

## Clasificación de complicaciones en diabetes mellitus mediante algoritmos genéticos

Mario Daniel Cervantes-Guerrero<sup>1</sup>, Miguel Cruz<sup>2</sup>,  
Adan Valladares-Salgado<sup>2</sup>, Jorge Issac Galván-Tejada<sup>1</sup>,  
Tania A. Gutiérrez-García<sup>3</sup>, Carlos Eric Galván-Tejada<sup>1</sup>

<sup>1</sup> Universidad Autónoma de Zacatecas,  
México

<sup>2</sup> Unidad de Investigación Médica en Bioquímica,  
Centro Médico Nacional Siglo XXI,  
México

<sup>3</sup> Universidad de Guadalajara,  
Centro Universitario de Ciencias Exactas e Ingenierías,  
México

{danielcervantesguerrero, gatejo, ericgalvan}@uaz.edu.mx,  
{adan.valladares, miguel.cruzlo}@imss.gob.mx  
tania.gutierrez@academicos.udg.mx

**Resumen.** La diabetes tipo 2 es una enfermedad metabólica que se caracteriza por la incapacidad de regular los niveles de glucosa en la sangre debido a la resistencia a la insulina y una limitada producción de la misma en el cuerpo, causando graves complicaciones a largo plazo, las cuales incluyen retinopatía, neuropatía y nefropatía. En este trabajo se presenta una clasificación de complicaciones de pacientes con diabetes tipo 2 realizada a partir de una base de datos con información clínica del Hospital de Especialidades del Centro Médico Nacional Siglo XXI, México. Se identificaron las características más relevantes mediante algoritmos genéticos (GA), para luego usarlos como información de entrada para la estimación de un modelo de clasificación basado en SVM (Support Vector Machine). Los resultados muestran la correlación entre las características estudiadas y las complicaciones de los pacientes, con un área bajo la curva ROC (Receiver Operating Characteristic) de 0.69. Así, se propone el uso de ésta combinación de estimadores para el futuro desarrollo de herramientas predictivas y de diagnóstico para la enfermedad.

**Palabras clave:** Máquina de soporte de vectores, algoritmos genéticos, diabetes tipo 2.

### Classification of Complications in Diabetes Mellitus Using Genetic Algorithms

**Abstract.** Type 2 diabetes is a metabolic disease characterized by the inability to regulate blood glucose levels due to insulin resistance and limited insulin production in the body, causing serious long-term complications, which they include retinopathy, neuropathy, and nephropathy. This paper presents a

classification of complications of patients with type 2 diabetes based on a database with clinical information from the Hospital de Especialidades del Centro Médico Nacional Siglo XXI, Mexico. The most relevant features were identified using genetic algorithms (GA), to later use them as input information for the estimation of a classification model based on SVM (Support Vector Machine). The results show the correlation between the characteristics studied and the complications of the patients, with an area under the ROC curve (Receiver Operating Characteristic) of 0.69. Thus, the use of this combination of estimators is proposed for the future development of predictive and diagnostic tools for the disease.

**Keywords:** Support vector machine, genetic algorithms, diabetes type 2.

## 1. Introducción

De acuerdo a la Organización Mundial de la Salud (WHO), aproximadamente más de 422 millones de personas padecían diabetes en 2014; se estima que para el 2030, esta enfermedad será la séptima causa de mortalidad a nivel mundial [15]. La diabetes es una enfermedad que se desarrolla cuando el páncreas no sintetiza o secreta la insulina que el cuerpo humano necesita o cuando la insulina no se usa de manera eficiente [13], lo que provoca un aumento en los valores normales de glucosa en la sangre.

Como consecuencia, este cambio de niveles produce un aumento excesivo en los niveles de glucosa en la sangre, esto se relaciona con complicaciones a largo plazo tales como la disfunción e insuficiencia de órganos, especialmente ojos, riñones, nervios, corazón y vasos sanguíneos [1].

La diabetes tipo 2 es una afección común y grave relacionada con la reducción de la esperanza de vida, representa a más del 90 % de los pacientes con diabetes y causa diferentes complicaciones microvasculares y macrovasculares. A pesar del conocimiento cada vez mayor sobre los factores de riesgo de la diabetes tipo 2 y la evidencia de programas de prevención exitosos, la incidencia y prevalencia de la enfermedad continúa en aumento a nivel mundial.

La detección temprana mediante programas de detección y la disponibilidad de terapias seguras y eficaces reducen la morbilidad y la mortalidad al prevenir o retrasar las complicaciones asociadas a la enfermedad [3]. En otros trabajos se ha identificado que algunas características como el tipo de insulina suministrada al paciente y el tiempo transcurrido desde que el paciente contrajo diabetes, ayudan a identificar el riesgo de retinopatía diabética [21].

En otros estudios, utilizando parámetros no invasivos que no se basan en el laboratorio, pudo predecir con éxito el desarrollo de diabetes e hipertensión en pacientes [5]. En trabajos similares, se muestra cómo la extracción de datos y los métodos computacionales se pueden adoptar de manera efectiva en la medicina clínica para calcular el factor de riesgo de un paciente con diabetes tipo 2 para desarrollar una complicación microvascular [4]. Se han probado diferentes clasificadores para medir su desempeño en la clasificación de pacientes con diabetes mellitus [8, 21, 5].

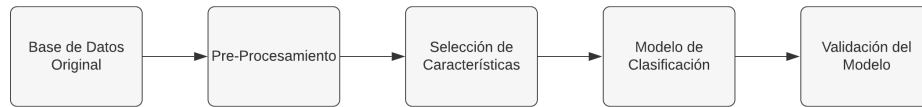


Fig. 1. Diagrama de la metodología.

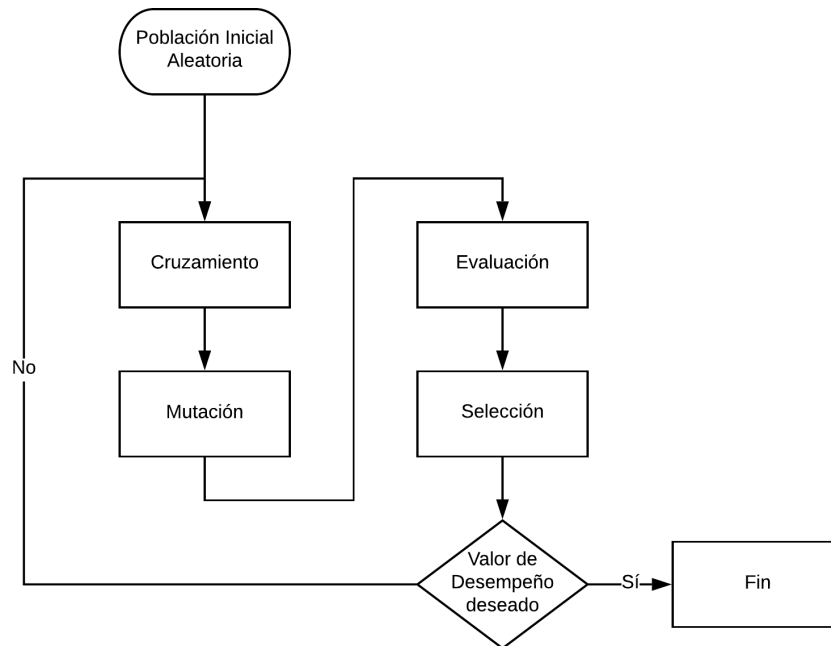


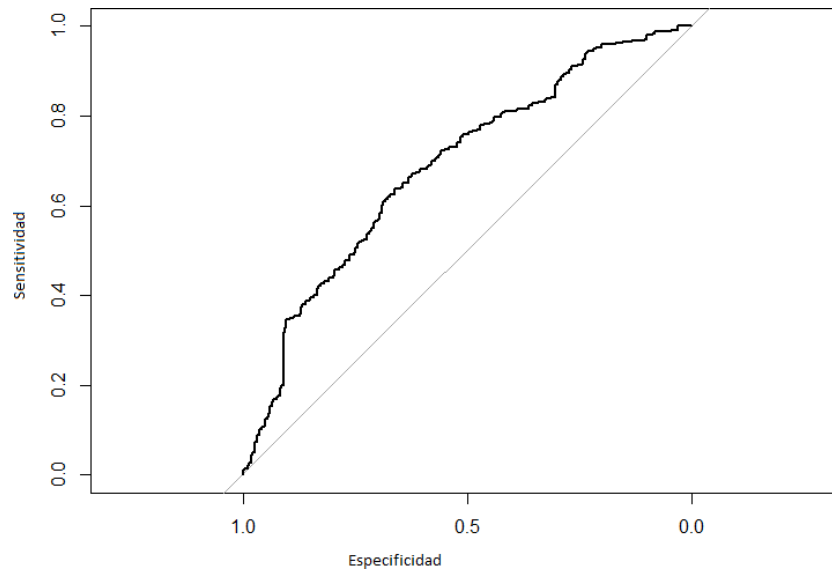
Fig. 2. Diagrama general de un algoritmo genético.

El objetivo de este trabajo es diseñar un modelo predictivo para pacientes diabéticos y anticipar la aparición de complicaciones de la enfermedad, así como detectar parámetros que proporcionen información relevante a un modelo de clasificación usando algoritmos genéticos como enfoque para la selección de características de un conjunto de datos clínicos. El modelo es útil para notificar el riesgo a los pacientes de forma temprana y evitar complicaciones que aumentan el riesgo de muerte.

## 2. Materiales y métodos

La metodología seguida se presenta en la Figura 1. Se inicia con la base de datos pura (sin procesar), que es la información sin ninguna modificación y que contiene todas las variables y observaciones capturadas en el Hospital de Especialidades del Centro Médico Nacional Siglo XXI, del Instituto Mexicano del Seguro Social (IMSS). Luego, el pre-procesamiento, consiste en la selección de muestras y observaciones del conjunto de datos original, la normalización de datos, la eliminación de casos incompletos y la eliminación de características sin suficientes observaciones.





**Fig. 4.** Curva ROC del modelo de clasificación.

Existen variables con suficientes observaciones para su análisis, pero con valores ausentes, un método sencillo para enfrentar este problema es la imputación por promedio aritmético de la variable [24], por lo tanto se realiza la selección de parámetros y observaciones, así como la eliminación de casos incompletos y características sin suficientes observaciones.

Finalmente, los datos totales se dividen en dos conjuntos, uno conformado por el 70 % de los datos, el cual es designado para el entrenamiento del modelo y el 30 % restante de los datos, que se utilizará para probar su eficiencia de clasificación [22, 18].

### 2.3. Selección de características

La selección de características se realiza con algoritmos genéticos [20]. Los GA son algoritmos inspirados en la teoría de la evolución. Los GA buscan una posible solución a un problema específico en una estructura de datos similar a un cromosoma, y luego estas estructuras se recombinan para preservar la información más relevante [23]. El diagrama general de un algoritmo genético se muestra en la figura 2 [19].

El procedimiento parte de una población aleatoria de subconjuntos variables de un tamaño determinado (definidos como cromosomas). Cada cromosoma se evalúa por su capacidad para predecir una variable dependiente, en este caso la presencia o ausencia de complicaciones.

El principio general es reemplazar la población inicial con una nueva población que incluye variantes de cromosomas con mayor precisión de clasificación y repetir el proceso suficientes veces para lograr el nivel deseado de precisión [20]. El objetivo de este proceso es encontrar aquellas características que describan mejor las observaciones en el conjunto de datos en relación con la salida.

**Tabla 2.** Parámetros de evaluación del modelo.

Parámetro de evaluación	Value
Exactitud	0.6528
Sensitividad	0.6325
Especificidad	0.6696
PPV	0.6130
NPV	0.6877
Exactitud balanceada	0.6510

Una vez que se han determinado las características más significativas, estas se pueden utilizar como entrada para el paso de clasificación. R [17] es un lenguaje de programación de licencia abierta y un entorno de software libre para análisis estadístico respaldado por R Foundation. El lenguaje R se usa universalmente entre los científicos de datos para desarrollar modelos estadísticos.

Este lenguaje fue seleccionado por su accesibilidad y su desempeño para el modelado estadístico. Para el análisis de algoritmos genéticos se utilizó GALGO [20], un paquete de software R para entrenar algoritmos genéticos y así seleccionar subconjuntos de variables.

#### 2.4. Modelo de clasificación

Se utilizan las características más relevantes como entrada para un modelo de clasificación de máquina de soporte de vectores (SVM).

Un SVM es un algoritmo de aprendizaje automático supervisado en el cual se entrena un modelo de aprendizaje mediante un ejemplo de datos, los datos deben clasificarse como una de muchas categorías y el algoritmo SVM construye un modelo que predice la categoría de un nuevo ejemplo dado, un SVM construye un hiperplano óptimo para clasificar patrones que maximizan la distancia desde el hiperplano al punto más cercano de cada patrón.

Su objetivo principal es maximizar el margen para que pueda clasificar correctamente los patrones dados; cuanto mayor sea el tamaño del margen, mejor será el rendimiento [16].

#### 2.5. Validación del modelo

La validación de modelos matemáticos y computacionales es un paso importante porque sin este procedimiento, las razones para creer que el modelo funciona de manera óptima son inexistentes [22].

Para medir la capacidad de clasificación del modelo de este estudio, se calculan las métricas estadísticas que normalmente son útiles para tal fin, como son especificidad, sensibilidad, curva ROC, AUC, valor predictivo positivo (VPP), valor predicho negativo (VPN) y precisión equilibrada. La sensibilidad está ligada a la verdadera razón positiva, que en este caso sería la capacidad del modelo para clasificar correctamente a los pacientes con presencia de complicaciones.

**Tabla 3.** Métodos de clasificación.

<b>Método de clasificación</b>	<b>Desempeño</b>
Máquina de soporte de vectores	0.694
Random forest	0.607
Regresión lineal	0.6308
Regresión logística	0.6384

Por otro lado, la especificidad está ligada a la razón negativa, que se referiría a la capacidad del modelo para clasificar correctamente a los pacientes sin complicaciones. La especificidad está ligada a la capacidad de los modelos para clasificar correctamente a los pacientes sin complicaciones.

Los valores graficados de la sensibilidad y la especificidad se denominan curva ROC, y se ha utilizado ampliamente para medir o visualizar el rendimiento de un clasificador junto con su valor de área bajo la curva (AUC), con el fin de seleccionar un punto de operación adecuado, llamado como umbral de decisión [10]. El análisis ROC proporciona dos resultados principales: la precisión de la clasificación de complicaciones de la prueba y el valor de punto de corte óptimo para la prueba.

El valor VPP es la probabilidad de que un paciente con clasificación positiva en el modelo realmente presente una complicación mientras que el valor de VPN es la probabilidad de que un paciente con una prueba de clasificación negativa realmente no tenga una complicación.

La precisión equilibrada se reduce al rendimiento de predicción promedio entre VPP y VPN. Una validación hold-out es un tipo de validación cruzada. Esta separa las observaciones en dos conjuntos, uno utilizado para entrenar al modelo y otro conjunto de datos diferentes para realizar su prueba de validación [2]. Para calcular estas métricas, se utilizó el paquete Caret [9] para el software estadístico R [17].

### **3. Resultados**

Las características utilizadas para este estudio se muestran en el Anexo 1. De estas, las que se eliminaron durante el Pre-procesamiento por presentar una ausencia de observaciones de más del 80 %, están en la Tabla 1. Tampoco se muestran las variables que se refieren al identificador del paciente o su estatus como caso o control.

Como se mencionó anteriormente, se utilizó GALGO para implementar el algoritmo genético, los parámetros de este se configuraron para que tome de forma aleatoria una muestra entre todo el conjunto de datos, el método de clasificación evaluador fue por máquinas de soporte de vectores.

Para reducir la dimensionalidad del modelo se eligió un tamaño de cromosoma de 5 genes, probando la eficiencia de clasificación del modelo usando conjuntos de cinco parámetros hasta encontrar el conjunto con el mayor desempeño.

El ajuste a alcanzar es de 0.7 a lo largo de 100 generaciones. Esto quiere decir que el algoritmo genético repitió 100 veces sus interacciones hasta acercarse lo más posible al valor deseado. En la Figura 3 se muestra la frecuencia de aparición de los parámetros más significativos en todas las generaciones entrenadas del GA.

**Tabla 4.** Parámetros eliminados.

Parámetro	Definición
ACARBO	Miligramos de acarbosa tomados al día
ROSIGLI	Miligramos de rosiglitazona tomados al día
PIOGLI	Miligramos de pioglitazona tomados al día
INSUL	Miligramos de insulina tomados al día
GLIBEN	Miligramos de glibenclamina tomados al día
METFOR	Miligramos de metmorfina tomados al día

La Figura presenta el cromosoma más fuerte luego de 100 generaciones en color negro, las características que explican el gen más fuerte del algoritmo son GLIBENCLAMIDA, TG-B (mg.dl), CHOLU, INSULINA y DBPU. Para la validación hold-out utilizando el 30 % de los datos para evaluar el rendimiento, los sujetos se dividieron aleatoriamente en dos grupos, utilizando un total de 446 observaciones para el conjunto de prueba.

El conjunto de datos de validación hold-out de pruebas contenía 446 observaciones, 192 de los pacientes en el conjunto de datos de prueba presentaron una complicación y 254 no presentaron ninguna complicación. El modelo de predicción tuvo un rendimiento promedio de 118 observaciones correctamente clasificadas como pacientes con complicaciones y 175 observaciones correctamente clasificadas como pacientes sin complicaciones. La curva ROC del modelo se muestra en la Figura 4:

El AUC de la curva ROC de la Fig 4. es de 0.69. Un AUC de 0.5 se traduce en la posibilidad de clasificar aleatoriamente a los pacientes; un valor mayor indica que el modelo tiene capacidad predictiva [11]. Los parámetros de calidad para evaluar el modelo validado se muestran en la Tabla 2.

#### 4. Discusión

La selección de características en el algoritmo genético usada en estudio permitió identificar las principales variables correlacionadas con el desarrollo de complicaciones en pacientes con diabetes tipo 2, por ejemplo, el tipo de hormona que el paciente usa.

En conjunto, el resultado de clasificación y la estimación del modelo permiten resaltar, al paciente o al médico, la prioridad que tienen las variables, por ejemplo, administrar el tipo correcto de la hormona al paciente o monitorear la presión vascular para el desarrollo de síntomas crónicos de gravedad.

Otras de las variables correlacionadas con el desarrollo de complicaciones fueron la presión arterial diastólica, los niveles de triglicéridos y los niveles de colesterol. De esta forma, es posible aportar directamente a la prevención y además al conocimiento del efecto a largo plazo de las características de la base de datos utilizada.

Un valor de 0.5 para AUC indica que la salida del modelo es aleatoria y, por lo tanto, el modelo no tiene capacidad para clasificar a los pacientes [12]. En este estudio, un número importante de pacientes se clasifican correctamente. Los resultados muestran un área ROC bajo la curva de 0.69.



**Tabla 5.** Lista de variables.

Nombre de la variable	Definición
Age	Edad del paciente
Age.Dx	Tiempo desde el diagnóstico de diabetes al paciente
BMI	Índice de masa corporal
Chol.mg.dl	Esteroide que modula la fluidez de las membranas biológicas, se encuentra en los tejidos. Precursor de hormonas esteroides, vitamina D y ácidos biliares. Valor de referencia colesterol total normal inferior a 200 mg / dL.
CHOLU	Concentraciones plasmáticas de colesterol total
Crea.Mg.Dl	Un anhidrido de creatina desechado en la orina. Su valor de referencia es de 0,5 a 1,3 mg / dL.
DBPU	Presión arterial diastólica Valor de referencia: 80 mm Hg
Gen.Bi	El sexo biológico del paciente
GLIBENCLA	La glibenclámda, también conocida como gliburida, es un medicamento que se usa para tratar la diabetes mellitus tipo 2. Presencia o ausencia en el tratamiento del paciente
Glu.Mg-Dl	Monosacárido, hexosa con eficiencia energética. Usado por los tejidos como una forma de energía. Valores de referencia 70 a 100 mg / dL
HA.TX	Hipertensión arterial. Se reduce a la presión arterial. Es una medida de la fuerza que se ejerce contra las paredes de las arterias cuando el corazón bombea sangre al cuerpo. Hipertensión es el término utilizado para describir la presión arterial alta. Presión normal y rango: 120/80 mm Hg.
HbA1c	Hemoglobina glucosilada, es una forma de hemoglobina que está químicamente ligada a un azúcar
HDLc.mg.dl	lipoproteínas de alta densidad. Limpian el colesterol de los tejidos del hígado y lo lleva al hígado. Valor de referencia mayor < 50 mg / dL en mujeres y 40 mg / dL en hombres.
HDLU	Lipoproteínas de alta densidad. Niveles antes de recibir medicamentos para controlarlos
INSULINA	Tipo de insulina que el paciente recibe
LDLc.mg.dl	Lipoproteína de baja densidad que transporta el colesterol desde el hígado a los tejidos extrahepáticos. Valor de referencia inferior a 100 mg / dL
LDLU	Lipoproteínas de baja densidad. Niveles antes de recibir medicamentos para controlarlos
Lípids.Tx	Moléculas orgánicas, insolubles en agua pero solubles en algunos disolventes polares. Tienen funciones estructurales, regulatorias y de reserva energética.
TG-B	Es un lípido que actúa como reserva energética. Valor de referencia de menos de 150 mg / dL
METFORMI	Metformina, es el medicamento más común para el tratamiento de la diabetes tipo 2, especialmente en personas con sobrepeso. hace referencia a si el paciente está siendo tratado con Metformina o no
PC1- PC5	Puntuaciones de PC (PC corresponde a la ascendencia NAM / EUR). Es una herramienta útil para el análisis de datos genéticos
SBPU	Presión arterial sistólica Valor de referencia: 120 mm Hg. La U se refiere al valor ajustado debido al tratamiento.
TG-B (mg/dl)	Niveles de triglicéridos

El número de pacientes que se clasificaron correctamente en este estudio indican que los algoritmos genéticos son una alternativa para realizar una clasificación. Los genes en el cromosoma de mayor desempeño encontrado por el algoritmo coinciden con las variables estudiadas en otros trabajos para predecir el desarrollo de algunas complicaciones tales como cardiopatías o retinopatías [8, 21].

Considerando distintos métodos de clasificación, las máquinas de soporte de vectores tuvieron un valor más alto de clasificación a comparación de la regresión lineal, regresión logística o random forest . Se hace la comparación del desempeño de estos métodos de clasificación considerando las variables del cromosoma más fuerte como características de entrada para cada modelo en la Tabla 3.

## 5. Trabajo futuro

Se planea diseñar una herramienta de predicción con este modelo estadístico, con el fin de hacer una detección temprana y/o predecir. También, se propone utilizar la metodología presentada para entrenar un modelo con un conjunto de datos más grande y considerando un tamaño de cromosoma mayor, con el fin de usar más parámetros al momento de la clasificación y esperando un desempeño aún mejor. Finalmente, se sugiere la comparación de diferentes algoritmos de clasificación para mejorar el rendimiento y además separar el tipo de complicaciones en diferentes grupos.

## Referencias

1. American Diabetes Association: Diagnosis and classification of diabetes mellitus. *Diabetes Care*, vol. 36, pp. 62–67 (2009) doi: 10.2337/dc09-S062
2. Celisse, A., Robin, S.: A cross-validation based estimation of the proportion of true null hypotheses. *Journal of Statistical Planning and Inference*, vol. 140, no. 11, pp. 3132–3147 (2010) doi: 10.1016/j.jspi.2010.04.014
3. Chatterjee, S., Khunti, K., Melanie J. D.: Type 2 diabetes. *The Lancet*, vol. 53, pp. 2239–2251 (2017) <https://www.thelancet.com/clinical/diseases/diabetes-type2>
4. Dagliati, A., Marini, S., Sacchi, L., Cogni, G., Teliti, M., Tibollo, V., De Cata, P., Chiovato, L., Bellazzi, R.: Machine learning methods to predict diabetes complications. *Journal of Diabetes Science and Technology*, vol. 12, no. 2, pp. 295–302 (2017) doi: 10.1177/1932296817706375
5. Farran, B., Channanath, A. M., Behbehani, K., Thanaraj, T. A.: Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: Machine-learning algorithms and validation using national health data from Kuwait-a cohort study. *BMJ Open*, vol. 3, no. 5 (2013) doi: 10.1136/bmjopen-2012-002457
6. Ghorbani, A., Zou, J.: Data Shapley: Equitable valuation of data for machine learning. In: *Proceedings of the International Conference on Machine Learning* (2019) doi: 10.48550/ARXIV.1904.02868
7. Gutierrez, D. D.: *Machine learning and data science: An introduction to statistical learning methods with R*. Technics Publications (2015)
8. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I.: Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal*, vol. 15, pp. 104–116 (2017) doi: 10.1016/j.csbj.2016.12.005
9. Kuhn, M.: *Caret: Classification and regression training*. Astrophysics Source Code Library (2015)
10. Kumar, R., Indrayan, A.: Receiver operating characteristic (ROC) curve for medical researchers. *Indian Pediatrics, Springer Science and Business Media LLC*, vol. 48, no. 4, pp. 277–287 (2011) doi: 10.1007/s13312-011-0055-4

11. Lasko, T. A., Bhagwat, J. G., Zou, K. H., Ohno-Machado, L.: The use of receiver operating characteristic curves in biomedical informatics. *Journal of Biomedical Informatics*, vol. 38, no. 5, pp. 404–415 (2005) doi: 10.1016/j.jbi.2005.02.008
12. Mandrekar, J. N.: Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, vol. 5, no. 9, pp. 1315–1316 (2010) doi: 10.1097/jto.0b013e3181ec173d
13. Nathan, D. M.: Long-Term complications of diabetes mellitus. *New England Journal of Medicine*, vol. 328, no. 23, pp. 1676–1685 (1993) doi: 10.1056/nejm199306103282306
14. Nguyen, L. H., Holmes, S.: Ten quick tips for effective dimensionality reduction. *PLOS Computational Biology*, vol. 15, no. 6, (2019) doi: 10.1371/journal.pcbi.1006907
15. Ogurtsova, K., da Rocha Fernandes, J. D., Huang, Y., Linnenkamp, U., Guariguata, L., Cho, N. H., Cavan, D., Shaw, J. E., Makaroff, L. E.: IDF Diabetes atlas: Global estimates for the prevalence of diabetes for 2015 and 2040. *Diabetes Research and Clinical Practice*, vol. 128, pp. 40–50 (2017) doi: 10.1016/j.diabres.2017.03.024
16. Pradhan, A.: Support vector machine-a survey. *International Journal of Emerging Technology and Advanced Engineering* 2.8, pp. 82–85 (2012).
17. R Core Team: A language and environment for statistical computing. R Foundation for Statistical Computing (2018) [www.R-project.org/](http://www.R-project.org/)
18. Refaeilzadeh, P., Tang, L., Liu, H.: Cross-validation. *encyclopedia of database systems*, pp. 532–538 (2009) doi: 10.1007/978-0-387-39940-9\_565
19. Scrucca, L.: GA: A package for genetic algorithms in R. *Journal of Statistical Software, Foundation for Open Access Statistic*, vol. 53, no. 4 (2013) doi: 10.18637/jss.v053.i04
20. Trevino, V., Falciani, F.: GALGO: An R package for multivariate variable selection using genetic algorithms. *Bioinformatics*, vol. 22, no. 9, pp. 1154–1156 (2006) doi: 10.1093/bioinformatics/btl074
21. Tsao, H. Y., Chan, P. Y., Su, E. C.: Predicting diabetic retinopathy and identifying interpretable biomedical features using machine learning algorithms. *BMC Bioinformatics*, vol. 19, no. S9 (2018) doi: 10.1186/s12859-018-2277-0
22. Vanslette, K., Tohme, T., Youcef-Toumi, K.: A general model validation and testing tool. *Reliability Engineering and System Safety*, vol. 195 (2020) doi: 10.1016/j.res.2019.106684
23. Whitley, D.: A genetic algorithm tutorial. *Statistics and Computing*, Springer Science and Business Media LLC, vol. 4, no. 2 (1994) doi: 10.1007/bf00175354
24. Zhang, Z.: Missing data imputation: Focusing on single imputation. *Annals of Translational Medicine*, vol. 4, no. 1 (2016) doi: 10.3978/j.issn.2305-5839.2015.12.38