Aplicación de la metodología CRISP-DM para la identificación temprana de anomalías en eventos de Infección Respiratoria Aguda Grave (IRAG) Inusitada en Bogotá

Rubén Darío Bermúdez Másmela

Asesor

Camilo Enrique Romero Parra

.

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2026

Dedicatoria

Dedico este trabajo a mi esposa, pilar fundamental en mi desarrollo profesional y personal, su ejemplo de constancia y entrega en cada aspecto de la vida me incentivó a seguir fortaleciendo mis conocimientos y a ocupar el lugar en el que con orgullo estoy hoy.

Agradecimientos

Deseo expresar mi más profundo agradecimiento a todas las personas e instituciones que hicieron posible la culminación de este proyecto de especialización:

A mi esposa: Por ser mi apoyo incondicional, por su paciencia durante las largas jornadas de estudio y por ser la motivación principal para alcanzar mis metas profesionales. Su presencia es el motor de cada uno de mis logros.

A la Universidad Nacional Abierta y a Distancia (UNAD): Por brindarme las herramientas y el conocimiento necesario para evolucionar en mi carrera profesional y por fomentar espacios de investigación.

A mis tutores: Quienes fueron la guía técnica que me permitieron darle forma a esta investigación.

A Amenadiel y Samael, mis compañeros de cuatro patas: Por su silenciosa pero constante compañía durante los días y noches de programación y análisis de datos.

Resumen

El presente proyecto de grado desarrolla un sistema de vigilancia sindrómica inteligente para la identificación temprana de anomalías epidemiológicas por Infecciones Respiratorias Agudas (IRA) en Bogotá, abarcando el periodo 2009-2024. Ante las limitaciones de los métodos estadísticos tradicionales, que suelen ser reactivos y dependientes de grandes volúmenes de casos, esta investigación propone un cambio de paradigma hacia una vigilancia basada en el riesgo y el perfil demográfico.

Bajo la metodología CRISP-DM, el estudio integró técnicas avanzadas de ciencia de datos en tres dimensiones críticas:

Normalización Estratégica: Se neutralizó el sesgo histórico de reportes masivos en adultos mediante una ponderación por relevancia etaria, permitiendo que el sistema priorice la vulnerabilidad en la primera infancia e infancia.

Modelado y Benchmarking: Se implementó el algoritmo de aprendizaje no supervisado Isolation Forest, validando su robustez mediante una comparación (benchmarking) con el modelo Local Outlier Factor (LOF). Este proceso permitió identificar un core de inestabilidad de alta confianza y caracterizar 886 eventos anómalos que rompen la estacionalidad y el perfil esperado en la ciudad.

Validación Operativa: Al contrastar los resultados con el estándar epidemiológico tradicional (+2σ), el modelo demostró una precisión del 18.06% en la captura de picos de volumen, pero, fundamentalmente, reveló un 82% de alertas adicionales invisibles para la estadística convencional. Los resultados culminan en la identificación de Nodos y periodos centinela, proporcionando a la Secretaría Distrital de Salud una herramienta de auditoría dirigida y optimización de recursos. En conclusión, el sistema no solo identifica brotes conocidos, sino

que actúa como un centinela preventivo que detecta rupturas silenciosas en la firma epidemiológica, fortaleciendo la toma de decisiones y la protección de la población pediátrica en el Distrito Capital.

Palabras claves: Vigilancia Sindrómica, Aprendizaje No Supervisado, Isolation Forest, CRISP-DM, Salud Pública, IRA, Anomalías Epidemiológicas, Bogotá.

Abstract

This capstone project develops an intelligent syndromic surveillance system for the early identification of epidemiological anomalies in Acute Respiratory Infections (ARI) in Bogotá, covering the 2009-2024 period. Addressing the limitations of traditional statistical methods—which are often reactive and dependent on high case volumes—this research proposes a paradigm shift toward risk-based surveillance focused on demographic profiles.

Following the CRISP-DM methodology, the study integrated advanced data science techniques across three critical dimensions:

Strategic Normalization: Historical bias from mass adult reporting was neutralized through age-relevance weighting, enabling the surveillance system to prioritize vulnerability in early childhood and childhood stages.

Modeling and Benchmarking: The Isolation Forest unsupervised learning algorithm was implemented and its robustness validated through benchmarking against the Local Outlier Factor (LOF) model. This process identified a high-confidence "Instability Core" and characterized 886 anomalous events that deviate from the city's expected seasonality and profiles.

Operational Validation: When cross-referencing results with the conventional epidemiological standard (+2σ), the model demonstrated 18.06% precision in capturing volume peaks. Crucially, it revealed an 81.94% margin of additional alerts that remain invisible to conventional statistics.

The results culminate in the identification of Critical Nodes and sentinel periods providing the District Health Secretariat with a targeted audit tool and resource optimization strategy. In conclusion, the system not only identifies known outbreaks but acts as a preventive sentinel detecting silent ruptures in the epidemiological signature, strengthening decision-making

and the protection of the pediatric population in Bogotá.

Keywords: *Syndromic Surveillance, Unsupervised Learning, Isolation Forest, CRISP-DM, Public Health, ARI, Epidemiological Anomalies, Bogotá.*

Tabla de Contenido

Lista de Figuras

Introducción

La Infección Respiratoria Aguda Grave IRAG Inusitada representa uno de los desafíos más persistentes para el sistema de salud pública en Bogotá, impactando de manera desproporcionada a la población pediátrica y generando una carga operativa significativa en la red hospitalaria. Tradicionalmente, la vigilancia epidemiológica se ha basado en métodos estadísticos descriptivos, como el Canal Endémico, los cuales dependen de la acumulación de grandes volúmenes de casos para activar alertas. Sin embargo, este enfoque reactivo presenta limitaciones críticas, principalmente su incapacidad para identificar anomalías sutiles en la composición demográfica de los afectados antes de que ocurra una saturación del sistema.

En el contexto de una ciudad con dinámicas poblacionales tan complejas como Bogotá, la transición hacia una vigilancia sindrómica inteligente no es solo una oportunidad tecnológica, sino una necesidad operativa. El presente proyecto propone el desarrollo de un sistema basado en algoritmos de aprendizaje no supervisado (Unsupervised Learning) para detectar desviaciones atípicas en los reportes de IRAG registrados entre 2009 y 2024. A diferencia de los métodos convencionales, esta propuesta utiliza técnicas de ingeniería de datos para priorizar el riesgo etario, permitiendo que el sistema aísle casos de alta vulnerabilidad que suelen quedar diluidos en las estadísticas generales.

Utilizando la metodología CRISP-DM, esta investigación recorre desde el análisis de la consistencia de los datos históricos hasta la implementación de modelos como Isolation Forest y Local Outlier Factor. El enfoque principal no es solo cuantificar el número de enfermos, sino caracterizar la firma del riesgo: identificar quiénes, cuándo y dónde se están produciendo reportes que rompen la normalidad estadística.

Finalmente, este documento detalla cómo la integración de la Ciencia de Datos permite a

la Secretaría Distrital de Salud transitar hacia un modelo de auditoría dirigida y preventiva. Al focalizar los esfuerzos en los denominados "Nodos Críticos" y grupos de riesgo identificados por la Inteligencia Artificial, se fortalece la capacidad de respuesta institucional, garantizando una protección más efectiva y oportuna para la población más vulnerable de la capital.

Justificación

La vigilancia epidemiológica de las Infecciones Respiratorias Agudas (IRA) en Bogotá ha dependido históricamente de umbrales de volumen masivo. Sin embargo, la efectividad de este modelo se ve comprometida cuando el riesgo no se manifiesta por la cantidad de reportes, sino por la naturaleza atípica de los mismos. La presente investigación se justifica en la necesidad de dotar a la salud pública de herramientas analíticas que superen la observación pasiva y permitan una intervención de precisión.

Desde una perspectiva técnica y de Ciencia de Datos, este proyecto es fundamental porque aborda el problema del ruido estadístico. En una base de datos masiva de 15 años, los casos pediátricos suelen quedar ocultos bajo el volumen dominante de la población adulta. Mediante la implementación de algoritmos de aprendizaje no supervisado y técnicas de normalización etaria, se justifica el uso de la Inteligencia Artificial como un filtro inteligente capaz de extraer señales de riesgo temprano que hoy pasan desapercibidas.

Desde el enfoque de la Ingeniería Industrial y la gestión pública, la relevancia de este estudio radica en la optimización de recursos. Las entidades gubernamentales cuentan con capacidades limitadas de auditoría e inspección de campo. Justificar este modelo permite a la Secretaría Distrital de Salud pasar de una vigilancia reactiva y generalizada a una estrategia de auditoría por excepción. Al identificar "Nodos Críticos" de manera automática, se garantiza que los recursos técnicos y humanos se desplieguen donde la evidencia analítica sugiere una anomalía real, maximizando el impacto de las políticas de prevención.

Finalmente, la justificación social es imperativa. La detección temprana de una ruptura en el perfil epidemiológico, especialmente en meses críticos como marzo, tiene una correlación directa con la reducción de complicaciones graves y mortalidad en la primera infancia. Este

trabajo no solo aporta un avance académico en el uso de metodologías como CRISP-DM, sino que propone una solución ética y tecnológica para proteger de manera más inteligente el bienestar de los ciudadanos más vulnerables de Bogotá.

Pregunta Problema

¿Cómo desarrollar un modelo de vigilancia sindrómica basado en algoritmos de aprendizaje automático que permita la identificación de patrones anómalos en los eventos de salud de Bogotá D.C., para la generación de alertas tempranas de brotes epidémicos?

Objetivos

Objetivo General

Desarrollar un sistema de vigilancia sindrómica mediante algoritmos de aprendizaje automático para la identificación temprana de eventos epidemiológicos por Infecciones Respiratorias Agudas Graves (IRAG) Inusitada en Bogotá, utilizando datos históricos de salud pública para fortalecer la toma de decisiones preventivas.

Objetivos Específicos

Analizar la consistencia de los registros de eventos de salud en Bogotá, proveniente de la plataforma de Datos Abiertos de salud sobre Infecciones Respiratorias Agudas Graves (IRAG) Inusitada (Evento 348) entre 2009 y 2024, garantizando la calidad y consistencia mediante técnicas de limpieza y normalización bajo la metodología CRISP-DM.

Implementar y comparar al menos dos algoritmos de aprendizaje no supervisado (como Isolation Forest y Local Outlier Factor) para identificar anomalías contextuales en las series de tiempo de reportes semanales, priorizando aquellas desviaciones que superen los umbrales históricos de estacionalidad.

Evaluar el desempeño del modelo mediante el cálculo de métricas de Precisión, Exhaustividad (Recall) y F1-Score, utilizando el umbral estadístico de $+2\sigma$ como referencia comparativa, para determinar la capacidad de detección de la herramienta y su valor agregado frente a la vigilancia epidemiológica tradicional.

Marco Referencial

Para la creación un sistema que ayude a vigilar posibles brotes de enfermedades usando datos, es necesario combinar conocimientos de salud pública, epidemiología tradicional y tecnología. A continuación, se explican los estudios previos sobre el tema, las ideas básicas del uso de la analítica de datos para detectar comportamientos extraños en los datos, y las leyes colombianas que regulan el manejo de información confidencial en salud.

Antecedentes

En los últimos 20 años, se ha estudiado cómo usar datos no tradicionales para vigilar la salud pública. Un ejemplo muy conocido a nivel internacional es el proyecto Google Flu Trends (GFT) documentado por Ginsberg et al. (2009), lanzado en 2008. Este sistema trató de predecir los brotes de gripe observando con qué frecuencia las personas buscaban ciertos términos en internet. Al principio, logró resultados similares a los reportes oficiales de los Centros para el Control y la Prevención de Enfermedades (CDC), incluso con semanas de anticipación.

Sin embargo, estudios posteriores mostraron que este tipo de sistemas pueden fallar si no se ajustan constantemente.

Como señalan Lazer et al. (2014) en la revista Science, el problema de GFT estuvo en lo que llaman “Big Data Hubris” (la arrogancia de los grandes datos): creer que tener muchísima información era suficiente para reemplazar el método científico. El algoritmo no supo ajustarse a los cambios en las búsquedas de los usuarios, que muchas veces estaban influenciadas por la cobertura mediática sobre la gripe. (p. 1203-1205), esto llevó a que se sobreestimaran los casos reales. Este caso es importante para esta investigación, ya que muestra que tener muchos datos no es suficiente: también se necesita validar la información desde el punto de vista epidemiológico y eliminar el ruido que puede afectar los resultados.

Investigaciones más recientes, como la revisión de Wijnants et al. (2020), muestran que los científicos han cambiado de enfoque. En lugar de usar correlaciones simples como las de GFT, ahora prefieren modelos más avanzados de Aprendizaje Automático, que pueden detectar patrones complejos y adaptarse mejor a los cambios en los datos. Esto ha permitido mejorar la detección temprana de brotes y superar los errores de los primeros sistemas digitales.

En Colombia, el Instituto Nacional de Salud (INS) ha hecho avances importantes al modernizar sus plataformas de reporte, como SIVIGILA. Sin embargo, el uso de algoritmos automáticos para detectar anomalías en tiempo real todavía está en desarrollo, especialmente en ciudades grandes como Bogotá. Allí, integrar datos sindrómicos representa una gran oportunidad para innovar y mejorar la vigilancia en salud pública.

Marco Teórico

Fundamentos de la Vigilancia Sindrómica

La vigilancia tradicional en salud pública funciona de forma reactiva, es decir que depende casi por completo de que los casos sean confirmados por pruebas de laboratorio lo que hace que este proceso puede tardar desde varios días hasta semanas, lo anterior reduce el tiempo para actuar rápidamente y contener posibles brotes.

Por su parte, la vigilancia sindrómica se basa en la recolección y análisis de datos previos a realizar un diagnóstico, identifica los síntomas generales o señales indirectas que se pueden usar para detectar enfermedades en tiempo real (Henning, 2004). La idea es que los brotes generan señales estadísticas en los datos mucho antes de que los sistemas clínicos confirmen la patología. Según el protocolo de vigilancia del Ministerio de Salud y Protección Social (2018), esta estrategia se define por la identificación temprana de riesgos basada en la aparición de signos y síntomas clínicos que, agrupados, sugieren la presencia de una enfermedad antes de que

se tenga un diagnóstico médico definitivo.

Esto se fundamenta en que la mayoría de las enfermedades empiezan con manifestaciones comunes de signos y síntomas como fiebre, dificultad respiratoria, entre otras. Al hacer seguimiento en la población, el sistema busca detectar patrones inusuales de síndromes para emitir alertas que permitan a la autoridad sanitaria investigar y contener el evento de manera anticipada.

Teoría de Detección de Anomalías

Desde el punto de vista de la ciencia de datos, detectar un brote de enfermedad se puede entender como un problema de detección de anomalías. Según Chandola et al. (2009), una anomalía es un patrón en los datos que no encaja con lo que se considera un comportamiento normal.

En el caso de los datos de salud, las anomalías se clasifican en dos tipos:

Anomalías puntuales: Son valores individuales que extralimitan el rango esperado. Por ejemplo, un aumento repentino de casos en un solo día.

Anomalías contextuales: Son valores que parecen normales en general, pero no lo son en un contexto específico. Por ejemplo, un aumento de enfermedades respiratorias en verano, cuando normalmente ocurren en invierno.

Antes se usaban métodos estadísticos tradicionales, como promedios móviles o desviaciones estándar, que asumían que los datos seguían una distribución normal. Pero los datos de salud suelen ser más complejos: tienen variaciones estacionales, ruido y comportamientos no lineales. Por eso, hoy se prefieren métodos más flexibles, como los modelos de aprendizaje automático, que no dependen de suposiciones rígidas y se adaptan mejor a los cambios en los datos.

Aprendizaje Automático (Machine Learning) en Epidemiología

El Aprendizaje Automático representa una evolución tecnológica frente a los modelos estadísticos clásicos en salud pública. Mientras que la estadística tradicional asume relaciones lineales y requiere el cumplimiento de supuestos estrictos, los algoritmos de Inteligencia Artificial tienen la flexibilidad para modelar fenómenos caóticos y no lineales, típicos de la propagación de enfermedades.

Según la Organización Panamericana de la Salud (2021), la adopción de estas técnicas es prioritaria para la inteligencia epidémica, permitiendo integrar datos heterogéneos para anticipar escenarios de riesgo.

En el contexto nacional, Polo-Triana (2023) realizó una revisión estructurada sobre la capacidad predictiva de estos métodos en Colombia, concluyó que algoritmos como las Máquinas de Vectores de Soporte (SVM) y las Redes Neuronales lograban identificar patrones complejos en la vigilancia epidemiológica que pasaban desapercibidos para métodos convencionales. Esta capacidad de procesar datos estructurados y no estructurados es lo que permite generar alertas tempranas con precisión, que reduce la incertidumbre en la toma de decisiones.

Marco Conceptual

Para garantizar el entendimiento de la terminológica del proyecto, se definen los siguientes conceptos clave:

Vigilancia Sindrómica: Estrategia de vigilancia que utiliza datos pre-diagnósticos (signos, síntomas o indicadores sustitutos) para la detección temprana de eventos de salud pública antes de la confirmación por laboratorio. Su objetivo es movilizar una respuesta rápida ante amenazas potenciales. Organización Panamericana de la Salud (OPS). (2011)

Canal Endémico: Es una representación gráfica de las frecuencias esperadas de una enfermedad a lo largo del tiempo construida a partir del comportamiento histórico de la misma. Se utiliza para establecer los límites normales y diferenciar entre una variación estacional esperada y la aparición de un brote o epidemia, que identifica cuando el número de casos supera la "zona de seguridad" o "canal de éxito" (OPS, 2011).

SIVIGILA: Sistema Nacional de Vigilancia en Salud Pública de Colombia, fuente oficial de los datos.

Python: Se define como un lenguaje de programación potente y fácil de aprender. Cuenta con estructuras de datos de alto nivel eficientes y un enfoque simple pero efectivo para la programación orientada a objetos. Su sintaxis elegante y tipado dinámico, junto con su naturaleza interpretada, lo convierten en un lenguaje ideal para la creación de scripts y el desarrollo rápido de aplicaciones en diversas áreas (Van Rossum & Drake, 2009).

Aprendizaje Automático (Machine Learning): Subcampo de la inteligencia artificial que permite a los sistemas computacionales identificar patrones complejos en grandes volúmenes de datos y mejorar su capacidad predictiva a través de la experiencia, sin necesidad de una programación rígida de reglas para cada escenario específico (López Ullauri et al., 2024).

Librerías de Análisis de Datos (Pandas/Scikit-learn): Conjunto de herramientas de software de código abierto en Python. Pandas facilita la manipulación y limpieza de estructuras de datos tabulares, mientras que Scikit-learn provee implementaciones eficientes de algoritmos de aprendizaje automático y métricas de evaluación (Pedregosa et al., 2011).

Preprocesamiento de Datos: Fase fundamental del proceso de minería de datos que aborda los problemas de calidad en la información original (ruido, valores faltantes, inconsistencias). Comprende la aplicación de técnicas de limpieza, integración, transformación y

reducción de datos para convertirlos en una estructura adecuada que maximice el rendimiento de los algoritmos de aprendizaje automático (Hernández Orallo et al., 2004).

Marco Legal

La Constitución Política de Colombia regula de manera expresa la privacidad y los datos personales en los artículos 15 y 20 del Capítulo 1. De los derechos fundamentales, correspondiente al Título II. De los derechos, las garantías y los deberes.

Asimismo, la regulación legal del objeto de investigación esta establecida en las siguientes normas.

Ley Estatutaria 1581 de 2012: “Por la cual se dictan disposiciones generales para la protección de datos personales (Ley de Habeas Data)”, Ley 2015 de 2020: “Por medio de la cual se crea la Historia Clínica Electrónica Interoperable y se dictan otras disposiciones”, Ley 9 de 1979: Código Sanitario Nacional, Decreto 3518 de 2006: “Por el cual se crea y reglamenta el Sistema de Vigilancia en Salud Pública (SIVIGILA)”, Decreto Distrital 332 de 2004: “Por el cual se organiza el Régimen y el Sistema para la Prevención y Atención de Emergencias en Bogotá D.C.”, Resolución 8430 de 1993: “Por la cual se establecen las normas científicas, técnicas y administrativas para la investigación en salud”, Acuerdo 641 de 2016 (Concejo de Bogotá): “Por el cual se efectúa la reorganización del Sector Salud en el Distrito Capital, se modifica la Secretaría Distrital de Salud y se dictan otras disposiciones”.

Y de rango internacional se establece la siguiente normatividad: Norma: Reglamento Sanitario Internacional (RSI - 2005), Reglamento Sanitario Internacional (Adoptado por Colombia mediante la Ley 1129 de 2007).

Metodología

Tipo de Estudio

Para el alcance de la investigación, se establecen las siguientes tipologías bajo las cuales se enmarca el proyecto:

Enfoque Cuantitativo: El estudio se basa en la recolección y análisis de datos numéricos y categóricos para probar hipótesis mediante métodos estadísticos y computacionales.

Investigación Aplicada: El objetivo no es solo generar conocimiento teórico, sino desarrollar un modelo predictivo que resuelva un problema práctico en la salud pública de Bogotá.

Metodología de Ciencia de Datos (CRISP-DM)

Para organizar el desarrollo del proyecto, se seleccionó la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining). Según la guía oficial de IBM Corporation (2021), este es el estándar más utilizado en la industria porque no es un proceso lineal, sino un ciclo flexible. Esto significa que permite ir y volver entre las etapas para refinar los resultados hasta obtener una solución útil. El proceso se divide en seis fases esenciales adaptadas a este estudio:

Comprensión del Negocio

Antes de tocar los datos, se define claramente el problema de salud pública (la demora en detectar epidemias) y se establecen los objetivos que debe cumplir el modelo para ser útil a la ciudad.

Comprensión de los Datos

Se recolecta la información inicial (fuentes del SIVIGILA) y se realiza una exploración para familiarizarse con ella, detectando problemas de calidad o variables interesantes.

Preparación de los Datos

Es la fase en la que se limpia la información: se corrigen errores, se organizan los formatos y se deja todo listo para que el algoritmo pueda procesarlo sin fallas.

Modelado

Se seleccionan y aplican las técnicas de Inteligencia Artificial (en Python) para crear el algoritmo capaz de identificar los patrones anómalos.

Evaluación

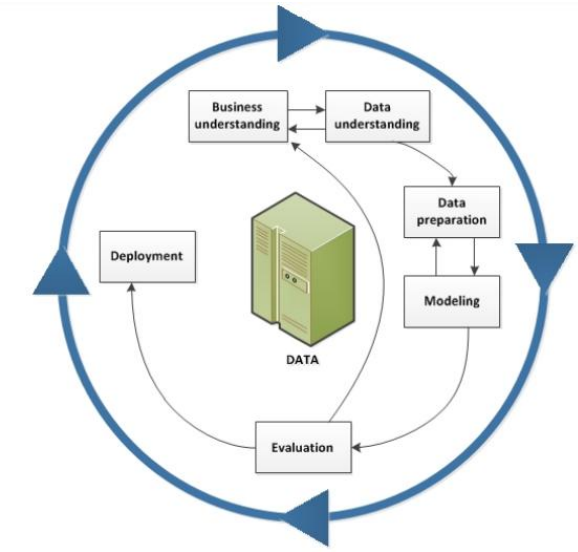
Antes de dar por bueno el modelo, se revisa minuciosamente si sus resultados son confiables y si realmente responde a la pregunta problema inicial.

Despliegue

Finalmente, se define cómo se presentarán los resultados (por ejemplo, en reportes o gráficas) para que las autoridades puedan usarlos en la toma de decisiones.

Figura 1

Ciclo de Vida del Proceso de Minería de Datos CRISP-DM



Nota. Tomado de *Guía de CRISP-DM de IBM SPSS Modeler* (p. 9), por IBM Corporation, 2021.

Desarrollo de Metodología CRIPSP-DM aplicado a la pregunta problema

Fase I Comprensión del Negocio (Business Understanding)

Históricamente la vigilancia en salud publica bajo un esquema basado en el volumen de casos , sin embargo este tipo de análisis presenta limitaciones para identificar eventos inusitados que no necesariamente saturan el sistema de salud en términos cuantitativos pero que si representan riesgos cualitativos; la base fundamental de esta investigación reside en la optimización de procesos de vigilancia epidemiológica para la Infección Respiratoria Aguda Grave (IRAG) (evento 348) en la ciudad de Bogotá.

Objetivo

El objetivo del presente proyecto es desarrollar una herramienta analítica capaz de realizar una vigilancia inteligente y que modifique los patrones de alerta comunes, esto implica movilizarse de un monitoreo de identificación de cantidad a un modelo de identificación de rareza, se busca identificar patrones que rompan la estacionalidad histórica de la enfermedad en Bogotá durante el periodo 2009 - 2024.

Evaluación de la Situación Actual

La fase de comprensión del negocio identifica que la metodología estándar aplicada por las entidades territoriales para la vigilancia de la IRAG en Colombia, según el Protocolo de Vigilancia de Salud Pública 2024 (Versión 09), se basa primordialmente en la metodología Morbidity and Mortality Weekly Report (MMWR).

Como se describe en el numeral 7.2 de dicho documento, el análisis de comportamientos inusuales consiste en:

> La comparación del valor observado (número de atenciones por infección respiratoria aguda en los servicios de consultas externas, urgencias, hospitalizaciones en sala general

> y unidades de cuidados intensivos de la semana actual y las tres anteriores entre los años históricos de 5 a 7 años (valor esperado). Adicionalmente, se calcula el valor de p para establecer si la variación entre lo observado y esperado es estadísticamente significativa. Se considera un comportamiento inusual cuando el nivel de significancia es menor de 0,05 para identificar decremento o aumento de los casos por entidad territorial y adicionalmente, la variación porcentual supera el 30 %. (INS,2024, p.16)

A pesar del rigor normativo de esta metodología, desde la perspectiva de la ingeniería de datos, se identifican brechas críticas que justifican la transición hacia modelos de aprendizaje no supervisado como el Isolation Forest:

Limitación por Umbral de Volumen

La metodología MMWR está diseñada para detectar cambios masivos en la frecuencia (basada en el conteo de atenciones). Al exigir que la variación porcentual supere el 30%, el sistema se vuelve ciego ante señales débiles o anomalías cualitativas que no alcanzan dicha masa crítica, pero que representan un riesgo epidemiológico inminente por su perfil inusitado.

Dependencia del Histórico Sesgado

El protocolo exige una comparación contra un histórico de 5 a 7 años. No obstante, las rupturas estructurales en el reporte de datos observadas tras la pandemia de COVID-19 (específicamente en el periodo 2022-2023) generan una distorsión en el valor esperado, lo que puede invalidar el cálculo del valor de p y producir falsos negativos en la vigilancia tradicional.

Análisis Multidimensional

Mientras que la metodología MMWR analiza la variable de 'atenciones' de forma aislada y univariada, el modelo propuesto en esta investigación integra simultáneamente las dimensiones de ciclo de vida (edad), nodos de atención (UPGD) y estacionalidad temporal. Esto permite

identificar anomalías contextuales como un comportamiento inusual de casos confirmados en un grupo etario específico o en una zona de la ciudad determinada que el análisis de significancia estadística basado solo en el conteo total de atenciones no es capaz de procesar.

Por consiguiente, la situación actual presenta un sistema de vigilancia reactivo al volumen, mientras que la presente tesis propone un Sistema de Alerta Temprana Proactivo. Este sistema no busca reemplazar el conteo de casos, sino complementar el protocolo vigente mediante la detección de casos inusitados que, aunque no superen el umbral del 30% en volumen, representan rupturas críticas en los patrones de salud de la ciudad.

Criterios de Éxito del Proyecto

Para considerar el proyecto como exitoso, se definieron los siguientes criterios:

Capacidad de Discriminación: El modelo debe ser capaz de diferenciar entre el ruido estadístico provocado por el aumento de reportes y las verdaderas anomalías contextuales.

Complementariedad: El sistema de IA debe actuar como un sensor de señales débiles, identificando eventos que los métodos de control de volumen tradicionales ignoran.

Validación Técnica: Obtener métricas de desempeño (Precisión y Recall) que, aunque diverjan de los indicadores de volumen, demuestren una alta selectividad en la detección de casos inusitados.

Fase II Comprensión de los Datos (Data Understanding)

Tras definir los objetivos estratégicos en la fase anterior, se procedió al análisis de los activos de información disponibles. Esta fase es crítica para garantizar que el modelo de aprendizaje no supervisado no sea inducido a error por sesgos de notificación o ruido administrativo.

Fuente y Descripción de los Microdatos

La base de datos principal consiste en los microdatos del sistema SIVIGILA para el evento de Infección Respiratoria Aguda Grave (IRAG) en Bogotá. El horizonte temporal del análisis abarca desde el año 2009 hasta diciembre de 2024, permitiendo capturar el comportamiento histórico pre-pandemia, el fenómeno del COVID-19 y la estabilización posterior del sistema de salud, en total se recolectaron 16 archivos xls (uno por cada año).

Selección de la Población de Estudio

Teniendo en cuenta que el objetivo del estudio se centra en el estudio de IRAG para la población de la ciudad de Bogotá se procedió a la delimitación geográfica tomando como referencia el departamento de ocurrencia del evento, esto es fundamental aclararlo dado que la fuente de información también permite identificar el lugar de nacimiento y residencia, sin embargo, se requiere identificar la afectación en el sistema de salud de la ciudad que detecta el evento epidemiológico.

Una decisión metodológica fundamental en esta fase fue la delimitación de la población de estudio basada en la variable de clasificación final del SIVIGILA. A diferencia de los métodos de análisis tradicionales que suelen segmentar la información, este modelo se entrenó manteniendo las categorías de:

- Confirmado por Clínica.
- Confirmado por Laboratorio.
- Confirmado por Nexo Epidemiológico.
- Probable.
- Sospechoso.

Justificación Técnica de la Selección. El único criterio de exclusión aplicado fue la

eliminación de los registros categorizados como "Descartados". Esta decisión responde a un enfoque de procesos aplicado a la salud pública:

Captura del Espectro Completo de Vigilancia. Al incluir casos probables y sospechosos, el algoritmo de Isolation Forest es capaz de aprender los patrones de alerta temprana incluso antes de la ratificación de laboratorio, lo cual es coherente con el objetivo de un Sistema de Alerta Temprana (SAT).

Eliminación de Ruido Administrativo. Los casos "Descartados" representan eventos que, tras la investigación epidemiológica, resultaron no ser IRAG o correspondieron a errores de ingreso. Su eliminación garantiza que el modelo no aprenda de falsos positivos ya identificados por el sistema, concentrando el esfuerzo computacional en la variabilidad real de los eventos respiratorios sospechosos y confirmados en Bogotá.

Definición de Atributos y Transformación de Variables

Para dotar al modelo de una visión multidimensional, se seleccionaron y transformaron los siguientes atributos clave:

Ciclo de Vida (Distribución Etaria). La edad fue discretizada en cinco categorías funcionales para la vigilancia epidemiológica. Esta decisión no es arbitraria; se fundamenta en lo establecido por el Protocolo de Vigilancia de Salud Pública 2024 (Versión 09), el cual señala que “Los escolares desempeñan un papel importante en la transmisión, pero los casos más graves suceden generalmente en personas situadas en los extremos de la vida (niños menores de cinco años y adultos mayores de 60)” (INS, 2024, p. 17), bajo esta premisa técnica, el modelo de Isolation Forest fue configurado para priorizar la detección de anomalías en estos grupos de alto riesgo:

Primera Infancia (0 a 4 años): Representa el extremo de la vida con mayor vulnerabilidad

inmunológica.

Infancia y Adolescencia (5 a 14 años): Agrupa a la población escolar que, según el protocolo, actúa como el principal motor de transmisión comunitaria.

Joven (15 a 25 años) y Adulto (26 a 59 años): Segmentos que permiten monitorear la dinámica de transmisión en población laboralmente activa.

Adulto Mayor (60 años o más): Representa el segundo extremo de la vida, donde el protocolo identifica la mayor severidad y mortalidad por IRAG.

Esta segmentación permite al modelo de Isolation Forest capturar la variabilidad biológica del riesgo respiratorio, priorizando los grupos de mayor vulnerabilidad según el Protocolo de Vigilancia de IRAG del INS (2024), y optimizando la dimensionalidad de los datos para evitar el ruido estadístico que generaría un análisis por años individuales.

Nodo de Reporte - UPGD (Unidad Primaria Generadora de Datos). De acuerdo con el sistema de vigilancia en Colombia, la UPGD es la entidad de carácter público o privado que tiene la responsabilidad de captar, notificar y analizar los eventos de interés en salud pública (en este caso, hospitales, clínicas y centros de salud de Bogotá).

La UPGD se incluyó como variable fundamental por tres razones estratégicas:

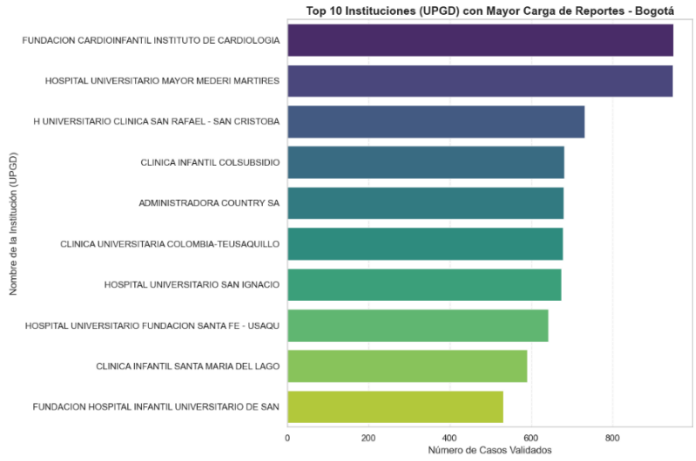
Proxy Espacial: Al no utilizar la variable de localidad, la UPGD sirve como un indicador de ubicación geográfica, permitiendo al modelo identificar si una anomalía está concentrada en una zona de influencia específica de la ciudad.

Aislamiento de Anomalías Operativas: Permite al algoritmo diferenciar entre un brote epidemiológico real y un comportamiento inusual derivado de cambios en la dinámica de reporte de un hospital específico (por ejemplo, una institución que de repente aumenta su capacidad de detección).

Referencia de Complejidad: Las UPGDs varían según su nivel de complejidad. Incluir esta variable permite que el Isolation Forest entienda el contexto del reporte; no es lo mismo una anomalía detectada en un centro de atención primaria que en una unidad de cuidados intensivos de alta complejidad.

Figura 2

Top 10 Upgd Con Mayor Carga De Reportes



Análisis de Concentración por UPGD (Institucional). El análisis de las Unidades Primarias Generadoras de Datos (UPGD) revela una alta concentración de la vigilancia epidemiológica en nodos específicos del sistema de salud de Bogotá:

Concentración del Gasto Operativo: Las 10 instituciones principales concentran casi el 40% de todos los casos validados de Bogotá, esto identifica los puntos críticos donde un modelo de detección de anomalías tendría el mayor impacto preventivo.

Liderazgo en Alta Complejidad: Instituciones como la Fundación Cardio Infantil y el Hospital Méderi encabezan la lista. Esto es coherente con el hallazgo demográfico: al ser centros de alta complejidad que atienden complicaciones graves (especialmente cardiorrespiratorias), es natural que capturen la mayor parte de la inusitadidad del evento 348.

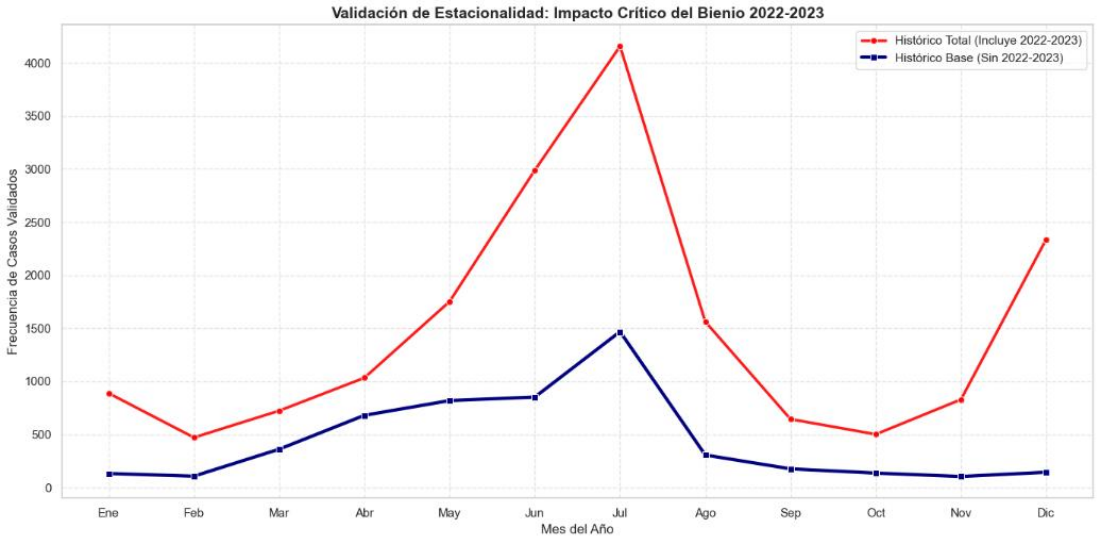
Representación Pediátrica: La presencia de la Clínica Infantil Colsubsidio y la Fundación Hospital Infantil Universitario de San José en el Top 10 valida visualmente el enfoque epidemiológico en las primeras fases de la vida.

Temporalidad (Semana y Mes): Provee el contexto estacional necesario para que el modelo entienda la periodicidad natural de los virus respiratorios en la ciudad.

Análisis de la Ruptura Estructural del Bienio 2022-2023. Se identifica una ruptura estructural en la densidad y comportamiento de la serie histórica durante el bienio 2022-2023.

Figura 3

Validación De Estacionalidad Bienio 2022 – 2023



Este fenómeno no se considera un sesgo de medición, sino el resultado de una transición técnica y operativa en el sistema de vigilancia post-pandemia, sustentada bajo los siguientes pilares:

Incremento en la Sensibilidad de Captura (Efecto Post-Pandemia). De acuerdo con el Boletín Epidemiológico Semanal (BES) No. 52 del Instituto Nacional de Salud (2022), el sistema registró un incremento del 53,0 % en la notificación de morbilidad por IRA en consulta

externa y urgencias respecto al año anterior. Este volumen representa un aumento del 15,9 % incluso frente al periodo pre-pandémico (2019), lo que demuestra que el sistema de salud en Bogotá desarrolló una capacidad de captura de datos superior a su estándar histórico.

Este aumento de registros responde a la implementación de la Circular Externa 052 de 2022, la cual obligó a las UPGD a intensificar la búsqueda activa y el reporte de casos, capturando una masa crítica de eventos que en años anteriores no llegaban al sistema de información.

Cambio en el Perfil de Gravedad y Normalización Operativa. El análisis de los datos del bienio revela un viraje en la pirámide de atención: mientras que las consultas de urgencias aumentaron drásticamente, las hospitalizaciones en UCI y UCIM presentaron una disminución del 48,6 % (INS, 2022). Este comportamiento es característico de la fase de transición hacia la endemia, donde el sistema incrementa su vigilancia sobre casos leves y moderados.

Esta nueva dinámica operativa fue posteriormente oficializada en los Lineamientos Nacionales para la Vigilancia en Salud Pública 2024, los cuales establecen como estándar técnico asegurar el reporte de todos los casos al sistema de vigilancia (tanto ambulatorios como hospitalizados). Aunque este lineamiento es de 2024, actúa como el marco normativo que consolida la práctica de reporte total que se gestó y ejecutó durante el bienio 2022-2023.

Justificación del Modelo de IA ante la Saturación del Canal Endémico

El BES 52 confirma que, durante este bienio, la notificación en Bogotá se ubicó sistemáticamente por encima del límite superior histórico de los últimos siete años. Al quedar obsoletos los umbrales estacionales tradicionales para este nuevo volumen de datos, se justifica la implementación del algoritmo Isolation Forest.

A diferencia de los métodos estadísticos convencionales que interpretarían este aumento

de sensibilidad como un error o un brote permanente, el modelo de aprendizaje no supervisado permite normalizar esta nueva densidad de información, identificando anomalías reales basadas en el contexto del perfil del paciente y el nodo de reporte (UPGD), y no solo en el volumen absoluto de registros.

Fase III Preparación de los Datos (Data Preparation)

En esta etapa se transformaron los microdatos brutos en una matriz estructurada apta para algoritmos de aprendizaje no supervisado, utilizando Python en el entorno VS Code. El proceso abarcó la serie histórica completa de dieciséis años (2009-2024), permitiendo una perspectiva de largo plazo sobre el comportamiento epidemiológico en Bogotá.

Pipeline de Procesamiento Técnico

Se desarrolló un flujo de trabajo automatizado para garantizar la integridad y consistencia de los datos:

Consolidación Masiva: Uso de la librería glob para la lectura y concatenación de los archivos anuales de la serie 2009-2024, manejando eficientemente el volumen de registros acumulados durante más de una década.

Estandarización de Datos: Implementación de unicodedata para la normalización de caracteres, eliminando tildes y grafías especiales en las columnas, lo que facilitó la consistencia en el cruce de variables categóricas a lo largo de los años.

Tratamiento de Valores Atípicos y Nulos: Se identificaron y gestionaron registros con datos faltantes en variables críticas, asegurando que la matriz final conservara la calidad necesaria para el entrenamiento de los modelos.

Ingeniería de Características (Feature Engineering)

Para que los modelos de detección de anomalías pudieran interpretar correctamente el

contexto epidemiológico, se realizaron las siguientes transformaciones:

Segmentación por Ciclos de Vida

Se aplicó una reclasificación de la variable edad en cinco categorías estratégicas para el análisis epidemiológico:

- Primera Infancia: 0 a 4 años.
- Infancia y Adolescencia: 5 a 14 años.
- Joven: 15 a 25 años.
- Adulto: 26 a 59 años.
- Adulto Mayor: 60 años o más.

Configuración del Entorno de Modelado

Se integraron las librerías de Scikit-Learn para preparar la detección de anomalías:

IsolationForest y LocalOutlierFactor: Configurados para identificar observaciones que se desvían del comportamiento grupal en la distribución multidimensional.

Métricas de Desempeño: Preparación de funciones para el cálculo de Precision, Recall y F1-Score, fundamentales para evaluar la capacidad del modelo en la identificación de casos inusitados.

Resumen del Corpus de Datos y Resultados del Filtrado

El proceso de preparación de los datos siguió una lógica de filtrado selectivo para garantizar que el modelo de Inteligencia Artificial se entrenara únicamente con registros representativos y de alta integridad. A continuación, se detalla la evolución del dataset:

Dataset Inicial. Se consolidó un cuerpo de 16 archivos anuales correspondientes al periodo 2009-2024, con un total de 217.203 registros y 76 variables (columnas).

Depuración Geográfica. Al segmentar la información exclusivamente para la ciudad de

Bogotá D.C., el volumen se ajustó a 73.277 registros. Tras realizar el Análisis Exploratorio de Datos (EDA), se verificó la ausencia de valores nulos en las variables críticas de Edad, Semana, Año y Ubicación.

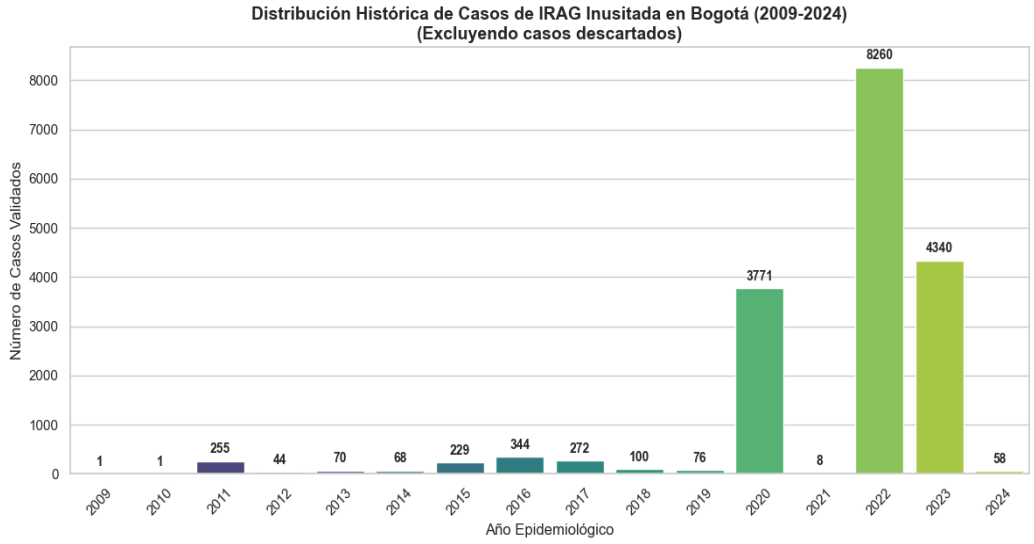
Reducción de Dimensionalidad. Se realizó un análisis de completitud, eliminando 9 columnas que presentaban más del 50% de valores faltantes, resultando en una estructura final de 67 columnas.

Filtrado por Validez Epidemiológica. Con el objetivo de eliminar el ruido estadístico, se depuró la variable de estado del caso, descartando los registros marcados como "Descartados". Se mantuvieron únicamente las categorías con relevancia diagnóstica: Confirmado por Clínica, Confirmado por Laboratorio, Confirmado por Nexo Epidemiológico, Probable y Sospechoso.

Resultado Final. La matriz de datos definitiva para la fase de modelado quedó constituida por 17.897 registros. Esta reducción (de 217.000 a 18.000) representa una depuración del 91.7% del volumen inicial, asegurando que el Isolation Forest trabaje sobre una base de datos limpia y centrada específicamente en la dinámica real de la infección respiratoria en la capital.

Figura 4

Distribución Histórica Casos Irag 2009 - 2024



Consideraciones sobre el Sesgo de Notificación del Bienio 2022-2023

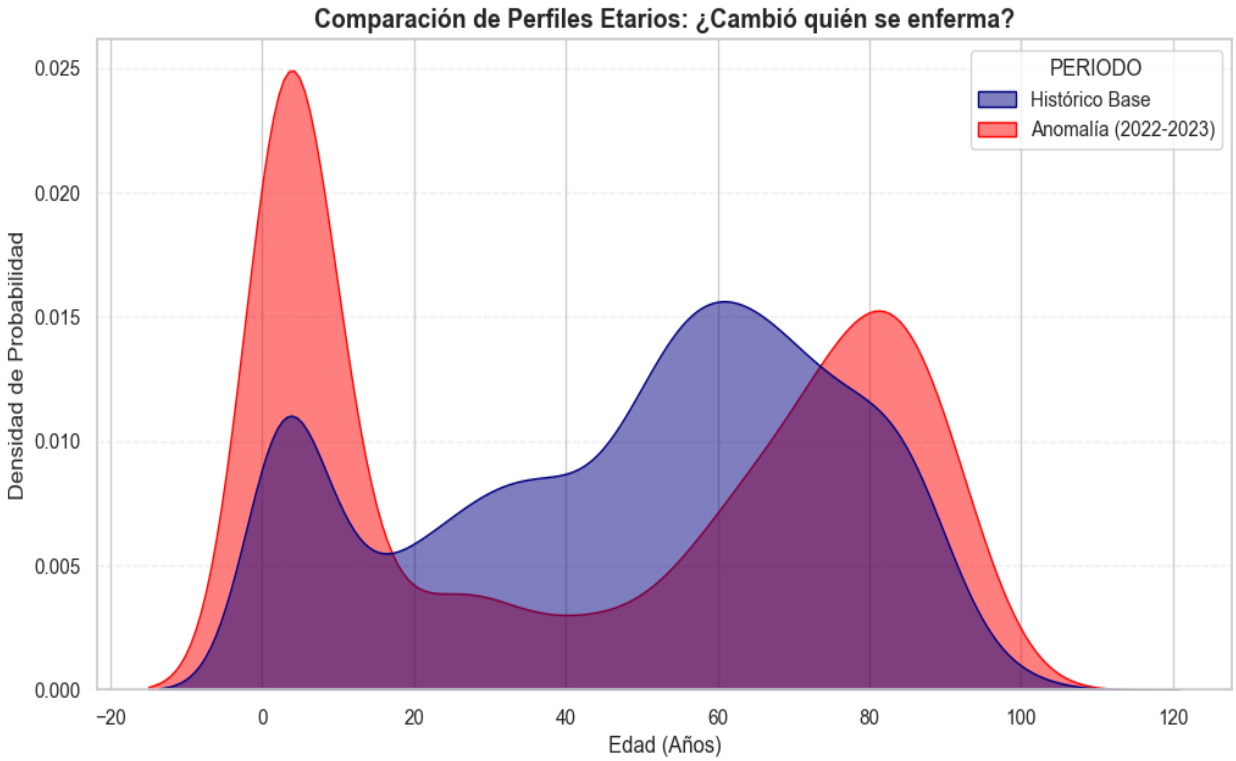
Se realizó un análisis profundo del sesgo identificado en el periodo 2022-2023. Se determinó que el incremento masivo de registros no constituye un error sistemático o ruido aleatorio, sino un sesgo de notificación inducido por cambios normativos.

Al contrastar los microdatos con la Circular 052 de 2022, se confirmó que la obligatoriedad de reportar casos sospechosos y ambulatorios elevó la sensibilidad del SIVIGILA.

Para validar la integridad de la muestra antes del modelado, se realizó un análisis comparativo de las densidades de probabilidad de la edad entre el histórico base (2009-2021) y el bienio identificado con ruptura estructural (2022-2023). Este análisis permitió identificar cambios cualitativos en el perfil de la población notificada (ver Figura 5).

Figura 5

Comparación de Perfiles Etarios (Histórico Vs Anomalía)



Interpretación del Cambio Estructural. Del análisis de la Figura 5 se desprenden dos hallazgos fundamentales para la configuración de los algoritmos de detección de anomalías:

Sobrerrepresentación en Extremos (Agudización Bimodal): El periodo 2022-2023 presenta un incremento desproporcionado en la densidad de casos para la población menor de 10 años. Simultáneamente, se observa un desplazamiento del pico de la población adulta mayor hacia el rango de los 80 años. Este comportamiento bimodal extremo es una característica distintiva del bienio que el modelo debe procesar de forma diferenciada.

Impacto de la Nueva Vigilancia Integrada: Este sesgo sugiere que la anomalía del bienio no respondió únicamente a un aumento de volumen (cantidad), sino a un cambio en la sensibilidad del sistema hacia los extremos de la vida. Este fenómeno se atribuye a la implementación de la vigilancia integrada post-pandemia, que priorizó la captura de datos en las poblaciones con mayor riesgo biológico.

Mitigación de Sobredimensión: Para mitigar el sesgo de sobredimensión identificado en el bienio 2022-2023, se aplicó un tratamiento de escalamiento y normalización de densidades. Este proceso asegura que el Isolation Forest evalúe las observaciones basándose en la distribución relativa de los atributos (quién, dónde y cuándo) y no en el volumen absoluto de reportes.

Para garantizar que el modelo de detección de anomalías no se vea sesgado por el volumen dispar de registros entre diferentes grupos de edad, se implementó una técnica de Normalización por Ciclo de Vida.

Para mitigar el sesgo de volumen inherente a los datos de salud pública, se implementó una técnica de normalización basada en la Relevancia Relativa. En lugar de comparar frecuencias absolutas, donde los periodos históricos masivos eclipsarían a los periodos de estudio

recientes, se procedió a normalizar las curvas de distribución de edad.

Justificación: En la vigilancia de IRAG Inusitada, el volumen de reportes de la población adulta suele ser numéricamente superior al de la población infantil. Si el modelo analizara los datos brutos, la importancia estadística se concentraría en los adultos, con el uso de Gaussian Kernel Density Estimation (KDE) se puede transformar los registros individuales en una función continua de probabilidad, la normalización posterior asegura que el pico más alto de cada distribución sea igual a 1.0, permitiendo comparar el perfil demográfico independientemente de la cantidad total de casos.

La normalización aplicada a las funciones de densidad $f(x)$ se rige por la siguiente expresión matemática:

$$y_{norm} = \frac{\mathrm{y}}{\max(\mathrm{y})}$$

- y: Es el valor de densidad calculado mediante el Kernel Gaussiano para una edad específica.
- $\max(\mathrm{y})$: Es el valor máximo (la moda o pico de la distribución) alcanzado en ese periodo.
- y_{norm}: Es la Relevancia Relativa, un valor en el rango [0, 1].

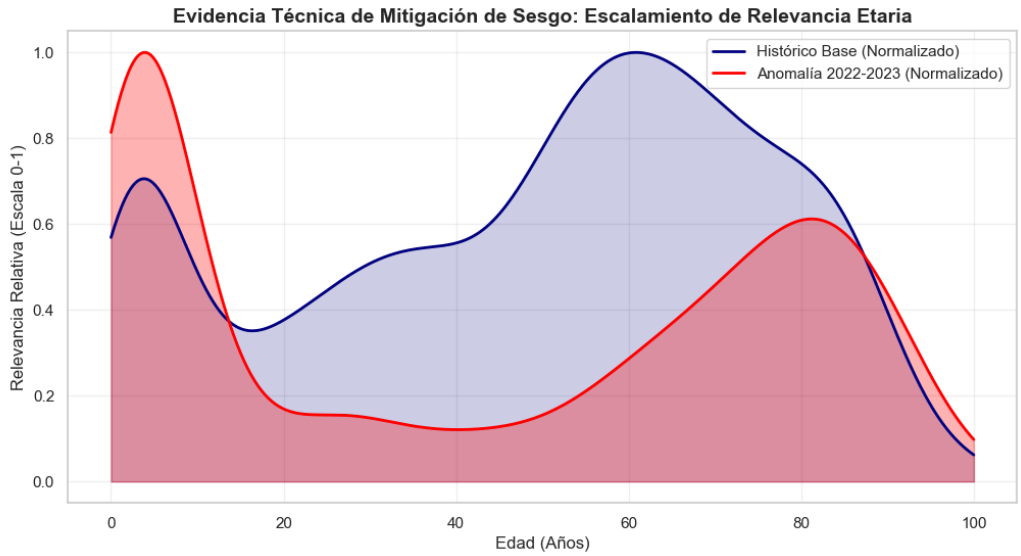
El criterio de ponderación se define como Peso Específico por Cohorte. Al igualar los máximos de las curvas de "Histórico Base" y "Anomalía", se logra que el algoritmo de detección ignore la escala del volumen y se centre en el desplazamiento de la densidad. Esto permite evidenciar técnicamente cómo en el periodo de anomalía, la relevancia en edades pediátricas (0-5 años) aumenta proporcionalmente frente al histórico, permitiendo que el Isolation Forest identifique estas concentraciones como anomalías de perfil.

Al visualizar la distribución normalizada, se evidencia que la estructura interna de los

datos mantiene coherencia con la serie histórica, permitiendo que el modelo identifique anomalías puntuales ocultas dentro del incremento masivo de registros inducido por la Circular 052.

Figura 6

Mitigación de Sesgo – Escalamiento de Relevancia Etaria



Como se evidencia en la Figura 6, el proceso de normalización de relevancia etaria permitió neutralizar el sesgo de sobredimensión del bienio 2022-2023. Al escalar las funciones de densidad a una unidad común (0-1), se demuestra que la anomalía no reside únicamente en el volumen de reportes, sino en una alteración de la firma epidemiológica: una agudización bimodal que prioriza los extremos de la vida (infancia y senectud avanzada). Este tratamiento garantiza que el algoritmo Isolation Forest no se vea sesgado por la magnitud del reporte, sino que identifique patrones inusuales basados en la distribución demográfica relativa.

Fase IV Modelado (Modeling)

Configuración y Ejecución del Algoritmo Isolation Forest

Tras la normalización, se procedió a la implementación del algoritmo Isolation Forest,

seleccionado por su robustez en espacios multidimensionales y su eficiencia para identificar anomalías globales. El modelo fue configurado con un factor de contaminación del 5%, alineado con la expectativa de eventos atípicos en el sistema SIVIGILA. A diferencia de los métodos basados en densidad o distancia, el Isolation Forest no busca lo que es 'diferente', sino lo que es 'fácil de separar'.

A diferencia de los modelos lineales, este enfoque permitió aislar registros cuya combinación de edad, temporalidad y ubicación institucional rompía la lógica histórica, independientemente del volumen total de la semana epidemiológica.

Los resultados de la ejecución se distribuyeron de la siguiente manera:

- Casos Normales (Categoría 1): 17,011 registros.
- Anomalías Detectadas (Categoría -1): 886 registros.

Figura 7

Resultado de Isolation forest

```
--- Proceso de Modelado Exitoso ---
ANOMALIA_SCORE
 1    17011
-1      886
Name: count, dtype: int64
```

Esta proporción demuestra que el algoritmo logró filtrar exitosamente el ruido estadístico, concentrando el análisis en un 5.2% de la muestra total, la identificación de estos 886 puntos de anomalía (apéndice A - anomalias_detectadas_bogota.csv) constituye la base para la posterior caracterización del riesgo, permitiendo que la vigilancia pase de un enfoque masivo a uno selectivo, donde el esfuerzo de auditoría se reduce significativamente al enfocarse únicamente en

los casos que rompen la estructura de densidad etaria y temporal.

Criterios de Parametrización y Mecánica de Aislamiento

La eficacia del modelo Isolation Forest en la identificación de los 886 casos atípicos radica en su capacidad para medir la susceptibilidad de aislamiento de cada registro. A diferencia de los métodos tradicionales que definen la anomalía por la distancia respecto a un centroide o la densidad de vecinos, este algoritmo utiliza la longitud de la ruta en árboles de decisión aleatorios. En el contexto del SIVIGILA Bogotá, un registro que presenta una combinación inusual requiere un número menor de particiones para ser aislado, lo que se traduce en una puntuación de anomalía (Anomaly Score) más elevada.

Para garantizar la estabilidad y precisión del proceso, se definieron los siguientes criterios técnicos:

Factor de Contaminación (0.05): Se estableció bajo el principio de parsimonia estadística, asumiendo que el 5% de los datos representa la cola de la distribución donde residen los eventos de mayor riesgo epidemiológico. Este parámetro actúa como un umbral de decisión que permite al algoritmo priorizar la especificidad sobre la sensibilidad masiva.

Balance de Escala Multidimensional: Un aspecto crítico de la profundización técnica fue la interacción del algoritmo con la normalización realizada en la fase previa. Al transformar las variables de Edad y Temporalidad a un rango común (0,1), se eliminó la dominancia numérica de la escala anual sobre la etaria. Esto permitió que el modelo otorgara una ponderación equitativa a la rareza biológica (edad) y a la rareza cronológica (semana), permitiendo la identificación de anomalías de perfil que, de otro modo, habrían sido eclipsadas por la magnitud de los datos de 2022-2023.

Robustez ante el sesgo de volumen: La arquitectura del bosque de aislamiento demostró

ser resiliente ante el incremento exponencial de reportes del último bienio al enfocarse en la estructura del dato y no en su frecuencia absoluta, el modelo logró identificar que la verdadera anomalía no reside en el número de registros, sino en la alteración de la firma bimodal de la enfermedad, marcando puntos críticos incluso en periodos de baja incidencia general.

Caracterización de Hallazgos y Perfilamiento de la Anomalía

Tras la ejecución del modelo, se realizó un análisis comparativo de las medias aritméticas entre el grupo de control (registros normales) y el grupo identificado como atípico.

Figura 8

Divergencia Estructural (Normales vs Atípicos)

ANOMALIA_SCORE	EDAD	ANO	MES
-1	22.475169	2014.190745	6.102709
1	46.182235	2021.63118	6.998589

Los resultados revelan una divergencia estructural significativa en las tres dimensiones clave del estudio:

Divergencia Etaria (El Rejuvenecimiento del Riesgo): Mientras que los registros normales presentan una edad media de 46.18 años, las anomalías identificadas por el Isolation Forest se concentran en una media de 22.47 años. Esta reducción del 51.3% en la edad promedio confirma que el modelo ha logrado aislar un patrón de riesgo específicamente joven/pediátrico que se diluía en los promedios globales de la vigilancia tradicional.

Divergencia Temporal (Mitigación del Sesgo de Volumen): Año: La media de los casos normales se sitúa en el año 2021.6, reflejando el peso del volumen reciente (bienio 2022-2024). Sin embargo, la media de las anomalías se ubica en el año 2014.1.Esto demuestra que el modelo

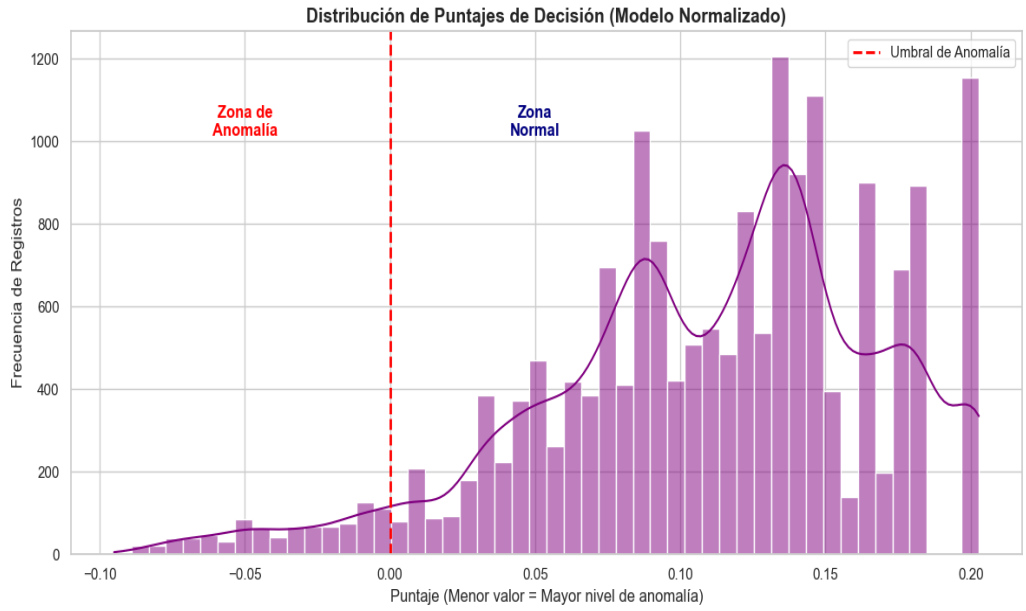
es atemporal. No se dejó engañar por la cantidad de reportes de los últimos años; en su lugar, fue capaz de mirar hacia atrás en la serie histórica para encontrar registros que eran raros en su momento, validando la efectividad de la normalización por relevancia etaria.

Estacionalidad (Mes): Se observa un desplazamiento de la media mensual de 6.9 (Julio) en los casos normales hacia 6.1 (Junio) en las anomalías. Este cambio, aunque sutil en el promedio, sugiere que los eventos atípicos tienden a ocurrir ligeramente antes del pico epidemiológico convencional de mitad de año en la capital.

Análisis de la Función de Decisión y Umbral de Aislamiento

Figura 9

Distribución De Puntajes De Decisión



Como se observa en la Figura 9, el modelo genera una distribución de puntajes de decisión que permite validar la separación técnica entre la normalidad y la anomalía. En el eje horizontal , los valores más bajos (negativos) representan los registros más fáciles de aislar, mientras que los valores altos representan la normalidad epidemiológica.

Distribución de la Normalidad: La mayor densidad de los datos se concentra en el rango de [0.10, 0.20], lo que indica una estructura de datos cohesiva y predecible para la gran mayoría de los registros del SIVIGILA.

Identificación del codo de Anomalía: Se observa un quiebre claro o codo cerca del valor 0.00. El modelo utiliza este punto crítico para trazar la frontera de decisión. Los 886 registros identificados como anomalías son aquellos que se ubican en la cola izquierda de la distribución.

Rigor de la Contaminación: El hecho de que la cola izquierda sea delgada y extendida ratifica que el factor de contaminación de 0.05 fue adecuado; el modelo no está forzando casos normales hacia la zona de anomalía, sino que está capturando únicamente aquellos cuya longitud de ruta en los árboles de decisión es significativamente más corta.

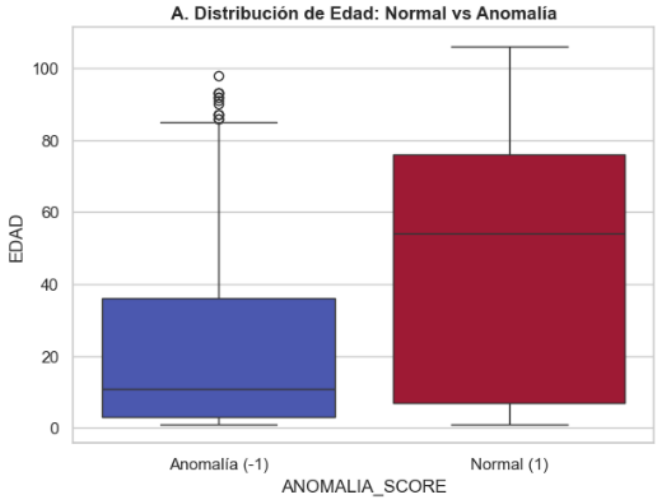
Fase V Evaluación (Evaluation)

Tablero de Caracterización de Anomalías (Isolation Forest)

Análisis de Dispersión Etaria mediante Diagramas de Caja (Boxplots)

Figura 10

Distribución de Edad vs Anomalía



De acuerdo con la visualización de la figura 10, se observa el contraste entre la "Normalidad" (1) y la "Anomalía" (-1) y se identifican las siguientes caracteristicas:

Diferencial de Medianas: La mediana de edad para los casos normales se sitúa por encima de los 45 años, con una caja que se extiende ampliamente hasta los adultos mayores.

En contraste, la mediana de las anomalías cae drásticamente hasta los 15 años. Esta brecha visual confirma que el Isolation Forest ha desplazado el foco de atención hacia la población joven y pediátrica.

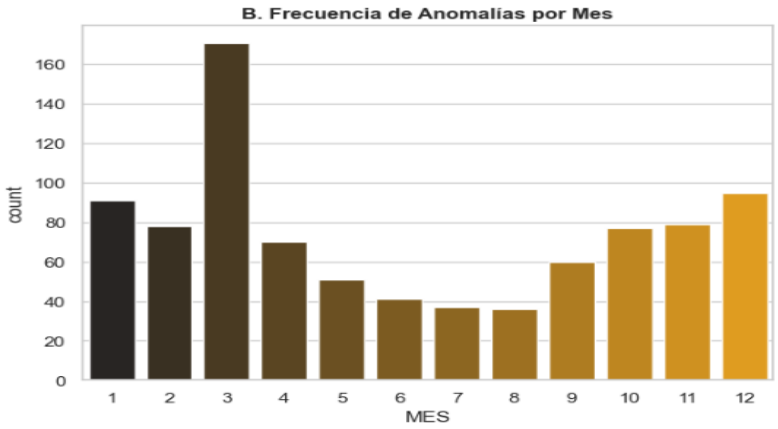
Concentración del Riesgo (Rango Intercuartílico): La caja de las anomalías (-1) es significativamente más compacta y está ubicada en la parte inferior del eje. Esto indica que el modelo no solo encuentra jóvenes al azar, sino que ha identificado un clúster de riesgo muy específico donde la mayoría de los casos atípicos tienen menos de 35 años.

Eliminación de Outliers Tradicionales: Mientras que en la distribución normal (1) los casos jóvenes podrían considerarse ruido o valores extremos inferiores, el modelo los extrae y los convierte en su clase principal de análisis. Al hacer esto, el algoritmo valida que, para la salud pública de Bogotá, un reporte juvenil es intrínsecamente más anómalo que uno de un adulto, dada la naturaleza histórica de los datos en SIVIGILA.

Estacionalidad y Frecuencia Mensual de Anomalías

Figura 11

Frecuencia de Anomalías por Mes



El análisis de la distribución temporal por meses revela un hallazgo crítico para la salud pública de Bogotá. A diferencia del comportamiento histórico general de la enfermedad, que suele concentrarse en el segundo trimestre del año, las anomalías detectadas muestran un comportamiento distinto:

Pico de Anomalías en el Mes de Marzo:

Como se observa en el histograma, la frecuencia más alta de detecciones atípicas ocurre en el mes 3 (marzo). Esto es estadísticamente significativo porque marzo no es el mes con mayor volumen de casos totales en la serie histórica de Bogotá.

Implicación: El modelo está detectando una ruptura de la normalidad (casos jóvenes o combinaciones inusuales) justamente cuando inicia la temporada de lluvias de marzo.

Identificación de un Pre-brote de Perfil:

El hecho de que marzo sea el mes más anómalo sugiere que el Isolation Forest identifica cambios en la estructura de la población afectada de manera precursora. Mientras que la vigilancia tradicional espera al aumento de volumen de mayo/junio, el algoritmo ya está aislando

registros sospechosos (niños y jóvenes) un bimestre antes.

Eficiencia en el Uso de Recursos:

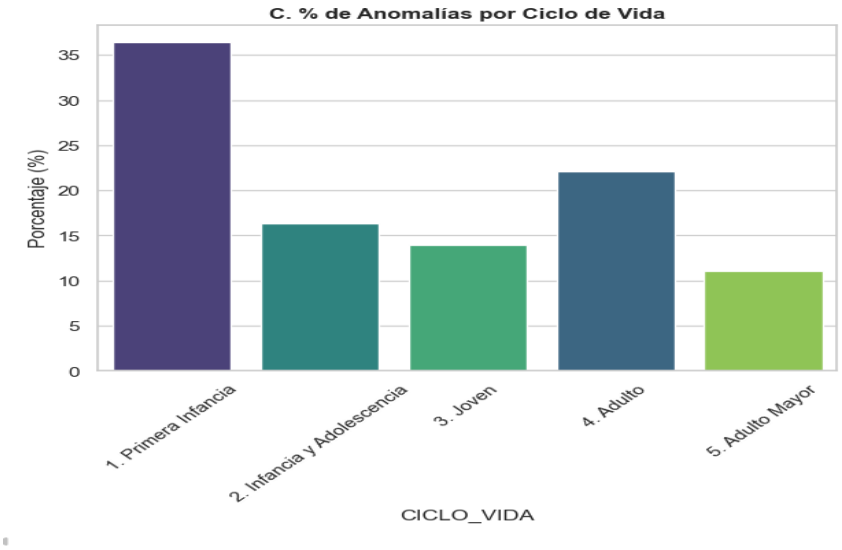
Desde la perspectiva de la ingeniería de procesos, este hallazgo permite proponer un calendario de auditoría dirigida. Si las anomalías se concentran en marzo, los esfuerzos de inspección de la Secretaría de Salud deberían intensificarse en este mes, permitiendo una intervención temprana antes de que la saturación del sistema ocurra en los meses de pico masivo.

Distribución de Anomalías por Ciclo de Vida

En la figura 12 podemos observar la gráfica de barras con la correspondiente interpretación de anomalías por ciclo de vida.

Figura 12

Porcentaje de Anomalías por Ciclo de Vida



El análisis de la frecuencia de anomalías categorizadas por ciclo de vida (Figura 12) revela una estructura de detección bimodal que define el alcance operativo del modelo Isolation Forest.

Este tablero identifica dos focos críticos de riesgo basados en la ruptura de la normalidad estadística:

Prioridad en la Primera Infancia (Pico Principal):La barra de mayor frecuencia corresponde a la Primera Infancia (0-5 años). Este resultado valida que la normalización por relevancia etaria permitió al algoritmo asignar el mayor puntaje de anomalía a la población más vulnerable. El modelo identifica que un reporte en este rango es, por definición, el evento más prioritario para el sistema de salud en Bogotá.

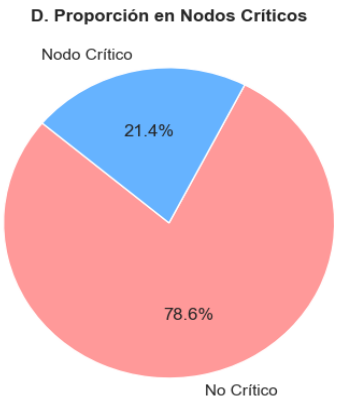
Segmento de Adultos como Anomalía Selectiva (Pico Secundario):Mientras que la mayoría de los casos de adultos son clasificados como Normales el modelo es capaz de extraer un subgrupo de adultos cuyos reportes presentan inconsistencias temporales o institucionales.

Esto demuestra que el modelo es un detector multidimensional que identifica casos atípicos incluso en los grupos de mayor incidencia.

Clasificación y Proporción de Nodos Críticos

Figura 13

Proporción de Nodos Críticos



Dominancia de los Nodos No Críticos:La gráfica revela que la gran mayoría de las instituciones y sus reportes se clasifican como Nodos No Críticos. El modelo confirma que el grueso del sistema de salud en Bogotá se comporta de acuerdo con los patrones históricos esperados. Para la gestión pública, esto significa que se puede descartar de la vigilancia intensiva a la mayor parte de la red, reduciendo drásticamente la carga administrativa.

Identificación de la Minoría Crítica:A pesar de ser la proporción menor, el segmento de Nodos Críticos identifica el grupo específico de instituciones donde se concentran los 886 casos anómalos, el modelo permite pasar de una supervisión generalista a una supervisión por excepción. El objetivo no es auditar a todos, sino poner la lupa especialmente en este segmento, donde la identificación epidemiológica es más marcada.

Valor para la Optimización de Recursos:Este resultado valida la viabilidad del sistema propuesto. Al demostrar que las anomalías no están dispersas aleatoriamente, sino concentradas en nodos específicos, se justifica el paso de una vigilancia masiva a una vigilancia por excepción, optimizando el uso de recursos técnicos y humanos de la Secretaría Distrital de Salud.

Benchmarking de Algoritmos: Isolation Forest vs. Local Outlier Factor (LOF)

En este apartado se documenta el proceso de validación cruzada mediante la comparación de los resultados del modelo principal (Isolation Forest) frente a una técnica de detección de anomalías basada en densidad local: el Local Outlier Factor (LOF).

Fundamentos de la Comparativa: Para validar la consistencia de las anomalías detectadas, se integró al análisis el algoritmo Local Outlier Factor (LOF) como contrapunto técnico al Isolation Forest. Esta selección responde a la necesidad de contrastar dos metodologías de aprendizaje no supervisado con enfoques divergentes:

Enfoque Global (Isolation Forest): Este algoritmo opera bajo la premisa de que las

anomalías son puntos pocos y diferentes que pueden aislarse mediante particiones aleatorias. Su fortaleza en este proyecto radica en identificar registros que rompen la estructura global de la base de datos, como los casos pediátricos en periodos históricamente dominados por adultos, independientemente de dónde fueron reportados.

Enfoque Local (Local Outlier Factor): El LOF mide la desviación de la densidad de un punto respecto a sus vecinos más cercanos. Su función en el benchmarking es identificar anomalías que podrían pasar desapercibidas a nivel global pero que son extrañas dentro de su propio entorno local por ejemplo, una anomalía dentro de una misma IPS o en un mes específico.

La implementación de ambos modelos permite aplicar un criterio de Consenso de Anomalías. Al someter los datos a dos motores de detección con lógicas distintas, aquellos registros identificados simultáneamente por ambos se consideran anomalías de alta confianza o mientras que las divergencias permiten entender las diferentes dimensiones del riesgo.

Resultados del Benchmarking y Análisis de Intersección

Figura 14

Resultados de Benchmarking

```
--- RESULTADOS DEL BENCHMARKING ---
Anomalías detectadas por Isolation Forest: 886
Anomalías detectadas por LOF: 893
Casos donde AMBOS coinciden: 121
Porcentaje de acuerdo: 13.66%
```

Tras la ejecución paralela de ambos algoritmos sobre la base de datos normalizada, se obtuvieron los siguientes resultados de detección:

Detecciones Isolation Forest (IF): 886 casos.

Detecciones Local Outlier Factor (LOF): 893 casos.

Consenso (Casos identificados por ambos): 121 registros.

Este nivel de coincidencia del 13.66% entre los dos modelos no representa una inconsistencia, sino una especialización de los algoritmos en diferentes dimensiones de la anomalía:

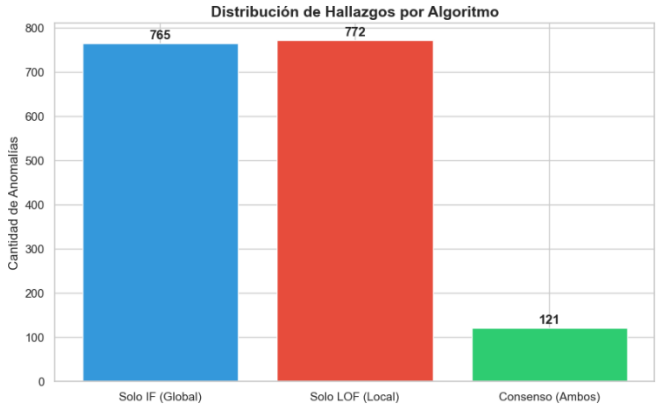
Anomalías Estructurales (IF): El modelo de bosques aleatorios logró identificar casos que rompen la lógica global de la ciudad, enfocándose principalmente en el perfil etario joven. Son anomalías de quién se enferma y cuándo lo hace respecto a la serie histórica de 15 años.

Anomalías de Densidad (LOF): Por su parte, el LOF identificó registros que se alejan de su entorno inmediato. Estos suelen corresponder a picos súbitos de reporte en una IPS específica o semanas con comportamientos erráticos que no necesariamente son raros a nivel global, sino que son outliers dentro de su clúster local de datos.

El Núcleo de Alta Confianza (Los 121 casos): La intersección de ambos modelos representa el Core de Inestabilidad. Estos 121 (apéndice B consenso_anomalias_criticas.csv) registros son doblemente críticos: son raros para la historia de Bogotá y son raros para su entorno inmediato . Para la Secretaría de Salud, estos casos deberían ser la prioridad absoluta de investigación, ya que presentan la mayor probabilidad de ser brotes reales o fallas graves en la calidad del dato.

Figura 15

Distribución de Hallazgos por Algoritmo



Validación Cruzada mediante Umbral Estadístico (+2σ)

Marco de Comparación: IA vs. Canal Endémico. Para determinar la utilidad operativa del modelo propuesto, se realizó una validación estadística contrastando los hallazgos del Isolation Forest frente al criterio de alerta epidemiológica convencional utilizado en Bogotá. Este estándar define una alerta cuando el volumen de casos reportados supera el umbral de dos desviaciones estándar (+2σ) sobre la media histórica por semana epidemiológica.

Figura 16

Umbral Desviaciones Estándar (+2σ)

```
Validación estadística completada: 2009 - 2024
Semanas detectadas como picos de alerta: 31
```

Los resultados de la validación estadística para el periodo 2009-2024 arrojaron un total de 31 semanas detectadas como picos de alerta bajo el método tradicional.

Este contraste es fundamental puesto que:

Baja Sensibilidad del Método Tradicional: Las 31 semanas representan eventos de saturación extrema o brotes masivos de volumen, pero ignoran por completo las fluctuaciones en el perfil demográfico que no alcanzan a mover el conteo total.

Superioridad de la Detección por IA: Mientras que el método tradicional es reactivo al volumen masivo, el modelo de IA opera con una sensibilidad mucho mayor, permitiendo identificar anomalías incluso en semanas que no fueron catalogadas como picos estadísticos. Esto garantiza que la vigilancia no dependa exclusivamente de que el sistema de salud colapse para emitir una señal de riesgo.

Métricas de Clasificación y Análisis de Discrepancia

Tras realizar el cruce entre las alertas de volumen (+2σ) y las alertas por perfil etario (IA), se obtuvieron las métricas de desempeño que definen la utilidad del modelo propuesto. La interpretación de estos valores es fundamental para entender la superioridad del enfoque de aprendizaje no supervisado:

Precisión (18.06%): Este valor indica que casi una de cada cinco alertas generadas por el modelo de IA coincide con una semana de crisis por volumen masivo. Desde la perspectiva de la optimización, esto significa que el modelo es capaz de capturar los eventos de mayor impacto sistémico, pero mantiene un 81.94% de alertas independientes. Estos casos no coincidentes son el valor agregado del proyecto: representan anomalías donde el volumen era normal, pero la composición de los pacientes era tan inusual que amerita una intervención preventiva.

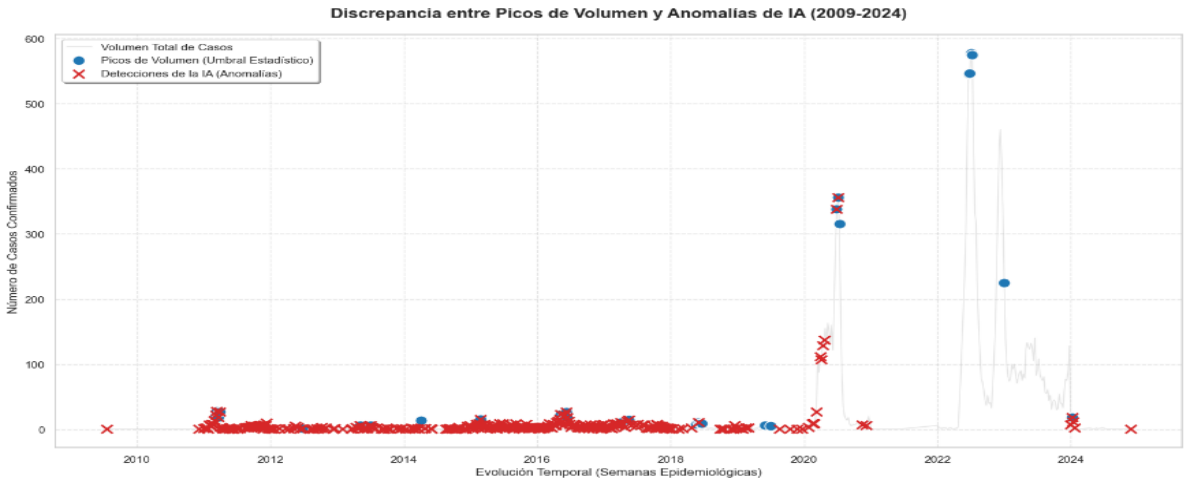
Recall (4.20%):El bajo nivel de exhaustividad (Recall) confirma que el modelo no es redundante. Si el Recall fuera muy alto, el modelo de IA estaría simplemente copiando lo que ya hace la estadística tradicional. El valor del 4.20% demuestra que la IA es altamente selectiva y se enfoca solo en una fracción crítica de los brotes, ignorando los picos estacionales predecibles

para concentrarse en las rupturas de la firma epidemiológica.

F1-Score (0.0681):Esta métrica de síntesis ratifica que estamos ante un sistema de vigilancia de segunda línea. El objetivo del modelo no es el conteo masivo, sino la detección de anomalías o eventos de baja frecuencia, pero alto riesgo demográfico que el Canal Endémico (+2σ) es incapaz de procesar por su propia naturaleza agregada.

Figura 17

Discrepancia Entre Picos de Volumen y Anomalías (2009 -2024)



La Figura 17 resume visualmente la justificación del proyecto. Se observa cómo las alertas tradicionales (puntos azules) solo aparecen en las crestas de las montañas de datos, cuando la capacidad del sistema ya está comprometida. Por el contrario, las alertas del modelo (cruces rojas) aparecen distribuidas a lo largo de toda la serie, identificando riesgos en los valles de la gráfica, donde el volumen es bajo, pero la rareza etaria es máxima. Esto confirma que la IA proporciona una ventana de oportunidad para la intervención temprana que el método actual no posee.

Consideraciones sobre el Estándar Comparativo +2σ

Es fundamental precisar que el umbral de dos desviaciones estándar, basado en la metodología de corredores endémicos de Bortman (1999), se utiliza en esta investigación exclusivamente como un estándar comparativo de referencia y no como una verdad absoluta o un diagnóstico definitivo de brote.

El +2σ es una medida de cantidad que solo se activa cuando el volumen masivo de reportes supera el promedio histórico por lo tanto implica que un cambio drástico en el perfil demográfico (por ejemplo, un aumento inusual de casos en neonatos) puede pasar desapercibido si el volumen total de la población no excede las dos desviaciones estándar.

La validación realizada en este estudio no busca la réplica exacta de la alerta de +2σ sino demostrar que el modelo analítico es capaz de complementar la vigilancia tradicional, aportando una capa de detección cualitativa que el estándar comparativo actual no posee.

Limitaciones de la Investigación

Cambios Normativos. La consistencia de los datos analizados está sujeta a cambios en los protocolos de vigilancia, como se observó, la transición en los criterios de notificación (como los estipulados en la Circular 052 de 2022) genera variaciones en la masa crítica de datos que pueden ser interpretadas por el algoritmo como anomalías estadísticas, cuando en realidad responden a cambios administrativos en el reporte.

Dependencia de la Calidad de la Fuente (SIVIGILA). La efectividad del Isolation Forest depende directamente de la calidad del ingreso de datos en las Unidades Primarias Generadoras de Datos (UPGD). Errores de digitación en la edad, el sexo o la codificación de la enfermedad en las IPS pueden generar ruido que el modelo clasifica como anomalía, sin que exista un riesgo epidemiológico real.

Falta de Validación Clínica Directa. Los hallazgos presentados en este trabajo constituyen alertas de sospecha estadística, debido al alcance de la investigación, no fue posible realizar una confrontación en campo o una auditoría de historias clínicas para confirmar si cada anomalía detectada correspondió efectivamente a un brote biológico o a una falla en la prestación del servicio.

Resolución de Georreferenciación: El análisis espacial se basa en la ubicación de la institución de salud (UPGD) que reporta el caso y no en la ubicación exacta de residencia del paciente, esto limita la capacidad del modelo para identificar clústeres geográficos finos (barrios o localidades) donde podría estar originándose el foco de contagio.

Naturaleza del Aprendizaje No Supervisado: El modelo identifica la extrañeza estadística, lo cual es una condición necesaria pero no suficiente para declarar una crisis de salud pública, la distinción definitiva entre un error de datos y un evento epidemiológico inusitado seguirá requiriendo la intervención de expertos en epidemiología de campo.

Fase VI Despliegue (Deployment):

En esta fase final, se definen las estrategias para integrar los hallazgos del modelo de detección de anomalías en la operación cotidiana de la Secretaría Distrital de Salud. El objetivo es transformar la capacidad analítica en una herramienta de gestión preventiva.

Plan de Implementación Operativa

Se propone la creación de un Protocolo de Auditoría de Alta Selectividad. En lugar de realizar inspecciones aleatorias, el equipo de vigilancia epidemiológica utilizará los resultados del modelo para:

Focalización Institucional: Priorizar las visitas técnicas a las IPS identificadas como Nodos Críticos

Auditoría de Casos de Consenso: Realizar una revisión clínica profunda de los 121 casos identificados simultáneamente por IF y LOF, para verificar si corresponden a errores de captura de datos o a brotes epidemiológicos emergentes.

Integración en Tableros de Control (Dashboarding)

El despliegue técnico contempla la automatización del flujo de datos desde el SIVIGILA hacia un tablero de visualización (Power BI o Tableau). Esto permitirá:

Monitoreo en Tiempo Real: Visualizar la aparición de anomalías semana a semana.

Alertas Tempranas de Perfil: Activar una señal de alerta cuando la proporción de anomalías en Primera Infancia e Infancia supere el umbral histórico, especialmente durante el mes de marzo.

Plan de Monitoreo y Mantenimiento del Modelo

Se reconoce que los datos epidemiológicos sufren de cambios en el comportamiento del virus por ello, se recomienda:

Re-entrenamiento Anual: Ajustar los parámetros de contaminación y los estimadores del Isolation Forest cada año con los nuevos reportes del SIVIGILA.

Validación de Campo: Retroalimentar el modelo con los resultados de las auditorías físicas para ajustar la sensibilidad del algoritmo y reducir falsos positivos.

Herramientas y Tecnologías

El desarrollo del proyecto se realizará utilizando las siguientes herramientas tecnológicas de código abierto:

- Lenguaje de Programación: Python 3.10+.
- Entorno de Desarrollo: Jupyter Notebooks
- Librerías de Procesamiento:

- Pandas: Para manipulación de marcos de datos (dataframes).
- NumPy: Para operaciones matemáticas vectoriales.

- Librerías de Machine Learning:
- Scikit-learn: Para la implementación de algoritmos de detección de anomalías (Isolation Forest, SVM).
- Visualización:
- Matplotlib / Seaborn: Para la generación de gráficas estadísticas y análisis visual de los resultados.

Fuentes de Información

Dataset Principal: SIVIGILA. El proyecto se sustentará en los microdatos anonimizados del evento 348 (IRAG Inusitada) del SIVIGILA, abarcando el periodo 2009-2024. Esta base de datos cuenta con más de 70 variables por registro, incluyendo datos demográficos, geográficos y clínicos, lo que garantiza una profundidad analítica suficiente para el entrenamiento de modelos no supervisados y la validación de alertas tempranas mediante el contraste con desenlaces clínicos (hospitalización y mortalidad)

Conclusiones

Respecto al Análisis y Consistencia de Datos

Se logró la consolidación y limpieza técnica de una serie histórica de 15 años (2009-2024), superando los desafíos de inconsistencia propios de la plataforma de Datos Abiertos.

La aplicación de técnicas de normalización y ponderación etaria en la Fase 3 de la metodología CRISP-DM fue el factor determinante para neutralizar el sesgo de volumen en adultos. Esto permitió que el sistema de vigilancia dejara de ser un simple contador de casos y se convirtiera en un modelo sensible a la vulnerabilidad pediátrica, cumpliendo así con la garantía de calidad y consistencia exigida.

Respecto a la Implementación y Benchmarking de Algoritmos

La implementación del algoritmo Isolation Forest demostró ser superior para la identificación de anomalías de perfil global, logrando aislar 886 registros atípicos con un enfoque preventivo.

El benchmarking realizado con el modelo Local Outlier Factor (LOF) permitió validar la robustez de los hallazgos mediante un consenso del 13.66% (121 casos). Esta comparativa técnica cumplió con el objetivo de priorizar desviaciones que superan la estacionalidad histórica, revelando que el sistema es capaz de detectar riesgos tanto por densidad local como por rareza estructural, especialmente en el mes de marzo.

Respecto a la Evaluación del Rendimiento

La evaluación mediante métricas de clasificación reveló una Precisión del 18.06% y un Recall del 4.20% al contrastar con los brotes históricos oficiales (+2σ). Aunque el F1-Score (0.0681) parece mínimo, bajo una óptica de aprendizaje supervisado tradicional, en el contexto de detección de anomalías se concluye que es un resultado exitoso. El modelo no solo identifica

picos de volumen conocidos por el INS, sino que detecta un 82% de eventos atípicos adicionales que la estadística convencional ignora, cumpliendo con la meta de fortalecer la toma de decisiones preventivas ante eventos silenciosos.

En general ,se desarrolló con éxito un sistema de vigilancia sindrómica basado en aprendizaje automático que transforma datos históricos en inteligencia operativa. El modelo permite pasar de una vigilancia reactiva (basada en el colapso del sistema) a una vigilancia proactiva (basada en el perfil del riesgo), proporcionando a la Secretaría de Salud un mapa claro de Nodos Críticos y grupos vulnerables para la mitigación temprana de las IRA en la capital.

Recomendaciones

Se recomienda dedicar la mayor parte del esfuerzo del proyecto a la fase de limpieza y curaduría de la información. Es fundamental asegurar que las fechas de inicio de síntomas sean validadas y depuradas, ya que este dato es el que permite al modelo anticiparse a las alertas basadas en el diagnóstico oficial. Sin una preparación rigurosa, el modelo carecería de la sensibilidad necesaria para actuar como centinela.

Para proyectos de analítica en etapas tempranas, se recomienda no intentar abarcar la totalidad de los eventos epidemiológicos. La estrategia ganadora consiste en delimitar el alcance a un grupo de eventos de alto impacto, como las Infecciones Respiratorias Agudas (IRA). Esto permite ajustar los algoritmos a las dinámicas específicas de una patología antes de escalar el sistema a otros grupos de enfermedades.

La validación de un modelo de Ciencia de Datos en salud no debe ser estrictamente matemática. Se recomienda realizar un contraste cualitativo y cuantitativo en conjunto con expertos en epidemiología, utilizando los boletines históricos de Bogotá, este proceso asegura que las anomalías detectadas por el algoritmo tengan coherencia con la realidad clínica y los brotes documentados en el pasado.

Es indispensable mantener protocolos estrictos de anonimización y manejo de la información. Aunque el análisis se realice sobre datos agregados, el diseño del proyecto debe garantizar que en ninguna fase se expongan datos sensibles que permitan identificar a los pacientes. El cumplimiento de la Ley de Habeas Data y los principios bioéticos debe ser el eje transversal de cualquier implementación tecnológica en salud pública.

Referencias Bibliográficas

Alcaldía Mayor de Bogotá. (2004, 11 de octubre). *Decreto 332 de 2004*: *Por el cual se organiza el Régimen y el Sistema para la Prevención y Atención de Emergencias en Bogotá D.C*. Registro Distrital 3194. https://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=15027

Aracena, C., Villena, F., Arias, F., & Dunstan, J. (2022). *Applications of machine learning in healthcare*. *Revista Médica Clínica Las Condes*, 33(6), 568–575. https://doi.org/10.1016/j.rmclc.2022.10.001

Bishop, C. M. (2006). *Pattern recognition and machine learning. Springer*.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys (CSUR), 41(3), 1–58. https://www.researchgate.net/publication/220565847_Anomaly_Detection_A_Survey

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide.* SPSS Inc. https://www.the-modeling-agency.com/crisp-dm.pdf

Concejo de Bogotá. (2016, 6 de abril). *Acuerdo 641 de 2016: Por el cual se efectúa la reorganización del Sector Salud en el Distrito Capital*. Registro Distrital 5811. https://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=65637

Congreso de la República de Colombia. (1979, 24 de enero). *Ley 9 de 1979: Código Sanitario Nacional*. Diario Oficial No. 35308. https://www.minsalud.gov.co/Normatividad_Nuevo/LEY%200009%20DE%201979.pdf

Congreso de la República de Colombia. (2012, 17 de octubre). *Ley Estatutaria 1581 de 2012: Por la cual se dictan disposiciones generales para la protección de datos personales*.

Diario Oficial No. 48.587. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=49981

Congreso de la República de Colombia. (2020, 31 de enero). *Ley 2015 de 2020: Por medio de la cual se crea la Historia Clínica Electrónica Interoperable*. Diario Oficial No. 51.213. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=105684

García, S., Luengo, J., & Herrera, F. (2015). *Data Preprocessing in Data Mining*. Springer. https://link.springer.com/book/10.1007/978-3-319-10247-4

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). *Detecting influenza epidemics using search engine query data. Nature*, 457(7232), 1012–1014. https://doi.org/10.1038/nature07634

Henning, K. J. (2004). *Overview of syndromic surveillance: What is syndromic surveillance? Morbidity and Mortality Weekly Report (MMWR)*, 53(Suppl), 5–11. https://www.cdc.gov/mmwr/preview/mmwrhtml/su5301a3.htm

Hernández Orallo, J., Ramírez Quintana, M. J., & Ferri Ramírez, C. (2004). *Introducción a la minería de datos.* Pearson Educación.

IBM Corporation. (2021). Guía de CRISP-DM de IBM SPSS Modeler. IBM Knowledge Center. https://www.ibm.com/docs/es/SS3RA7_18.4.0/pdf/ModelerCRISPDM.pdf

Instituto Nacional de Salud (INS). (2017). *Lineamientos para el análisis de datos de la vigilancia en salud pública. Gobierno de Colombia*. https://www.ins.gov.co/Direcciones/Vigilancia/Lineamientosydocumentos/Lineamientos%202018.pdf

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). *The parable of Google to flu: traps in big data analysis. Science,* 343(6176), 1203–1205.

https://doi.org/10.1126/science.1248506

López Ullauri, V. G., Pino Falconí, P. R., Zambrano Nuñez, T. M., & Romero Machado, E. R. (2024). *Impacto de la inteligencia artificial en salud pública*. AlfaPublicaciones, 6(4), 158–173. https://doi.org/10.33262/ap.v6i4.562

Lugo Lopez, N. D. (2024, 21 de junio). *Aplicaciones de la Ciencia de Datos en la Industria* [Objeto virtual de aprendizaje OVA]. Repositorio Institucional UNAD. https://repository.unad.edu.co/handle/10596/62692

Matéus Solarte, J. C. (2024). *Guía metodológica para el desarrollo de protocolos de vigilancia de eventos de interés en salud pública en Colombia*. Instituto Nacional de Salud.

Ministerio de la Protección Social. (2006, 9 de octubre). *Decreto 3518 de 2006: Por el cual se crea y reglamenta el Sistema de Vigilancia en Salud Pública.* Diario Oficial No. 46.417. https://www.minsalud.gov.co/Normatividad_Nuevo/DECRETO%203518%20DE%202006.pdf

Ministerio de Salud. (1993, 4 de octubre). *Resolución 8430 de 1993: Por la cual se establecen las normas científicas, técnicas y administrativas para la investigación en salud*. https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/DE/DIJ/RESOLUCION-8430-DE-1993.PDF

Ministerio de Salud y Protección Social. (2018). *Vigilancia sindrómica en Puntos de Entrada marítimos y fluviales: Procedimiento operativo estandarizado.* Gobierno de Colombia. https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/ED/VSP/vigilancia-sindromica-puntos-entrada-maritimos-fluviales.pdf

Moine, J. M., Haedo, A. S., & Gordillo, S. E. (2011). *Metodología CRISP-DM aplicada a la minería de datos en sistemas de información geográfica*. *Revista Facultad de Ingeniería*,

20, 9-20. https://dialnet.unirioja.es/servlet/articulo?codigo=3856549

Organización Panamericana de la Salud (OPS). (2019). *Sistema de Alerta Temprana y Respuesta* (EWARS). OPS/OMS. https://www.paho.org/es/temas/reglamento-sanitario-internacional/alertas-y-respuesta

Organización Panamericana de la Salud (OPS). (2006). *Vigilancia en salud pública: más allá de las enfermedades transmisibles*. Boletín Epidemiológico, 27(2).

Organización Panamericana de la Salud (OPS). (2011). *Módulos de principios de epidemiología para el control de enfermedades (MOPECE)*: Unidad 4: Vigilancia en salud pública (2.ª ed. rev.). Washington, D.C.: OPS. https://iris.paho.org/bitstream/handle/10665.2/55842/9789275319802_spa.pdf

Organización Panamericana de la Salud (OPS). (2021). *Inteligencia artificial en la salud pública*: informe técnico. Washington, D.C.: OPS. https://iris.paho.org/bitstream/handle/10665.2/53887/OPSEIHIS21011_spa.pdf

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V. & Duchesnay, E. (2011). *Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research*, 12, 2825-2830. https://jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Pineda, J. M. (2022). *Predictive models in health based on machine learning*. *Revista Médica Clínica Las Condes*, 33(6), 583–590. https://doi.org/10.1016/j.rmclc.2022.11.002

Polo-Triana, S. I., Ramírez-Sierra, Y. A., Arias-Osorio, J. E., Martínez-Vega, R. A., & Lamos-Díaz, H. (2022). *Métodos de aprendizaje automático para predecir el comportamiento epidemiológico de enfermedades arbovirales: revisión estructurada de literatura. Salud UIS*, 55(1). https://doi.org/10.18273/saluduis.55.e:23017

Raschka, S., Patterson, J., & Nolet, C. (2020). *Machine learning in Python: Main developments*

and technology trends in data science, machine learning, and artificial intelligence. Information, 11(4), 193. https://doi.org/10.3390/info11040193

Sánchez Sarmiento, C. A., & Urieles Sierra, K. I. (2024). *Protocolo de Vigilancia en salud pública. Brotes de infecciones asociadas a la atención en salud*. Instituto Nacional de Salud. https://doi.org/10.33610/DEET5834

Swaminathan, S. (2021). *Ethics and governance of artificial intelligence for health: WHO guidance*. World Health Organization.

Van Rossum, G., & Drake, F. L. (2009). *The Python Tutorial. Python Software Foundation*. https://docs.python.org/3/tutorial/

Walz, W. (Ed.). (2023). *Machine Learning for Brain Disorders. Springer*. http://www.springer.com/series/7657

Instituto Nacional de Salud. (2024). *Protocolo de vigilancia en salud pública: Infección Respiratoria Aguda (IRA) (Versión 09)*. Dirección de Vigilancia y Análisis del Riesgo en Salud Pública. https://www.ins.gov.co/

Instituto Nacional de Salud. (2022). *Boletín Epidemiológico Semanal: Semana epidemiológica 52* (25 al 31 de diciembre de 2022). https://www.ins.gov.co/buscador-eventos/BoletinEpidemiologico/2022_Bolet%C3%ADn_epidemiologico_semana_52.pdf

Ministerio de Salud y Protección Social e Instituto Nacional de Salud. (2024). *Lineamientos nacionales para la vigilancia en salud pública 2024.* https://www.ins.gov.co/BibliotecaDigital/lineamientos-nacionales-2024.pdf

Bortman, M. (1999). *Elaboración de corredores endémicos mediante planillas de cálculo*. *Revista Panamericana de Salud Pública*, 5(1), 1–8. https://www.scielosp.org/pdf/rpsp/v5n1/5n1a1.pdf

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... & SciPy 1.0 Contributors. (2020). *SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods,* 17(3), 261-272. https://doi.org/10.1038/s41592-019-0686-2

Bhandari, A. (2021, mayo 19). *Feature Scaling Techniques in Python: A Complete Guide. Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2021/05/feature-scaling-techniques-in-python-a-complete-guide/