UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE MEDELLÍN

Estudio estadístico con enfoque bayesiano sobre la Diabetes.

Brahian Steven Serna Restrepo[(1)], Julián Saavedra Echavarría[(1)],
Nataly García Osorio[(1)], Johnatan Cardona Jiménez.[(2)]

(1) Carrera de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia- Sede Medellín
(2) Profesor Titular, Escuela de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, sede Medellín

Resumen:
La diabetes es una enfermedad crónica que afecta a millones de personas en todo el mundo y representa una carga significativa para la salud pública. En las últimas décadas, se han realizado importantes avances en la comprensión de los factores que contribuyen al desarrollo y progresión de la diabetes, sobre todo en la diabetes tipo 2 en donde factores genéticos, ambientales y de estilo de vida interactúan de manera compleja.
Bajo este contexto, en el presente artículo se propone un modelo que permita predecir la probabilidad de que una persona promedio con características específicas padezca de diabetes sacarina o no. Es aquí, donde se emplea un modelo de regresión logístico donde a diferencia de estimar los parámetros como tradicionalmente se hace con el paradigma frecuentista, se emplea el paradigma bayesiano y se observa significancia en las variables y capacidades predictivas del modelo.

Palabras clave: Diabetes tipo 2, modelo logístico y bayesiano, promedios, predicción.

1. Introducción

La diabetes sacarina, diabetes mellitus o tipo 2 (que aquí denominaremos, para simplificar, «diabetes»), es una enfermedad crónica y metabólica que representa un desafío significativo para la salud pública en todo el mundo. Caracterizada por la resistencia a la insulina y una disminución en la producción de esta hormona, la diabetes tipo 2 afecta a millones de personas y su incidencia continúa en aumento.[(3)]

El desarrollo de la diabetes es resultado de la interacción compleja entre factores genéticos y ambientales, incluyendo el sedentarismo, la dieta poco saludable y la obesidad. “El sobrepeso / obesidad y la inactividad física son los principales factores de riesgo de diabetes tipo 2”.[(4)] A medida que la prevalencia de la obesidad aumenta a nivel global, la incidencia de la diabetes también se incrementa de manera preocupante pues en la actualidad “aproximadamente 62 millones de personas en las Américas (422 millones de personas en todo el mundo) tienen diabetes, la mayoría vive en países de ingresos bajos y medianos, y 244 084 muertes (1.5 millones en todo el mundo) se atribuyen directamente a la diabetes cada año”.[(5)]

Esta enfermedad conlleva graves complicaciones a largo plazo, como enfermedades cardiovasculares, enfermedad renal crónica, neuropatía y retinopatía. Estas complicaciones no solo afectan la calidad de vida de los pacientes, sino que también imponen una carga significativa en los sistemas de atención médica.

Con base en estas definiciones previas, se pretende abordar la problemática aplicando los conocimientos adquiridos en Estadística Bayesiana. Se realizará un análisis con el conjunto de datos extraídos de “National Institute of Diabetes and

Digestive and Kidney Diseases"[6],en donde se contiene información médica y demográfica de los pacientes junto con su estado de diabetes, todo esto con el fin de predecir si un paciente con ciertas características preestablecidas en el caso de estudio padece de diabetes o no, también con el fin de saber si dicho paciente se encuentra en riesgo de desarrollarla.

2. Descripción de los datos

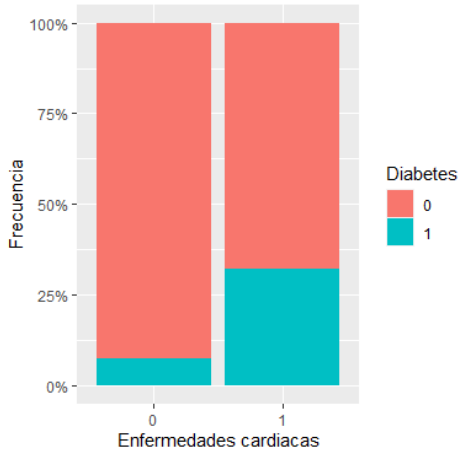
Este conjunto de datos sacados del registro de datos de "National Institute of Diabetes and Digestive and Kidney Diseases" anteriormente mencionado, contiene datos médicos y demográficos de los pacientes junto a su estado de diabetes. Esta base de datos cuenta con 9 variables y aproximadamente 100.000 observaciones para cada una de ellas. Estos datos serán utilizados para construir modelos de aprendizaje automático que puedan predecir la probabilidad de diabetes en los pacientes en función de su historial médico y detalles demográficos. A continuación, las variables utilizadas:

- **Gender:** Hace referencia al sexo biológico del individuo, ya que este puede tener un impacto en su propensión a la diabetes.
- **Age:** Factor muy importante, ya que la diabetes se diagnostica con más frecuencia en adultos mayores. La edad oscila entre 0 y 80 años en el conjunto de datos.
- **Hypertension:** Condición médica en la que la presión arterial en las arterias se eleva de forma persistente. Tiene valores 0 (No tiene hipertensión) o 1 (si tiene hipertensión).
- **Heart_disease:** Condición médica que se asocia con un mayor riesgo de desarrollar diabetes. Tiene valores 0 (No tiene enfermedad cardiaca) o 1 (si tiene enfermedad cardiaca).
- **Smoking_history:** Historial de tabaquismo del individuo según el tiempo de consumo. Tiene valores (nunca, sin información, actual, antiguo, alguna vez, no actual).
- **Bmi:** El IMC (índice de masa corporal) es una medida de la grasa corporal basada en el peso y la altura. Los valores más altos de IMC están vinculados a un mayor riesgo de padecer diabetes.
- **HbA1c_level:** El nivel de HbA1c (hemoglobina A1c) es una medida del nivel promedio de azúcar en la sangre de una persona durante los últimos 2 a 3 meses. Los niveles más altos indican un mayor riesgo de padecer diabetes.
- **Blood_glucose_level:** El nivel de glucosa en sangre se refiere a la cantidad de glucosa en el torrente sanguíneo en un momento dado. Los niveles altos de glucosa en la sangre son un indicador clave de la diabetes.
- **Diabetes:** Es la variable objetivo que se predice, con valores de 1 que indican la presencia de diabetes y 0 que indican que el paciente no padece de diabetes.

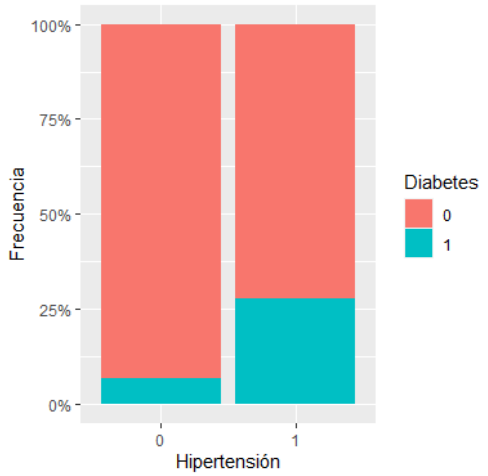
3. Análisis descriptivo

Es de gran importancia realizar un análisis descriptivo de la base de datos cruzando la variable de interés, lo que se busca con esta parte es saber qué comportamiento tienen algunas variables que se consideraron más fuertes con dicha variable «diabetes».

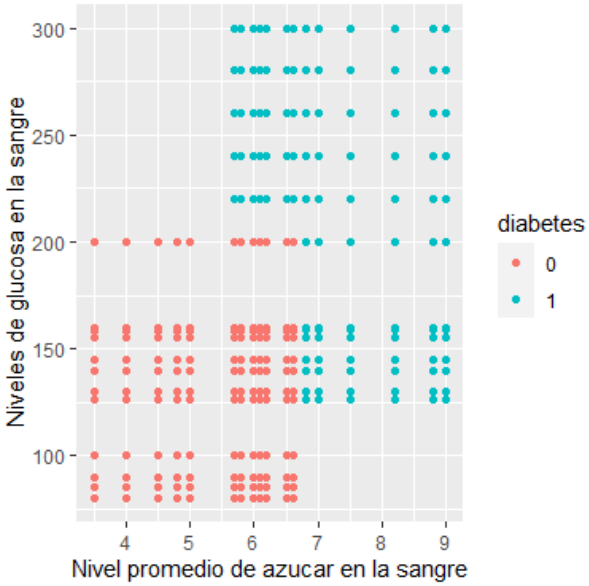
Nota: es de vital importancia saber que las gráficas presentadas a continuación corresponden a proporciones en lugar de conteos, esto se hace por la cantidad de observaciones en la base de datos, ya que si se grafican en términos de conteo las gráficas no serían nada informativas.



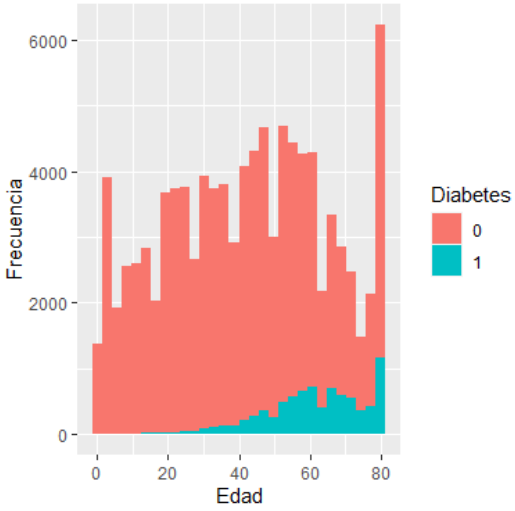
Grafica 1, Relación Diabetes – Enfermedad Cardiaca.



Grafica 2, Relación Diabetes – Hipertensión.



Grafica 3, Relación Diabetes – Nivel promedio de azúcar– Niveles glucosa.



Grafica 4, Relación Diabetes – Edad.

4. Metodología

En este estudio, se recopiló un conjunto de datos relacionados con una investigación médica en la que se buscaba determinar si la presencia de ciertos factores de riesgo se asociaba con la ocurrencia de una enfermedad específica. Las variables predictoras incluyeron características demográficas, antecedentes médicos y hábitos de vida de los participantes como si fumaba con regularidad o no lo hacía en absoluto. La variable respuesta fue la presencia o ausencia de la diabetes.

Se hizo un balanceo en los datos para evitar tener modelos sesgados por la clase mayoritaria y las variables categóricas de la base de datos se convirtieron en variables discretas para que fueran computacionalmente más manejables, obteniendo la siguiente tabla de variables:

Variables	Tipo	Niveles
Genero	Discreta	Hombre=0 Mujer=1 Otros=2
Edad	Continua	
Hipertensión	Discretas	Si=1 No=2
Enfermedad Cardiovascular	Discreta	Si=1 No=0
Tabaquismo	Categórica	Nunca=0 Sin información=1 Actual=2 Antiguo=3 Alguna vez= 4 No actual=5
IMC	Continua	
Hemoglobina	Continua	
Nivel de Glucosa en Sangre	Continua	
Diabetes	Discreta	Si=1 No=0

Tabla 1, Variables después del procesamiento

Se seleccionará un algoritmo para implementar los modelos desde un enfoque bayesiano para hacer predicciones a la variable objetivo, posteriormente tratará de encontrar un modelo parsimonioso, para esto se emplearán métodos para comparación de los

modelos y se evaluará la capacidad predictiva de ellos.

5. Modelo Propuesto

Se selecciona emplear el algoritmo de la regresión logística para dar solución al problema objetivo de este proyecto ya que permite clasificar y predecir la presencia o ausencia de una característica, es decir, cuando se tiene como variable objetivo una variable dicotómica, esta se puede modelar como una distribución Bernoulli y así procedemos a construir los pasos para construir los modelos:

$$Y_i = \begin{cases} 1 & Persona\ i\ padece\ diabetes \\ 0 & Persona\ i\ no\ padece\ diabetes \end{cases}$$

Donde:

$$Y_i \sim Bernoulli(\theta_i), i = 1,2,3,..,17000$$

Y está dado por la siguiente expresión:

$$\theta_i = \frac{e^{\boldsymbol{X_i}\vec{\boldsymbol{\beta}}}}{1 + e^{\boldsymbol{X_i}\vec{\boldsymbol{\beta}}}}$$

Ahora para proponer estimadores de los parámetros usando el enfoque bayesiano, se tiene la siguiente distribución de probabilidad posterior.

$$f(\vec{\beta}|y, X)\ \alpha \prod_{i=1}^{17000} \frac{e^{\boldsymbol{X_i}\vec{\boldsymbol{\beta}}}}{1 + e^{\boldsymbol{X_i}\vec{\boldsymbol{\beta}}}} * f(\vec{\beta})$$

Esta distribución de probabilidad no tiene conjugación para obtener una distribución posteriori de una distribución conocida por lo que se emplean métodos de aproximación Monte Carlo y así obtener una aproximación y estimar los parámetros.

6. Comparativa de los modelos propuestos

Para efectos del estudio, se consideraron dos modelos; el primero con todas las variables propuestas y el segundo con solo cuatro como se muestra a continuación:

Modelo 1

Para el modelo 1, se tienen las siguientes variables:

β_i	media	se_media	des_estandar	2.5%	50%	97.5%	tamaño efectivo	Convergencia
β_1	0.04	0.00	0.05	-0.06	0.04	0.14	4337	1
β_2	-0.29	0.00	0.05	-0.40	-0.29	-0.18	4766	1
β_3	1.03	0.00	0.03	0.97	1.03	1.10	6680	1
β_4	0.86	0.00	0.08	0.70	0.86	1.02	6416	1
β_5	0.85	0.00	0.11	0.64	0.85	1.06	5987	1
β_6	0.06	0.00	0.02	0.02	0.06	0.09	5664	1
β_7	0.77	0.00	0.03	0.71	0.77	0.84	6671	1
β_8	2.76	0.00	0.07	2.63	2.76	2.89	6167	1
β_9	1.80	0.00	0.05	1.71	1.80	1.90	6230	1

Tabla 2, resultados obtenidos con el primer modelo

En donde la variable respuesta es dicotómica (Y: el paciente tiene diabetes, si o no) y todas son significativas menos x_1 que representa el género.

Modelo 2

Para el segundo modelo propuesto, se tienen las siguientes variables:

β_i	media	se_media	des_estándar	2.5%	50%	97.5%	tamaño efectivo	convergencia
β^1	0.06	0.00	0.02	0.01	0.06	0.10	5495	1
β^2	0.97	0.00	0.09	0.81	0.97	1.15	5777	1
β^3	1.13	0.00	0.03	1.08	1.13	1.18	5581	1
β^4	1.74	0.00	0.03	1.67	1.74	1.81	5391	1

Tabla 3, resultados obtenidos con el segundo método

En donde se redujeron las variables a solo 4 y todas resultan significativas.

7. Selección del modelo

Una vez evaluados ambos modelos (modelo 1 tabla 2 y modelo 2 tabla 3), se procede a elegir el modelo que mejor se ajusta a los requerimientos de este caso de estudio. Para ello, se utilizan los siguientes métodos de decisión:

7.1 Factor de bayes

El factor de Bayes es un concepto clave en el análisis estadístico bayesiano. Se utiliza para comparar la evidencia proporcionada por diferentes hipótesis o modelos en función de los datos observados, donde el i-ésimo modelo tiene verosimilitud $f_i(\beta_i|y, X)$ y βi tiene densidad $f(\beta_i)$, para $i = 1,...,k$

Por lo que el factor de bayes para el modelo 1 y modelo 2 es:

$$B_{1,2} = \frac{\int f(Y|M_1, \beta) * f(\beta) d\beta}{\int f(Y|M_2, \beta) * f(\beta) d\beta}$$

Haciendo aproximación por *Monte Carlo*, en el numerador y denominador, el factor de bayes es aproximadamente: 79.37021, lo que significa que hay evidencia solida a favor del modelo 1.

7.2 DIC

DIC es un método bayesiano para la comparación de modelos y también se va a tener en cuenta para la selección de uno de los dos modelos que se tienen en estudio, se calcula de la siguiente forma.

$DIC = -2log(f(y|\widehat{\theta_{Bayes}}) + 2P_{DIC}$

Donde un $\widehat{\theta_{Bayes}} = E[(\theta|\boldsymbol{y})]$, se selecciona el modelo con el DIC más bajo siempre y cuando la diferencia entre estos sea mayor a 10, esto permitirá descartar el modelo con el DIC más alto.

DIC del modelo 1	DIC del modelo 2
38 309.7525	40 380.5564

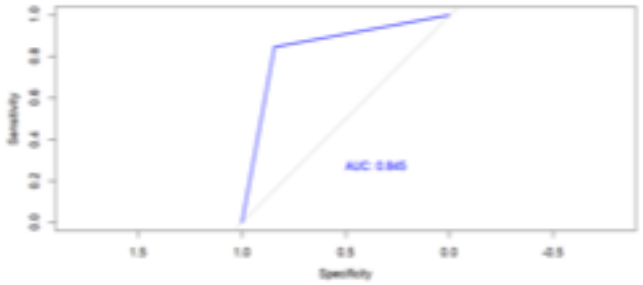
Tabla 4, comparación DIC en modelo 1 y 2.

De la tabla se observa que la diferencia entre el DIC de ambos modelos es de 2070.797, por tanto, según este criterio se selecciona el modelo 1.

9. Capacidad predictiva del modelo

Curva ROC

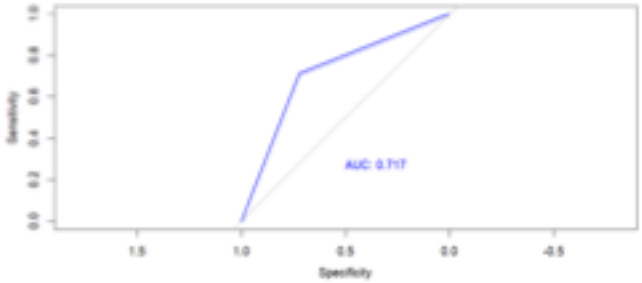
La curva ROC es un método que nos permite evaluar la capacidad predictiva de los modelos. A continuación, se observa la curva ROC del modelo 1 o modelo completo:



Grafica 5, Curva ROC modelo 1

En donde el área bajo curva es de 0.845, por lo que es un indicio de que el modelo está ajustando bien las predicciones y no lo está haciendo de forma aleatoria, es decir, tiene un buen criterio para tomar decisiones.

Ahora se observará a continuación la curva ROC del modelo 2.

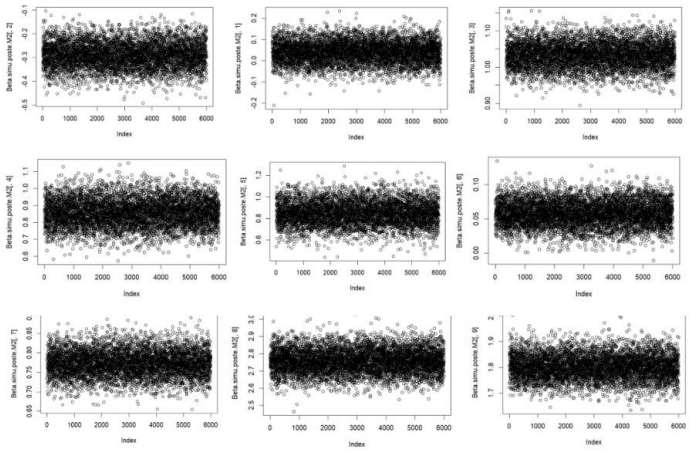


Grafica 6, Curva ROC modelo 2

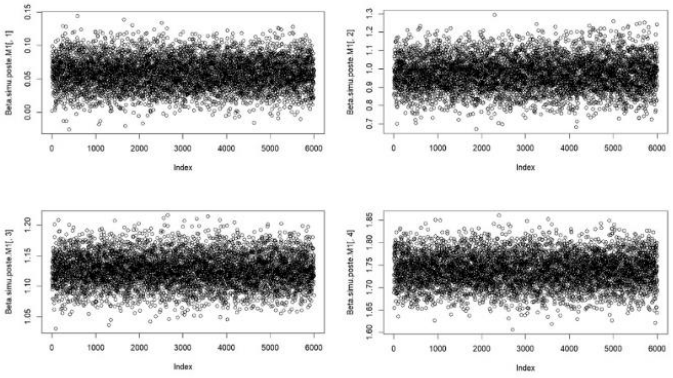
Acá el área bajo la curva es de 0.717, por lo que también es un indicio de que esta tomado bien las decisiones, aunque no mejor modelo 1, es decir, si se toma esto como un criterio para seleccionar el mejor modelo, se selecciona nuevamente el modelo 1, el modelo con todas las variables.

10. Estacionariedad en los parámetros de los modelos

Modelo 1



Modelo 2



11. Conclusiones

- Para el *modelo 1*, todas las variables, a excepción del intercepto, son significativas. Esto afirma la importancia de la presencia de dichas variables en el modelo para estimar la probabilidad de que un paciente padezca diabetes.
- Para el *modelo 2*, todas las variables son significativas, incluyendo el intercepto. De la misma manera, nos indica que las variables presentes en el modelo aportan información necesaria para la predicción de la enfermedad.
- Para ambos modelos, la estacionariedad de los parámetros del modelo es buena y se refleja la independencia entre ellas. Esto indica que el modelo fue, de una manera correcta, estimado.
- Con respecto a las variables predictoras, las que mostraron más influencia con el padecimiento de diabetes son hipertensión, enfermedad cardiovascular y el IMC. Siempre que dichas variables mostraban de manera individual una alerta, no era una excepción para el padecimiento de la enfermedad.
- Ambos modelos, en términos de convergencia mostraron un buen comportamiento, esto confirmó que la cantidad de iteraciones usadas, la cantidad de cadenas y el número de saltos fueron considerablemente buenos.
- Cabe recalcar que ambos modelos funcionan considerablemente bien para la predicción de diabetes en un paciente con la información médica recopilada, solo el *modelo 1* produce una mayor confianza y un mejor rendimiento con respecto al *modelo 2*.
- El modelo seleccionado es el *modelo 1,* el cual contiene todas las variables de la base de datos. Esta decisión se toma con referencia a los métodos se selección de modelos expuestos, donde siempre se muestra una evidencia fuerte a favor de este modelo.
- El *modelo 1* contiene a la presencia de diabetes como variable objetivo y como variables predictoras a género, edad, hipertensión, enfermedad cardiovascular, tabaquismo, IMC, hemoglobina A1C y nivel de glucosa en la sangre.

12. Referencias

(3) Organización Panamericana de la Salud. Diabetes. [Internet]. Washington, D.C.: Organización Panamericana de la Salud; c2021 [citado el 20 de junio de 2023]. Disponible en: https://www.paho.org/es/temas/diabetes.

(4) Rovalino M, et al. Revision Bibliografica Sobre Diagnostico y Tratamiento de Diabetes Millitus Tipo 2 en Pacientes con Sindrome Metabolico.[internet].Ambato, Ecuador:Mayo 2023[citado el 20 de junio de 2023].Disponible en:https://repositorio.uta.edu.ec/jspui/handle/123456789/38787

(5) Organización Panamericana de la Salud. Diabetes. [Internet]. Washington, D.C.: Organización Panamericana de la Salud; c2021 [citado el 20 de junio de 2023].

(6) National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Health Information: Diabetes [Internet]. Bethesda, MD: NIDDK; c2021 [citado el 20 de junio de 2023]. Disponible en: https://www.niddk.nih.gov/health-information/diabetes