

La ciencia de datos en la detección de diabetes

**Diego Alexis Morales Martínez,
Iván Andrés Burgos Castro y
Pedro Javier García Ramírez**

¿Te imaginas ser diagnosticado por una computadora? ¿Crees que conociendo tu expediente médico podría decirte si tienes diabetes o no? Esto no es el futuro, es el presente y toca hablar de él. Actualmente nos encontramos en una era donde la digitalización de la información, también conocida como *industria 4.0*, se encuentra consolidada. Esto significa que gran parte de la información que anteriormente era almacenada en expedientes formulados a partir de tinta y papel, hoy está en servidores, los cuales son equipos de *hardware* (como una computadora o un equipo que contenga componentes electrónicos físicos, como una impresora) dedicados al almacenamiento de recursos digitales.

La medicina no ha estado exenta de esto, históricamente antes de la introducción de las computadoras, los médicos generaban expedientes clínicos de los pacientes a mano, luego empleando máquinas de escribir mecánicas, las cuales plasmaban las letras y números a partir de mecanismos conformados por palancas, resortes y engranajes.

Con el paso del tiempo, el área médica fue adoptando herramientas digitales hasta que hoy en día, las recetas son generadas a partir de computadoras y la entrega de resultados de análisis de sangre de laboratorio se hace vía correo electrónico, o la gestión del inventario de medicamentos se realiza mediante aplicaciones específicas. Pero eso no es lo único que se ha incorporado.

Es posible que hayas escuchado el término de *machine learning* (aprendizaje automático), pero si no, este artículo lo aborda de una manera tal que pueda ser entendido su funcionamiento y su importancia aplicado en campos como la medicina,

por ejemplo, en la detección de diabetes en todo el mundo.

Conociendo un poco acerca de la diabetes *mellitus*

De acuerdo con la Organización Panamericana de la Salud, la diabetes mellitus es una enfermedad que se caracteriza por altos niveles de glucosa en la sangre los cuales, de mantenerse por un tiempo prolongado, provocan daños en distintas partes del cuerpo como en los ojos, los riñones y los nervios, entre otras. Esta enfermedad se divide en tres tipos: tipo 1, tipo 2 y gestacional. La *diabetes tipo 1* es una afección crónica que no se puede prevenir. En pacientes con este diagnóstico (4%), generalmente niños o jóvenes, su páncreas produce poca o nula insulina por sí mismo. Sin embargo, el padecimiento más común es la *diabetes tipo 2*, presente en el 95% de los pacientes con esta afección, en su mayoría adultos. Esta se presenta cuando el cuerpo desarrolla una resistencia a la insulina o no produce la suficiente.

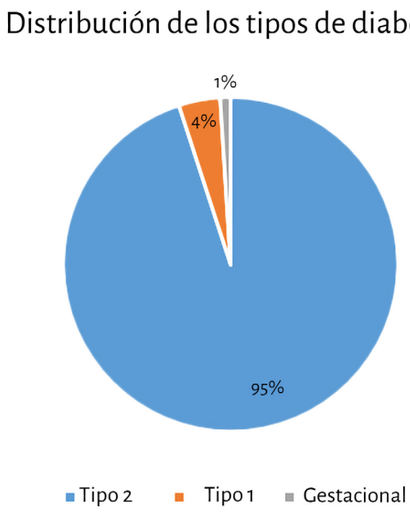


Figura 2. Distribución de los tipos de diabetes mellitus en la población con este padecimiento.

A diferencia de la tipo 1, la tipo 2 es prevenible. Por otro lado, la diabetes gestacional ocurre durante el tiempo de embarazo (1%). Esto se muestra en la Figura 2.

Se estima que 112 millones de adultos tienen diabetes en América y 43 millones de diabéticos en el rango de 30 años o más no reciben tratamiento. En México, de acuerdo con la Encuesta Nacional de Salud y Nutrición 2022, la prevalencia de la diabetes fue de 18.3% lo que equivale a 14.6 millones de habitantes. Ante este escenario, destaca la importancia de incorporar acciones innovadoras en la detección temprana y la prevención.

La mayoría de nosotros tiene algún familiar, amigo o conocido que sufre de este padecimiento y las complicaciones de salud que enfrentan.

Pero a pesar de conocer su condición, pocas acuden a una valoración médica por el miedo de 1) enfrentar su estado de salud ante la diabetes y 2) realizar cambios en sus hábitos de vida, por lo que la detección muchas veces se realiza ya en un estado avanzado de la enfermedad cuando se presenta un malestar. Además, existe un desconocimiento generalizado de otros factores que provocan la diabetes, más allá de la genética. A pesar de que está comprobado que vivir sanamente previene la diabetes tipo 2, casi nadie lo pone en práctica o le toma la suficiente importancia.

¿La ciencia de datos puede detectar casos de diabetes?

La ciencia de datos es una disciplina muy popular en los últimos años. Esta ciencia trabaja con la información: la recopila, la depura, la analiza y la visualiza para fundamentar decisiones informadas en cualquier área que imagines. Dentro de este proceso bastante general tenemos que resaltar el concepto de *modelado*, ya que aquí es donde se realiza la "magia": los expertos toman los datos y con ellos hacen predicciones. Contextualizándonos en el tema, existen registros médicos de pacientes con diabetes con características particulares: los niveles de glucosa, registros de familiares que padecieron diabetes, niveles de insulina, etcétera. Durante años, esta información ha permanecido guardada en archivos físicos.

Hoy surge una pregunta interesante: ¿es posible saber si alguien tiene diabetes solo con esos datos médicos, sin necesidad de una prueba de sangre? Aunque pueda parecer sorprendente, la respuesta es sí.

Ciencia y tecnología trabajan de la mano para que este tipo de predicciones formen parte de la vida diaria, tal como ocurrió con la digitalización de los expedientes clínicos.

La Figura 3 muestra el problema de clasificación para diagnóstico de diabetes *mellitus*.

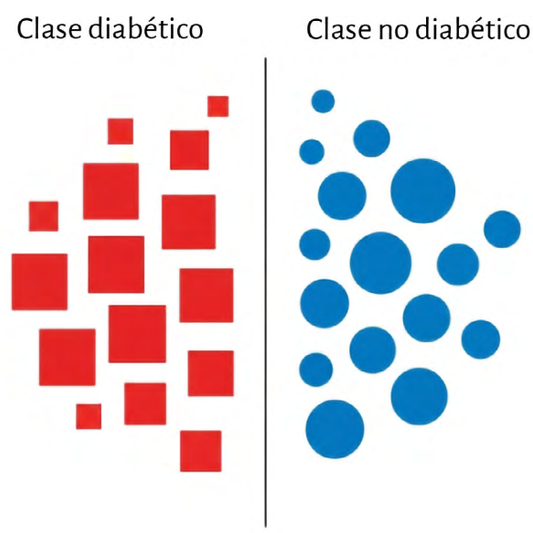


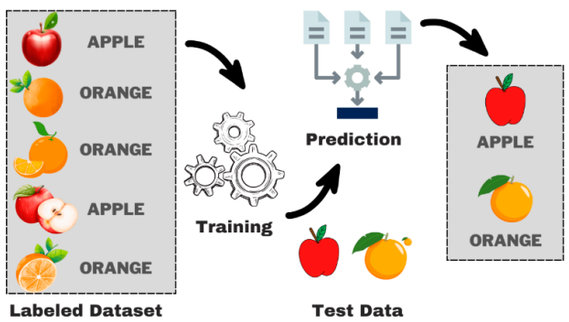
Figura 3. Problema de clasificación para diagnóstico de padecimiento de diabetes mellitus. Imagen generada con Gemini 2.5 Flash.

Enseñando a las máquinas ¿Qué es el *machine learning*?

Así como enseñamos a un niño el nombre de las frutas para que las identifique intuitivamente por su forma, color y sabor, podemos enseñar a una computadora el perfil de una persona con diabetes a través de sus datos personales como edad, el nivel de glucosa en sangre, su presión arterial, así como su nivel de insulina, por mencionar algunos factores. Cuando la computadora ha practicado con distintos casos, se le entrega un paciente nuevo sin etiqueta alguna que indique "diabético" o "no diabético". Con base en los datos que se tienen, los cuales se almacenan en algo conocido como conjunto de datos (*dataset*), el sistema predice si esa persona padece o no la enfermedad.

Esto es lo que se conoce como aprendizaje máquina, aprendizaje automático o *machine learning*.

La Figura 4 muestra este proceso en formato visual.



Fgura 4. Representación del funcionamiento del aprendizaje automático supervisado. Priyadarsini D. (2024) en https://www.linkedin.com/pulse/supervised-learning-priyadarsini-dhanasekhar-g1qic/

En este aprendizaje existen diferentes algoritmos que tienen, al menos, dos finalidades. La primera es identificar a qué grupo pertenece un elemento, lo que se conoce como *problema de clasificación*, y el segundo, es calcular un valor numérico aproximado, *problema de regresión*. Como el diagnóstico de diabetes consiste en determinar si una persona es diabética o no, se hace referencia a un problema de clasificación. Para explicarlo mejor, analicemos dos algoritmos básicos:

Regresión logística: imagina una especie de semáforo que, según la probabilidad, dice "sí" o "no". Si la probabilidad supera el 50%, el modelo marca al paciente como diabético; si no, lo clasifica como sano.

Árbol de clasificación: funciona bajo un esquema de preguntas y respuestas. Cada paso pregunta algo, por ejemplo: "¿su nivel de glucosa es mayor a 120?". Dependiendo de si la respuesta es sí o no, se sigue otra rama hasta llegar al resultado final para clasificar si es diabético o no diabético.

Si bien el texto se enfoca en la clasificación de diabetes con categorías (diabético y no diabético) a esto se le conoce como *clasificación binaria*, no obstante, es posible tener más de dos clases, por ejemplo: no diabético, prediabético y diabético, a esto se le conoce como *clasificación multiclase* donde se pueden tener tantas clases como se requiera, sobre esto se exhorta al lector a profundizar más.

¿El *machine learning* busca reemplazar al diagnóstico médico tradicional?

La respuesta corta y directa es: no. Estas herramientas no reemplazan al personal médico, sino que sirven como apoyo basado en la teoría y conocimiento clínico existente. En cuanto a su fiabilidad, debemos recordar que ni siquiera las computadoras son infalibles, estas son programadas por humanos, lo que a su vez implica un error asociado.

Los modelos funcionan a partir de un conjunto de datos divididos típicamente en dos subconjuntos de información: uno para aprender (entrenamiento) y otro para comprobar lo aprendido (prueba).

Por lo general, entre el 70% y 80% de los datos se usan para entrenar, y el resto para validar.

La evaluación del modelo se realiza brindando toda la información requerida para la predicción como se mencionó por ejemplo la edad, el nivel de glucosa en sangre, etcétera (esta viene del conjunto de prueba) y se observa la predicción del modelo, que se compara contra el valor real, entonces se mide qué tan bien clasifican los casos. Un valor entre 0.8 y 1 significa que el modelo predice con buena precisión y si baja de 0.8, es mejor explorar otros enfoques, revisar estudios previos que puedan mejorar su rendimiento, o reflexionar sobre si los datos realmente tienen una correlación con el fenómeno que se está tratando de predecir.

¿Cuáles son los avances en el diagnóstico y predicción de diabetes usando *machine learning*?

Cuando los científicos investigan un tema, siempre revisan primero lo que otros ya han descubierto: qué resultados obtuvieron, cómo lo hicieron y qué contribuyeron. Este conjunto de investigaciones previas se conoce como "*estado del arte*". En el campo de la diabetes, existen miles de estudios publicados que intentan predecir esta enfermedad usando datos de pacientes, con la esperanza de detectar lo más temprano posible a las personas en riesgo o que, sin saberlo aún, ya son diabéticas.

Por ejemplo, Bhat y su equipo en 2023 probaron tres métodos para predecir diabetes: dos de ellos similares a los que ya mencionamos (modelo de probabilidades y árbol de decisiones) y uno más llamado *Gradient Boost* el cual funciona a partir de mejorar el resultado anterior. Esto es, supongamos que el primer intento predijo bien algunos casos, pero en otros no por distintas razones; sin embargo, el segundo intento se enfoca en dónde falló y de esta manera se hace una sumatoria de las contribuciones de todos los intentos hasta tener un modelo mejorado.

Después de organizar bien los datos y eliminar información incompleta o problemática, estos métodos lograron predecir correctamente en más del 80% de los casos. El método del árbol de decisiones fue el mejor, acertando en el 91% de los casos. Este estudio no solo buscó un buen método de predicción, sino que también mostró lo importante que es tener datos limpios y organizados, ya que, si la información tiene errores o faltan datos, los modelos no funcionan correctamente.

Otro estudio, es el de Lee y su equipo en 2025, investigó qué factores aumentan el riesgo de diabetes en personas mayores de 60 años en Seúl, Corea del Sur. Usando métodos de aprendizaje automático, descubrieron que la presión arterial alta, la edad y el ritmo cardíaco eran algunos de los factores fundamentales para predecir la diabetes en este grupo. ¿Por qué es importante este estudio? Además de identificar qué factores aumentan la predisposición para padecer diabetes en las personas mayores, abre la puerta a que se realicen investigaciones similares en otras partes del mundo para comparar resultados y ver si hay diferencias o similitudes entre distintas poblaciones. Como estos estudios, hay miles de trabajos disponibles en internet.

La información está al alcance de todos, y ofrece la oportunidad de profundizar e incluso proponer nuevas ideas que mejoren lo que ya se ha hecho.

Conclusión

Ahora sabemos el panorama global de la diabetes mellitus alrededor del mundo y el gran problema al que se enfrenta la humanidad para su prevención y tratamiento. También exploramos cómo la ciencia y tecnología han desarrollado soluciones que apoyan a los profesionales médicos, como fue desde la escritura a mano, el uso de la máquina de escribir hasta las computadoras y los expedientes digitales. Por otro lado, aprendimos la definición de aprendizaje automático y cómo esta nueva tecnología busca apoyar a los profesionales para el diagnóstico de diabetes más allá de reemplazarlos. Aún queda mucho trabajo por delante, pero los avances han sido significativos y los esfuerzos científicos muestran tendencias hacia estas tecnologías.

Se exhorta a los lectores a informarse más, así como a realizarse chequeos constantes de diabetes y, conociendo cuáles son los factores de riesgo, hacer cambios paulatinos en su estilo de vida.

Referencias

Basto Abreu, A., López Olmedo, N., Rojas Martínez, R., Aguilar Salinas, C. A., Moreno Banda, G. L., Carnalla, M., Rivera, J. A., Romero Martínez, M., Barquera, S. & Barrientos Gutiérrez, T. (2023). Prevalencia de prediabetes y diabetes en México: Ensanut 2022. *Salud Pública de México, 65*, s163-s168. https://doi.org/10.21149/14832

Bhat, S. S., Banu, M., Ansari, G. A. & Selvam, V. (2023). A risk assessment and prediction framework for diabetes mellitus using machine learning algorithms. *Healthcare Analytics, 4*, 1-12. https://doi.org/10.1016/j.health.2023.100273

Blanco Naranjo, E. G., Chavarría Campos, G. F. & Garita Fallas, Y. M. (2021). Estilo de vida saludable en diabetes mellitus tipo 2: Beneficios en el manejo crónico. *Revista Médica Sinergia, 6*(2). https://doi.org/10.31434/rms.v6i2.639

Lee, H., Park, M. B. & Won, Y. J. (2025). AI machine learning-based diabetes prediction in older adults in South Korea: Cross-sectional analysis. *JMIR Formative Research, 9*, 1-9. https://doi.org/10.2196/57874

Organización Panamericana de la Salud. (s. f.). Diabetes. https://www.paho.org/es/temas/diabetes

