Predicción de Enfermedad Cardiaca Como Medida de Prevención.

Proyecto del Modulo 4 (Machine Learning) del Diplomado de Ciencias de Datos 2024.

Autor: Cinthya Simoneen

Fecha: Octubre-2024

Objetivo:

La enfermedad cardiaca cobra la vida de más personas cada día. Tan solo en México en 2023, las enfermedades del corazón fueron la principal causa de defunción, con 97,187 casos, lo que representa el 25% de las muertes registradas. A nivel mundial, las enfermedades cardiovasculares (ECV) son la principal causa de muerte. Se estima que 17,9 millones de personas fallecieron por ECV, lo que representa el 32% de todas las muertes a nivel global.

Basándonos en indicadores como glucosa en sangre, medición de la presión arterial y otros más, se realiza una predicción de riesgo que nos indica si un paciente está en riesgo de tener enfermedad cardiaca y de ser así derivarlo a un tratamiento preventivo.

Información del Dataset

El dataset con los datos necesarios para el proyecto se descarga de Keaggle, el dataset está conformado por:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1025 entries, 0 to 1024
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       1025 non-null   int64  
 1   sex       1025 non-null   int64  
 2   cp        1025 non-null   int64  
 3   trestbps  1025 non-null   int64  
 4   chol      1025 non-null   int64  
 5   fbs       1025 non-null   int64  
 6   restecg   1025 non-null   int64  
 7   thalach   1025 non-null   int64  
 8   exang     1025 non-null   int64  
 9   oldpeak   1025 non-null   float64
 10  slope     1025 non-null   int64  
 11  ca        1025 non-null   int64  
 12  thal      1025 non-null   int64  
 13  target    1025 non-null   int64  
dtypes: float64(1), int64(13)
memory usage: 112.2 KB
```

En donde las columnas contienen la siguiente información:

Indicadores de la columnas

Indicadores (columnas)

Columna	Información	Valores
age	Edad del paciente en años	
sex	Género del paciente	0=mujer,1=hombre
cp	Tipo de dolor de Pecho	0 = Typical Angina, 1 = Atypical Angina, 2 = Non-anginal Pain, 3 = Asymptomatic
trestbps	Presión arteria en descanso (mm hg)	
chol	cholesterol en suero (mgdl)	
fbs	Glucosa en sangre	> 120 mg/dl (1 = true; 0 = false).
restecg	Resultados del electrocardigrama en reposo	0=Normal,1=ST-T normal,2=Hipertrofia ventriculo izquierdo
thalach	Frecuencia máxima en reposo	
exang	Angina de pecho inducida por ejercicio	0=False, 1=True
oldpeak	Depresión del segmento ST al hacer ejercicio	
slope	Pendiente ST en pico al ejercitarse	Valores del 1 al 3
caa	Vasos mayores encontrados	Valores del 0 al 3
thal	Thalassemia	1 = normal; 2 = tratado; 3 = reversible
target	Diagnostico de enfermedad cardiaca	1 = presente; 0 = ausente

Ejemplo de los datos:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
5	58	0	0	100	248	0	0	122	0	1.0	1	0	2	1
6	58	1	0	114	318	0	2	140	0	4.4	0	3	1	0
7	55	1	0	160	289	0	0	145	1	0.8	1	1	3	0
8	46	1	0	120	249	0	0	144	0	0.8	2	0	3	0
9	54	1	0	122	286	0	0	116	1	3.2	1	2	2	0
10	71	0	0	112	149	0	1	125	0	1.6	1	0	2	1
11	43	0	0	132	341	1	0	136	1	3.0	1	0	3	0
12	34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
13	51	1	0	140	298	0	1	122	1	4.2	1	3	3	0
14	52	1	0	128	204	1	1	156	1	1.0	1	0	0	0

Información de los Datos

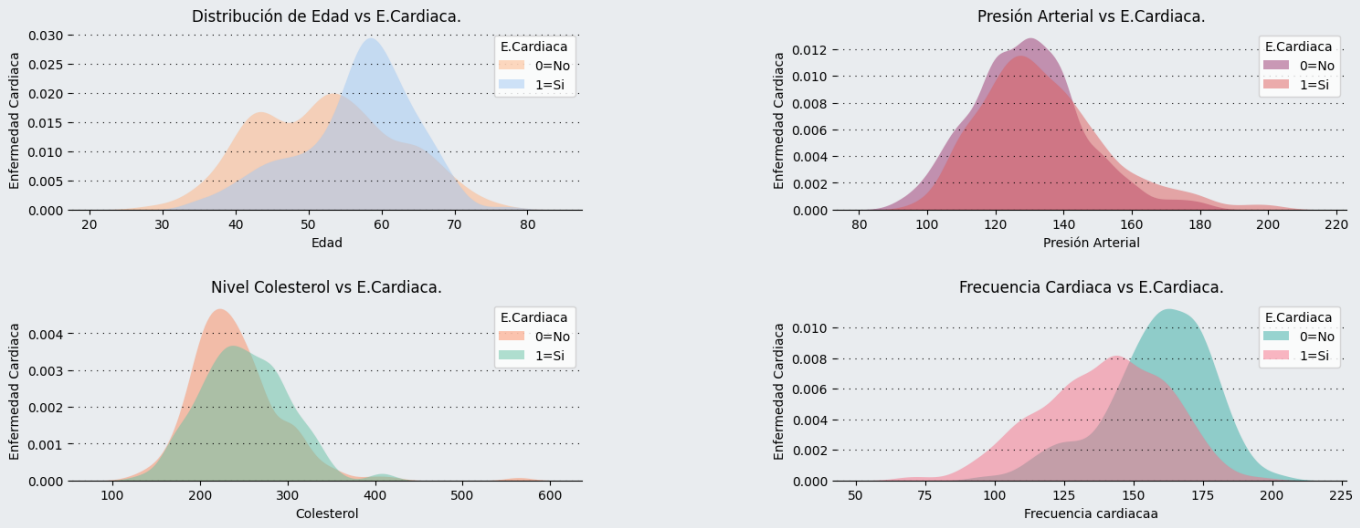
Información estadística del Dataset

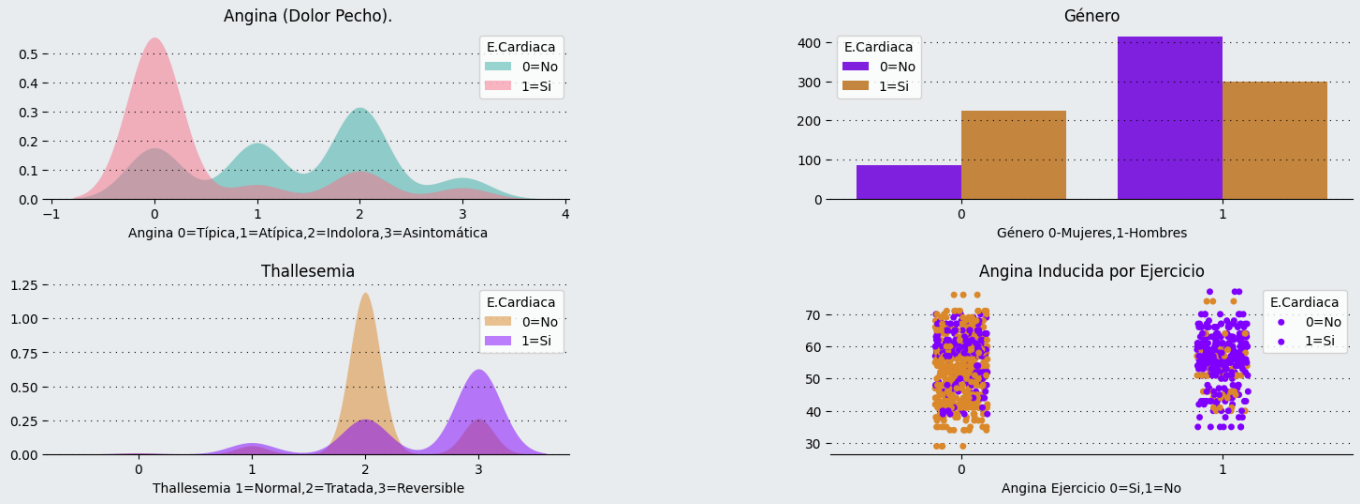
	count	mean	std	min	25%	50%	75%	max
age	1025.0	54.434146	9.072290	29.0	48.0	56.0	61.0	77.0
sex	1025.0	0.695610	0.460373	0.0	0.0	1.0	1.0	1.0
cp	1025.0	0.942439	1.029641	0.0	0.0	1.0	2.0	3.0
trestbps	1025.0	131.611707	17.516718	94.0	120.0	130.0	140.0	200.0
chol	1025.0	246.000000	51.592510	126.0	211.0	240.0	275.0	564.0
fbs	1025.0	0.149268	0.356527	0.0	0.0	0.0	0.0	1.0
restecg	1025.0	0.529756	0.527878	0.0	0.0	1.0	1.0	2.0
thalach	1025.0	149.114146	23.005724	71.0	132.0	152.0	166.0	202.0
exang	1025.0	0.336585	0.472772	0.0	0.0	0.0	1.0	1.0
oldpeak	1025.0	1.071512	1.175053	0.0	0.0	0.8	1.8	6.2
slope	1025.0	1.385366	0.617755	0.0	1.0	1.0	2.0	2.0
ca	1025.0	0.754146	1.030798	0.0	0.0	0.0	1.0	4.0
thal	1025.0	2.323902	0.620660	0.0	2.0	2.0	3.0	3.0
target	1025.0	0.513171	0.500070	0.0	0.0	1.0	1.0	1.0

Edad: Media entre 29 y 77 años
Género de los pacientes:1 = hombre; 0 = mujer
Tipo de dolor de Pecho: 0 = Angina Tipica, 1 = Angina Atipica, 2 = Sin dolor , 3 = Asintomático
Presión Arterial: Media entre 94 y 200 mm Hg.
Colesterol: Media entre 126 y 564 mg/dl.
Glucosa en sangre: > 120 mg/dl 1 = True, 0 = False.
Máxima Frecuencia Cardiaca Durante Ejercicio: Media entre 71 y 202 latidos por minuto
Angina inducida por ejercicio: 1 = si, 0 = no.
Segmento ST deprimido al hacer ejercicio: Media entre 0 y 6.2.
Tipo de Thalassemia: 1 = normal, 2 = tratamiento, 3 = reversible).

Gráficas de los datos

Distribución de Enfermedad Cardiaca de acuerdo a Indicadores





Resultado de la revisión del dataset y de los datos

1.) Se cuenta con un total de 1025 registros

2.) El dataset no contiene registros en NaN

3.) Podría pensarse que entre mas edad se tiene mayor es el riesgo de padecer Enfermedad Cardiaca pero no es así.

4.) Los pacientes con mayor frecuencia cardiaca tienen mayor riesgo de infarto

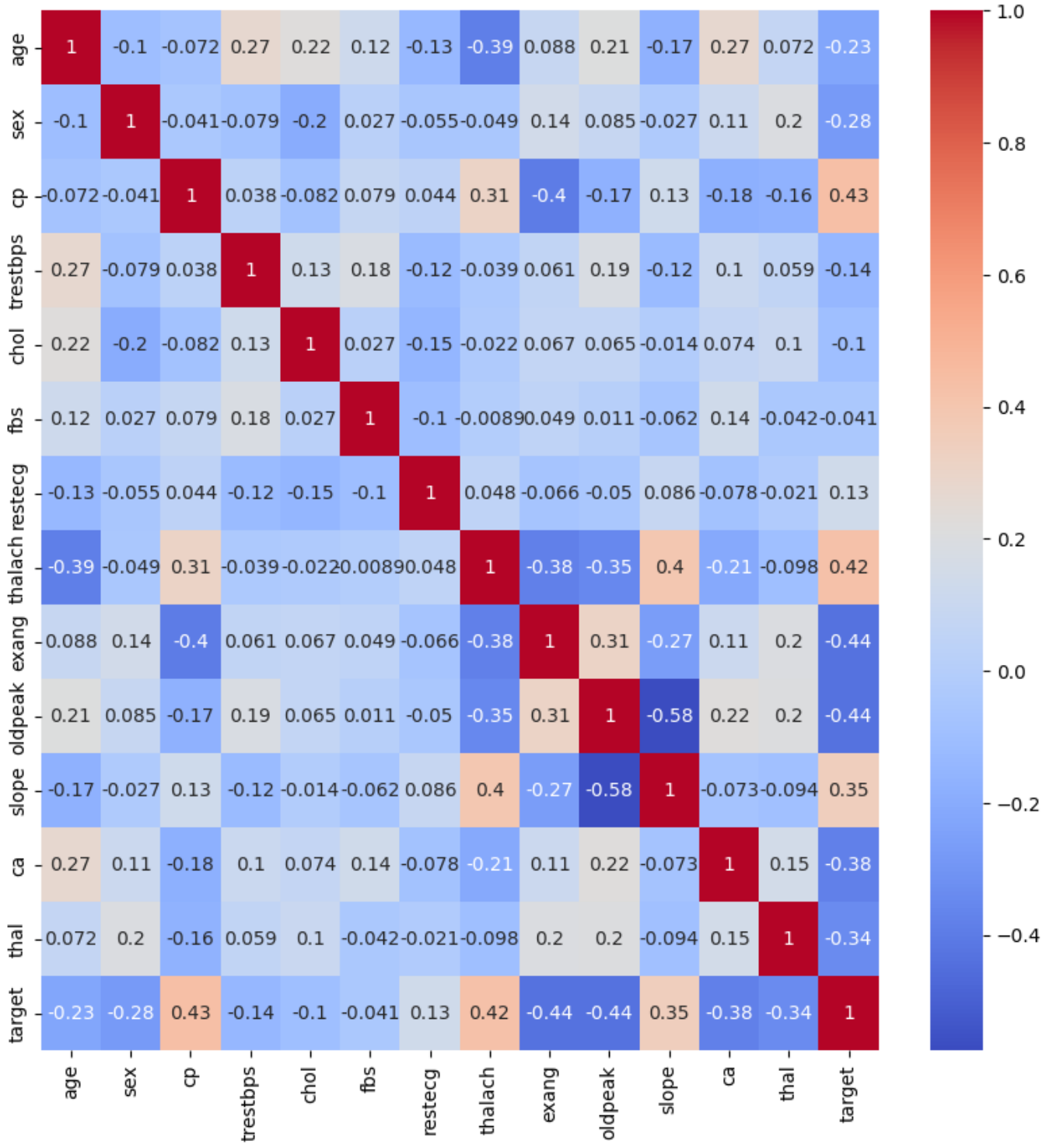
5.) Los pacientes que no presentan dolor (angina) tienen mayores posibilidades de infartarse

6.) Las mujeres tienen más probabilidad de infarto

7.) Niveles altos de colesterol aumentan el riesgo de infarto

8.) Los pacientes con Thalassemia tienen un riesgo muy alto de infarto

Correlación

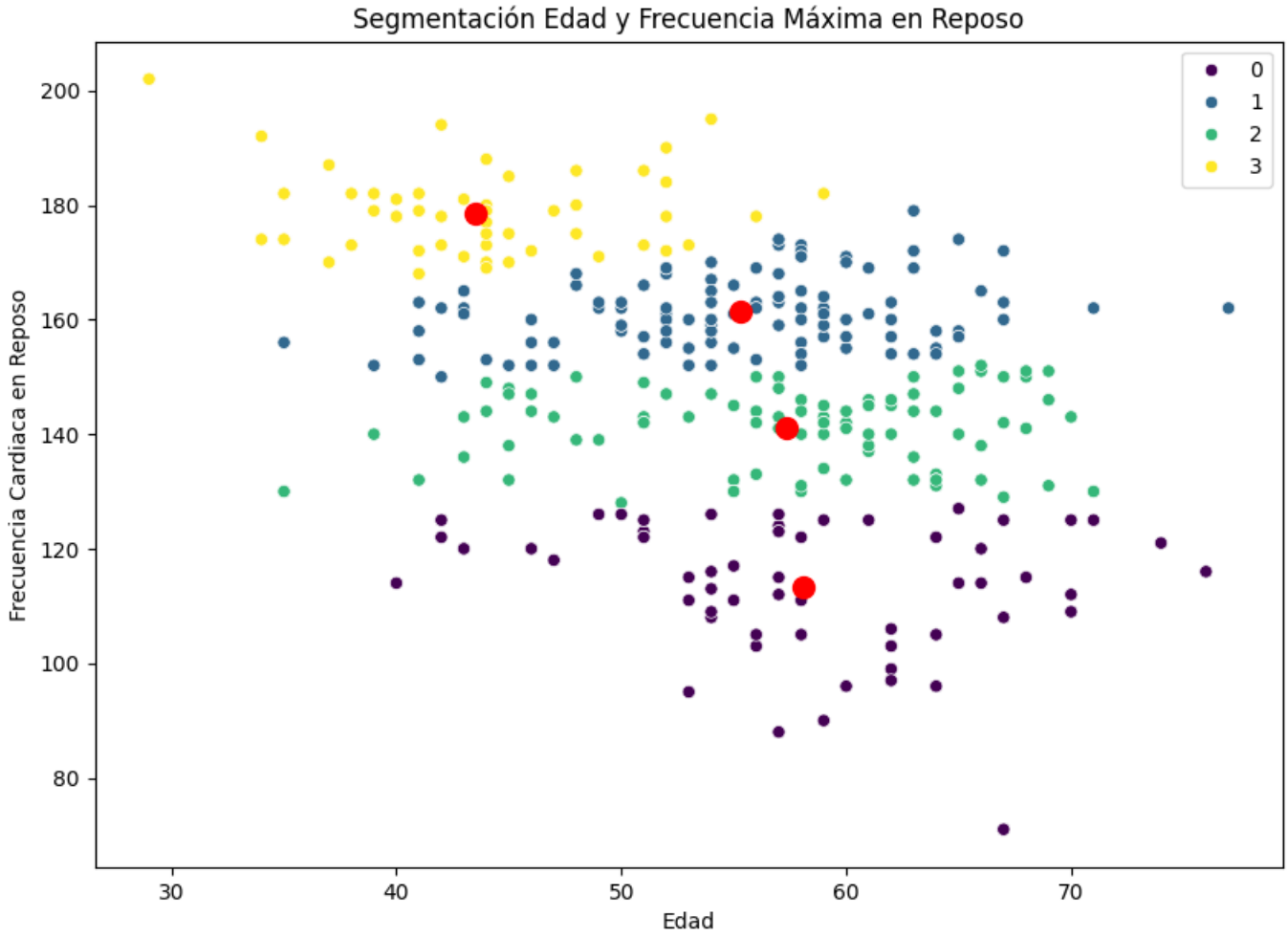


De la gráfica de correlación (mapa de calor) se deduce que hay correlación entre el dolor de pecho (angina), la frecuencia cardiaca elevada y la elevación del segmento ST al ejercitarse (Isquemia). Tener estos tres factores aumentan el riesgo de tener un ataque cardiaco.

KMEANS

SCORE=0.8349990481629546

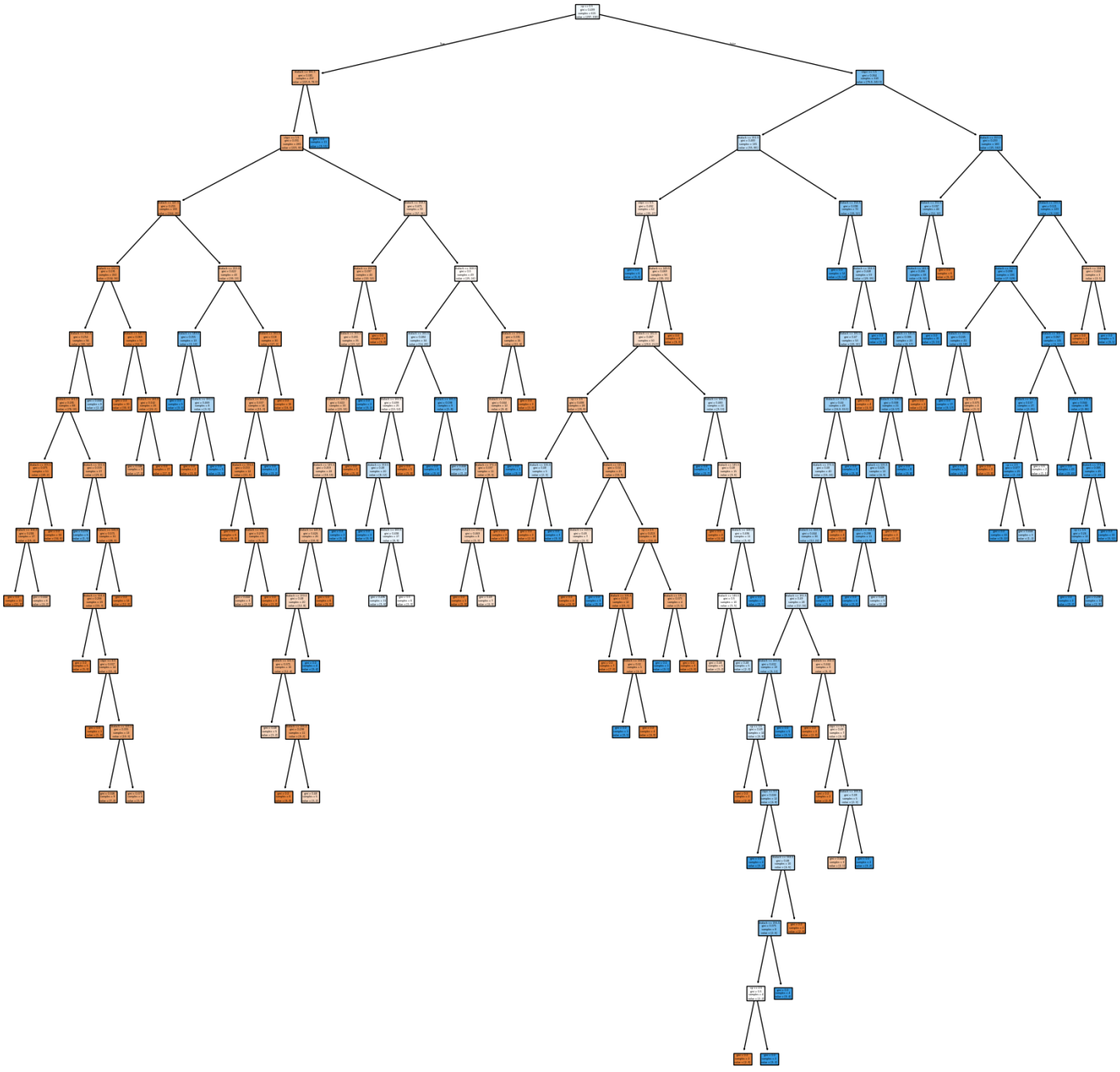
Utilizando KMEANS se revisa la segmentación de los pacientes por su edad y frecuencia cardiaca en reposo.



En la segmentación se puede observar que el grupo de pacientes que presentan mayor frecuencia cardiaca en reposo y con esto mayor riesgo de infarto se encuentra entre 35 y 52 años aproximadamente .

Árboles Decisión

SCORE=0.9164683222585133



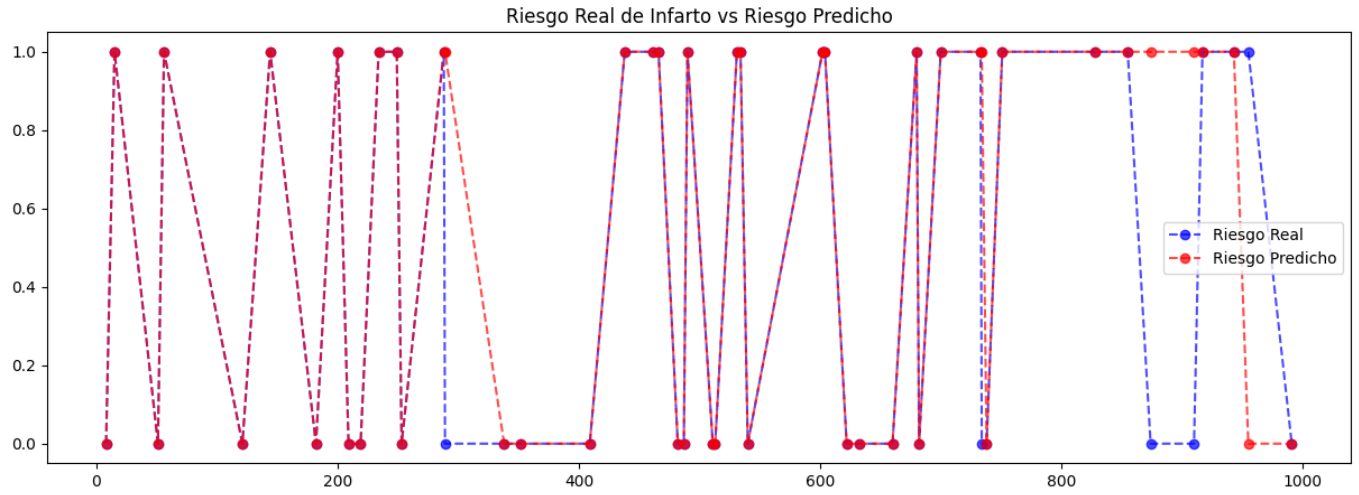
El modelo ha identificado el inicio de la clasificación con el indicador angina(dolor de pecho)

Regresión Logística

SCORE =0.8512195121951219

Con datos de prueba se corre el modelo para predecir el riesgo de infarto cardiaco o enfermedad cardiaca los resultados son:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target	target_pred	count
534	54	0	2	108	267	0	0	167	0	0.0	2	0	2	1	1	1
943	65	1	0	120	177	0	1	140	0	0.4	2	0	3	1	1	2
209	62	1	1	120	281	0	0	103	0	1.4	1	1	3	0	0	3
289	40	1	0	152	223	0	1	181	0	0.0	2	0	3	0	1	4
144	47	1	0	112	204	0	1	143	0	0.1	2	0	2	1	1	5
200	62	0	0	124	209	0	1	163	0	0.0	2	0	2	1	1	6
680	42	1	1	120	295	0	1	162	0	0.0	2	0	2	1	1	7
56	56	1	3	120	193	0	0	162	0	1.9	1	0	3	1	1	8
991	60	1	0	117	230	1	1	160	1	1.4	2	2	3	0	0	9
855	46	1	1	101	197	1	1	156	0	0.0	2	0	3	1	1	10
182	60	1	0	140	293	0	0	170	0	1.2	1	2	3	0	0	11
513	54	1	0	110	206	0	0	108	1	0.0	1	1	2	0	0	12
700	41	1	2	130	214	0	0	168	0	2.0	1	0	2	1	1	13
734	52	1	0	128	204	1	1	156	1	1.0	1	0	0	0	1	14
51	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0	0	15
482	51	1	0	140	298	0	1	122	1	4.2	1	3	3	0	0	16
910	50	1	2	140	233	0	1	163	0	0.6	1	1	3	0	1	17
751	65	0	2	160	360	0	0	151	0	0.8	2	0	2	1	1	18



Conclusión

Con los modelos podemos predecir a que indicadores se les debe prestar mayor atención para prevención de la enfermedad cardiaca o prevenir un infarto cardiaco por ejemplo glucosa en sangre o colesterol.El análisis de está información sirve para seleccionar pacientes que al tener un riesgo de infarto cardiaco puedan seer turnados al Área de Prevención de los Sistemas de Salud.