practica

May 17, 2025

Estudiante :

- Escriba Flores, Daniel Agustin

1 Primer Caso de Estudio: Predicción Temprana de Diabetes mediante Regresión Logística

---

1.1 Contexto

Se desea construir un modelo predictivo que permita estimar el riesgo de padecer diabetes en función de características relacionadas con la salud de una persona (por ejemplo: presión alta, colesterol alto, actividad física, entre otras). Para ello, se ha utilizado el conjunto de datos CDC Diabetes Health Indicators, disponible en el UCI Machine Learning Repository. Este dataset contiene información de más de 250.000 registros, con variables numéricas y categóricas relacionadas con hábitos y condiciones de salud.

La variable objetivo del modelo es binaria: - 0 = No presenta diabetes - 1 = Presenta diabetes

Debido a que el número de personas diagnosticadas con diabetes es considerablemente menor al de personas sin diagnóstico, este caso representa un escenario con clases desbalanceadas, frecuente en contextos de medicina preventiva.

1.2 Preguntas de Análisis e Interpretación

1. **¿El dataset presenta un problema de desbalance de clases? Justifique su respuesta con base en los porcentajes observados.**

Sí, el dataset presenta un claro desbalance de clases.La distribución muestra que el 86.07% de los registros corresponden a personas sin diabetes (Clase 0) , mientras que solo el 13.93% son casos positivos (Clase 1). Se hace mas visible con el grafico de Distribucion encontrado

> Esta gran diferencia indica que el modelo puede tener dificultades para aprender patrones asociados a la clase minoritaria, afectando su capacidad predictiva en esa categoría.

2. **Explique cómo el desbalance de clases afectó al modelo de regresión logística sin SMOTE, especialmente en su capacidad para detectar personas con diabetes.**

Sin SMOTE, el modelo priorizo la clase mayoritaria (0), que son precisamente las personas que no tienen diabetes. Esto se refleja claramente en el bajo recall de la Clase 1 (0.16) , lo que significa que solo identificó correctamente al 16% de los casos reales de diabetes .Aunque el accuracy fue alto (0.86) , este valor es engañoso, ya que solo se enfoca en la clase mayoriyaria. Por eso mismo, el F1-score para la Clase 1 fue muy bajo (0.24) , reflejando una mala combinación entre precisión y recall. > En resumen, el desbalance provocó que el modelo fallara en detectar la mayoría de los casos de diabetes.

3. **Luego de aplicar SMOTE, ¿qué cambios se observan en las métricas para la clase 1 (diabetes)? Comente los beneficios y las posibles consecuencias de este cambio.**

Al aplicar SMOTE, se observa una mejora significativa en el recall para la Clase 1 (de 0.16 a 0.76) , lo que implica que ahora se detecta casi el 76% de los casos reales de diabetes , reduciendo notablemente los falsos negativos .

Sin embargo, esta mejora viene con un costo: la precisión disminuye considerablemente (de 0.53 a 0.31) , lo que genera más falsos positivos (personas sin diabetes clasificadas erróneamente como diabéticas).

Sobre el F1-score este Aumentó ligeramente, pasando de 0.24 a 0.44 , aunque sigue siendo bajo debido al desequilibrio entre precisión y recall.

> Esto podría traducirse en diagnósticos preliminares incorrectos que generen falsas alarmas, lo que podría requerir revisiones adicionales o exámenes complementarios.

4. **¿Cuál de los dos modelos considera más apropiado para un contexto de salud pública, en el que es fundamental identificar la mayor cantidad posible de personas con diabetes, aunque se cometan algunos falsos positivos? Justifique su respuesta con base en las métricas.**

Aunque ninguno de los dos modelos alcanza un desempeño óptimo para un uso clínico directo, el modelo con SMOTE resulta más adecuado en un contexto de salud pública donde la detección temprana de diabetes es prioritaria.

Este modelo logra un recall del 0.76 para la clase 1 (diabetes) , lo cual implica que ahora se identifica correctamente al 76% de los casos reales , un gran salto desde el 16% del modelo sin balanceo. Este aumento en el recall es clave para minimizar los falsos negativos , es decir, dejar pasar por alto a personas que sí tienen la enfermedad.

> Aunque como se observa esto trae como consecuencia un aumento en los falsos positivos , estos pueden gestionarse mediante exámenes adicionales o revisiones médicas, lo cual es preferible a no detectar casos reales. En este tipo de contextos preventivos, detectar más casos potenciales, incluso con algunas alertas falsas, es generalmente más deseable que no detectar los verdaderos casos .

5. **Explique por qué la métrica de accuracy puede ser engañosa en problemas con clases desbalanceadas. ¿Qué métricas deben priorizarse en este tipo de problemas y por qué?**

Cuando hay desbalanceo en clases, La accuracy puede ser engañosa porque mide la proporción total de predicciones correctas sin distinguir entre tipos de error. Un modelo puede alcanzar un alto accuracy simplemente prediciendo siempre la clase mayoritaria, ignorando completamente la

clase minoritaria. Por ejemplo, en nuestro caso, el modelo sin SMOTE tuvo un accuracy de 0.86 , pero solo identificó al 16% de los casos reales de diabetes.

En escenarios desbalanceados, es preferible usar métricas que evalúen el desempeño en cada clase:

- Recall : Mide cuántos verdaderos positivos fueron identificados. Es clave cuando queremos minimizar los falsos negativos .
- Precision : Evalúa cuántas de las predicciones positivas son realmente correctas. Útil para controlar los falsos positivos .
- F1-score : Combina precisión y recall, ofreciendo un equilibrio útil en escenarios desbalanceados.
- AUC-ROC : Evalúa el desempeño global del modelo en distintos umbrales, permitiendo comparar modelos incluso en presencia de desbalance.

Estas métricas ofrecen una visión más realista del comportamiento del modelo, especialmente en la clase minoritaria, lo que es crucial en contextos médicos.

Codigos del Primer Caso