

1INF03 - Análisis de Datos

Lucio Cornejo

2022-04-08

Table of contents

Metodología KDD	4
¿Qué es un dato?	4
Descripción de la metodología KDD	4
Etapas de la metodología KDD	5
Otras metodologías	9
Step 1: “Entendimiento del negocio”	10
Describir problema o situación a analizar	10
Definir los objetivos	10
Delimitar la población de análisis	11
Identificar recursos necesarios	11
Identificar limitaciones	11
Output del paso 1: Ficha técnica del proyecto de análisis de datos	12
Tipos de datos según su origen	13
Tipo de variable	13
Medidas de resumen	14
Tendencia central	15
Variabilidad	16
Asimetría	17
Curtosis	17
Distribución marginal	19
Distribución condicional	20
Hands-on: step2_entendimiento_de_datos.ipynb	21
Mis soluciones del laboratorio	22
Semana 04/04	23
Descripción	

Apuntes del curso **Análisis de Datos**, dictado en la *Pontificia Universidad Católica del Perú*.

Clases

Semana 03/21

- Será necesario hacer un grupo con otros estudiantes del curso, con quienes se comparta afinidad de investigación, para el proyecto final del curso, el cual se irá desarrollando a lo largo del curso.
- Python y R son complementarios, no es que uno sea *mejor* que el otro.
- En el curso, usaremos Python en su mayoría, pero también se compartirá, después de clase, el código análogo, en R, de lo que trabajemos.
- En las unidades 4 y 5, es donde más podremos contrastar el uso de Python y R. De esa manera, uno tendría más claro qué lenguaje escoger al momento de iniciar algún proyecto particular.
- Fechas de laboratorio
 - 9 abril
 - 23 abril
 - 7 mayo
 - 11 junio
 - 25 junio
- Las dirigidas (perhaps a veces pcs) de IOP se me cruzan con todos los labs, excepto por el primero.

Metodología KDD

¿Qué es un dato?

- El dato es el valor de una característica/variable/atributo (edad, sexo, etc) de la población (población delimitada en espacio, tiempo, etc).
- Procesos paralelos

Variable \Rightarrow Variable aleatoria \Rightarrow Dato

Población \Rightarrow Muestra \Rightarrow Observación

- La **información** parte de la unión de los datos recopilados.
 - Es de utilidad para tomar decisiones.
 - Un solo dato, por su cuenta, no nos da información.
- El **conocimiento** es un conjunto de informaciones aplicadas, que permite prever y planificar.
 - La información asociada a un **contexto** y una **experiencia** se convierte en conocimiento.

Descripción de la metodología KDD

- KDD: Knowledge Discovery in Databases
- Algunas definiciones:

Knowledge Discovery in Databases is the non trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

- *Nivel bajo de datos* se refiere a datos que no nos dice nada, pero que podría servir para generar conocimiento a partir de estos datos.

Etapas de la metodología KDD

- Esta metodología nos da pasos para cómo convertir **datos** en **conocimiento**.
- Estas etapas no son obligatorias ... sirven de **guía**.

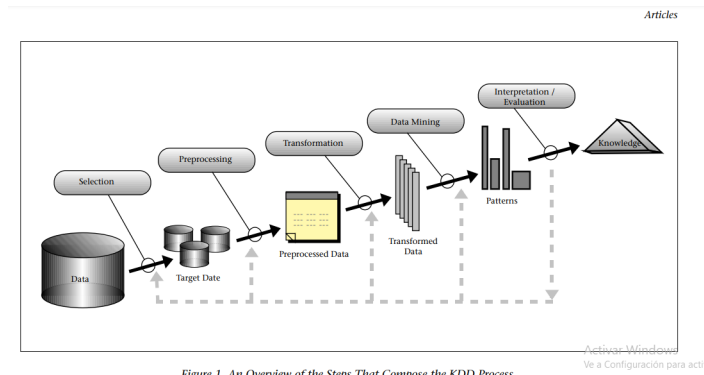


Figure 1: Etapas de la metodología KDD

- En la etapa **selection**, se reduce la cantidad de data, quedándonos con la data que **nos va a servir** para lograr el objetivo de nuestro análisis. Implica filtrar filas y/o columnas/variables de la data (entendida como data frame). Requiere el entendimiento del objetivo del análisis.
- La parte de información surge en la etapa **Patterns** de la metodología KDD. Esa información requiere del bloque *interpretation/evaluation* (ver imagen) para convertirse en **knowledge**.
- El paso de **Transformed data** a patterns es vía “Descriptive methods”.

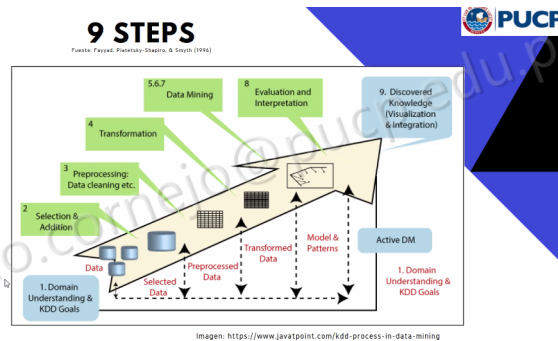


Figure 2: Etapas (más a detalle) de la metodología KDD

- Las flechas verticales indican que, a medida que avanzamos en las etapas, podemos volver al inicio para poder obtener nueva data que haya surgido la necesidad de requerir para el análisis.
- El bloque **Active DM (Data Mining)** se refiere a que el proceso *Data mining* forma parte de TODO el proceso de 9 pasos (es otro enfoque).
Regresar a cualquier paso es válido.

1. Paso 1

- **Es el paso principal.**
- Reunirse con los expertos del tema en que se va a trabajar. Se discuten cosas como
 - ¿Cómo sucede el fenómeno?
 - ¿Qué agentes intervienen con el fenómeno?
 - ¿Qué datos se recolectan (variables disponibles) o se pueden recolectar para el fenómeno?
 - ¿Para qué población se va a construir el proyecto?
- Se habla en lenguaje entendible para todos los expertos, no usando, por ejemplo, palabras particulares de Estadística.
- Se busca **entender el negocio/problema**.
- Se busca identificar la **meta** del proceso KDD desde la perspectiva del **customer**.
- Es más que nada un proceso *cualitativo* que servirá para formalizar el análisis futuro.
- Es recomendable crear una ficha resumen sobre este paso, donde se anota la información recopilada en la reunión (o reuniones) con el customer.
 - Asignar un experto del negocio como encargado del proyecto. Esta persona debe ir validando el avance del proyecto, en cada uno de los 9 pasos.
 - Anotas una meta principal y las secundarias.
 - Una vez completa esta ficha resumen es que podemos pasar al siguiente paso; debe redactarse, quedar como evidencia.

2. Paso 2

- **Creating a target data set.**
- Filtramos la data para obtener un subconjunto, tanto en variables (columnas) y data samples (filas), al cual se le analizará durante pasos siguientes.
- No se trata de la selección de variable que se realiza con código, por ejemplo la que busca explicar un fenómeno con las variables *independientes*.
- Esta selección **no** tiene que ver con la **calidad de datos**. Esa selección ocurrirá más adelante.
- Formalmente, estos filtros se realizan en base a **criterios de inclusión/exclusión**.

3. Paso 3

- **Data cleaning and preprocessing.**
- Se le dice también *remover el ruido*. Donde, el *ruido* hace referencia a los **datos atípicos**.

- Se ve la forma de trabajar los *datos perdidos*.
 - Para construir un modelo, necesitamos lidiar primero con los datos perdidos.
 - Dependiendo del contexto, y requiriendo fundamento, se pueden imputar/reemplazar los datos vacíos por **cero**, **la mediana de esa variable**, etc.
 - Desde el punto de vista de la profesora, máximo se debería imputar el 30% de los valores vacíos de una misma variable (que tiene varios valores vacíos). Pues, sino, se estaría trabajando con una variable *ficticia*, y podría así generar ruido en los resultados obtenidos.
 Pero eso **no es una regla**. La decisión de imputación dependerá del contexto/fenómeno, y debe estar fundamentada **numéricamente**, además de tener sentido respecto al negocio.
 (Por ejemplo, si imputar una variable por cero tiene sentido en cierto contexto particular).
- Debido, en parte, a estas razones, es importante la comunicación constante con un experto del negocio.

4. Paso 4


- **Data reduction and projection.**
- **La transformación de la data debe suceder después de la limpieza de esta.**

Semana 03/28

- Step 4
 - Data reduction and projection.
 - Using methods to reduce the number of variables.
 - * Análisis factorial
 - * Análisis por componentes
 - * Etc
- Step 5
 - Resumir, clasificar, regresión, etc, para las variables.
- Step 6
 - **Choosing the data mining algorithms**
 - Se recomienda establecer mínimo **tres** modelos para poder compararlos tras su funcionamiento.
 - No escoger solo un modelo.
- Step 7
 - **Data mining: Buscando patrones de interés**

- Step 8
 - **Identificar e interpretar los patrones encontrados.**
 - La primera identificación es matemática/numérica/estadística.
- Step 9
 - **Combinar la interpretación numérica del paso 8 junto a la expertise sobre el negocio, con el fin de poder darle utilidad a lo hallado.**

Otras metodologías



Comparativo de Metodologías

Data Mining Process Models		KDD	CRISP-DM	SEMMA
No. of Steps		9	6	5
Name of Steps	Developing and Understanding of the Application		Business Understanding	Sample
	Creating a Target Data Set		Data Understanding	Explore
	Data Cleaning and Pre-processing		Data Preparation	Modify
	Data Transformation			
	Choosing the suitable Data Mining Task		Modeling	Model
	Choosing the suitable Data Mining Algorithm			
	Employing Data Mining Algorithm			
	Interpreting Mined Patterns		Evaluation	Assessment
	Using Discovered Knowledge		Deployment	

Imagen: <https://medium.com/analitics-volhy/knowledge-discovery-data-kdd-8b641509b7f6>
Asívar Wind Configuración

Figure 3: Otras metodologías conocidas

Step 1: “Entendimiento del negocio”

Describir problema o situación a analizar

- El problema debe expresar una **relación entre dos o más variables**.
- Debe estar **formulado claramente**, sin ambigüedad, como pregunta.
- Debe **implicar** la **posibilidad** de realizar una **prueba empírica** o una **recolección de datos**.

Definir los objetivos

- ¿Qué se desea lograr?
- ¿**Cómo ayudará** al negocio?
- Principales áreas interesadas
- Otros objetivos a tener en cuenta.
- ¿Qué características debe tener para ser considerado **factible**?
- ¿Qué esperan recibir?
- ¿Cómo están pensando utilizar el resultado del análisis de datos?
- ¿Con **cuánto tiempo** contamos?
- Objetivos de análisis de datos
 - Traducir los objetivos del negocio en **objetivos para el análisis**.
 - Establecer las métricas o criterios de evaluación de resultados, que serán útiles para el negocio.
 - Diseñar un **Plan de Análisis de Datos** , considerando tiempos, hitos de desarrollo, responsables y fechas para presentación de avances.
 - Validar cada paso con el negocio.

Delimitar la población de análisis

- La delimitación principal es en **espacio y tiempo**.
- Uso de los siguientes criterios:
 - Caso *retrospectivo*:
 - * **Inclusión:** Características que deben reunir las unidades de observación.
 - * **Exclusión:** Características que deben estar ausentes en las unidades de observación.
 - Caso *prospectivo*:
 - * **Eliminación:** Son aquellas características que aparecen una vez que ya han sido seleccionadas las unidades de observación (surgen en la medida que se realiza el análisis)

Identificar recursos necesarios

- Personas
 1. **Experto del negocio**
 2. Líder analítico del proyecto
 3. Equipo especialista de analistas de datos
 4. Equipo de acceso e ingeniería de datos
- Datos
 - Identificar fuentes y dueños de los datos
 - Preguntar por la **calidad de datos** por recibir
 - * ¿Cómo se recolectaron los datos?
 - * ¿Cómo se guardaron los datos?
 - * ¿Cómo se llenó la tabla de datos?
- Herramientas
 - **Softwares disponibles** (libres o con licencia)
 - Entorno para selección y preprocesamiento
 - Entorno para entrenamiento de modelos
 - Entorno para despliegue de modelos

Identificar limitaciones

- Limitaciones del **negocio**

- Posibles restricciones de capacidad operativa
- Poder de acción para utilizar los resultados
- Normativas de la institución o empresa
- Limitaciones respecto a **datos**
 - Si tendremos acceso a todos los datos
 - ¿Se tendrá acceso a toda la población definida?
 - * Acceso para que el modelo pueda ser usado por los usuarios relevantes
- Limitaciones respecto al **tiempo**
 - Restricciones en el tiempo de análisis
 - Tiempo para el despliegue del modelo

Output del paso 1: Ficha técnica del proyecto de análisis de datos

- Archivo en Paideia
 - Las partes **role** y **area** no serán necesario llenarlas.
 - Problemática
 - * Colocar como pregunta, tipo, “¿Se puede blah ...?”
 - Importante plasmar las limitaciones.
 - Acciones de negocio con los resultados
 - * Cómo van a desplegar el modelo creado

Semana 04/04

Tipos de datos según su origen

Tipo de variable

1. Según naturaleza

- Cualitativa (categorías)
- Cuantitativa
 - Discreta
 - Continua

1. Según escala de medida

- Nominal
 - Cualitativa
- Ordinal
 - Cualitativa
- De Intervalo
 - Cuantitativa
 - Solo existe un zero relativo, cuyo valor **no significa ausencia**. Simplemente es una referencia dentro de una escala de medida.
 - Ejemplo: Temperatura
- De razón
 - Cuantitativa
 - Existe un cero absoluto que **significa ausencia de la unidad**.

Medidas de resumen

Medidas de Resumen



Figure 4: Medidas básicas de resumen

Tendencia central

La media trabaja con la **magnitud** de los elementos. Se le puede entender como un *punto de equilibrio* de una distribución.

La mediana trabaja con la **posición** de los elementos (ordenados). Esta divide a la variable en *dos partes iguales*.

Una variable ordinal podría no tener asociada una mediana, en caso que la cantidad de datos sea par, pues requeriría realizar un promedio entre ambas categorías *centrales*.

La **moda** no es muy útil para variables continuas, ya que la frecuencia de un único valor está no definida. Incluso para variables discretas la moda no es tan útil, comparado a la media o mediana.

Debe tenerse cuidado con transformar variables ordinales en numéricas, si es que se desea aprovechar en un cierto modelo, pero **no es recomendable**.

Moda, es el valor **más frecuente** de la variable. *Pico más alto* de la distribución.

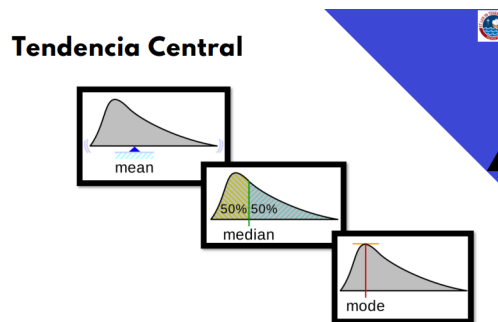


Figure 5: Tendencia central

- Cuantil
 - Dividen al total de observaciones en m partes iguales.
 - $\text{Cuantil}_i = \frac{i*n}{m}$
- Cuartil
 - Dividen al total de observaciones en 4 partes iguales.
 - $Q_i = \frac{i*n}{4}$

- Decil
 - Dividen al total de observaciones en 10 partes iguales.
 - $D_i = \frac{i \cdot n}{10}$

Variabilidad

- **Rango: máximo - mínimo**
- **Rango Intercuartil:**
 - $R = Q_3 - Q_1$
 - R pequeño implica poca dispersión.
 - Es la medida de dispersión recomendada cuando la distribución presenta **datos atípicos**.
- **Varianza muestral:**
 - $S^2(x) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$
 - Es casi como un *promedio de distancias*.
 - El término $n - 1$ es para que la varianza muestral tienda a la varianza poblacional (*estimador insesgado*).
- **Desviación estándar muestral**
 - $s(x) = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$
- **Coefficiente de variación**
 - $CV(x) = \frac{s(x)}{\bar{x}}$
 - **Carece de unidades.** Está *estandarizado*, así que se puede emplear para comparar distribuciones con diferentes escalas de medida.
 - Menor coeficiente de variación implica mayor homogeneidad.
 - Este valor está sesgado cuando existen valores atípicos en la data. En ese caso, se debe hacer un tratamiento especial para los datos atípicos; no basta con usar algo como $\frac{IQR(x)}{median(x)}$
- Cuando la distribución no presenta valores atípicos (o tiene muy pocos (percentage wise)), y es simétrica, se recomiendan como medidas de resumen la media, varianza y desviación estándar.
- Cuando la distribución presenta datos atípicos, y es asimétrica, se recomiendan como medidas de resumen la mediana y el rango intercuartil.

Asimetría

- Trata sobre la *deformación horizontal*.
- **Coefficiente de asimetría (AS)**
 - *Asimetría negativa*: $AS < 0$
 - * Cola jalada hacia la izquierda.
 - *Simétrica*: $AS = 0$
 - *Asimetría positiva*: $AS > 0$
 - * Cola jalada hacia la derecha.

Asimetría

- **Coefficiente de Asimetría (AS) :**
Mide la simetría o asimetría de la distribución de una determinada variable, si presenta una deformación hacia la izquierda o hacia la derecha.

$$AS = \frac{3 * (\bar{Y} - Me)}{s}$$

Otras formas de calcular la asimetría:

$$AS = \frac{\bar{Y} - Mo}{s}$$
$$AS = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}$$

Activar Win
Ve a Configuración

Figure 6: Asimetría

Curtosis

- Trata sobre *deformación vertical*.
- No hay límite establecido, solo referencias, para saber qué tipo de curtosis está presente.
- Mayor coeficiente de curtosis implica mayor deformación vertical.

La curtosis mesocúrtica hace referencia a cuando la distribución se asemeja a la **normal**.

En general, cuando vamos a presentar estadísticas, estas deben estar acompañadas de su gráfico respectivo.

Curtosis

- **Coefficiente de Curtosis (K) :**

Mide el grado de apuntamiento o de deformación vertical de la distribución de una variable.

$$Curtosis = a_4 = \frac{\left(\frac{\sum_{i=1}^m (X_i - \bar{X})^4 n_i}{n} \right)}{s^4}$$

Otra forma de calcular el apuntamiento:

$$\text{Coeficiente Percentil de Curtosis: } K = \frac{(Q_3 - Q_1)^2}{P_{90} - P_{10}}$$

Figure 7: Curtosis

Curtosis

TIPOS DE CURTOSIS

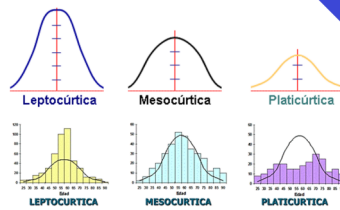


Figure 8: Tipos de curtosis

Distribución marginal

- El estudio de una variable, **no** analizándola por casos/categorías.

Distribución condicional

- Análisis de una variable en función de otras, tratadas como casos/categorías.
- Si una variable X presenta una misma distribución marginal que sus distribuciones condicionales, respecto a una colección Y de categorías, entonces se concluye que estas variables (X e Y) **no están relacionadas**.

Hands-on:

step2_entendimiento_de_datos.ipynb

- Uno de los primeros pasos tras cargar la data es verificar que cada columna tenga el tipo de dato que le corresponde. Por ejemplo, variables cuantitativas como float, int, etc.
- También hay que omitir duplicados en el dataset.
- Al omitir los datos vacíos, no calcular solo cuántos datos vacíos hay, sino también el **porcentaje que representan** por columna/variable.
- Comparando media y mediana de una misma lista de datos, podemos deducir información sobre la presencia de valores atípicos.

Hasta entendimiento de datos entrará en el laboratorio 1 de mañana, el cual será individual. Pero, como alumno libre, no tendré nota calificada.

El grupo del proyecto final solo puede ser de 4 personas, máximo.

Laboratorios

Laboratorio 01

En los laboratorios 3 y 5 se debe presentar avances del proyecto final.

Mis soluciones del laboratorio

- [Python version](#)
- [R version](#)

Extra

Paralelismo Python R

La notación `df` hace referencia a *data frame*.

Primero mostraré el código en Python, después en R, separando cada pareja de códigos *análisis* por un segmento de recta.

Semana 04/04

```
df.shape
```

```
dim(df)
```

```
df.head()
```

```
head(df)
```

```
pd.read_csv("file.csv")
```

```
read.csv("file.csv")
```

```
df.dtypes
```

```
str(df)
```

```
# More precise alternative  
sapply(df,class)
```

```
df.describe()
```

```
skimr::skim(df)
```

```
df.nombre_columna
```

```
df$nombre_columna
```

```
variable.astype(variable_type)
```

```
# Example  
x.astype('float')
```

```
as.variable_type(variable)
```

```
# Example  
as.numeric(x)
```

```
df.drop_duplicates(inplace=True)
```

```
df <- dplyr::distinct(df, .keep_all = TRUE)
```

```
df.drop_duplicates(subset = "column_name", inplace = True)
```

```
df <- df %>% dplyr::distinct(column_name, .keep_all = TRUE)
```

```
df.sort_values(by=['column_name'], inplace=True, ascending = False)
```

```
df <- df %>% arrange(desc(column_name))
```

```
df.dropna(inplace=True)
```

```
df <- na.omit(df)
```

```
df2 = pd.DataFrame(df.isnull().sum())
```

```
# Aplicar una misma función a cada columna de un data frame  
# (plus some other stuff)  
df2 <- data.frame(  
  t(summarise_all(data.frame(is.na(df)), sum))  
)
```

```
df2 = df.groupby(["column_name_1"])  
df2 = df2["column_name_2"].std() / df2["column_name_2"].mean()  
df2.sort_values()
```

```
df2 <- df %>%  
  group_by(column_name_1) %>%  
  summarize(some_name = sd(column_name_2) / mean(column_name_2)) %>%  
  arrange(some_name)
```

```
numpy.mean(df["column_name"])
```

```
mean(df["column_name"])
```

```
df["column_name"].median()
```

```
median(df["column_name"])
```

```
df["column_name"].quantile([0.25, 0.50, 0.75])
```

```
quantile(df["column_name"], c(0.25, 0.50, 0.75))
```

```
df["column_name"].min()
```

```
min(df["column_name"])
```

```
df["column_name"].max()
```

```
max(df["column_name"])
```

```
(df["column_name"].quantile([0.75])).values - (df["column_name"].quantile([0.25])).values
```

```
IQR(df["column_name"])
```

```
df["column_name"].var()
```

```
var(df["column_name"])
```

```
df["column_name"].std()
```

```
sd(df["column_name"])
```

```
df2 = df[["column_name_1", "column_name_2"]].std() / df[["column_name_1", "column_name_2"]].
```

```
df2 <- df %>%  
  summarize(  
    across(  
      c(column_name_1, column_name_2),  
      function(x) { sd(x) / mean(x) }  
    )  
  )
```

```
# Coeficientes de deformación  
df['column_name'].skew()  
df['column_name'].kurt()
```

```
moments::skewness(df['column_name'])  
moments::kurtosis(df['column_name'])
```

```
df.column_name.unique()
```

```
unique(df$column_name)
```

References