Sí tiene sentido dividir y comparar si hay **una motivación clara** detrás del umbral. Pero si el umbral es arbitrario o fue elegido luego de mirar los datos, **debes ser muy cuidadoso con las inferencias que haces** y considerar métodos más robustos o justificar bien tu elección.

Si me cuentas más sobre tu caso (datos, objetivo, cómo elegiste el umbral), te puedo ayudar a decidir el mejor enfoque.

mi objetivo es separar en diabéticos y no diabéticos en función de la hemoglobina glicosilada y luego ver si otras variables correlacionan con el diagnóstico

¡Perfecto! En ese caso **sí tiene mucho sentido usar un umbral**, porque:

**✅ Estás usando un umbral clínicamente validado:**

Para **hemoglobina glicosilada (HbA1c)**, las guías clínicas (como la ADA y la OMS) definen:

| **Diagnóstico** | **HbA1c (%)** |
| --- | --- |
| Normal | < 5.7 |
| Prediabetes | 5.7–6.4 |
| Diabetes | ≥ 6.5 |

Así que si decides usar **HbA1c ≥ 6.5% como umbral para definir “diabéticos”**, estás actuando conforme a la evidencia médica.

**✅ Tiene sentido hacer comparaciones estadísticas**

Después de dividir tu conjunto en:

* **Grupo 1: No diabéticos (HbA1c < 6.5%)**
* **Grupo 2: Diabéticos (HbA1c ≥ 6.5%)**

Puedes aplicar análisis como:

* **Comparación de medias**: t-test o Mann-Whitney para edad, colesterol, presión, etc.
* **Comparación de proporciones**: chi-cuadrado si analizas género, fumadores sí/no, etc.
* **Correlaciones**: Ver si alguna variable continua (como glucosa en ayunas) se relaciona con el grupo (codificado como 0/1), aunque puede ser más potente usar regresión.

**🛠️ Consejo extra: prueba con regresión logística**

Si quieres ir más allá de simples comparaciones, puedes modelar:

text

CopiarEditar

P(Diabetes) = f(edad, colesterol, IMC, etc.)

Con una **regresión logística**, puedes:

* Ver qué variables se asocian fuertemente al diagnóstico,
* Predecir probabilidad de diabetes según otras variables.

**📌 Conclusión**

Tu enfoque es **correcto y estadísticamente válido**, ya que usas un **punto de corte con base clínica**, y luego investigas **diferencias entre grupos definidos por una condición médica real**.