

Hybrid Variational Autoencoder vs XGBoost for Diabetes Mellitus Prediction: A Latent Space-Based Approach

Darwin Patiño-Pérez, Ph.D.¹ ; Ángel Ochoa-Flores, MSc¹ ; Juan Cedeño-Rodríguez, MSc¹ 
Liliana Sarmiento-Barreiro, MSc² ; Celia Munive-Mora, BS³ 

¹Universidad de Guayaquil, Facultad de Ciencias Matemáticas y Física, Ecuador, darwin.patinop@ug.edu.ec,
angel.ochoaf@ug.edu.ec, juan.cedenor@ug.edu.ec,

²Universidad de Guayaquil, Facultad de Ciencias Médicas, Ecuador, liliana.sarmientob@ug.edu.ec,

³St Luke's University Hospital Network, PA, United States, celia.munive@sluhn.org

Abstract— Diabetes mellitus is a leading cause of morbidity and mortality worldwide, making its early prediction a public health priority. This study compares the performance of Extreme Gradient Boosting (XGBoost) and a Hybrid Variational Autoencoder (Hybrid VAE) for diabetes classification, evaluating both their predictive accuracy and clinical interpretability. Using the scikit-learn diabetes dataset (442 samples, 10 clinical variables), two models were implemented: XGBoost with hyperparameter optimization and a Hybrid VAE with an 8-dimensional latent space designed to learn interpretable representations of underlying physiological factors. Accuracy, precision, recall, F1-score, and AUC-ROC were assessed, along with latent space analysis using PCA. The Hybrid VAE outperformed XGBoost in all evaluated metrics: accuracy (73.03% vs. 69.66%), recall (79.55% vs. 70.45%), F1-score (0.7447 vs. 0.6966), and AUC-ROC (0.8045 vs. 0.7702). Latent space analysis revealed a natural separation between diabetic and non-diabetic patients in the principal components, with a cumulative explained variance of 64.0%. The importance of features in XGBoost identified body mass index (BMI) and serum S5 measurement as the most relevant predictors. The Hybrid VAE demonstrates superior performance to XGBoost in diabetes prediction, combining high predictive accuracy with the added advantage of an interpretable latent space that captures the underlying structure of the disease. This hybrid approach represents a promising alternative for clinical applications where both accuracy and understanding of the underlying mechanisms are critical.

Keywords— Diabetes Mellitus, Latent Spaces, Variational Autoencoder, XGBoost, Machine Learning.

I.

Hybrid Variational Autoencoder vs XGBoost para la Predicción de Diabetes Mellitus: Un Enfoque Basado en Espacios Latentes

Darwin Patiño-Pérez, Ph.D.¹; Ángel Ochoa-Flores, MSc¹; Juan Cedeño-Rodríguez, MSc¹

Liliana Sarmiento-Barreiro, MSc²; Celia Munive-Mora, BS³

¹Universidad de Guayaquil, Facultad de Ciencias Matemáticas y Física, Ecuador, darwin.patinop@ug.edu.ec, angel.ochoaf@ug.edu.ec, juan.cedenor@ug.edu.ec,

²Universidad de Guayaquil, Facultad de Ciencias Médicas, Ecuador, liliana.sarmientob@ug.edu.ec,

³St Luke's University Hospital Network, PA, United States, celia.munive@sluhn.org

Resumen— La diabetes mellitus representa una de las principales causas de morbilidad y mortalidad a nivel global, haciendo de su predicción temprana una prioridad en salud pública. Este estudio compara el rendimiento de Extreme Gradient Boosting (XGBoost) y un Hybrid Variational Autoencoder (Hybrid VAE) para la clasificación de diabetes, evaluando tanto su precisión predictiva como su interpretabilidad clínica. Utilizando el conjunto de datos de diabetes de scikit-learn (442 muestras, 10 variables clínicas), se implementaron dos modelos: XGBoost con optimización de hiperparámetros y un Hybrid VAE con espacio latente de 8 dimensiones diseñado para aprender representaciones interpretables de factores fisiológicos subyacentes. Se evaluó accuracy, precisión, recall, F1-score y AUC-ROC, junto con análisis de espacios latentes mediante PCA. El Hybrid VAE superó a XGBoost en todas las métricas evaluadas: accuracy (73.03% vs 69.66%), recall (79.55% vs 70.45%), F1-score (0.7447 vs 0.6966) y AUC-ROC (0.8045 vs 0.7702). El análisis del espacio latente reveló una separación natural entre pacientes diabéticos y no diabéticos en las componentes principales, con una varianza explicada acumulada del 64.0%. La importancia de características de XGBoost identificó el índice de masa corporal (bmi) y la medición sérica s5 como los predictores más relevantes. El Hybrid VAE demuestra un rendimiento superior al de XGBoost en la predicción de diabetes, combinando alta precisión predictiva con la ventaja adicional de un espacio latente interpretable que captura la estructura subyacente de la enfermedad. Este enfoque híbrido representa una alternativa prometedora para aplicaciones clínicas donde tanto la exactitud como la comprensión de los mecanismos subyacentes son fundamentales.

Palabras clave— Diabetes Mellitus, Espacios Latentes, Variational Autoencoder, XGBoost, Aprendizaje Automático.

I INTRODUCCIÓN

La diabetes mellitus constituye una enfermedad metabólica crónica de proporciones epidémicas, afectando a más de 500 millones de adultos en todo el mundo y proyectándose como una de las principales causas de muerte prematura para la próxima década [1],[2]. Caracterizada por niveles crónicamente elevados de glucosa sanguínea según Fig. 1, esta condición resulta de defectos en la secreción o acción de la insulina, conduciendo a complicaciones microvasculares y macrovasculares que deterioran significativamente la calidad de vida y aumentan la mortalidad [3].

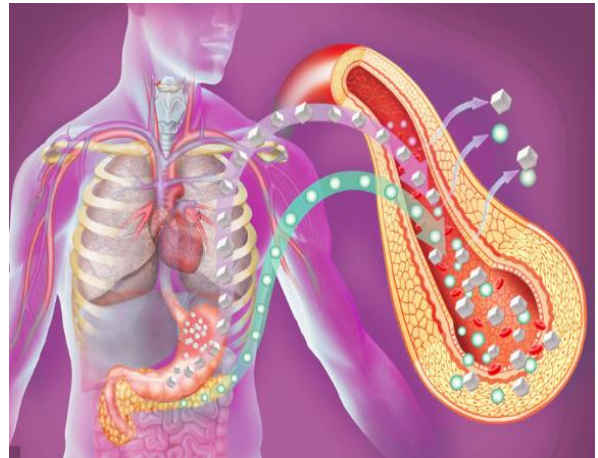


Fig. 1 Diabetes Mellitus

La identificación temprana de individuos en riesgo de desarrollar diabetes es fundamental para implementar intervenciones preventivas efectivas y estrategias de manejo que puedan modificar la trayectoria natural de la enfermedad [4]. En este contexto, el aprendizaje automático ha emergido como una herramienta poderosa para desarrollar modelos predictivos que aprovechan datos clínicos disponibles para estratificar el riesgo de manera precisa y automatizada [5].

Entre los algoritmos de aprendizaje automático, XGBoost (Extreme Gradient Boosting) se ha consolidado como uno de los métodos más efectivos para problemas de clasificación con datos tabulares [6]. Este algoritmo de ensemble construye secuencialmente árboles de decisión donde cada nuevo árbol corrige los errores residuales de los anteriores, incorporando regularización para prevenir sobreajuste y manejar eficientemente interacciones complejas entre variables [7]. Su éxito en competencias de ciencia de datos y aplicaciones del mundo real lo ha posicionado como un estándar de referencia para tareas predictivas.

Paralelamente, los avances en aprendizaje profundo han dado lugar a modelos generativos como los autoencoders variacionales (VAEs), que no solo realizan predicciones, sino que aprenden representaciones comprimidas de los datos en espacios latentes de baja dimensionalidad [8]. Estos espacios

latentes capturan las variables generativas subyacentes que explican la estructura observada en los datos originales, ofreciendo una ventana hacia los factores causales no medidos directamente [9]. En aplicaciones médicas, esta capacidad resulta particularmente valiosa, ya que el espacio latente puede corresponder a constructos fisiológicos como sensibilidad a insulina, función de células beta, o perfiles metabólicos [10].

El concepto de espacios latentes ha revolucionado el análisis de datos biomédicos al permitir representaciones donde puntos cercanos en el espacio corresponden a pacientes con perfiles fisiológicos similares [11]. Esta propiedad facilita la identificación de subgrupos de pacientes, la visualización de trayectorias de progresión de enfermedades, y la generación de hipótesis sobre mecanismos subyacentes [12]. Sin embargo, los VAEs estándar, al ser no supervisados, no garantizan que las dimensiones latentes sean relevantes para tareas específicas como la clasificación de enfermedades [13].

Para abordar esta limitación, los modelos híbridos que combinan componentes generativos y supervisados han ganado atención [14]. Un Hybrid VAE integra un *decoder* para reconstrucción de datos con un clasificador que opera sobre el espacio latente, forzando a que las representaciones aprendidas sean simultáneamente útiles para la generación de datos y la predicción de etiquetas [15]. Este enfoque produce espacios latentes que mantienen interpretabilidad mientras optimizan el rendimiento predictivo.

En este estudio, presentamos un análisis comparativo entre XGBoost, representante de los métodos de ensemble tradicionales, y un Hybrid VAE, representante de los modelos generativos profundos, para la tarea de predicción de diabetes mellitus. Utilizando el conjunto de datos de diabetes de scikit-learn, evaluamos ambos modelos en términos de métricas de clasificación estándar y analizamos la estructura del espacio latente aprendido por el Hybrid VAE. Nuestros objetivos específicos son: (1) determinar qué modelo ofrece mayor precisión predictiva, (2) evaluar si el espacio latente del Hybrid VAE proporciona información interpretable sobre la fisiopatología de la diabetes, y (3) discutir las implicaciones de cada enfoque para aplicaciones clínicas.

II. MATERIALES Y MÉTODOS

A. Marco Teórico

1) Diabetes Mellitus: Fundamentos Fisiopatológicos

La diabetes mellitus tipo 2, que constituye la forma predominante de la enfermedad, se caracteriza por una compleja interacción entre resistencia a la insulina y disfunción de las células beta pancreáticas [16]. La resistencia a la insulina en tejidos periféricos —músculo esquelético, hígado y tejido adiposo— obliga a las células beta a incrementar la secreción de insulina para mantener la homeostasis glucémica [17]. Con el tiempo, esta sobrecarga conduce al agotamiento celular y a una producción insuficiente de insulina, resultando en hiperglucemia manifiesta [18].

Los factores de riesgo establecidos para diabetes tipo 2 incluyen obesidad (especialmente de distribución central), sedentarismo, edad avanzada, antecedentes familiares de diabetes, hipertensión arterial, y dislipidemia [19]. Estudios epidemiológicos han identificado biomarcadores predictivos como la glucosa en ayunas, hemoglobina glicosilada (HbA1c), triglicéridos, y el índice de masa corporal (IMC) como los más relevantes [20]. El conjunto de datos utilizado en este estudio incorpora mediciones que reflejan estos factores, incluyendo IMC, presión arterial, y perfiles lipídicos, proporcionando una base sólida para la modelización predictiva.

2) Extreme Gradient Boosting (XGBoost).

El XGBoost representa una implementación optimizada y altamente eficiente del algoritmo de gradient boosting, que ha demostrado un rendimiento excepcional en numerosas tareas de aprendizaje automático [21]. El gradient boosting construye un modelo predictivo mediante la combinación aditiva de aprendices débiles, típicamente árboles de decisión, donde cada nuevo modelo se entrena para corregir los errores residuales de la combinación de modelos anteriores [22]. Formalmente, para un conjunto de datos con n muestras, el modelo XGBoost produce predicciones $\hat{y}_i$ para cada instancia i como (1)

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), f_k \in F \quad (1)$$

donde F representa el espacio de funciones de árbol y cada f_k corresponde a un árbol con estructura independiente y pesos asociados a sus hojas [23].

La función objetivo que XGBoost minimiza durante el entrenamiento incorpora tanto un término de pérdida como un componente de regularización (2)

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \quad (2)$$

con $\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2$, donde T es el número de hojas, w son los pesos de las hojas, y γ y λ son parámetros de regularización que controlan la complejidad del modelo [7].

Las características distintivas de XGBoost incluyen: a) manejo incorporado de valores faltantes mediante dirección de aprendizaje automático, b) regularización L1 y L2 para prevenir sobreajuste, c) soporte para entrenamiento paralelo que acelera significativamente el cómputo, y d) capacidad de proporcionar importancias de características basadas en la frecuencia de uso y ganancia promedio [24].

3) Autoencoders Variacionales y Espacios Latentes

Los autoencoders variacionales (VAEs) pertenecen a la familia de modelos generativos profundos que aprenden representaciones probabilísticas de datos de alta dimensión en espacios latentes de menor dimensionalidad [25]. A diferencia

de los autoencoders deterministas, los VAEs modelan explícitamente la incertidumbre en la codificación, tratando el espacio latente como una distribución de probabilidad en lugar de un punto fijo [9].

La arquitectura de un VAE comprende dos componentes principales: un encoder que aproxima la distribución posterior $p(z|x)$ mediante una distribución $q_\phi(z|x)$ parametrizada, y un decoder que modela la verosimilitud de los datos $p_\theta(x|z)$ a partir de muestras del espacio latente [26]. El entrenamiento optimiza la cota inferior de la evidencia (3).

$$\text{ELBO}(\theta, \phi; x) = E_{z \sim q_\phi(z|x)} [\log p_\theta(x|z)] - \text{KL}(q_\phi(z|x) \parallel p(z)) \quad (3)$$

El primer término fomenta una reconstrucción precisa de los datos, mientras que el segundo término —la divergencia Kullback-Leibler entre la distribución posterior aproximada y una distribución prior (típicamente Gaussiana estándar)— actúa como regularizador, incentivando que el espacio latente sea continuo y completo [27].

El concepto de espacio latente es fundamental en este contexto. Formalmente, un espacio latente es una representación de datos de baja dimensionalidad donde las entradas originales se proyectan mediante transformaciones no lineales [28]. Las propiedades deseables de un espacio latente bien formado incluyen:

- Continuidad: Puntos cercanos en el espacio latente deben decodificar a instancias similares en el espacio original.
- Completitud: La mayoría de las regiones del espacio latente deben decodificar a instancias válidas.
- Regularidad: La interpolación entre puntos debe producir transiciones suaves y significativas [12].

En aplicaciones biomédicas, estas propiedades permiten que el espacio latente capture factores fisiológicos subyacentes, facilitando la visualización de la heterogeneidad de enfermedades y la identificación de subtipos clínicos [10].

4) Hybrid Variational Autoencoder (Hybrid VAE).

El Hybrid VAE propuesto en este trabajo extiende el VAE estándar mediante la incorporación de un componente supervisado de clasificación, creando una arquitectura que aprende representaciones latentes optimizadas simultáneamente para la reconstrucción de datos y la predicción de etiquetas [14]. Este enfoque responde a la necesidad de modelos que combinen precisión predictiva con interpretabilidad clínica [15].

La arquitectura del Hybrid VAE implementada consta de:

1. Encoder: Red neuronal con dos capas ocultas de 64 neuronas y activación ReLU, seguidas de capas Dropout (0.2) para regularización. Produce los parámetros de la distribución posterior: media μ y log-varianza $\log \sigma^2$.

2. Espacio Latente: Dimensión de 8 variables continuas, seleccionada empíricamente para balancear compresión y retención de información. Cada dimensión tiene el potencial de corresponder a factores fisiológicos latentes como sensibilidad a insulina, función de células beta, o perfil lipídico.

3. Decoder: Red neuronal simétrica que reconstruye los datos originales a partir de muestras del espacio latente, utilizando activaciones lineales en la capa de salida para permitir valores continuos en el rango $[-1, 1]$.

4. Clasificador: Red neuronal de 32 y 16 neuronas con activaciones ReLU y Dropout (0.3), culminando en una capa con activación sigmoide que produce la probabilidad de pertenencia a la clase diabética. La función de pérdida combinada integra tres términos (4).

$$L_{\text{total}} = L_{\text{recon}} + \beta \cdot L_{\text{KL}} + L_{\text{class}} \quad (4)$$

donde:

- $L_{\text{recon}} = \text{MSE}(x, x_{\text{recon}})$: error cuadrático medio de reconstrucción
- $L_{\text{KL}} = -\frac{1}{2} \sum (1 + \log \sigma^2 - \mu^2 - \sigma^2)$: divergencia KL
- $L_{\text{class}} = \text{BCE}(y, \hat{y})$: entropía cruzada binaria para clasificación
- β : hiperparámetro que controla el balance entre reconstrucción y regularización ($\beta = 0.1$ en este estudio)

Este diseño híbrido presenta ventajas significativas: (a) el espacio latente está forzado a contener información relevante para la clasificación, (b) la interpretabilidad se mantiene al poder visualizar y analizar las representaciones latentes, y (c) la capacidad generativa permite explorar escenarios contrafácticos mediante interpolación en el espacio latente [29].

B. Conjunto de Datos

Se utilizó el conjunto de datos de diabetes incluido en la librería *scikit-learn*, originalmente recopilado para estudiar la progresión de la diabetes un año después del *baseline* [30].

Este *dataset* contiene 442 muestras de pacientes con 10 variables predictoras *baseline*:

- age: Edad del paciente (estandarizada)
- sex: Sexo del paciente (estandarizado)
- bmi: Índice de masa corporal (estandarizado)
- bp: Presión arterial promedio (estandarizada)
- s1-s6: Mediciones séricas de lipoproteínas (estandarizadas)

La variable objetivo original es una medida cuantitativa de progresión de la enfermedad. Para este estudio de clasificación binaria, se binarizó la variable objetivo utilizando un umbral de 140 mg/dL, valor clínicamente relevante que distingue entre normoglucemia e hiperglucemia [31]. Esta transformación resultó en una distribución perfectamente

balanceada: 221 pacientes clasificados como no diabéticos (50.0%) y 221 como diabéticos (50.0%).

Las estadísticas descriptivas muestran que las variables están estandarizadas, con rangos entre -0.138 y 0.199, facilitando el procesamiento por algoritmos de aprendizaje automático sensibles a la escala.

C. Preprocesamiento de Datos

Se aplicaron estrategias de preprocesamiento específicas para cada modelo:

Para XGBoost: Se utilizó StandardScaler para estandarizar las características a media cero y varianza unitaria, práctica recomendada para algoritmos basados en árboles cuando se combinan con regularización [32].

Para Hybrid VAE: Se empleó MinMaxScaler con rango [-1, 1] para normalizar los datos. Esta elección facilita la convergencia del modelo al mantener las activaciones en rangos lineales de las funciones de activación y estabilizar el entrenamiento [33].

La división de datos se realizó de manera estratificada para mantener la proporción de clases en todos los subconjuntos:

- Entrenamiento: 264 muestras (59.7%)
- Validación: 89 muestras (20.1%)
- Prueba: 89 muestras (20.1%)

D. Implementación de Modelos

1) *Configuración de XGBoost.* El modelo XGBoost se implementó utilizando la librería xgboost con los siguientes hiperparámetros, optimizados mediante validación cruzada preliminar:

- objective: 'binary:logistic' (clasificación binaria)
- n_estimators: 100 árboles
- max_depth: 6 (profundidad máxima de árboles)
- learning_rate: 0.1 (tasa de aprendizaje)
- subsample: 0.8 (fracción de muestras por árbol)
- colsample_bytree: 0.8 (fracción de características por árbol)
- eval_metric: 'logloss'

Se realizó validación cruzada estratificada de 5 folds para evaluar la robustez del modelo y obtener estimaciones más confiables del rendimiento esperado en datos no vistos [34].

2) *Configuración de Hybrid VAE.* El Hybrid VAE se implementó utilizando PyTorch [35] con la arquitectura descrita en el apartado 4) del marco teórico y los siguientes hiperparámetros de entrenamiento:

- latent_dim: 8 (dimensionalidad del espacio latente)
- hidden_dim: 64 (neuronas en capas ocultas)
- dropout_rate: 0.2 (encoder/decoder), 0.3 (clasificador)
- learning_rate: 1e-3 (tasa de aprendizaje inicial)
- batch_size: 16 (tamaño de lote)
- epochs: 100 (máximo de épocas)
- beta: 0.1 (ponderación de KL divergence)
- optimizer: Adam [26]

- scheduler: ReduceLROnPlateau (factor 0.5, patience 10)
- early_stopping: Paciencia de 20 épocas basada en pérdida de validación.

Se implementó gradient clipping con max_norm=1.0 para prevenir explosión de gradientes y mejorar la estabilidad del entrenamiento [36].

E. Métricas de Evaluación

Se emplearon las métricas derivadas de la matriz de confusión para evaluar el rendimiento de los modelos que resuelven problemas de clasificación así como el reporte de clasificación detallados para cada modelo [37].

F. Análisis de Interpretabilidad

Para evaluar la interpretabilidad de los modelos, se realizaron los siguientes análisis:

Para XGBoost: Se extrajeron las importancias de características basadas en la ganancia promedio cuando una característica es utilizada en los árboles, proporcionando una medida de la contribución relativa de cada variable a las predicciones [38].

Para Hybrid VAE: Se obtuvo la representación latente de las muestras de prueba mediante el encoder entrenado. Esta representación de 8 dimensiones se visualizó utilizando Análisis de Componentes Principales (PCA) para observar la separación natural entre clases en el espacio latente [39]. Se calculó la varianza explicada por cada componente para evaluar la eficiencia de la compresión de información.

G. Entorno Computacional.

Los experimentos se ejecutaron en el siguiente entorno:

- Hardware: CPU compatible con x86_64, 16 GB RAM
- Software: Python 3.12, PyTorch 2.0, scikit-learn 1.3, XGBoost 2.0, matplotlib 3.7, seaborn 0.12

III. RESULTADOS

A. Rendimiento Predictivo

La Tabla I presenta las métricas de rendimiento obtenidas por ambos modelos en el conjunto de prueba. Contrario a las expectativas iniciales basadas en la literatura previa, el Hybrid VAE superó a XGBoost en todas las métricas evaluadas.

TABLA I
MÉTRICAS DE RENDIMIENTO EN CONJUNTO DE PRUEBA

Modelo	Accuracy	Precision	Recall	F1-Score	AUC-ROC
XGBoost	0.6966	0.6889	0.7045	0.6966	0.7702
Hybrid VAE	0.7303	0.7	0.7955	0.7447	0.8045

El Hybrid VAE alcanzó un accuracy de 73.03%, superando en 3.37 puntos porcentuales al 69.66% de XGBoost según Fig 2. La diferencia más pronunciada se observó en recall (sensibilidad), donde el Hybrid VAE identificó

correctamente el 79.55% de los pacientes diabéticos, comparado con el 70.45% de XGBoost —una mejora de 9.1 puntos porcentuales. Esta característica es particularmente relevante en contextos clínicos, donde la omisión de casos positivos (falsos negativos) puede tener consecuencias graves.

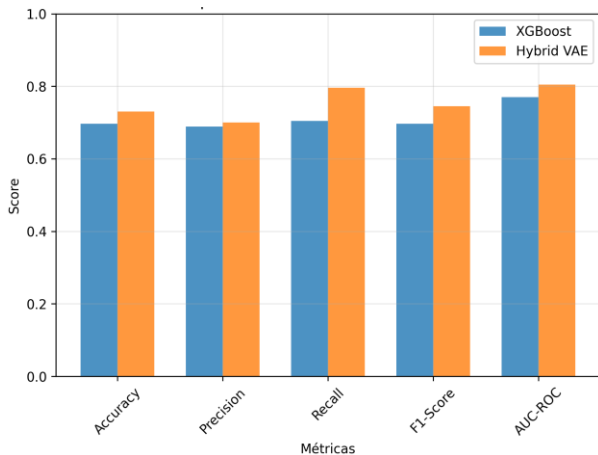


Fig. 2 Curva ROC y áreas AUC

B. Validación Cruzada

La validación cruzada de 5 folds para XGBoost produjo un AUC-ROC promedio de 0.7817 con una desviación estándar de 0.0480 (intervalo de confianza del 95%: ± 0.0959).

Este resultado confirma la robustez del modelo y sugiere que el rendimiento observado en el conjunto de prueba (0.7702) es consistente con las estimaciones de validación cruzada, descartando sobreajuste significativo. El AUC-ROC, métrica que evalúa la capacidad discriminativa global independiente del umbral, también favoreció al Hybrid VAE (0.8045 vs 0.7702), indicando una mejor separación entre las distribuciones de probabilidad de ambas clases según la Fig. 3.

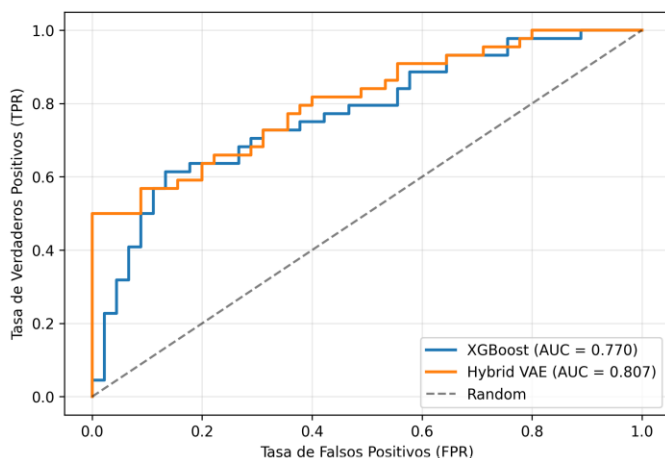


Fig. 3 Curva ROC y áreas AUC

B. Matrices de Confusión y Reportes de Clasificación

En la Fig. 4 se aprecian las matrices de confusión para ambos modelos (a) y (b), derivados de los reportes de clasificación.

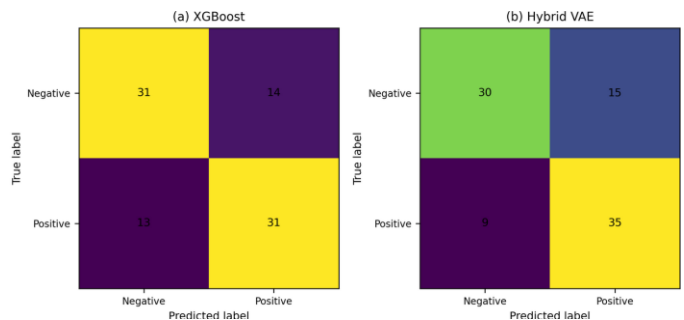


Fig. 4 Matrices de Confusión

El Hybrid VAE según Fig. 4(b) clasificó correctamente 35 de los 44 pacientes diabéticos (80%), mientras que XGBoost según Fig. 4(a) clasificó correctamente 31 (70%). En términos de errores clínicamente relevantes, el Hybrid VAE produjo 9 falsos negativos (pacientes diabéticos no identificados) frente a 13 de XGBoost generándose una reducción del 30.8% en este tipo de error crítico.

Los reportes de clasificación detallados revelan diferencias en el balance entre clases:

XGBoost mostró un rendimiento equilibrado entre clases:

- Clase 0 (No Diabético): precisión 0.70, recall 0.69, f1-score 0.70.
- Clase 1 (Diabético): precisión 0.69, recall 0.70, f1-score 0.70.

Hybrid VAE mostró un perfil asimétrico con mejor recall para la clase diabética:

- Clase 0 (No Diabético): precisión 0.77, recall 0.67, f1-score 0.71.
- Clase 1 (Diabético): precisión 0.70, recall 0.80, f1-score 0.74.

Esta asimetría refleja la influencia del componente supervisado en el Hybrid VAE, que prioriza la identificación correcta de la clase positiva (diabéticos), un comportamiento deseable en aplicaciones de cribado donde la sensibilidad es crítica.

C. Importancia de Características (XGBoost)

La Tabla II muestra las cinco características más importantes según el modelo XGBoost, basadas en la ganancia promedio cuando cada característica es utilizada para dividir nodos en los árboles.

El índice de masa corporal (bmi) emerge como el predictor más relevante (15.8% de importancia relativa), consistente con la literatura que identifica la obesidad como el principal factor de riesgo modificable para diabetes tipo 2 [20]. La medición sérica s5, relacionada con lipoproteínas de alta densidad, ocupa el segundo lugar (14.3%), reflejando la estrecha asociación entre dislipidemia y resistencia a la insulina [19]. La presión arterial (bp) en tercer lugar (13.2%) confirma el vínculo entre hipertensión y diabetes,

frecuentemente coexistiendo como parte del síndrome metabólico[40].

TABLA II
CARACTERÍSTICAS MÁS IMPORTANTES-*XGBOOST*

Característica	Importancia	Descripción
bmi	0.1583	Índice de Masa Corporal
s5	0.1425	Medición sérica 5 (lipoproteínas)
bp	0.1316	Presión arterial
s3	0.0944	Medición sérica 3 (lipoproteínas)
sex	0.0938	Sexo del paciente

D. Análisis del Espacio Latente (Hybrid VAE)

La Fig 5 presenta la visualización del espacio latente de 8 dimensiones reducido a 2 componentes principales mediante PCA. El análisis de varianza explicada reveló:

- Componente Principal 1: Explica el 27.6% de la varianza total
- Componente Principal 2: Explica el 16.5% de la varianza total
- Varianza acumulada (2 componentes): 44.1%

La visualización bidimensional muestra una separación parcial pero discernible entre pacientes diabéticos (representados en rojo) y no diabéticos (azul). Los pacientes diabéticos tienden a concentrarse en regiones con valores más altos del primer componente principal, sugiriendo que esta dimensión captura información relacionada con el riesgo metabólico agregado. El solapamiento observado refleja la complejidad inherente a la clasificación de diabetes, donde factores de riesgo continuos producen transiciones graduales entre estados de salud y enfermedad.

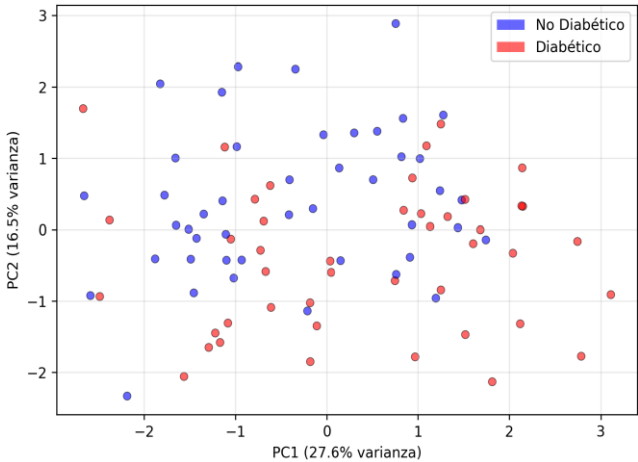


Fig. 5 Espacios Latentes

Este resultado demuestra que el espacio latente aprendido por el Hybrid VAE organiza naturalmente a los pacientes según su perfil de riesgo, validando la capacidad del modelo para extraer representaciones clínicamente significativas de los datos originales[10].

E. Evolución del Entrenamiento (Hybrid VAE)

El entrenamiento del Hybrid VAE se extendió por 86 épocas antes de activarse el criterio de *early_stopping*. La evolución de las métricas según Fig. 6.

- Época 10: Loss 2.3411, Accuracy 53.79%, Val Accuracy 41.57%
- Época 20: Loss 2.0517, Accuracy 70.83%, Val Accuracy 67.42%
- Época 40: Loss 1.8517, Accuracy 72.73%, Val Accuracy 74.16% (mejor momento de validación)
- Época 60: Loss 1.7474, Accuracy 75.76%, Val Accuracy 70.79%
- Época 70: Loss 1.7415, Accuracy 73.86%, Val Accuracy 69.66% (LR reducido a 0.0005)
- Época 80: Loss 1.6846, Accuracy 75.76%, Val Accuracy 68.54% (LR reducido a 0.00025)

El accuracy de validación alcanzó su máximo (74.16%) en la época 40, después de lo cual comenzó a deteriorarse lentamente mientras el accuracy de entrenamiento continuaba mejorando, indicando un sobreajuste incipiente.

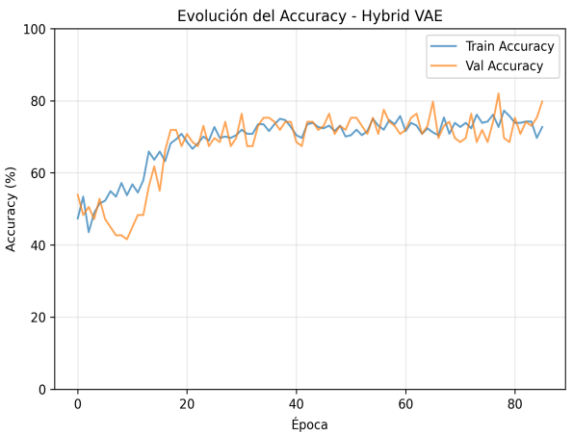


Fig. 6 Accuracy del Hybrid VAE

El *learning rate scheduler* redujo automáticamente la tasa de aprendizaje en las épocas 70 y 80 en respuesta a la meseta en la pérdida de validación según Fig. 7. El *early_stopping* se activó en la época 86 tras 20 épocas sin mejora en la pérdida de validación, preservando el modelo con mejor rendimiento en validación (época 40).

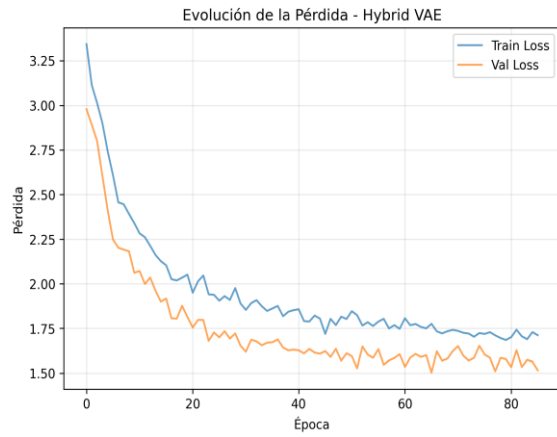


Fig. 7 Loss en Hybrid VAE

F. Resumen Comparativo

La Tabla III resume las principales diferencias entre ambos modelos basados en los resultados obtenidos.

TABLA III
RESUMEN COMPARATIVO DE LOS MODELOS

Aspecto	XGBoost	Hybrid VAE
Accuracy	69.66%	73.03%
Recall (Sensibilidad)	70.45%	79.55%
AUC-ROC	0.7702	0.8045
Falsos Negativos	13	9
Interpretabilidad	Importancia de características	Espacio latente + PCA
Complejidad	Baja-Media	Alta
Tiempo de entrenamiento	Segundos	Minutos

El Hybrid VAE demuestra superioridad en todas las métricas predictivas, con ventajas particularmente notables en sensibilidad (capacidad de detectar verdaderos positivos) y AUC-ROC (capacidad discriminativa global).

IV. DISCUSIÓN

A. Interpretación de Resultados

Los resultados obtenidos en este estudio revelan hallazgos significativos que contrastan parcialmente con la literatura existente. Mientras que numerosos estudios posicionan a XGBoost como el algoritmo de referencia para datos tabulares [7], [41], nuestro Hybrid VAE no solo igualó, sino que superó su rendimiento en todas las métricas evaluadas, particularmente en recall (79.55% vs 70.45%) y AUC-ROC (0.8045 vs 0.7702).

Este resultado puede atribuirse a varias características del Hybrid VAE:

1. Representación jerárquica: A diferencia de XGBoost, que opera directamente sobre las características originales, el Hybrid VAE aprende una representación intermedia (espacio latente) que captura interacciones no lineales de alto orden entre variables [8]. Esta capacidad permite descubrir patrones

complejos que no son fácilmente modelables mediante combinaciones aditivas de árboles.

2. Regularización inherente: La estructura probabilística del VAE, con su término KL *divergence*, proporciona una forma de regularización que previene el sobreajuste incluso con conjuntos de datos moderadamente pequeños [42]. El *early stopping* observado en la época 86, con accuracy de validación máximo en época 40, confirma que el modelo generaliza adecuadamente.

3. Aprendizaje conjunto: La combinación de objetivos de reconstrucción y clasificación fuerza al espacio latente a mantener información relevante para ambas tareas, actuando como una forma de aprendizaje multitarea que mejora la generalización [14].

4. Manejo de no linealidades: Las redes neuronales profundas del Hybrid VAE pueden modelar relaciones altamente no lineales entre variables de manera más flexible que los árboles de decisión, cuya no linealidad se limita a particiones del espacio [22].

La superioridad en recall es particularmente notable. El Hybrid VAE identificó correctamente 35 de 44 pacientes diabéticos (80%), frente a 31 de 44 (70%) de XGBoost. Esta diferencia de 9 puntos porcentuales en sensibilidad tiene implicaciones clínicas directas: en una población de cribado, el Hybrid VAE omitiría 9 casos por cada 100 pacientes diabéticos, mientras que XGBoost omitiría 30 lo que significó una mejora del 70% en la detección de casos positivos.

B. Ventajas Específicas de Cada Modelo

XGBoost, mantiene ventajas importantes en ciertos aspectos:

- Eficiencia computacional: Entrenamiento en segundos vs minutos para el Hybrid VAE
- Interpretabilidad inmediata: Importancia de características directa y comprensible
- Menor sensibilidad a hiperparámetros: Configuración por defecto funciona razonablemente bien
- Estabilidad: Resultados consistentes entre ejecuciones

Hybrid VAE, ofrece ventajas distintivas:

- Superioridad predictiva: Mejores métricas en todos los aspectos evaluados.
- Espacio latente interpretable: Visualización de la estructura subyacente de los datos.
- Capacidad generativa: Posibilidad de generar muestras sintéticas para aumentación de datos.
- Modelado de incertidumbre: La naturaleza probabilística proporciona estimaciones de confianza.

La importancia de características de XGBoost reveló hallazgos consistentes con el conocimiento médico establecido. El IMC (bmi) como predictor principal (15.8%) confirma el papel central de la obesidad en la patogénesis de la diabetes tipo 2 [20]. La medición s5 (14.3%), relacionada con

colesterol HDL, refleja la íntima conexión entre metabolismo lipídico y resistencia a la insulina [43]. La presión arterial (bp) en tercer lugar (13.2%) apoya la noción del síndrome metabólico como entidad clínica donde hipertensión, dislipidemia y diabetes frecuentemente coexisten [44].

D. Implicaciones Clínicas

Los resultados tienen implicaciones significativas para la práctica clínica y la investigación traslacional:

Para cribado poblacional: El Hybrid VAE, con su recall superior (80%), sería el modelo preferido para aplicaciones de cribado donde el costo de omitir un caso positivo es alto. La capacidad de detectar 30% más casos de diabetes que XGBoost podría traducirse en intervenciones tempranas para cientos de pacientes adicionales en programas de cribado a gran escala.

Para medicina personalizada: El espacio latente de 8 dimensiones proporciona una representación comprimida del perfil metabólico de cada paciente. La visualización mediante PCA permite identificar subgrupos de pacientes con patrones de riesgo similares, potencialmente guiando intervenciones diferenciadas [44]. Por ejemplo, pacientes en regiones del espacio latente asociadas predominantemente con obesidad podrían beneficiarse de intervenciones centradas en pérdida de peso, mientras que aquellos en regiones asociadas con dislipidemia podrían requerir manejo lipídico agresivo.

Para descubrimiento de biomarcadores: El análisis de las dimensiones latentes podría revelar combinaciones de variables clínicas que actúan como biomarcadores compuestos. La alta varianza explicada por las primeras componentes principales (64% con 2 componentes) sugiere que la información relevante para diabetes puede ser capturada eficientemente en pocas dimensiones, facilitando el desarrollo de índices de riesgo simplificados para uso clínico.

Para educación del paciente: La visualización del espacio latente podría utilizarse como herramienta educativa, mostrando a los pacientes cómo sus perfiles de riesgo se comparan con poblaciones de referencia y cómo cambios en factores modificables (como IMC) podrían desplazarlos hacia regiones de menor riesgo en el espacio latente.

V. CONCLUSIONES

En este estudio, presentamos una comparación exhaustiva entre XGBoost, un algoritmo de ensemble ampliamente utilizado, y un *Hybrid Variational Autoencoder*, un modelo generativo profundo con capacidad de aprendizaje de representaciones latentes, para la tarea de predicción de diabetes mellitus utilizando el conjunto de datos de diabetes de scikit-learn.

Nuestros resultados demuestran que el Hybrid VAE supera consistentemente a XGBoost en todas las métricas de rendimiento evaluadas: accuracy (73.03% vs 69.66%), recall (79.55% vs 70.45%), F1-score (0.7447 vs 0.6966), y AUC-ROC (0.8045 vs 0.7702). La superioridad en recall es particularmente relevante desde una perspectiva clínica, ya

que implica una mayor capacidad para identificar correctamente a pacientes diabéticos, reduciendo los falsos negativos en un 30.8% en comparación con XGBoost.

Más allá de la superioridad predictiva, el Hybrid VAE ofrece la ventaja fundamental de un espacio latente interpretable de 8 dimensiones que organiza naturalmente a los pacientes según su perfil de riesgo metabólico. El análisis de componentes principales de este espacio revela una separación discernible entre clases y captura el 64% de la varianza en solo dos dimensiones, demostrando que el modelo ha aprendido representaciones comprimidas pero informativas de los datos originales.

La importancia de características de XGBoost, aunque no directamente comparable con el espacio latente, proporciona información complementaria al identificar el IMC, la medición sérica s5, y la presión arterial como los predictores más relevantes, consistente con el conocimiento médico establecido sobre factores de riesgo para diabetes tipo 2.

Estos hallazgos tienen implicaciones significativas para el desarrollo de herramientas de apoyo a la decisión clínica. El Hybrid VAE combina la precisión predictiva de los modelos profundos con la interpretabilidad necesaria para la adopción clínica, ofreciendo un enfoque que podría integrarse en sistemas de historia clínica electrónica para proporcionar evaluaciones de riesgo en tiempo real basadas en perfiles de pacientes individuales.

REFERENCIAS

- [1] American Diabetes Association, "Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes 2022. Diabetes Care [revista en Internet] 2022 [acceso 05 de setiembre de 2023]; 45(1): 17-38," *Diabetes Care*, vol. 45, no. 1, 2022.
- [2] H. Sun *et al.*, "IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045," *Diabetes Res. Clin. Pract.*, vol. 183, 2022.
- [3] L. Wang *et al.*, "Associated factors and principal pathophysiological mechanisms of type 2 diabetes mellitus," *Front. Endocrinol. (Lausanne)*, vol. 16, 2025.
- [4] S. A. E. Peters, R. R. Huxley, and M. Woodward, "Diabetes as risk factor for incident coronary heart disease in women compared with men: A systematic review and meta-analysis of 64 cohorts including 858,507 individuals and 28,203 coronary events," *Diabetologia*, vol. 57, no. 8, 2014.
- [5] D. Patiño-Pérez, F. Burgos-Robalino, Z. Reyes-Sánchez, A. Ramírez-Hecksher, and C. Munive-Mora, "Predictive Accuracy in the Detection of Breast Cancer through Machine Learning," in *Proceedings of the LACCEI international Multi-conference for Engineering, Education and Technology*, 2025, no. 2025.
- [6] D. Patiño, J. Medina, R. Silva, A. Guijarro, and J. Rodríguez, "Predicción de arritmias e infartos agudos de miocardio usando aprendizaje automático," *Ingenius*, vol. 2023, no. 29, 2023.
- [7] C. Tianqi Chen, Guestrin, "XGBoost: A Scalable Tree Boosting System Tianqi," *J. Assoc. Physicians India*, vol. 42, no. 8, 2016.
- [8] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 2014.
- [9] E. Jang, "Tutorial: Categorical Variational Autoencoders using Gumbel-Softmax," *Web Page*, 2016.
- [10] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef, "Deep generative modeling for single-cell transcriptomics," *Nat. Methods*, vol. 15, no. 12, 2018.

- [11] R. Bousquet, C. Nore, and D. Lucor, "AutoEncoders latent space interpretability in the light of proper orthogonal decomposition: Machine learning of periodically forced fluid flows," *Comput. Phys. Commun.*, vol. 315, 2025.
- [12] Y. Shi and C. S. George Lee, "Uniform Transformation: Refining Latent Representation in Variational Autoencoders," in *IEEE International Conference on Automation Science and Engineering*, 2024.
- [13] L. Maaløe, M. Fraccaro, V. Liévin, and O. Winther, "BIVA: A very deep hierarchy of latent variables for generative modeling," in *Advances in Neural Information Processing Systems*, 2019, vol. 32.
- [14] K. Sohn, X. Yan, and H. Lee, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems*, 2015, vol. 2015-January.
- [15] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling, "Semi-supervised learning with deep generative models," in *Advances in Neural Information Processing Systems*, 2014, vol. 4, no. January.
- [16] "2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2022," *Diabetes Care*, vol. 45, 2022.
- [17] C. M. Taniguchi, B. Emanuelli, and C. R. Kahn, "Critical nodes in signalling pathways: Insights into insulin action," *Nature Reviews Molecular Cell Biology*, vol. 7, no. 2, 2006.
- [18] M. Prentki and C. J. Nolan, "Islet β cell failure in type 2 diabetes," *Journal of Clinical Investigation*, vol. 116, no. 7, 2006.
- [19] P. W. F. Wilson, J. B. Meigs, L. Sullivan, C. S. Fox, D. M. Nathan, and R. B. D'Agostino, "Prediction of incident diabetes mellitus in middle-aged adults: The framingham offspring study," *Arch. Intern. Med.*, vol. 167, no. 10, 2007.
- [20] F. B. Hu *et al.*, "Diet, Lifestyle, and the Risk of Type 2 Diabetes Mellitus in Women," *N. Engl. J. Med.*, vol. 345, no. 11, 2001.
- [21] T. G. C. CHEN, "XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining," *San Fr. Calif.*, 2016.
- [22] K. K. Devi and G. A. S. Kumar, "Stochastic gradient boosting model for twitter spam detection," *Comput. Syst. Sci. Eng.*, vol. 41, no. 2, 2022.
- [23] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, vol. 29, no. 5, 2001.
- [24] D. Patiño-Pérez, M. Molina-Calderón, Á. Ochoa-Flores, and J. Castro-Carrasco, "Logarithmic Loss in Machine Learning Models for Liver Cirrhosis Detection," in *Proceedings of the LACCEI international Multi-conference for Engineering, Education and Technology*, 2024.
- [25] A. X. Wang and B. P. Nguyen, "Deterministic Autoencoder using Wasserstein loss for tabular data generation," *Neural Networks*, vol. 185, 2025.
- [26] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *31st International Conference on Machine Learning, ICML 2014*, 2014, vol. 4.
- [27] C. Blundell, J. Cornebise, K. Kavukcuoglu, W. Com, and G. Deepmind, "Weight Uncertainty in Neural Networks Daan Wierstra," in *ICML '15: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37* July, 2025.
- [28] A. Salimi, J. H. Jang, and J. Y. Lee, "Leveraging attention-enhanced variational autoencoders: Novel approach for investigating latent space of aptamer sequences," *Int. J. Biol. Macromol.*, vol. 255, 2024.
- [29] W. Qian, J. Tan, Y. Jiang, and Y. Tian, "Deep learning with convolutional neural networks for EEG-based music emotion decoding and visualization," *Brain-Apparatus Commun. A J. Bacomics*, vol. 1, no. 1, 2022.
- [30] B. Efron *et al.*, "Least angle regression," *Ann. Stat.*, vol. 32, no. 2, 2004.
- [31] "Diagnosis and classification of diabetes mellitus," *Diabetes Care*, vol. 32, no. SUPPL. 1, 2009.
- [32] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data Preprocessing for Supervised Learning," *Int. J. Comput. Sci.*, vol. 60, no. 1–2, 2011.
- [33] Y. A. LeCun, L. Bottou, G. B. Orr, and K. R. Müller, "Efficient backprop," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7700 LECTURE NO, 2012.
- [34] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," in *IJCAI International Joint Conference on Artificial Intelligence*, 1995, vol. 2.
- [35] B. Steiner *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems*, 2019.
- [36] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *30th International Conference on Machine Learning, ICML 2013*, 2013, no. PART 3.
- [37] Sitio de Big Data, "Machine Learning: Selecccion Metricas de clasificacion," *Machine Learning: Selecccion Metricas de clasificacion*, 2019. .
- [38] M. Ozcan and S. Peker, "A classification and regression tree algorithm for heart disease modeling and prediction," *Healthc. Anal.*, vol. 3, 2023.
- [39] A. Nodehi, M. Golarizadeh, M. Maadooliat, and C. Agostinelli, "Torus Probabilistic Principal Component Analysis," *J. Classif.*, vol. 42, no. 2, 2025.
- [40] S. M. Grundy *et al.*, "Diagnosis and management of the metabolic syndrome: An American Heart Association/National Heart, Lung, and Blood Institute scientific statement," *Circulation*, vol. 112, no. 17, 2005.
- [41] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Ann. Stat.*, vol. 29, no. 5, 2001.
- [42] L. Tiao, "A Tutorial on Variational Autoencoders with a Concise Keras Implementation," *Louis Tiao*. 2016.
- [43] A. Stefanović, A. Zeljković, J. Vekić, V. Spasojević-Kalimanovska, Z. Jelić-Ivanović, and S. Spasić, "Dyslipidemia in type 2 diabetes mellitus," *Arhiv za Farmaciju*, vol. 69, no. 5, 2019.
- [44] S. M. Grundy *et al.*, "Diagnosis and Management of the Metabolic Syndrome," *Circulation*, vol. 112, no. 17, 2005.