Diseño e Implementación de una Aplicación para el Diagnóstico y Detección Temprana de Diabetes Tipo 2

Sonya Valentina Castro Gomez
Ingeniería de Sistemas
Universidad del Norte
Barranquilla, Colombia
sonyac@uninorte.edu.co

Jeffrey Andres Felix Carvajal
Ingeniería de Sistemas
Universidad del Norte
Barranquilla, Colombia
felixj@uninorte.edu.co

Dario Jose Mejia Caballero
Ingeniería de Sistemas
Universidad del Norte
Barranquilla, Colombia
jdario@uninorte.edu.co

Wilson Nieto Bernal
Ingeniería de Sistemas
Universidad del Norte
Barranquilla, Colombia
wnieto@uninorte.edu.co

Eduardo David Angulo Madrid
Ingeniería de Sistemas
Universidad del Norte
Barranquilla, Colombia
edangulo@uninorte.edu.co

***Keywords*—Diabetes, Machine Learning, Diseases, Diagnostic**

I. Introducción

La diabetes es una enfermedad crónica que afecta a millones de personas en todo el mundo, con un aumento preocupante en su prevalencia, según datos de la Organización Mundial de la Salud (OMS) (OMS, 2024). Entre 1980 y 2014, el número de personas con diabetes aumentó de 108 millones a 422 millones, con un crecimiento más pronunciado en los países de ingresos bajos y medianos (OMS, 2024). Esta enfermedad conlleva riesgos graves para la salud, como ceguera, insuficiencia renal, infarto de miocardio, accidente cerebrovascular y amputación de miembros inferiores. Además, las tasas de mortalidad por diabetes han aumentado en un 3% entre 2000 y 2019, causando dos millones de defunciones en 2019, según las últimas estadísticas de la OMS (2024).

Por otro lado, en los últimos años, el avance acelerado de la tecnología ha generado un impacto significativo en diversos sectores, transformando la manera en que se llevan a cabo las operaciones y procesos administrativos en organizaciones públicas y privadas. Esta revolución tecnológica ha encontrado un campo de innovación en el ámbito del Machine Learning (ML), una rama de la Inteligencia Artificial (IA) que se centra en crear algoritmos y modelos para que las computadoras aprendan de los datos y realicen predicciones o juicios de forma autónoma (Gracious et al., 2023).

En el contexto de la salud, ML ha demostrado ser una herramienta poderosa que puede mejorar significativamente la precisión y eficiencia en el diagnóstico y tratamiento de enfermedades. Uno de los avances más significativos ha sido el desarrollo del Machine Learning (ML). Esta técnica permite a las computadoras aprender de los datos y realizar predicciones o juicios de manera autónoma (Gracious et al., 2023; Hamsagayathri & Vigneshwaran, 2021)

En particular, ML juega un papel fundamental en el desarrollo de sistemas de Apoyo a la Toma de Decisiones Clínicas (CDSS); estos sistemas ayudan a los a los profesionales de la salud a tomar decisiones informadas basadas en evidencia (Gracious et al., 2023). Desempeñando un papel crucial en la predicción de enfermedades, la estratificación de riesgos, la recomendación de tratamientos y el monitoreo en tiempo real, contribuyendo así a una atención más personalizada y eficaz para los pacientes (Gracious et al., 2023).

Estos modelos de ML han sido implementados para diversos propósitos, tal como se describe en (Azar et al., 2015), donde se desarrolla un modelo capaz de clasificar o predecir trastornos mentales basándose en una serie de síntomas ingresados a través de texto. En este caso, se emplea una variante del algoritmo genético. Del mismo modo, Mangal and Jain (2022) desarrollaron un modelo capaz de predecir la enfermedad de la diabetes utilizando algoritmos de Random Forest para clasificar los síntomas médicos de los pacientes, logrando precisiones del 99%.

Este proyecto tiene como objetivo desarrollar un modelo de diagnóstico de la diabetes tipo 2 utilizando principalmente técnicas avanzadas de ML y algunas de DL. Al aprovechar la capacidad del ML para aprender de datos pasados y DL para extraer características complejas de los datos, se busca crear una herramienta que pueda ayudar a los profesionales de la salud en la detección temprana y precisa de esta enfermedad. Adicionalmente, se busca que el modelo final sea implementado en un prototipo de aplicación web interactiva que permitirá a los usuarios ingresar una serie de síntomas relacionados con la diabetes tipo 2 y recibir una evaluación de riesgo de padecer dicha enfermedad basada en el análisis de

datos clínicos y médicos. Con esto se busca que la herramienta contribuya a una atención médica más efectiva y personalizada, facilitando la detección temprana y el tratamiento oportuno de la diabetes tipo 2 .

II. Definición del problema

El campo de la ingeniería de sistemas enfrenta el reto de aplicar avances tecnológicos como el Machine Learning (ML) y el Deep Learning (DL) para resolver problemas complejos en diversos sectores. En el ámbito de la salud, estos avances prometen revolucionar el proceso diagnóstico y terapéutico, ofreciendo herramientas precisas para la toma de decisiones basadas en datos. El problema de investigación que se aborda es: ¿Cómo puede una página web, apoyada en técnicas de ML y DL, mejorar la detección temprana y el tratamiento de la diabetes tipo 2, una enfermedad crónica que representa uno de los mayores desafíos para la salud pública mundial, y en qué medida puede asistir a los profesionales de la salud en la toma de decisiones informadas y basadas en la evidencia? Este problema refleja la necesidad de una solución computacional tangible que se materialice en un prototipo de página web interactivo donde los usuarios puedan ingresar síntomas y recibir una evaluación de riesgo basada en datos clínicos y médicos.

Para ilustrar las diferentes dimensiones de este desafío y cómo se interrelacionan en el contexto de nuestro proyecto, se ha desarrollado el siguiente Árbol del Problema (ver Figura 1). Este diagrama visual detalla la jerarquía de los factores implicados en la detección y manejo de la diabetes tipo 2, desde la recopilación y análisis de datos hasta el desarrollo de modelos predictivos y su implementación en una solución de software accesible.

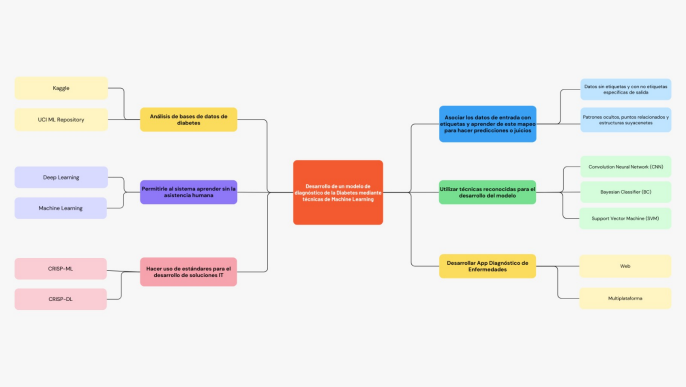

Fig. 1. Árbol del Problema

El desarrollo de Sistemas de Apoyo a la Decisión Clínica (CDSS) se presenta como una solución prometedora en este contexto. Estudios como el de Gracious et al. (2023) destacan la importancia de estas herramientas que asisten en la toma de decisiones médicas minimizando los errores de diagnóstico y tratamiento, sin embargo, reconocen que estos sistemas no pueden reemplazar a los profesionales de la salud sino complementar su juicio clínico.

En este sentido, ML y DL se consolidan como una poderosa herramienta que mejora significativamente la precisión y eficiencia en el diagnóstico y tratamiento de enfermedades. Su capacidad para aprender de los datos y realizar predicciones o juicios de manera autónoma ha sido un avance significativo en el desarrollo de los CDSS, como lo señalan Hamsagayathri and Vigneshwaran (2021), quienes describen cómo estos sistemas pueden desempeñar un papel crucial en la predicción de enfermedades, la estratificación de riesgos, la recomendación de tratamientos y el monitoreo en tiempo real, contribuyendo a una atención más personalizada y efectiva para los pacientes

Además, el uso de técnicas de aprendizaje supervisado como la regresión lineal, Support Vector Machine (SVM), Random Forest (RF) y Decision Tree (DT), aplicadas en la salud para la categorización de enfermedades y predicción de riesgos, demuestra la aplicabilidad y relevancia de ML en el diagnóstico médico, lo cual respalda la factibilidad de la propuesta del proyecto.

En resumen, este proyecto plantea un reto multidisciplinario: integrar los avances de ML y DL en una plataforma web que pueda transformar la calidad de la atención médica en el tratamiento de la diabetes tipo 2. Este desafío no solo requiere una comprensión profunda de las tecnologías involucradas sino también una evaluación cuidadosa de su implementación y efectividad en entornos reales, lo que implica la posibilidad de realizar pruebas empíricas para observar su impacto en la realidad única y objetiva.

III. Justificación

Las técnicas de Machine Learning proveen herramientas para asistir a los practicantes de medicina en la prevención, diagnóstico y tratamiento (Arun Bhavsar et al., 2021; Caballé-Cervigón, Castillo-Sequera, Gómez-Pulido, Gómez-Pulido, & Polo-Luque, 2020). Un diagnóstico erróneo puede llevar a negar el cuidado necesario del paciente o a la aplicación errónea (Arun Bhavsar et al., 2021). Esto puede causar complicaciones en la salud del paciente además de aumentar los gatos médicos (Arun Bhavsar et al., 2021). Bajo este contexto, las herramientas de diagnóstico ML buscan reducir los errores humanos, mejorar los servicios médicos y reducir el costo y tiempo de estos (Arun Bhavsar et al., 2021; Alowais et al., 2023). Sin embargo, estas herramientas aún están en las etapas iniciales de integración a la práctica clínica (Alowais et al., 2023). Además, a pesar de la inmensa disponibilidad de datos disponibles, los modelos tienden a ser lineales con un rango reducido de variables (Caballé-Cervigón et al., 2020)

El presente proyecto propone una solución de diagnóstico basado en síntomas para asistencia de personal médico con un enfoque a la enfermedad de la diabetes tipo 2. Este proyecto busca crear un modelo para usarse mediante una interfaz sencilla y amigable con el usuario que permita ingresar una serie de síntomas relacionados con la diabetes tipo 2 para la obtención de un posible diagnóstico.

IV. Objetivos

A. Objetivo General

Diseñar e implementar una aplicación web basada en Machine Learning para diagnosticar y detectar tempranamente la Diabetes tipo 2, utilizando datos del Behavioral Risk Factor Surveillance System (BRFSS) de 2015. Esta herramienta estará diseñada para facilitar la identificación de individuos en riesgo de desarrollar esta condición, basándose en análisis predictivos avanzados.

B. Objetivos Específicos

1) Realizar una revisión sistemática de los avances en el uso de Machine Learning y Deep Learning para la detección y diagnóstico temprano de la Diabetes tipo 2, evaluando la efectividad de estas tecnologías en el análisis de grandes conjuntos de datos de salud.
2) Desarrollar la arquitectura de la solución para procesar el conjunto de los datos BRFSS 2015 asociados, identificando los factores de riesgo y patrones asociados con la diabetes tipo 2.
3) Implementar una aplicación web que integre el modelo predictivo para ofrecer una solución práctica que ayude en la detección temprana de la Diabetes tipo 2. La aplicación proporcionará una interfaz intuitiva para que los usuarios (investigadores, posiblemente profesionales de la salud) puedan fácilmente ingresar datos y recibir evaluaciones de riesgo.
4) Validar la precisión del modelo y la funcionalidad de la aplicación a través de pruebas que involucren un análisis detallado de los resultados y comparaciones con diagnósticos conocidos, para asegurar que la herramienta es efectiva en identificar correctamente a individuos en riesgo.

V. Metodología

Para el desarrollo del presente trabajo, se utilizará la metodología de desarrollo CRISP-ML (Q) (Cross-Industry Standard Process model for the development of Machine Learning applications with Quality assurance) (Studer et al., 2021). Inspirado en CRISP-DM (minería de datos), CRISP-ML(Q) es un modelo iterativo (por lo cual un paso anterior es fundamental proceso) que añade una capa de monitoreo y de calidad en cada etapa y tareas (Studer et al., 2021).

A. Comprensión de los datos

Studer et al. (2021) define las siguientes tareas de esta fase:

1) Definir el alcance de la aplicación ML.
2) Definir los criterios de éxito.
3) Definir la factibilidad.
4) Colectar los datos.
5) Verificar la calidad de los datos.
6) Revisar resultados de la fase.

Para esta primera fase, se busca comprender el problema a resolver y realizar la búsqueda de datos para definir los objetivos del proyecto. Para ello, los datos se buscarán en plataformas abiertas como *Kaggle*.

B. Preparación de los datos

Esta fase provee de un dataset para la siguiente fase (entrenamiento) y, si es necesario, se puede volver a esta etapa en cualquier momento; no es estática (Studer et al., 2021). Studer et al. (2021) define las tareas de esta fase a continuación:

1) Limpieza de datos.
2) Selección de variables (*feature selection*).
3) Selección de datos.
4) Resolver el desequilibrio de clase.
5) Construcción de datos (*feature engineering*).
6) Estandarización de datos.

Para la limpieza de datos, se resolverán los problemas de datos faltantes y errores de formato en caso de llegar a encontrarse. También se representarán las variables categóricas utilizando representación binaria y se normalizarán los datos utilizando un *standard scalar*. Los datos se dividirán en 3 subconjuntos: datos de entrenamiento (70%), datos de validación (20%) y datos de evaluación (10%). Con estos datos se pasará a la fase de entrenamiento para realizar una prueba inicial. Luego, se regresará a esta fase para realizar cambios y mejorar el modelo.

Para la selección de variables se utilizará la matriz de correlación para identificar las variables que expliquen mejor la variable objetivo (presentar o no diabetes). De igual forma, se buscará una relación entre variables presentes con el fin de intentar reducir el número de variables de entrada en los modelos.

En el caso del desequilibrio de clase, se realizarán pruebas sin resolver el desequilibrio, luego resolviendo con *undersampling* y finalmente resolviendo con *oversampling*. Haciendo esto, se busca encontrar qué método da resultados más favorables. De igual manera, se realizarán pruebas con datos duplicados y no duplicados. Finalmente, antes de pasar a la fase de entrenamiento, se configurarán los *hiperparámetros* de cada modelo.

C. Entrenamiento

Studer et al. (2021) menciona que la técnica de modelado dependerá de los objetivos y datos disponibles para el desarrollo de la aplicación. El conjunto de entrenamiento será usado para entrenar el modelo utilizando diferentes algoritmos:

- Ridge logistic regression
- Naive-bayes Classifier
- Decision tree
- Random Forest
- MLP classifier

D. Evaluación del modelo

Tomando los datos de validación y evaluación, se evaluará el rendimiento de cada uno de los algoritmos utilizados para su posterior comparación. Primero se hará la validación y si esta es desfavorable, se ajustarán los hiperparámetros del modelo y se regresará a la fase de modelado. Si la validación es favorable, se evaluará el modelo con los datos de evaluación.

Para la medición del rendimiento del modelo, se medirá su exactitud (*accuracy*), su precisión, *recall* (sensibilidad o tasa de verdaderos positivos), *F1-Score* y el *roc-auc*.

E. Despliegue o implementación

Esta fase se encarga de llevar el modelo a un uso práctico en un campo de aplicación Studer et al. (2021). Para esta fase, se llevará el modelo a una aplicación web que permita a los usuarios finales utilizar el modelo de forma sencilla e intuitiva.

F. Monitoreo y mantenimiento

Con el propósito de evaluar la usabilidad y la experiencia del usuario en la aplicación web, se llevará a cabo una encuesta dirigida a diferentes usuarios de la aplicación, con un tamaño mínimo de 25 participantes. Esta encuesta tiene como objetivo determinar qué tan intuitiva y fácil de utilizar es la aplicación, además de evaluar la percepción de los participantes sobre la fiabilidad del modelo.

VI. Marco Conceptual

A. Diabetes

La diabetes puede ser clasificada en varias categorías. La **diabetes tipo 1** es generada por la destrucción auto-inmune de las células β, causando deficiencia de insulina (ElSayed et al., 2022). La **diabetes tipo 2** ocurre a una perdida progresiva (no auto-inmune) de segregación de insulina las células β (ElSayed et al., 2022). Además de los anteriores, se incluye la diabetes gestacional, que se diagnostica en el segundo o tercer trimestre de embarazo, y otros tipos de diabetes ocurren por otras causas, como la diabetes inducida por químicos, diabetes neonatal, entre otras (ElSayed et al., 2022). La **prediabetes** no es una entidad clínica por su cuenta, sino un factor de riesgo para denominar a individuos cuyos niveles de glucosa no entran en la categoría de diabetes pero tienen un metabolismo anormal de clorhidratos que puede progresar en diabetes y enfermedades cardiovasculares (ElSayed et al., 2022).

B. Sistemas de apoyo a la decisión clínica

Los sistemas de apoyo a la decisión clínica (CDSS por sus siglas en ingles Clinical Decision Support Systems) son utilizados para asistir a los practicantes de salud en la toma de decisiones, tomando como base evidencias obtenidas mediante técnicas de ML (Machine Learning) y DL (Deep Learning) (Gracious et al., 2023). El uso de Ml en la asistencia de toma de decisiones médicas, se busca minimizar los errores de diagnóstico y tratamiento (Arun Bhavsar et al., 2021). Sin embargo, estos sistemas no pueden remplazar a los practicantes de salud, solo asistirlos en la toma de decisiones (Arun Bhavsar et al., 2021).

C. Machine Learning

El Machine Learning tiene como objetivo la creación de algoritmos que puedan utilizar datos y utilizarlos para aprender automáticamente (sin asistencia humana) a hacer predicciones y juicios (Gracious et al., 2023; Hamsagayathri & Vigneshwaran, 2021). El principal objetivo es crear un sistema con la capacidad aprender y mejorar sin la asistencia humana (Hamsagayathri & Vigneshwaran, 2021). A diferencia de los acercamientos convencionales, en el cual se programan el modelo en su totalidad, ML permite aprender de experiencias pasadas para mejorar (Arun Bhavsar et al., 2021). Para ello, el modelo aprende del mapeo de predictores (características de entrada) y el objetivo (variable salida), y lo generaliza para funcionar con datos no observados (Arun Bhavsar et al., 2021). ML tiene un gran impacto en diferentes áreas (Park et al., 2021) y aplicaciones donde juega un rol vital, desde procesamiento de lenguaje natural hasta diagnóstico de enfermedades (Hamsagayathri & Vigneshwaran, 2021). Existen diferentes técnicas de ML que pueden ser agrupadas en aprendizaje supervisado, no supervisado y por refuerzo (Gracious et al., 2023; Dahiwade, Patle, & Meshram, 2019; Ahsan, Luna, & Siddique, 2022).

El aprendizaje supervisado asociar los datos de entrada con las etiquetas y aprende este mapeo para hacer predicciones o juicios (Gracious et al., 2023). Técnicas bajo la categoría de aprendizaje supervisado incluyen regresión linear (LR), Support Vector Machine (SVM), Random Forest (RD) y Decision Tree (DT), que son utilizadas en el área de la salud para la categorización de enfermedades, predicción de riesgos, entre otros (Gracious et al., 2023).

El aprendizaje no supervisado aprende de datos de entada no etiquetados sin una salida en particular para buscar patrones ocultos, agrupar datos relacionados y descubrir estructuras subyacentes (Gracious et al., 2023). Técnicas de aprendizaje no supervisado tienes la capacidad de agrupar basado en características similares, con esto, se puede mejorar el tratamiento y terapias personalizadas (Gracious et al., 2023). Técnicas bajo la categoría de aprendizaje no supervisado incluyen el agrupamiento k-medias, agrupamiento jerárquico y modelo de mezcla gaussiana (Gracious et al., 2023).

El aprendizaje por refuerzo se basa en aprender con fallo y error para obtener retroalimentación (en formas de recompensas o penalizaciones) para maximizar las recompensas (Gracious et al., 2023). En el área de la salud, se utiliza para la optimización de terapia y tratamiento adaptativo (Gracious et al., 2023).

Además de los acercamientos anteriores, también se habla del aprendizaje profundo (Deep learning; DL) como una rama de ML (Gracious et al., 2023). En DL, se modelan relaciones complejas y se extraen características de alto nivel utilizando redes neuronales con muchas capas (Gracious et al., 2023). Estas redes pueden aprender jerarquías de datos permitiendo capturar patrones y conexiones complejas (Gracious et al., 2023).

1) Bayesian Classifier (BC) y Naïve Bayes (NB)

BC es uno de los algoritmos de clasificación más poderos (Hamsagayathri & Vigneshwaran, 2021). El BC utiliza técnicas de modelado probabilísticas para la representación de variables y dependencias condicionales entre ellas, permitiendo crear modelos para hacer predicciones rápidas (Arun Bhavsar et al., 2021; Hamsagayathri & Vigneshwaran, 2021). Entre los clasificadores bayesianos, se destaca Naïve

Bayes (NB) (Arun Bhavsar et al., 2021; Hamsagayathri & Vigneshwaran, 2021).

2) Random Forests (RF)

RF es un algoritmo de clasificación que contine arboles de decisión (Hamsagayathri & Vigneshwaran, 2021).

D. Medida de rendimiento de modelos

Algunas de las medidas de rendimiento incluyen la exactitud (Accuracy), la precisión (P), el Recall (R) y el F1-Score (F1) que son ampliamente usadas en el diagnostico de enfermedades (Ahsan et al., 2022).

1) Matriz de confusión

Verdadero Positivo (VP) ocurre cuando el modelo predice correctamente un diagnóstico positivo basado en los síntomas y ese diagnóstico es realmente positivo. Falso Positivo (FP) ocurre cuando el modelo predice incorrectamente un diagnostico positivo basado en los síntomas, pero en realidad el diagnóstico es negativo. Verdadero Negativo (VN) ocurre cuando el modelo predice correctamente un diagnóstico negativo basado en los síntomas y ese diagnóstico es realmente negativo. Falso Negativo (VF) ocurre cuando el modelo predice incorrectamente un diagnóstico negativo basado en los síntomas, pero en realidad el diagnóstico es positivo.

2) Exactitud

La exactitud (E) se define como se define como la proporción de predicciones correctas sobre el total de predicciones realizadas (Ahsan et al., 2022).

$$E = \frac{VP + VN}{VP + FP + VN + FN} \tag{1}$$

3) precisión

La precisión (P) es la proporción de verdadero positivos (instancias correctamente clasificadas como positivas) sobre el total de instancias clasificadas como positivas (Ahsan et al., 2022).

$$E = \frac{VP}{VP + FP} \tag{2}$$

4) Recall

El recall (R) es la proporción de verdaderos positivos sobre el total de instancias que son realmente positivas (Ahsan et al., 2022).

$$R = \frac{VP}{VP + FN} \tag{3}$$

5) F1-Score

El F1-Score combina la precisión y el recall (Ahsan et al., 2022).

$$F1 = \frac{P \times R}{P + R} \tag{4}$$

VII. Estado del arte

A. Revisión sistemática de la literatura

Criterios De Búsqueda

1) Machine Learning & Diseases & Symptoms
2) Deep Learning & Diseases & Symptoms
3) Machine Learning & Diabetes
4) Machine Learning & Diagnostic & Diabetes

TABLE I: Criterios de búsqueda

Criterios	Fuente				
	IEEE	Zu Scholars	Nature	MDPI	Springer
ML & Diseases	2,826	14	3050	264	62,549
DL & Diseases	1,774	32	8026	151	100,212
ML & Diabetes	2,618	3	3440	420	41,620
ML & Diagnostic & Diabetes	395	14	1540	34	19,507

B. Antecedentes

En un estudio realizado por Tan citado por (Hamsagayathri & Vigneshwaran, 2021) se consideró la idea de unir 2 algoritmos de Machine Learning, el algoritmo genético y SVM, como una estrategia hibrida. En este se utilizó LIBSVM y weka, una herramienta de extracción de datos. Para llevar al cabo dicho estudio se hizo uso de dos sets de datos enfocados en enfermedades de diabetes y enfermedades cardiacas tomados de UC Irvine ML repository. De la evaluación de este modelo hibrido para la diabetes se obtuvo una precisión del 84.07% y para las enfermedades del corazón se obtuvo una precisión del 78.26% (Hamsagayathri & Vigneshwaran, 2021) de lo cual se destacó la correcta clasificación, así como la disminución del sobre ajuste en los datos, sin embargo, el costo computacional que requiere el modelo es alto y por ende el modelo trabaja de manera lenta (Hamsagayathri & Vigneshwaran, 2021).

Por otra parte, Fathima realizo un estudio para la predicción de la enfermedad del dengue haciendo uso del modelo de ML SVM. Para la toma de datos se hizo uso de diversas encuestas de varios hospitales y laboratorios de Chennai y Tiruelveli en la india, así como datos obtenidos del instituto king de medicina preventiva (Hamsagayathri & Vigneshwaran, 2021). De estos datos se utilizaron 28 características de 5006 muestras. La precisión alcanzada de este estudio por el modelo SVM fue de 0.9043 (Hamsagayathri & Vigneshwaran, 2021). En un estudio reciente, se ha demostrado que el uso del big data en el sector de la salud ha permitido realizar predicciones más precisas y descubrir patrones ocultos (Silahtaroğlu & Yılmaztürk, 2019). En este se propone un modelo de prediagnóstico de aprendizaje automático para departamentos de emergencia, el cual utiliza las quejas verbales de los

pacientes como datos de entrada. Este modelo alcanza una precisión mínima del 75.5% y se basa en dos enfoques de aprendizaje automático: Red neuronal probabilística y Árbol de decisión de bosque aleatorio. La información utilizada para este estudio fue proporcionada por un hospital privado en Estambul, Turquía. Todos los registros son anónimos y están relacionados con el departamento de emergencias. Los datos contienen las quejas de los pacientes y el diagnóstico final de cada caso, con un total de 5,394 registros después de la limpieza necesaria (Silahtaroğlu & Yılmaztürk, 2019). Este estudio destaca la importancia de utilizar datos generados por los usuarios para mejorar la precisión en la predicción de diagnósticos y proporcionar un mejor soporte en la toma de decisiones en los departamentos de emergencia.

En un artículo de la revista Scientific Reports, se describe un modelo de red neuronal (DL) entrenado con 88 parámetros diferentes, que incluyen 86 características de pruebas de laboratorio, sexo y edad, con el fin de asistir a los médicos en el diagnóstico de enfermedades (Park et al., 2021). Para validar este algoritmo DL, se llevó a cabo una validación cruzada estratificada de cinco veces y se empleó el criterio TOP5 (cinco enfermedades más probables) para evaluar el modelo. Además, se evaluó el rendimiento de cada modelo utilizando puntuaciones F1, dado un desequilibrio en el número de 39 enfermedades. Las puntuaciones F1 y la precisión del modelo DL fueron del 80% y 91%, respectivamente. Para las cinco categorías de enfermedades más comunes, la precisión y el recuerdo fueron del 77% y el 87%, respectivamente (Park et al., 2021).

También en una investigación dirigida a mejorar la precisión de los modelos de predicción del riesgo de enfermedad, especialmente en el caso del infarto cerebral, se exploró la combinación de modelos para obtener resultados más efectivos en comparación con los enfoques convencionales. Para ello, se utilizaron diversos algoritmos de aprendizaje automático que abordaron tanto datos estructurados como no estructurados, obtenidos de registros hospitalarios (Chen, Hao, Hwang, Wang, & Wang, 2017). Los datos estructurados comprendían información de laboratorio y datos básicos del paciente, como la edad, el género y los hábitos de vida. Mientras tanto, los datos no estructurados incluían la narración del paciente sobre su enfermedad, los registros de interrogatorio del médico y los diagnósticos, entre otros. Para los datos estructurados, se aplicaron algoritmos tradicionales como Naive Bayes (NB), K-Nearest Neighbors (KNN) y Decision Trees (DT) para predecir el riesgo de infarto cerebral. Además, se propuso un nuevo algoritmo de predicción de riesgo de enfermedad multimodal basado en una red neuronal convolucional (CNN-MDRP), que integraba datos estructurados y no estructurados. Se destacó que este enfoque proporcionaba una precisión de predicción del 94.8%, superando la velocidad de convergencia de otros modelos unimodales basados en CNN (Chen et al., 2017). A partir de estas investigaciones, se concluyó que la precisión de la predicción del riesgo de enfermedad depende de la diversidad y la calidad de las características de los datos hospitalarios. Por lo tanto, se propuso la combinación de datos estructurados y no estructurados como una estrategia efectiva para mejorar la precisión en la predicción del riesgo de infarto cerebral y posiblemente otras enfermedades complejas (Chen et al., 2017).

En una investigación previa, se exploró un enfoque de predicción general de enfermedades basado en los síntomas del paciente. Este estudio hizo uso de algoritmos de aprendizaje automático como K-Nearest Neighbor (KNN) y redes neuronales convolucionales (CNN) para llevar a cabo la predicción. Para este propósito, se utilizó un conjunto de datos que comprendía síntomas de enfermedades, junto con información sobre los hábitos de vida y los resultados de los chequeos médicos para mejorar la precisión de la predicción (Dahiwade et al., 2019). Los resultados obtenidos revelaron que la precisión de la predicción general de enfermedades mediante el uso de CNN fue del 84,5%, superando al algoritmo KNN. Además, se observó que el algoritmo KNN requería más tiempo y memoria en comparación con CNN. Después de la predicción de la enfermedad general, el sistema desarrollado en este estudio pudo proporcionar información sobre el riesgo asociado, que variaba según la enfermedad detectada (Dahiwade et al., 2019).

TABLE II: Comparación de la precisión de diferentes modelos por tema y autores

Autores	Tema	Modelo	Precisión
Tan (citado por Gracious et al. (2023))	Predicción de enfermedades del corazón.	SVM y Algoritmo genético	84.07%
	Predicción de enfermedades de diabetes.	SVM	78.26%
Fathima (citado por Hamsagayathri and Vigneshwaran (2021))	Predicción de la enfermedad del dengue	SVM	90.43%
Silahtaroğlu and Yılmaztürk (2019)	Prediagnóstico según quejas de los pacientes	Random forest	77.65%
		Deep Learning Model	91.0%

Chen et al. (2017)	Predicción de enfermedades (infarto cerebral)	convolutional neural network based multimodal disease risk prediction. (CNN-MDRP)	94.8%
Dahiwade et al. (2019)	Predicción de enfermedades basado en los síntomas del paciente	Convolutional neural network (CNN)	84.5%

VIII. Arquitectura

A. Arquitectura de la solución

La primera etapa de la arquitectura se centra en la adquisición de datos, los cuales están disponibles en Kaggle en formato CSV. La siguiente fase aborda la ingestión de estos datos, donde se lleva a cabo la transferencia de los datos desde Kaggle a BigQuery, un servicio de almacenamiento en la nube de Google, para su posterior utilización en las fases subsiguientes. En la fase de transformación de datos, se preparan y procesan los datos, se realiza el entrenamiento de los modelos y se evalúan para su ajuste. Una vez que se ha obtenido un modelo satisfactorio, se exporta para su consumo en la fase siguiente. Por último, en la etapa final, el modelo creado se emplea para desarrollar una API utilizando FastAPI en Python y React, todo ello orquestado eficientemente mediante Docker.

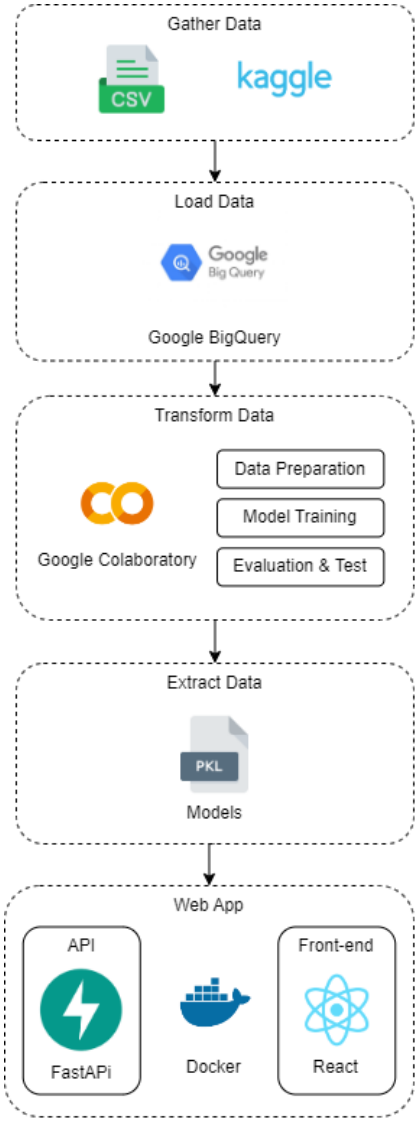

Fig. 2. Arquitectura de la Solución

B. Obtención del datos

Los datos están disponibles de forma abierta en Kaggel *Diabetes Health Indicators Dataset*.

En las variables binarias, 0 representa no y 1 representa si.

TABLE III: Descripción del dataset

Variable	Rol	Tipo	Descripción
ID	ID	Integer	ID del paciente
Diabetes _binary	Target	Binario	0 = No diabetes 1 = Pre-diabetes o diabetes
HighBP	Feature	Binario	¿Tiene presión arterial alta?
HighChol	Feature	Binario	¿Tiene colesterol alto?

CholCheck	Feature	Binario	¿Se ha realizado un chequeo de colesterol en 5 años?
BMI	Feature	Integer	Índice de masa corporal
Smoker	Feature	Binario	¿Ha fumado al menos 100 cigarros en su vida entera?
Stroke	Feature	Binario	¿(Alguna vez le dijeron) que tuvo un derrame cerebral?
HeartDiseaseorAttack	Feature	Binario	¿Tiene enfermedad coronaria o infarto de miocardio?
PhysActivity	Feature	Binario	¿Ha realizado actividad física en los últimos 30 días?(Sin incluir el trabajo)
Fruits	Feature	Binario	¿Consume fruta 1 o más veces al día?
Veggies	Feature	Binario	¿Consume verduras 1 o más veces al día?
HvyAlcoholConsump	Feature	Binario	¿Bebedores empedernidos? (hombres adultos que toman más de 14 tragos por semana y mujeres adultas que toman más de 7 tragos por semana)
AnyHealthcare	Feature	Binario	¿Tener cualquier tipo de cobertura de atención médica, incluidos seguros médicos, planes prepagos como HMO, etc.?
NoDocbcCost	Feature	Binario	¿Hubo algún momento en los últimos 12 meses en el que necesitó consultar a un médico pero no pudo debido al costo?
GenHlth	Feature	Integer	Diría usted que en general su salud es: escala 1-5 1 = excelente 2 = muy buena 3 = buena 4 = regular 5 = mala
MentHlth	Feature	Integer	Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? scale 1-30 days
PhysHlth	Feature	Integer	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? scale 1-30 days
DiffWalk	Feature	Binario	Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes
Sex	Feature	Binario	Sex 0 = female 1 = male
Age	Feature	Integer	Age 13-level age category (_AGEG5YR see codebook) 1 = 18-24 9 = 60-64 13 = 80 or older
Education	Feature	Integer	Education Level Education level (EDUCA see codebook) scale 1-6 1 = Never attended school or only kindergarten 2 = Grades 1 through 8 (Elementary) 3 = Grades 9 through 11 (Some high school) 4 = Grade 12 or GED (High school graduate) 5 = College 1 year to 3 years (Some college or technical school) 6 = College 4 years or more (College graduate)
Income	Feature	Integer	Income Income scale (INCOME2 see codebook) scale 1-8 1 = less than $ 10,000 5 = less than $35,000 8 = $75,000 or more

C. Transformación de datos

Para la fase de transformación de datos, se extraerán los datos de BigQuery para llevar a cabo la preparación de los mismos. En esta etapa, se llevarán a cabo diversas tareas, que incluyen la limpieza de los datos, la selección de variables (feature selection), la búsqueda de soluciones para resolver el desequilibrio de clases, la realización de ingeniería de características y la estandarización de los datos. Los datos preparados se dividirán en tres subconjuntos: datos de entrenamiento (70%), datos de validación (20%) y datos de prueba (10%).

Posteriormente, se procederá a configurar los hiperparámetros de cada tipo de modelo y se iniciará el proceso de entrenamiento. Después de esta fase, se realizará la validación y, en caso de no obtener resultados satisfactorios, se ajustarán los hiperparámetros o se podrá incluso regresar a la fase de preparación. Si los resultados de la validación son positivos,

se avanzará a la fase de evaluación. En caso contrario, se podrá volver a la fase de entrenamiento o preparación según sea necesario. En el caso de obtener resultados favorables, se concluye la fase de transformación de datos.

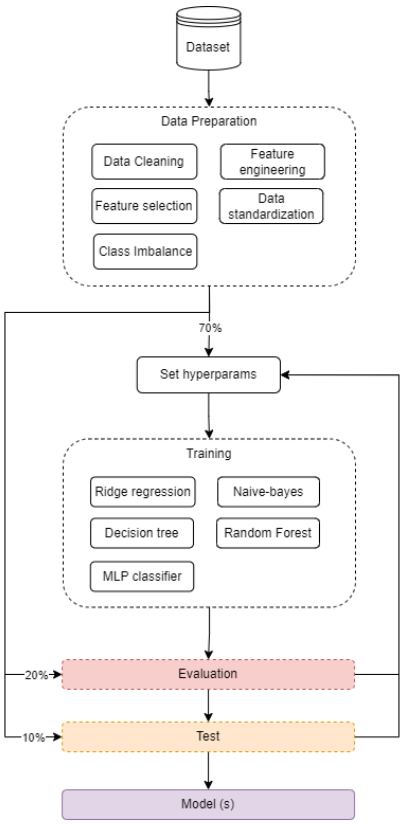

Fig. 3. Arquitectura de entrenamiento

REFERENCES

Ahsan, M. M., Luna, S. A., & Siddique, Z. (2022, March). Machine-learning-based disease diagnosis: A comprehensive review. *Healthcare*, *10*(3), 541. doi: 10.3390/healthcare10030541

Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., … Albekairy, A. M. (2023, September). Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Medical Education*, *23*(1). doi: 10.1186/s12909-023-04698-z

Arun Bhavsar, K., Abugabah, A., Singla, J., Ali AlZubi, A., Kashif Bashir, A., & Nikita. (2021). A comprehensive review on medical diagnosis using machine learning. *Computers, Materials amp; Continua*, *67*(2), 1997–2014. doi: 10.32604/cmc.2021.014943

Azar, G., Gloster, C., El-Bathy, N., Yu, S., Neela, R. H., & Alothman, I. (2015, May). Intelligent data mining and machine learning for mental health diagnosis using genetic algorithm.
doi: 10.1109/eit.2015.7293425

Caballé-Cervigón, N., Castillo-Sequera, J. L., Gómez-Pulido, J. A., Gómez-Pulido, J. M., & Polo-Luque, M. L. (2020, July). Machine learning applied to diagnosis of human diseases: A systematic review. *Applied Sciences*, *10*(15), 5135. doi: 10.3390/app10155135

Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. *IEEE Access*, *5*, 8869–8879. doi: 10.1109/access.2017.2694446

Dahiwade, D., Patle, G., & Meshram, E. (2019, March). Designing disease prediction model using machine learning approach. In *2019 3rd international conference on computing methodologies and communication (iccmc).* IEEE. doi: 10.1109/iccmc.2019.8819782

ElSayed, N. A., Aleppo, G., Aroda, V. R., Bannuru, R. R., Brown, F. M., Bruemmer, D., … Association, A. D. (2022, 12). 2. Classification and Diagnosis of Diabetes: Standards of Care in Diabetes—2023. *Diabetes Care*, *46*(Supplement$_1$), $S19 - S40$.

Gracious, L. A., Jasmine, R. M., Pooja, E., Anish, T., Johncy, G., & Subramanian, R. S. (2023, October). Machine learning and deep learning transforming healthcare: An extensive exploration of applications, algorithms, and prospects. In *2023 4th ieee global conference for advancement in technology (gcat).* IEEE. doi: 10.1109/gcat59970.2023.10353476

Hamsagayathri, P., & Vigneshwaran, S. (2021, February). Symptoms based disease prediction using machine learning techniques. In *2021 third international conference on intelligent communication technologies and virtual mobile networks (icicv).* IEEE. doi: 10.1109/icicv50876.2021.9388603

Mangal, A., & Jain, V. (2022). Performance analysis of machine learning models for prediction of diabetes. In *2022 2nd international conference on innovative sustainable computational technologies (cisct)* (p. 1-4). doi: 10.1109/CISCT55310.2022.10046630

OMS. (2024). *Diabetes.* World Health Organization. Retrieved from `https://www.who.int/es/news-room/fact-sheets/d`

Park, D. J., Park, M. W., Lee, H., Kim, Y.-J., Kim, Y., & Park, Y. H. (2021, April). Development of machine learning model for diagnostic disease prediction based on laboratory tests. *Scientific Reports*, *11*(1). doi: 10.1038/s41598-021-87171-5

Silahtaroğlu, G., & Yılmaztürk, N. (2019, June). Data analysis in health and big data: A machine learning medical diagnosis model based on patients' complaints. *Communications in Statistics - Theory and Methods*, *50*(7), 1547–1556. doi: 10.1080/03610926.2019.1622728

Studer, S., Bui, T. B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., & Müller, K.-R. (2021, April). Towards crisp-ml(q): A machine learning process model with quality assurance methodology. *Machine Learning and Knowledge Extraction*, *3*(2), 392–413. doi:

10.3390/make3020020

IX. Anexos

A. Tabla de revisión Sistemática de la literatura

TABLE IV: Revisión sistemática de la literatura

No.	Titulo	Palabras Clave	Fuente
1	Machine Learning and Deep Learning Transforming Healthcare: An Extensive Exploration of Applications, Algorithms, and Prospects	Deep learning, Surveys, Drugs, Decision support systems, Ethics, Data privacy, Precision medicine	IEEE
2	Symptoms Based Disease Prediction Using Machine Learning Techniques	Heart, Solid modeling, Decision making,Machine learning, Diabetes, Medical diagnostic imaging, Diseases	IEEE
3	A comprehensive review on medical diagnosis using machine learning	Diagnostic system, Healthcare applications, Machine learning, medical diagnosis	Zu Scholars
4	Development of machine learning model for diagnostic disease prediction based on laboratory tests	Experimental models of disease, Information theory and computation, Machine learning	Nature
5	Designing Disease Prediction Model Using Machine Learning Approach	Diseases, Prediction algorithms, Machine learning algorithms, Machine learning, Classification algorithms, Biomedical monitoring, Cloud computing	IEEE
6	Machine Learning Applied to Diagnosis of Human Diseases: A Systematic Review	human disease, machine learning, data mining, artificial intelligence, big data	MDPI
7	Revolutionizing healthcare: the role of artificial intelligence in clinical practice	AI, Healthcare, Patient care, Quality of life, Clinicians, Decision-making, Personalized treatment plans	Spinger
8	Data analysis in health and big data: A machine learning medical diagnosis model based on patients' complaints	Machine learning, text mining, patients' complaintsdiagnosis	Taylor & Francis
9	Disease Prediction by Machine Learning Over Big Data From Healthcare Communities	Diseases, Hospitals, Prediction algorithms, Machine learning algorithms, Big Data, Data models, Big data analytics, machine learning, healthcare	IEEE
10	Machine Learning-based Diabetes Prediction: A Cross-Country Perspective	Machine learning algorithms;Predictive models;Prediction algorithms;Boosting;Data models;Diabetes;Decision trees;Diabetes;Machine Learning Models;Missing values;Outliers	IEEE
11	Performance analysis of machine learning models for prediction of diabetes	Machine learning algorithms;Computational modeling;Forestry;Predictive models;Prediction algorithms;Medical tests;Diabetes;Predictive Analysis;Machine Learning;Diabetes Prediction;Random Forest;Logistic Regression	IEEE