Predicción inteligente de diabetes mellitus tipo 2 usando R y mapas cognitivos difusos *

Kenia Hoyos *Clínica Salud Social, Sincelejo, Colombia*
Rander Ruíz *Universidad de Antioquia, Caucasia, Colombia*
William Hoyos *Universidad Cooperativa de Colombia, Montería, Colombia*

Resumen: El diagnóstico tardío de la diabetes mellitus tipo 2 (DM2) conlleva al aumento de las tasas de morbilidad y mortalidad a nivel mundial. El uso de herramientas computacionales para la predicción de DM2 podría acelerar el diagnóstico. Por tanto, el objetivo de la presente investigación fue desarrollar un modelo de predicción de DM2 basado en un mapa cognitivo difuso (MCD) implementado en R. El MCD puede predecir la DM2 con un 99% de exactitud, 100% de sensibilidad y 98% de especificidad y un indice Kappa de 0.98. Los resultados demuestran la capacidad del modelo para predecir y evaluar el comportamiento de las variables de interés en la DM2.

Keywords: diabetes mellitus tipo 2, mapa cognitivo difuso, diagnóstico, predicción

Introducción

En los últimos años, la diabetes ha incrementado su prevalencia convirtiéndose en un problema de salud pública a nivel global (World Health Organization, 2023). Se calcula que en 2021 había 537 millones de personas con diabetes en el mundo y se produjeron 6,7 millones de muertes (International Diabetes Federation, 2021). Más del 95% de la población con diabetes corresponde al tipo 2 y un diagnóstico tardío puede generar complicaciones que llevan a la muerte (World Health Organization, 2023). Para abordar esta problemática se han desarrollado estudios de predicción de diabetes, empleando técnicas de aprendizaje automático implementadas en R, como árboles de decisión Kalange et al. (2022) y Naive Bayes Marathe, Gawade and Kanekar (2021). Estos modelos obtuvieron una buena exactitud, sin embargo, no evaluaron el comportamiento de las variables. Por lo tanto, nuestro objetivo fue desarrollar un modelo de predicción inteligente de DM2 basado en MCD implementado en R que permite analizar la evolución de las variables usando gráficas de iteraciones simuladas. La siguiente sección describe la metodología empleada, luego se presentan los resultados y la ultima sección concluye el artículo.

Metodología

Utilizamos la base de datos de DM2 del Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales de la India (PIMA) (Smith et al., 1988). La información corresponde a 768 pacientes (500 sin DM2 y 268 con DM2) y 9 variables (ver Tabla 1). Aplicamos la técnica SMOTE (Chawla et al., 2002) para balancear las clases. Para el entrenamiento y validación del modelo empleamos el 70% de los datos y el 30% restante, se utilizó para las pruebas de rendimiento. Construimos un MCD, el cual está compuesto por nodos que representan conceptos y flechas que representan las

*Autor de contacto: william.hoyos@campusucc.edu.co.

influencias o relaciones entre ellos (Kosko, 1986). Para encontrar las relaciones entre los conceptos usamos un algoritmo genético. Posteriormente, realizamos un proceso de inferencia usando el paquete de R (R Core Team, 2019) llamado *fcm* (Dikopoulou and Papageorgiou, 2017). El mejor modelo estaba conformado por los siguientes hiperparámetros, para algoritmo genético: *población inicial* = 200, *probabilidad de cruce* = 0.5, *probabilidad de mutación* = 0.5; para el MCD, *función de activación* = *sigmoidea*, *función de inferencia* = *rescaled*. Finalmente, el modelo seleccionado se aplicó al conjunto de datos de prueba (30%) para evaluar su desempeño. La figura 1 muestra el código principal de R para la predicción de DM2 y evaluación de variables involucradas. Después de importar las librerías necesarias, se carga la matriz de pesos del modelo y el conjunto de datos de prueba (30%) como se muestra en la sección A de la figura 1. El desempeño predictivo del modelo fue evaluado a través de métricas como exactitud, sensibilidad, especificidad y el índice de Kappa como se muestra en la sección B de la figura 1. Por último, realizamos iteraciones simuladas usando el paquete *ggplot2* (Wickham, 2016) para visualizar el comportamiento de las variables involucradas en la predicción. La sección C de la figura 1 muestra el código para generar esta gráfica.

Tabla 1: Variables incluidas en el conjunto de datos PIMA

Concepto	Nombre de la variable
C1	Número de embarazos
C2	Glucosa plasmática
C3	Presión sanguínea diastólica
C4	Espesor del pliegue cutáneo del tríceps
C5	Insulina sérica
C6	Índice de Masa Corporal (IMC)
C7	Función del pedigrí de diabetes
C8	Edad
C9	Diagnóstico

Resultados

El modelo obtuvo una exactitud del 99%, sensibilidad del 100% y especificidad del 98% en la predicción de DM2. Con relación al grado de concordancia entre la predicción de nuestro modelo y el diagnóstico real, se obtuvo un índice Kappa de Cohen de 0,98, un valor casi perfecto según Landis and Koch (1977) que indica que la fiabilidad del modelo es alta. Con respecto al comportamiento de las variables, la figura 2 muestra la simulación de un caso con antecedentes de embarazo, presión diastólica e IMC alterado. Conceptos como la concentración de glucosa, niveles de insulina sérica y espesor del pliegue cutáneo del tríceps, que no estaban activos inicialmente, posteriormente se activan y aparecen como curvas que incrementan sus valores a través de las iteraciones. De esta manera, estos resultados representados en forma gráfica le permiten al profesional médico realizar un análisis mas completo de la DM2 para cualquier paciente.

A

```r
#Leemos la base de datos de prueba
pima_test_dataset <- read.arff("diabetes_test_dataset.arff")

#Creamos un data frame con una fila de ceros (0) para rellenar el dataframe
inference_results <- data.frame(matrix(ncol = 9, nrow = 0))
#Le colocamos nombres a las columnas
names(inference_results) <- names(pima_test_dataset)

#Convertimos todas las columnas a tipo numerico
#pima_test_dataset_2 <- pima_test_dataset %>% mutate_if(is.factor, as.numeric)

#Convertimos las categorías de la clase
#pima_test_dataset_2$class <- ifelse(pima_test_dataset_2$class == 2,1,0)

for(i in 1:nrow(pima_test_dataset)){
        cat("Paciente: ", i)
        results <- fcm_infer(activation_vec = pima_test_dataset[i,],
                             weight_mat = adj_matrix,
                             infer = "rc",
                             transform = "s",
                             iter = 200)
        inference_results <- rbind(inference_results, results$values[results$iter_convergencia,
})
```

B

```r
#Extraemos las predicciones de inference_results
predicciones <- inference_results[,"c9"]
predicciones_2 <- ifelse(predicciones <= 0.5, 0, 1)

#Agregamos el vector de predicciones al dataframe de pacientes
resultados_finales <- data.frame(pima_test_dataset$c9, predicciones_2)
names(resultados_finales) <- c("Diagnostico", "Predicción")

#Sacamos la matriz de confusión
(matriz_confusion <- table(resultados_finales$Diagnostico,
resultados_finales$Predicción))
#Calculamos la exactitud
accuracy = 0
for (i in 1:nrow(matriz_confusion)){
        accuracy = accuracy + matriz_confusion[i, i] / sum(matriz_confusion)
}

#Calculamos la exactitud, sensibilidad, especificidad y el indice Kappa
cat("La exactitud del modelo es: ", accuracy * 100, "%")
cat("La sensibilidad del modelo es: ", sensitivity(matriz_confusion) * 100, "%")
cat("La especificidad del modelo es: ", specificity(matriz_confusion) * 100, "%")
Kappa.test(matriz_confusion)
```

C

```r
#Creamos un vector inicial que simula ser un paciente
initial_vector <- c(0,1,0,0,0,0,0,0,0)

#Realizamos la inferencia
results <- fcm_infer(activation_vec = initial_vector,
                     weight_mat = adj_matrix,
                     infer = "mk",
                     transform = "s",
                     iter = 10,
                     lambda = 1,
                     e = 0.001)

#Extraemos los resultados de las predicciones
resultados <- results$values[1:results$iter_convergencia,]

# Crea un vector llamado iteraciones
iterations <- as.numeric(rownames(results$values))
# Añadimos el vector de iteraciones al marco de datos df
df <- data.frame(iterations, results$values)
#Transforma el marco de datos df en formato largo
df2 <- melt(df, id="iterations")

#Creamos el grafico usando ggplot
ggplot(data = df2,
       aes(x = iterations, y = value, group = variable,
           colour = variable)) +
        theme_bw() +
        geom_line(size = 0.5) +
        geom_point(size = 2) +
  geom_vline(xintercept = 9, color = "black",
             linetype = "dashed", size = 1) +
        labs(x = "Iteraciones", y = "Valor del concepto",
             colour = "Conceptos") +
        theme(legend.position="bottom",
              legend.title = element_text(size = 15, face =
"bold"),      legend.text = element_text(size = 15),
              axis.title=element_text(size = 15, face = "bold"),
              axis.text = element_text(size = 15)) +
  scale_x_continuous(breaks = seq(0, 10, by = 1))
```

Figura 1: Código usado para la predicción de DM2 y evaluación de comportamiento de variables. A representa la carga de datos y el proceso de inferencia, B representa la evaluación del modelo usando metricas de desempeño y C representa el código para la generación de la gráfica que muestra el comportamiento de las variables predictoras.

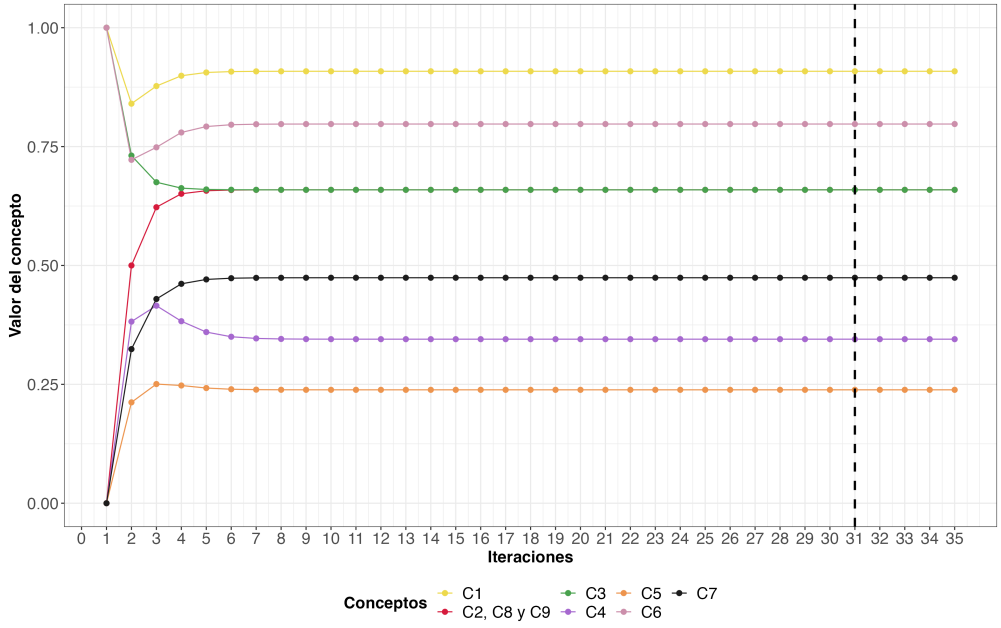

Figura 2: Representación gráfica del comportamiento de las variables predictoras de DM2

Conclusión

En este estudio, hemos propuesto un modelo de predicción inteligente de DM2 basado en MCD implementado en R. Los resultados demuestran el potencial de los MCD para la predicción de la DM2 con un excelente desempeño. Además de la predicción, el modelo permite el análisis del comportamiento de las variables predictoras. La presente investigación se podría ampliar mediante el uso de otras variables de importancia en el diagnóstico de la enfermedad e involucrando expertos que asignen y evalúen las relaciones entre los conceptos del MCD.

Referencias

Chawla, N. V., K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer. 2002. "SMOTE: Synthetic Minority Over-sampling Technique." 16:321–357.
URL: *https://doi.org/10.1613/jair.953*

Dikopoulou, Zoumpoulia and Elpiniki Papageorgiou. 2017. *fcm: Inference of Fuzzy Cognitive Maps (FCMs)*. R package version 0.1.3.
URL: *https://CRAN.R-project.org/package=fcm*

International Diabetes Federation. 2021. *IDF Diabetes Atlas 10th edition*. Vol. 10 of *Diabetes Research and Clinical Practice*.
URL: *https://diabetesatlas.org/idfawp/resource-files/2021/07/IDF_Atlas_10th_Edition_2021.pdf*

Kalange, Omkar, Tejaswini Katale, Atharv Kale and Juwairia Sayyed. 2022. "Prediction of Diabetes using R." 4(12).
URL: *https://doi.org/10.35629/5252-0412885890*

Kosko, Bart. 1986. "Fuzzy cognitive maps." 24(1):65–75.
URL: *https://doi.org/10.1016/S0020-7373(86)80040-2*

Landis, J. R. and G. G. Koch. 1977. "The measurement of observer agreement for categorical data." 33(1):159–174.
URL: *https://pubmed-ncbi-nlm-nih-gov.udea.lookproxy.com/843571/*

Marathe, Ninad, Sushopti Gawade and Adarsh Kanekar. 2021. "Prediction of Heart Disease and Diabetes Using Naive Bayes Algorithm." pp. 447–453.
URL: *https://doi.org/10.32628/CSEIT217399*

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
URL: *https://www.R-project.org/*

Smith, Jack W, JE Everhart, WC Dicksont, WC Knowler and RS Johannes. 1988. "Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus." pp. 261–265.
URL: *https://www-ncbi-nlm-nih-gov.udea.lookproxy.com/pmc/articles/PMC2245318/*

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
URL: *https://ggplot2.tidyverse.org*

World Health Organization. 2023. "Diabetes.".
URL: *https://www.who.int/news-room/fact-sheets/detail/diabetes*