UNIVERSIDAD
NACIONAL
DE COLOMBIA

PROYECTO **CULTURAL, CIENTÍFICO Y COLECTIVO** DE NACIÓN

Modelos de Clasificación en la Identificación Temprana de Riesgo de Fallo Cardiaco

Ivan Santiago Roja Martinez

Introducción a Análisis Multivariado
Facultad de ciencias - Estadistica
Sede Medellín

Universidad Nacional de Colombia
PROYECTO **CULTURAL, CIENTÍFICO Y COLECTIVO** DE NACIÓN

Temario

- Contextualización de la Base de Datos

- Análisis Descriptivo y Exploratorio de los datos

- Modelos de clasificación: LDA - Discriminante Lineal de fisher, Regresión Logística y KNN.

Introducción

Las enfermedades cardiovasculares son la primera causa de muerte en el mundo: se calcula que cada año se cobran 17,9 millones de vidas, lo que representa el 31% de todas las muertes en el mundo. Cuatro de cada cinco muertes por enfermedades cardiovasculares se deben a infartos de miocardio y accidentes cerebrovasculares, y un tercio de estas muertes se producen prematuramente en personas menores de 70 años. Las personas con enfermedades cardiovasculares o que corren un alto riesgo cardiovascular (debido a la presencia de uno o más factores de riesgo como hipertensión, diabetes, hiperlipidemia o enfermedad ya establecida) necesitan una detección y gestión temprana.

Base de Datos

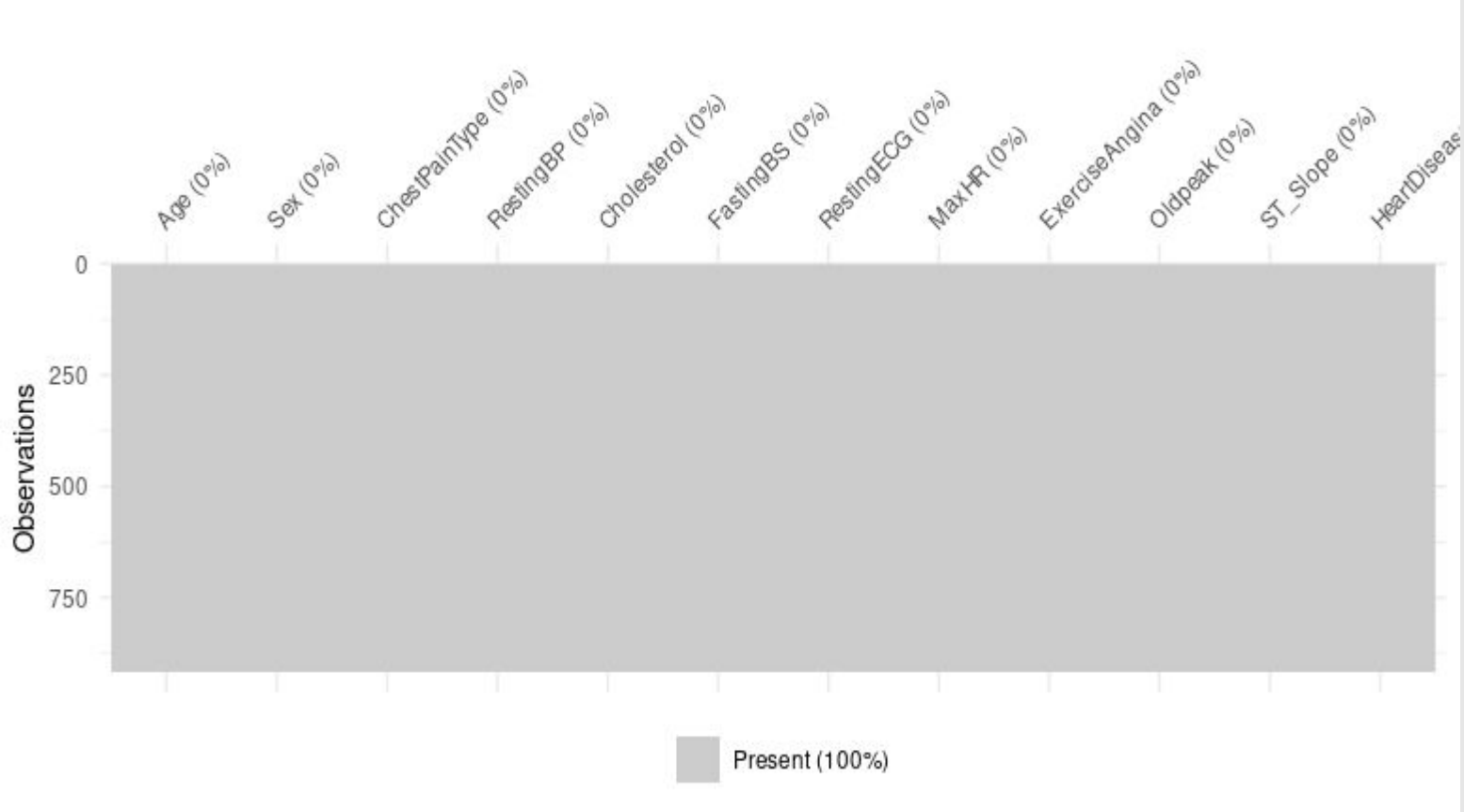
- **Kaggle:** Heart Failure Prediction Dataset[1]
- **Dimensión:** 12 variables - 918 Observaciones

Variables de la Base de Datos

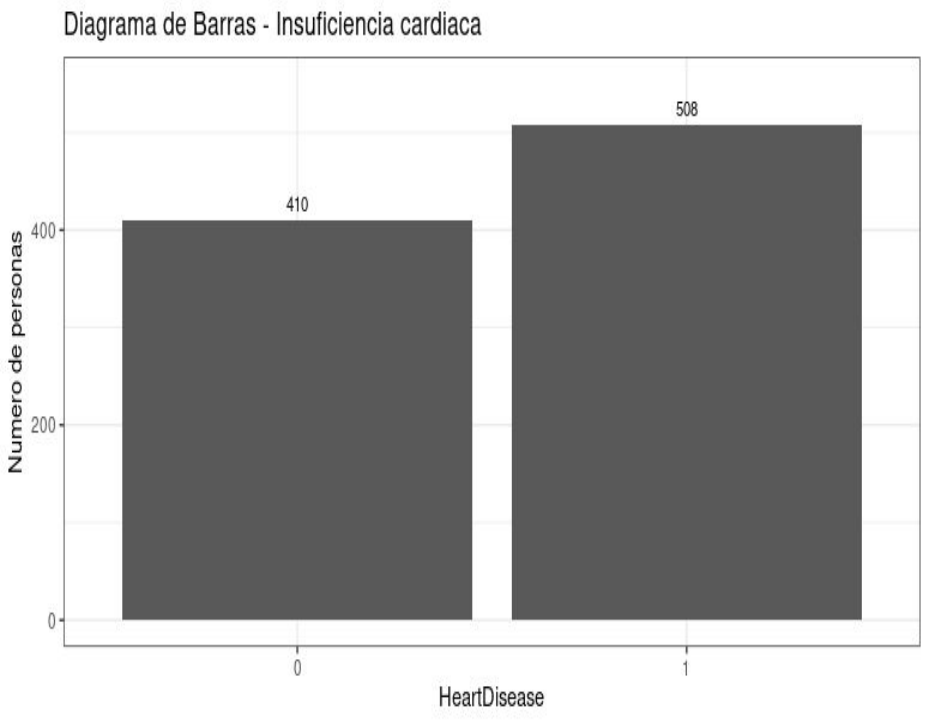
1. **Age:** Edad del paciente [años]
2. **Sex:** sexo del paciente [M: Masculino, F: Femenino].
3. **ChestPainType:** Tipo de dolor torácico [AT: Angina típica, ATA: Angina atípica, PAN: Dolor no Anginoso, ASY: Asintomático].
4. **RestingBP:** Presión arterial en reposo [mm Hg]
5. **Cholesterol:** Prueba colesterol [mm/dl].
6. **FastingBS:** Azúcar en la sangre en ayunas[1: if FastingBS > 120 mg/dl, 0: Otro caso]
7. **RestingECG:** Resultados del electrocardiograma en reposo [Normal: Normal, ST: con anomalía de la onda ST-T (inversiones de la onda T y/o elevación o depresión del ST de > 0,05 mV), HVI: que muestra hipertrofia ventricular izquierda probable o definida según los criterios de Estes].
8. **MaxHR:** Frecuencia cardíaca máxima alcanzada [Valor numérico entre 60 y 202].
9. **ExerciseAngina:** Angina inducida por el ejercicio [S: Sí, N: No].
10. **Oldpeak:** *Depresión* del segmento *ST(signo de daño miocárdico)* [Valor numérico medido en depresión].
11. **ST_Slope:** la pendiente del segmento ST de ejercicio máximo [Up: pendiente ascendente, Plano: flat, Down: pendiente descendente].
12. **HeartDisease:** Variable Clasificadora [1: cardiopatía, 0: normal].

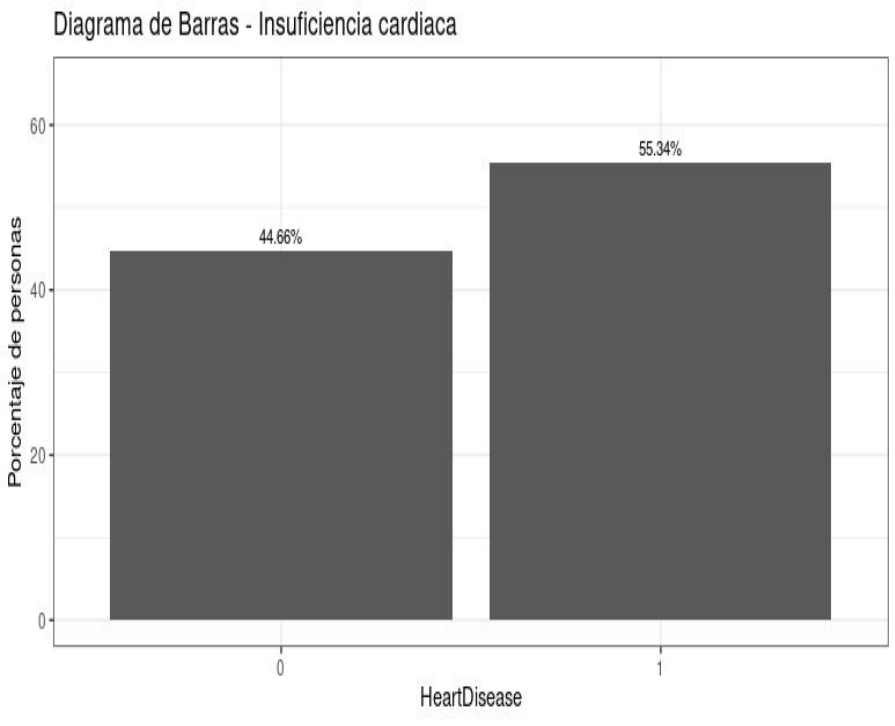
Análisis Descriptivo y Exploratorio de los datos

Análisis de datos faltantes

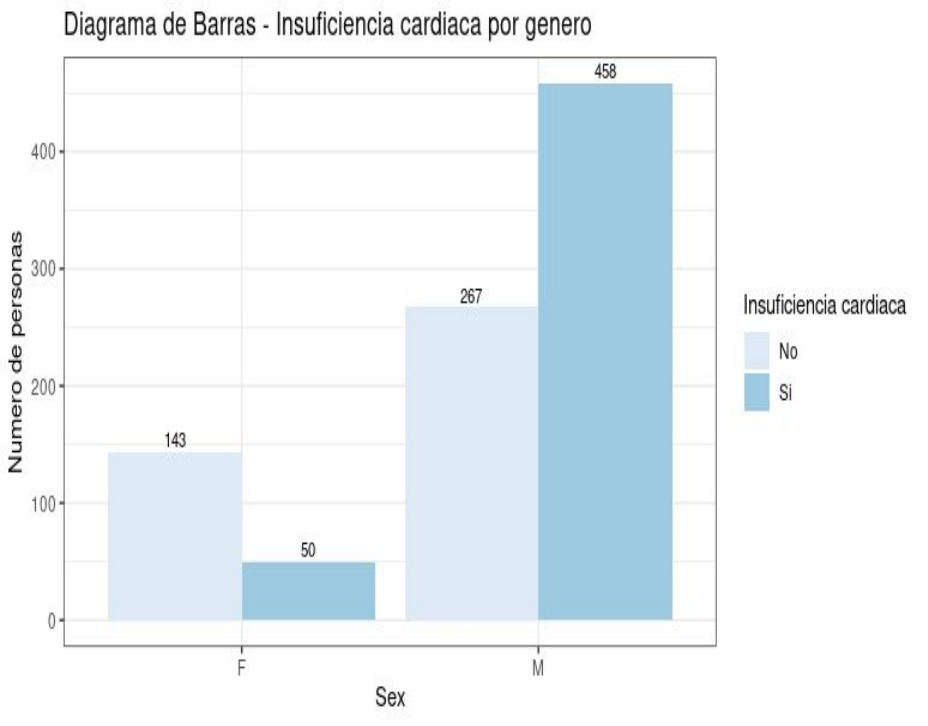


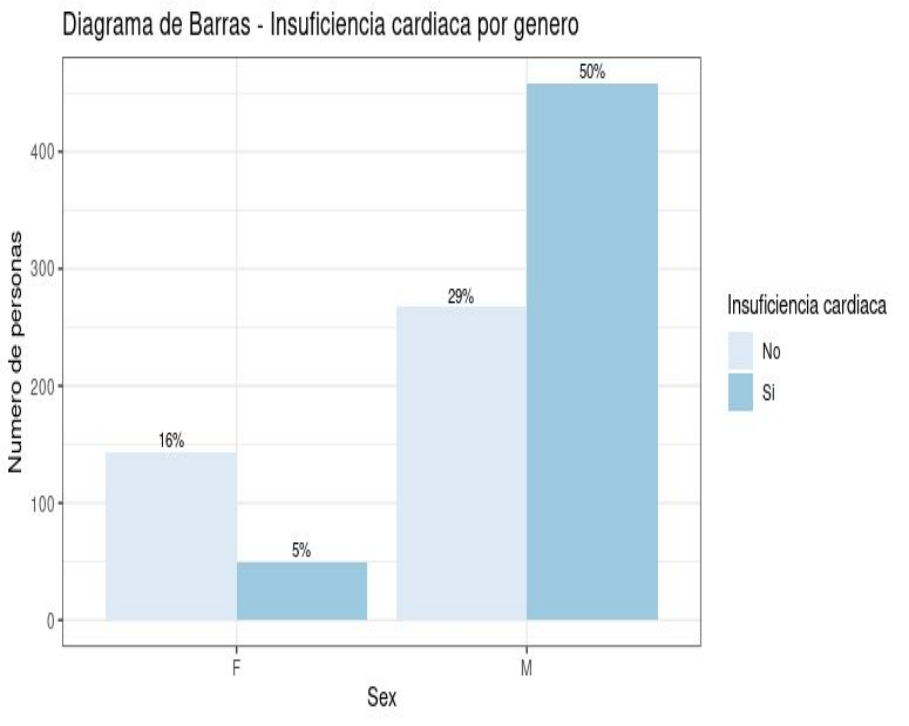
Gráficos Descriptivos



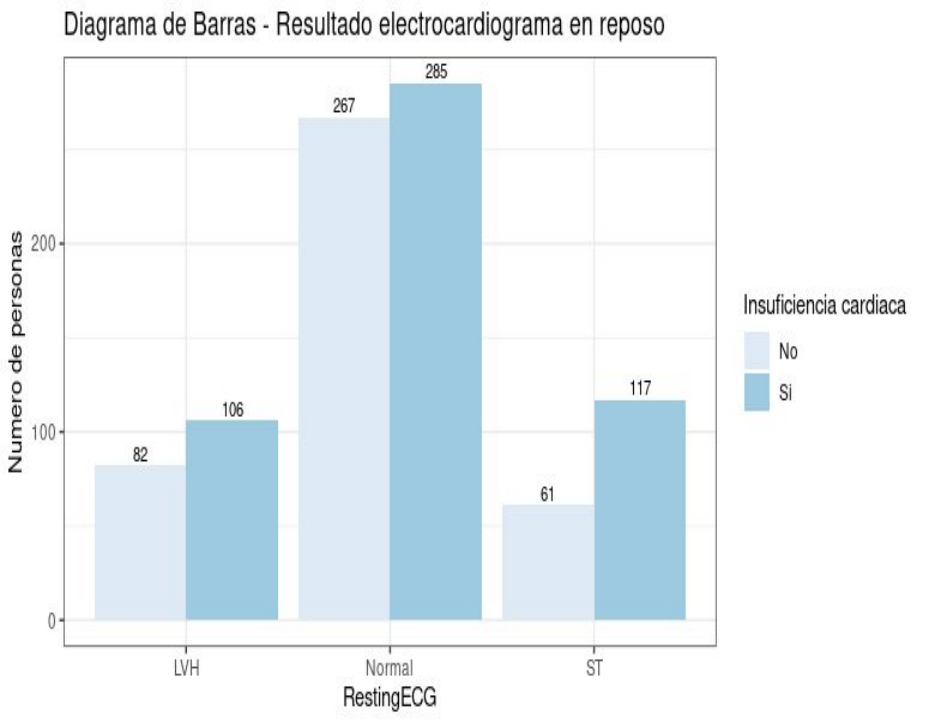


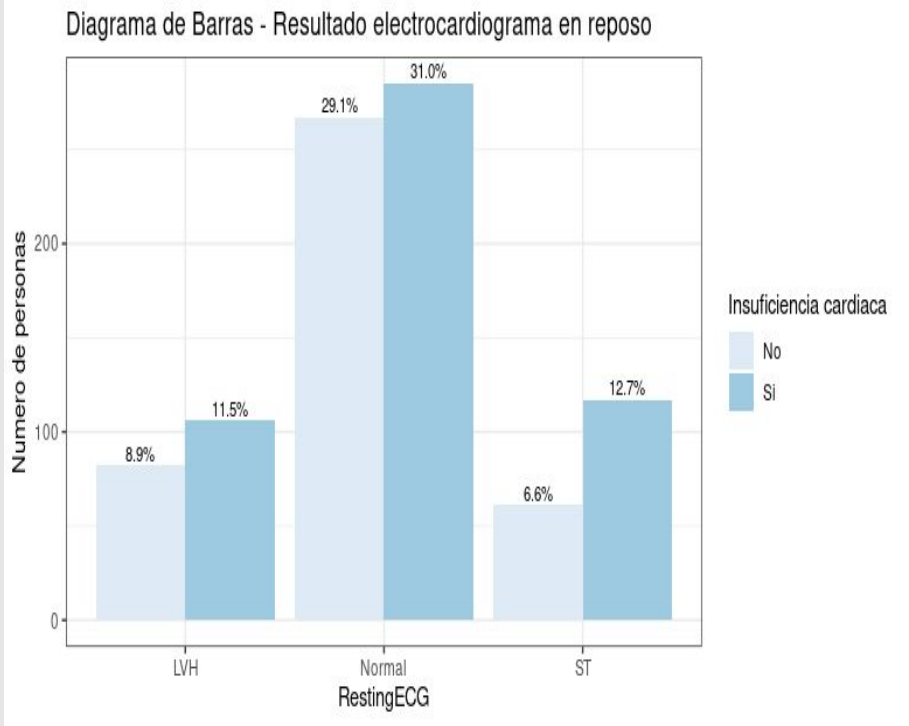
Gráficos Descriptivos



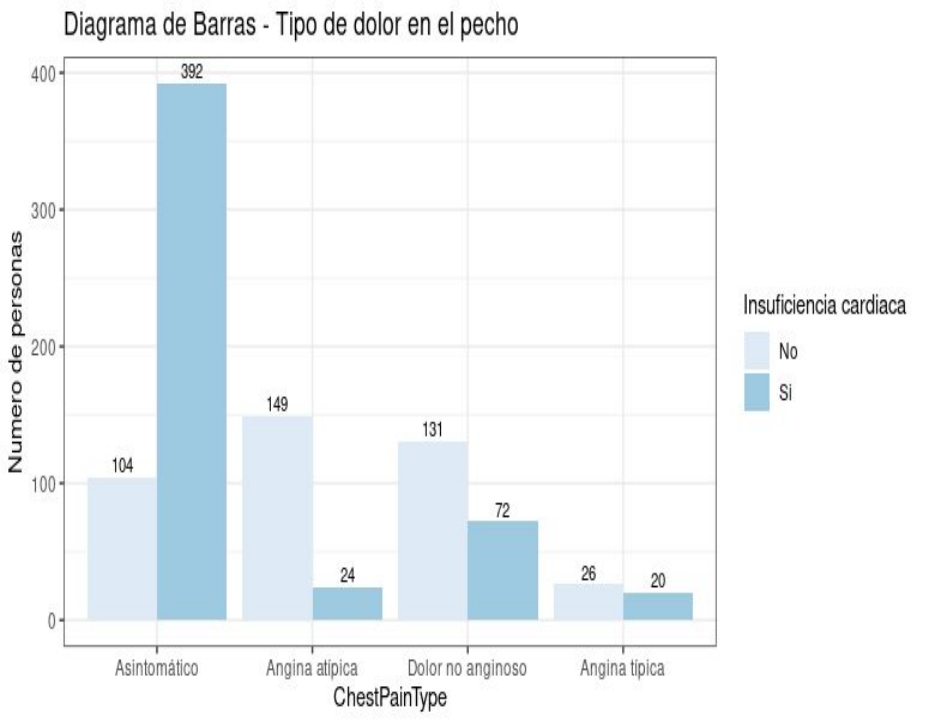


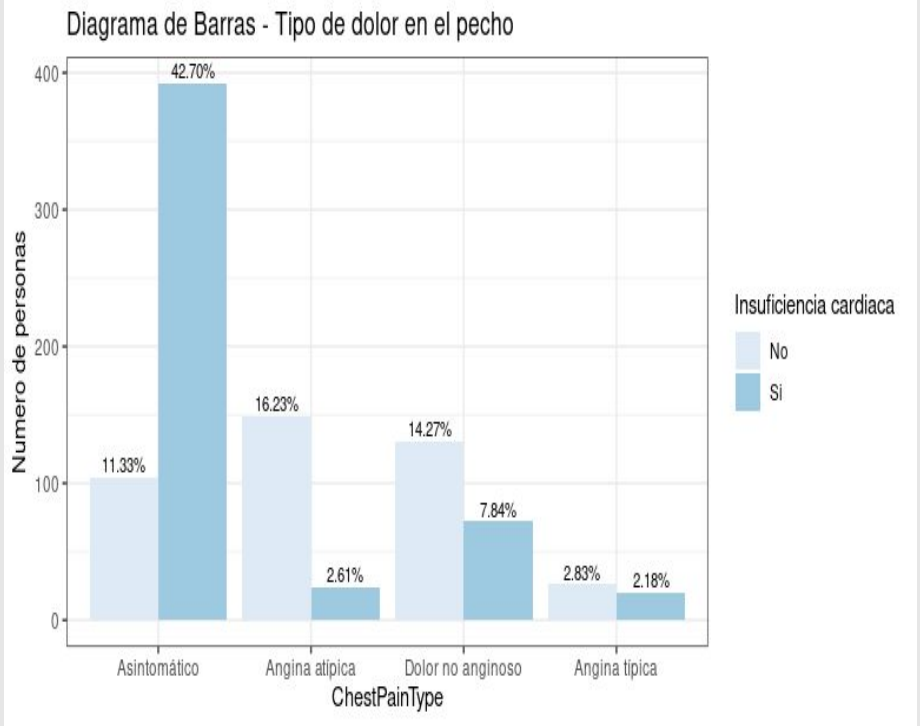
Gráficos Descriptivos



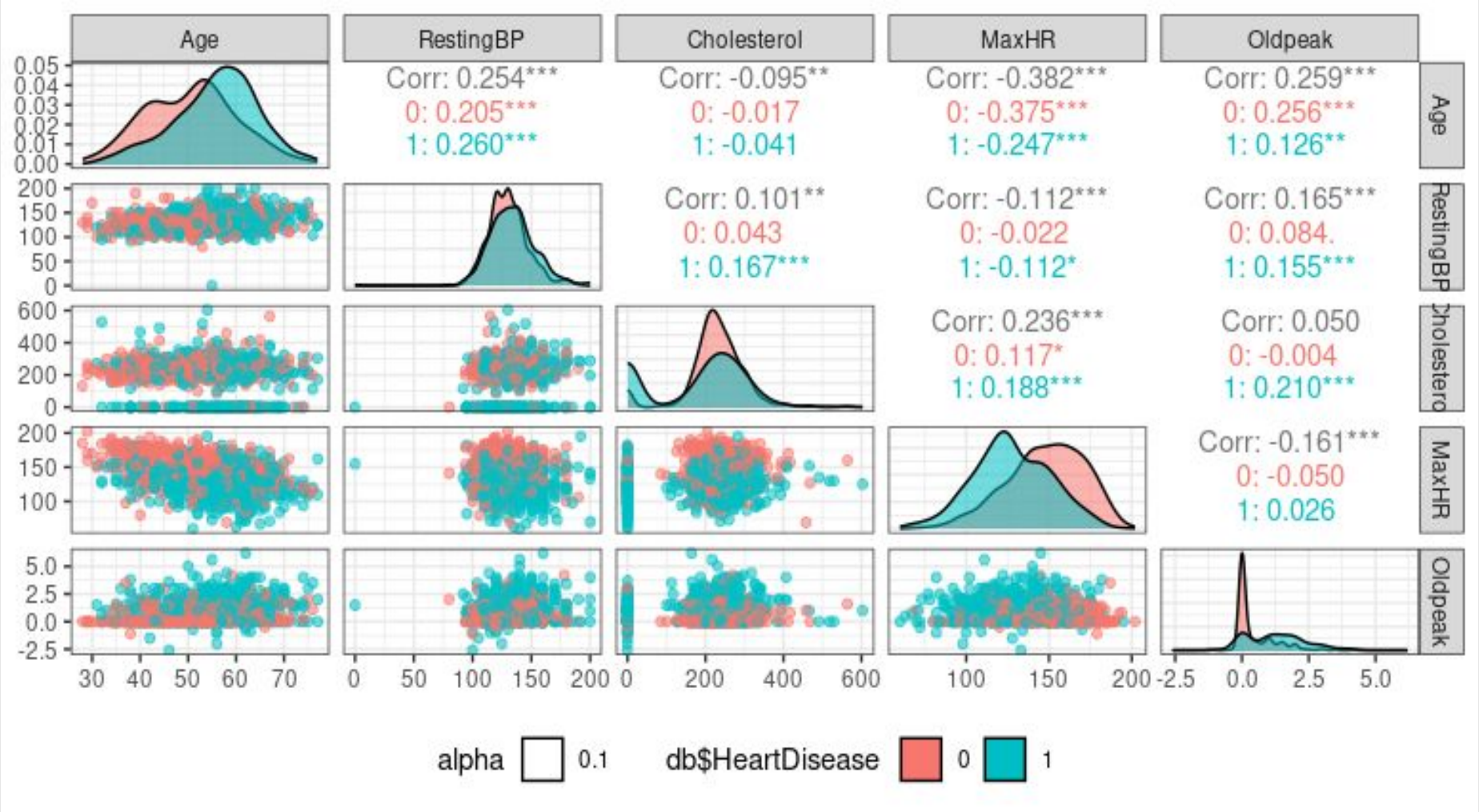


Gráficos Descriptivos





Gráficos Descriptivos



Resumen Numérico (Promedios)

Pacientes con Fallo Cardiaco

Edad	Presión arterial en reposo	Prueba colesterol	Frecuencia cardíaca máxima	*Depresión* del segmento *ST(Miocárdico)*
55.899606	134.185039	175.940945	127.655512	1.274213

Pacientes Normales

Edad	Presión arterial en reposo	Prueba colesterol	Frecuencia cardíaca máxima	*Depresión* del segmento *ST(Miocárdico)*
50.5512195	130.1804878	227.1219512	148.1512195	0.4080488

Prueba Colesterol: HDL, LDL, triglicéridos

Selección de variables

- Factores de riesgo cardiovascular. Perspectivas derivadas del Framingham Heart Study[2]
- A Fuzzy Expert System for Heart Disease Diagnosis[3]
- Valoración del Segmento ST[4]
- Serum cholesterol[5]

Modelamiento

- Train: 643 = 70%
- Test: 275 = 30 %

Modelos de Clasificación
Discriminante Lineal de Fisher (LDA)

Prueba de Normalidad Multivariada

```
        Generalized Shapiro-Wilk test for Multivariate Normality by Villasenor-Alva and Gonzalez-Estrada

data:  as.matrix(db_num)
MVW = 0.95295, p-value < 2.2e-16
```

Modelos de Clasificación
Discriminante Lineal de Fisher (LDA)

Matriz de Covarianza - Pacientes con Fallo cardiaco

	Age	RestingBP	Cholesterol	MaxHR	Oldpeak
Age	76.161499	45.028475	-45.65683	-50.3522069	1.2693981
RestingBP	45.028475	393.176738	418.77426	-52.0091087	3.5485681
Cholesterol	-45.656828	418.774263	15974.78546	554.3977620	30.5994603
MaxHR	-50.352207	-52.009109	554.39776	546.9481550	0.7049057
Oldpeak	1.269398	3.548568	30.59946	0.7049057	1.3268090

Matriz de Covarianza - Pacientes Normales

	Age	RestingBP	Cholesterol	MaxHR	Oldpeak
Age	89.206417	31.9100483	-11.9695867	-82.4845369	1.6899290
RestingBP	31.910048	272.2362932	52.6845369	-8.5090226	0.9733604
Cholesterol	-11.969587	52.6845369	5570.3322798	203.1502177	-0.2222554
MaxHR	-82.484537	-8.5090226	203.1502177	542.3340450	-0.8092886
Oldpeak	1.689929	0.9733604	-0.2222554	-0.8092886	0.4895928

Modelos de Clasificación
Discriminante Lineal de Fisher (LDA)

Prueba de hipótesis para homogeneidad de covarianza

$$H_0 : \Sigma_1 = \Sigma_2 \quad vs \quad H_0 : \Sigma_1 \neq \Sigma_2 .$$

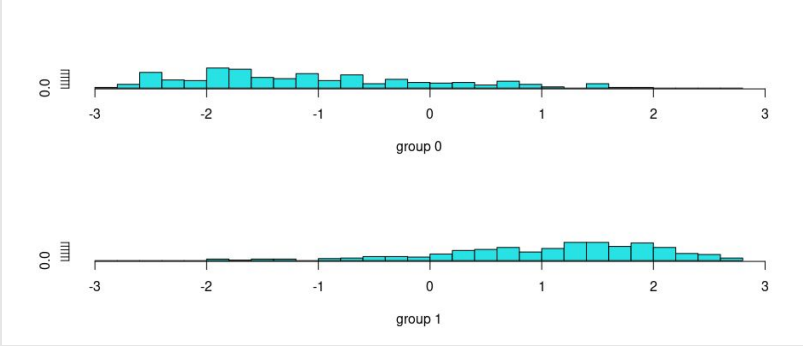
```
        Box's M-test for Homogeneity of Covariance Matrices

data:  db_num
Chi-Sq (approx.) = 254.57, df = 15, p-value < 2.2e-16
```

Modelos de Clasificación
Discriminante Lineal de Fisher (LDA)

```
model_lda <- lda(HeartDisease ~ ., data = train_data)
plot(model_lda)
```

Regla de discriminación

$$\widehat{w} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' S_p^{-1} \left[\mathbf{x} - \frac{1}{2} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \right]$$



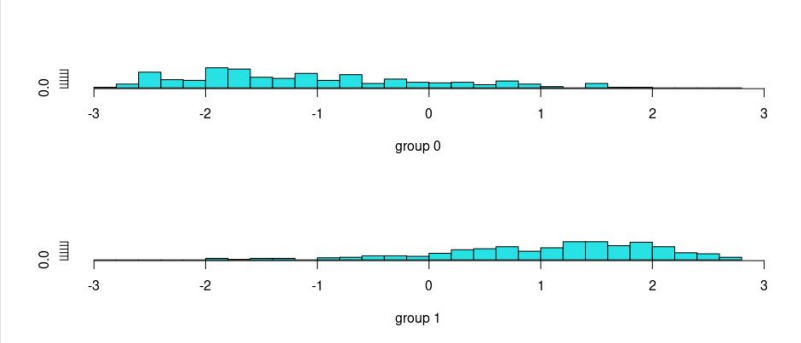
Matriz de confusión - Datos de prueba

Real	Predicha 0	Predicha 1
0	108	15
1	16	136

Tasa de mala clasificación: 11.27%

Modelos de Clasificación
Discriminante Cuadrático (QDA)

```
model_qda <- qda(HeartDisease ~ ., data = train_data)
plot(model_qda)
```



Matriz de confusión - Datos de prueba

Real	Predicha 0	Predicha 1
0	110	13
1	25	127

Tasa de mala clasificación: 13.81%

Modelos de Clasificación
Logístico

```
model_logistic <- glm(HeartDisease ~ ., family=binomial, data=train_data)
summary(model_logistic)
```

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.842221   1.675548  -0.503 0.615208
Age              0.015273   0.015495   0.986 0.324290
SexM             1.507029   0.323270   4.662 3.13e-06 ***
ChestPainTypeATA -1.806232   0.385128  -4.690 2.73e-06 ***
ChestPainTypeNAP -1.922537   0.312419  -6.154 7.57e-10 ***
ChestPainTypeTA  -1.904039   0.546820  -3.482 0.000498 ***
RestingBP         0.004104   0.006789   0.605 0.545510
Cholesterol      -0.004185   0.001235  -3.388 0.000703 ***
FastingBS1        1.183915   0.321640   3.681 0.000232 ***
RestingECGNormal -0.389396   0.315018  -1.236 0.216419
RestingECGST     -0.191001   0.394204  -0.485 0.628014
MaxHR            -0.004912   0.006014  -0.817 0.414069
ExerciseAnginaY   0.648619   0.281296   2.306 0.021121 *
Oldpeak           0.286305   0.140361   2.040 0.041373 *
ST_SlopeFlat      1.596272   0.500953   3.186 0.001440 **
ST_SlopeUp       -0.867916   0.525467  -1.652 0.098594 .
```

Regla de discriminación

$$\frac{\widehat{p}(\mathbf{x}_0)}{1-\widehat{p}(\mathbf{x}_0)} > 1 \quad o \quad \widehat{p}(\mathbf{x_0}) = \frac{e^{\widehat{\beta}_0+\mathbf{x}_0'\widehat{\boldsymbol{\beta}}}}{1+e^{\widehat{\beta}_0+\mathbf{x}_0'\widehat{\boldsymbol{\beta}}}} > 0.5$$

Matriz de confusión - Datos de prueba

	Predicha	
Real	0	1
0	107	16
1	16	136

Tasa de mala clasificación: 11.63%

Modelos de Clasificación
KNN

```
> train.kknn(train_data$HeartDisease ~ ., train_data, kmax = 20)

Call:
train.kknn(formula = train_data$HeartDisease ~ ., data = train_data,     kmax = 20)

Type of response variable: nominal
Minimal misclassification: 0.1213064
Best kernel: optimal
Best k: 20
```

Matriz de confusión - Datos de prueba

Real	Predicha 0	Predicha 1
0	109	14
1	14	138

Tasa de mala clasificación: 10.18%

```
model_knn <- kknn(HeartDisease ~ ., train_data, test_data, k = 2)
```

Matriz de confusión - Datos de prueba

Real	Predicha 0	Predicha 1
0	101	22
1	24	128

Tasa de mala clasificación: 16.72%

Selección del Modelo

Modelo	Tasa de Mala Clasificación
LDA	11.27%
QDA	13.81%
Logístico	11.63%
KNN, k = 20	10.18 %

Bibliografía

[1] **Kaggle:** Heart Failure Prediction Dataset
https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

[2]Factores de riesgo cardiovascular. Perspectivas derivadas del Framingham Heart Study
https://www.sciencedirect.com/science/article/abs/pii/S0300893208733888

[3]A Fuzzy Expert System for Heart Disease Diagnosis
https://www.researchgate.net/publication/44260568_A_Fuzzy_Expert_System_for_Heart_Disease_Diagnosis

[4]Valoración del Segmento ST
https://www.my-ekg.com/como-leer-ekg/segmento-st.html

[5]Serum cholesterol
https://www.medicalnewstoday.com/articles/321519

Gracias