Endocrinol Diabetes Nutr. 2022;69(Espec Cong 1):210

Endocrinología, Diabetes y Nutrición

P-158 - Validación Clínica y Análisis de Datos Sintéticos de Mujeres con Diabetes Generados con Técnicas de Inteligencia Artificial: prueba de concepto

A.J. Rodríguez Almeida[a], *A. Déniz*[b], *H. Fabelo*[c,a], *S. Ortega*[a], *C. Soguero*[d], *A. Wägner*[b] *y G. Marrero Callicó*[a]

[a]*Instituto Universitario de Microelectrónica Aplicada, Universidad de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria.*[b]*Departamento de Endocrinología y Nutrición, Complejo Hospitalario Universitario Insular Materno-Infantil de Gran Canaria, Instituto Universitario de Investigaciones Biomédicas y Sanitarias (IUIBS), ULPGC, Las Palmas de Gran Canaria.*[c]*Fundación Canaria del Instituto de Investigación Sanitaria de Canarias (FIISC), Las Palmas de Gran Canaria.*[d]*Departamento de Teoría de la Señal y las Comunicaciones y Sistemas Telemáticos y Computación, Universidad Rey Juan Carlos, Fuenlabrada.*

Resumen

Objetivos: El acceso a datos clínicos para el desarrollo y evaluación de tecnologías médicas se ve dificultado por el alto coste de la creación de bases de datos y las limitaciones asociadas a la preservación de la privacidad de los pacientes. El uso de datos sintéticos generados mediante técnicas de Inteligencia Artificial (IA) podría facilitar y acelerar la investigación en tecnologías médicas. Las *Conditional Tabular Generative Adversarial Networks* (CTGANs), una técnica de IA diseñada para la generación de datos sintéticos a partir de datos reales, han surgido como posible solución a este problema. El objetivo de este trabajo es la validación clínica de los pacientes sintéticos creados a partir de una CTGAN por parte de varios expertos clínicos, así como evaluar los cambios en la estructura subyacente de los datos sintéticos con respecto a los datos originales

Material y métodos: Se desarrolló un flujo de trabajo utilizando una CTGAN y la base de datos pública *PIMA Indian Database*, que contiene 8 variables numéricas de pacientes diabéticas. Se generaron 20 pacientes sintéticos y se combinaron de forma aleatoria con 20 pacientes reales. El conjunto de 40 pacientes fue evaluado de forma ciega por dos expertos (sin conocer la cantidad y la etiqueta de los datos sintéticos), identificando los pacientes que no concordaban con un paciente real basándose en los valores de las variables. Las métricas utilizadas para evaluar los cambios en la estructura estadística de los datos sintéticos han sido: *Pairwise Correlation Difference* (PCD), *Kullback-Leibler Divergence* (KLD), *Maximum Mean Discrepancy* (MMD) y la importancia de las variables usando *Random Forest* (RF).

Resultados: La tabla muestra el análisis de los expertos. Una identificación correcta corresponde con identificar el “Real” como “Verdadero” y el “Sintético” como “Falso”. Los resultados estadísticos obtenidos sugieren que se conserva fielmente la estructura de los datos reales. Sin embargo, hay diferencias en la importancia de las variables entre los pacientes reales y los pacientes sintéticos.

		Exp.I	Exp.II	Exp.III
Real	Verdadero	13 (65%)	20 (100%)	15 (100%)

Falso	7 (70%)	0 (0%)	5 (0%)	
Sintético	Verdadero	14 (70%)	20 (0%)	14 (0%)
	Falso	6 (30%)	0 (100%)	6 (100%)

Conclusiones: Como prueba de concepto, se demuestra que el uso de la CTGAN, para esta base de datos, permite generar pacientes sintéticos diabéticos que, además de mantener razonablemente la estructura estadística de los datos, mantienen la coherencia clínica, lo cual es crucial en el ámbito médico. Se espera, en trabajos futuros, validar este método con un conjunto mayor de pacientes, diferentes bases de datos y un mayor número de expertos. De esta manera, podría demostrarse si esta herramienta puede ser útil para sustituir o complementar pacientes reales por pacientes sintéticos sin perder validez clínica ni la estructura de los datos a la hora de desarrollar tecnologías médicas basadas en IA.