Preprocesado de la información

Máster en Data Science y Big Data

Rocío Parrilla
rocio.parrilla@atresmedia.com
Enero 2022

Índice

El proceso de KDD

Preprocesado de datos

El proceso de KDD

Proceso de Extracción del Conocimiento (Knowledge Discovery in Databases) es:

- El proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles e inteligibles en datos.

- Un análisis exploratorio más o menos automático de bases de datos grandes.

.

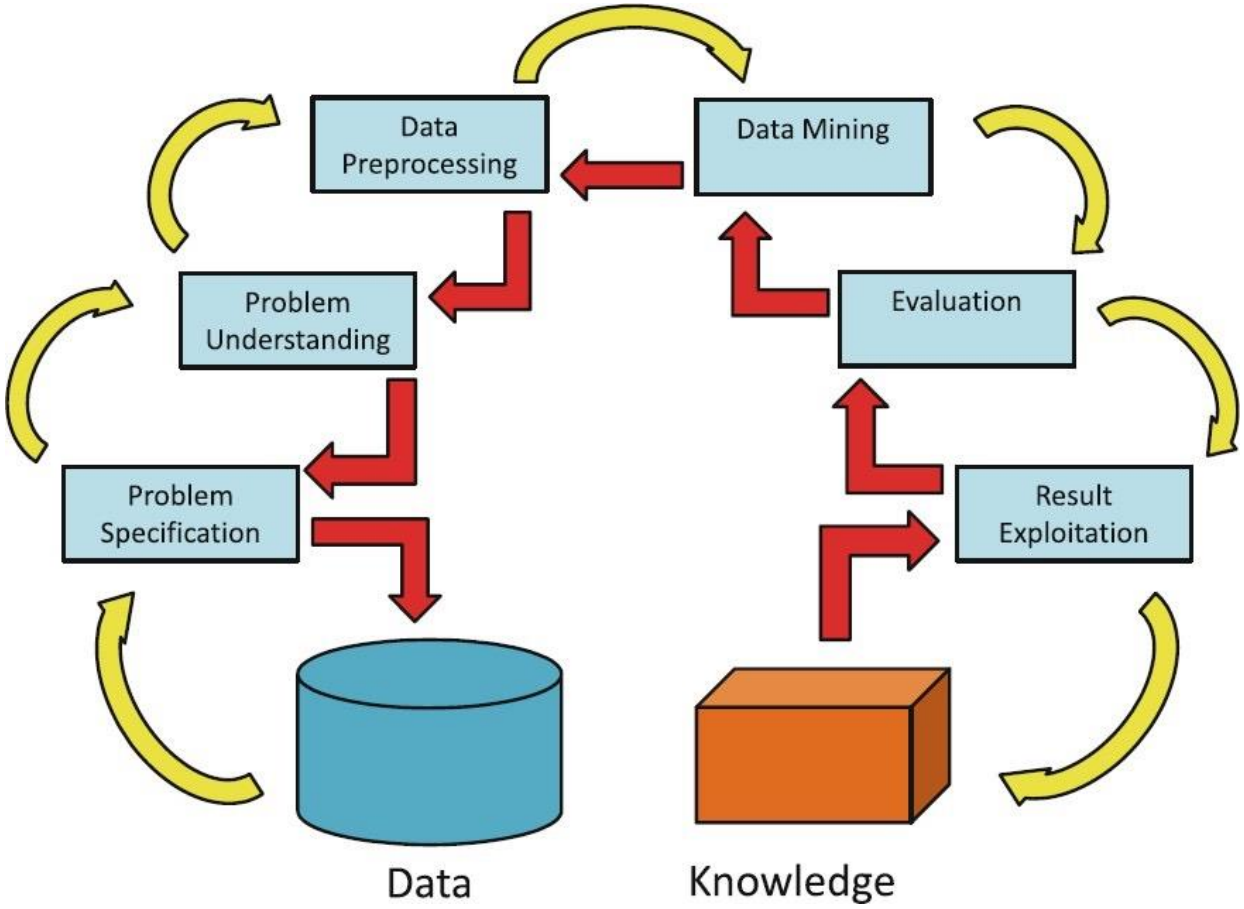
El proceso de KDD

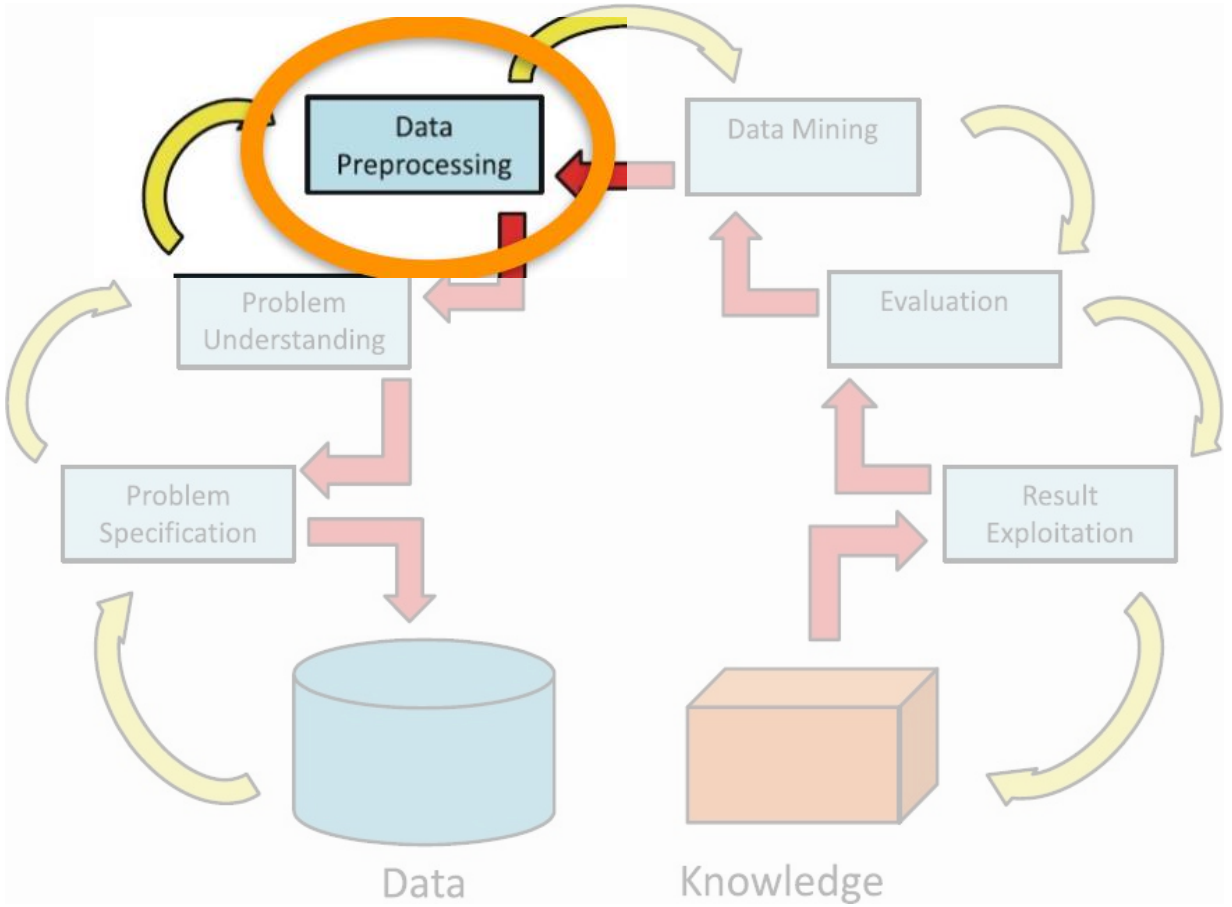
Proceso de Extracción del Conocimiento (Knowledge Discovery in Databases) es:

- El proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles e inteligibles en datos.

- Un análisis exploratorio más o menos automático de bases de datos grandes.

Algunos autores identifican KDD con Data Science o Data Mining (la capacidad de resolver problemas mediante el análisis de datos presentes en bases de datos reales), mientras que muchos otros consideran KDD como un proceso más amplio que engloba a aquellos.




Data Preprocessing
Data Mining
Problem Understanding
Evaluation
Problem Specification
Result Exploitation
Data
Knowledge

Preprocesado de Datos

1 Introducción

Importancia de la preparación de datos

Los datos del mundo real *"son/están sucios"*:

- **Incompletos**: ciertos atributos carecen de valores, los atributos carecen de interés o contienen solo datos agregados, *Missing values…*

Importancia de la preparación de datos

Los datos del mundo real *"*son/están sucios*"*:

- **Incompletos**: ciertos atributos carecen de valores, los atributos carecen de interés o contienen solo datos agregados, *Missing values…*
- **Ruidosos**: contienen errores o "*outliers*".

Importancia de la preparación de datos

Los datos del mundo real *"son/están sucios"*:

- **Incompletos**: ciertos atributos carecen de valores, los atributos carecen de interés o contienen solo datos agregados, *Missing values…*
- **Ruidosos**: contienen errores o "*outliers*".
- **Inconsistentes**: contienen discrepancias en códigos o nombres.

Importancia de la preparación de datos

Los datos del mundo real *"son/están sucios"*:

- **Incompletos**: ciertos atributos carecen de valores, los atributos carecen de interés o contienen solo datos agregados, *Missing values…*
- **Ruidosos**: contienen errores o "*outliers*".
- **Inconsistentes**: contienen discrepancias en códigos o nombres.
- **Tamaño excesivo**: en filas y/o columnas.

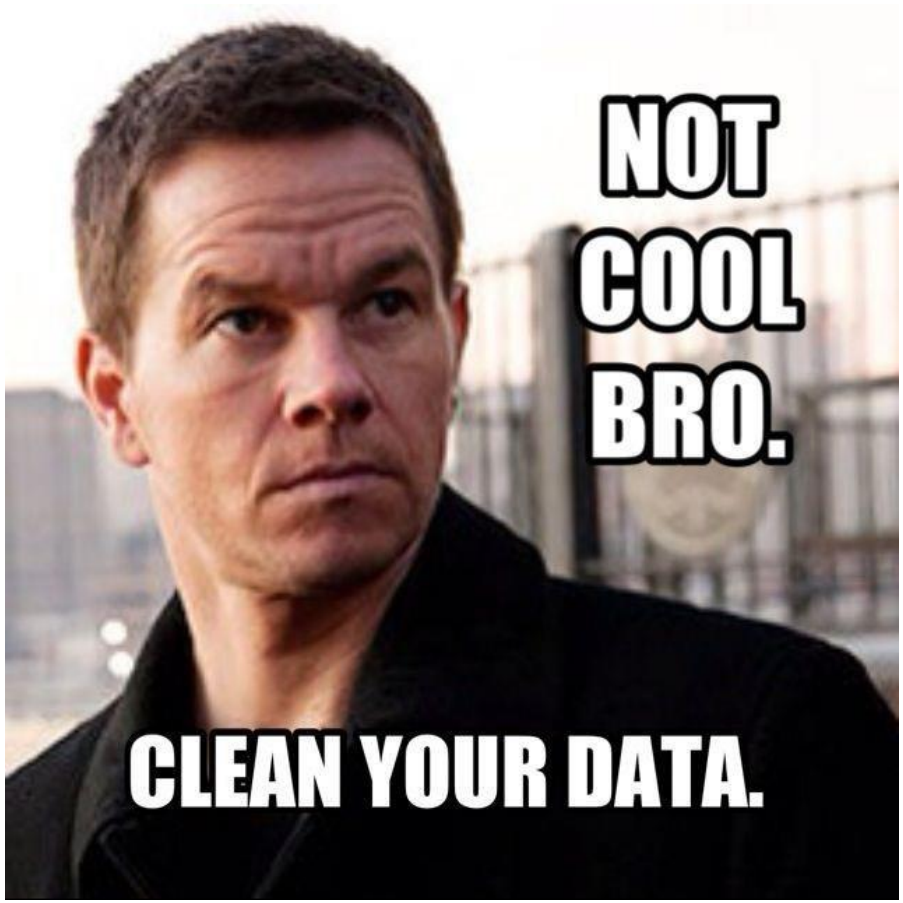
Importancia de la preparación de datos

Los datos del mundo real *"son/están sucios"*:

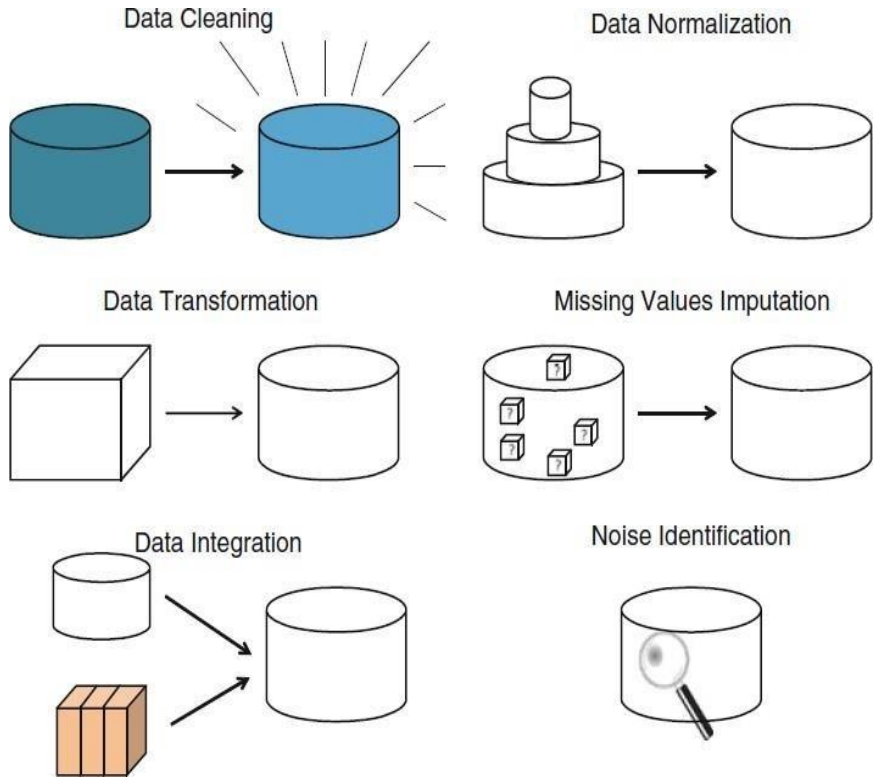
- **Incompletos**: ciertos atributos carecen de valores, los atributos carecen de interés o contienen solo datos agregados, *Missing values…*
- **Ruidosos**: contienen errores o "*outliers*".
- **Inconsistentes**: contienen discrepancias en códigos o nombres.
- **Tamaño excesivo**: en filas y/o columnas.

Para que el proceso de *Data Mining* sea exitoso, los datos de entrada se deben proporcionar en la forma (cantidad, estructura y formato) adecuada.

Sin datos de calidad, no hay calidad en los resultados



El **preprocesado de datos** es el conjunto de técnicas que dejan los datos listos para que sirvan de entrada al proceso de Data Mining.



El preprocesado de datos suele ser una necesidad cuando se trabaja con una aplicación real, con datos obtenidos directamente del problema.

El preprocesado de datos suele ser una necesidad cuando se trabaja con una aplicación real, con datos obtenidos directamente del problema.

Una **ventaja**: **permite aplicar los modelos de forma más rápida y sencilla**, obteniendo modelos y/o resultados de más calidad, en cuanto a precisión e interpretabilidad.

El preprocesado de datos suele ser una necesidad cuando se trabaja con una aplicación real, con datos obtenidos directamente del problema.

Una **ventaja**: **permite aplicar los modelos de forma más rápida y sencilla**, obteniendo modelos y/o resultados de más calidad, en cuanto a precisión e interpretabilidad.

Un **inconveniente**: **no es un área totalmente estructurada** o con una metodología concreta de actuación para todos los problemas.
Cada problema puede requerir una actuación diferente, utilizando diferentes herramientas de preprocesamiento.

2 Limpieza de datos

La **limpieza de datos** es el conjunto de operaciones que

- Corrigen datos erróneos.
- Filtran datos incorrectos.
- Reducen un nivel innecesario de detalle en los datos.
- Detectan y resuelven discrepancias.

La **limpieza de datos** es el conjunto de operaciones que

- Corrigen datos erróneos.
- Filtran datos incorrectos.
- Reducen un nivel innecesario de detalle en los datos.
- Detectan y resuelven discrepancias.

La **limpieza de datos** **no es lo mismo** que la **validación de datos**, que casi siempre rechaza los registros erróneos durante la entrada al sistema. El proceso de limpieza de datos incluye tanto la **validación** como la **corrección** de datos, para obtener datos de calidad.

La **calidad de los datos** se consigue cuando se cumplen los siguientes requisitos:

- **Integridad**: deben cumplir los requisitos de entereza (se consigue al corregir datos con anomalías) y validez.
- **Consistencia**
- **Uniformidad**
- **Densidad**.
- **Unicidad**

La **calidad de los datos** se consigue cuando se cumplen los siguientes requisitos:

- **Integridad**: deben cumplir los requisitos de entereza (se consigue al corregir datos con anomalías) y validez.
- **Consistencia**: corrección de contradicciones
- **Uniformidad**
- **Densidad**.
- **Unicidad**

La **calidad de los datos** se consigue cuando se cumplen los siguientes requisitos:

- **Integridad**: deben cumplir los requisitos de entereza (se consigue al corregir datos con anomalías) y validez.
- **Consistencia**: corrección de contradicciones
- **Uniformidad:** relacionado con irregularidades
- **Densidad**.
- **Unicidad**

La **calidad de los datos** se consigue cuando se cumplen los siguientes requisitos:

- **Integridad**: deben cumplir los requisitos de entereza (se consigue al corregir datos con anomalías) y validez.
- **Consistencia**: corrección de contradicciones
- **Uniformidad:** relacionado con irregularidades
- **Densidad:** conocer el cociente de valores omitidos sobre el número de valores totales.
- **Unicidad**

La **calidad de los datos** se consigue cuando se cumplen los siguientes requisitos:

- **Integridad**: deben cumplir los requisitos de entereza (se consigue al corregir datos con anomalías) y validez.
- **Consistencia**: corrección de contradicciones
- **Uniformidad:** relacionado con irregularidades.
- **Densidad:** conocer el cociente de valores omitidos sobre el número de valores totales.
- **Unicidad:** no tener datos duplicados e inconsistentes.

Problema de la limpieza de datos

Corrección de errores y pérdida de información

En ocasiones, la información que disponemos sobre los datos anómalos es insuficiente, por lo que es difícil determinar las transformaciones necesarias que debemos llevar a cabo, lo que nos puede llevar a perder información.

3 Transformación de datos

Una vez los datos están limpios, suele ser necesario realizar **transformaciones** para que el proceso de *Data Mining* sea más eficiente (o, directamente, aplicable). Entre otras, destacan:

- **Suavizado.**
- **Agregación y resumen.**
- **Discretización.**
- **Generación de variables.**

Suavizado (I)

El **suavizado** es, en cierto sentido, **eliminación de algo del ruido** de una variable.

Suavizado (I)

El **suavizado** es, en cierto sentido, **eliminación de algo del ruido** de una variable.

- El ruido se conoce como los datos distorsionados y sin sentido dentro de un conjunto de datos.
- El suavizado utiliza algoritmos para resaltar las características especiales de los datos.
- Cualquier modificación o tendencia de los datos puede identificarse mediante este método.

Suavizado (II)

En procesamiento de imágenes, vendría a ser el equivalente a la aplicación de un desenfoque y un enfoque sucesivamente.

Agregación y resumen (I)

Las técnicas de **agregación y resumen** no se emplean solamente para hacerse una idea rápida de cómo es la distribución de una variable, sino también cuando los objetos sobre los que se quiere entrenar un modelo *están compuestos* de otros, de los que tenemos las observaciones.

Agregación y resumen (I)

Las técnicas de **agregación y resumen** no se emplean solamente para hacerse una idea rápida de cómo es la distribución de una variable, sino también cuando los objetos sobre los que se quiere entrenar un modelo *están compuestos* de otros, de los que tenemos las observaciones.

Por ejemplo, si queremos hacer un modelo de scoring de crédito a clientes de un banco, no tiene sentido utilizar todas las transacciones o movimientos de los clientes como variables (que será lo que podremos *descargar* del *Data Warehouse* con relativa facilidad).

Agregación y resumen (II)

Se pueden considerar variables como:

- Saldo medio en cuenta.
- Número de transferencias al mes.
- Importe medio de pagos con tarjeta de crédito de los últimos tres meses.

Discretización

La **discretización** es el proceso por el cual se convierten variables continuas en categóricas.

Discretización

La **discretización** es el proceso por el cual se convierten variables continuas en categóricas.

A priori, uno nunca quiere hacer esto, ya que se pierde información, pero **hay algoritmos** (por ejemplo, algoritmos de clasificación) que **necesitan que los datos de entrada** sean **variables categóricas**.

Discretización

La **discretización** es el proceso por el cual se convierten variables continuas en categóricas.

A priori, uno nunca quiere hacer esto, ya que se pierde información, pero **hay algoritmos** (por ejemplo, algoritmos de clasificación) que **necesitan que los datos de entrada** sean **variables categóricas**.

Además, esto hace que los **datos sean más fáciles de estudiar y analizar**, y **mejora la eficiencia** de las tareas que queramos realizar con ellos.

Discretización

La **discretización** es el proceso por el cual se convierten variables continuas en categóricas.

A priori, uno nunca quiere hacer esto, ya que se pierde información, pero **hay algoritmos** (por ejemplo, algoritmos de clasificación) que **necesitan que los datos de entrada** sean **variables categóricas**.

Además, esto hace que los **datos sean más fáciles de estudiar y analizar**, y **mejora la eficiencia** de las tareas que queramos realizar con ellos.

Este método también se denomina mecanismo de **reducción de datos**, ya que transforma un gran conjunto de datos en un conjunto de datos categóricos.

4 | Integración

La **integración de datos** es el proceso que consiste en unir datos provenientes de diversas fuentes en un único *dataset* (o tabla). Esto debe hacerse con cierta cautela para evitar redundancias e inconsistencias.

La **integración de datos** es el proceso que consiste en unir datos provenientes de diversas fuentes en un único *dataset* (o tabla). Esto debe hacerse con cierta cautela para evitar redundancias e inconsistencias.

Es especialmente importante tener **cuidado** con:

- **Atributos redundantes**: El set de datos es más grande de lo que debería y los tiempos de entrenamiento de modelos son más largos. Además, puede llevar a *overfitting*.

- **Registros duplicados e inconsistencias**: Aparte del problema de tamaño, puede inducir a errores.

Un atributo es **redundante** cuando sus valores se pueden deducir de otro atributo o de un conjunto de ellos.

Identificar redundancias es relativamente sencillo para pares de atributos:

- Análisis de correlaciones (o *scatterplots*) para variables numéricas.
- Test χ^2 para variables categóricas: contrasta la hipótesis de que las variables son independientes, frente a la hipótesis alternativa de que una variable se distribuye de modo diferente para diversos valores de la otra.

5 Normalización

El proceso de **normalización** trata de conseguir que todas las variables de un cierto conjunto de datos estén expresadas en una escala similar, para que todas ellas tengan un peso comparable *a priori* a la hora de desarrollar un modelo.

El proceso de **normalización** trata de conseguir que todas las variables de un cierto conjunto de datos estén expresadas en una escala similar, para que todas ellas tengan un peso comparable *a priori* a la hora de desarrollar un modelo.

Algunas de las normalizaciones más utilizadas son:

- La normalización min-max:

$$x = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- La normalización *z*-score:

$$z = \frac{x - \bar{x}}{s}$$

Recomendación: Leer esto.

Nota: En Python, es conocido el submódulo *sklearn.preprocessing* para llevar a cabo escalado y codificación de variables.

6 Imputación de missing values

Un ***missing value*** (valor faltante, perdido o desconocido) es un valor de un atributo que no está proporcionado para alguna observación. La mayoría de datasets reales contienen missing values, que es necesario identificar y, en ocasiones, rellenar.

Normalmente, se representan mediante NULL o NA.

Un ***missing value*** (valor faltante, perdido o desconocido) es un valor de un atributo que no está proporcionado para alguna observación. La mayoría de datasets reales contienen missing values, que es necesario identificar y, en ocasiones, rellenar.

Normalmente, se representan mediante NULL o NA.

La presencia de missing values puede deberse a multitud de motivos, y las formas de *missingness* pueden presentarse de las siguientes maneras:

- MCAR (*missing completely at random*).
- MAR (*missing at random*).
- MNAR (*missing not at random*).

MCAR

Se considera que los *missing values* son *MCAR* cuando las **características de los sujetos con información son las mismas que las de los sujetos sin información.**

Dicho de otra manera, la probabilidad de que un sujeto presente un valor ausente en una variable no depende ni de otras variables del cuestionario ni de los valores de la propia variable con *missing values*. Las observaciones con *missing values* son una muestra aleatoria del conjunto de observaciones.

MCAR

Se considera que los *missing values* son *MCAR* cuando las **características de los sujetos con información son las mismas que las de los sujetos sin información.**

Dicho de otra manera, la probabilidad de que un sujeto presente un valor ausente en una variable no depende ni de otras variables del cuestionario ni de los valores de la propia variable con *missing values*. Las observaciones con *missing values* son una muestra aleatoria del conjunto de observaciones.

Ejemplo:

Imaginemos que hacemos una encuesta a n individuos. Se les preguntan por datos demográficos y también su salario.

Cuando las características estadísticas (media, porcentajes) del resto de las variables son las misma para los sujetos que nos proporcionan su salario y para los que no lo proporcionan, se considera que los valores que faltan son MCAR.

MAR

La perdida de datos es *MAR* cuando **los sujetos con datos incompletos son diferentes significativamente de los que presentan datos completos en alguna variable**, y el **patrón de ausencia de datos puede ser predecible** a partir de variables con datos observados en la base de datos del estudio que no muestran ausencia de datos.

La probabilidad de que se produzca la ausencia de una observación depende de otras variables, pero no de los valores de la variable con el valor ausente.

MAR

La perdida de datos es *MAR* cuando **los sujetos con datos incompletos son diferentes significativamente de los que presentan datos completos en alguna variable**, y el **patrón de ausencia de datos puede ser predecible** a partir de variables con datos observados en la base de datos del estudio que no muestran ausencia de datos.

La probabilidad de que se produzca la ausencia de una observación depende de otras variables, pero no de los valores de la variable con el valor ausente.

Ejemplo:

La pérdida de valores en la variable sueldo es MAR si depende del estado civil, pero dentro de cada categoría, la probabilidad de missing no está relacionada con el sueldo.

MNAR

La perdida de datos es *MNAR* cuando **la probabilidad de los *missing values* sobre una variable Y depende de los valores de dicha variable,** una vez que se han controlado el resto de las variables.

Ejemplo:

Si son los hogares de renta mayor los que con menos probabilidad nos proporcionan el salario, entonces la perdida de datos no es aleatoria.

Dependiendo de las circunstancias, los *missing values* se tratan de distintas formas:

- Eliminando filas: Cuando la mayoría de atributos de una cierta observación son *missing values.*
- Eliminando columnas: Cuando el atributo no presenta valores para la mayoría de observaciones.
- Rellenando (*imputando* su valor): En situaciones intermedias (y cuando sea necesario).

Dependiendo de las circunstancias, los *missing values* se tratan de distintas formas:

- Eliminando filas: Cuando la mayoría de atributos de una cierta observación son *missing values.*
- Eliminando columnas: Cuando el atributo no presenta valores para la mayoría de observaciones.
- Rellenando (*imputando* su valor): En situaciones intermedias (y cuando sea necesario).

Lo más habitual a la hora de imputar *missing values* es rellenar con alguna medida resumen del atributo en cuestión (media, mediana, moda. . .). Esto tiene **dos problemas**:

- Reduce la dispersión del atributo (lo que puede no tener sentido).
- Se está utilizando (mucha) menos información de la "disponible".

Otra posibilidad para **imputar *missing values*** de un cierto atributo **es emplear los valores del resto de atributos** para tratar de inferir el valor del que nos falta.

Dado un set de datos con atributos $X_1, \ldots, X_p$ y otro atributo Y para el que hay *missing values*:

- Se entrena un modelo de predicción de $\hat{Y} \approx Y$ respecto a $X_1, \ldots, X_p$ sobre las observaciones para las que Y es conocido.

- Se imputan los valores de Y mediante los valores predichos de $\hat{Y}$ en el resto de observaciones.

Paquetes más populares para tratamiento de MV con R

Los paquetes más populares en R para el tratamiento de datos son:

- **MICE** (Multivariate Imputation via Chained Equations): es de los paquetes más usados para esta finalidad.

Este paquete realiza múltiples imputaciones y las agrupa de alguna forma, para asegurar que sean estimaciones insesgadas.

MICE asume que los valores perdidos son MAR. Por defecto, se usa regresión lineal para predecir los valores perdidos. Para los *missing values* categóricos se usa regresión logística.

Paquetes más populares para tratamiento de MV con R

Los paquetes más populares en R para el tratamiento de datos son:

- **MICE** (Multivariate Imputation via Chained Equations): es de los paquetes más usados para esta finalidad.

Este paquete realiza múltiples imputaciones y las agrupa de alguna forma, para asegurar que sean estimaciones insesgadas.

MICE asume que los valores perdidos son MAR. Por defecto, se usa regresión lineal para predecir los valores perdidos. Para los *missing values* categóricos se usa regresión logística.

- **Amelia.**

Amelia Earhart, fue la primera mujer aviadora en volar sóla cruzando el Océano Atlántico. La historia dice que desapareció misteriosamente (*missing*) mientras volaba sobre el océano Pacífico en 1937, por lo que el paquete **Amelia** para resolver problemas de valores perdidos, lleva su nombre.

También hace uso de imputación múltiple para disminuir el sesgo.

Otros paquetes para tratamiento de MV con R

- **missForest**
- **Hmisc**
- **mi**

Paquetes más populares para tratamiento de MV con Python

Los paquetes más populares en Python para el tratamiento de datos son:

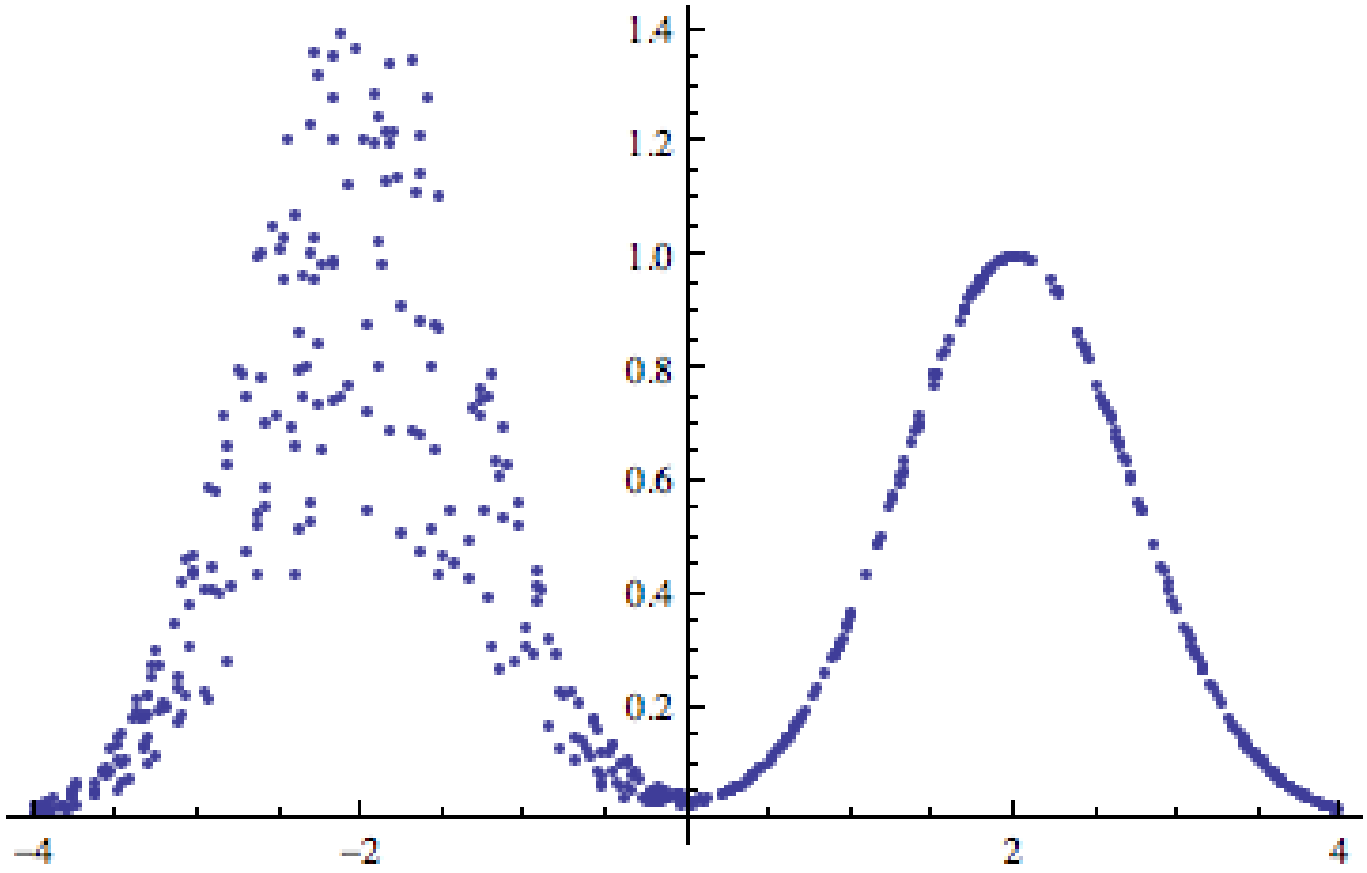
- **sklearn.impute** (submódulo de *sklearn*): tiene métodos para imputación de *missing values* **simple**, es decir, se atiende sólo a la variable en la que están los valores faltantes, o imputación **múltiple,** que tiene en cuenta todo el set de datos.

- **impyte:** este módulo se centra en **visualización** de patrones para reconocer fácilmente *missing values,* y en la imputación de los mismos.

- **missingpy:** realiza imputación de *missing values* aplicando *KNN o Random Forest.*

7 Identificación del ruido

El **ruido** en los datos se ve como un error aleatorio en las variables que se están midiendo.



El **ruido** en los datos se ve como un error aleatorio en las variables que se están midiendo.

Tenemos valores incorrectos debido a:

- Instrumentos de medición erróneos.

El **ruido** en los datos se ve como un error aleatorio en las variables que se están midiendo.

Tenemos valores incorrectos debido a:

- Instrumentos de medición erróneos.
- Problemas en la entrada o transmisión de datos.

El **ruido** en los datos se ve como un error aleatorio en las variables que se están midiendo.

Tenemos valores incorrectos debido a:

- Instrumentos de medición erróneos.
- Problemas en la entrada o transmisión de datos.
- Limitaciones tecnológicas.

El **ruido** en los datos se ve como un error aleatorio en las variables que se están midiendo.

Tenemos valores incorrectos debido a:

- Instrumentos de medición erróneos.
- Problemas en la entrada o transmisión de datos.
- Limitaciones tecnológicas.

¿Cómo podemos suavizar los datos, para minimizar el ruido?

El **ruido** en los datos se ve como un error aleatorio en las variables que se están midiendo.

Tenemos valores incorrectos debido a:

- Instrumentos de medición erróneos.
- Problemas en la entrada o transmisión de datos.
- Limitaciones tecnológicas.

¿Cómo podemos suavizar los datos, para minimizar el ruido?

- Binning
- Clustering
- Regresión

Binning (Método de “cubas”)

Ordenamos primero los datos y agrupamos en “cubas” de igual profundidad (misma cantidad de valores). Luego se suaviza cada porción (de esta forma, lo que se hace es un tratamiento local del ruido, ya que se actúa de manera individual en cada porción).

Binning (Método de “cubas”)

Ejemplo

4,6,11,23,27,34,36,39,42

Lo separamos en tres *bins* de la misma longitud:

B1: 4,6,11
B2: 23,27,34
B3: 36,39,42

Luego se puede suavizar (*smooth*) por media de cubas, mediana de cubas, frontera de cubas, etc.

Por las medias:
B1:7,7,7
B2: 28,28,28
B3: 39,39,39

Por la frontera:
B1:4,4,11
B2: 23,23,34
B3: 36,36,42

Regresión

Podemos suavizar lo datos mediante una regresión lineal.

- Dado un conjunto de tuplas (registros) representado por dos variables, hallamos la línea recta que mejor se ajusta a esos datos.
- Minimizamos un error entre puntos de la recta y reales.
- Si el número de variables es mayor que dos tenemos regresión lineal múltiple.

En realidad buscamos una expresión matemática con la que reproducir los datos y eliminar así el error en la medida de lo posible.

Clustering (I)

Permite detectar "*outliers*". Los valores que quedan fuera de los *clusters,* se pueden considerar anómalos.

Un ***outlier*** es una observación que es distante del resto de observaciones.

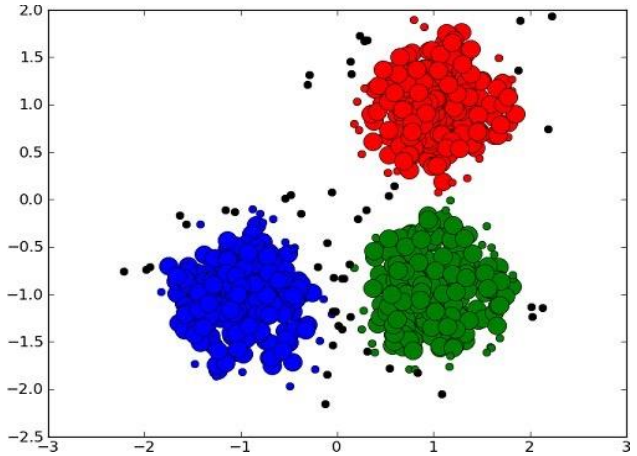
La distancia puede deberse a la propia variabilidad de las variables de estudio (OK) o a errores de medida (en este caso, hay que eliminarlos).

Una buena opción para detectar *outliers,* de forma gráfica, son los boxplots.

Clustering (II)

No hay una definición matemática rígida de lo que es un *outlier*, y determinar si una observación es un *outlier* o no es, en el fondo, un ejercicio subjetivo. Algunas de las técnicas que se emplean son:

- Histogramas y *boxplots*.
- Diagramas Q-Q.
- La regla 3σ.



Las técnicas anteriores sólo pretenden detectar *outliers* univariantes. Detectar *outliers* multivariantes es una tarea mucho más complicada. Algunos de los elementos que se emplean son:

- Distancia de *Mahalanobis*.

  Describe la distancia entre cada punto de datos y el centro de masa. Cuando un punto se encuentra en el centro de masa, la distancia de *Mahalanobis* es cero y cuando un punto de datos se encuentra distante del centro de masa, la distancia es mayor a cero. Por lo tanto, los puntos de datos que se encuentran **lejos del centro de masa** se consideran **valores atípicos.**

- Reducción de dimensiones (por ejemplo, PCA).

Ejercicio práctico

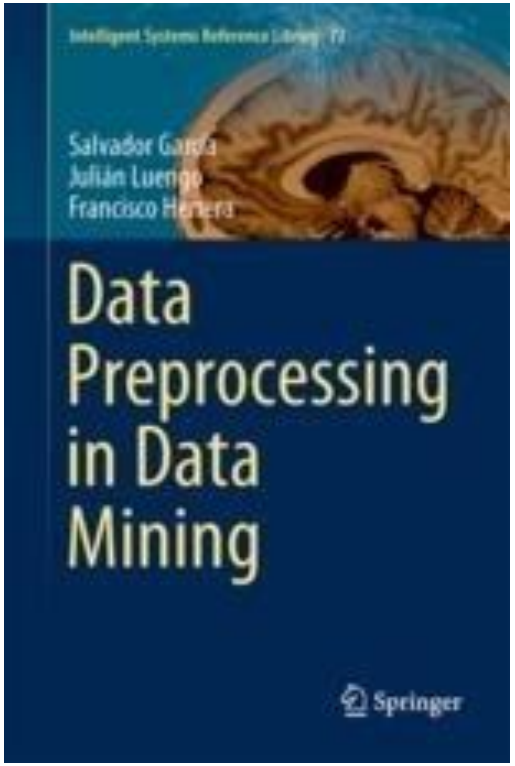
8 Materiales

Referencias:

García, S., Luengo, J., Herrera, F. (2015).

Data Preprocessing in Data Mining.

Springer. ISBN: 978-3-319-10246-7.



Referencias en la web:

- https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide/

- https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825

Afi Escuela de Finanzas