Machine Learning in Genomics - Proyecto Final

Victoria Lelis & Renata Sandoval

Contents

1 Descarga de datos y carga de librerías

```r
rm(list=ls())
library(DataExplorer)
library(skimr)
library(ggplot2)
library(GGally)
library(tidyverse)
library(tidymodels)
library(corrplot)
library(dplyr)
library(nnet)
library(yardstick)
library(viridis)
library(vip)
library(recount3)
library(tidyr)
library(GSVA)
library(GSVAdata)
library(msigdbr)
library(SummarizedExperiment)
library(pheatmap)
library(patchwork)
theme_set(theme_minimal(base_size = 8))
```

Ocupamos la librería `recount3 ()` para hacer la descarga de nuestros datos, ocupamos el proyecto **SRP026208**.

1.1 Reducción de dimensionalidad del dataset con GSVA

Dado que nuestro data frame es demasiado extenso para los análisis posteriores, reducimos su dimensionalidad utilizando GSVA (Gene Set Variation Analysis). Esta herramienta permite asociar los datos de expresión génica con diversos **pathways biológicos** y asignar un puntaje que refleja la actividad del conjunto de genes en cada ruta.

2 Parte 1: Selección del conjunto de datos y análisis inicial

2.1 Documentación del conjunto de datos

2.1.1 - Describa detalladamente el conjunto de datos genómicos seleccionado:

2.1.1.1 ¿Cuál es la fuente y el número de acceso (si procede)?c El accession number del dataset utilizado para este proyecto es GSE48166. Aunque este dataset no cuenta con un artículo asociado, es posible acceder a los datos directamente mediante la herramienta `recount3`. Puedes acceder a los datos del experimento en el siguiente link: GEO-GSE48166.

2.1.1.2 ¿De qué tipo de datos se trata (RNA-seq, ChIP-seq, microarrays, etc.)?
El dataset contiene datos de **RNA-seq**, específicamente, **expr_matrix** incluye los perfiles de expresión génica de cada paciente involucrado en este estudio.

2.1.1.3 ¿Cuántas muestras y características contiene?

En este dataset tenemos un total de 29 **biological pathways** y una variable de la condición del paciente (columnas), y en el caso de las muestras tenemos 30 pacientes (columnas), de los cuales 15 pacientes presentan **miocardiopatía isquémica**, y los otros 15 pacientes son del grupo control (personas saludables).

```
## Número de columnas: 30
## Número de filas: 30
```

2.1.1.4 ¿Cuál es la cuestión biológica que se aborda?

La insuficiencia cardíaca es un problema grave que afecta a millones de personas en todo el mundo. Esta enfermedad puede clasificarse en diferentes tipos según sus causas, las cuales suelen ser variadas y complejas de identificar. Comprender los diversos tipos de insuficiencia cardíaca podría aportar información valiosa para desarrollar tratamientos más especializados, mejorando así el pronóstico de los pacientes. Siendo un primer acercamiento en este proyecto el detectar si los pacientes presentan **miocardiopatía isquémica**, un tipo de insuficiencia cardíaca.

2.2 Análisis Exploratorio de Datos

2.2.1 - Realice y documente un AED exhaustivo:

2.2.1.1 Generar e interpretar gráficos de distribución de características clave:

Los plots nos dan información sobre cómo se distribuyen los datos de cada una de las variables (genes) de nuestro dataset. Para realizar estos análisis ocupamos:

```
#create_report(final_data)
```

En este caso podemos observar cómo están dsitribuidos los valores de cada una de los pathways biológicos.

De la misma exploración obtuvimos un Análisis de Componentes Principales realizado por la misma función `create_report()`. Esta revisión nos permitió darnos una idea de cómo esperamos que se organizen los datos; una opción sería comparar nuestros resultados con este plot para identificar posibles diferencias entre ambos.

Por otro lado, `skim ()` nos da información como el número de columnas (genes) y filas (pacientes); al igual que el tipo de varibles, en su mayoría numéricas, excepto `condition`. También nos da si es que hay valores faltantes (los cuales no hay), el promedio de cada variable, su desviación estándar, entre otros datos.

```
# skim(final_data)
```

2.2.1.2 Analiza potenciales batch effects o artefactos técnicos

Para analizar posibles batch effects o artefactos técnicos es bueno primero comenzar con la evaluación de la variabilidad de los datos, por medio de métodos como el Principal Component Analysis (PCA), que se realizará más adelante. Como vimos durante clase, el PCA nos permite identificar patrones en los datos que pueden no estar relacionados con las condiciones biológicas, sino que se trata de artefactos técnicos, como podría ser variaciones en el procesamiento que se les dio a las muestras.

2.2.1.3 Evaluar la calidad de los datos y los valores que faltan.

Al evaluar los datos faltantes de nuestro dataset, obtuvimos el siguiente plot, igualmente proviene de la función `create_report()`. Como podemos observar, no tenemos ningún valor faltante, lo que es un muy

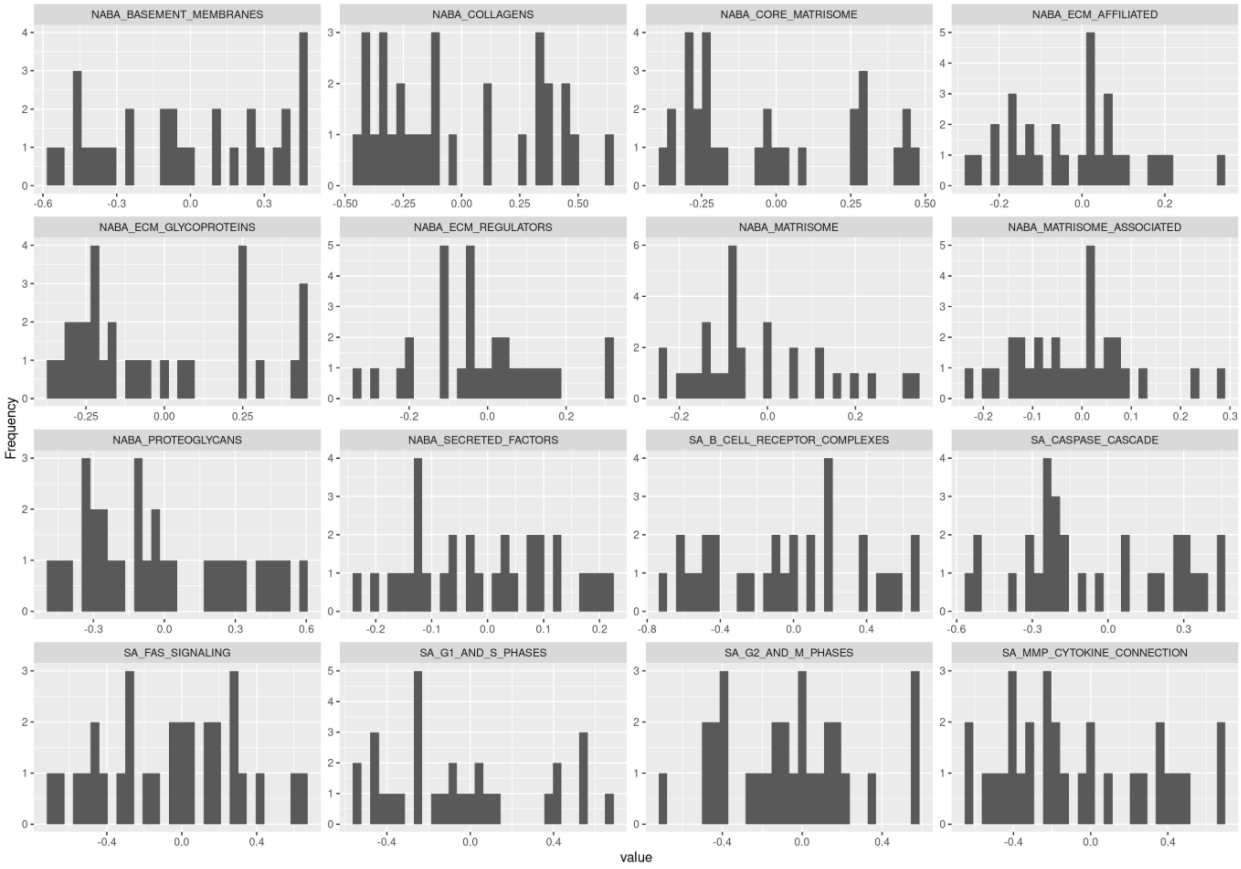


Figure 1: Histograma 1 de la distribución de los datos de los genes

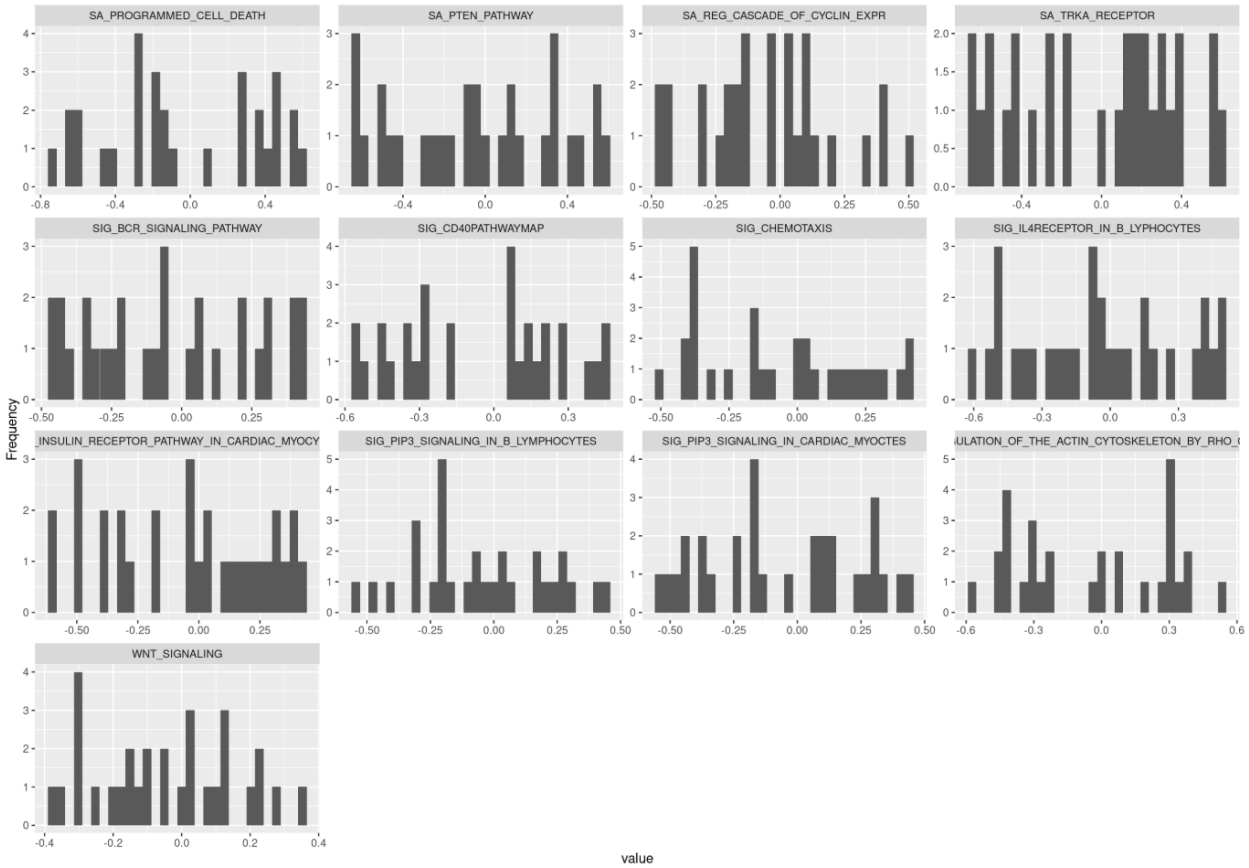


Figure 2: Histograma 2 de la distribución de los datos de los genes

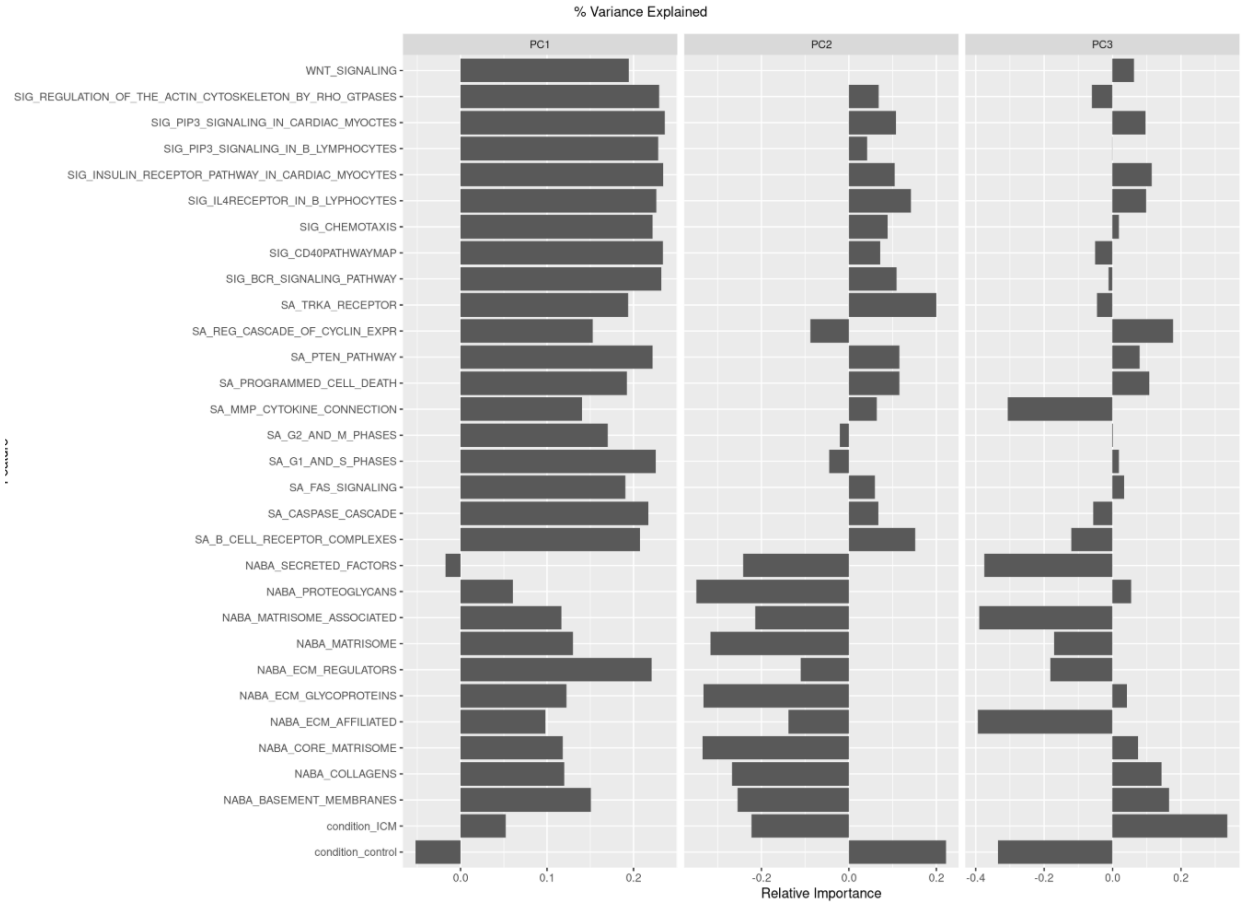
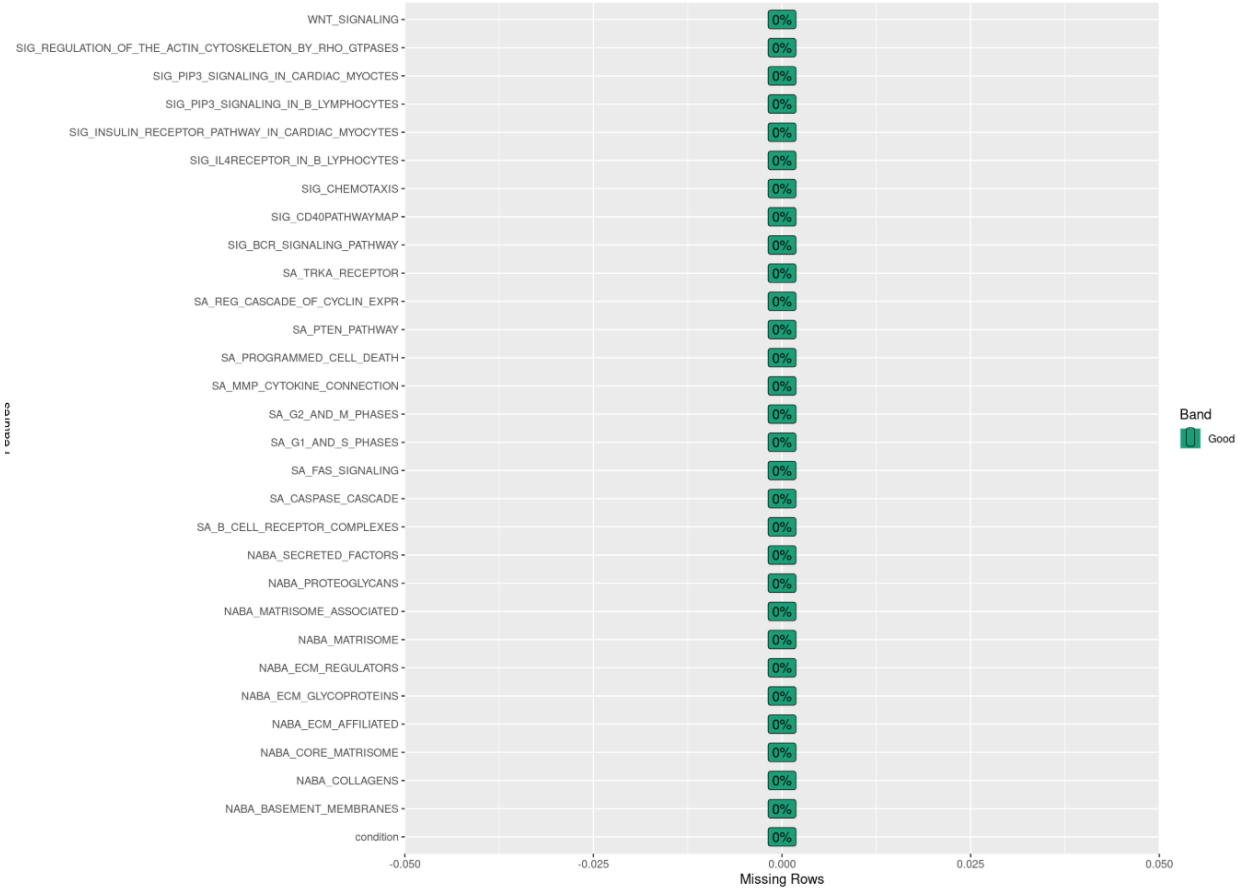


Figure 3: PCA realizado con DataExplorer

buen indicador de la calidad de nuestros datos.



2.2.1.4 Visualizar las relaciones entre características

En la siguiente matriz tenemos una vista general de las relaciones que hay entre las variables (genes). Observamos una barra de color que nos indica el tipo de correlación, si es roja, indica que hay una correlación positiva (valores cercanos a 1). Cuando se tiene un color azul, indica una correlación negativa (valores cercanos a -1); finalmente cuando se obseva de un color blanco, señala que no se detectó ninguna correlación. A pesar de que en el eje X no podemos diferenciar de qué pathway biológico estamos hablando, en una vista general podemos observar que la mayoría de las regiones son de color rojo, es decir, presentan correlaciones positivas.

2.3 Análisis de reducción de dimensionalidad

2.3.1 Realizar el análisis de componentes principales

```
## # A tibble: 841 x 4
##    terms                      value component id
##    <chr>                      <dbl> <chr>     <chr>
##  1 NABA_BASEMENT_MEMBRANES   -0.147  PC1       pca
##  2 NABA_COLLAGENS            -0.116  PC1       pca
##  3 NABA_CORE_MATRISOME       -0.113  PC1       pca
##  4 NABA_ECM_AFFILIATED       -0.0990 PC1       pca
##  5 NABA_ECM_GLYCOPROTEINS    -0.118  PC1       pca
##  6 NABA_ECM_REGULATORS       -0.220  PC1       pca
```

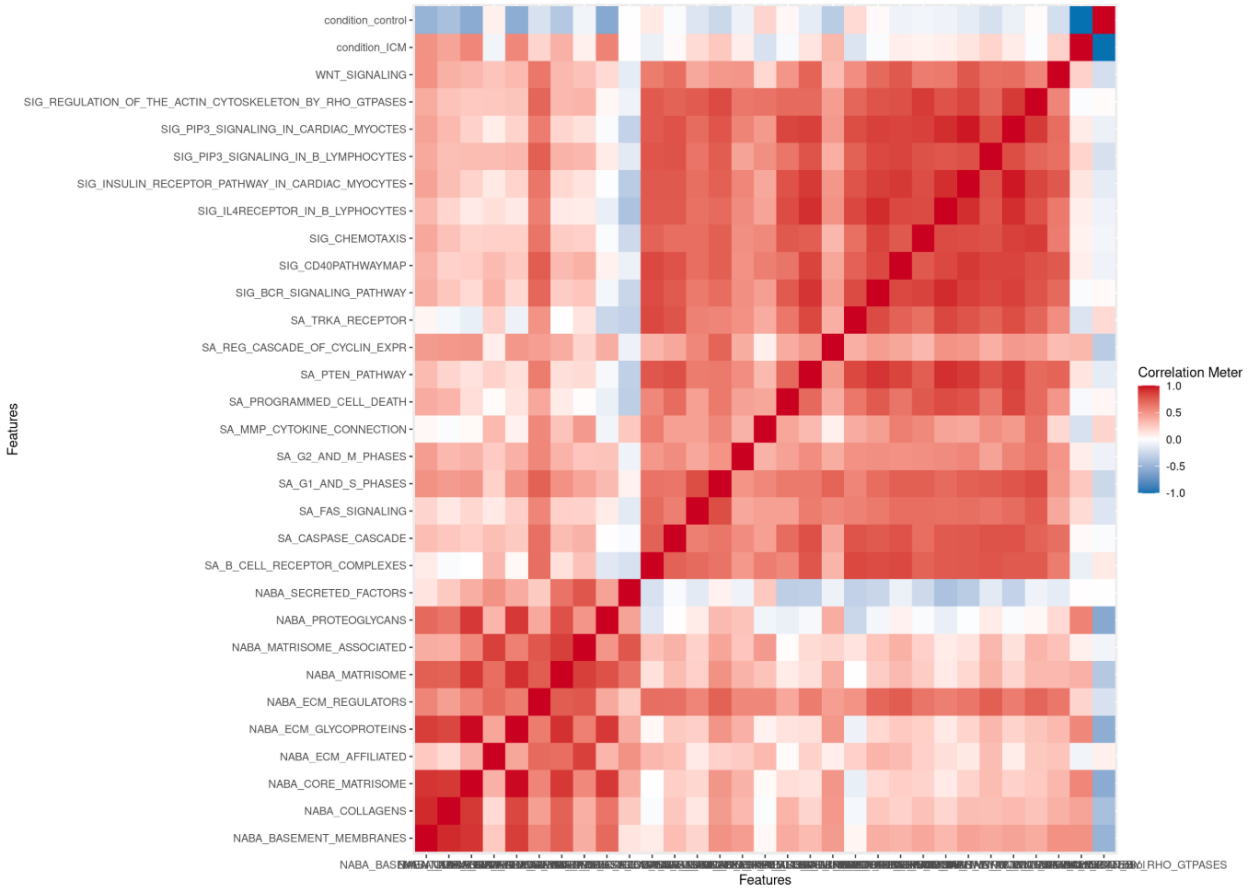
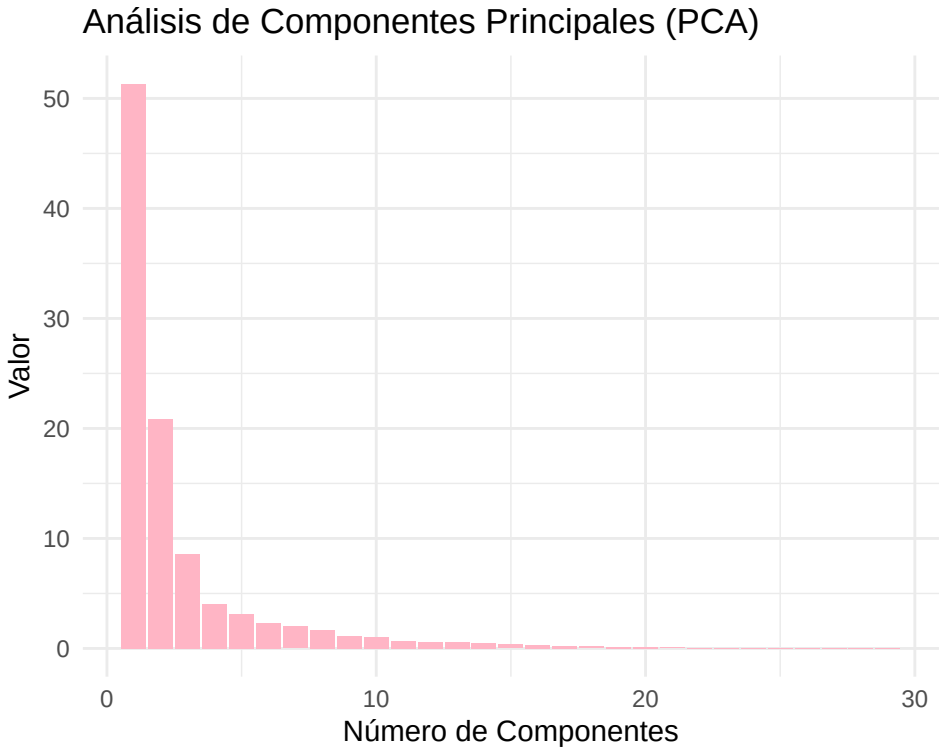


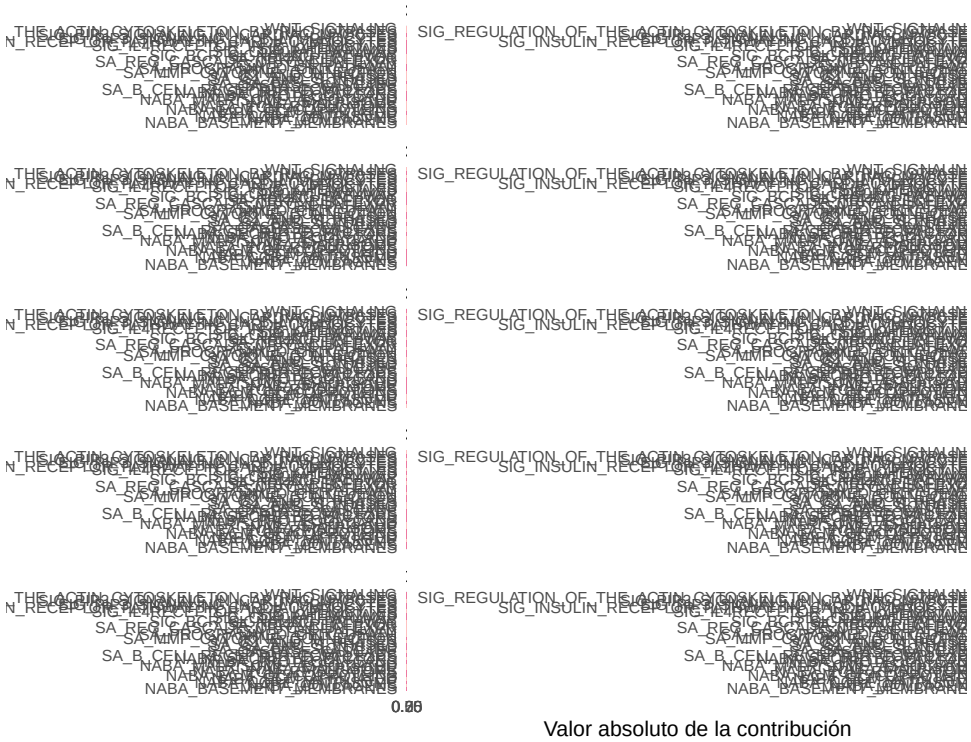
Figure 4: Matriz de correlacion

```
##  7 NABA_MATRISOME            -0.127  PC1       pca
##  8 NABA_MATRISOME_ASSOCIATED -0.116  PC1       pca
##  9 NABA_PROTEOGLYCANS        -0.0548 PC1       pca
## 10 NABA_SECRETED_FACTORS      0.0188 PC1       pca
## # i 831 more rows
```

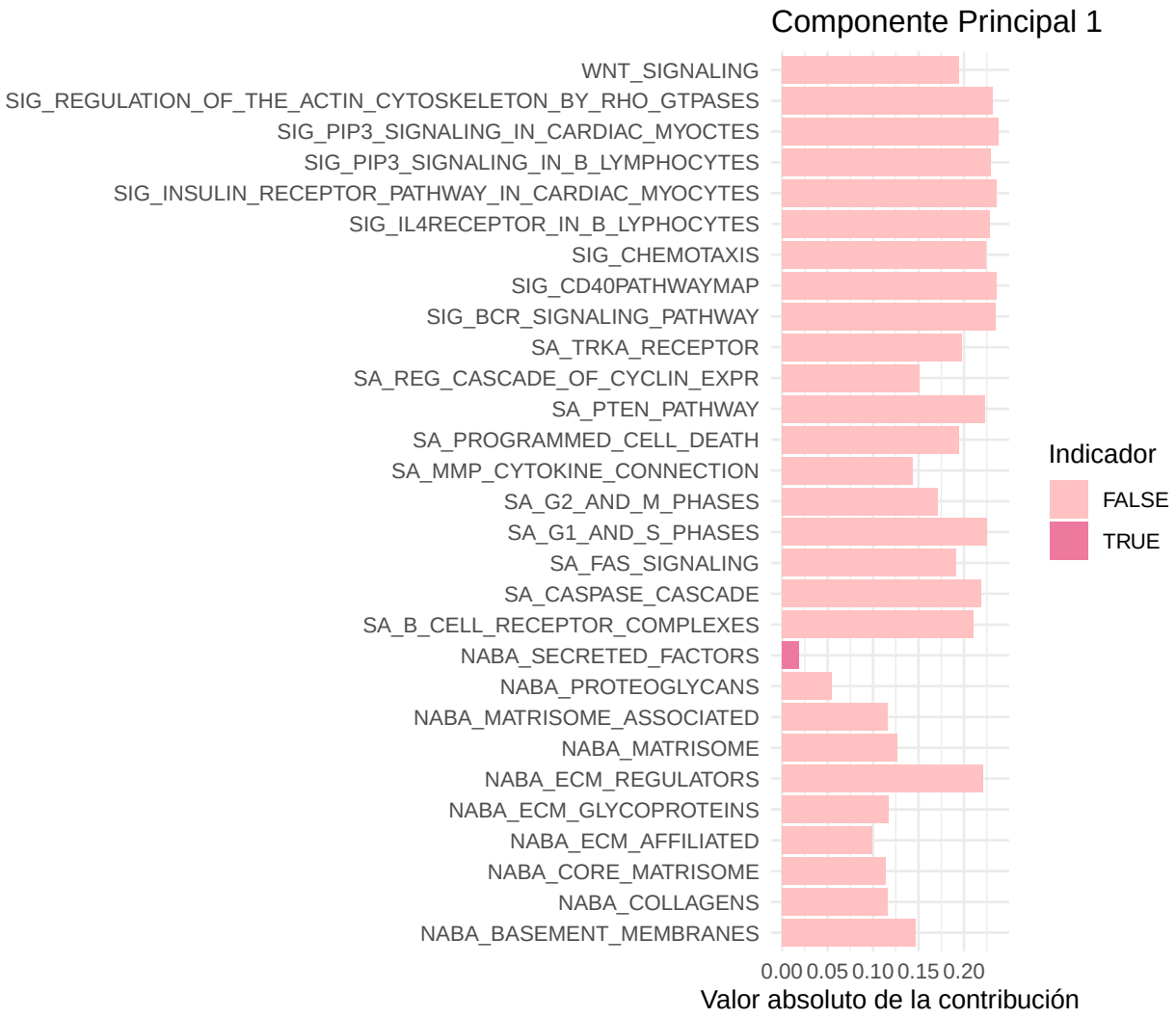
2.3.1.1 Creación de un "scree plot" para observar la proporción de varianza explicada por cada componente principal.

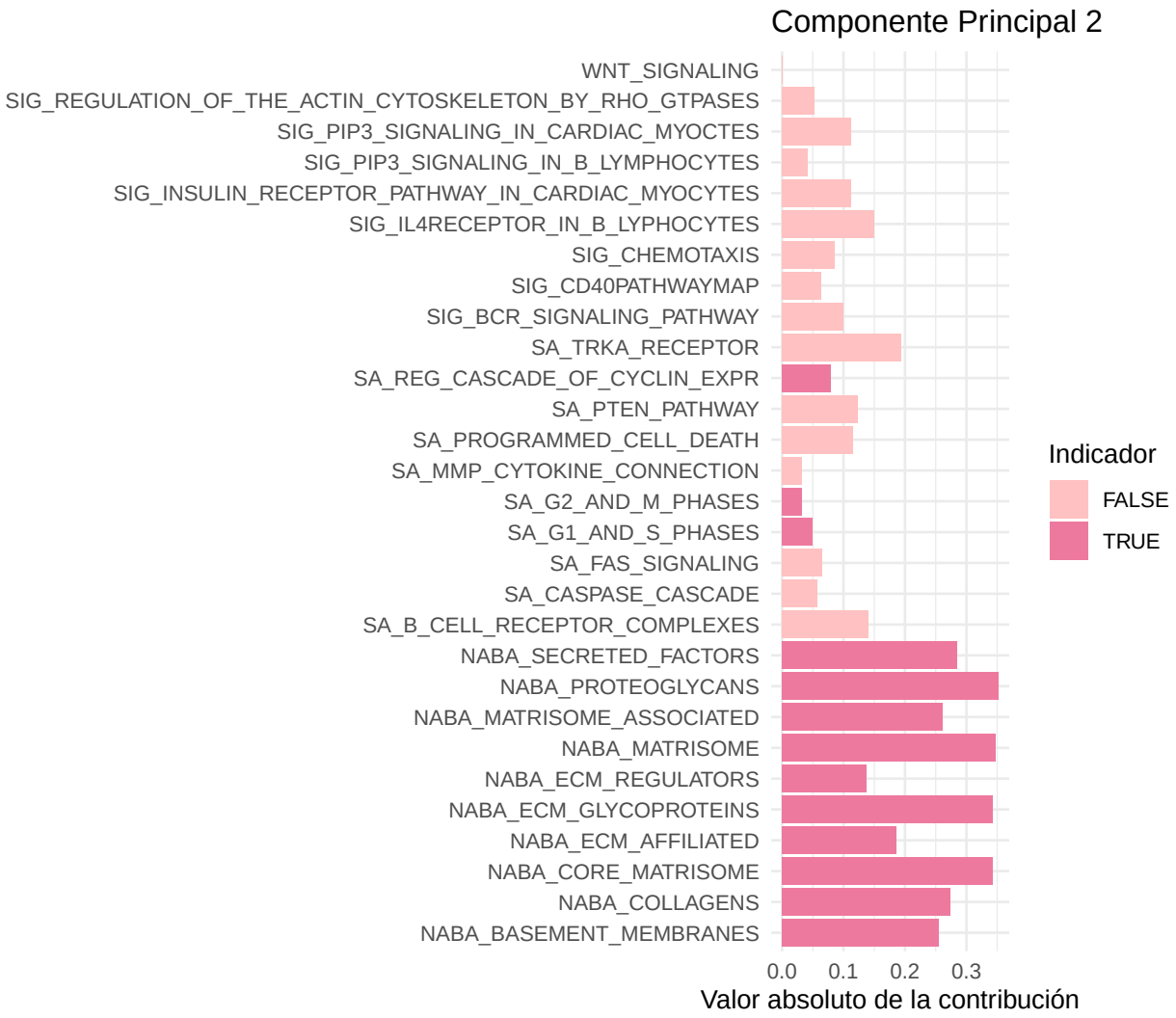


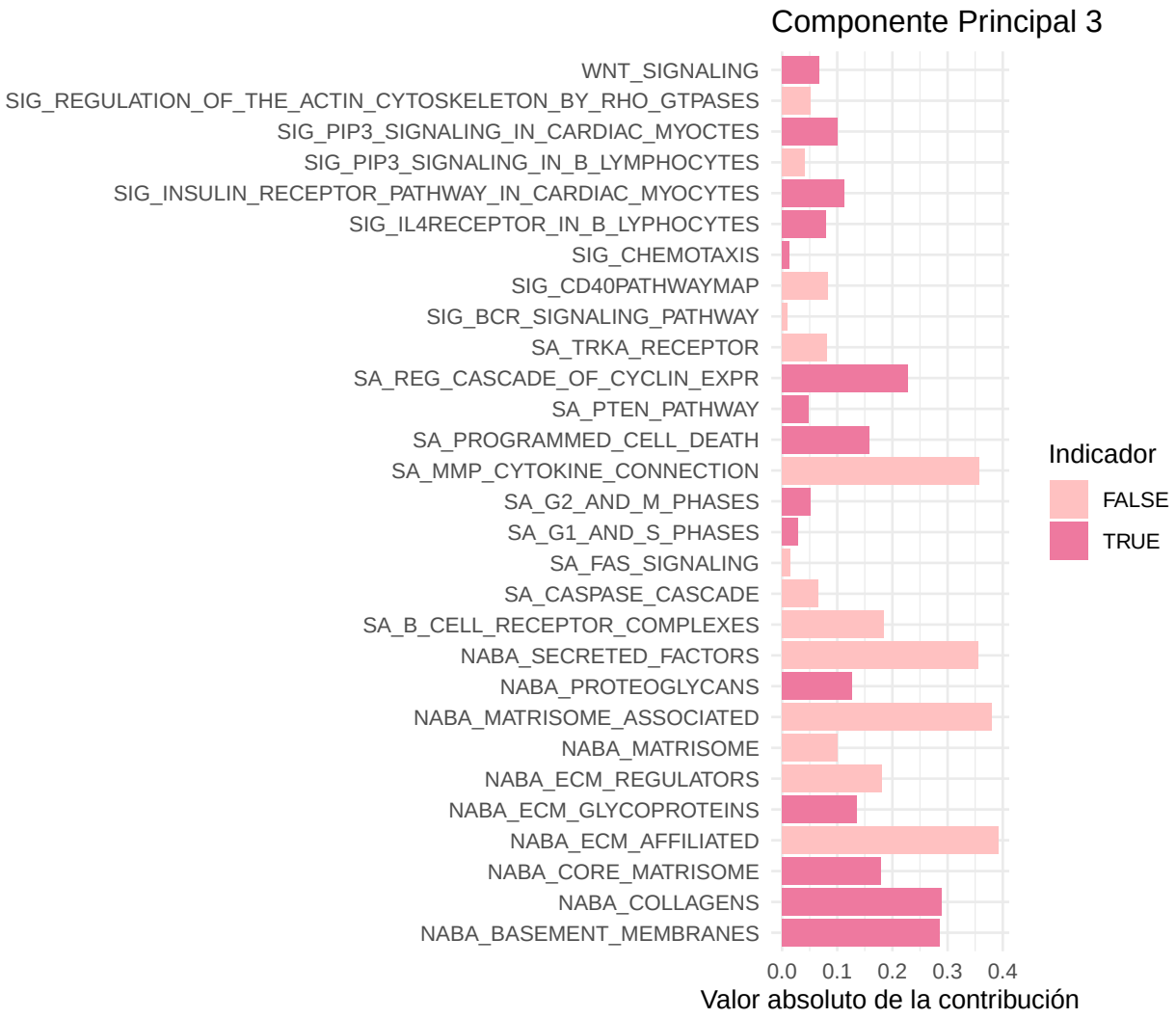
2.3.1.2 Visualizar la contribución de cada variable en cada PC.

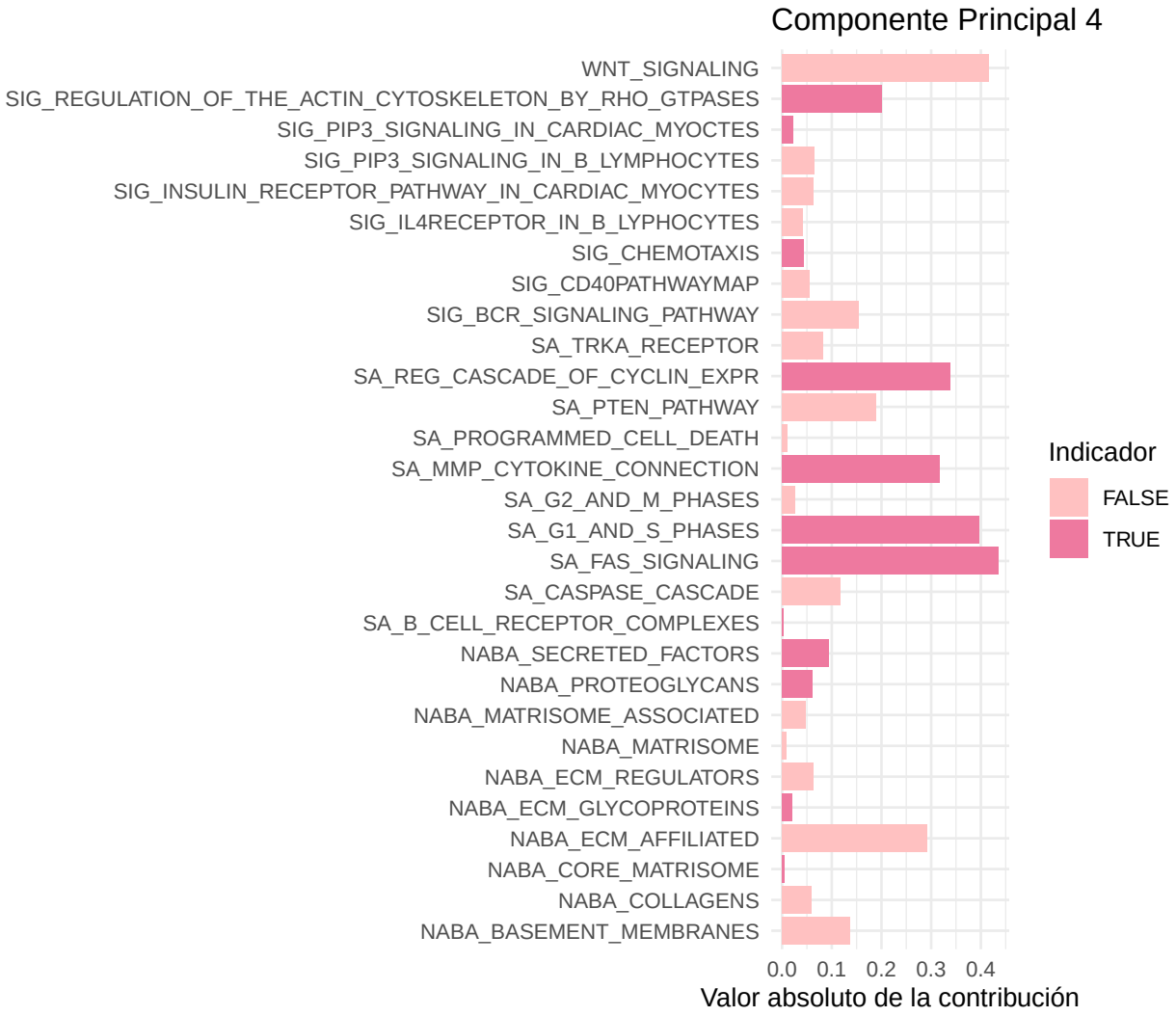


Dado que es poco posible visualizar todos los PC, se seleccionó los primeros 5 componentes principales, ya que entre estos parece estar aproximadamente el 90% de la variabilidad total de los datos (90% de la información total):


Componente Principal 1
WNT_SIGNALING
SIG_REGULATION_OF_THE_ACTIN_CYTOSKELETON_BY_RHO_GTPASES
SIG_PIP3_SIGNALING_IN_CARDIAC_MYOCTES
SIG_PIP3_SIGNALING_IN_B_LYMPHOCYTES
SIG_INSULIN_RECEPTOR_PATHWAY_IN_CARDIAC_MYOCYTES
SIG_IL4RECEPTOR_IN_B_LYPHOCYTES
SIG_CHEMOTAXIS
SIG_CD40PATHWAYMAP
SIG_BCR_SIGNALING_PATHWAY
SA_TRKA_RECEPTOR
SA_REG_CASCADE_OF_CYCLIN_EXPR
SA_PTEN_PATHWAY
SA_PROGRAMMED_CELL_DEATH
SA_MMP_CYTOKINE_CONNECTION
SA_G2_AND_M_PHASES
SA_G1_AND_S_PHASES
SA_FAS_SIGNALING
SA_CASPASE_CASCADE
SA_B_CELL_RECEPTOR_COMPLEXES
NABA_SECRETED_FACTORS
NABA_PROTEOGLYCANS
NABA_MATRISOME_ASSOCIATED
NABA_MATRISOME
NABA_ECM_REGULATORS
NABA_ECM_GLYCOPROTEINS
NABA_ECM_AFFILIATED
NABA_CORE_MATRISOME
NABA_COLLAGENS
NABA_BASEMENT_MEMBRANES
0.00 0.05 0.10 0.15 0.20
Valor absoluto de la contribución
Indicador
FALSE
TRUE


Componente Principal 2
WNT_SIGNALING
SIG_REGULATION_OF_THE_ACTIN_CYTOSKELETON_BY_RHO_GTPASES
SIG_PIP3_SIGNALING_IN_CARDIAC_MYOCTES
SIG_PIP3_SIGNALING_IN_B_LYMPHOCYTES
SIG_INSULIN_RECEPTOR_PATHWAY_IN_CARDIAC_MYOCYTES
SIG_IL4RECEPTOR_IN_B_LYPHOCYTES
SIG_CHEMOTAXIS
SIG_CD40PATHWAYMAP
SIG_BCR_SIGNALING_PATHWAY
SA_TRKA_RECEPTOR
SA_REG_CASCADE_OF_CYCLIN_EXPR
SA_PTEN_PATHWAY
SA_PROGRAMMED_CELL_DEATH
SA_MMP_CYTOKINE_CONNECTION
SA_G2_AND_M_PHASES
SA_G1_AND_S_PHASES
SA_FAS_SIGNALING
SA_CASPASE_CASCADE
SA_B_CELL_RECEPTOR_COMPLEXES
NABA_SECRETED_FACTORS
NABA_PROTEOGLYCANS
NABA_MATRISOME_ASSOCIATED
NABA_MATRISOME
NABA_ECM_REGULATORS
NABA_ECM_GLYCOPROTEINS
NABA_ECM_AFFILIATED
NABA_CORE_MATRISOME
NABA_COLLAGENS
NABA_BASEMENT_MEMBRANES
0.0
0.1
0.2
0.3
Valor absoluto de la contribución
Indicador
FALSE
TRUE


Componente Principal 3
WNT_SIGNALING
SIG_REGULATION_OF_THE_ACTIN_CYTOSKELETON_BY_RHO_GTPASES
SIG_PIP3_SIGNALING_IN_CARDIAC_MYOCTES
SIG_PIP3_SIGNALING_IN_B_LYMPHOCYTES
SIG_INSULIN_RECEPTOR_PATHWAY_IN_CARDIAC_MYOCYTES
SIG_IL4RECEPTOR_IN_B_LYPHOCYTES
SIG_CHEMOTAXIS
SIG_CD40PATHWAYMAP
SIG_BCR_SIGNALING_PATHWAY
SA_TRKA_RECEPTOR
SA_REG_CASCADE_OF_CYCLIN_EXPR
SA_PTEN_PATHWAY
SA_PROGRAMMED_CELL_DEATH
SA_MMP_CYTOKINE_CONNECTION
SA_G2_AND_M_PHASES
SA_G1_AND_S_PHASES
SA_FAS_SIGNALING
SA_CASPASE_CASCADE
SA_B_CELL_RECEPTOR_COMPLEXES
NABA_SECRETED_FACTORS
NABA_PROTEOGLYCANS
NABA_MATRISOME_ASSOCIATED
NABA_MATRISOME
NABA_ECM_REGULATORS
NABA_ECM_GLYCOPROTEINS
NABA_ECM_AFFILIATED
NABA_CORE_MATRISOME
NABA_COLLAGENS
NABA_BASEMENT_MEMBRANES
0.0
0.1
0.2
0.3
0.4
Valor absoluto de la contribución
Indicador
FALSE
TRUE

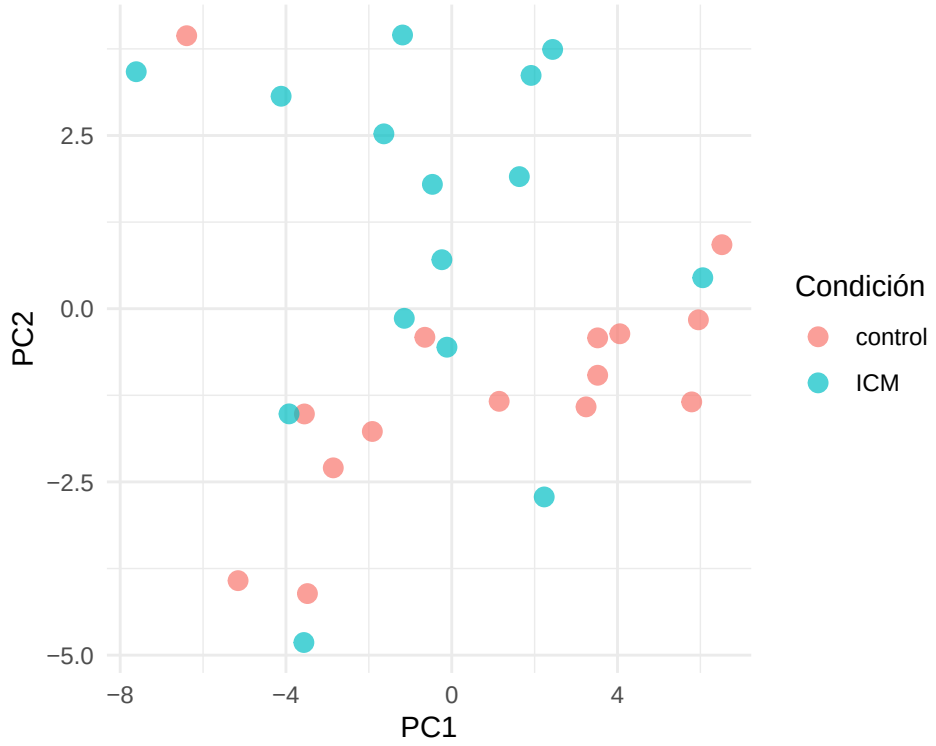


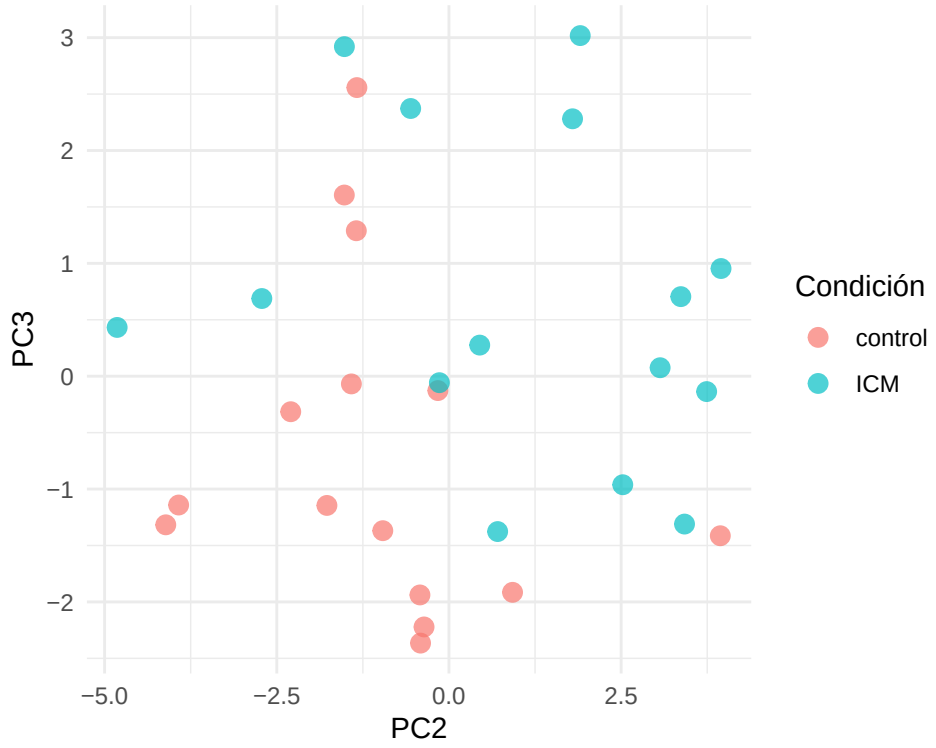
2.3.1.3 Justifica el número de componentes seleccionados:

Se seleccionó los primeros 4 componentes principales, ya que entre estos parece estar aproximadamente el 90% de la variabilidad total de los datos (90% de la información total). Aunque también podríamos conservar solo los primeros 3 componentes, ya que en estos 3 se encuentra entre el aproximadamente el 80% de la información total.

2.3.1.4 Vizualiza e interpreta los primeros 2-3 componentes principales:

En los componentes principales 1-2 se observa una buena separación de los grupos, se podría pensar que dicha separación no es tan obvia pero esto puede ser una consecuencia de la cantidad de muestras, sin embargo, como ya se mencionó se ve una clara separación de los dos grupos. De igual modo, dicha distinción entre los grupos se observa entre los componentes principales 2-3. Con estos resultados parece que incluso podríamos quedarnos con 3 componentes.





2.3.1.5 Calcular y graficar la proporción de varianza explicada:

Como se observa en el punto 2, el PC1 explica aproximadamente el 50% de la varianza, en el PC2 se explica aproximadamente el 120%, y en los siguentes PCs esta varianza explicada sigue disminuyendo, con

esto podemos supoener que el 80-90% de la varianza se encuentra entre los primeros 3 o 4 componentes principales.

2.3.1.6 Identificar las características con las mayores cargas en los componentes principales:

Observando el punto 4 podemos observar; primero que el PC1 explica el 50% de la varianza. Centrandonos es este PCs podemos observar que las rutas que más aportan a este PC son:

```
## # A tibble: 5 x 4
##   terms                                                 value component id
##   <chr>                                                 <dbl> <chr>     <chr>
## 1 SIG_PIP3_SIGNALING_IN_CARDIAC_MYOCTES                -0.238 PC1       pca
## 2 SIG_INSULIN_RECEPTOR_PATHWAY_IN_CARDIAC_MYOCYTES     -0.236 PC1       pca
## 3 SIG_CD40PATHWAYMAP                                   -0.235 PC1       pca
## 4 SIG_BCR_SIGNALING_PATHWAY                            -0.234 PC1       pca
## 5 SIG_REGULATION_OF_THE_ACTIN_CYTOSKELETON_BY_RHO_GTPASES -0.232 PC1       pca
```

2.3.1.7 Discutir la importancia biológica de los resultados del PCA :

El análisis de componentes principales (PCA) muestra una separación clara entre los dos grupos que queremos predecir en modelos posteriores. Esto sugiere que las rutas biológicas utilizadas como variables en el análisis tienen diferentes niveles de actividad dependiendo de la condición del paciente, lo que ayuda a distinguir entre los grupos. También podemos observar que dos de las rutas biológicas que resultaron importantes en el PC1 están relacionadas directamente con funciones cardíacas: **SIG_PIP3_SIGNALING_IN_CARDIAC_MYOCYTES** que contiene **genes relacionados con la señalización de PIP3 en miocitos cardíacos**, esta vía "regula múltiples procesos clave en los miocitos cardíacos, incluyendo el tamaño celular, la supervivencia, la angiogénesis y la inflamación, tanto en casos de hipertrofia cardíaca fisiológica como patológica" (Aoyagi & Matsui, 2011). y **SIG_INSULIN_RECEPTOR_PATHWAY_IN_CARDIAC_MYOCYTES** o **la señalización de la insulina**, esta vía "regula el crecimiento cardíaco, la supervivencia, la captación y uso de sustratos y el metabolismo mitocondrial. Además, modula las respuestas del corazón frente a estresores tanto fisiológicos como patológicos" (Abel, 2021). Estas vías tienen roles cruciales en la fisiología y homeostasis del tejido cardíaco.

2.3.1.8 Identidica batch effects en el espacio visible del PCA:

Al analizar la proyección de las muestras en los primeros tres componentes principales (PCs), no se observan patrones inesperados o agrupaciones extrañas que sugieran la presencia de batch effects. Las muestras se agrupan según lo esperado, con una diferencia entre los grupos de control y los de ICM (insuficiencia cardíaca).

Además, las variables que contribuyen significativamente al PC1 están estrechamente relacionadas con procesos biológicos cardíacos. Esto refuerza la idea de que el PCA está capturando diferencias biológicas relevantes en lugar de artefactos técnicos.

En resumen, no se evidencian efectos de batch en el análisis de componentes principales, lo que indica que los datos son consistentes y aptos para su uso en análisis posteriores.

2.4 Relevancia biológica

2.4.0.1 1. Explique cómo este conjunto de datos y su análisis contribuirán al campo.

Este conjunto de datos y su análisis con un enfoque con Machine Learning, contribuirán a identificar patrones transcriptómicos que ayuden a diferenciar entre paciente sanos y aquellos que padecen cardiomiopatía isquémica (ICM). Ocupando datos de expresión génica, este trabajo busca clasificar entre estos dos gurpos, además que esperamos que provee información sobre genes y vías de regularización que sean clave en esta enfermedad, lo que contribuirá a investigaciones y tratamientos futuros.

2.4.0.2 2. Describa las posibles implicaciones clínicas o de investigación.

El trabajo que deasrrollamos no buscar hacer un diagnóstico clínico en el que pueda basarse para iniciar un tratamiento. La implicación clínica de este proyecto es ser capaces de identificar a pacaiente sanos de aquellos que presentan insuficiencias cardíacas. Haremos la clasificación entre paciente "sano" y con "ICM" en base a los genes, los cuales ocuparemos como predictores.

2.4.0.3 3. Identifique cualquier limitación del conjunto de datos que pueda afectar a su análisis.

La principal limitación que observamos fue el tamaño de la muestra, ya que solamente tenemos datos ded 30 pacientes. Esto podría significar un potencial problema, ya que un tamaño de muestra tan pequeño puede limitar la generalización de los resultados, lo que a su vez aumentaría la probabilidad de sesgo en las predicciones.

3 Parte 2: Enfoque de Machine Learning

3.1 Formulación del problema

3.1.0.1 Exponga claramente su objetivo de Machine Learning.

En objetivo de este proyecto de Machine Learning es desarrollar un modelo que se capaz de predecir si una persona es sana o si presenta cardiomiopatía isquémica (ICM). Ocuapremos datos de expresión génica como predictores.

3.1.0.2 Justifique por qué el Machine Learning es apropiado para este problema.

El enfoque de Machine Learning es bueno porque nos permite hacer una reducción en la dimensionalidad del dataset, ocupando técnicas como el PCA, lo que nos permite manejar los datos que tienen dimensiones grandes sin perder información importante. Además, nos permite hacer clasificación, que es el enfoque que vamos a realizar, entrenando modelos predictivos en conjuntos de entrenamiento y prueba. Por último, algoritmos de Machine Learning nos permiten identificar las variables más importantes para realizar las predicciones, lo que nos proporcionaría información sobre los genes que podrían estar relacionados con la ICM.

3.1.0.3 Explique su elección de enfoque supervisado/sin supervisión.

Nosotras decidimos ocupar un enfoque supervisado, esto porque se buscó entrenar un modelo en base a los genes. Sumando que, la clasificación (identificar si la persona es sana o presenta cardiomiopatía isquémica) forma parte de este enfoque.

3.1.0.4 Describa cómo evaluará el éxito.

Nosotras evaluamos el éxito como: dados los genes, el modelo será capaz de hacer una predicción para identificar si la persona está sana o muestra ICM.

3.2 Implementación del modelo

3.2.1 Regresión Logística Regularizada

Realizamos tres modelos de regresión logística rgularizada:
- Lasso
- Ridge
- Elastic Net

Tras realizar nuestros modelos de `Regresón Logística Regularizada` podemos hacer una comparación entre los tres modelos a partir de tres métricas:

3.2.1.1 Accuracy: proporción de predicciones correctas. Proporción de cuántas veces el modelo predijo correctamente. Esperamos que el modelo esté haciendo buenas predicciones cuando toma valores muy cercanos a 1.

a. Lasso Model: el valor de L1 incia en ~1e-08 es el 71% de sus predicciones son correctas, y se mantiene así hasta que L1 alcanza un valor de 1e-02, cuando comienza a crecer la proporción de predicciones correctas, alcanzando un máximo en un 90% de predicciones correctas, después de esto, disminuye drásticamente.

b. Ridge Model: durante la primera parte de la gráfica, tenemos una alta proporción (< 0.8) de predicciones correctas, hasta que llega a ~1e+0.1, donde comienza a disminuir la proporción, llegando a valores de 0.5.

c. Elastic Net: a comparación de los otros dos modelos, aquí tenemos parte de las dos regularizaciones (L1 y L2), así que cada línea de diferente color mostrará la proporción de L1 que estuvo presente dentro de la evaluación. La recta naranja, tiene una mejor proporción de predicciones correctas (~0.84), pero también es la que no tiene regularización L1. En general podemos observar que conforme aumenta la proporción de L1 que tien el modelo, el valor de ‘accuracy‘ decrementa. Esto nos dice que el modelo Ridge, hace un mejor papel en la predicción (para esta métrica).

3.2.1.2 Brier class: mide la calidad de las predicciones. Compara las probabilidades predichas por el modelo con el valor real. Esperamos que este valor sea muy cercano a 0, esto nos indicaría que el modelo está haciendo un buen trabajo.

a. Lasso Model: el valor del Brier Score inicia en ~0.19, que en sí podemos pensar que no es muy alto, pero, comparado con otros modelos realizados en clase y tareas (Brier Score de 0.05), es considerablemente grande. Este valor se mantiene constante hasta que L1 toma valores de ~1e-0.2, donde hay una caída en este valor, llegando a un coeficiente de ~ 0.15, para despues crecer y llegar a 0.25. Este mínimo coincide cuando el valor accuracy también toma mejores valores (0.8).

b. Ridge Model: iniciamos con un valor de 0.20 y se mantiene constante hasta ~1e+0.1, donde comienza a creer esta medida, lo que indica que las predicciones son malas.

c. Elastic Net: por otro lado, el modelo que tiene una mejor calidad en las predicciones (Brier Class pequeño) es aquel que tiene una proporción de L1 = 0.25, con un valor en esta métrica de ~0.16, con un mínimo que llega a ~0.125.

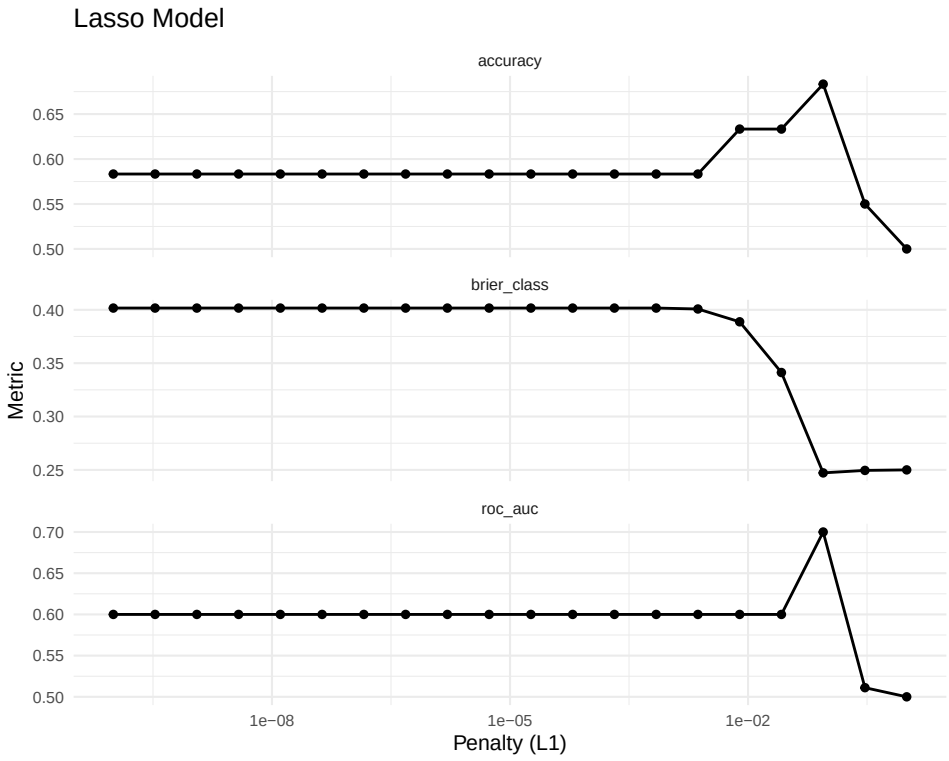
3.2.1.3 ROC-AUC: esta curva nos muestra la relación que hay entre los verdaderos positivos y los falsos positivos. Valores cercanos a 1 indican un muy buen modelo, mientras que valores cercanos 0, son señalan de un bajo rendimiento en el modedlo.

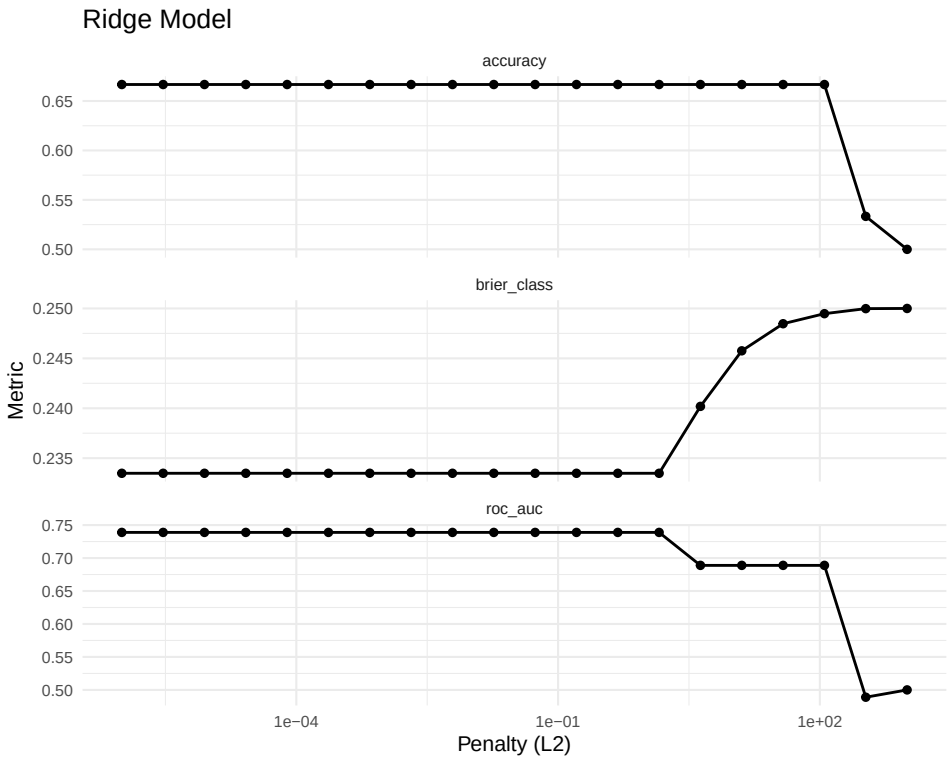
a. Lasso Model: en este modelo, iniciamos con un muy buen valor (0.9), el cual se mantiene conforme crece L1, hasta llegar a ~1e-0.2, donde, al igual que en las otras métricas, toma un mejor valor, para luego decaer. En conclusión sobre este modelo, esperamos que el mejor modelo esté entre valores de L1 cercanos a 1e-0.2.

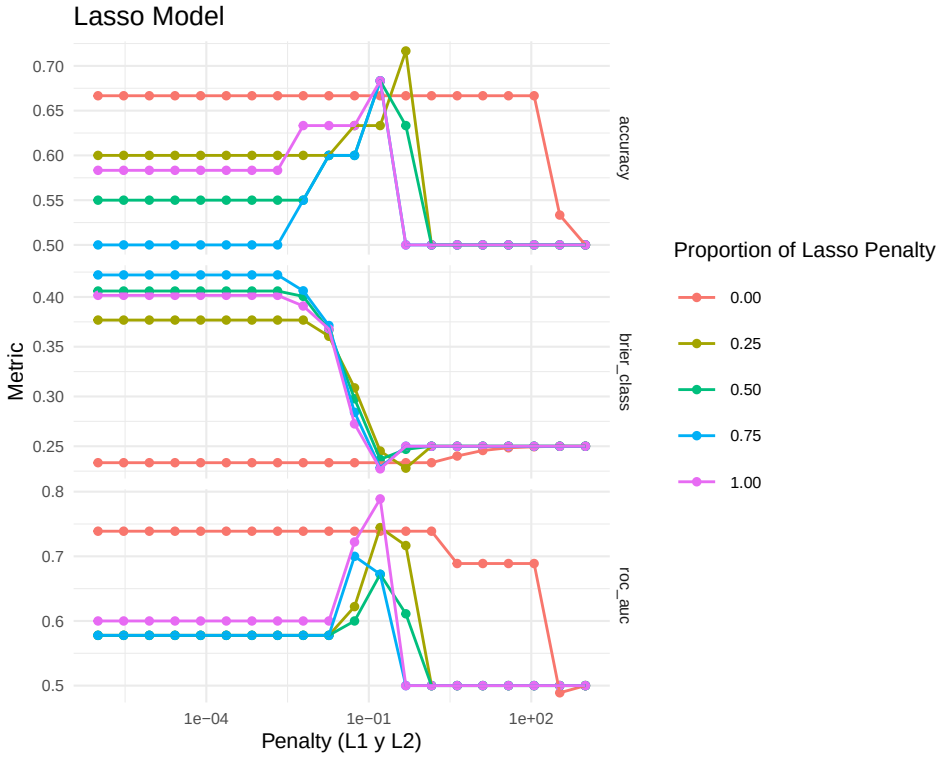
b. Ridge Model: al igual que el modelo anterior, el valor bajo la curva es alto (0.9), hasta

que llegamos a ~1e+0.1, donde igualmente comienza a dismnuir el valor de ROC-AUC, lo cual no es lo ideal. En conclusión, esperamos que un buen modelo esté por debajo de un valor de L2 de ~1e+0.1.

c. Elastic Net: por otro lado, para esta métrica, el mejor modelo es aquel que tiene la proporción de L1 de 1, seguido del modelo de proporción L1 = 0. En este caso, resulta confuso qué modelo escoger.\newline







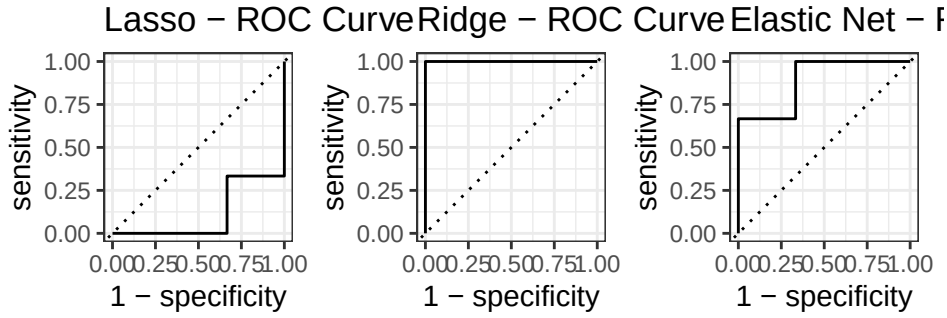
Luego realizamos la evaluación de los modelos en el dataset de entrenamiento. Para poder comparar las tres técnicas obtuvimos las curvas ROC-AUC. Como se puede observar, en general, no son muy buenas y esto se debe a que se ocuparon solamente 6 datos para hacer la evaluación.

```
## # A tibble: 2 x 3
```

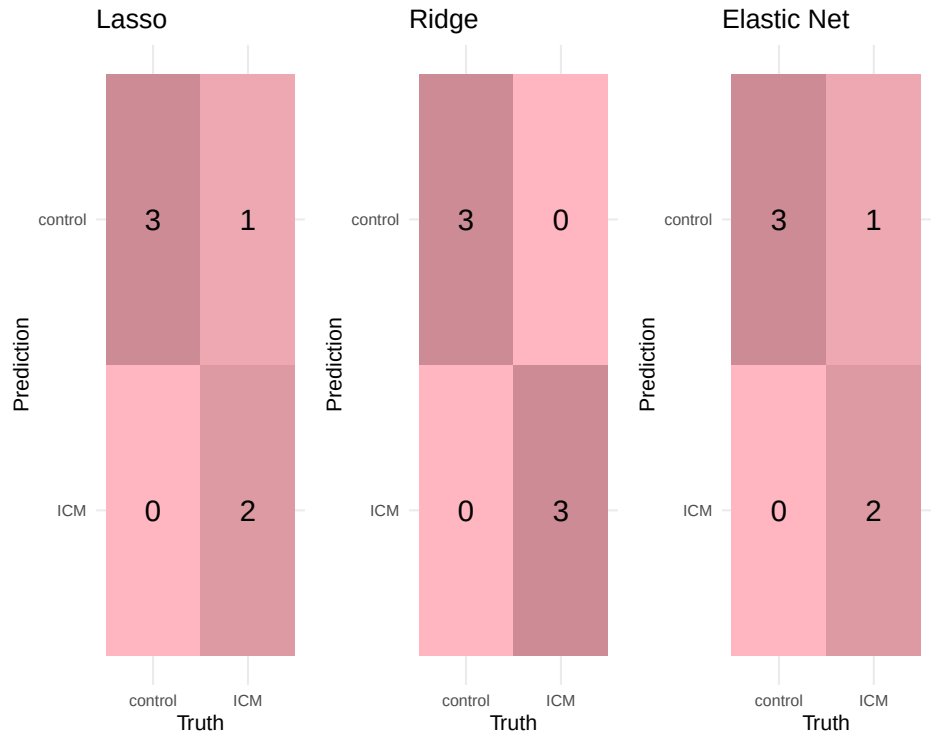
```
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.833
## 2 kap      binary         0.667
```

```
## # A tibble: 2 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary             1
## 2 kap      binary             1
```

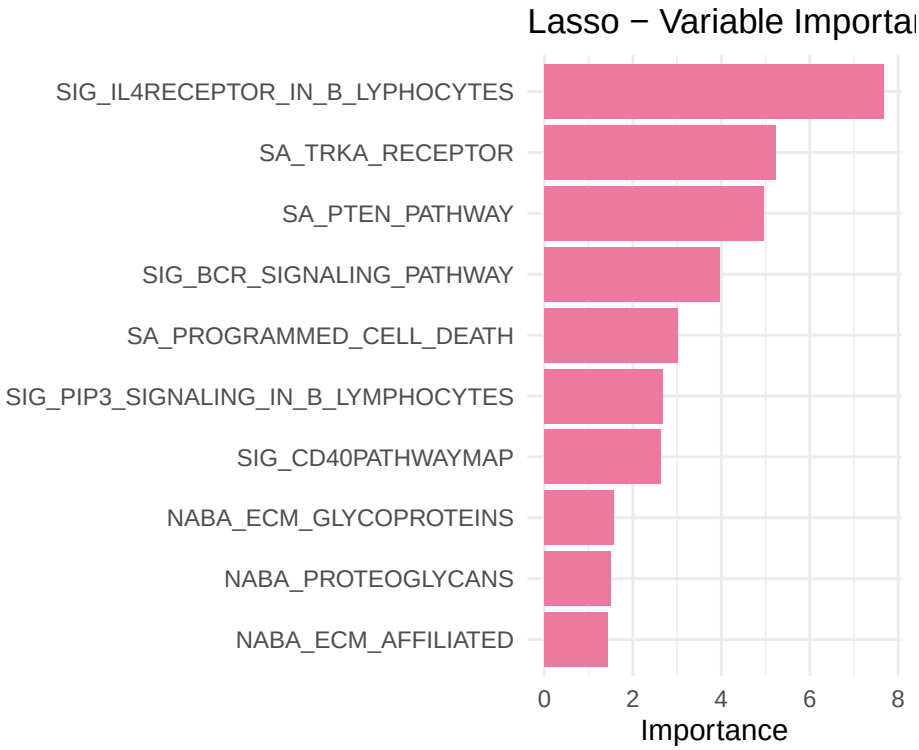
```
## # A tibble: 2 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.833
## 2 kap      binary         0.667
```

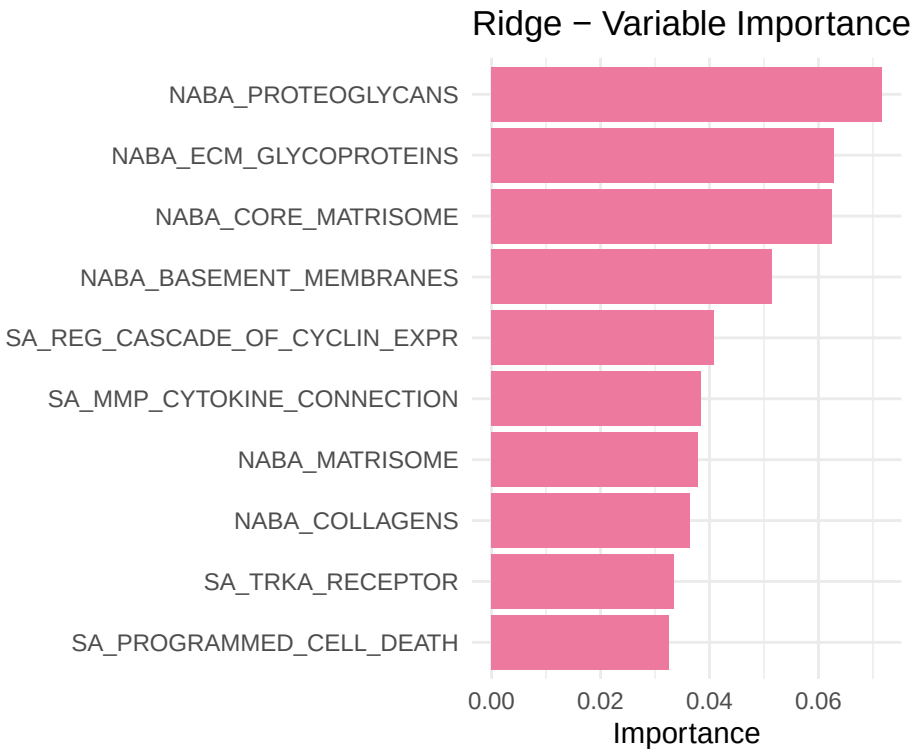


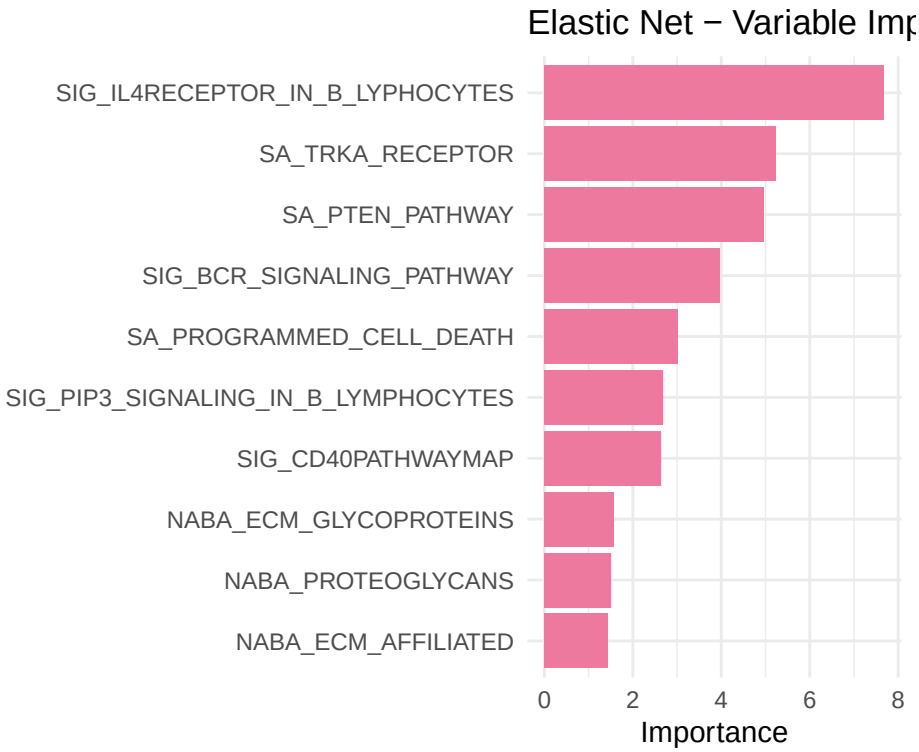
La matrices de confusión nos permiten observas las predicciones que son correctas de las que no lo son, como mencionamos anteriormente, solo tenemos 6 datos, por lo que las tres matrices son iguales. Si tuviéramos más datos, sí esperaríamos ver diferencias entre los modelos.



En los siguientes plots podemos ver las variables que contribuyen más a cada uno de los modelos. Estos nos permitirán compararlos con las técnicas de Decision Tree y Random Forest.







3.2.2 Decision Tree y Random Forest

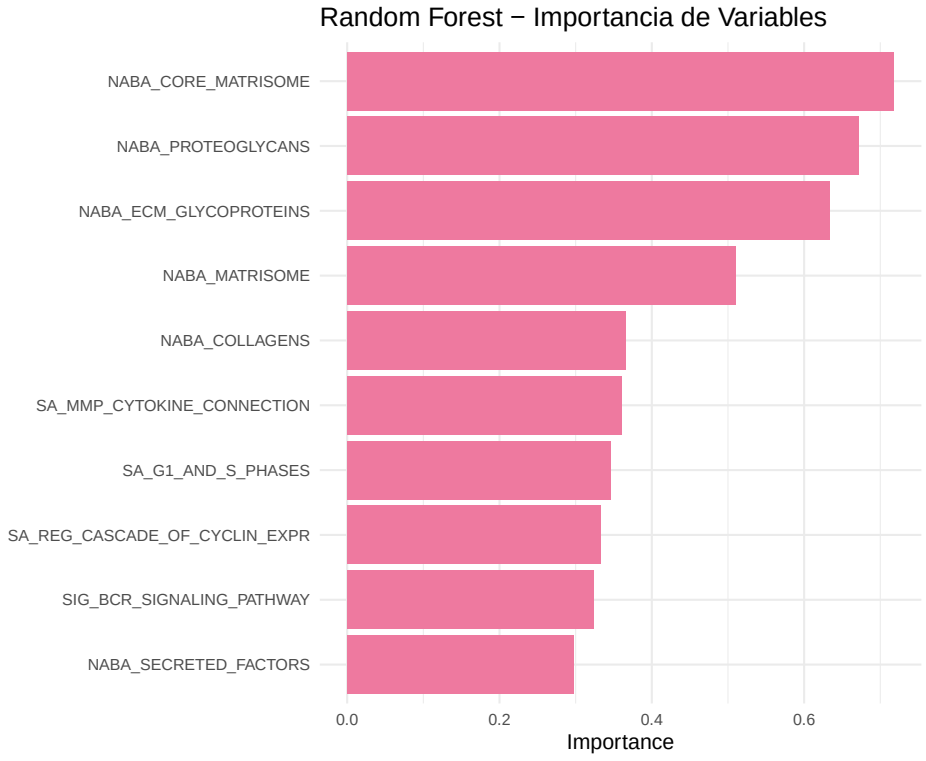
Implementamos ambos modelos de clasificación.

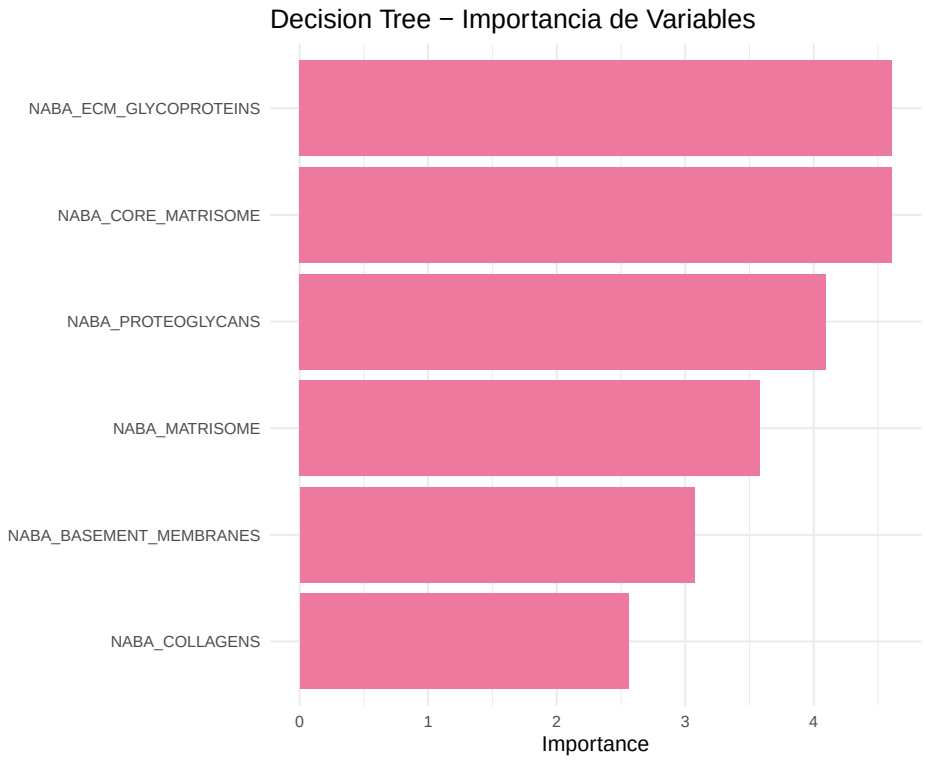
Luego, evaluamos los modelos en base a sus métricas y observamos que **Random Forest** tiene un mejor desempeño al realizar las predicciones en comparación con **Decision Tree**.

```
## # A tibble: 4 x 8
##   .estimator     n std_err .config              model    accuracy    kap roc_auc
##   <chr>      <int>   <dbl> <chr>                <chr>       <dbl>  <dbl>   <dbl>
## 1 binary         5  0.0972 Preprocessor1_Model1 Random ~    0.783 NA      NA
## 2 binary         5  0      Preprocessor1_Model1 Decisio~    0.5    0       0.5
## 3 binary         5  0.194  Preprocessor1_Model1 Random ~   NA      0.567  NA
## 4 binary         5  0.0849 Preprocessor1_Model1 Random ~   NA     NA       0.811
```

En el modelo de **Random Forest** tenemos como variables importantes a varias rutas de **NABA_PROTEOGLYCANS**, **NABA_ECM_GLYCOPROTEINS**, **NABA_CORE_MATRISOME**, **NABA_MATRISOME**, etc. En el modelo de **Decision Tree** tenemos estas mismas cuatro al principio aunque en diferente orden, y otras dos rutas que no se encuentran en las variables de **Random Forest**. Estas rutas están altamente relacionadas con la estructura de la Matriz EXtracelular, esto se puede observar en los siguientes plots:

```
## Warning: `pull_workflow_fit()` was deprecated in workflows 0.2.3.
## i Please use `extract_fit_parsnip()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



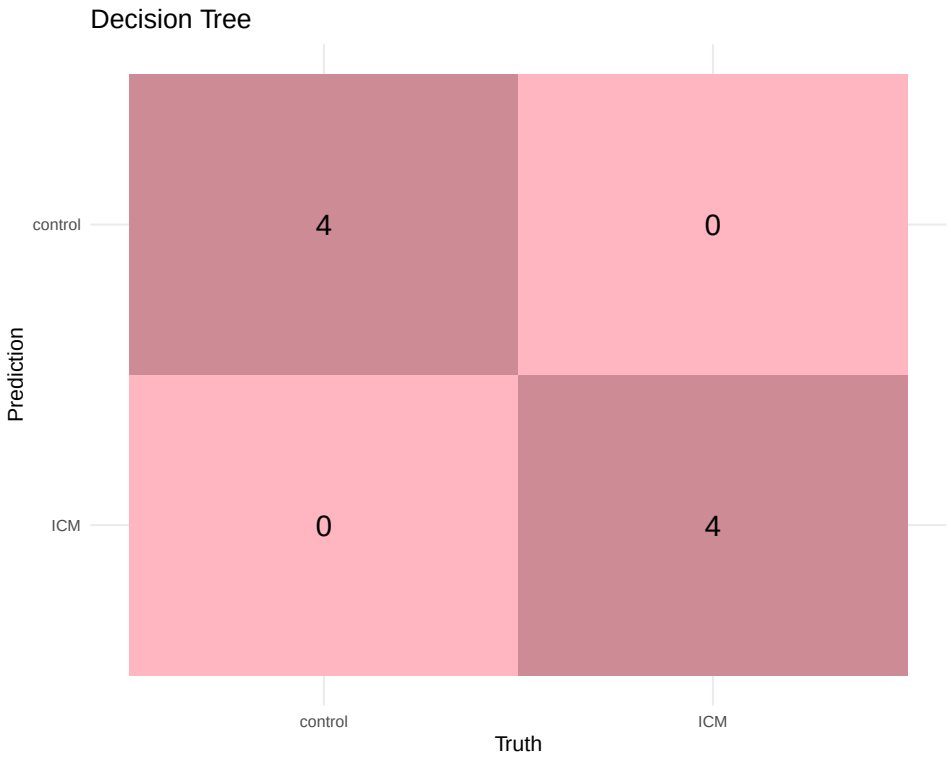



```
## # A tibble: 3 x 4
##   .metric     .estimator .estimate .config
##   <chr>       <chr>          <dbl> <chr>
## 1 accuracy    binary         1     Preprocessor1_Model1
## 2 roc_auc     binary         1     Preprocessor1_Model1
## 3 brier_class binary         0.105 Preprocessor1_Model1
```

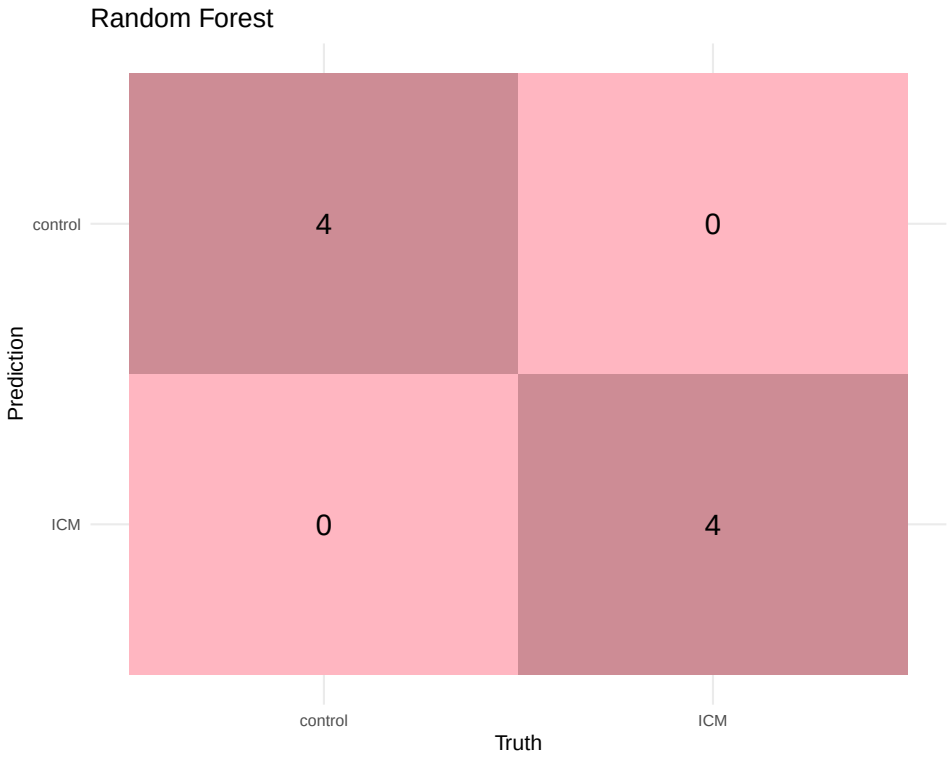
```
## # A tibble: 3 x 4
##   .metric     .estimator .estimate .config
##   <chr>       <chr>          <dbl> <chr>
## 1 accuracy    binary        1      Preprocessor1_Model1
## 2 roc_auc     binary        1      Preprocessor1_Model1
## 3 brier_class binary        0.0328 Preprocessor1_Model1
```

En términos generales, parece ser que el modelo de **Random Forest** y el de **Decision Tree** tienen un desempeño similar. Sin embargo, es posible que observemos este resultado debido a la limitación en la cantidad de muestras del dataset. Con base en las métricas y si se tuviera un dataset más grande en cuanto al número de pacientes, se esperaría que **Random Forest** tuviera un mejor desempeño en la predicción en comparación con **Decision Tree**.

```
## Scale for fill is already present.
## Adding another scale for fill, which will replace the existing scale.
```

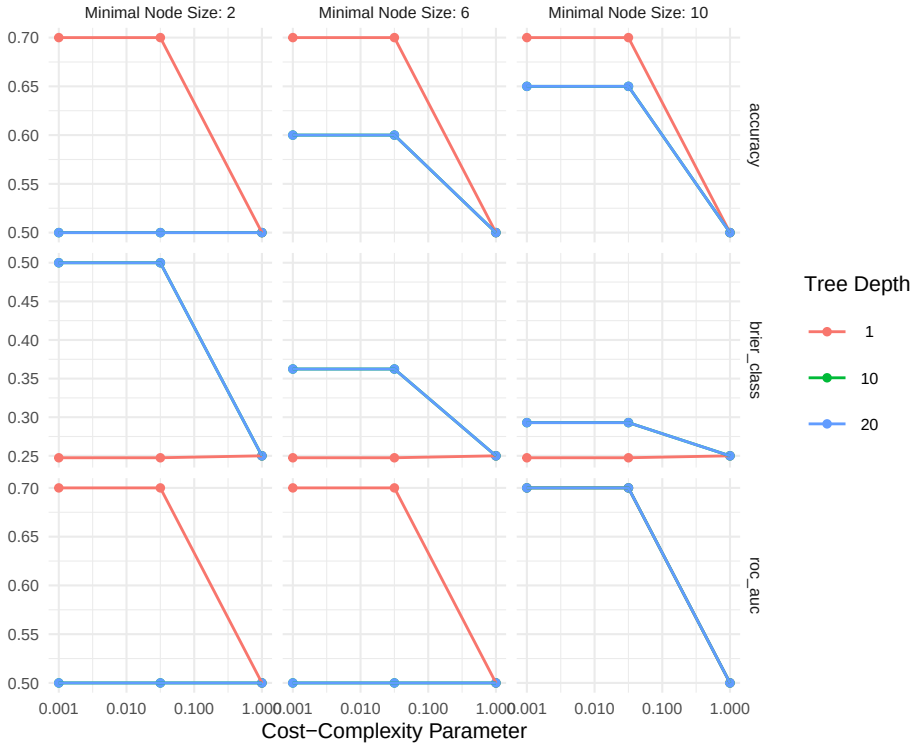


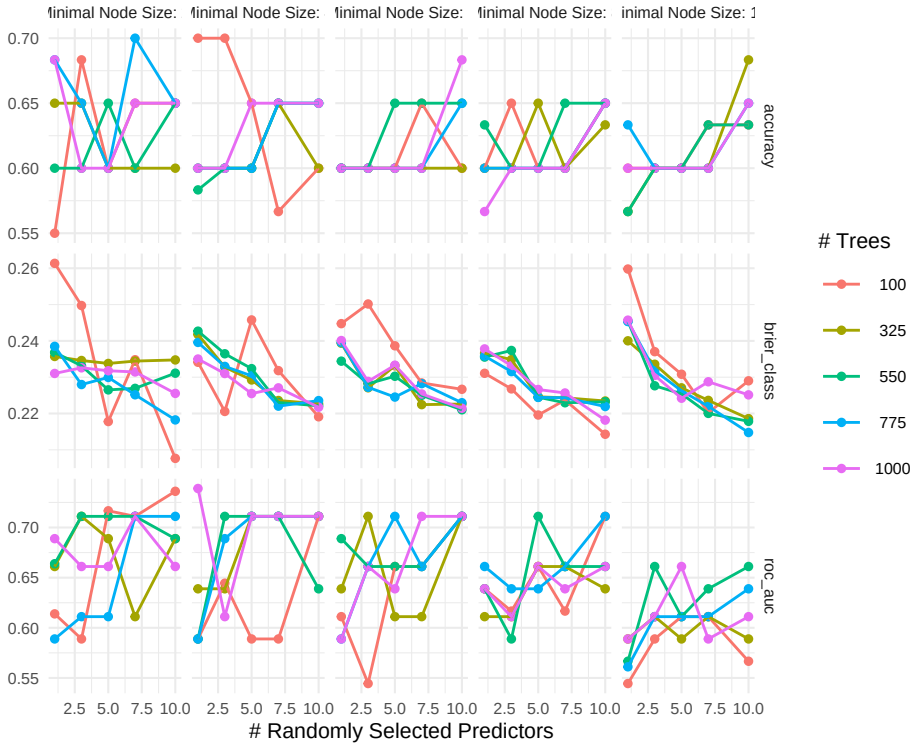

```
## Scale for fill is already present.
## Adding another scale for fill, which will replace the existing scale.
```



Para estos modelos optimizamos parámetros como tree depth o number of trees.

En las siguientes gráficas podemos observar el desempeño de el modelo en relación al valor de sus hiperparámetros.





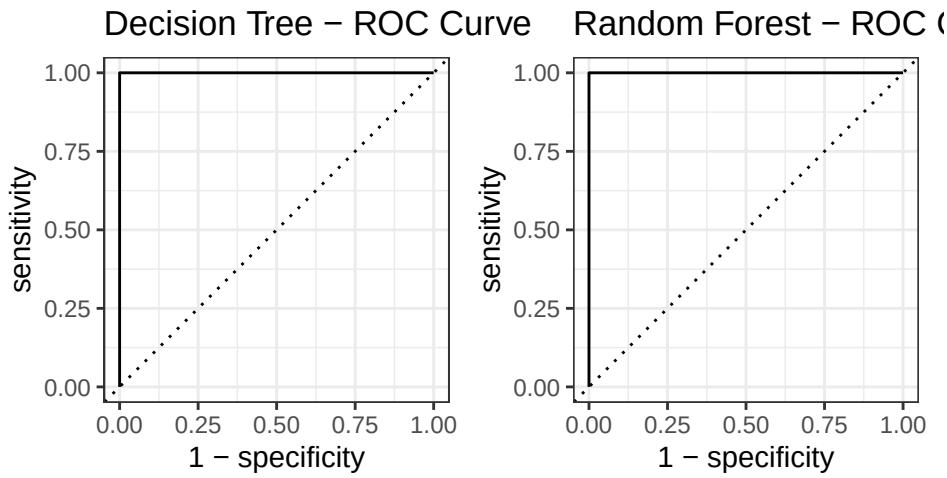
Ahora seleccionaremos los mejores hiperparámetros, para **Random Forest** fueron: mtry = 1, trees = 1000 y min_n = 4; mientras que para **Decision Tree** fueron: cost_complexity = 0.001, tree_depth = 1 y min_n

= 2.

```
## # A tibble: 1 x 4
##    mtry trees min_n .config
##   <int> <int> <int> <chr>
## 1     1  1000     4 Preprocessor1_Model046
```

```
## # A tibble: 1 x 4
##   cost_complexity tree_depth min_n .config
##             <dbl>      <int> <int> <chr>
## 1           0.001          1     2 Preprocessor1_Model01
```

Finalmente comparamos el desempeño de los modelos (ocupando los hiperparámetros ya seleccionados). Parece ser que nuevamente ambos modelos tienen un desempeño muy similar, y como se mencionó anteriormente es posible que sus predicciones pudieran variar si se tuvieran más muestras.



Por último, las matrices de confusión nos permiten observar las predicciones realizadas por ambos modelos. Nuevamente obtuvimos predicciones perfectas, esto también se debe a los pocos datos con los que se evaluó el modelo (dataset de prueba).

3.3 Ingeniería de características

3.3.0.1 Describa su estrategia de selección/ingeniería de características.

Dado que el final_data no contenía una gran cantidad de variables (pathways), como ocurría con el dataset inicial que incluía los transcritos, decidimos utilizar todas las variables como predictoras. Esta decisión se tomó considerando la posibilidad de que las variables importantes para explicar la variabilidad en el PCA no fueran necesariamente las mismas relevantes para la predicción de la condición de los pacientes. Como resultado se observó que las variables clave para cada análisis diferían.

3.3.0.2 Justificar la relevancia biológica de las características seleccionadas

De manera general, la mayoría de los modelos incluyen rutas relacionadas con la estructura y función de la matriz extracelular (ECM). Como se ha mencionado, "La red de la matriz extracelular (ECM, por sus siglas en inglés) juega un papel crucial en la homeostasis cardíaca, no solo proporcionando soporte estructural, sino también facilitando la transmisión de fuerzas y transmitiendo señales clave a los cardiomiocitos, células vasculares y células intersticiales. Los cambios en el perfil y la bioquímica de la ECM pueden estar involucrados de manera crítica en la patogénesis tanto de la insuficiencia cardíaca con fracción de eyección reducida como de la insuficiencia cardíaca con fracción de eyección preservada' (Frangogiannis, 2019, traducido)".

3.3.0.3 Aborde la "maldición de la dimensionalidad" en los datos genómicos.

La "maldición de la dimensionalidad" es un desafío inevitable al trabajar con datos genómicos, dados los miles de elementos que pueden analizarse, como genes, transcritos o proteínas. Esto debido al inmenso tamaño del genoma y la complejidad de sus interacciones. Aunque existen diferentes tipos de datos y enfoques para procesarlos, gran parte del tiempo se estará lidiando con un volumen masivo de variables. Dichos problemas los observamos tanto en el PCA como en los modelos antes de el uso de GSVA.

3.3.0.4 Explica cómo manejarás los " batch effects" o variaciones técnicas.

Una forma en que se abordó este problema en este proyecto fue mediante un análisis exploratorio de los datos, particularmente observando el PCA. Se verificó que las agrupaciones de las muestras y las variables más relevantes del análisis fueran coherentes con el contexto biológico del estudio "ICM". Este proceso permitió detectar posibles "batch effects" como lo hubiera sido si las muestras se agrupaban de manera inesperada, lo cual podría indicar que la variación observada no estaba relacionada con las condiciones biológicas de los pacientes.

4 Parte 3: Revisión de la Literatura

4.1 Análisis de la literatura primaria

4.1.1 Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure

Cameron R. Olsen, MD, a Robert J. Mentz, MD, a Kevin J. Anstrom, PhD, b David Page, PhD, b and Priyesh A. Patel, MDc Durham, and Charlotte, NC

4.1.1.1 ¿Cuál fue el objetivo principal y el enfoque?

La relevancia clínica de este artículo de investigación se centra en el número de personas, en Estados Unidos, que son afectadas por algun tipo de insuficiencia cardíaca. Un diagnóstico a tiempo es clave, sin embargo las herramientas convencionales de predicción tiene un bajo rendimiento y pueden llegar a ser poco accesibles. Por lo anterior, es necesario implementar nuevos métodos y formas de diagnóstico, los cuales sean capaces de procesar grandes cantidades de datos clínicos. Ellxs propusieron un enfoque con Inteligencia Artificial (IA), ocupando Machine Learning, con enfoque supervisado y no supervisado, al igual que redes neuronales.

4.1.1.2 ¿Cuáles fueron las principales innovaciones metodológicas?

Hablando de la capacida de predicción de los modelos, estos algoritmos fueron capaces de diferenciar entre personas sanas y personas enfermas, incluso en un dataset pequeño.

Para la medicina de precisión, consdieramos que esta investigación tuvo las siguientes innovaciones:

a. Clasificación de tipos de IC: se identificaron fenotipos específicos, lo que permitió clasificar las distinas condiciones de problemas cardíacos.

b. Respuesta a terapias: con técnicas supervisadas y de k-means, exploraron la capacidad de predicción de respuesta a diferentes terapias, como la resincronización cardíaca y dispositivos de asistencia ventricular, en donde los resultados fueron prometedores.

4.1.1.3 ¿Cómo validaron sus resultados?

Solo mencionaremos dos modelos realizados en esta investigación:

a. Machine Learning para puntuación de riesgo: identificaron seis variables predictivas clave para realizar un puntaje de riesgo. Ocuparon métricas scomo AUC para evaluar el resultado, el cual tuvo un valor de 0.841, el cual lo podríamos considerar como un puntaje relativamente alto.

b. Convolutional Neural Network: buscaron detectar disfunción ventricular. Para validar este modelo dividieron los datos en conjunto y entrenamiento, de aquí, las predicciones obtenidas fueron bastante buenas, esto lo podemos ver en la métria AUC, la cula tuvo un puntaje de 0.93, el cual es bastante bueno.

Por último, en otros tres modelos que realizaron, hicieron comparaciones con otras investigaciones.

4.1.1.4 ¿Cuáles fueron las limitaciones

Algunas de las limitaciones que notamos fueron:

a. Los datos ocuapdos para los estudios iniciales fueron limitados, lo que se podría ver reflejado en una falta de representación, lo que limitaría la capacidad del algoritmo para generalizar a poblaciones más grandes.
b. Igualmente consideramos que en la parte de validación, hubo resultados que no se validaron con información externa, sino que, apesar de que eran pocos datos, se limitaron a los resultados obtenidos.
c. La calidad y consistencia de los datos, como en los reportes de ecocardiografía y registros clínicos mencionados, pueden introducir ruido y sesgos en los modelos.

4.1.1.5 ¿De qué manera su planteamiento es útil para su proyecto?

Consideramos que este planteamiento es útil porque nos da la perspectiva de que es posible usar algoritmos de Machine Learning para hacer predicciones sobre insuficiencias cardíacas, ya que nuestros resultados no fueron tan prometedores como esperábamos; pero esta investigación mostró que sí es posible, pero con características y variable particulares. Además que también aborda este tema desde las Redes Neuronales, que creemos que dan otra pespectiva muy valiosa.

4.1.2 Applications of artificial intelligence and machine learning in heart failure

Tauben Averbuch, Kristen Sullivan, Andrew Sauer, Mamas A Mamas, Adriaan A. Voors, Chris P. Gale, Marco Metra, Neal Ravindra,and Harriette G.C. Van Spal

4.1.2.1 ¿Cuál fue el objetivo principal y el enfoque?

El objetivo principal de este artículo fue desarrollar un algoritmo de Deep Learning para predecir la mortalidad de pacientes con insuficiencia cardíaca aguda. Ocuparon información de 12 hospitales. El modelo se entrenó ocupando los datos de dos hospitales y se validó con información de diferentes hospitales de diferentes lugares.

4.1.2.2 ¿Cuáles fueron las principales innovaciones metodológicas?

Las características que notamos ded esta investigación fueron que se desarrollaron varios algoritmos, tanto de Machine Learning como de Deep Learning, lo que consideramos que hace da más robustez. Algunos de los modelos de MLS ocupados fueron: Random Forest, Regresión Logística, Máquinas de Soporte Vectorial y Redes Bayesianas. También es importante mencionar que se realizó un buen pre-procesamiento de los datos, además de que se realizaron varios conjuntos de entrenamiento, algo que solo notamos en este artículo.

4.1.2.3 ¿Cómo validaron sus resultados?

Hubo varios puntos que se realizaron:

a. Los datos ocupados para hacer la evaluación del modelo fueron independientes a los empleados para entrenar los modelos. Estos datos provenían de otros hospitales, centros y clínicas.

b. Los modelos desarrollados fueorn comparados con otros modelos de predicción que son más *convencionales* para evaluar la mortalidad de personas que tienen alguna insuficiencia cardíaca. Para estas comparaciones se ocupó la métrica ROC-AUC, la cual permite evaluar el rendimiento que tuvo el modelo de clasificación para distinguir entre las clases de la respuesta variable. El puntaje obtenido fue de:

i. 0.88 para mortalidad hospitalaria.
ii. 0.782 para mortalidad a los 12 meses.
iii. 0.813 para mortalidad a los 36 meses.

En general; el ROC-AUC (mencionados anteriormente) de los modelos no convencionales superó al de los modelos convencionales.

4.1.2.4 ¿Cuáles fueron las limitaciones

Observamos algunas limitantes:

a. Para entrenar al modelo se le dio información de solamente dos hospitales, lo que creemos que podría haber una falta de representación de otros centros, en especial, si estos datos provienen de diferentes lugares. Incluso puede ser que las condiciones ambientales y/o sociales puedan influenciar en el momento en que se entrene el algoritmo.

b. Durante la recolección de datos, se omitieron registros de pacientes que estuvieran incompletos; esto puede provocar que se excluyan características significativas.

4.1.2.5 ¿De qué manera su planteamiento es útil para su proyecto?

A pesar de que este enfoque no se alinee completamente con el planteado en nuestro proyecto final, coincidemos en que es crucial entender ćomo, en la actualidad, se están utilizando las herramientas de Inteligencia Artificial dentro del área de salud humana, en este caso, en cardiología. Examinar esta investigación nos permitió observar el proceso detallado de cómo implementar todas estas herramientas, desde la recolección y limpieza de datos hasta la selección de la información que se ocupará para entrenar los algoritmos de ML.

4.1.3 Machine learning to identify a composite indicator to predict cardiac death in ischemic heart disease

Alessandro Pingitore, Chenxiang Zhang, Cristina Vassalle, Paolo Ferragina, Patrizia Landi, Francesca Mastorci, Rosa Sicari, Alessandro Tommasi, Cesare Zavattari, Giuseppe Prencipe, Alina Sîrbu

4.1.3.1 ¿Cuál fue el objetivo principal y el enfoque?

En el artículo se menciona cómo la *enfermedad isquémica del corazón* es una de las principales causas de muerte en el mundo. Sin embargo, esta enfermedad presenta diversas variables relacionadas, lo que dificulta identificar a los pacientes con alto riesgo. El objetivo de este estudio es utilizar *ML* para identificar a pacientes con enfermedad isquémica del corazón (EIC) que tienen un alto riesgo de muerte cardíaca (MC).

4.1.3.2 ¿Cuáles fueron las principales innovaciones metodológicas?

Para lograr esta meta, se utilizaron diversos métodos de *Machine Learning* combinados, formando así un modelo de ensamble, el cual mejora el rendimiento predictivo en comparación con los modelos individuales. Los modelos utilizados fueron los siguientes:

- Regresión Logística (Logistic Regression, LR)
- Bosques Aleatorios (Random Forest, RF)
- Adaboost.

4.1.3.3 ¿Cómo validaron sus resultados?

Los autores realizaron diversos análisis para validar sus modelos, pero uno que nos gustaría mencionar fue el uso de las curvas ROC-AUC. Este enfoque es similar al que utilizamos nosotras para validar nuestros modelos, ya que permite comparar y evaluar el desempeño de diferentes modelos en cuanto a su capacidad para clasificar correctamente los casos. A través de este análisis, los autores pudieron identificar qué modelo tenía un mejor desempeño y llegaron a la conclusión de que el modelo de ensamble. De este modo, al igual que nosotros, los autores utilizaron las curvas ROC para determinar cuál era el modelo más adecuado, observando cómo las predicciones de los modelos diferían en cuanto a precisión y capacidad de clasificación.

4.1.3.4 ¿Cuáles fueron las limitaciones

Los autores mencionan que los métodos de Machine Learning (ML) utilizados en el estudio resultaron ser "black-box methods" debido a que, aunque los modelos ofrecieron buenas predicciones, era complicado interpretar lo que realmente estaba sucediendo dentro del modelo. A pesar de contar con un buen desempeño predictivo, la falta de explicabilidad es una limitación importante en el uso de estos métodos. Otra limitación fue el uso de datos recolectados hace varios años, lo que impidió la inclusión de variables más recientes y relevantes que podrían haber mejorado el modelo. Finalmente, se excluyeron ciertos datos durante la fase de entrenamiento, lo que introdujo la posibilidad de sesgos, ya que la exclusión de información podría haber afectado la representatividad y la precisión del modelo.

4.1.3.5 ¿De qué manera su planteamiento es útil para su proyecto?

Este artículo es relevante para nuestro proyecto, ya que aborda varios aspectos clave que también hemos enfrentado. Al igual que nosotras, los autores utilizaron Random Forest (RF) y las métricas de ROC-AUC para evaluar el rendimiento de los modelos. También destacan cómo la exclusión de datos puede generar sesgos, un problema similar al que nos enfrentamos debido al tamaño reducido de nuestro conjunto de datos. Entender los procedimientos y desafíos que los autores superaron nos ayuda a manejar problemas similares y a interpretar mejor nuestros resultados.

4.2 Comparación de Métodos

4.2.0.1 Comparar y Contrastar Diferentes Enfoques de ML Utilizados en la Literatura.

Sin duda, algo que se destaca al revisar estos artículos es que, aunque todos se enfocan en la insuficiencia cardíaca, abordan diferentes aspectos y objetivos dentro del campo. Algunos artículos se centran en predecir la presencia de condiciones específicas en pacientes, como la insuficiencia cardíaca, mientras que otros están más orientados a predecir el riesgo de muerte o el pronóstico de los pacientes con esta enfermedad. A pesar de los diferentes enfoques, el uso de métodos de ML está presentes en todos

4.2.0.2 Justificación del Método Elegido Basado en Esta Revisión.

Un dato relevante es que, en dos de los artículos presentados, tanto Random Forest como Regresión Logística fueron utilizados como modelos principales para la predicción de insuficiencia cardíaca. Estos enfoques, ampliamente utilizados en la literatura, también demostraron un alto rendimiento en nuestro estudio, lo que resalta la efectividad de estos métodos para manejar datos en contextos clínicos y biológicos.

4.2.0.3 Identificación de Potenciales Mejoras sobre Enfoques Existentes.

Consideramos que, dado que este es un tema ampliamente conocido y ya abordado en la literatura, el enfoque podría redirigirse principalmente hacia la identificación de los subtipos. Esta aproximación podría tener importantes implicaciones médicas, ya que un diagnóstico más preciso y específico de los subtipos permitiría un tratamiento más adecuado y personalizado para los pacientes, mejorando potencialmente los resultados clínicos.

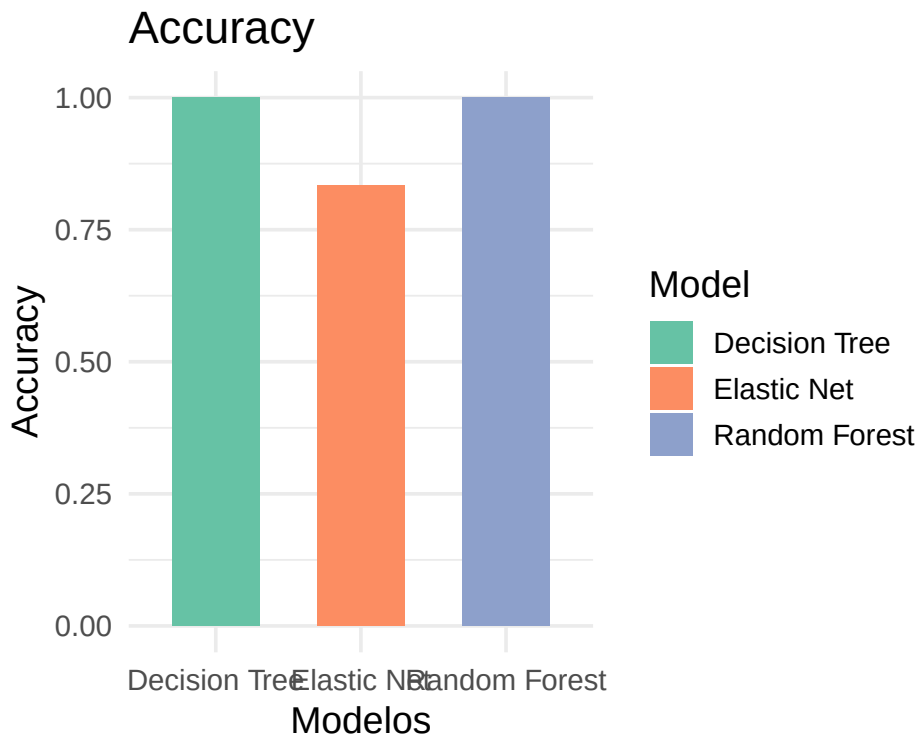
5 Parte 4: Resultados e Implementación

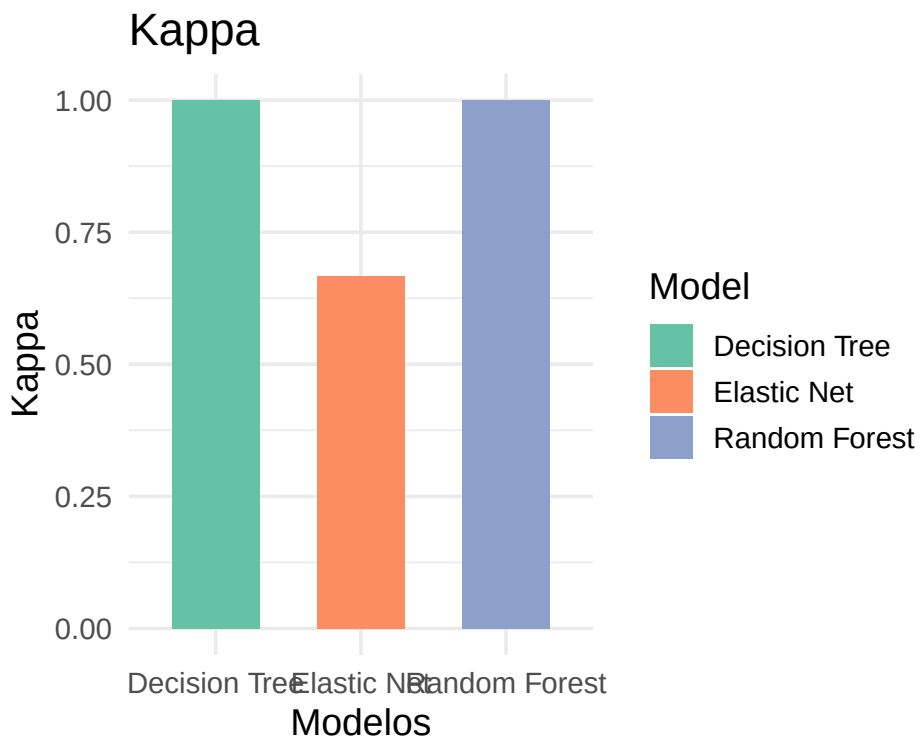
5.1 Aplicación Técnica

5.1.0.1 Métricas de comparación entre distintas variaciones del modelo y compara con métodos existentes

El mejor modelo que obtuvimos en Regresión Logística Regularizada fue **Elastic Net**, por lo que lo comparamos con **Decision Tree** y **Random Forest** para seleccionar el mejor modelo de entre los tres.

Como podemos observar en las siguientes dos gráficas, en dos de tres modelos obtuvimos muy buenas métricas, tanto en Accuracy como en Kap. Esto se debe a los pocos datos con los que se entrenaron los modelos, y puede resultar confuso para escoger un único modelo. Para poder hacer descarte de modelos y seleccionar aquel que tenga un mejor rendimiento, es necesario evaluarlos en un conjunto de datos más grande. En la revisión de literatura encontramos un artículo en donde realizaron un `modelo de ensamble`, en donde juntaron tres modelos diferentes; tal vez se pdoría hacer esto mismo pero con los tres modelos con mayor poder predictivo que obtuvimos.





5.2 Análisis de resultados

5.2.0.1 Interpretación biológica de los resultados

Retomando los puntos previamente discutidos sobre las variables importantes, podemos concluir que, en la mayoría de los modelos, las rutas relacionadas con la estructura y función de la matriz extracelular (ECM) resultaron ser de suma importancia. Dado que una de las principales funciones de la ECM es proporcionar soporte estructural a las células, ayudándolas a mantener su forma y permitir que se organicen en un tejido funcional, esto podría estar vinculado a la miocardiopatía isquémica, en particular en lo que respecta a la estructura de las arterias, ya que esta condición reduce la capacidad del corazón para bombear sangre.

5.2.0.2 Discutir las limitaciones y posibles mejoras

La principal limitación que tuvimos fue el número de datos que utilizamos (30 pacientes); esto lo vimos reflejado en todos los modelos realizados, ya que las métricas de evaluación (por ejemplo: ROC-AUC) salían o muy buenas o muy malas, debido a que en el dataset de prueba había muy pocos datos. De esta misma limitación podemos mencionar que en un número pequeño de datos es poco probable que se identifiquen patrones o relaciones entre las variables, lo que podría cambiar la interpretación biológica; además que si decidiéramos escoger otro enfoque, como serían las Redes Neuronales, necesitaríamos muchos más datos, o incluso para un enfoque de Machine Learning no supervisado.

El otro problema que encontramos durante el ajuste de hiperparámetros de los modelos de Random Forest (RF) y Decision Tree (DT) fue que las métricas obtenidas para estos modelos eran de 1, lo que nos llevó a plantear dos posibles explicaciones. Primero, que las variables utilizadas como predictoras pudieran estar demasiado correlacionadas con el outcome, lo que provocaría una predicción perfecta. La segunda posible explicación, que consideramos más probable debido al tamaño limitado del dataset, es que no se capturó toda la variabilidad de los datos. A partir de la matriz de correlación, supusimos que el problema podría deberse al tamaño de muestra.

Una mejora sería recabar más información de pacientes que tengan cardiomaptía isquémica, esto nos permitirá generalizar los modelos y reducir el sesgo que hayamos obtenido. Probar nuevos (o cambios) modelos

también podría ampliar nuestro cocnocimiento sobre esta área de la salud humana.

5.3 Perspectiva biológica

5.3.0.1 Explica los nuevos conocimientos biológicos adquiridos

Aunque ya sabemos que la matriz extracelular (ECM) está relacionada con la miocardiopatía isquémica, identificar las rutas biológicas más asociadas nos permitirá comprender de manera más precisa cómo la ECM influye en el desarrollo y progresión de esta enfermedad. Este conocimiento más detallado podría ayudar a identificar mecanismos específicos de la ECM que podrían ser objetivos terapéuticos, mejorando así el enfoque en el tratamiento y diagnóstico de la miocardiopatía isquémica.

5.3.0.2 Discute potenicales aplicaciones clínicas o de investigación.

Inicialmente no esperamos que nuestras prediccione se ocupen para hacer diagnósticos médicos, pero se pueden realizar para hacer una primera exploración sobre la salud del paciente. Permitirá identificar los pathways biológicos más relevantes y se podrían brindar terapias dirigidas, lo que aumentaría la efectividad del diagnóstico y del tratamiento. Del lado de investigación, la identificación de estas vías biológicas podría iniciar nuevos enfoques en el área farmacológica o explorar el campo de la terapias génicas.

5.4 Futuras direcciones de investigación

Inicialmente este dataset es de una población concreta, pero se puede expandir ocupando información de más pacientes de diferentes lugares, esto nos permitirá tomar en cuenta factores genómicos, ambientales e incluso sociales, que hagan al modelo más robusto. Con los resultados obtenidos, podemos explorar más sobre las principales pathways biológicas que están relacionadas con esta condición, para poder entender qué papel juegan e incluso pensar en alguna terapia a futuro. También podemos evaluar las predicciones de nuestros modelos en un contexto clínico real (no ocupar la predicciones para realizar una decisión médica), esto nos permitirá expandir el modelo y haceer un análisis más riguroso.

6 Referencias

- Abel, E. D. (2021). Insulin signaling in the heart. American Journal of Physiology-Endocrinology and Metabolism, 321(1), E130–E145. https://doi.org/10.1152/ajpendo.00158.2021
- Aoyagi, T., & Matsui, T. (2011). Phosphoinositide-3 kinase signaling in cardiac hypertrophy and heart failure. Current Pharmaceutical Design, 17(18), 1818–1824. https://doi.org/10.2174/138161211796390976
- Averbuch, T., Sullivan, K., Sauer, A., Mamas, M. A., Voors, A. A., Gale, C. P., Metra, M., Ravindra, N., & Van Spall, H. G. C. (2022). Applications of artificial intelligence and machine learning in heart failure. European Heart Journal - Digital Health, 3(2), 311–322. https://doi.org/10.1093/ehjdh/ztac025
- Frangogiannis, N. G. (2019). The extracellular matrix in ischemic and nonischemic heart failure. Circulation Research, 125(1), 117–146. https://doi.org/10.1161/CIRCRESAHA.119.311148
- Kwon, J. M., Kim, K. H., Jeon, K. H., Lee, S. E., & Lee, H. Y. (2019). Artificial intelligence algorithm for predicting mortality of patients with acute heart failure. PLOS ONE, 14(7), e0219302. https://doi.org/10.1371/journal.pone.0219302

- Pingitore, A., Zhang, C., Vassalle, C., Ferragina, P., Landi, P., Mastorci, F., Sicari, R., Tommasi, A., Zavattari, C., Prencipe, G., & Sîrbu, A. (2024). Machine learning to identify a composite indicator to predict cardiac death in ischemic heart disease. International Journal of Cardiology, 404, 131981. https://doi.org/10.1016/j.ijcard.2024.131981