Trabajo Práctico Final

Seminario *Big Data y Minería de Datos*

Leandro Pisaroni

10 de octubre de 2021

1. INTRODUCCIÓN

Se analiza un conjunto de datos proveniente de un estudio sobre enfermedades cardíacas donde se evaluaron distintas variables de la salud de 270 pacientes y el desarrollo o no de enfermedades coronarias. La Tabla 1 muestra las 14 variables recolectadas en el estudio junto a sus respectivas descripciones.

Tabla 1: Descripción de las variables para el estudio de las enfermedades cardíacas.

Variable	Descripción
age	Edad en años
sex	Sex del paciente (0: Femenino, 1: Masculino)
chest_pain_type	Tipo de dolor de pecho (1, 2, 3 o 4)
resting_blood_pressure	Presión arterial en reposo
serum_cholesterol	Colesterol sérico [mg/dl]
fasting_blood_sugar_gt_120	¿Nivel de azúcar en sagre en ayunas > 120 mg/dl? (0: No, 1: Si)
resting_ekg_results	Resultados electrocardiográficos en reposo (0, 1, 2)
max_heart_rate_achieved	Frecuencia cardíaca máxima alcanzada [latidos/min]
exercise_induced_angina	¿Dolor en el pecho inducido por el ejercicio? (0: No, 1: Si)
oldpeak_eq_st_depression	Depresión del ST inducida por el ejercicio en relación con el reposo
slope_of_peak_exercise	Calidad del flujo sanguíneo al corazón
num_major_vessels	Número de vasos principales
thal	Flujo sanguíneo al corazón (3: 'Normal', 6: 'Defecto fijo', 7: 'Defecto reversible')
heart_disease	¿Enfermedad coronaria? (1: No, 2: Si)

El **objetivo** del trabajo es *construir modelos que permitan predecir si un paciente tendrá o no enfermedades cardíacas*. Para esto se emplearán tres técnicas diferentes: árboles de clasificación, regresiones regularizadas y modelos lineales generalizados.

2. DESCRIPCIÓN DE LOS DATOS

Antes de evaluar los distintos modelos se describe brevemente los datos seleccionados. En la Tabla 2 se presentan diferentes medidas resumen para las variables del estudio, en la Figura 1 se muestra la distribución de las variables categóricas y en la Figura 2 la distribución de los pacientes con enfermedades cardíacas. Finalmente, en la Figura 3 se presenta la matriz de correlaciones lineales.

Tabla 2: Medidas resumen de las variables cuantitativas del estudio de enfermedades cardíacas.

	Media	Mediana	Desvío	Mínimo	Máximo
age	54.43	55	9.11	29	77
resting_blood_pressure	131.34	130	17.86	94	200
serum_cholesterol	249.66	245	51.69	126	564
max_heart_rate_achieved	149.68	153.5	23.17	71	202
oldpeak_eq_st_depression	1.05	0.8	1.15	0	6.2
slope_of_peak_exercise	1.59	2	0.61	1	3
num_major_vessels	0.67	0	0.94	0	3

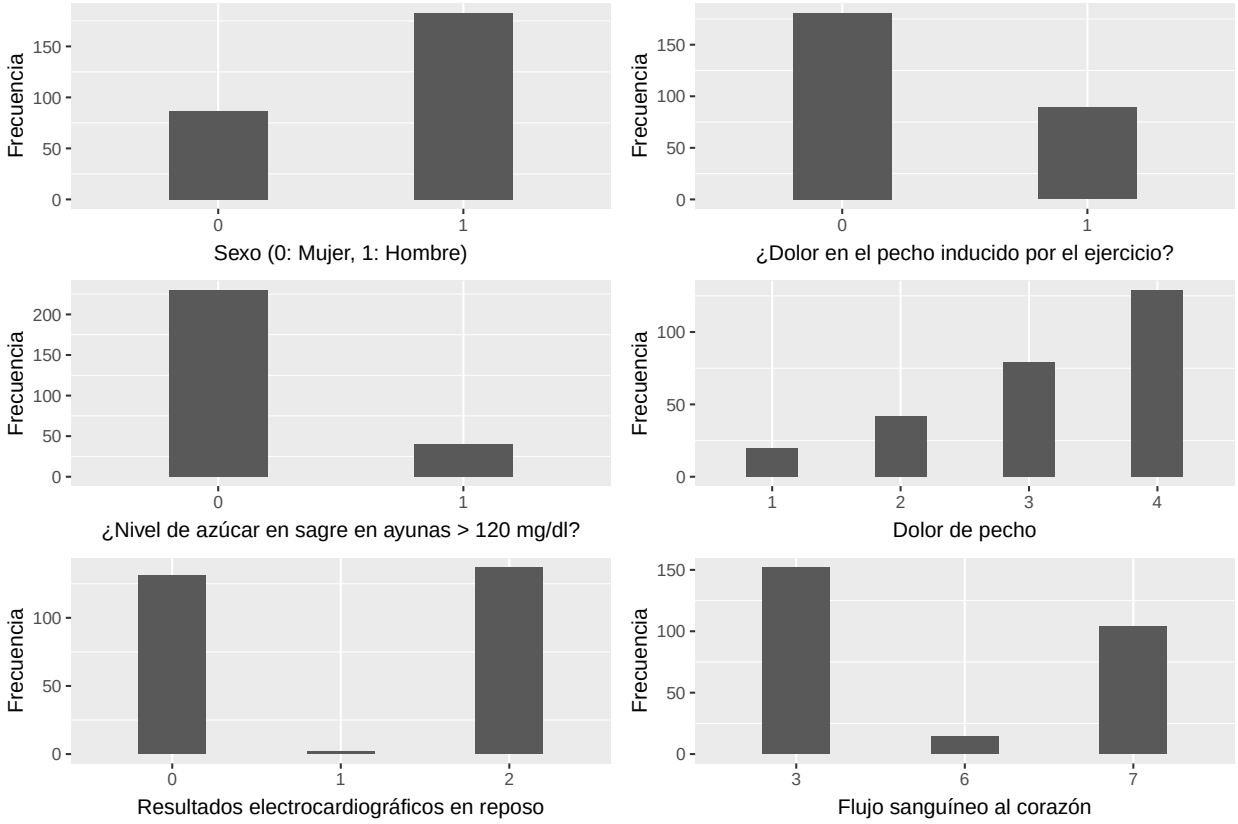

Figura 1: Gráfico de barras para la distribución de las variables categóricas.

Lo primero que se observa en la Figura 2 es que las categorías de la variable de **respuesta** están balanceadas, sin que haya un nivel que domine la mayor parte de las observaciones. Teniendo en cuenta, además, que esta variable es categórica, se puede optar como **métricas de comparación de modelos** para evaluar la *capacidad predictiva* de las distintas técnicas a aplicar la *sensibilidad*, la *especificidad*, el *área bajo la curva ROC* (*AUC*), el F_1 *Score* y el *Log Loss*, entre otras.

Por otro lado, en la Figura 3 se observa que algunas correlaciones lineales entre las **variables de respuesta** son moderadas (como por ejemplo entre la edad y la frecuencia cardíaca máxima, o entre la calidad del flujo cardíaco y la depresión del ST inducida por el ejercicio en relación con el reposo). Por tal motivo, cuando se construyan árboles de clasificación se optará por una técnica que controle estas correlaciones, como son los *bosques aleatorios* (*random forest*).

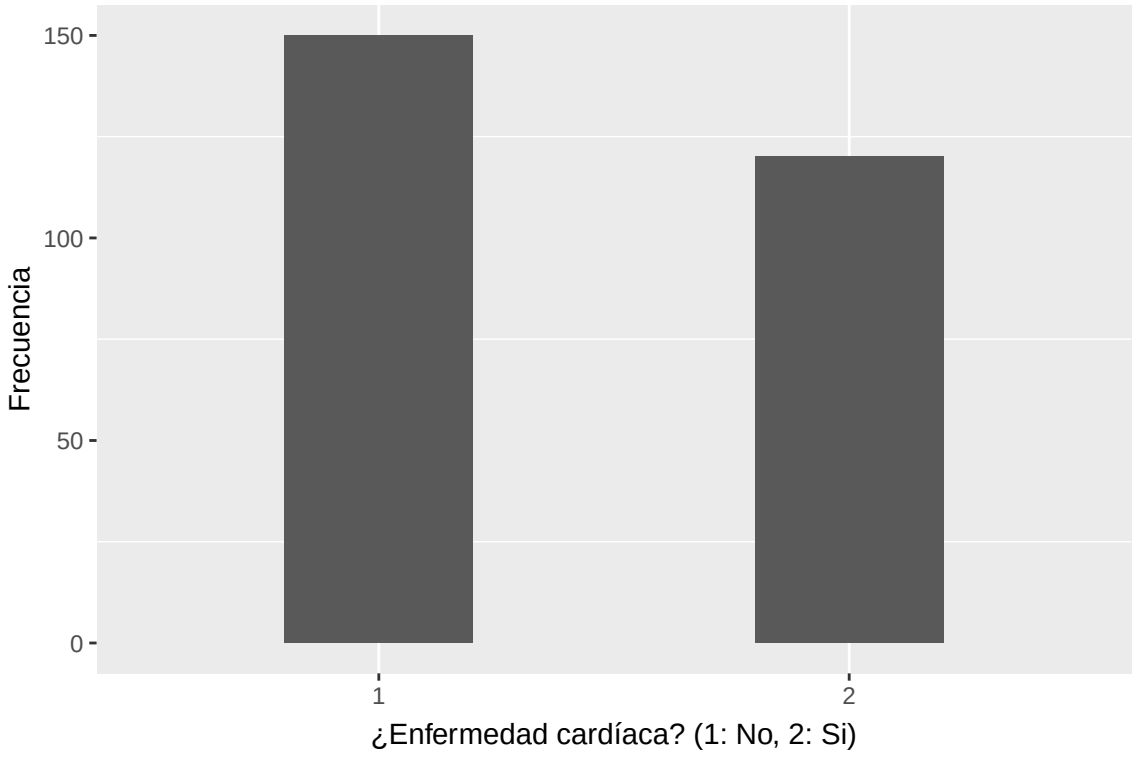

Figura 2: Gráfico de barras para la distribución de los pacientes con enfermedades cardíacas en el estudio.

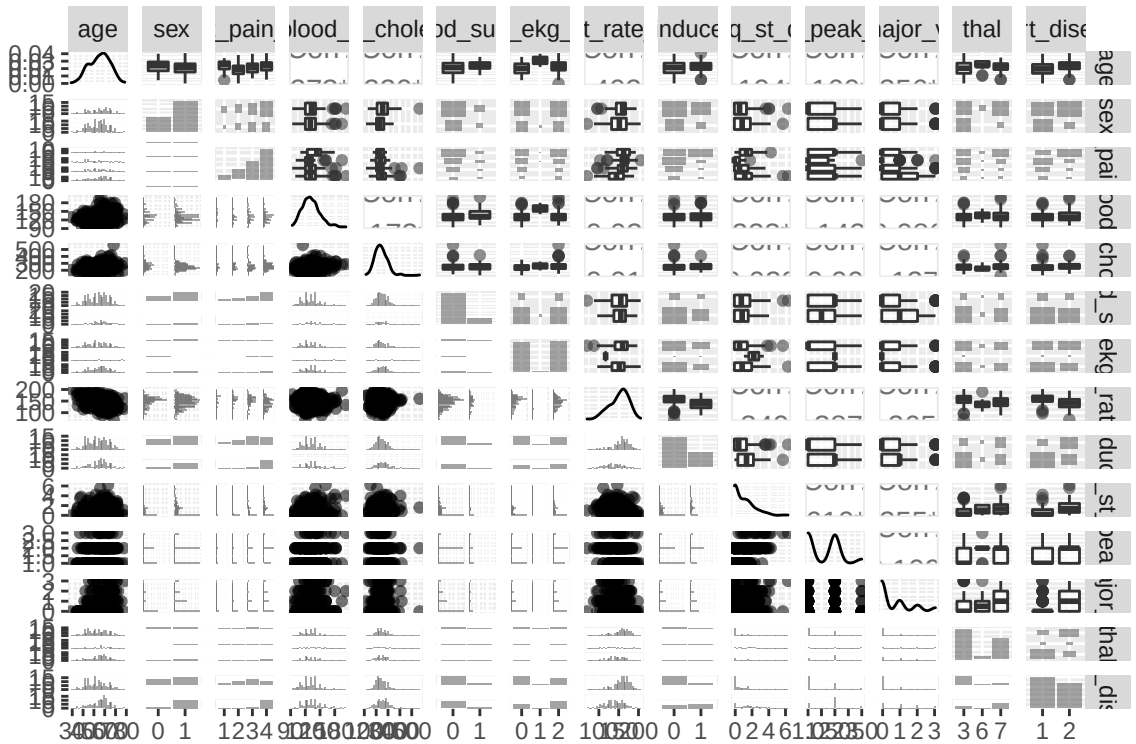

Figura 3: Matriz de correlaciones para las variables del estudio.

3. MODELOS PROPUESTOS

Antes de construir los modelos se divide la totalidad de los pacientes en dos grupos: un *subconjunto de entrenamiento* (con el 70 % de las observaciones) y un *subconjunto de prueba* (con el 30 % restante).

El primer conjunto de datos se utilizará para obtener los valores óptimos de los parámetros de cada técnica, mientras que el segundo se usará para estimar las métricas que luego se empleará en la comparación de los modelos.

Además, en todos los casos se utilizará la misma fórmula: se considera como variable de respuesta a *heart_disease* y como variables regresoras a todas las demás.

3.1. RANDOM FOREST

El primero de los modelos que se construye es un **Random Forest**. Para esto se decide trabajar con 50, 75 y 100 árboles y un tamaño mínimo de nodo igual a 1 (por defecto así lo considera la función), y aumentar el número máximo de *features* a muestrear en cada paso de 2 a 10 para determinar el valor óptimo de los mismos. La Tabla 3 muestra diferentes métricas para cada una de estas combinaciones.

Tabla 3: Métricas para los Bosques Aleatorios calculados.

Nº Árboles	Nº Variables	F_1	AUC	Especificidad	Sensibilidad	Exactitud	LogLoss
50	2	0.813	0.795	0.771	0.86	0.79	19.897
50	3	0.817	0.799	0.76	0.884	0.79	19.334
50	4	0.809	0.789	0.745	0.884	0.778	19.641
50	5	0.809	0.791	0.783	0.837	0.79	19.334
50	6	0.767	0.752	0.767	0.767	0.753	18.6
50	7	0.83	0.816	0.765	0.907	0.802	20.076
50	8	0.826	0.81	0.776	0.884	0.802	19.76
50	9	0.769	0.743	0.729	0.814	0.741	19.351
50	10	0.804	0.784	0.755	0.86	0.778	19.581
75	2	0.8	0.78	0.766	0.837	0.778	19.541
75	3	0.826	0.81	0.776	0.884	0.802	19.53
75	4	0.848	0.835	0.796	0.907	0.827	19.877
75	5	0.783	0.758	0.735	0.837	0.753	19.001
75	6	0.804	0.784	0.755	0.86	0.778	19.973
75	7	0.809	0.789	0.745	0.884	0.778	19.831
75	8	0.804	0.784	0.755	0.86	0.778	19.518
75	9	0.787	0.766	0.761	0.814	0.765	19.734
75	10	0.783	0.758	0.735	0.837	0.753	19.763
100	2	0.804	0.784	0.755	0.86	0.778	19.445
100	3	0.826	0.81	0.776	0.884	0.802	19.636
100	4	0.813	0.795	0.771	0.86	0.79	19.474
100	5	0.804	0.784	0.755	0.86	0.778	19.389
100	6	0.848	0.835	0.796	0.907	0.827	19.44
100	7	0.804	0.784	0.755	0.86	0.778	19.56
100	8	0.8	0.78	0.766	0.837	0.778	19.398
100	9	0.8	0.78	0.766	0.837	0.778	19.41
100	10	0.8	0.78	0.766	0.837	0.778	19.342

Como puede observarse en la tabla anterior, el modelo que considera 75 árboles y 4 variables presenta una de los mayores valores de F_1 y, simultáneamente, uno de los menores valores de Log Loss.

Por otro lado, en la Figura 4 se muestra la *tasa de error estimada* versus el número de árboles utilizados para el random forest. En el mismo, se puede observar que el error disminuye muy poco a partir de los 50

árboles. De esta manera, el modelo a utilizar considera 4 variables y 50 árboles. En la Tabla 4 se muestran las métricas para el modelo final de random forest.

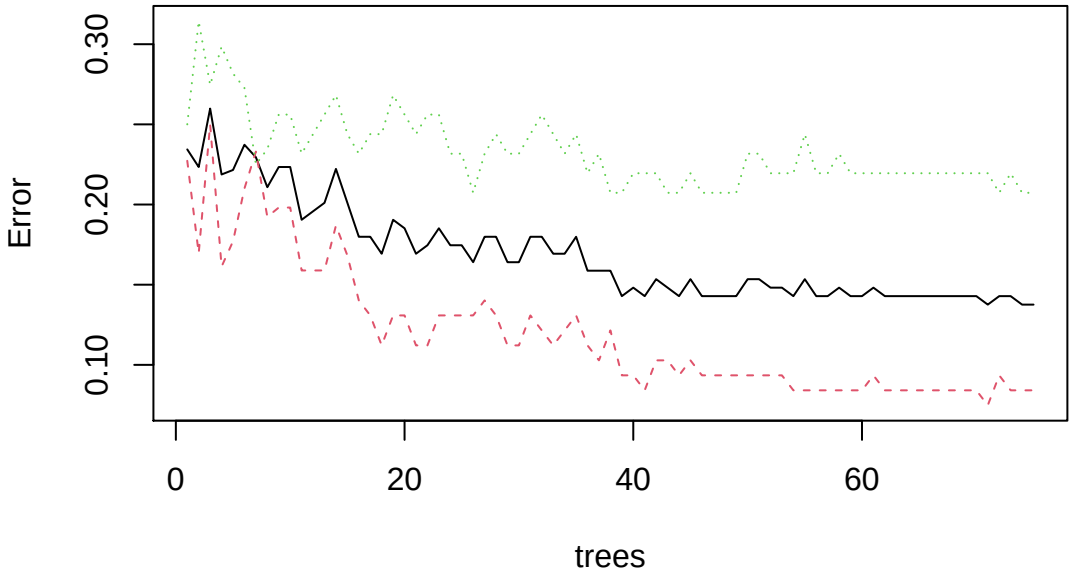

Figura 4: Tasa de error estimada en función de la cantidad de árboles para los random forests.

Tabla 4: Métricas para el Bosque Aleatorio de 40 árboles y 6 variables.

F_1	AUC	Especificidad	Sensibilidad	Exactitud	LogLoss
0.804	0.787	0.722	0.907	0.765	20.306

3.1.1. ANÁLISIS DEL MODELO

El modelo anterior puede utilizarse para obtener las probabilidades de que un individuo padezca una enfermedad conoraria, tal como muestra la Tabla 5 (para los primeros 6 pacientes).

Tabla 5: Probabilidad de que el paciente sufra una enfermedad cardíaca (primeros valores).

	Probabilidad
2	0.48
3	0.10
12	0.54
15	0.34
18	0.72
19	0.42

Según este modelo, por ejemplo, el paciente 2 tiene una probabilidad de 0.48 de padecer una enfermedad

cardíaca.

3.2. REGRESIÓN REGULARIZADA

La segunda técnica que se utiliza son las **regresiones regularizadas**. Se evalúan regresiones Ridge, LASSO y SCAD, en donde la determinación del valor de λ se realiza minimizando el error de estimación por medio de *validación cruzada*.

En la Tabla 6 se muestran los valores de diferentes métricas obtenidas para cada uno de los modelos.

Tabla 6: Métricas para las regresiones regularizadas.

Regresión	RMSE	MAE	RMSLE	Gini
Ridge	0.400	0.358	0.161	0.582
LASSO	0.385	0.340	0.154	0.606
SCAD	0.393	0.338	0.158	0.564

3.2.1. COMPARACIÓN DE LAS REGRESIONES

Para evaluar el desempeño de los tres modelos, podría utilizarse como métrica de comparación alguna de las presentadas en la tabla anterior. Sin embargo, como la respuesta ("heart_disease") es una variable binaria, se propone usar las curvas ROC de la Figura 5 para cada uno de los tres modelos.

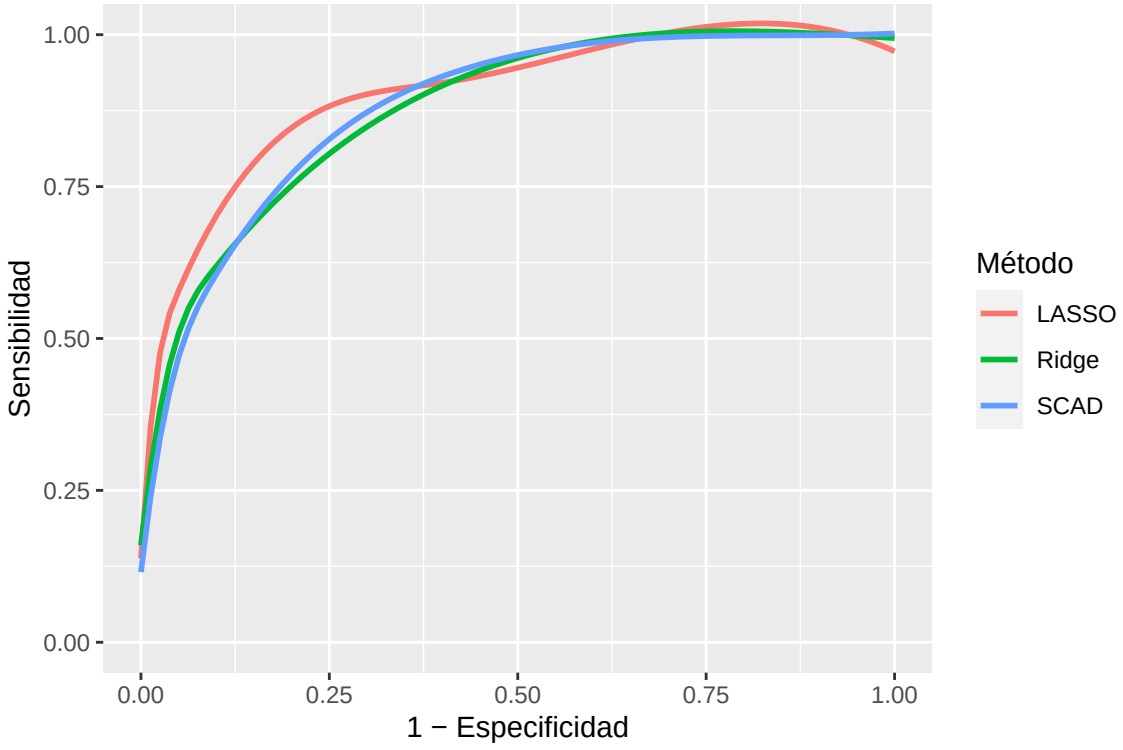

Figura 5: Curvas ROC para cada uno de los tres modelos de regresión regularizada.

Como puede observarse en la figura anterior, la curva ROC de la regresión Ridge es muy similar tanto en forma como en superficie bajo a ella a la curva ROC de la regresión SCAD. Por su parte, la curva ROC de la regresión LASSO tiene un comportamiento extraño en cuanto a su forma (puede deberse a un sobreajuste del método que construye la curva que la define), por lo que se decide no seguir trabajando con ella.

En la Tabla 7 se presentan los valores calculados del *área bajo la curva* (AUC) para cada una de las dos regresiones restantes. A partir de estos valores, la mejor de las regresiones es la SCAD En la Tabla 8 se muestran las métricas de comparación para este modelo.

Tabla 7: AUC de las curvas ROC para cada una de las regresiones regularizadas.

Ridge	SCAD
0.878	0.879

Tabla 8: Métricas para la Regresión SCAD.

	F_1	AUC	Especificidad	Sensibilidad	Exactitud	LogLoss
.	0.8	0.779	0.731	0.884	0.765	19.805

3.2.2. ANÁLISIS DEL MODELO

Al igual que lo que se hizo con el bosque aleatorio, el modelo anterior puede utilizarse para obtener las probabilidades de que un individuo padezca una enfermedad conoraria, tal como muestra la Tabla 9 (para los primeros 6 pacientes).

Tabla 9: Probabilidad de que el paciente sufra una enfermedad cardíaca (primeros valores).

	Probabilidad
2	0.510
3	0.322
12	0.651
15	0.357
18	0.768
19	0.454

Según este modelo, por ejemplo, el paciente 2 tiene una probabilidad de 0.51 de padecer una enfermedad cardíaca.

3.3. MODELO LOGIT

Como última técnica para estudiar los datos se propone un **modelo lineal generalizado** para respuestas binarias que utilice la distribución logística (*modelo logit*).

Para seleccionar las variables que se incluirán en el modelo se utiliza un *proceso de selección automático* sobre el *modelo logit*. Luego, las variables seleccionadas son *resting_blood_pressure*, *serum_cholesterol*, *max_heart_rate_achieved*, *oldpeak_eq_st_depression*, *num_major_vessels*, *sex*, *chest_pain_type*, *exercise_induced_angina* y *thal*.

En la Tabla 10 se muestran las métricas de comparación para este modelo.

Tabla 10: Métricas para el Modelo Logit

F_1	AUC	Especificidad	Sensibilidad	Exactitud	LogLoss
0.887	0.872	0.862	0.913	0.87	19.189

3.3.1. ANÁLISIS DEL MODELO

Al igual que en los dos caso anteriores, el modelo logit puede utilizarse para obtener las probabilidades de que un individuo padezca una enfermedad conoraria, tal como muestra la Tabla 11 (para los primeros 6 pacientes).

Tabla 11: Probabilidad de que el paciente sufra una enfermedad cardíaca (primeros valores).

	Probabilidad
2	0.458
3	0.411
13	0.034
15	0.199
18	0.937
19	0.119

Según este modelo, por ejemplo, el paciente 2 tiene una probabilidad de 0.458 de padecer una enfermedad cardíaca.

3.4. COMPARACIÓN DE LOS MODELOS

Finalmente, para determinar cuál de los métodos utilizados es el que mejor predice la posibilidad de desarrollar enfermedades cardíacas se comparan las métricas calculadas para cada modelo. Estos valores se presentan en la Tabla 12.

Tabla 12: Métricas de comparación para los modelos propuestos.

Modelo	F_1	AUC	Especificidad	Sensibilidad	Exactitud	LogLoss
Random Forest	0.804	0.787	0.722	0.907	0.765	20.306
SCAD	0.800	0.779	0.731	0.884	0.765	19.805
Logit	0.887	0.872	0.862	0.913	0.870	19.189

De acuerdo con la tabla anterior, el mejor de los modelos es el *Logit*, ya que tiene el mayor valor de F_1 y el menor valor de Log Loss.

Por otro lado, también se pueden comparar las probabilidades estimadas con cada uno de estos modelos, tal como se muestra en la Tabla 13.

Tabla 13: Probabilidad de que el paciente sufra una enfermedad cardíaca (primeros valores).

Paciente	Random Forest	SCAD	Logit	¿Padece enfermedad?
2	0.48	0.510	0.458	No
3	0.10	0.322	0.411	Si
13	0.54	0.651	0.034	No
15	0.34	0.357	0.199	No
18	0.72	0.768	0.937	Si
19	0.42	0.454	0.119	No

Nuevamente, de acuerdo con los valores estimados, el modelo que mejor predice es el Logit.