Estadística Multivariada Computacional
20/12/2024

Joana Auriello
5.004.559-1

Introducción

Análisis descriptivo/exploratorio de datos

Previo a aplicar distintas técnicas estadísticas realizaremos un breve análisis descriptivo y exploratorio de los datos.

Se presenta una base de datos de dimensiones 2278x9, que corresponde a una encuesta sobre el estado de salud de personas adultas y mayores en Estados Unidos para los años 2013 y 2014.
Las columnas presentes en el conjunto de datos son las siguientes:

- GE (grupo etario)= puede ser adulto (adulto <65) o senior (>=65).
- Ed (edad)= edad del paciente
- Ge (género)= 1 si el paciente es masculino, 2 si es femenino.
- AF (actividad física)= vale 1 si el encuestado participa en deportes, actividades físicas o actividades recreativas de intensidad moderada o vigorosa en la semana típica, en caso contrario vale 2.
- IMC (índice de masa corporal)= índice de masa corporal del encuestado.
- Glu (glucosa)= nivel de glucosa en sangre después del ayuno.
- Diag (diagnóstico)= 1 si el encuestado no fue diagnosticado como diabético y 0 en caso contrario (cuando sí lo fue o no sabe).
- GLT (glucosa total)= nivel de glucosa de todas las líneas sanguíneas.
- In (insulina)= nivel de insulina en la sangre del encuestado.

Se presenta un resumen de los tipos de datos presentes en cada columna y las dimensiones del DF:

```
Tipos de datos:
GE        object
Ed       float64
Ge       float64
AF       float64
IMC      float64
Glu      float64
Diag     float64
GLT      float64
In       float64
dtype: object

Dimensiones del DataFrame:
Filas: 2278, Columnas: 9
```

A su vez, también se analizó una muestra de observaciones (primeras filas) del subconjunto de datos que se utiliza para realizar PCA, y se presenta un resumen de las estadísticas descriptivas. No se observa valores faltantes en ninguna de las variables que se utilizaran en PCA.

```
Primeras filas del subconjunto de datos:
     Ed   IMC    Glu    GLT     In
0  61.0  35.7  110.0  150.0  14.91
1  26.0  20.3   89.0   80.0   3.85
2  16.0  23.2   89.0   68.0   6.14
3  32.0  28.9  104.0   84.0  16.15
4  38.0  35.9  103.0   81.0  10.92

Resumen de valores faltantes:
Ed     0
IMC    0
Glu    0
GLT    0
In     0
dtype: int64

Estadísticas descriptivas:
                Ed          IMC          Glu          GLT           In
count  2278.000000  2278.000000  2278.000000  2278.000000  2278.000000
mean     41.795874    27.955180    99.553117   114.978929    11.834794
std      20.156111     7.248962    17.889834    47.061239     9.718812
min      12.000000    14.500000    63.000000    40.000000     0.140000
25%      24.000000    22.800000    91.000000    87.000000     5.860000
50%      41.000000    26.800000    97.000000   105.000000     9.040000
75%      58.000000    31.200000   104.000000   130.000000    14.440000
max      80.000000    70.100000   405.000000   604.000000   102.290000
```

PCA

El Análisis de Componentes Principales (PCA) es una técnica estadística que permite reducir la dimensionalidad de los datos al transformar las variables originales en un conjunto más pequeño de variables no correlacionadas llamadas componentes principales. Este método conserva la mayor parte de la variabilidad presente en los datos, y facilita la interpretación y visualización.

Antes de aplicar PCA, es fundamental centrar y estandarizar los datos. Esto se debe a que las variables suelen estar medidas en diferentes escalas, y PCA se basa en la matriz de covarianzas o correlaciones, lo que puede llevar a que las variables con escalas más grandes dominen el análisis. Al centrar los datos (restar la media) y estandarizarlos (dividir por la desviación estándar), garantizamos que todas las variables contribuyan de manera equitativa al análisis.

La matriz de rotación (o matriz de coeficientes de componentes principales) refleja cómo cada variable original se proyecta en los componentes principales. Los valores de esta matriz indican la contribución de cada variable original a cada componente principal, lo que nos permite interpretar las relaciones entre las variables y los componentes.

La matriz de coeficientes obtenida para los primeros tres componentes principales se presenta a continuación:

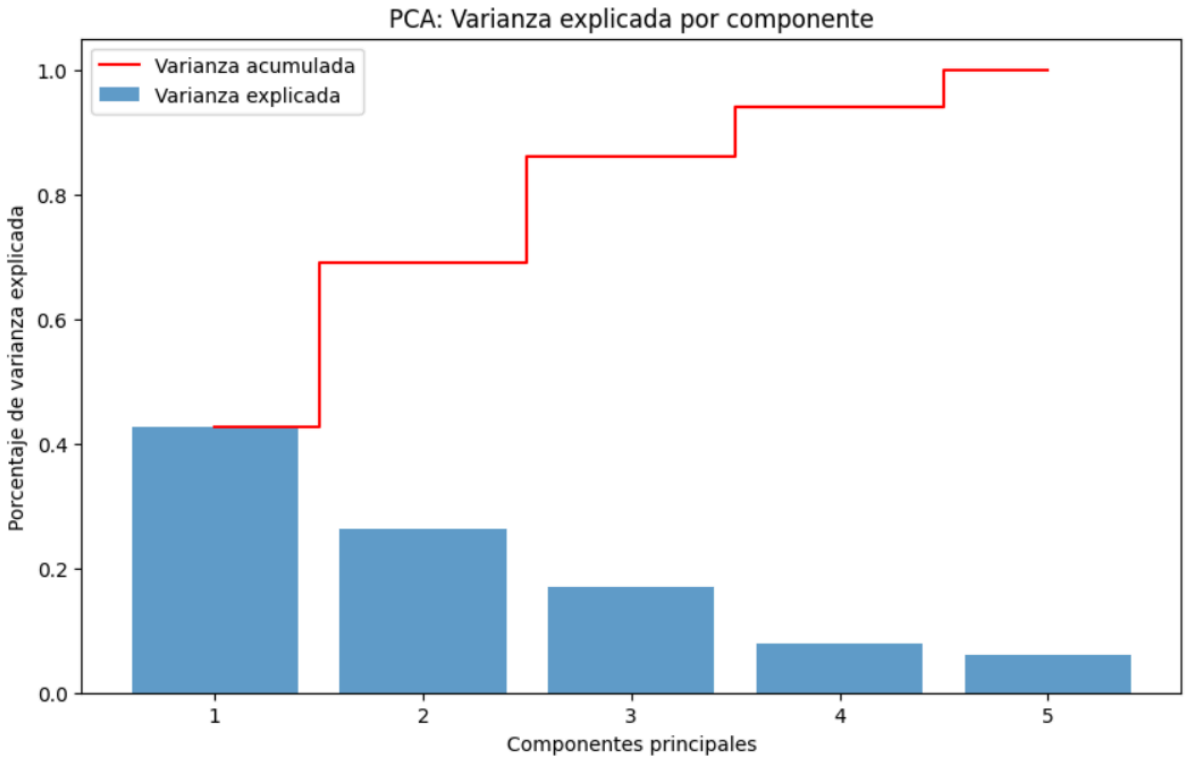
```
Matriz de coeficientes de los primeros 3 componentes principales:
                    Ed       IMC       Glu       GLT        In
Componente 1  0.286243  0.417884  0.539243  0.550846  0.386295
Componente 2 -0.455247  0.498118 -0.266529 -0.310936  0.613929
Componente 3  0.773575  0.406440 -0.392960 -0.278202 -0.067638

Porcentaje de varianza explicada (primeros 3 componentes):
     Componente  Varianza Explicada (%)  Varianza Acumulada (%)
0  Componente 1                   42.71                   42.71
1  Componente 2                   26.29                   69.00
2  Componente 3                   17.09                   86.09
```

En la matriz de coeficientes, cada fila representa un componente principal y cada columna el peso de una variable. Los signos de estos coeficientes permiten interpretar cómo cada variable contribuye a las componentes principales, ya sea de forma positiva o negativa. A continuación, se detalla una lectura breve:

1. Componente 1:
    - Las variables con mayores contribuciones positivas son glucosa total (GLT) (0.551), glucosa (Glu) (0.539) y índice de masa corporal (IMC) (0.418).
    - El signo positivo indica que este componente está asociado a un aumento simultáneo en estos indicadores. Este componente parece capturar una dimensión metabólica relacionada con la glucosa y el IMC.
2. Componente 2:
    - Las contribuciones más significativas son de nivel de insulina (In) (0.614) y índice de masa corporal (IMC) (0.498), ambos con signos positivos.
    - Por otro lado, edad (Ed) (-0.455) tiene un peso negativo. Esto sugiere que este componente representa una relación inversa entre la edad y los indicadores metabólicos como insulina y IMC.
3. Componente 3:
    - Edad (Ed) tiene el peso más alto y positivo (0.774), lo que indica que este componente está fuertemente relacionado con la edad.
    - Por otro lado, las variables glucosa (Glu) (-0.393) y glucosa total (GLT) (-0.278) presentan coeficientes negativos, indicando que el tercer componente podría capturar un efecto inverso entre la edad y ciertos indicadores metabólicos.

Varianza Explicada y Varianza Explicada Acumulada.



Análisis de la varianza explicada por componente

El gráfico muestra el porcentaje de la varianza explicada por cada componente principal (barras azules) y la varianza acumulada (línea roja) en el análisis PCA. Este análisis ayuda a determinar cuántos componentes principales son necesarios para representar la mayoría de la información contenida en los datos.

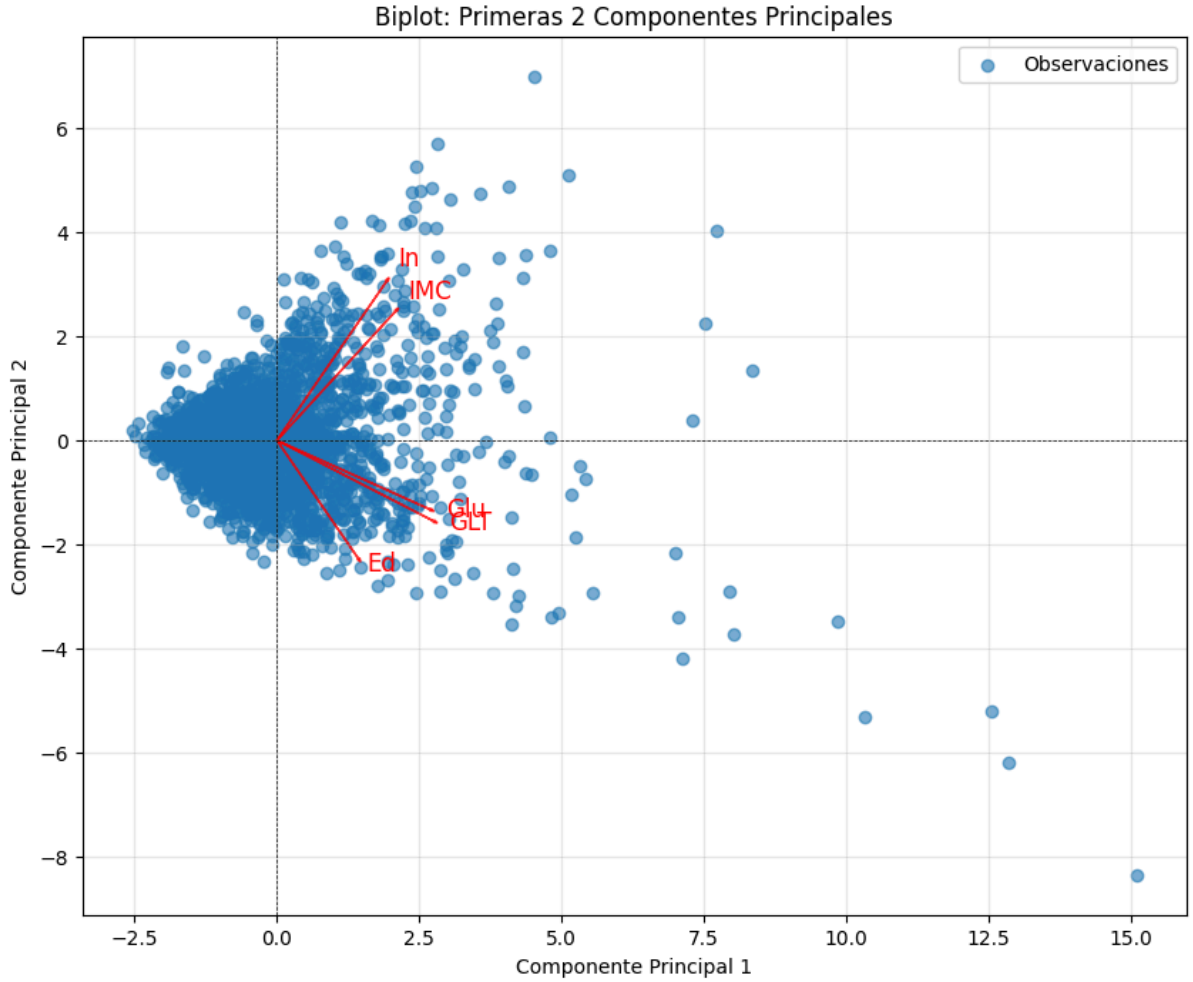
1. Primer componente (PC1):
    - El primer componente explica aproximadamente el 42% de la varianza total.
    - Esto indica que una sola dimensión capta casi la mitad de la información contenida en los datos originales, combinando variables como glucosa, glucosa total e IMC, que tienen altos pesos en este componente.
2. Segundo componente (PC2):
    - El segundo componente añade un 26% adicional de varianza explicada.
    - Combinado con el primer componente, ambos explican un total de aproximadamente el 69% de la varianza, lo que significa que estas dos dimensiones capturan la mayor parte de las relaciones entre las variables originales.
3. Tercer componente (PC3):
    - El tercer componente explica aproximadamente un 17% adicional de la varianza.
    - Con este componente, el total acumulado alcanza el 86%, lo que implica que los primeros tres componentes son suficientes para retener la mayoría de la información del conjunto de datos.
4. Cuarto y quinto componentes (PC4 y PC5):

- Los últimos dos componentes explican un porcentaje de varianza mucho menor, siendo aproximadamente 8% y 7%, respectivamente.

En conclusión, los primeros dos componentes principales explican el 69% de la varianza total, lo que indica que gran parte de la información puede resumirse en estas dos dimensiones. Agregar un tercer componente eleva la varianza acumulada al 86%, capturando casi toda la variabilidad relevante. Los últimos componentes (4 y 5) tienen contribuciones menores y no son críticos para el análisis.

Biplot

El biplot muestra las observaciones proyectadas en el espacio definido por las dos primeras componentes principales, junto con los vectores que representan las variables originales. Este gráfico combina la información de las variables y las observaciones, facilitando la interpretación de las relaciones en los datos.



¿Es informativo el Biplot?

A simple vista, las observaciones se concentran en torno al origen, lo que indica que muchas tienen valores promedio en las dos primeras componentes principales. Sin embargo, hay algunas observaciones que se alejan significativamente del centro,

especialmente hacia el eje de la primera componente, lo que podría indicar patrones específicos o individuos con características atípicas(datos atípicos sobre la derecha del gráfico).

Por otro lado, la longitud de los vectores indica la importancia de cada variable en las dos primeras componentes principales. Variables como GLT, Glu e In tienen vectores más largos, lo que sugiere que estas variables tienen una mayor contribución a las componentes principales y explican una mayor proporción de la varianza en los datos.

Además, la dirección de los vectores refleja cómo las variables están alineadas con las componentes principales. Por ejemplo:

- GLT y Glu están orientadas hacia la misma dirección, indicando que están positivamente correlacionadas.
- Ed (edad) está orientada en una dirección opuesta a In y IMC, lo que indica una relación inversa entre la edad y estas variables metabólicas.

Finalmente, los ángulos entre los vectores permiten interpretar las correlaciones entre las variables originales:

- Vectores como Glu y GLT, que forman ángulos pequeños, tienen una correlación alta y positiva.
- Por el contrario, Ed tiene ángulos amplios con In y IMC, indicando correlaciones negativas.

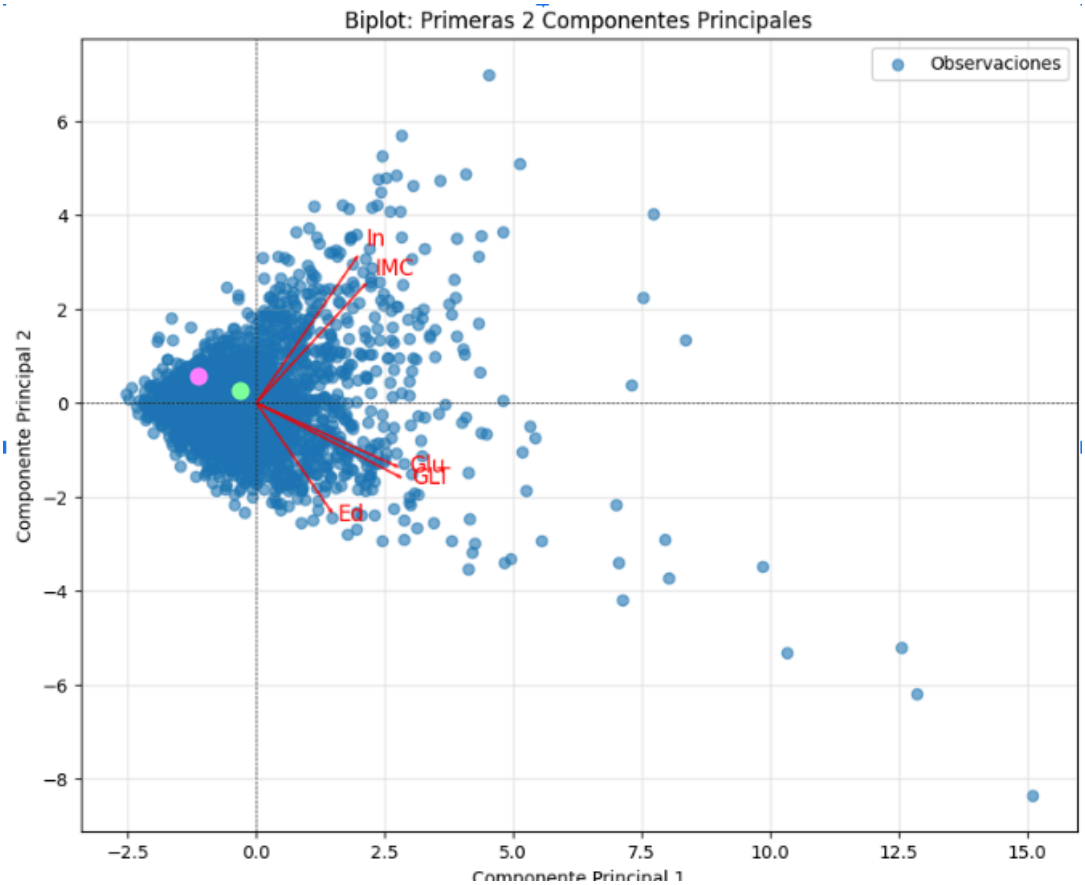
El biplot es informativo en este caso porque:

1. Las dos primeras componentes explican el 69% de la varianza, lo que asegura que el gráfico captura la mayoría de la estructura de los datos.
2. Permite visualizar patrones clave, como las correlaciones entre variables y las relaciones entre las observaciones.
3. Se observan posibles outliers o grupos de datos que se podrían analizar más profundamente.

En resumen, el biplot permite interpretar las relaciones entre las variables originales y observar cómo estas contribuyen a las principales dimensiones de variación en los datos. También ofrece una herramienta visual para explorar posibles patrones en las observaciones.

¿Cómo interpretar los datos 1225 y 893?

```
Observación 1225: PC1 = -1.17, PC2 = 0.63
Observación 893: PC1 = -0.05, PC2 = 0.33
```



El análisis del biplot nos permite interpretar cómo las observaciones 1225 y 893 se posicionan en el espacio definido por las dos primeras componentes principales (PC1 y PC2). Estas observaciones están marcadas en el gráfico: la observación 1225 en color violeta y la observación 893 en color verde.

Observación 1225

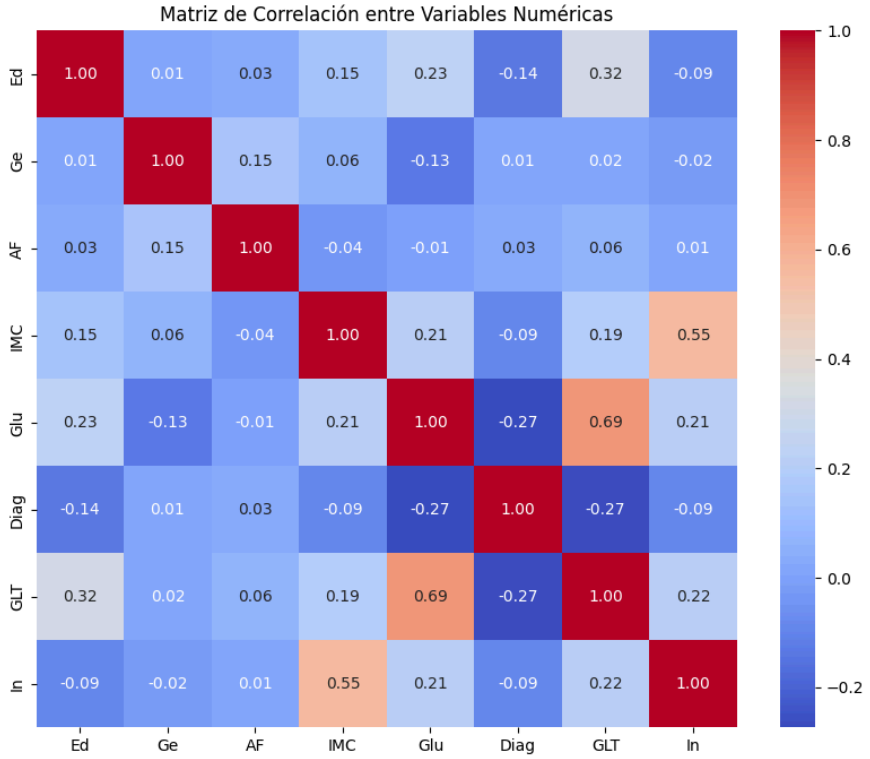
Las coordenadas de esta observación son PC1=−1.17 y PC2=0.63. Su posición en el biplot indica que está ligeramente alejada hacia el lado negativo de la primera componente principal (PC1), lo que sugiere que esta observación tiene valores bajos en variables que contribuyen positivamente a PC1, como GLT, Glu e IMC. Por otro lado, tiene un valor moderadamente positivo en PC2, lo que indica que posee valores algo altos en variables asociadas positivamente a esta componente, como In e IMC, y valores bajos en Ed, que está negativamente correlacionada con PC2. Esto sugiere que la observación 1225 podría representar un individuo más joven, con niveles bajos de glucosa y glucosa total, pero con niveles moderados de insulina.

Observación 893

Las coordenadas de esta observación son PC1=−0.05 y PC2=0.33. Esta posición, cercana al origen en PC1, indica que tiene valores promedio en las variables asociadas a este componente, como GLT y Glu. En PC2, su valor positivo más pequeño sugiere que tiene una contribución moderada de In e IMC, mientras que su valor de Ed podría ser ligeramente más bajo que el promedio. En general, esta observación parece representar un individuo con características metabólicas parecidas al promedio en las variables analizadas.

En conclusión, ambas observaciones están ubicadas en el cuadrante superior izquierdo del biplot, lo que indica que comparten algunas similitudes, especialmente en su relación con PC2. Sin embargo, la observación 1225 está más alejada hacia el lado negativo de PC1, indicando mayores diferencias en términos de glucosa y glucosa total, mientras que la observación 893 presenta un perfil cercano al promedio en ambas componentes.

Relaciones entre las variables en base al biplot VS correlaciones entre variables



No hay contradicción entre las relaciones observadas en el biplot y la matriz de correlaciones. En el biplot, las direcciones y ángulos de los vectores reflejan las correlaciones entre las variables: vectores cercanos entre sí, como GLT y Glu, indican una correlación positiva alta, lo cual concuerda con la matriz de correlaciones (r=0.69). Asimismo, Ed muestra una orientación opuesta a In y IMC, reflejando correlaciones negativas que también están presentes en la matriz (r=−0.32 y r=−0.14, respectivamente).

El biplot utiliza los dos primeros componentes principales para representar estas relaciones, y aunque puede haber cierta pérdida de información debido a la reducción de dimensiones, las tendencias generales coinciden con las correlaciones numéricas.

CLUSTERING

Indicios de Agrupamiento y medidas de tendencia

Observando los primeros dos componentes principales:

- La mayoría de las observaciones se concentran en torno al origen, lo que sugiere que tienen características similares.
- Sin embargo, hay algunas observaciones que se alejan significativamente del centro, especialmente hacia los extremos de la primera componente principal (eje de PC1). Estas podrían ser indicios de posibles grupos o valores atípicos.

Por otro lado, el estadístico de Hopkins es una medida utilizada para evaluar si existe evidencia de agrupamiento en un conjunto de datos. Este estadístico compara las distancias entre puntos reales y puntos generados aleatoriamente en el mismo espacio de las variables. Un valor cercano a 0.5 indica que los datos están distribuidos de manera aleatoria (sin agrupamientos), mientras que un valor cercano a 1 sugiere que los datos tienen una estructura clara de agrupamiento.

```
Estadístico de Hopkins: 1.0000
```

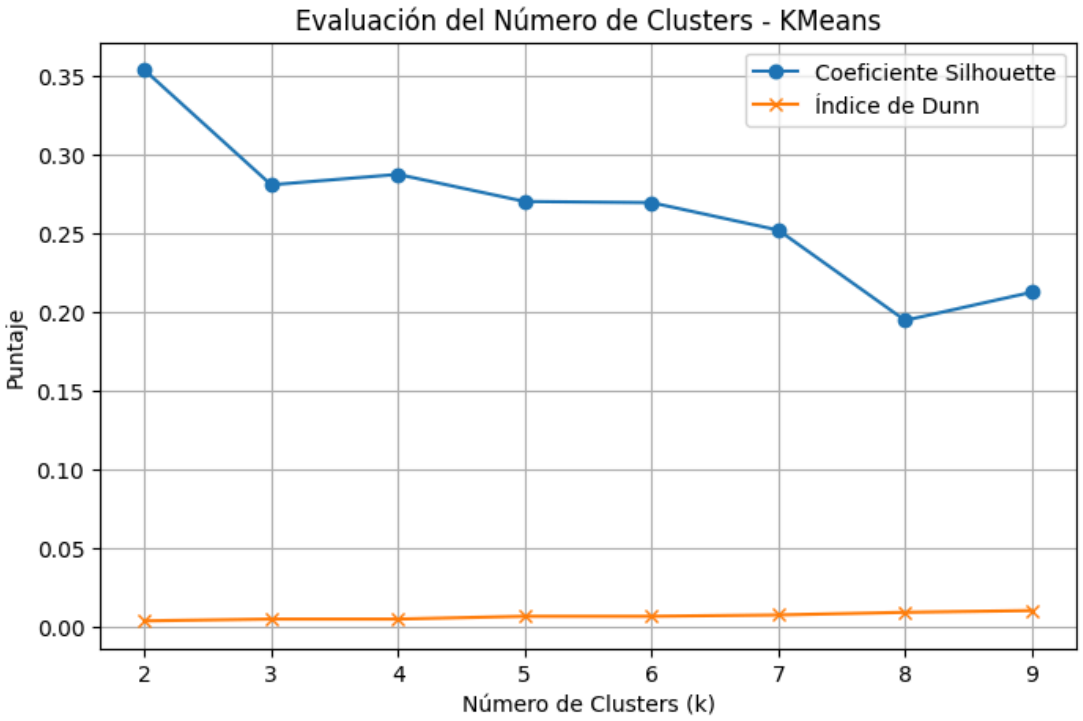
En este caso, el estadístico de Hopkins resultó ser 1, lo cual indica una fuerte evidencia de que las observaciones presentan agrupamientos naturales y no están distribuidas de manera uniforme en el espacio de las variables analizadas. Esto justifica la aplicación de métodos de clustering para identificar y analizar estos grupos.

También podríamos utilizar métodos de clustering jerárquicos para hacer un análisis estadístico preliminar y observar los dendrogramas. No procederemos a realizar este tipo de análisis porque la cantidad de observaciones es alta, y consideramos que ya contamos con evidencia suficiente en base a las conclusiones esbozadas por el biplot y el estadístico de Hopkins.

Elección de K: Silhouette y Dunn para KMeans y Ward

Se evaluó el número óptimo de clusters utilizando los índices de Silhouette e Índice de Dunn en el rango de k=2 a k=9 para el método de agrupamiento KMeans. Los resultados se presentan a continuación.

```
Método: KMeans
  - Clusters: 2 | Silhouette: 0.3537 | Índice de Dunn: 0.0041
  - Clusters: 3 | Silhouette: 0.2808 | Índice de Dunn: 0.0052
  - Clusters: 4 | Silhouette: 0.2874 | Índice de Dunn: 0.0052
  - Clusters: 5 | Silhouette: 0.2702 | Índice de Dunn: 0.0070
  - Clusters: 6 | Silhouette: 0.2695 | Índice de Dunn: 0.0070
  - Clusters: 7 | Silhouette: 0.2521 | Índice de Dunn: 0.0078
  - Clusters: 8 | Silhouette: 0.1948 | Índice de Dunn: 0.0095
  - Clusters: 9 | Silhouette: 0.2127 | Índice de Dunn: 0.0106
```



Coeficiente de Silhouette

El coeficiente de Silhouette mide la cohesión dentro de los clusters y la separación entre ellos. En este caso:

- El valor más alto se obtuvo con k=2 (Silhouette=0.3537), lo que indica que dividir los datos en dos clusters maximiza la separación entre grupos y minimiza la dispersión dentro de los mismos.
- A medida que k aumenta, el coeficiente de Silhouette disminuye progresivamente, lo que sugiere que un mayor número de clusters no mejora la calidad del agrupamiento.

Índice de Dunn

El índice de Dunn evalúa la relación entre la distancia mínima entre clusters y la máxima dispersión dentro de un cluster. En este caso:

- Los valores del índice de Dunn son bajos en general.
- El valor máximo se observa con k=9 (Dunn=0.0106), pero su incremento con respecto a valores más bajos de k es marginal, lo que dificulta justificar este número de clusters como el óptimo.

Conclusión para KMeans

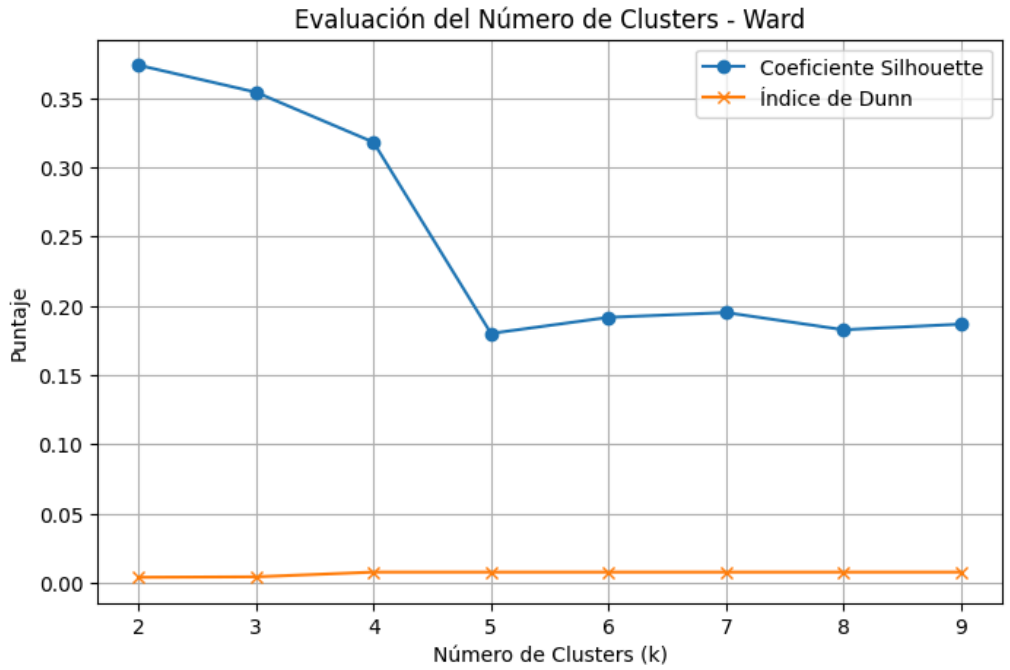
- Según el coeficiente de Silhouette, la mejor opción es k=2, ya que maximiza la separación entre clusters.

- Según el Índice de Dunn, no hay un valor claro que sobresalga significativamente, pero el incremento con k=9 es leve, por lo que k=2 o k=3 podrían ser elecciones razonables considerando ambos criterios.

Se concluye que el método KMeans es más estable con k=2, ya que este valor es respaldado claramente por el coeficiente de Silhouette.

Se evaluó el número óptimo de clusters utilizando los índices de Silhouette e Índice de Dunn en el rango de k=2 a k=9 para el método de agrupamiento Ward. Los resultados se presentan a continuación.

```
Método: Ward
  - Clusters: 2 | Silhouette: 0.3736 | Índice de Dunn: 0.0041
  - Clusters: 3 | Silhouette: 0.3541 | Índice de Dunn: 0.0044
  - Clusters: 4 | Silhouette: 0.3180 | Índice de Dunn: 0.0078
  - Clusters: 5 | Silhouette: 0.1800 | Índice de Dunn: 0.0078
  - Clusters: 6 | Silhouette: 0.1916 | Índice de Dunn: 0.0078
  - Clusters: 7 | Silhouette: 0.1951 | Índice de Dunn: 0.0078
  - Clusters: 8 | Silhouette: 0.1827 | Índice de Dunn: 0.0078
  - Clusters: 9 | Silhouette: 0.1867 | Índice de Dunn: 0.0078
```



Coeficiente de Silhouette

El coeficiente de Silhouette indica que:

- El valor más alto se obtiene con k=2 (Silhouette=0.3736), lo que sugiere que dividir los datos en dos clusters ofrece la mejor separación entre grupos y la menor dispersión dentro de ellos.
- Para k>2, el coeficiente de Silhouette disminuye progresivamente. Sin embargo, con k=3 (Silhouette=0.3541), el valor sigue siendo relativamente alto y podría considerarse como una opción alternativa razonable.

Índice de Dunn

El Índice de Dunn muestra los siguientes resultados:

- Aunque los valores del índice son bajos para todos los k, el valor más alto se observa con k=4 y k=9 (Dunn=0.0078).
- Sin embargo, la diferencia entre los índices es muy pequeña en todo el rango de k, lo que indica que el índice de Dunn no muestra una clara preferencia por un número de clusters específico.

Conclusión para Ward

- Según el coeficiente de Silhouette, la opción más razonable es k=2, ya que maximiza la separación entre los grupos.
- El Índice de Dunn no aporta una conclusión clara, pero podría respaldar k=4 o k=9 con valores marginalmente más altos.

En general, k=2 es la opción más sólida para el método Ward debido a la consistencia de los resultados del coeficiente de Silhouette.

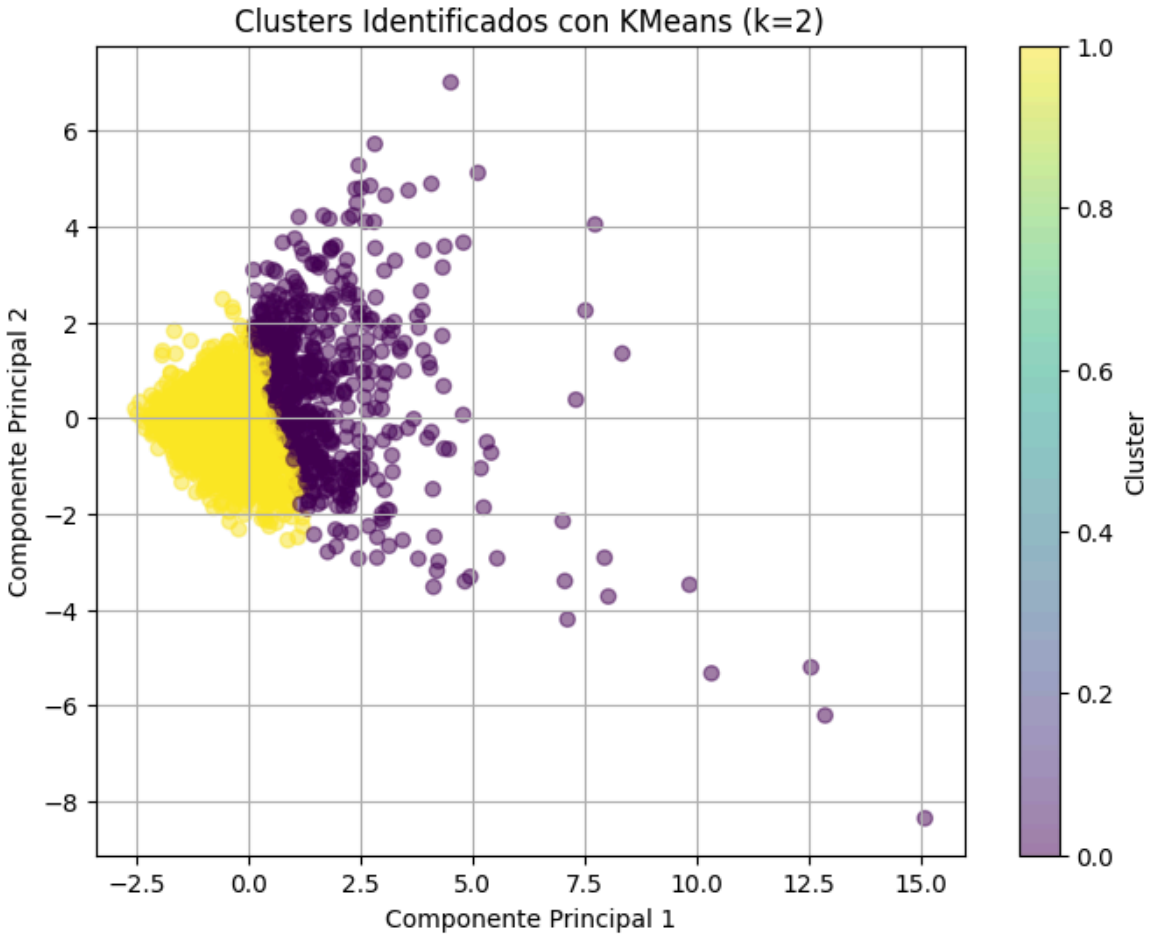
Teniendo en cuenta los resultados obtenidos con ambos métodos de agrupamiento:

- Tanto para KMeans como para Ward, el coeficiente de Silhouette sugiere que el número óptimo de clusters es k=2.
- El Índice de Dunn no aporta una preferencia clara, pero no contradice la elección de k=2.

Por lo tanto, se concluye que el número más razonable de clusters para aplicar KMeans es k=2, ya que está respaldado de manera consistente por el coeficiente de Silhouette y tiene sentido desde una perspectiva práctica.

Aplicación de KMeans

El gráfico a continuación presenta los resultados del modelo de KMeans aplicado con k=2, el número óptimo de clusters seleccionado previamente con base en los criterios de Silhouette y Dunn. Las observaciones fueron proyectadas en el espacio de las dos primeras componentes principales obtenidas mediante PCA, lo que permite una visualización clara de la separación entre los clusters.



En el gráfico:

- Cada punto representa una observación en el espacio reducido a dos dimensiones.
- Los colores indican la asignación de las observaciones a uno de los dos clusters identificados por el modelo.
- La barra de colores muestra las etiquetas de los clusters (0 y 1).

El modelo de KMeans con k=2 logró identificar dos clusters bien definidos en el conjunto de datos, como se observa en el gráfico. Aunque existe cierta superposición entre los grupos en la región central, la separación entre los clusters es clara, especialmente a lo largo de la primera componente principal. Esto sugiere que las diferencias principales entre las observaciones están asociadas a las variables que tienen mayor peso en este componente, como Glu, GLT e IMC.

Los dos clusters muestran características distintas en cuanto a su dispersión. El cluster representado en amarillo tiene una distribución más concentrada en torno a valores bajos de la primera componente principal, lo que indica una menor variabilidad en las características de las observaciones dentro de este grupo. Por otro lado, el cluster violeta presenta una mayor dispersión tanto en la primera como en la segunda componente principal, reflejando una mayor heterogeneidad en las características de sus observaciones.

En general, la segmentación en dos clusters parece adecuada y consistente con los patrones visuales. Este resultado proporciona una base para interpretar las diferencias entre los grupos en términos de las variables originales.

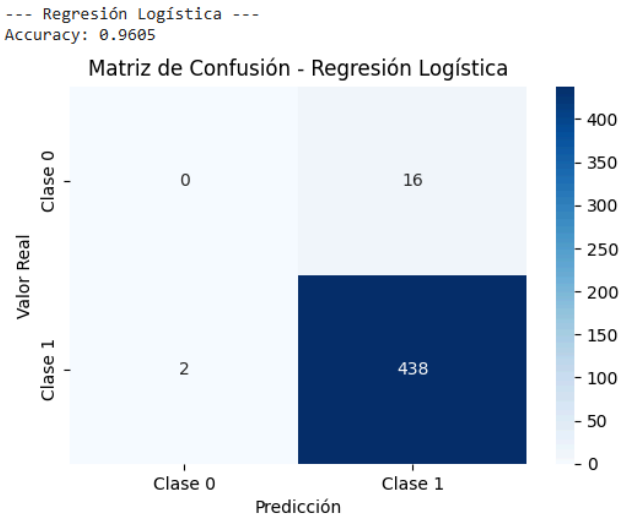
CLASIFICACIÓN

La clasificación es una técnica supervisada de aprendizaje automático utilizada para asignar etiquetas o categorías a nuevas observaciones, basándose en un conjunto de datos previamente conocido. Este enfoque tiene como objetivo aprender una función que relacione las características de las observaciones (variables predictoras) con su etiqueta o categoría (variable objetivo), de manera que se pueda predecir correctamente la clase de nuevas observaciones.

En este caso, se busca predecir la presencia o ausencia de diabetes utilizando como variables predictoras indicadores clave relacionados con la salud y el metabolismo, como la edad, el índice de masa corporal (IMC), y niveles de glucosa e insulina. Para evaluar el desempeño de los modelos de clasificación, se dividió el conjunto de datos en un 80% para entrenamiento y un 20% para prueba, y se calculó el porcentaje de error en el conjunto de prueba como métrica principal de evaluación.

Los métodos utilizados incluyen enfoques clásicos como la Regresión Logística, modelos basados en distancias como K-Nearest Neighbors (KNN), análisis discriminante (LDA/QDA), y técnicas más avanzadas como Random Forest y Naive Bayes, lo que permite comparar el desempeño entre diferentes enfoques y seleccionar el modelo más adecuado para este problema.

MODELO LOGÍSTICO



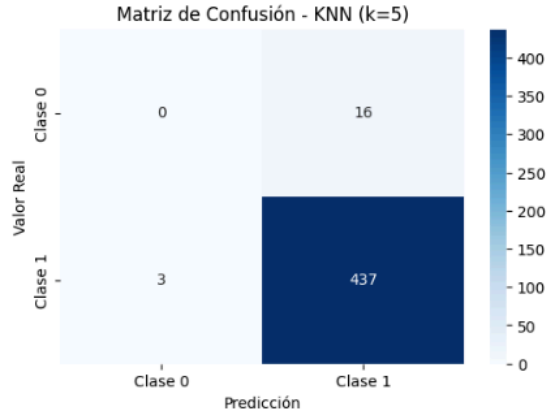
El modelo de Regresión Logística obtuvo un accuracy del 96.05%, lo que indica que predijo correctamente la mayoría de las observaciones en el conjunto de prueba. Sin embargo, al analizar la matriz de confusión, se pueden extraer las siguientes observaciones clave:

- Desempeño general:
    - De las 456 observaciones en el conjunto de prueba, el modelo clasificó correctamente 438 observaciones de la clase 1 (diabetes presente).
    - Hubo 16 falsos positivos, es decir, observaciones que fueron clasificadas como clase 1 cuando en realidad pertenecían a la clase 0 (diabetes ausente).
    - Se cometieron 2 falsos negativos, donde el modelo predijo clase 0 para observaciones que en realidad eran de clase 1.
- Errores en la clase 0:
    - No se logró clasificar correctamente ninguna observación de la clase 0 (diabetes ausente), lo que indica que el modelo tiene dificultades para distinguir las observaciones de esta clase.
    - Esto podría deberse a un posible desbalance en las clases, donde la clase 1 tiene muchas más observaciones que la clase 0, haciendo que el modelo esté sesgado hacia la clase mayoritaria.
- Porcentaje de error: Error Rate = errores totales/obs totales = 18/456 = 3.95%

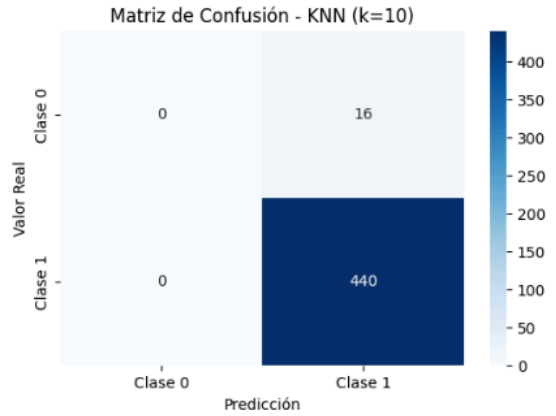
El modelo de Regresión Logística es efectivo para predecir la clase mayoritaria (diabetes presente), pero presenta dificultades significativas para identificar correctamente las observaciones de la clase minoritaria (diabetes ausente). Este resultado sugiere la necesidad de ajustar el modelo, por ejemplo, utilizando técnicas para manejar el desbalance de clases, como aplicar pesos balanceados o estrategias de submuestreo/sobremuestreo.

KNN

```
--- KNN (k=5) ---
Accuracy: 0.9583
```




```
--- KNN (k=10) ---
Accuracy: 0.9649
```



KNN con k=5

El modelo de KNN con k=5 alcanzó un accuracy del 95.83%, con los siguientes resultados destacados:

- Desempeño general:
  - Clasificó correctamente 437 observaciones de la clase 1 (diabetes presente).
  - Hubo 16 falsos positivos, donde observaciones de la clase 0 (diabetes ausente) fueron clasificadas como clase 1.
  - Se cometieron 3 falsos negativos, donde observaciones de la clase 1 fueron clasificadas incorrectamente como clase 0.
- Errores en la clase 0:
  - Similar a la Regresión Logística, el modelo no logró clasificar correctamente ninguna observación de la clase 0, lo que evidencia dificultades para identificar esta clase.
- Porcentaje de error: Error Rate = 19/456 = 4.17%

KNN con k=10

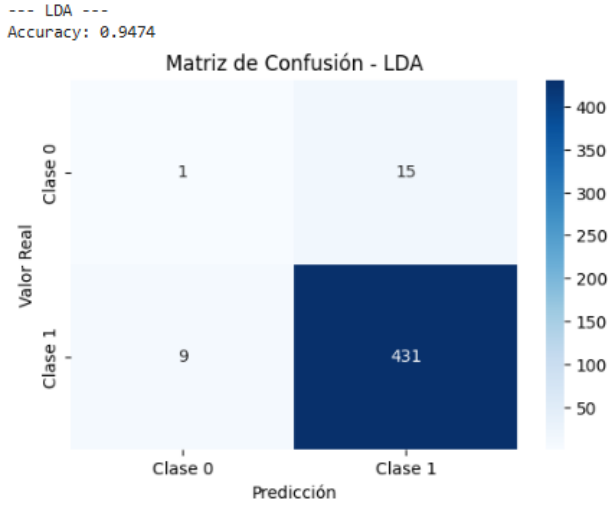
El modelo KNN con k=10 mejoró ligeramente, logrando un accuracy del 96.49%. Los resultados fueron:

- Desempeño general:
    - Clasificó correctamente 440 observaciones de la clase 1.
    - Hubo nuevamente 16 falsos positivos para la clase 0.
    - El número de falsos negativos se redujo a 0, mostrando que todas las observaciones de la clase 1 fueron clasificadas correctamente.
- Errores en la clase 0:
    - Al igual que con k=5, no se clasificaron correctamente observaciones de la clase 0.
- Porcentaje de error: Error Rate = 16/456 = 3.51%

Tanto k=5 como k=10 muestran un desempeño similar, con una leve ventaja para k=10, ya que redujo los falsos negativos y logró un menor porcentaje de error. Sin embargo, ambos modelos tienen dificultades importantes para identificar correctamente observaciones de la clase 0, posiblemente debido al desbalance en las clases.

Este comportamiento sugiere que, aunque KNN es efectivo para predecir la clase mayoritaria, su capacidad para clasificar la clase minoritaria es limitada, y podría beneficiarse de técnicas de balanceo o ajustes en los datos.

LDA



El modelo de Análisis Discriminante Lineal (LDA) obtuvo un accuracy del 94.74%, con los siguientes resultados destacados:

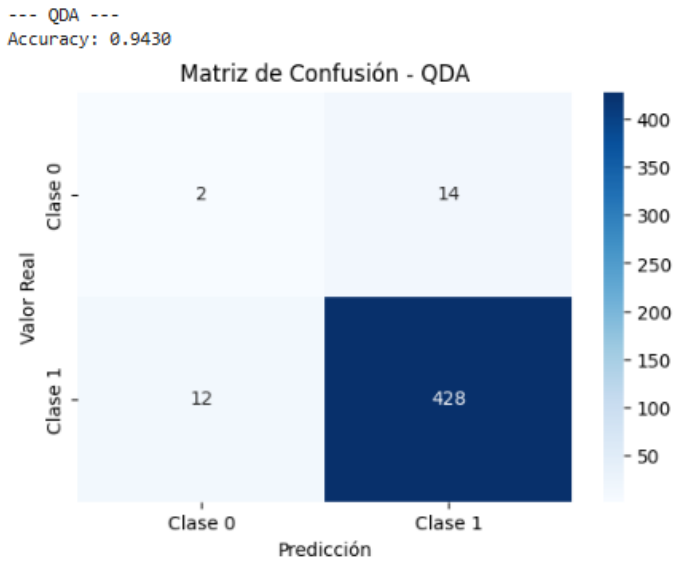
- Desempeño general:
    - Clasificó correctamente 431 observaciones de la clase 1 (diabetes presente).
    - Clasificó correctamente 1 observación de la clase 0 (diabetes ausente).

    - Se registraron 15 falsos positivos, donde observaciones de la clase 0 fueron clasificadas como clase 1.
    - Hubo 9 falsos negativos, donde observaciones de la clase 1 fueron clasificadas como clase 0.
- Errores en la clase 0:
    - A diferencia de los modelos previos, LDA logró identificar correctamente una observación de la clase 0, aunque el número de errores en esta clase sigue siendo alto.
- Porcentaje de error: Error Rate = 24/456 = 5.26%

LDA mostró un desempeño general sólido, clasificando correctamente la mayoría de las observaciones de la clase 1. Aunque logró identificar una observación de la clase 0, sigue teniendo dificultades significativas para clasificar correctamente esta clase, con 15 falsos positivos y 9 falsos negativos.

Comparado con otros modelos, LDA muestra un mayor porcentaje de error y menos precisión en la clasificación, especialmente en la clase 0, posiblemente debido al desbalance en las clases o a limitaciones del modelo en este conjunto de datos.

QDA



El modelo de Análisis Discriminante Cuadrático (QDA) alcanzó un accuracy del 94.30%, con los siguientes resultados destacados:

- Desempeño general:
    - Clasificó correctamente 428 observaciones de la clase 1 (diabetes presente).
    - Clasificó correctamente 2 observaciones de la clase 0 (diabetes ausente).
    - Hubo 14 falsos positivos, donde observaciones de la clase 0 fueron clasificadas como clase 1.
    - Se registraron 12 falsos negativos, donde observaciones de la clase 1 fueron clasificadas como clase 0.
- Errores en la clase 0:

    - Aunque QDA logró clasificar correctamente 2 observaciones de la clase 0, la mayoría de las observaciones de esta clase fueron clasificadas incorrectamente, lo que refleja una dificultad persistente para diferenciar esta clase.
- Porcentaje de error: Error Rate = 26/456 = 5.70%

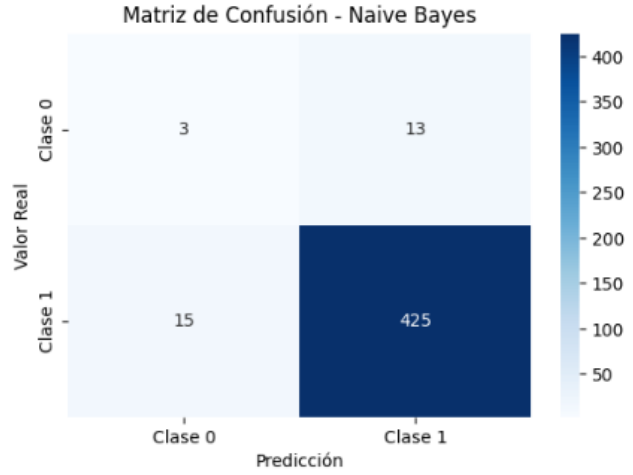
Conclusión General para QDA

- QDA mostró un desempeño adecuado para la clasificación de la clase 1, con un buen nivel de accuracy general.
- Aunque logró identificar más observaciones de la clase 0 que otros modelos como KNN y LDA, aún tiene problemas para diferenciar correctamente esta clase, lo que se traduce en un alto número de falsos positivos y falsos negativos.
- Comparado con LDA, QDA parece más adecuado para conjuntos de datos con una distribución no lineal, aunque su desempeño sigue limitado por el desbalance en las clases.

NAIVE BAYES

```
5. Naive Bayes:

--- Naive Bayes ---
Accuracy: 0.9386
```



El modelo de Naive Bayes alcanzó un accuracy del 93.86%, con los siguientes resultados destacados:

- Desempeño general:
    - Clasificó correctamente 425 observaciones de la clase 1 (diabetes presente).
    - Clasificó correctamente 3 observaciones de la clase 0 (diabetes ausente).
    - Hubo 13 falsos positivos, donde observaciones de la clase 0 fueron clasificadas como clase 1.
    - Se registraron 15 falsos negativos, donde observaciones de la clase 1 fueron clasificadas como clase 0.
- Errores en la clase 0:
    - Naive Bayes logró identificar correctamente 3 observaciones de la clase 0, lo que es una ligera mejora respecto a modelos como KNN y QDA. Sin

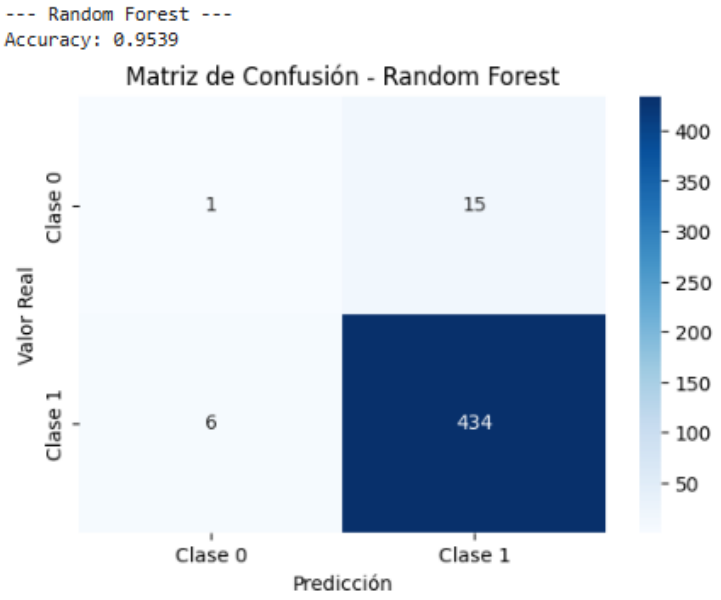
embargo, la mayoría de las observaciones de esta clase aún fueron clasificadas incorrectamente.

- Porcentaje de error: Error Rate = 28/456 = 6.14%

Naive Bayes mostró un desempeño sólido para la clasificación de la clase 1, aunque con un accuracy ligeramente menor en comparación con otros modelos como QDA o KNN. Su capacidad para identificar la clase 0 sigue siendo limitada, pero logró clasificar correctamente más observaciones de esta clase que algunos modelos previos.

Este modelo es rápido y eficiente, pero su supuesto de independencia entre las variables predictoras podría estar limitando su desempeño en este conjunto de datos.

RANDOM FOREST



El modelo de Random Forest alcanzó un accuracy del 95.39%, con los siguientes resultados:

- Desempeño general:
  - Clasificó correctamente 434 observaciones de la clase 1 (diabetes presente).
  - Clasificó correctamente 1 observación de la clase 0 (diabetes ausente).
  - Hubo 15 falsos positivos, donde observaciones de la clase 0 fueron clasificadas como clase 1.
  - Se cometieron 6 falsos negativos, donde observaciones de la clase 1 fueron clasificadas como clase 0.
- Errores en la clase 0:
  - Similar a otros modelos, Random Forest tiene dificultades para identificar correctamente la clase 0, logrando clasificar correctamente solo una observación de esta clase.
- Porcentaje de error: Error Rate = 21/456 = 4.6%

Random Forest mostró un excelente desempeño general, clasificando correctamente la mayoría de las observaciones de la clase 1 y logrando un bajo porcentaje de error. Aunque clasifica correctamente la clase mayoritaria (clase 1) con gran precisión, su capacidad para identificar correctamente la clase 0 sigue siendo limitada, con solo una observación correctamente clasificada.

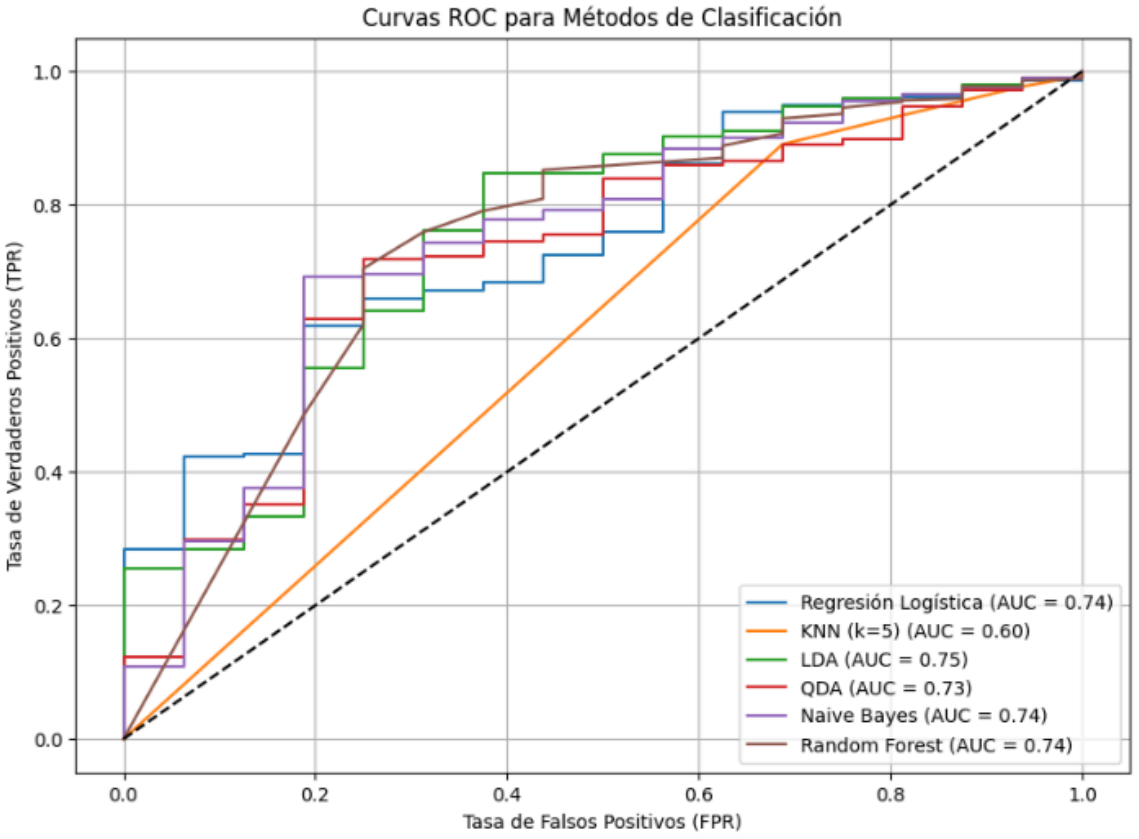
Este modelo es robusto y tiende a ser menos propenso al sobreajuste que otros métodos, pero aún podría beneficiarse de técnicas para manejar el desbalance en las clases. Además, la afinación de hiper parámetros podría mejorar su desempeño en la clase minoritaria.

CURVAS ROC

Explicación de las Curvas ROC y AUC

La curva ROC (Receiver Operating Characteristic) es una herramienta utilizada para evaluar el desempeño de un modelo de clasificación. Representa la relación entre la tasa de verdaderos positivos (sensibilidad) y la tasa de falsos positivos (1 - especificidad) a diferentes umbrales de decisión.

El área bajo la curva (AUC, Area Under the Curve) mide la capacidad del modelo para distinguir entre las clases. Un AUC de 1 indica un modelo perfecto, mientras que un AUC de 0.5 indica un modelo que no es mejor que una predicción aleatoria.



El modelo con mayor AUC es LDA, lo que indica que tiene la mejor capacidad para distinguir entre las clases (diabetes presente y diabetes ausente) en comparación con los demás modelos. Este resultado sugiere que LDA puede ser particularmente efectivo en este problema, incluso si el accuracy general no es el más alto, y es por eso que resulta relevante considerar métricas adicionales como el AUC para evaluar la calidad de los modelos.