Análisis Exploratorio de Datos, Extracción de Características

Universidad Politécnica Salesiana

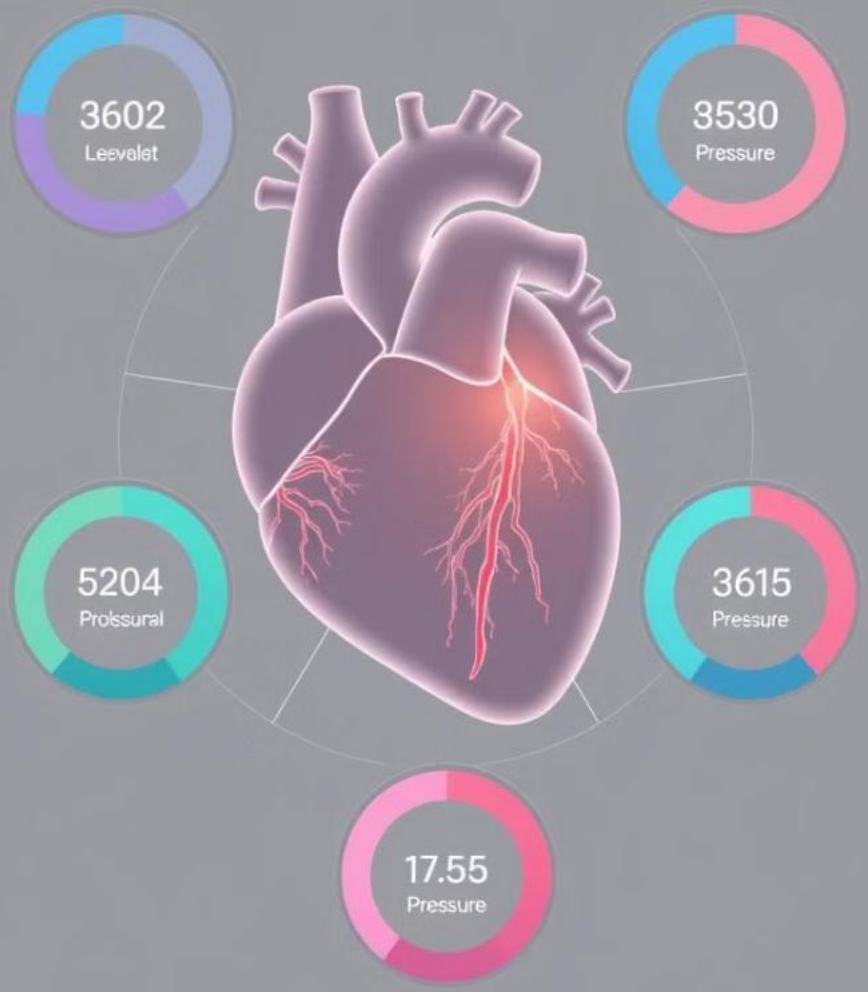
Michael Israel Lata Zambrano

Computación

Inteligencia Artificial

Marzo, 2025

Indice



Heart Disease Dataset

Fuente: UCI Machine Learning Repository

Enlace: https://archive.ics.uci.edu/dataset/45/heart+diseaseq

Numero de variables: 14 atributos.

Observaciones : 303 registros. registros.

Variable objetivo (TARGET): Predicción de la enfermedad cardiaca.

Descripción del Dataset

Tiene como finalidad de evaluar la predicción de la enfermedad cardiaca. Está compuesta por

Número de Variables: 14
Número de instancias(Pacientes): 303
Variable de estudio (Objetivo): TARGET
Sin enfermedad:1
Con enfermedad:2

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   AGE       303 non-null    float64
 1   SEX       303 non-null    float64
 2   CP        303 non-null    float64
 3   TRESTBPS  303 non-null    float64
 4   CHOL      303 non-null    float64
 5   FBS       303 non-null    float64
 6   RESTECG   303 non-null    float64
 7   THALACH   303 non-null    float64
 8   EXANG     303 non-null    float64
 9   OLDPEAK   303 non-null    float64
 10  SLOPE     303 non-null    float64
 11  CA        299 non-null    float64
 12  THAL      301 non-null    float64
 13  TARGET    303 non-null    int64
dtypes: float64(13), int64(1)
memory usage: 33.3 KB
```

Resumen Estadistico

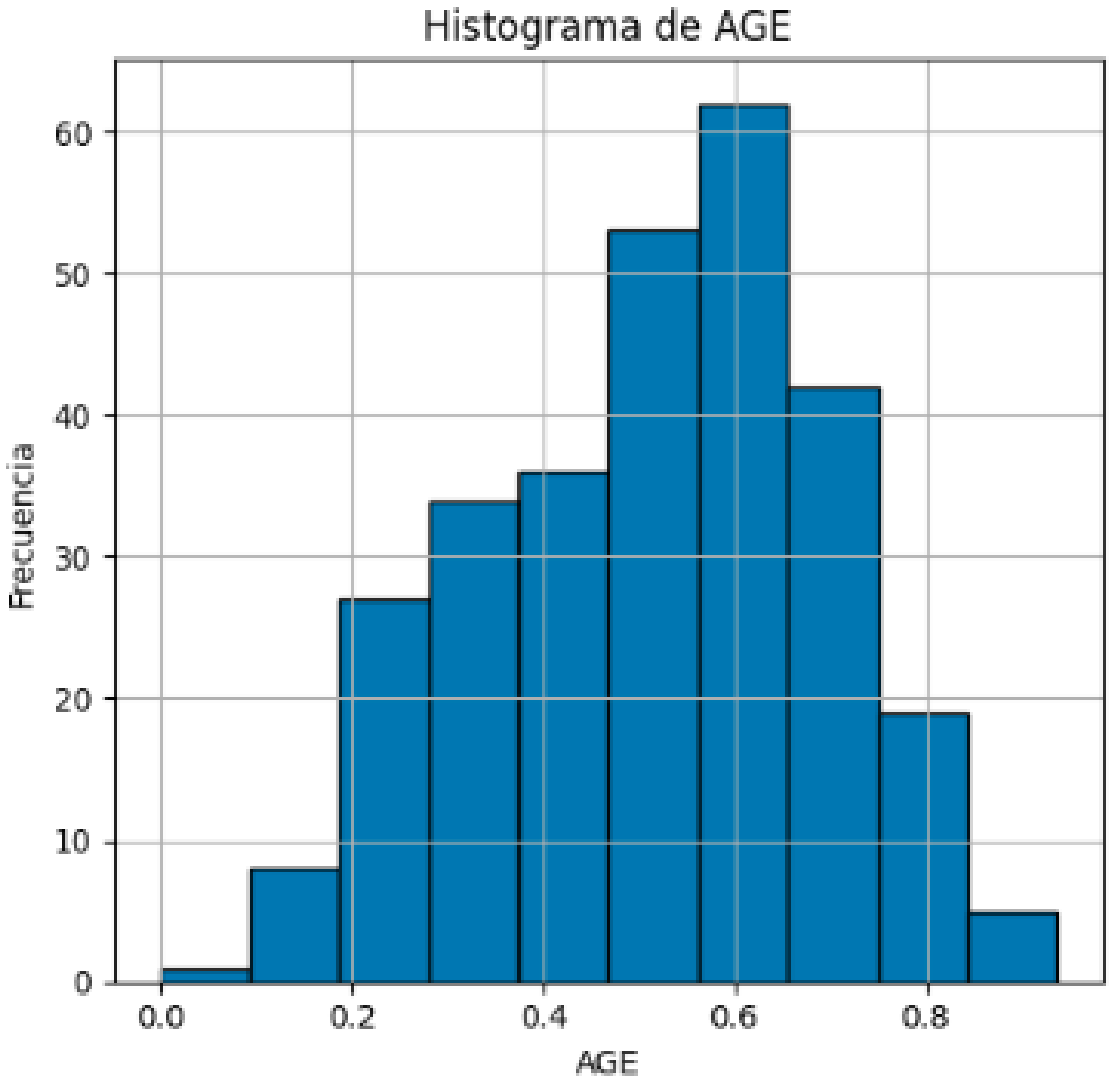
- Edad(AGE)
- Frecuencia Cardiaca Máxima(THALACH)
- Colesterol(CHOL)
- Presión Arterial en Reposo(TRESTBPS)
- Vasos Coloreados(CA)

	AGE	SEX	CP	TRESTBPS	CHOL	FBS	RESTECG	THALACH	EXANG	OLDPEAK	SLOPE	CA	THAL	TARGET
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	299.000000	301.000000	303.000000
mean	54.438944	0.679868	3.158416	131.689769	246.693069	0.148515	0.990099	149.607261	0.326733	1.039604	1.600660	0.672241	4.734219	0.937294
std	9.038662	0.467299	0.960126	17.599748	51.776918	0.356198	0.994971	22.875003	0.469794	1.161075	0.616226	0.937438	1.939706	1.228536
min	29.000000	0.000000	1.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	1.000000	0.000000	3.000000	0.000000
25%	48.000000	0.000000	3.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	3.000000	0.000000
50%	56.000000	1.000000	3.000000	130.000000	241.000000	0.000000	1.000000	153.000000	0.000000	0.800000	2.000000	0.000000	3.000000	0.000000
75%	61.000000	1.000000	4.000000	140.000000	275.000000	0.000000	2.000000	166.000000	1.000000	1.600000	2.000000	1.000000	7.000000	2.000000
max	77.000000	1.000000	4.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	3.000000	3.000000	7.000000	4.000000

Figuras Más Relevantes

Histograma

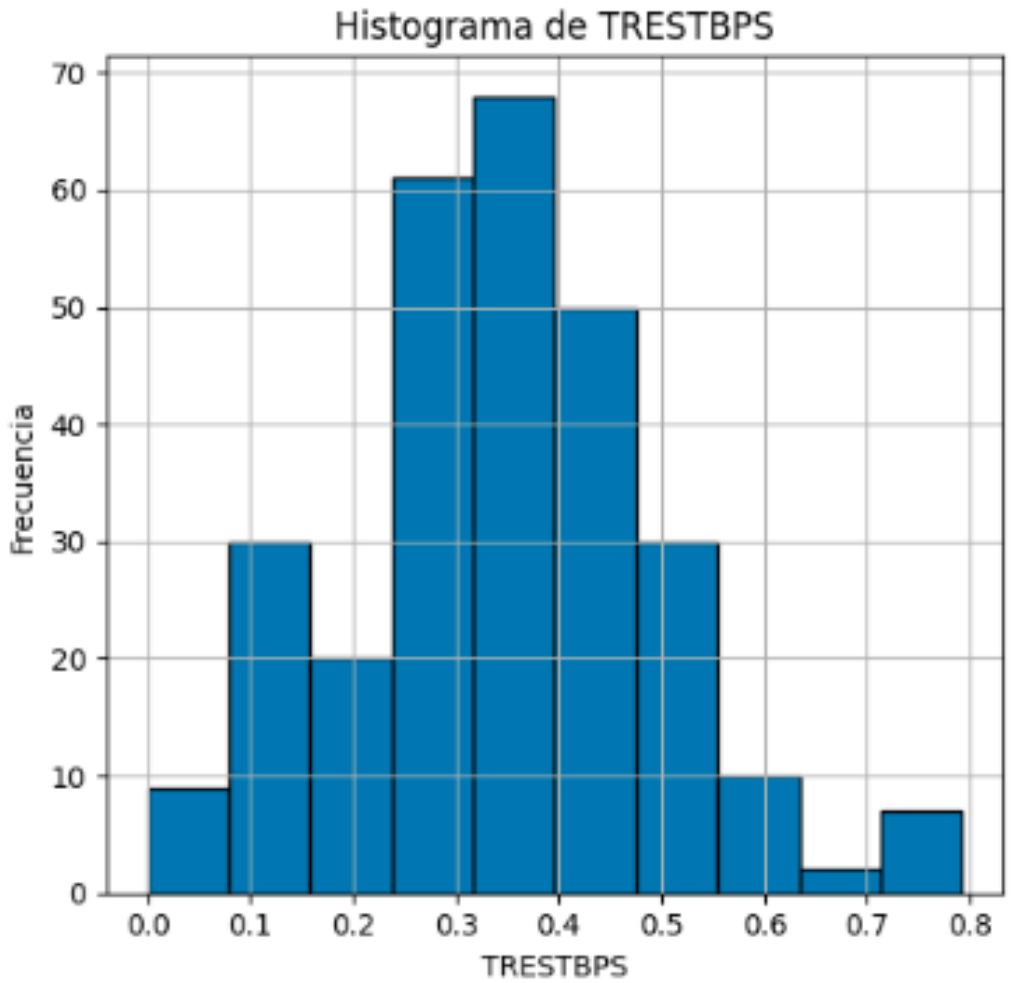
Representa la mayoría de los pacientes tienen entre 40 y 60 años, lo que sugiere que esta población debe ser prioritaria para intervenciones preventivas.



Figuras Más Relevantes

Histograma

Representa a los pacientes con valores bajos de frecuencia cardíaca máxima tienden a presentar una mayor probabilidad de padecer enfermedad cardíaca, representada por valores altos en la columna (cercanos a 1).



Figuras Más Relevantes

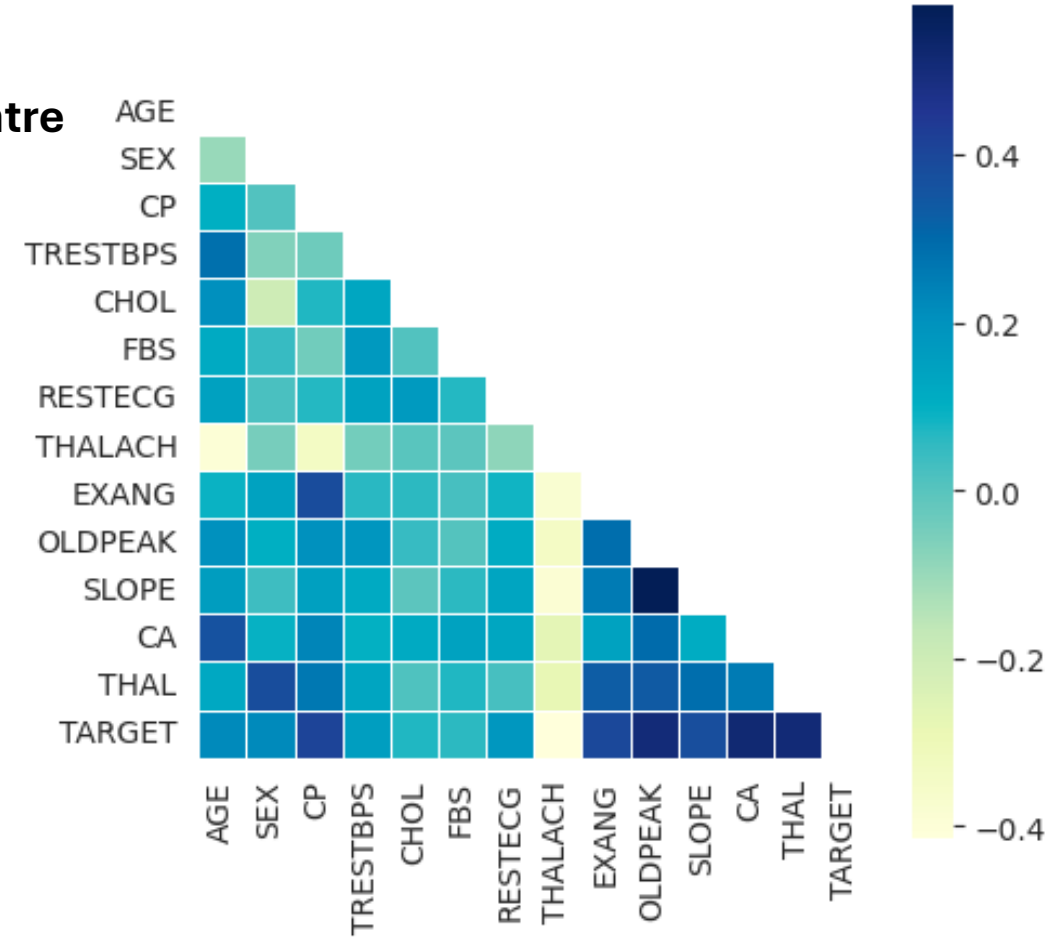
Mapa de Calor

¿Cómo están relacionadas las distintas variables entre sí?

Edad (AGE) y Frecuencia Cardíaca Máxima (THALACH): A mayor edad, menor frecuencia cardíaca máxima.

Presión Arterial (TRESTBPS) y Colesterol (CHOL) : Altos niveles de colesterol suelen asociarse con presión arterial elevada.

Vasos Coloreados (CA) y Enfermedad (TARGET): Más vasos coloreados, mayor probabilidad de enfermedad cardíaca.



Datos que pueden influir en la predicción de enfermedad cardíaca

Variables Categóricas Ordinales:

Edad (AGE): El rango de edad puede ser clave para identificar grupos de riesgo.

Frecuencia Cardíaca Máxima (THALACH): Indicador ordinal de la capacidad del corazón ante estrés.

Variables Categóricas Nominales:

Presión Arterial (TRESTBPS): Un factor crítico que define el estado cardiovascular.

Colesterol (CHOL) : Niveles altos pueden asociarse con mayor riesgo cardíaco.

```
Frecuencias de la columna 'CA':
CA
0.000000    169
0.333333     62
0.666667     35
1.000000     17
Name: count, dtype: int64
```

```
Porcentajes:
CA
0.000000    59.717314
0.333333    21.908127
0.666667    12.367491
1.000000     6.007067
Name: proportion, dtype: float64
```

```
Frecuencias de la columna 'TARGET':
TARGET
0.00    157
0.25     54
0.50     34
0.75     31
1.00     11
Name: count, dtype: int64
```

```
Porcentajes:
TARGET
0.00    54.703833
0.25    18.815331
0.50    11.846690
0.75    10.801394
1.00     3.832753
Name: proportion, dtype: float64
```

Variables Más Correlacionadas

- **Frecuencia Cardíaca Máxima (THALACH):**

Los pacientes con frecuencias cardíacas bajas tienden a tener mayor probabilidad de enfermedad cardíaca.
Relación inversa moderada con la variable objetivo **TARGET**

- **Número de Vasos Coloreados (CA):**

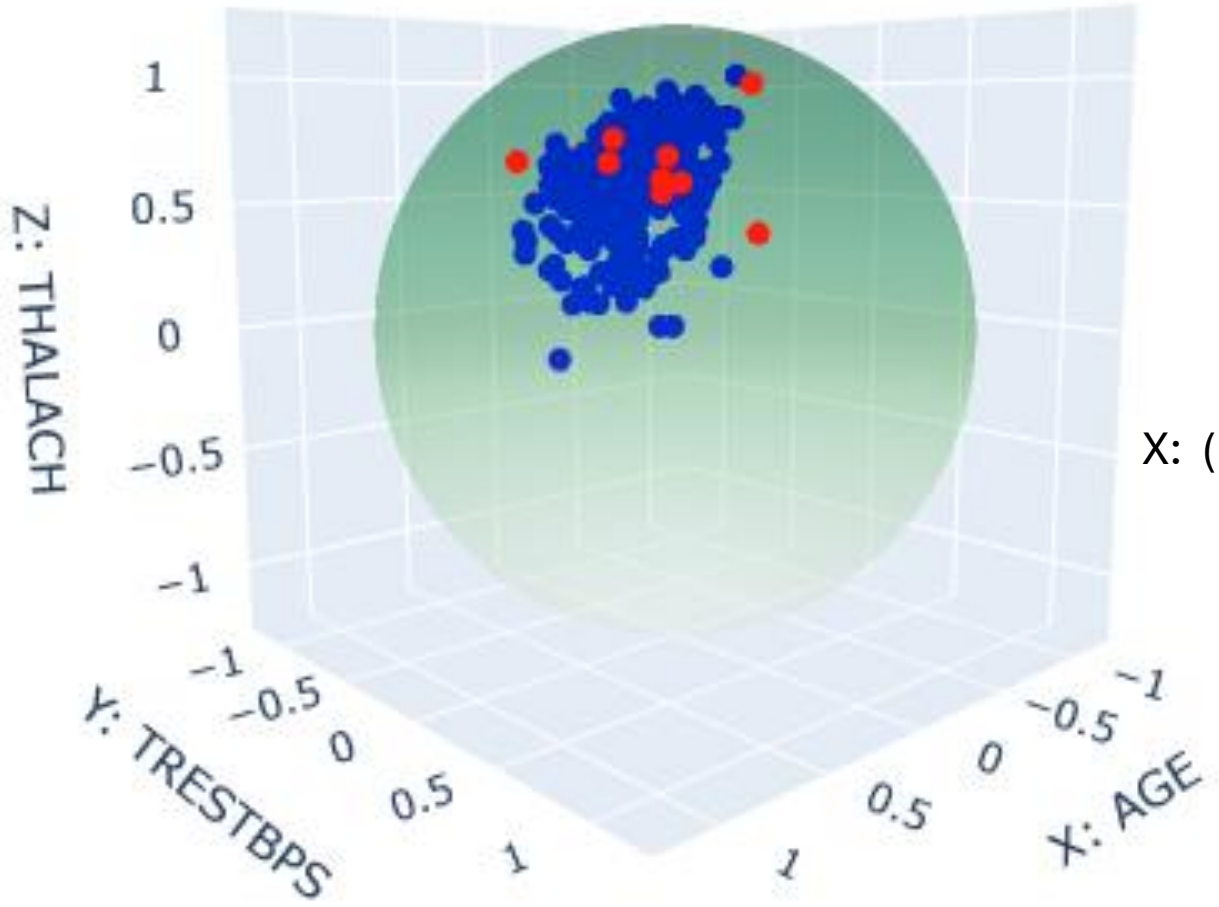
Más vasos coloreados están correlacionados con mayor probabilidad de enfermedad cardíaca.
Relación positiva con **TARGET**.

- **Colesterol Total (CHOL):**

Los niveles altos de colesterol suelen estar asociados con mayor riesgo cardíaco.
Relación positiva leve con **TARGET** .

Análisis De Outliers

Z: (Frecuencia Cardíaca Máxima) – Variable dependiente.



X: (Edad) – Edad de los pacientes.

Y: (Colesterol Total) – Indicador del nivel de colesterol.

Hipotesis

.

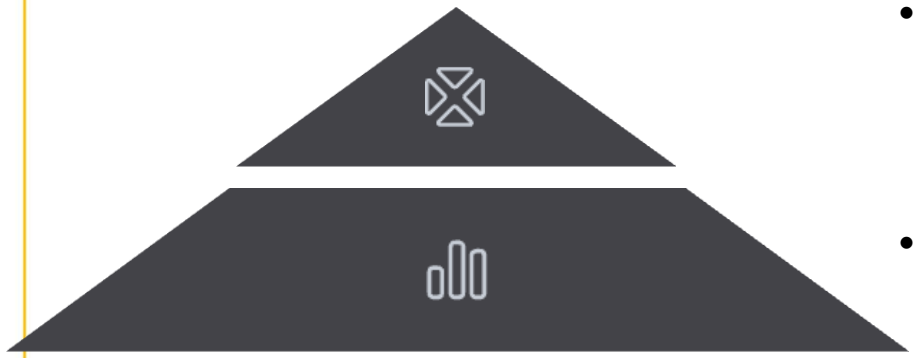
Aumentar la frecuencia cardíaca máxima mediante actividad física puede reducir el riesgo de enfermedad cardíaca en pacientes mayores.

La reducción de los niveles de colesterol mediante intervenciones médicas o cambios en el estilo de vida podría tener un impacto significativo en la prevención de enfermedades cardiovasculares.

Una frecuencia cardíaca máxima más alta está asociada con un menor riesgo de enfermedad cardíaca. Intervenciones físicas regulares podrían mejorar esta métrica y reducir el riesgo.

Conclusiones

.



- El conjunto de datos Heart Disease Dataset permite construir modelos predictivos para determinar la probabilidad de enfermedad cardíaca en pacientes, utilizando variables clave como frecuencia cardíaca máxima, número de vasos coloreados y niveles de colesterol.
- El análisis exploratorio revela correlaciones significativas y patrones generales, como la relación negativa entre frecuencia cardíaca máxima y riesgo de enfermedad, además de posibles sesgos en los datos que podrían influir en futuros modelos predictivos.

UNIVERSIDAD POLITÉCNICA SALESIANA ECUADOR

Más que formar Profesionales, *Transformamos Vidas*

@upsalesianaec

www.ups.edu.ec