MACC
Matemáticas Aplicadas y
Ciencias de la Computación

Universidad del Rosario | Escuela de Ingeniería, Ciencia y Tecnología

Proyecto Final
Insuficiencia Cardíaca

Dafne Valeria Castellanos Rosas
Diryon Yonith Mora Romero
Fabio Andres Rizo Montoya
Laura Valentina Gonzalez Rodriguez

Prof. Juan Camilo Yepes Borrero

Análisis Estadístico de Datos
Matemáticas Aplicadas y Ciencias de la Computación
Universidad del Rosario
Mayo 2023

1. Introducción

Las enfermedades cardiovasculares se refieren a una variedad de trastornos que afectan el corazón y los vasos sanguíneos. Estas enfermedades incluyen condiciones como enfermedades coronarias, enfermedades de las arterias periféricas, arritmias cardíacas, enfermedades de las válvulas cardíacas y accidentes cerebrovasculares. En esencia, cualquier afección que afecte el corazón o los vasos sanguíneos se considera una enfermedad cardiovascular.

Este tipo de enfermedades son la principal causa de muerte a nivel mundial, siendo incluso más letales que el cáncer, cobrando la vida de alrededor de 18 millones de personas cada año. Particularmente, las enfermedades coronarias y los accidentes cerebrovasculares, representan más del 80 % de los casos. Además, un tercio de estas muertes ocurren en personas menores de 70 años, lo que indica una carga significativa de mortalidad prematura. [1]

La identificación de factores de riesgo específicos es esencial para la prevención y tratamiento de las enfermedades cardiovasculares. Sin embargo, existe una amplia lista de factores de riesgo, como la hipertensión, la diabetes, condiciones hereditarias, factores alimentarios y el estilo de vida de los pacientes. Esta diversidad de factores dificulta el análisis médico y la identificación precisa de la causa del desarrollo de una enfermedad cardiovascular.

En este proyecto, se abordarán los factores de riesgo y los marcadores biológicos utilizados en la evaluación médica de las enfermedades cardiovasculares. Entre los factores de riesgo considerados se encuentran la diabetes, la hipertensión, la anemia, el hábito de fumar y el género del paciente. Además, se analizarán marcadores biológicos como la fosfokinasa-creatinina, la fracción de expulsión, la cantidad de plaquetas, la creatinina y el sodio en el suero sanguíneo, que brindan información relevante sobre la enfermedad cardiovascular.

2. Problema a Analizar

Las enfermedades cardiovasculares (ECV) es la principal causa de muerte a nivel mundial, cobrando la vida de aproximadamente 17.9 millones de personas cada año, lo que representa el 31 % de todas las muertes en el mundo. Las ECV duplican la tasa de mortalidad de todos los tipos de cáncer combinados, lo que coloca a las ECV como la primera causa de muerte en el mundo y se estima que para 2030, cerca de 23,6 millones de personas podrían morir a causa de una de estas enfermedades. [2]

Según se puede apreciar en la figura 1, durante el período comprendido entre 1990 y 2019, las enfermedades cardiovasculares han sido la enfermedad con más casos a nivel mundial. Se estima que más de 350 millones de personas padecieron de estas enfermedades hasta el año 2019. Esta cifra resalta la importancia y el impacto significativo que las enfermedades cardiovasculares han tenido en la mortalidad global. [3]

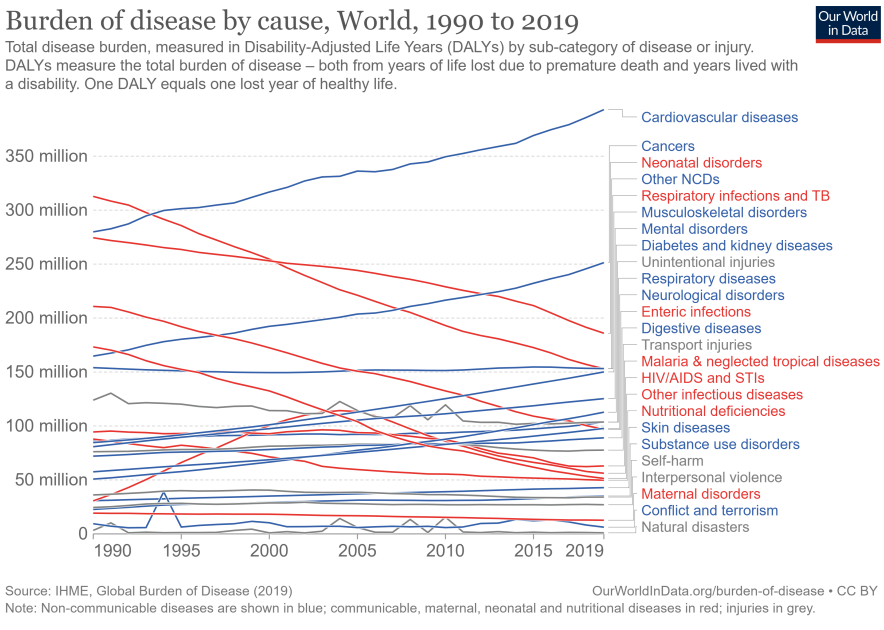


Figura 1: Carga de enfermedad por causa desde 1990 a 2019 a nivel global.

Las enfermedades cardiovasculares representan una grave problemática a nivel mundial. Existe un alto riesgo para muchas personas asintomáticas que presentan múltiples factores de riesgo, como se ven representados en la figura 2 [4], pero desafortunadamente, en la mayoría de los casos, estos factores no son adecuadamente controlados. En países como España, más de un tercio de los pacientes con infarto agudo de miocardio fallecen antes de recibir un tratamiento eficaz, lo que indica que la prevención y el tratamiento se han abordado tardíamente. [5] Por lo tanto, es crucial promover la prevención primaria de las enfermedades cardiovasculares y dar prioridad a la identificación y control de los factores de riesgo específicos.

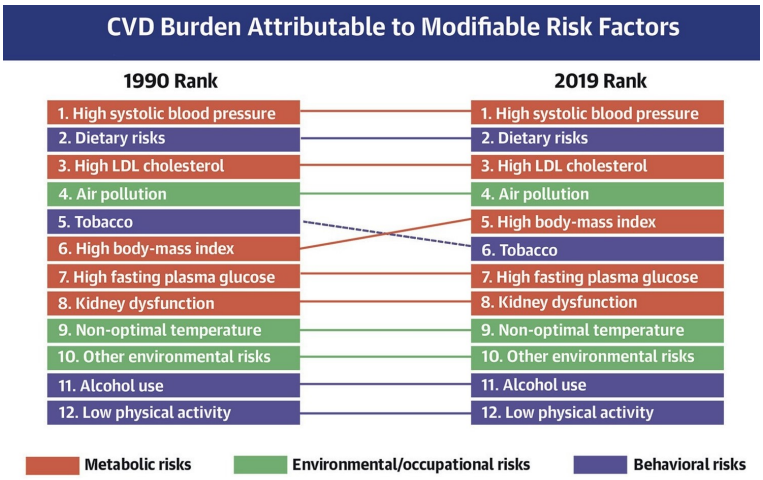


Figura 2: Carga de ECV atribuible a factores de riesgo modificables.

Por lo expuesto anteriormente, el presente proyecto se centrará en analizar los datos de adultos nacidos en Pakistán entre 40 y 95 años que en el pasado habían presentado disfunción sistólica del ventrículo izquierdo y habían sufrido insuficiencia cardíaca. Y se plantea la siguiente pregunta de investigación: ¿Cuáles son los factores de riesgo que más contribuyen a la mortalidad por enfermedades cardiovasculares?

3. Objetivos

3.1. Objetivo General

Informar sobre los factores asociados a un mayor riesgo de mortalidad en casos de insuficiencia cardiaca, a través de la identificación de estos y la comprensión de su influencia en estas enfermedades. Para que con base en esta información se logre el desarrollo de estrategias efectivas de prevención y tratamiento, lo que permitirá aumentar las probabilidades de una vida más prolongada y saludable para los pacientes afectados por esta condición.

3.2. Objetivos Específicos

1. Analizar las correlaciones entre variables numéricas en pacientes con insuficiencia.

2. Determinar si existen diferencias significativas en los niveles de CPK, fracción de eyección, plaquetas, creatinina y sodio entre diferentes poblaciones de pacientes, como los pacientes fumadores, de diferentes edades y géneros.

3. Predecir la mortalidad por insuficiencia cardíaca utilizando múltiples variables predictoras de carácter numérico, por medio de diferentes métodos estadísticos, encontrando el mejor modelo que ayude en la toma de decisiones médicas.

4. Base de datos

Se analizó un conjunto de datos que contiene los registros médicos de 299 pacientes con insuficiencia cardíaca recopilados en el Instituto de Cardiología de Faisalabad y en el Hospital Aliado de Faisalabad (Punjab, Pakistán) entre abril y diciembre de 2015. Los pacientes incluyen a 105 mujeres y 194 hombres, con edades comprendidas entre 40 y 95 años. Todos los pacientes tenían disfunción sistólica del ventrículo izquierdo y habían sufrido insuficiencia cardíaca previa, lo que los ubicaba en las clases III o IV de la clasificación de la Asociación de Corazón de Nueva York (NYHA) de las etapas de la insuficiencia cardíaca. [6]

Columna	Explicación	Medida	Rango
Age	Edad del paciente	Años	[40;95]
Anaemia	Si el paciente tiene anemia.	Booleano	[0,1]
High blood pressure	Si el paciente tiene hipertensión.	Booleano	[0,1]
CPK	Nivel de la enzima CPK en la sangre.	mcg/L	[23;7861]
Diabetes	Si el paciente sufre diabetes.	Booleano	[0,1]
Ejection fraction	Cantidad de sangre que el corazón expulsa del ventrículo izquierdo en cada contracción.	Porcentaje	[14;80]
Sex	Hombre (1) o mujer (0).	Binario	[0,1]
Platelets	Plaquetas en la sangre.	kiloplaquetas/mL	[25.01;850.00]
Serum creatinine	Nivel de creatinina en la sangre.	mg/dL	[0.50;9.40]
Serum sodium	Nivel de sodio en la sangre.	mEq/L	[114;148]
Smoking	Si el paciente es fumador.	Booleano	[0,1]
Time	Período de seguimiento.	Días	[4;25]
Death event	Si el paciente falleció durante el periodo de seguimiento.	Booleano	[0,1]

Tabla 1: Resumen de la información

El conjunto de datos contiene 12 características que pueden utilizarse para predecir la mortalidad por insuficiencia cardíaca. [7] Dicha información se encuentra en la Tabla 1.

5. Metodología

Objetivo 1: Relaciones entre conjuntos de variables.

Con el fin de explorar las posibles relaciones entre las variables numéricas y la mortalidad en pacientes con insuficiencia cardíaca, se utilizó el análisis de correlación canónica (CCA). Una técnica estadística multivariada que permite estudiar las relaciones lineales entre dos conjuntos de variables. Investigando la asociación conjunta entre dos grupos de variables, en lugar de considerar cada grupo por separado.

El objetivo principal del CCA es encontrar combinaciones lineales de las variables en cada grupo que estén altamente correlacionadas entre sí y maximicen la correlación entre los dos grupos. En otras palabras, busca identificar los patrones de asociación más fuertes entre los dos conjuntos de variables. Cuando se aplica el CCA, se generan una serie de correlaciones canónicas. Cada correlación canónica representa una combinación lineal de las variables de ambos grupos que maximiza la correlación entre los dos grupos.

Con lo cual, se procedió a implementar el código en R. En primer lugar, se realizó una estandarización de las variables seleccionadas del conjunto de datos 'hfc_data', utilizando la función 'scale()', la cual estandariza las variables numéricas restando la media y dividiendo por la desviación estándar. El resultado se almacenó en el objeto 'hfc_data_scaled'. A continuación, se crearon dos grupos de variables llamados 'grupo1' y 'grupo2'. El 'grupo1' incluyó las variables 'ejection_fraction', 'serum_sodium' y 'platelets', mientras que el 'grupo2' contuvo las variables 'age', 'serum_creatinine' y 'creatinine_phosphokinase'.

Después de definir los grupos, se aplicó el análisis de correlación canónica (CCA) utilizando la función 'cc()'. Esta función tomó como argumentos los dos grupos de variables. Para obtener información sobre las correlaciones canónicas resultantes, se accedió al atributo 'cor' del objeto 'cca'. Esto se realizó mediante 'cca$cor', lo cual devolvió una matriz que

mostraba las correlaciones canónicas entre los grupos de variables. Para después, obtener los vectores de carga canónica (loadings) del CCA utilizando el atributo 'scores'. Estos vectores representaban las combinaciones lineales de las variables en cada grupo que maximizaban las correlaciones canónicas.

Por último, se utilizó la función 'plt.cc()' para generar un gráfico de correlación canónica (CCA plot). Este gráfico visualizó la relación entre los grupos de variables y mostró la ubicación relativa de las variables en un espacio de correlación canónica. Se incluyeron etiquetas de variables en el gráfico para identificar cada variable.

Objetivo 2: Diferencia de medias significativa entre diversos grupos

Con el interés de determinar si existían diferencias significativas en los niveles de CPK (creatinine phosphokinase), fracción de eyección, plaquetas, creatinina y sodio entre diferentes grupos de pacientes, específicamente evaluando las diferencias asociadas al hábito de fumar, los rangos de edad y el género de los pacientes, se empleó el Análisis de Varianza Multivariante (MANOVA). El MANOVA es una técnica estadística utilizada para evaluar si existen diferencias significativas en múltiples variables dependientes continuas entre diferentes grupos o niveles de una variable independiente o predictor. Para llevar a cabo esto, se utilizo la función 'manova()'.

En el primer análisis, se crea el modelo llamado 'smoking_model' utilizando la fórmula 'cbind(creatinine_phosphokinase, ejection_fraction, platelets, serum_creatinine, serum_sodium) ∽ smoking', donde 'smoking' representa el hábito de fumar y las variables separadas por comas representan las variables dependientes. Utilizamos el conjunto de datos 'hfc_data' para realizar el análisis. Luego, se utiliza la función 'summary()' para obtener un resumen de los resultados del MANOVA y examinar las diferencias significativas entre los grupos de fumadores y no fumadores en relación con las variables dependientes.

El mismo proceso se repitió con los rangos de edad y el género. En el segundo análisis, se crea el modelo llamado 'age_group_model' utilizando la fórmula 'cbind(creatinine_phosphokinase, ejection_fraction, platelets, serum_creatinine, serum_sodium) ∽ age_group', donde 'age_group' representa los rangos de edad de los pacientes. Y en el tercer análisis, se crea un modelo llamado 'sex_model' utilizando la fórmula 'cbind(creatinine_phosphokinase, ejection_fraction, platelets, serum_creatinine, serum_sodium) ∽ sex', donde 'sex' representa el género.

Los resultados de los MANOVAs proporcionaron información importante sobre las diferencias en la fisiología y las características clínicas de los pacientes fumadores, de diferentes rangos de edad y géneros. Una vez se encuentra evidencias de diferencia en la media a través de un p-value inferior a un $\alpha = 0{,}05$, se establecen qué variables contribuyeron significativamente a estas diferencias, utilizando para este fin un Análisis de la Varianza (ANOVA), aislando por separado cada una de las variables y seleccionando nuevamente aquellas donde su p-value sea inferior a un $\alpha = 0{,}05$.

Objetivo 3: Probabilidad de mortalidad

Con el objetivo de comprender la probabilidad de mortalidad en pacientes con insuficiencia cardíaca y determinar qué variables podrían ser predictivas de dicho desenlace, se decidió utilizar técnicas de regresión logística y discriminante de Fisher. Iniciando con la regresión logística, es una técnica ampliamente utilizada para modelar la probabilidad de un evento binario, en este caso, la mortalidad. Donde se relaciona un conjunto de variables independientes (predictores) con una variable dependiente binaria (variable de respuesta) y estimar la probabilidad de que ocurra el evento de interés.

Para esto, se selccionaron las variables de interés, que incluían el nivel de CPK, la fracción de eyección, las plaquetas, el nivel de creatinina y el sodio, como predictores potenciales de la mortalidad. Para ello, se crea el conjunto 'hfc_data_fish', donde se eliminan las variables

categóricas, excepto 'DEATH_EVENT'. A continuación, se ajusta un modelo de regresión logística utilizando la función 'glm()', especificando 'DEATH_EVENT ∽ .' en la fórmula para indicar que se quiere predecir la variable de respuesta 'DEATH_EVENT' utilizando todas las demás variables numéricas.

Posteriormente, se realiza un análisis adicional utilizando el Análisis Discriminante de Fisher (ADF). El objetivo de este análisis era encontrar una función discriminante que nos permitiera clasificar correctamente nuevas observaciones en las categorías de "muere.º "vive". El discriminante de Fisher busca maximizar la separación entre los grupos mientras minimiza la variación dentro de cada grupo. Esta técnica utiliza la matriz de covarianza de las variables independientes y las medias de cada grupo para calcular la función discriminante.

Para realizar esto, se utiliza la función 'lda()' en R, que corresponde al Análisis Discriminante Lineal, que permite implementar tanto el ADL y el ADF. Se deben cambiar los parámtros de 'prior' y 'storecov' para asumir que las matrices de covarianza se consideren iguales entre las clases. Retomando la función 'lda()' se especifica 'DEATH_EVENT ∽ .' en la fórmula para indicar la predicción de la variable 'DEATH_EVENT' utilizando todas las demás variables numéricas en el conjunto de datos 'hfc_data_fish'.

Una vez que se obtuvieron los resultados de ambos modelos, se utiliza la función 'summary()' para obtener un resumen detallado de los coeficientes, las estadísticas de prueba y otros diagnósticos relevantes. Esto permitió evaluar la significancia relativa de cada variable en la predicción de la mortalidad y comprender mejor la naturaleza de su influencia en el desenlace. Además, se realizó una última regresión logística utilizando todas las variables en el conjunto de datos original 'hfc_data', excepto la variable 'DEATH_EVENT'. Este análisis ampliado permitió evaluar si había otras variables, más allá de las seleccionadas previamente, que podrían ser predictivas de la mortalidad en pacientes con insuficiencia cardíaca.

6. Análisis de Resultados

Primero se carga la base de datos y se hacen los correspondientes ajustes:

```
hfc_data = read.csv("dataset/heart_failure_clinical_records.csv")
hfc_data = na.omit(hfc_data) # Eliminar datos nulos
hfc_data$age = round(hfc_data$age) # Redondear ed
unused_columns = c("time") # Eliminar columnas no necesarias
hfc_data = hfc_data[, !(names(hfc_data) %in% unused_columns)]
# Subgrupo sin categóricas
cat_columns <- c("anaemia", "diabetes","high_blood_pressure", "sex",
"smoking","age_group", "DEATH_EVENT")
hfc_data_no_cat <- hfc_data[, !(names(hfc_data) %in% cat_columns)]
intervalos = c(0,69,Inf)
etiquetas = c(0, 1)
hfc_data$age_group = cut(hfc_data$age, breaks = intervalos, labels = etiquetas)
```

Objetivo 1: Correlación entre las variables.

Las variables numéricas se separaron en dos grupos aleatorios, el primer grupo con ejection_fraction, serum_sodio y platelets, mientras que el segundo grupo se separo en age, serum_creatinine y cratinine_phosphokinase. A estos grupos se les aplico CCA.

```
hfc_data_scaled = as.data.frame(scale(hfc_data[, c("ejection_fraction", "serum_sodium","platelets","age","serum_creatinin
e","creatinine_phosphokinase")]))
grupo1 <- hfc_data_scaled[, c("ejection_fraction", "serum_sodium","platelets")]
grupo2 <- hfc_data_scaled[, c("age","serum_creatinine","creatinine_phosphokinase")]
cca=cc(grupo1, grupo2)
cca$cor
```

Los valores presentados a continuación representan las correlaciones canónicas obtenidas en el análisis. En un CCA, se generan varias dimensiones canónicas, y estos valores muestran la correlación entre las dimensiones correspondientes en los dos conjuntos de variables. Estos valores de correlación canónica son bajos, oscilando entre 0.207 y 0.014. Esto indica una relación débil entre los conjuntos de variables o una falta de estructura clara en los datos.

```
## [1] 0.20725755 0.08453181 0.01445662
```

La matriz a continuación muestra las correlaciones entre las variables originales del primer conjunto de datos y los puntajes canónicos del primer conjunto de datos. Los puntajes canónicos son las variables transformadas que representan las dimensiones canónicas. Las variables ejection_fraction, serum_sodium y platelets están correlacionadas con los puntajes canónicos de la primera dimensión en -0.07460701, 0.94471986 y 0.26097321 respectivamente. De manera similar, estas variables también están correlacionadas con los puntajes canónicos de la segunda y tercera dimensión. Observando que para la primera dimensión, la variable más importante es serum_sodio.

```
## $corr.X.xscores
##                          [,1]       [,2]        [,3]
## ejection_fraction -0.07460701 -0.8621680  0.50109896
## serum_sodium       0.94471986 -0.3219248 -0.06219978
## platelets          0.26097321  0.4034243  0.87700729
```

La matriz a continuación muestra las correlaciones entre las variables originales del segundo conjunto de datos y los puntajes canónicos del primer conjunto de datos. Las variables age, serum_creatinine y creatinine_phosphokinase están correlacionadas con los puntajes canónicos de la primera dimensión en -0.07216118, -0.19085814 y 0.07512076 respectivamente. De manera similar, estas variables también están correlacionadas con los puntajes canónicos de

la segunda y tercera dimensión. Así, para la primera dimensión, la variable más importante es serum_creatinine.

```
## $corr.Y.xscores
##                                [,1]        [,2]         [,3]
## age                      -0.07216118 -0.06752670 -0.007091769
## serum_creatinine         -0.19085814  0.02787822 -0.003005640
## creatinine_phosphokinase  0.07512076  0.03775257 -0.011825927
```

La matriz a continuación muestra las correlaciones entre las variables del primer conjunto de datos X y los puntajes canónicos del segundo conjunto de datos. Las variables ejection_fraction, serum_sodium y platelets están correlacionadas con los puntajes canónicos de la primera dimensión en -0.01546287, 0.19580032 y 0.05408867 respectivamente. De manera similar, estas variables también están correlacionadas con los puntajes canónicos de la segunda dimensión. Así, para la primera dimensión, la variable más importante es serum_sodio.

```
## $corr.X.yscores
##                            [,1]        [,2]          [,3]
## ejection_fraction -0.01546287 -0.07288062  0.0072441983
## serum_sodium       0.19580032 -0.02721288 -0.0008991987
## platelets          0.05408867  0.03410219  0.0126785629
```

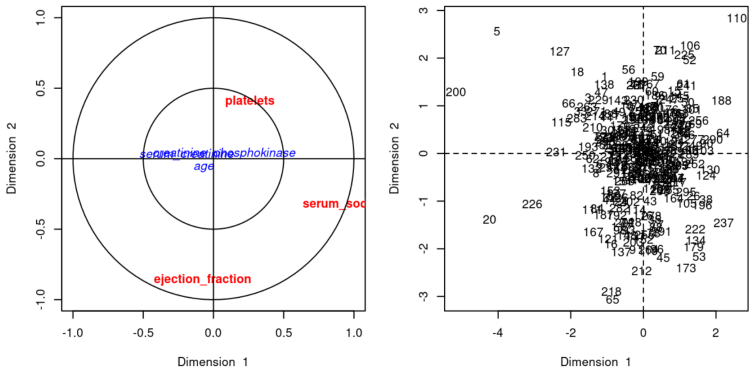
La matriz a continuación muestra las correlaciones entre las variables del segundo conjunto de datos y los puntajes canónicos del segundo conjunto de datos. Las variables age, serum_creatinine y creatinine_phosphokinase están correlacionadas con los puntajes canónicos de la primera dimensión en -0.3481715, -0.9208742 y 0.3624513 respectivamente. De manera similar, estas variables también están correlacionadas con los puntajes canónicos de la segunda y tercera dimensión. Así, para la primera dimensión, la variable más importante es serum_creatinine.

```
## $corr.Y.yscores
##                                [,1]       [,2]       [,3]
## age                      -0.3481715 -0.7988319 -0.4905550
## serum_creatinine         -0.9208742  0.3297956 -0.2079075
## creatinine_phosphokinase  0.3624513  0.4466079 -0.8180284
```

Se pudo observar que algunos de los coeficientes de correlación son altos, como 0.94471986 (serum_sodio en el primer grupo) y -0.9208742 (serum_creatininte el el segundo grupo), lo que indica una fuerte relación entre esas variables y los puntajes canónicos correspondientes. Sin embargo, también hay coeficientes de correlación más bajos, como -0.07460701 (ejection_fraction), lo que sugiere una relación más débil.

Mientras que las correlaciones entre las variables originales de un conjunto de datos con los puntajes canónicos del otro conjunto de datos indican cómo se relacionan los dos conjuntos de variables. En particular, hay algunas correlaciones significativas, como -0.19085814 (serum_creatinine) y 0.19580032(serum_sodio), lo que sugiere una relación entre las variables del un grupo y los puntajes canónicos del otro grupo de datos. Sin embargo, es importante tener en cuenta que las correlaciones son relativamente bajas en general.

```
## [1] "plt.cc(cca, var.label = TRUE)"
```



Tras analizar dos pares de correlaciones canónicas, se observa una correlación muy baja entre los datos, aproximadamente del 30 %. Al graficar la dimensión 1 contra la dimensión 2,

se aprecia que las variables platelets, ejection_fraction y serum_sodio muestran una correlación baja entre sí. Por otro lado, las variables age, serum_creatinine y creatinine_phosphokinase quedaron justo en el centro del gráfico. Cuando esto sucede, puede ser porque tienen una correlación promedio con el resto de las variables o que no presentan patrones claros de correlación con otras variables específicas.

Objetivo 2: Diferencias significativas entre diversos grupos

6.0.1. Fumadores

Con un $\alpha = 0{,}05$, podemos apreciar que no hay diferencia significativa entre las personas fumadoras y no fumadoras en lo que respecta a la media de sus registros de CPK, la fracción de expulsión, sus plaquetas, creatinina y sodio ($0{,}8613 > \alpha$). Así, no es necesario indagar en las diferencias de cada una de las variables.

```
smoking_model = manova(cbind(creatinine_phosphokinase, ejection_fraction,
platelets, serum_creatinine, serum_sodium) ~ smoking, data=hfc_data)
summary(smoking_model)

##            Df   Pillai approx F num Df den Df Pr(>F)
## smoking     1 0.006468  0.38149      5    293 0.8613
## Residuals 297
```

6.0.2. Mayores de 70 años

Con un $\alpha = 0{,}05$, se puede observar nuevamente que no hay diferencia significativa entre las personas mayores de 70 y aquellas que no en lo que respecta a la media de las mediciones

de sus factores biológicos ($0{,}5424 > \alpha$). Por tanto, no se indaga en la comparación individual de estos.

```
age_group_model <- manova(cbind(creatinine_phosphokinase, ejection_fraction,
platelets, serum_creatinine, serum_sodium) ~ age_group , data=hfc_data)
summary(age_group_model)

##            Df   Pillai approx F num Df den Df Pr(>F)
## age_group   1 0.013654  0.81122      5    293 0.5424
## Residuals 297
```

6.0.3. Género

Finalmente, con un $\alpha = 0{,}05$ vemos que sí hay diferencia significativa entre las personas de género masculino y femenino ($0{,}0341 < \alpha$):

```
sex_model <- manova(cbind(creatinine_phosphokinase, ejection_fraction,
platelets, serum_creatinine, serum_sodium) ~ sex , data=hfc_data)
summary(sex_model)

##            Df   Pillai approx F num Df den Df  Pr(>F)
## sex         1 0.041038   2.5077      5    293 0.03041 *
## Residuals 297
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Por tanto, procedemos a verificar si hay diferencia significativa en este grupo en cada una de las variables.

Creatinine Phosphokinase Con CPK, podemos apreciar que no hay diferencia significativa ($0{,}169 > \alpha$)

```
summary(aov(creatinine_phosphokinase ~ sex, hfc_data))

##              Df    Sum Sq Mean Sq F value Pr(>F)
## sex           1   1786164 1786164   1.903  0.169
## Residuals   297 278768491  938614
```

Ejection Fraction Con la fracción de expulsión es visible una diferencia en las medias ($0{,}0102 < \alpha$)

```
summary(aov(ejection_fraction ~ sex, hfc_data))

##              Df Sum Sq Mean Sq F value Pr(>F)
## sex           1    919   919.0   6.687 0.0102 *
## Residuals   297  40820   137.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Platelets Sí se puede observar diferencia significativa ($0{,}0305 < \alpha$)

```
summary(aov(platelets ~ sex, hfc_data))

##              Df    Sum Sq   Mean Sq F value Pr(>F)
## sex           1 4.463e+10 4.463e+10   4.724 0.0305 *
## Residuals   297 2.806e+12 9.448e+09
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Serum Creatinine No hay diferencia significativa ($0{,}904 > \alpha$)

```
summary(aov(serum_creatinine ~ sex, hfc_data))

##              Df Sum Sq Mean Sq F value Pr(>F)
## sex           1    0.0  0.0155   0.014  0.904
## Residuals   297  318.9  1.0738
```

Serum Sodium No hay diferencia significativa ($0{,}635 > \alpha$)

```
summary(aov(serum_sodium ~ sex, hfc_data))

##              Df Sum Sq Mean Sq F value Pr(>F)
## sex           1      4   4.409   0.226  0.635
## Residuals   297   5798  19.521
```

Así, tenemos que entre personas hay suficiente evidencia estadística para apoyar la hipótesis de que en las poblaciones género masculino y femenino hay una diferencia significativa en sus niveles de fracción de expulsión y plaquetas. Además, estos indicios de diferencia biológica podría despertar el interés de centrar análisis haciendo la distinción entre estas dos poblaciones.

Objetivo 3: Probabilidad de mortalidad

6.0.4. Resultados de los ajustes

Primero se ajustan los datos para obtener los conjuntos de entrenamiento y prueba.

```
hfc_data_fish = hfc_data_no_cat;
hfc_data_fish$DEATH_EVENT = hfc_data$DEATH_EVENT;
set.seed(1)
train_index = sample(c(TRUE, FALSE), nrow(hfc_data_fish), replace=TRUE,
prob=c(0.7, 0.3));
hfc_data_fish_train = hfc_data_fish[train_index, ]
hfc_data_fish_test = hfc_data_fish[!train_index, ]
hfc_data_train = hfc_data[train_index, ]
hfc_data_test = hfc_data[!train_index, ]
```

Iniciamos con una regresion logistica, solo considerando las variables numericas de la base de datos.

```
modelo5=glm(DEATH_EVENT ~., data = hfc_data_fish_train)
summary(modelo5)

##
## Call:
## glm(formula = DEATH_EVENT ~ ., data = hfc_data_fish_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -0.8387  -0.2861  -0.1356   0.3551   1.0427
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.821e+00  9.223e-01   1.974  0.04966 *
## age                       1.040e-02  2.344e-03   4.437 1.47e-05 ***
## creatinine_phosphokinase  5.604e-05  3.311e-05   1.692  0.09206 .
## ejection_fraction        -1.012e-02  2.494e-03  -4.056 7.05e-05 ***
## platelets                 2.788e-07  2.800e-07   0.996  0.32056
## serum_creatinine          8.189e-02  2.956e-02   2.770  0.00611 **
## serum_sodium             -1.449e-02  6.688e-03  -2.167  0.03138 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1637064)
##
##     Null deviance: 45.530  on 216  degrees of freedom
## Residual deviance: 34.378  on 210  degrees of freedom
## AIC: 232
##
## Number of Fisher Scoring iterations: 2
```

Al realizar la regresión logística se evidencia que los factores que más influyen en determinar si un paciente con ECV va a morir son la edad, la fracción de expulsión y los niveles de creatinina en la sangre.

Una vez realizado el modelo, ya se pueden integrar datos de prueba con el fin de determinar

la probabilidad que tiene un paciente de morir debido a su ECV.

Finalmente, utilizando ahora el discriminante de Fischer seleccionando únicamente las variables más significativas obtenidas de las regresiones logísticas (edad, creatinina en la sangre y fracción de expulsión cardiaca), obtenemos:

```
modelo6 <- lda(DEATH_EVENT ~ age + ejection_fraction + serum_creatinine,
data=hfc_data_fish_train);
modelo6
```

```
## Call:
## lda(DEATH_EVENT ~ age + ejection_fraction + serum_creatinine,
##     data = hfc_data_fish_train)
##
## Prior probabilities of groups:
##         0         1
## 0.7004608 0.2995392
##
## Group means:
##        age ejection_fraction serum_creatinine
## 0 58.36842          40.56579         1.203882
## 1 66.49231          33.69231         1.805846
##
## Coefficients of linear discriminants:
##                             LD1
## age                  0.05448980
## ejection_fraction -0.05720335
## serum_creatinine    0.50347946
```

Podemos apreciar las diferencias de las medias entre el grupo de los pacientes que realizaron una recuperación completa y aquellos que fallecieron. De las variables significativas, nos damos cuenta que las personas que murieron eran longevos, poseían un mayor nivel de creatinina en la sangre y su fracción de expulsión cardíaca era menor comparado a aquellos pacientes que se recuperaron.

6.0.5. Evaluación de los ajustes

Naturalmente los modelos de regresión logística no producen predicciones binarias. En su lugar, producen una posibilidad de pertenecer justamente a una categoría binaria. Así, no existe en un inicio el concepto de predecir si un individuo definitivamente morirá o no. No obstante, si asumimos que una probabilidad mayor al 50 % le generará la muerte al individuo y análogamente con la condición contraria, obtenemos una forma de evaluar un modelo de este tipo, con lo que obtenemos:

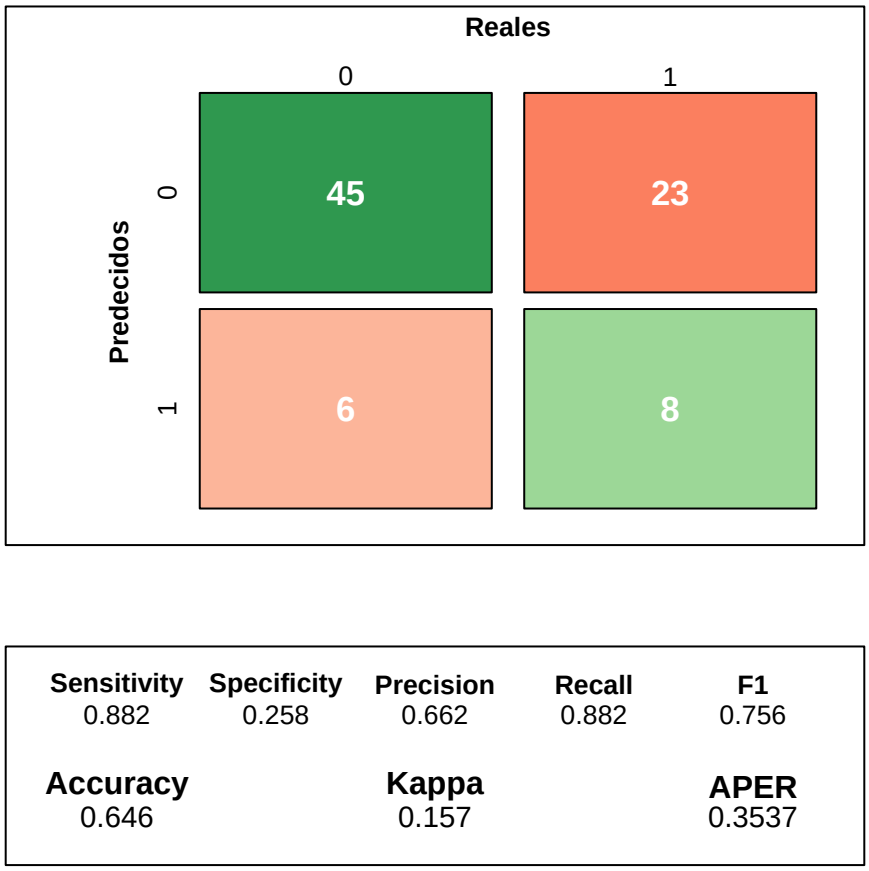


Figura 3: Regresión Logística I

Donde se puede observar una especificidad anormalmente baja, indicándonos la tendencia de clasificar la conjunción de los predictores como no letales.

Por otro lado, un ajuste obtenido de un discriminante lineal de Fisher sí busca, valga la redundancia, discriminar las categorías que comprende su salida, en este caso el estar muerto o vivo. Así, inherentemente sus predicciones tendrán una clase binaria asociada, permitiendo evaluar el modelo fácilmente. Con él, obtuvimos:

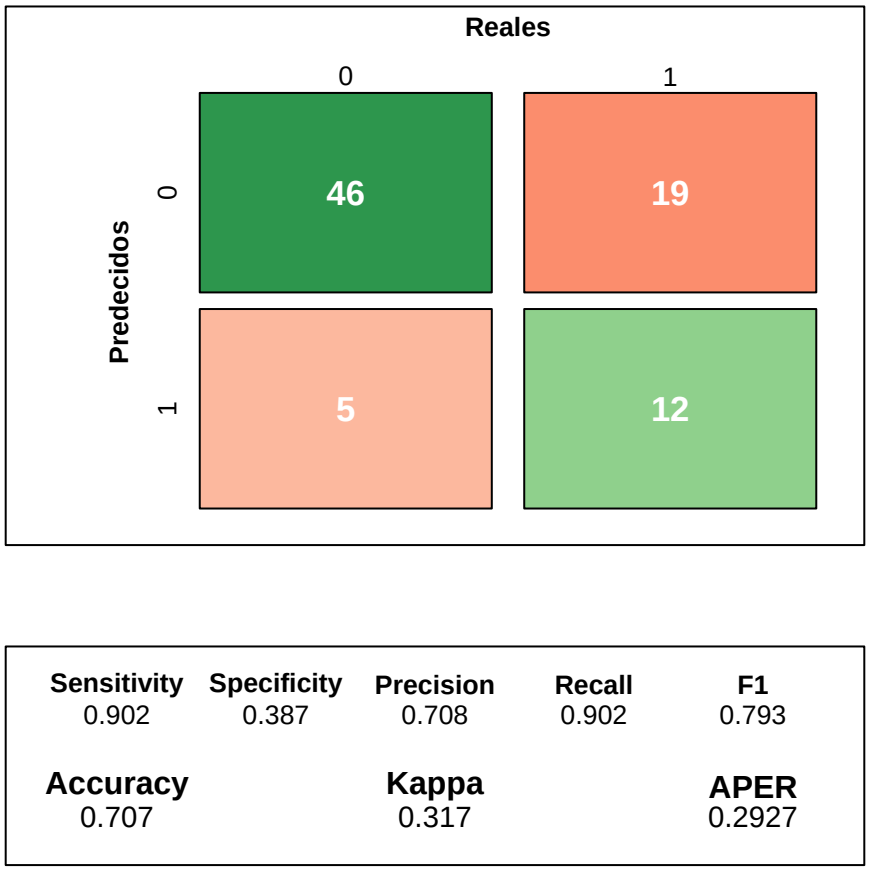


Figura 4: Discriminante de Fisher

Así, podemos observar que los ajustes obtenidos logran, en el mejor de los casos, predecir con una precision del 70 % si un individuo va a morir o no. No obstante, estos tienen una especificidad muy baja, lo que significa que una persona que muy probablemente morirá, no está bajo riesgo de mortalidad según tales modelos.

7. Conclusiones

Se encontraron diferencias significativas entre las poblaciones de mujeres y hombres al comparar sus marcadores biológicos, específicamente en los factores de plaquetas y fracción de expulsión. Esto indica que se podría hacer una investigación separando estos dos grupos para encontrar nuevos modelos de predicción que se ajusten a sus diferencias.

En cambio, las poblaciones de fumadores y edades no presentan suficiente evidencia estadística como para aceptar la hipótesis de que se presentan diferencias entre sus factores biológicos. Debido a esto, no tendría sentido, aunque sea con estos datos, realizar análisis por separado en estas poblaciones.

Se encontró que los factores más determinantes en la muerte por insuficiencia cardiaca son la fracción de eyección, la edad y el nivel de creatinina en la sangre, esto se logra evidenciar gracias a la regresión logística realizada entre las variables de la base y el Death-event. También se utilizó el discriminante de Fischer para generar un modelo más preciso de predicción y obtener los promedios de las variables significativas para los dos grupos, muertos y vivos. Esto con el fin de identificar factores de alarma al registrar los datos de un paciente.

Al evaluar los modelos, obtuvimos una precisión del 70 %, lo cual es un buen indicativo para la eficacia de estos. No obstante, cuentan con una muy baja especificidad, implicando así que tienen un sesgo a calificar como no letal la mayoría de predictores. Así, con el conjunto de datos suministrado no es posible plantear un modelo lineal que logre predecir la mortalidad satisfactoriamente.

Adicionalmente, el análisis de correlación revela diferentes niveles de relación entre los conjuntos de variables analizados. Algunos coeficientes de correlación muestran una fuerte asociación entre ciertos conjuntos de variables y los puntajes canónicos correspondientes, mientras que otros coeficientes indican una relación más débil. Esto sugiere que las variables pueden tener una influencia limitada en los puntajes canónicos y que otros factores podrían

estar influyendo en los resultados.

Finalmente, una vez se consideran todos estos análisis y resultados obtenidos, se pueden realizar varias recomendaciones en el tratamiento de pacientes con enfermedad cardiovascular (ECV). Sin embargo, es importante destacar que el modelo de regresión mencionado anteriormente tiene una especificidad muy baja. Por ello, para mejoras en el proyecto a futuro, es fundamental analizar las variables más significativas y utilizar métodos más adecuados, como modelos de clasificación o técnicas de aprendizaje automático, que puedan proporcionar una predicción más precisa y confiable. Al igual, se sugiere utilizar otros enfoques y modelos más precisos para determinar el riesgo de mortalidad de los pacientes con ECV.

Estos modelos deben enfocarse en identificar a aquellos pacientes que requieren atención más urgente y brindarles el tratamiento adecuado. Así mismo, es importante aclarar que estos modelos pueden predecir la probabilidad de morir una vez que se tiene una ECV, pero no evalúan la probabilidad de desarrollarla. Además, no se deben utilizar para determinar quiénes no necesitan tratamiento, ya que todos los pacientes con ECV requieren atención médica y seguimiento adecuado. Los resultados de este estudio no implican que las personas con una baja probabilidad de morir no deban recibir atención médica o ser descuidadas en su tratamiento.

Referencias

[1] O. M. de la Salud, "Enfermedades cardiovasculares," 2021. [Online]. Available: https://www.who.int/es/health-topics/cardiovascular-diseases#tab=tab_1

[2] O. P. de la Salud. Día mundial del corazón: Enfermedades cardiovasculares causan 1,9 millones de muertes al año en las américas. (2012, Sep). [Online]. Available: https://www3.paho.org/hq/index.php?option=com_content&view=article&id=7257

[3] M. Roser, H. Ritchie, and F. Spooner, "Burden of disease," *Our World in Data*, 2021, https://ourworldindata.org/burden-of-disease.

[4] "Global burden of cardiovascular diseases and risk factors, 1990–2019: Update from the gbd 2019 study," *Journal of the American College of Cardiology*, vol. 76, DOI https://doi.org/10.1016/j.jacc.2020.11.010, no. 25, pp. 2982–3021, 2020.

[5] J. R. Banegas, F. Villar, A. Graciani, and F. Rodríguez-Artalejo, "Epidemiología de las enfermedades cardiovasculares en españa," *Revista Española de Cardiología Suplementos*, vol. 6, DOI https://doi.org/10.1016/S1131-3587(06)75324-9, no. 7, pp. 3G–12G, 2006, tratamiento de las hiperlipemias en pacientes con riesgo cardiovascular elevado. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1131358706753249

[6] G. J. Davide Chicco, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Med Inform Decis Mak.*, vol. 20, DOI https://doi.org/10.1186/s12911-020-1023-5, 2020. [Online]. Available: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5#citeas

[7] S. H. B. Tanvir Ahmad, Assia Munir. Conjunto de datos. (2017). [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data?resource=download