

Evidencia 1: Automatización de la gestión de ayuda

Valeria Arciga Valencia¹ y Abril Álvarez Mercado¹

^aInstituto Tecnológico y de Estudios Superiores de Monterrey

Andrés Saldaña Rodríguez

1. Introducción

En un contexto donde la automatización de servicios se ha posicionado como un componente estratégico en la transformación digital de las organizaciones, la implementación de sistemas expertos en ámbitos como el de la salud surge como respuesta a desafíos de operación.

En ese sentido, esta evidencia aborda la problemática de una empresa que desea implementar un servicio de diagnóstico basado en reglas que simule el criterio clínico de profesionales.

Para esto, se propone el desarrollo de este sistema basado en los contenidos del curso como agentes lógicos, lógica proposicional, inferencia y que realice diagnósticos médicos simulados con información de una base de datos.

2. Objetivos

Para el planteamiento del sistema basado en regla ha sido necesario el desarrollo de objetivos para el funcionamiento de la aplicación para la gestión de ayuda como se describen a continuación

- Diseño de un sistema y motor de inferencia que involucre el análisis de datos.
- Desarrollar un sistema experto capaz de interactuar mediante preguntas, procesar síntomas y emitir recomendaciones basadas en reglas.

3. Metodología

El enfoque y metodología adaptada para el desarrollo del sistema experto de orientación médica automatizada fue basada en principios de análisis de datos.

En primer lugar se hizo un procesamiento previo de la información, en donde se verifique la integridad del conjunto de datos, posterior a esto se realizó un análisis exploratorio de datos para la identificación de patrones clínicamente útiles que sean útiles como base para las reglas de inferencia.

Para el diseño e implementación del motor de inferencia ha sido necesario utilizar un enfoque de inferencia basada en reglas, de forma que se hiciera una filtración de información para obtener un diagnóstico lo más preciso posible.

Posterior al diseño, el proyecto tiene como finalidad la evaluación diagnóstica y la generación de recomendaciones a los usuarios, de modo que se observe la precisión del sistema. Al obtener un resultado, se realiza el desarrollo de la interfaz en Streamlit con el fin de crear una aplicación web interactiva, atractiva para el usuario y que principalmente sea intuitiva y fácil de utilizar.

4. Análisis Exploratorio de Datos

Siguiendo la metodología planteada anteriormente, y para el cumplimiento de los objetivos planteados para el desarrollo del

sistema, en primer lugar resulta necesario hacer una exploración de la base de datos para la identificación de información clave para el diseño de las reglas de inferencia o motor de inferencia y la forma de interacción del sistema con el usuario.

4.1. Descripción del dataset

El dataset se encuentra en formato .csv y contiene registros de consultas médicas. Cada fila representa un caso individual, en donde se identifica la enfermedad diagnosticada en la columna (diseases), y el resto de columnas son binarias indicando presencia (1) o ausencia (0) de síntomas.

Durante el análisis exploratorio de los datos se realizaron cálculos de estadísticos de la base de datos, en donde se obtuvo que la base cuenta con 246945 registros y 377 síntomas.

De esta información se ha obtenido que hay 773 enfermedades únicas, no existen valores nulos en la base de datos, en todas las columnas de síntomas existen valores numéricos, por lo que no se requirió hacer modificaciones a la base de datos.

Además de esto, se hace énfasis en que la cantidad de registros por enfermedad no es la misma, ya que existen enfermedades con un solo registro, lo que limita la información del sistema.

4.2. Análisis y visualizaciones

En primer lugar se analizó la distribución de cantidad de síntomas por enfermedad, en donde se halló que la distribución si bien no es normal, como se observa en la Figura 1 a continuación, hay un rango de síntomas de 2 a 12, lo que quiere decir un rango amplio para el diagnóstico de enfermedades.

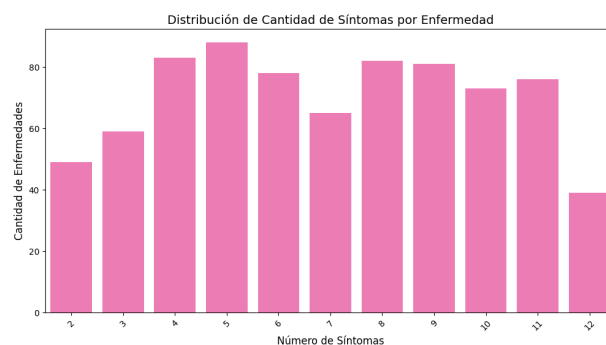


Figura 1. Distribución de cantidad de síntomas por enfermedad

Se realizó un diagrama de correlación como se observa en la Figura 2, entre síntomas para identificar grupos altamente asociados, en donde se observa que los síntomas de una misma región corporal están más asociados, además de que algunos síntomas se asocian a enfermedades comunes, como resfriado o afecciones respiratorias.

Además de este análisis, se realizaron histogramas de aparición de síntomas y de enfermedades como se observan en las Figuras 3 y 4 a continuación. En el caso de las enfermedades, la que tiene más registros es *cystitis* con 1219 registros, seguido

de *vulvodynia* y *nose disorder* con 1218 registros. Por otra parte, en cuanto a los síntomas, *sharp abdominal pain* apareció en 32307 del total de registros, seguido de *vomiting* con 27874 registros, lo que ha sido clave para priorizar las preguntas a realizar por el sistema experto.

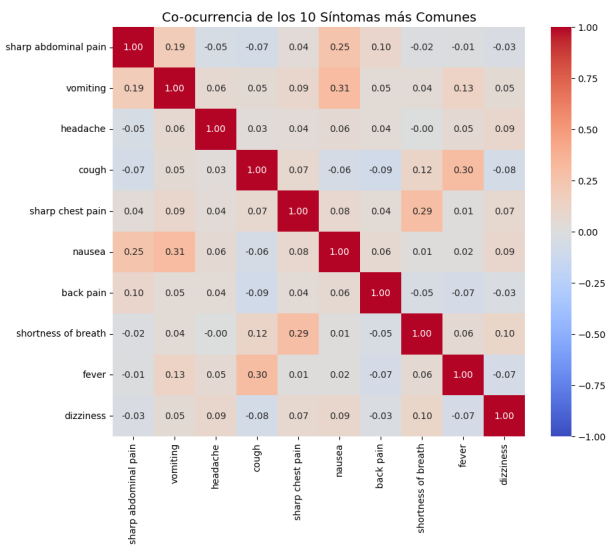


Figura 2. Correlación de síntomas más comunes

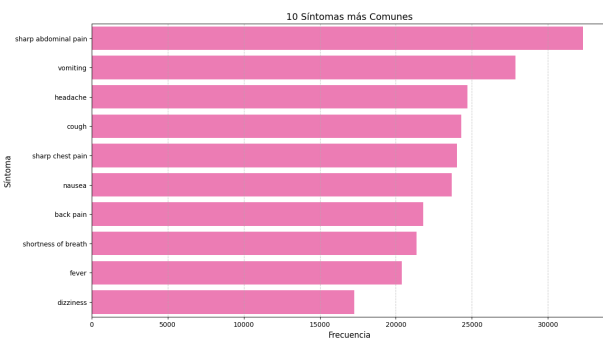


Figura 3. Síntomas más comunes en los registros de datos

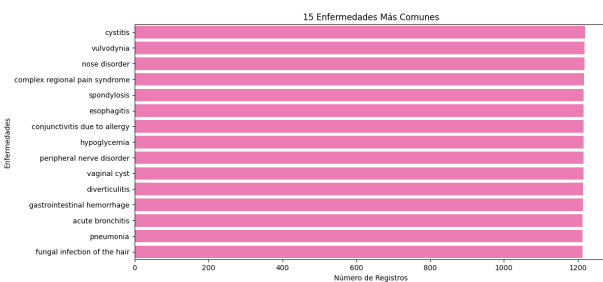


Figura 4. Enfermedades con más registros

5. Diseño e Implementación del Sistema Basado en Reglas

Considerando lo anterior, tomando en cuenta que la base de datos muestra que los síntomas se solapan entre enfermedades, se pensó en el desarrollo de un modelo dinámico, en donde se le diera importancia a los síntomas más frecuentes de cada enfermedad para hacer un diagnóstico más preciso.

En ese sentido, el sistema fue diseñado para ser una guía interactiva que, mediante preguntas estratégicas, haga una reducción de las posibles enfermedades de la base de datos y que realice un acercamiento a un diagnóstico asociado a los síntomas que presenta el usuario.

De este modo, es que el sistema utiliza un motor de inferencia en donde se emplean dos enfoques: Forward Chaining y Backward Chaining.

5.1. Forward Chaining

Dado que el concepto de un motor de inferencia Forward Chaining parte de los hechos para la obtención de conclusiones, se han desarrollado las funciones *best_symptom* y *update_candidates*.

En un inicio del diagnóstico, todas las enfermedades son posibles debido a que no se cuenta con información del usuario, por lo que la función tiene como finalidad el realizar un filtrado de enfermedades a partir de las respuestas del usuario.

El sistema sigue un proceso de selección de síntoma a preguntar, al obtener el resultado hace un filtrado de enfermedades y hacer la repetición hasta obtener un diagnóstico claro.

En primer lugar, la función *best_symptom* resulta ser la principal en el sistema al hallar el síntoma que maximiza la discriminación de enfermedades posibles de diagnosticar en el sistema.

Para cada síntoma no preguntado se realiza un cálculo de frecuencia relativa, se obtiene el puntaje de utilidad, a modo de priorizar los síntomas con una frecuencia cercana al 50 % para maximizar la información a obtener. La función elige el síntoma con mayor *score* con el fin de reducir las opciones de forma efectiva.

Esto resulta ser un motor de inferencia forward adaptativo, ya que dependiendo de las respuestas va a dar síntomas que reduzcan las opciones.

De este modo es que, el sistema inicia con preguntas sobre el padecimiento de síntomas que afectan al mayor número de enfermedades, por lo que, al obtener la información sobre si el paciente tiene o no una enfermedad, la función *update_candidates* va a realizar una filtración de enfermedades que tengan o no los síntomas que indique el paciente. Para esto, hay que considerar que se considera que los síntomas en una enfermedad son significativos al estar presentes en los registros con una frecuencia del 30 %

Esta función, en caso de obtener un síntoma padecido por el usuario, se va a enfocar en las enfermedades que consideren significativo ese síntoma; de otro modo, descartará esas enfermedades. Al final de ambos casos se hace una actualización de las enfermedades que son posibles diagnósticos para el usuario.

5.2. Backward Chaining

Por otra parte, igualmente se ha implementado un enfoque de motor de inferencia Backward Chaining, en donde a partir de una hipótesis se realiza una búsqueda para determinar si los hechos la justifican.

Esto se refleja en la función *find_specific_symptom*, en donde, posteriormente a la actualización de posibles diagnósticos para el usuario, y cuando quedan pocas enfermedades similares, se hace el cálculo de qué síntoma tiene una mayor variación entre los candidatos para preguntar de forma específica

por ese síntoma a modo de confirmar o descartar esa hipótesis. Este enfoque es el que hace que el sistema no realice preguntas genéricas y el diagnóstico sea más preciso.

5.3. Ventajas del diseño

La implementación de estas tres funciones es lo que da a lugar a la realización de preguntas en un inicio generales para la reducción de posibilidades de diagnóstico, un ajuste en el enfoque basándose en respuestas previas, y finalmente, al reducir significativamente los posibles diagnósticos, permite una diferenciación mediante síntomas poco frecuentes.

Este sistema se considera basado en reglas de forma implícita debido a las reglas de filtrado con la función *update_candidates* en donde se considera significativo un síntoma presente en el 30 % de los casos. La segunda regla es la de mejora de preguntas con la función *best_symptom* que selecciona los síntomas que dividan las enfermedades candidatas para una mayor ganancia de información. Finalmente, la tercera regla *find_specific_symptom* que prioriza síntomas con máxima diferencia en enfermedades candidatas en caso de que el primer filtrado no reduzca lo suficiente las opciones.

5.4. Entorno de pruebas

Para comprobar la efectividad del sistema basado en reglas se han elegido 50 enfermedades aleatorias de la base de datos agrupadas, de donde se han extraído todos los síntomas, con los que se han respondido las preguntas, de esta muestra se han obtenido resultados de un diagnóstico acertado más del 90 % de los casos, de modo que hay una demostración de la efectividad, que se asocia al motor de inferencia híbrido.

Algunas de estas enfermedades que se han obtenido como muestra para la prueba del sistema se agregan en el anexo del presente informe.

6. Desarrollo de la Aplicación en Streamlit

El sistema se construyó inicialmente en Python sin interfaz para la realización de ajustes en el funcionamiento de forma inmediata, posterior a esto, la aplicación, que se incluye en el repositorio se construyó y se adaptaron las funciones en *Streamlit*, un framework de *Python* para la creación de aplicaciones web interactivas.

En ese sentido, el código inicia con configuraciones clave, como la personalización del nombre de la pestaña, y posteriormente se implementó la estructura de la interfaz con su cabecera para las instrucciones del funcionamiento del sistema y el nombre del sistema, *Medical Diagnosis*, esto se ha realizado con el uso de HTML y CSS para la personalización de colores, además de la alineación.

El motor de inferencia, es decir, la lógica central del sistema se mantuvo idéntico, en donde se elige el síntoma general para las preguntas iniciales y el descarte de enfermedades incompatibles con la respuesta del usuario.

Para el almacenamiento de la información y el seguimiento del cuestionario se han empleado métodos para la gestión de estado y así mantener la lista dinámica de los posibles diagnósticos y determinar las funciones a emplear.

En la interfaz se colocó la dinámica de preguntas, en donde son generadas por las funciones descritas anteriormente, y el usuario puede interactuar con el sistema y responder a la

plataforma mediante botones de respuesta y con la posibilidad de reiniciar el diagnóstico.

Al finalizar el cuestionario, se muestran los diagnósticos, el principal en una ficha para destacar el que tiene una mayor coincidencia, y otros dos posibles resultados, todos incluyendo su porcentaje de coincidencia.

En ese sentido, se muestra la interfaz en funcionamiento en las figuras 5 y 6 a continuación.

Medical Diagnosis System

Welcome to the intelligent medical diagnosis system. Please answer the following questions to receive a preliminary analysis.

Do you have this symptom?

Sharp Abdominal Pain

Yes

No

Restart Diagnosis

⚠ This system is a diagnostic support tool. Does not replace professional medical consultation.

Figura 5. Interfaz inicial para el desarrollo de preguntas

Final Results

Primary Diagnosis

conjunctivitis due to allergy

Match: 33%

Other Possibilities

abdominal aortic aneurysm

0%

abdominal hernia

0%

Restart Diagnosis

Figura 6. Despliegue de resultados

7. Aplicaciones futuras y mejoras

El sistema actual es una versión funcional para el diagnóstico de enfermedades, sin embargo, hay oportunidades para ampliar el alcance de este proyecto.

En primer lugar se puede mencionar una ampliación de cobertura diagnóstica, una interfaz más avanzada en donde tenga mayor personalización por perfil del paciente e incluir variables como edad, género, historial médico, entre otros, a modo de ajustar las reglas para cada usuario. Además de

estas áreas de oportunidad para el desarrollo avanzado de este sistema, igualmente se considera reelevante la integración de modelos de aprendizaje automático para reforzar las reglas y mejorar la precisión diagnóstica.

Entre otros posibles escenarios para la mejora del sistema es una mayor base de datos, validación con expertos de la salud y una mejora en la interfaz en cuanto a recomendaciones para el diagnóstico recibido o realizar un chatbot.

8. Conclusiones

El desarrollo de este sistema experto ha dado lugar a la aplicación de los conocimientos adquiridos durante el curso *Análisis de sistemas basados en conocimiento*, en donde se ha demostrado efectividad al aplicar análisis estadístico con un motor de inferencia basado en el filtrado iterativo de síntomas.

La metodología implementada ha permitido generar diagnósticos preliminares con precisión, al mismo tiempo que reduce preguntas innecesarias.

El uso de Streamlit como plataforma para el desarrollo de la aplicación ha permitido una experiencia de usuario más fluida, con controles más intuitivos y que sea minimalista.

Finalmente, la lógica desarrollada y la documentación de este programa dan a lugar a agregar mejoras y expansiones del programa, como la posible integración de aprendizaje automático, personalización y el desarrollo de historiales clínicos. Este proyecto ha sido de gran utilidad para aplicar los conocimientos adquiridos en la introducción de sistemas basados en conocimiento.

Anexos

.1. Extracto de enfermedades empleadas para la prueba del sistema de diagnóstico

- **Enfermedad:** cystitis. **Síntomas:** retention of urine, suprapubic pain, sharp abdominal pain, painful urination, involuntary urination, frequent urination, lower abdominal pain, blood in urine, back pain, pelvic pain, side pain, symptoms of bladder
- **Enfermedad:** vulvodynia. **Síntomas:** sharp abdominal pain, nausea, painful urination, lower abdominal pain, vaginal discharge, back pain, pain during pregnancy, pelvic pain, burning abdominal pain, vaginal pain, side pain, cramps and spasms
- **Enfermedad:** nose disorder. **Síntomas:** sore throat, cough, nasal congestion, headache, facial pain, ear pain, fever, difficulty breathing, coryza, sinus congestion, painful sinuses, nosebleed
- **Enfermedad:** conjunctivitis due to allergy. **Síntomas:** cough, nasal congestion, diminished vision, symptoms of eye, pain in eye, eye redness, lacrimation, itchiness of eye, eye burns or stings, allergic reaction, swollen eye, sneezing
- **Enfermedad:** hypoglycemia. **Síntomas:** depressive or psychotic symptoms, dizziness, abnormal involuntary movements, fainting, feeling ill, nausea, problems with movement, weakness, decreased appetite, seizures, sleepiness, sweating
- **Enfermedad:** acute bronchitis. **Síntomas:** shortness of breath, sharp chest pain, sore throat, cough, nasal congestion, headache, wheezing, fever, difficulty breathing,

coughing up sputum, coryza, congestion in chest

- **Enfermedad:** pneumonia. **Síntomas:** shortness of breath, sharp chest pain, sore throat, cough, nasal congestion, vomiting, wheezing, weakness, fever, difficulty breathing, chills, coryza
- **Enfermedad:** fungal infection of the hair. **Síntomas:** skin swelling, abnormal appearing skin, skin lesion, acne or pimples, skin growth, irregular appearing scalp, pelvic pain, itching of skin, skin dryness, peeling, scaliness, or roughness, skin irritation, itchy scalp, skin rash

.2. Repositorio de GitHub

El análisis exploratorio de datos, código fuente del proyecto integrador implementado a Streamlit está disponible en el siguiente repositorio de GitHub, además de su versión beta en Streamlit:

<https://github.com/valeriaarciga/SistemaBasadoEnConocimiento.git>.

Autores: Valeria Arciga Valencia y Abril Álvarez Mercado