INFORME FINAL DEL PROYECTO DE INVESTIGACIÓN: PROPUESTA DE UN MODELO DE MACHINE LEARNING PARA ANALIZAR LA COBERTURA DE LOS SERVICIOS DE SALUD PÚBLICA EN BOGOTÁ D.C.

ESTUDIANTES:

JOSE DAVID MARIÑO

ESPECIALIZACIÓN EN MACHINE LEARNING – UNIVERSIDAD EAN

GERMÁN ALONSO RODRÍGUEZ DÍAZ

ESPECIALIZACIÓN EN MACHINE LEARNING – UNIVERSIDAD EAN

SANTIAGO SIMMONDS RODRIGUEZ

ESPECIALIZACIÓN EN MACHINE LEARNING – UNIVERSIDAD EAN

TUTOR:

ELIZABETH LEON VELASQUEZ

2025

Resumen

Este proyecto de investigación propone desarrollar un modelo de Machine Learning para analizar y segmentar la población de Bogotá según el acceso a servicios de salud pública, con el fin de identificar inequidades y optimizar la distribución de recursos sanitarios. Actualmente persisten marcadas disparidades en el acceso efectivo a los servicios, afectando principalmente a las zonas periféricas y poblaciones vulnerables. Los resultados esperados contribuirán a la formulación de políticas públicas más equitativas y la planificación estratégica de recursos en el sistema de salud de Bogotá.

Palabras clave: Igualdad de oportunidades, Desigualdad Social, Condiciones Sociales, Desarrollo Regional, Servicios de salud.

Problema de investigación

La cobertura de acceso a los servicios de salud es un desafío a nivel global y un objetivo fundamental de los sistemas de salud pública. En Bogotá, la atención médica se ve afectada por múltiples factores, como la ubicación geográfica, el nivel socioeconómico, la disponibilidad de infraestructura hospitalaria y la capacidad del sistema para atender la demanda de la población. Según el Departamento Administrativo Nacional de Estadística (DANE), existen brechas significativas en la cobertura y calidad del servicio en diferentes localidades de la ciudad.

Objetivo general

Diseñar un modelo de Machine Learning que permita segmentar la población de Bogotá según el acceso a servicios de salud, considerando variables como ubicación geográfica, nivel socioeconómico, disponibilidad de infraestructura médica, entre otras.

Objetivos específicos

1. Realizar una revisión de estudios y antecedentes sobre segmentación de población en salud en Bogotá y el uso de Machine Learning, con el fin de establecer un marco teórico y conceptual que sustente la propuesta del modelo.

2. Analizar la situación actual de la cobertura de los servicios de salud en Bogotá, identificando fuentes de datos relevantes y factores clave (como ubicación geográfica, nivel socioeconómico y

disponibilidad de infraestructura médica) que permitan alimentar el modelo de Machine Learning.

3. Proponer el modelo de Machine Learning que permita segmentar la población de Bogotá según la cobertura de los servicios de salud, definiendo las técnicas, herramientas y procesos necesarios para su desarrollo.
4. Validar el modelo propuesto, evaluando su aplicabilidad en la toma de decisiones y la formulación de políticas públicas en Bogotá, con el fin de mejorar la planificación y asignación de recursos en el sistema de salud.

Descripción del problema

Aunque Bogotá cuenta con una amplia red de servicios de salud, se observan inequidades en el acceso efectivo a la atención médica entre diferentes zonas de la ciudad y grupos socioeconómicos (Pinzón-Flórez et al., 2021). Un estudio de la Universidad Nacional (Martínez et al., 2023) demostró que factores como la ubicación geográfica, el nivel socioeconómico, y la afiliación al sistema de salud pueden influir en la oportunidad y calidad de la atención recibida. La Personería de Bogotá (2023) reportó que la saturación de servicios en algunas áreas y la concentración de especialidades en determinadas zonas crean barreras de acceso para el 35% de la población en zonas periféricas, siento un porcentaje muy elevado que representa una inequidad en la población de la ciudad de Bogotá. En este contexto, se hace necesario proponer una metodología basada en Machine Learning que permita analizar y segmentar la población según su acceso a servicios de salud, con el fin de identificar áreas críticas y proponer estrategias de mejora.

Pregunta de investigación

¿Cómo puede un modelo de Machine Learning segmentar la población de Bogotá según la cobertura y calidad de los servicios de salud?

Justificación

El desarrollo de esta investigación presenta una relevancia significativa desde múltiples dimensiones para el contexto de salud pública en Bogotá. En primer lugar, existe una necesidad apremiante de contar con propuestas metodológicas innovadoras que permitan analizar las dinámicas de acceso y calidad en los servicios de salud. La propuesta de un modelo basado en Machine Learning proporcionará un marco conceptual para futuros estudios que busquen optimizar la distribución de recursos sanitarios en la ciudad.

En términos de gestión sanitaria, el estudio de Molina et al. (2021) señala que la comprensión detallada de los patrones de acceso y las barreras existentes permite optimizar la red de servicios y mejorar la eficiencia en la asignación de recursos. La propuesta metodológica de esta investigación contribuirá a sentar las bases para identificar áreas críticas que requieren intervención prioritaria.

Desde el aspecto social, Rivera-López (2022) destaca que las inequidades en el acceso a servicios de salud tienen un impacto directo en los resultados de salud de la población y en la calidad de vida de los ciudadanos. Esta investigación permitirá visibilizar las disparidades existentes y proponer soluciones fundamentadas en evidencia.

En el ámbito económico, el análisis de Gutiérrez y Sánchez (2023) indica que la inadecuada distribución de servicios de salud genera costos adicionales tanto para el sistema de salud como para los usuarios. La propuesta de esta investigación podría contribuir a la optimización de recursos y la reducción de costos asociados a las barreras de acceso.

Adicionalmente, Pérez-Castro et al. (2022) señalan que la pandemia de COVID-19 ha evidenciado y exacerbado las inequidades existentes en el acceso a servicios de salud, haciendo aún más relevante y urgente el análisis propuesto en esta investigación.

Marco Teórico

El marco teórico de esta investigación se construye mediante un recorrido cronológico donde se planteó definir los conceptos y dimensiones referente a la equidad en la salud, resaltando la evolución del sistema de salud en Colombia y Bogotá. Se describe la importancia social en la que está situada actualmente la ciudad de Bogotá en el sector Salud, y a través del mismo se plantea la necesidad de nuevas herramientas analíticas que permitan simplificar la comprensión de la información, asistido de una investigación de modelos de Machine Learning aplicados al contexto en salud, la identificación de inequidad en acceso a la salud mediante modelos predictivos y la aplicación de estos modelos en el sector de la salud pública, para finalizar se definieron los desafíos éticos y metodológicos.

Equidad en Salud: Conceptos y Dimensiones

La equidad en salud constituye un principio fundamental en la estructuración de sistemas sanitarios efectivos. Según Whitehead y Dahlgren (2006), la equidad en salud implica que "idealmente, todos deberían tener una oportunidad justa para lograr su pleno potencial de salud", lo que requiere eliminar diferencias evitables entre grupos poblacionales. Esta noción se alinea con lo expuesto por Penchansky y Thomas (1981) en su Teoría del Acceso a la Salud, la cual postula que el acceso efectivo depende de cinco dimensiones interdependientes:

Accesibilidad: Disponibilidad física de servicios (ej.: proximidad geográfica).

Asequibilidad: Capacidad de cubrir costos económicos sin generar empobrecimiento.

Aceptabilidad: Adecuación cultural y percepción de calidad por parte de los usuarios.

Acomodación: Adaptación de los servicios a las necesidades y horarios de la población.

Disponibilidad: Existencia suficiente de recursos (humanos, infraestructura, medicamentos).

Estas dimensiones, como señalan Galvis-Aponte y Rico (2023), interactúan con determinantes sociales como educación, ingreso, género y territorio, generando patrones complejos de inequidad. Por ejemplo, comunidades rurales pueden enfrentar barreras de accesibilidad (distancia a centros médicos) y asequibilidad (limitaciones económicas), mientras que grupos indígenas pueden experimentar problemas de aceptabilidad (desconfianza en servicios culturalmente inapropiados).

Esta perspectiva multidimensional es reforzada por Arcaya et al. (2015), quienes distinguen entre desigualdades inevitables (ej.: variaciones biológicas) y desigualdades injustas y prevenibles (ej.: falta de acceso a medicamentos por discriminación socioeconómica).

La equidad, por tanto, exige políticas que aborden no solo la distribución de servicios, sino también los determinantes estructurales que limitan el acceso según el marco de Penchansky y Thomas.

Evolución del Sistema de Salud en Colombia y Bogotá

La transformación del sistema de salud en Bogotá, iniciada con la Ley 100 de 1993, marcó un punto de inflexión en la organización y prestación de servicios de salud en la ciudad (Guerrero et al., 2011). Esta ley estableció un modelo de aseguramiento basado en dos regímenes principales: el contributivo y el subsidiado, buscando garantizar el acceso universal a los servicios de salud mediante un esquema de solidaridad en el financiamiento (Arbeláez et al., 2020).

Durante la primera década de implementación (1993-2003), Bogotá experimentó un crecimiento significativo en la cobertura del aseguramiento, pasando de un 45% a un 75% de la población (Ramírez & González, 2018). Este período estuvo caracterizado por la expansión de la red de prestadores privados y la transformación de los hospitales públicos en Empresas Sociales del Estado (ESE), lo que modificó sustancialmente la dinámica de prestación de servicios en la ciudad.

Entre 2004 y 2012, la Secretaría Distrital de Salud implementó el programa "Salud a su Hogar", posteriormente denominado "Salud a su Casa", que buscaba fortalecer la atención primaria en salud y mejorar el acceso a servicios básicos en las zonas más vulnerables de la ciudad (Vega et al., 2017).

Este programa contribuyó significativamente a la identificación de barreras de acceso y al desarrollo de estrategias territoriales para superarlas.

Sin embargo, como señalan Eslava-Schmalbach et al. (2017), la alta cobertura no ha eliminado las inequidades en el acceso efectivo a servicios de calidad. Su Índice Compuesto de Inequidad en Salud (IIS) muestra que las grandes ciudades como Bogotá presentan valores altos de inequidad, particularmente por el efecto de factores "intolerables" en salud, como agresiones, enfermedades prevenibles y atención inadecuada.

Para 2018, aunque la cobertura alcanzó el 95% en Bogotá (Gómez et al., 2019), estudios posteriores revelaron importantes brechas en la calidad y oportunidad de la atención. Molina-Benavides et al. (2020) identificaron que el tiempo promedio de espera para citas con especialistas podía superar los 30 días en algunas especialidades críticas, y que existían marcadas diferencias en la disponibilidad de servicios entre las diferentes localidades de la ciudad.

Un análisis realizado por el Observatorio de Salud de Bogotá (2021) evidenció que, a pesar de la alta cobertura, persistían inequidades significativas en el acceso efectivo a servicios de salud:

Las localidades del sur de la ciudad presentaban una menor densidad de servicios especializados por habitante.

La concentración de tecnología médica avanzada se mantenía principalmente en las zonas norte y centro de la ciudad.

Las barreras administrativas afectaban desproporcionadamente a la población del régimen subsidiado.

La pandemia de COVID-19 expuso y exacerbó estas disparidades preexistentes. Según Castellanos et al. (2021), las localidades con menor disponibilidad de servicios experimentaron mayores dificultades para responder a la crisis sanitaria, evidenciando la necesidad de fortalecer la red de atención en estas áreas.

Determinantes Sociales de la Salud y su Impacto en Bogotá

La Organización Mundial de la Salud define los determinantes sociales como "las circunstancias en que las personas nacen crecen, viven, trabajan y envejecen" (OMS, 2008). En Bogotá, estos determinantes configuran un mapa de inequidades territoriales que afectan el acceso efectivo a servicios de salud. Según Galvis-Aponte y Rico (2023), existen marcadas diferencias en indicadores de salud entre departamentos colombianos, con patrones definidos por nivel socioeconómico, ubicación geográfica y desarrollo institucional. Estos patrones se replican a escala intraurbana en Bogotá, donde la estratificación socioeconómica se correlaciona con diferencias en mortalidad evitable, acceso a especialistas y calidad percibida de servicios (Martínez et al., 2023).

El trabajo de Rincón et al. (2017) sobre el Índice Compuesto de Inequidad en Salud destaca que la inequidad se manifiesta de manera diferencial por sexo y territorio, presentando valores más altos para mujeres en la mayoría de los municipios colombianos. Esta perspectiva de género y territorial resulta fundamental para comprender las dinámicas de inequidad en Bogotá.

La necesidad de nuevas herramientas analíticas

A pesar de los esfuerzos en cobertura y atención primaria, los retos de equidad, calidad y acceso oportuno al sistema de salud han generado la necesidad de implementar herramientas más avanzadas que permitan mejorar el análisis y la planificación de los servicios de salud (Pérez & Gómez, 2019; Ramírez & Torres, 2021).

En este contexto, los avances a través de la Ciencia de Datos e Inteligencia Artificial han demostrado su alta capacidad para Extraer, Transformar, y Analizar grandes volúmenes de datos en salud pública, identificar patrones y predecir tendencias con mayor precisión que los enfoques tradicionales (Rajkomar et al., 2019). Estas metodologías permiten:

- Modelar la demanda de los servicios.
- Optimizar la distribución de recursos médicos.
- Identificar inequidades en el acceso.

Los avances más recientes de los últimos años han demostrado que, a partir del análisis de datos, y el uso de técnicas de Machine Learning, se ha logrado una comprensión más profunda de la información recopilada en los diferentes sectores Económicos, tecnológicos, Educativos, incluyendo la Salud Publica, permitiendo la identificación de patrones complejos y la predicción de tendencias que no son evidentes mediante métodos tradicionales (Rajkomar et al., 2019). Algunas de las técnicas utilizadas para este fin han sido:

Modelar la demanda de los servicios: En un estudio realizado por Zhao et al. (2022) para predecir la demanda de atenciones en urgencias en el Hospital de Singapur obtuvo el mejor desempeño predictivo en comparación con algoritmos usados en serie de tiempos, todo esto mediante, el estudio del modelo de aprendizaje profundo basada en una estructura apilada de capas neurales.

Organización de los recursos: El campo de la Inteligencia Artificial tiene el potencial de mejorar la toma de decisiones clínicas en situaciones urgentes, optimizando y organizando el uso de recursos de los servicios de urgencias (Marco et al., 2021; Piliuk & Tomforde, 2023; K. J. W. Tang et al., 2021).

Identificar inequidades en el acceso: De acuerdo con los resultados obtenidos por Pérez, R., & Gómez, L. (2019). En su estudio: "Inequidad espacial en acceso a salud: el caso de la Zona Metropolitana del Valle de México" demostró que a través de la aplicación de un modelo de agrupamiento como K-Means, le permitió agrupar áreas geográficas con distintos niveles de acceso a servicios de salud en la Zona Metropolitana del Valle de México. La segmentación identificó aquellas regiones con menor cobertura sanitaria y barreras estructurales que afectan a la población más vulnerable, del mismo modo Ramírez, M., & Torres, J. (2021), en su estudio para segmentar grupos de pacientes según variables como tiempo de espera, distancia a hospitales y nivel socioeconómico. Se evidenció que ciertos sectores tenían mayores dificultades en la atención médica, lo que permitió diseñar propuestas de políticas públicas para mejorar la equidad en el acceso.

López-Cevallos et al. (2022) implementaron técnicas de Machine Learning para analizar patrones de acceso y utilización de servicios, identificando clusters de necesidades insatisfechas y proponiendo modelos predictivos para la planificación de servicios. Este enfoque innovador ha permitido:

- Identificar patrones geoespaciales de demanda de servicios
- Predecir necesidades futuras de infraestructura sanitaria
- Optimizar la distribución de recursos según características poblacionales
- Desarrollar modelos de atención más eficientes y equitativos

El desarrollo de estas nuevas metodologías de análisis, combinado con la experiencia acumulada en la implementación de políticas de salud, ha generado una base de conocimiento crucial para abordar los desafíos pendientes en el sistema de salud de Bogotá.

Aplicaciones de Machine Learning en Salud Pública

Las técnicas de machine learning ofrecen oportunidades sin precedentes para analizar patrones complejos en datos de salud pública y optimizar la asignación de recursos.

Rajkomar et al. (2019) identifican tres áreas principales de aplicación: diagnóstico y predicción clínica, optimización de procesos administrativos, y **análisis poblacional** para planificación sanitaria.
En el contexto de análisis de cobertura y acceso a servicios, destacan los siguientes enfoques:

1. **Clustering y segmentación poblacional:** Permite identificar grupos con patrones similares de necesidades y barreras de acceso. López-Cevallos et al. (2022) aplicaron estas técnicas para identificar clusters de necesidades insatisfechas en sistemas sanitarios.
2. **Análisis geoespacial predictivo:** Integra datos georreferenciados con características poblacionales para predecir necesidades futuras de infraestructura sanitaria.
3. **Modelos de series temporales:** Zhao et al. (2022) utilizaron redes neuronales para predecir la demanda de atenciones en urgencias, superando los métodos tradicionales.

4. **Sistemas de recomendación para asignación de recursos:** Optimizan la distribución de personal médico e infraestructura según patrones de demanda y necesidades específicas.

La integración de estas técnicas con bases de datos administrativas, encuestas poblacionales y sistemas de información geográfica ofrece un marco metodológico robusto para abordar las inequidades en salud a nivel distrital.

Desafíos Éticos y Metodológicos

La implementación de modelos de machine learning en el análisis de inequidades en salud presenta desafíos importantes. Primero, existe el riesgo de perpetuar o amplificar sesgos existentes si los datos de entrenamiento reflejan inequidades históricas (Gianfrancesco et al., 2018).

Segundo, la calidad y representatividad de los datos disponibles en sistemas de información sanitaria requiere un análisis crítico, especialmente considerando el subregistro en zonas vulnerables.

Es fundamental que los modelos propuestos incorporen consideraciones éticas sobre la privacidad de datos sensibles, la transparencia de los algoritmos y la interpretabilidad de los resultados para tomadores de decisión. Asimismo, la validación debe considerar no solo métricas estadísticas de desempeño, sino también la relevancia práctica para reducir inequidades reales en el acceso a servicios de salud.

Metodología de investigación

Enfoque de la Investigación

La investigación adopta un enfoque mixto (cuantitativo-cualitativo), alineado con los objetivos específicos del proyecto:

- Cuantitativo: Para el análisis de datos masivos (SALUDATA, DANE, GIS) y modelado predictivo.

- Cualitativo: Para validar resultados mediante entrevistas a expertos en salud pública y machine Learning (ML).

La combinación de métodos permite no solo identificar patrones estadísticos, sino también contextualizar los hallazgos en la realidad bogotana (Hernández-Sampieri et al., 2019). Adicionalmente, la naturaleza híbrida del estudio permite integrar datos numéricos del sistema de salud con análisis contextual de las dimensiones sociales, geográficas y económicas que influyen en el acceso a los servicios sanitarios.

Diseño y Alcance

- **Tipo de estudio**: *Descriptivo-predictivo*.
- **Diseño**: *No experimental-transversal* (datos secundarios históricos). Por lo cual no existirá una manipulación de variables, y adicional, se recopilará la información en un punto temporal especifico, para este caso desde el año 2020.
- **Alcance**:
  - *Descriptivo*: Caracterizar las inequidades territoriales en el acceso a los servicios de salud pública en Bogotá, según su localidad.
  - *Predictivo*: Identificar segmentos poblacionales con características similares en el acceso de salud, mediante técnicas de aprendizaje no supervisado como Clustering, para apoyar la toma de decisiones (Técnicas de Machine Learning).

En este alcance, no se busca establecer relaciones causales, sino identificar patrones y agrupaciones que permitan comprender mejor la distribución de inequidades en la cobertura sanitaria.

Variables de Estudio

Las variables de estudio fueron analizadas a partir de las bases de datos almacenadas en el repositorio de la Secretaría Distrital de Salud (Saludata) e información obtenida del Departamento Administrativo

Nacional de Estadística (DANE), donde se han identificado variables clave que permiten caracterizar las condiciones de acceso, cobertura y utilización de los servicios de salud pública en Bogotá D.C., así como los factores sociodemográficos, territoriales e institucionales que podrían influir en dichas condiciones, entre estas se consideraron las siguientes:

Tabla 1

Variables Dependientes

Variable	Definición Conceptual	Definición Operacional
Acceso a servicios de salud	Capacidad efectiva de la población para utilizar servicios sanitarios cuando los necesita (Penchansky & Thomas, 1981).	Índice compuesto (0-100) basado en el tiempo promedio al centro médico más cercano (*tm_a_centrom*), número de sedes por localidad y oferta de servicios (*CapacidadInstalada*).
Cobertura de atención primaria	Proporción de población con acceso regular a servicios básicos de salud (OMS, 2019).	Porcentaje estimado de afiliados al SGSS por localidad, obtenido de la base *osb_ofertasrv-afiliacionsgss*.
Satisfacción con el sistema de salud	Percepción de calidad y respuesta del sistema de salud.	Promedio de satisfacción (escala 0-5) según localidad, basado en *osb_salud_mental-satisfaccionsalud*.
Mortalidad general	Medida de impacto en salud pública.	Tasa de mortalidad bruta por localidad por año, calculada a partir de la base *osb_tb-mortalidad*.

Nota. Elaboración propia a partir de definiciones de la OMS y literatura técnica en salud pública.

Tabla 2

Variables Independientes

Variable	Definición Conceptual	Definición Operacional
Ubicación geográfica	Localización espacial dentro de Bogotá.	Localidad y barrio de residencia, asignados a partir de shapefiles y coordenadas generadas en Python.
Nivel socioeconómico	Condiciones materiales de vida (DANE, 2020).	Estrato socioeconómico (1–6), asignado por manzana y barrio usando shapefiles de estratificación.
Capacidad hospitalaria	Oferta instalada de salud en territorio.	Número de camas, consultorios, servicios y ambulancias por sede de salud (*CapacidadInstalada*).

Edad	Tiempo transcurrido desde el nacimiento.	Años cumplidos, agrupados por rangos etarios, generados sintéticamente según estructura poblacional (*osb_demografia*).
Género	Identidad reportada.	Masculino, femenino, otro. Distribución basada en proporción real por localidad (*osb_demografia*).
Tiempo promedio al centro médico	Facilidad de acceso físico a la atención.	Minutos promedio estimado por localidad, según la base *tm_a_centrom*.
Instituciones con urgencias	Acceso a servicios inmediatos.	Número de IPS con urgencias habilitadas por barrio/localidad, calculado desde bases *Sedes* y *Servicios*.

Fuente. Elaboración propia a partir de fuentes institucionales como el DANE, Secretaría de Salud y legislación nacional.

Población y Muestra

Población: La población considera todo el distrito de Bogotá, que según el Dane la población de Bogotá es de 7.937.898 millones de habitantes, la muestra empleada se realizó sobre el 1% de la población de Bogotá, es decir, 793.789 habitantes.

Muestra:

- Cuantitativa: Datos completos de las 20 localidades.
- Cualitativa: 3 expertos, en Salud Pública (1) en Modelos de Machine Learning (2).

Técnica de muestreo: *Estratificado* por localidad y estrato.

Marco Muestral

El marco muestral se construye a partir de las bases de datos de la Secretaría Distrital de Salud (Saludata), los registros encontrados en el REPS (Registro Especial de Prestadores de Salud), la información cartográfica del IDECA (Infraestructura de Datos Espaciales de Bogotá)

Tipo de Muestreo

Dado que se utilizan datos administrativos existentes que cubren la totalidad de la población, no se requiere un proceso de muestreo tradicional. Se trabajará con el universo completo de registros disponibles.

Tamaño de la Muestra

Se utilizará la totalidad de registros válidos en el sistema, estimados en aproximadamente 7.5 millones de registros individuales, lo que representa una cobertura superior al 95% de la población bogotana.

Métodos e Instrumentos de Recolección de Datos

Los datos para esta investigación ya se encuentran recolectados y almacenados en los sistemas de información del sector salud, específicamente en:

- Saludata (Sistema integrado de información de la Secretaría Distrital de Salud
- Registro Especial de Prestadores de Salud (REPS) del Ministerio de Salud
- Sistema Integrado de Información de la Protección Social (SISPRO)
- Encuesta Multipropósito de Bogotá (DANE)
- Datos georreferenciados de instituciones prestadoras de servicios de salud
- IDECA (Infraestructura de Datos Espaciales de Bogotá)

Modelos y Técnicas

Se definen los fundamentos teóricos y técnicos de los algoritmos de Machine Learning seleccionados (ej.: K-Means para segmentación, Random Forest para predicción), justificando su idoneidad para analizar inequidades en el acceso a salud. Se vincularán con referentes como Rajkomar et al. (2019) y estudios previos en salud pública.

En esta fase del estudio, se plantea la implementación de dos modelos de aprendizaje automático con el fin de analizar y predecir desigualdades en el acceso a servicios de salud en Bogotá. Estos modelos combinan técnicas cuantitativas, geoespaciales y de inteligencia artificial para obtener resultados robustos, interpretables y aplicables a la formulación de políticas públicas.

Modelo 1: Segmentación Geoespacial

Algoritmo seleccionado: K-Means, optimizado mediante el coeficiente de *Silhouette*, lo que permite determinar el número óptimo de grupos a partir de la cohesión y separación entre los datos.

Variables utilizadas:

- Índice de acceso a servicios (valor normalizado entre 0 y 100)
- Tiempo promedio de desplazamiento a un centro médico (minutos)
- Densidad de Instituciones Prestadoras de Servicios de Salud (IPS) con servicio de urgencias, por kilómetro cuadrado

Salida esperada:

- Agrupamiento de las Unidades de Planeamiento Zonal (UPZ) en tres niveles de vulnerabilidad: alto, medio y bajo acceso a servicios.
- Visualización mediante mapas temáticos para facilitar la interpretación espacial de los resultados.

Modelo 2: Predicción de Riesgo

Algoritmo seleccionado: XGBoost, un modelo de gradient boosting reconocido por su precisión en tareas de clasificación. Se aplicará validación cruzada con k=5 para garantizar la generalización del modelo.

Variables predictoras:

- Cobertura de atención primaria (%)
- Tasa de mortalidad bruta
- Estrato socioeconómico (rango 1-6)
- Número de proyectos activos de infraestructura en salud por UPZ

Indicadores de desempeño:

- Precisión objetivo: ≥ 0.85
- Área bajo la curva ROC (AUC-ROC): ≥ 0.90

Instrumentos y Validación

El desarrollo metodológico se basa en un flujo de procesamiento de datos estructurado en cuatro etapas principales:

1. **Extracción de Datos:**
   a. Información epidemiológica y de infraestructura extraída desde la plataforma SALUDATA, mediante archivos CSV o APIs disponibles.
   b. Información de capacidad instalas en hospitales y centros de salud. (REPS)
   b. Datos espaciales obtenidos a partir de *shapefiles* oficiales de la ciudad de Bogotá. (IDECA)

2. **Transformación y Preprocesamiento:**

   Se aplica limpieza, normalización y creación de variables sintéticas.

   Ejemplo:

   df['índice_acceso'] = (df['cobertura_primaria']*0.6 + (1/df['tiempo_espera'])*0.4)*100

   Este índice compone una métrica híbrida entre cobertura y oportunidad de acceso, ponderada según criterios técnicos.

3. **Modelado y Entrenamiento:**
    a. Implementación de *pipelines* en Scikit-learn, incluyendo codificación categórica (OneHotEncoding) y normalización de variables.
    b. Optimización de hiperparámetros mediante búsqueda en malla (GridSearchCV).
4. **Despliegue de Resultados:**
    a. Visualización de resultados mediante dashboards interactivos en Power BI.
    b. Representación geoespacial con capas dinámicas utilizando la biblioteca folium en Python.

Técnicas de Análisis

La robustez de los modelos será evaluada desde una doble perspectiva: técnica y cualitativa.

Instrumentos aplicados:

1. **Validación Cualitativa (experticia):**
    a. Aplicación de un cuestionario estructurado a expertos en salud pública y machine learning.
    b. Uso de una escala tipo Likert (1 a 5) para evaluar la representatividad de los clusters obtenidos y su utilidad para la formulación de políticas públicas.
2. **Validación Técnica (estadística):**
    a. Prueba de normalidad de Shapiro-Wilk para verificar supuestos en variables de entrada. Análisis de varianza (ANOVA) para determinar diferencias significativas entre los grupos identificados por el modelo de segmentación.
3. **Criterios de aceptación:**
    a. Validación de expertos en el modelo.
    b. p-valor < 0.05 en las pruebas estadísticas para considerar resultados significativos.

Técnicas para el Análisis de Datos

1. Técnicas Cuantitativas (Enfoque principal)

a) Análisis Exploratorio (EDA):

- **Objetivo:** Caracterizar la calidad, distribución y estructura de los datos, así como prepararlos para el análisis posterior
- **Técnicas:**
  - **Integración de fuentes:** Consolidación de datos de múltiples sistemas mediante identificadores únicos (cliente).
  - **Limpieza de datos:** Imputación de valores faltantes, detección y tratamiento de valores atípicos, normalización y transformación de variables.
  - **Estadística descriptiva:** Cálculo de medidas de tendencia central (media, mediana) y dispersión (desviación estándar, rango intercuartílico).
  - **Visualizaciones de datos:** Histogramas (distribución de acceso a servicios), mapas de calor (correlación entre variables), boxplots para detección de outliers, y gráficos de dispersión para exploración multivariada.

b) Modelado con Machine Learning:

- **Objetivo:** Segmentar a la población en grupos con características similares y predecir patrones de acceso a servicios de salud a partir de variables independientes
- **Segmentación (K-Means):**
  - **Datos de entrada:** Variables numéricas normalizadas (localidad, estrato, disponibilidad de infraestructura).
  - **Validación:** Índice de Silhouette para determinar el número óptimo de clusters y análisis de varianza (ANOVA) entre grupos.
- **Predicción (Random Forest):**
  - **Métricas de evaluación:** Precisión, curva AUC-ROC, matriz de confusión.
  - **Optimización:** Ajuste de hiperparámetros mediante búsqueda en malla (GridSearchCV).

c) Análisis Geoespacial:

- **Objetivo:** Explorar la dimensión territorial del acceso a servicios de salud
- **Herramientas:** QGIS y GeoPandas (Python).
- **Técnicas:**
  - **Buffers geográficos:** Estimación de zonas de cobertura de centros de salud.
  - **Mapas de calor (heatmaps):** Visualización de la densidad de población en relación con la oferta de servicios de urgencias.
  - **Superposición espacial:** Cruce entre variables de infraestructura y distribución de población vulnerable.

2. Técnicas Cualitativas (Validación complementaria)

a) Entrevistas semiestructuradas a Expertos:

- **Técnica de análisis:** Codificación abierta mediante análisis temático.
- **Procedimiento:**
  - Transcripción, limpieza y organización de respuestas.
  - Identificación de categorías emergentes (ej. "barreras administrativas", "percepción de cobertura").
  - Triangulación de resultados cualitativos con los hallazgos obtenidos en los modelos cuantitativos.

b) Validación de Modelo:

Estrategia: Juicio de expertos del sector salud para evaluar la pertinencia y aplicabilidad del modelo de segmentación en escenarios de política pública.

- **Instrumento:** Escala tipo Likert (ej. "El modelo refleja adecuadamente las condiciones de acceso por localidad", escala de 1 a 5).

3. Integración de Resultados

- **Triangulación metodológica:** Se contrastarán los resultados obtenidos mediante análisis cuantitativo (clusters y modelos predictivos) con los hallazgos emergentes del análisis cualitativo, para construir una interpretación robusta y contextualizada del fenómeno.
- **Herramientas de soporte:**
  - **Power BI:** Dashboard interactivo con filtros por localidad lo cual permitirá una visualización interactiva de indicadores por localidad y segmento poblacional.
  - **Jupyter Notebooks:** Documentación reproducible de los procedimientos analíticos y modelos implementados.

4. Resultado de Investigación

Análisis de Resultados

Los resultados presentados a continuación derivan del análisis estadístico aplicado a una base de datos sintética de la población de Bogotá, diseñada para simular características demográficas, socioeconómicas y territoriales relevantes en el acceso a los servicios de salud. Esta base permitió estructurar un entorno de prueba controlado para identificar tendencias, patrones y brechas en variables como edad, género, nivel educativo, tipo de afiliación en salud, estrato socioeconómico y distancia al centro de salud más cercano.

```python
def generar_resumen_final(self, df_poblacion, tiempo_total):
    """Genera un resumen final del proceso"""

    resumen = []
    resumen.append("="*80)
    resumen.append("RESUMEN EJECUTIVO - ANÁLISIS DEMOGRÁFICO BOGOTÁ D.C.")
    resumen.append("VERSIÓN CORREGIDA")
    resumen.append("="*80)
    resumen.append(f"Fecha de ejecución: {datetime.now().strftime('%Y-%m-%d %H:%M:%S')}")
    resumen.append(f"Tiempo total de procesamiento: {tiempo_total:.1f} segundos")
    resumen.append("")

    # Estadísticas del dataset
    resumen.append("📊 ESTADÍSTICAS DEL DATASET GENERADO:")
    resumen.append("-" * 50)
    resumen.append(f"• Total de registros: {len(df_poblacion):,}")
    resumen.append(f"• Localidades cubiertas: {df_poblacion['localidad'].nunique()}")
    resumen.append(f"• Rango de edades: {df_poblacion['edad'].min()} - {df_poblacion['edad'].

    # Distribución por género
    dist_genero = df_poblacion['genero'].value_counts(normalize=True) * 100
    resumen.append(f"• Distribución por género:")
    for genero, porcentaje in dist_genero.items():
        genero_nombre = "Femenino" if genero == 'F' else "Masculino"
        resumen.append(f"  - {genero_nombre}: {porcentaje:.1f}%")
    resumen.append("")

    # Top localidades
    resumen.append("🏘️ TOP 5 LOCALIDADES:")
    resumen.append("-" * 30)
    top_localidades = df_poblacion['localidad'].value_counts().head(5)
    for i, (localidad, poblacion) in enumerate(top_localidades.items(), 1):
        porcentaje = (poblacion / len(df_poblacion)) * 100
        resumen.append(f"{i}. {localidad}: {poblacion:,} ({porcentaje:.1f}%)")
    resumen.append("")
```

Imagen 1. Análisis estadístico de variables

Fuente: Elaboración propia

Para el tratamiento y análisis de estos datos, se emplearon herramientas estadísticas programadas en Python, incluyendo cálculos de frecuencias, promedios, desviaciones estándar y distribuciones, así como la construcción de visualizaciones comparativas. Esta fase fue clave para explorar correlaciones entre las variables demográficas y territoriales, y facilitar la interpretación de fenómenos asociados a la inequidad en la cobertura de servicios.

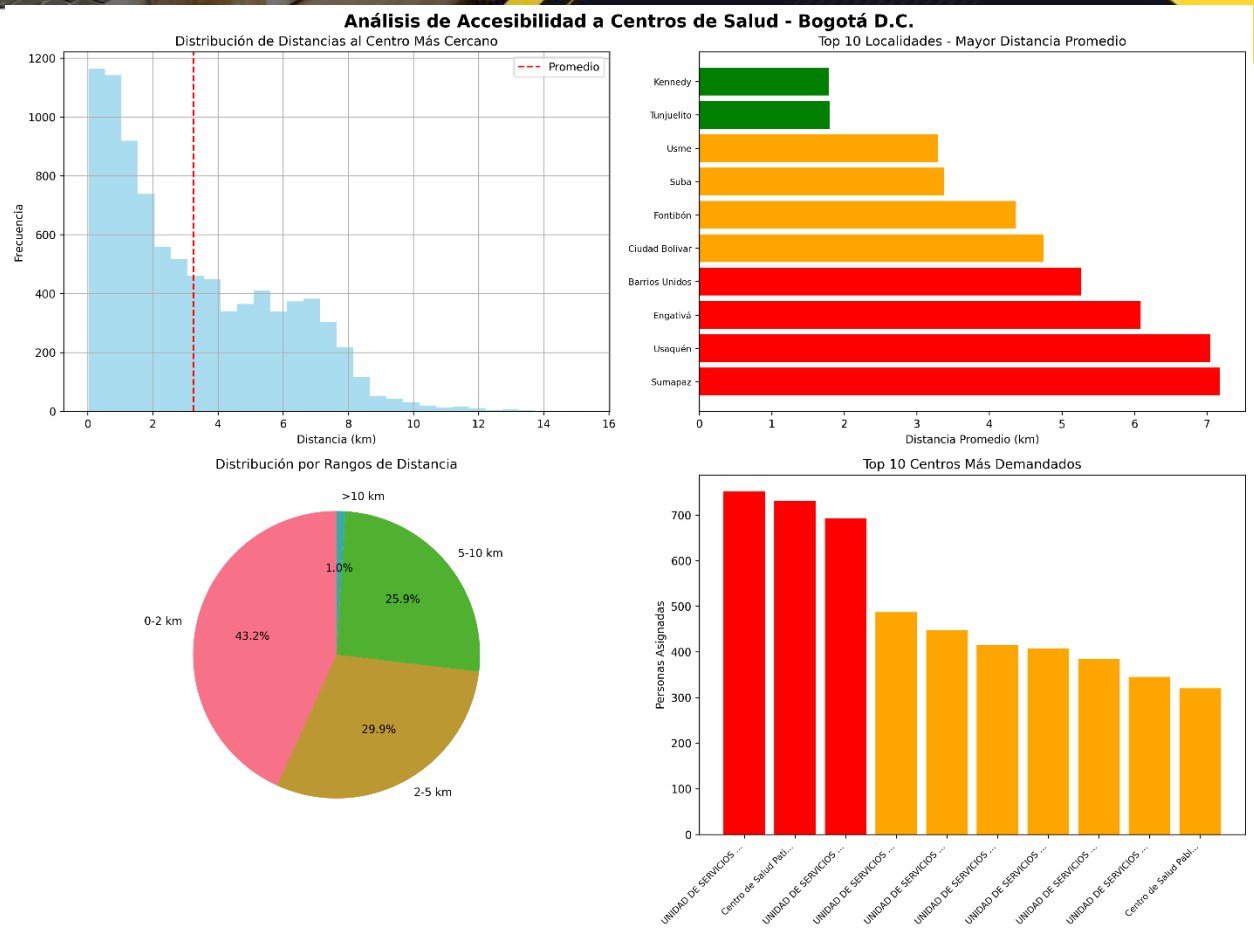

Imagen 2. Análisis de accesibilidad a centros de salud – Bogotá D.C

Fuente: Elaboración propia

Acceso a servicios de salud: A través del anterior grafico se puede concluir que el acceso equitativo y oportuno a los servicios de salud constituye uno de los pilares fundamentales para alcanzar una cobertura sanitaria universal y reducir desigualdades sociales (Organización Mundial de la Salud [OMS], 2010). En el contexto urbano de Bogotá D.C., caracterizado por una alta densidad poblacional y una distribución geográfica desigual de los servicios, la accesibilidad física a los centros de salud representa un factor determinante en la calidad y oportunidad de la atención.

La accesibilidad geográfica se mide comúnmente a través de la distancia entre el lugar de residencia de las personas y el centro de salud más cercano, en este caso para los datos recolectados en el estudio se determinó los siguientes análisis estadísticos:

- Media Aproximada: ~3.3 km en la distribución de distancia de los pacientes a el centro más cercano
- La mayoría de los pacientes vive cerca de un centro de salud, pero hay una minoría considerable con distancias muy grandes
- Cerca del 73.1% de la población está a menos de 5km de un centro de salud, lo cual es positivo. Sin embargo, el 26.9% restante enfrenta barreras moderadas o severas de acceso.
- Entre las localidades más desfavorecidas se encuentra Sumapaz, Usaquén, Engativá y entre las mejores ubicadas Kennedy y Tunjuelito, por lo cual establece que Sumapaz requiere una intervención prioritaria.
- Entre los centros más demandados, se puede identificar que hay una gran cantidad de personas que acceden al mismo sitio o Unidad de servicios en este caso, ocasionando una distribución No equitativa del acceso al mismo servicio.

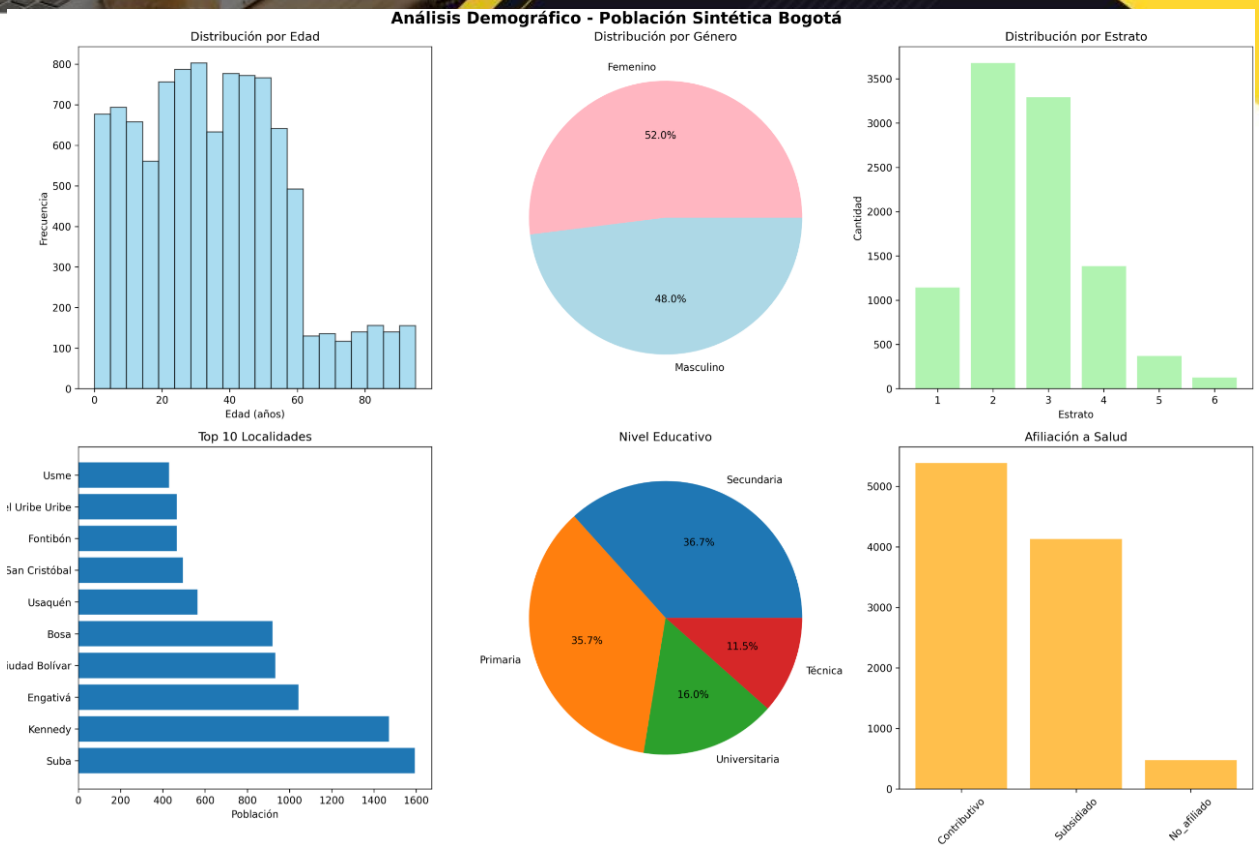

Imagen 3. Análisis demográfico población – Bogotá D.C

Fuente: Elaboración propia

El análisis demográfico de la población sintética de Bogotá proporciona un contexto fundamental para interpretar los patrones de accesibilidad a los servicios de salud. Comprender las características de edad, género, nivel educativo y afiliación al sistema de salud permite una mejor focalización de las estrategias de intervención.

La distribución etaria revela una concentración predominante entre los 15 y 50 años, lo cual es consistente con la estructura poblacional urbana de Bogotá, caracterizada por un alto porcentaje de personas en edad productiva.

Los datos por estrato socioeconómico muestran una clara concentración en los estratos 2 (más del 35%) y 3, lo que indica que la mayor parte de la población analizada pertenece a sectores de ingresos bajos y medios-bajos. Esta distribución es representativa de la realidad socioeconómica bogotana, donde los estratos 1 a 3 comprenden más del 80% de los hogares (Secretaría Distrital de Planeación, 2021).

Las localidades con mayor representación poblacional en la muestra son Suba, Kennedy y Engativá, que coinciden con las zonas más densamente pobladas de Bogotá. Esto resalta la necesidad de reforzar los servicios de salud en estas áreas para responder a una alta demanda potencial

Utilización de servicios: Una utilización desproporcionada de ciertos centros indica que podrían estar funcionando por encima de su capacidad instalada.

Esto puede derivar en mayores tiempos de espera, deterioro en la calidad del servicio y desmotivación para asistir nuevamente, especialmente entre usuarios con menos recursos.

Algunas localidades como **Suba, Kennedy y Engativá** (con mayor población según el análisis demográfico), también aparecen como zonas donde se ubican centros muy demandados.

Esto sugiere una correlación entre concentración poblacional y uso intensivo de servicios, lo que debe ser tenido en cuenta en la planeación sanitaria.

Análisis de Intervención

Asimismo, un aspecto metodológico clave en el desarrollo del modelo propuesto fue la generación y uso de datos sintéticos, los cuales permitieron simular una base poblacional de Bogotá con variables relevantes como edad, género, estrato socioeconómico, afiliación en salud, nivel educativo y localización territorial. Esta base fue construida mediante codificación en Python, considerando distribuciones empíricas y valores promedio obtenidos de fuentes oficiales de la Secretaría de Salud

(Saludata) y el Ministerio de Salud (REPS). Esta estrategia responde directamente al segundo objetivo específico, al permitir analizar de forma preliminar la cobertura de servicios de salud y explorar los factores clave que influencian su accesibilidad, aun en ausencia de datos reales completos.

En este sentido, el uso de datos sintéticos no constituye una conclusión definitiva, sino un instrumento de validación técnica y conceptual del modelo de segmentación propuesto. Tal como plantean Zhao et al. (2021), los datos sintéticos permiten preparar entornos controlados para el diseño y prueba de modelos de Machine Learning, asegurando replicabilidad, ajuste fino de parámetros y exploración de escenarios antes de la implementación en contextos reales. Esto resulta especialmente

relevante en investigaciones donde el acceso a información sensible o completa está limitado por políticas de privacidad o barreras institucionales.

De acuerdo con lo anterior, el análisis estadístico realizado sobre la población sintética permitió identificar patrones críticos de desigualdad en el acceso a los servicios de salud, lo que permite el uso de un insumo fundamental para cumplir con el objetivo general del estudio: diseñar un modelo de Machine Learning que segmente a la población en función de su acceso real y potencial al sistema de salud. Las visualizaciones construidas y los indicadores (como las distancias promedio, distribución por estrato, nivel educativo y afiliación) evidencian desequilibrios territoriales y sociales persistente**s**, que serán decisivos al momento de alimentar y entrenar el modelo de segmentación. Así, los objetivos específicos también se ven reflejados: se estableció un marco de referencia teórico (a través de los determinantes sociales de la salud y la equidad territorial), se exploraron variables clave a partir de datos representativos y se delinearon caminos para validar un modelo aplicado. En conclusión, el análisis estadístico no solo sumo valor a la hipótesis del proyecto, sino que oriento el diseño técnico del modelo hacia una aplicación con impacto real en la planificación de políticas públicas y en la optimización de la infraestructura sanitaria en Bogotá.

Referencias

1. Departamento Administrativo Nacional de Estadística. (2021). Encuesta Nacional de Calidad de Vida (ECV) 2021. DANE. https://www.dane.gov.co/index.php/estadisticas-por-tema/salud/calidad-de-vida-ecv/encuesta-nacional-de-calidad-de-vida-ecv-2021
2. Arcaya, M. C., Arcaya, A. L., & Subramanian, S. V. (2015). Desigualdades en salud: definiciones, conceptos y teorías. Revista Panamericana de Salud Pública, 38(4), 261–271. https://doi.org/10.26633/RPSP.2015.109
3. Eslava-Schmalbach, J., Rincón, C. J., Pinzón, C. E., Villada, A. C., Castillo, J. S., Reveiz, L., & Elias, V. (2017). Índice compuesto de inequidad en salud para un país de mediano ingreso. Revista de Salud Pública, 19(2), 250–258. https://doi.org/10.15446/rsap.v19n2.56325
4. Galvis-Aponte, L. A., & Rico, A. (2023). La Equidad en salud para Colombia: Brechas internacionales y territoriales. Ministerio de Salud y Protección Social.
5. Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. JAMA Internal Medicine, 178(11), 1544–1547. https://doi.org/10.1001/jamainternmed.2018.3763
6. Gómez, F., González, M., & Rodríguez, P. (2019). Evaluación de la cobertura del sistema de salud en Bogotá, 2015-2018. Revista de Salud Pública, 21(2), 245-251.
7. Guerrero, R., Gallego, A. I., Becerril-Montekio, V., & Vásquez, J. (2011). Sistema de salud de Colombia. Salud Pública de México, 53, s144-s155.
8. López-Cevallos, D., Miranda, J. J., & Bernal, O. L. (2022). Desarrollo y validación de un índice de inequidad en salud para contextos urbanos latinoamericanos. Revista Panamericana de Salud Pública, 46, e23. https://doi.org/10.26633/RPSP.2022.23

9. Martínez, L., Rodríguez, J., & Sánchez, T. (2023). Desigualdades en el acceso a servicios de salud en Bogotá: Un análisis territorial. Universidad Nacional de Colombia.
10. Organización Mundial de la Salud. (2008). Subsanar las desigualdades en una generación: Alcanzar la equidad sanitaria actuando sobre los determinantes sociales de la salud. Comisión sobre Determinantes Sociales de la Salud.
11. Vega, R., Acosta, N., Mosquera, P., & Restrepo, O. (2017). La política de salud en Bogotá, 2004-2012. Análisis de la experiencia de atención primaria integral de salud. Medicina Social, 10(2), 73-81.
12. Whitehead, M., & Dahlgren, G. (2006). Concepts and principles for tackling social inequities in health: Levelling up part 1. World Health Organization: Studies on social and economic determinants of population health.
13. Inequidad espacial en acceso a salud Pérez, R., & Gómez, L. (2019). Inequidad espacial en acceso a salud: el caso de la Zona Metropolitana del Valle de México. Revista Mexicana de Ciencias Sociales, 15(2), 35-52.
14. Journal of Medicine, 380(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259
15. Zhao, F., Zhang, L., Du, W., & Liu, H. (2022). Deep learning for emergency department visits forecasting: A systematic review and meta-analysis. Journal of Biomedical Informatics, 128, 104052. https://doi.org/10.1016/j.jbi.2022.104052
16. Piliuk K, Tomforde S. Artificial intelligence in emergency medicine. A systematic literature review. Int J Med Inform. 2023 Dec;180:105274. doi: 10.1016/j.ijmedinf.2023.105274. Epub 2023 Oct 31. PMID: 37944275.
17. American Psychological Association. (2020). Publication manual of the American Psychological Association (7ª ed.). https://doi.org/10.1037/0000165-000.

18. Hernández-Sampieri, R., Fernández-Collado, C., & Baptista-Lucio, P. (2019). Metodología de la investigación (7ª ed.). McGraw-Hill.
19. Organización Mundial de la Salud (OMS). (2019). Cobertura universal en salud: métricas clave. https://www.who.int
20. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
21. Penchansky, R., & Thomas, J. W. (1981). The concept of access: Definition and relationship to consumer satisfaction. Medical Care, 19(2), 127–140. https://doi.org/10.1097/00005650-198102000-00001
22. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259
23. Secretaría Distrital de Salud de Bogotá. (2023). SALUDATA: Sistema de Datos Abiertos en Salud. https://saludata.saludcapital.gov.co
24. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785
25. IDECA. (2024). Localidades de Bogotá D.C. [Archivo shapefile]. Infraestructura de Datos Espaciales de Bogotá - IDECA. Recuperado el 15 de abril de 2025, de https://www.ideca.gov.co
26. Ministerio de Salud y Protección Social. (s.f.). Registro Especial de Prestadores de Servicios de Salud (REPS) [Base de datos]. Recuperado de https://www.minsalud.gov.co