Gabriela Chimali Nava Ramírez | A01710530
31/08/2025

1. INTRODUCCIÓN

La regresión logística es un modelo estadístico ampliamente utilizado en el ámbito del Machine Learning para abordar problemas de clasificación. A diferencia de la regresión lineal, que predice valores continuos, este modelo emplea la función sigmoide para estimar la probabilidad de que un registro pertenezca a una clase determinada, generalmente expresada en términos binarios (0 o 1). Gracias a esta característica, resulta una herramienta especialmente útil en contextos donde se busca identificar la presencia o ausencia de una condición específica.

En este trabajo se desarrolla un modelo de regresión logística para clasificar a los individuos de acuerdo con su condición diabética, utilizando un conjunto de indicadores de salud. A lo largo del documento se describen las etapas de preparación de datos, selección de variables, construcción del modelo y entrenamiento, con el objetivo de demostrar la aplicación de esta técnica en un problema de clasificación real.

2. DESCRIPCIÓN DEL DATASET

El análisis se realizó utilizando el conjunto de datos "diabetes_binary_5050split_health_indicators_BRFSS2015.csv" [1], el cual contiene 70,692 respuestas a la encuesta BRFSS2015 de los CDC. Este dataset está balanceado, con una distribución equitativa (50-50%) entre personas sin diabetes y aquellas con prediabetes o diabetes.

2a.- Features

El conjunto de datos incluye la variable de clasificación u objetivo y 21 variables predictoras que abarcan indicadores de salud física y mental, hábitos de vida, antecedentes médicos y características sociodemográficas.

Variable objetivo:

- **Diabetes_binary**: Indica si el encuestado es prediabético o diabético (1) o no (0).

Variables numéricas:

- **BMI**: Índice de Masa Corporal del encuestado.
- **MentHlth**: Indica el número de días en los últimos 30 en que la salud mental del encuestado no fue buena, incluyendo estrés, depresión y problemas con las emociones (0-30).
- **PhysHlth**: Indica el número de días en los últimos 30 en que la salud física del encuestado no fue buena, incluyendo enfermedades físicas y lesiones (0-30).

Variables categóricas binarias:

- **HighBP**: El encuestado ha sido diagnosticado con presión arterial alta (1) o no (0).
- **HighChol**: El encuestado ha sido diagnosticado con colesterol alto (1) o no (0).
- **CholCheck**: El encuestado se realizó un chequeo de colesterol en los últimos 5 años (1) o no (0).

- **Smoker**: Indica si el encuestado fumó al menos 100 cigarrillos en toda su vida (1) o no (0).
- **Stroke**: Indica si el encuestado ha sido diagnosticado con un derrame cerebral (1) o no (0).
- **HeartDiseaseorAttack**: Indica si el encuestado ha sido diagnosticado con una enfermedad coronaria o infarto de miocardio (1) o no (0).
- **PhysActivity**: Indica si el encuestado realizó actividad física en los 30 días pasados (1) o no (0).
- **Fruits**: Indica si el encuestado consume al menos una fruta diariamente (1) o no (0).
- **Veggies**: Indica si el encuestado consume al menos una verdura diariamente (1) o no (0).
- **HvyAlcoholConsump**: Indica si el encuestado bebe excesivamente (1), esto es en hombres más de 14 tragos semanales y en mujeres más de 7 tragos semanales. O si no excede estos valores (0).
- **AnyHealthcare**: Indica si el encuestado cuenta con algún tipo de cobertura de atención médica (1), incluyendo seguro de salud, planes prepagados como HMO, etcétera. O si no lo tiene (0).
- **NoDocbcCost**: Indica si el encuestado no pudo consultar a un médico en el último año por motivos de costo (1), o si no la necesitó o pudo acudir (0).
- **DiffWalk**: Indica si el encuestado experimenta dificultad seria para caminar o subir escaleras (1) o no (0).
- **Sex**: Indica el sexo del encuestado, femenino (0) o masculino (1).

Variables categóricas ordinales:

- **GenHlth**: Indica la autocalificación del encuestado de su salud en general en escala de 1 a 5, donde:
    - 1 = excelente
    - 2 = muy buena
    - 3 = buena
    - 4 = regular
    - 5 = mala
- **Age**: Representa la edad del encuestado agrupada en 13 rangos de 5 años cada uno. Cada valor numérico corresponde a un grupo de edad específico, por ejemplo:
    - 1 = 18-24 años
    - 2 = 25-29 años
    - …
    - 9 = 60-64 años
    - …
    - 13 = 80 años o más
- **Education**: Representa el nivel educativo más alto alcanzado por el encuestado, codificada en una escala del 1 al 6:
    - 1 = Nunca asistió a la escuela o solo kínder
    - 2 = Grados 1 a 8 (Primaria)
    - 3 = Grados 9 a 11 (Secundaria incompleta)
    - 4 = Grado 12 o GED (Preparatoria terminada)
    - 5 = 1 a 3 años de universidad o escuela técnica (Universidad/tecnológico incompleto)
    - 6 = 4 años o más de universidad (Universidad terminada)
- **Income**: Representa el rango de ingresos anuales del hogar del encuestado, codificada en una escala del 1 al 8:
    - 1 = Menos de $10,000

- 2 = $10,000 a menos de $15,000
- 3 = $15,000 a menos de $20,000
- 4 = $20,000 a menos de $25,000
- 5 = $25,000 a menos de $35,000
- 6 = $35,000 a menos de $50,000
- 7 = $50,000 a menos de $75,000
- 8 = $75,000 o más

3. EXTRACCIÓN DE DATOS

3a.- Carga y filtrado de datos

El conjunto de datos utilizado no requirió procesos de limpieza, ya que se obtuvo en formato depurado, sin valores nulos ni registros inválidos, como se indica en la fuente original. Esto permitió trabajar con todas las observaciones y variables disponibles.

3b.- Selección de variables mediante correlación

Para identificar las variables más representativas para generar el modelo de predicción de la condición diabética, se realizó un análisis de correlación entre las variables independientes y la variable objetivo. A partir de este análisis, se seleccionaron las 9 variables con mayor grado de asociación absoluta: GenHlth, HighBP, BMI, HighChol, Age, DiffWalk, Income, PhysHlth y HeartDiseaseorAttack.

* GenHlth: 0.407612

* HighBP: 0.381516

* BMI: 0.293373

* HighChol: 0.289213

* Age: 0.278738

* DiffWalk: 0.272646

* Income: 0.224449

* PhysHlth: 0.213081

* HeartDiseaseorAttack: 0.211523

Estas variables destacan por su relevancia clínica; por ejemplo, la autopercepción de salud general (GenHlth), la presencia de hipertensión (HighBP), el índice de masa corporal (BMI) y el colesterol alto (HighChol) son factores ampliamente reconocidos en la investigación médicca como asociados al riesgo de diabetes. La edad (Age) y la dificultad para caminar (DiffWalk) reflejan el impacto de la condición física y el envejecimiento en la salud metabólica. El nivel de ingresos (Income) y la salud física reportada (PhysHlth) aportan una perspectiva socioeconómica y de bienestar general, mientras que los antecedentes de enfermedades (HeartDiseaseorAttack) refuerzan la relación entre diabetes y ciertos problemas de salud, principalmente cardiovasculares.

La selección se fundamentó exclusivamente en los valores de correlación observados, lo cual se visualiza en el heatmap de correlaciones [Figura 1]. Este gráfico permite apreciar no sólo la relación

entre las variables seleccionadas y la objetivo, sino también la interacción entre otros indicadores que son relevantes, pero presentaron menor grado de asociación por lo que se excluyeron del modelo. De este modo, se prioriza la inclusión de variables que podrían presentar mayor peso pedictivo.

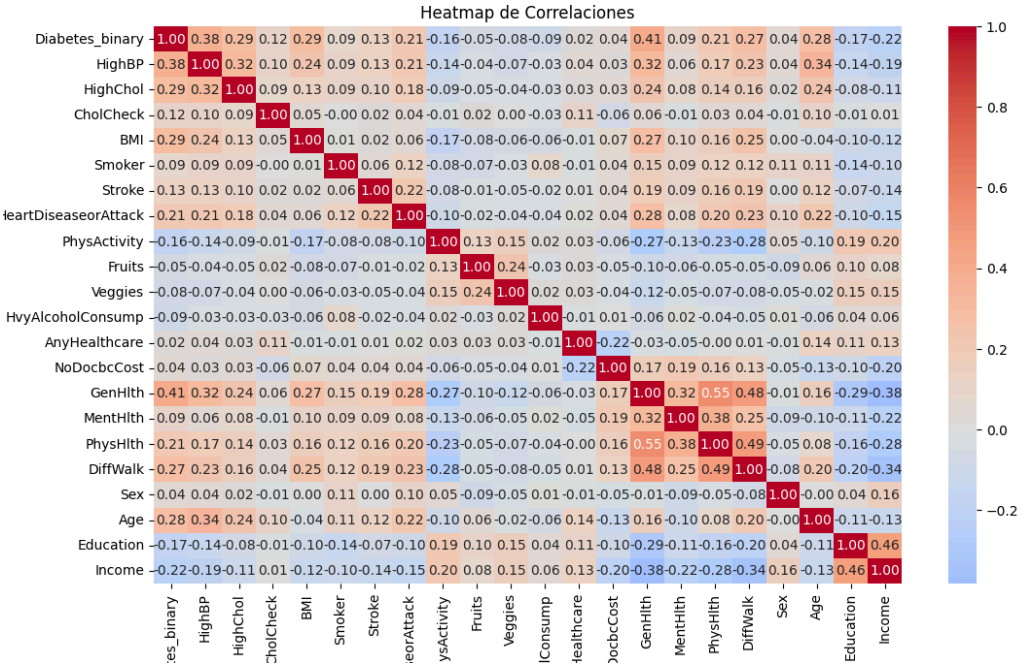

Figura 1. Mapa de correlaciones de variables para predecir diabetes

4. TRANSFORMACIÓN DE DATOS

4a.- Estandarización de variables

Como parte del preprocesamiento, se aplicó estandarización z-score a las variables numéricas y a las categóricas ordinales. Este procedimiento consiste en transformar cada valor restando la media de la variable y dividiendo entre su desviación estándar, lo que da como resultado una distribución con media cero y desviación estándar uno. Esta transformación no altera la relación entre los valores, pero sí permite que todas las variables estén en una escala comparable.[2]

La estandarización ayuda a evitar que variables con magnitudes mayores dominen el comportamiento del modelo, lo que puede afectar la interpretación de los coeficientes y la estabilidad del ajuste.

Las variables binarias se mantuvieron sin transformación, dado que ya se encuentran en una escala adecuada para el análisis.

4b.- División de Dataset en Conjuntos para Entrenamiento y Prueba

La división del conjunto de datos en subconjuntos de entrenamiento y prueba es una práctica que permite evaluar el desempeño del modelo sobre datos no vistos, garantizando que los resultados sean representativos y generalizables.

En este caso, se aprovechó la característica particular del dataset, que presenta una distribución balanceada (50% sin diabetes y 50% con prediabetes o diabetes, como se mencionó previamente) y

está ordenado, con la primera mitad correspondiente a personas sin diabetes y la segunda mitad a personas con prediabetes o diabetes. Para evitar sesgos dado este orden, se realizó una selección aleatoria de cada clase, y se dividió de forma que el 80% de los datos de cada grupo se destinara al entrenamiento y el 20% restante a la prueba. Finalmente, estos subconjuntos se mezclaron nuevamente para eliminar cualquier patrón que podría haber quedado.

Este procedimiento garantiza que ambos conjuntos mantengan el equilibrio entre clases, evitando el sobreajuste.

5. CONSTRUCCIÓN DEL MODELO

5a.- Función de hipótesis

En la regresión logística, la función de hipótesis integra las variables independientes y sus parámetros para generar una predicción lineal. Esta predicción se transforma mediante la función sigmoide, que convierte cualquier valor real en una probabilidad entre 0 y 1. Así, la hipótesis permite estimar la probabilidad de que un individuo pertenezca a una clase específica (por ejemplo, tener o no diabetes), facilitando la clasificación binaria a partir de múltiples factores.

Matemáticamente, la función se expresa como:

$$h(x) = \sigma(w_1x_1 + w_2x_2 + \ldots + w_nx_n + b)$$

donde σ representa la función sigmoide:

$$\sigma(z) = 1 / (1 + e^{(-z)})$$

5b.- Función de costo: Cross Entropy

La función de costo utilizada es la entropía cruzada (cross entropy), que mide la discrepancia entre las probabilidades predichas por el modelo y los valores reales observados. Esta función penaliza fuertemente las predicciones incorrectas, propiciando que el modelo ajuste sus parámetros para minimizar el error y mejorar la precisión en la clasificación.

La fórmula para el cálculo es la siguiente:

$$J = -(1/N)\ \Sigma\ [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

donde N es el número de muestras, y_i es el valor real, y p_i la probabilidad predicha para cada muestra.

5c.- Función de optimización: Gradiente descendiente

El gradiente descendiente es el método de optimización empleado para ajustar los parámetros del modelo. Consiste en calcular la dirección en la que la función de costo disminuye más rápidamente y actualizar los parámetros en esa dirección. Esta iteración permite que el modelo aprenda progresivamente, mejorando su capacidad predictiva en cada ciclo de entrenamiento.

Las fórmulas seguidas para actualiza los parámetros son:

$$w_n := w_n - \alpha\ (\partial J/\partial w_n)$$

$$b := b - \alpha\ (\partial J/\partial b)$$

donde α es la tasa de aprendizaje, y $\partial J/\partial w$ y $\partial J/\partial b$ son los gradientes calculados para los pesos y el sesgo, respectivamente. Hablaremos sobre estos parámetros a continuación.

6d.- Hiperparámetros

En el proceso de entrenamiento, tenemos dos hiperparámetros relevantes: el término de sesgo (bias) y la tasa de aprendizaje (learning rate). Su adecuada calibración es fundamental para lograr un desempeño óptimo que evite tanto el sobreajuste como la convergencia demasiado lenta.

El bias se recomienda inicializar en cero, permitiendo que el modelo comience sin preferencia por una clase y que se ajuste progresivamente durante el entrenamiento.

El learning rate se suele iniciar con un valor pequeño, como 0.1, para asegurar que los cambios en los parámetros sean graduales y evitar saltos bruscos que puedan dificultar la convergencia. Este valor se puede ajustar posteriormente acorde a los resultados obtenidos, buscando un equilibrio entre la velocidad de aprendizaje y la estabilidad del modelo.

6. ENTRENAMIENTO DEL MODELO

El proceso de entrenamiento consiste en ajustar los parámetros del modelo de regresión logística para que pueda predecir con precisión la probabilidad de que un individuo pertenezca a una clase específica. Para ello, se utiliza el conjunto de entrenamiento, sobre el cual el modelo realiza iteraciones sucesivas, aplicando la función de hipótesis, calculando el error mediante la función de costo (entropía cruzada) y optimizando los parámetros con el método de gradiente descendiente.

En cada época, el modelo actualiza los pesos y el sesgo a partir de los gradientes calculados, buscando minimizar el error y mejorar la precisión de las predicciones. El proceso se repite durante un número determinado de épocas, o hasta que el error de entrenamiento alcance un valor suficientemente bajo. Durante el entrenamiento, se monitorean métricas como la pérdida y la precisión tanto en el conjunto de entrenamiento como en el de prueba, lo que permite evaluar el desempeño y ajustar los hiperparámetros si es necesario.

7. REFERENCIAS

[1] CDC (2021, November 8). Diabetes Health Indicators Dataset. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?resource=download&select=diabetes_binary_5050split_health_indicators_BRFSS2015.csv

[2] Bhandari, A. (2025, April 23). What is Feature Scaling and Why is it Important? Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/