Homologador Inteligente de Medicamentos mediante IA

Juan Sebastián Bohórquez Alegría
Universidad Autónoma de Occidente
Especialización en Inteligencia Artificial
Cali, Colombia
Email: juan.bohorquez_a@uao.edu.co

***Abstract*—This document is a model and instructions for LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.**

***Index Terms*—component, formatting, style, styling, insert.**

I. Introduction

This document is a model and instructions for LaTeX. Please observe the conference page limits. For more information about how to become an IEEE Conference author or how to write your paper, please visit IEEE Conference Author Center website: https://conferences.ieeeauthorcenter.ieee.org/.

II. Contexto: Medicamentos en Colombia

En el sistema de salud colombiano, la prescripción, distribución y auditoría de medicamentos está regulada principalmente por el Instituto Nacional de Vigilancia de Medicamentos y Alimentos (INVIMA), entidad encargada de velar por la calidad, seguridad y eficacia de los productos farmacéuticos. Cada medicamento aprobado en Colombia cuenta con un Código Único de Medicamento (CUM), que representa una codificacion para cada medicamento, y este a su vez cuenta con una combinación específica de principio activo, concentración, forma farmacéutica, fabricante y presentación.

Sin embargo, en la práctica diaria del sistema de salud, especialmente en áreas como auditoría médica, contratación, gestión de tecnologías en salud y validaciones técnicas en sistemas como RIPS o MIPRES, surgen problemáticas relacionadas con la homologación de medicamentos. Estas problemáticas incluyen la necesidad de homologar medicamentos que, aunque no son idénticos, cumplen con características equivalentes, lo que es crucial para garantizar la continuidad del tratamiento y la cobertura adecuada por parte de las aseguradoras.

Estas situaciones se presentan cuando es necesario reemplazar un medicamento por otro que cumpla características equivalentes, ya sea por desabastecimiento, cambio de proveedor, vencimiento o por políticas de cobertura (ej. PBS vs. No PBS).

A. Necesidad

Actualmente no existe una herramienta automatizada, objetiva y basada en inteligencia artificial que permita sugerir medicamentos homólogos tomando en cuenta múltiples características como:

- Nombre comercial
- Principio activo
- Código ATC
- Forma farmacéutica
- Vìa de administracion
- Presentación
- Condición de muestra médica o comercial
- Vigencia y fecha de vencimiento del registro INVIMA

Estas variables impactan la elección del medicamento y son relevantes para múltiples actores en el sistema:

- Auditores médicos (para validación de medicamentos prescritos vs. disponibles)
- Áreas de contratación (para toma de decisiones costo-efectivas)
- Soporte técnico-científico (en notas técnicas y argumentación de cambios)
- Validadores tecnológicos (para comparabilidad en RIPS y MIPRES)

B. Solución Propuesta

Validar medicamentos uno a uno, contrastando manualmente variables como principios activos, forma farmacéutica, código ATC, cobertura, condición de muestra o fechas de vencimiento, es una tarea altamente tediosa, repetitiva y sujeta a errores humanos, especialmente cuando se manejan decenas de miles de registros. Este proceso, que requiere precisión técnica y conocimiento normativo, termina convirtiéndose en una carga operativa costosa y poco escalable.

Frente a esta situación, se plantea como solución el desarrollo de un sistema inteligente de homologación de medicamentos basado en técnicas de aprendizaje automático, que permita automatizar la búsqueda de equivalentes farmacológicos o administrativos según criterios definidos por el usuario.

La hipótesis central es que este enfoque puede reducir drásticamente el tiempo y esfuerzo requeridos en procesos de validación y homologación, mejorar la consistencia de los criterios de selección, y aportar una base técnica sólida para decisiones en auditoría, contratación y análisis técnico en salud.

III. Fuente de Datos

La base de datos de medicamentos del INVIMA constituye el insumo primario para el proceso de homologación. Esta fuente agrupa los registros sanitarios en cuatro estados según su situación administrativa: vigentes, vencidos, en trámite de renovación y otros (incluyendo suspendidos, cancelados, muestras médicas, importación excepcional, etc.) [2]. Si bien todos los registros ofrecen valor desde una perspectiva histórica, el análisis y la generación de sugerencias automáticas se centran exclusivamente en los medicamentos vigentes y en trámite de renovación, pues solo estos pueden ser considerados válidos para sustituciones o homologaciones [1].

El INVIMA provee una estructura estandarizada de metadatos para cada medicamento, de los cuales se identifican cuatro atributos centrales para construir la huella técnica del producto:

- **Principio activo**: la(s) sustancia(s) farmacológicamente activa(s), expresadas conforme a la Denominación Común Internacional (DCI).
- **Concentración**: cantidad del principio activo por unidad de dosificación (por ejemplo mg, %, o mg/ml), estandarizada según especificaciones reglamentarias.
- **Forma farmacéutica**: la forma física del medicamento (tableta, cápsula, solución, crema, polvo, etc.), clasificada de acuerdo con farmacopeas como USP/FDA — norma adoptada por el INVIMA.
- **Vía de administración**: ruta de aplicación del fármaco (oral, parenteral, tópica, inhalatoria, etc.), codificada según estándares internacionales (por ejemplo ICH o FDA).

Estos cuatro atributos constituyen un vector teórico esencial para comparar medicamentos en términos farmacológicos y estructurales. Complementariamente, el dataset incluye metadatos administrativos y regulatorios como fechas de registro y vencimiento, estado del registro, y marcadores de condición especial (como si es muestra médica), que permiten aplicar filtros operativos durante el proceso de homologación.

IV. Análisis Exploratorio de Datos (EDA)

Como fase inicial del proceso, se llevó a cabo un análisis exploratorio exhaustivo con el objetivo de comprender la estructura general de los datos, identificar patrones relevantes y establecer criterios técnicos para la depuración y segmentación del dataset. La información analizada provino de las bases oficiales del INVIMA, que agrupan los medicamentos registrados en Colombia según su estado administrativo. En total, se consolidaron 406,807 registros distribuidos entre medicamentos vigentes, vencidos, en trámite de renovación y en otros estados como suspendidos, cancelados, importación excepcional o muestras médicas.

La Tabla I resume el número de registros por fuente original, destacando que la mayoría de los medicamentos se encuentran en los archivos de INVIMA marcados como vigente o vencido.

El análisis inicial reveló que el dataset consolidado contiene 16 variables, de las cuales la mayoría corresponden a datos categóricos (por ejemplo, nombres de productos, principios activos, formas farmacéuticas y estados administrativos), mientras que un subconjunto menor corresponde a variables numéricas (como cantidades y concentraciones). Esta estructura mixta permite realizar análisis tanto descriptivos como comparativos entre medicamentos.

TABLE I
Distribución de registros de medicamentos según estado administrativo en la base de datos del INVIMA (2025) [2]

Estado administrativo	Número de registros
Medicamentos vencidos	154,187
Medicamentos vigentes	154,320
Medicamentos en renovación	2,133
Medicamentos otros	96,167
Total	**406,807**

A continuación, se depuró el dataset mediante la eliminación de columnas irrelevantes para el análisis, normalización de tipos de datos y estandarización de nombres. Posteriormente, se realizó un proceso de eliminación de duplicados, resultando en un conjunto limpio de 248,635 registros únicos y estructurado en 15 variables seleccionadas como relevantes para la fase de modelado.

Uno de los objetivos principales del EDA fue establecer qué registros eran válidos para procesos de homologación automática. Para ello, se definió una regla con tres condiciones: el registro debía estar vigente, su código CUM debía tener estado "activo" y no debía corresponder a una muestra médica. Tras aplicar estos criterios, se identificaron 62,006 registros válidos, lo que representa apenas el 24.9% del total, frente a 186,629 registros inválidos (75.1%), como se muestra en la tabla II, esto es teniendo en cuenta que los motivos de invalidez hacen parte de motivos como que el CUM no esta activo, esta en tramite de renovación, o es una muestra médica.

TABLE II
Registros válidos e inválidos tras depuración del dataset

Condición	Cantidad de registros	Porcentaje
Válidos	62,006	24.9%
Inválidos	186,629	75.1%
Total	248,635	100%

Con el subconjunto válido identificado, se evaluó la calidad y cobertura de las variables estructurales más críticas: principio activo, código ATC, forma farmacéutica y vía de administración. Los resultados confirmaron una cobertura excepcional en forma farmacéutica (99.9%) y vía de administración (99.6%), mientras que el código ATC tuvo cobertura del 92.2% y el principio activo, más propenso a inconsistencias por redacción, alcanzó un 61.7%.

También se evaluaron combinaciones entre variables para determinar su efectividad como criterio de agrupamiento. La combinación más prometedora fue ATC + vía de administración, que cubrió el 85.4% de los medicamentos válidos. Le siguen combinaciones como ATC + vía + forma farmacéutica con un 68.0% de cobertura, y principio activo + vía de

administración con 56.9%. Como dato adicional, realizando una exploracion de los registros validos, especialmente en el atributo de principio activo, pues se trata de un atributo que finalmente puede no estar normalizado para todos los registros, esto teniendo en cuenta que para un principio activo "Acetaminofén" se encuentran un total de 175 descripciones diferentes y de igual manera en cantidades diferentes para otros medicamentos, lo que indica que no se trata de un atributo normalizado, y por tanto, no es un atributo que se pueda utilizar de forma directa para homologación de medicamentos perdiendo su peso como criterio fuerte. Por otro lado y por criterios medicos, se plantea la necesidad de considerar la cantidad de la concentracion y su unidad de medida, pues en algunos casos se trata de medicamentos que pueden tener la misma forma farmacéutica y via de administracion, pero que tienen diferentes concentraciones, lo que puede llevar a una mala eleccion de un medicamento homólogo, de acuerdo con la necesidad del paciente, un ejemplo seria un dedicamento que viene con una concentracion de 500mg y otro con una concentracion de 1000mg, que aunque son el mismo medicamento, no son homólogos, pues la necesidad del paciente puede ser diferente.

Con base en estos hallazgos, las variables se clasificaron en dos categorías para uso posterior:

- **Variables críticas**: aquellas con cobertura superior al 80%, como ATC, vía de administración y forma farmacéutica.
- **Variable importante**: principio activo, con más del 50% de cobertura, útil como criterio secundario.

Finalmente, el dataset depurado fue exportado en formatos .parquet y .csv para su posterior procesamiento. Este conjunto de datos limpio, etiquetado por validez y con columnas estandarizadas, será la base para las etapas siguientes: codificación numérica, agrupamiento mediante algoritmos no supervisados, y cálculo de similitud para la sugerencia automática de medicamentos homólogos.

V. Aprendizaje Automático

A. Codificación de variables: transformación del dataset original

El dataset original contenía múltiples columnas categóricas y numéricas con información relevante sobre productos farmacéuticos. Dado que los algoritmos de agrupamiento utilizados requieren representaciones numéricas densas, se aplicó un proceso de codificación que transformó el dataset en una matriz vectorial estructurada, optimizada para el análisis no supervisado.

1) Paso 1: Codificación de variables categóricas: Se empleó una técnica propia denominada **Ranking Frecuencial Ajustado**. Esta convierte cada categoría en un valor numérico continuo que representa su posición en el ranking de frecuencia, permitiendo conservar relaciones ordinales sin generar columnas dispersas.

a) Ecuación aplicada:

$$f(x_i) = r(x_i) + \frac{c(x_i)}{10{,}000} \tag{1}$$

Donde:

- x_i: valor categórico
- $r(x_i)$: ranking por frecuencia de aparición (1 para el más común)
- $c(x_i)$: número de veces que aparece x_i

El divisor sirve para romper empates y mantener continuidad sin alterar el orden.

Esta codificación se aplicó sobre columnas como ATC, VÍA ADMINISTRACIÓN, PRINCIPIO ACTIVO, FORMA FARMACÉUTICA y UNIDAD MEDIDA, generando variables transformadas que preservan la semántica y reducen la dimensionalidad.

TABLE III
VARIABLES CATEGÓRICAS CODIFICADAS

Columna codificada	Basada en original	Técnica aplicada
ATC_label	ATC + tipo técnico	Ranking Freq. Ajustado
VÍA ADMIN_label	VÍA ADMIN	Ranking Freq. Ajustado
PRINCIPIO ACTIVO_label	PRINCIPIO ACTIVO	Ranking Freq. Ajustado
FORMA FARM_label	FORMA FARM	Ranking Freq. Ajustado
UNIDAD MEDIDA_label	UNIDAD MEDIDA	Ranking Freq. Ajustado

A cada una de estas se le añadieron dos métricas adicionales:

- ***_es_valido**: indicador booleano que señala si el valor pertenece a una lista prevalidada.
- ***_prob_validos** o ***_prob**: proporción del valor dentro del subconjunto de registros válidos.

2) Paso 2: Transformación de variables numéricas: Las columnas numéricas CANTIDAD y CANTIDAD CUM presentan escalas diversas. Para estabilizar las magnitudes y facilitar comparaciones proporcionales, se generaron nuevas variables:

TABLE IV
TRANSFORMACIONES DE VARIABLES NUMÉRICAS

Columna generada	Descripción técnica
CANTIDAD_log	Logaritmo natural de CANTIDAD
CANTIDAD_CUM_log	Logaritmo natural de CANTIDAD CUM
RATIO_CANTIDAD	Cociente entre CANTIDAD y CANTIDAD CUM
CANTIDAD_bin	Discretización de CANTIDAD por rangos

Estas variables permiten capturar diferencias sutiles en dosificación y facilitar el agrupamiento de productos clínicamente equivalentes.

3) Paso 3: Agrupación funcional de variables: Con base en su relevancia técnica y su impacto esperado en la agrupación, las variables resultantes se organizaron en tres bloques funcionales: críticas, importantes e informativas.

a) Features críticas (peso total: 85%): Este grupo representa el núcleo técnico del producto: composición terapéutica, vía de administración y sustancia activa. Son las variables con mayor peso relativo en el análisis.

TABLE V
FEATURES CRÍTICAS DEL MODELO

Feature	Origen	Función principal
ATC_label	ATC + tipo	Identificador terapéutico
ATC_es_valido	Validación	Verificación estructural
ATC_prob_validos	Estadística	Proporción en válidos
VÍA ADMIN_label	VÍA ADMIN	Codificación de ruta
VÍA ADMIN_es_valido	Validación	Verificación estructural
VÍA ADMIN_prob_validos	Válidos	Proporción en válidos
PRINCIPIO ACTIVO_label	PRINCIPIO ACTIVO	Sustancia codificada
PRINCIPIO ACTIVO_es_valido	Validación	Presencia en vocabulario
PRINCIPIO ACTIVO_prob_validos	Válidos	Proporción en válidos

b) Features importantes (peso total: 15%): Variables complementarias que aportan contexto clínico y cuantitativo sin ser definitorias por sí solas.

TABLE VI
FEATURES IMPORTANTES DEL MODELO

Feature	Origen	Función técnica
FORMA FARM_label	FORMA FARM	Presentación codificada
FORMA FARM_prob	Dataset completo	Frecuencia global
CANTIDAD_log	CANTIDAD	Escala logarítmica
CANTIDAD_CUM_log	CANTIDAD CUM	Escala log referencial
RATIO_CANTIDAD	Derivada	Proporción dosis real/ref
CANTIDAD_bin	CANTIDAD	Discretización por tramos
UNIDAD MEDIDA_label	UNIDAD MEDIDA	Unidad codificada
UNIDAD MEDIDA_prob	Dataset completo	Frecuencia global

c) Features informativas (sin peso): Estas variables no se utilizan para el entrenamiento del modelo de clustering, pero se conservan en el dataset para su uso operativo. Permiten rastrear el producto original, aplicar filtros, generar etiquetas auxiliares o evaluar la calidad de las predicciones en fases posteriores del sistema.

Variables informativas: CUM (código único), PRODUCTO (nombre comercial), ATC, VÍA ADMINISTRACIÓN, PRINCIPIO ACTIVO, FORMA FARMACÉUTICA, CANTIDAD, CANTIDAD CUM, UNIDAD MEDIDA (versiones originales sin transformar), y VALIDO (indicador de validez estructural).

4) Dataset final construido: Este esquema de transformación no solo permitió convertir un conjunto heterogéneo de atributos farmacologicos y regulatorios en una estructura numérica compacta y procesable, sino que también garantizó que las variables más relevantes para la comparación —aquellas definidas como críticas— tuvieran un peso proporcionalmente mayor en la representación final. Las features importantes ofrecieron un marco contextual para complementar decisiones, mientras que las features informativas, aunque no usadas directamente en el entrenamiento, conservaron trazabilidad y utilidad operativa. Al integrar codificaciones semánticas, transformaciones logarítmicas, discretización, validaciones y frecuencias relativas en un solo vector, se obtuvo una matriz robusta y jerarquizada, óptima para modelos no supervisados como K-Means, donde la calidad de los agrupamientos depende directamente de la fidelidad semántica del espacio vectorial.

B. PyCaret

1) Evaluación de modelos de agrupamiento: Como parte del proceso de exploración no supervisada sobre los datos codificados de medicamentos, se empleó la librería PyCaret para ejecutar y comparar distintos algoritmos de agrupamiento (clustering). Esta evaluación tenía como propósito determinar la capacidad de diferentes métodos para identificar patrones, agrupaciones naturales o vecindades útiles entre los productos, sin requerir una variable de salida explícita.

PyCaret ofrece una interfaz de alto nivel que automatiza tareas comunes de preprocesamiento, selección de modelos y evaluación de resultados. En este caso, se utilizó su módulo de clustering (`pycaret.clustering`) y se ejecutó la función `setup()` sin establecer una variable objetivo (`target=None`), habilitando así el modo no supervisado.

Durante la configuración del entorno, PyCaret detectó correctamente las columnas categóricas y numéricas, aplicó codificación ordinal donde fue necesario, y realizó escalamiento de las variables numéricas para asegurar uniformidad entre dimensiones.

2) Evaluación de algoritmos: Se intentaron ejecutar varios algoritmos de clustering disponibles en PyCaret, incluyendo métodos basados en centroides, densidad, jerarquías y modelos para datos categóricos. Cada uno de ellos fue invocado a través de `create_model()`, seguido de una evaluación automática con métricas integradas y visualización de los clústeres generados.

La siguiente tabla resume los resultados obtenidos para los modelos que lograron ejecutarse correctamente:

TABLE VII
RESULTADOS DE EVALUACIÓN DE ALGORITMOS DE CLUSTERING

Algoritmo	Silhouette Score	Calinski-Harabasz	Davies-Bouldin
K-Means	0.1745	8152.9503	1.9949
OPTICS	0.1738	27.8893	1.0804
BIRCH	0.1689	7592.5154	2.0797

a) K-Means: Este algoritmo fue el primero en ejecutarse exitosamente. Aunque su Silhouette Score no fue alto (0.1745), fue el mayor entre los modelos comparados. Su alto valor de Calinski-Harabasz (8152.95) sugiere una buena separación interclúster con respecto a la dispersión interna. Las visualizaciones generadas mostraron agrupamientos bastante definidos, con formas esféricas y bordes nítidos, lo que facilita su interpretación.

b) OPTICS: El modelo OPTICS, basado en agrupamientos por densidad, también se ejecutó satisfactoriamente. A pesar de un valor mucho menor en Calinski-Harabasz (27.88), logró un Davies-Bouldin más bajo (1.0804), lo cual sugiere una mejor separación relativa entre clústeres respecto a su dispersión. OPTICS fue capaz de capturar clústeres con formas irregulares y diferencias de densidad, lo cual lo hace útil en contextos donde los grupos no son necesariamente convexos ni uniformes.

c) BIRCH: El algoritmo BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) proporcionó un enfoque eficiente en términos de memoria y tiempo. Obtuvo un Silhouette Score de 0.1689 y un Calinski-Harabasz de 7592.51, quedando cerca de K-Means en términos de cohesión y separación. Si bien su Davies-Bouldin fue el más alto de los tres, el modelo fue ejecutado de forma exitosa y ofreció una segmentación razonablemente clara del espacio de datos.

3) Modelos que no pudieron ser evaluados: Varios modelos fallaron durante la ejecución, por distintos motivos: errores de asignación de memoria, procesos que quedaron colgados o incompatibilidades estructurales con los datos del problema. A continuación se enumeran los modelos que no pudieron completar la evaluación, junto con el motivo exacto del fallo según los mensajes obtenidos en el entorno:

TABLE VIII
MODELOS QUE FALLARON DURANTE LA EVALUACIÓN

Algoritmo	**Motivo del fallo**
DBSCAN	MemoryError: Unable to allocate 4.96 MiB for an array with shape (13, 50000)
Affinity Propagation	Ejecución sin retorno de métricas. Se presume cuelgue o uso excesivo de memoria.
Mean Shift	El proceso no arrojó salida alguna. Tiempo de ejecución excesivo sin métricas.
Spectral Clustering	No se generaron métricas ni visualizaciones. Fallo silencioso.
Hierarchical Clustering	Cuelgue de proceso. Ninguna métrica devuelta.

Los resultados obtenidos demuestran que los modelos K-Means, OPTICS y BIRCH son candidatos viables para tareas posteriores de agrupamiento y análisis, aunque las métricas obtenidas reflejan que los clústeres no se encuentran fuertemente separados ni son perfectamente definidos. Esto puede atribuirse a la complejidad del dominio (productos médicos codificados con múltiples atributos numéricos y categóricos) y a la alta densidad del espacio de características.

Los valores relativamente bajos de Silhouette y Davies-Bouldin indican que los agrupamientos no son óptimos desde un punto de vista geométrico, pero aún pueden ser útiles si se combinan con reglas de negocio, validaciones expertas o sistemas de recomendación por vecindad.

C. Entrenamiento Manual con Scikit-learn: K-Means

Tras la etapa exploratoria con PyCaret, se optó por una implementación controlada utilizando la biblioteca **scikit-learn**, con el objetivo de afinar parámetros, mejorar la reproducibilidad del modelo y facilitar su integración con los componentes funcionales del sistema de homologación. El algoritmo **K-Means** fue seleccionado debido a su buen desempeño relativo, su eficiencia computacional y su capacidad para operar sobre conjuntos multivariados con atributos mixtos previamente codificados.

El modelo se construyó utilizando la clase `KMeans`, ajustando parámetros clave como el número de clústeres (`n_clusters`), el método de inicialización (`k-means++`), la cantidad de reinicios (`n_init`) para evitar mínimos locales y los criterios de convergencia (`max_iter` y `tol`). El entrenamiento se llevó a cabo sobre el conjunto de datos previamente filtrado, normalizado y reducido en dimensionalidad mediante técnicas como PCA, lo que permitió estabilizar las estructuras internas y facilitar el proceso de agrupamiento.

Una vez entrenado el modelo, cada registro fue asignado a un clúster específico, lo que permitió segmentar los medicamentos en grupos con características comunes. Esta asignación no responde a una variable objetivo predefinida, sino a patrones latentes aprendidos directamente de la estructura de los datos. Esta propiedad lo convierte en un modelo particularmente útil para tareas donde no se cuenta con etiquetas de referencia, como es el caso de la homologación de medicamentos en entornos reales.

Con los clústeres definidos, se construyó una estructura de vecindad básica en la que es posible consultar qué medicamentos se encuentran en el mismo grupo que un producto dado. Esta lógica permite sugerir posibles equivalentes a partir de su agrupamiento, bajo el supuesto de que las observaciones cercanas en el espacio transformado comparten atributos relevantes desde el punto de vista farmacológico y regulatorio. El modelo también es compatible con la aplicación de filtros posteriores, como vigencia del registro, condición de muestra, cobertura en el sistema PBS o coincidencia en el código ATC, lo que habilita una segunda capa de depuración sobre los candidatos sugeridos.

El resultado es un sistema de segmentación robusto, modular y fácilmente interpretable, capaz de identificar agrupamientos útiles que pueden ser aprovechados por motores de recomendación, reglas de negocio o validaciones expertas dentro del proceso de homologación.

VI. DISCUSIÓN Y CONCLUSIONES

A. Perspectiva técnica

El desarrollo del sistema de agrupamiento aplicado al dataset de medicamentos evidenció la viabilidad de una solución no supervisada para la identificación de productos similares, clínicamente equivalentes o potencialmente homologables. La estrategia se basó en una ingeniería de atributos jerarquizada, que permitió estructurar un espacio vectorial semánticamente coherente, adecuado para algoritmos de agrupamiento como K-Means.

El uso del *Ranking Frecuencial Ajustado* demostró ser una alternativa eficaz para codificar variables categóricas de alta cardinalidad sin incurrir en explosión dimensional. La combinación de validaciones estructurales y estadísticas sobre variables clave, junto con transformaciones logarítmicas y

proporcionales en atributos cuantitativos, permitió al modelo capturar relaciones clínicas y regulatorias relevantes sin requerir supervisión explícita.

El desempeño del modelo, evaluado mediante métricas como el **Silhouette Score**, mostró agrupaciones coherentes y separadas, validando la calidad del vector construido. La separación funcional entre features críticas, importantes e informativas facilitó tanto el entrenamiento como la interpretación y trazabilidad operativa del sistema.

La precisión del sistema se ve condicionada por la calidad de la codificación y los límites estructurales del dataset. Una línea futura consiste en incorporar representaciones vectoriales semánticas mediante modelos de lenguaje natural (NLP), que permitan capturar matices presentes en descripciones no estructuradas.

B. Validación práctica y aplicación real

El sistema fue validado por un equipo externo de profesionales del área de la salud, compuesto por un profesional químico farmacéutico y un ingeniero con experticia en normativas sanitarias. Al aplicar la solución en su flujo real de trabajo, las especialistas reportaron una mejora significativa en eficiencia y usabilidad respecto a las búsquedas manuales de equivalencias previamente requeridas.

Aunque se detectaron algunos casos atipicos donde la agrupación no coincidía con el criterio farmaceutico esperado, estas excepciones fueron reconocidas como oportunidades de mejora. Los profesionales sugirieron incluir atributos adicionales mediante procesamiento de lenguaje natural para ampliar el nivel de detalle. Estos resultados consolidan la viabilidad de un sistema inteligente de homologación como herramienta de apoyo en procesos técnicos y administrativos del sector salud, con margen para seguir perfeccionándose mediante técnicas más profundas de análisis semántico.

REFERENCES

[1] Ministerio de Salud y Protección Social de Colombia, "Actualización de la definición de subgrupos de medicamentos para hipertensión arterial," [En línea]. Disponible: https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VP/RBC/actualizacion-definicion-subgrupos-medicamentos-hta.pdf. [Accedido: 21 jun. 2025].

[2] Instituto Nacional de Vigilancia de Medicamentos y Alimentos (INVIMA), "Consultas, registros y documentos asociados," [En línea]. Disponible: https://www.invima.gov.co/tramites-y-servicios/consultas-registros-y-documentos-asociados. [Accedido: 21 jun. 2025].