Proyecto de Investigación

Hugo Missael González Cruz

19 de febrero de 2026

1. Introducción al Análisis de Datos

Recientemente ha habido una explosión en la cantidad de datos que se recopilan sobre diversos aspectos de nuestra vida. Las actividades que realizamos en línea como leer noticias, escuchar música o comprar, están siendo registradas constantemente y generan una cantidad inmensa de datos. Además, actualmente, contamos con gran poder computacional accesible y a bajo costo, en este contexto, ha surgido la necesidad de procesar los datos para convertirlos ya sea en productos o en decisiones accionables. Puede definirse la ciencia de datos como la intersección de matemáticas, estadística y programación, con el objetivo de descubrir patrones en los datos, esto —claro– sin ser únicamente una o la otra. Según Garzon y col. (2022):

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

Tratándose de la materia prima de la disciplina, nuestro punto de partida son los datos. Empezando por distinguir los objetos que consideramos datos, estableceremos nuestro objeto de análisis. Según Garzon y col. (2022):

> 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.

Es decir, un dato es un registro de un fenómeno del mundo real que es susceptible a ser analizado.

Ejemplo 1 *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.*

Podemos clasificar a los datos de acuerdo a los valores que puede tomar. Un dato es *cualitativo* si describe información que no puede ser medida, por ejemplo, grado de estudios, género o religión. Es ordinal si es susceptible a ser ordenado y nominal en otro caso. Por otro lado, un dato es *cuantitativo* si representa información que puede ser medida y representada naturalmente con números, por ejemplo, edad, estatura o salario. Es discreto si toma valores de un conjunto finito o infinito numerable y es continua si toma valores sobre un conjunto no numerable. Podemos resumir la clasificación anterior en la siguiente tabla:

1. Cualitativa
    a) Ordinal
    b) Nominal
2. Cuantitativa
    a) Discreta
    b) Continua

Decimos que los datos recolectados de un fenómeno del mundo real están en "crudo". Los datos en dicho estado son susceptibles para ser analizados, pero tal vez no estén en una forma útil para nuestro análisis. A esta parte del proceso, se le conoce como limpieza de datos y consiste en transformarlos de modo que sean apropiados para su 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 al paso anterior, es decir, limpiar de nuevo el conjunto de datos. Enseguida, dependiendo del problema que deseamos resolver, podemos usar algún modelo o algoritmo que nos ayude a resolver el problema. Finalmente, se interpretan los resultados y opcionalmente se comunican mediante reportes o visualizaciones. A partir de la cuales se toman decisiones, acciones o se desarrollan productos.

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 que describen **atributos** de los datos. A su vez, los renglones en la tabla son un vector de dimensión p conformado por valores especifícos de los atributos. Llamamos a una columna **atributo** (*feature*) y a un renglón, **registro** (*record*). Una tabla de este estilo o un conjunto de ellas se llama conjunto de datos (*dataset*).

De lo anterior, es claro que nuestro interés no son los datos aislados, sino un conjunto de ellos —extraídos de un fenómeno del mundo real— en el que podemos comenzar a buscar patrones. Las razones para analizar datos son muchas y

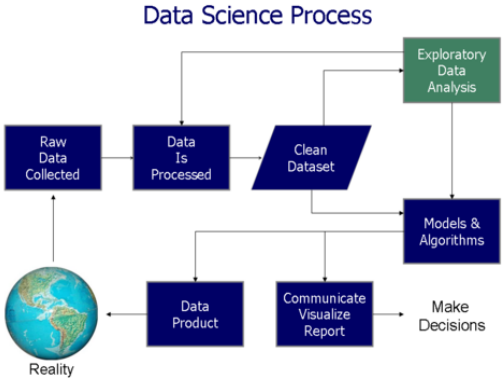

Figura 1: El proceso de análisis de datos. Farcaster at English Wikipedia, CC BY-SA 3.0 https://creativecommons.org/licenses/by-sa/3.0, via Wikimedia Commons.

muy variadas, pero —a grandes rasgos– se pueden clasificar en tres problemas: clasificación, predicción y *clustering*.

Finalmente, presentamos una clasificación típica de los tres principales tipos de problemas que podemos resolver con análisis de datos. Para ello, llamamos Ω a la población concerniente al problema y $D \subset \Omega$ una muestra relativamente pequeña (respecto a la población). Entonces, resolver el problema, consiste en obtener un modelo M en términos de D que soluciona el problema en los términos esperados.

Ejemplo 2 *Considera el problema de clasificar flores Iris en tres categorías: Setosa, Versicolor y Virginica. En este caso, Ω es el conjunto de flores Iris, D es el conjunto de flores Iris de las que hemos registrado datos como la longitud y anchura de sus pétalos. Entonces, una solución sería un algoritmo que, a partir de la longitud y anchura de los pétalos, sea capaz de clasificar una flor Iris en la especie correcta.*

El anterior es un ejemplo de un problema de **clasificación**. Un problema de este tipo 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 $f : \Omega \to \mathbb{R}$ que le asigna un valor numérico a alguna característica de $x \in \Omega$ y cuyo valor es difícil de medir directamente. El problema es hallar el valor de $f(x)$. Una solución es un modelo, basado en otras características de x, capaz de predecir el valor de $f(x)$.

Finalmente, un problema de *clustering* o de conglomerados, consiste en construir una partición Π en Ω, dada una métrica d 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.

Con esto, tenemos una idea general del flujo de trabajo. Sin embargo, el enfoque de este trabajo son los algoritmos de reducción de dimensión que típicamente suelen considerarse concerniente a la fase de exploración de datos o a la fase de visualización.

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. Según Tukey (1977), se trata de observar los datos para ver qué es lo que parecen decir. Se considera que cualquier apariencia que observemos son descripciones parciales y trata de ver más allá. Está interesado en la apariencia y no en la confirmación.

Por lo anterior, el análisis exploratorio de datos no puede ser todo el análisis, es un primer paso. Una manera de encontrar pistas que guíen el análisis confirmatorio. 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.

El conjunto de herramientas para hacer análisis exploratorio de datos incluye:

- Estadísticas que resumen el estado global del conjunto de datos.
- Histogramas, Diagramas de caja, diagramas de tallo y hoja.
- Visualizaciones que ayudan a entender la relación entre las variables.

Un problema que surge cuando se tienen una gran cantidad de datos es hacerla entendible. La visualización de datos es un enfoque del análisis exploratorio que utiliza métodos gráficos para hacer sentido del conjunto de datos. En especial, cuando el conjunto es grande o tienen una cantidad de variables. En dichos casos, las herramientas mencionadad anteriormente podrían no ser suficientes, entonces, se utilizan herramientas más sofisticadas como el *clustering* y la reducción de dimensión.

3. Reducción de Dimensión

En general, el objetivo de la Reducción de Dimensión es encontrar una representación en una dimensión más baja de los datos y, que a su vez, preserve las

propiedades clave para un problema dado. Dependiendo del enfoque usado, las propiedades que deseamos preservar son distintas. Nosotros, nos enfocaremos en el enfoque estadístico y en el enfoque geométrico.

Suponiendo que nuestras observaciones son mediciones de d características, es decir, $x \in \mathbb{R}^d$. Entonces, nuestro conjunto de datos es un conjunto de observaciones de dimensión d. Usualmente, no todas las características son igual de significativas, de modo que podríamos prescindir de ellas y conservar una cantidad significativa de la información original. En otras palabras, el conjunto de datos no cubre todo el espacio $\mathbb{R}^d$, pero yace en una estructura embebida de dimensión menor que d. En especial, la hipótesis de la variedad postula que los puntos de un conjunto de datos, yacen en una variedad (subvariedad o subespacio) de dimensión menor que d. Por lo anterior, a la reducción de dimensión también se le conoce como aprendizaje de variedades (subvariedades).

Podemos clasificar las técnicas de reducción de dimensión en tres: métodos espectrales, métodos probabilísticos y métodos basados en redes neuronales. Los métodos espectrales intentan encontrar la subvariedad, ya sea lineal o no-lineal, de los datos. Usualmente, el problema se reduce a un problema de eigenvectores generalizados. En los métodos probabilísticos, se considera que los datos son variables aleatorias multidimensionales. Es posible hallar una variable latente en la cual, la variable aleatoria, está condicionada. El problema es hallar una representación de menor dimensión de dicha variable latente. Finalmente, los métodos basados en redes neuronales utilizan un *autoencoder* donde los datos se comprimen entre el codificador y el decodificador, en ese punto los datos pasan por un cuello de botella de menor dimensión para luego ser reconstruidos, lo que hace al cuello de botella ideal para reducir la 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.