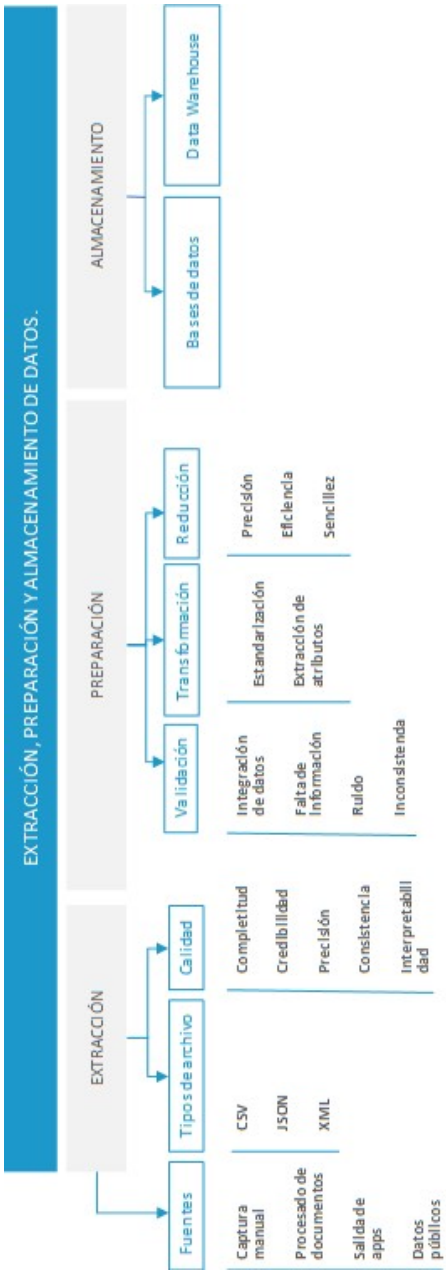
Análisis de Datos Masivos para el Negocio

Tema 2. Extracción, preparación y almacenamiento de datos

Índice

Esquema



2.1. Introducción y objetivos

Antes de comenzar cualquier análisis de datos es necesario **extraerlos** de alguna fuente, procesarlos y almacenarlos en algún lugar para que estos sean accesibles.

En la actualidad, cada vez son más numerosas las fuentes de datos que existen teniendo la posibilidad de simplemente con algoritmos de búsqueda, por ejemplo, extraer información de las páginas webs que sea útil para la empresa.

La primera fase de una estrategia de negocio basada en datos es **entender cuáles son las necesidades y qué tipos de datos se necesitan** extraer o recolectar y en qué forma. Son muchas las alternativas disponibles para extraer datos y dependiendo de cuáles se utilicen se obtendrán unos resultados u otros.

Por otro lado, una vez recolectados y almacenados los datos en función de las necesidades específicas de cada empresa, es importante **prepararlos y realizar las modificaciones que sean necesarias** para que su posterior análisis se realice sobre la información adecuada. Dentro de este apartado de preparación de técnicas existen infinidad de alternativas con un objetivo concreto.

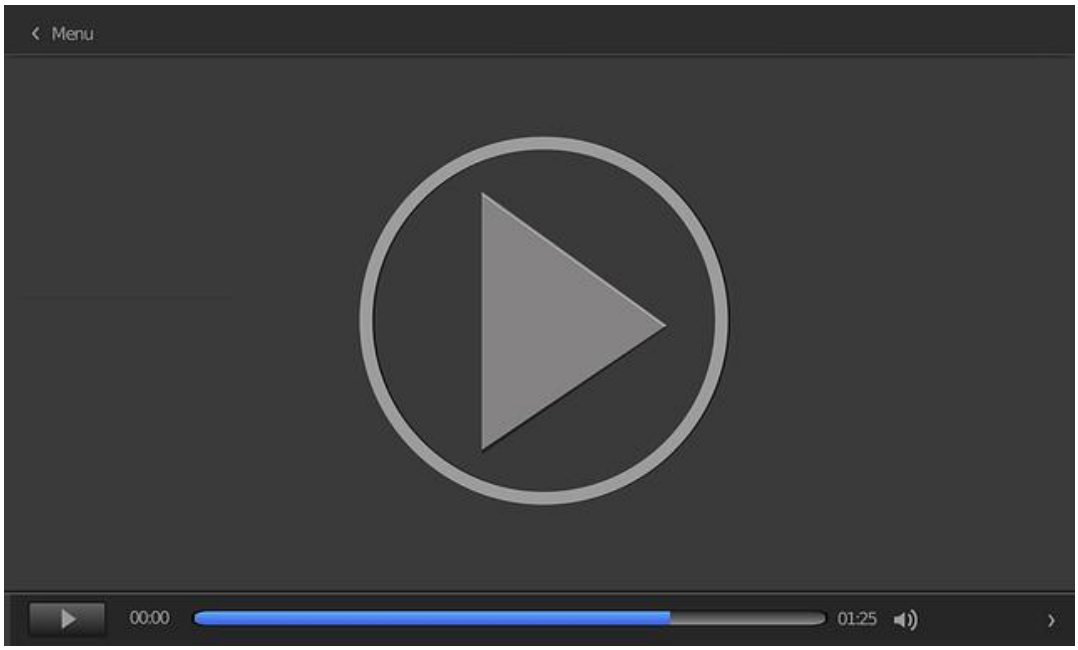
Para finalizar, una vez obtenidos y procesados los datos que serán objeto del análisis para extraer información y generar conocimiento en la empresa, debe decidirse de entre las alternativas disponibles dónde y cómo se van a **almacenar.**

Por tanto, los principales **objetivos** de este tema son:

- Conocer las **fuentes de datos y los tipos de archivos** de almacenamiento.
- Conocer todas las alternativas disponibles para **validar, transformar y reducir** los datos.
- Conocer cuáles son y qué función tienen los recursos disponibles para **almacenar datos.**

Ideas clave

Vídeo *Extracción, preparación y almacenamiento de datos.*



Accede al vídeo:

https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=b4e6821f-de4f-434d-9a9e-b15d00aa2d76

2.2. Extracción de datos

Una parte esencial en el proceso de inteligencia de negocio basada en datos es recolectar los datos necesarios para obtener información de ellos.

Es importante realizar una distinción entre datos e información. El conocimiento basado en datos es fácilmente representable a través de una pirámide como muestra la figura 1.

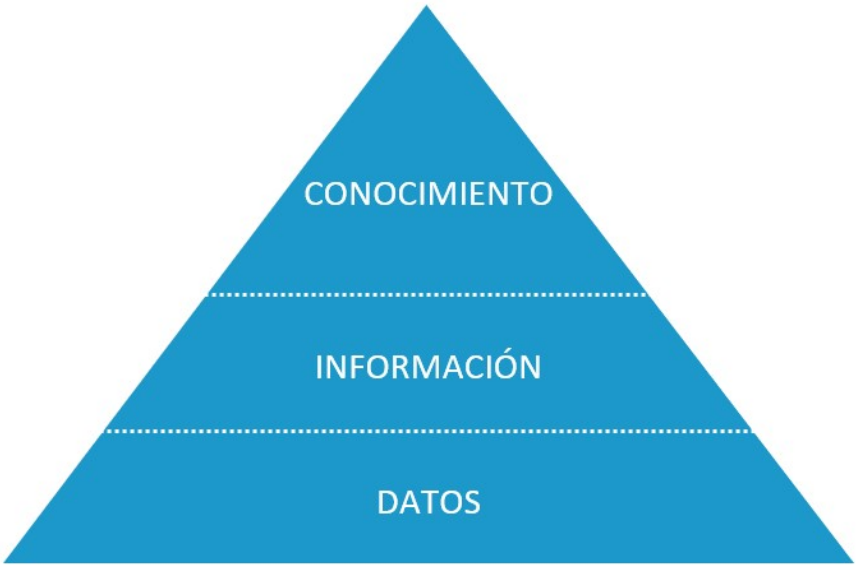


Figura 1. Pirámide del conocimiento.

En esta figura se observa cómo a través de la información obtenida de los datos se consigue el conocimiento en la empresa. Distinguir estos tres conceptos básicos es clave para entender el proceso de análisis de datos.

El hecho de que un dato sea erróneo o no válido no significa que no se pueda extraer información del conjunto de datos. Es clave conocer qué tipo de información deseamos obtener antes de comenzar con la extracción de los datos ya que en función de nuestro objetivo debemos focalizar la búsqueda de los datos de forma

distinta.

Cuando generamos conocimiento erróneo a través de la información extraída de los datos es importante saber si se debe a un error en la fuente de datos, a un problema al realizar el análisis o a una confusión por parte del usuario final de los mismos.

Fuentes de información

Cuando comenzamos un proyecto de conocimiento basado en datos es necesario conocer cuáles son todas las fuentes disponibles de las cuales podemos extraer información útil para generar conocimiento en la empresa.

Las cinco fuentes más utilizadas hoy en día en la empresa son:

Captura manual

Es el método más tradicional y uno de los más frecuentes en la investigación. Dentro de este método se encuentra el uso de encuestas y las observaciones que son útiles para recopilar información al detalle de lo que necesitamos conocer. Aunque no sea necesario el uso de tecnologías de la información para la extracción de datos mediante este método, normalmente la información requiere ser digitalizada para su análisis y almacenamiento.

Procesado de documentos

El objetivo principal de este método es la extracción de datos de documentos cuyo fin no es aportar datos. Ejemplos de este método son el web scraping (extracción de datos de las páginas HTML de los sitios webs) o el análisis de los logs (ficheros que contienen eventos secuenciales que ocurren en un sistema).

Salida de aplicaciones

Dentro de este método de extracción de datos se encuentran todos los procedimientos que implican el acceso a almacenes de datos tradicionales como

bases de datos relacionales o ficheros de valores separados por comas (.csv).

Acceso a datos públicos

Por último, muchas entidades públicas publican sus datos y crean aplicaciones para que sea más sencilla su descarga y manejo. Estos datos son normalmente fácilmente accesibles y descargables y en muchas ocasiones pueden contener información muy útil acerca de características de la población en diferentes territorios.

Por otro lado, existen también empresas privadas como Facebook, Twitter o Instagram que permiten la descarga de datos que contienen la información de un determinado usuario.

Tipos de archivos

Cuando extraemos información esta puede venir almacenada en distintos tipos de ficheros. Las hojas de cálculos Excel son la forma más común de almacenamiento de datos entre usuarios básicos, pero cuando la cantidad de información que se almacena es grande y contienen muchos datos la forma más común de almacenar e intercambiar información entre sistemas son los **ficheros planos.**

Estos ficheros tienen la ventaja que permiten ver y editar el contenido de un fichero tan solo con una herramienta de edición de texto. Entre los formatos más comunes se encuentran los CSV, JSON y XML.

CSV

Los ficheros CSV (Comma Separated Values o valores separados por comas en español) son uno de los formatos más comunes para almacenar datos y como su nombre indica los datos están separados por comas en un texto plano. Un ejemplo de formato CSV se muestra en la figura 2.

```
Nombre, Edad, Cargo
Alejandro, 35, Director
Antonio, 35, "Gestor de proyectos"
Juan, 34, "Analista"
Pepe, 32, "Administrador de bases de datos"
```

Figura 2. Formato CSV

JSON

Los ficheros JSON (JavaScript Object Notation o notación de objetos en JavaScript) están basados en el lenguaje de programación JavaScript y se basa en la creación de objetos para identificar los datos. Un ejemplo de este tipo de ficheros se muestra en la figura 3.

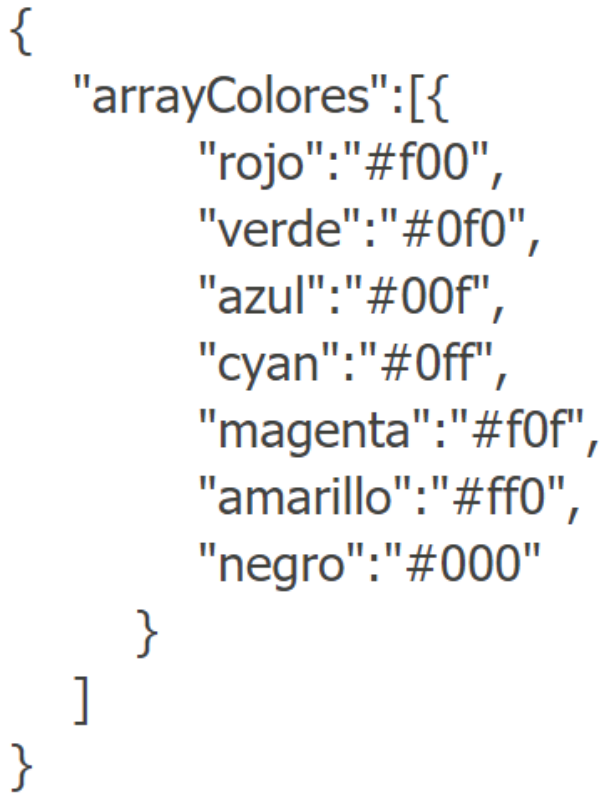


Figura 3. Formato JSON

XML

Por último, el formato XML (eXtended Markup Languaje o lenguaje de marcas extensible por sus siglas en español) es el que más detalles incorpora a la hora de definir los datos y utiliza etiquetas. Este formato permite que el documento contenga información heterogénea al igual que el formato JSON. Un ejemplo de formato XML se muestra en la figura 4.

```
<contact-info>

   <contact1>
      <name>Tanmay Patil</name>
      <company>TutorialsPoint</company>
      <phone>(011) 123-4567</phone>
   </contact1>

   <contact2>
      <name>Manisha Patil</name>
      <company>TutorialsPoint</company>
      <phone>(011) 789-4567</phone>
   </contact2>

</contact-info>
```

Figura 4. Formato XML.

Calidad de los datos

Al extraer datos de cualquier fuente debemos analizar la calidad de los mismos para ver con qué información vamos a trabajar y cuál es su información real.

Siguiendo a Jarke *et al.* (1998) pueden distinguirse cinco propiedades que deben presentar un conjunto de datos para que sean de calidad.

- La **completitud o cobertura** describe el porcentaje de datos de los que se dispone con respecto a la población total de dichos datos. Esto hace referencia a la cantidad de datos que disponemos del total real.

Ideas clave

- La **credibilidad** indica si nuestros datos provienen de una fuente fiable. Esto puede ser contrastable de muchas formas, por ejemplo, comprobando si existen valores lógicos.
- La **precisión** indica la cantidad de datos correctos que disponemos. Puede representarse en porcentaje sobre el total.
- La **consistencia** representa el nivel de coherencia de los datos. Por ejemplo, que una determinada observación tenga determinada la localización en Sevilla (Asturias) muestra la inconsistencia de los datos.
- La **interpretabilidad** define el grado en el que los datos pueden ser entendidos por el usuario final. Entre los factores que pueden afectar está la información adicional explicativa como los metadatos.

2.3. Validación de datos

Una vez extraídos los datos debemos comprobar la calidad de los mismos, ya que esta puede no ser la esperada debido a la **integridad, falta de información, el ruido o la inconsistencia de los datos.**

Integración de datos

En muchas ocasiones los datos no provienen de una sola fuente, sino que deben utilizarse diferentes fuentes para poder completar el set de datos necesario para nuestro análisis.

Si el **proceso de integración** de los datos provenientes de las diferentes fuentes no se realiza correctamente pueden aparecer redundancias o inconsistencias que reducirán la precisión y velocidad de los análisis y crearán problemas a la hora de utilizarlos.

En general debemos centrarnos en dos problemas que pueden surgir:

- Encontrar atributos redundantes

La **redundancia** es un problema típico que debe ser eliminada en la mayor media posible. Normalmente causa un incremento en el tamaño de los datos lo que provoca que el posterior tratado de estos sea más complejo.

En general, **un atributo es redundante cuando** puede ser derivado de otro o de un conjunto de ellos. Por ejemplo, si tenemos las ventas *per cápita* de una determinada localización, incluir la población en el análisis o las ventas totales puede ser redundante ya que esa información ya se encuentra incorporada.

Para detectar la redundancia normalmente se utiliza un **análisis de correlación** que indica la fuerza de la relación y proporcionalidad entre dos variables.

- Detectar datos duplicados e inconsistencia

Tener datos duplicados no solo es un **gasto excesivo de espacio y de tiempo de computación** en los análisis, también puede ser una **fuente de inconsistencia.**

Debido a los errores en el proceso de captura, almacenamiento y definición de variables puede producir datos que contengan la misma información y sean tratados como diferentes.

Estos datos normalmente son muy difíciles de identificar, aunque existen múltiples técnicas que pueden ser utilizadas para identificarlos tales como aproximaciones probabilísticas, técnicas basadas en las distancias o algoritmos de *clustering*.

Falta de información

Los datos que extraemos pueden **no estar completos** y uno o varios atributos pueden aparecer como vacíos. Esto se puede deber a que algunos datos no se registraron en su momento o no estaban disponibles cuando los datos se extrajeron. Otra posible explicación puede ser que los datos se hayan eliminado en el proceso de recolección por considerarse incorrectos o que un error a la hora de conexión con la base de datos no haya introducido esa información en los datos.

Para corregir la falta de datos se pueden utilizar las siguientes técnicas:

- Eliminación. Una de las técnicas y quizás la más sencilla es eliminar todos los registros para los cuales uno o más valores faltan. En concreto, algunas técnicas de análisis necesitan de todos los datos para que puedan ser utilizadas y en caso de que faltase algún valor estas no podrían ser aplicadas lo que podría dificultar el análisis y utilización de estos.

Esta política basada en la eliminación sistemática de registros incompletos puede ser ineficaz cuando existen muchos valores incompletos y puede existir una pérdida sustancial de información que dificulte el análisis.

- Inspección. Alternativamente, también se puede inspeccionar cada registro incompleto para intentar obtener valores que puedan sustituirlos. Esta técnica sufre un problema de alto grado de arbitrariedad y subjetividad a la hora de intentar descubrir el posible valor perdido y puede costar demasiado tiempo y esfuerzo en set de datos de gran tamaño. Por otra parte, si la cantidad de datos perdidos no es excesiva esta técnica es una de las más eficaces para corregir la falta de información.

- Identificación. La tercera posibilidad es identificar aquellos valores perdidos reemplazándolos por un valor que los codifique. Por ejemplo, en un conjunto de valores que se asumen necesariamente mayores que cero, colocar un -1 a todos aquellos valores incompletos los identifica correctamente. Utilizando el mismo procedimiento se puede sustituir el lugar del valor perdido por un valor categórico.

- Sustitución. Existen muchos criterios para reemplazar automáticamente un valor perdido en un conjunto de datos. Por ejemplo, este valor perdido puede ser reemplazado por la media del atributo en cuestión, calculada a partir de los datos existentes. Esta técnica puede ser solo aplicada a valores numéricos y puede ser errónea en caso de distribuciones asimétricas.

Otra técnica es realizar un análisis supervisado donde el valor perdido se reemplaza con la media de los atributos de las entidades similares a el valor perdido.

Finalmente, se pueden aplicar técnicas de regresión o métodos bayesianos para reemplazarlos, sin embargo, estos métodos pueden ser complicados y muy tediosos para grandes sets de datos con muchos valores perdidos.

Ruido

Los datos pueden contener **errores o valores anormales** a los cuales se les conoce como ***outliers*.** Estos datos pueden provocar anomalías y falta de exactitud en nuestros análisis.

Estos *outliers* deben ser identificados y corregidos o eliminar el atributo entero que los contengan. La forma más sencilla de identificar estos *outliers* está basada en el concepto estadístico de **dispersión.** Una distribución que, con media y varianza que sigue a una distribución parecida a la normal, puede definir como *outliers* a aquellos valores fuera de un intervalo alrededor de la media.

Otra técnica es la ilustrada en la figura 5, donde estableciendo un concepto de distancia entre valores, técnicas de **clustering** pueden ser aplicadas para identificar clúster de datos. Todo aquel valor que quede fuera de estos clústers puede ser considerado como *outliers.*

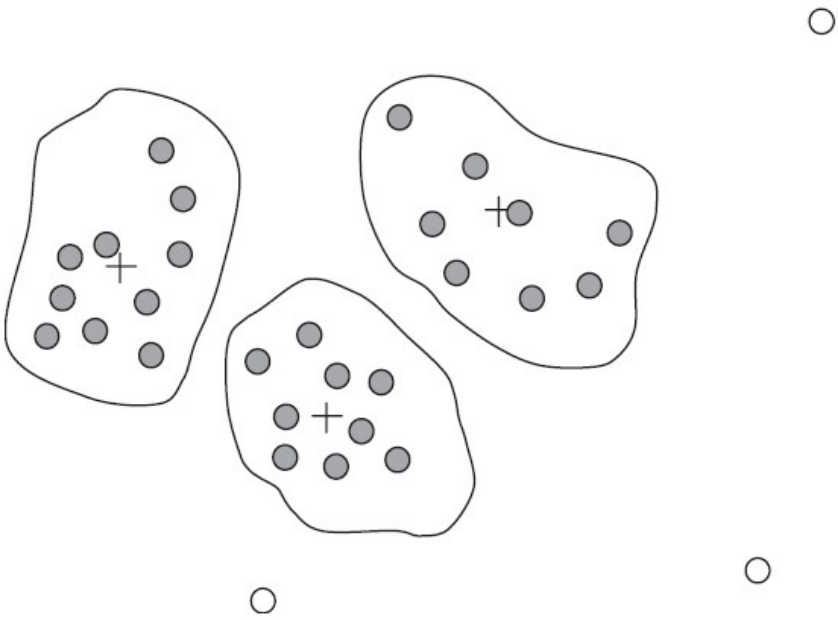


Figura 5. Clustering. Fuente: Vercellis (2009)

Todas estas técnicas deben ser **combinadas con la opinión de un experto** en los datos que pueda identificar si realmente esos valores pueden ser tratados como *outliers,* aunque estos valores no sean normales en la muestra.

También existen otras técnicas de regularización que automáticamente corrigen valores anómalos en la muestra. Por ejemplo, a través de técnicas de regresión con

información de otras variables se pueden obtener valores lógicos de un determinado atributo con un intervalo de confianza que puede ser computado para predecir la curva normal de valores que este debería tener.

Inconsistencia

Por último, los datos pueden contener discrepancias debido a un cambio en la forma de codificarlos o de obtenerlos. Por ejemplo, un cambio en la codificación de los productos que se venden deberá hacer que la recolección de datos se revise ya que los datos hasta ahora almacenados no son consistentes con los que se almacenarán en adelante y estos deberán ser transformados para que se traten por igual como productos.

Esto hace necesario que cualquier cambio en la medición o recolección de los datos deba tomarse con cautela y deba ser tenido en cuenta a la hora de preparar validar los datos para su análisis.

2.4. Transformación de datos

En la mayoría de las ocasiones es necesario realizar algún tipo de transformación a los datos para que estos puedan ser utilizados en el análisis. Son muchas las transformaciones que se pueden realizar a los datos y sirven para que sea más fácil su tratamiento.

Estandarización

En muchas ocasiones los datos deben ser estandarizados o normalizados debido a que la escala en la que se encuentran puede dificultar la aplicación de determinados modelos de análisis.

Por ejemplo, si queremos medir la cantidad de productos higiénicos que representa la compra mensual de unas familias en un supermercado y tenemos los datos de productos higiénicos en % del consumo total, la renta de la familia en miles de euros y el coste de la compra en euros, es posible que se necesite un proceso de normalización para determinados análisis.

Entre las técnicas más conocidas de estandarización o normalización de datos se encuentran:

- Normalización Min-Max
- Normalización Z-score
- Normalización de escalamiento decimal

Extracción de atributos de los datos

A veces, necesitamos extraer información de los datos que estos no representan de forma directa. Por ejemplo, supongamos que tenemos el nivel de ventas de nuestra página web por horas y necesitamos conocer a qué horas del día se produce el mayor incremento de compras para así planificar el mantenimiento de la página web

de forma eficiente. En este caso, deberíamos calcular las variaciones entre horas para nuestro análisis.

Este tipo de transformaciones y otra más complejas deben ser realizadas en **ocasiones para que nuestro análisis se centre en aquellas características de los datos que son nuestro objetivo.** Este tipo de transformaciones que se realizan a los datos pueden ser realizados con técnicas más complejas como las funciones kernel o las transformaciones de Fourier.

La extracción de atributos también puede suponer la creación de nuevas variables que contengan la información relevante para nuestro análisis. Por ejemplo, si queremos conocer las ventas que nuestra empresa realiza en una determinada ciudad, quizás sea más útil conocer el dato de ventas *per cápita* = Ventas / Población que el número total de ventas para toda la población.

Además de estas técnicas existen otra variedad de técnicas que pueden ser aplicadas para realizar transformaciones a los datos como son:

- Transformaciones lineales
- Transformaciones cuadráticas
- Aproximaciones no polinomiales
- Aproximaciones polinomiales
- Transformaciones de rango
- Transformaciones Box-Cox
- Transformación de nominal a binario

2.5. Reducción de los datos

Cuando trabajamos con conjuntos de datos pequeños las transformaciones que se han descrito hasta ahora son muy útiles para preparar los datos para el análisis. Si embargo, cuando el **conjunto de datos** es **grande** es necesario **reducir su tamaño** perdiendo la menor información posible para hacer que las **técnicas de análisis sean más eficientes.**

Antes de estudiar qué técnicas pueden ser aplicadas para reducir el tamaño de los conjuntos de datos perdiendo la menor información posible debemos conocer **tres criterios** que determinan qué tipo de reducción de datos debe aplicarse.

- **Precisión.** En la mayoría de los casos la precisión de los modelos es un factor crítico para que el análisis sea considerado como bueno y para seleccionar un método por encima de otro. Por esto, las técnicas de reducción no pueden comprometer la precisión de los datos y como se verá más adelante alguna de estas técnicas generalizan demasiado la información lo que puede provocar pérdida de precisión.

- **Eficiencia.** A la hora de realizar análisis de datos y aplicar modelos a los datos, el tiempo que tarda en computarse esos modelos es un factor importante que mide la eficiencia del análisis. Por utilizar esos modelos en sets de datos pequeños es más eficiente que en sets de datos mayores. La reducción de datos permite trabajar con una cantidad menor de información a la cual se le pueden aplicar más técnicas en menos tiempo.

- **Sencillez.** Algunas aplicaciones del análisis de datos están más centradas en la interpretación de estos. En muchas ocasiones, para que sea más sencillo su interpretación y análisis por parte de los expertos, estos prefieren sacrificar precisión para obtener resultados más simples e interpretables. La reducción de datos es una técnica para permitir que los modelos sean más sencillos.

Ya que es muy difícil desarrollar una técnica que tenga una solución óptima para los tres criterios anteriormente explicados, el analista debe decidir cuál de ellos es más importante para el propósito del análisis. La reducción de datos, por tanto, puede ser realizada a través de tres conjuntos de técnicas diferentes, la reducción del número de **observaciones** a través del **muestreo,** la reducción del número de **atributos** a través de la **selección y proyección** y la reducción del número de **valores** a través de la **discretización y la agregación.**

Muestreo

Una forma de reducción del tamaño del set de datos es **extraer una muestra** de esta que sea significativa desde un punto de vista estadístico, esto es, basándose en el razonamiento inferencial clásico.

Un paso previo es **determinar el tamaño de la muestra** que es requerido para que la precisión de los métodos de análisis posteriores no se vea afectado. Una vez determinado el tamaño óptimo ha de decidirse el **tipo de muestreo,** simple o estratificado dependiendo de si se desea mantener en la muestra los porcentajes respecto a un atributo del set inicial.

Un procedimiento típico es configurar muestras independientes cada una de un tamaño diferente para ver cuál es la que mejor precisión consigue en los modelos y poder compararlas.

Las conclusiones que se extraigan de los posteriores análisis pueden ser consideradas como robustas y válidas siempre y cuando estas se mantengan relativamente estables con todo el conjunto de muestras utilizado.

Selección de características

El propósito de esta técnica es **eliminar las variables que no son relevantes para el objetivo del análisis** a realizar. Este es uno de los aspectos más críticos del análisis de datos, donde seleccionar qué variables son las adecuadas para investigar el fenómeno en cuestión es clave.

Este procedimiento permite reducir la cantidad de datos a procesar en los análisis, aumentar la precisión de análisis al no introducir variables que no sean útiles y normalmente son más sencillas de interpretar, por lo que cumplen los tres atributos esenciales.

En general los métodos de selección de características o atributos clave se pueden clasificar en tres grandes grupos:

- Método de filtrado. Estos métodos seleccionan los atributos relevantes antes de continuar por lo que son independientes al análisis que se vaya a realizar.

Los atributos seleccionados se utilizan para el análisis y el resto se eliminan. Son algunos los métodos que existen para que este filtrado sea eficiente y no se eliminen atributos que podrían ser útiles para el análisis.

El método más simple implica el estudio de las correlaciones entre la variable objetivo y las variables a seleccionar con el fin de seleccionar aquellas que posean un mayor grado de correlación.

- Método «wrapper». El método «wrapper» o método de envoltura en español tienen como objetivo mantener un alto nivel de precisión y es útil como procedimiento previo para aquellos análisis que necesiten una alta precisión.

Estos métodos realizan la selección de atributos dependiendo del nivel de precisión que generen en los modelos de análisis. Prueban con todas las posibilidades y seleccionan aquellos atributos que mejores resultados generan.

Hay que tener en cuenta que este método es muy costoso en términos de eficiencia, ya que normalmente son métodos que emplean mucho tiempo y capacidad computacional.

- Métodos integrados. Los métodos integrados seleccionan los atributos dentro del propio algoritmo de análisis, de modo que la selección del conjunto óptimo de atributos se hace directamente en la fase de generación del modelo.

Un ejemplo de esto son los métodos de clasificación en forma de árboles donde en cada nodo del árbol se evalúan la capacidad de cada atributo para generar valor en el modelo.

Cada método ofrece unos beneficios diferentes y deben ser aplicados de forma correcta en función de la cantidad de datos, el tipo de datos y el objetivo del análisis.

Análisis de componentes principales

Cuando hemos hablado de agregar para reducir el número de valores a analizar nos referimos al método de Análisis de Componentes Principales (PCA, por sus siglas en inglés) que es una de las técnicas más conocidas para **reducir el número de atributos.**

El objetivo de este método es obtener un subset de atributos que **reduzca el número de datos a analizar sin perder información** ninguna a través de la transformación e integración de los atributos originales.

Para entender esto mejor imaginemos que tenemos un set de datos con 11 atributos y realizamos un análisis de componentes principales. El resultado del análisis se puede resumir en la figura 6 que muestra el porcentaje de la varianza que es explicada por cada componente.

De esta forma podemos encontrar cuál es el número de componentes principales a seleccionar para nuestro análisis mediante los cuales recogemos casi la totalidad de la varianza.

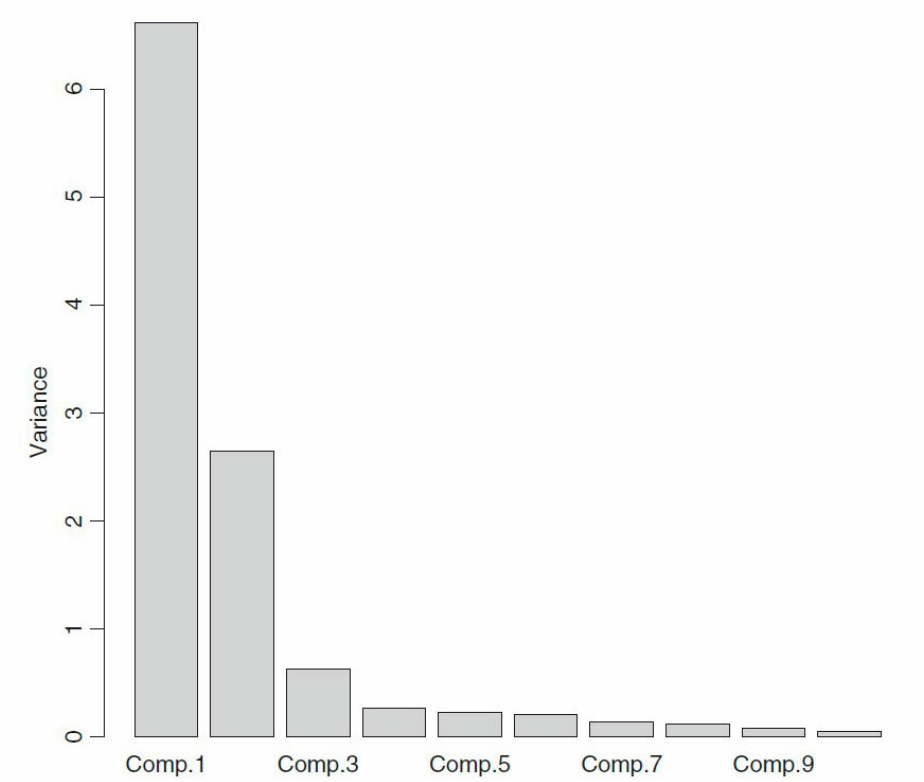


Figura 6. Análisis de componentes principales.

Discretización de datos

El objetivo principal de los métodos de reducción es **reducir el número de valores** que se encuentran en la muestra de datos. Discretizar los datos consigue reducir el número de valores puesto que incorpora dentro de un atributo categórico varios números y reduce el número de posibles valores diferentes. Por ejemplo, si tenemos la cantidad de ventas por vendedor y día quizás nos interese resumir las ventas en tres categorías: bajas, medias y altas.

Esta discretización puede mejorar la capacidad analítica de los modelos haciendo además más fácil su interpretación.

Ideas clave

Las técnicas de discretización más conocidas son la subdivisión subjetiva, la subdivisión en clases y la discretización jerárquica.

- **Subdivisión subjetiva.** Es la más popular e intuitiva ya que las clases se definen basándose en la experiencia y la opinión de un experto en el tema en cuestión.
- **Subdivisión en clases.** Esta técnica consiste en dividir los datos siguiendo algún patrón, por ejemplo en función de la media o del tamaño de estos.
- **Discretización jerárquica**. Este tipo de discretización se basa en la jerarquía que existe entre diferentes categorías. En concreto cuando un conjunto de categorías pertenece a una categoría superior es posible reemplazar cada valor de un atributo por el correspondiente valor de la categoría superior en el nivel jerárquico, por ejemplo, provincias en comunidades o comunidades en países.

2.6. Almacenamiento de datos

Para almacenar y gestionar los datos existen distintas alternativas conocidas como **bases de datos.**

Una base de datos es un conjunto de datos relacionados duraderos en el tiempo y con un significado implícito que se utilizan a través de un software. Toda base de datos dispone de al menos cuatro componentes:

- **Datos,** que son el motivo por el cual se crean. Estos pueden estar almacenados y relacionados de diferentes formas y pueden ser accesibles por una sola persona (integrados) o por varias (compartidos).
- **Hardware,** necesario para poder almacenar y gestionar los datos y el software asociado. Cuenta con volúmenes de almacenamiento, procesadores y una memoria principal.
- **Software,** se encarga de relacionar al usuario con la base de datos y generalmente se le conoce como SGBD (Sistema de Gestión de Bases de Datos).
- **Usuarios,** entre los que se encuentran los programadores, encargados de crear aplicaciones para interaccionar con la base de datos, los usuarios finales de los datos almacenados y los administradores de la base de datos que se encarga de gestionar la estructura, disponibilidad y eficiencia de esta.

Ideas clave

La alternativa clásica y la más extendida en las compañías son las **bases de datos relaciones** que utilizan el lenguaje **SQL.** En estas bases de datos los archivos se representan en tablas donde cada columna representa un atributo diferente y cada fila un registro. Las tablas están relacionadas entre sí mediante una clave creando un entramado de relaciones que representa la realidad de la estructura a representar mediante datos.

Con la llegada del conocido *big data* se han popularizado los ***data warehouse*** o «almacenes de datos» que están diseñados para responder a consultas de todo tipo sin tener en cuenta el rendimiento, las consultas en paralelo o el espacio optimizado. Es una herramienta útil para conocer el estado de una organización en un determinado momento del tiempo.

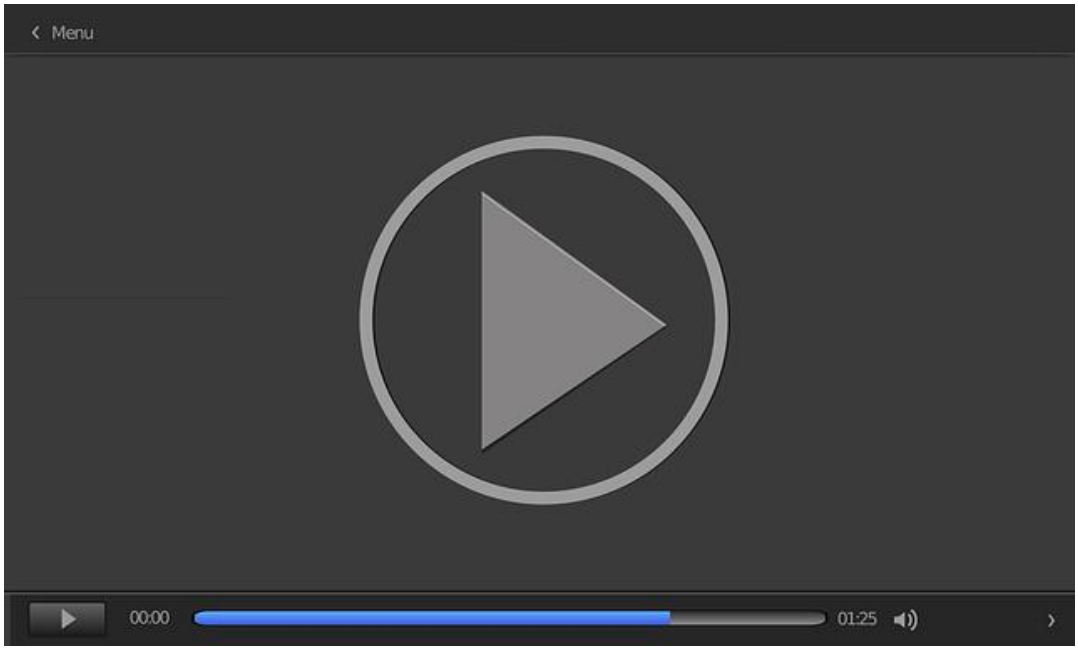
2.7. Referencias bibliográficas

Jarke, M., Jeusfeld, M. A., Quix, C. y Vassiliadis, P. (1998). Architecture and quality of data warehouses: An extended repository approach. *Advanced Information Systems Engineering, Lecture Notes in Computer Science*, 1413, 243-260.

Internet de las cosas: sensores, sistemas embebidos y vestibles como fuente del dato

Ministerio de Industria, Energía y Turismo. (2016, 28 abril). *Internet de las cosas: sensores, sistemas embebidos y vestibles como fuente del dato* [Archivo de vídeo]. https://www.youtube.com/watch?v=vYXH2WLFDdM

En este vídeo encontrarás cómo muchos dispositivos tecnológicos pueden captar información y almacenarla como datos que pueden ser útiles para la toma de decisiones de la empresa.



Accede al vídeo:

https://www.youtube.com/embed/vYXH2WLFDdM

Test

1. El conocimiento en la empresa:

   A. Es generado directamente de los datos.

   B. Es la información que se extrae de los datos mediante un proceso de análisis.

   C. Se extrae a partir de la información obtenida de los datos.

   D. Se genera sin necesidad de datos.

2. ¿Qué criterios podemos seguir a la hora de reducir el tamaño de los datos?

   A. Precisión.

   B. Eficiencia.

   C. Sencillez.

   D. Todas las anteriores.

3. ¿Qué técnica se puede utilizar para reducir el número de observaciones?

   A. Muestreo.

   B. Selección y proyección.

   C. Discretización y agregación.

   D. Ninguna de las anteriores.

4. La discretización de los datos sirve para:

   A. Conseguir datos más consistentes.

   B. Eliminar los datos erróneos.

   C. Reducir el número de valores distintos.

   D. Entender mejor la información contenida en los datos.

5. En qué parte del proceso de obtención de datos se debe tener en cuenta el ruido:

A. En la extracción.

B. En la validación.

C. En el almacenamiento.

D. En la transformación.

6. ¿Cuál de las siguientes opciones es un ejemplo de archivo plano usado para almacenar datos?

A. SQL.

B. .docx.

C. .csv.

D. .exe.

7. ¿Qué formato de archivo permite estructurar datos mediante etiquetas jerárquicas?

A. JSON.

B. XML.

C. CSV.

D. SQL.

8. ¿Qué método de tratamiento de valores perdidos puede introducir sesgos si los datos están distribuidos de forma asimétrica?

A. Eliminación.

B. Reemplazo por media.

C. Inspección manual.

D. Sustitución por categoría.

9. ¿Cuál de estas técnicas se basa en crear nuevas variables a partir de las existentes para facilitar el análisis?

A. PCA.

B. Discretización.

C. Extracción de atributos.

D. Eliminación de duplicados.

10. ¿Qué tipo de base de datos se caracteriza por organizar la información en tablas relacionadas entre sí mediante claves?

A. NoSQL.

B. Data warehouse.

C. Relacional.

D. En la nube.