Análisis y Modelado Predictivo de Indicadores de Salud para la Detección de Diabetes mediante Machine Learning

1st Emanuel Vasquez Yepes
Facultad de Ingeniería
Universidad de Antioquia
Medellín, Colombia
emanuel.vasquea@udea.edu.co

2nd Jose Andres Echavarria Rios
Facultad de Ingeniería
Universidad de Antioquia
Medellín, Colombia
jose.echavarria1@udea.edu.co

I. Descripción del problema

1. Contexto del problema

La diabetes constituye una de las enfermedades crónicas más prevalentes a nivel mundial, afectando a millones de personas y generando un impacto económico y sanitario significativo. Este padecimiento se caracteriza por la incapacidad del organismo para regular adecuadamente los niveles de glucosa en sangre, lo cual puede derivar en complicaciones severas como enfermedades cardiovasculares, pérdida de visión o daño renal. La detección temprana es un factor crítico para mejorar la calidad de vida y reducir la mortalidad asociada.

En este contexto, el desarrollo de modelos predictivos basados en *Machine Learning* (ML) permite identificar individuos en riesgo y apoyar la toma de decisiones médicas y de salud pública, optimizando recursos y facilitando estrategias preventivas más eficaces.

Esto puede ser de gran utilidad para la detección temprana y prevención de la enfermedad, así como para apoyar tomas de decisiones médicas basadas en datos.

2. Composición de la base de datos

El conjunto de datos utilizado proviene del *Behavioral Risk Factor Surveillance System* (BRFSS) 2015, recopilado por el *Centers for Disease Control and Prevention* (CDC). La base procesada contiene 22 variables, de las cuales 21 son predictoras y una corresponde a la variable objetivo (`Diabetes_012`), con tres clases: 0 para no diabetico, 1 para prediabetico y 2 para diabetico. Inicialmente se contaba con 253,680 registros, observandose un marcado desbalance de clases, predominando la categoria sin diabetes. Para mitigar este problema y mejorar la capacidad de generalizacion del modelo, se aplico una estrategia de submuestreo, resultando en un conjunto final de 38,052 instancias. Todas las variables son de tipo numerico, pero algunas variables tienen valores numéricos que en realidad representan categorías, por lo que necesitan una codificación adicional antes de ser utilizadas en los modelos.

El conjunto de datos "Diabetes Health Indicators Dataset" (extraído de la base Behavioral Risk Factor Surveillance System — BRFSS 2015) contiene información proveniente de encuestas realizadas en Estados Unidos, enfocadas en hábitos de vida y condiciones de salud de los participantes.

A continuación, presentaremos el dataset antes y después del proceso de submuestreo donde explicaremos a fondo la distribución del mismo y haremos algunos énfasis en los resultados obtenidos en el análisis exploratorio.

#	Variable	Tipo de dato	Valores no nulos
0	Diabetes_012	float64	253,680
1	HighBP	float64	253,680
2	HighChol	float64	253,680
3	CholCheck	float64	253,680
4	BMI	float64	253,680
5	Smoker	float64	253,680
6	Stroke	float64	253,680
7	HeartDiseaseorAttack	float64	253,680
8	PhysActivity	float64	253,680
9	Fruits	float64	253,680
10	Veggies	float64	253,680
11	HvyAlcoholConsump	float64	253,680
12	AnyHealthcare	float64	253,680
13	NoDocbcCost	float64	253,680
14	GenHlth	float64	253,680
15	MentHlth	float64	253,680
16	PhysHlth	float64	253,680
17	DiffWalk	float64	253,680
18	Sex	float64	253,680
19	Age	float64	253,680
20	Education	float64	253,680
21	Income	float64	253,680

Cuadro I: Estructura del conjunto de datos BRFSS 2015. Todas las variables son de tipo `float64` y no presentan valores nulos.

A continuación, explicaremos cada una de estas variables y como se encuentran codificadas actualmente en el dataset, esto nos ayudara a contextualizar como se tomarán decisiones y dará un mejor contexto del problema.

Variable	Tipo	Descripción
Diabetes_012	Categórica (0, 1, 2)	**Variable objetivo.** Indica el estado de diabetes: 0: No diabetes, 1: Prediabetes, 2: Diabetes.
HighBP	Binaria (0, 1)	Diagnóstico de presión arterial alta.
HighChol	Binaria (0, 1)	Diagnóstico de colesterol alto.
CholCheck	Binaria (0, 1)	Revisión de colesterol en los últimos 5 años.
BMI	Numérica (float)	Índice de Masa Corporal (peso / altura^2).
Smoker	Binaria (0, 1)	Ha fumado al menos 100 cigarrillos en su vida.
Stroke	Binaria (0, 1)	Ha sufrido un accidente cerebrovascular.
HeartDiseaseorAttack	Binaria (0, 1)	Enfermedad cardíaca o ataque al corazón previo.
PhysActivity	Binaria (0, 1)	Realizó actividad física en los últimos 30 días.
Fruits	Binaria (0, 1)	Consume frutas al menos una vez al día.
Veggies	Binaria (0, 1)	Consume vegetales al menos una vez al día.
HvyAlcoholConsump	Binaria (0, 1)	Consumo excesivo de alcohol: más de 14 bebidas/semana (hombres) o más de 7 (mujeres).
AnyHealthcare	Binaria (0, 1)	Tiene cobertura médica o acceso a servicios de salud.
NoDocbcCost	Binaria (0, 1)	No pudo acudir al médico en el último año por costo.
GenHlth	Ordinal (1–5)	Salud general autoevaluada: 1 = Excelente, 5 = Mala.
MentHlth	Numérica (0–30)	Días con mala salud mental en el último mes.
PhysHlth	Numérica (0–30)	Días con mala salud física en el último mes.
DiffWalk	Binaria (0, 1)	Dificultad para caminar o subir escaleras.
Sex	Binaria (0, 1)	Sexo biológico: 0 = Mujer, 1 = Hombre.
Age	Categórica (1–13)	Grupo de edad: 1 = 18–24, 2 = 25–29, ..., 13 = 80+.
Education	Ordinal (1–6)	Nivel educativo: 1 = Ninguno, 6 = Título universitario.
Income	Ordinal (1–8)	Nivel de ingreso: 1 = <10k USD, 8 = 75k USD.

Cuadro II: Descripción de las variables del conjunto de datos de indicadores de salud (BRFSS).

Es importante aclarar que, aunque todas las variables del conjunto de datos se encuentran almacenadas con tipo `float64`, la mayoría representan categorías discretas o respuestas binarias (0 o 1). Por tanto, se consideran variables categóricas codificadas numéricamente. La única variable verdaderamente continua es `BMI`, que corresponde al índice de masa corporal y presenta una distribución de valores reales en el rango aproximado de 12 a 98.

El siguiente paso será validar y revisar la distribución de los datos en la variable objetivo:

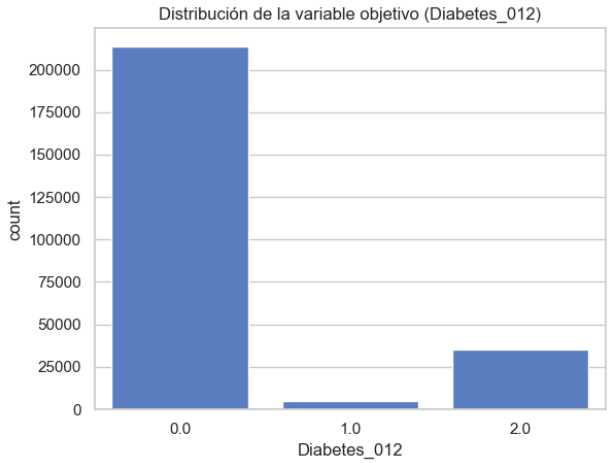

Figura 1: Distribución de la variable objetivo `Diabetes_012`. Se observa un claro desbalance entre las clases, predominando la categoría 0 (no diabético).

La variable objetivo `Diabetes_012` presenta un marcado desbalance entre clases, lo que puede afectar negativamente el desempeño de los modelos de clasificación. Como se observa en la Tabla IV, la mayoría de los registros corresponden a individuos sin diagnóstico de diabetes, mientras que las clases de prediabetes y diabetes representan una proporción mucho menor.

Clase (`Diabetes_012`)	Proporción (%)	Cantidad
0 (No diabético)	84.24	213703
1 (Prediabético)	1.83	4631
2 (Diabético)	13.93	35346

Cuadro III: Distribución de clases en la variable objetivo `Diabetes_012`.

La matriz de correlación muestra cómo se relacionan entre sí las variables numéricas del dataset. Cada valor en la matriz representa un coeficiente de correlación pearson entre dos variables. A continuación se presenta la matriz obtenida para el conjunto de datos

La matriz de correlación permite identificar relaciones lineales entre las variables del conjunto de datos. Se observan correlaciones positivas moderadas entre variables relacionadas con condiciones de salud, como HighBP y HeartDiseaseorAttack, lo que sugiere que la hipertensión puede coexistir con enfermedades cardíacas. Por otro lado, variables como Age, BMI y GenHlth presentan correlaciones más bajas, indicando que su relación con el diagnóstico de diabetes puede ser más compleja y no estrictamente lineal.

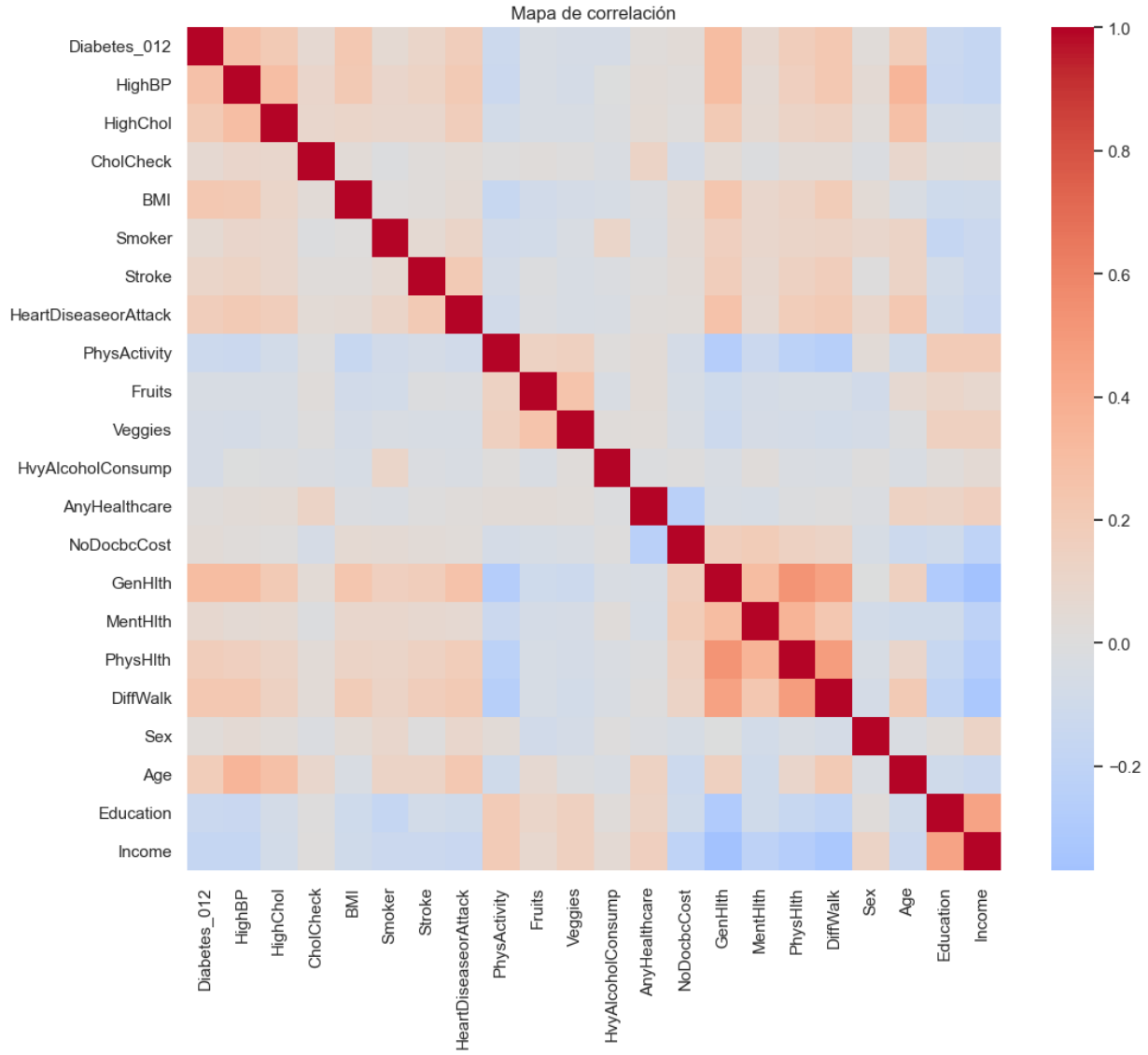

Figura 2: Matriz de correlación `Diabetes_012`.

Para finalizar la exploracion del dataset se presentara de manera analoga al dataset completo los datos obtenidos mediante el submuestreo y se dara una breve justificación del por que de este:

Para reducir el costo computacional sin modificar la estructura estadística del conjunto de datos, se aplicó un submuestreo estratificado, el cual selecciona aleatoriamente una fracción del conjunto original manteniendo las proporciones de la variable objetivo. De esta forma, se preserva el desbalance natural del problema, lo que permite evaluar posteriormente el impacto de dicho desbalance en el rendimiento del modelo.

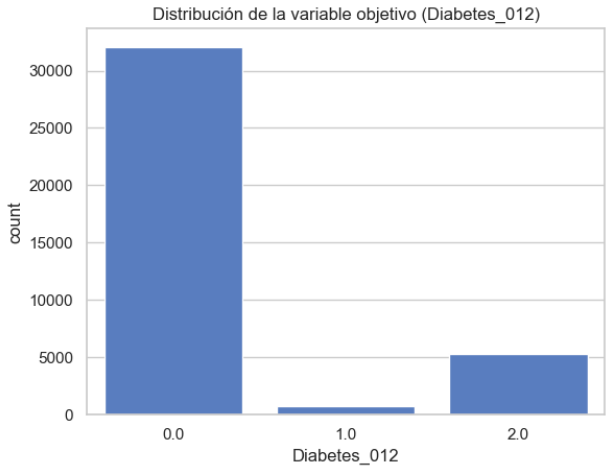

Figura 3: Distribución de la variable objetivo despues del submuestreo `Diabetes_012`. Se observa que se mantiene el desbalance en las tres clases

Clase (`Diabetes_012`)	Proporción (%)	Cantidad
0 (No diabético)	84.24	32055
1 (Prediabético)	1.82	695
2 (Diabético)	13.93	5302

Cuadro IV: Distribución de clases en la variable objetivo `Diabetes_012` despues del submuestreo

Podemos observar que después del submuestreo los datos conservan las mismas características del dataset original, manteniendo la correlación igual y el desbalance. Con esto garantizamos que, aunque estemos usando una porción de los datos el problema original que consiste en enfrentarse al efecto del desbalance se mantiene.

En resumen, el análisis exploratorio del conjunto de datos permitió comprender la naturaleza y composición de las variables empleadas para la predicción de diabetes. Se identificó que la mayoría de las variables corresponden a características categóricas codificadas numéricamente, con excepción del índice de masa corporal (`BMI`), que presenta valores continuos. El análisis de la variable objetivo evidenció un marcado desbalance entre clases, predominando los casos sin diagnóstico de diabetes.

Asimismo, la matriz de correlación permitió observar asociaciones moderadas entre variables relacionadas con la salud cardiovascular, como `HighBP` y `HeartDiseaseorAttack`, lo que respalda su relevancia en la detección del riesgo de diabetes. Finalmente, mediante el submuestreo estratificado se logró reducir el tamaño del conjunto de datos sin alterar su distribución original, conservando el desbalance inherente del problema.

3. Paradigma de aprendizaje y justificación

Dado que la variable objetivo representa tres categorías mutuamente excluyentes, el problema se aborda como una tarea de clasificación supervisada multiclase. Este paradigma permite al modelo aprender patrones asociados con diferentes niveles de riesgo de diabetes a partir de las características sociodemográficas y de comportamiento reportadas en la encuesta. La elección de este enfoque se fundamenta en la necesidad de discriminar entre distintos estados de salud (sin diabetes, prediabetes y diabetes) de manera precisa, con el objetivo de contribuir al desarrollo de herramientas predictivas que puedan integrarse en sistemas de apoyo a la decisión médica y en programas de monitoreo poblacional.

Además, debido a la naturaleza multiclase del problema y a la mezcla de variables numéricas y categóricas, se consideran adecuados modelos de Machine Learning como la regresión logística multinomial, los árboles de decisión y los clasificadores basados en ensambles, como Random Forest o Gradient Boosting. Estos algoritmos permiten capturar relaciones no lineales entre los indicadores de salud y el riesgo de diabetes, manteniendo un equilibrio entre interpretabilidad y capacidad predictiva. La selección final del modelo dependerá del rendimiento obtenido en las métricas de evaluación y de su capacidad para manejar el desbalance de clases presente en los datos.

II. Estado del arte

Artículo 1: Diabetic Mellitus Prediction with BRFSS Data Sets [1]

- **Paradigma:** Aprendizaje supervisado. El estudio aplica distintos algoritmos de clasificación para predecir la diabetes tipo 2 en función de variables de salud obtenidas de los conjuntos de datos BRFSS 2014 y 2015.
- **Técnicas:** Se utilizaron múltiples modelos de Machine Learning: Árbol de Decisión, Regresión Logística, SVM (lineal, RBF y polinomial), Naïve Bayes, Random Forest y Redes Neuronales. Además, se aplicó SMOTE para balancear las clases (diabético, prediabético y no diabético).
- **Validación:** Se realizó una división del conjunto de datos en entrenamiento (dos tercios) y prueba (un tercio). No se especifica validación cruzada, pero se comparan los modelos usando los mismos subconjuntos.
- **Métricas:** Exactitud (Accuracy), Sensibilidad (Sensitivity), Especificidad (Specificity) y Área Bajo la Curva (AUC).
- **Resultados:** En el conjunto BRFSS 2014, la Red Neuronal obtuvo la mayor exactitud (82.4 %) y especificidad (90.2 %). El Árbol de Decisión logró la mayor sensibilidad (51.6 %). Para 2015, los modelos CatBoost y XGBoost alcanzaron 86 % de exactitud, superando a los demás clasificadores.
- **Fuente:** M. H. Mohamed et al., "Diabetic Mellitus Prediction with BRFSS Data Sets," *Journal of Theoretical and Applied Information Technology*, vol. 102, no. 3, 2024. Disponible en: www.jatit.org/volumes/Vol102No3/10Vol102No3.pdf

Artículo 2: An investigation of machine learning algorithms and data augmentation techniques for diabetes diagnosis using class imbalanced BRFSS dataset [2]

- **Paradigma:** Aprendizaje supervisado enfocado en el manejo del desbalance de clases. El estudio utiliza el conjunto de datos BRFSS (Behavioral Risk Factor Surveillance System) para diagnosticar diabetes, aplicando técnicas de aumento y balanceo antes del entrenamiento de los modelos.
- **Técnicas:** Se emplearon métodos de muestreo como oversampling (SMOTE-N, una versión adaptada de SMOTE para datos nominales), undersampling (Edited Nearest Neighbour, ENN) y enfoques híbridos (SMOTE-Tomek, SMOTE-ENN). Los algoritmos de Machine Learning utilizados incluyen Regresión Logística, Random Forest, Gradient Boosting y AdaBoost, todos evaluados sobre los diferentes conjuntos de datos generados.
- **Validación:** Los autores enfatizan la prevención de *data leakage* (fuga de datos) al aplicar las técnicas de muestreo antes de la división de datos de entrenamiento y prueba. Esto garantiza que los datos aumentados no contaminen el conjunto de validación. La comparación de modelos se realiza bajo las mismas condiciones experimentales.

- **Métricas:** Se usaron métricas estándar de clasificación, principalmente Exactitud (Accuracy) y Sensibilidad (Recall), para evaluar el efecto del balanceo y comparar el rendimiento entre algoritmos.
- **Resultados:** Los resultados mostraron que la técnica ENN (Edited Nearest Neighbour) produjo un rendimiento superior en comparación con los demás métodos de balanceo. Entre los factores más influyentes para el diagnóstico se identificaron la edad, el índice de masa corporal (IMC) y la presión arterial alta, coherentes con hallazgos clínicos previos.
- **Fuente:** M. M. Chowdhury, R. S. Ayon, M. S. Hossain, "An investigation of machine learning algorithms and data augmentation techniques for diabetes diagnosis using class imbalanced BRFSS dataset," *Healthcare Analytics*, vol. 5, p. 100297, 2024. DOI: doi.org/10.1016/j.health.2023.100297

REFERENCIAS

[1] Chowdhury, M. M., Ayon, R. S., & Hossain, M. S. (2024). *An investigation of machine learning algorithms and data augmentation techniques for diabetes diagnosis using class imbalanced BRFSS dataset*. *Healthcare Analytics*, 5, 100297. ISSN: 2772-4425. Disponible en: doi.org/10.1016/j.health.2023.100297

[2] Mohamed, M. H., Kamel, M. H. K., Said, W., & Mohamed, N. (2024). *Diabetic Mellitus Prediction with BRFSS Data Sets*. *Journal of Theoretical and Applied Information Technology*, 102(3), 15th February 2024. Little Lion Scientific. ISSN: 1992-8645, E-ISSN: 1817-3195. Disponible en: www.jatit.org/volumes/Vol102No3/10Vol102No3.pdf

[3] Kaggle. (2015). *Diabetes Health Indicators Dataset (BRFSS 2015)*. Recuperado de: www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset