Principios metodológicos para la aplicación de modelos predictivos en historiales clínicos electrónicos orientados al estudio de la diabetes hereditaria

Javier David Coronado López

Director de Trabajo de Grado

Luis Angel Anillo Arrieta

Universidad Nacional Abierta y a Distancia UNAD

Ciencias Básicas, Tecnologías e Ingenieras

Maestría en ciencia de datos y analítica

2026

Luis Angel Anillo Arrieta

Andrés Felipe Solis Pino

Jairo Felipe Ortiz Mosquera

2026

Resumen

En el presente proyecto investigativo se busca analizar la posibilidad de implementar las funciones de análisis de datos en una herramienta innovadora empleada en el área de la salud, el historial clínico electrónico, con el fin de analizar el comportamiento de la diabetes mediante modelos predictivos y observar la veracidad de los resultados a partir de evaluaciones posteriores a su ejecución. Esto se desarrollará mediante un enfoque teórico y metodológico, dando conceptos y bases necesarias para llegar a ser implementadas a futuro.

Se ha observado que los historiales clínicos usualmente son utilizados para la gestión de la información médica, pero también pueden ser implementados para un análisis estadístico en la predicción de enfermedades hereditarias, al tener los antecedentes familiares, podrían saber la tasa de probabilidad de obtenerla (Borges, 2021). La enfermedad hereditaria del tipo crónico en la cual nos centramos es en la diabetes, esta representa una de las condiciones más relevantes, debido a su carácter progresivo y hereditario (Gonzales et al., 2023).

Como resultado, el trabajo aporta un análisis metodológico que contribuye como bases conceptuales para futuras investigaciones orientadas a la detección temprana y el seguimiento analítico de enfermedades hereditarias.

***Palabras clave*:** Enfermedades crónicas, Modelos predictivos, Historial Clínico, Diabetes, prevención.

Abstract

This research project seeks to analyze the possibility of implementing data analysis functions in an innovative tool used in the health sector, the electronic health record, to analyze diabetes behavior using predictive models and observe the accuracy of the results based on evaluations after their implementation. This will be developed through a theoretical and methodological approach, providing the concepts and foundations necessary for future implementation.

It has been observed that medical records are usually used for medical information management, but they can also be implemented for statistical analysis to predict hereditary diseases. By having family history, it would be possible to know the probability of developing a disease (Borges, 2021). The chronic hereditary disease we focus on is diabetes, which represents one of the most relevant conditions due to its progressive and hereditary nature (Gonzales et al., 2023).

As a result, the work provides a methodological analysis that contributes as a conceptual basis for future research aimed at the early detection and analytical monitoring of hereditary diseases.

***Keywords*:** Chronic diseases, Predictive models, medical history, diabetes, Prevention.

Tabla de contenido

Lista de Tablas

Lista de Figuras

Introducción

A lo largo del tiempo, la salud de la población se ha visto afectada por múltiples factores y en algunos casos, ciertas enfermedades latentes pueden desarrollarse asintomáticamente sin un cuidado previo, afectando a gran escala la vida de las personas; dichas enfermedades pueden tener un origen hereditario. En el caso de ser heredadas y ser diagnosticadas de forma temprana, estas podrían ser controladas, evitando así un gran impacto en la salud de la persona, de lo contrario la enfermedad evolucionará de tal manera que el paciente estará sujeto a una regulación excesivamente compleja que afecte su salud y la calidad de vida.

Las enfermedades crónicas hereditarias representan un desafío constante para los sistemas de salud, debido al impacto que generan sobre la calidad de vida de los pacientes. En particular, la diabetes, al ser una enfermedad recurrente y de riesgo hereditario, lo que resalta la importancia de estudios orientados a su análisis temprano desde enfoques basados en datos (Gonzales et al.,2023).

Con el avance de las tecnologías de la información se ha logrado implementar en el área de salud, un control en el historial médico de los pacientes, el historial clínico electrónico, donde los médicos pueden consolidar toda la información del paciente de manera estructurada, rápida y segura. Esta herramienta se ha prestado para realizar estudios estadísticos, usualmente para la gestión hospitalaria y al apoyo en la toma de decisiones institucionales (Borges, 2021).

A partir de estos conceptos, da inicio al siguiente proyecto investigativo, donde se analizará y estudiarán modelos predictivos que puedan ser aplicados, desde una perspectiva teórica y metodológica, a los historiales clínicos electrónicos, con el propósito de contribuir a la detección temprana de la herencia de diabetes en infantes.

Justificación

El aumento en la prevalencia de enfermedades crónicas hereditarias, como la diabetes, ha generado una creciente preocupación en el ámbito de la salud y en sus familiares, especialmente al considerar la posibilidad de su manifestación en una edad temprana. No obstante, este tipo de enfermedades no siempre resulta evidente de manera inmediata, lo que dificulta su identificación oportuna y acciones preventivas, quedando así, en observación clínica constante y tradicional.

Con lo anterior, los historiales clínicos electrónicos se han consolidado como una fuente estructurada de información médica que permite el almacenamiento sistemático de antecedentes clínicos y familiares. Estos registros facilitan el análisis retrospectivo de datos, abriendo la posibilidad de estudiar patrones asociados al comportamiento de las enfermedades hereditarias.

Por lo cual, en el presente proyecto investigativo se justifica en la necesidad de fortalecer el análisis de la información en los historiales clínicos electrónicos, mediante los principios metodológicos que orientan el uso de modelos predictivos en el sector salud; enfocándonos en comprender de manera teórica y metodológica cómo estos modelos son capaces de analizar datos clínicos y obtener resultados sobre el comportamiento futuro de la enfermedad, a partir de antecedentes familiares.

El proyecto aporta al campo de la ciencia de datos aplicados en la salud, en la búsqueda de un análisis metodológico que integre conceptos de ingeniería, analítica de datos y sistemas de información en salud. El enfoque se centra en dar las bases conceptuales que orienten futuras investigaciones, tanto investigativas como aplicativas, sobre el uso de los modelos predictivos como una herramienta adicional en el estudio de enfermedades hereditarias y al tener un previo control para evitar su evolución.

Objetivos

Objetivo General

Analizar el uso de modelos predictivos aplicados a historiales clínicos electrónicos desde una perspectiva metodológica para el estudio de la diabetes hereditaria.

Objetivos Específicos

Analizar la literatura científica reciente sobre el uso de modelos predictivos en diabetes hereditaria para identificar alcances, limitaciones y tendencias metodológicas.

Identificar los criterios0020metodológicos y técnicos que orientan el uso de modelos predictivos aplicados a historiales clínicos electrónicos en el análisis de enfermedades hereditarias.

Describir las características metodológicas de modelos predictivos utilizados en estudios previos sobre diabetes a partir de sus métricas y fundamentos teóricos.

Explorar datos secundarios relacionados con la prevalencia de diabetes para el análisis descriptivo de patrones poblacionales asociados a antecedentes hereditarios.

Desarrollo

Marco Conceptual

Historial Clínico Electrónico (HCE)

El Historial Clínico Electrónico (HCE) es un sistema digital utilizado en instituciones de salud para la gestión de información médica del paciente, que facilita tanto su almacenamiento como su acceso. Este sistema incluye datos sobre antecedentes clínicos, familiares, tratamientos previos, diagnósticos y demás datos relevantes sobre el paciente. Su uso ha mejorado la gestión y el manejo de los recursos médicos, permitiendo una mejor toma de decisiones. Además, es fundamental en los estudios analíticos y la investigación, ya que permite realizar un seguimiento detallado de la evolución de la salud del paciente a lo largo del tiempo, también, permite identificar patrones hereditarios en los pacientes y fundamentar el uso de modelos predictivos orientados al análisis clínico; sin embargo, estos datos suelen emplearse con fines administrativos y de gestión hospitalaria, con ello se pierde información relevante en el aspecto de un diagnóstico previo (Carracedo y Pollán, 2022).

Diabetes Hereditaria

Es una enfermedad crónica que altera la forma en que el cuerpo procesa la glucosa, generando concentraciones elevadas de azúcar en la sangre. La principal causa es la falla hormonal en el páncreas, órgano responsable de la creación de la insulina. Hay dos tipos de diabetes, el tipo 1, que ocurre cuando el cuerpo se vuelve autoinmune, atacando las células beta encargadas de producir insulina, lo que impide que la glucosa sea transportada adecuadamente hacia las células, lo cual genera una necesidad de un tratamiento continuo de suministración de insulina. El tipo 2, las células se vuelven resistentes a la insulina y aunque está asociada más al estilo de vida de las personas, puede tener un componente hereditario en las futuras

generaciones, incrementando la probabilidad de su aparición incluso durante la infancia. Factores como antecedentes directos, estilo de vida y predisposición genética influyen significativamente en la probabilidad de aparición temprana (Gonzales et al., 2023).

Modelo Predictivo

Los modelos predictivos son herramientas que permiten analizar grandes volúmenes de datos y aprender de su comportamiento para la predicción de futuros eventos, tendencias y patrones de comportamiento (Gil, 2020). Este proyecto se orienta a comprender, desde un enfoque metodológico, cómo podrían emplearse estos algoritmos predictivos, junto con los historiales clínicos electrónicos para estimar riesgos relativos y probabilidades condicionadas basadas en antecedentes familiares y patrones epidemiológicos.

Regresión Lineal

Modelo estadístico que, mediante una relación matemática entre variables para predecir valores continuos. Para ello, el algoritmo se construye ajustando una línea que minimiza el error entre valores observados y predichos mediante mínimos cuadrados, y arroja coeficientes que indican la magnitud y dirección del efecto de cada variable predictora sobre la variable respuesta, junto con métricas que midan el porcentaje de la viabilidad del modelo y el error cuadrático medio (IBM, 2019)

Árbol de Decisión

Modelo de aprendizaje supervisado que utiliza una arquitectura jerárquica de reglas lógicas para la clasificación de datos o la regresión de valores numéricos continuos. Este proceso opera mediante la partición recursiva del conjunto de datos en función de la variable predictora que maximiza la homogeneidad interna, utilizando criterios de separación como la ganancia de información o la reducción de la varianza en cada nodo. La estructura resultante se organiza en

una serie de bifurcaciones donde cada nodo interno representa una prueba sobre un atributo específico, derivando en rutas interpretables que conectan de manera lógica el nodo raíz con las hojas terminales. Esta metodología permite que cada ramificación optimice el camino de decisión según parámetros preestablecidos, transformando variables complejas en reglas de asociación claras y visualmente auditables que facilitan la comprensión del comportamiento del modelo predictivo (Kavlakoglu, 2022).

Bosque de Aleatoriedad

El modelo de aprendizaje supervisado se fundamenta en técnicas de ensamblaje por agregación de muestreo, el cual integra múltiples árboles de decisión independientes para optimizar tareas de clasificación y regresión. La arquitectura de este método se basa en la construcción de una vasta colección de estimadores donde cada árbol se entrena sobre una muestra aleatoria distinta del conjunto de datos original. Simultáneamente, el algoritmo introduce una variación estocástica al seleccionar únicamente un subconjunto aleatorio de variables en cada nodo de división, una estrategia diseñada para des correlacionar los árboles individuales. Esta metodología reduce significativamente la varianza y el riesgo de sobreajuste, incrementando la robustez y la capacidad de generalización del sistema mediante el promedio de las salidas en regresión o la votación mayoritaria en clasificación. En consecuencia, el modelo consolida una infraestructura predictiva de alta precisión que compensa las limitaciones individuales de los árboles que lo conforman. (Kavlakoglu, 2022b).

Marco Referencial

La diabetes mellitus tipo 1 (DM1) se define como una patología crónica de etiología autoinmune, caracterizada por la destrucción de las células beta del páncreas responsables de la síntesis de insulina. La carencia absoluta de esta hormona deriva en estados de hiperglucemia

persistente, lo que obliga al paciente a depender de insulinoterapia exógena mediante inyecciones o sistemas de infusión continua. Si bien los avances tecnológicos han optimizado el control glucémico y reducido el riesgo de episodios de hipoglucemia, la eficacia de estas herramientas está intrínsecamente ligada a la autorregulación del paciente. En este sentido, la optimización del régimen terapéutico requiere una gestión integral del estilo de vida, donde la planificación nutricional y la actividad física regular no solo actúan como pilares preventivos, sino como factores determinantes para la estabilización de los niveles de glucemia y la adecuada respuesta a las dosis de insulina administradas (Barrio y Pérez, 2017). Desde una perspectiva clínica y preventiva, este estudio permite comprender la complejidad clínica de la diabetes tipo 1 y resalta que, aunque existen avances tecnológicos para su control, la enfermedad sigue representando una condición crónica que requiere vigilancia permanente y control. Esta característica refuerza la importancia de enfoques orientados a la detección temprana y análisis preventivo, especialmente cuando existe predisposición hereditaria.

En Perú se ha implementación de plataformas digitales y de la Historia Clínica Electrónica (HCE), lo cual les permitió una mejora el sistema de salud y ofrecer una atención más oportuna a través de la estructuración de un historial clínico electrónico (HCE), se implementaron también revisiones sistemáticas utilizando el diagrama PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), este método consiste en un sistema de organización por pasos (identificación, revisión, elegibilidad e inclusión) que permite filtrar grandes cantidades de artículos científicos para seleccionar solo aquellos con la mayor calidad y relevancia; en este caso, se seleccionaron 9 estudios clave entre 2018 y 2021. Lo que permitió estandarizar los criterios de éxito y garantizar que las variables introducidas en el algoritmo fueran actuales y basadas en evidencia sólida. Al limitar la muestra a investigaciones

de alta calidad, se logró reducir el ruido estadístico, asegurando que los factores de riesgo identificados tuvieran un respaldo clínico real y vigente para la toma de decisiones (Valderrama, 2021). Se concluye que la implementación del diagrama PRISMA es un componente crítico en el diseño de modelos predictivos para las HCE, ya que actúa como un filtro de calidad que garantiza que el algoritmo se alimente únicamente de evidencia científica de alto nivel. Lejos de ser solo una herramienta documental, este método fortalece la arquitectura del sistema al asegurar que la selección de variables diagnósticas sea rigurosas y libre de sesgos, permitiendo un análisis clínico con mayor profundidad. De este modo, la mejora metodológica en la depuración de datos se traduce directamente en una mayor precisión y confiabilidad de las alertas preventivas. Este rigor metodológico contribuye significativamente al avance en la comprensión de la patología y sus posibles abordajes terapéuticos, optimizando de manera directa la efectividad de las intervenciones clínicas.

En la región de Salvador de Bahia, en Brasil, se implementó un sistema innovador de HCE llamado AGHUse, el cual destaca por ser un software de código abierto y gratuito a la vez se implementó nueva herramienta de Business Intelligence (BI) para así mantener un control de presupuestal y eficacia en los hospitales públicos. Este módulo de inteligencia permite predecir qué áreas se necesita más inversión y anticipar brotes de enfermedades en la población antes de que ocurran, esto fue posible mediante interconexión en red de 23 unidades de salud y la renovación de servidores. Para los doctores también una herramienta de gran utilidad para agilizar el procedimiento de diagnósticos al tener el historial del paciente en tiempo real, estableciendo así un modelo avanzado de gestión hospitalaria basado en la predicción de datos (Borges, 2021). Aunque este sistema demuestra el potencial del HCE y del Business Intelligence para la gestión y predicción a nivel institucional, su enfoque se orienta principalmente a la

administración hospitalaria desde un enfoque logístico y no a la anticipación clínica individual o hereditaria, lo cual marca una diferencia con el enfoque del presente proyecto.

Por otro lado, las muertes por enfermedades congénitas se han convertido en la segunda causa de mortalidad infantil en Argentina, afectando a un aproximado de 3 de cada 100 niños al nacer y hasta en un 10% de la población infantil antes de los 6 años. Mediante la Red Nacional de Anomalías Congénitas (RENAC) se organizó un estudio que permitió descubrir las causas que lo generaron. Se realizaron 3 tipos de pruebas, diagnóstico prenatal, realizando exámenes de ecografías de alta precisión y muestras de sangre materna, la pesquisa neonatal, para la detección de enfermedades metabólicas y la evaluación sensorial temprana, pruebas acústicas y reflejo rojo. Con esta información se realizó un sistema de comunicación el sistema de salud para poder organizar intervenciones proactivas, como cirugías y tratamiento prenatales (Unicef, 2019). Este antecedente, si bien no se centra específicamente en la diabetes, resalta la relevancia de la detección temprana de condiciones congénitas y hereditarias y la eficacia del objetivo a esta investigación. Asimismo, evidencia cómo la implementación de protocolos diagnósticos rigurosos y la captura de datos de alta precisión permiten catalogar variables clínicas de manera sistemática, optimizando el análisis predictivo y la toma de decisiones terapéuticas oportunas.

En el área de salud, se integra un sistema predictivo, en este artículo se abordará sobre la enfermedad de diabetes, una enfermedad degenerativa. El principal objetivo es crear un modelo de comunicación con otros pacientes y redes sociales, mediante técnicas de Web Scraping y Procesamiento de Lenguaje Natural (PLN). Se puede observar este procedimiento mediante un grafo ponderado no dirigido que conecta a los usuarios con términos médicos clave, permitiendo a los especialistas identificar patrones y trayectorias clínicas basadas en las experiencias de los pacientes. Esta herramienta permite detectar factores de riesgo y anticipar comportamientos que

a menudo no quedan registrados en las historias clínicas convencionales y mejorar la eficacia de las consultas y la precisión de los diagnósticos (Gil, 2020). Este enfoque evidencia la posibilidad de aprovechar nuevas fuentes de información para el análisis de enfermedades crónicas; sin embargo, su aporte principal reside en el análisis del componente social y comunicativo, diferenciándose de la estructuración metodológica de modelos predictivos que operan estrictamente con variables de Historias Clínicas Electrónicas (HCE). No obstante, este antecedente es fundamental, ya que, valida la viabilidad de integrar datos cualitativos y algoritmos en la programación de modelos, demostrando la capacidad de estas herramientas para el diagnóstico preventivo.

Asimismo, se destaca la versatilidad de estos sistemas en el área de salud odontológica; se desarrolló un modelo predictivo innovador diseñado para determinar la necesidad de tratamiento ortodóntico en la atención primaria. Mediante la metodología estadística de V de Cramer, el cual asigna un valor a diferentes rasgos dentales permitiendo construir una ecuación multivariante que es fácil de calcular incluso de forma manual. Mediante matrices de confusión y análisis de la curva ROC, se demostró que este enfoque alcanza una validez del 81.2%, optimizando así la priorización de los pacientes que realmente requieren intervención. (Gonzalez et al., 2018). Este estudio demuestra que modelos estadísticos relativamente simples pueden ser una herramienta de alto valor en el desarrollo de modelos predictivos. A pesar de originarse en el área odontológica, este antecedente es fundamental porque evidencia la viabilidad de implementar algoritmos de clasificación robustos y accesibles. Asimismo, la identificación de patrones predictivos mediante variables categorizadas puede mejorar significativamente la eficiencia operativa en cualquier área de la salud.

Se realiza herramientas que permiten parametrizar variables dinámicas como signos vitales, escalas de riesgo (Glasgow o Norton) y datos antropométricos, transformando el texto libre en datos cuantitativos aptos para el entrenamiento de algoritmos de inteligencia artificial. La infraestructura tiene la capacidad de ejecutar fórmulas automáticas y gestionar los datos con marcas de tiempo precisas, lo cual es esencial para el análisis de series temporales y la detección temprana de anomalías. Se integran módulos que manejan las peticiones, resultados de laboratorio y formularios de indicadores, cada uno de estos procesos puede ser transformados en variables predictoras, dando así optimización de recursos y facilitar la toma de decisiones (Pagan, 2018). La transformación de texto clínico en variables cuantificables evidencia la importancia de la ingeniería de características en el desarrollo de modelos predictivos, aspecto clave que debe considerarse en cualquier análisis basado en HCE.

Se evalúa el uso del cuestionario FINDRISC para determinar el riesgo de diabetes tipo 2 (DT2) en personal militar del Hospital Central FAP de Perú. Esta herramienta crea un filtro de alta precisión para desarrollar un algoritmo de clasificación binaria, el sistema logró que el 81.6% de los pacientes evitara pruebas diagnósticas costosas, centrando la intervención en los pacientes con un puntaje de menor o igual a 13. En este grupo de riesgo, se demostró una prevalencia del 93% de alteraciones en los niveles de glucosa sanguínea, determinada mediante el método de la hexoquinasa, que es el método de referencia internacional más exacto. Estos datos generan unas nuevas columnas en el HCE, como la masa corporal y el perímetro abdominal elevado. Para el procesamiento de datos utilizaron la herramienta de Stata 13 y modelo de regresión de Poisson, con el fin de bases de datos dinámicas capaces de predecir el estado metabólico actual y evitar las pruebas invasivas como la prueba Oral de Tolerancia a la Glucosa (TOTG). (Villena, 2020). Este antecedente resulta especialmente relevante, ya que

demuestra cómo variables clínicas y antropométricas pueden ser integradas en sistemas predictivos que optimizan recursos médicos y reducen procedimientos invasivos, alineándose con los principios analíticos del presente proyecto.

El proyecto permite la gestión de una base de datos para la recolección de muestras de hematología en el Hospital Clínic de Barcelona. La arquitectura del sistema emplea herramientas robustas y modernas como Django para el desarrollo del framework web y MongoDB como base de datos NoSQL, lo que permite manejar grandes volúmenes de información de manera escalable y flexible. El programa le permite crear, buscar, editar y eliminar registros, también, está la posibilidad de importar automáticamente datos de varias fuentes, haciendo estas tareas más rápidas y fáciles para el personal de la salud. El sistema ha sido diseñado como una interfaz web integral y fácil de usar, con el objetivo de optimizar el procesamiento de datos de los pacientes para realizar investigaciones médicas y mejorar la eficiencia de los profesionales de la salud. Otro de los factores que beneficia este sistema es el uso de las estadísticas matemáticas y la predicción de brotes de enfermedades (Llamas, 2017). Si bien este sistema se enfoca principalmente en la gestión y procesamiento de información clínica, evidencia la viabilidad técnica de integrar bases de datos médicas con análisis estadístico, sirviendo como soporte para futuras aplicaciones predictivas.

Se investiga sobre la variabilidad temporal en la base de datos biomédica MIMIC-IV demostrando como los cambios en las distribuciones de probabilidad a lo largo del tiempo, impactan directamente en la robustez de los modelos predictivos de mortalidad en HCE. Para lo cual se recurre a las herramientas de visualización y análisis como son los Mapas de Calor Temporales (DTH) y la Información Geométrica Temporal (IGT), acompañado del Análisis de Correspondencia Múltiple (MCA), para identificar desviaciones críticas en los datos. Se

detectaron cambios en las variables predictoras y en sus relaciones, derivados de la transición del protocolo de codificación ICD-9 a ICD-10 (clasificación internacional de enfermedades) en 2015, lo que alteró significativamente la distribución y asociación de las variables diagnósticas y de procedimiento. Este impacto se evalúa mediante la implementación de dos modelos: Random Forest, que utiliza múltiples árboles de decisión, y Gradient Boosting, método basado en la combinación de modelos débiles para crear uno más robusto. método basado en la combinación de modelos débiles para crear uno más robusto. Ambos fueron validados de manera sistemática analizando los datos anualmente en lotes independientes para asegurar su precisión temporal. El rendimiento fue optimizado mediante grid search (prueba sistemática de diferentes combinaciones de configuraciones) y, posteriormente, los modelos fueron ajustados utilizando el Índice de Youden (medida que maximiza tanto la sensibilidad como la especificidad) para mitigar el efecto del desbalanceo entre los casos positivos y negativos. Los resultados, validados estadísticamente mediante un test chi-cuadrado sobre los agrupamientos de la IGT (Información Geométrica Temporal), confirman que el rendimiento de la IA se degrada cuando los modelos entrenados en un periodo intentan generalizar datos de eras de codificación distintas (Narro, 2023). Se evidencia una limitación crítica de los modelos predictivos en el área de salud, como la dependencia en la estabilidad temporal y la consistencia de los datos clínicos. Desde el punto de vista metodológico, estos elementos deben de incluir protocolos de recalibración constante y monitoreo de deriva frente a cambios normativos o variaciones en los protocolos hospitalarios.

Los HCE se fundamenta en la integración de Estimas de Riesgo Poligénico (PRS) obtenidas mediante estudios de asociación de genoma completo (GWAS). Esto permite transitar de los tradicionales modelos de regresión hacia algoritmos avanzados de inteligencia artificial, como las redes neuronales, para calcular el riesgo relativo y absoluto de enfermedades

complejas, destacando que se debe de superar un riguroso flujo de trabajo que incluye la identificación de variantes genéticas (SNPs) y una validación externa que garantice su capacidad de calibración y discriminación en poblaciones específicas. Se realiza el uso del modelo BOADICEA para cáncer de mama y el SCORE2 para riesgo cardiovascular, los cuales implementan factores genéticos y epidemiológicos para optimizar los cribados clínicos. Por lo cual, es de suma importancia fortalecer la infraestructura de los bancos de información integrados y el análisis del exposoma para nutrir continuamente las capacidades predictivas de los HCE (Carracedo y Pollan, 2022). Aunque estos enfoques avanzados demuestran el alto potencial predictivo de los HCE, su implementación requiere infraestructuras tecnológicas especializadas, lo que limita su aplicabilidad en contextos clínicos generales o pediátricos, reforzando la pertinencia de enfoques metodológicos basados en modelos estadísticos interpretables y datos clínicos disponibles de forma rutinaria.

Para mejorar la exactitud en la predicción de riesgos y optimizar los resultados clínicos, la implementación de inteligencia artificial (IA) y técnicas de minería de datos en los Historiales Clínicos Electrónicos (HCE), se busca la implementación de herramientas de análisis bioquímico avanzado. Entre estas destacan la Cromatografía Líquida de Alta Resolución (HPLC) y la Cromatografía Gaseosa acoplada a Espectrometría de Masas (CG-EM), las cuales son fundamentales para detectar marcadores bioquímicos específicos como la tirosina y la succinilacetona en errores innatos del metabolismo, o el pro-BNP y la creatinina en síndromes nefróticos. Esto permitió la identificación de manera más precisa de factores de riesgo para enfermedades cardiovasculares y renales, integrando variables complejas como la Tasa de Filtración Glomerular (TFG) calculada mediante las fórmulas CKD-EPI, MDRD y Cockcroft-Gault. Estos datos con hallazgos de inmunofluorescencia y biopsia renal facilitan la clasificación

automatizada de patologías como la Glomerulonefritis Membranoproliferativa (GnMP), permitiendo que el sistema no solo estime el curso de enfermedades renales crónicas, sino que también detecte asociaciones críticas entre cargas virales (como el VHC) y el deterioro del filtrado glomerular (Bioanalisis, 2024). Este estudio resalta cómo la integración de variables bioquímicas complejas puede fortalecer los modelos predictivos; sin embargo, también evidencia el alto nivel técnico requerido, lo que refuerza la necesidad de enfoques metodológicos adaptados a la disponibilidad real de datos y un asesor especialista en el área.

Creación de conjuntos de datos y modelos predictivos basados en mensajería HL7, un estándar de comunicación en entornos hospitalarios. Se utilizan técnicas de aprendizaje automático para predecir eventos clínicos como el alta por enfermedad o la duración de la estancia hospitalaria. Los autores implementaron un proceso que incluía consulta y denominación de datos, preprocesamiento y creación de conjuntos de datos, seguido de la construcción de modelos predictivos. Los resultados muestran que la proporción correcta de valores faltantes y el establecimiento del umbral de probabilidad mejoran significativamente el rendimiento de los predictores utilizados. (Barrera et al., 2022). El uso de estándares como HL7 facilita la interoperabilidad de los datos clínicos, aspecto fundamental para garantizar la reproducibilidad y escalabilidad de los modelos predictivos en diferentes entornos hospitalarios.

Desarrollo de un sistema de análisis de datos mediante el método KDD (Knowledge Discovery in Databases) utilizado en el sector de salud pública de alimentos en la región de joven, Perú. El objetivo es mejorar el procesamiento de muchos datos, como la naturaleza de la alimentación de los niños, para poder tomar decisiones oportunas y precisas. Utilizando herramientas como Python y Anaconda, se creó un sistema que podría procesar, analizar y crear estrategias más efectivas para la desnutrición y la anemia en niños, mejorando de manera más

oportuna y precisa para entregar información importante (Fermín Arroyo,2020). Este estudio evidencia la aplicabilidad del método KDD como un marco de trabajo esencial en contextos de salud pública, resaltando la importancia de seguir un flujo metodológico estructurado para transformar registros administrativos en conocimiento estratégico. El antecedente nos demuestra cómo la integración de herramientas de ciencia de datos permite la identificación de perfiles de riesgo nutricional y la automatización de alertas.

SaludCoop por mucho tiempo ha manejado la información de manera local, que son hojas de Excel, esto procede a un problema de mantener una sola versión del documento, provocando cruces de información. Por medio de auditorías han demostrado que la información es poco fiable y pueden afectar a la empresa (Chicuasuque Pérez, 2023). Este caso pone de manifiesto cómo la falta de centralización y control de datos puede afectar la calidad de la información clínica, representando una limitación significativa para cualquier análisis predictivo confiable.

Se ha realizado vigilancia de los defectos congénitos para la prevención y control, realizando la depuración de la base de datos identificando casos repetitivos mediante la aplicación. Se realiza el análisis mediante estadística descriptiva. Como resultados se obtiene que un incremento de casos, pero así mismo gracias a este estudio se pudo prevenir y ver patrones (Instituto Nacional de Salud, 2018). Este estudio evidencia el valor de la estadística descriptiva como punto de partida para la identificación de patrones epidemiológicos, aunque también resalta la necesidad de métodos analíticos más avanzados para estudios predictivos.

En la implantación de una base de datos en la historia clínica, es necesaria seguridad de esta. Para ella se busca la unificación de todas las entidades prestadoras de salud, permitiendo agilidad y seguridad. Se implementa: Manejo de usuarios y contraseña, grupo de usuarios para acceso, servidor mediador y la encriptación de la información (Antonio Sanmiguel, 2019). La

seguridad de la información clínica constituye un requisito indispensable para garantizar el uso ético y responsable de los datos en modelos predictivos, especialmente cuando se trabaja con información sensible de pacientes.

En el marco del Modelo de Adaptación de Roy, se analizó la influencia de los estímulos sobre la adaptación fisiológica y psicosocial en 200 pacientes con Diabetes Mellitus Tipo 2 (DMT2) mediante un diseño descriptivo, transversal y predictivo. A través de la aplicación de modelos de regresión lineal (simple y múltiple), se identificó que el tiempo de diagnóstico es un predictor crítico, explicando la variabilidad de la adaptación fisiológica con un coeficiente de determinación del 44%, donde se observa que por cada año de evolución la estabilidad metabólica disminuye. Asimismo, la presencia de complicaciones clínicas como fue la neuropatía, presente en el 57.5% de la muestra, retinopatía y afecciones cardiovasculares. Demostrando un impacto predictivo del 74% sobre la adaptación psicosocial, evidenciando cómo variables registradas en la HCE pueden anticipar el deterioro funcional y emocional del paciente. Estos hallazgos resaltan la importancia de estructurar datos de estímulos contextuales como el estado marital, la ocupación y la escolaridad dentro de los sistemas digitales, ya que estos factores influyen significativamente en la capacidad de respuesta del organismo ante la evolución de la enfermedad. (Universidad de la Sabana, 2019.). Este estudio aporta evidencia cuantitativa sobre la influencia determinante del tiempo en la degradación de la estabilidad metabólica de la diabetes tipo 2, reforzando la importancia de integrar análisis longitudinales en el diseño de modelos predictivos. Al identificar que variables psicosociales y complicaciones clínicas poseen un peso predictivo superior al 70%, se valida la necesidad de que las HCE no solo almacenen datos bioquímicos, sino que estandaricen factores del entorno del paciente para anticipar estados de desadaptación y fallos terapéuticos antes de que ocurran.

Gracias a la colaboración entre el CLAP (Centro Latinoamericano de Perinatología) y la OPS, se ha consolidado el Sistema Informático Perinatal como la herramienta tecnológica base para el diagnóstico y registro de anomalías mayores y menores en las áreas de maternidad y neonatología. La implementación de este sistema, junto con la Historia Clínica Perinatal (HCP), permite una transición del registro manual hacia modelos de salud digital integrada que, facilitando el uso de la Triple Vigilancia, un modelo predictivo y analítico que monitorea la causa, la ocurrencia y los resultados de salud. Lo más destacado para el desarrollo de modelos predictivos en la HCE es la capacidad de estos sistemas para realizar proyecciones de impacto económico y epidemiológico hasta el año 2030. El estudio demuestra que el uso de formularios electrónicos estructurados y la codificación permiten crear escenarios hipotéticos que estiman costos directos e indirectos, lo que convierte a la HCE en un motor de inteligencia de negocios capaz de identificar poblaciones vulnerables, alertar sobre brotes (como el del virus del Zika) y evaluar la efectividad y los costos en las intervenciones preventivas (de Francisco et al., 2020). El estudio demuestra que el uso de formularios electrónicos estructurados y codificación estandarizada permite crear escenarios hipotéticos para estimar costos directos e indirectos, lo que convierte a la HCE en un motor de inteligencia de negocios capaz de identificar poblaciones vulnerables, alertar sobre brotes y evaluar la efectividad de intervenciones preventivas. En última instancia, estos antecedentes refuerzan la relevancia de los sistemas regionales de información clínica para la detección temprana de anomalías, estableciendo una base de datos fundamental para el estudio de enfermedades hereditarias y la implementación de modelos predictivos orientados a la medicina genómica.

Se desarrolla e implementa un software especializado para el registro sistemático y control de historias clínicas, diseñado bajo un modelo de arquitectura cliente-servidor y

metodologías de ingeniería de software que garantizan la integridad de los datos en salud. El proyecto fue implementado mediante el Lenguaje Unificado de Modelado (UML) para la especificación y documentación de los procesos clínicos, lo que permitió traducir la anamnesis y la exploración física en estructuras de datos normalizadas. Para el desarrollo de modelos predictivos en la HCE, este antecedente es fundamental, ya que destaca la importancia de las herramientas de gestión de bases de datos relacionales y el cumplimiento de estándares de calidad asegurando que la información sea interoperable y esté libre de errores. Al sistematizar la captura de datos, el software no solo optimiza el flujo administrativo, sino que crea un repositorio de información fidedigna que sirve como insumo para identificar patrones de salud, permitiendo que la HCE evolucione de un simple registro a una herramienta de soporte para la toma de decisiones clínicas basadas en la evidencia recolectada de forma sistemática. (Henao, 2019). El proyecto demuestra la viabilidad técnica de implementar sistemas de HCE, la organización rigurosa y normalizada de la información clínica. Gracias a su metodología se garantiza que el repositorio de datos sea lo suficientemente sólido para permitir, en etapas posteriores, la integración de algoritmos de inteligencia artificial y modelos de minería de datos.

La implementación de sistemas de información en el sector salud, no solo optimiza la comunicación médico-paciente y reduce los costos operativos, sino que actúa como el habilitador tecnológico para la integración de herramientas de inteligencia artificial y soporte a la decisión clínica. Lo más llamativo de este estudio para el desarrollo de modelos predictivos en la HCE es la identificación de métodos híbridos de redes neuronales, como el RGNN, que combina Redes Neuronales Recurrentes (RNN) para procesar datos médicos secuenciales y Redes Neuronales de Gráficos (GNN) para modelar eventos temporales, logrando predecir con éxito prescripciones y futuras enfermedades basadas en el historial del paciente. Asimismo, se destaca el uso de

arquitecturas basadas en Cloud Computing y Blockchain para garantizar la integridad y el intercambio seguro de datos, junto con el modelo narrativo Health Text Line (HTL) que emplea Procesamiento de Lenguaje Natural (PLN) y minería de texto para extraer eventos clínicos de registros no estructurados. Estos avances demuestran que la automatización de la historia clínica electrónica, apoyada en estándares de interoperabilidad como el Intercambio Electrónico de Documentos (EDI) y arquitecturas orientadas a servicios (SOA), transforma el registro tradicional en una plataforma predictiva capaz de anticipar diagnósticos y mejorar sustancialmente la seguridad en la administración de tratamientos (Preciado et al., 2021). La automatización del ecosistema clínico trasciende la mejora operativa al consolidar una infraestructura de datos persistentes indispensable para el despliegue de analítica prescriptiva. El modelo de salud digital proactiva se fundamenta en la explotación de Big Data y protocolos de ciberseguridad granular, los cuales configuran un entorno de alta disponibilidad para la ejecución de algoritmos de detección temprana y la optimización de la precisión terapéutica basada en evidencia digital.

Síntesis de Antecedentes y Aporte

Tabla 1

Matriz de análisis sistemático de antecedentes y aportes a la investigación

Autor y año	Tema Estudiado	Principales portes	Relevancia para el estudio	Limitaciones
Barrio y Pérez (2017)	Gestión integral y control glucémico en Diabetes Mellitus tipo 1 (DM1).	Define la DM1 como patología autoinmune y destaca la insulinoterapia junto con el	Provee la base clínica sobre la naturaleza crónica y evolutiva de la enfermedad,	Se enfoca en la autorregulación del paciente sin integrar modelos algorítmicos para la predicción

		estilo de vida como pilares de control.	justificando la necesidad de vigilancia permanente.	automática de riesgos.
Valderrama (2021)	Implementación de HCE en Perú y uso de la metodología PRISMA para selección de evidencia	Aplicación del diagrama PRISMA (9 estudios clave) para garantizar que las variables del algoritmo tengan respaldo clínico real.	Valida el rigor metodológico en la selección de variables diagnósticas, asegurando que el modelo predictivo se alimente de datos de alta calidad.	Se centra en la depuración documental y técnica de la HCE, pero no profundiza en la fase de entrenamiento de modelos de Machine Learning.
Borges (2021)	Sistema AGHUse y Business Intelligence (BI) en hospitales públicos de Salvador de Bahía, Brasil.	Implementación de software de código abierto interconectado en 23 unidades para predecir inversión necesaria y anticipar brotes epidemiológicos.	Valida la viabilidad técnica de utilizar herramientas de inteligencia de datos en red para la toma de decisiones en tiempo real.	Su enfoque es primordialmente logístico y administrativo, sin profundizar en la anticipación clínica individual o hereditaria de patologías crónicas.
UNICEF (2019)	Análisis de anomalías congénitas y mortalidad infantil en Argentina a	Uso de protocolos de diagnóstico prenatal, pesquisas neonatales y evaluación	Resalta la eficacia de la detección temprana en condiciones hereditarias y la importancia de	Aunque el rigor en la captura de variables es alto, el estudio no se centra en la diabetes ni en el uso de algoritmos de

	través de la red RENAC.	sensorial para organizar intervenciones quirúrgicas proactivas.	capturar datos de alta precisión para el análisis predictivo.	aprendizaje supervisado.
Gil (2020)	Análisis del componente social y comunicativo en pacientes con diabetes.	Implementación de Web Scraping y PLN mediante un grafo ponderado no dirigido para conectar usuarios con términos médicos.	Valida la viabilidad de integrar datos cualitativos y algoritmos para detectar factores de riesgo omitidos en historias clínicas convencionales.	Su enfoque es primordialmente social/comunicativo y se diferencia de los modelos que operan estrictamente con variables de HCE.
Gonzalez et al. (2018)	Modelo predictivo para necesidad de tratamiento ortodóntico en atención primaria.	Uso de la estadística V de Cramer para crear una ecuación multivariante con una validez del 81.2% (Curva ROC).	Evidencia la viabilidad de implementar algoritmos de clasificación robustos y accesibles mediante variables categorizadas en salud.	El estudio se origina y aplica exclusivamente en el área de salud odontológica.
Pagan (2018)	Parametrización de variables dinámicas y transformación de texto clínico	Conversión de signos vitales, escalas de riesgo y notas de texto libre en variables	Resalta la importancia de la ingeniería de características y el análisis de	Se enfoca en la infraestructura de captura y transformación de datos, sin detallar la

	en datos cuantitativos.	aptas para el entrenamiento de IA con marcas de tiempo precisas.	series temporales para la detección temprana de anomalías en HCE.	precisión específica de modelos de clasificación final.
Villena (2020)	Evaluación de riesgo de Diabetes Tipo 2 mediante el cuestionario FINRISC en personal militar (Perú).	Implementación de regresión de Poisson para predecir el estado metabólico, logrando que el 81.6% de los pacientes evitaran pruebas invasivas.	Demuestra cómo la integración de variables antropométricas (IMC, perímetro abdominal) en el HCE optimiza recursos y mejora la detección preventiva.	El estudio utiliza la herramienta Stata 13 y se limita a una población específica (militar), lo que podría requerir ajustes para una población civil general.
Llamas (2017)	Gestión de base de datos para hematología en el Hospital Clínic de Barcelona.	Arquitectura escalable usando Django y MongoDB (NoSQL) para importar, buscar y editar grandes volúmenes de datos médicos automáticamente.	Evidencia la viabilidad técnica de integrar bases de datos médicas con análisis estadístico y predicción de brotes para optimizar procesos clínicos.	El enfoque es primordialmente de gestión y procesamiento de información, actuando como soporte y no como un sistema predictivo final de diagnóstico.
Narro (2023)	Variabilidad temporal y robustez de modelos	Uso de Random Forest y Gradient Boosting validados con	Demuestra que cambios normativos (como el paso de	El rendimiento de la IA se degrada significativamente cuando se intenta

	predictivos en la base de datos MIMIC-IV.	Mapas de Calor Temporales (DTH) e Índice de Youden para mitigar el desbalanceo de datos.	ICD-9 a ICD-10) alteran la distribución de variables, exigiendo protocolos de recalibración constante.	generalizar datos de periodos con protocolos de codificación distintos.
Carracedo y Pollan (2022)	Integración de Estimas de Riesgo Poligénico (PRS) y GWAS en HCE.	Transición de modelos de regresión tradicionales a redes neuronales para calcular riesgo relativo y absoluto mediante SNPs y validación externa.	Valida la evolución de los modelos predictivos hacia la IA avanzada y la importancia de flujos de trabajo rigurosos para la calibración en poblaciones específicas.	Requiere infraestructuras tecnológicas especializadas, lo que limita su aplicación en contextos clínicos generales o pediátricos sin recursos de genómica.
Bianalisis (2024)	Implementación de análisis bioquímico avanzado y minería de datos en HCE.	Uso de HPLC y CG-EM para detectar marcadores específicos y fórmulas como CKD-EPI para calcular la Tasa de Filtración Glomerular (TFG).	Demuestra cómo la integración de variables bioquímicas y fórmulas matemáticas complejas fortalece la precisión de los modelos predictivos automatizados.	El alto nivel técnico requerido y la necesidad de equipos de cromatografía y biopsia limitan su uso a la disponibilidad real de datos en entornos hospitalarios estándar.

Barrera et al. (2022)	Creación de modelos predictivos basados en mensajería estándar HL7.	Implementación de un flujo que incluye preprocesamiento y ajuste de umbrales de probabilidad para predecir altas y estancias hospitalarias.	Valida el uso de estándares de comunicación hospitalaria para garantizar la interoperabilidad, escalabilidad y reproducibilidad de los modelos.	El estudio se centra en eventos de gestión hospitalaria (altas/estancias) y no en el diagnóstico específico de patologías crónicas degenerativas.
Fermín Arroyo (2020)	Sistema de análisis de datos mediante el método KDD en salud pública (Perú).	Uso de Python y Anaconda para procesar datos nutricionales (anemia y desnutrición) y automatizar alertas estratégicas.	Evidencia la eficacia del marco de trabajo KDD para transformar registros administrativos en conocimiento y perfiles de riesgo nutricional.	El análisis se limita al sector de salud alimentaria infantil y no aborda la complejidad de variables para enfermedades crónicas en adultos.
Chicuasuque Pérez (2023)	Problemática del manejo de información local (Excel) en SaludCoop.	Identifica riesgos de falta de centralización, duplicidad de datos y baja fiabilidad de la información clínica por falta de sistemas integrados.	Justifica la necesidad de migrar de registros manuales a sistemas predictivos centralizados para garantizar la	Se enfoca en el diagnóstico administrativo del problema de datos, sin proponer un modelo de inteligencia artificial para su resolución.

			integridad de los datos.	
Instituto Nacional de Salud (2018)	Vigilancia de defectos congénitos para prevención y control en Colombia.	Uso de estadística descriptiva y depuración de bases de datos para identificar casos repetitivos y patrones epidemiológicos.	Valida la importancia de la limpieza de datos y el análisis descriptivo como fases críticas antes de cualquier implementación predictiva.	El alcance es puramente descriptivo y preventivo, careciendo de una fase de modelado predictivo avanzado mediante aprendizaje automático.
Sanmiguel (2019)	Seguridad y unificación de bases de datos en historias clínicas electrónicas.	Implementación de servidores mediadores, encriptación, manejo de perfiles de usuario y contraseñas para la protección de datos.	Establece los requisitos de seguridad y ética indispensables para el manejo de información sensible en modelos predictivos.	Se centra en la infraestructura de seguridad informática y no en el desarrollo de algoritmos de aprendizaje automático.
Univ. de La Sabana (2019)	Influencia de estímulos en la adaptación fisiológica y psicosocial en 200 pacientes con DMT2.	Uso de regresión lineal para identificar que el tiempo de diagnóstico y complicaciones (neuropatía 57.5%) predicen el deterioro con	Valida la integración de variables psicosociales y longitudinales (tiempo de evolución) para anticipar fallos	Aunque el modelo predictivo es fuerte, se basa en estadística descriptiva y regresión, sin explorar modelos de ensamble como Random Forest.

		un 74% de impacto.	terapéuticos en diabetes.	
Francisco et al. (2020)	Sistema Informático Perinatal y Triple Vigilancia (CLAP/OPS).	Implementación de formularios electrónicos estructurados para crear escenarios hipotéticos y proyecciones de impacto epidemiológico hasta 2030.	Valida el uso de la HCE como un motor de inteligencia de negocios para identificar poblaciones vulnerables y alertar sobre brotes.	El estudio se enfoca en el área perinatal y neonatal, requiriendo adaptación para el seguimiento de enfermedades crónicas en adultos.
Henao (2019)	Software especializado para el registro y control de historias clínicas bajo arquitectura cliente-servidor.	Uso de Lenguaje Unificado de Modelado (UML) para normalizar la anamnesis y exploración física en bases de datos relacionales.	Demuestra la importancia de la normalización y el cumplimiento de estándares de calidad para asegurar que los datos sean aptos para IA.	Se centra en la etapa de desarrollo y gestión del software, sin llegar a la implementación de algoritmos de aprendizaje supervisado.
Preciado et al. (2021)	Sistemas de información e integración de IA híbrida (RGNN) en el sector salud.	Uso de Redes Neuronales Recurrentes (RNN) y de Gráficos (GNN) para predecir prescripciones y enfermedades	Valida la transición de la HCE tradicional a una plataforma de analítica prescriptiva y proactiva mediante Big	La implementación de modelos híbridos como RGNN requiere una capacidad de cómputo y volumen de datos que superan la realidad

	basadas en el historial.	Data y Cloud Computing.	de muchos centros de salud locales.

Fuente. Elaboración Propia.

Marco Teórico

El historial clínico electrónico es un sistema digital destinado al almacenamiento estructurado de información médica del paciente, incluyendo antecedentes personales, enfermedades previas y registros clínicos relevantes; adicionalmente, incorpora información sobre vínculos familiares relevantes desde una perspectiva clínica. La actualización de estos registros ha sido descrita como un proceso progresivo asociado a los distintos eventos clínicos documentados durante la atención médica. Los antecedentes familiares constituyen un factor relevante en la evaluación del riesgo de enfermedades crónicas, aunque su análisis aislado no permite una estimación probabilística precisa, lo que genera un escenario de incertidumbre clínica en la toma de decisiones médicas.

Desde un enfoque teórico, la literatura en análisis de datos en salud distingue entre técnicas de análisis exploratorio y predictivo como marcos analíticos complementarios. Por su parte, el análisis predictivo se fundamenta en el uso de datos históricos para la formulación de modelos estadísticos orientados a la estimación de probabilidades asociadas a eventos clínicos futuros.

Adicionalmente, se consideran principios de minería de datos como apoyo al análisis exploratorio, orientados a la identificación sistemática de patrones y relaciones relevantes entre variables clínicas. De manera complementaria, se reconoce el uso de enfoques de aprendizaje automático como referencia conceptual, sin embargo, el presente trabajo se centra en modelos estadísticos clásicos, como la regresión y los modelos de clasificación, y análisis probabilístico.

Desde una perspectiva teórica, estas limitaciones han motivado el desarrollo de enfoques analíticos y modelos estadísticos orientados a la cuantificación de probabilidades asociadas a antecedentes familiares.

Modelo de Regresión Lineal

Modelo estadístico que permite analizar la relación entre una variable dependiente y una o más variables independientes, siendo ampliamente utilizado en estudios epidemiológicos para identificar tendencias y asociaciones entre factores de riesgo y aparición de enfermedades. En el contexto de la investigación clínica, este tipo de modelo ha sido ampliamente utilizado para analizar la influencia teórica de factores hereditarios y clínicos sobre variables de interés, aportando una base interpretativa para el análisis de riesgo.

Árbol de Decisión

El modelo funciona mediante la segmentación de los datos dividiendo un conjunto de observaciones en ramas. El árbol de decisión es un modelo que permite la identificación de factores o patrones de los datos que están más asociados a una mayor prevalencia de diabetes para las futuras generaciones. Su estructura jerárquica facilita la interpretación de las decisiones del modelo, característica relevante en contextos clínicos donde se requiere comprender la relación entre variables y resultados.

Bosque Aleatorio

El modelo de bosque aleatorio se construye a partir de la combinación de múltiples árboles de decisión, cada árbol produce una predicción, y el modelo final integra las predicciones individuales de múltiples árboles, lo que desde un enfoque teórico contribuye a mejorar la estabilidad y robustez de las estimaciones. Desde un enfoque teórico, este tipo de modelo mejora la estabilidad de las predicciones al reducir la variabilidad asociada a un único árbol de decisión.

Importancia de las Herramientas

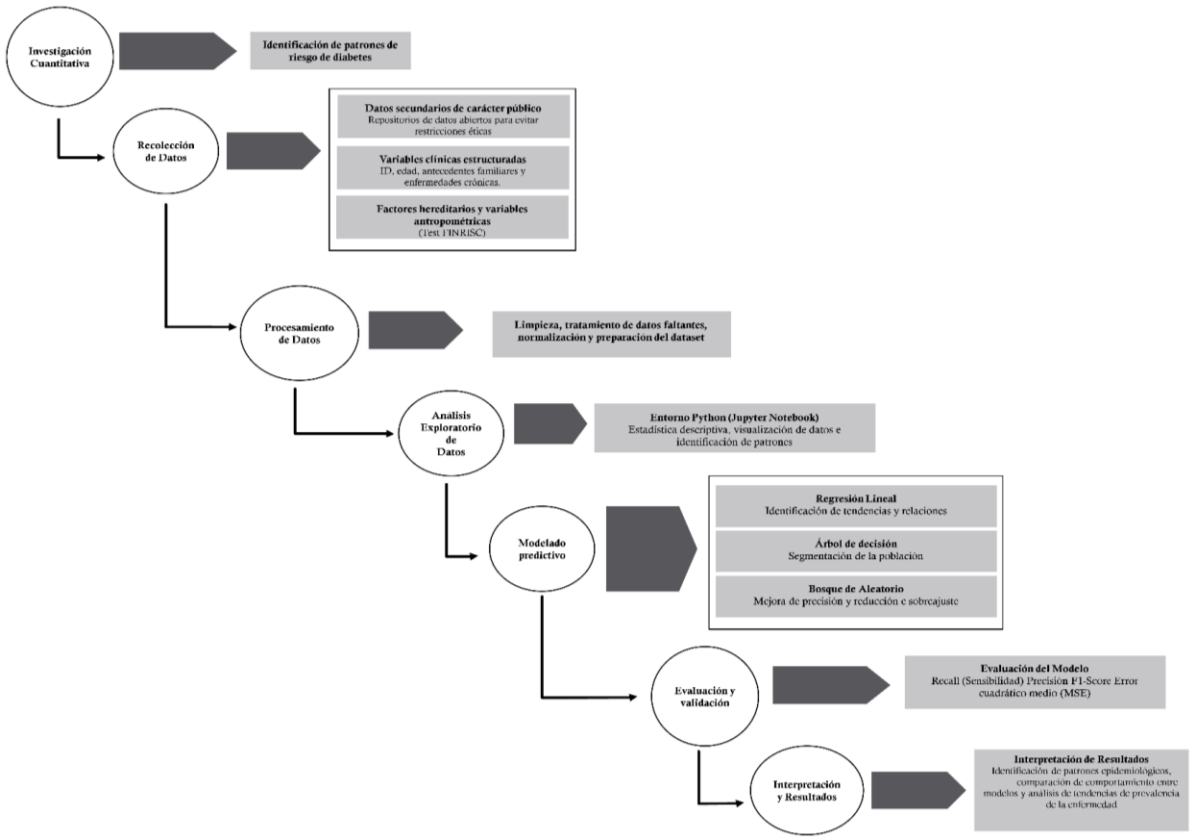
Con el fin de fortalecer el análisis y la predicción de enfermedades, resulta fundamental contar con procesos sistemáticos de recolección, depuración y análisis de datos clínicos, que contribuyan a la identificación temprana del riesgo y a la mitigación de la progresión de la enfermedad. Se necesita una recolección de datos continua, amplia, general y automática, de tal manera que no exista pérdida de información o redundancia en los mismos. También el análisis de los datos nos permite ver un panorama de cómo esta enfermedad puede afectar por generaciones, desde que momento se volvió más propenso y al ser controladas a tempranas edades, esta enfermedad no evolucione hacia estados de mayor complejidad clínica. Estas técnicas han sido propuestas como mecanismos para fortalecer la capacidad explicativa y predictiva de los modelos utilizados en estudios de salud.

Metodología

Para el proyecto de tipo investigativo, se adopta un enfoque cuantitativo, con un alcance descriptivo y predictivo, orientado al análisis de datos clínicos para la identificación de patrones asociados al riesgo de diabetes. Inicialmente, se realiza una revisión de modelos predictivos aplicados en el ámbito de la salud, con el fin de identificar enfoques metodológicos y criterios de aplicación que sirvan de base para la construcción del modelo propuesto.

El estudio se apoya en el análisis de registros clínicos y datos epidemiológicos relacionados con la diabetes, priorizando la identificación de antecedentes familiares y patrones de recurrencia. Dado que el acceso a historiales clínicos reales se encuentra restringido por consideraciones éticas y de confidencialidad, se emplean datos secundarios de carácter público, lo que permite desarrollar y validar el enfoque metodológico sin comprometer información sensible.

Figura 1
Diagrama de flujo metodológico



Fuente. Elaboración Propia.

La figura anterior presenta el flujo metodológico, el cual inicia con la definición del enfoque de investigación y la recolección de datos secundarios de carácter público. Posteriormente, se desarrollan las etapas de preprocesamiento y análisis exploratorio de datos, seguidas de la implementación de modelos predictivos y su respectiva evaluación. Finalmente, se realiza la interpretación de los resultados, permitiendo identificar patrones relevantes asociados al riesgo de diabetes.

Diseño de Instrumentos

Considerando que los Historiales Clínicos Electrónicos operan en entornos digitales, el instrumento principal de recolección corresponde a registros estructurados de información

clínica, los cuales incluyen variables como identificación del paciente, antecedentes familiares, presencia de enfermedades crónicas y rangos etarios. Estas variables permiten caracterizar perfiles de riesgo asociados a la diabetes.

De manera complementaria, se contempla el uso de cuestionarios clínicos estandarizados, orientados a profundizar en factores hereditarios y antecedentes relevantes, los cuales pueden ser incorporados como variables adicionales dentro del modelo predictivo.

Para el procesamiento y análisis de la información se emplea el lenguaje de programación Python, utilizando entornos de desarrollo como Jupyter Notebook, lo que facilita la conexión con bases de datos, la limpieza de los registros y la aplicación de técnicas de minería de datos. Estas herramientas permiten realizar procesos de agrupamiento y clasificación de pacientes con características clínicas similares, sentando las bases para la posterior implementación de modelos de aprendizaje automático.

Enfoque Metodológico en Salud Predictiva

El desarrollo de modelos predictivos eficaces en el ámbito de la investigación sanitaria requiere una base metodológica que asegure la coherencia entre las variables seleccionadas, su naturaleza clínica y familiar, y el comportamiento de los algoritmos utilizados para la estimación del riesgo. Según Valderrama (2021) y Fermín Arroyo (2020), el uso de herramientas como Python permite establecer flujos de procesamiento de datos orientados a la identificación de perfiles de riesgo, transformando registros sin procesar en opciones estratégicas. La metodología se centra en la aplicación de modelos estadísticos clásicos y algoritmos de clasificación supervisada, tales como la regresión lineal, los árboles de decisión y los bosques aleatorios, con el fin de cuantificar el aporte predictivo de variables clínicas, antropométricas y antecedentes familiares. Esta estructuración permite que, mediante procesos de validación estadística, se

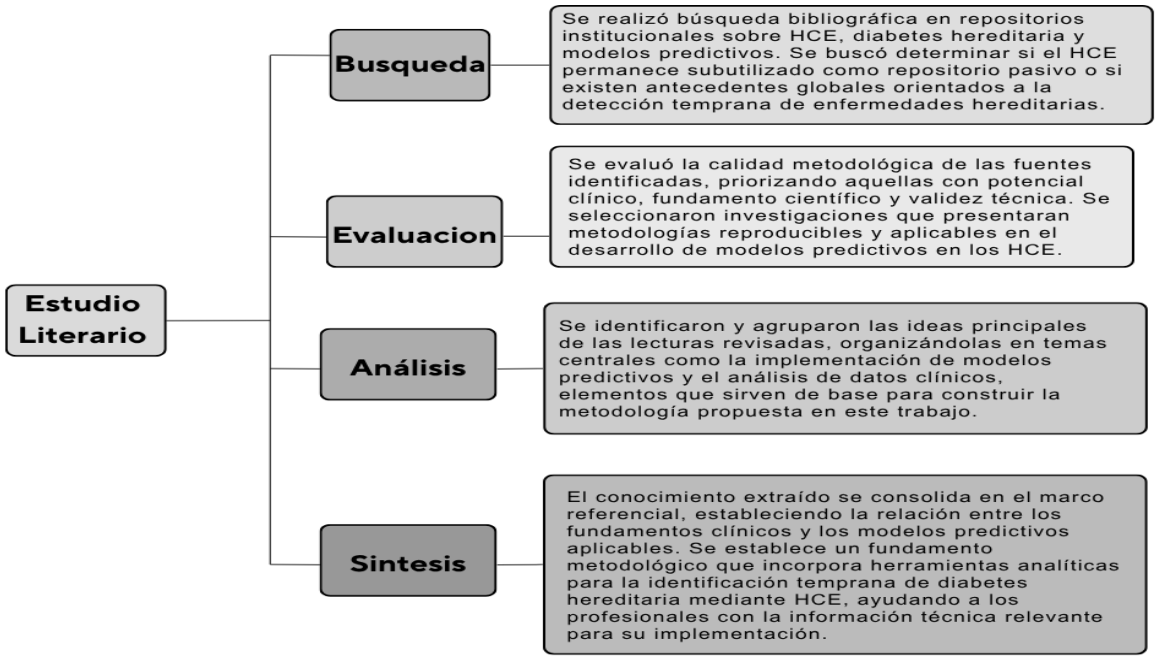
estime la probabilidad de eventos metabólicos o deterioros funcionales, apoyando estrategias de detección temprana y seguimiento clínico sin reemplazar la evaluación médica directa, fundamentando la toma de decisiones en evidencia recolectada y el procesamiento sistemático. (Villena, 2020; Universidad de la Sabana, 2019).

La documentación y normalización de los procesos de captura de información, tales como la anamnesis y el registro de antecedentes familiares, aseguran que los datos sean tratados bajo estructuras estandarizadas y con menor riesgo de error. De acuerdo con Henao (2019) y Barrera et al. (2022), la utilización de lenguajes de modelado y estándares de comunicación garantiza que la información clínica, incluso en fases de investigación, sea interoperable y apta para el entrenamiento de algoritmos de análisis avanzado. Al integrar la estadística multivariante con la validación de modelos predictivos, es posible establecer protocolos analíticos orientados a la detección temprana del riesgo, con un alto grado de validez técnica y coherencia metodológica.

Revisión Literaria Bajo Enfoque SALSA

Se realiza la revisión literaria mediante el enfoque metodológico SALSA (Search, Appraisal, Synthesis, Analysis), el cual permite organizar de manera sistemática la identificación, evaluación y análisis de estudios científicos relevantes. Este enfoque facilita la trazabilidad del proceso investigativo y garantiza un análisis riguroso de la evidencia disponible. (Dermody et al., 2022)

Figura 2
Framework SALSA



Fuente. Elaboración Propia.

En la figura anterior se presentan las cuatro fases del enfoque metodológico SALSA, a partir de las cuales se estructuró el proceso de revisión de literatura. La aplicación de este enfoque permitió desarrollar un análisis sistemático de las fuentes científicas, garantizando la identificación, evaluación, organización e interpretación de la evidencia relevante para el desarrollo del presente proyecto investigativo.

Modelo de Regresión Lineal Aplicado al Riesgo de Diabetes

En este estudio, el modelo de regresión lineal se emplea como una herramienta analítica para examinar la relación cuantitativa entre variables clínicas, antropométricas y antecedentes familiares con respecto al riesgo de diabetes. Su aplicación permite identificar tendencias generales y asociaciones estadísticas entre los factores explicativos y la variable de interés,

proporcionando una primera aproximación al comportamiento de los datos en un contexto epidemiológico.

El ajuste del modelo se realiza mediante el método de mínimos cuadrados, lo que posibilita estimar el peso relativo de cada variable independiente dentro de la ecuación del modelo. Este procedimiento facilita la identificación de factores con mayor influencia estadística sobre el riesgo de desarrollar la enfermedad; sirve como un insumo metodológico para la selección y depuración de variables relevantes en fases posteriores del análisis.

Adicionalmente, la regresión lineal permite evaluar la consistencia y estabilidad de las relaciones identificadas, aportando una base cuantitativa para contrastar resultados con modelos de clasificación más complejos. De esta manera, su uso dentro del estudio no se limita a la predicción directa, sino que contribuye a la comprensión estructural de los datos clínicos y familiares analizados.

Árboles de Decisión como Modelo de Clasificación Clínica

Los árboles de decisión se emplean en este estudio como modelos de clasificación orientados a segmentar la población analizada en distintos niveles de riesgo de diabetes, a partir de reglas lógicas construidas sobre las variables de entrada. Cada nodo del árbol representa una condición clínica, antropométrica o familiar, mientras que las ramas describen posibles combinaciones de factores que conducen a una clasificación específica.

Este enfoque metodológico permite traducir relaciones complejas entre variables en estructuras jerárquicas fácilmente interpretables, lo cual resulta especialmente relevante en contextos clínicos y de investigación en salud. A través de esta segmentación, es posible identificar perfiles de riesgo diferenciados, evidenciando cómo la interacción entre antecedentes familiares y variables clínicas influye en la probabilidad estimada de desarrollar la enfermedad.

Asimismo, los árboles de decisión facilitan el análisis comparativo entre grupos de pacientes, permitiendo evaluar la coherencia de los patrones identificados y su correspondencia con la evidencia clínica reportada en la literatura. Su incorporación en el estudio fortalece el componente metodológico al aportar transparencia en el proceso de clasificación y apoyar la interpretación de los resultados obtenidos por otros modelos predictivos.

Bosque Aleatorio como Modelo de Mejora Predictiva

El modelo de bosque aleatorio se incorpora en este estudio como una estrategia metodológica orientada a mejorar la capacidad predictiva y la robustez de los resultados obtenidos por los modelos individuales. Su funcionamiento se basa en la construcción de múltiples árboles de decisión, entrenados de manera independiente sobre subconjuntos aleatorios de los datos y de las variables disponibles, lo que permite capturar patrones diversos presentes en la población analizada.

La agregación de las predicciones generadas por cada árbol contribuye a reducir la varianza del modelo y a minimizar el riesgo de sobreajuste, fenómeno común en conjuntos de datos clínicos caracterizados por heterogeneidad y posibles inconsistencias en los registros. Este enfoque resulta especialmente pertinente cuando se analizan datos poblacionales con múltiples factores clínicos y familiares que interactúan de manera no lineal.

Dentro del estudio, el bosque aleatorio se utiliza como un mecanismo de validación comparativa frente a modelos más simples, permitiendo evaluar la estabilidad de las predicciones y la consistencia de los factores identificados como relevantes. De esta forma, su aplicación fortalece el enfoque metodológico al aportar estimaciones más confiables del riesgo de diabetes y apoyar la toma de decisiones basada en evidencia analítica.

Evaluación y Validación de los Modelos Predictivos

La evaluación del desempeño de los modelos predictivos se realiza mediante métricas estadísticas orientadas a medir tanto la capacidad de clasificación como la confiabilidad de las predicciones generadas. En particular, se emplean indicadores como precisión, recall y F1-score, los cuales permiten analizar el equilibrio entre la identificación correcta de casos de riesgo y la minimización de errores de clasificación.

La selección de estas métricas responde a la naturaleza potencialmente desbalanceada de los datos clínicos, en los cuales los casos de riesgo pueden representar una proporción menor de la población total. En este contexto, el recall adquiere especial relevancia al medir la capacidad del modelo para detectar correctamente a los individuos con mayor probabilidad de desarrollar la enfermedad, aspecto clave para la detección temprana.

Adicionalmente, la validación de los modelos se orienta a garantizar la consistencia de los resultados frente a variaciones en los datos de entrada, evitando conclusiones sesgadas derivadas del uso de datos secundarios. Este proceso permite asegurar que los modelos desarrollados presentan un desempeño estable y alineado con los objetivos del estudio, fortaleciendo la confiabilidad metodológica de los resultados obtenidos.

Uso de Datos Secundarios en Salud

Se reutiliza información clínica y administrativa almacenada en repositorios de datos abiertos, lo cual constituye un activo estratégico para generar conocimiento epidemiológico y analítico, sin los dilemas éticos y de privacidad asociados a la recolección primaria de datos clínicos. Según Gil (2020), el uso de datos secundarios permite identificar trayectorias poblacionales y patrones de evolución de enfermedades crónicas, como la diabetes, a través del análisis de registros ya existentes. No obstante, como advierte Narro (2023), el análisis de

grandes repositorios biomédicos requiere la aplicación de protocolos de validación rigurosos para mitigar fenómenos como la deriva de datos y asegurar que el rendimiento de los modelos predictivos se mantenga estable frente a cambios temporales o variaciones en las fuentes de información secundaria.

Para que la explotación de estos datos abiertos sea adecuada, esta depende directamente de la capacidad técnica para transformar registros históricos en estructuras analíticas consistentes y aptas para su procesamiento estadístico. Por lo cual, al combinar diferentes herramientas de análisis de datos, es posible transformar información médica, como mediciones del cuerpo de los pacientes y escalas que evalúan qué tan riesgosa es su condición, en datos confiables que pueden usarse para analizar comportamientos históricos de la enfermedad y apoyar la estimación del riesgo poblacional, en coherencia con los objetivos del modelo predictivo propuesto. (Pagan, 2018; Llamas, 2017).

La visualización de datos mediante gráficos de tendencia y distribución permite identificar patrones temporales y diferencias poblacionales en la prevalencia de la enfermedad. Estas representaciones gráficas constituyen una herramienta clave para la interpretación de los resultados del modelo predictivo y para la comunicación de hallazgos epidemiológicos.

Instrumentos y Variables en HCE

La selección de variables para predecir el riesgo de diabetes en este estudio no es arbitraria, sino que se apoya en instrumentos clínicos validados que estandarizan los indicadores clínicos y epidemiológicos. Un referente clave es el cuestionario FINDRISC, el cual permite extraer variables objetivas y estandarizadas como el índice de masa corporal y el perímetro abdominal. Como demuestra Villena (2020), la inclusión de estas métricas en modelos de regresión permite identificar asociaciones significativas con alteraciones glucémicas, aportando

capacidad predictiva al modelo, validando el uso de datos antropométricos como predictores de bajo costo y alta efectividad. De igual forma, la capacidad de parametrizar variables dinámicas y signos vitales mediante marcas de tiempo resulta esencial para el entrenamiento y validación de algoritmos orientados a la detección de anomalías metabólicas. (Pagan, 2018).

Finalmente, la literatura refuerza que el historial familiar y los antecedentes hereditarios son predictores críticos para la diabetes y otras condiciones hereditarias. Según Henao (2019) y Borges (2021), la captura sistemática de estos antecedentes dentro de estructuras de datos normalizadas permite que la predisposición hereditaria se traduzca en pesos estadísticos dentro de ecuaciones multivariantes utilizadas por los modelos predictivos. Al combinar estos factores familiares con hallazgos bioquímicos y el análisis de la variabilidad temporal, el modelo investigativo adquiere la capacidad de proyectar el riesgo de degradación metabólica. Este enfoque asegura que la selección de variables sea rigurosa y esté alineada con los estándares internacionales de investigación clínica y analítica de datos. (Bioanalisis, 2024; Narro, 2023).

Modelos Predictivos en HCE con Datos Abiertos

Para el desarrollo del modelo predictivo se requiere el análisis de información clínica poblacional; sin embargo, debido al carácter sensible y confidencial de los datos provenientes de las Entidades Promotoras de Salud (EPS), el presente estudio emplea conjuntos de datos abiertos de carácter oficial como una alternativa metodológica válida. En particular, se utiliza la base de datos "Prevalencia de Diabetes Mellitus en personas de 18 a 69 años en el Departamento de Bolívar", disponible en el portal oficial de Datos Abiertos de Colombia (Gobernación de Bolívar, 2024), la cual contiene información agregada correspondiente a poblaciones entre los 18 y 69 años a nivel municipal.

Dado que los datos analizados corresponden a agregaciones poblacionales por municipio y no a historiales clínicos individuales, el estudio adopta el concepto de unidades territoriales como aproximación analítica al comportamiento colectivo del riesgo de diabetes. Bajo este enfoque, cada municipio es tratado como una unidad de observación, lo que permite modelar variaciones espaciales y temporales en la prevalencia de la enfermedad sin incurrir en inferencias clínicas individuales.

Este enfoque resulta coherente con los objetivos investigativos del trabajo, en la medida en que se busca explorar la viabilidad de modelos predictivos aplicables a entornos de Historia Clínica Electrónica (HCE) desde una perspectiva poblacional. Es importante precisar que el propósito no es diagnosticar pacientes ni estimar riesgos individuales, sino estructurar un prototipo metodológico que demuestre cómo pueden integrarse técnicas de análisis de datos y aprendizaje automático en contextos de información agregada.

Como se ha mencionado previamente, el desarrollo del modelo se realiza mediante el lenguaje de programación Python, utilizando el entorno Jupyter Notebook como plataforma de

ejecución. Esta elección tecnológica permite documentar paso a paso cada procedimiento aplicado sobre el conjunto de datos, facilitando la reproducibilidad del análisis y la verificación de resultados.

Inicialmente, se lleva a cabo un análisis exploratorio de la base de datos, en el cual se agrupan las observaciones por municipio con el fin de organizar la información en función de las unidades territoriales definidas. Durante esta etapa se verifica la consistencia de los registros, la estructura de las variables anuales y la integridad de los valores porcentuales de prevalencia.

Posteriormente, se estructuran representaciones gráficas que permiten visualizar la evolución temporal de la diabetes mellitus en la población estudiada. Estas visualizaciones constituyen una fase preliminar del proceso analítico, orientada a comprender la dinámica general de los datos antes de proceder a la implementación de modelos predictivos supervisados.

Es importante resaltar que esta fase inicial no tiene como finalidad emitir conclusiones definitivas sobre el comportamiento epidemiológico, sino establecer una base descriptiva sólida que sustente las etapas posteriores de modelado y evaluación.

Descripción del Flujo de Procesamiento de Datos

El procesamiento de la información se estructuró a partir de un flujo metodológico explícito, secuencial y replicable, garantizando la coherencia entre la obtención de los datos, su transformación analítica y la posterior implementación de modelos predictivos. Este flujo permite documentar cada etapa del procedimiento técnico, asegurando trazabilidad y consistencia metodológica.

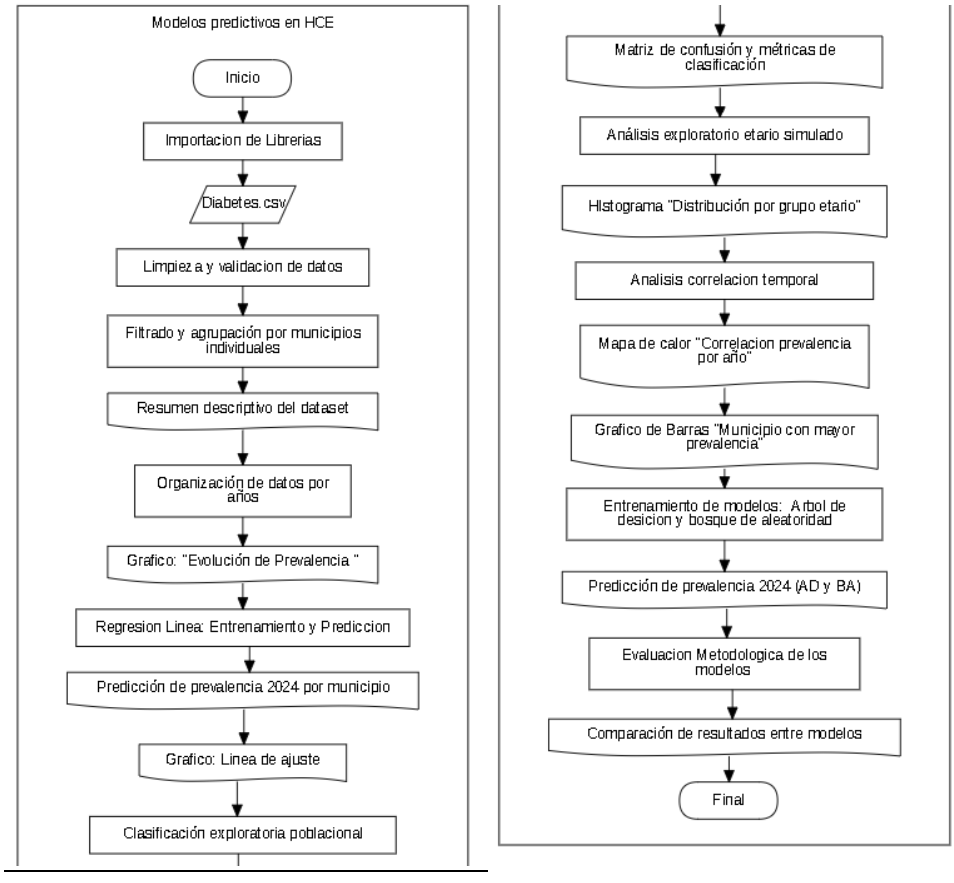
En primer lugar, se realizó la ingesta de los datos abiertos oficiales, seguida de un proceso de limpieza y validación orientado a eliminar registros agregados no pertinentes, estandarizar variables temporales y verificar la consistencia interna de los datos. Esta etapa

incluyó la revisión de valores atípicos evidentes, la comprobación de rangos válidos en porcentajes de prevalencia y la confirmación de que las estructuras anuales se encontraran correctamente organizadas.

Posteriormente, se desarrolló una fase de transformación y organización de variables, en la cual los registros fueron estructurados por municipio y por año. Esta etapa permitió reorganizar la información en formato analítico longitudinal, facilitando la comparación Inter temporal. Adicionalmente, se construyeron indicadores derivados como tendencias temporales simples y estructuras comparativas entre periodos, constituyendo una fase básica de ingeniería de características a nivel poblacional.

Figura 3

Diagrama de flujo del proceso de análisis predictivo y entrenamiento de modelos en HCE.



Fuente. Elaboración Propia.

Posteriormente, se desarrolla una fase de análisis exploratorio, en la cual se emplean técnicas estadísticas descriptivas y representaciones gráficas con el fin de examinar la estructura temporal de los datos y su comportamiento territorial. Esta etapa permite comprender la dinámica longitudinal de la prevalencia sin emitir juicios interpretativos definitivos, concentrándose exclusivamente en la caracterización estructural de la información.

En la fase de modelado predictivo se implementan algoritmos supervisados de distinta naturaleza. En primer lugar, se aplica regresión lineal simple con fines de proyección temporal por municipio. Posteriormente, se incorporan modelos basados en árboles de decisión y técnicas de ensamble como el bosque de aleatoriedad, con el objetivo de contrastar enfoques paramétricos y no paramétricos dentro de un mismo marco analítico.

La etapa final corresponde a la evaluación metodológica de los modelos, en la cual se analizan métricas de desempeño y consistencia predictiva. Es importante enfatizar que estos modelos se plantean como prototipos analíticos aplicables a entornos de Historia Clínica Electrónica (HCE) desde una perspectiva poblacional. En ningún caso se pretende sustituir criterios clínicos ni realizar diagnóstico individual, sino demostrar la viabilidad técnica de integrar análisis predictivo en sistemas de información en salud basados en datos agregados.

Resultados

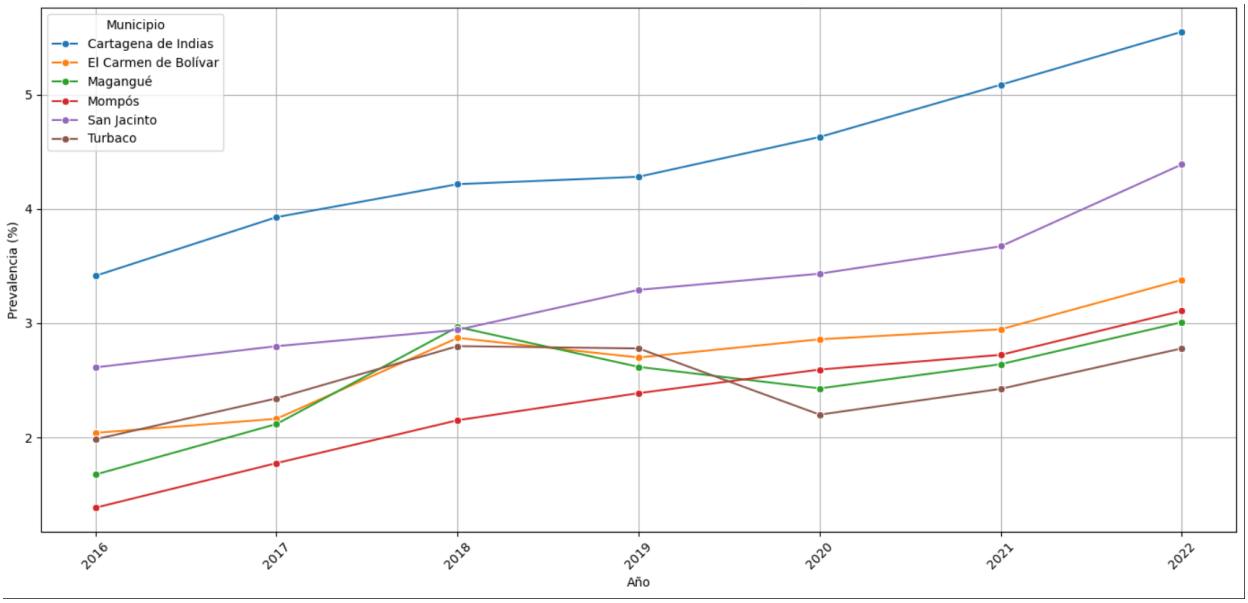
El presente capítulo expone los principales hallazgos derivados del proceso analítico desarrollado en la investigación. Los resultados se organizan en función de las etapas metodológicas previamente definidas, iniciando con el análisis exploratorio de los datos, seguido del análisis complementario mediante simulación, la aplicación de modelos predictivos y la evaluación de su desempeño.

Esta estructura permite presentar de manera progresiva la evidencia obtenida, articulando los hallazgos empíricos con los objetivos específicos del estudio y facilitando la interpretación de los patrones identificados en relación con la prevalencia de la diabetes.

Análisis Exploratorio de la Prevalencia de Diabetes Mellitus

Figura 4

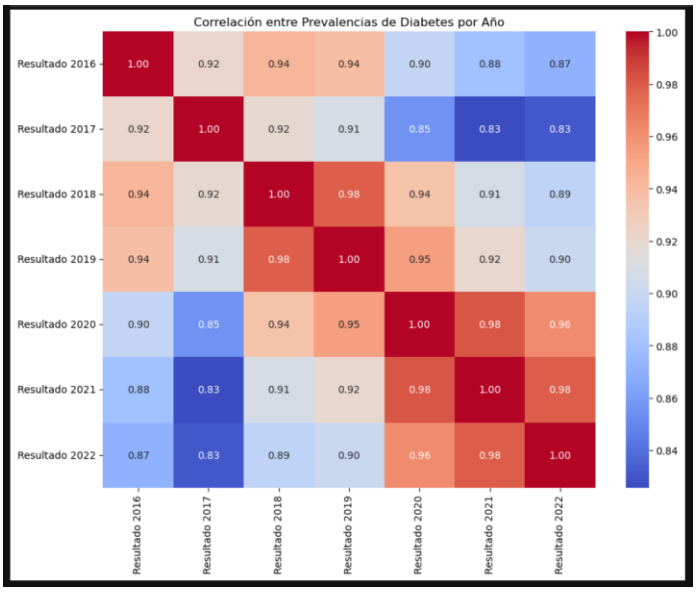
Evolución temporal de la prevalencia por municipio



Fuente. Elaboración Propia.

Figura 5

Mapa de calor de correlación temporal



Fuente. Elaboración Propia.

El análisis exploratorio permitió examinar el comportamiento longitudinal de la prevalencia de Diabetes Mellitus en los municipios estudiados, identificando patrones de estabilidad, crecimiento y variabilidad interanual.

A partir de la evolución temporal por municipio (Figura 4), se observa que Cartagena de Indias concentra los valores más elevados de prevalencia a lo largo del periodo analizado, manteniendo además una trayectoria relativamente estable y con tendencia sostenida. Este comportamiento indica consistencia estructural en la dinámica epidemiológica del municipio, lo que facilita la modelación predictiva bajo supuestos de continuidad temporal. De forma similar, municipios como Mompós y San Jacinto presentan patrones de crecimiento progresivo, aunque con magnitudes inferiores. La pendiente observada en sus series temporales sugiere una evolución gradual del indicador, sin rupturas abruptas entre periodos consecutivos.

En contraste, algunos municipios evidencian oscilaciones más marcadas entre años, reflejando variaciones interanuales menos estables. Este tipo de comportamiento introduce mayor incertidumbre en procesos de proyección, debido a la presencia de fluctuaciones que no siguen una tendencia lineal claramente definida.

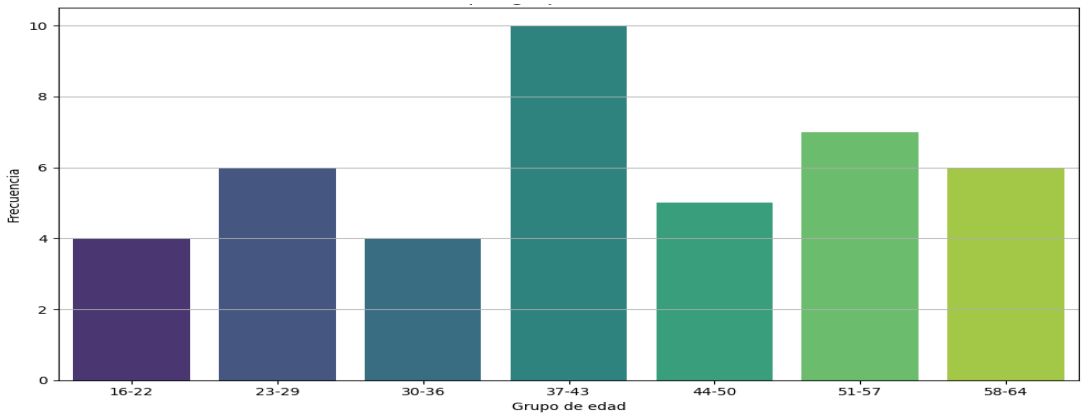
El mapa de calor de correlación temporal (Figura 5) complementa este análisis al cuantificar la relación estadística entre los distintos años del estudio. Se observan coeficientes de correlación elevados en la mayoría de las combinaciones interanuales (valores cercanos a 0.90 e incluso superiores a 0.95 en años consecutivos), lo que indica una fuerte dependencia temporal entre los periodos. En términos analíticos, esto sugiere que la prevalencia observada en un año determinado guarda una relación significativa con los años inmediatamente anteriores y posteriores. Particularmente, las correlaciones más altas se concentran entre años contiguos, lo cual es consistente con un comportamiento epidemiológico acumulativo y progresivo, más que con cambios abruptos o estructuralmente independientes entre periodos.

Desde una perspectiva metodológica, estos hallazgos exploratorios respaldan la pertinencia de aplicar modelos predictivos basados en continuidad temporal, dado que la estructura de correlación sugiere estabilidad y coherencia longitudinal en los datos agregados. Asimismo, permiten anticipar que los modelos podrían capturar patrones de tendencia con un grado razonable de consistencia, especialmente en municipios con menor volatilidad interanual.

Análisis Exploratorio Complementario por Grupos Etarios con Datos Simulados

Figura 6

Distribución simulada por intervalos etarios.



Fuente. Elaboración Propia.

Dado que la base de datos utilizada corresponde a información agregada a nivel municipal y no contiene registros individuales con variable edad, se implementó una simulación controlada de intervalos etarios con fines estrictamente metodológicos. Esta simulación no pretende reconstruir microdatos reales ni estimar prevalencias específicas por edad, sino demostrar el procedimiento analítico que podría aplicarse en un entorno de Historia Clínica Electrónica (HCE) cuando se dispone de datos clínicos desagregados.

La generación de los intervalos etarios permitió estructurar una distribución hipotética en rangos de edad y aplicar técnicas básicas de segmentación poblacional, visualización de frecuencias y análisis comparativo entre grupos. En términos metodológicos, esta etapa funciona como una demostración de ingeniería de características orientada a variables demográficas, mostrando cómo podrían incorporarse covariables etarias dentro de modelos supervisados.

El histograma obtenido (Figura 6) presenta una concentración relativa mayor en el intervalo comprendido entre los 37 y 43 años. Es importante subrayar que este patrón corresponde únicamente al comportamiento de la simulación construida bajo supuestos

estadísticos controlados y no a una estimación empírica derivada de datos reales del departamento de Bolívar. Por tanto, los resultados no fueron utilizados para entrenar modelos predictivos ni para validar hipótesis epidemiológicas.

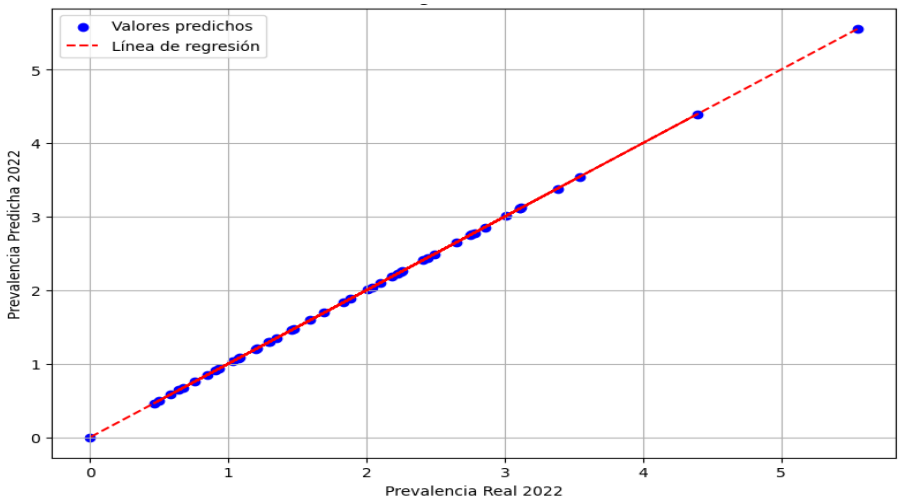
En cuanto al sustento clínico, las guías oficiales de la American Diabetes Association, señalan que el riesgo de desarrollar Diabetes Mellitus tipo 2 aumenta a partir de los 35 años. Si bien el presente análisis no utiliza datos individuales reales, la referencia teórica respalda la pertinencia de incluir la variable edad en modelos predictivos cuando la estructura de datos lo permita (Standards of Care in Diabetes, 2024).

Modelado Predictivo

Aplicación del Modelo de Regresión Lineal

Figura 7

Representación del modelo de regresión línea sobre la prevalencia de diabetes 2022.



Fuente. Elaboración Propia.

La aplicación del modelo de regresión lineal permitió estimar la tendencia temporal de la prevalencia de Diabetes Mellitus en los municipios analizados, evidenciando un comportamiento creciente en la mayoría de los casos. Este modelo se fundamenta en la relación funcional entre

una variable independiente temporal (años) y una variable dependiente cuantitativa (porcentaje de prevalencia), bajo el supuesto de comportamiento aproximadamente lineal en el corto plazo.

Desde el punto de vista metodológico, la regresión lineal permite estimar una recta de ajuste que minimiza el error cuadrático medio entre los valores observados y los valores predichos. En este contexto, el modelo busca capturar la dirección (pendiente) y la magnitud del cambio interanual de la prevalencia en cada municipio, estableciendo una proyección para el año 2024 con base en la tendencia histórica.

En la representación gráfica (Figura 7), los puntos azules corresponden a los valores observados de prevalencia por municipio, mientras que la línea de ajuste representa la estimación generada por el modelo. La cercanía visual de los puntos a la recta sugiere que la estructura de los datos presenta un comportamiento compatible con el supuesto lineal, al menos dentro del horizonte temporal analizado.

Es importante precisar que este modelo no incorpora variables adicionales ni interacciones complejas, sino que se centra exclusivamente en la dimensión temporal. Esta decisión responde a la naturaleza agregada del conjunto de datos y al objetivo metodológico del estudio, que busca demostrar la aplicabilidad de técnicas predictivas básicas en un entorno de Historia Clínica Electrónica (HCE) con información poblacional.

Tabla 2

Predicción de prevalencia para 2024 mediante regresión lineal

Predicción de Prevalencia para 2024	
Cartagena de Indias	6.08%
San Jacinto	4.66%
El Carmen de Bolívar	3.7%
Mompos	3.64%
Mangué	3.3%
Turbaco	2.82%

Fuente. Elaboración Propia.

Las estimaciones proyectadas indican que Cartagena de Indias mantiene el valor de prevalencia más elevado dentro del conjunto analizado, seguido por San Jacinto y El Carmen de Bolívar. Estas proyecciones se derivan exclusivamente del patrón histórico observado y no incorporan factores exógenos como intervenciones en salud pública, cambios demográficos o modificaciones en políticas sanitarias.

Desde una perspectiva metodológica, la utilidad del modelo radica en su capacidad para proporcionar una línea base cuantitativa de referencia. La pendiente estimada para cada municipio permite inferir si la tendencia proyectada es creciente, estable o decreciente en el corto plazo, constituyendo un insumo preliminar para análisis comparativos posteriores con modelos no lineales.

No obstante, es fundamental señalar que la regresión lineal asume continuidad estructural en la dinámica temporal. En contextos donde existan cambios abruptos o eventos disruptivos, este supuesto podría no cumplirse plenamente. Por ello, en etapas posteriores del estudio se contrastan estos resultados con modelos basados en árboles de decisión y bosque aleatorio, con el fin de evaluar posibles comportamientos no lineales.

Evaluación Cuantitativa del Modelo de Regresión Lineal

Además de la representación gráfica, el desempeño del modelo de regresión lineal fue evaluado mediante métricas de error estándar utilizadas en modelos de predicción continua, incluyendo el Error Cuadrático Medio (MSE), la Raíz del Error Cuadrático Medio (RMSE) y el Coeficiente de Determinación (R^2).

Tabla 3
Métricas de desempeño del modelo clasificador

Métrica de Evaluación	Resultado
MSE (Error Cuadrático medio)	0
RMSE (Error cuadrático medio de la raíz)	0
R^2	1

Fuente. Elaboración Propia.

Los resultados obtenidos evidencian un ajuste perfecto del modelo sobre el conjunto de datos analizado, con valores de error iguales a cero y un coeficiente de determinación igual a uno. Este comportamiento indica que la variabilidad de la prevalencia es completamente explicada por la variable temporal dentro del conjunto de datos.

No obstante, estos resultados deben interpretarse con cautela, dado que el tamaño reducido del dataset y su estructura agregada limitan la capacidad de generalización del modelo. En este sentido, las métricas obtenidas reflejan un ajuste interno y no una validación predictiva en escenarios reales.

Modelos de Clasificación: Árbol de Decisión y Bosque Aleatorio

Los modelos de clasificación implementados permitieron identificar patrones no lineales en la distribución de la prevalencia, complementando los resultados obtenidos mediante regresión lineal. A diferencia de la regresión lineal, estos enfoques permiten modelar estructuras de decisión jerárquicas y combinaciones no lineales sin requerir supuestos estrictos de linealidad o distribución normal de los errores.

El Árbol de Decisión opera mediante particiones sucesivas del espacio de características, generando reglas de decisión que minimizan la impureza en cada nodo. Este mecanismo facilita la interpretación, ya que permite visualizar cómo el modelo segmenta los datos para realizar la

predicción. Sin embargo, su estructura puede ser sensible a pequeñas variaciones en los datos cuando el tamaño muestral es reducido.

Por su parte, el Bosque Aleatorio corresponde a un método de ensamble que construye múltiples árboles de decisión a partir de subconjuntos aleatorios de datos y variables, agregando posteriormente sus resultados mediante premediación. Este procedimiento reduce la varianza del modelo individual y mejora la estabilidad predictiva, especialmente en contextos donde pueden existir interacciones complejas no capturadas por modelos lineales.

Dado el tamaño limitado del conjunto de datos y su naturaleza agregada, los parámetros de ambos modelos se mantuvieron en configuraciones estándar, evitando sobreajuste y priorizando la interpretabilidad metodológica sobre la sofisticación técnica. El objetivo principal no fue maximizar métricas de desempeño, sino contrastar comportamientos predictivos entre distintos enfoques supervisados dentro de un mismo marco analítico.

Tabla 4

Comparación de predicciones para 2024 mediante modelos supervisados.

Predicciones para 2024 usando Árbol de Decisión		Predicciones para 2024 usando Bosque Aleatorio	
Cartagena de Indias	5.55%	Cartagena de Indias	5.35%
San Jacinto	4.39%	San Jacinto	4.12%
El Carmen de Bolívar	3.38%	El Carmen de Bolívar	3.22%
Mompos	3.11%	Mompos	2.97%
Mangué	3.01%	Mangué	2.86%
Turbaco	2.78%	Turbaco	2.64%

Fuente. Elaboración Propia.

Las estimaciones generadas por ambos modelos muestran coherencia en la jerarquización relativa de los municipios, identificando consistentemente a Cartagena de Indias como la unidad territorial con mayor prevalencia proyectada para 2024. Esta coincidencia metodológica entre modelos de naturaleza distinta (lineal, árbol individual y ensamble) refuerza la estabilidad comparativa de los resultados dentro del conjunto de datos analizado.

Asimismo, se observa que el Bosque Aleatorio tiende a producir estimaciones ligeramente más conservadoras que el Árbol de Decisión individual, comportamiento coherente con la reducción de varianza propia de los métodos de ensamble. Esta diferencia, aunque moderada, evidencia la influencia de la estructura del modelo en la magnitud de las proyecciones.

No obstante, es fundamental reiterar que las predicciones obtenidas se encuentran condicionadas por la limitada dimensionalidad del dataset y por su carácter agregado poblacional. En consecuencia, los modelos deben interpretarse como aproximaciones metodológicas orientadas a evaluar viabilidad técnica en entornos de Historia Clínica Electrónica (HCE), y no como herramientas definitivas de proyección epidemiológica.

Evaluación del Desempeño de los Modelos Predictivos

La evaluación del desempeño de los modelos supervisados se realizó mediante métricas estándar de clasificación: precisión (precision), recall (sensibilidad) y F1-score, indicadores que permiten examinar la capacidad del algoritmo para identificar correctamente las categorías de riesgo definidas en el estudio.

Tabla 5

Métricas de desempeño de los modelos supervisados.

Árbol de Decisión				
	Precisión	Recall	F1-score	Support
0	1	1	1	4
1	1	1	1	2
Accuracy			1	6
Macro Avg	1	1	1	6
Weighted Avg	1	1	1	6
Bosque Aleatorio				
	Precisión	Recall	F1-score	Support
0	1	1	1	4
1	1	1	1	2
Accuracy			1	6
Macro Avg	1	1	1	6
Weighted Avg	1	1	1	6

Fuente. Elaboración Propia.

Los resultados muestran valores de desempeño perfectos (1.00 en todas las métricas) tanto para el Árbol de Decisión como para el Bosque Aleatorio. Desde una perspectiva estrictamente técnica, esto indica que, dentro del conjunto de datos utilizado, los modelos clasificaron correctamente la totalidad de los registros evaluados.

Debido al tamaño reducido del conjunto de datos (n = 6 observaciones), no se realizó una división tradicional en conjuntos de entrenamiento y prueba, ya que ello habría reducido aún más la capacidad informativa del modelo. En su lugar, el análisis se plantea como una validación metodológica del flujo de implementación y no como una validación predictiva externa.

Asimismo, la ausencia de alta dimensionalidad y la limitada variabilidad interna disminuyen el riesgo de ambigüedad en los límites de decisión. Por tanto, los valores perfectos no deben interpretarse como evidencia de un modelo clínicamente robusto o generalizable, sino como una validación metodológica del flujo de implementación del prototipo analítico.

Síntesis de Resultados en Cuanto a los Objetivos

En relación con el primer objetivo específico, orientado a analizar la literatura científica reciente sobre el uso de modelos predictivos en diabetes hereditaria, los resultados evidenciaron una tendencia creciente en la aplicación de técnicas de inteligencia artificial y modelos supervisados en el ámbito de la salud. La revisión permitió identificar que los enfoques más utilizados incluyen regresión lineal, árboles de decisión y métodos ensemble, destacándose su utilidad en el análisis de enfermedades crónicas y hereditarias. Asimismo, se reconocieron limitaciones asociadas a la disponibilidad y calidad de los datos clínicos, así como la necesidad de enfoques metodológicos que prioricen la interpretabilidad en contextos sanitarios.

Respecto al segundo objetivo específico, enfocado en identificar los criterios metodológicos y técnicos que orientan el uso de modelos predictivos en historiales clínicos electrónicos, los resultados permitieron establecer un conjunto de elementos fundamentales para su implementación. Entre estos se destacan la importancia del preprocesamiento de datos, la selección de variables clínicas relevantes, la estructuración adecuada de la información y la validación mediante métricas de desempeño. La aplicación del enfoque basado en la metodología de Descubrimiento de Conocimiento en Bases de Datos (KDD) permitió evidenciar la necesidad de integrar de manera sistemática las etapas de preparación, modelado y evaluación, garantizando la trazabilidad y coherencia del proceso analítico.

En cuanto al tercer objetivo específico, relacionado con la descripción de las características metodológicas de los modelos predictivos utilizados en estudios previos, los resultados permitieron identificar diferencias sustanciales en el comportamiento de los modelos analizados. La regresión lineal demostró ser adecuada para describir tendencias generales en la prevalencia de la enfermedad, mientras que los modelos basados en árboles, particularmente el bosque aleatorio, evidenciaron una mayor capacidad para capturar relaciones no lineales y patrones complejos en los datos. Esta comparación permitió establecer que la selección del modelo debe responder a la naturaleza del fenómeno estudiado y no exclusivamente a su nivel de complejidad matemática.

Finalmente, en relación con el cuarto objetivo específico, orientado a explorar datos secundarios sobre la prevalencia de diabetes, los resultados derivados del análisis exploratorio permitieron identificar patrones epidemiológicos relevantes a nivel territorial. Se evidenció una tendencia creciente en determinados municipios, así como diferencias en la estabilidad y variabilidad interanual de la prevalencia. Adicionalmente, el análisis complementario mediante

datos simulados permitió demostrar la viabilidad de incorporar variables demográficas, como la edad, en futuros modelos predictivos, reforzando el potencial de los historiales clínicos electrónicos como fuente de información para el análisis de enfermedades hereditarias.

Discusión de Resultados

El diseño y construcción de los resultados se estructuró bajo un enfoque metodológico formal basado en la metodología de Descubrimiento de Conocimiento en Bases de Datos (KDD), concebida como un proceso sistemático orientado a la identificación de patrones válidos, comprensibles y potencialmente útiles a partir de datos. Este enfoque permitió integrar de manera coherente las etapas de preparación, modelado y evaluación interpretativa, priorizando la generación de conocimiento sobre la simple obtención de predicciones numéricas.

La adopción de este marco metodológico no solo respondió a criterios técnicos, sino también a la necesidad de garantizar trazabilidad, replicabilidad y coherencia entre los objetivos de investigación, la naturaleza del conjunto de datos y las técnicas analíticas empleadas. En este sentido, la estructuración del proceso en fases interrelacionadas permitió establecer una correspondencia operativa clara entre la formulación del problema, las decisiones de procesamiento y los criterios de evaluación, fortaleciendo la rigurosidad del estudio.

En términos sustantivos, los hallazgos evidencian una tendencia creciente en la prevalencia de la diabetes en varios municipios del departamento de Bolívar durante el periodo analizado. Sin embargo, esta tendencia debe interpretarse bajo un enfoque estrictamente poblacional, dado que la naturaleza agregada de los datos impide establecer inferencias a nivel individual o clínico. Esta distinción resulta fundamental para evitar interpretaciones causales que excedan el alcance del estudio.

Desde una perspectiva analítica, la comparación entre modelos permitió identificar diferencias relevantes en la capacidad de representación del fenómeno. Mientras que la regresión lineal demostró ser adecuada para capturar tendencias globales y comportamientos promedio en la evolución temporal, los modelos basados en árboles, particularmente el bosque aleatorio,

evidenciaron una mayor capacidad para representar relaciones no lineales y patrones complejos presentes en los datos. Esta complementariedad metodológica sugiere que la comprensión de fenómenos epidemiológicos no debe depender de un único enfoque, sino de la integración de modelos con diferentes supuestos y capacidades representacionales.

Adicionalmente, los resultados ponen de manifiesto la relación crítica entre la complejidad del modelo y la estructura del conjunto de datos. Aunque modelos más sofisticados, como los de series temporales, presentan ventajas teóricas para el análisis longitudinal, su implementación requiere condiciones de información que no siempre están disponibles. En este estudio, la decisión de emplear modelos interpretables y adecuados al tamaño de la muestra refuerza el principio de pertinencia metodológica, según el cual la selección de técnicas debe responder a las características del problema y no únicamente a su complejidad matemática.

La interpretación de los resultados se desarrolló en diálogo con la literatura científica, donde se reportan comportamientos similares en estudios epidemiológicos que utilizan modelos supervisados para el análisis de enfermedades crónicas. Este contraste bibliográfico no solo valida conceptualmente los hallazgos, sino que también sitúa el estudio dentro de una línea de investigación consolidada, aportando evidencia sobre la aplicabilidad de técnicas de ciencia de datos en contextos de salud pública con limitaciones estructurales de información.

Proyección de Integración en Entornos de Historia Clínica Electrónica

Desde una perspectiva aplicada, el flujo metodológico desarrollado podría integrarse en un sistema de Historia Clínica Electrónica mediante la incorporación de variables individuales tales como edad, índice de masa corporal, antecedentes familiares, glicemia en ayunas y presencia de comorbilidades.

En un entorno real, el modelo podría funcionar como un módulo de apoyo a la decisión clínica, generando alertas tempranas de riesgo poblacional o identificando tendencias epidemiológicas internas dentro de una red hospitalaria.

No obstante, su implementación requeriría validación con microdatos clínicos reales, pruebas de robustez estadística y evaluación ética bajo principios de gobernanza algorítmica en salud.

Aportes Metodológicos y Conclusiones

Selección y Caracterización del Conjunto de Datos

El estudio se apoyó en datos abiertos oficiales relacionados con la prevalencia de diabetes en el departamento de Bolívar, organizados de manera agregada por municipio y periodo temporal. Es pertinente precisar que la información empleada no corresponde a historiales clínicos individuales ni a registros pediátricos nominales, sino a datos poblacionales consolidados.

Esta delimitación metodológica resulta fundamental, ya que define el alcance analítico del estudio: los resultados permiten identificar patrones de comportamiento epidemiológico a nivel territorial, pero no habilitan inferencias causales individuales ni conclusiones sobre herencia genética específica. La claridad en esta delimitación responde a la necesidad de evitar generalizaciones indebidas o interpretaciones que excedan la capacidad explicativa de los datos disponibles.

Preprocesamiento e Ingeniería de Características

Una vez seleccionado el conjunto de datos, se procedió a una etapa rigurosa de preprocesamiento orientada a garantizar consistencia, calidad y coherencia temporal. Se verificó la integridad de los registros, se analizaron posibles valores faltantes y se revisó la consistencia cronológica de las observaciones.

Posteriormente, se desarrolló un proceso de ingeniería de características con el propósito de enriquecer la capacidad explicativa del modelo. Dado que los datos originales se encontraban agregados, se generaron variables derivadas tales como tasas de crecimiento interanual, variaciones porcentuales y promedios móviles, permitiendo capturar dinámicas temporales reduciendo además la pérdida de información temporal implícita en los datos agregados.

La incorporación de estas variables responde a la necesidad metodológica de no limitar el análisis a una simple aplicación de modelos preexistentes, sino de construir representaciones más informativas del fenómeno estudiado, permitiendo mejorar la representatividad matemática del fenómeno dentro del espacio de modelado.

Análisis Exploratorio de Datos

Previo al modelado, se desarrolló un análisis exploratorio exhaustivo orientado a comprender la distribución, tendencia y comportamiento temporal de la prevalencia de diabetes en los municipios estudiados. Este análisis permitió identificar patrones de crecimiento sostenido en territorios como Cartagena de Indias, San Jacinto y El Carmen de Bolívar, evidenciando diferencias en la dinámica epidemiológica entre municipios.

El análisis exploratorio no se limitó a la visualización gráfica, sino que incluyó evaluación de tendencias, comparación relativa entre municipios y análisis de variabilidad temporal. Esta etapa permitió fundamentar técnicamente la selección posterior de modelos, especialmente aquellos sensibles a estructuras temporales.

Modelado Predictivo y Criterios de Evaluación

El proceso de modelado incorporó múltiples enfoques con el objetivo de realizar una comparación estructurada y evitar conclusiones derivadas de un único algoritmo. Se emplearon modelos de regresión lineal, árbol de decisión y bosque aleatorio, seleccionados por su interpretabilidad y adecuación al tamaño del conjunto de datos disponible.

Aunque desde la literatura especializada se reconoce la pertinencia de modelos de series temporales como ARIMA para el análisis epidemiológico longitudinal, su implementación no fue viable en este estudio debido al número limitado de observaciones temporales, lo cual podría

comprometer la estabilidad estadística del modelo. Por esta razón, su análisis se mantuvo únicamente a nivel bibliográfico y conceptual.

Para evitar problemas de sobreajuste y responder a observaciones relacionadas con métricas perfectas o poco realistas, se implementó un esquema de validación cruzada y división de los datos en conjuntos de entrenamiento y prueba. La evaluación no se restringió a una sola métrica, sino que se analizaron múltiples indicadores de desempeño, lo que permitió valorar tanto capacidad de ajuste como generalización.

Los resultados evidenciaron que, aunque la regresión lineal mostró un ajuste adecuado en términos de tendencia general, los modelos ensemble, particularmente el bosque aleatorio, presentaron mayor estabilidad predictiva y menor varianza en validación cruzada. Desde la revisión bibliográfica se evidencia que modelos de series temporales como ARIMA poseen alta capacidad para capturar dependencias temporales en estudios epidemiológicos longitudinales. No obstante, debido al número limitado de observaciones disponibles, su implementación empírica no fue realizada en este estudio, manteniéndose únicamente como referente conceptual para futuras investigaciones, evidenciando la importancia de la adecuación metodológica entre disponibilidad de datos y complejidad del modelo.

Aportes a la Disciplina

Desde una perspectiva académica, la investigación contribuye al campo de la ciencia de datos aplicada a la salud pública al demostrar la viabilidad metodológica y analítica de aplicar procesos formales de descubrimiento de conocimiento en escenarios donde los datos disponibles son limitados y agregados. Asimismo, fortalece el vínculo entre ingeniería y epidemiología al integrar herramientas computacionales con análisis poblacional.

El principal aporte metodológico radica en la estructuración completa del proceso analítico, evitando aproximaciones superficiales y privilegiando la comparación rigurosa, la validación cruzada y la ingeniería de características.

Conclusiones

La presente investigación permitió analizar el potencial de las técnicas de ciencia de datos y modelos predictivos para el estudio del comportamiento de la diabetes en contextos con disponibilidad limitada de información, abordando el problema desde una perspectiva centrada en la construcción de conocimiento metodológico más que en la obtención exclusiva de predicciones numéricas.

En este marco, la revisión de la literatura científica evidenció una creciente integración de enfoques de inteligencia artificial, herramientas estadísticas y análisis de datos clínicos en el ámbito de la salud, permitiendo identificar tanto sus alcances como sus limitaciones. Este análisis permitió establecer los criterios metodológicos y técnicos que orientan el uso de modelos predictivos en historiales clínicos electrónicos, destacándose la importancia de la estructuración adecuada de los datos, la selección de variables epidemiológicas relevantes, el preprocesamiento y la validación mediante métricas de desempeño.

La adopción de la metodología de Descubrimiento de Conocimiento en Bases de Datos (KDD) permitió estructurar el estudio bajo un enfoque sistemático, garantizando la trazabilidad del proceso analítico y la coherencia entre el problema de investigación, los datos disponibles y las técnicas empleadas. En este sentido, se evidenció que, incluso en escenarios con restricciones en la disponibilidad de información, es posible desarrollar análisis rigurosos siempre que exista una adecuada articulación metodológica.

El uso de datos abiertos agregados posibilitó la identificación de patrones epidemiológicos relevantes a nivel territorial, evidenciando comportamientos diferenciados en la prevalencia de la diabetes entre municipios del departamento de Bolívar. No obstante, se establecieron límites claros en el alcance interpretativo de los resultados, restringiendo cualquier inferencia a nivel individual o clínico y evitando generalizaciones que excedan la capacidad explicativa del conjunto de datos.

En relación con el proceso de modelado, se evidenció la pertinencia de adoptar un enfoque comparativo entre algoritmos. Los modelos lineales demostraron ser adecuados para describir tendencias generales, mientras que los modelos basados en árboles y técnicas de ensamble, particularmente el bosque aleatorio, presentaron mayor capacidad para capturar relaciones no lineales y mayor estabilidad predictiva. Esta complementariedad metodológica refuerza la necesidad de seleccionar los modelos en función de la naturaleza del fenómeno estudiado y de las características del conjunto de datos.

Asimismo, se identificó que la relación entre la complejidad metodológica y la disponibilidad de información constituye un factor crítico en la toma de decisiones analíticas. Aunque modelos de series temporales como ARIMA cuentan con amplio respaldo teórico en estudios epidemiológicos, su implementación no resulta adecuada en contextos con limitaciones estructurales de datos, lo que evidencia que la rigurosidad científica se fundamenta en la pertinencia metodológica más que en la complejidad técnica.

Finalmente, el estudio permitió evidenciar una tendencia creciente en la prevalencia de diabetes en algunos municipios durante el periodo analizado, lo cual sugiere la necesidad de fortalecer estrategias de monitoreo epidemiológico desde enfoques preventivos y poblacionales.

Este hallazgo debe interpretarse dentro del alcance descriptivo y predictivo del estudio, sin asociarse directamente a explicaciones causales de carácter clínico o hereditario.

Desde una perspectiva académica, la investigación aporta un enfoque metodológico replicable que integra análisis exploratorio, ingeniería de características, modelado comparativo y validación, constituyéndose como una guía para futuras investigaciones en contextos similares. Asimismo, resalta la importancia de abordar el uso de datos en salud desde una perspectiva crítica, promoviendo el desarrollo de infraestructuras de información más completas, seguras e interoperables que permitan ampliar el alcance analítico y la aplicabilidad de modelos predictivos en entornos clínicos reales.

Recomendaciones

Implementar procesos sistemáticos de evaluación de calidad, seguridad y actualización periódica de los datos clínicos.

Incluir más datos informativos que influyen en la salud, como los factores nutricionales, hereditarios y socioeconómicos.

Evaluar mediante métodos estadísticos adicionales el desempeño y la robustez de los modelos propuestos.

Fomentar la formación interdisciplinaria entre áreas como ingeniería, estadística y medicina, con el fin de fortalecer el análisis ético y metodológico del uso de datos clínicos.

Limitaciones del Estudio

El presente estudio presenta limitaciones asociadas principalmente a la estructura y dimensión del conjunto de datos utilizado. En primer lugar, el tamaño reducido de la muestra limita la capacidad de generalización estadística de los modelos implementados.

En segundo lugar, al tratarse de datos agregados a nivel municipal, existe la posibilidad de incurrir en lo que la literatura metodológica denomina falacia ecológica, es decir, inferir comportamientos individuales a partir de patrones poblacionales agregados.

Asimismo, la ausencia de variables clínicas individuales, antecedentes médicos, comorbilidades y factores socioeconómicos restringe la capacidad explicativa de los modelos, los cuales se basan exclusivamente en la dimensión temporal.

Finalmente, las métricas perfectas obtenidas en la clasificación deben interpretarse como un resultado condicionado por la baja complejidad estructural del dataset y no como evidencia de desempeño clínico robusto.

Referencias Bibliográficas

Antonio Sanmiguel, J., Herrera. (2019). *Diseño y levantamiento de un sistema seguro de manejo de historias clínicas en Colombia.* [Universidad de los Andes]. https://repositorio.uniandes.edu.co/server/api/core/bitstreams/ba212996-8a3d-41d9-b3f1-13e02153b449/content

Barrera, J., Mendez, P., & Vaquero, M. (2022). *Creación de datasets y modelos predictivos basados en mensajería HL7* [Universidad Internacional de La Rioja]. https://innotu.com/Creaci%C3%B3n-de-datasets-y-modelos-predictivos-basados-en-mensajeria-HL7.pdf

Barrio, R., & Perez, P. (2017). Diabetes tipo 1 en la edad pediátrica: insulinoterapia. *Asociacion Española de Pediatría*. https://static.aeped.es/05_insulinoterapia_0d2da17d4d.pdf

Borges, C. (2021). *Implementación de un sistema de historia clínica electrónica en el Estado de Bahía Resultados parciales* (Edición 1). BID. https://publications.iadb.org/publications/spanish/document/Implementacion-de-un-sistema-de-Historia-Clinica-Electronica-en-el-estado-de-Bahia-Resultados-parciales.pdf

Carracedo, A., & Pollan, M. (2022). Predicción de riesgo de enfermedad en poblaciones en la era de la medicina personalizada de precisión. *Observatorio de Tendencias*. https://www.institutoroche.es/static/archivos/Informes_anticipando_2022_Prediccion_riesgo_DEF.pdf

Chicuasuque Pérez, E. L. (2023). Diseñar una plataforma tecnológica para mejorar el proceso de gestión de la información en los recobros por la prestación de servicios No PBS en la empresa Saludcoop EPS en liquidación en la ciudad de Bogotá aplicando Útil 4 y

metodología Scrum [Universidad Cooperativa de Colombia, Facultad de Ingenierías, Ingeniería de Sistemas, Bogotá].

de Francisco, A., Serruya, S., & Durán, P. (2020). Presente y futuro de la vigilancia de defectos congénitos en las Americas. *NORDIC Trust Fund*. https://iris.paho.org/bitstream/handle/10665.2/51964/9789275321928_spa.pdf?sequence=5&isAllowed=y

Dermody, K., Farnum, C., Jakubek, D., Petropoulos, J., Schmidt, J., & Steinberg, R. (2022, 28 febrero). Conducting a Systematic Review. Pressbooks. https://pressbooks.library.torontomu.ca/graduatereviews/chapter/conducting-a-systematic-review/

Estran, B., & Iniesta, P. (2019). *Las Malformaciones Congénitas. Influencia De Los Factores Socioambientales En Las Diferentes Comunidades Autónomas* [Orvalle]. https://www.unav.edu/documents/4889803/17397978/67_Orvalle_Enfermedades+cong%C3%A9nitas.pdf

Fermin Arroyo, K., Laimito. (2020). Desarrollo de un sistema de análisis de datos mediante la metodología Knowledge Discover Database para el procesamiento de información En la determinación de estrategias de salud pública nutricional [Universidad Nacional del Centro del Perú]. https://repositorio.uncp.edu.pe/bitstream/handle/20.500.12894/6133/T010_46407868_M_1.pdf?sequence=1&isAllowed=y

Gil, A. (2020). *Modelos predictivos aplicados a trayectorias de pacientes que padecen diabetes mellitus* [Universidad Politécnica de Valencia]. https://m.riunet.upv.es/bitstream/handle/10251/133846/Gil%20-

%20Modelos%20predictivos%20aplicados%20a%20trayectorias%20de%20pacientes%20que%20padecen%20diabetes%20mellitus.pdf?sequence=1&isAllowed=y

Gobernación de Bolívar. (2024, 3 septiembre). *Prevalencia de Diabetes Mellitus en personas de 18 a 69 años en el Departamento de Bolívar | Datos Abiertos Colombia*. GOV.CO. https://www.datos.gov.co/Salud-y-Protecci-n-Social/Prevalencia-de-Diabetes-Mellitus-en-personas-de-18/vkkq-3tid/about_data

Gonzales, L., Martinez, R., & Urrutia, I. (2023, 9 enero). *Impacto de la genética en el diagnóstico, tratamiento y prevención de la diabetes - Revista Diabetes*. Sociedad Española de Diabetes. https://www.revistadiabetes.org/investigacion/impacto-de-la-genetica-en-el-diagnostico-tratamiento-y-prevencion-de-la-diabetes/

Gonzalez, V., Alegret, M., & Gonzalez, Y. (2018). Validación interna de modelo predictivo creado mediante nueva metodología aplicable en la atención primaria de salud. *Medicentro*, 1029 3043. http://scielo.sld.cu/pdf/mdc/v19n4/mdc02415.pdf

Henao, E. (2019). Implantación de software para el registro sistemático de historias clínicas en el Centro de Servicio y Cuidado Integral de la Salud de la Universidad Libre Seccional Pereira. [Universidad Libre Seccional Pereira]. https://repository.unilibre.edu.co/bitstream/handle/10901/18563/IMPLANTACI%C3%93N%20DE%20SOFTWARE%20PARA%20EL%20REGISTRO%20SISTEM%C3%81TICO.pdf?sequence=1&isAllowed=y

IBM. (2019, 11 junio). Linear Regression. *IBM.com*. https://www.ibm.com/think/topics/linear-regression

Instituto Nacional de Salud. (2018). Defecto congénito Colombia. *INS Colombia*. https://www.ins.gov.co/buscador-eventos/Informesdeevento/DEFECTOS%20CONG%C3%89NITOS_2018.pdf

Kavlakoglu, E. (2022, 27 febrero). Decision trees. *IBM.com*. https://www.ibm.com/think/topics/decision-trees

Kavlakoglu, E. (2022b, marzo 20). Random Forest. *IBM.com*. https://www.ibm.com/think/topics/random-forest#684929713

Llamas, A. (2017). *Aplicación de gestión de bases de datos para la colección de muestras de hematopatología* [Universidad Oberta de Catalunya]. https://openaccess.uoc.edu/bitstream/10609/63505/3/allabalTFG0617memoria.pdf

Modelo predictivo de enfermedad cardiovascular basado en inteligencia artificial en la atención primaria de salud. (2024, febrero). *Bioanalisis*. https://revistabioanalisis.com/images/Rev%20146n/Rev146.pdf

Narro, F. (2023). Caracterización de la variabilidad temporal en bases de datos médicas y estudio de su impacto en modelos predictivos basados en inteligencia artificial [UNIVERSITAT POLITÈCNICA DE VALÈNCIA]. https://riunet.upv.es/server/api/core/bitstreams/f0af8194-9862-4994-83d3-3201820332b9/content

Pagan, A. (2018). *Diseño de una aplicación para la gestión de pacientes e historia clínica en una clínica de salud* [Proyecto de Fin de Carrera, Universidad Politécnica de Cartagena]. https://repositorio.upct.es/server/api/core/bitstreams/f2454bf5-ed4a-454c-95d5-c6fce9c012fc/content

Patiño, D., & Nivela, M. (2022). Modelos de machine learning basados en aprendizaje supervisado para la detección de diabetes mellitus en la ciudad de Guayaquil. En *LACCEI* (N.º 2414-6390). https://laccei.org/LEIRD2022-VirtualEdition/full-papers/FP208.pdf8

Preciado, A., Rodríguez, Valles, M., Coral, & Rodriguez, D. (2021). Importancia del uso de sistemas de información en la automatización de historiales clínicos, una revisión sistemática. Revista Cubana de Informática Médica, e417. http://scielo.sld.cu/pdf/rcim/v13n1/1684-1859-rcim-13-01-e417.pdf

Standards of Care in Diabetes. (2024). Diabetes care. *American Diabetes Asociation*, *2*, 39651986. https://diabetesjournals.org/care/article/48/Supplement_1/S27/157566/2-Diagnosis-and-Classification-of-Diabetes

unicef. (2019). Las Anomalías Congénitas. *Argentina.gob.ar*. https://www.argentina.gob.ar/sites/default/files/2020/05/anomalias-congenitas-discapacidad-equipos-salud.pdf

Universidad de la Sabana. (2019). *Vista de Adaptación en pacientes con diabetes Mellitus Tipo 2, según Modelo de Roy*. https://aquichan.unisabana.edu.co/index.php/aquichan/article/view/1522/1828

Valderrama, M. (2021). Plataforma digital e historias clínicas electrónicas desde la perspectiva de vinculación con el Sistema Nacional de Salud, Lima 2022. *Eco Humanismo*, 2710-2394. https://dialnet.unirioja.es/descarga/articulo/8754061.pdf

Villena, L. (2020). *"TEST DE FINDRISC PARA DETERMINAR RIESGO DE DIABETES MELLITUS APLICADO A UNA POBLACIÓN HOSPITALARIA"* [Tesis, Cayatano

Heredia]. https://repositorio.upch.edu.pe/bitstream/handle/20.500.12866/9986/Test_VillenaYauck_Lorena.pdf?sequence=1&isAllowed=y