Heart Failure Clinical Records

Maximiliano Arévalo Sáez and Benjamín Muñoz Tapia

Universidad de Santiago de Chile

Abstract. La base de datos correspondiente a Heart Failure Clinical Records contiene información de registros clínicos realizados a pacientes que han tenido insuficiencia cardíaca, los cuales se recopilaron durante el tiempo de seguimiento de estos. Con el objetivo de verificar la incidencia de estas características en la salud del paciente, y si es posible predecir si el paciente sobrevivirá a una falla cardíaca a partir de los valores de las mediciones. En esta experiencia se utiliza el método de agrupamiento basado en modelos, para obtener información y conocimiento de los datos a través del método.

Keywords: Agrupamiento · incidencia

1 Introducción

Actualmente un 31% de las personas que muere padecen una enfermedad cardiovascular, las que también se ven afectadas por otros factores como la diabetes, anemia, hipertensión, entre otros. Para un estudio de esta problemática, se elaboró una base de datos con el registro de 299 pacientes, 12 factores, y su estado (vivo o muerto)[1]. Entre las variables presentadas, las cuales son de caracter numérico y categórico en cuanto a los antecedentes del paciente, por lo que para la experiencia actual se tienen los siguientes objetivos:

- Analizar la base de datos Heart Failure Clinical Records y sus variables.
- Aplicar el método de agrupamiento basado en modelos para obtener conocimiento.

2 Métodos y datos

2.1 Métodos utilizados

El método a aplicar es el Método de agrupación basado en modelos, el cual además de distribuir los datos en grupos, supone un modelo estadístico en base a cada grupo. Para aplicar este método hay variados modelos, por lo que se debe utilizar el Criterio de Información Bayesiano (BIC) que entrega un ranking de los mejores modelos con la cantidad de grupos a realizar.

Sin embargo, antes de realizar el agrupamiento, se deben revisar los datos disponibles y ver cuales son más relevantes para el estudio a través de un análisis estadístico y una revisión de la literatura.

2.2 Datos utilizados

El conjunto de datos contiene los registros clínicos de 299 pacientes que han tenido insuficiencia cardíaca. Originalmente fueron recopilados por Tanvir Ahmad, Assia Munir, Sajjad Haider Bhatti, Muhammad Aftab y Muhammad Ali Raza del Government College University, Faisalabad, Pakistan. La versión actual del dataset fue elaborada por Davide Chicco del Krembil Research Institute, Toronto, Canadá.
El conjunto de datos contiene 299 instancias con 13 atributos y la clase, la cual indica si el paciente sobrevivió o no durante el período de seguimiento. La 4 muestra la descripción de los atributos del conjunto de datos.

Table 1. Definición del conjunto de datos.

Variable	**Descripción**	**Medida**
Age	Edad del paciente	Años
Anemia	Disminución de glóbulos rojos o hemoglobina	Booleano
High blood pressure	Si el paciente tiene hipertensión	Booleano
Creatinine phosphokinase (CPK)	Nivel de enzima CPK en la sangre	mcg/L
Diabetes	Si el paciente tiene diabetes	Booleano
Ejection fraction	Porcentaje de sangre que sale del corazón en cada contracción	Porcentaje
Sex	Mujer u hombre	Binario
Platelets	Plaquetas en la sangre	kiloplaquetas/mL
Serum creatinine	Nivel de creatinina en la sangre	mg/dL
Serum sodium	Nivel de sodio en la sangre	mEq/L
Smoking	Si el paciente fuma	Booleano
Time	Período de seguimiento	Días
Death event	Si el paciente murió durante el período de seguimiento	Booleano

3 Resultados

Luego de realizar el análisis estadístico y correlaciones correspondientes (ver anexo), se dejaron las 6 variables más significativas para el dataset, omitiendo las variables categóricas también. En cuanto a la cantidad de grupos y modelo adecuado, el indicador BIC muestra que el mejor modelo es el "VVI" con 7 grupos. A continuación se muestra el resultado del indicador BIC para el dataset inicial, el BIC después de haber hecho el análisis estadístico y reducción de variables, y finalmente el agrupamiento resultante.

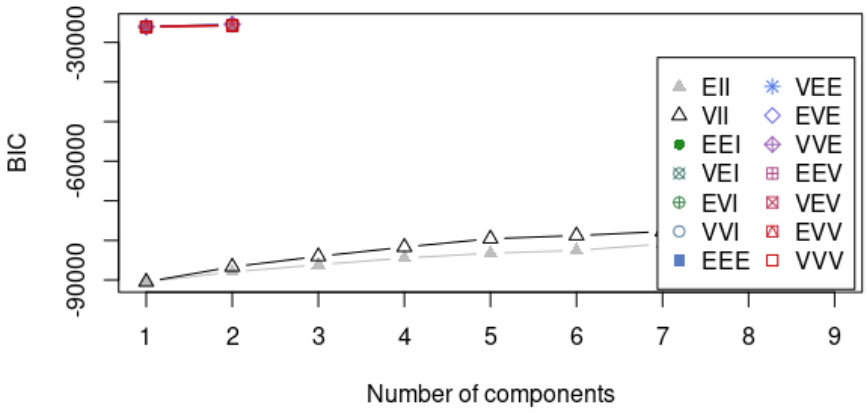

Fig. 1. BIC inicial

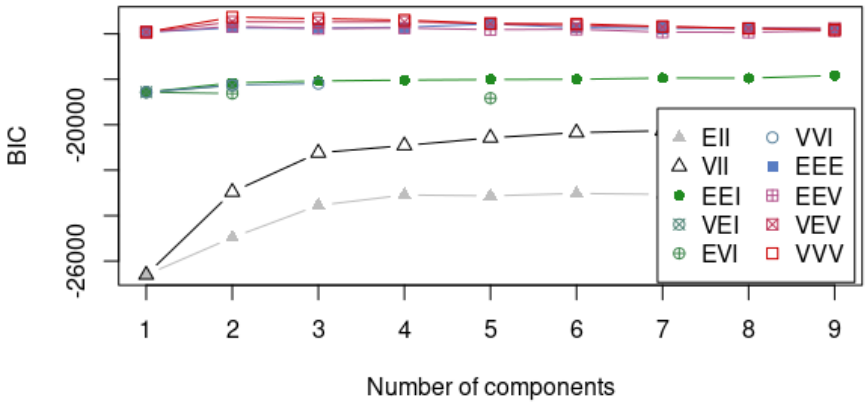

Fig. 2. BIC para 6 variables

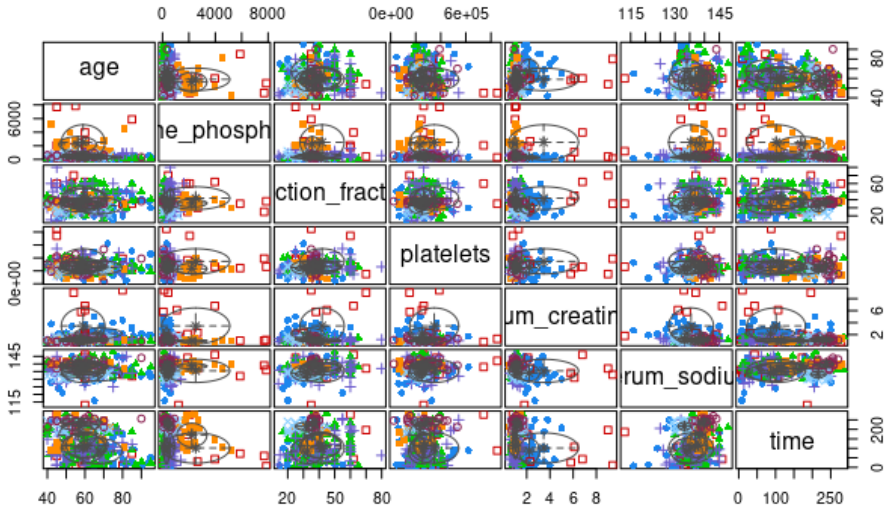

Fig. 3. Agrupamiento resultante

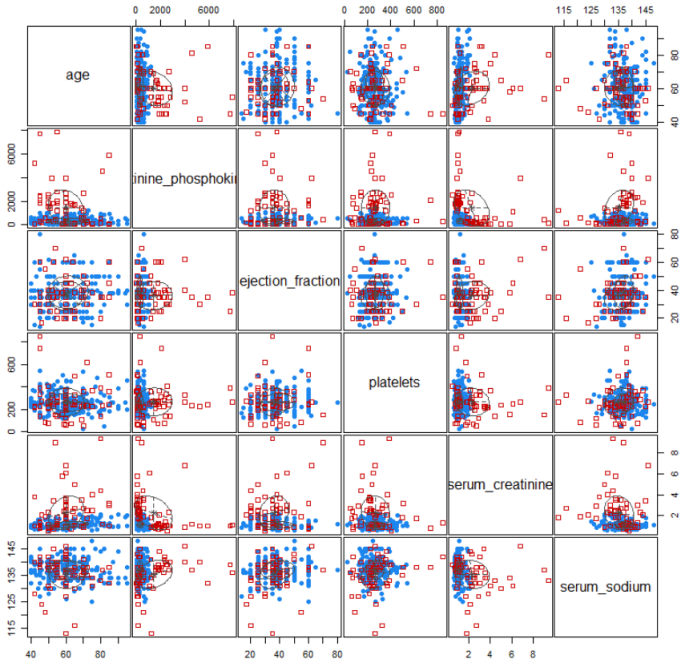

Fig. 4. Segundo agrupamiento resultante

4 Discusión

Al realizar el análisis estadístico y reducir la cantidad de variables se logra ver cuales influyen más en el comportamiento de los datos al momento de agrupar. En cuanto al criterio de agrupamiento, se obtuvo un BIC de -15265.46, el cual inicialmente era de -25406.61, con los modelos VVV y VVE con dos y siete grupos respectivamente. El BIC mejoró cuando se dejó de considerar el tiempo de atención al paciente, dejando finalmente el análisis para 6 variables.

En cuanto al resultado de agrupamiento, al tener dos grupos se puede inferir que hay una buena asociación de los datos, la cual resulta en 234 de un grupo y 65 del otro, mientras que los datos originales cuentan con 203 pacientes vivos y 96 muertos. Finalmente, si se compara con el método de K-medias, hay una gran diferencia, ya que se tuvieron 4 grupos resultantes que puede ser debido al modelo con el que se realizó este agrupamiento, dando a ver que el modelo VVV fue más preciso.

5 Conclusiones

A través de la experiencia se pudo realizar un estudio del problema, junto con un análisis estadísticos de las variables involucradas, tanto categóricas como numéricas. Gracias a dicho análisis se pudieron considerar las variables más relevantes y aplicar el criterio BIC y posteriormente realizar un agrupamiento por modelos, viendo que el mejor modelo para este caso es el VVI con dos grupos y un BIC de -15256.46.

Por otro lado, se hace un agrupamiento también con el método k-medias para comparar resultados, viendo que agrupamiento con modelos es más preciso. Además, al no presentar un gran número de grupos, se puede decir que el modelo es menos complejo como en el caso de k-medias con 4 grupos. Es importante destacar que los resultados logrados no podrían haber sido obtenidos sin haber hecho el análisis estadístico a las variables correspondientes, para lograr entender la naturaleza del problema y del dataset como tal.

References

1. Ahmad T., Munir A., Haider Bhatti S., Aftab M., Ali Raza M.(2017). Survival analysis of heart failure patients: A case study. Plos One. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0181001
2. Chicco, D., Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5.
3. UCI Heart failure clinical records Data Set, https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records.

6 Anexo

6.1 Clasificación de variables

Para realizar un análisis estadístico, es necesario identificar los tipos de variable que se encuentran en el dataset, para aplicar diferentes métodos que permitan obtener información de la distribución de los datos dentro de este. El dataset actual cuenta con variables numéricas y variables categóricas, las cuales se agrupan como muestra la siguiente tabla:

Table 2. Clasificación de variables.

Numéricas	**Categóricas**
Age	Anaemia
Creatinine phosphokinase	Diabetes
Ejection fraction	High blood pressure
Platelets	Sex
Serum creatinine	Smoking
Serum sodium	

6.2 Análisis de las variables

Se construyen gráficos para verificar la relación entre las observaciones de cada variable, correspondiente a una característica médica, y el estado de vida del paciente luego de finalizar el período de seguimiento. Para el caso de las variables numéricas se elaboran gráficos de densidad para realizar la comparación, en los cuales se puede visualizar el comportamiento de cada característica médica en relación a si el paciente sobrevive o no, donde se identifica que pareciera ser que las que entregan mayor cantidad de información corresponden a la fracción de ejección y serum creatinine, en base a los gráficos obtenidos:

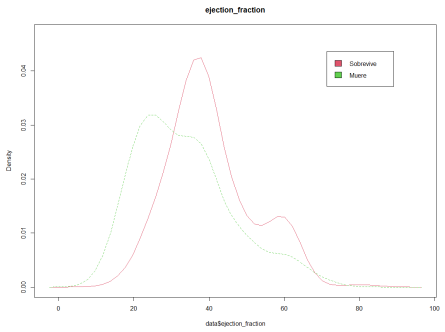

Fig. 5. Gráfico fracción de ejección y muerte

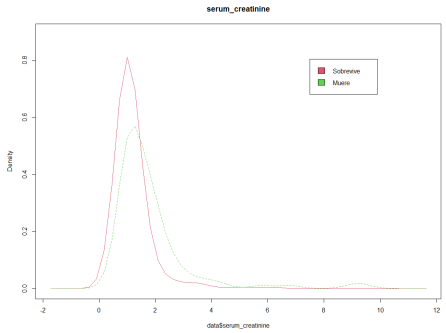

Fig. 6. Gráfico serum creatinine y muerte

Para el caso de las variables categóricas se elaboran gráficos de barras para realizar la comparación, los cuales son útiles para identificar la cantidad de pacientes en base a las características médicas del tipo categórico, sobreviven o no a lo largo del período de seguimiento. A modo de ejemplo se muestran a continuación los gráficos correspondientes para la anemia y la diabetes.

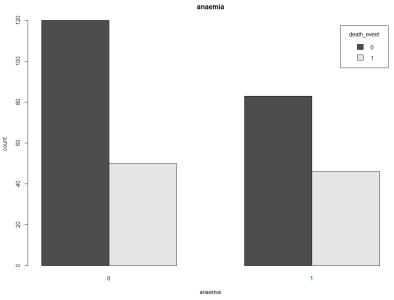

Fig. 7. Gráfico anemia y muerte

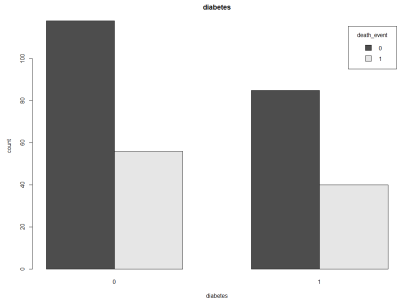

Fig. 8. Gráfico diabetes y muerte

6.3 Matriz de correlación

Se elabora una matriz de correlación de Pearson, para identificar la relación lineal entre cada una de las variables, obteniendo los resultados indicados en la siguiente tabla:

	age	*anaemia*	*CPK*	*diabetes*	*ejection fraction*	*high blood pressure*	*platelets*	*serum creatinine*	*serum sodium*	*sex*	*smoking*
age	1.00	0.09	-0.08	-0.10	0.06	0.09	-0.05	0.16	-0.05	0.07	0.02
anaemia	0.09	1.00	-0.19	-0.01	0.03	0.04	-0.04	0.05	0.04	-0.09	-0.11
CPK	-0.08	-0.19	1.00	-0.01	-0.04	-0.07	0.02	-0.02	0.06	0.08	0.00
diabetes	-0.10	-0.01	-0.01	1.00	0.00	-0.01	0.09	-0.05	-0.09	-0.16	-0.15
ejection fraction	0.06	0.03	-0.04	0.00	1.00	0.02	0.07	-0.01	0.18	-0.15	-0.07
high blood pressure	0.09	0.04	-0.07	-0.01	0.02	1.00	0.05	0.00	0.04	-0.10	-0.06
platelets	-0.05	-0.04	0.02	0.09	0.07	0.05	1.00	-0.04	0.06	-0.13	0.03
serum creatinine	0.16	0.05	-0.02	-0.05	-0.01	0.00	-0.04	1.00	-0.19	0.01	-0.03
serum sodium	-0.05	0.04	0.06	-0.09	0.18	0.04	0.06	-0.19	1.00	-0.03	0.00
sex	0.07	-0.09	0.08	-0.16	-0.15	-0.10	-0.13	0.01	-0.03	1.00	0.45
smoking	0.02	-0.11	0.00	-0.15	-0.07	-0.06	0.03	-0.03	0.00	0.45	1.00

Fig. 9. Matriz de correlación de Pearson

Se puede apreciar que las características no están correlacionadas entre sí, sin embargo, pareciera que sex y smoking indican la presencia de una correlación ligeramente positiva.

6.4 Prueba de Shapiro-Wilk

Para estudiar la distribución de las variables se utiliza el test de normalidad de Shapiro-Wilk para verificar si las variables numéricas siguen una distribucion normal o no. Los resultados indican que las variables no se distribuyen de manera normal, ya que al obtener p-valores muy pequeños, no existe suficiente evidencia estadística que permita rechazar esta hipótesis, esto al considerar un valor hipotético de alfa igual a 0.05.

6.5 Valores obtenidos para el BIC

Inicialmente al considerar las variables del dataset estipulada en la sección de resultados, se obtuvo la siguiente configuración para el BIC:

Table 3. Valores de BIC I.

Configuración	**N° Grupos**	**BIC**
VVI	7	-22626.63
VVI	4	-22652.78
VVE	4	-22692.86

Luego de realizar las modificaciones indicadas, se volvió a obtener los valores para la nueva configuración del agrupamiento basado en los nuevos valores de BIC recomendandos, los cuales son los siguientes:

Table 4. Valores de BIC I.

Configuración	**N° Grupos**	**BIC**
VVV	2	-15265.46
VVV	3	-15331.99
VVV	4	-15396.11

Finalmente, se puede mencionar que al considerar determinadas variables que tengan mayor relación con lo que se desea estudiar, es posible acotar las que son requeridas y aportan información útil para el estudio. Lo que se traduce en un menor valor del BIC, en comparación a otras posibles consideraciones de variables.