Data Science

www.datascience.pe

Data Science

Modelo predictivo para cuantificar el riesgo de mortalidad causada por insuficiencia cardiaca.

Alumnos:

Kevin Rivera Vergaray

Gustavo Coronel Berrospi

Alfonso Pastor Carreño

Problemática

Contexto

- *Las enfermedades cardiovasculares (ECV) son la principal causa de muerte a nivel mundial , cobrando aproximadamente 17,9 millones de vidas cada año , lo que representa el 31% de todas las muertes en todo el mundo .*

- *La mayoría de las enfermedades cardiovasculares se pueden prevenir abordando los factores de riesgo conductuales, como el consumo de tabaco, la dieta poco saludable y la obesidad, la inactividad física y el consumo nocivo de alcohol mediante estrategias de población.*

Antecedentes

- *Existe trabajos anteriores que lograban predecir la deserción y/o permanencia del estudiante.*

1) **MODELO PREDICTIVO PARA LA SUPERVIVENCIA Y LA MORTALIDAD PERIOPERATORIA EN PACIENTES CON CARCINOMA RENAL Y EXTENSIÓN VENOSA TUMORAL (D. Juan Ignacio Martínez Salamanca - Universidad Autónoma de Madrid).**
2) **Predicción de muerte súbita en pacientes con insuficiencia cardíaca crónica mediante el estudio de la dinámica periódica de la repolarización(S. Palacios, I. Cygankiewicz , A. Bayés-de-Luna , J.P. Martínez, E. Pueyo– XXXVIII Congreso Anual de la Sociedad Española de Ingeniería Biomédica)**
3) **Aplicación del puntaje de riesgo en Síncope de Boston para la predicción de mortalidad y desenlaces cardiovasculares en pacientes adultos(Mauricio Andrés Quintero Betancur - Universidad Nacional de Colombia)**

Planteamiento del problema

¿Cómo el modelo predictivo cuantifica el riesgo de mortalidad causada por insuficiencia cardiaca?

Objetivo

Cuantificar la mortalidad causada por la insuficiencia cardiaca utilizando un modelo predictivo.

Hipótesis

El modelo predictivo cuantifica el riesgo de mortalidad causada por la insuficiencia cardiaca.

Propuesta

Extracción y preparación de datos

- *Se explora la data de DATA Heart Failure Prediction que se encuentra en kaggle, el cual cuenta con 12 características clínicas para predecir eventos de muerte por insuficiencia cardiaca*
- *Se utilizará Python para extraer, hacer el preprocesamiento de la información y generar un modelo predictivo.*
- *Se utilizará un repositorio para guardar la información y sea consumida para el modelamiento.*

Desarrollo

VARIABLES

N°	VARIABLE	DESCRIPCIÓN DE LA VARIABLE
1	age	Edad del paciente.
2	anemia	Si el paciente sufre de anemia.
3	creatinine_phosphokinase	nivel de la encima CPK en la sangre (mcg/L). Cuando el nivel total de CPK es muy alto, a menudo significa que ha habido lesión o estrés en el corazón, el cerebro o el tejido muscular. 10 to 120 micrograms per liter (mcg/L)
4	diabetes	Si el paciente tiene diabetes.
5	ejection_fraction	porcentaje de sangre que sale del corazon en cada contracción. Una fracción de eyección de 55 por ciento o más se considera normal. Una fracción de eyección de 50 por ciento o menos se considera reducida. Una fracción de eyección de entre 50 y 55 por ciento generalmente se considera «limítrofe».
6	high_blood_pressure	Si el paciente tiene hipertensión.
7	platelets	Plaquetas en la sangre (kPlatelets/mL). Valores normales Hombre: 135-317 billones/L (De 135,000 a 317,000/mcL) ,Mujer: 157-371 billones/L (157,000-371,000/mcL)
8	serum_creatinine	nivel de creatinina en la sangre (mg/dL). Niveles normales suelen ser 0.7 a 1.3 mg/dL en hombres 0.5 a 1.2 mg/dL en mujeres
9	serum_sodium	nivel de sodio en la sangre (mEq/L). Un nivel normal de sodio en la sangre oscila entre 135 y 145 miliequivalentes por litro (mEq/L). La hiponatremia se produce cuando el sodio en el cuerpo se encuentra por debajo de 135 mEq/L.
10	sex	sexo.
11	smoking	si el pacientes es fumador.
12	time	tiempo de seguimiento (days).
13	DEATH_EVENT	si el paciente muere en el tiempo de seguimiento.

Diccionario de datos

Dentro de la DATA obtenida no se encontraron valores nulos.

Análisis estadístico variables Categóricas

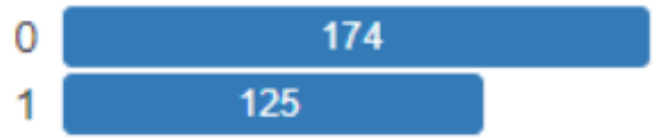


Fig. 1 Análisis de variable "anaemia"

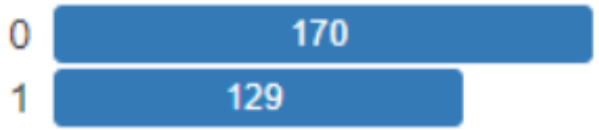


Fig. 2 Análisis de variable "diabetes"

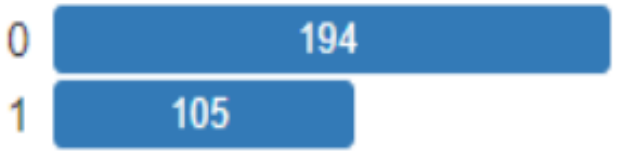


Fig. 3 Análisis de variable "high_blood_pressure"

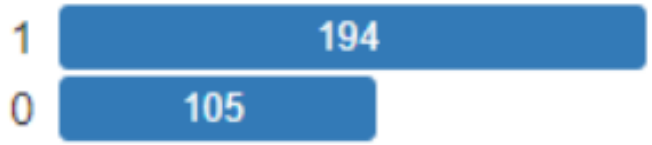


Fig. 4 Análisis de variable "sex"

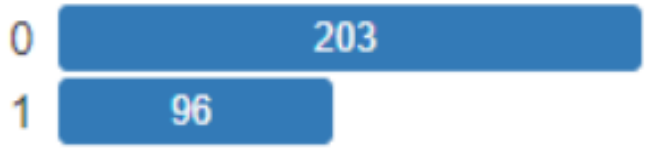
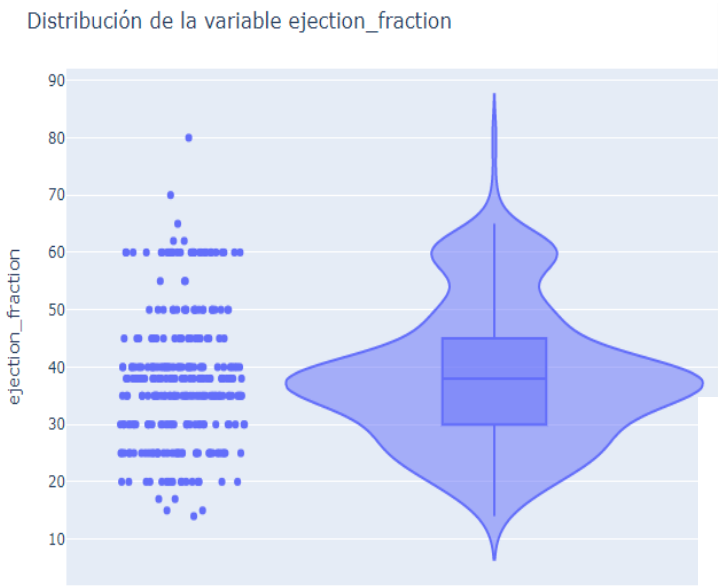


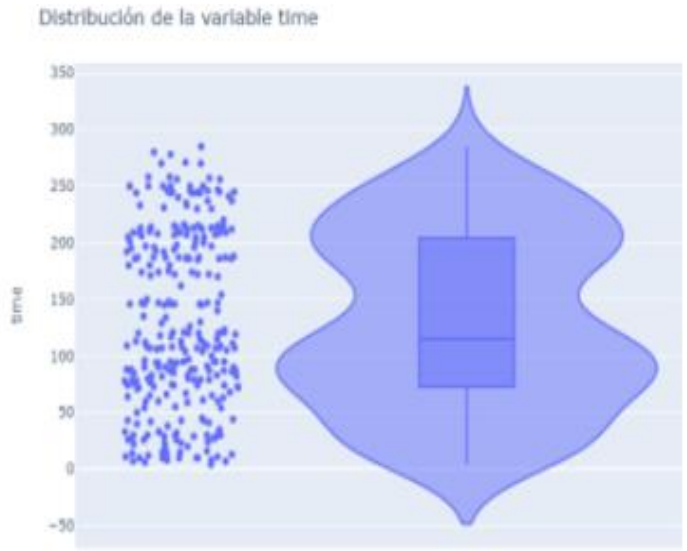
Fig. 5 Análisis de variable "smoking"

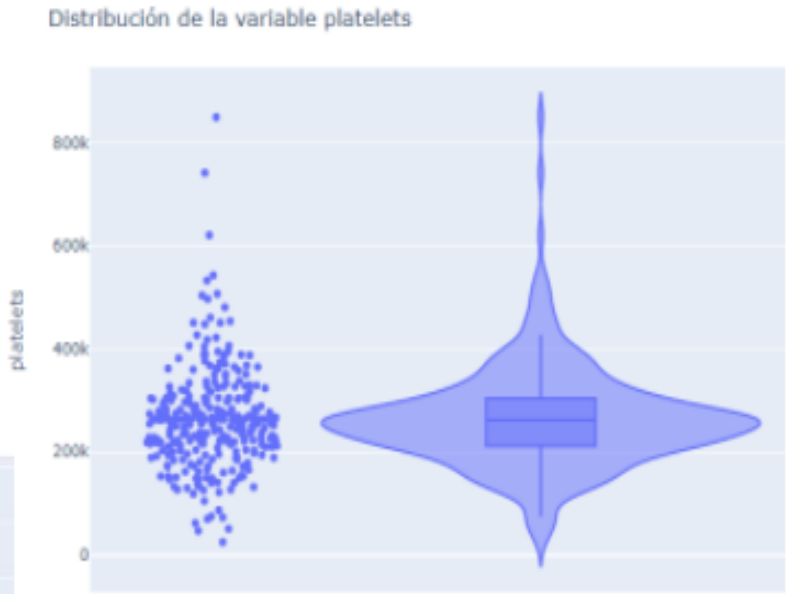
Análisis de percentiles

	age	creatinine_phosphokinase	ejection_fraction	platelets	serum_creatinine	serum_sodium	time
count	299	299	299	299	299	299	299
mean	60.834	581.839465	38.08361	263358	1.39388	136.625418	130.2609
std	11.895	970.287881	11.83484	97804.24	1.03451	4.412477	77.61421
min	40	23	14	25100	0.5	113	4
25%	51	116.5	30	212500	0.9	134	73
50%	60	250	38	262000	1.1	137	115
75%	70	582	45	303500	1.4	140	203
max	95	7861	80	850000	9.4	148	285

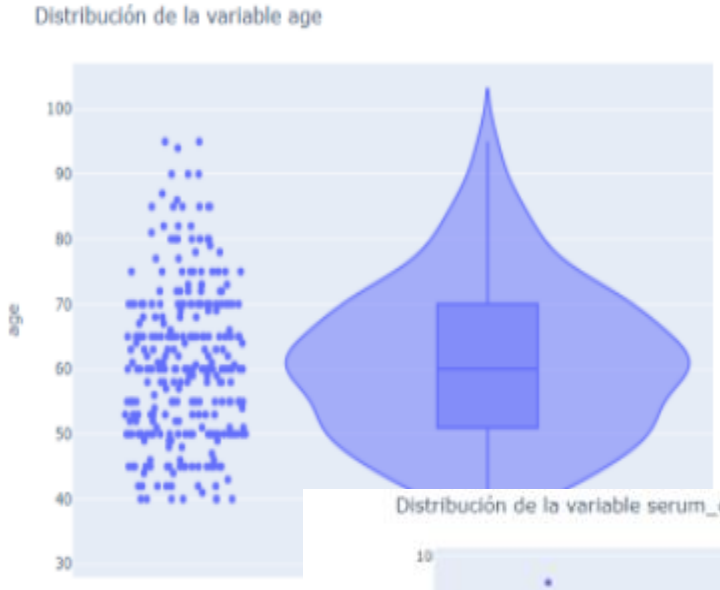
Análisis estadístico variables numéricas

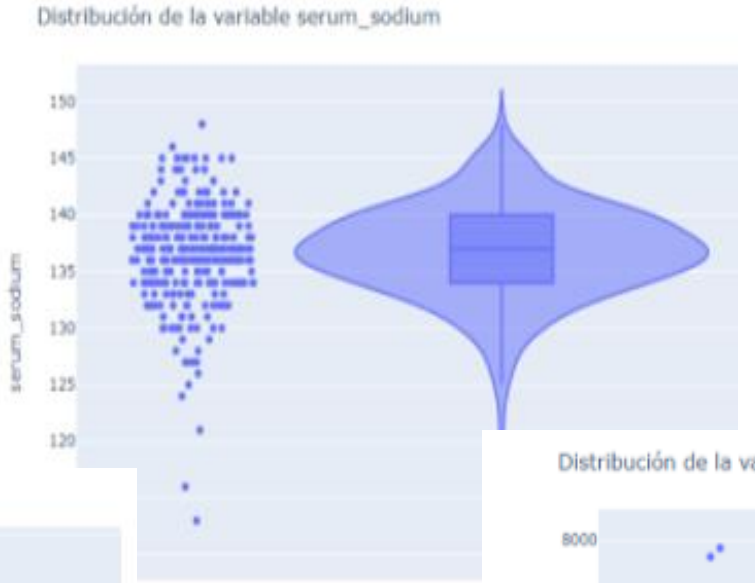


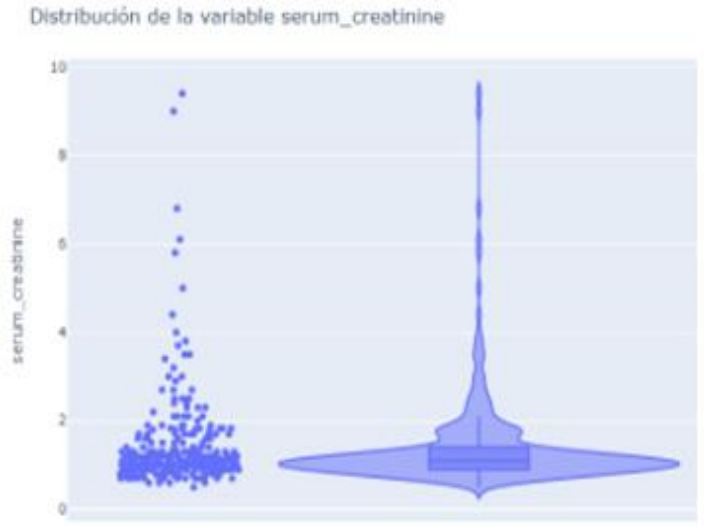


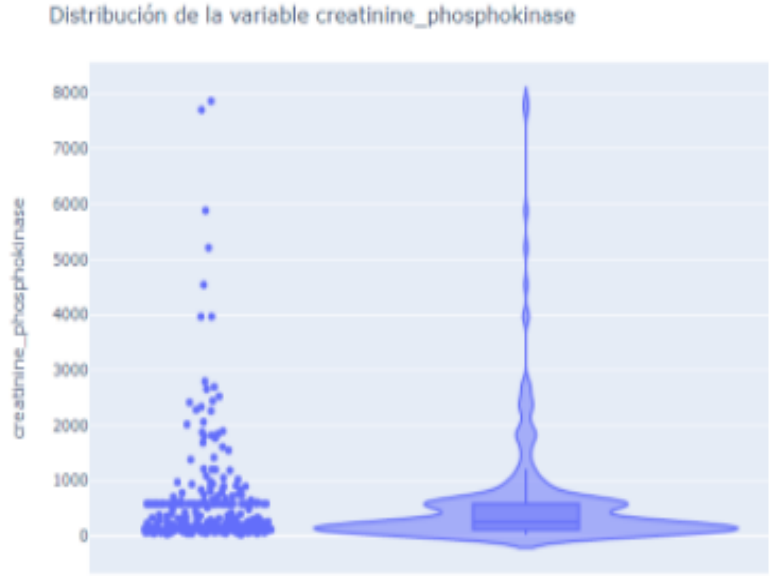


Análisis estadístico variables numéricas

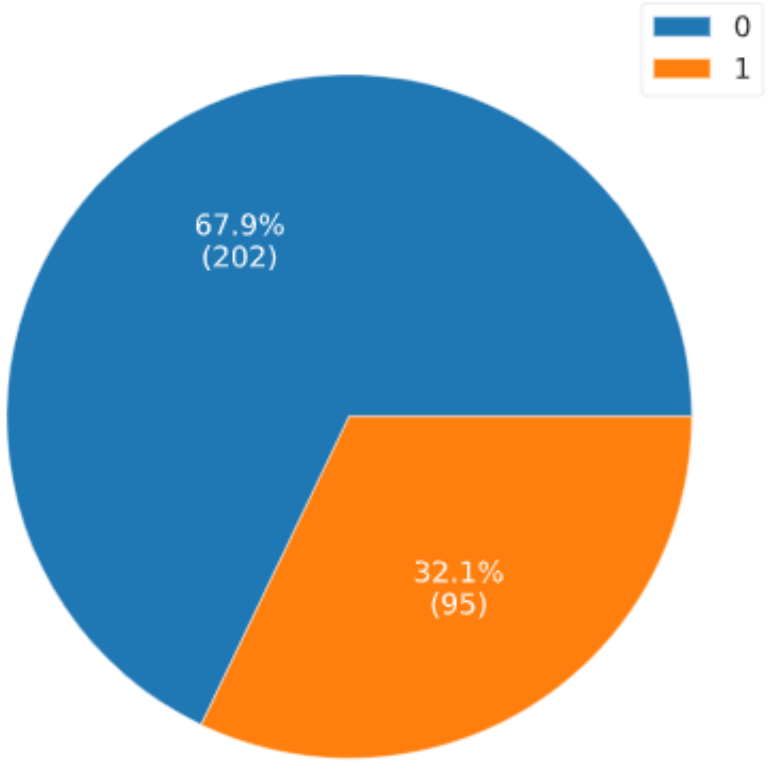








Análisis estadístico de la Target (DEATH_EVENT)



Análisis Bivariante Variables Categóricas

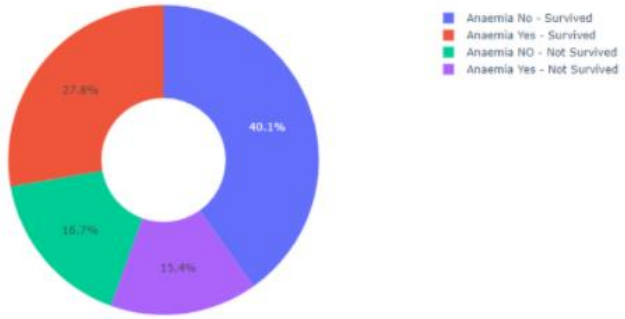


Fig. 14 Análisis de "anaemia vs DEATH_EVENT "

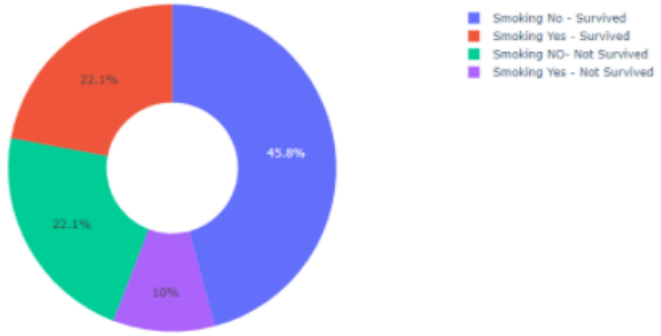


Fig. 18 Análisis de smoking vs DEATH_EVENT

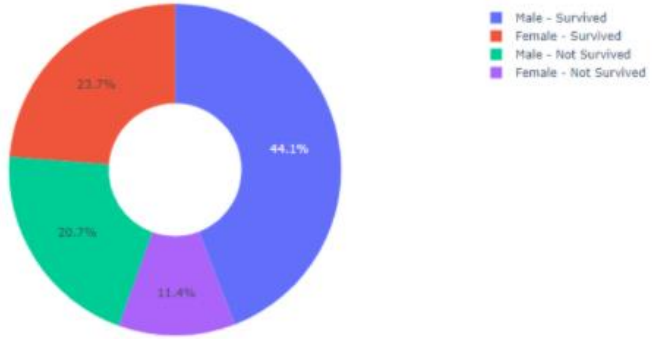


Fig. 17 Análisis de "sex vs DEATH_EVENT"

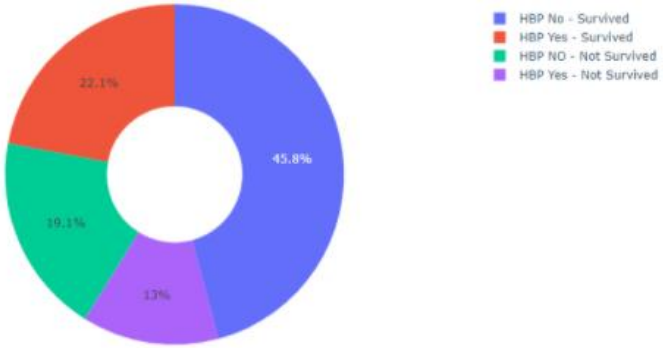


Fig. 16 Análisis de "high_blood_pressure vs DEATH_EVENT"

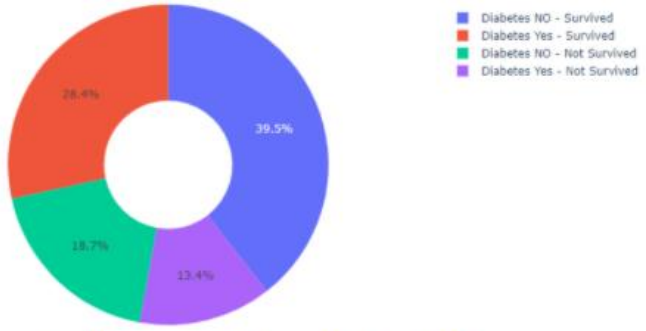
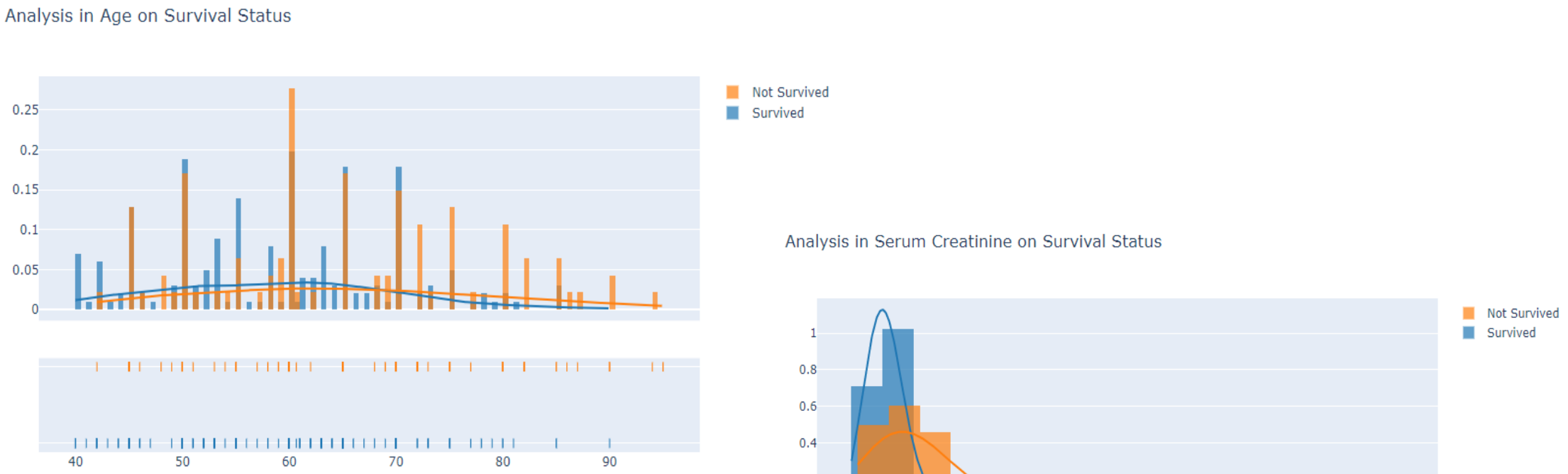
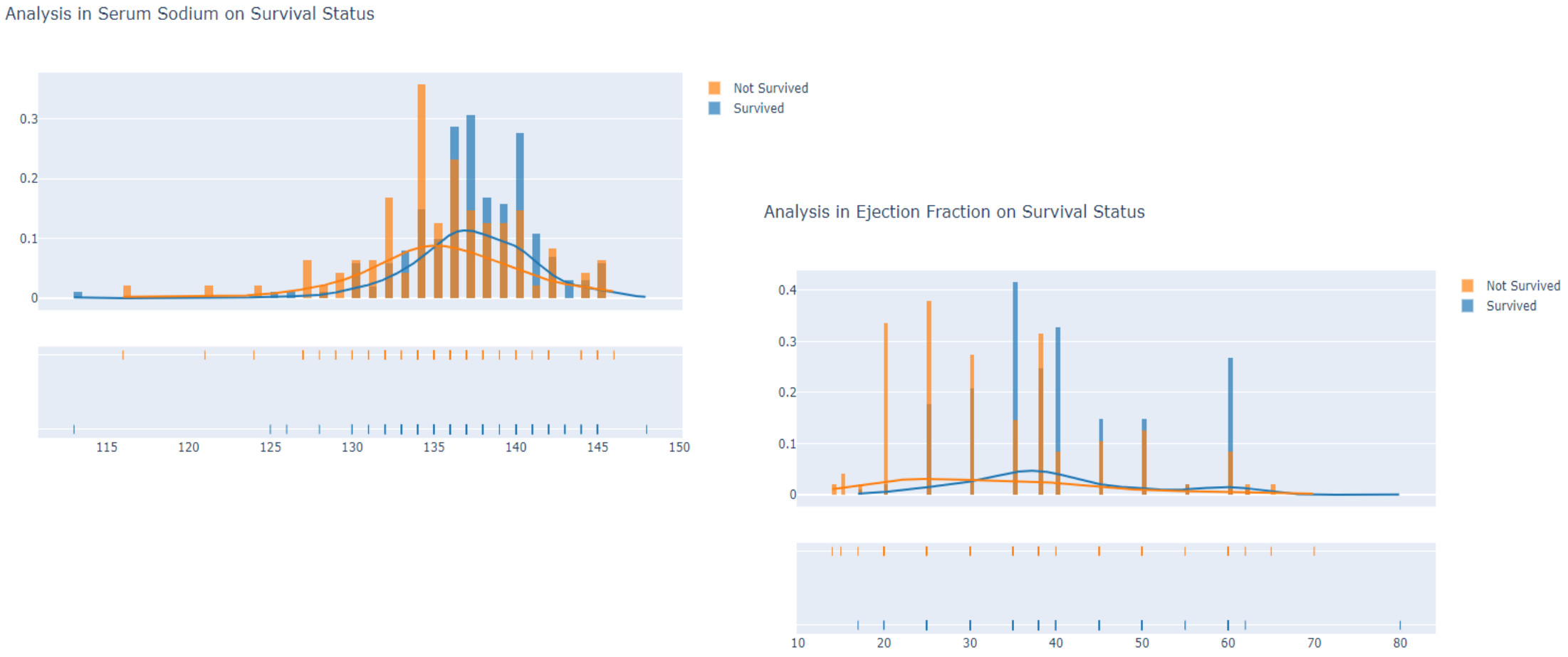


Fig. 19 Análisis de "diabetes vs DEATH_EVENT"

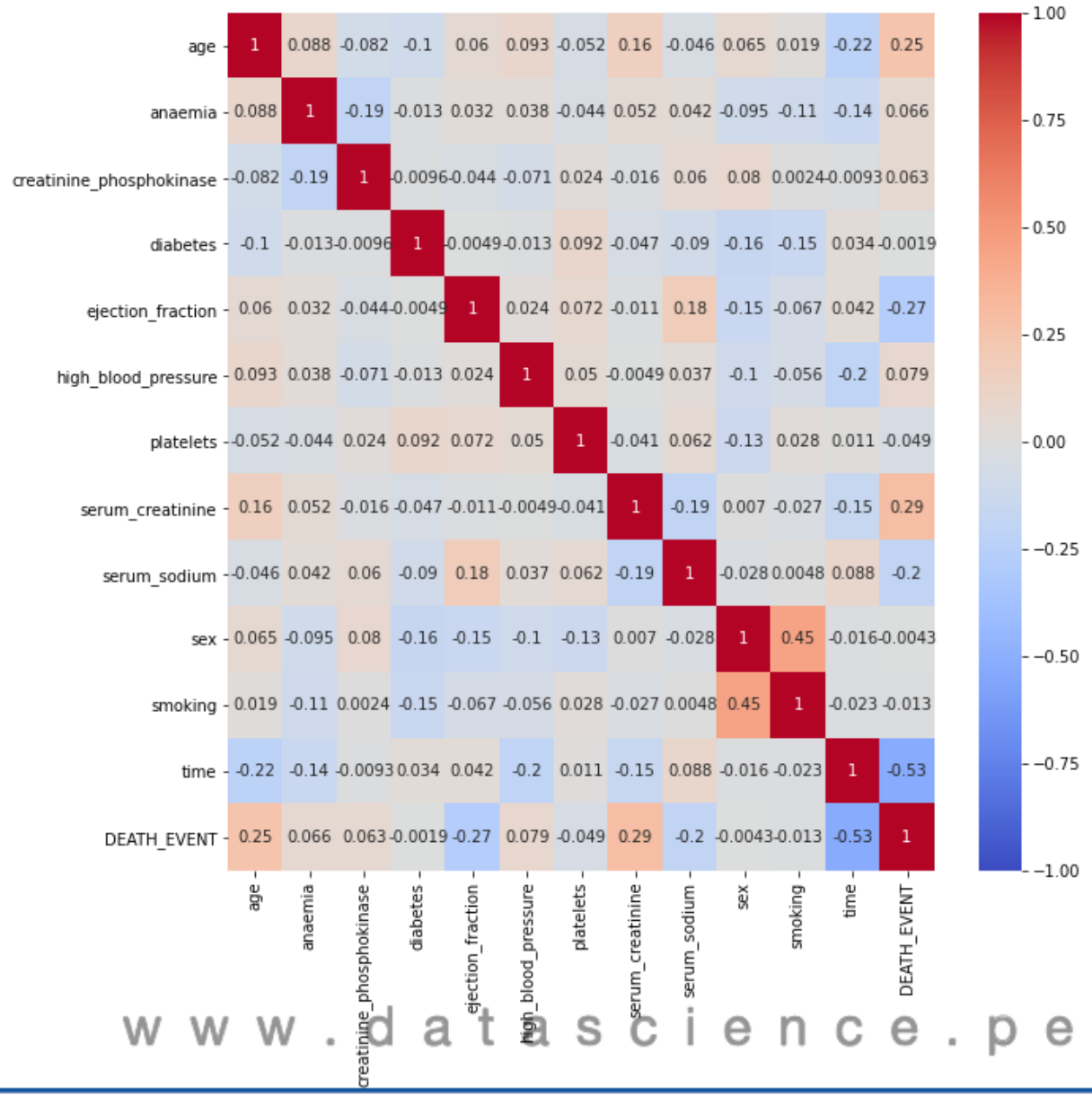
Análisis Bivariante Variables Numéricas



Análisis Bivariante Variables Numéricas



Análisis de correlación de datos



Preparación de los datos

- *Selección de datos: Predictoras y la Target*

  predictoras = ['anaemia','diabetes','high_blood_pressure','sex','smoking','age','creatinine_phosphokinase',

  'ejection_fraction', 'platelets', 'serum_creatinine', 'serum_sodium', ' time','total']

  target = 'DEATH_EVENT"

- *Limpieza de datos:*

  La data la encontramos bastante limpia no había datos nulos.

- *Ingeniería de variables*

  A Partir de la data se generaron las siguientes variables:

  cpk_ok, ejection_fraction_ok, platelets_ok,serum_creatinine_ok, serum_sodium_ok, total

Matriz de confusión (Regresión Logística)

Matriz de confusión (train)		
Predicción/Realidad	**Realmente Falleció**	**Realmente No Falleció**
Predicción de mortalidad	128	16
Predicción de No de no mortalidad	21	14

Matriz de confusión (test)		
Predicción/Realidad	**Realmente Falleció**	**Realmente No Falleció**
Predicción de mortalidad	55	4
Predicción de No de no mortalidad	12	19

Matriz de confusión (Árboles de decisión)

Matriz de confusión (train)		
Predicción/Realidad	**Realmente Falleció**	**Realmente No Falleció**
Predicción de mortalidad	141	9
Predicción de No de no mortalidad	15	44

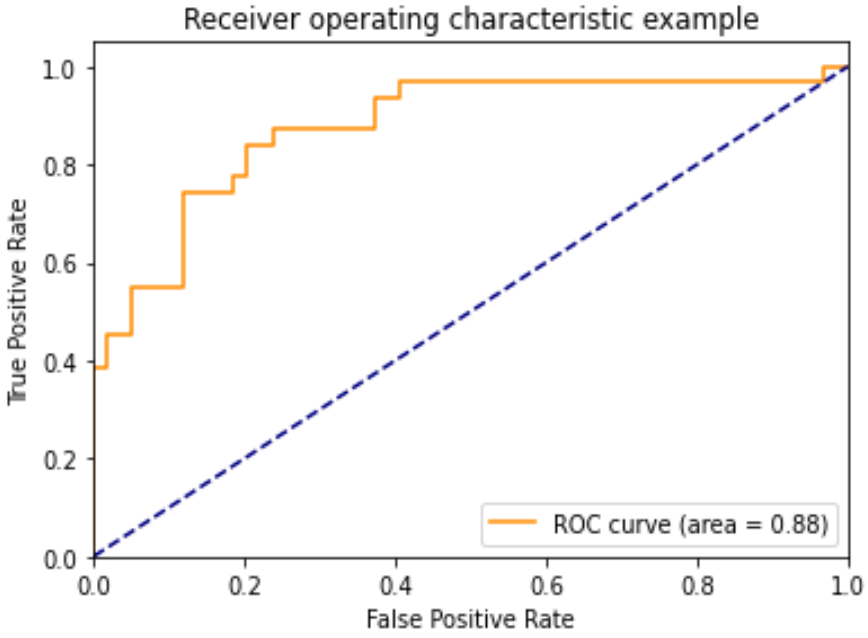
Matriz de confusión (test)		
Predicción/Realidad	**Realmente Falleció**	**Realmente No Falleció**
Predicción de mortalidad	51	2
Predicción de No de no mortalidad	16	21

Matriz de confusión (KNN)

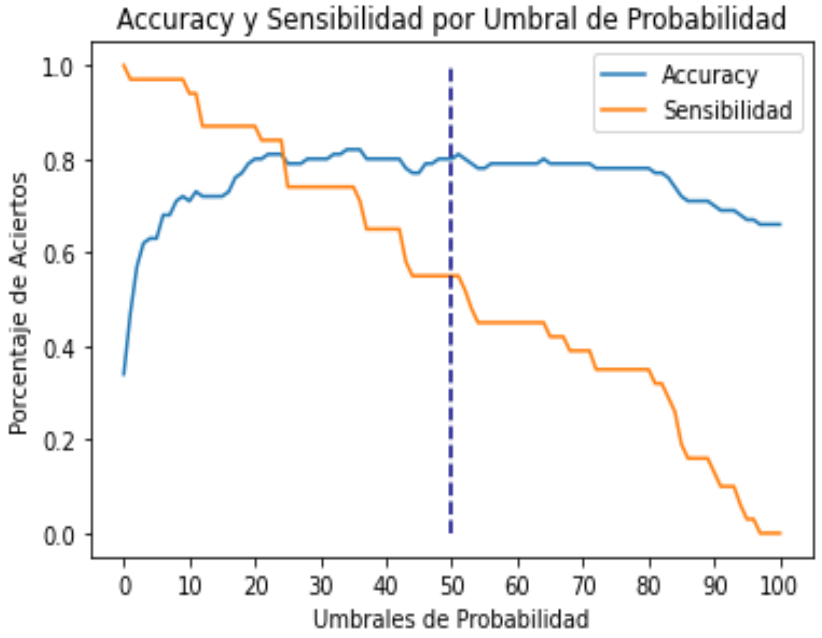
Matriz de confusión (train)		
Predicción/Realidad	**Realmente Falleció**	**Realmente No Falleció**
Predicción de mortalidad	147	13
Predicción de No de no mortalidad	45	34

Matriz de confusión (test)		
Predicción/Realidad	**Realmente Falleció**	**Realmente No Falleció**
Predicción de mortalidad	36	7
Predicción de No de no mortalidad	13	4

AUC Regresión Logística (MV)

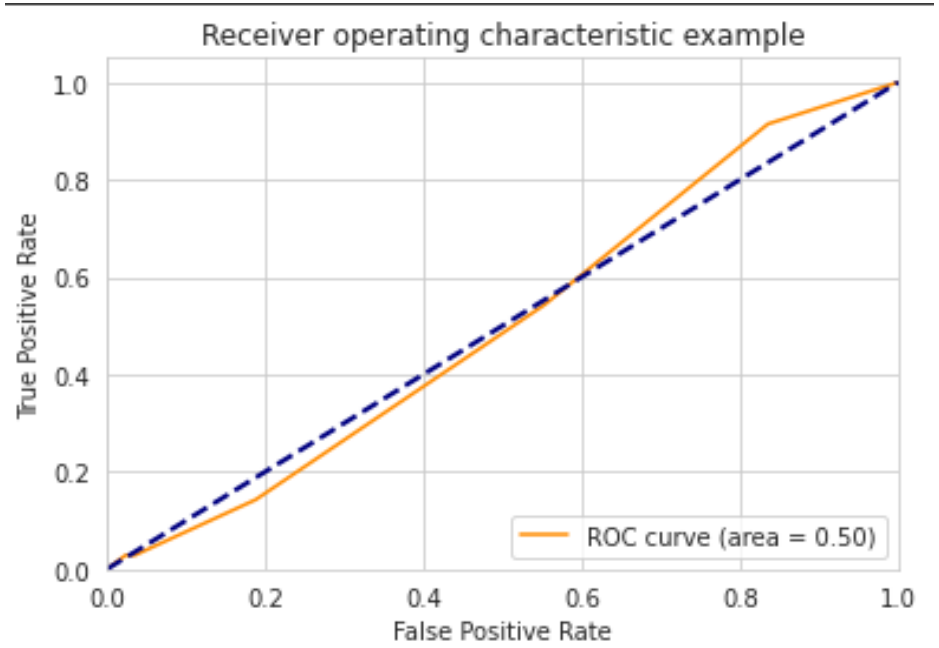


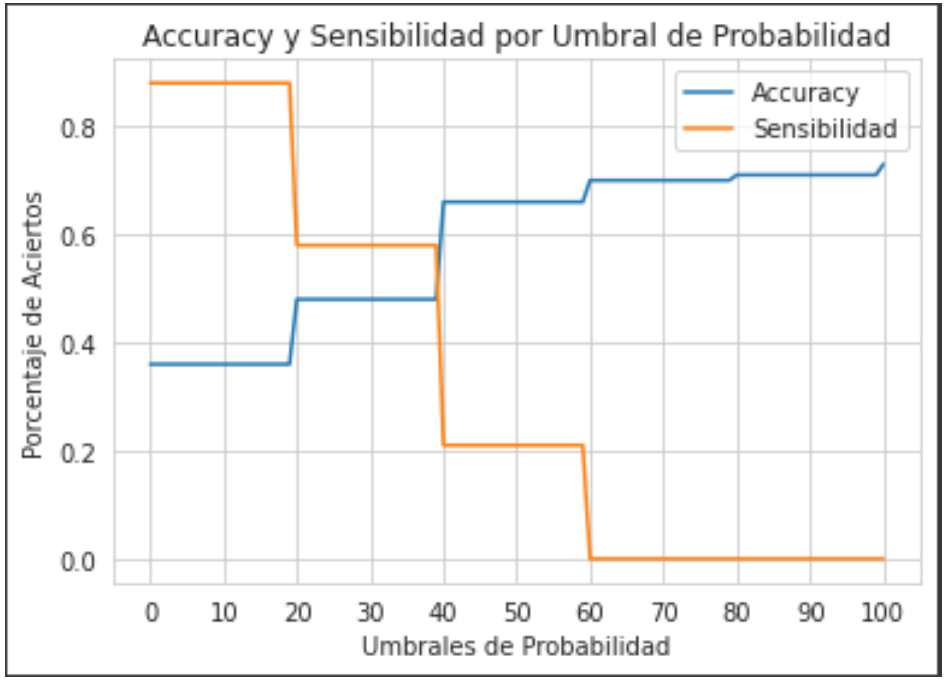
Al calcular la curva ROC se obtiene el área bajo la curva es 0.88 lo cual indica que tenemos un buen modelo.



Se visualiza que el mejor punto de corte es 0.09 ya que Accuracy en este punto es 0.72 y la Sensibilidad es 0.97.

AUC(KNN)





Al calcular la curva ROC se obtiene el área bajo la curva es 0.50 lo cual indica que no se cuenta con un buen diagnóstico .

COMPARACIÓN DE LOS MODELOS UTILIZADOS

Indicador Del Modelo	Modelos utilizados							
	Regresión Logística (GD)		Regresión Logística (MV)		Árboles de Decisión		KNN	
	Train	Test	Train	Test	Train	Test	Train	Test
Accuracy	0.82	0.82	0.86	0.86	0.88	0.8	0.75	0.66
Sensibilidad	0.67	0.61	0.72	0.68	0.74	0.57	0.43	0.23
Precisión	0.73	0.83	0.81	0.88	0.83	0.91	0.72	0.36
AUC	--	--	0.89	0.88	-	-	0.79	0.50
GINI	--	--	0.79	0.75	-	-	0.59	0.00

Conclusiones

- Habiendo aplicado el modelo de regresión logística por gradiente del descenso se obtuvo los siguientes resultados: El accuracy salió 82% y 82% para el train y el test respectivamente, la sensibilidad salió 67% y 61% para el traint y test respectivamente y la precisión 73% y 83% para el traint y el test respectivamente, se concluye que se obtuvo un accuray y una precisión bastante aceptable a diferencia de la sensibilidad que es poco aceptable ya que salió con porcentajes bajos.

Conclusiones

- Habiendo aplicado el modelo de regresión logística por Máxima Verosimilitud se obtuvo los siguientes resultados: El accuracy salió 86% y 86% para el train y el test respectivamente, la sensibilidad salió 72% y 68% para el traint y test respectivamente, la precisión 81% y 88% para el traint y el test respectivamente, el AUC salió 89% y 88% para el traint y el tes respectivamente, el GINI salió 79% y 75% respectivamente, se concluye que se obtuvo un accuray y una precisión bastante aceptable, una la sensibilidad que es regularmente aceptable ya que salió con porcentajes no muy altos, un AUC y un GINI que superan el 75% lo cual los hace aceptables.

Conclusiones

- Habiendo aplicado el modelo de árboles de decisión se obtuvo los siguientes resultados: El accuracy salió 88% y 80% para el train y el test respectivamente, la sensibilidad salió 74% y 57% para el traint y test respectivamente y la precisión 91% y 72% para el traint y el test respectivamente, se concluye que se obtuvo un accuray y una precisión bastante aceptable a diferencia de la sensibilidad que es poco aceptable ya que salió con un bajo porcentaje para el test.

Conclusiones

- Habiendo aplicado el modelo de KNN se obtuvo los siguientes resultados: El accuracy salió 75% y 66% para el train y el test respectivamente, la sensibilidad salió 43% y 23% para el traint y test respectivamente, la precisión 72% y 36% para el traint y el test respectivamente, el AUC salió 79% y 50% para el traint y el tes respectivamente, el GINI salió 59% y 75% respectivamente, se concluye que se obtuvo un aceptable, una precisión y una sensibilidad no aceptable ya que salieron con porcentajes muy bajos, un AUC poco aceptable para el test y un GINI muy poco aceptable.
- Habiendo ejecutado 4 modelos predictivos se concluye que el mejor modelo predictivo para cuantificar la mortalidad por insuficiencia cardiaca es el modelo de regresión logística por máxima verosimilitud.

¡Gracias!