APLICACIÓN DEL MODELO DE REGRESIÓN LOGÍSTICA PARA LA PREDICCIÓN DE DIABETES EN PACIENTES MUJERES INDÍGENAS PIMA

HUERTA GARCÍA M. [a], HERNANDEZ LUGO R. [a], GOMEZ ESTEVA B. [a], CASTRO ALVA J. [a]

[a]Facultad de Ciencias Físico Matematicas
Benemerita Universidad Autónoma de Puebla
Avenida San Claudio y 18 Sur, Ciudad Universitaria, C.P. 72570, Puebla, Puebla, Mexico.
e-mail: melissa.huerta@alumno.buap.mx, rafael.hernandezlu@alumno.buap.mx, bacaanda.gomez@alumno.buap.mx, jose.castroalva@correo.buap.mx

La regresión logística es un tipo de regresión no lineal que permite estimar la probabilidad de una variable categórica en función de una variable cuantitativa. En el presente trabajo, analizamos una base de datos que contiene información de pacientes mujeres indígenas Pima de 21 años en adelante, con ayuda de este modelo nos permitimos predecimos si el individuo tiene o no diabetes, considerando factores como la Glucosa, Insulina, Embarazos, IMC, entre otros. Así mismo, para obtener un mejor ajuste en el modelo se uso el método de selección de modelos, en específico, stepwise y asi obtenemos las variabkes significativas.

Keywords: Modelo de regresión logística, Diabetes, Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales, Predicción, Análisis de correlación, Pacientes indígenas Pima

1. Introducción

La diabetes es una enfermedad crónica que afecta la forma en que el cuerpo convierte los alimentos en energía. La detección tardía y maneja poco adecuadamente, puede llevar a complicaciones graves, como enfermedades cardíacas, daño renal y problemas en la visión. Por lo que, la detección temprana de la diabetes es esencial para prevenir estas complicaciones y mejorar la calidad de vida de las personas afectadas.

Los avances en tecnología y ciencia de datos han permitido desarrollar modelos predictivos que pueden ayudar en la identificación temprana de la diabetes. Estos modelos, como la regresión logística, analizan factores de riesgo y patrones en los datos de salud para predecir la probabilidad de que una persona desarrolle la enfermedad. En este ámbito, los modelos predictivos juegan un rol crucial al proporcionar herramientas que pueden ayudar en la identificación precisa de esta enfermedad.

El dataset es ampliamente utilizado en la comunidad de ciencia de datos para proyectos relacionados con la salud, ya que permite aplicar técnicas de machine learning para la predicción de enfermedades. Además, este conjunto de datos es un recurso valioso para quienes desean experimentar con diferentes modelos de regresión y clasificación, como regresión logística, máquinas de soporte vectorial, y redes neuronales, entre otros.

La regresión logística es una herramienta estadística muy útil cuando se trata de predecir eventos binarios, como en el caso del diagnóstico temprano en pacientes con o sin diabetes.

Este tipo de modelo permite evaluar cómo diferentes características en el historial clínico (o variables independientes) de las muestras de pacientes afectan la probabilidad de que una muestra presente o no diabetes.

El preprocesamiento exhaustivo de los datos y la selección de características relevantes mediante análisis de correlación son pasos clave en la construcción de un modelo predictivo robusto. Estos pasos ayudan a mejorar la precisión del modelo al enfocarse en las variables que tienen mayor impacto en la predicción, y también permiten abordar algunas de las limitaciones que podrían haber afectado estudios previos.

El desarrollo del modelo incluye pruebas de hipótesis tanto globales como individuales. Las pruebas globales verifican la significancia del modelo en su conjunto, mientras que las pruebas individuales se centran en la relevancia estadística de cada variable predictora. En resumen, este enfoque permite crear un modelo más preciso y confiable para el diagnóstico de la diabetes.

2. Regresión logística

El modelo de regresión lineal, desarrollado por Legendre, Gauss, Galton y Pearson, postula que, dado un conjunto de observaciones $\{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$, la media μ de la variable dependiente y se relaciona de manera lineal con una o más variables independientes $x_1, \ldots, x_k$, según la siguiente ecuación:

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \qquad (1)$$

Donde la interpretación de los elementos del modelo es la siguiente:

- β_0: representa la intersección con el eje y, indicando el valor promedio de la variable respuesta y cuando todas las variables predictoras son cero.
- β_k: indica el efecto promedio sobre la variable respuesta al aumentar en una unidad la variable predictora x_k, manteniendo constantes las demás variables.

El método más utilizado es el ajuste por mínimos cuadrados ordinarios (OLS), que determina el mejor modelo como la línea (o el hiperplano en el caso de la regresión múltiple) que minimiza la suma de las desviaciones verticales al cuadrado entre los datos de entrenamiento y la línea de ajuste.

2.1. Variable de Respuesta Binaria. Si una variable cualitativa binaria se codifica como 0 y 1, es matemáticamente posible ajustar un modelo de regresión lineal utilizando el método de mínimos cuadrados. No obstante, esta aproximación presenta dos problemas principales. En primer lugar, al generar una línea (o un hiperplano en el caso de variables múltiples), es posible obtener valores predichos que no se limitan a 0 y 1, lo que contradice la definición de una variable respuesta binaria. En segundo lugar, si se desea interpretar las predicciones del modelo como probabilidades de pertenencia a cada clase, no se cumpliría la condición de que todas las probabilidades deben estar en el intervalo [0,1], ya que podrían resultar valores fuera de este rango.

Para resolver estos problemas,la regresión logística, desarrollada por David Cox en 1958, aplica una transformación a los valores generados por la regresión lineal mediante una función que siempre produce resultados entre 0 y 1. Una de las funciones más comunes para este propósito es la función logística, también conocida como función sigmoide:

$$\text{sigmoide} = \sigma(y) = \frac{1}{1 + e^{-y}} \qquad (2)$$

Además, se cumplen los siguientes resultados:

$$\lim_{y \to \infty} \sigma(y) = 1$$

$$\lim_{y \to -\infty} \sigma(y) = 0$$

Sustituyendo y de la ecuación (2) por (1),

$$\mathbb{P}(y = 1|\mathbf{X} = \mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}}$$

$$= \frac{1}{\frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}} + \frac{1}{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}}$$

$$= \frac{1}{\frac{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}}$$

obtenemos:

$$\mathbb{P}(y = 1|\mathbf{X} = \mathbf{x}) = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}} \qquad (3)$$

donde $\mathbb{P}(y = 1|\mathbf{X} = \mathbf{x})$ puede interpretarse como la probabilidad de que la variable respuesta y sea 1 (la clase de referencia), dados los predictores $x_1, \cdots, x_k$.

2.2. Transformación logarítmica para linearizar el modelo. Este modelo tiene los coeficientes de regresión en los exponentes, por lo que no es un modelo lineal y no puede ajustarse con las técnicas descritas inicialmente. Para solucionar este inconveniente, se utiliza el logaritmo de la expresión obtenida.

La probabilidad de que $y = 1$ dado los predictores $\mathbf{X} = x_1, \cdots, x_k$ es:

$$\mathbb{P}(y = 1|\mathbf{X} = \mathbf{x}) = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}} = \frac{e^{\sum \beta_i x_i}}{1 + e^{\sum \beta_i x_i}}$$

Al ser un caso de clasificación binaria, la probabilidad de que $y = 0$ es el complemento, esto es:

$$\mathbb{P}(y = 0|\mathbf{X} = \mathbf{x}) = 1 - \mathbb{P}(y = 1|\mathbf{X} = \mathbf{x})$$

$$= 1 - \frac{e^{\sum \beta_i x_i}}{1 + e^{\sum \beta_i x_i}}$$

$$\mathbb{P}(y = 0|\mathbf{X} = \mathbf{x}) = \frac{1}{1 + e^{\sum \beta_i x_i}} \qquad (4)$$

La función exponencial siempre produce valores positivos, por lo que el cociente de estos valores es

también positivo, permitiendo aplicar el logaritmo. Haciendo un cociente entre (3) y (4),

$$\frac{\mathbb{P}(y=1|\mathbf{X}=\mathbf{x})}{\mathbb{P}(y=0|\mathbf{X}=\mathbf{x})} = \frac{\frac{e^{\sum \beta_i x_i}}{1+e^{\sum \beta_i x_i}}}{\frac{1}{1+e^{\sum \beta_i x_i}}} = e^{\sum \beta_i x_i}$$

aplicando logaritmo, obtenemos:

$$\ln\left(\frac{\mathbb{P}(y=1|\mathbf{X}=\mathbf{x})}{\mathbb{P}(y=0|\mathbf{X}=\mathbf{x})}\right) = \ln\left(e^{\sum \beta_i x_i}\right)$$

$$\ln\left(\frac{\mathbb{P}(y=1|\mathbf{X}=\mathbf{x})}{\mathbb{P}(y=0|\mathbf{X}=\mathbf{x})}\right) = \sum \beta_i x_i$$

$$\ln\left(\frac{\mathbb{P}(y=1|\mathbf{X}=\mathbf{x})}{\mathbb{P}(y=0|\mathbf{X}=\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \quad (5)$$

Esta transformación convierte el lado derecho en una ecuación lineal. El término del lado izquierdo es el logaritmo de un cociente de probabilidades, conocido como la razón de probabilidades (*log of odds*).

Este proceso permite transformar un problema de clasificación no lineal en uno de regresión lineal, que puede ser ajustado con métodos convencionales. Una vez obtenidos los coeficientes del modelo ($\beta_0, \beta_1, \cdots, \beta_k$), es posible calcular la probabilidad de que una nueva observación pertenezca a la clase $y = 1$ utilizando la ecuación (3).

3. Descripción de la Base de Datos

Para el proyecto se utilizó la base de datos denominada Pima Indians Diabetes Database (PIDD) (1990), cuya propiedad original pertenece al National Institute of Diabetes and Digestive and Kidney Diseases de los Estados Unidos de América (EUA), la población para este estudio era la población india Pima cerca de Phoenix, Arizona.

3.1. Procedencia. Estos datos fueron obtenidos de la University of California in Irving (UCI) Machine Learning Repository – Pima Indians Diabetes Data Set.

3.2. Dimensión. Las unidades de análisis consistieron en 768 mujeres pertenecientes a la etnia Pima y con al menos 21 años de edad.

3.3. Variables Predictoras. Fueron registradas 9 variables: 8 numéricas detalladas a continuación:

1. `Glucose`: Concentración de glucosa plasmática a las 2 horas de una prueba de tolerancia oral a la glucosa (G120 mg/dl).
2. `Insuline`: Concentración de insulina sérica a las 2 horas de una prueba de tolerancia oral a la glucosa (I120 mU/ml).
3. `BloodPressure`: Presión arterial diastólica (PAD mmHg).
4. `Skin Thickness`: Grosor del pliegue de la piel del tríceps (GPPT mm).
5. `BMI`: Índice de masa corporal (IMC= peso /altura al cuadrado= kg/m2).
6. `Diabetes Pedigree Fuction`: Antecedentes Familiares o función de pedigrí de diabetes (FPD).
7. `Pregnancies`: Número de embarazos.
8. `Age`: Edad (En años).

3.4. Diagnóstico. El diagnóstico estuvo basado en el criterio de la OMS (i.e.: $G_{120} \geq 200$ mg/dl en cualquier examen o evaluación de rutina médica). Donde:

$$0 \Longleftrightarrow \text{No tiene diabetes.}$$

$$1 \Longleftrightarrow \text{Sí tiene diabetes.}$$

4. Construcción del Modelo

A continuación, se ofrece una descripción detallada del proceso llevado a cabo para desarrollar, entrenar y evaluar el modelo.

4.1. Importacion de Bibliotecas. El modelo fue implementado en `Python` utilizando diversas bibliotecas clave para el manejo de datos, visualización y modelado. Para la manipulación de datos se utilizaron `pandas` y `numpy`. Las visualizaciones se realizaron con `matplotlib` y `seaborn`. El preprocesamiento y modelado se llevaron a cabo con `sklearn`, utilizando módulos como `LogisticRegression` y `train_test_split`, además de `statsmodels` para análisis estadísticos más avanzados. Finalmente, se empleó la biblioteca `warnings` para gestionar advertencias y garantizar una ejecución fluida del código.

4.2. Carga de Datos. Los datos fueron cargados desde un archivo CSV utilizando `pandas`, lo cual permitió la conversión eficiente de los datos en un `DataFrame`. Este paso es fundamental para estructurar los datos de manera que puedan ser fácilmente manipulados y analizados. La base de datos es libre, puede ser consultada y descargada en `https://www.kaggle.com/uciml/pimaindians-diabetes-database`.

4.3. Preprocesamiento de Datos. Antes de proceder con el modelado, se llevó a cabo un preprocesamiento de los datos. Este proceso incluyó la limpieza de datos para eliminar o imputar valores faltantes, así como la codificación de variables de respuesta en un formato numérico adecuado para el análisis.

4.4. Creación del Modelo. Se construyó un modelo de regresión logística utilizando `statsmodels`. La regresión logística es un modelo estadístico utilizado para predecir la probabilidad de una variable categórica binaria basada en una o más variables independientes. Para la evaluación del modelo, los datos fueron divididos en un 80 % para el entrenamiento y un 20 % para la prueba. Esta partición permite que el modelo se ajuste a una porción significativa de los datos mientras se evalúa su rendimiento en un conjunto independiente.

4.5. Selección del Modelo. Se empleó el método *stepwise* para garantizar la selección de las variables predictoras con mayor significancia, asegurando a su vez un modelo parsimonioso que se ajuste adecuadamente a los requerimientos de nuestro problema.

4.6. Entrenamiento del Modelo. Durante el entrenamiento del modelo, se utilizaron los datos de entrenamiento que fueron previamente escalados para asegurar que todas las variables tuvieran una escala comparable. La variable de respuesta se trató como binaria, facilitando la implementación de la regresión logística para la clasificación binaria. Las variables predictoras se consideraron cuantitativas, y el modelo ajustó los coeficientes para minimizar el error de clasificación y mejorar la precisión de las predicciones.

4.7. Significancia del Modelo. Uno de los primeros resultados que hay que evaluar al ajustar un modelo de regresión logística es el resultado del test de significancia *likelihood ratio* (LLR). Este contraste responde a la pregunta de si el modelo en su conjunto es capaz de predecir la variable respuesta mejor de lo esperado por azar, o lo que es equivalente, si al menos uno de los predictores que forman el modelo contribuye de forma significativa.

Con frecuencia, la hipótesis nula y alternativa de este test se describen como:

$$H_0 : \beta_1 = \cdots = \beta_k = 0 \quad \text{vs} \quad H_a : \text{al menos un } \beta_i \neq 0$$

Si el test resulta significativo, implica que el modelo es útil, pero no que sea el mejor. Podría ocurrir que alguno de sus predictores no fuese necesario.

4.8. Significancia de los Predictores. En la mayoría de casos, aunque el estudio de regresión logística se aplica a una muestra, el objetivo último es obtener un modelo que explique la relación entre las variables en toda la población. Esto significa que el modelo generado es una estimación de la relación poblacional a partir de la relación que se observa en la muestra y, por lo tanto, está sujeto a variaciones. Para cada uno de los coeficientes de la ecuación de regresión logística (β_k) se puede calcular su significancia (p-value) y su intervalo de confianza. El test estadístico más empleado es el *Wald chi-test*.

El test de significancia para los coeficientes (β_k) del modelo logístico considera como hipótesis:

H_0: el predictor x_k NO contribuye al modelo ($\beta_k = 0$), en presencia del resto de predictores.

vs

H_a: el predictor x_k SÍ contribuye al modelo ($\beta_k \neq 0$), en presencia del resto de predictores.

4.9. Validación del Modelo. Finalmente, se llevó a cabo una validación exhaustiva del modelo utilizando diversas métricas de evaluación. Estas incluyeron la precisión (accuracy), el recall, el F1-score, la matriz de confusión y las curvas ROC-AUC. La precisión proporciona una medida general del rendimiento del modelo, mientras que el recall y el F1-score ofrecen una visión más detallada de su capacidad para identificar correctamente las instancias positivas. La matriz de confusión permite una evaluación más granular de las predicciones, y las curvas ROC-AUC ayudan a entender la capacidad del modelo para distinguir entre las clases positivas y negativas en diferentes umbrales de decisión.

5. Resultados

5.1. Matriz de Confusión Sin el Método Stepwise. Las matrices de confusión son herramientas que evalúan modelos de clasificación mostrando el número de predicciones correctas e incorrectas. Representan Verdaderos Positivos (TP), Verdaderos Negativos (TN), Falsos Positivos (FP) y Falsos Negativos (FN), permitiendo así analizar la precisión y el rendimiento del modelo.

5.1.1. Conjunto de Entrenamiento Sin StepWise. La matriz de confusión presentada en la Figura 1 muestra los resultados del modelo de clasificación entrenado para distinguir entre pacientes sin diabetes (etiquetados como 0) y pacientes con diabetes (etiquetados como 1).

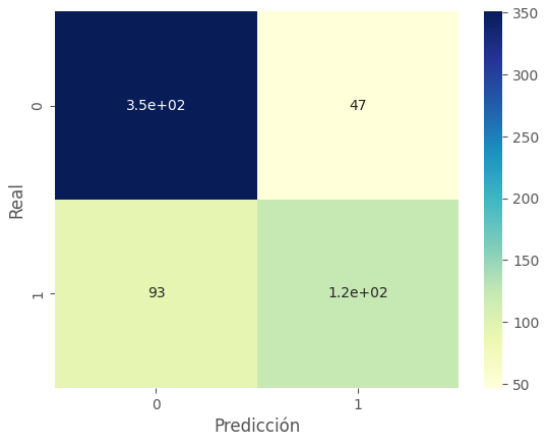

Fig. 1. Matriz de Entrenamiento Sin StepWise

- **Verdaderos Negativos (TN):** El modelo predijo correctamente 350 casos sin diabetes.
- **Falsos Positivos (FP):** El modelo tomo incorrectamente 47 casos con diabetes cuando en realidad no tenian diabetes.
- **Falsos Negativos (FN):** El modelo clasifico incorrectamente 93 casos sin diabetes cuando en realidad si tenia diabetes.
- **Verdaderos positivos (TP):** El modelo predijo correctamente 120 casos con diabetes.

5.1.2. Conjunto de Prueba Sin StepWise. La matriz de confusión presentada en la Figura 2 muestra los resultados del modelo de clasificación en el conjunto de datos de prueba para distinguir entre pacientes sin diabetes (etiquetados como 0) y con diabetes (etiquetados como 1).

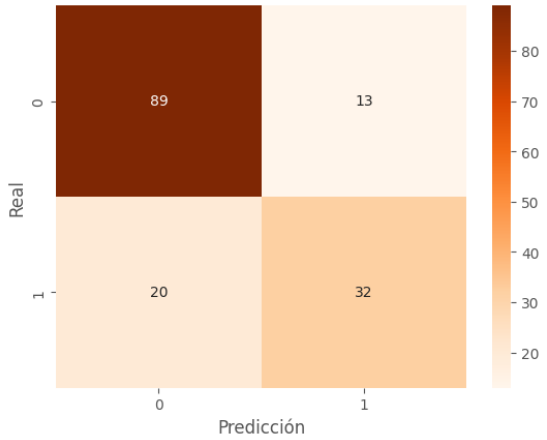

Fig. 2. Matriz de Prueba Sin StepWise

- **Verdaderos Negativos (TN):** El modelo predijo correctamente 89 casos sin diabetes.
- **Falsos Positivos (FP):** El modelo tomo incorrectamente 13 casos con diabetes cuando en realidad no tenian diabetes.
- **Falsos Negativos (FN):** El modelo clasifico incorrectamente 20 casos sin diabetes cuando en realidad si tenia diabetes.
- **Verdaderos positivos (TP):** El modelo predijo correctamente 32 casos con diabetes.

5.2. Matriz de Confusión con StepWise. A continuación se muestran las matrices de confusión obtenidas aplicando el método StepWise para la selección de nuestra variables predictoras que construyen nuestro mejor modelo de manera parsimoniosa.

5.2.1. Conjunto de Entrenamiento Con StepWise. La matriz de confusión presentada en la Figura 3 muestra los resultados del modelo de clasificación en el conjunto de datos de prueba para distinguir entre pacientes sin diabetes (etiquetados como 0) y con diabetes (etiquetados como 1).

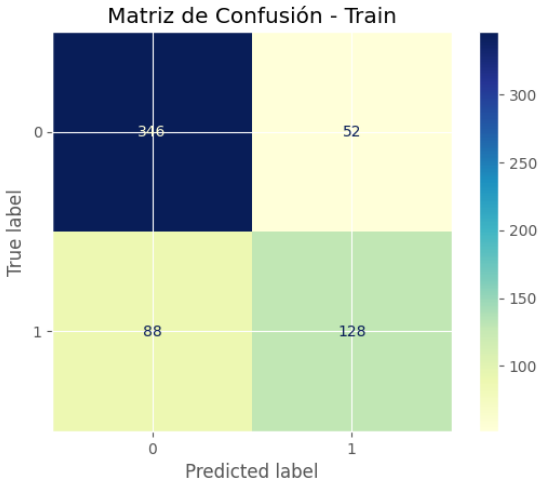

Fig. 3. Matriz de Entrenamiento Con StepWise

- **Verdaderos Negativos (TN):** El modelo predijo correctamente 346 casos sin diabetes.
- **Falsos Positivos (FP):** El modelo tomo incorrectamente 52 casos con diabetes cuando en realidad no tenian diabetes.
- **Falsos Negativos (FN):** El modelo clasifico incorrectamente 88 casos sin diabetes cuando en realidad si tenia diabetes.

- **Verdaderos positivos (TP):** El modelo predijo correctamente 128 casos con diabetes.

5.3. Interpretación de las métricas de validación. Mostraremos las metricas calculadas de los modelos por separado así como sus interpretaciones.

5.3.1. Conjunto de Prueba Con StepWise. La matriz de confusión presentada en la Figura 2 muestra los resultados del modelo de clasificación en el conjunto de datos de prueba para distinguir entre pacientes sin diabetes (etiquetados como 0) y con diabetes (etiquetados como 1).

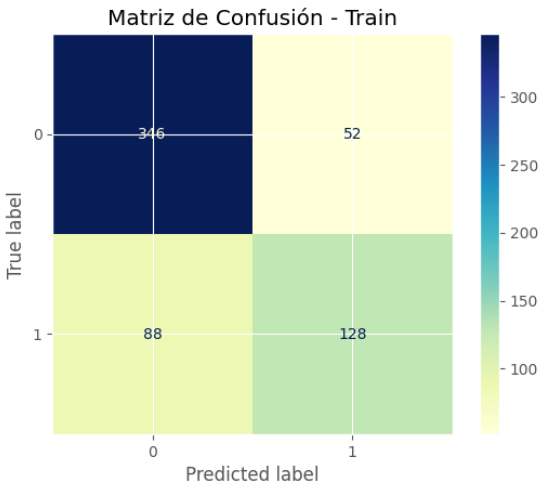

Fig. 4. Matriz Conjunto de Prueba Con StepWise

- **Verdaderos Negativos (TN):** El modelo predijo correctamente 88 casos sin diabetes.
- **Falsos Positivos (FP):** El modelo tomo incorrectamente 14 casos con diabetes cuando en realidad no tenian diabetes.
- **Falsos Negativos (FN):** El modelo clasifico incorrectamente 25 casos sin diabetes cuando en realidad si tenia diabetes.
- **Verdaderos positivos (TP):** El modelo predijo correctamente 27casos con diabetes.

5.3.2. Métricas de Validación para el Modelo de Entrenamiento sin StepWise.

Métrica	Valor
Exactitud (Accuracy)	0.771987
Precisión (Precision)	0.723529
Recall (Sensitivity)	0.569444
F1-score	0.637306
AUC-ROC	0.725677

Tabla 1: Resultados de las métricas del modelo

- **Exactitud (0.772)**: El modelo clasifica correctamente aproximadamente el 77.2 % de los casos, lo que sugiere una moderada capacidad predictiva general.
- **Precisión (0.724)**: De todas las mujeres predichas como diabéticas, el 72.4 % realmente lo son, indicando un buen manejo de falsos positivos.
- **Recall (0.569)**: El modelo detecta el 56.9 % de las mujeres que realmente tienen diabetes, lo que indica un nivel de sensibilidad relativamente bajo.
- **F1-score (0.637)**: Este balance entre precisión y recall indica un desempeño moderado en la predicción de casos positivos.
- **AUC-ROC (0.726)**: Este valor indica que el modelo tiene un buen rendimiento para diferenciar entre mujeres con y sin diabetes, siendo 1 un desempeño perfecto y 0.5 un desempeño aleatorio.

5.3.3. Métricas de Validación para el Modelo de Prueba sin StepWise.

Métrica	Valor
Exactitud (Accuracy)	0.785714
Precisión (Precision)	0.711111
Recall (Sensitivity)	0.615385
F1-score	0.659793
AUC-ROC	0.743967

Tabla 2: Resultados de las métricas del modelo de prueba

- **Exactitud (0.786)**: El modelo clasifica correctamente aproximadamente el 78.6 % de los casos, lo que indica una capacidad predictiva moderadamente alta.
- **Precisión (0.711)**: De todas las mujeres predichas como diabéticas, el 71.1 % realmente lo son, lo que refleja un buen manejo de falsos positivos.
- **Recall (0.615)**: El modelo identifica correctamente el 61.5 % de las mujeres que tienen diabetes, lo que indica una mejora en la sensibilidad comparado con el modelo anterior, aunque sigue siendo un área a mejorar.
- **F1-score (0.660)**: Este balance entre precisión y recall indica un desempeño moderado en la predicción de casos positivos, mostrando que el modelo es equilibrado pero puede seguir mejorando.
- **AUC-ROC (0.744)**: Este valor indica un buen rendimiento para diferenciar entre mujeres con y sin diabetes, lo que demuestra una mejora en la capacidad del modelo para discriminar entre ambas clases.

5.3.4. Métricas de Validación para el Modelo de Entrenamiento con StepWise.

Métrica	Valor
Train Accuracy	0.771987
Train Precision	0.711111
Train Recall	0.592593
Train F1-score	0.646464
Train AUC-ROC	0.730970

Tabla 3: Resultados de las métricas del modelo de entrenamiento

- **Train Accuracy (0.772)**: El modelo clasifica correctamente aproximadamente el 77.2 % de los casos en el conjunto de entrenamiento, lo que sugiere una capacidad predictiva adecuada después de la selección de variables.
- **Train Precision (0.711)**: De todas las mujeres predichas como diabéticas, el 71.1 % realmente lo son, lo que indica un buen manejo de falsos positivos con las variables seleccionadas.
- **Train Recall (0.593)**: El modelo identifica correctamente el 59.3 % de las mujeres que tienen diabetes, lo que indica que la selección de variables ha mejorado ligeramente la sensibilidad.
- **Train F1-score (0.646)**: Este balance entre precisión y recall indica un desempeño moderado en la predicción de casos positivos, mostrando que el modelo es equilibrado.
- **Train AUC-ROC (0.731)**: Este valor indica un buen rendimiento para diferenciar entre mujeres con y sin diabetes en el conjunto de entrenamiento, lo que demuestra que la selección de variables ha mantenido una capacidad discriminativa sólida.

5.3.5. Métricas de Validación para el Modelo de Prueba con StepWise.

Métrica	Valor
Test Accuracy	0.746753
Test Precision	0.658537
Test Recall	0.519231
Test F1-score	0.580645
Test AUC-ROC	0.690988

Tabla 4: Resultados de las métricas del modelo de prueba

- **Test Accuracy (0.747)**: El modelo clasifica correctamente aproximadamente el 74.7 % de los casos en el conjunto de prueba, lo que indica una capacidad predictiva adecuada.
- **Test Precision (0.659)**: De todas las mujeres predichas como diabéticas en el conjunto de prueba, el 65.9 % realmente lo son, lo que sugiere que el modelo maneja falsos positivos de manera moderada.
- **Test Recall (0.519)**: El modelo identifica correctamente el 51.9 % de las mujeres que tienen diabetes en el conjunto de prueba, lo que indica un desempeño moderado en la detección de casos positivos.
- **Test F1-score (0.581)**: Balance entre precisión y recall, indica un desempeño moderado en la predicción de casos positivos.
- **Test AUC-ROC (0.691)**: Indica una capacidad moderada del modelo para diferenciar entre mujeres con y sin diabetes en el conjunto de prueba.

6. Notas y conclusiones

Primero mostraremos una tabla comparativa entre los dos modelos realizados, con selección de modelo y sin selección de modelo.

Métrica	Sin StepWise	Con StepWise
Entrenamiento		
Exactitud (Accuracy)	0.771987	0.771987
Precisión (Precision)	0.723529	0.711111
Recall (Sensitivity)	0.569444	0.592593
F1-score	0.637306	0.646464
AUC-ROC	0.725677	0.730970
Prueba		
Exactitud (Accuracy)	0.785714	0.746753
Precisión (Precision)	0.711111	0.658537
Recall (Sensitivity)	0.615385	0.519231
F1-score	0.659793	0.580645
AUC-ROC	0.743967	0.690988

Tabla 5: Comparativa de Métricas entre Modelos (Con y Sin StepWise)

1. El modelo con StepWise seleccionó las variables *Glucose*, *Pregnancies*, y *BMI* debido a su alta correlación con la diabetes y su capacidad para explicar gran parte de la variabilidad en los datos. Estas variables fueron seleccionadas porque están estrechamente relacionadas con los factores de riesgo conocidos de la diabetes.
2. Aunque las métricas del modelo con StepWise son ligeramente inferiores en comparación con el modelo completo, el nuevo modelo es más parsimonioso, lo que reduce el riesgo de sobreajuste y mejora la interpretabilidad. La simplificación del modelo permite un análisis más claro y evita la

complejidad innecesaria introducida por variables menos significativas.

3. A pesar de la ligera disminución en algunas métricas, el AUC-ROC cercano a 0.691 en el conjunto de prueba sigue indicando una buena capacidad discriminativa, lo que respalda la elección de un modelo más simple y robusto. Este modelo logra un equilibrio entre simplicidad y rendimiento, lo que lo hace más adecuado para la predicción de diabetes en mujeres de la tribu Pima.

4. **Resumen**: El uso de StepWise en este modelo permitió reducir el número de variables predictoras, eliminando aquellas que no aportaban significativamente al poder predictivo del modelo. Aunque las métricas como la precisión y el recall se redujeron ligeramente, el modelo resultante es más simple y menos propenso a errores de autocorrelación y ruido, lo que lo hace más robusto y generalizable. La selección de variables clave, como *Glucose*, *Pregnancies*, y *BMI*, refleja un enfoque en factores bien conocidos y relevantes en la predicción de la diabetes, manteniendo un rendimiento aceptable en la discriminación entre casos positivos y negativos.

Referencias

[1] Gareth, J., W. Daniela, H. Trevor, and T. Robert (2013). *An introduction to statistical learning: with applications in R. Spinger.*

[2] Sheather, S. (2009). *A modern approach to regression with R. Springer Science & Business Media.*

[3] Diabetes Dataset. (2022, 6 octubre). Kaggle. `https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset`

Appendix

Apéndice A

A continuación, podrán encontrar como recurso adicional el código QR para visualizar el Notebook que documenta todo el proceso y explicación paso a paso de la creación del Modelo con Regresión Logística.