Dashboard Interactivo con Modelo de Red Neuronal LSTM para la Segmentación y Análisis de la Población Afectada por Infección Respiratoria Aguda Grave (IRAG) en Bogotá

Baquero Criollo Diego Alejandro, Sabogal Suarez Cristian Felipe y Franco Calderón José Alejandro
Universidad de Bogotá Jorge Tadeo Lozano

Abstract - This work paper presents an epidemiological monitoring system for Severe Acute Respiratory Infection (SARI; or Infección Respiratoria Aguda Grave, IRAG) in Bogotá that integrates the cleansing and exploratory analysis of 71.326 SIVIGILA-INS records (2018-2023) with a predictive model based on multilayer LSTM neural network. After benchmarking decision trees and random forests on weekly time-series data enriched with demographic variables and sine–cosine seasonal components, the LSTM (optimized with normalization, dropout, and early stopping) offered the best balance between fit and generalization.

The model projects approximately 5.000 cases for 2024-2025, concentrated (60%) in vulnerable populations and (67%) in lower socioeconomic strata, with the peak shifting toward infants and young adults. Results are deployed in an interactive Power BI dashboard that enables segmentation by year, sex, age group, socioeconomic stratum, and vulnerability, providing clear evidence to prioritize health interventions and reduce public-health inequalities.

Keywords: LSTM neural networks, Interactive dashboard, Infección Respiratoria Aguda Grave (IRAG), Power BI y Epidemiological prediction.

I. INTRODUCCIÓN

Las infecciones respiratorias agudas graves (IRAG) constituyen una de las principales causas de morbilidad y mortalidad a nivel global, con un impacto especialmente grande en niños menores de cinco años, adultos mayores y poblaciones vulnerables. En Bogotá, estas patologías presentan una alta prevalencia y exhiben patrones estacionales complejos, lo cual exige el desarrollo de estrategias eficaces de monitoreo, prevención y respuesta oportuna. En este contexto, la creciente disponibilidad de datos abiertos y los avances en inteligencia artificial (IA) ofrecen nuevas oportunidades para transformar el análisis y la gestión de estos eventos de salud pública.

La persistencia de las IRAG como amenaza sanitaria en la ciudad limita la capacidad de las entidades para realizar una vigilancia epidemiológica oportuna y dificulta la detección temprana de patrones emergentes o de poblaciones en riesgo. La escasa disponibilidad de información analítica y la fragmentación de los datos existentes constituyen, por tanto, un obstáculo significativo para el fortalecimiento de las estrategias de salud pública en Bogotá.

El presente artículo tiene como objetivo diseñar un sistema interactivo en Power BI que permita visualizar de forma detallada los casos de IRAG en Bogotá, incorporando un modelo de inteligencia artificial para predecir la evolución futura de los posibles casos confirmados segmentando grupos etarios, estrato socioeconómico y población vulnerable a partir del análisis de datos históricos entre 2018 y 2023. La motivación central es ofrecer una herramienta práctica y accesible para apoyar la planeación sanitaria, priorizando información que pueda contribuir a la toma de decisiones preventivas.

El artículo se organiza de la siguiente manera: en primer lugar, se presenta el marco referencial, que contextualiza los antecedentes teóricos y tecnológicos del problema. Luego, se describe el proceso de recopilación, limpieza y análisis de los datos utilizados. Posteriormente, se detallan los modelos de inteligencia artificial implementados y los criterios empleados para su comparación. A continuación, se expone el diseño del dashboard interactivo desarrollado en Power BI. Finalmente, se discuten los resultados obtenidos, se comparan los modelos predictivos y se concluye el trabajo realizado.

II. MARCO REFERENCIAL

A. Estado del Arte

Entre 2015 y 2016, se realizó un estudio descriptivo, retrospectivo y de corte transversal para analizar factores clínicos y epidemiológicos en pacientes con neumonía. Se reportaron 610 casos en total, con un aumento del 15% en 2015 (366 casos) respecto a 2014 (253 casos). La información se recopiló durante un periodo de 2 años con un muestreo en

pacientes, realizando informes epidemiológicos semanalmente, se procesó con SPSS y Excel, y se presentó en tablas y gráficos. Los resultados destacan la utilidad del análisis sistemático para identificar tendencias epidemiológicas en la población infantil [1].

Entre el 2012 y 2016, se utilizó minería de datos con el algoritmo de clustering K-Means para analizar enfermedades respiratorias agudas (IRA) en Bogotá. El estudio generó un modelo descriptivo con cuatro clústeres, destacando patrones como la mayor prevalencia en niños de 3 a 5 años y una distribución de género equilibrada. Se emplearon herramientas como RapidMiner, SAS, software IBM y datos públicos de SISPRO, SIVIGILA y DANE para el análisis [2].

En [3] el trabajo evidenció un desarrolló de un modelo estadístico basado en el método de mínimos cuadrados para pronosticar casos de infecciones respiratorias agudas (IRA) y enfermedades diarreicas agudas (EDA) en un hospital, utilizando datos históricos como base. El modelo asocia años como variable independiente y número de casos como variable dependiente, permitiendo identificar periodos de mayor incidencia. Aplicado en la sede principal y municipios de su red, el modelo mostró un error promedio del 3% al comparar los casos pronosticados con los reales. La herramienta utilizada para el análisis y proyecciones fue Microsoft Excel.

Este estudio [4] analizó las infecciones respiratorias agudas en la provincia de Cienfuegos, considerando variables como edad, municipio, demanda médica y evolución clínica. Se identificó mayor incidencia en menores de 5 años y mayores de 60, siendo la neumonía la principal causa de muerte. Los virus predominantes fueron similares al respiratorio, parainfluenza, influenza A y coronavirus, con una tendencia ascendente. A partir de la semana 11, se observó un incremento en consultas, posiblemente relacionado con acciones de vigilancia. Los datos fueron procesados en SPSS y presentados en gráficos y tablas, utilizando registros del CPHEM como fuente principal.

En el año 2020, se realizó un estudio descriptivo en Colombia para analizar las características epidemiológicas y clínicas de los casos de infecciones respiratorias agudas (IRA) durante las semanas epidemiológicas 01 a 53. Utilizando la metodología de Bortman, se construyeron canales endémicos con datos de morbilidad por IRA. A nivel nacional, se observó una disminución en la notificación de casos en consulta externa, urgencias y hospitalizaciones generales, pero un aumento en las hospitalizaciones por IRA en unidades de cuidados intensivos e intermedios. Los datos fueron obtenidos de los registros del Sistema Sivigila y la base de datos SISMUESTRAS, consolidados por el Instituto Nacional de Salud (INS) [5].

En Valledupar, Cesar, se realizó un estudio cuantitativo no experimental y descriptivo para analizar la incidencia de enfermedades respiratorias en la Unidad de Cuidados Pediátricos. Los datos se recolectaron mediante cuestionarios estructurados según grupos etarios, como lactantes menores y mayores. La neumonía fue la enfermedad más prevalente, seguida del asma y la laringotraqueitis, asociadas al déficit en el tratamiento oportuno en la red ambulatoria del programa IRA. Los datos fueron procesados y analizados en Epi-info, Excel y SPSS, presentándose en tablas y gráficos con interpretación detallada [6].

SPREAD, una aplicación web de código abierto que integra datos genómicos, espaciales y temporales de patógenos para analizar y visualizar la propagación de enfermedades infecciosas de manera intuitiva. Utilizando tecnologías como JavaScript, Leaflet y GrapeTree, facilita el rastreo de rutas de transmisión y la identificación de clusters genéticos, sin requerir infraestructura informática compleja. Probada con patógenos como Listeria monocytogenes y SARS-CoV-2, combina herramientas bioinformáticas como fastp, shovill y chewBBACA para un análisis integral, mientras su arquitectura autónoma garantiza privacidad de los datos y fomenta la colaboración científica [7].

El panel interactivo del Estudio de Cohorte Comunitaria de Rakai (RCCS) en Uganda es un ejemplo innovador de democratización de datos epidemiológicos. Utilizando Microsoft SQL Server y Tableau, permite explorar tendencias del VIH de forma accesible y segura mediante técnicas de anonimización basadas en la Regla de Privacidad de HIPAA. Este tablero, integrado en la web del RHSP, analiza incidencia, prevalencia y estrategias preventivas, facilitando la generación de hipótesis y promoviendo la reproducibilidad y colaboración científica. Desde el 2019, ha recibido más de 3,000 visitas, destacando su impacto en la investigación y acceso público a datos [8].

El NIMR-MDB (National Institute of Malaria Research - Malaria Dashboard) es una herramienta digital innovadora desarrollada en R mediante el paquete Shiny, diseñada para analizar datos epidemiológicos de malaria en India. Este panel interactivo permite visualizar 16 indicadores clave, como la incidencia anual de parásitos (API) y la distribución de especies de Plasmodium, a nivel nacional, estatal y distrital. Con 14 pestañas y soporte para uso en línea o local, facilita la planificación de estrategias y evaluación de intervenciones. Además, su flexibilidad y capacidad de autenticación lo posicionan como un recurso clave para la investigación y control de la malaria, con perspectivas de expansión hacia datos ambientales y resistencia a insecticidas. [9]

Un estudio sobre la mortalidad hospitalaria en pacientes con Infección Respiratoria Aguda Grave (IRAG) utilizó modelos de aprendizaje automático aplicados a datos administrativos de 86 hospitales en Alemania, analizando 241,988 casos entre 2016 y 2019. Los algoritmos empleados incluyeron Regresión Logística (GLM), Bosque Aleatorio (RF), Red Neuronal (NNET) y XGBoost, mostrando un rendimiento superior de NNET y XGBoost frente a métodos tradicionales. Técnicas como validación cruzada, análisis SHAP y curvas ROC/AUPRC destacaron la relevancia de las comorbilidades en la predicción de mortalidad. Los modelos permiten estratificar riesgos y facilitar evaluaciones comparativas, aunque los autores subrayan la necesidad de validaciones externas para confirmar su aplicabilidad clínica [10].

El pronóstico de hospitalizaciones por Infección Respiratoria Aguda Grave (SARI) se realizó utilizando AutoGluon-TS en Python, con modelos como Temporal Fusion Transformer (TFT) y DeepAR. Empleando validación de un paso hacia adelante y datos de laboratorio como covariables, los resultados mostraron que estos enfoques superaron los métodos tradicionales. El TFT destacó por su precisión y robustez, evidenciando su utilidad en aplicaciones prácticas de pronóstico.[11]

En [12] en Brasil se muestra cómo se desarrollaron modelos predictivos para brotes de Infección Respiratoria Aguda Grave (SRAG) utilizando redes neuronales LSTM (Long Short-Term Memory) y datos de hospitalizaciones entre el 2013 y 2020, excluyendo casos de COVID-19. Los modelos lograron alta precisión en predecir picos estacionales, volumen de notificaciones y el inicio del período pre-epidémico, con un R^2 de 0.97 en la región Sur para 2019. Se emplearon herramientas como Knime para preparación de datos y R para análisis, junto con técnicas como SARIMA y algoritmos alternativos como Random Forest y Naive Bayes. Los resultados destacan la utilidad de estos modelos en la planificación sanitaria, aunque reconocen limitaciones para eventos excepcionales como pandemias.

El artículo propuesto en [13] describe el protocolo del estudio COLEV, un proyecto interdisciplinario en Colombia que utiliza inteligencia artificial (IA) y ciencia de datos para enfrentar los retos de la pandemia de COVID-19. El estudio busca desarrollar modelos de pronóstico, analizar el impacto de la COVID-19, analizar las opiniones y discusiones de las personas y abordar la información falsa, como en temas de vacunación, salud mental y poblaciones vulnerables. Se utilizan métodos mixtos, combinando datos cualitativos y cuantitativos, y tecnologías como aprendizaje automático, procesamiento del lenguaje natural (NLP) y análisis de redes sociales. Aunque aún no presentan resultados concretos, se resalta la importancia de usar la inteligencia artificial de forma responsable para crear políticas públicas que tengan en cuenta las desigualdades sociales en Colombia.

B. Marco conceptual

A continuación, se presentan una serie de conceptos relevantes de acuerdo con el contexto del tema que será desarrollado a lo largo del trabajo.

Morbilidad: Se refiere a la presentación de una enfermedad o síntoma de una enfermedad, o a la proporción de enfermedad en una población [14].

Mortalidad: Término que se refiere a la cualidad o el estado de mortal (destinado a morir). En el campo de la medicina, este término también se usa para la tasa de muertes, tasa de mortalidad o el número de defunciones en cierto grupo de personas en determinado período. Es posible notificar la mortalidad de personas con cierta enfermedad, que viven en un área del país o que son de determinado sexo, edad o grupo étnico [15].

Inteligencia artificial: En [16] lo definen como un campo de la ciencia que busca desarrollar máquinas capaces de razonar, aprender y actuar como los humanos o procesar grandes volúmenes de datos. Abarca disciplinas como informática, estadística, neurociencia y psicología. En el ámbito empresarial, se basa en tecnologías como el aprendizaje automático y profundo para analizar datos, generar predicciones, procesar lenguaje natural y optimizar diversas tareas.

Modelos de IA: En [17] se definen como un programa que ha sido entrenado en un conjunto de datos para reconocer ciertos patrones o tomar ciertas decisiones sin más intervención humana. Los modelos de inteligencia artificial aplican distintos algoritmos a las entradas de datos relevantes para lograr las tareas, u outputs, para los que han sido programados. Los autores también definen que los modelos de regresión predicen valores continuos (como el precio, la antigüedad, el tamaño o el tiempo) y se utilizan principalmente para determinar la relación entre una o más variables independientes (x) y una variable dependiente (y): dado x, predecir el valor de y. Por último, explican que los modelos de clasificación predicen valores discretos, se utilizan principalmente para determinar una etiqueta adecuada o para categorizar (es decir, clasificar). Puede ser una clasificación binaria, como “sí o no”, “aceptar o rechazar”, o una clasificación multiclase (como un motor de recomendación que sugiere el producto A, B, C o D).

Algunos ejemplos comunes de modelos son:

- *Naïve Bayes*: Un algoritmo generativo de aprendizaje supervisado utilizado habitualmente en el filtrado de spam y la clasificación de documentos.
- *Regresión Logística*: Predice probabilidades continuas que luego se utilizan como proxy para los rangos de clasificación.
- *K-means:* Es un algoritmo de agrupamiento iterativo basado en centroides que divide un conjunto de datos en grupos similares en función de la distancia entre sus centroides. El centroide, o centro del clúster, es la media o la mediana de todos los puntos dentro del clúster, según las características de los datos [18].
- *Random Florest (Bosque Aleatorio):* Es un algoritmo de aprendizaje automático de uso común, registrado por Leo Breiman y Adele Cutler, que combina el resultado de múltiples árboles de decisión para llegar a un resultado único. Su facilidad de uso y flexibilidad han impulsado su adopción, ya que maneja problemas de clasificación y regresión [19].
- *Árbol de decisión*: Su objetivo es encontrar la mejor división para los subconjuntos de datos, por lo general,

se entrenan a través del algoritmo del árbol de clasificación y regresión (CART). Las métricas, como la impureza de Gini, la ganancia de información o el error cuadrático medio (MSE), pueden utilizarse para evaluar la calidad de la división [19].

Dashboard: En [20] se indica como la herramienta ideal para las empresas que desean tener acceso rápido y en tiempo real a los principales indicadores de gestión de negocios. Este dispositivo es fundamental para identificar buenos resultados o métricas que deben mejorar.

Procesamiento de lenguaje natural: El procesamiento del lenguaje natural (PLN) [21], como rama de la inteligencia artificial, utiliza el aprendizaje automático para procesar e interpretar textos y datos. El reconocimiento y la generación de lenguaje natural son tipos de PLN.

Red neuronal: Una red neuronal, [22] es un programa o modelo de aprendizaje automático que toma decisiones de manera similar al cerebro humano, mediante el uso de procesos que imitan la forma en que las neuronas biológicas trabajan juntas para identificar fenómenos, sopesar opciones y llegar a conclusiones.

XGBoost: En [23] lo definen como un método de aprendizaje automático supervisado para clasificación y regresión, se utiliza en la herramienta para entrenar con AutoML. XGBoost es la abreviatura de las palabras inglesas "extreme gradient boosting" (refuerzo de gradientes extremo). Este método se basa en árboles de decisión y supone una mejora sobre otros métodos, como el bosque aleatorio y refuerzo de gradientes. Funciona bien con datasets grandes y complejos al utilizar varios métodos de optimización.

API: Las API [24] son mecanismos que permiten a dos componentes de software comunicarse entre sí mediante un conjunto de definiciones y protocolos. Por ejemplo, el sistema de software del instituto de meteorología contiene datos meteorológicos diarios. La aplicación meteorológica de su teléfono “habla” con este sistema a través de las API y le muestra las actualizaciones meteorológicas diarias en su teléfono.

DANE: Manejar los recursos y financiar la realización por parte del Departamento Administrativo Nacional de Estadística, DANE, de los censos nacionales y de las encuestas que servirán de base para los programas y proyectos de carácter tecnológico y de desarrollo que establezcan las normas legales vigentes sobre la materia y el Gobierno Nacional [25].

IRAG: Infección Respiratoria Aguda Grave (IRAG) se usa para monitorear a las personas con enfermedad respiratoria más grave que han sido admitidas a un hospital [26].

RapidMiner: Es una de las herramientas de minería de datos o data mining más utilizada y sencilla por lo que es muy recomendada para su uso en entornos menos técnicos. Su sistema de programación visual (Drag&Drop) requiere de una menor curva de aprendizaje logrando mayor productividad en menos tiempo [27].

SIVIGILA: Es el Sistema Nacional de Vigilancia en Salud Pública -SIVIGILA, que se ha creado para realizar la provisión en forma sistemática y oportuna de información sobre la dinámica de los eventos que afecten o puedan afectar la salud de la población colombiana [28].

SISPRO: Información oportuna, suficiente y estandarizada para la toma de decisiones del Sector Salud y Protección Social, centrada en el Ciudadano. El SISPRO está conformado por bases de datos y sistemas de información del Sector sobre oferta y demanda de servicios de salud, calidad de los servicios, aseguramiento, financiamiento, promoción social [29].

SPSS: Software de análisis estadístico avanzado, cuenta con una amplia biblioteca de algoritmos de aprendizaje automático, análisis de texto, extensibilidad de código abierto, integración con big data y una implementación fluida en aplicaciones [30].

Metodología de bortman: Método para el cálculo de los canales endémicos, el método sistémico en la determinación de las herramientas para el desarrollo del software, la metodología XP en la modelación, definición de las etapas de desarrollo del sistema de gestión y la Norma ISO/IEC 9126 para evaluar la calidad [31].

SARS-CoV-2: En [32] Virus que causa una enfermedad respiratoria llamada enfermedad por coronavirus de 2019 (COVID-19). El SARS-CoV-2 es un virus de la gran familia de los coronavirus.

chewBBACA: Es un paquete de software para la creación y evaluación de esquemas y resultados de tipificación de secuencias de múltiples loci del genoma central y del genoma completo [33].

SQL: Lenguaje de programación para almacenar y procesar información en una base de datos relacional. Una base de datos relacional almacena información en forma de tabla, con filas y columnas que representan diferentes atributos de datos y las diversas relaciones entre los valores de datos [34].

Tableau: Plataforma de análisis visual que transforma la manera en que usamos los datos para resolver problemas. Además, permite a las personas y las organizaciones sacar el máximo partido de los datos [35].

HIPAA: En [36] es la Ley de Portabilidad y Responsabilidad del Seguro Médico. Esta ley estadounidense se aprobó en 1996 para garantizar la protección de los datos personales de salud, incluyendo las copias impresas y la información compartida verbal o digitalmente.

Análisis SHAP: Es un marco teórico y una biblioteca de Python que se utiliza para explicar el resultado de cualquier modelo de Machine Learning [37].

Curvas ROC/AUPRC: AUC-ROC evalúa los modelos de clasificación binaria, centrándose en el rendimiento a través de umbrales, especialmente en conjuntos de datos desequilibrados. Utiliza las bibliotecas de Python para calcular los valores AUC y comparar clasificadores en un solo flujo de trabajo [38].

AutoGluon-TS: En [39] se indica que es una biblioteca AutoML de código abierto para la previsión probabilística de series temporales. Centrada en la facilidad de uso y la robustez, permite a los usuarios generar previsiones puntuales y cuantílicas precisas con solo 3 líneas de código Python.

Temporal Fusion Transformer: Este modelo integra características como redes neuronales recurrentes, mecanismos de atención y selección de características, para abordar los desafíos de la previsión de múltiples pasos, en particular en dominios donde los datos de series temporales se ven influenciados por numerosas variables [40].

deepAR: Una metodología para producir previsiones probabilísticas precisas, basada en el entrenamiento de un modelo de red recurrente autorregresivo en una gran cantidad de series temporales relacionadas [41].

SARIMA: Modelo Autorregresivo Integrado de Media Móvil Estacional, es un modelo de pronóstico de series temporales versátil y ampliamente utilizado. Captura tanto las dependencias a corto plazo como a largo plazo dentro de los datos, lo que lo convierte en una herramienta robusta para la predicción [42].

III. DESCRIPCIÓN DEL DESARROLLO

A. Contexto

Se analizarán datos del Instituto Nacional de Salud sobre infecciones respiratorias agudas graves inusitadas (IRAG), correspondientes al periodo comprendido entre los años 2018 y 2023, con una frecuencia de registro anual. La información fue recopilada de la página web del Sistema de Vigilancia en Salud Pública (SIVIGILA), garantizando la trazabilidad y transparencia del proceso de captura de datos. Estos datos serán utilizados para probar distintos modelos de inteligencia artificial, con el objetivo de generar predicciones para los próximos años, seleccionando el modelo más adecuado para este propósito.

B. Limpieza de datos

El proceso de limpieza de datos comenzó con la unificación de los archivos individuales en formato .xls y .xlsx, correspondientes a los años 2010 hasta 2023. Estos datos fueron combinados en un único archivo denominado datos _juntos.csv para facilitar su procesamiento. Posteriormente, se aplicaron técnicas de limpieza, incluyendo la eliminación de registros duplicados, el manejo de valores faltantes mediante eliminación, la conversión de formatos de fecha y la detección de valores atípicos.

1) Datos faltantes

Para abordar los datos faltantes en el archivo datos_juntos.csv, inicialmente se realizó un análisis exploratorio con *data.info()* para obtener una visión general del conjunto de datos, incluyendo la cantidad de valores no nulos por columna. Se identificaron las columnas con mayor cantidad de datos faltantes, entre ellas "nacionalidad", "nombre_nacionalidad" y "estrato". Debido a la alta proporción de valores ausentes en estas variables, se decidió eliminarlas utilizando *data.dropna()*, lo que permitió reducir el conjunto de datos de (215.597 filas y 75 columnas) a (204.619 filas y 75 columnas), asegurando que los datos restantes fueran más consistentes y adecuados para el análisis posterior.

2) Filas repetidas

Se realizó una depuración del conjunto de datos enfocada en la ciudad de Bogotá. Para ello, se filtraron únicamente las filas donde el municipio de ocurrencia fuera "BOGOTÁ", lo que redujo el set de datos a (71.326 filas y 75 columnas). Posteriormente, se llevó a cabo la eliminación de filas duplicadas con *data.drop_duplicates()*, manteniendo el mismo número de registros, lo que indica que no se encontraron datos repetidos dentro del conjunto de datos filtrado para Bogotá.

3) Columnas irrelevantes

Una vez finalizado el proceso de eliminación de filas repetidas, se procedió a la depuración de las columnas irrelevantes para el estudio. Dado que el análisis se centra en algunos grupos poblacionales, casos confirmados, años de registro, sexo y edad, se identificaron variables que no aportaban información relevante. Se eliminaron columnas relacionadas con códigos administrativos, identificadores internos, datos de residencia y otros atributos no esenciales. Para ello, se utilizó *data.drop()*, lo que redujo el número de columnas de 75 a 26, dejando un set de datos de 71.326 filas, 26 columnas.

4) Identificación de valores atípicos

Tras el análisis de las variables numéricas en el conjunto de datos, no se identificaron valores atípicos, lo que indica que la información es consistente y no requiere ajustes en esta etapa.

5) Errores tipográficos en variables categóricas

Para garantizar la uniformidad, se realizó un análisis de los subniveles presentes en cada variable categórica utilizando el método value_counts(). Tras la validación visual de los datos, no se identificaron errores tipográficos, lo que confirma la coherencia en la representación de las categorías.

C. Análisis exploratorio

Tras la limpieza de datos, que incluyó la eliminación de columnas irrelevantes, la detección y eliminación de filas duplicadas, la revisión y tratamiento de valores atípicos, así como la corrección de errores tipográficos. Se realiza el proceso que incluyó la identificación de variables categóricas y numéricas, para así poder generar gráficos de distribución que nos permita analizar mejor los datos. Para las variables categóricas, se utilizaron gráficos de barras para visualizar la distribución de los datos, mientras que para las variables numéricas se aplicaron histogramas para evaluar la dispersión y posibles sesgos en la distribución.

1) Análisis de cada variable de manera individual

Con cada variable del conjunto de datos se realizó un gráfico de barras para visualizar la distribución de los casos según diferentes factores. Para la variable estrato socioeconómico [Fig. 1] muestra que la mayoría de los casos se concentran en los estratos 1 y 2, mientras que los estratos más altos presentan una menor frecuencia. Por otro lado, la distribución de casos por sexo [Fig. 2] muestra una proporción relativamente equilibrada entre hombres y mujeres.

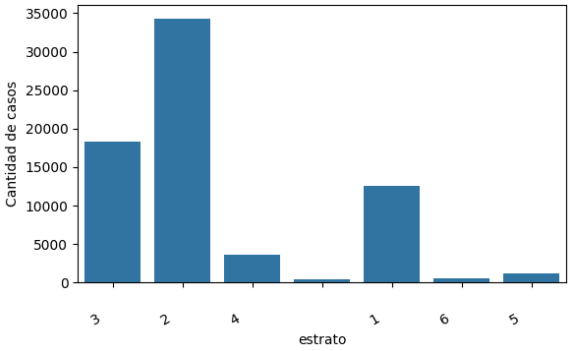

Fig. 1. Gráfico de Barras: Distribución de casos por estrato. Fuente: Los autores.

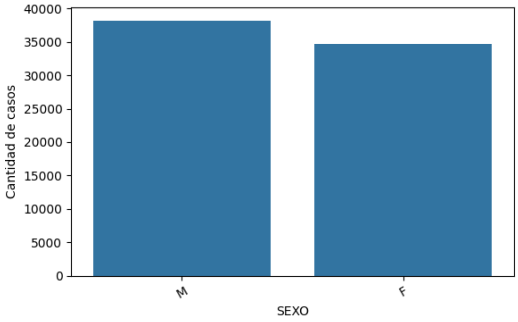

Fig. 2. Gráfico de Barras: Distribución de casos por sexo. Fuente: Los autores.

2) Visualización de variables numéricas

Para comprender mejor la distribución de los datos, se han generado gráficos de barras para todas las variables numéricas. A partir de estos análisis, se han obtenido las siguientes observaciones: la mayor parte de la población afectada se encuentra en el rango de 0 a 10 años, con un segundo grupo significativo entre 60 y 90 años [Fig. 3]. Además, los años con mayor número de casos registrados fueron 2022 y 2023 [Fig. 4], lo que sugiere un posible aumento en la incidencia durante este período.

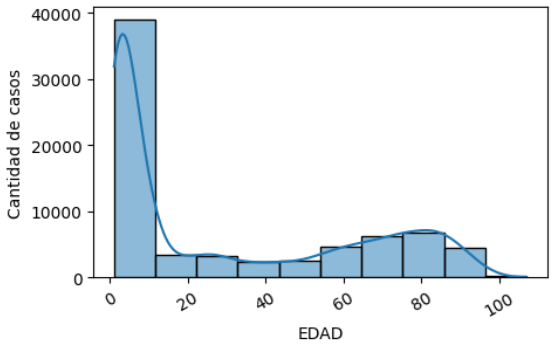

Fig. 3. Histograma de la distribución de edad en casos de IRAG. Fuente: Los autores.

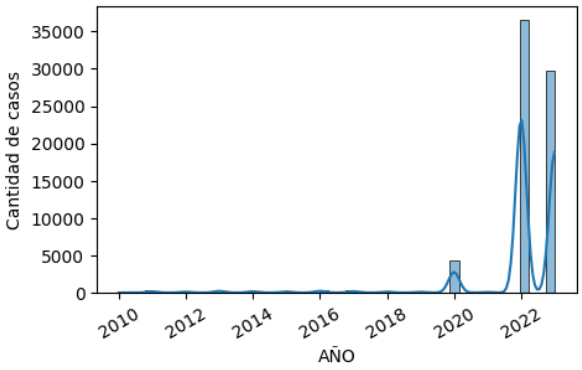

Fig. 4. Histograma de la distribución de años en casos de IRAG. Fuente: Los autores.

3) Análisis de correlación

Para identificar relaciones entre variables, se realizó un análisis de correlación utilizando casos confirmados como variable independiente. Se graficaron diversas variables, como año y la cantidad de casos confirmados mediante un gráfico de dispersión [Fig. 5], se observa un incremento drástico en la cantidad de casos confirmados a partir del año 2020, con un pico en 2022. Además, se analizó la distribución de casos confirmados por edad y año, la [Fig. 6] muestra que los casos se concentran principalmente en las edades de 0 a 10 y de 60 a 90 en los años 2022 y 2023.

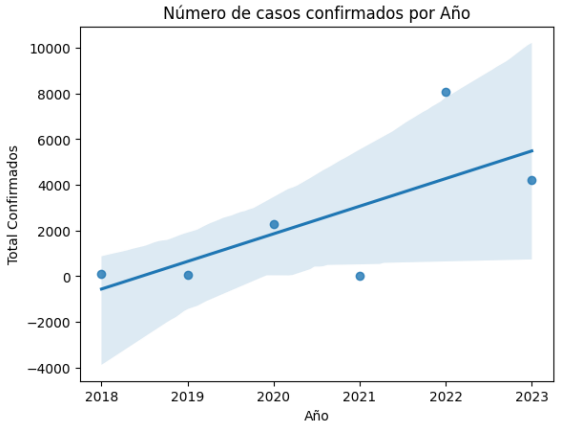

Fig. 5. Gráfica de dispersión sobre la evolución anual de casos confirmados de IRAG. Fuente: Los autores.

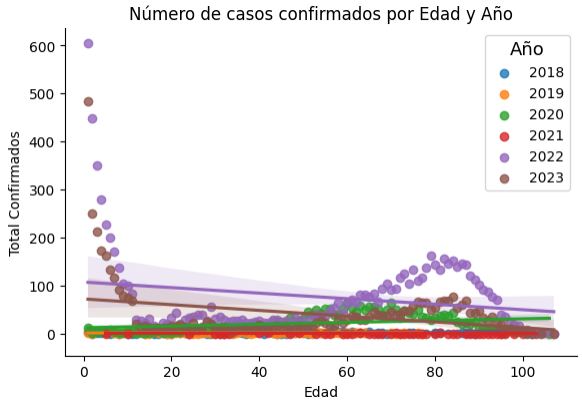

Fig. 6. Gráfica de dispersión sobre la evolución anual de casos de IRAG por edad y confirmación. Fuente: Los autores.

4) Columnas relevantes después del análisis de correlación

Luego de realizar el análisis de correlación, se identificaron las variables más relevantes para el estudio, eliminando aquellas que no aportaban valor significativo al objetivo del análisis. Se seleccionaron únicamente las columnas esenciales: ANO, SEMANA, EDAD, SEXO, estrato, GP_DISCAPA, GP_DESPLAZ, GP_MIGRANT, GP_GESTAN y confirmados), asegurando que el dataset contuviera la información clave. Esta reducción dejó un conjunto de datos de (71.326 filas y 9 columnas), listo para su uso en las siguientes etapas del estudio.

D. Modelos de inteligencia artificial

Para la predicción de enfermedades IRAG en Bogotá en los próximos años, se realizó una investigación del estado del arte [Fig. 7], en la que se analizaron 10 proyectos de investigación afines a la temática propuesta en este trabajo.

	TITULO	FUENTE	MODELOS UTILIZADOS
2	MINERIA DE DATOS PARA EL DESCUBRIMIENTO DE PATRONES EN ENFERMEDADES RESPIRATORIAS EN BOGOTÁ, COLOMBIA	https://repository.ucatolica.edu.co/server/api/core/bitstreams/671a216d-16f5-4ba7-a154-e2c0fb849202/content	Algoritmo K-Means
3	IMPLEMENTAR UN MODELO ESTADÍSTICO PARA EL ESTUDIO DE LAS ENFERMEDADES ENDÉMICAS Y CRÓNICAS ATENDIDAS EN EL HOSPITAL SAN JUAN DE DIOS DEL MUNICIPIO DE PAMPLONA, NORTE DE SANTANDER	http://repositoriodspace.unipamplona.edu.co/jspui/bitstream/20.500.12744/1861/1/Guerrero_2016_TG.pdf	modelo de regresión lineal por el método de mínimos cuadrados
4	FORECASTING SEVERE RESPIRATORY DISEASE HOSPITALIZATIONS USING MACHINE LEARNING ALGORITHMS	s12911-024-02702-0.pdf	"Algoritmos de Pronóstico: ARIMA ETS Theta CrostonSBA NPTS Redes Neuronales: PatchTST DeepAR Temporal Fusion Transformer (TFT)"
5	A MIXED-METHODS STUDY ON THE DESIGN OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE-BASED STRATEGIES TO INFORM PUBLIC HEALTH RESPONSES TO COVID-19 IN DIFFERENT LOCAL HEALTH ECOSYSTEMS: A STUDY PROTOCOL FOR COLEV	f1000research-11-122623.pdf	Procesamiento del lenguaje natural (NLP)

Fig. 7. Cuadro comparativo sobre modelos IA en trabajos de investigación. Fuente: Los autores.

A partir de estos estudios, se identificaron los modelos de inteligencia artificial utilizados [Fig. 8], por medio de un conteo de frecuencia de uso. Como resultado, se seleccionaron cuatro modelos más utilizados para abordar problemas afines al que aquí se propone, estos son: redes neuronales, regresión logística, árboles aleatorios y árboles de decisión.

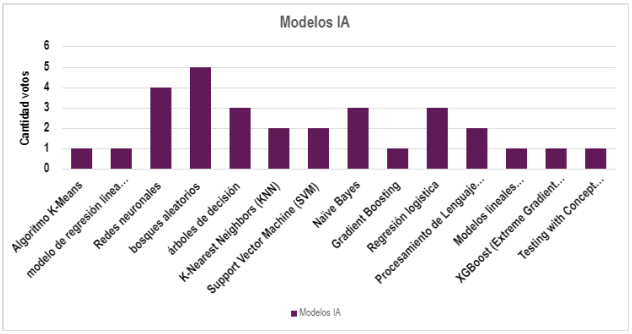

Fig. 8. Gráfica de barras sobre el conteo para la selección de modelos IA. Fuente: Los autores.

Teniendo en cuenta los 4 modelos mencionados anteriormente, se descarta el modelo de regresión logística, ya que es un algoritmo de clasificación y no es adecuado para el enfoque que aborda el presente trabajo.

Posteriormente, junto con el set de datos adecuado, estos modelos han sido implementados y comparados con el objetivo de evaluar su desempeño en términos de precisión, interpretabilidad y capacidad de generalización. Para ellos, se utilizó un conjunto de datos semanales con variables

demográficas y de vulnerabilidad social para modelar la serie temporal de casos confirmados de IRAG. Se seleccionaron las variables EDAD, AÑO, SEXO, GP_DISCAPA, GP_DESPLAZ, GP_MIGRANT, GP_GESTAN, y se agregaron semanalmente junto con funciones seno y coseno para capturar estacionalidad temporal (mes y semana). Las variables numéricas fueron normalizadas mediante MinMaxScaler, la variable objetivo (confirmados_sum) fue transformada con log1p. Se dividieron los datos en entrenamiento (80 %) y prueba (20 %).

1) Árboles de decisión

La construcción de este modelo fue mediante la clase DecisionTreeRegressor de scikit-learn, bajo una estructura jerárquica no paramétrica, permitiendo así segmentar los datos de entrada a través de decisiones binarias que maximizan la reducción de varianza en cada nodo. La configuración empleada incluyó una profundidad máxima de 5 niveles, un mínimo de 10 muestras para dividir un nodo y al menos 3 observaciones por hoja, con el objetivo de limitar la complejidad del árbol y controlar el sobreajuste.

La arquitectura del modelo se basó en una matriz de características que incluye variables epidemiológicas y sociodemográficas semanales, junto con componentes temporales construidos a partir de transformaciones cíclicas (seno y coseno) sobre las semanas y los meses del año, lo que permitió representar estacionalidad de forma continua dentro de una estructura tabular. Antes del entrenamiento, todas las variables numéricas fueron normalizadas con MinMaxScaler para asegurar uniformidad en la escala y evitar sesgos en las particiones del árbol.

El modelo fue validado mediante una estrategia de validación cruzada con particiones temporales (TimeSeriesSplit con 8 divisiones), manteniendo la integridad cronológica de los datos. Esta técnica permite medir el desempeño del árbol sobre distintas ventanas de tiempo, replicando condiciones reales de predicción retrospectiva. La variable objetivo fue transformada previamente mediante logaritmo natural ajustado (log1p) para estabilizar la varianza y mejorar la capacidad de ajuste del modelo sobre datos altamente sesgados.

2) Bosques aleatorios

Se implementó un ensamble de 300 árboles de decisión (RandomForestRegressor) con configuración intencionada para equilibrar sesgo y varianza en un contexto epidemiológico caracterizado por ruido semanal y picos estacionales. Cada árbol se limita a una profundidad máxima de 5 niveles, con al menos 10 muestras requeridas para dividir un nodo y 3 muestras mínimas en cada hoja, lo cual es análogo a la configuración del árbol único previo pero aplicado de forma coordinada a múltiples estimadores para reducir sobreajuste.

El uso de muestreo bootstrap con el 80% de los datos por árbol (max_samples=0.8) introduce diversidad en los subconjuntos de entrenamiento, reforzando la robustez frente a variaciones abruptas en los casos semanales y a posibles valores atípicos vinculados a eventos epidemiológicos puntuales. Al igual que en el árbol de decisión, la estructura de partición binaria segmenta las variables cíclicas (seno y coseno de semana/mes) y sociodemográficas, pero el ensamble promedia interpretaciones locales, mitigando el alto sesgo de un solo árbol y capturando relaciones no lineales entre los determinantes poblacionales y la dinámica de contagio.

3) Redes neuronales

El modelo fue implementado en TensorFlow utilizando una arquitectura de red neuronal recurrente del tipo LSTM (Long Short-Term Memory), que fue seleccionada por su capacidad para modelar secuencias temporales con dependencias de largo plazo y capturar la estacionalidad presente en datos epidemiológicos. La estructura del modelo consistió en una red secuencial compuesta por dos capas LSTM. La primera capa tiene 32 unidades y utiliza return_sequences=True, lo cual permite transmitir la secuencia completa a una segunda capa LSTM con 16 unidades. Ambas capas están seguidas de capas de Dropout con tasas de 0.4 y 0.3, respectivamente, como técnica de regularización para mitigar el sobreajuste durante el entrenamiento.

Después de las capas LSTM se añadió una capa densa con 8 neuronas y función de activación ReLU, seguida por una capa de salida con una sola neurona y activación lineal, utilizada para generar la predicción final de casos confirmados por semana. La compilación del modelo se realizó utilizando el optimizador Adam con una tasa de aprendizaje de 0.0007. Se entrenó el modelo durante un máximo de 150 épocas, con un tamaño de lote de 16 y la incorporación de EarlyStopping con espera de 15 épocas, lo que permitió detener el entrenamiento de forma anticipada al no observar mejoras en la validación.

E. Dashboard interactivo

Para facilitar la exploración y análisis de los datos relacionados con las Infecciones Respiratorias Agudas Graves (IRAG) en Bogotá, se desarrolló un dashboard interactivo utilizando una herramienta BI. Este apartado describe el proceso técnico de construcción, desde la selección de la herramienta hasta la integración de datos históricos y proyectados.

1) Selección de la herramienta de BI

Para la construcción del tablero interactivo se eligió Power BI como herramienta de visualización, debido a la experiencia previa del equipo con esta plataforma, su curva de aprendizaje accesible y su capacidad para generar visualizaciones interactivas de manera eficiente. Power BI ofrece una opción directa para publicación web mediante iframe o etiquetas

HTML simples, su potencia en el análisis visual y su interoperabilidad fueron factores decisivos. El Dashboard interactivo desarrollado como parte de este proyecto fue publicado en línea y puede consultarse libremente [43]. Además, no fue necesaria la integración con servicios externos, ya que el conjunto de datos definitivo se obtuvo tras el proceso de limpieza, transformación y análisis de correlación, tal como se explicó en apartados anteriores.

2) *Extracción*

Esta información se sometió a un proceso de limpieza y unificación automatizada utilizando Google Colab. El resultado fue un único archivo .CSV llamado *datos_listos.csv*, que sirvió como fuente única de datos para el análisis en Power BI. Cabe aclarar que no se contempla actualización automática de los datos, ya que el objetivo es analizar el histórico y proyectar valores para los años 2024 y 2025.

3) *Transformación y modelado de datos*

La [Fig. 9] muestra como en Power BI se han aplicado transformaciones adicionales para facilitar la segmentación y el análisis. Entre ellas se incluyó la conversión del número de mes a su nombre abreviado (ej. "01" → "Ene") y la estandarización del campo SEXO de "M"/"F" a "Masculino"/"Femenino". Se clasificaron los grupos etarios según rangos de edad en categorías como lactante, niñez, adolescencia, adultez y adulto mayor. Los estratos socioeconómicos se agruparon en una nueva columna llamada “categoría estrato”, con tres niveles: Bajo, Medio y Alto.

Para identificar a la población vulnerable, se creó una columna condicional basada en variables como gestación, migración, desplazamiento y discapacidad. Además, se generaron columnas DAX para estimar fechas a partir del año y la semana epidemiológica, lo cual facilitó el análisis temporal.

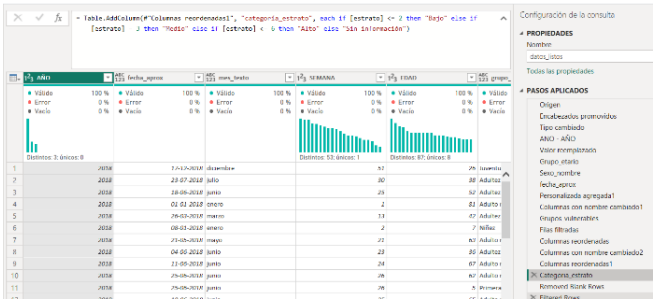

Fig. 9. Proceso de transformación de datos desde Power BI. Fuente: Los autores.

Gracias a la limpieza previa del archivo .csv, no fue necesario aplicar técnicas de imputación ni gestionar valores nulos dentro de Power BI, lo que mejoró el rendimiento general del tablero.

4) *Carga y modelado en el entorno BI*

Lo archivos con los nombres de: *datos_listos.csv* y *datos_predicciones_completo.csv* fueron cargado directamente a Power BI, sirviendo como base para la construcción de las visualizaciones tanto del análisis histórico (2018 – 2023) como de las predicciones (2024 – 2025). Estas fueron las únicas fuentes de datos permitidas para mantener la consistencia entre ambas vistas del tablero.

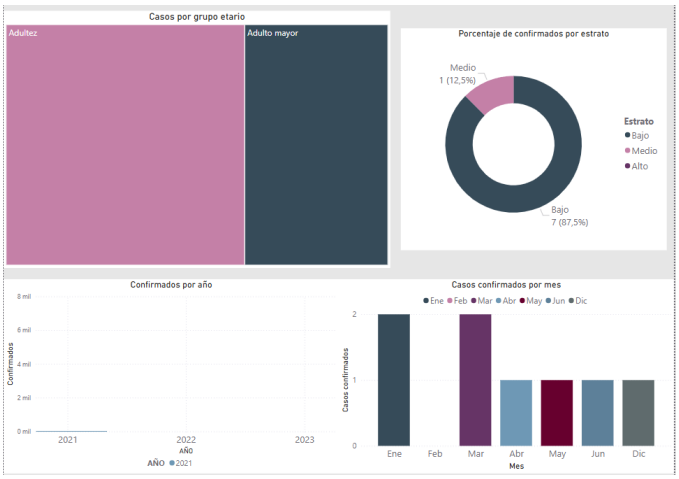

Fig. 10. Elaboración de gráficos Treemap, Anillos y Barras para el Dashboard interactivo IRAG Bogotá. Fuente: Los autores.

Se desarrollaron diversas visualizaciones que permitieron explorar la información desde múltiples dimensiones. En la [Fig. 10] se destacan: un *Treemap* para mostrar los casos confirmados por grupo etario; un gráfico de *Anillos* que representa el porcentaje de casos confirmados según la categoría de estrato socioeconómico; gráficos de *Barras* apiladas para los casos confirmados por año y por mes; y un gráfico de *Cintas* que ilustra la evolución de los casos confirmados por edad y año, permitiendo observar tendencias longitudinales.

Fig. 11. Segmentaciones y tarjetas creadas en el Dashboard interactivo IRAG Bogotá. Fuente: Los autores.

Además, la [Fig. 11] muestra cómo se incorporaron tarjetas dinámicas que resumen indicadores clave como el total de casos confirmados, los casos identificados en población vulnerable, los casos en adultos mayores, así como los casos confirmados en el último año con datos reales (2023) y en el año más reciente estimado por el modelo predictivo (2025).

Todas las visualizaciones están integradas con filtros de segmentación que permiten al usuario interactuar con el tablero y ajustar la visualización de los datos según año, sexo, grupo

etario y condición de vulnerabilidad.

5) Datos de predicciones

Las predicciones para 2024 y 2025 se integraron como una segunda página en el mismo Dashboard, denominada *Predicción*. Esta contiene visualizaciones equivalentes a las del histórico, permitiendo comparar ambos periodos.

El lienzo de diseño tiene un tamaño de 1720 px de alto por 1280 px de ancho, optimizado para visores de escritorio. Se mantuvo una estructura consistente entre ambas páginas del Dashboard. Se utilizó el tema de color *Orquídea Accesible* de Power BI, el cual ofrece buena visibilidad y contraste.

IV. RESULTADOS

Luego de comparar diferentes modelos para predecir los casos confirmados de IRAG, como muestra la [TABLA I], el modelo de árboles de decisión presentó el mejor rendimiento global con un (R^2) de 98.83% en prueba y el menor error (MSE de 0.000226), el modelo de redes neuronales (LSTM) destacó por su capacidad de aprender patrones y mantener una buena estabilidad entre fases de entrenamiento y prueba, con un R2 de 92.6% en entrenamiento y 84.6% en prueba.

TABLA I
Evaluación Comparativa del R^2 en Modelos de IA.

Modelo	R2 Train	R2 Train %	R2 Test	R2 Test %
Bosques aleatorios	0.98099	98.0876	0.9854	98.5396
Arboles de decisiones	0.99859	99.8587	0.98839	98.8394
Redes neuronales	0.926	92.598	0.84633	84.6336

TABLA II
Evaluación Comparativa del MSE en Modelos de IA.

Modelo	MSE Train	MSE Train %	MSE Test	MSE Test %
Bosques aleatorios	0.00066	0.06596	0.000284	0.02844
Arboles de decisiones	4.87E'05	0.00487	0.000226	0.02260
Redes neuronales	0.00255	0.25524	0.002993	0.29929

Modelos tradicionales como los árboles de decisión y los bosques aleatorios obtuvieron altos niveles de precisión en el entrenamiento, con coeficientes de determinación (R^2) de 98.85% y 98.08%, respectivamente, pero presentaron una caída importante en el rendimiento al predecir datos nuevos, como se evidencia en la [Fig. 12] y [Fig. 13], lo cual refleja una tendencia al sobreajuste. En contraste, el modelo de red neuronal demostró una mejor capacidad de generalización, al mantener valores más estables tanto en entrenamiento como en prueba, con un MSE de 0.002552 y 0.002993, y un (R^2) de 92.6% y 84.6%, respectivamente evidenciados en la [TABLA II].

Fig. 12. Predicción de casos IRAG para los próximos 2 años con árboles de decisión. Fuente: Los autores.

Fig. 13. Predicción de casos IRAG para los próximos 2 años con bosques aleatorios. Fuente: Los autores.

El uso de técnicas como normalización, Dropout y EarlyStopping permitió mejorar la capacidad de generalización del modelo LSTM, reduciendo el sobreajuste y manteniendo un rendimiento estable ante semanas con alta variabilidad. Este comportamiento se observa en la [Fig. 14], donde el modelo proyecta una tendencia coherente con intervalos de confianza ajustados.

Un aspecto clave en la efectividad del modelo de redes neuronales fue la transformación de variables temporales como la semana y el mes en componentes cíclicos mediante funciones seno y coseno. Esta codificación permitió que la red neuronal captara mejor los patrones estacionales de la serie temporal, como los picos recurrentes de casos. Mostraron una mayor capacidad para aprender dinámicas temporales complejas, lo que se reflejó en un mejor equilibrio entre entrenamiento y prueba y en una proyección más coherente.

Fig. 14. Predicción de Casos IRAG para los Próximos 2 Años con Redes Neuronales. Fuente: Los autores.

En cuanto a los resultados realizados a los datos históricos, la [Fig. 15] muestra que en el periodo de los años 2018 a 2023 se registraron cerca de 15.000 casos confirmados de IRAG en Bogotá, con 934 pertenecientes a población vulnerable. El grupo etario más afectado fue el de adultos mayores (6.972 casos), seguido por personas adultas y la primera infancia. Más del 50 % de los casos correspondieron al estrato socioeconómico bajo. En el comportamiento anual, los años 2022 y 2023 concentraron la mayor carga de casos. Estos hallazgos, junto con la identificación de desigualdades por edad y nivel socioeconómico, reflejan patrones epidemiológicos consistentes con determinantes sociales de la salud, útiles para priorizar acciones de intervención.

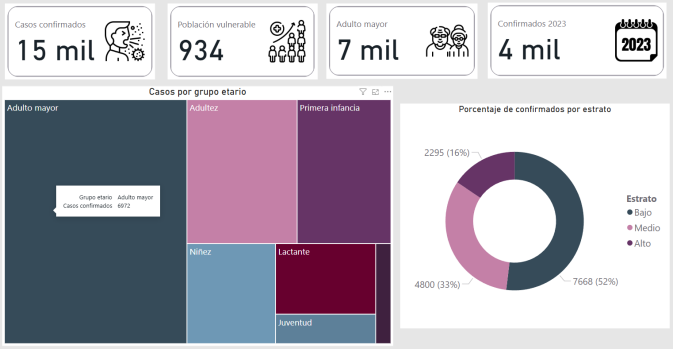

Fig. 15. Visualización de los casos confirmados de IRAG por grupo etario, condición de vulnerabilidad y nivel socioeconómico en el periodo 2018–2023. Fuente: Los autores.

Por otro lado, en cuanto a los resultados generados según lo que el modelo predice, se muestra en la [Fig. 16] en la sección (A), un estimado de aproximadamente 5.000 casos confirmados para los años 2024 y 2025, de los cuales 3.000 corresponderían a población vulnerable.

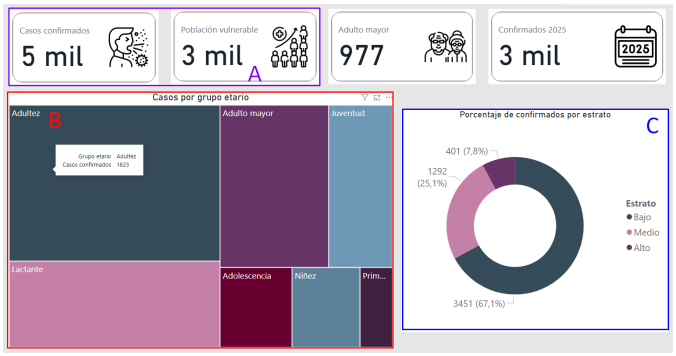

Fig. 16. Visualización de los casos de predicción de IRAG por grupo etario, condición de vulnerabilidad y nivel socioeconómico en el periodo 2024–2025. Fuente: Los autores.

Por grupo etario, la [Fig. 16], la sección (B) evidencia que la adultez será el grupo más afectado con (1.823) casos, seguido por lactantes (1.008), adultos mayores (977), juventud (573), adolescencia (319), niñez (301) y primera infancia (143). El descenso relativo en adultos mayores en comparación con el histórico podría reflejar mejoras en intervención.

En cuanto a los estratos socioeconómicos, la [Fig. 16] sección (C) demuestra que la desigualdad persiste: el estrato bajo agruparía el (67,09 %) de los casos proyectados, seguido por el medio (25,12 %) y alto (7,8 %). Estos resultados refuerzan la urgencia de intervenciones focalizadas en poblaciones de alta vulnerabilidad social.

Al comparar los resultados históricos desde el 2018 al 2023 con las predicciones de 2024 y 2025, se observa una disminución en el aumento proporcional de casos en población vulnerable y en el estrato bajo advierte sobre una persistente desigualdad. Otro aspecto llamativo es el cambio en la distribución por grupo etario. Mientras que en el histórico los adultos mayores dominaban ampliamente el total de casos, en la proyección se aprecia un crecimiento en los casos de personas adultas jóvenes y lactantes.

V. CONCLUSIONES

El presente trabajo logró recopilar y consolidar en un único dataset los registros históricos de casos de IRAG correspondientes al periodo 2018 – 2023, lo que facilitó la limpieza, la depuración de inconsistencias y el análisis exploratorio de datos. Este proceso fue complementado con una revisión exhaustiva de literatura, en la que se consultaron alrededor de quince fuentes científicas relevantes, seleccionadas a partir de palabras clave como morbilidad, mortalidad, inteligencia artificial, red neuronal, IRAG y SIVIGILA. Esta revisión permitió identificar experiencias previas en el uso de aprendizaje automático y visualización aplicada a la predicción de enfermedades respiratorias, que aportaron fundamentos para la metodología al enfoque propuesto en esta investigación.

En la fase de modelado se implementaron tres algoritmos de

aprendizaje supervisado: árboles de decisión, bosques aleatorios y redes neuronales LSTM. Aunque los modelos basados en árboles alcanzaron métricas de precisión superiores en entrenamiento (R^2 superiores al 98 %), presentaron un grado de sobreajuste considerable al compararse con los datos de prueba. Por el contrario, la red neuronal logró un equilibrio más estable entre precisión y capacidad de generalización (R^2 de 92,6 % en entrenamiento y 84,6 % en prueba), criterio que resultó determinante para su selección como modelo final. Este enfoque permitió obtener predicciones con menor riesgo de sesgo por sobreajuste.

Se diseñó un dashboard interactivo en Power BI con dos páginas diferenciadas: una dedicada al análisis histórico de los casos entre el 2018 y 2023, y otra enfocado en las predicciones para el 2024 y 2025. Este tablero incluye visualizaciones como treemaps, gráficos de anillos y de columnas apiladas, junto con tarjetas dinámicas que muestran indicadores clave de distribución por grupo etario, estrato socioeconómico, año, mes y condición de vulnerabilidad. Su carácter interactivo permite filtrar y segmentar la información, lo que facilita la identificación de patrones y tendencias epidemiológicas.

Al contrastar los datos históricos con las predicciones obtenidas, se identificó que durante el periodo 2024 – 2025 persistirá una concentración importante de casos en los estratos socioeconómicos bajos y un incremento proporcional en la población adulta. Esta dinámica sugiere la necesidad de ampliar el enfoque de intervención hacia este segmento, sin desatender la vigilancia en los menores y adultos mayores, que continúan siendo grupos de alta vulnerabilidad clínica. El análisis permite anticipar escenarios de riesgo y planificar acciones que reduzcan desigualdades en la incidencia de IRAG.

Como aprendizaje significativo, se destaca la importancia de asegurar un proceso riguroso de limpieza de datos, una adecuada interpretación de las métricas de desempeño y la elección de herramientas que permitan generar modelos reproducibles. Para investigaciones posteriores, se recomienda realizar una validación comparativa con los datos que el Instituto Nacional de Salud planea publicar durante julio de 2025, ya que, pese al contacto y búsqueda de estos, a la fecha de culminación de este trabajo no se encontraban disponibles en la plataforma oficial. Aunque se realizó la solicitud a través de canales electrónicos institucionales, no fue posible obtener una respuesta oportuna. Esta futura comparación permitirá comprobar qué tan cercanas fueron las proyecciones generadas por el modelo con respecto a los datos reales en este trabajo.

Referencias

[1] S. Aragón, “Perfil Clínico y Epidemiológico de los Pacientes con Diagnóstico de Neumonía Viral Incluidos en la Vigilancia Centinela del Hospital Nacional de Niños Benjamín Bloom, de enero 2014 a diciembre 2015”, Universidad de El Salvador, 2018. Consultado: el 6 de enero de 2025. [En línea]. Disponible en: https://docs.bvsalud.org/biblioref/2021/04/1177388/491-11105864.pdf

[2] E. Rojas y S. Aguilar, “Minería de Datos para el Descubrimiento de Patrones en Enfermedades Respiratorias en Bogotá, Colombia”, Universidad Católica de Colombia, Bogotá D.C, 2017. Consultado: el 6 de enero de 2025. [En línea]. Disponible en: https://repository.ucatolica.edu.co/server/api/core/bitstreams/671a216d-16f5-4ba7-a154-e2c0fb849202/content

[3] Carlos Guerrero, “Implementar un Modelo Estadístico para el Estudio de las Enfermedades Endémicas y Crónicas Atendidas en el Hospital San Juan de Dios del Municipio de Pamplona, Norte de Santander”, Universidad de Pamplona, 2016. Consultado: el 6 de enero de 2025. [En línea]. Disponible en: http://repositoriodspace.unipamplona.edu.co/jspui/bitstream/20.500.12744/1861/1/Guerrero_2016_TG.pdf

[4] B. Del Rosario, M. Valdés, J. Díaz, L. Duany, L. Santeiro, y S. Del Villar, “Caracterización del Comportamiento de las Infecciones Respiratorias Agudas. Provincia Cienfuegos. Primer Trimestre 2020”. Consultado: el 6 de enero de 2025. [En línea]. Disponible en: https://www.redalyc.org/journal/1800/180065014011/html/#B1

[5] D. Carolina, M. Sánchez, S. Milena, y A. Fuentes, “Informe de Evento Infección Respiratoria Aguda, Colombia, 2020”, pp. 1–28, 2020.

[6] A. En, M. De, A. E. El, y M. De Valledupar, “Estudio sobre la Morbimortalidad por Infección Respiratoria Carmenza González Fermín Leguizamo Marisela López Montilla Yesenia Patricia Ponce Hilvar Yamid Malaver Universidad Nacional Abierta y a Distancia Escuela de Ciencias Básicas, Tecnología e Ingeniería”, pp. 1–31, 2010.

[7] A. de Ruvo et al., “Spread: Spatiotemporal Pathogen Relationships and Epidemiological Analysis Dashboard”, Vet Ital, vol. 60, núm. 4 Special Issue, pp. 1–13, mar. 2024, doi: 10.12834/VetIt.3476.23846.1.

[8] K. Footer et al., “Using Publicly Available, Interactive Epidemiological Dashboards: An Innovative Approach to Sharing Data from the Rakai Community Cohort Study”, JAMIA Open, vol. 7, núm. 3, oct. 2024, doi: 10.1093/jamiaopen/ooae069.

[9] C. P. Yadav y A. Sharma, “National Institute of Malaria Research-Malaria Dashboard (NIMR-MDB): a Digital Platform for Analysis and Visualization of Epidemiological Data”, The Lancet Regional Health - Southeast Asia, vol. 5, p. 100030, 2022, doi: 10.1016/j.

[10] J. Leiner et al., “Machine Learning-Derived Prediction of In-Hospital Mortality in Patients with Severe Acute Respiratory Infection: Analysis of Claims Data from the German-Wide Helios Hospital Network”, Respir Res, vol. 23, núm. 1, dic. 2022, doi: 10.1186/s12931-022-02180-w.

[11] S. Albrecht et al., “Forecasting Severe Respiratory Disease Hospitalizations Using Machine Learning Algorithms”, BMC Med Inform Decis Mak, vol. 24, núm. 1, dic. 2024, doi: 10.1186/s12911-024-02702-0.

[12] A. Duarte Da Silva, M. Ferreira Da Costa Gomes, T. S. Gregianini, L. G. Martins, A. B. Gorini Da Veiga, y C. A. D. Silva, “Machine Learning in Predicting Severe Acute Respiratory Infection Outbreaks”, Cad. Saúde Pública, vol. 40, núm. 1, p. 122823, 2024, doi: 10.1590/0102-311XEN122823.

[13] C. González-Uribe et al., “A Mixed-Methods Study on the Design of Artificial Intelligence and Data Science-Based Strategies to Inform Public Health Responses to COVID-19 in Different Local Health Ecosystems: A Study Protocol for COLEV [Version 1; Peer Review: 1 Approved, 2 Approved with Reservations]”, 2022, doi: 10.12688/f1000research.110958.1.

[14] A. En, M. De, A. E. El, y M. De Valledupar, “Estudio Sobre La Morbimortalidad Por Infección Respiratoria Carmenza González Fermín Leguízamo Marisela López Montilla Yesenia Patricia Ponce Hilvar Yamid Malaver Universidad Nacional Abierta Y A Distancia Escuela De Ciencias Básicas, Tecnología E Ingeniería”.

[15] Instituto Nacional del Cáncer, “Mortalidad”, Instituto Nacional del Cáncer. Consultado: el 26 de enero de 2025. [En línea]. Disponible en: https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/mortalidad

[16] Google Cloud, “¿Qué es la Inteligencia Artificial o IA?” Consultado: el 26 de enero de 2025. [En línea]. Disponible en: https://cloud.google.com/learn/what-is-artificial-intelligence?hl=es-419

[17] IBM, “¿Qué es un Modelo de IA?”, IBM. Consultado: el 26 de enero de 2025. [En línea]. Disponible en: https://www.ibm.com/es-es/topics/ai-model

[18] IBM, “¿Qué es la Agrupación en Clústeres K-Means?”, K-Means Clustering. Consultado: el 12 de marzo de 2025. [En línea]. Disponible en: https://www.ibm.com/mx-es/topics/k-means-clustering

[19] IBM, “¿Qué es el Bosque Aleatorio?”, Artificial Intelligence. Consultado: el 12 de marzo de 2025. [En línea]. Disponible en: https://www.ibm.com/mx-es/think/topics/random-forest

[20] Salesforce, “Dashboard: ¿Cómo Crear Uno para tu Estrategia?”, Salesforce Latam. Consultado: el 26 de enero de 2025. [En línea]. Disponible en: https://www.salesforce.com/mx/blog/dashboard/

[21] Google Cloud, “¿Qué es el Procesamiento del Lenguaje Natural?”, PLN. Consultado: el 27 de enero de 2025. [En línea]. Disponible en: https://cloud.google.com/learn/what-is-natural-language-processing?hl=es

[22] IBM, “¿Qué es una Red Neuronal?”, neural networks. Consultado: el 27 de enero de 2025. [En línea]. Disponible en: https://www.ibm.com/mx-es/topics/neural-networks

[23] esri, “Cómo Funciona el Algoritmo XGBoost”. Consultado: el 27 de enero de 2025. [En línea]. Disponible en: https://pro.arcgis.com/es/pro-app/latest/tool-reference/geoai/how-xgboost-works.htm#:~:text=XGBoost%20es%20la%20abreviatura%20de,aleatorio%20y%20refuerzo%20de%20gradientes.

[24] AWS, “¿Qué es una Interfaz de Programación de Aplicaciones (API)?” Consultado: el 27 de enero de 2025. [En línea]. Disponible en: https://aws.amazon.com/es/what-is/api/#:~:text=API%20significa%20%E2%80%9Cinterfaz%20de%20programaci%C3%B3n,de%20servicio%20entre%20dos%20aplicaciones.

[25] Función Pública, “Organigrama Sector de Estadística”, jul. 2021.

[26] E. Inmunoprevenibles et al., “Documento Elaborado por”, 2017.

[27] LIS Data Solutions, “¿Qué es RapidMiner?” Consultado: el 27 de enero de 2025. [En línea]. Disponible en: https://www.lisdatasolutions.com/es/que-es-rapidminer/

[28] Instituto Nacional de Salud, “SIVIGILA”, INS. Consultado: el 27 de enero de 2025. [En línea]. Disponible en: https://www.ins.gov.co/Direcciones/Vigilancia/Paginas/SIVIGILA.aspx

[29] Sistema Integrado de Información de la Proyección Social, “¿Qué es SISPRO?”, SISPRO. Consultado: el 27 de enero de 2025. [En línea]. Disponible en: https://www.sispro.gov.co/Pages/Home.aspx

[30] “Software IBM SPSS”. Consultado: el 26 de enero de 2025. [En línea]. Disponible en: https://www.ibm.com/es-es/spss

[31] E. M. Cáceres, F. P. Moya, D. A. Á. Concepción, y L. M. Cáceres, “Sistema de Gestión para la Información de los Canales Endémicos”, Revista Cubana de Tecnología de la Salud, vol. 12, núm. 2, pp. 40–49, may 2021, Consultado: el 26 de enero de 2025. [En línea]. Disponible en: https://revtecnologia.sld.cu/index.php/tec/article/view/1921

[32] “Definición de SARS-CoV-2 - Diccionario de Cáncer del NCI - NCI”. Consultado: el 26 de enero de 2025. [En línea]. Disponible en: https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/sars-cov-2

[33] M. Silva et al., “chewBBACA: A complete suite for gene-by-gene schema creation and strain identification”, Microb Genom, vol. 4, núm. 3, mar. 2018, doi: 10.1099/MGEN.0.000166.

[34] “¿Qué es SQL? - Explicación de Lenguaje de Consulta Estructurado (SQL) - AWS”. Consultado: el 26 de enero de 2025. [En línea]. Disponible en: https://aws.amazon.com/es/what-is/sql/

[35] “¿Qué es Tableau?” Consultado: el 26 de enero de 2025. [En línea]. Disponible en: https://www.tableau.com/es-es/why-tableau/what-is-tableau

[36] “Glosario | HIPAA”. Consultado: el 26 de enero de 2025. [En línea]. Disponible en:

https://start.docuware.com/es/glosario-de-terminos/hipaa

[37] "SHAP, una Librería de Python para la Interpretabilidad de Modelos de Machine Learning | by Skillsbox | Medium". Consultado: el 26 de enero de 2025. [En línea]. Disponible en: https://medium.com/@Skillsbox/shap-una-liberia-de-python-para-la-interpretabilidad-de-modelos-de-machine-learning-d111c4bed8a8

[38] "AUC y Curva ROC en Aprendizaje Automático | DataCamp". Consultado: el 26 de enero de 2025. [En línea]. Disponible en: https://www.datacamp.com/es/tutorial/auc

[39] O. Shchur et al., "AutoGluon–TimeSeries: AutoML for Probabilistic Time Series Forecasting", 2023. Consultado: el 27 de enero de 2025. [En línea]. Disponible en: https://www.amazon.science/publications/autogluon-timeseries-automl-for-probabilistic-time-series-forecasting

[40] "Una Exploración en Profundidad de los Transformadores de Fusión Temporal para la Predicción de Series Temporales | por Mirza Samad | AI Simplified in Plain English | Medium". Consultado: el 27 de enero de 2025. [En línea]. Disponible en: https://medium.com/ai-simplified-in-plain-english/an-in-depth-exploration-of-temporal-fusion-transformers-for-time-series-forecasting-91e74040a079

[41] D. Salinas, V. Flunkert, J. Gasthaus, y T. Januschowski, "DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks", Int J Forecast, vol. 36, núm. 3, pp. 1181–1191, abr. 2017, doi: 10.1016/j.ijforecast.2019.07.001.

[42] "¿Qué es el Modelo SARIMA? | Técnicas de Trading". Consultado: el 27 de enero de 2025. [En línea]. Disponible en: https://www.tecnicasdetrading.com/2024/07/modelo-sarima.html

[43] A. Baquero y C. Sabogal, "IA IRAG EN BOGOTÁ". Consultado: el 1 de junio de 2025. [En línea]. Disponible en: https://dashboard-irag.github.io/IA-IRAG/