UNIVERSIDAD NACIONAL DE COLOMBIA

PREGRADO EN ESTADISTICA

Departamento de Estadistica
Facultad de Ciencias

— Introduccion al Analisis Multivariado —

Tarea 6
Metodos de clasificacion

Integrantes:

Santiago Vera Quiceno C.C. 1.000.205.497

Medellin, Colombia

Agosto 30 de 2024

Índice

Índice de figuras

Índice de cuadros

1 Introduccion

1.1 Objetivo del estudio

en el presente analisis queremos comparar las dos grandes clases de de metodos de clasificación siendo estos los clasificadores jerarquicos tales como: ward o simple linkage y los clasificafores no jerarquicos tales como : K-means

estos metodos buscan el descubrimiento de patrones en los datos dando forma a grupos bien diferenciados los cuales contengan individuos homogeneos en su interior

1.2 Descripción de los datos

Este conjunto de datos *(Credit Card Fraud Detection Dataset 2023)* contiene transacciones con tarjetas de crédito realizadas por titulares de tarjetas europeos en el año 2023. Comprende más de 550,000 registros, y los datos han sido anonimizados para proteger la identidad de los titulares de las tarjetas. El objetivo principal de este conjunto de datos es facilitar el desarrollo de algoritmos y modelos de detección de fraude para identificar transacciones potencialmente fraudulentas

Tabla 1: Ficha Técnica

VARIABLE	SIGNIFICADO
Id	Identificador único para cada transacción.
V1-V28	Características anonimizadas que representan varios atributos de la transacción (por ejemplo, tiempo, ubicación, etc).
Amount	El monto de la transacción.
Class	Etiqueta binaria que indica si la transacción es fraudulenta (1) o no (0).

Debido a la variable *Amount* la cual resulta en el monto por el cual se realiza la transaccion debemos estandarizar las variables para un correcto analisis a posteriori

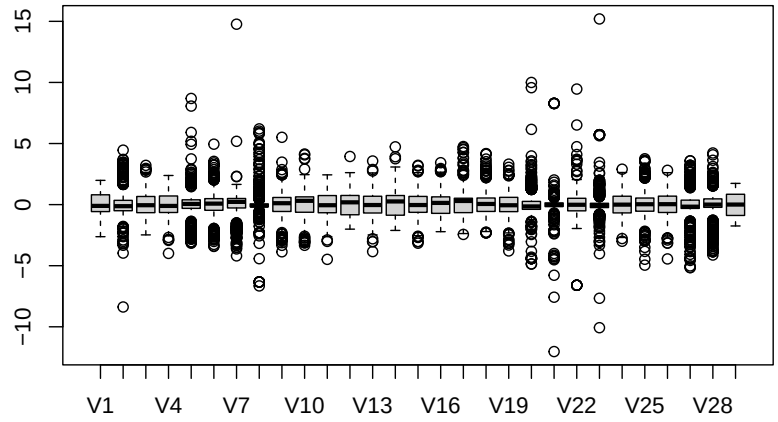

Figura 1: Boxplot datos escalados

1.3 Metricas de Clasificación

Dado que en nuestra base de datos a través de la variable: *Class* conocemos la clasificación de cada dato es fraude o no por lo tanto nos serviremos de las metricas de clasificación para comparar la capacidad predictiva los modelos de clasificación a evaluar.

Las métricas de clasificación son herramientas fundamentales en el análisis de modelos de clasificación, utilizadas para evaluar la precisión y efectividad de estos modelos en predecir las clases correctas de los datos. las metricas a evaluar serán:

- **Exactitud** $= \frac{VP+VN}{VP+FP+FN+VN}$ Mide el porcentaje de predicciones correctas sobre el total de predicciones.

- **precision** = $\frac{VP}{VP+FP}$ Indica el porcentaje de verdaderos positivos entre todas las instancias que el modelo predijo como positivas
- **sensibilidad** = $\frac{VP}{VP+FN}$ Mide el porcentaje de verdaderos positivos que fueron correctamente identificados por el modelo.
- **especificidad** = $\frac{VN}{VN+FP}$ Mide el porcentaje de verdaderos negativos que fueron correctamente identificados por el modelo

2 Clasificacion Jerarquica

Los metodos de clasificacion jerarquico son de dos tipos: aglomerativo y divisivo. Siendo los aglomerativos los mas usados este metodo costruye una serie de particiones anidadas empezando por los n individios uniendo los dos mas cercanos para tener una particion de n-1 clases, calculando las distancias entre el nuevo grupo y los demas individuos, seleccionando de nuevo los dos mas cercanos y continuar aglomerando hasta una particion de una clase con los n individuos. al conocer las categorias a clasificar se forzará a particionar en 2 grupos, siendo esto no lo recomendado

2.1 Ward.D

Para lograr grupos que tengan inercia minima interclases se debe tener una distancia euclidia y unir en cada paso del procedimiento los dos grupos que conlleven a un menor aumento de la inercia interclases siendo esto conocido como el metodo de Ward. mostraremos acontinuación el algoritmo para su construcción:

- Calcular la matriz de distancias de ward entre parejas de individuos de la siguiente manera: $W(i,l) = \frac{p_i p_l}{p_i + p_l} d^2(i,l)$, dado el caso que los pesos sean iguales a $1/n$ para los dos individuos la anterior expresión se reduce a: $W(i,l) = \frac{1}{2n} d^2(i,l)$
- seleccionamos la pareja de grupos(individuos en el primer paso) que presente la menor distancia de ward para conformar el nuevo grupo
- Calcular las distancias entr todos los grupos y el grupo recien conformado utilizando la formula de distacia de ward o la formula de recurrencia(Pardo,1992):

$$d(AB,C) = \frac{(P_a + P_c)W(A,C) + (P_b + P_c)W(B,C) - (P_a + P_b)W(A,B)}{P_a + P_b + P_c}$$

donde A, B y C son tres grupos presentes en el mismo paso de construccion del arbol, uniendo A y B para formar AB calculando la discacia de ward entre AB y C

- Eliminar las filas y columnas correspondietes a los individuos o grupos unidos y adicionar una fila y una columna para registrar la distancia entre el nuevo grupo y los demas repetir el proceso hasta llegar a una sola clase

2.1.1 Visualizacion de clusters

A continuacion podemos observar los resultados del algoritmo de clasificación *Ward.D*

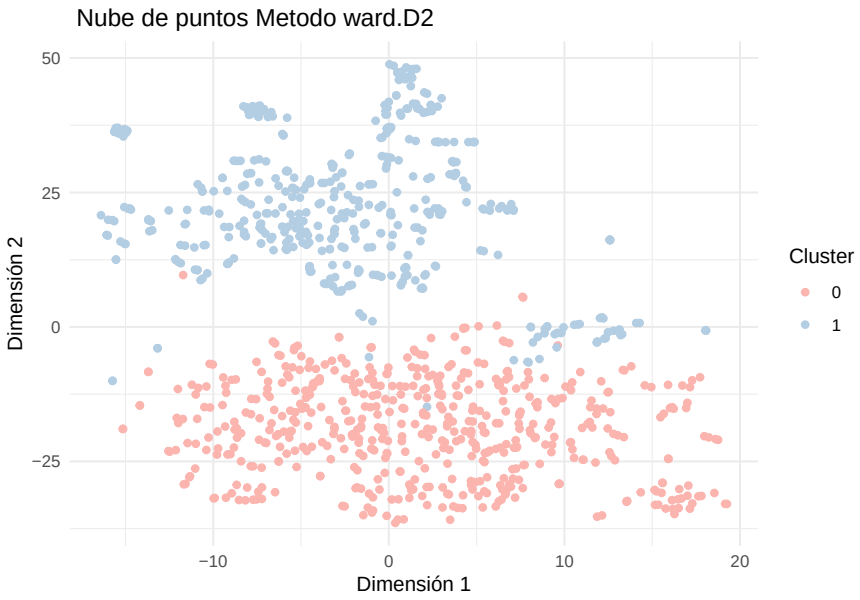

Figura 2: Nube de puntos Metodo ward.D2

2.1.2 MAtriz de confusion

Tabla 2: Matriz de Confusion metodo ward.D2

	Not Fraud	Fraud
Cluster NF	505	32
Cluster F	5	458

observamos en la matriz de confución que obtenemos: VP=505(Fraudes clasificados como Fraudes), VN=458(No fraudes clasificados como No fraudes), FP=5(No frayde clasificados como Fraude) y FN=32(Fraudes clasificados como No fraudes)

2.1.3 Evaluacion de metricas

- Exactitud

$$\frac{VP+VN}{VP+FP+FN+VN} = \frac{505+458}{1000} = 0.963$$

- Precisión

$\frac{VP}{VP+FP} = \frac{505}{505+5} = 0.9901$

- Sensibilidad

$\frac{VP}{VP+FN} = \frac{505}{505+32} = 0.940$

- Especificidad

$\frac{VN}{VN+FP} = \frac{458}{458+5} = 0.9892$

3 Clasificacion No-Jerarquica

Los métodos de clasificación no jerárquicos son técnicas utilizadas en el análisis de datos para agrupar un conjunto de elementos en clústeres o grupos, basándose en características similares. A diferencia de los métodos jerárquicos, estos métodos no crean una estructura de árbol o jerarquía, sino que agrupan los datos directamente en un número predefinido de clústeres.

Características Principales

- Predefinición de Clústeres: El número de clústeres debe ser especificado de antemano, lo que implica un conocimiento previo de los datos o una decisión basada en experimentación.
- Asignación Directa: Cada elemento es asignado directamente a un clúster sin seguir una estructura jerárquica.
- Iterativos: Estos métodos suelen ser iterativos, ajustando la asignación de los elementos en cada paso hasta alcanzar una configuración óptima.

3.1 K-means

Conocido en la literatura francesa como *Agregación alrededor de medias móviles*. El algoritmo busca una partición en K clases de un conjunto I de n individuos descritos por p variables continuas teniendo una nube de n puntos-individuos en R^p denotada por una distancia euclidea d.

se crean K centros iniciales induciendo a una particion de I en las K clases creadas de manera que el individuo i pertenece pertenecera a una clase con un centro mas cercano, repitiendo este proceso de manera iterativa hasta crear una partición no mejor a la inmediatamente anterior (la inercia interclases deja de disminuir) o hasta alcanzar el numero maximo de iteraciones previamente fijado generalmente la partición obtenida depende de la seleccion inicial de los centros

3.1.1 Eleccion de centros

Como se explicó al inicio del presente trabajo se sabe que la base de datos cuenta con dos categorias pero de igual manera se presentará el metodo de siluetas de forma que confirme el uso de dos centros para la clasificación de nuestra base de datos

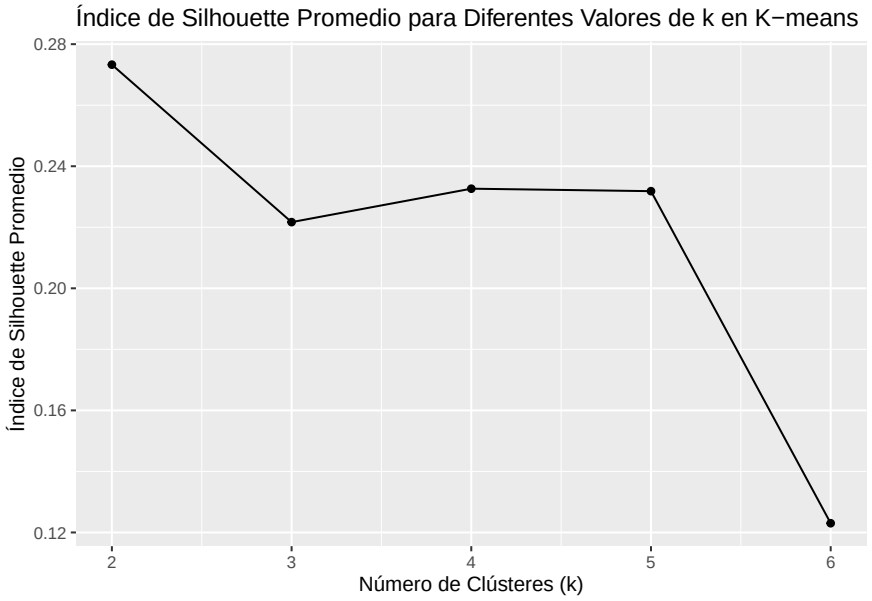

Figura 3: Índice de Silhouette Promedio para Diferentes Valores de k

El índice de Silhouette es una medida que evalúa la calidad de una partición en un análisis de clustering, como el de k-means. Este índice cuantifica qué tan bien se agrupan los puntos dentro de sus clústeres y qué tan separados están de otros clústeres.

La fórmula del índice de Silhouette para un punto i es

$$S(i) = \frac{b_i - a_i}{\mathbf{max}(a_i, b_i)}$$

- a_i es la distancia promedio del punto i a los otros puntos dentro de su propio clúster.
- b_i es la distancia promedio del punto i al clúster más cercano al que no pertenece.

El valor del índice de Silhouette varía entre -1 y 1:

- Un valor cercano a 1 indica que el punto está bien agrupado.
- Un valor cercano a 0 indica que el punto está en el borde entre dos clústeres.
- Un valor negativo indica que el punto está probablemente en el clúster incorrecto.

el valor mas alto obtenido es cuando usamos 2 clusters obteniendo: 0.2732839 el cual será usado para el analis de K-means

3.1.2 Visualizacion de clusters

A continuacion podemos observar los resultados del algoritmo de clasificación *K-meas*

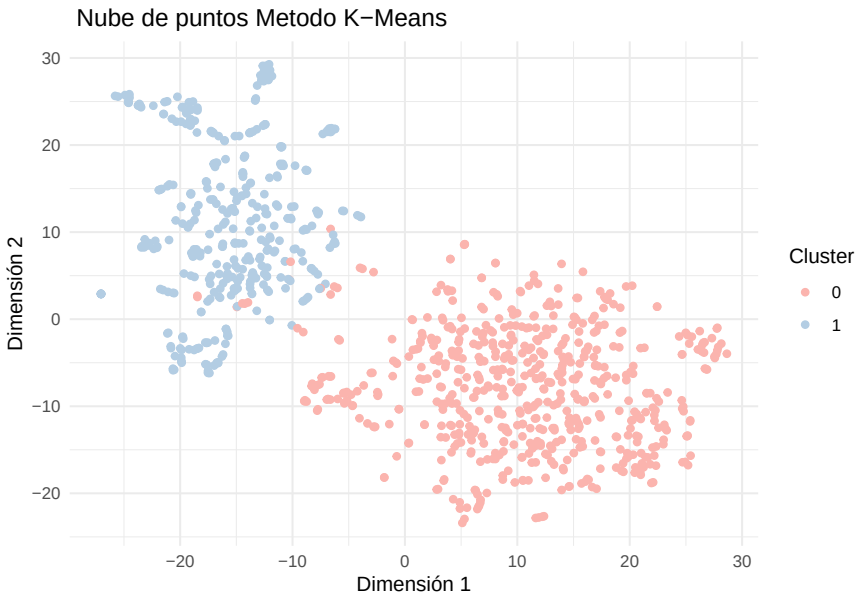

Figura 4: Nube de puntos Metodo K-Means

3.1.3 Matriz de confusion

Tabla 3: Matriz de Confusion metodo K-means

	Not Fraud	Fraud
Cluster NF	510	91
Cluster F	0	399

observamos en la matriz de confución que obtenemos: VP=510(Fraudes clasificados como Fraudes), VN=399(No fraudes clasificados como No fraudes), FP=0(No frayde clasificados como Fraude) y FN=91(Fraudes clasificados como No fraudes)

3.1.4 Evaluación de metricas

- Exactitud

$\frac{VP+VN}{VP+FP+FN+VN} = \frac{510+399}{1000} = 0.909$

- Precisión

$\frac{VP}{VP+FP} = \frac{510}{510} = 1$

- Sensibilidad

$\frac{VP}{VP+FN} = \frac{510}{510+91} = 0.8485$

- Especificidad

$\frac{VN}{VN+FP} = \frac{399}{399} = 1$

4 Comparación de metodos

Compararemos el rendimiento de cada tipo de modelo a traves de las metricas de medición antes descritas las cuales podemos comparar a traves de esta tabla:

$$\begin{pmatrix} & \text{Ward.D} & \text{K-means} \\ \text{Exactitud} & 0.963 & 0.909 \\ \text{Precisión} & 0.9901 & 1 \\ \text{Sensibilidad} & 0.940 & 0.8485 \\ \text{Especificidad} & 0.9892 & 1 \end{pmatrix}$$

vemos que la mayoria de las metricas son muy similares excepto en la sensibilidad en la cual el metodo Ward.D obtiene una amplia ventaja frente al metodo k-means y en el contexto en el que estamos será preferible usar el metodo Ward.D ya que el tener una clasificación tan grande de no fraudes como si fueran frudes puede resultar en un gran problema logistico dentro de una entidad financiera

5 Conclusiones

- obtuvimos un gran desempeño de cada uno de los metodos de clasificacion tanto jerarquico como no jerarquico lo cual muestra la potencia de cada uno de los algoritmos
- se muestra la importancia de usar algoritmos los cuales busquen la reduccion de la inercia interclases para un desarrollo oprtimo de las tecnicas de clasificacion
- se sugiere el uso del metodo Ward.D ya que su gran capacidad para clasificar correctamente los casos de Verdaderos positivos la hacen destacar frente a la
- Esta superioridad puede darse debido a la sensibilidad de los metodos no jerarquicos a la presencia de datos atipicos los cuales por la naturaleza da la base de datos la cual busca clasificar en fraude o no fraude son escenciales para la deteccion de estos ademas de la baja muestra usada para este estudio siendo solamente de 1000 observaciones cuando se sabe que los metodos nos jerarquicos trabajan mejor con muestras grandes

Referencias

Allaire, J., Xie, Y., Dervieux, C., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W. & Iannone, R. (2023). *rmarkdown: Dynamic Documents for R*.

Luque-Calvo, P.L. (2017). *Escribir un Trabajo Fin de Estudios con R Markdown*. Disponible en http://destio.us.es/calvo.

WID.world. (2024). URL https://wid.world/es/series/

Xie, Y. (2015). *Dynamic Documents with R and knitr*, 2nd edn. Chapman; Hall/CRC, Boca Raton, Florida.

Xie, Y. (2014). knitr: A Comprehensive Tool for Reproducible Research in R. *Implementing Reproducible Computational Research* (eds V. Stodden, F. Leisch & R.D. Peng). Chapman; Hall/CRC.

Xie, Y. (2023). *knitr: A General-Purpose Package for Dynamic Report Generation in R*.

Xie, Y., Allaire, J.J. & Grolemund, G. (2018). *R Markdown: The Definitive Guide*. Chapman; Hall/CRC, Boca Raton, Florida.

Xie, Y., Dervieux, C. & Riederer, E. (2020). *R Markdown Cookbook*. Chapman; Hall/CRC, Boca Raton, Florida.