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 Random Forest con el fin de obtener información acerca de los pacientes que mueren y los que sobreviven a la deficiencia cardiaca.

Keywords: Agrupamiento · Clasificación

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 Random Forest para obtener un clasificador que determine si un paciente sobrevivirá o no ante una incidencia cardiaca
- Aplicar para el apoyo de Random Forest Coordenadas Paralelas, Mtry, entre otros.

2 Métodos y datos

2.1 Métodos utilizados

El método a aplicar es el Método de Bosques Aleatorios (Random Forest) es una técnica que combina Árboles de Decisión mediante Bagging, donde cada subelemento es un árbol de decisión que considera solamente algunas variables. Al momento de clasificar, se elige la clase mayor votada por los árboles de decisión del bosque presente, clasificación que se puede medir con el error Out-Of-Bag (OOB).

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 2 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

Inicialmente se selecciona un conjunto de datos que contiene las variables numéricas correspondientes a *Age, Creatinine Phosphokinase, Ejection Fraction, Platelets, Serum Creatinine, Serum Sodium* y la clase *Death Event*.
Ya que estas características entregan mayor cantidad de información al momento de realizar el análisis, debido a los diferentes niveles y valores asociados a las variables.

Se aplica por primera vez el método de Random Forest con las variables mencionadas en el párrafo anterior, generando un conjunto de entrenamiento en base a la totalidad existente de datos, obteniendo como resultado el siguiente modelo:

```
Call:
 randomForest(x = data3[training.ids, 1:6], y = data3[training.ids,      7], ntree = 500, keep.forest = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 27.49%
Confusion matrix:
    0  1 class.error
0 120 23   0.1608392
1  35 33   0.5147059
```

Fig. 1. Primer Random Forest obtenido

Del cual se puede identificar la existencia de un error Out Of Bound de 27.49%, y al analizar la matriz de confusión obtenida, se puede apreciar que para el caso en el que los pacientes sobreviven se tiene un error de 16.08% aproximadamente, clasificando correctamente 120 pacientes y fallando en 23 pacientes. Mientras que para el caso de los pacientes que fallecen, se tiene un error de 51.48% aproximadamente, clasificando correctamente a 35 pacientes y fallando en 33 pacientes. Este nivel tan alto de error se debe al desbalanceo existente en el dataset *Heart Failure Clinical Records*, ya que dentro de los datos existen un total de 206 pacientes que sobreviven, mientras para el caso de los que fallecen se tiene un total de 93. Por lo que la clase tendrá un mayor momento al clasificar a este tipo de pacientes al contar con menos datos que aporten información a los Random Forest generados, obteniendo la siguiente curva ROC para este caso:

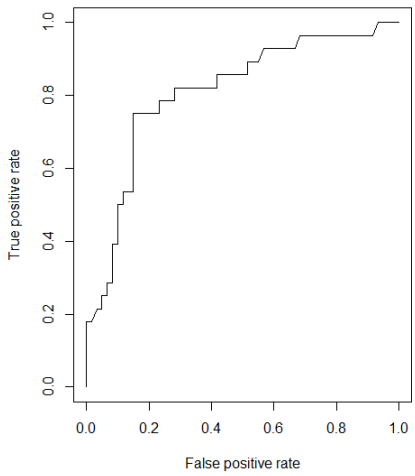

Fig. 2. Curva ROC del Random Forest obtenido

De la curva anterior se puede apreciar que si bien la curva ROC muestra resultados positivos, debido a la tendencia que tiene y el área bajo la curva obtenida. Existe un error de más del 50% para clasificar a los pacientes que fallecen, en base a una serie de Random Forest generados para identificar la variación de este porcentaje. Por lo que se piensa que este error se compensa en la curva ROC por la gran cantidad de pacientes que sobreviven clasificados correctamente, cuya cantidad es mucho mayor.

Para mejorar los resultados del método se realiza un balanceo del dataset eliminando un total de 70 pacientes que sobreviven, cuyo criterio se basó en los que tenían características más similares para algunas características de las indicadas anteriormente para el estudio. Contando así con 133 pacientes que sobreviven y 96 pacientes que fallecen, se realiza un serie de Random Forest para identificar los resultados obtenidos, identificando una reducción en el error OOB al obtener un valor de 25.66%. Mientras que para el caso de la matriz de confusión se obtuvo un error de 21.05% aproximadamente para la clasificación de pacientes que sobreviven, y un error de 32,3% aproximadamente para los pacientes que fallecen, obteniendo mejores resultados que para el caso del Random Forest con el dataset sin balanceo.

```
Call:
 randomForest(formula = DEATH_EVENT ~ ., data = data3, importance = TRUE,      proximity = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 25.66%
Confusion matrix:
    0  1 class.error
0 105 28   0.2105263
1  30 63   0.3225806
```

Fig. 3. Segundo Random Forest obtenido

También se obtiene la importancia del Random Forest de la figura 4, para realizar un gráfico que permita identificar cuales son las variables que aportan más al proceso de clasificación, obteniendo como resultado las características correspondientes a *Age, Ejection Fraction* y *Serum Creatinine*, las cuales se conservan para la siguiente implementación.

Al realizar todas las consideraciones comentadas anteriormente, se obtiene un tercera implementación de Random Forest, obteniendo un OOB de 22,12%, un error de clasificación para los pacientes sobrevivientes de un 18,8% aproximadamente, y para el caso de los pacientes que fallecen un error de un 26,89% aproximadamente.

```
Call:
 randomForest(formula = DEATH_EVENT ~ ., data = importantData,      ntree = 500, importance = TRUE, proximity = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 1

        OOB estimate of  error rate: 22.12%
Confusion matrix:
    0  1 class.error
0 108 25   0.1879699
1  25 68   0.2688172
```

Fig. 4. Tercer Random Forest obtenido

Finalmente, se implementan dos nuevos Random Forest realizando variaciones en el parámetro *mtry*, donde para el primero caso se considera un valor correspondiente a 2, obteniendo un OOB de 23,89% con un error de 19,55% aproximadamente para pacientes que sobreviven y un 30% de error aproximado para pacientes que fallecen. Mientras que para un *mtry* igual a 3 se obtiene un ligero aumento del OOB a un 25,22%, con un error de 20,3% aproximadamente para pacientes que sobreviven y un 32,3% para pacientes que fallecen.

4 Discusión

En el primer Random Forest se ve la presencia de un alto porcentaje de error que ronda el 50% para clasificación de pacientes fallecidos, lo cual es producido por el desbalanceo existente en el dataset, aunque de igual manera se desarrolló un clasificador teniendo éxito en determinados casos. Como resultado de esta aplicación se intenta reducir el error OOB, y el error de clasificación para cada caso, se realiza una eliminación de los pacientes que sobreviven en base a los que más se parecen para no afectar en gran medida la clasificación realizada, obteniendo una reducción en el OBB, y una reducción en el porcentaje de error para el caso de los pacientes que fallecen aproximadamente en un 20% para este caso, y un 5% para el OOB.

Para obtener mejores resultados se realiza un análisis de importancia de las variables, donde se pudieron eliminar algunas que no aportaban tanta información al proceso de clasificación, las cuales corresponden a *Age, Ejection Fraction* y *Serum Creatinine*, coincidiendo con los papers encontrados en la literatura. Por lo que al balancear el dataset y realizar el Random Forest considerando estas variables más importantes, se pueden verificar mejores resultados al reducir el porcentaje de error (ver figura 14, Anexo).

Al variar el parámetro *mtry* se pudo identificar que al asignarle el valor de 2, se obtienen mejores resultados para el caso del valor 3, obteniendo un menor error para el OOB en el primer caso, junto con un menor porcentaje de error para la clasificación de los pacientes que sobreviven y fallecen (ver figuras 19 y 20, Anexo).

En cuanto al escalamiento dimensional (ver figuras 16 y 17, Anexo), al balancear el Dataset se apercia que los datos están bastante dispersos y no se ve una tendencia clara entre ambos grupos. Mientras que con las curvas paralelas (ver figura 18, Anexo) se ve como se comportan las clases con las distintas variables, siendo la clase de personas que fallecen más dispersa, mientras que la clase de los pacientes que viven se ven más agrupadas para cada variable.

5 Conclusiones

En la primera aplicación del método Random Forest se tiene un error OOB 27.49%, pero para la clasificación de pacientes que mueren se tiene un error de un 50% el cual es elevado, por lo que se necesitó hacer un balance al Dataset con el fin de mejorar el clasificador.

Después de realizar el balance de los datos y hacer un análisis de importancia y proximidad, se puede ver cuales eran las variables más significativas al momento de clasificar con el método, siendo estas: *Age, Ejection Fraction y Serum Creatinine*. Aún así, el error de clasificación sigue siendo alto para los pacientes que mueren, permaneciendo cerca del 30%, pero el seguir alterando el Dataset puede significar un sobreajuste que puede afectar a los resultados. El escalamiento multidimensional por su parte, permite ver como están de dispersos los datos ante dos variables, y para este caso, se encontraban bastante dispersas ambas clases, que también se corrobora con coordenadas paralelas.

Finalmente se puede ver una correcta aplicación de Random Forest y se cumplen los objetivos del trabajo. Se ve que este método es bastante sensible al balanceo de los datos, por lo que se deben tener clases más equitativas si se quiere tener mejores resultados de clasificación.

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:

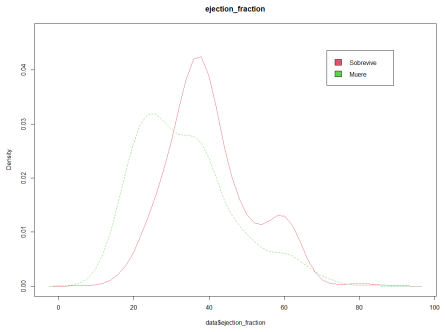

Fig. 5. Gráfico fracción de ejección y muerte

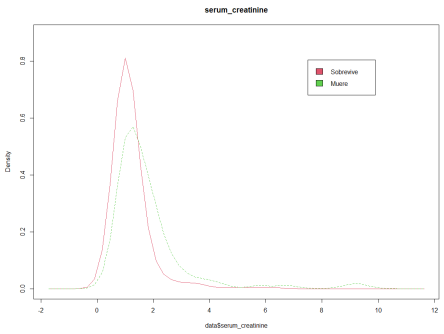

Fig. 6. Gráfico serum creatinine y muerte

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.

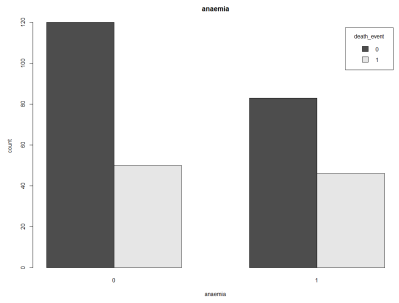

Fig. 7. Gráfico anemia y muerte

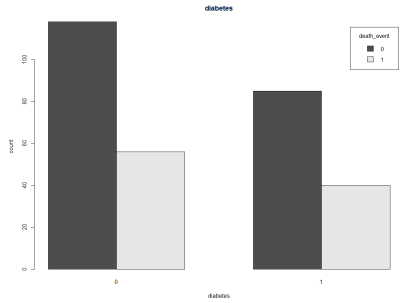

Fig. 8. Gráfico diabetes y muerte

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 Gráficos

```
Call:
 randomForest(x = data3[training.ids, 1:6], y = data3[training.ids,      7], ntree = 500, keep.forest = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 27.49%
Confusion matrix:
    0  1 class.error
0 120 23   0.1608392
1  35 33   0.5147059
```

Fig. 10. Primer Random Forest obtenido

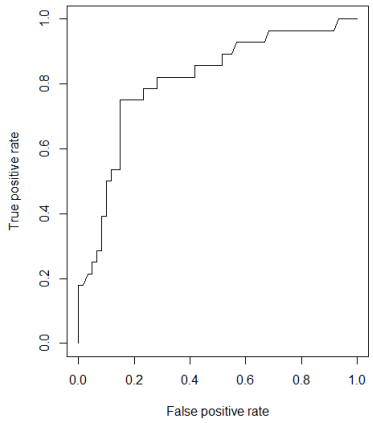

Fig. 11. Curva ROC Primer Random Forest obtenido

```
Call:
 randomForest(formula = DEATH_EVENT ~ ., data = data3, importance = TRUE,      proximity = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 25.66%
Confusion matrix:
    0  1 class.error
0 105 28   0.2105263
1  30 63   0.3225806
```

Fig. 12. Segundo Random Forest obtenido

```
Call:
 randomForest(formula = DEATH_EVENT ~ ., data = importantData,      ntree = 500, importance = TRUE, proximity = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 1

        OOB estimate of  error rate: 22.12%
Confusion matrix:
    0  1 class.error
0 108 25   0.1879699
1  25 68   0.2688172
```

Fig. 13. Tercer Random Forest obtenido

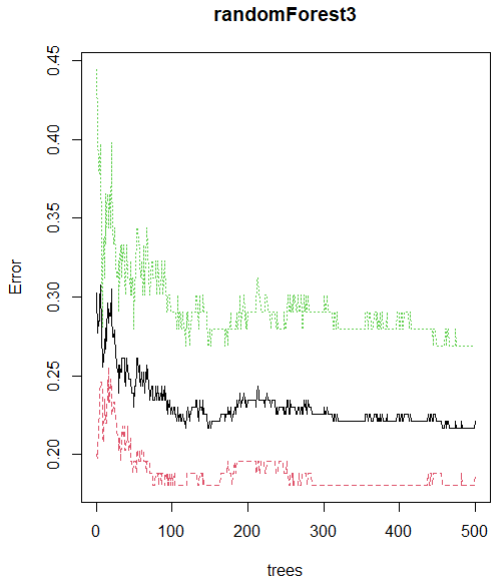

Fig. 14. Rendimiento Tercer Random Forest obtenido

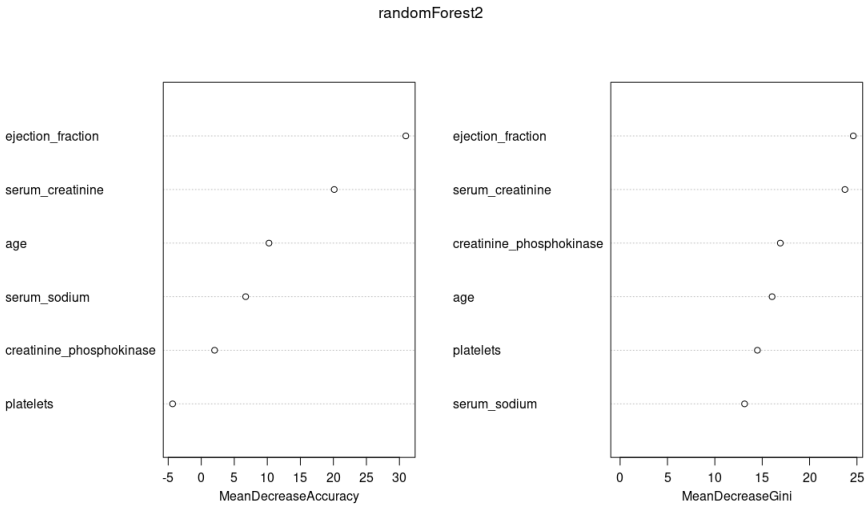

Fig. 15. Importancia de las Variables

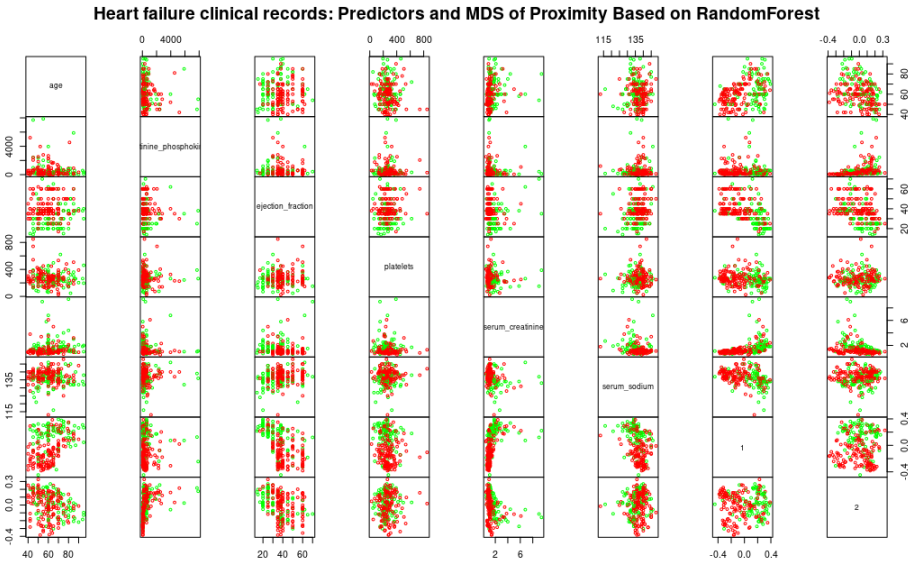

Fig. 16. Escalamiento Multidimensional

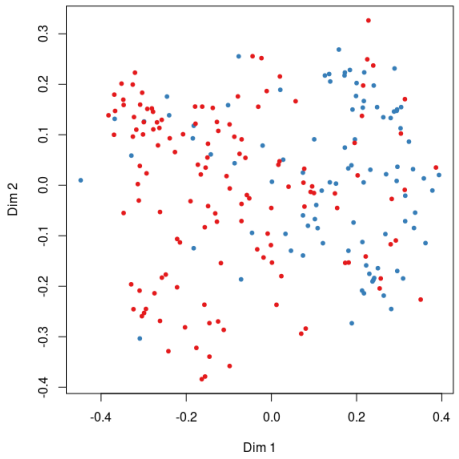

Fig. 17. Escalamiento Multidimensional de la clase

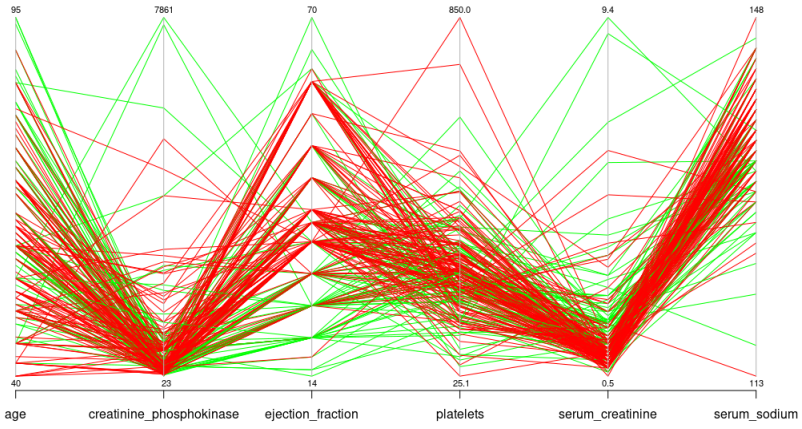

Fig. 18. Coordenadas paralelas

```
Call:
 randomForest(formula = DEATH_EVENT ~ ., data = importantData,      ntree = 500, mtry = 2, importance = TRUE, proximity = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 23.89%
Confusion matrix:
    0  1 class.error
0 107 26   0.1954887
1  28 65   0.3010753
```

Fig. 19. Random Forest con mtry = 2

```
Call:
 randomForest(formula = DEATH_EVENT ~ ., data = importantData,      ntree = 500, mtry = 3, importance = TRUE, proximity = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 25.22%
Confusion matrix:
    0  1 class.error
0 106 27   0.2030075
1  30 63   0.3225806
```

Fig. 20. Random Forest con mtry = 3