ICAP
Instituto Centroamericano de Administración Pública

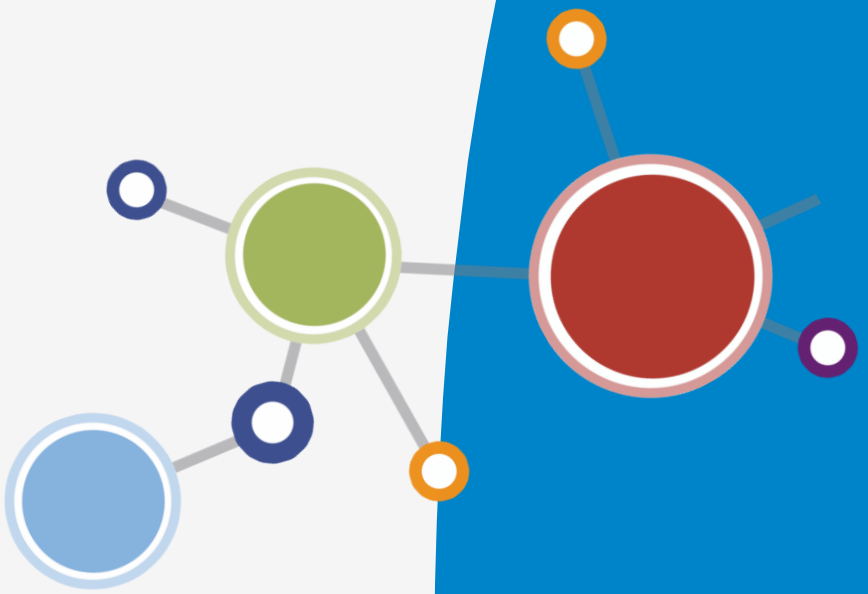
Inteligencia de Datos

Prof. Catalina Artavia Pereira

Temas:

2. 1 Análisis Exploratorios de datos

2.2 Modelos de aprendizaje automático supervisados y no supervisados



Temas relevantes

01 **Análisis Exploratorio de Datos (EDA)**

02 **Aprendizaje Automático Supervisado**

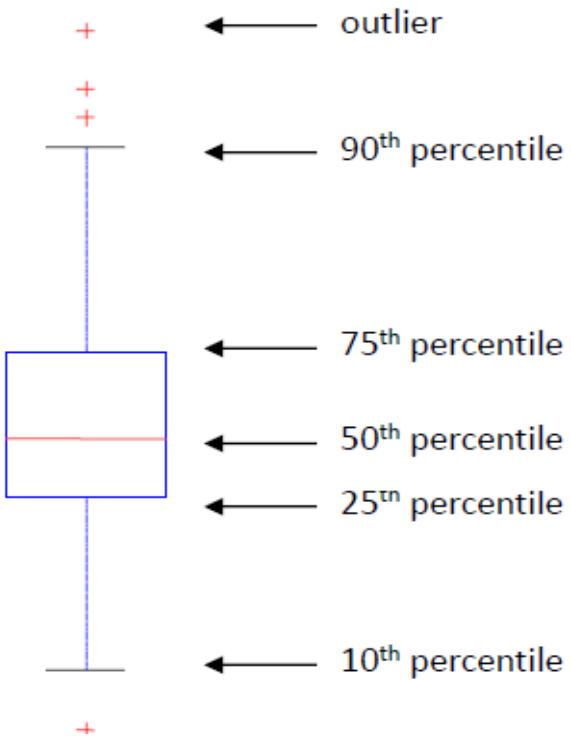
03 **Aprendizaje Automático No Supervisado**

Conceptos básicos

Análisis Exploratorio de Datos:

- Analiza e investiga un conjuntos de datos.
- Resumir sus características principales.
- Permite descubrir patrones, valores atípicos, anomalías, probar una hipótesis o verificar suposiciones.
- Permite comprender de una mejor forma las variables del conjunto de datos y las relaciones entre ellas.
- Es uno de los primeros pasos para realizar modelación de datos, porque nos permite garantizar que los resultados que se vayan a generar sean válidos y se puedan aplicar , así como que respondan a las necesidades identificadas.

Diagrama de caja (Identificación valores atípicos)



El gráfico de boxplot muestra la distribución de una variable numérica, valores por encima del percentil 90 o por debajo del percentil 10 se considera un valor atípico o extremo

Aprendizaje No Supervisado

-Análisis en Componentes Principales.

-Clustering Jerárquico.

Aprendizaje Supervisado

-Árboles de decisión

-Método de Bayes

-Bosques Aleatorios

-Potenciación

-Redes Neuronales

-Maquinas de Soporte Vectorial

-XGBoosnting

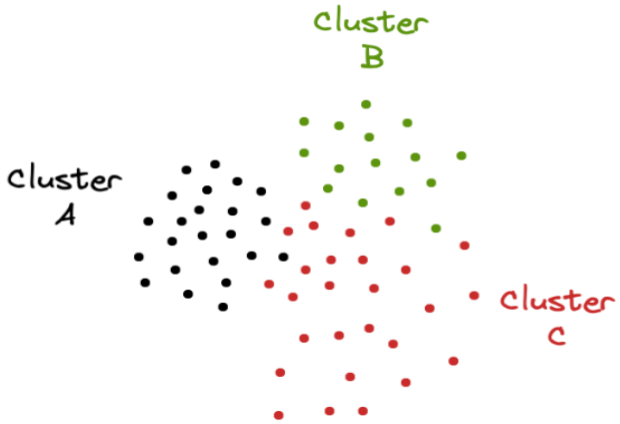
Aprendizaje No Supervisado

Se considera como técnicas de análisis exploratorio

No se tiene conocimiento a priori del análisis.

El objetivo es encontrar patrones "ocultos" de acuerdo agrupaciones de datos similares

Ejemplo:



Aprendizaje No Supervisado: Análisis de Componentes Principales.

-Es una de las técnicas de aprendizaje no supervisado, las cuales suelen aplicarse como parte del análisis exploratorio de los datos (Martínez, 2018).

-Solamente se cuenta con un número de variables de las cuales nos interesa conocer o de las que queremos extraer información.

Funcionalidad:

1. Reducción de dimensionalidad permite reducirlas a un número menor de variables transformadas (componentes principales) que expliquen gran parte de la variabilidad en los datos.
2. **Agrupación de los datos**

Aprendizaje No Supervisado: Coeficiente de correlación.

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y},$$

Donde:

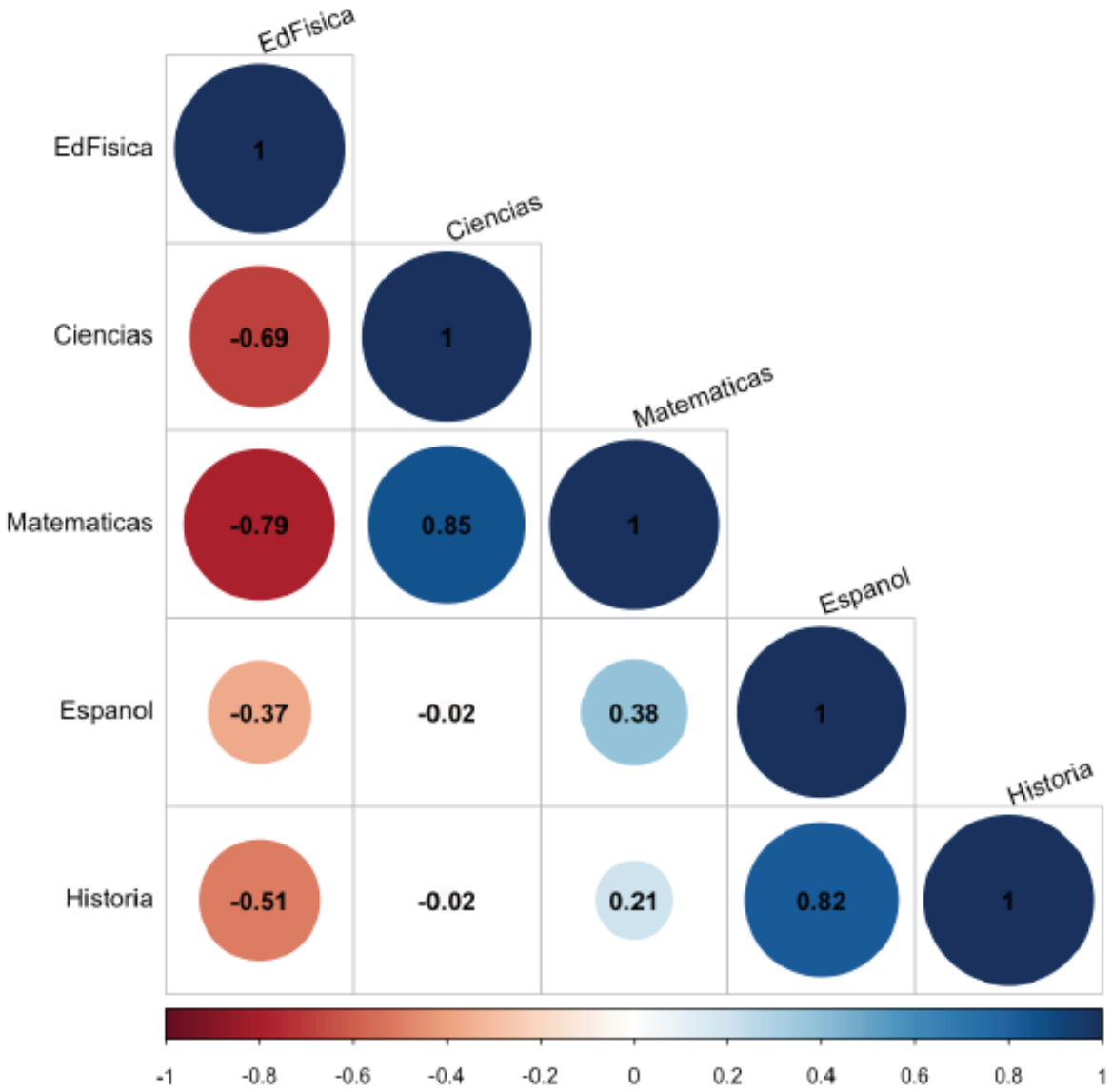
- σ_{XY} es la covarianza de (X, Y)
- σ_X es la desviación típica de la variable X
- σ_Y es la desviación típica de la variable Y

El valor del índice de correlación varía en el intervalo [-1,1], indicando el signo el sentido de la relación:

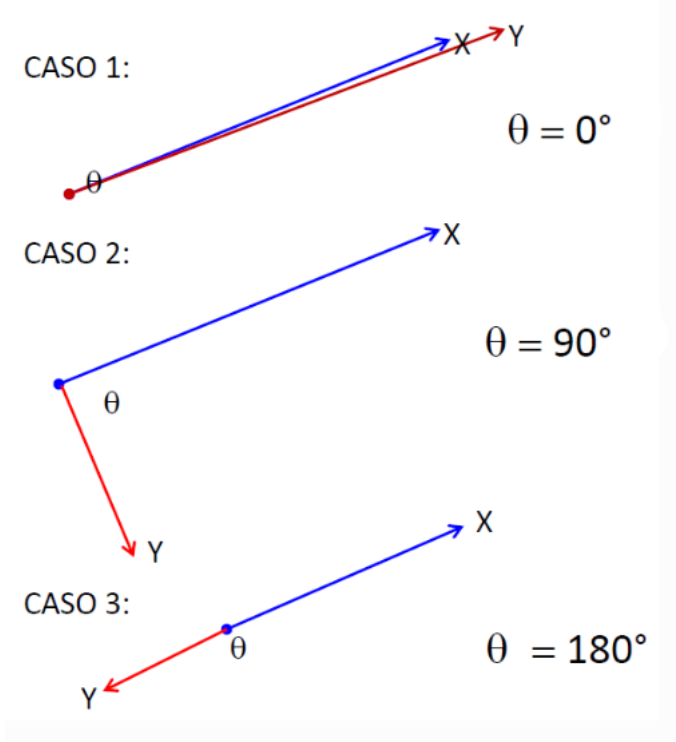
- Si $r = 1$, existe una correlación positiva perfecta. El índice indica una dependencia total entre las dos variables denominada *relación directa*: cuando una de ellas aumenta, la otra también lo hace en proporción constante.
- Si $0 < r < 1$, existe una correlación positiva.
- Si $r = 0$, no existe relación lineal. Pero esto no necesariamente implica que las variables son independientes: pueden existir todavía relaciones no lineales entre las dos variables.
- Si $-1 < r < 0$, existe una correlación negativa.
- Si $r = -1$, existe una correlación negativa perfecta. El índice indica una dependencia total entre las dos variables llamada *relación inversa*: cuando una de ellas aumenta, la otra disminuye en proporción constante.

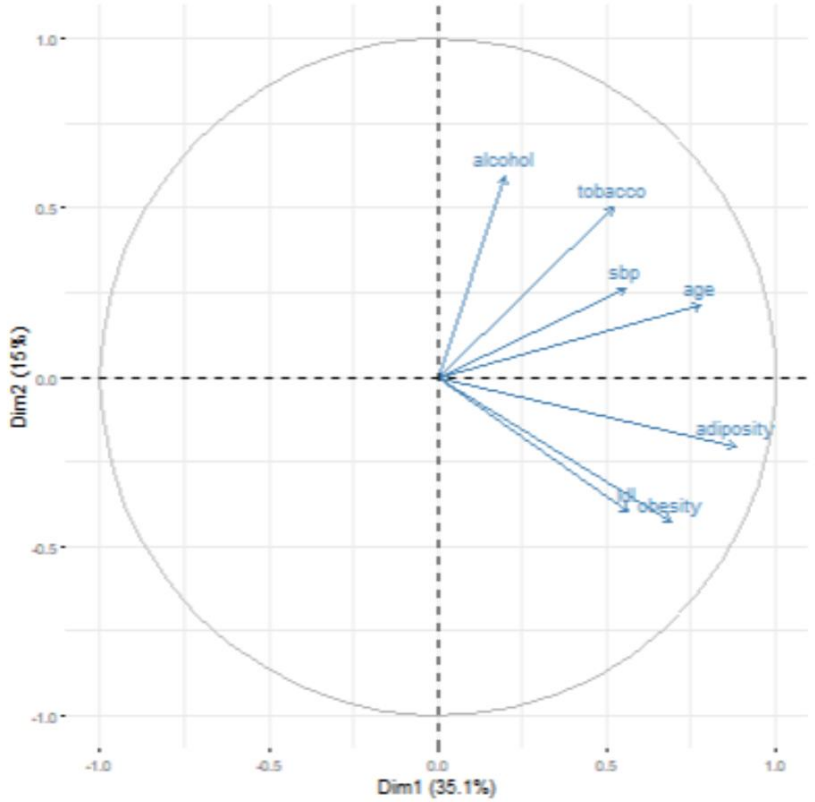
1.Correlaciones altas positivas implican que si una variable crece la otra también crece.
2.Correlaciones altas negativas implican que si una variable crece la otra decrece y a la inversa.
3.Correlaciones cercanas a cero.

EJEMPLO


EdFisica
Ciencias
Matematicas
Espanol
Historia
1
-0.69
1
-0.79
0.85
1
-0.37
-0.02
0.38
1
-0.51
-0.02
0.21
0.82
1
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1

EJEMPLO





- Se presenta una correlación fuerte y positiva entre las variables **alcohol y tabaco**, esto debido a que se presenta un ángulo relativamente pequeño, es decir, a mayor consumo de alcohol de los hombres en una región de alto riesgo de enfermedad cardíaca de la Provincia Occidental del Cabo, Sudáfrica, mayor es el consumo de tabaco en kilogramos. O bien a mayor consumo de tabaco, mayor consumo de alcohol.
- Se presenta una correlación fuerte y positiva entre las variables **tabaco y edad**, e
  a que se presenta un ángulo relativamente pequeño, es decir, a mayor consum
  de los hombres en una región de alto riesgo de enfermedad cardíaca de l
  Occidental del Cabo, Sudáfrica, mayor es la edad de esos individuos. O bien a r
  de los individuos, mayor consumo de tabaco.

Aprendizaje Supervisado

-Árboles de decisión

-Redes Neuronales

Existe conocimiento previo del análisis y por tanto se tienen datos etiquetados de entrenamiento para entrenar los modelos o algoritmos generados

Aprendizaje Supervisado

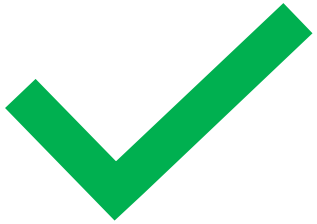
-Árboles de decisión

Es una representación gráfica por medio de mapa de posible respuestas de una serie de decisiones relacionadas.

Permite que la organización realice una comparación de posibles acciones o decisiones a tomar entre sí según sus costos, probabilidades y beneficios entre otros.

Aprendizaje Supervisado

-Redes Neuronales



Aprendizaje Supervisado

-Redes Neuronales

Es un modelo que busca imitar el funcionamiento del cerebro humano, sobre como este procesa la información.

La red aprende analizando los registros individuales y desarrolla una predicción para cada registro.

Veamos ejemplos de cada uno de los modelos