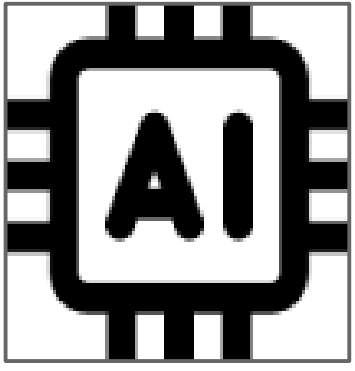
Introducción a la Inteligencia Artificial



Reducción de la dimensionalidad

Reducción de Dimensionalidad

En esta Presentación

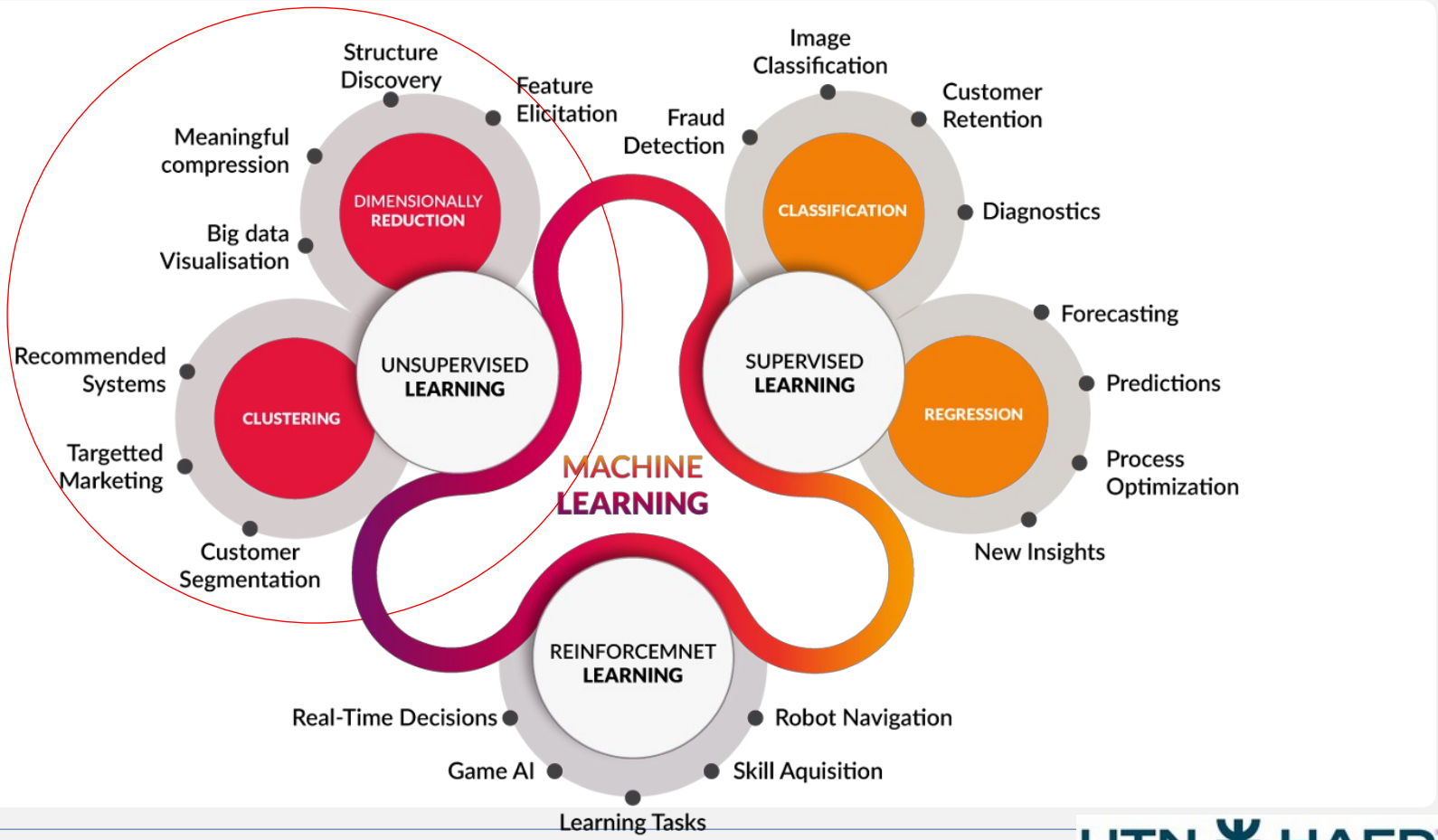
1. El problema de la dimensionalidad

- La maldición de la dimensionalidad
- La dimensionalidad y el tamaño del dataset

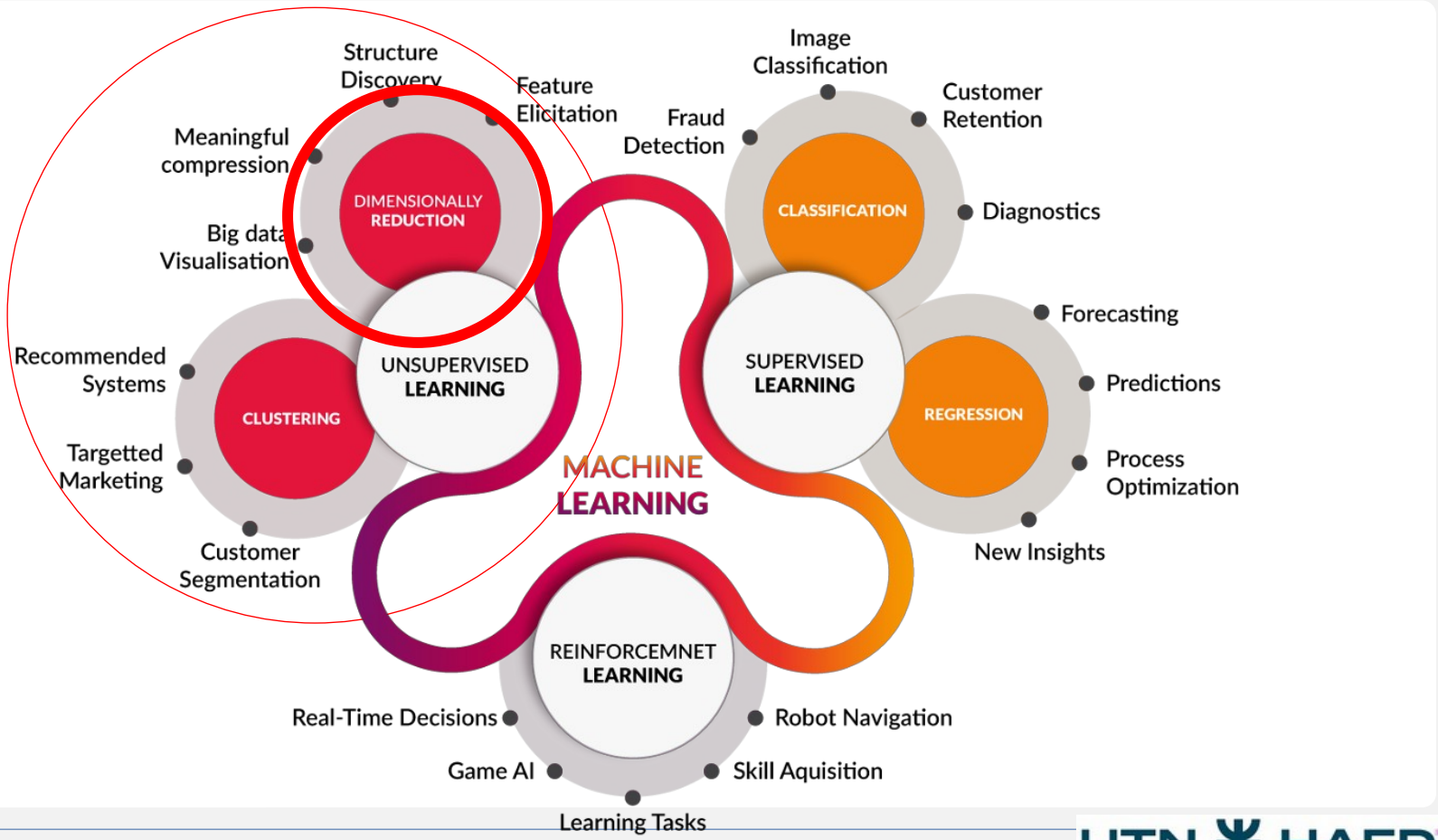
2. Técnicas para la Reducción de Dimensionalidad

- Enfoques
- Técnicas más comunes

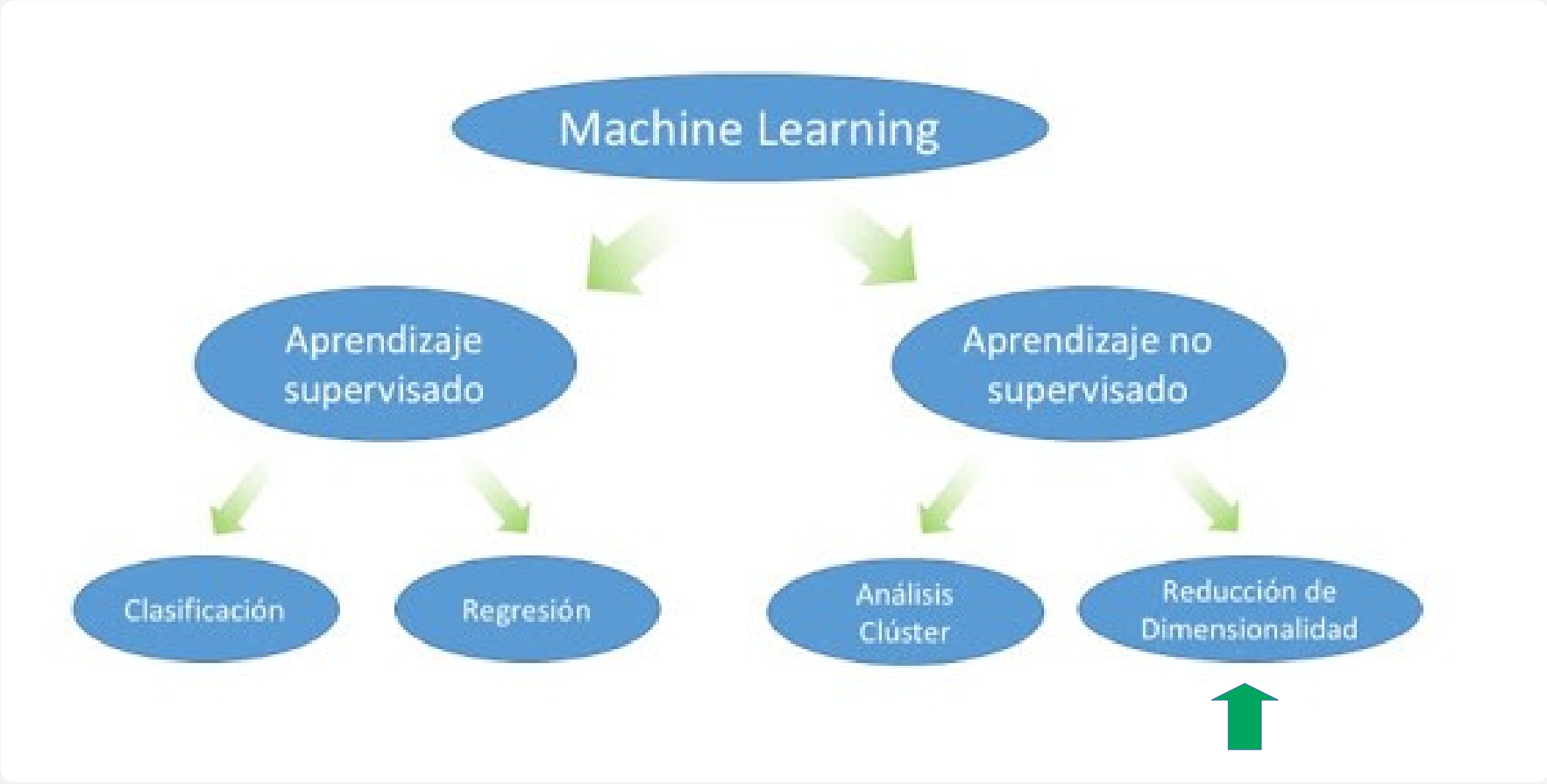
Reducción de Dimensionalidad



Reducción de Dimensionalidad



Reducción de Dimensionalidad



El problema de la dimensionalidad

¿Qué es la dimensionalidad de un dataset?

Es la cantidad de atributos que posee un dataset.

Fuente: https://www.feedingthemachine.ai/machine-learning-reduccion-de-dimensionalidad/

El problema de la dimensionalidad

¿Qué es la dimensionalidad de un dataset?

Es la cantidad de atributos que posee un dataset.

Los atributos de un dataset están asociados con las columnas del mismo.

Fuente: https://www.feedingthemachine.ai/machine-learning-reduccion-de-dimensionalidad/

El problema de la dimensionalidad

La maldición de la dimensionalidad:

Ocurre cuando la cantidad de muestras (o registros) es muy pequeña y la dimensionalidad es muy alta.

El problema de la dimensionalidad

La maldición de la dimensionalidad:

Ocurre cuando la cantidad de muestras (o registros) es muy pequeña y la dimensionalidad es muy alta.

La creación de modelos precisos requiere reusar datos y el uso de todas las características. De ese modo se crean modelos que se ajustan bien al conjunto de datos pero no generalizan adecuadamente.

Fuente: https://aprendeia.com/reduccion-de-la-dimensionalidad-machine-learning/

El problema de la dimensionalidad

La maldición de la dimensionalidad:

Ejemplo:

Si se necesita presentar todas las combinaciones de las n variables, se requiere cálculo combinatorio:

$$Q = \sum_{m=1}^{n} \frac{n!}{m!.(n-m)!}$$ (grupos n elementos agrupados de a m)

El problema de la dimensionalidad

La maldición de la dimensionalidad:

Ejemplo:

Si se necesita presentar todas las combinaciones de las n variables, se requiere cálculo combinatorio:

$$Q = \sum_{m=1}^{n} \frac{n!}{m!.(n-m)!}$$ (grupos n elementos agrupados de a m)

Si n=33 ==> Q $\simeq$ $8{,}57.10^9$

Obtener Q modelos en una PC Normal tardaría: 1,33 años.

El problema de la dimensionalidad

La dimensionalidad y el tamaño del dataset

A un dataset se lo puede entender como una matriz de mxn con:

- m: número de filas (o registros)
- n: número de columnas (o variables)

El problema de la dimensionalidad

La dimensionalidad y el tamaño del dataset

A un dataset se lo puede entender como una matriz de mxn con:

- m: número de filas (o registros)
- n: número de columnas (o variables)

Lo ideal en ML es tener m grande y n chico

El problema de la dimensionalidad

La dimensionalidad y el tamaño del dataset

Problemas con gran cantidad de atributos hacen que la complejidad, los tiempos y/o los recursos computacionales necesarios imposibiliten su resolución.

Fuente: https://www.feedingthemachine.ai/machine-learning-reduccion-de-dimensionalidad/

El problema de la dimensionalidad

La dimensionalidad y el tamaño del dataset

Ejemplo:

10.000.000 de registros, con 10 columnas de 16 bits c/u: memoria=10.000.000*10*16=1,6Gb

El problema de la dimensionalidad

La dimensionalidad y el tamaño del dataset

Ejemplo:

10.000.000 de registros, con 10 columnas de 16 bits c/u: memoria=10.000.000*10*16=1,6Gb

¿Y si el dataset tuviera más atributos? p/ej. El doble...

Introducción a Reducción de Dimensionalidad

Enfoques

- Extracción de características destacadas
- Compresión de atributos

Enfoques

Extracción de Características Destacadas

Al intentarse hallar la solución de un problema, muchas veces deben analizarse los resultados de todas las combinaciones de modelos obtenidos, para quedarse con aquel que ofrece mejores resultados (el resultado tiene el mejor indicador).

Enfoques

Compresión de atributos

En muchos casos es conveniente o requerido reducir o comprimir las dimensiones del dataset.

Técnicas más comunes

- **PCA:** Análisis de Componentes principales
- **Análisis de correlación**
- **La entropía de Shannon:** Teoría de la información

PCA: Componentes Principales

Introducción:

PCA funciona analizando datos que contienen múltiples variables.

PCA Busca correlaciones entre las variables y determina la combinación de valores que captura mejor las diferencias en los resultados.

Estos valores de características combinados se utilizan para crear un espacio de características más compacto denominado componentes principales.

PCA: Componentes Principales

Fundamentos:

Dado un conjunto de m datos (o registros) de n variables predictoras (o columnas), puede modelarse al dataset como una matriz de m x n, con normalmente: m»n.

PCA: Componentes Principales

Fundamentos:

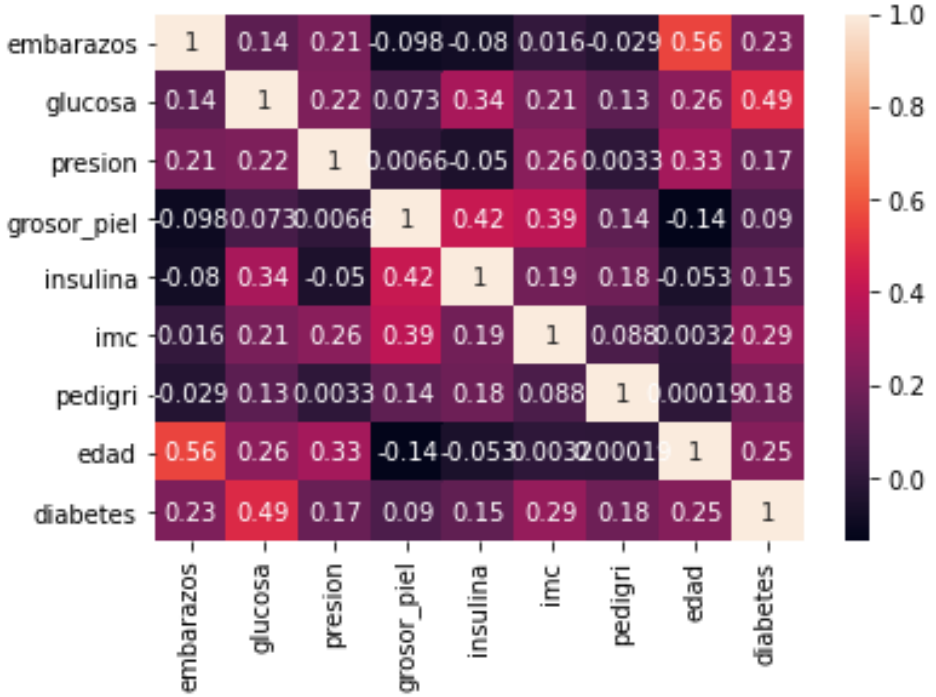
Dado un conjunto de m datos (o registros) de n variables predictoras (o columnas), puede modelarse al dataset como una matriz de m x n, con normalmente: m»n.

A partir de dicha matriz de m x n puede hallarse la correspondiente matriz de correlación entre variables. Dicha matriz es una matriz cuadrada de n x n.

Técnicas más comunes para la Reducción de Dimensionalidad

PCA: Componentes Principales

Fundamentos:



PCA: Componentes Principales

Fundamentos:

Dicha matriz es simétrica, por lo que es diagonalizable.

PCA: Componentes Principales

Fundamentos:

Dicha matriz es simétrica, por lo que es diagonalizable.

Entonces, es posible hallar para cada autovalor, un espacio representativo más compacto de vectores: el espacio de los autovectores asociados a los autovalores.

PCA: Componentes Principales

Fundamentos:

Dicha matriz es simétrica, por lo que es diagonalizable.

Entonces, es posible hallar para cada autovalor, un espacio representativo más compacto de vectores: el espacio de los autovectores asociados a los autovalores.

Donde los autovalores indican el peso de la variable: Así, pueden seleccionarse las variables de mayor peso.

Análisis de correlación

El análisis de correlación entre las variables numéricas y la variable objetivo, consiste en identificar si un predictor incide fuertemente o no sobre el objetivo.

Análisis de correlación

El análisis de correlación entre las variables numéricas y la variable objetivo, consiste en identificar si un predictor incide fuertemente o no sobre el objetivo.

$$\begin{cases} \text{Correlación} \rightarrow |1| : \text{incidencia fuerte} \\ \text{Correlación} \rightarrow |0| : \text{incidencia débil} \end{cases}$$

Análisis de correlación

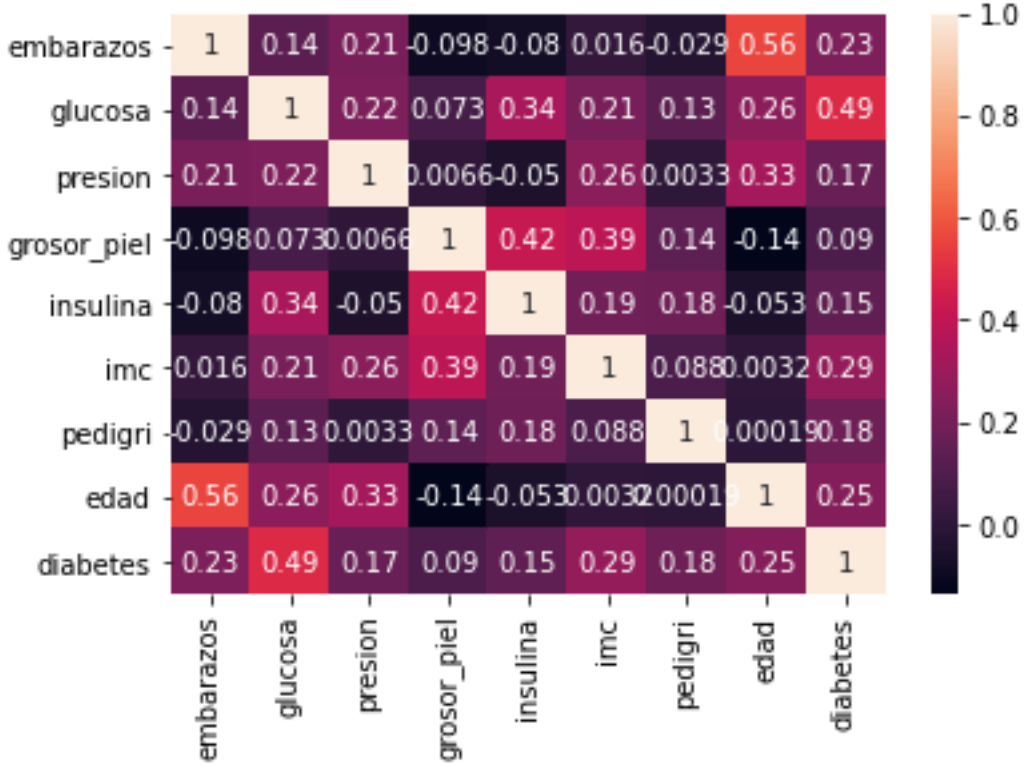
El análisis de correlación entre las variables numéricas y la variable objetivo, consiste en identificar si un predictor incide fuertemente o no sobre el objetivo.

$$\begin{cases} \text{Correlación} \rightarrow |1|: \text{incidencia fuerte} \\ \text{Correlación} \rightarrow |0|: \text{incidencia débil} \end{cases}$$

Así, puede reducirse el análisis armando modelos con aquellas variables fuertemente correlacionadas con la variable objetivo.

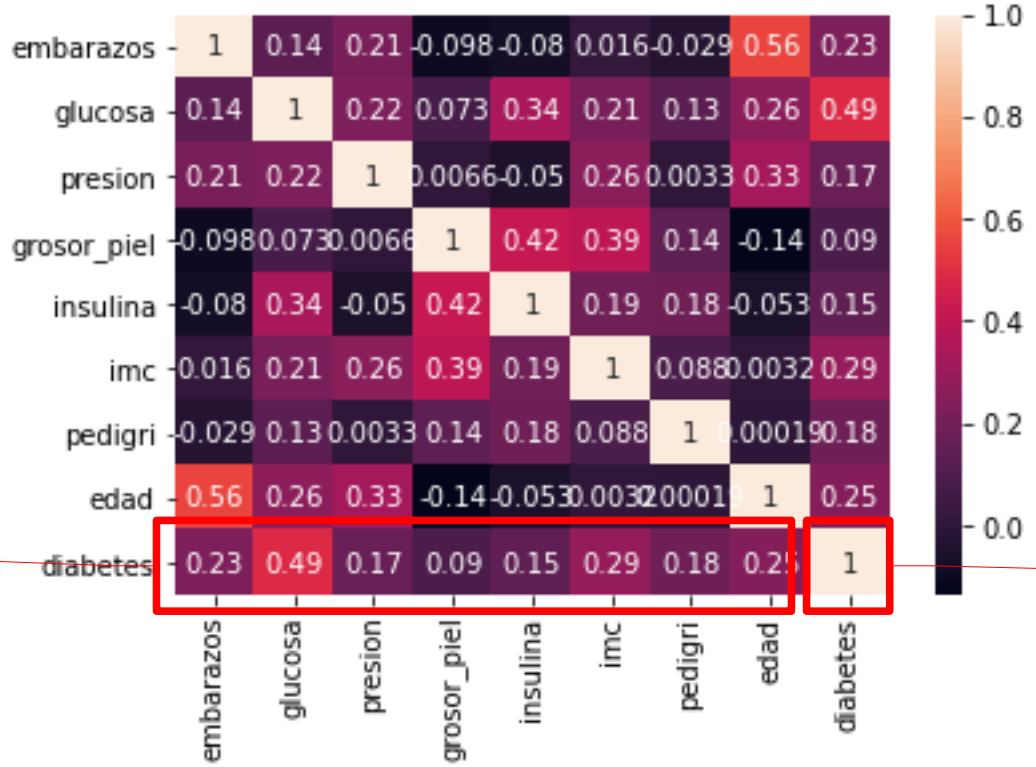
Técnicas más comunes para la Reducción de Dimensionalidad

Análisis de correlación



Técnicas más comunes para la Reducción de Dimensionalidad

Análisis de correlación



Variables predictoras

Variable objetivo

La entropía de Shannon

Entropía: En el contexto de la teoría de la información (Shannon, 1948) es una medida del desorden de un conjunto datos.

La entropía de Shannon

Entropía: En el contexto de la teoría de la información (Shannon, 1948) es una medida del desorden de un conjunto datos.

Ganancia de Información: La ganancia de información a causa de una variable, es cuánto aporta cada variable a la variable Target.

$$GI(H|Y) = H(X) - H(X|Y)$$

La entropía de Shannon

Puede armarse un listado ordenado con la ganancia de información de la variable Target a causa de cada variable predictora y elegir aquellas que realizan mayor aporte.

La entropía de Shannon

Ejemplo: La Ganancia de Información en el dataset diabetes.csv

id_variable	variable	entropia acum
6	pedigri	0.652797110218937 2
5	imc	0.993733168684338 4
1	glucosa	1.299326659155487 3
4	insulina	1.590173461974592 6
7	edad	1.733863275233733 7
3	grosor_piel	1.817605991340953 7
0	embarazos	1.879402156985345 6
2	presion	1.940263389222092 4

Algunos Usos

Reducción de dimensionalidad:

Teléfonos con decenas de sensores usan PCA para predecir nuestras acciones, Filtros de ruido, etc.

Selección de atributos:

Obtención de causales de un fenómeno (p/ej: daño genético).

Reducción de Dimensionalidad

¿Preguntas?