

Proyecto de Investigación

Hugo Missael González Cruz

January 2026

1. Introducción al Análisis de Datos

La ciencia de datos es la intersección de matemáticas, estadística y programación, con el objetivo de descubrir patrones en los datos. Según Garzon y col. (2022):

Data science is about solving problems based on observations of factors (referred to as co-variables, predictors, or just features) that may determine a solution.

(p.1).

El proceso básico se sigue para el análisis de datos es, a grandes rasgos, como sigue:

1. Se elige un fenómeno en el mundo real de donde obtenemos los datos en “crudo”.
2. Se procesan y limpian los datos, es decir, se transforman de tal forma que sean apropiados para el análisis.

A partir de aquí el proceso no es lineal y depende del objetivo para el que se hace el análisis. Una opción es hacer un análisis exploratorio de datos. En tal caso, podríamos hallar valores duplicados, faltantes o mal registrados; de forma que deberíamos regresar a limpiar el conjunto de datos.

Dependiendo de nuestro objetivo —una vez los datos están limpios— podemos usar algún algoritmo. Finalmente, se interpretan los resultados y opcionalmente se comunican mediante reportes o visualizaciones.

Podemos agrupar los problemas concernientes al análisis de datos en tres: clasificación, predicción y *clustering*.

Dada la población concerniente a un problema Ω y una muestra $D \subset \Omega$, resolver el problema, generalmente, consiste en obtener un modelo en términos de D que soluciona el problema en los términos esperados.

Un problema de **clasificación** consiste en definir categorías sobre la población y, dada una entrada, decidir a cuál de ellas pertenece. Esto es, si Π es una partición de Ω , dada una entrada x , encontrar $C \in \Pi$ tal que $x \in C$.

Una solución a un problema de clasificación es un modelo que coloca a x en la categoría correcta.

Un problema de **predicción** consiste en, dada una función $f : \Omega \rightarrow \mathbb{R}$ que le asigna un valor numérico a alguna característica de $x \in \Omega$ y cuyo valor es difícil de medir directamente, hallar el valor de $f(x)$. Una solución es un modelo, basado en otras características de x , capaz de predecir el valor de $f(x)$.

Un problema de *clustering* consiste en construir una partición Π en Ω , dada una métrica $d : \Omega \times \Omega \rightarrow \mathbb{R}$ que mide el grado de similitud entre elementos de la población. Una solución es un modelo que produce una partición tal que elementos en un *cluster* son similares.

Data is an objective recording of one or several event(s) in the real world that is accessible at later times for perusal and analysis by at least one person.

Garzon y col., 2022

Ejemplo 1 Según la definición anterior, son datos:

- Los contenidos de una página web
- Mediciones (precisos o no) de algún fenómeno físico
- Estadísticas recolectadas por sitios web

No son datos:

- Los recuerdos de una persona sobre un evento
- La ocurrencia de un evento (si no se registra)
- Los contenidos de la memoria RAM de una computadora

Una forma simple de organizar datos es en una tabla. En un arreglo de este tipo, las columnas se pueden considerar como vectores de dimensión n (nD) que describen atributos de los datos. Llamamos records a las columnas. A su vez, los renglones en la tabla son un vector de dimensión p (pD) conformado por valores de los atributos. Llamamos a un renglón feature. A los valores específicos en una feature se llaman datums (plural data). De esta manera, una tabla es una vector nD de vectores pD .

Aún más, una tabla puede considerarse como una muestra donde las columnas son variables aleatorias.

Podemos clasificar a las features por el tipo de dato que contienen. Tenemos la siguiente clasificación:

1. Cualitativa

- a) Ordinal
- b) Nominal

2. Cuantitativa

- a) Discreta
- b) Continua

Un dato es cualitativo si toma un pequeño número de valores. Es ordinal si es susceptible a ser ordenado y nominal en otro caso. De forma similar, un dato es Cuantitativo si toma un número grande de valores. Es Discreto si toma una cantidad de valores finito o infinito numerable y es continua si toma valores sobre un conjunto no numerable.

2. Análisis Exploratorio de Datos

Usualmente, se presenta al análisis exploratorio de datos en contraparte al análisis confirmatorio de datos, este último, dedicado a las hipótesis y al modelado. Por esta razón, en el análisis exploratorio no hay un modelo explícito. Según Tukey (1977), un problema básico sobre cualquier conjunto de datos es hacerlo más entendible para la mente. Para dicho fin es deseable:

- Tener una descripción simple
- Ser capaces de describir una capa más profundo (aun si no hallamos nada)

El análisis exploratorio de datos es un método sistemático para hacerse de una visión global del conjunto de datos. Para esto, utiliza gráficas, transforma las variables y presenta estadísticas que resumen el estado de los datos. Se trata de entender los datos, de hacerse una idea de su forma y ganar intuición. Sin embargo, no habría gran valor en explorar los datos si descubrimos lo que ya sabíamos. Un buen análisis exploratorio nos fuerza a ver lo que no esperábamos.

3. Reducción de Dimensión

3.1. PCA

Referencias

- Garzon, M., Yang, C.-C., Venugopal, D., Kumar, N., Jana, K., & Deng, L.-Y. (Eds.). (2022). *Dimensionality reduction in data science*. Springer International Publishing.
- Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.