Facultad de Ciencias

IMPLEMENTACIÓN Y EVALUACIÓN DE ALGORITMOS DE CLUSTERING DE DATOS MIXTOS

(IMPLEMENTATION AND EVALUATION OF MIXED DATA CLUSTERING ALGORITHMS)

Trabajo de Fin de Grado para acceder al

GRADO EN INGENIERÍA INFORMÁTICA

Autor: Fernando Sáenz Villaverde

Director: Camilo Palazuelos Calderón

Noviembre - 2023

1. Resumen

Este documento contiene información sobre el desarrollo de un código para la implementación y evaluación de algoritmos de clustering para un conjunto de datos mixto.
Normalmente, los algoritmos de clustering utilizan cierta métrica por la que considerar si dos datos cualesquiera son lo suficientemente similares como para agruparlos. Estas métricas usualmente son la distancia euclídea en el caso de datos de tipo numérico y el índice de Jaccard en el caso de datos categóricos. Sin embargo, en este caso se trabaja con un conjunto de datos mixtos, con datos tanto numéricos como categóricos, por lo que la métrica seleccionada es la distancia de Gower.
El objetivo principal del trabajo es implementar un algorimo de clustering de datos mixtos, evaluando su desempeño con algoritmos de clustering de datos numéricos o categóricos exclusivamente.
Para la realización de este proyecto se ha utilizado un conjunto de datos mixtos proveniente del proyecto ARCEUS (Universidad de Cantabria e Instituto de Investigación Marqués de Valdecilla).

Palabras clave: Distancia de Gower, clúster, regresión logística, predicción, curva ROC

2. Abstract

This document contains information about the development of a code for the implementation and evaluation of clustering algorithms for a mixed dataset.
Usually, clustering algorithms use a certain metric to determine if two data points are sufficently similar to be grouped together. These metrics are tipically Euclidean distance for numerical data and the Jaccard index for categorical data. However, a mixed dataset, that contains both numerical and categorical data, it is used in this project, so the selected metric is Gower's distance.
The main objective of thw project is the implementation of a clusteing algorithm for mixed data and evaluate its performance against clustering algorithms designed for exclusively numerical or categorical data.
To carry out this project, a mixed data from the ARCEUS project (Universidad de Cantabria and Instituto de investigación Marqués de Valcecilla) has been used.

Keywords: Gower Distance, clustering, logistic regression, prediction, ROC curve, R

Índice

3. Introducción

3.1. Motivación

El cáncer colorrectal es uno de los cánceres con mayor incidencia en todo el mundo. Según datos de la SEOM (Sociedad Española de Oncología Médica), en el año 2020, 1.931.590 pacientes en todo el mundo fueron diagnosticados con esta patología, constituyendo el $10,7\,\%$ de los cánceres más frecuentemente diagnosticados. Constituye también uno de los cánceres son mayor número de fallecimientos en el mundo, con una estimación del $9,5\,\%$.

En España es el cáncer más diagnosticado, con 34331 casos detectados en pacientes entre 50 y 69 años en el año 2017, constituyendo un $15\,\%$ del total, según datos del Observatorio del Cáncer de la AECC (Asociación Española contra el cáncer), siendo el segundo cáncer más diagnosticado tanto en hombres como en mujeres (tras el cáncer de próstata y el de mama, respectivamente).

En la mayoría de casos, el tratamiento para el cáncer colorrectal es la cirugía. El objetivo de dicha cirugía es la completa extirpación del segmento de colon o recto donde se halla el tumor. Hay diversos enfoques a la hora de realizar la cirugía, cuya elección depende de donde se encuentra el tumor, del estado del paciente en el momento de la operación y de la edad del mismo.

Al tratarse de una enfermedad que afecta a tanta gente, es interesante la clasificación de los pacientes, utilizando una métrica para calcular las disimilitudes entre pares de observaciones. Con ese agrupamiento, se puede tratar de realizar una predicción sobre si el paciente puede llegar a sufrir una complicación después de la cirugía, ya sea de índole infecciosa o no infecciosa.

3.2. Objetivo

Como se ha mencionado en el resumen previo, el objetivo principal de este proyecto es la implementación de un algoritmo de clustering de datos mixtos, evaluando y comparando su desempeño con el de algoritmos de clustering de datos numéricos o categóricos exclusivamente.
Este objetivo general se puede concretar en tres objetivos más específicos:

- Implementar el algoritmo k-means en su versión para datos mixtos, utilizando como métrica de disimilaridad la distancia de Gower, que permite cuantificar la distancia entre dos vectores de datos mixtos.
- Realizar el estudio teórico y empírico del coste temporal de dicho algoritmo, haciendo énfasis en la complejidad computacional del mismo.

- Evaluar el algoritmo en términos de interpretabilidad y explicabilidad de los resultados obtenidos con un conjunto de datos mixtos real y compararlo con algoritmos de clustering de datos numéricos o categóricos exclusivamente, como PAM (partitioning around medoids).

3.3. Herramientas

A continuación, se da una breve explicación de las tecnologías y herramientas utilizadas a lo largo del proyecto.

3.3.1. Lenguaje

El estudio se ha realizado utilizando el lenguaje de programación R, el cual se distribuye bajo la licencia GNU GPL (GNU General Public License). Esta es una licencia de derecho de autor, que garantiza a los usuarios la posibilidad de usar, compartir y modificar el software libremente.
La versión de R utilizada en el proyecto es la 4.3.1, lanzada el 16/06/2023. Se puede comprobar utilizando el comando "sessionInfo()" en la consola del entorno. Este comando proporciona también información sobre la plataforma y la versión de Windows utilizadas en el desarrollo, que en este caso se corresponden con un sistema Windows 11 de 64 bits. La decisión de utilizar dicho lenguaje de programación reside en características como su gran colección de herramientas para el análisis de datos, su amplia variedad de utilidades gráficas para la visualización de los datos o su efectividad a la hora del manejo de datos.

3.3.2. Entorno

El entorno utilizado es RStudio, un entorno construido exclusivamente para el uso de R, permitiendo la ejecución del código desde el propio editor del código fuente, el resaltado de sintaxis, autocompletado de código y sangría inteligente. Además, permite la depuración del código mediante un depurador interactivo, que permite corregir los errores de manera rápida. Estas características han sido importantes a la hora de seleccionar este entorno para el estudio.
La versión de RStudio utilizada es la 2023.06.1+524. Esto se puede comprobar dentro del propio entorno desde la barra de herramientas, en el botón "$Help''$ > "$AboutRStudio''$.

4. Conceptos básicos

En este apartado se van a explicar en profundidad algunos aspectos básicos muy presentes en este proyecto, y cuyo entendimiento es importante a la hora de comprender el mismo.

4.1. Conjunto de datos mixto

El conjunto de datos utilizado en este proyecto es mixto, es decir, no todas las variables son del mismo tipo.
Por un lado, se encuentran variables cualitativas, que son aquellas variables que indican una cualidad del objeto o sujeto a estudiar (en este caso, cada paciente tendrá su identificador), y que se pueden representar con palabras.
En el otro lado, se observan las variables cuantitativas, que conforman la inmensa mayoría del conjunto de datos, que son aquellas que solo se pueden representar mediante nuúmeros. Estas, a su vez, pueden dividirse en los siguientes grupos:

- Variables categóricas: son aquellas variables que contienen un número finito de categorías o grupos distintos. Si solo contienen dos grupos se conocen como variables binomiales o dicotómicas (por ejemplo, el sexo del paciente).
- Variables discretas: son aquellas variables numéricas que contienen un valor entre dos valores observables (por ejemplo, el número de pacientes operados en un determinado año).
- Variables continuas: son aquellas variables que contienen un número infinito de valores en un determinado rango (por ejemplo, la fecha en la que se ha realizado la cirugía).

4.2. Clustering

El siguiente concepto a explicar es el de clustering. El clustering es es una técnica o conjunto de técnicas que tienen como objetivo agrupar ítems distintos pero que comparten características entre sí en grupos denominados clústeres.
Esta agrupación se realiza en función de la similitud de los elementos, que se cálcula utilizando el concepto de distancia. Para ello, se considerará a los individuos objetos de estudio como vectores en el espacio de las variables, de manera que cuanto mayor sea el espacio entre ellos, mayor será su grado de disimilaridad.
Hay diversos métodos para calcular la distancia (distancia euclídea, distancia de Manhattan, distancia de Mahalanobis), sin embargo, como se ha mencionado en el subapartado anterior, se ha utilizado un conjunto de datos mixto, por lo que se utilizará una técnica distinta, la distancia de Gower.

4.3. Distancia de Gower

Seǵun se menciona en [Everitt et al., 2011], la distancia de Gower es una métrica utilizada para medir la similitud entre dos puntos de datos que contienen variables numéricas y categóricas.
Aplica distintas métricas de similitud en función del tipo de dato, usando la distancia Manhattan en el caso de las variables numéricas y ordinales, mientras que se usa la distancia de Jaccard en el caso de las variables categóricas.
Estas características hacen que sea la técnica adecuada para conjuntos de datos mixtos y, por tanto, para este proyecto.

4.4. Algoritmos para clustering de datos mixtos

El análisis de clústeres se utiliza por regla general en el contexto de conjuntos de datos categóricos o continuos, sin embargo, la dificultad aumenta sensiblemente cuando el conjunto de datos es mixto, tal y como se muestra en este proyecto.
Existen una gran variedad de formas de agrupar los datos, y la decisión sobre si escoger una u otra se puede tomar en base a tres cuestiones:

- La forma en la que se calculan las similitudes.
- La metodología para fusionar las partes numéricas con las partes categóricas.
- La elección de un algoritmo para la construcción de clústeres óptimos.

Los algoritmos encargados de realizar esta tarea se pueden dividir en dos grupos: métodos basados en la distancia y métodos basados en el modelo. Algunos de estos algoritmos son:

- K-Means: realiza el agrupamiento minimizando la suma de distancias entre cada objeto y el centroide de su clúster. Para ello, se selecciona el número de centroides, inicializando estos en coordenadas aleatorias. Una vez establecidos los centroides, se asocia cada punto con el centroide más cercano y se recalcula el centroide de cada clúster.
  Estos dos últimos pasos se repiten hasta que se llega a cierto criterio de parada, que puede venir dado por llegar a cierto límite de iteraciones o simplemente porque los centroides y puntos dejan de cambiar su posición.
- PAM: este algoritmo se basa en la búsqueda de k objetos representativos entre el conjunto de datos, donde se utilizan medoides para representar los grupos. Tras encontrar k objetos representativos, se construyen k clústeres, asignando cada objeto del conjunto de datos al objeto representativo más cercano.
  Acto seguido, se determina un nuevo medoide, que puede representar más adecuadamente al grupo, alterando así su ubicación. Esto se repite hasta que los medoides dejan de moverse.
- Kamila: es una adaptación basada en modelos del algoritmo k-means. Este algoritmo comienza con un conjunto de centroides para las variables continuas y un conjunto de parámetros para las variables categóricas.
  Para las variables continuas, se calcula la distancia euclidiana con el centroide más cercano, utilizando el conjunto de distancias mínimas para estimar la distribución de mezcla de las variables continuas. En el caso de las variables categóricas se calculan las probabilidades de observar los datos en función del clúster.
  Acto seguido, se utiliza el logaritmo de la verosimilitud de la suma de los componentes para encontrar el clúster apropiado para cada sujeto. Estos pasos se repiten hasta que los clústeres sean estables.
- K-prototypes: este algoritmo define G individuos virtuales como los centros de los clústeres, construidos a partir de las medias para las variables numéricas y las modas para las variables categóricas.

En la práctica, el algoritmo es muy similar a k-means: se eligen G prototipos iniciales y se asigna a cada sujeto al prototipo más cercano. Con todos los sujetos asignados, se actualizan los prototipos, repitiéndose esto hasta que la partición sea estable.

Estos son algunos de los algoritmos más utilizados para la agrupación de datos mixtos, en función de cual de ellos consigue optimizar mejor los clústeres, se seleccionará uno u otro.

4.5. Curva ROC

Según se menciona en [Pérez and Martín, 2022], una curva ROC es una herramienta estadística utilizada para evaluar la capacidad discriminativa de una variable.
Se ha establecido un porcentaje umbral, considerándose "positivos" aquellos pacientes con una probabilidad de complicación superior al umbral y "negativos" en caso contrario. Si se comparan los resultados con los datos del conjunto surgen cuatro alternativas:

- Pacientes que sufren una complicación y cuya predicción los considera "positivos". Reciben el nombre de "verdaderos positivos, VP".
- Pacientes que sufren una complicación y cuya predicción los considera "negativos". Reciben el nombre de "falsos negativos, FN".
- Pacientes que no sufren ninguna complicación y cuya predicción los considera "negativos". Reciben el nombre de "verdaderos negativos, VN".
- Pacientes que no sufren ninguna complicación y cuya predicción los considera "positivos". Reciben el nombre de "falsos positivos, FP".

La curva ROC se construye en base a la unión de distintos puntos de corte, correspondiendo el eje Y a la sensibilidad y el eje X a 1-especificidad de cada uno de ellos, incluyendo ambos ejes valores entre 0 y 1.
Es posible calcular la sensibilidad y la especificidad para el punto de corte que las origina mediante las siguientes fórmulas:
$Sensibilidad = \frac{VP}{VP+FN}$

$Especificidad = \frac{VN}{FP+VN}$

4.6. Área bajo la curva

El área bajo la curva representa es un parámetro para evaluar la precisión de una prueba diagnóstica que produce resultados continuos. Este área puede interpretarse como la probabilidad de que ante un par de individuos, uno con una complicación infecciosa y otro sin ella, el modelo los clasifique correctamente, según menciona en [de Ullibarri Galparsoro and Fernández, 2001].

4.7. Validación cruzada

La validación cruzada es un método o conjunto de métodos que permite probar el rendimiento de un modelo predictivo de Machine Learning. Estas técnicas, conocidas como "resampling", son estrategias que permiten estimar la capacidad predictiva de los modelos cuando se aplican a nuevas observaciones, según se menciona en [Rodrigo, 2020].
Todos estos métodos se basan en lo siguiente: el modelo es ajustado empleando un subconjunto de observaciones del conjunto de entrenamiento y se evalúa con las observaciones restantes. Este proceso es iterativo, agregando y promediando de esta forma los resultados. Algunos de los métodos para realizar validación cruzada son:

- División del conjunto. En este método, el conjunto de datos se divide en datos de entrenamiento y datos de prueba según cierto criterio.
- Leave One Out. Consiste en un método iterativo que se inicia empleando como conjunto de entrenamiento todas las observaciones disponibles, a excepción de una, que es la que se excluye para utilizarla como validación.
- K-fold: Consiste en tomar el conjunto de datos y dividirlo en dos conjuntos separados, datos de entrenamiento y datos de prueba.
  Acto seguido, este conjunto de entrenamiento se va a dividir en k subconjuntos, de manera que a la hora de realizar del entrenamiento se toma cada k como conjunto de prueba del modelo, tomándose el resto como conjunto de entrenamiento.

5. Diseño e implementación

En este apartado se van a desarrollar todas las funciones implementadas, así como una estimación de la complejidad temporal de cada una de ellas.

5.1. K-Means

La primera implementación a analizar es el desarrollo de un grupo de funciones para la clasificación en clústeres utilizando el algoritmo K-Means. Este desarrollo incluye varias funciones, que son llamadas desde una función principal. Para este desarrollo se ha utilizado como modelo la implementación planteada en la página web Domino, donde se explica cómo implementar esta funcionalidad en python paso a paso.

5.1.1. Iniciar centroides

La primera función a analizar se encarga de iniciar los k centroides aleatoriamente. Para ello se ha desarrollado la función "initiate_centroids(k, df)" que toma como parámetros el número de centroides y el conjunto de datos utilizado. Esta función toma k centroides aleatorios del conjunto de datos y los retorna como los centroides iniciales del algoritmo.

```
initiate_centroids(k, df) {
    n = numero de filas de df
    indices = seleccion de k indices aleatorios entre 1 y n filas
    initial_centroids = df[indices,]
    return initial_centroids
}
```

La generación de los índices tiene una complejidad temporal de $O(k)$, mientras que la creación de los centroides iniciales tiene una complejidad de $0(k * m)$, siendo m el número de columnas de df.
Por ello, la complejidad temporal de esta función es de $0(k) + 0(k * m)$.

5.1.2. Distancia de Gower

A continuación se va a proceder con el análisis de la función que calcula la distancia de Gower entre dos puntos. Esta función recorre todas las variables de ambos puntos, y realiza una operación u otra en función del tipo de variable. Si ambas variables son númericas se calcula la distancia euclídea entre ellas, mientras que si las variables son categóricas se suma 0 si son iguales y 1 (diferencia máxima) en caso contrario. El pseudocódigo planteado es el mostrado a continuación:

```
gower_distance(p1, p2) {
   n = length(p1)
   gower = 0
   Para cada i entre 1 y n {
      Si p1[i] y p2[i] son variable categoricas {
         Si p1[i] y p2[i] son iguales {
        gower += 1
         } Si no {
        gower += 0
         }
      } Si p1[i] y p2[i] son variables numericas {
          gower = gower  + Distancia euclidea(p1, p2)
      }
   }
   gower_distance = gower / n
   return gower_distance
}
```

En cuanto a la complejidad temporal, la función consiste en un bucle que itera sobre todas las variables de n, siendo n el número de variables de los puntos. Por ello, la complejidad temporal de la función es de $0(n)$.

5.1.3. Asignación de centroides

Para la asignación de centroides se ha creado la función "centroid_assignation(df, centroids)", que se encarga de asignar cada observación del conjunto de datos al centroide más cercano entre los proporcionados en la matriz "centroids", que se pasa como parámetro.
La función itera a través de las observaciones de df, calculando la distancia entre dicha observación y cada uno de los centroides llamando a la función "gower_distance()", almacenando las distancias en el vector "errors".
Con esto, encuentra el índice del centroide más cercano a la observación, asignando esta a dicho centroide.

```
centroid_assignation(df, centroids) {
   j = numero de filas de centroids
   n = numero de filas de df
   assignation = lista vacia
   assign_errors = lista vacia
   Para cada obs entre 1 y n {
      errors <- lista vacia
      Para cada centroide entre 1 y j {
         nearest_centroid = menor indice en el vector de distancias
         nearest_centroid_error = menor indice en el vector de errores
         assignation.append(nearest_centroid)
         assign_errors.append(nearest_centroid_error)
      }
```

```
    }
}
```

La complejidad temporal de la función depende del número de centroides y observaciones de los datos. Por lo general, será de $0(n * j)$, siendo n el número de observaciones y j el número de centroides.

5.1.4. Actualización de los centroides

Una vez hecha la asignación de las observaciones a cada uno de los clústeres se han de recalcular los centroides. Para ello, se ha desarrollado la función "update_centroids(df, assignation, k)", que recibe como parámetros el conjunto de datos, el vector de asignaciones de las observaciones a los clústeres y el número de clústeres.
Esta función crea una matriz inicializada con ceros e itera a través de los clústeres, calculando los nuevos centroides como el promedio de las observaciones en cada cluster, almacenándolos en la matriz creada.

```
update_centroids(df, assignation, k) {
   new_centroids = matriz(k, nº columnas(df)) inicializada a 0
   Para cada i entre 1 y k {
      cluster_points = df[assignation == i, ]
      Si el numero de filas de cluster_points > 0 {
         new_centroids[i,] = media de la columna(cluster_points)
      }
   }
   return new_centroids
}
```

La complejidad temporal de esta función es $0(n * k * m)$, donde n es el número de observaciones, k es el número de clústeres y m es el número de variables de los datos.

5.1.5. K-Means

Esta es la función principal del algoritmo, que utiliza todas las mencionadas anteriormente para la agrupación en clústeres. Esta función toma cuatro argumentos: el conjunto de datos "df", el número de clústeres "k", la tolerancia para converger "tol" y el número máximo de iteraciones "max_iter".
A continuación se inicializan las variables "continue", "j" y "err", además de los centroides llamando a la función "initiate_centroids()".
El siguiente paso consiste en un bucle while, que se ejecuta hasta que la variable booleana "continue" sea "FALSE", lo cual indica que el algoritmo ha convergido o que se ha llegado al número máximo de iteraciones. Dentro de este bucle se realiza la asignación de puntos al clúster más cercano mediante la función "centroid_assignation()", se registra el error de asignación en la lista "err", se verifica si se ha alcanzado los criterios de parada, ya sea en función de la tolerancia o del número máximo de iteraciones y se actualizan los centroides utilizando la función "update_centroids()".

Finalmente, se almacenan los resultados en una lista llamada "result", que contiene los siguientes elementos:

- "centroids": los centroides finales de cada clúster.
- "assignments": las asignaciones de puntos a cada clúster.
- "errores": los errores de asignación en cada iteración.

```
kmeans(df, k, tol=1e-6, max_iter=100) {
   continue = TRUE
   j = 1
   err = lista vacia
   centroids = initiate_centroids(k,df)
   Mientras continue sea TRUE {
      centrds = centroid_assignation(df, centroids)[[1]]
      assign_err = centroid_assignation(df, centroids)[[2]]
      err.append(assign_err)
      Si j > 1 {
        Si err[j-1] y err[j] son numeros {
            Si err[j-1] - err[j] <= tol o j >= max_iter {
                continue = FALSE
            }
        } Si no {
            continue = FALSE
        }
      }
      centroids = update_centroids(df, centrds, k)
      j += 1
   }
   result = lista(centroids, centrds, err)
   return result
}
```

La complejidad temporal de esta función depende de la complejidad de todas las funciones anteriores, explicadas en sus respectivos subapartados.

5.2. Análisis de los clústeres

En este apartado se va a realizar el estudio de la función planteada para el análisis de los clústeres.

```
stats_cluster(num_cluster) {
   Inicializacion de variables
   total_length = tamaño del cluster
   Para cada fila en df {
      Si la fila pertenece a num_cluster {
         age = age + edad del paciente
         Incrementar contador del resto de variables
```

```
        }
    }
    Calcular porcentajes para cada variable
    Imprimir edad promedio
    Crear graficos para los porcentajes calculados
}
```

Respecto a la complejidad temporal del código, lo primero que se encuentra es un bucle que engloba operaciones que tienen una complejidad lineal en función de 'n'. Por ello, el bucle principal tiene una complejidad temporal de $O(n)$.
Acto seguido, se consideran las operaciones para el cálculo de estadísticas y la representación mediante gráficos. En este caso, depende del número de estadísticas y clústeres, teniendo una complejidad constante de $O(1)$.
Por tanto, la complejidad de esta función será de $O(n)$, donde n es el número de objetos en el clúster.

5.3. Modelo de clasificación

A continuación, se muestra el pseudocódigo de la función "show_roc_curve(complicacion)", que recibe como parámetro el tipo de complicación que se quiere predecir.

```
show_roc_curve(complicacion) {
    length_train = 0
    Para cada fila en df$Year {
        Si el año es menor o igual a 2017 {
            length_train += length_train
        }
    }
    porcentaje_train = length_train / longitud de df$Year
    train_index = createDataPartition
    data_train = df[train_index,]
    data_test = df[-train_index,]
    Si complicacion == 0 {
        Entrenar modelo de regresion logistica para IAI
        Realizar prediccion en los datos de prueba
        Crear un histograma con las probabilidades de IAI
    } Si complicacion == 1 {
        Entrenar modelo de regresion logistica para NonIAI
        Realizar prediccion en los datos de prueba
        Crear un histograma con las probabilidades de NonIAI
    }

    Ajustar opciones para la presentacion de resultados
    Mostrar resumen del modelo
    Si pred >= 0.4 {
        pred_final = 1
    } sino {
```

```
        pred_final = 0
    }
    Calcular la curva ROC
    Graficar la curva ROC
    Imprimir el area bajo la curva
}
```

Respecto a la complejidad temporal de la función, lo primero que se realiza es un conteo de las operaciones menores o iguales a 2017, iterando a través de la columna Year. Esto implica una complejidad de $O(n)$ donde n es el número de filas.
La función "createDataPartition" es lineal, con una complejidad $O(n)$, así como la separación de datos en conjunto de datos de prueba y de entrenamiento.
En el caso del entrenamiento del modelo, la complejidad temporal es de $O(f)$, donde f es la cantidad de características en los datos. Tanto las predicciones, los histogramas, la clasificación de las predicciones finales como el cálculo de la curva ROC tienen una complejidad lineal $O(n)$.
Con todo esto se puede estimar que la complejidad de la función es $O(n)$.

5.4. Validación cruzada: Leave One Out

A continuación se procede con la explicación de la función creada para el cálculo de la precisión mediante el método de validación cruzada Leave One Out.
El pseudocódigo propuesto es el siguiente:

```
loocv() {
    precision = lista vacia
    Para todo i entre 1 y length(df$Age) {
        train = df excluyendo la fila i
        test = df con solo la fila i
        model = Entrenar el modelo de regresion logistica para IAI
        pred = Realizar prediccion en el modelo de prueba
        Si pred >= 0.4 {
            pred_final = 1
        } sino {
            pred_final = 0
        }
        respuestas = Obtener respuestas verdaderas del conjunto de prueba
        error = Calcular el error de clasificacion
        precision[i] = 1 - error
    }
    media = media de precision sin valores nulos
    return media
}
```

De nuevo, el código contiene un bucle principal, que recorre todas las filas del conjunto de datos, el cual tiene una complejidad temporal de $O(n)$, siendo n el número de filas en el conjunto de datos.
Dentro de este bucle, se crea el conjunto de entrenamiento y el conjunto de datos

mediante la exclusión de una fila, lo cual se puede considerar que tiene una complejidad $O(1)$. En cuanto al entrenamiento del modelo y las predicciones, al igual que en el caso anterior tendrán una complejidad de $O(f)$, donde f es el número de características de los datos.
Por otra parte, el cálculo de la precisión media tiene una complejidad de $O(n)$, ya que se suma un número lineal de valores y se divide por n.
Con estos datos, se puede determinar que la complejidad temporal de la función es $O(n)$.

5.5. Modelo de clasificación para los clústeres

En este caso se va a analizar la función para establecer el modelo de regresión logística, pero en vez de usando el conjunto de datos al completo, se aplica sobre los datos de un clúster, cuyo índice se pasa por parámetro.

```
roc_curve_cluster(num_cluster) {
   indices = indices de las filas de los pacientes del cluster
   cluster_data = subconjunto que contiene los pacientes del cluster
   length_train = 0
   Para cada fila i entre 1 y cluster_data$Year {
      Si cluster_data[i,]$Year <= 2017 {
     length_train++
      }
   }
   porcentaje_train = length_train/longitud de cluster_data$Year
   train_index = crear un indice basado en cluster_data$Year
   data_train = cluster_data[train_index,]
   data_test = cluster_data[-train_index,]
   modelo = Entrenar modelo de regresion logistica para IAI
   pred = Realizar prediccion en los datos de prueba
   Crear un histograma con las probabilidades de IAI
   Si pred >= 0.4 {
      pred_final = 1
   } sino {
      pred_final = 0
   }
   Calcular la curva ROC
   Graficar la curva ROC
   Imprimir el area bajo la curva
}
```

El primer paso que realiza esta función es la obtención de los índices del clúster, el cual tiene un coste lineal, por lo que tiene una complejidad de $O(m)$, siendo m la cantidad de elementos en el clúster.
El cálculo de la longitud del conjunto de entrenamiento es el siguiente paso, el cual tiene una complejidad lineal $O(m)$. El entrenamiento del modelo tiene una complejidad temporal de $O(f)$, donde f es el número de características en los datos.

Finalmente, el cálculo de la curva ROC tiene complejidad $O(m)$, por lo que la complejidad temporal de la función al completo $O(m + f)$.

6. Experimentación

En esta sección se van a explicar las decisiones tomadas respecto a la experimentación y métodos utilizados, así como el conjunto de datos y librerías utilizado en el proyecto.

6.1. Conjunto de datos

En esta sección se explicará brevemente el conjunto de datos utilizado, así como todas sus variables.
El conjunto de datos utilizado consta de 1047 filas y 18 columnas, lo que da lugar a 1046 pacientes distintos (ya que la primera fila se usa para especificar el nombre de cada uno de los campos). Es interesante destacar que se trata de un conjunto de datos mixto, que tiene variables numéricas discretas, continuas y categóricas, así como variables de texto.

Las columnas definidas en el conjunto de datos son:

- ID: variable de texto que actúa como identificador del paciente.
- Date: variable de texto que indica la fecha de la operación.
- Age: variable numérica que indica la edad del paciente.
- ASA: variable numérica categórica que indica el estado físico del paciente según la clasificación ASA de [WD et al., 1978]. Puede tomar valores entre 1 y 5.

- CCI: variable numérica que indica el estado del paciente según el índice de Comorbilidad de Charlson. Este sistema permite la evaluación de la esperanza de vida a los diez años, en función de la edad del paciente y de las comorbilidades de este, según se explica en [ME et al., 1987] y [ME et al., 2008]. Puede tomar valores entre 1 y 15, siendo este el máximo valor que un paciente de nuestro conjunto de datos presenta.

- Sex: variable numérica categórica binaria que india el sexo del paciente. Toma valor 1 si el paciente es mujer y valor 0 si el paciente es hombre.
- Anastomosis: variable numérica categórica que indica el tipo de anastomosis realizado en la operación. Según se menciona en [Slieker et al., 2013], la anastomosis es la conexión entre dos vasos sanguíneos, de forma espontánea o como resultado de una intervención quirúrgica.
  Se encuentran tres tipos distintos:

    - 0: Anastomosis enterocólica. Conecta segmentos del intestino delgado y el colon o el intestino grueso.
    - 1: Anastomosis colorrectal. Conecta el colon y el recto.
    - 2: Anastomosis colorrectal con ileostomía protectora. Conecta el colon y el recto con una abertura en el abdomen (ileostomía) que permite la eliminación de desechos.
- Approach: variable numérica continua que indica el enfoque quirúrgico utilizado. Hay tres tipos de enfoques, que toman los siguientes valores:
    - 0: Abierto. Se realiza una incisión en la piel y tejidos circundantes para acceder al área quirúrgica y tener una visualización completa de las estructuras y órganos involucrados.
    - 1: Laparoscopia. Se realiza a través de orificos en la cavidad abdominal. Una mínima incisión en un pliegue del ombligo, que permite la introducción del endoscopio con una micro-cámara, según [Álvarez Fernández-Represa et al., 2000]
    - 2: Asistida robóticamente. Utiliza pequeñas herramientas que van fijadas a un brazo robótico dirigido por el cirujano, como se menciona en [C and M, 2012]
- CRP1: variable numérica continua que indica la cantidad de proteína C-Reactiva en sangre el primer día después de la operación. Esta proteína mide los niveles de inflamación en el cuerpo.
- CRP3: variable numérica continua que indica la cantidad de proteína C-Reactiva en sangre el tercer día después de la operación.
- CRP5: variable numérica continua que indica la cantidad de proteína C-Reactiva en sangre el quinto día después de la operación.
- PCT1: variable numérica continua que indica el nivel de procalcitonina en sangre el primer día después de la operación. Un nivel alto de este biomarcador puede indicar la aparición de una infección bacteriana.
- PCT3: variable numérica continua que indica el nivel de procalcitonina en sangre el tercer día después de la operación.
- PCT5: variable numérica continua que indica el nivel de procalcitonina en sangre el quinto día después de la operación.
- IAI: variable numérica categórica binaria que indica si el paciente ha sufrido una complicación por infección intraabdominal. Toma valor 0 si no hay complicación y 1 en caso contrario.
- GradeIAI: variable numérica categórica que indica el grado de infección en caso de haber complicación infecciosa. Puede tomar valores entre 0 y 6.

- NonIAI: variable numérica categórica binaria que indica si el paciente ha sufrido una complicación de índole no infecciosa. Toma valor 0 si no hay complicación y 1 en caso contario.
- GradeNonIAI: variable numérica categórica que indica el grado de la complicación de índole no infecciosa, en caso de haberla. Puede tomar valores entre 0 y 6.

6.2. Librerías

En este apartado se mencionan todas las librerías utilizadas y una breve descripción de su funcionalidad.

- **Cluster:** incluye métodos para el ánalisis de los clústeres.
- **Readxl:** lee el conjunto de datos de un archivo Excel.
- **Factoextra:** incluye funciones para la visualización y análisis de datos. Depende de la librería ggplot2.
- **Lubridate:** contiene funciones para trabajar con fechas y espacios de tiempo.
- **Caret:** funciones para la creación de modelos predictivos. Depende de la librería lattice.
- **PROC:** proporciona herramientas para la visualización y análisis de curvas ROC.
- **Boot:** incluye conjuntos de datos y funciones para muestreo.

Como se ha mencionado, algunas de estas librerías tienen dependencias en otras. Estas son:

- **Ggplot2:** proporciona funciones para la creación y visualización de gráficas. De ella depende la librería factoextra.
- **Lattice:** proporciona funciones para la visualización de datos. De ella depende la librería caret.

6.3. Experimentos

Como se ha mencionado, el primer paso realizado en el proyecto ha sido la agrupación de los pacientes del conjunto de datos en clústeres, mediante el utilización del algoritmo k-means. Para ello, primero se ha calculado el número óptimo de clústeres en los que dividir los datos. Dichos clústeres han sido analizados, detectando cuáles son las variables clave a la hora de realizar la agrupación.
Dicho algoritmo utiliza la distancia de Gower entre dos puntos como medida de disimilaridad, que como se ha mencionado anteriormente, se trata de un conjunto de datos mixto, con variables de distintos tipos.
Acto seguido se ha propuesto el modelo de regresión logística para la predicción de variables dicotómicas, que en este caso son la aparición de complicaciones, ya sean

de índole infecciosa o no. Se ha representado este modelo mediante su representación con una curva ROC y se ha validado utilizando Leave One Out.
Por último, se ha propuesto de nuevo el modelo de regresión logística, esta vez para cada uno de los clústeres en lugar de para todo el conjunto de datos, representándolo de nuevo mediante su curva ROC.

7. Resultados

En este apartado se van a comentar los resultados obtenidos por los experimentos realizados.

7.1. Cálculo del número de clústeres

El cálculo del número óptimo de clústeres se puede realizar representando el número de clústeres en función del total dentro de la suma del cuadrado. Para ello se utilizará la función "fviz_nbclust(data, función, método)". En este caso, el parámetro data será el data frame de la matriz de distancias, la función será kmeans y el método de representación wss.

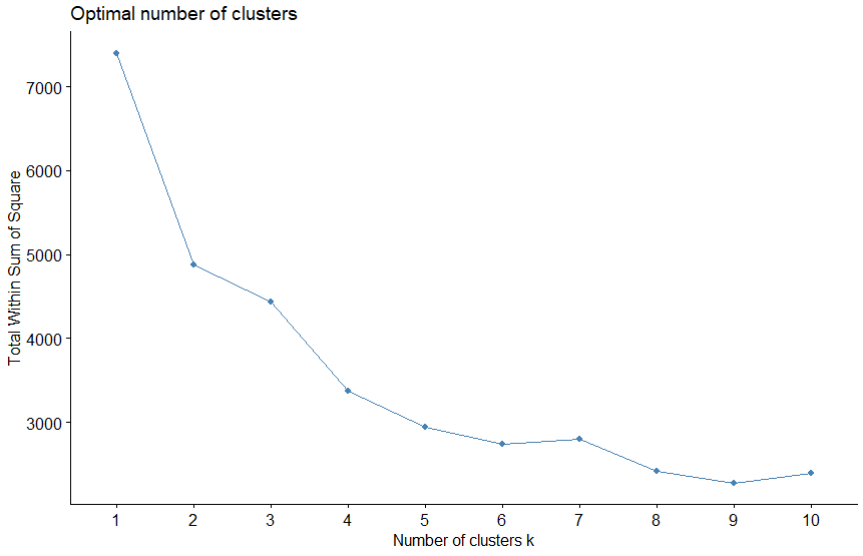


Figura 1: Número de clústeres

Como se puede observar en la figura 1, entre 1 y 4 clústeres el valor de la suma disminuye notablemente. Sin embargo, a partir de 4 clústeres la diferencia entre las sumas es mínima, por lo que el número óptimo de clústeres para este desarrollo será 4.

7.2. Comparación entre clústeres

En este apartado se va a realizar el estudio de los clústeres para cada una de las variables. Cabe destacar que el ID no se utiliza y que la columna Date se ha sustituido

por una columna Year, que solo contiene el año en el que se realizó la cirugía. para una mayor simplicidad.

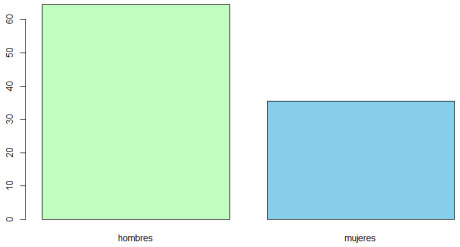
7.2.1. Edad

El primer parámetro a analizar será la edad media de los pacientes de cada clúster. Como se puede observar en la figura a continuación, la edad media en cada clúster es muy similar, siendo la máxima 69,62 en el clúster 1 y la mínima 67,43 en el clúster 2. La diferencia entre estos valores es mínima, por lo que se puede deducir que no es un factor relevante a la hora de agrupar los datos.

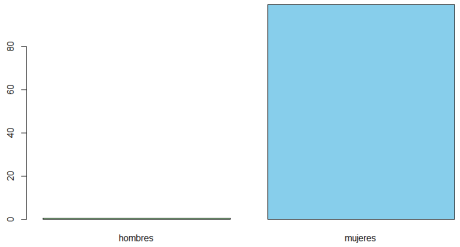
Clúster	Edad
1	69.62
2	67.43
3	67.93
4	68.63

7.2.2. Sexo

En este apartado se analizará la distribución por sexo de cada clúster. Tal y como se observa en la figura 2, hay una gran heterogeneidad en cuanto a los porcentajes de hombres y mujeres en cada clúster.
Se puede observar que los clústeres 1 y 4 son muy similares, con más de un 60 % de hombres en ambos clústeres. Por otro lado, los otros dos clústeres representan los dos extremos, teniendo el clúster 2 prácticamente el 100 % de mujeres, ocurriendo el caso totalmente contrario en el clúster número 3.

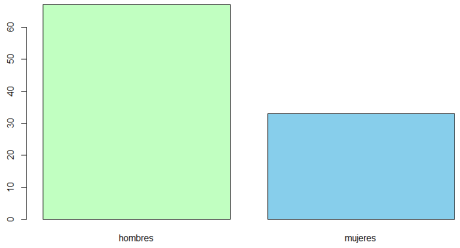


(a) Clúster 1



(b) Clúster 2

(c) Clúster 3

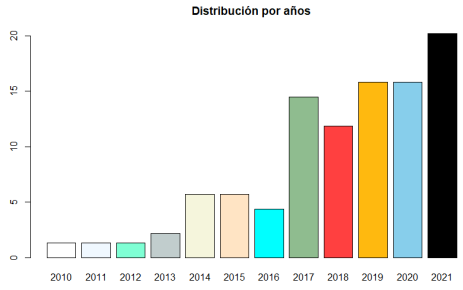


(d) Clúster 4

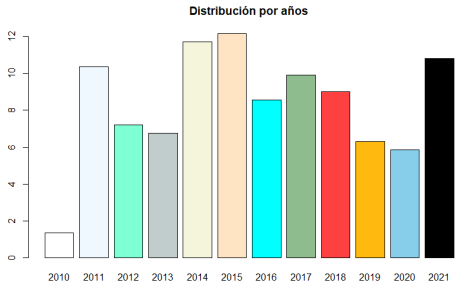
Figura 2: Comparación de sexo entre clústeres

7.2.3. Año

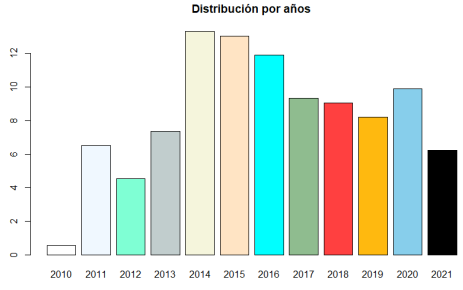
Otro factor que puede resultar interesante analizar es el año en el que se realizó la cirugía. Se puede observar un alto grado de heterogeneidad entre los clústeres, siendo el clúster 2 y el clúster 3 los más similares entre si.
En el caso del clúster 1, se observa que la gran mayoría de los pacientes han sido operados en los últimos 4 años, especialmente en el 2021, que es el último año que se refleja en el conjunto de datos utilizado. En los clústeres 2 y 3, el grueso de los pacientes fueron operados entre los años 2014 y 2016, siendo este año más representado en el clúster 3. Destacar también que en el clúster 2 hay gran representación de pacientes operados tanto en 2011 como en 2021.
Por último, en el clúster 4 encontramos un mayor reparto en lo que al año de cirugía se refiere, con mayoría de cirugías en el año 2021, seguido muy de cerca de los años 2014 y 2016.
Como se ha mencionado, los clústeres son muy distintos entre sí en cuanto al reparto del año de la cirugía, por lo que es un factor relevante a la hora de la agrupación de los datos.



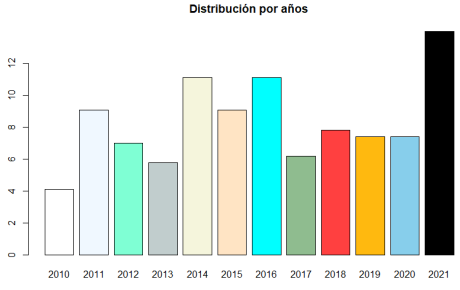
(a) Clúster 1



(b) Clúster 2



(c) Clúster 3

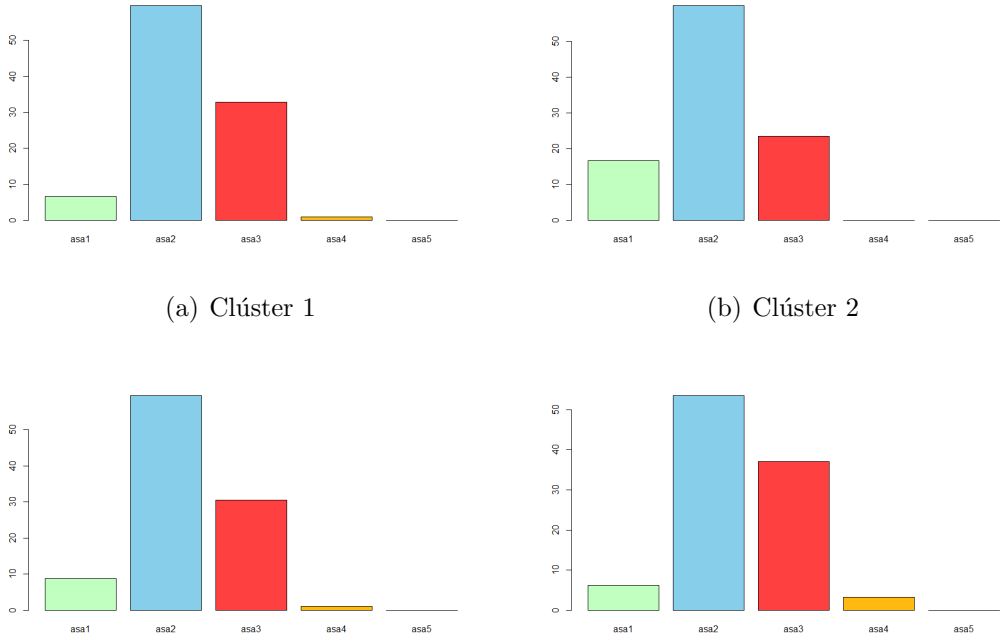


(d) Clúster 4

Figura 3: Comparación del año de la operación entre clústeres

7.2.4. ASA

A continuación se procede con el estudio de la división en función del estado del paciente respecto a la clasificación ASA. Es importante destacar que los pacientes con un grado ASA 5 son pacientes moribundos, que necesitan una cirugía en menos de 24 horas para sobrevivir, por lo que no se encuentra ninguno con ese grado en el conjunto de datos.
Se puede observar un patrón general en todos los clústeres, que se corresponde con que el grado 2 es el más extendido por el conjunto de datos, constituyendo más del 50 % en todos los clústeres.
El siguiente grado más extendido por los clústeres es el 3, constituyendo alrededor del 25 % del total de cada clúster. Seguido se puede encontrar el grado 1, presente en menor medida en todos los clústeres.
Por último se encuentra el grado 4, con una representación mínima en todos los clústeres, a excepción del último, en el que tiene algo más de relevancia.
La clasificación ASA no es un parámetro relevante a la hora de la clasificación de los datos, debido a la similitud en todos los clústeres.



(a) Clúster 1

(b) Clúster 2

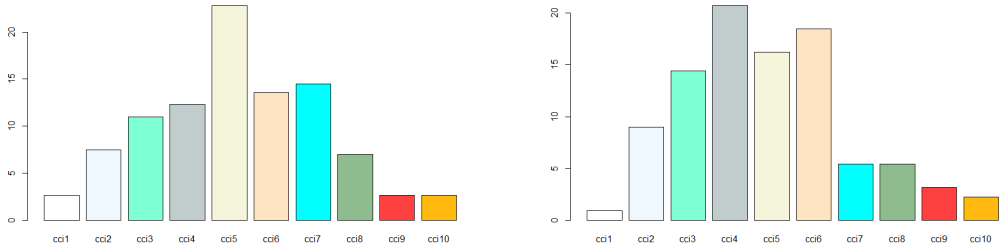
(c) Clúster 3

(d) Clúster 4

Figura 4: Comparación de grado ASA entre clústeres

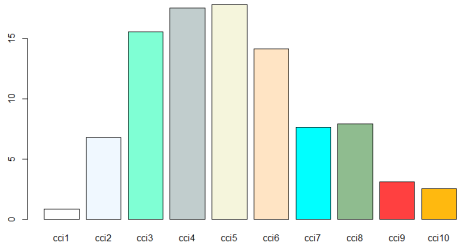
7.2.5. CCI

La clasificación CCI es la siguiente a analizar. Cabe destacar que el estudio se ha realizado con un grado máximo de CCI de 15, sin embargo, su porcentaje en los clústeres era ínfimo, por lo que, por claridad a la hora de analizar las gráficas, solo se han representado aquellos pacientes cuyo grado en la clasificación CCI es como máximo 10.
Observando las gráficas, se encuentra que los clústeres presentan cierto grado de heterogeneidad en cuanto a la clasificación de los pacientes según su CCI. Si bien en todos los clústeres el grueso de los pacientes se encuentra entre el CCI 3 y el 7, la distribución es distinta en todos los clústeres.
Por ejemplo, se observa que en el clúster 1 hay una gran mayoría de pacientes con un grado CCI de 5, mientras en el clúster 3 los grados de los pacientes se encuentran repartidos entre el 3 y el 5.
Estas diferencias hacen del CCI un factor relevante a la hora de la agrupación de los datos.

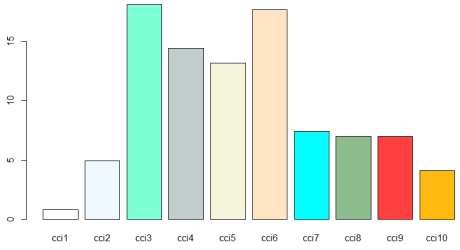


(a) Clúster 1

(b) Clúster 2



(c) Clúster 3

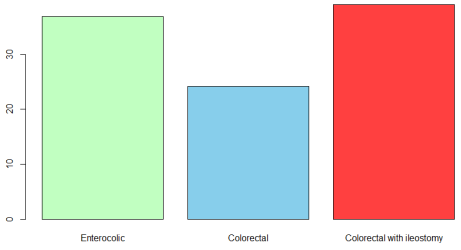


(d) Clúster 4

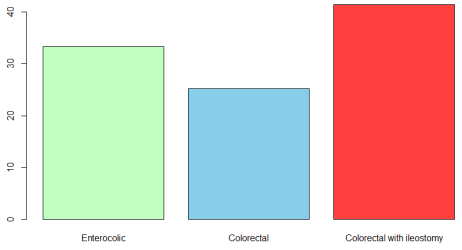
Figura 5: Comparación de grado CCI entre clústeres

7.2.6. Anastomosis

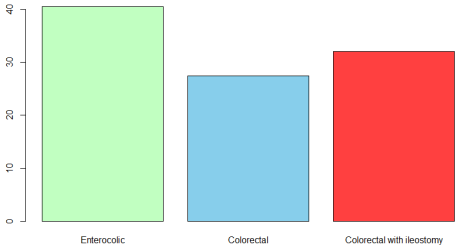
A continuación se procede con la clasificación en función del tipo de anastomosis requerida en la cirugía. En esta clasificación se puede observar poca diferencia entre los valores de cada uno de los clústeres, siendo la anastomosis enterocólica la más predominante en los clústeres 1 y 3, mientras que en los dos clústeres restantes predomina la anstomosis colorrectal con ileostomía protectora. Todos los clústeres coinciden en que la anastomosis de tipo 1 es la menos extendida.
Por estos motivos no se considera el tipo de anastomosis como un factor relevante, a la hora de la agrupación de los datos.



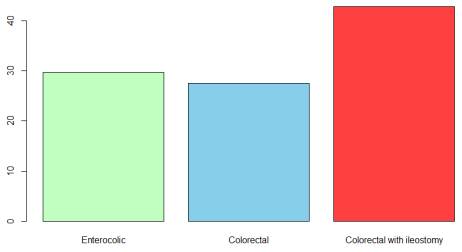
(a) Clúster 1



(b) Clúster 2



(c) Clúster 3

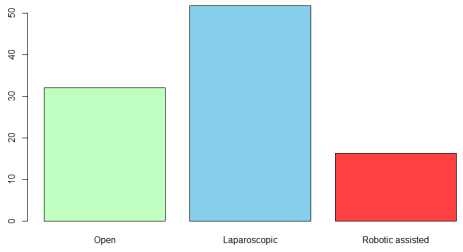


(d) Clúster 4

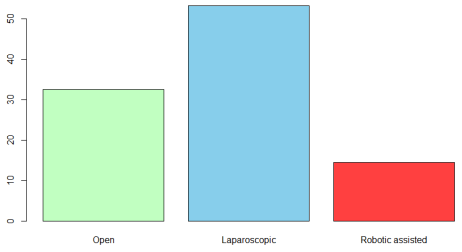
Figura 6: Comparación del tipo de anastomosis entre clústeres

7.2.7. Enfoque

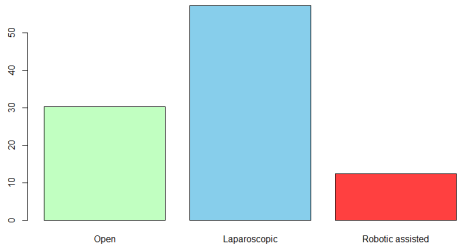
En este subapartado se compararán los clústeres en función del enfoque quirúrgico utilizado.
En este caso, se observa con un rápido repaso a las gráficas que el método más utilizado es la laparoscopia, constituyendo alrededor del 50 % de las operaciones de todos los clústeres.
El reparto en todos los clústeres es muy similar, como se ha mencionado, el método más extendido es la laparoscopia, seguido de la cirugía abierta y de la cirugía asistida por robot. El porcentaje de operaciones con estas dos últimas técnicas mencionadas cambia ligeramente entre los clústeres, pero no lo suficiente como para considerar al enfoque quirúrgico como un factor relevante a la hora de realizar la agrupación de los datos.



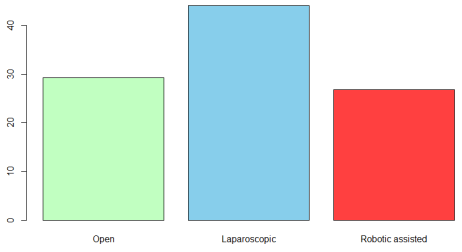
(a) Clúster 1



(b) Clúster 2



(c) Clúster 3



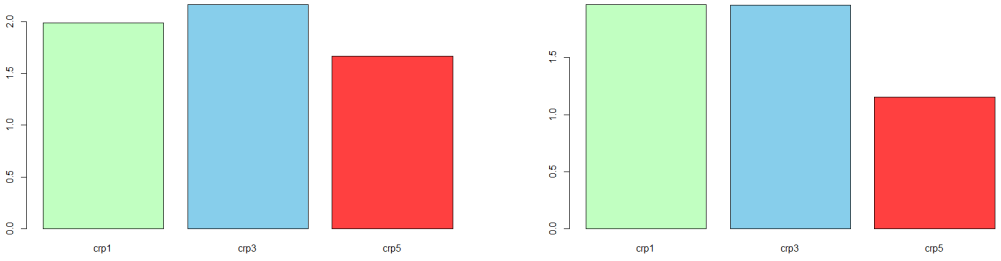
(d) Clúster 4

Figura 7: Comparación del tipo de enfoque entre clústeres

7.2.8. CRP

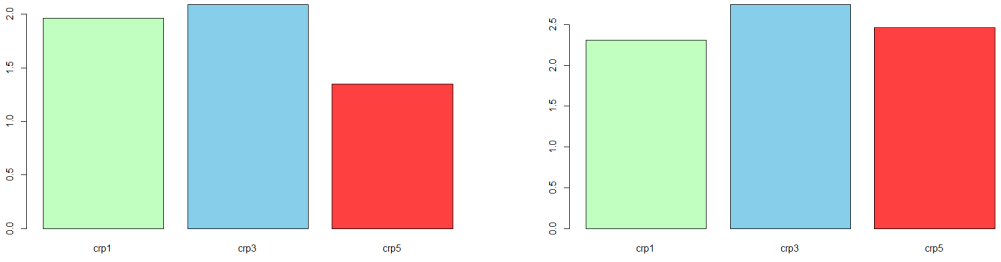
En esta subsección se va a estudiar la evolución del nivel de proteína C-reactiva en los pacientes los 5 primeros días después de la cirguía. Según se menciona en [Black et al., 2004]los niveles de CRP, que en una persona sana son bajos, sufren una subida al haber una reacción inflamatoria en el organismo, en este caso producida por la cirugía.
En este caso, se puede observar que el nivel se encuentra alrededor de 2,0 el primer día en los clústeres 1,2 y 3, aumentando ligeramente el tercer día, exceptuando el clúster 2, en el que se mantiene constante. El nivel de CRP el quinto día difiere un poco en estos tres clústeres, tomando un valor superior a 1,5 en el primero, de 1,0 en el segundo y alrededor de 1,25 en el tercero.

En el cuarto clúster los niveles son superiores a los mencionados en el párrafo anterior. El primer día se encuentra un nivel alrededor de 2,25. El pico más alto llega el tercer día, superando el 2,5, acabando el quinto día con un nivel superior al primer día. Estos niveles tan elevados nos indican que este clúster agrupa a pacientes que han tenido una mayor inflamación tras la cirugía.



(a) Clúster 1

(b) Clúster 2



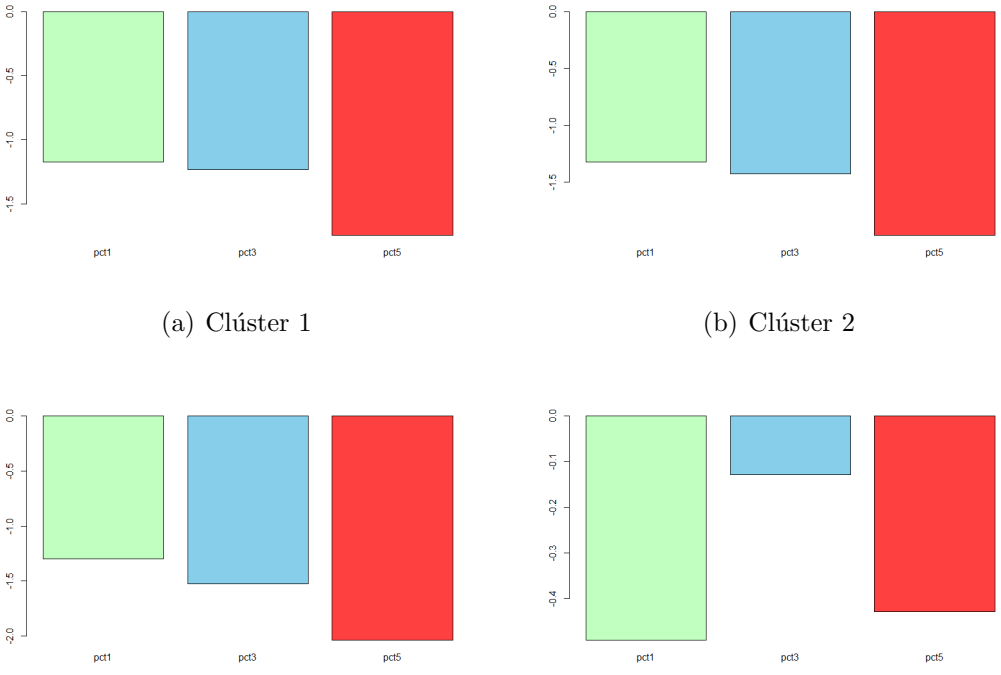
(c) Clúster 3

(d) Clúster 4

Figura 8: Comparación de la evolución del CRP entre clústeres

7.2.9. PCT

Al igual que en el apartado anterior, se va a analizar la evolución de un biomarcador durante el primer, tercer y quinto días después de la cirugía. En este caso se va a analizar el nivel de procalcitonina. Al igual que en el caso de la proteína C-reactiva, una persona sana tiene un nivel bajo de procalcitonina, nivel que puede aumentar en caso de existir una infección bateriana, según [Esper and Calatayud, 2013].
En este caso ocurre algo muy similar al caso anterior, se encuentran tres clústeres con números muy similares entre si y un cuarto con niveles y distribución distintos. Por una parte, se encuentran los clústeres 1, 2 y 3 con unos niveles y una evolución de los mismos muy similares. El día uno el nivel se encuentra ligeramente por debajo de $-1{,}0$, disminuyendo ligeramente el tercer día y volviendo a disminuir, esta vez de manera más significativa, el quinto, llegando a niveles inferiores a $-2{,}0$.
En el otro lado, se encuentra el clúster 4, con un nivel alrededor de $-0{,}5$ el primer día, que aumenta significativamente el tercero, llegando prácticamente a $-0{,}1$, volviendo a disminuir hasta $-0{,}4$ el quinto día. Estos niveles indican que los pacientes en este clúster pueden tener alguna infección de índole bacteriana.



(a) Clúster 1

(b) Clúster 2

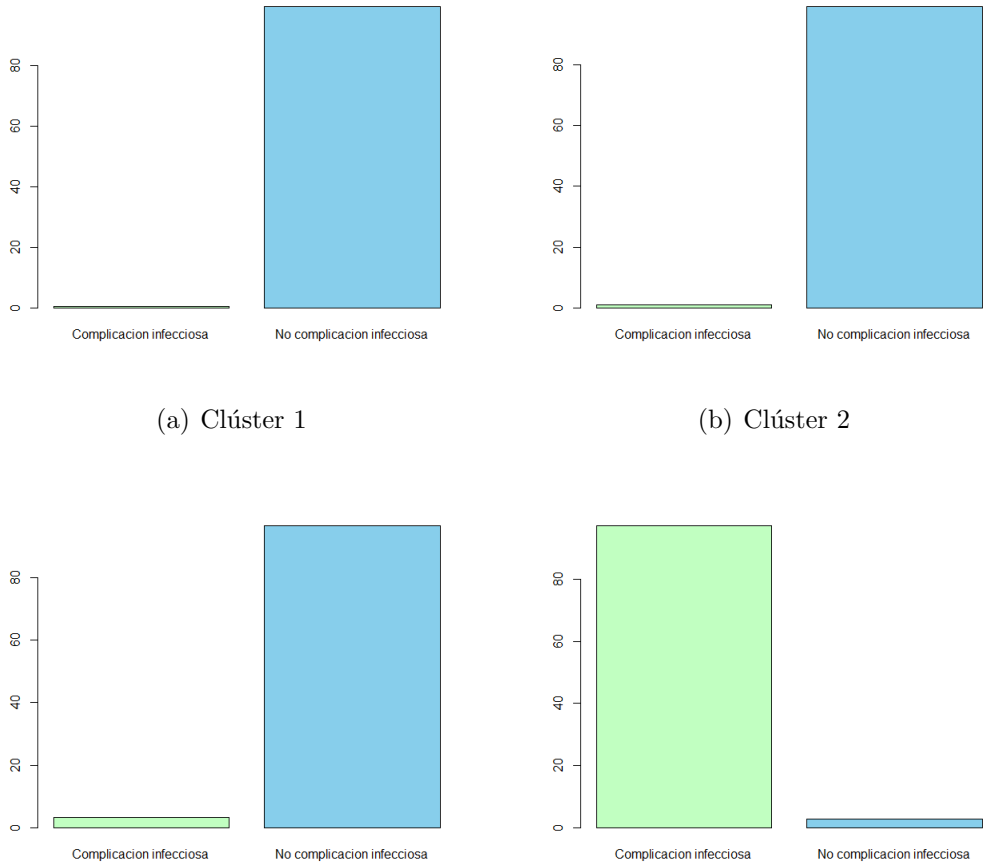
(c) Clúster 3

(d) Clúster 4

Figura 9: Comparación de la evolución del PCT entre clústeres

7.2.10. Complicación infecciosa

El siguiente parámetro objeto de estudio es la aparición de complicaciones de índole infecciosa en los pacientes después de la operación.
En los dos primeros clústeres encontramos que la gran mayoría de los pacientes no han sufrido ningún tipo de complicación de este tipo. Esto también puede aplicarse al tercer clúster, aunque en este caso el porcentaje de pacientes que han sufrido alguna complicación aumenta.
En el cuarto clúster se encuentran la gran mayoría de pacientes que han sufrido una complicación infecciosa, correspondiéndose con alrededor del 90 % del cluster.
De nuevo, es el cuarto clúster el que presenta mayor nivel de diferencia con el resto, haciendo de la aparición de complicaciones de índole infecciosa un factor relevante para la agrupación de los datos.



(a) Clúster 1

(b) Clúster 2

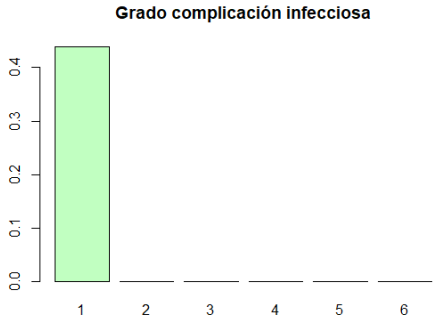
(c) Clúster 3

(d) Clúster 4

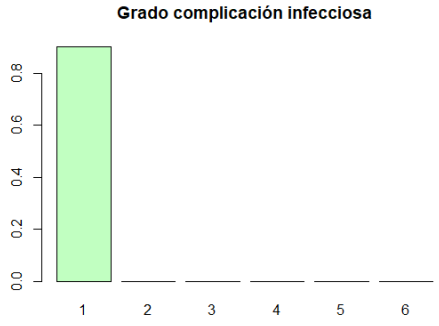
Figura 10: Comparación de presencia de complicación infecciosa entre los clústeres

7.2.11. Grado de complicación infecciosa

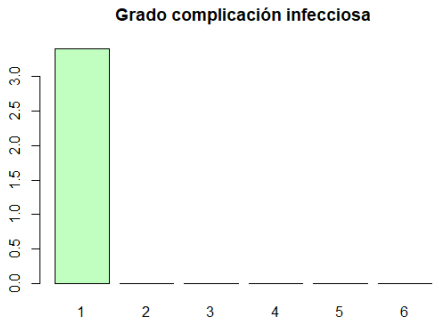
Este subapartado está fuertemente enlazado con el anterior. Como se mencionó en el anterior apartado, solo en el clúster 4 había un alto porcentaje de pacientes que sufrieron alguna complicación infecciosa.
Se puede observar que en los clústeres 1,2 y 3, solo hay pacientes con complicaciones infecciosas de grado 1, mientras que en el clúster 4, hay mayor diversidad de grados, desde 2 hasta 5.
Estos resultados son lógicos, ya que el grado de complicación más extendido es el 1, por lo que es normal que en los clústeres en los que apenas haya pacientes con complicaciones, el único grado sea este. Por otra parte, en el clúster 4 hay gran mayoría de pacientes con complicación infecciosa, por lo que es normal que haya una mayor variedad de grados.



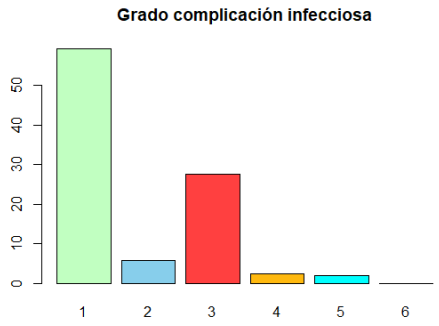
(a) Clúster 1



(b) Clúster 2



(c) Clúster 3

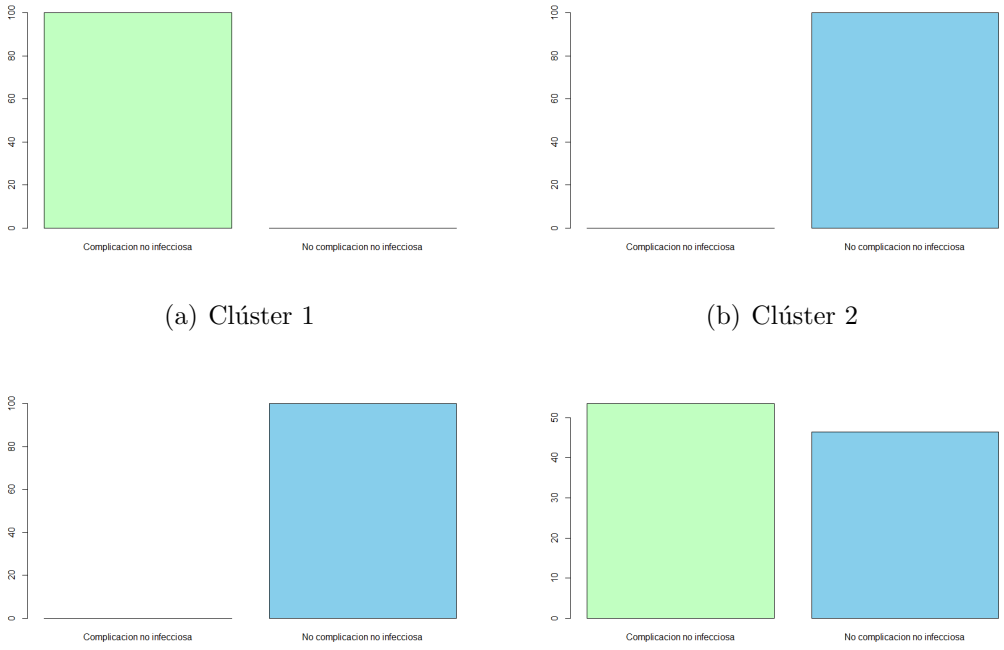


(d) Clúster 4

Figura 11: Comparación del grado de complicación infecciosa entre los clústeres

7.2.12. Complicación no infecciosa

En este apartado se va a analizar de nuevo la aparición de complicaciones, en este caso de índole no infecciosa.
Se observa que en los clústeres 2 y 3 no hay ningún paciente con complicaciones de este tipo.
Por otra parte, el clúster 1 representa el caso totalmente contrario, el 100 % de los pacientes han sufrido complicaciones no infecciosas.
Por último, en el clúster 4 el porcentaje está más repartido, constituyendo los pacientes con complicación no infecciosa poco más del 50 % del clúster.
Los clústeres presentan muchas disimilitudes entre sí, haciendo de la aparición de complicaciones no infecciosas un factor importante a la hora de la agrupación de los datos.



(a) Clúster 1

(b) Clúster 2

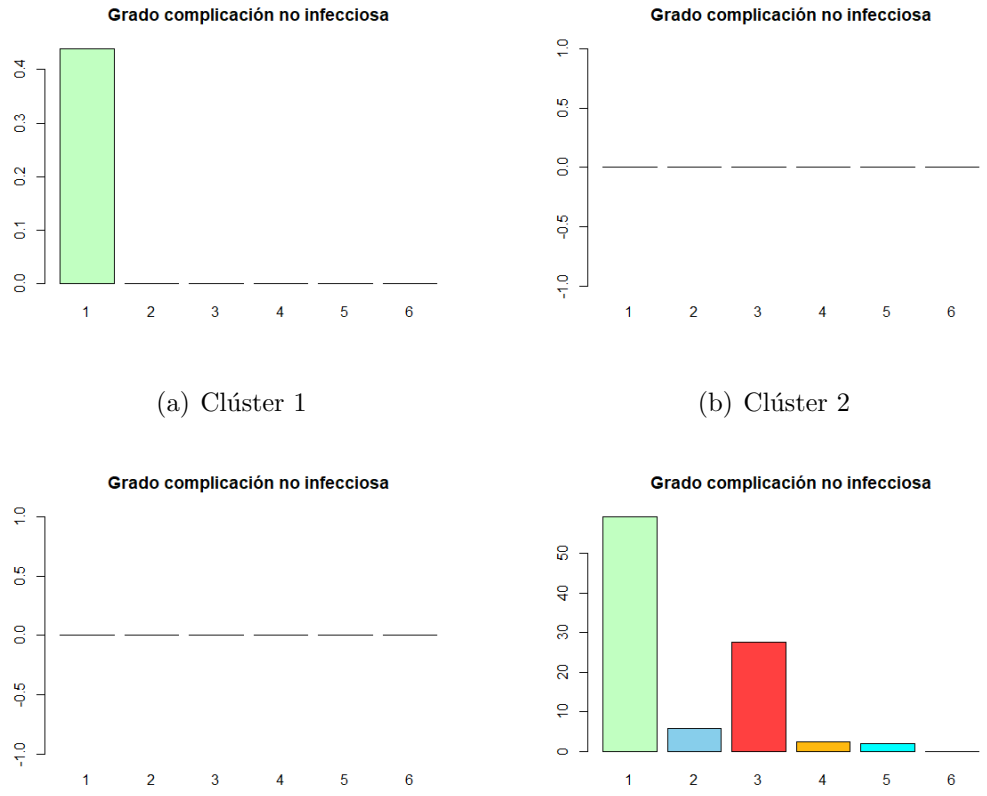
(c) Clúster 3

(d) Clúster 4

Figura 12: Comparación de aparición de complicación no infecciosa entre los clústeres

7.2.13. Grado de complicación no infecciosa

Al igual que ocurría en el caso de las complicaciones de índole infecciosa, esta subsección está fuertemente ligada a la anterior.
Como se observó en la subsección anterior, en los clústeres 2 y 3 el porcentaje de pacientes con complicaciones no infecciosas es de 0 %, lo cual se corresponde con las gráficas vacías de la figura 13.
Por otra parte, se puede ver que en el clúster 1 el 100 % de los pacientes tienen una complicación no infecciosa de grado 1.
Por último, es el clúster 4 el que contiene una mayor variedad de pacientes con distintos grados de complicaciones, siendo de nuevo el grado 1 el más extendido, seguido por el grado 3 y teniendo los grados 2, 4 y 5 unos porcentajes muy similares.



(a) Clúster 1

(b) Clúster 2

(c) Clúster 3

(d) Clúster 4

Figura 13: Comparación del grado de complicación no infecciosa entre los clústeres

7.3. Comparación con división en medoides

En el subarpatado anterior, se ha estudiado la división en clústeres utilizando centroides, pero también es interesante estudiar la distribución de los datos en función de medoides. Para ello, se ha utilizado el algoritmo PAM (partitioning around medoids), mediante la función de R "pam(data,k,nstart)"

Clúster	K-means	PAM
1	228	241
2	222	331
3	353	271
4	243	203

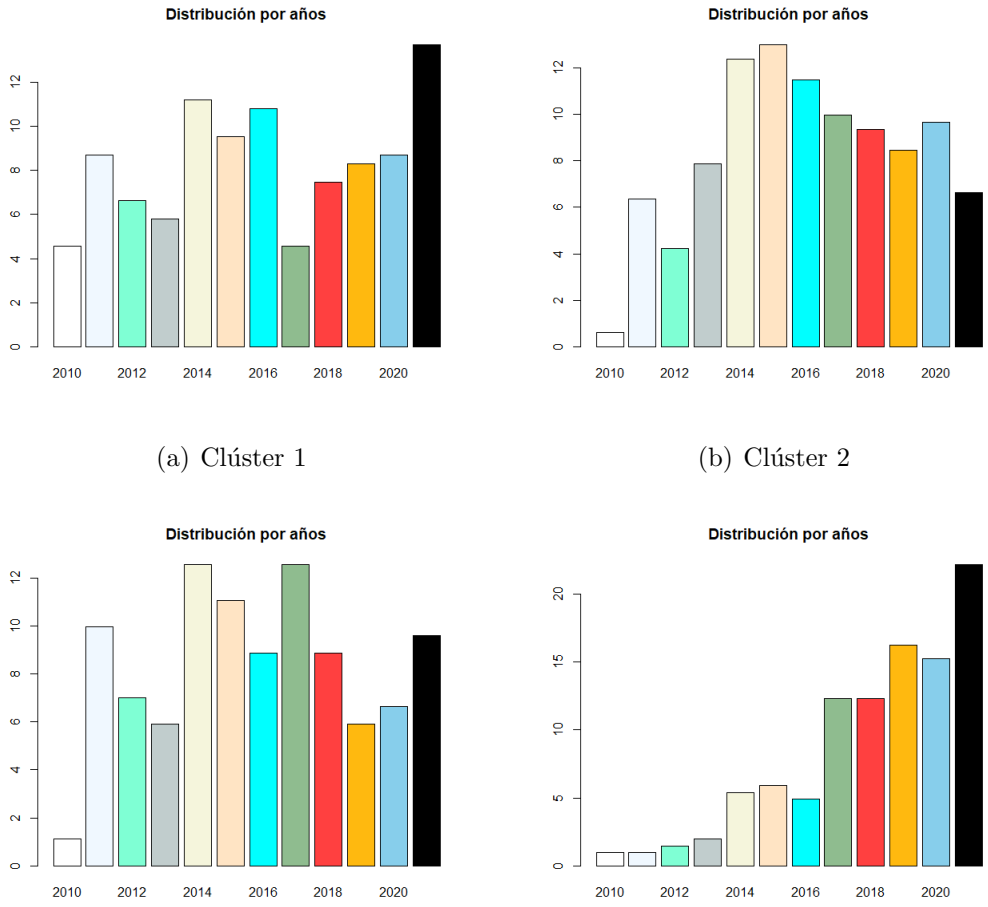
Cuadro 1: Comparación del tamaño de los clústeres

Comparando el tamaño de cada clúster en ambos modelos, se observa que la distribución difiera en gran medida del modelo planteado por el algoritmo K-Means, siendo el clúster número 2 el que tiene un mayor número de observaciones y el 4 el contrario.
A continuación, se muestra la representación de dichos clústeres:

Figura 14: Distribución de los clústeres

A continuación se van a analizar y comentar aquellas variables que son claves a la hora de la clasificación de los datos según el algoritmo PAM, en comparación con los que se han mencionado anteriormente del algoritmo K-Means.

7.3.1. Año



(a) Clúster 1

(b) Clúster 2

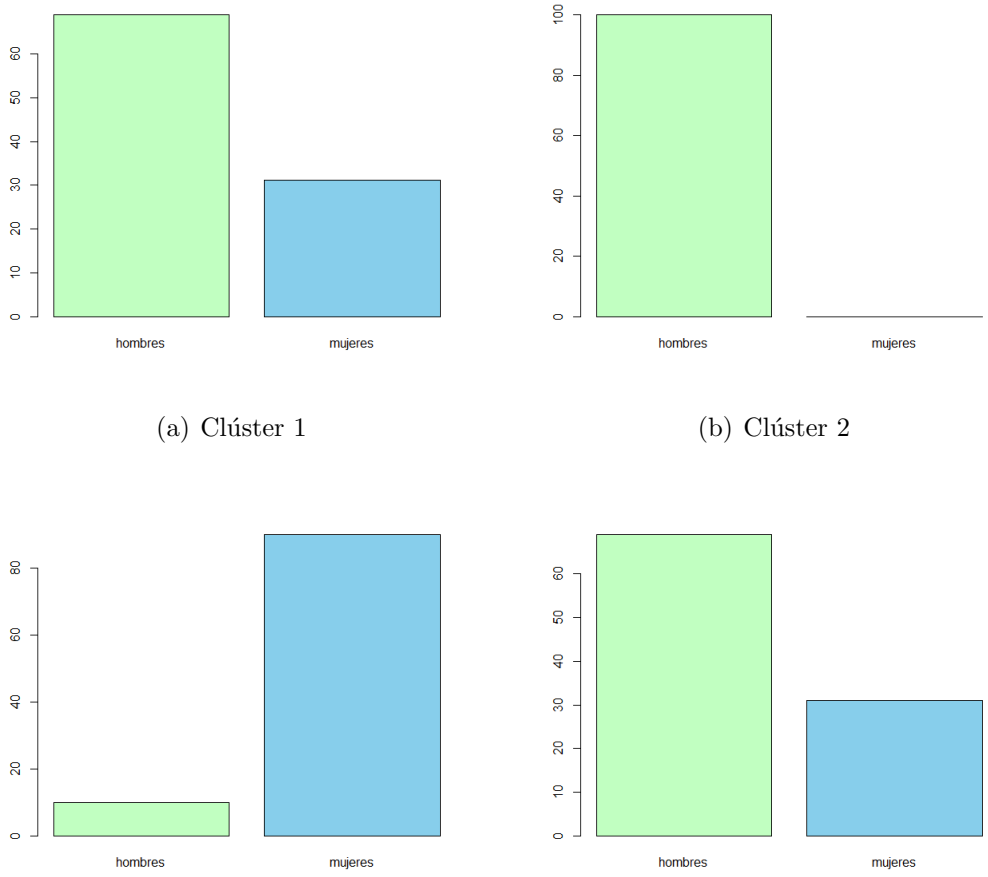
(c) Clúster 3

(d) Clúster 4

Figura 15: Comparación del año de la operación entre clústeres

El año en el que se ha realizado la cirugía es uno de esos factores previamente mencionados, ya que, como se observa en la figura 15, el reparto entre clústeres es heterogéneo, siendo el año 2020 el más recurrente en los clústeres 1 y 4, mientras que está más repartido en los dos clústeres restantes.

7.3.2. Sexo



(a) Clúster 1

(b) Clúster 2

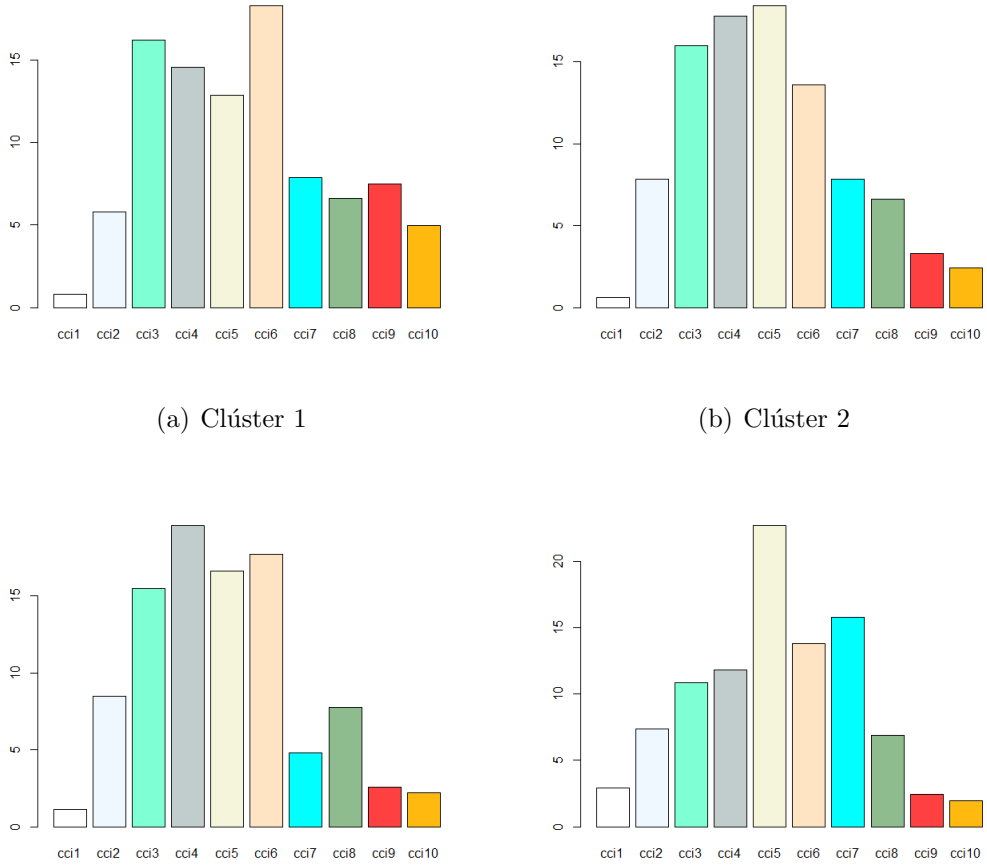
(c) Clúster 3

(d) Clúster 4

Figura 16: Comparación del sexo entre clústeres

Se puede observar un reparto muy similar al analizado previamente en el algoritmo K-Means, siendo los clústeres 1 y 4 prácticamente idénticos. Las diferencias residen en los clústeres 2 y 3, que tienen un reparto diferente al analizado previamente.
Por ello, el sexo es un factor clave a la hora de realizar el reparto según el algoritmo PAM.

7.3.3. CCI



(a) Clúster 1

(b) Clúster 2

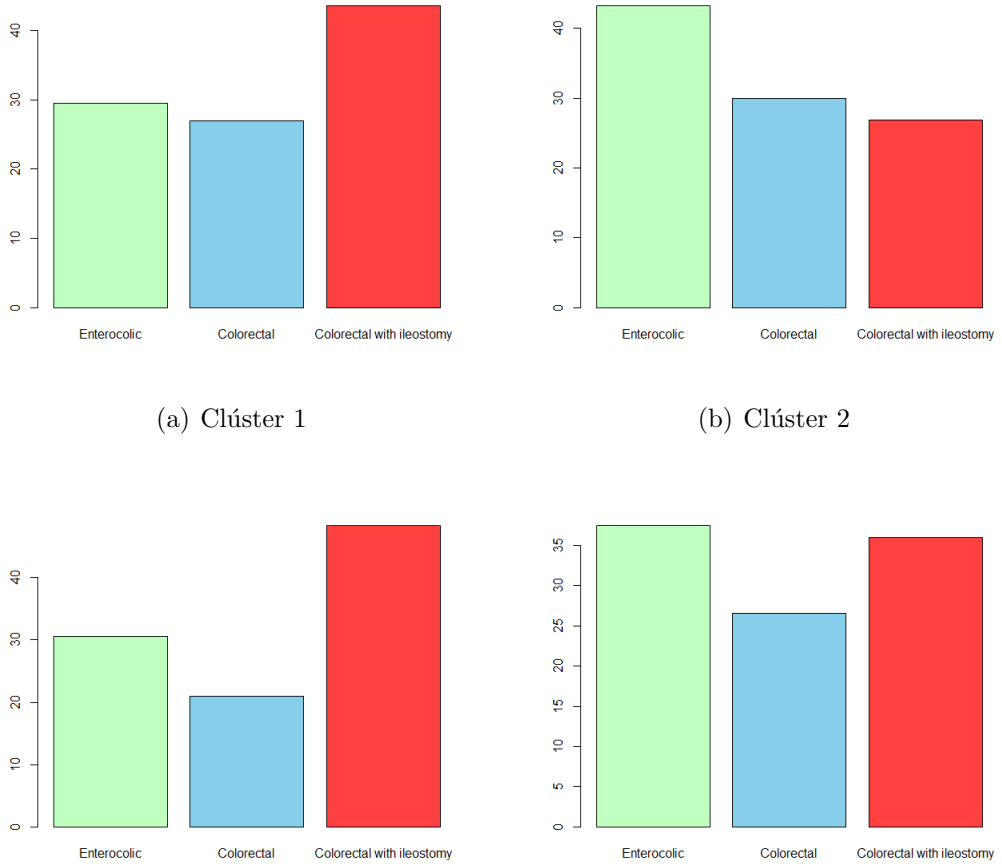
(c) Clúster 3

(d) Clúster 4

Figura 17: Comparación del cci entre clústeres

El índice de comorbilidad de Charlson es un factor relevante a la hora de realizar la clasificación según el algoritmo PAM. La distribución es muy similar a la realizada con el método k-means pero con un orden distinto. Es decir, el reparto es casi idéntico pero los clústeres no se identifican entre sí (por ejemplo, el clúster 1 del algoritmo K-Means se corresponde con el clúster 4 del algoritmo PAM)

7.3.4. Anastomosis



(a) Clúster 1

(b) Clúster 2

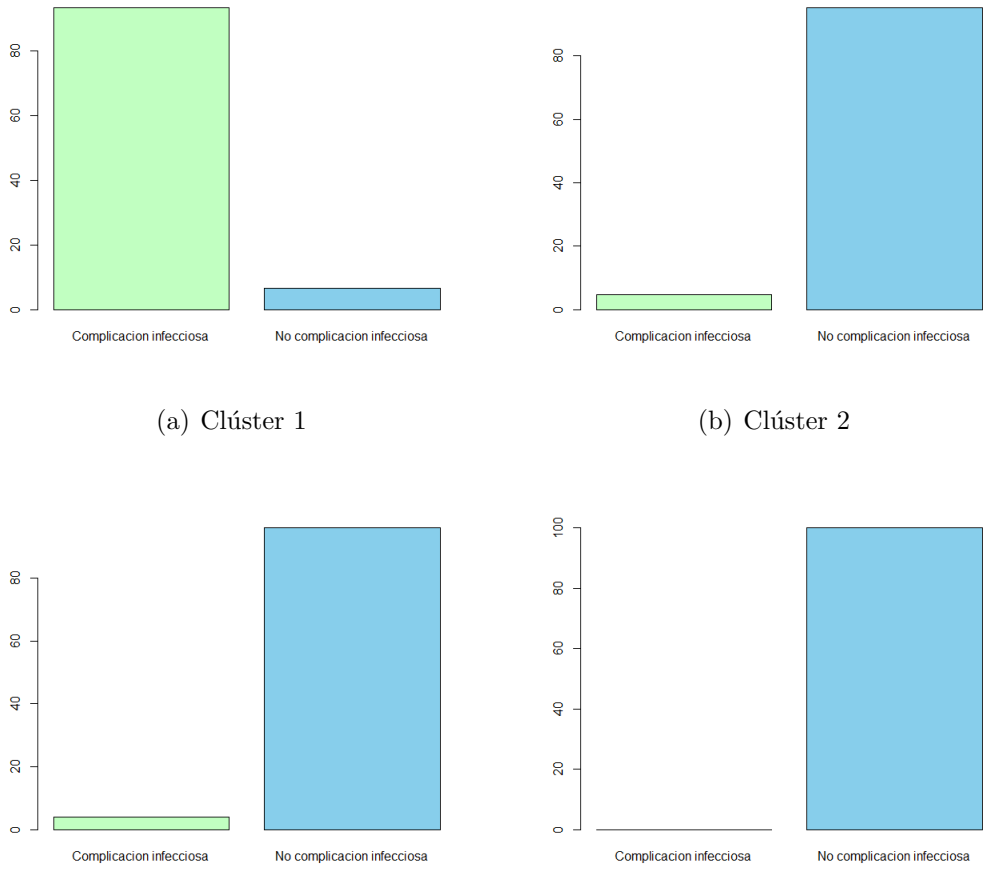
(c) Clúster 3

(d) Clúster 4

Figura 18: Comparación de la anastomosis entre clústeres

La anastomosis requerida durante la operación es otro de los factores importantes a la hora de la clasificación de los datos según este algoritmo. El método más común es la anastomosis colorrectal con ileostomía, siendo la más presente en los clústeres 1 y 3, teniendo también una gran influencia en el clúster 4.

7.3.5. Complicación infecciosa



(a) Clúster 1

(b) Clúster 2

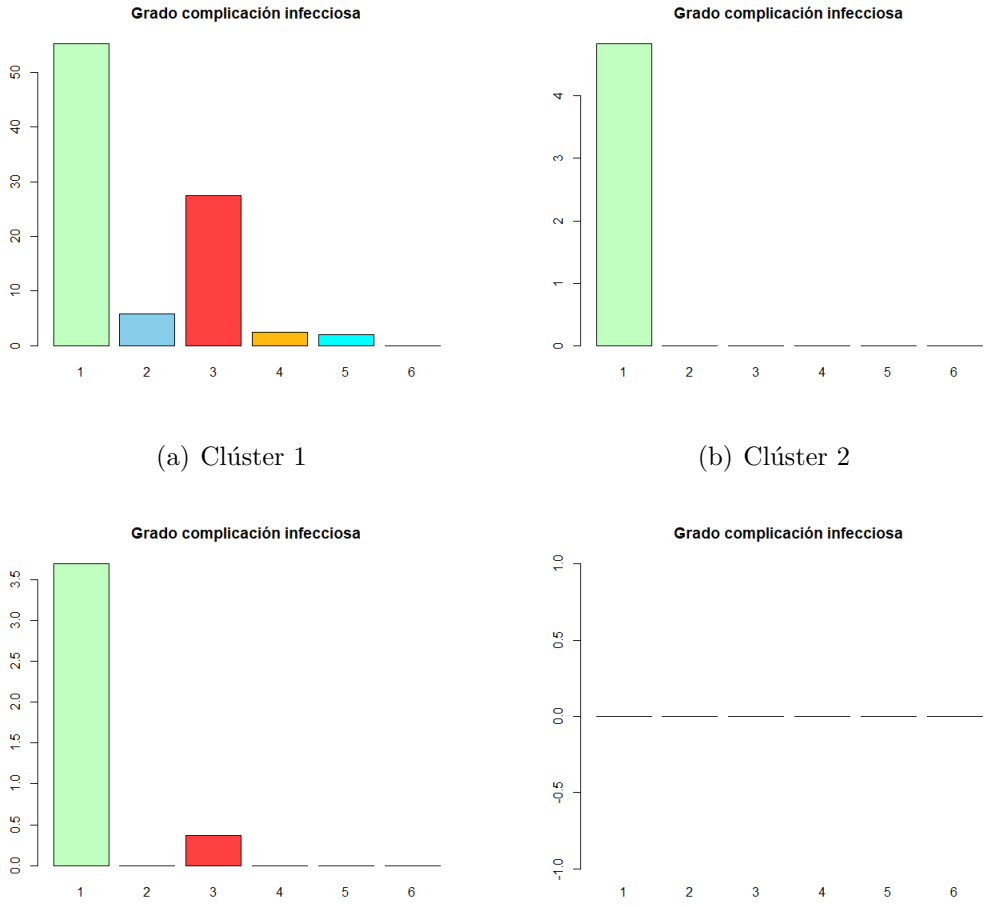
(c) Clúster 3

(d) Clúster 4

Figura 19: Comparación de complicación infecciosa entre clústeres

Otro de los factores claves de clasificación según este modelo es la existencia de complicaciones de índole infecciosa, siendo el clúster 1 el que mayor porcentaje de complicaciones de este tipo presenta, presentando en los otros tres clústeres un porcentaje ínfimo.

7.3.6. Grado de complicación infecciosa



(a) Clúster 1

(b) Clúster 2

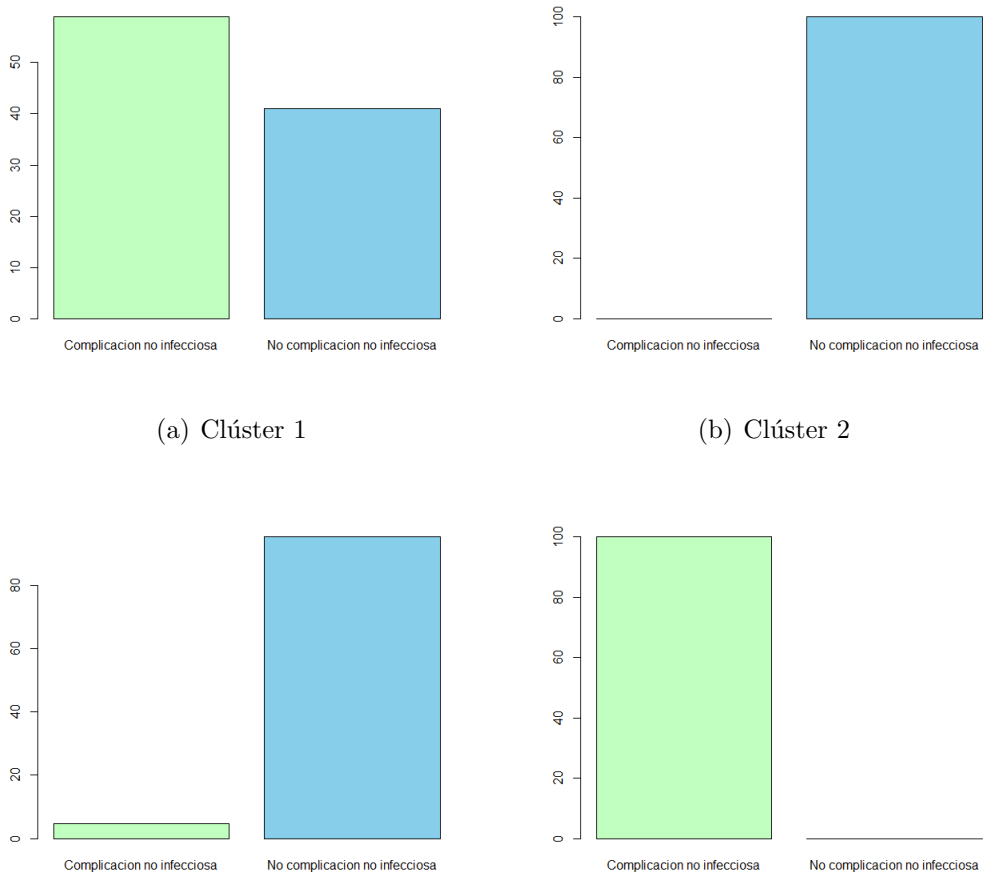
(c) Clúster 3

(d) Clúster 4

Figura 20: Comparación del grado complicación infecciosa entre clústeres

Este apartado está fuertemente relacionado con el anterior. Por ello, el clúster 1, que presentaba un gran porcentaje de complicación infecciosa, refleja una mayor variedad de grados de complicación. En el resto, al haber un menor porcentaje de complicación, la mayoría es de grado 1, con un pequeño porcentaje de grado 3 en el tercer clúster.
Por otra parte, como en el clúster 4 no hay pacientes con complicaciones de esta índole no pueden existir ningún grado para ellas.

7.3.7. Complicación no infecciosa



(a) Clúster 1

(b) Clúster 2

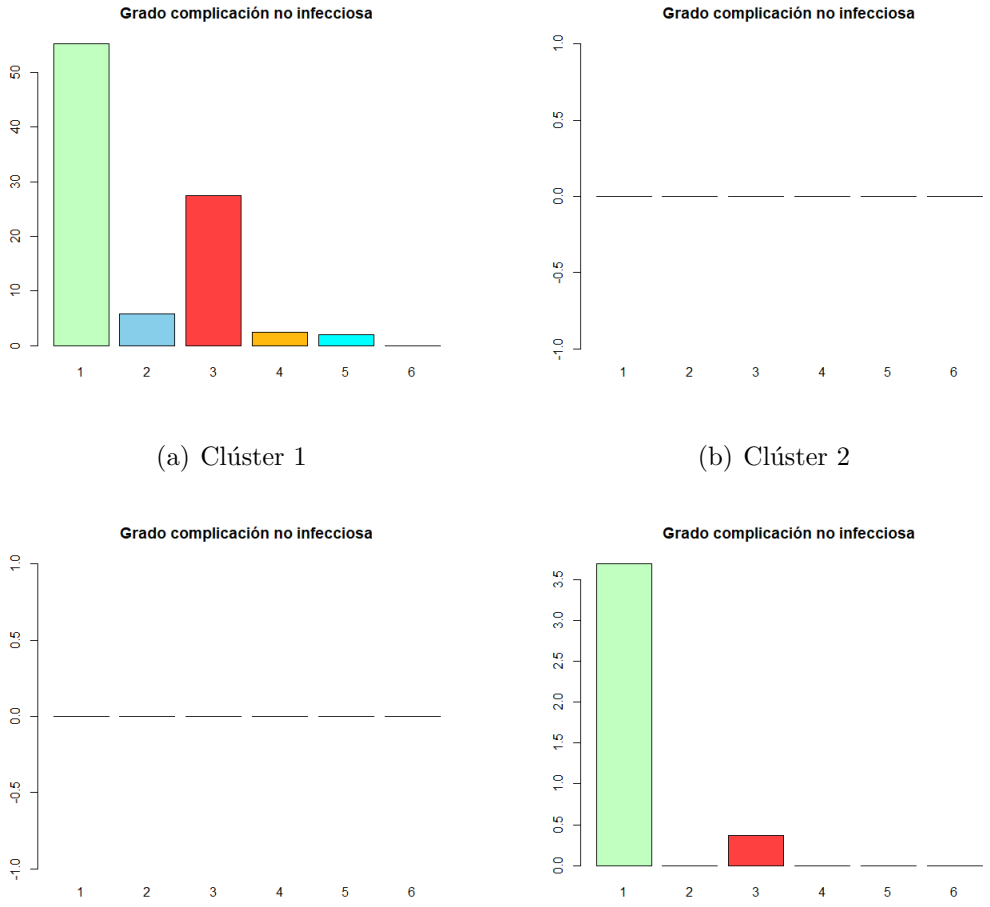
(c) Clúster 3

(d) Clúster 4

Figura 21: Comparación de complicación no infecciosa entre clústeres

La existencia de complicaciones de índole no infecciosa también es un factor importante en la división de los datos según el modelo propuesto por el algoritmo PAM. En este caso, se observa que los mayores porcentajes se observan en los clústeres 1 y 4, habiendo un mayor de pacientes sin complicaciones de este tipo en los clústeres 2 y 3.

7.3.8. Grado complicación no infecciosa



(a) Clúster 1

(b) Clúster 2

(c) Clúster 3

(d) Clúster 4

Figura 22: Comparación del grado de complicación no infecciosa entre clústeres

Al igual que en el caso de las complicaciones de índole infecciosa, el grado de complicación es un reflejo de lo que se ha visto en el apartado anterior. Los clústeres que presentaban mayor porcentaje de complicaciones no infecciosas son los mismos que tienen grados de los mismos.

7.3.9. Comparación entre ambos modelos

Una vez realizado el estudio de ambos modelos, se puede observar que, pese a uno utilizar centroides y otro medoides, la diferencia entre ambos modelos no es muy significativa, aunque encontramos diferencias en el reparto de ciertos datos por los clústeres. Es por ello que no se puede definir de forma certera cuál de los dos modelos es mejor para este conjunto de datos.

7.4. Modelo de clasificación

Una vez realizada la separación y análisis de los clústeres, se procede con la regresión logística.
Como se ha mencionado anteriormente, en este caso se aplica el modelo de regresión logística para todo el conjunto de datos, para tratar de predecir la aparición de complicaciones, ya sean de índole infecciosa o no.

7.4.1. Modelo para complicación infecciosa

El primer caso que se va a estudiar es el modelo para la predicción de complicaciones de índole infecciosa.
Como se ha mencionado en apartados anteriores, lo primero que se realiza en este método es el entrenamiento del conjunto de datos destinado a ello, mediante el comando "glm()", que recibe como parámetro la variable que se quiere predecir, las variables que se utilizan para dicha predicción, el conjunto de datos de entrenamiento sobre el que se quiere trabajar y el tipo de modelo que se quiere predecir, en este caso binomial, ya que solo puede tomar valor 0 o 1. Se utilizan todas las variables del conjunto para las predicciones, a excepción de las relacionadas con las propias variables a predecir(estas son, las complicaciones de ambos tipos y sus respectivos grados), así como el ID del paciente.
Se pueden analizar los resultados del modelo usando el comando "summary()", representándose de la siguiente manera:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3096  -0.6153  -0.2994   0.2994   3.0235

Coefficients:
             Estimate Std. Error z value        Pr(>|z|)
(Intercept) 158.169863  80.240121   1.971         0.04870 *
Age          -0.003721   0.011678  -0.319         0.75002
ASA           0.135144   0.213658   0.633         0.52704
Year         -0.080433   0.039861  -2.018         0.04361 *
CCI          -0.010537   0.060236  -0.175         0.86114
Sex          -0.328041   0.250465  -1.310         0.19029
Anastomosis   0.370889   0.157801   2.350         0.01876 *
Approach     -0.376042   0.164079  -2.292         0.02192 *
CRP1          0.050111   0.302437   0.166         0.86840
CRP3          0.243992   0.274241   0.890         0.37363
CRP5          1.408850   0.212841   6.619 0.0000000000361 ***
PCT1         -0.411158   0.128600  -3.197         0.00139 **
PCT3          0.360749   0.142397   2.533         0.01130 *
PCT5          0.236577   0.147938   1.599         0.10978
```

Figura 23: Resumen del entrenamiento del modelo para complicación infecciosa

Los valores más relevantes mostrados en la figura 23 son los siguientes:

- Deviance residuals: mide la diferencia entre los valores observados y los valores predichos por el modelo. Cuanto más cerca de cero estén los valores, mejor será el ajuste del modelo.
- Coefficients: representa las estimaciones de los efectos de las variables predictoras sobre la variable de respuesta.
- Estimate: representación de la estimación del efecto de cada variable en log-odds.
- Std.Error: es una medida de la incertidumbre asociada a la estimación de cada coeficiente.
- z value: se calcula dividiendo la estimación del coeficiente entre la desviación.
- Pr($> |z|$): indica la probabilidad de obtener un valor z igual o más extremo que el valor z observado en el modelo, si la relación entre variable predictora y variable de respuesta es nula.

Una vez realizado el entrenamiento de los datos, se realiza la predicción del modelo sobre el conjunto de prueba, utilizando el método "predict(objeto,tipo)", usando tanto los datos entrenados como los datos destinados a test. En cuanto al tipo, se utiliza "response", que proporciona las probabilidades de predicción. Es interesante representar estas probabilidades mediante un histograma:

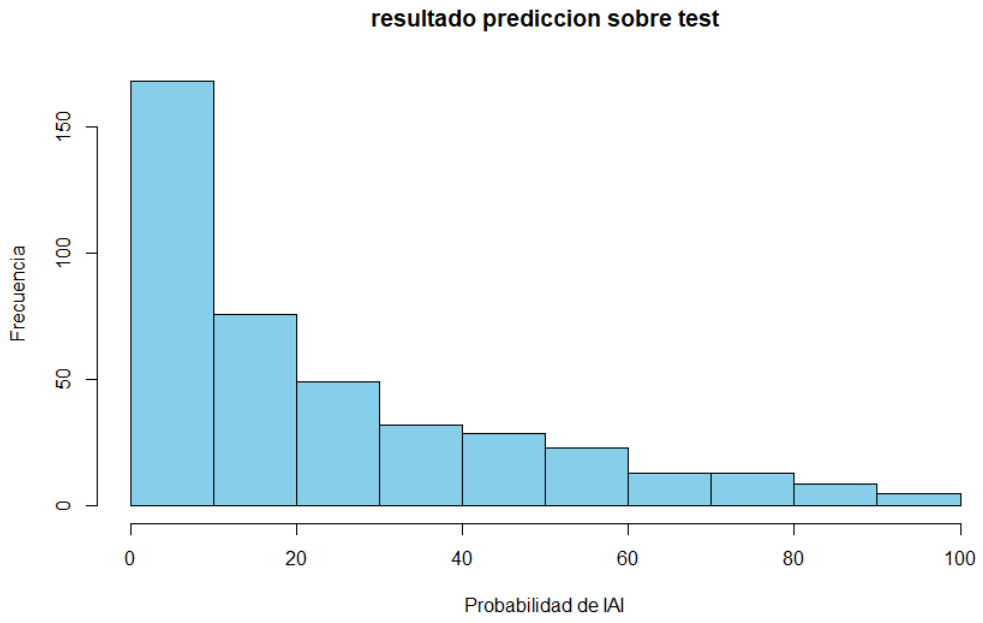


Figura 24: Histograma resultados de predicción de complicación infecciosa

Analizando el histograma, se observa que la mayoría de pacientes tienen entre un 0 % y un 10 %, con una frecuencia superior a 150, reduciéndose drásticamente entre

el 10 % y el 20 % y disminuyendo paulatinamente a partir de ese punto, hasta llegar al intervalo entre 90 % y 100 %, donde la frecuencia es prácticamente nula.
Sabiendo estos datos, se ha de establecer un valor porcentual a partir del cual se puede decir que la predicción es positiva o negativa para el paciente. Establecer un porcentaje del 50 % como umbral puede llegar a ser algo arriesgado, puesto que hará que la predicción sea ambigua.
Por ello, se ha decidido establecer un porcentaje umbral del 40 %, lo que significa que por debajo de ese porcentaje, la predicción establecerá que el paciente no ha sufrido ninguna complicación de índole infecciosa, estableciendo el caso contrario en caso de superar dicho umbral.
Las predicciones se almacenan en una variable denominada "pred_final", la cual toma los siguientes valores:

Predicción	Nº de elementos
Complicación	87
No complicación	332

Cuadro 2: Comparación del número de elementos para cada predicción

El siguiente paso es la representación y evaluación del modelo mediante una curva ROC y el área debajo de la curva.

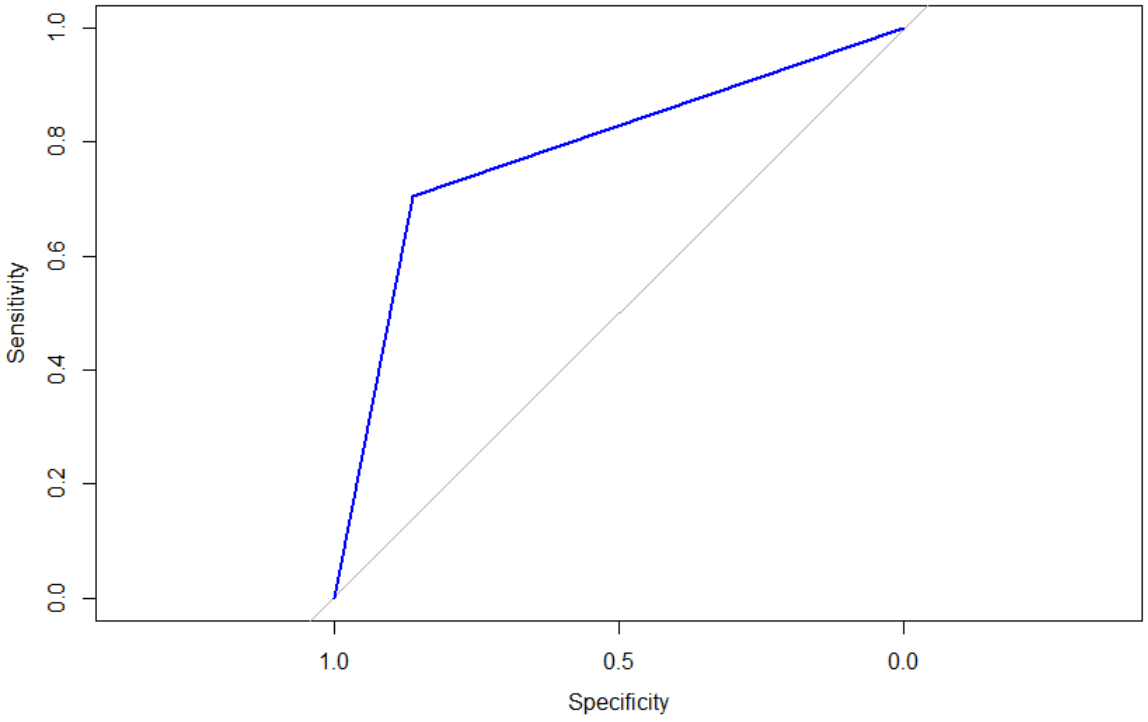


Figura 25: Curva ROC para modelo de complicación infecciosa

Calculando el valor del AUC se obtiene 0.778, un valor que se encuentra más cerca de 1 que de 0.5, por lo que, si bien el test no es excelente (se dice que es excelente cuando tiene un valor superior a 0.97), se puede considerar como un test bueno.
El último paso del proceso es la validación del modelo, en este caso aplicando el método de validación cruzada Leave One Out.
En este caso, aplicando este método de validación cruzada se obtienen los siguientes valores de precisión y error:

Precisión	0.821
Error	0.179

Cuadro 3: Valores de preción y error tras usar validación cruzada

7.4.2. Modelo para complicación no infecciosa

A continuación se van a mostrar los resultados del modelo aplicado para la predicción de la complicaciones de índole no infecciosa. En primer lugar, al igual que en el caso anterior, se realiza el entrenamiento del conjunto de datos destinado a ello, lo cual devuelve lo siguiente:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8054  -0.8341  -0.5289   0.9270   2.8182

Coefficients:
              Estimate  Std. Error z value            Pr(>|z|)
(Intercept) -554.069831   71.071521  -7.796 0.00000000000000639 ***
Age           -0.001787    0.009685  -0.184              0.8536
Year           0.274111    0.035232   7.780 0.00000000000000725 ***
ASA            0.136771    0.176251   0.776              0.4377
CCI            0.047135    0.051306   0.919              0.3582
Sex           -0.049762    0.202066  -0.246              0.8055
Anastomosis    0.195317    0.134151   1.456              0.1454
Approach      -0.174076    0.148109  -1.175              0.2399
CRP1           0.129932    0.238171   0.546              0.5854
CRP3          -0.552332    0.236496  -2.335              0.0195 *
CRP5           0.811691    0.167377   4.849 0.00000123789142546 ***
PCT1          -0.045166    0.113935  -0.396              0.6918
PCT3           0.107377    0.133575   0.804              0.4215
PCT5           0.126858    0.134268   0.945              0.3448
```

Figura 26: Resumen del entrenamiento del modelo para complicación no infecciosa

Comparando estos valores con los valores del modelo anterior, se puede entrever que este modelo va a ser menos preciso.
De nuevo, se representa los valores de la predicción mediante un histograma:

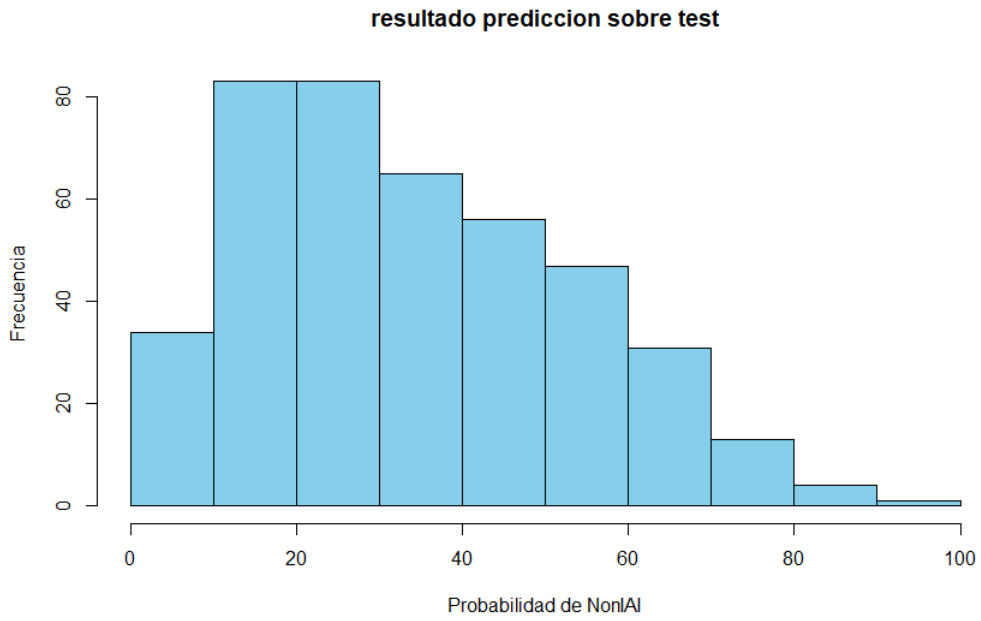


Figura 27: Histograma resultados de predicción de complicación no infecciosa

En el caso de las complicaciones de índole no infecciosa, se observa que las frecuencias entre intervalos de probabilidades tienen una mayor similitud que en el caso anterior. El grueso de las predicciones en este caso se encuentran entre el 10 % y el 40 %, rondando frecuencias entre 70 y 90.
Las frecuencias a partir del 40 % también son mayores que en el caso anterior, exceptuando los intervalos entre 80 % y 100 %, que en ambos casos son muy similares. Con estos datos y utilizando el mismo umbral de porcentaje que en el caso anterior, se puede llegar a predecir que la probabilidad de que exista una complicación de índole no infecciosa es considerablemente más alta que la probabilidad de que el paciente sufra una complicación infecciosa. De nuevo, se cotejan los resultados del histograma con la tabla de predicciones finales:

Predicción	Nº de elementos
Complicación	150
No complicación	269

Cuadro 4: Comparación del número de elementos para cada predicción

Tal como se refleja en el histograma de la figura 27, la predicción dice que el número de pacientes que sufrirán una complicación de índole no infecciosa es mucho mayor que en el caso contrario, con un total de 150 pacientes, conformando el 35, 79 % del conjunto de test.

A continuación se procede con el análisis de la curva ROC para el modelo de predicción de complicaciones de índole no infecciosa.

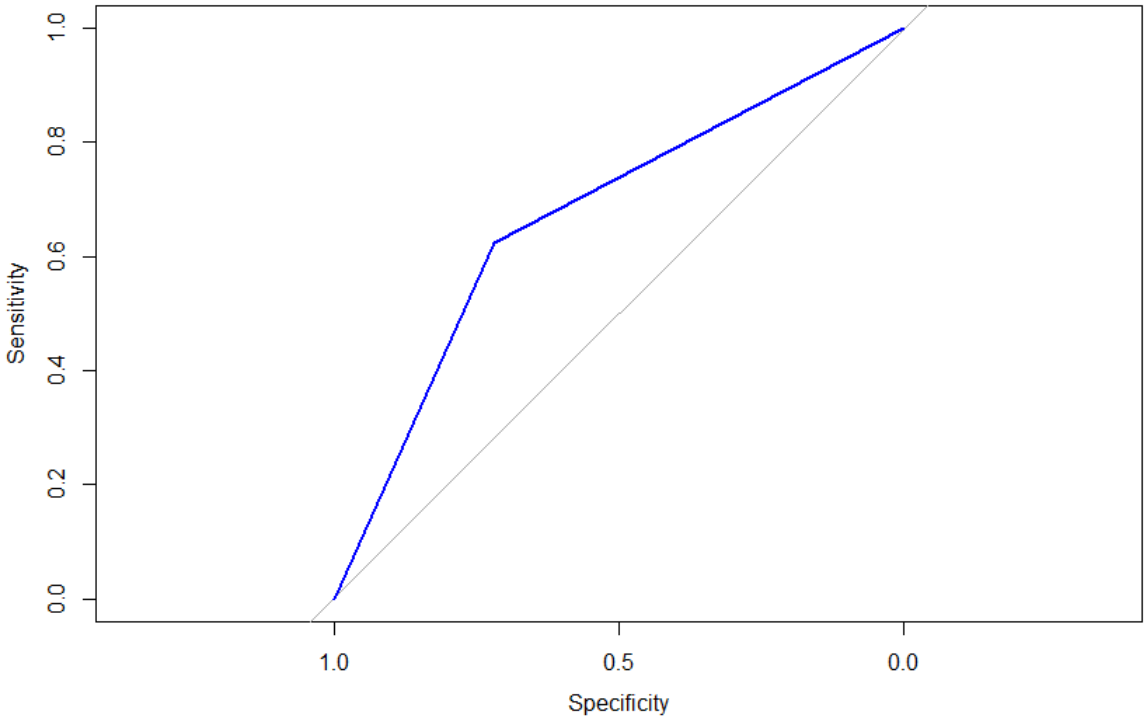


Figura 28: Curva ROC para modelo de complicación no infecciosa

El área bajo la curva en este caso es 0.672, valor que, tal como se esperaba, es menor que en el caso anterior, no llegando a considerarse como un test bueno sino como un test regular (ya que su valor se encuentra entre 0.6 y 0.75). Por último, se realiza la validación del modelo mediante Leave One Out:

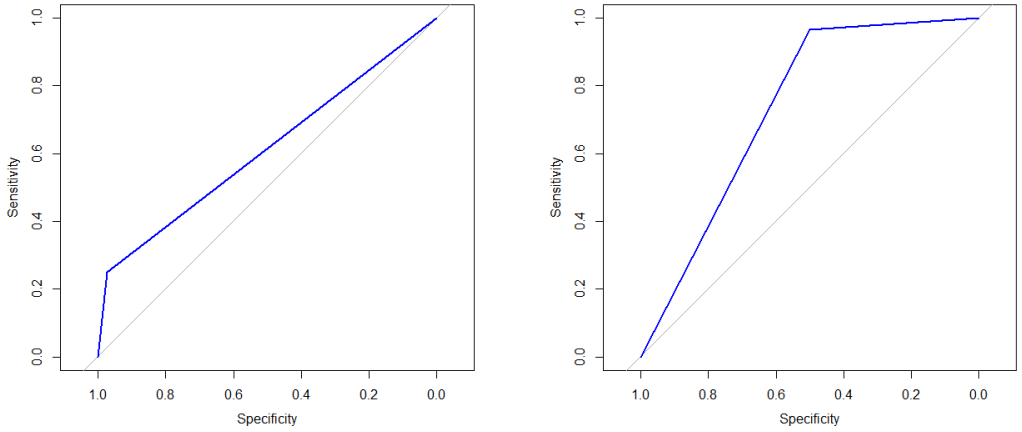
Precisión	0.686
Error	0.314

Cuadro 5: Valores de preción y error tras usar validación cruzada

Comparando los resultados, se obtiene que el modelo para la predicción de complicaciones de índole no infecciosas es menos preciso y fiable que en el otro caso.

7.5. Modelo de regresión logística sobre el conjunto de clústeres

En los apartados anteriores, se ha estudiado el modelo de regresión logística para todo el conjunto de datos proporcionado. En este apartado se va a realizar un estudio muy similar, esta vez analizando el modelo clúster por clúster, según la división realizada antes.
Destacar que en este estudio solo se va a trabajar el modelo para tratar de predecir la existencia de complicaciones de índole infecciosa, por lo que en este caso no es necesario un parámetro que indique el tipo de complicación a estudiar, como se hizo en el caso anterior.
Durante el análisis del modelo para cada uno de los clúster surge un inconveniente, en ciertos casos el tamaño de clúster es demasiado pequeño y el algoritmo no llega a converger correctamente. En estos casos no se aprecia la curva ROC, siendo la probabilidad de acierto del 50 %. Por ello, solo se representarán mediante una curva ROC aquellos clústeres en los que dicha curva sea significativa.



(a) Clúster 3 (b) Clúster 4

Figura 29: Comparación de las curvas ROC entre clústeres

Como se ha mencionado antes, el tamaño de los clústeres hace que el algoritmo no converja en ciertos casos, como son los clústeres 1 y 2. Si se calcula el AUC para estos dos casos se obtiene que para cada predicción del modelo hay un 50 % de probabilidades de que se corresponda con su valor real, es decir, misma probabilidad que si se lanzase una moneda al aire, lo que hace que para estos dos clústeres el modelo no sea válido.
Respecto a los clústeres 3 y 4, se observa que la la figura muestra una mayor curvatura que en los casos anteriores, por lo que el área debajo de la curva es mayor, siendo por tanto más preciso el modelo.

Clúster	AUC
1	0.5
2	0.5
3	0.612
4	0.732

Cuadro 6: Comparación del AUC entre los clústeres

Como se refleja en las curvas ROC, el área bajo la curva de ambos modelos es superior a la de los dos clústeres anteriores. En el caso del clúster 3, se obtiene un valor de 0.612, que si bien es superior al de los anteriores clústeres, no se puede considerar como un modelo bueno.
En el caso del clúster 4, se observa un valor mayor, pudiéndolo considerar como un test bueno, aunque lejos de ser excelente.

8. Conclusiones

Tal como se comentó al principio del documento, el objetivo principal del proyecto era la aplicación de un modelo de regresión logística para un conjunto de datos mixto, así como su división en clústeres. Al ser el conjunto de datos utilizado de este tipo, la división en clústeres se ha realizado utilizando el algoritmo K-means, y tomando como métrica de distancia la Gower.
Se ha demostrado que el modelo de regresión logística planteado funciona bien para predecir la aparición de complicaciones de índole infecciosa sobre el conjunto de datos en su totalidad, mientras que funciona peor a la hora de predecir complicaciones de índole no infecciosa. Esto se ha demostrado representando las predicciones mediante una curva ROC y calculando el área bajo la curva de la misma.
Se han comparado los resultados del modelo propuesto utilizando el modelo k-means con la implementación que proporciona el propio entorno de programación del algoritmo PAM (Partitioning around medoids), algoritmo de clustering basado en la distancia. Comparando los clústeres de ambos algoritmos, se observa que son muy similares, con diferencias mínimas en ciertas variables, por lo que es indiferente usar un algoritmo u otro.
Por último, se ha aplicado el modelo de predicción sobre cada uno de los clústeres creados con el algoritmo K-means. Analizando las curvas ROC y el área bajo la curva de cada uno de ellos se observa que el modelo no es bueno en la mayorá de los casos, siendo la probabilidad de acierto del 50 %. Esto ocurre debido al tamaño de los clústeres, el cual en estos casos es demasiado pequeño y hace que el algoritmo no llegue a converger como debiera.

Referencias

S. Black, I. Kushner, and D. Samols. C-reactive protein. *THE JOURNAL OF BIOLOGICAL CHEMISTRY*, 279(47):48487–48490, Agosto 2004. https://www.sciencedirect.com/science/article/pii/S0021925819322288?ref=pdf_download&fr=RR-2&rr=825ea06bf96a6675.

O. A. C. C and I. V. M. Cirugía robótica. *Revista chilena de cirugía*, (1):88–91, 2012. https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-40262012000100016.

I. L. de Ullibarri Galparsoro and S. P. Fernández. Curvas roc. *Unidad de Epidemiología Clínica y Bioestadística. Complexo Hospitalario Juan Canalejo*, pages 4–5, 2001. http://webpersonal.uma.es/~jmpaez/websci/BLOQUEI/DocuI/roc.pdf.

R. C. Esper and A. A. P. Calatayud. Procalcitonina como marcador de procesos infecciosos en cirugía. conceptos actuales. *Unidad de Terapia Intensiva. Fundación Clínica Médica Sur*, Julio 2013. https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-00992013000100009.

B. S. Everitt, S. Landau, M. Leese, and D. Stahl. Cluster analysis. *Wiley Series in probability and statistics*, pages 54–56, 2011. https://cicerocq.files.wordpress.com/2019/05/cluster-analysis_5ed_everitt.pdf.

C. ME, P. P. A. KL, and M. CR. A new method of clssifying prognostic comorbidity in longitudinal studies. *J Chronic Dis*, 40(5):373–383, 1987. https://pubmed.ncbi.nlm.nih.gov/3558716/.

C. ME, C. RE, and P. JC. The charlson comorbidity index is adapted to predict costs of chronic disease in primare care patients. *J Clin Epidemiol*, 61(12):1234–1240, 2008. https://pubmed.ncbi.nlm.nih.gov/18619805/.

J. M. Pérez and P. P. Martín. La curva roc. *SEMERGEN. Medicina de Familia*, (49), 2022. https://static.elsevier.es/ficheros/7.pdf.

J. A. Rodrigo. Validación de modelos predictivos: Cross-validation, oneleaveout, bootstraping. *Ciencia de datos*, Noviembre 2020. https://cienciadedatos.net/documentos/30_cross-validation_oneleaveout_bootstrap.

J. C. Slieker, F. Daams, I. M. Mulder, J. Jeekel, and J. F. Lange. Systematic review of the technique of colorectal anastomosis. *American Medical Association*, 2013. https://jamanetwork.com/journals/jamasurgery/article-abstract/1654856.

O. WD, F. JA, and S. E. Jr. Asa physical status classifications: a study of ratings. *Anesthesiology*, 49(4):239–243, October 1978. https://pubmed.ncbi.nlm.nih.gov/697077/.

J. Álvarez Fernández-Represa, J. de Diego Carmona, E. O. Oshiro, and J. M. Martínez. Cirugía laparoscópica. *Elsevier*, 68(4):304–308, 2000. https://tinyurl.com/yx28f88f.