement stiffness | Rigidez al moverse | Dificultad para desplazarse con fluidez |
| Mucoid sputum | Flema mucosa | Moco espeso y viscoso |
| Muscle pain | Dolor muscular | Dolor localizado o generalizado en los músculos |
| Muscle wasting | Pérdida muscular | Disminución de masa muscular por enfermedad o desnutrición |
| Muscle weakness | Debilidad muscular | Falta de fuerza en los músculos |
| Nausea | Náuseas | Sensación de querer vomitar |
| Neck pain | Dolor de cuello | Dolor en la parte posterior o lateral del cuello |
| Nodal skin eruptions | Erupciones nodulares | Erupciones en la piel con inflamación de ganglios linfáticos |
| Obesity | Obesidad | Acumulación excesiva de grasa corporal |
| Pain behind the eyes | Dolor ocular profundo | Dolor o presión en la parte posterior de los ojos |
| Pain during bowel movements | Dolor al evacuar | Dolor durante la defecación |
| Pain in anal region | Dolor anal | Molestia o dolor en la zona anal |
| Painful walking | Dolor al caminar | Molestia o dolor al desplazarse |
| Palpitations | Palpitaciones | Percepción intensa o acelerada del ritmo cardíaco |
| Passage of gases | Flatulencias | Expulsión de gases por el recto |
| Patches in throat | Placas en la garganta | Manchas o secreciones blanquecinas en la faringe |
| Phlegm | Flema | Moco espeso producido en vías respiratorias |
| Polyuria | Poliuria | Producción excesiva de orina |
| Prominent veins on calf | Venas marcadas en la pantorrilla | Venas visiblemente dilatadas |
| Puffy face and eyes | Cara y ojos hinchados | Hinchazón notable en el rostro y alrededor de los ojos |
| Pus-Filled pimples | Espinillas con pus | Lesiones cutáneas inflamadas con contenido purulento |
| Receiving blood transfusion | Transfusión sanguínea | Antecedente de haber recibido transfusión |
| Receiving unsterile injections | Inyecciones no estériles | Aplicación de inyecciones sin condiciones sanitarias adecuadas |
| Red sore around nose | Herida roja en la nariz | Lesión cutánea inflamada en la zona nasal |
| Red spots over body | Manchas rojas en la piel | Lesiones pequeñas y rojizas diseminadas |
| Redness of eyes | Ojos rojos | Inflamación ocular con vasos dilatados |
| Restlessness | Inquietud | Incapacidad de estar en reposo o relajado |
| Runny nose | Nariz mocosa | Flujo constante de mucosidad nasal |
| Rusty sputum | Flema marrón | Secreción bronquial con color oxidado o sanguinolento |
| Scurring | Costras | Formación de costras tras heridas o infecciones |
| Shivering | Temblores | Temblores involuntarios, por frío o fiebre |
| Silver-like Dusting | Polvo plateado | Apariencia brillante en la superficie de la piel |
| Sinus pressure | Presión sinusal | Sensación de presión en senos paranasales |
| Skin peeling | Descamación | Caída de capas superficiales de piel |
| Skin rash | Erupción cutánea | Enrojecimiento o inflamación de la piel |
| Slurred speech | Disartria | Dificultad para articular palabras |
| Small dents in nails | Hoyuelos en las uñas | Pequeñas depresiones o marcas en la superficie de las uñas |
| Spinning movements | Mareos con giros | Sensación de vértigo o vueltas |
| Spotting urination | Sangrado urinario leve | Presencia de gotas de sangre u orina anormal |
| Stiff neck | Cuello rígido | Dificultad para mover el cuello |
| Stomach bleeding | Sangrado estomacal | Pérdida de sangre en el tracto digestivo superior |
| Stomach pain | Dolor de estómago | Dolor abdominal de intensidad variable |
| Sunken eyes | Ojos hundidos | Apariencia deprimida en los ojos por deshidratación o fatiga |
| Sweating | Sudoración | Transpiración excesiva sin causa evidente |
| Swelled lymph nodes | Ganglios inflamados | Aumento de tamaño de los ganglios linfáticos |
| Swollen blood vessels | Vasos inflamados | Dilatación o inflamación de los vasos sanguíneos |
| Swollen extremities | Extremidades hinchadas | Manos y pies inflamados |
| Swollen legs | Piernas hinchadas | Hinchazón por acumulación de líquidos |
| Swelling joints | Articulaciones inflamadas | Hinchazón en una o más articulaciones |
| Swelling of stomach | Abdomen hinchado | Distensión abdominal perceptible |
| Throat irritation | Irritación de garganta | Molestia, picor o dolor leve en la garganta |
| Toxic look | Aspecto tóxico | Apariencia enfermiza grave |
| Ulcers on tongue | Úlceras en la lengua | Lesiones o llagas dolorosas en la lengua |
| Unsteadiness | Inestabilidad | Incapacidad de mantener el cuerpo firme |
| Visual disturbances | Alteraciones visuales | Pérdida de agudeza visual, visión borrosa o puntos ciegos |
| Vomiting | Vómitos | Expulsión del contenido del estómago por la boca |
| Watering from eyes | Lagrimeo excesivo | Producción continua de lágrimas |
| Weakness in limbs | Debilidad en extremidades | Sensación de falta de fuerza en brazos o piernas |
| Weakness of one body side | Hemiparesia | Debilidad localizada en un lado del cuerpo |
| Weight gain | Aumento de peso | Aumento corporal sin causa aparente |
| Weight loss | Pérdida de peso | Disminución de peso involuntaria |
| Yellow crust ooze | Costra amarilla supurante | Secreción amarilla con formación de costras en la piel |
| Yellow urine | Orina amarilla intensa | Tono amarillento fuerte en la orina |
| Yellowing of eyes | Ictericia ocular | Color amarillento en la parte blanca de los ojos |
| Yellowish skin | Piel amarilla | Tono amarillento por acumulación de bilirrubina |

Tabla 1: Tabla de síntomas

|  |  |  |
| --- | --- | --- |
| **Enfermedad (EN)** | **Traducción (ES)** | **Descripción** |
| Acne | Acné | Trastorno de la piel con espinillas por poros obstruidos. |
| AIDS | Síndrome de inmunodeficiencia adquirida | Enfermedad causada por VIH que debilita el sistema inmunológico. |
| Alcoholic hepatitis | Hepatitis alcohólica | Inflamación del hígado por consumo excesivo de alcohol. |
| Allergy | Alergia | Reacción exagerada del sistema inmunológico a sustancias externas como polvo, polen o alimentos. |
| Arthritis | Artritis | Inflamación articular con dolor e hinchazón. |
| Bronchial Asthma | Asma | Enfermedad crónica con inflamación y estrechamiento de las vías respiratorias. |
| Cervical spondylosis | Espondilosis | Desgaste de vértebras que causa rigidez y dolor. |
| Chicken pox | Varicela | Enfermedad viral con fiebre y ampollas. |
| Chronic colestasis | Colestasis crónica | Trastorno hepático donde el flujo de bilis se ve afectado, causando ictericia y fatiga. |
| Common cold | Resfriado común | Infección viral leve del tracto respiratorio. |
| Dengue | Dengue | Virus transmitido por mosquitos que provoca fiebre y dolor articular. |
| Diabetes | Diabetes | Trastorno en el que el cuerpo no regula bien el azúcar en sangre. |
| Dimorphic hemorrhoids | Hemorroides | Inflamación de venas en el ano que causa dolor al defecar. |
| Drug reaction | Reacción a medicamentos | Reacción adversa del cuerpo a medicamentos que puede provocar erupciones u otros efectos secundarios. |
| Fungal infection | Infección fúngica | Infección causada por hongos que puede afectar a la piel, uñas, boca o pulmones. |
| Gastroenteritis | Gastroenteritis | Inflamación del estómago e intestinos por infección o intoxicación. |
| GERD | Reflujo gastroesofágico | Enfermedad en la que el ácido del estómago regresa al esófago causando acidez. |
| Heart attack | Infarto de miocardio | Obstrucción coronaria que daña el corazón. |
| Hepatitis A | Hepatitis A | Inflamación del hígado por virus, transmitido por comida o agua. |
| Hepatitis B | Hepatitis B | Infección hepática viral que puede ser crónica. |
| Hepatitis C | Hepatitis C | Puede volverse crónica y causar daño hepático severo. |
| Hepatitis D | Hepatitis D | Variante que solo aparece con hepatitis B. |
| Hepatitis E | Hepatitis E | Transmitida por agua contaminada, generalmente autolimitada. |
| Hiperthyroidism | Hipertiroidismo | Exceso de hormonas tiroideas. |
| Hyperthension | Hipertensión arterial | Elevación persistente de la presión arterial. |
| Hypoglycemia | Hipoglucemia | Nivel bajo de azúcar en sangre. |
| Hypothyroidism | Hipotiroidismo | Poca producción de hormonas tiroideas. |
| Impetigo | Impétigo | Infección bacteriana de la piel con llagas y costras. |
| Jaundice | Ictericia | Color amarillento en piel y ojos por bilirrubina alta. |
| Malaria | Malaria | Enfermedad parasitaria transmitida por mosquitos. |
| Migraine | Migraña | Dolor de cabeza intenso con sensibilidad a la luz y náuseas. |
| Osteoarthritis | Osteoartritis | Degeneración articular con dolor y rigidez. |
| Paroxysmal Positional Vertigo | Vértigo posicional paroxístico | Trastorno del oído que provoca mareo súbito. |
| Paralysis (Brain hemorrhage) | Parálisis (hemorragia cerebral) | Pérdida de movimiento por daño cerebral. |
| Peptic ulcer disease | Úlcera péptica | Úlceras en el revestimiento del estómago o intestino delgado por acidez. |
| Pneumonia | Neumonía | Infección pulmonar grave con fiebre y dificultad respiratoria. |
| Psoriasis | Psoriasis | Enfermedad autoinmune con placas rojas en la piel. |
| Thyroid | Fiebre tifoidea | Infección bacteriana con fiebre alta y síntomas digestivos. |
| Tuberculosis | Tuberculosis | Infección pulmonar bacteriana crónica. |
| Urinary tract infection | Infección del tracto urinario | Infección bacteriana de la vejiga o uretra. |
| Varicose veins | Várices | Venas dilatadas por mala circulación. |

Tabla 2: Tabla de enfermedades

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **index** | **prognosis** | **num\_grupos\_duplicados** | **num\_individuos\_duplicados** | **ratio\_individuos\_grupos** |
| 0 | (vertigo) Paroymsal Positional Vertigo | 7 | 114 | 16.29 |
| 1 | AIDS | 5 | 116 | 23.2 |
| 2 | Acne | 5 | 116 | 23.2 |
| 3 | Alcoholic hepatitis | 8 | 113 | 14.12 |
| 4 | Allergy | 5 | 116 | 23.2 |
| 5 | Arthritis | 6 | 115 | 19.17 |
| 6 | Bronchial Asthma | 7 | 114 | 16.29 |
| 7 | Cervical spondylosis | 6 | 115 | 19.17 |
| 8 | Chicken pox | 10 | 111 | 11.1 |
| 9 | Chronic cholestasis | 8 | 113 | 14.12 |
| 10 | Common Cold | 9 | 112 | 12.44 |
| 11 | Dengue | 10 | 111 | 11.1 |
| 12 | Diabetes | 9 | 112 | 12.44 |
| 13 | Dimorphic hemmorhoids(piles) | 6 | 115 | 19.17 |
| 14 | Drug Reaction | 6 | 115 | 19.17 |
| 15 | Fungal infection | 5 | 116 | 23.2 |
| 16 | GERD | 7 | 114 | 16.29 |
| 17 | Gastroenteritis | 5 | 116 | 23.2 |
| 18 | Heart attack | 5 | 116 | 23.2 |
| 19 | Hepatitis B | 9 | 112 | 12.44 |
| 20 | Hepatitis C | 7 | 114 | 16.29 |
| 21 | Hepatitis D | 10 | 111 | 11.1 |
| 22 | Hepatitis E | 9 | 112 | 12.44 |
| 23 | Hypertension | 6 | 115 | 19.17 |
| 24 | Hyperthyroidism | 9 | 112 | 12.44 |
| 25 | Hypoglycemia | 9 | 112 | 12.44 |
| 26 | Hypothyroidism | 8 | 113 | 14.12 |
| 27 | Impetigo | 6 | 115 | 19.17 |
| 28 | Jaundice | 9 | 112 | 12.44 |
| 29 | Malaria | 8 | 113 | 14.12 |
| 30 | Migraine | 10 | 111 | 11.1 |
| 31 | Osteoarthristis | 7 | 114 | 16.29 |
| 32 | Paralysis (brain hemorrhage) | 5 | 116 | 23.2 |
| 33 | Peptic ulcer diseae | 7 | 114 | 16.29 |
| 34 | Pneumonia | 9 | 112 | 12.44 |
| 35 | Psoriasis | 7 | 114 | 16.29 |
| 36 | Tuberculosis | 9 | 112 | 12.44 |
| 37 | Typhoid | 9 | 112 | 12.44 |
| 38 | Urinary tract infection | 5 | 116 | 23.2 |
| 39 | Varicose veins | 8 | 113 | 14.12 |
| 40 | hepatitis A | 9 | 112 | 12.44 |

Tabla 3: Relación entre el número de grupos de individuos y los individuos repetidos por enfermedad

|  |  |
| --- | --- |
| **Enfermedad** | **frecuencia** |
| fatigue | 1949 |
| vomiting | 1931 |
| high\_fever | 1374 |
| loss\_of\_appetite | 1162 |
| nausea | 1156 |
| headache | 1144 |
| abdominal\_pain | 1041 |
| yellowish\_skin | 920 |
| yellowing\_of\_eyes | 823 |
| chills | 805 |

Tabla 4: Síntomas más frecuentes

|  |  |
| --- | --- |
| **Enfermedad** | **frecuencia** |
| sunken\_eyes | 109 |
| dehydration | 109 |
| swollen\_blood\_vessels | 109 |
| watering\_from\_eyes | 109 |
| dischromic \_patches | 109 |
| extra\_marital\_contacts | 109 |
| spinning\_movements | 109 |
| weakness\_of\_one\_body\_side | 109 |
| foul\_smell\_of urine | 103 |
| fluid\_overload | 0 |

Tabla 5: Síntomas menos frecuentes

|  |  |  |
| --- | --- | --- |
| **Síntoma p-valor Chi²** | **Celdas esperadas <5** | **Información Mutua** |
| fatigue | 0.0 | 0.59401 |
| vomiting | 0.0 | 0.57322 |
| high\_fever | 0.0 | 0.53266 |
| loss\_of\_appetite | 0.0 | 0.51049 |
| nausea | 0.0 | 0.49869 |
| headache | 0.0 | 0.49490 |
| abdominal\_pain | 0.0 | 0.47698 |
| yellowish\_skin | 0.0 | 0.43417 |
| yellowing\_of\_eyes | 0.0 | 0.42888 |
| chills | 0.0 | 0.41419 |
| malaise | 0.0 | 0.39703 |
| skin\_rash | 0.0 | 0.39369 |
| chest\_pain | 0.0 | 0.38949 |
| joint\_pain | 0.0 | 0.37254 |
| itching | 0.0 | 0.37220 |
| sweating | 0.0 | 0.36966 |
| dark\_urine | 0.0 | 0.33443 |
| cough | 0.0 | 0.33101 |
| diarrhoea | 0.0 | 0.33072 |
| muscle\_pain | 0.0 | 0.31349 |

Tabla 6: Tabla con los 20 síntomas más representativos frente a las enfermedades

|  |  |
| --- | --- |
| **Síntoma** | **Importancia (L1)** |
| itching | 14.959.820 |
| fatigue | 12.730.857 |
| muscle\_pain | 12.708.376 |
| pain\_behind\_the\_eyes | 12.537.228 |
| rusty\_sputum | 12.536.503 |
| blood\_in\_sputum | 12.519.389 |
| vomiting | 11.953.070 |
| dark\_urine | 11.511.376 |
| high\_fever | 10.752.723 |
| nausea | 10.224.352 |
| chest\_pain | 9.697.261 |
| stomach\_bleeding | 9.508.180 |
| coma | 9.508.180 |
| abdominal\_pain | 9.445.842 |
| joint\_pain | 9.080.094 |
| diarrhoea | 8.778.944 |
| altered\_sensorium | 8.640.790 |
| mild\_fever | 8.245.500 |
| sweating | 8.211.355 |
| stomach\_pain | 8.093.724 |

Tabla 7: Resultado de importancia de síntomas aplicando la regresión logística de nivel 1

|  |  |
| --- | --- |
| **Síntoma** | **Importancia (DT)** |
| fatigue | 0.159956 |
| vomiting | 0.094410 |
| loss\_of\_appetite | 0.068822 |
| high\_fever | 0.061109 |
| skin\_rash | 0.054040 |
| abdominal\_pain | 0.041459 |
| headache | 0.039233 |
| chest\_pain | 0.035304 |
| yellowing\_of\_eyes | 0.032850 |
| painful\_walking | 0.030975 |
| malaise | 0.030647 |
| loss\_of\_balance | 0.028034 |
| irritability | 0.025880 |
| itching | 0.022606 |
| muscle\_pain | 0.020022 |
| neck\_pain | 0.015460 |
| irritation\_in\_anus | 0.014357 |
| coma | 0.012319 |
| chills | 0.012129 |
| palpitations | 0.011834 |

Tabla 8: Síntomas más relevantes según el árbol de decisión

|  |  |
| --- | --- |
| **Síntoma** | **Importancia (RF)** |
| muscle\_pain | 0.017958 |
| high\_fever | 0.016849 |
| itching | 0.016831 |
| altered\_sensorium | 0.015538 |
| mild\_fever | 0.015095 |
| abdominal\_pain | 0.014947 |
| loss\_of\_appetite | 0.014567 |
| family\_history | 0.014283 |
| dark\_urine | 0.013883 |
| fatigue | 0.013855 |
| chest\_pain | 0.013354 |
| nausea | 0.013063 |
| neck\_pain | 0.012497 |
| vomiting | 0.012294 |
| weight\_loss | 0.012061 |
| breathlessness | 0.012019 |
| diarrhoea | 0.011819 |
| rusty\_sputum | 0.011749 |
| sweating | 0.011610 |

Tabla 9: Tabla con los 20 síntomas más significativos según la técnica del Random Forest

|  |  |  |  |
| --- | --- | --- | --- |
| **Información mutua** | **Regresión Logística** | **Árbol de decisión** | **Random Forest** |
| fatigue | itching | fatigue | fatigue |
| vomiting | fatigue | vomiting | vomiting |
| high\_fever | muscle\_pain | loss\_of\_appetite | loss\_of\_appetite |
| loss\_of\_appetite | pain\_behind\_the\_eyes | high\_fever | high\_fever |
| nausea | rusty\_sputum | skin\_rash | skin\_rash |
| headache | blood\_in\_sputum | abdominal\_pain | abdominal\_pain |
| abdominal\_pain | vomiting | headache | headache |
| yellowish\_skin | dark\_urine | chest\_pain | chest\_pain |
| yellowing\_of\_eyes | high\_fever | yellowing\_of\_eyes | yellowing\_of\_eyes |
| chills | nausea | painful\_walking | painful\_walking |
| malaise | chest\_pain | malaise | malaise |
| skin\_rash | stomach\_bleeding | loss\_of\_balance | loss\_of\_balance |
| chest\_pain | coma | irritability | irritability |
| joint\_pain | abdominal\_pain | itching | itching |
| itching | joint\_pain | muscle\_pain | muscle\_pain |
| sweating | diarrhoea | neck\_pain | neck\_pain |
| dark\_urine | altered\_sensorium | irritation\_in\_anus | irritation\_in\_anus |
| cough | mild\_fever | coma | coma |
| diarrhoea | sweating | chills | chills |
| muscle\_pain | stomach\_pain | palpitations | palpitations |

Tabla 10: Tabla conjunta de todas las técnicas de selección de variables con los respectivos síntomas por orden de importancia

|  |  |  |
| --- | --- | --- |
|  | Y = 0 | Y = 1 |
| X = 0 | A | B |
| X = 1 | C | D |

Tabla 11: Tabla generada para calcular el coeficiente de correlación phi

Gráfico

El contenido generado por IA puede ser incorrecto.

Ilustración 12: Dendograma del clúster jerárquico de enfermedades usando los 132 síntomas

|  |  |  |
| --- | --- | --- |
| **Síntoma 1** | **Síntoma 2** | **Phi** |
| loss\_of\_appetite | yellowing\_of\_eyes | 0.77 |
| abdominal\_pain | dark\_urine | 0.66 |
| yellowing\_of\_eyes | dark\_urine | 0.61 |
| high\_fever | chills | 0.54 |
| abdominal\_pain | yellowing\_of\_eyes | 0.53 |
| chills | malaise | 0.52 |
| muscle\_pain | chills | 0.49 |
| abdominal\_pain | loss\_of\_appetite | 0.49 |
| vomiting | abdominal\_pain | 0.48 |
| fatigue | malaise | 0.47 |
| high\_fever | malaise | 0.46 |
| loss\_of\_appetite | dark\_urine | 0.46 |
| dark\_urine | joint\_pain | 0.45 |
| fatigue | high\_fever | 0.41 |
| coma | dark\_urine | 0.41 |
| loss\_of\_appetite | malaise | 0.41 |
| chest\_pain | malaise | 0.41 |
| loss\_of\_appetite | joint\_pain | 0.40 |
| muscle\_pain | headache | 0.38 |
| coma | joint\_pain | 0.37 |

Tabla 12: Coeficiente de correlación phi entre pares de síntomas

|  |  |
| --- | --- |
| **Síntoma** | **Contribución Comp.1** |
| coma | 0.315649 |
| dark\_urine | 0.146871 |
| yellowing\_of\_eyes | 0.097492 |
| loss\_of\_appetite | 0.087005 |
| abdominal\_pain | 0.080423 |
| joint\_pain | 0.072071 |
| malaise | 0.047576 |
| loss\_of\_balance | 0.037742 |
| itching | 0.027609 |
| fatigue | 0.018251 |

Tabla 13: Síntomas con mayor contribución a la componente 1 del MCA

|  |  |
| --- | --- |
| **Síntoma** | **Contribución Comp.2** |
| muscle\_pain\_ | 0.392786 |
| chills\_ | 0.150796 |
| headache\_ | 0.124575 |
| malaise\_ | 0.074891 |
| high\_fever\_ | 0.039446 |
| itching\_ | 0.033573 |
| abdominal\_pain\_ | 0.030856 |
| dark\_urine\_ | 0.030007 |
| diarrhoea\_ | 0.025447 |
| loss\_of\_balance\_ | 0.023238 |

Tabla 14: Síntomas con mayor contribución a la componente 2 del MCA

Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.

Ilustración 13: MCA por enfermedades usando 20 síntomas sin agrupar

Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.

Ilustración 14: MCA por enfermedades usando todos los síntomas sin agrupar

Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.

Ilustración 15: Biplot conjunto de enfermedades agrupadas y síntomas resultado del MCA

Diagrama

El contenido generado por IA puede ser incorrecto.

Ilustración 16: MCA de enfermedades agrupadas con círculos delimitando el área de cada grupo

Calendario

El contenido generado por IA puede ser incorrecto.

Ilustración 17: UMAP usando todos los síntomas del conjunto de datos original

Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.

Ilustración 18: Resultado logistic PCA usando todos los síntomas del dataset

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

Ilustración 19: Precisión del modelo según la profundidad escogida usando solo 20 síntomas

Tabla

El contenido generado por IA puede ser incorrecto.

Ilustración 20: Resultados del árbol de decisión usando todos los síntomas sin profundidad máxima

Tabla

El contenido generado por IA puede ser incorrecto.

Ilustración 21: Resultado de la aplicación del Random Forest usando los 20 síntomas más significativos y con 200 estimadores

Tabla

El contenido generado por IA puede ser incorrecto.

Ilustración 22: Resultado de Random Forest usando todos los síntomas

Tabla

El contenido generado por IA puede ser incorrecto.

Ilustración 23: Resultado del algoritmo XGBoost para la ejecución con los 20 síntomas más representativos