dx.doi.org/10.24254/CNIB.17.85

Predicción de Diagnóstico de Diabetes Mellitus utilizando Máquinas de Soporte Vectorial en Pacientes de Baja California

B. Benítez[1], C. Castro[1], R. A. Castañeda-Martínez[1], A. Abaroa[1]
[1]Facultad de Ingeniería, Arquitectura y Diseño, Ensenada Baja California, México.

Resumen— **Los factores de mayor importancia para el diagnóstico de diabetes mellitus (DM) son edad, índice de masa corporal (IMC) y la concentración de glucosa en sangre. Baja California es el estado con el mayor porcentaje de diagnósticos positivos al año (23.2 %). El diagnóstico de la DM por parte de un médico resulta complicado, debido a que intervienen varios factores en la enfermedad, además de que el diagnóstico está sujeto al error humano. Un examen de sangre no proporciona información suficiente para llevar a cabo un diagnóstico correcto de la enfermedad. Se implementó una máquina de soporte vectorial (SVM) para predecir el diagnóstico de la DM basado en los factores mencionados en pacientes de Baja California. Las clases de la variable de salida son tres: sin diabetes, con predisposición a diabetes y con diabetes. Se obtuvo un SVM con una exactitud de 99.2 % con pacientes mexicanos y una exactitud de 65.6 % con un conjunto de datos de pacientes de un origen étnico diferente. Nuevos datos de pacientes mexicanos e incorporación de otros atributos son necesarios para aumentar la exactitud y capacidad de predicción del modelo, así como la capacidad de extrapolar predicciones a la población mexicana en general.**

Palabras clave—**Diagnóstico médico, diabetes mellitus, informática médica, *machine learning*, máquinas de soporte vectorial.**

I. Introducción

La diabetes mellitus (DM), por definición de la Organización Mundial de la Salud (OMS), es una enfermedad crónico-degenerativa causada por la producción insuficiente de insulina en el páncreas o por la incapacidad del organismo de utilizar eficazmente la insulina producida, teniendo como principal indicador la hiperglucemia (aumento de glucosa en la sangre).

En su etapa inicial, la DM generalmente no produce síntomas notorios, pero al ser detectada tardíamente y no recibir tratamiento adecuado puede conllevar serias complicaciones de salud, tales como: infarto al corazón, ceguera, falla renal, amputación de extremidades e incluso muerte prematura, ésta última representa una disminución en la esperanza de vida de entre 5 y 10 años respecto al promedio sano, convirtiéndola en la primera causa de muerte en México en el año 2011 [1, 2]. Un diagnóstico temprano de la enfermedad puede aumentar significativamente la calidad de vida del paciente.

De acuerdo a datos de la Secretaria de Salud de México, en el 2011, el 1.5% de las personas que se realizaron una prueba de detección obtuvieron un diagnóstico positivo, siendo Baja California (23.2%), Jalisco (19.1%), Estado de México (17%) y Zacatecas (16.3%) los estados con porcentajes más altos de diagnósticos positivos [3].

La Encuesta Nacional de Salud y Nutrición, en Baja California señalo que el 29% de las pruebas de detección resultaron positivas en adultos mayores a 20 años de edad, siendo el bloque con mayor incidencia los adultos, entre 40 - 59 años y 60 o más, para mujeres y hombres respectivamente, que comparado con cifras del 2006 existe un aumento del 7.6% en la incidencia de DM, ubicando al estado en un porcentaje por arriba de la media nacional [4].

Actualmente se invierte cerca del 30% del presupuesto total anual del Instituto Mexicano del Seguro Social (IMSS) al diagnóstico y tratamiento de DM, VIH/SIDA, hipertensión arterial y cáncer, por considerarse una prioridad para el sector salud y siendo un factor clave el diagnóstico temprano de las mismas para asegurar la calidad de vida del paciente [5].

El diagnóstico de la DM resulta complicado debido a que es una enfermedad multifactorial. Para realizar un diagnóstico, el médico debe evaluar los resultados de una prueba del paciente y compararlos con los de pacientes en condiciones similares para analizar previas decisiones. El análisis de los factores que influyen en el diagnóstico se puede ver afectado por el error humano ya que este está sujeto a la interpretación del médico. Otra cuestión importante es que los pacientes que no son diagnosticados no pueden ser tratados, por lo que su calidad de vida puede empeorar considerablemente.

Un examen de sangre no es suficiente por sí solo para llevar a cabo un diagnóstico correcto, ya que no es lo suficientemente discriminante, además de que su interpretación puede diferir entre poblaciones con características distintas. El diagnóstico de DM es aún más difícil debido a la falta de una prueba confiable, de bajo costo y con un alto desempeño que pueda ser universalmente aplicable (o en la población mexicana), y a la baja capacidad de los sistemas de salud de identificar y manejar nuevos casos de diabetes, especialmente en países en vías de desarrollo como México.

Lo anterior pone de manifiesto una ventaja del *machine learning* sobre la capacidad humana en el tópico de diagnóstico médico. Recientemente se ha comenzado a hablar de términos como *medical mining* y de informática médica para referirse a aplicaciones de la computación en la medicina. Estas áreas hacen uso de las herramientas computacionales para el procesamiento de datos médicos y así facilitar su interpretación.

De acuerdo a la referencia [6], no es necesario tomar en cuenta muchos parámetros para llevar a cabo un diagnóstico médico de la DM, que sólo aumentarían innecesariamente la dificultad de la predicción, ya que es posible llevar a cabo la predicción a partir de 5 parámetros clave, los cuales son medibles y no están sujetos a la interpretación humana ni al sesgo del paciente. La edad y la concentración de glucosa en el plasma sanguíneo son cruciales para una correcta identificación de la enfermedad.

Las máquinas de soporte vectorial (SVM, por sus siglas en inglés) son un conjunto de algoritmos computacionales capaces de identificar y representar relaciones no lineales en sistemas complejos [7]. Las SVM han tenido un desempeño efectivo en problemas tanto de regresión como de clasificación.

Varias técnicas se han utilizado para realizar una predicción en el diagnóstico de diabetes en pacientes [6], [8]–[12]. Sin embargo, no se encontraron referencias de trabajos de *machine learning* aplicado a la predicción de DM en pacientes mexicanos. Las SVM se han validado en trabajos previos como un algoritmo efectivo para predicción en diagnósticos médicos [6], [11]. Obtener una predicción del diagnóstico médico de la DM en pacientes permite una atención temprana al control de la enfermedad, además de que reduce los tiempos de diagnóstico y representa un ahorro económico para el sistema de salud y el paciente. En este trabajo, se propuso usar el ya conocido algoritmo SVM para analizar un conjunto de datos basado sólo en variables medibles de pacientes para llevar a cabo una predicción, siguiendo la propuesta de simplicidad de las entradas de [6].

II. Metodología

Se tomaron datos de índice de masa corporal (IMC), edad, concentración de glucosa en sangre (CG) y previo diagnóstico médico de DM (sin diabetes, predisposición a diabetes y con diabetes) de 500 pacientes del Hospital General de Ensenada, Baja California. En la Figura 1 se puede apreciar la relación entre el diagnóstico de los pacientes con el nivel de glucosa (1a) y con el IMC (1b). Se utilizó un 75% de este conjunto de datos para entrenar a un clasificador SVM no lineal para predecir el diagnóstico de DM en nuevos pacientes y el 25% restante para su validación. La edad, IMC y glucosa en sangre se fijaron como los indicadores y por lo tanto entradas para el SVM, mientras que el diagnóstico es la variable a predecir (clasificar). El kernel utilizado tanto para el entrenamiento como para la predicción fue de base radial. El conjunto de datos, así como los códigos usados para este trabajo, se pueden encontrar en el repositorio https://github.com/ruben1294/inteligluc; todos los códigos fueron escritos en el lenguaje de programación R versión 3.3.

Para validar el modelo computacional se utilizó el método de validación cruzada 10-*fold*. Como métricas de desempeño del SVM se utilizaron la exactitud, sensibilidad, especificidad, los valores positivo y negativo de predicción y la matriz de confusión, parámetros comúnmente utilizados en predicción de diagnósticos médicos.

El diagnóstico de cada paciente se agrupo en un conjunto de acuerdo a los siguientes criterios establecidos por la OMS; $CG \geq 126$ corresponde a un paciente con diabetes, $99 < CG < 126$ es un paciente con posible diabetes y $CG \leq 99$ un paciente sin diabetes.

Adicionalmente, se utilizaron el conjunto de datos *Pima Indians diabetes*, el cual se encuentra disponible en el repositorio de *machine learning* de la University of California - Irvine, y conjunto adicional correspondiente a pacientes de Tijuana, Baja California, para probar la capacidad de predicción del modelo entrenado con datos de pacientes de Baja California.

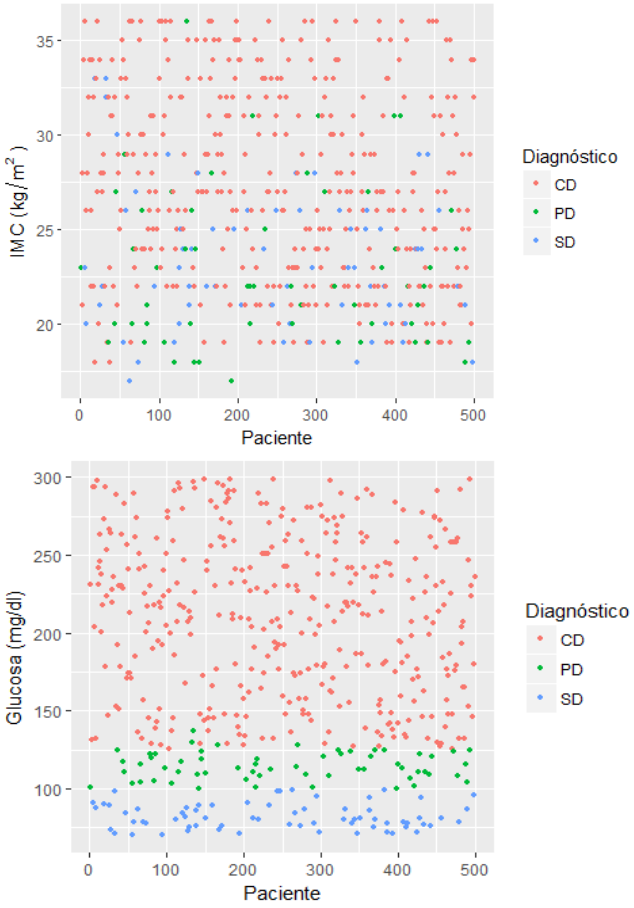

Fig. 1 a) Arriba: Relación entre el diagnóstico y valor de IMC; b) Abajo: Relación entre diagnóstico y nivel de glucosa en sangre.

III. RESULTADOS

La Tabla I muestra las métricas de desempeño del modelo. La Tabla II muestra las métricas de desempeño en la validación del modelo con datos de Pima Indians. La Tabla III muestra la el diagnostico real contra la predicción. En la Figura 2 se muestra la matriz de confusión para el modelo basado en SVM.

TABLA I
MÉTRICAS DE DESEMPEÑO DEL MODELO

Métrica	Porcentaje (%)
Exactitud	95.2
Sensibilidad	94.79
Especificidad	96.55
Valor de predicción positivo	98.91
Valor de predicción negativo	84.84

TABLA II
EXACTITUD PARA LA PRUEBA DE VALIDACIÓN CON UN SEGUNDO CONJUNTO DE DATOS

Métrica	Porcentaje (%)
Exactitud	64.09
Sensibilidad	50.84
Especificidad	70.7
Valor de predicción positive	46.39
Valor de predicción negative	74.26

TABLA III
EXACTITUD PARA LA PRUEBA DE VALIDACIÓN CON UN TERCER CONJUNTO DE DATOS

REAL	PREDICCIÓN
CD	CD
CD	PD
CD	CD
CD	CD
CD	PD
CD	PD

IV. DISCUSIÓN

Se aprecia una fuerte correlación lineal entre la CG en plasma sanguíneo y el diagnóstico de la DM (Fig. 1 a). Se obtuvo un SVM con una exactitud de 95.2 %, lo cual representa un valor aceptable para utilizar esta técnica en el diagnóstico de DM en pacientes de Baja California con la capacidad de que sea aplicada en pacientes de hospitales de todo el país, mejorando el proceso de detección de la enfermedad de forma rápida, económica y acertada.

Al probar el SVM en un conjunto de datos diferente al que se utilizó para el entrenamiento, tomando en cuenta las distintas características de la población diagnosticada (Tabla II). Se obtuvo una exactitud aceptable (64.09 %), lo cual valida el modelo desarrollado aun cuando existan diferencias entre los pacientes de ambos conjuntos de datos, principalmente de naturaleza étnica. Cabe mencionar que el conjunto de datos de Pima Indians toma como criterio una concentración de glucosa mayor a los 200 mg/dL (después de 2 horas transcurridas de la ingesta de un alimento) para agrupar al paciente en el grupo diagnosticado con diabetes, comparado con el criterio usado para este trabajo, la diferencia entre ambos es significativa y por ende la exactitud se ve afectada al momento de validar el modelo.

A pesar de que el diagnóstico real corresponde a pacientes con diabetes, el modelo asigna al tercer conjunto de datos con predisposición a diabetes (Tabla III), esto debido a que las concentraciones de glucosa en la sangre de estos pacientes se encuentran en el umbral en el límite de los criterios para ser asignados entre un grupo y otro. Con lo anterior se confirman las métricas del desempeño del modelo.

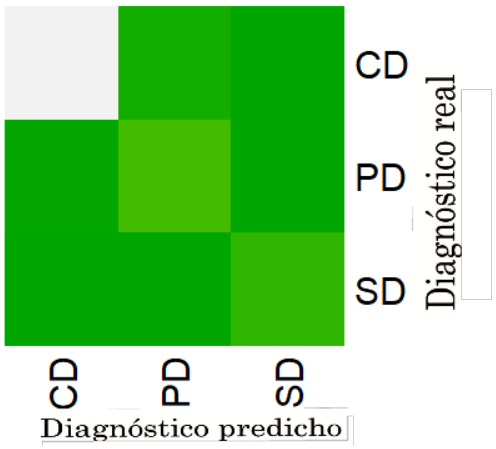

Fig. 2 Matriz de confusión para el modelo construido con el algoritmo Support Vector Machine. SD: sin diabetes, PD: predisposición a diabetes, y CD: con diabetes.

V. CONCLUSIÓN

Se obtuvo un clasificador eficaz de diagnóstico de DM basado en la edad, IMC y CG del paciente. Este clasificador es una herramienta potencial para contribuir a lograr un buen control sobre los nuevos casos de DM en México, además de ser una herramienta económica y universalmente aplicable. Nuevos datos y atributos relacionados con el diagnóstico de la DM son necesarios para probar y mejorar esta técnica.

Como trabajo futuro, es posible aumentar la exactitud y capacidad de predicción del clasificador utilizando diferentes algoritmos, o combinando estos con otras técnicas computacionales como algoritmos genéticos u optimización por enjambre de partículas. Aunado a esto, se puede aumentar el nivel de exactitud incorporando otros parámetros que contribuyan a un correcto diagnóstico, como la concentración de hemoglobina A glucosilada, un marcador biológico de alta importancia, que además da un indicio de la calidad de los cuidados que tiene el paciente para controlar su enfermedad y de su estado de salud [2].

VI. Bibliografía

[1] D. M. Hernández-Ávila, J. Gutiérrez, N. Reynoso-Noverón, "Diabetes mellitus en México. El estado de la epidemia," *Salud Publica México*, vol. 55, no. 2, pp. 129–136, 2013.

[2] Hernández-Romieu Alfonso Claudio, E.-O. Alejandro, H.-U. Nidia, and R.-N. Nancy, "Análisis de una encuesta poblacional para determinar los factores asociados al control de la diabetes mellitus en México," *Salud Publica Mex.*, vol. 53, no. 1, pp. 34–39, 2011.

[3] INEGI, "Estadistica a Proporsito del Día Mundial de la Diabetes," *Día Mund. la Diabetes.*, p. 18, 2013.

[4] Instituto Nacional de Salud Pública, *Encuesta Nacional de Salud y Nurtición: Resultados por Entidad Federativa*. 2012.

[5] Gobierno de la República Mexicana, "Programa Institucional del Instituto Mexicano del Seguro Social 2014-2018," pp. 1–81, 2014.

[6] T. Santhanam and M. S. Padmavathi, "Application of K-Means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis," *Procedia Comput. Sci.*, vol. 47, no. C, pp. 76–83, 2014.

[7] S. Li, H. Zhao, Z. Ru, and Q. Sun, "Probabilistic back analysis based on Bayesian and multi-output support vector machine for a high cut rock slope," *Eng. Geol.*, vol. 203, pp. 178–190, 2016.

[8] T. Zheng *et al.*, "A machine learning-based framework to identify type 2 diabetes through electronic health records," *Int. J. Med. Inform.*, vol. 97, pp. 120–127, 2017.

[9] Shankaracharya, D. Odedra, S. Samanta, and A. S. Vidyarthi, "Computational intelligence in early diabetes diagnosis: A review," *Rev. Diabet. Stud.*, vol. 7, no. 4, pp. 252–261, 2010.

[10] K. V. S. R. P. Varma, A. A. Rao, T. Sita Maha Lakshmi, and P. V. Nageswara Rao, "A computational intelligence approach for a better diagnosis of diabetic patients," *Comput. Electr. Eng.*, vol. 40, no. 5, pp. 1758–1765, 2014.

[11] D. Çalişir and E. Doğantekin, "An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier," *Expert Syst. Appl.*, vol. 38, no. 7, pp. 8311–8315, 2011.

[12] H. Temurtas, N. Yumusak, and F. Temurtas, "A comparative study on diabetes disease diagnosis using neural networks," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 8610–8615, 2009.