Toward Predicting Microvascular Complications in People with Type 2 Diabetes in Argentina

Gonzalo Hernán Domínguez[1], Jorge Elgart[1], Enzo Rucci[2,3]

[1] Centro de Endocrinología Experimental y Aplicada (CENEXA-UNLP-CONICET; CeAs CICPBA), La Plata, Provincia de Buenos Aires, Argentina.
[2] III-LIDI, Facultad de Informática, UNLP - CIC, La Plata (1900), Bs As, Argentina
[3] Comisión de Investigaciones Científicas (CIC), La Plata (1900), Bs As, Argentina
{ghdominguez,jelgart}@med.unlp.edu.ar,
erucci@lidi.info.unlp.edu.ar

Abstract. Type 2 diabetes mellitus (T2DM) is a growing public health problem associated with the development of microvascular complications that significantly impact patients' quality of life. In this context, the objective of this study is to build a robust clinical dataset that will enable the future development of machine learning models for predicting these complications in the Argentinian population. Data from the Qualidiab system were used, and a cohort of patients with T2DM with longitudinal records was selected. Pairs of records (baseline and 3–5-year follow-up) were defined, excluding patients with complications at baseline, resulting in a final set of 843 patients. Preprocessing included the removal of clinical inconsistencies, the detection and handling of outliers using expert criteria, the generation of derived variables, and the construction of binary treatment variables. Imputation of missing data combined temporal strategies, physiological relationships (HbA1c–blood glucose), clinical formulas (Friedewald), and multivariate methods. This study represents the first step toward the development of predictive tools for clinical practice.

Keywords: type 2 diabetes, data preprocessing, imputation, machine learning, microvascular complications.

Hacia la predicción de complicaciones microvasculares en personas con diabetes tipo 2 en Argentina

Resumen. La diabetes mellitus tipo 2 (DM2) constituye un problema de salud pública creciente, asociado al desarrollo de complicaciones microvasculares que impactan significativamente en la calidad de vida de los pacientes. En este contexto, el presente trabajo tiene como objetivo construir un *dataset* clínico robusto que permita el desarrollo futuro de modelos de aprendizaje automático para la predicción de dichas complicaciones en una población argentina. Se utilizaron datos provenientes del sistema Qualidiab, seleccionando una cohorte de pacientes con DM2 con registros longitudinales. Se definieron pares de fichas (ba-

sal y seguimiento a 3–5 años), excluyendo pacientes con complicaciones al inicio, obteniéndose un conjunto final de 843 pacientes. El preprocesamiento incluyó depuración de inconsistencias clínicas, detección y tratamiento de valores atípicos mediante criterios expertos, generación de variables derivadas y construcción de variables binarias de tratamiento. La imputación de datos faltantes combinó estrategias temporales, relaciones fisiológicas (HbA1c–glucemia), fórmulas clínicas (Friedewald) y métodos multivariados. Este trabajo representa el primer paso hacia el desarrollo de herramientas predictivas orientadas a la práctica clínica.

Palabras clave: diabetes tipo 2, preprocesamiento de datos, imputación, aprendizaje automático, complicaciones microvasculares.

1 Introducción

La diabetes *mellitus* tipo 2 (DM2) es una enfermedad metabólica crónica caracterizada por resistencia a la insulina y disfunción progresiva de las células beta pancreáticas, lo que conduce a hiperglucemia sostenida. Representa aproximadamente el 90% de los casos de diabetes a nivel mundial y su prevalencia ha crecido sostenidamente en las últimas décadas, impulsada por factores como el envejecimiento poblacional, la urbanización y los cambios en los estilos de vida (Quiñones Silva et al.,2024). En Argentina, esta tendencia global se replica, con prevalencias en torno al 12.7% en poblaciones adultas, consolidándose como un problema relevante de salud pública (Aparicio-Montenegro et al.,2025; Russo et al., 2023; INDEC, 2019).

Una de las principales preocupaciones asociadas a la DM2 es el desarrollo de complicaciones crónicas, particularmente las microvasculares, entre las que se incluyen la retinopatía, la nefropatía y la neuropatía diabética. Estas condiciones generan un deterioro progresivo en la calidad de vida de los pacientes y pueden derivar en consecuencias severas, como ceguera, insuficiencia renal o amputaciones. Estas complicaciones representan una carga significativa en términos de morbilidad y costos para los sistemas de salud (Quiñones Silva et al.,2024).

En este contexto, el aprendizaje automático (*machine learning*) ha emergido como una herramienta prometedora en el ámbito de la medicina, particularmente en tareas de predicción, diagnóstico temprano y medicina personalizada. Estudios recientes muestran que modelos basados en inteligencia artificial pueden alcanzar altos niveles de precisión en la identificación de patrones clínicos complejos, facilitando la detección temprana de enfermedades, y la estratificación del riesgo en pacientes (Aparicio-Montenegro et al.,2025). En el caso de la diabetes, estos enfoques han sido aplicados tanto para la predicción de la enfermedad como para la identificación de factores de riesgo y la anticipación de complicaciones.

Sin embargo, un desafío clave en el desarrollo de modelos predictivos robustos es la disponibilidad y calidad de los datos. En este sentido, este trabajo describe la construcción de un dataset específicamente diseñado para el entrenamiento y evaluación de modelos de aprendizaje automático orientados a la predicción de complicaciones microvasculares en pacientes con DM2 en Argentina. Este dataset busca integrar in-

formación clínica relevante, garantizando calidad, consistencia y representatividad, para facilitar el desarrollo de modelos predictivos más precisos y contextualizados.

2 Trabajo realizado y resultados preliminares

Para la construcción del dataset, se siguieron las dos primeras etapas del proceso clásico de minería de datos (Guerra-Hernández et al., 2008), las cuales se describen a continuación.

2.1 Selección de los datos

Los datos fueron seleccionados del programa Qualidiab, una iniciativa regional orientada al registro sistemático y estandarizado de información clínica de personas con enfermedades metabólicas, coordinada por el Centro de Endocrinología Experimental y Aplicada (CENEXA, CONICET–UNLP) (Gagliardino et al., 2001). Aunque inicialmente el registro sólo contemplaba pacientes con diabetes, la segunda versión incorporó registros de prediabetes, hipertensión arterial y obesidad. Para este trabajo, se usó el subconjunto perteneciente a la obra social argentina OSPERYH, el cual está compuesto por 24.944 fichas clínicas, correspondientes a múltiples registros longitudinales de pacientes. Cada ficha contiene 143 variables, que incluyen información demográfica, clínica, bioquímica y terapéutica. Para cada paciente se seleccionaron dos instancias temporales: (i) una ficha basal, correspondiente al primer registro disponible del paciente en el sistema, y (ii) una ficha de seguimiento, seleccionada dentro de un intervalo comprendido entre tres y cinco años posteriores al registro basal (admitiendo un margen temporal de tolerancia de tres meses).

El conjunto inicial consistió en 1.761 pares de fichas que, al excluir aquellos registros correspondientes a pacientes sin DM2, resultó en una cohorte de 992 pacientes. Adicionalmente, se estableció como criterio de inclusión que ningún paciente presentara complicaciones al momento basal. En la ficha clínica Qualidiab se registran diversas complicaciones asociadas a la diabetes, las cuales fueron integradas en una única variable binaria (dummy) que representa la presencia de complicaciones microvasculares. Aplicando este criterio, la cohorte final quedó conformada por 843 pacientes con DM2. Las variables seleccionadas fueron definidas a partir de este análisis inicial y en concordancia con trabajos previos, como el de Dagliati et al. (2018).

2.2 Preprocesamiento de los datos

Corrección de inconsistencias y outliers. Inicialmente, se eliminaron aquellas variables completamente nulas y las irrelevantes para el objetivo del estudio. Asimismo, se generaron nuevas variables derivadas de interés clínico, incluyendo la edad del paciente al momento del registro y los años de evolución de la enfermedad. Además, se construyeron variables binarias representativas de tratamientos clínicos relevantes (hipertensión arterial, diabetes con/sin insulina), integrando la información proveniente de múltiples fármacos registrados en la ficha clínica. Durante este proceso se iden-

tificaron y corrigieron inconsistencias clínicas, particularmente las relacionadas con la presencia de tratamientos para la diabetes o insulina en pacientes sin diagnóstico registrado. Asimismo, se resolvieron situaciones incompatibles, como la coexistencia de diagnósticos de diabetes tipo 1 y tipo 2 en un mismo registro.

La detección de valores atípicos (outliers) se realizó sobre las variables clínicas y bioquímicas, y se emplearon rangos clínicamente aceptables definidos a partir de la bibliografía especializada y del criterio experto. Los valores que excedían dichos rangos fueron recodificados como valores nulos.

Imputación de datos faltantes.

Imputación temporal. Este enfoque consistió en identificar el valor observado más cercano en el tiempo dentro de una ventana máxima de ±1,5 años y utilizarlo para completar valores faltantes. Para las fichas basales, se realizó una búsqueda hacia adelante en el tiempo (*trace forward*), mientras que para las fichas de seguimiento se efectuó una búsqueda hacia atrás (*trace back*).

Imputación basada en la relación entre HbA1c y glucemia en ayunas. Para las variables “glucemia en ayunas" y “hemoglobina glicosilada" (HbA1c), se aplicó una estrategia de imputación cruzada basada en la relación fisiológica entre ambas medidas. En particular, se utilizó la ecuación propuesta por Nathan et al. (2008), ampliamente empleada en estudios clínicos. Adicionalmente, se incorporó un término de corrección correspondiente al error medio observado entre los valores medidos y los valores estimados por la fórmula.

Imputación del perfil lipídico mediante la fórmula de Friedewald. Para completar valores faltantes del perfil lipídico, se empleó la fórmula de Friedewald (Friedewald et al. 1972), válida bajo la condición de triglicéridos inferiores a 400 mg/dL. Esta relación permite estimar el colesterol LDL a partir del colesterol total, el colesterol HDL y los triglicéridos. Al igual que en el caso anterior, se corrigió usando los valores medios.

Exclusión de registros con proporción alta de valores faltantes. Previo a la imputación final, se evaluó la cantidad de valores faltantes por paciente en el subconjunto de 16 variables clínicas provenientes de la exploración física y de laboratorio. Aquellos registros que presentaban ocho o más valores faltantes dentro de este conjunto de variables, fueron excluidos del análisis. Este criterio se adoptó para evitar procesos de imputación excesivos que pudieran introducir sesgos y comprometer la validez clínica y estadística de los análisis posteriores.

Imputación multivariada de valores faltantes. Los valores nulos remanentes fueron imputados utilizando enfoques complementarios, tanto estadísticos simples como métodos basados en aprendizaje automático. Inicialmente, se descartaron aquellas variables que presentaban más del 50% de valores faltantes. Tras este filtrado, el conjunto de datos resultante estuvo compuesto por 992 registros de 69 variables, de las cuales 27 presentaban valores faltantes en distintos porcentajes. En particular, las variables bioquímicas presentaron proporciones de valores nulos desde 10.18% hasta 21.67%. Entre las variables clínicas y demográficas, se observaron valores faltantes desde 0.60% hasta 33.67%. Para evaluar la calidad de las distintas estrategias de imputación, se construyó un conjunto de datos completo usando únicamente aquellas instancias sin valores faltantes. A partir de este conjunto de referencia, se aplicaron

tres métodos de imputación: por la media, por la mediana y multivariado basado en *Random Forest*; todos ellos usando la librería scikit-learn. El desempeño de cada estrategia se evaluó exclusivamente sobre variables numéricas continuas que presentaban valores faltantes en el conjunto original, utilizando como métricas el error cuadrático medio (RMSE) y el error cuadrático medio normalizado (RMSEN). En todas las variables evaluadas, el enfoque basado en *Random Forest* presentó sistemáticamente los valores más bajos de RMSE (promedio = 0.0138 ± 0.0275) y RMSEN (promedio = 0.0545 ± 0.0407), lo que indica una mayor precisión en la reconstrucción de los valores faltantes. En función de estos resultados, este método fue seleccionado como la estrategia final de imputación multivariada para el conjunto de datos propuesto.

2.3 Resultados del dataset construido

La Tabla 1 presenta las variables que integran el *dataset* construido y algunos datos que describen su distribución, las cuales fueron seleccionadas en base a los filtros aplicados durante el preprocesamiento (Sección 2.2) y en concordancia con lo reportado por Dagliati et al. (2018).

Tabla 1. Descripción del *dataset* construido

Categoría	Variable	Métrica	Resultados
Demográfica	edad_en_registro	Media ± DE	53.41 ± 6.97
	sexo	Femenino (%)	343 (40.7%)
		Masculino (%)	500 (59.3%)
Clínica	exploraciones_imc	Media ± DE	32.08 ± 5.54
	anios_con_enfermedad	Media ± DE	5.39 ± 6.38
	tabaquismo	No (%)	716 (84.9%)
		Sí (%)	127 (15.1%)
Laboratorio	indice_tg_hdl	Media ± DE	3.47 ± 2.52
	exploraciones_colesterol_valor	Media ± DE	202.53 ± 39.40
	exploraciones_hba1c_valor	Media ± DE	7.67 ± 1.89
	exploraciones_creatinina_valor	Media ± DE	1.01 ± 0.17
Tratamiento	trata_hipertension	No (%)	367 (43.5%)
		Sí (%)	476 (56.5%)

Resulta importante notar que la variable dependiente mostró un marcado desbalance de clases al final, con 731 (86.7%) personas sin complicaciones y 112 (13.3%) personas con complicaciones microvasculares. Esta situación deberá ser tenida en cuenta durante el futuro proceso de modelado.

3 Conclusiones y trabajo futuro

Ante la ausencia de alternativas, el dataset construido representa el primer paso hacia el entrenamiento y evaluación de modelos de aprendizaje automático orientados a la predicción de complicaciones microvasculares en pacientes con DM2 en Argentina. Su concreción sienta las bases para el desarrollo de herramientas de apoyo a la decisión clínica que podrían contribuir a la detección temprana de complicaciones, la personalización de tratamientos y la optimización de recursos en el sistema de salud.

Referencias

Quiñones Silva, J. B., Bayona Cebada, A., Escobar-Morreale, H. F. y Nattero Chávez, L. (2024). Diagnóstico de la diabetes mellitus tipo 2: situación epidemiológica, características de los pacientes, factores de riesgo y pronóstico. Medicine - Programa de Formación Médica Continuada Acreditado, 14(19), 1099–1106. https://doi.org/10.1016/j.med.2024.10.018

Aparicio-Montenegro, P. R., Narro-Andrade, M. G., León-Velarde, C. G., Morales-Romero, G. P. y Fernández-Flores, S. M. (2025). Modelos predictivos en la salud pública: el abordaje de la diabetes mediante la inteligencia artificial. Cuestiones Políticas, 43(82). https://doi.org/10.5281/zenodo.15565315

INDEC - Instituto Nacional de Estadística y Censos. (2019). 4° Encuesta Nacional de Factores de Riesgo: Resultados definitivos (1.ª ed.). https://www.indec.gob.ar

Russo, M. P., Grande-Ratti, M. F., Burgos, M. A., Molaro, A. A. y Bonella, M. B. (2023). Prevalence of diabetes, epidemiological characteristics and vascular complications. Archivos de Cardiología de México, 93(1), 30–36. https://doi.org/10.24875/ACM.21000410

Guerra-Hernández, A., Mondragón-Becerra, R. y Cruz-Ramírez, N. (2008). Explorations of the BDI multi-agent support for the knowledge discovery in databases process. https://www.researchgate.net/publication/236373188_Explorations_of_the_BDI_Multi-agent_support_for_the_Knowledge_Discovery_in_Databases_Process

Gagliardino, J. J., de la Hera, M., Siri, F. y Grupo de Investigación de la Red QUALIDIAB. (2001). Evaluación de la calidad de atención de pacientes diabéticos en América Latina. Revista Panamericana de Salud Pública, 10(5), 309–317. https://doi.org/10.1590/s1020-49892001001100003

Dagliati, A., Marini, S., Sacchi, L., Cogni, G., Teliti, M., Tibollo, V., De Cata, P., Chiovato, L. y Bellazzi, R. (2018). Machine learning methods to predict diabetes complications. Journal of Diabetes Science and Technology, 12(2), 295–302. https://doi.org/10.1177/1932296817706375

Nathan, D. M., Kuenen, J., Borg, R., Zheng, H., Schoenfeld, D. y Heine, R. J. (2008). Translating the A1C assay into estimated average glucose values. Diabetes Care, 31(8), 1473–1478. https://doi.org/10.2337/dc08-0545

Friedewald, W. T., Levy, R. I. y Fredrickson, D. S. (1972). Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. Clinical Chemistry, 18(6), 499–502. https://doi.org/10.1093/clinchem/18.6.499