2019
Data Scientist

CURSO BÁSICO

AleXandria
Business Solutions Based on Data

+54 9 11 4973-8574 | www.alexandria.com.ar | C.A.B.A., Argentina

CONTENIDO

MODULO 1

Analytics:

Es el uso de herramientas y técnicas que convierten los DATOS en importante INFORMACIÓN para el negocio. Por lo tanto, podemos encontrar ahí dentro, una cantidad de términos que ya conocemos, como, por ejemplo:

- Business Analytics: El análisis de negocios se refiere a las habilidades, tecnologías y prácticas para la exploración e investigación iterativas continuas del desempeño empresarial anterior para obtener una perspectiva y dirigir la planificación empresarial.
- Business Intelligence: es la habilidad para transformar los datos en información, y la información en conocimiento, de forma que se pueda optimizar el proceso de toma de decisiones en los negocios.
- Data Analytics: El análisis de datos es un proceso que consiste en inspeccionar, limpiar y transformar datos con el objetivo de resaltar información útil, lo que sugiere conclusiones, y apoyo en la toma de decisiones.
- Data Science: es la evolución de lo que hasta ahora se conocía como Analista de datos, pero a diferencia de éste que sólo se dedicaba a analizar fuentes de datos de una única fuente, el Data Scientist debe explorar y analizar datos de múltiples fuentes, a menudo inmensas (conocidas como Big Data), y que pueden tener formatos muy diferentes. Además, debe tener una fuerte visión de negocio para ser capaz de extraer y transmitir recomendaciones a los responsables de negocio de su empresa.
- Data Mining: Título de la maestría que obtiene un científico de datos.
- Inteligencia Artificial: es la combinación de algoritmos planteados con el propósito de crear máquinas que presenten las mismas capacidades que el ser humano.
- Deep Learning: es un conjunto de algoritmos de clase aprendizaje automático que intenta modelar abstracciones de alto nivel en datos usando arquitecturas compuestas de transformaciones no lineales múltiples.

Por esta razón, se pueden diferenciar tipos de analytics dadas las preguntas a las que responden cada uno de ellos.

Tipos de Analytics:

Cuando nos enfrentamos a los datos tenemos las siguientes preguntas a responder:

- ¿Qué pasó?
- ¿Por qué ocurrió?
- ¿Qué va a pasar?
- ¿Qué debería hacerse al respecto?

A estas preguntas podemos catalogarlas de la siguiente manera:

- Descriptivo: ¿Qué pasó?
- Diagnóstico: ¿Por qué ocurrió?
- Predictivo: ¿Qué va a pasar?
- Prescriptivo: ¿Qué debería hacerse al respecto?

Modelos de Analytics

Descriptivos

Los modelos descriptivos identifican patrones que explican o resumen los datos, es decir, sirven para explotar las propiedades de los datos examinados, no para predecir nuevos datos.

Ejemplo, un call center quiere identificar las personas que tienen gustos similares, con el objetivo de ofrecerles productos similares a cada grupo. Para ello, analiza los datos que tiene respecto a sus clientes e infiere un modelo descriptivo que caracteriza estos grupos.

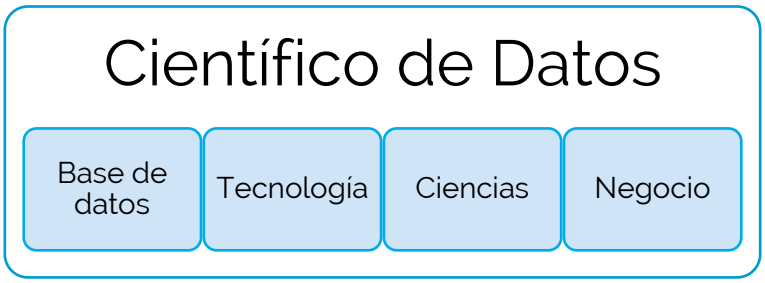
Predictivos

Los modelos predictivos, por otro lado, pretenden estimar valores futuros o desconocidos de variables de interés, que denominamos *variables objetivo* o *dependientes*, usando otras variables o campos de la base de datos, a las que nos referiremos como *variables independientes* o *predictivas*.

Ejemplo, estimación de la demanda de un producto en función del gasto de publicidad.

Perfil del Data Scientist

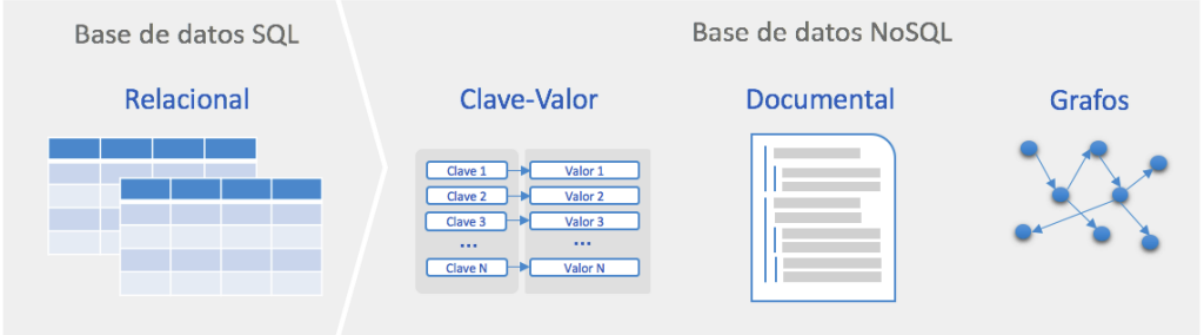
Se lo conoce como el unicornio, porque sus tareas le demandan conocimientos en los siguientes ámbitos:



Introducción a las bases de datos

Algunos tipos de bases de datos

A modo de acercarnos a los conceptos básicos que debemos tener incorporados para avanzar, es bueno poder diferenciar estos dos tipos de bases de datos.



Que pueden resultar de distintos tipos de fuentes:

- Fuentes internas:
  - Bases de datos de clientes
  - Bases de datos de transacciones
  - Archivos de campañas
  - Data Warehouse
  - Sistemas Operacionales
- Fuentes externas:
  - Bases de datos demográficos
  - Bases de datos comerciales
  - Padrones

Tipos de datos

Como a nadie se le escapa, estas bases de datos funcionan con datos de diferentes tipos. A modo de aclaración mencionamos las *principales clases de tipos de datos*:

- Según el momento de su tratamiento:
  - Datos brutos
  - Datos elaborados o derivados
  - Datos resultantes de un proceso de análisis o modelización
- Según su escala de medida:
  - Variables cuantitativas o cualitativas, escalas de medida:
    - Nominal
    - Ordinal
    - Intervalo
    - Razón
- Según su dimensionalidad:
  - Univariados
  - Bivariados
  - Multivariados
- Según su papel en un modelo:
  - ID: identificador de los casos
  - INPUT: variables de entrada
  - TARGET: variable objetivo
  - TIMEID: valor de la dimensión temporal

Para aplicar el conocimiento de esto, vamos a realizar el siguiente ejemplo:

IdE	ENombre	Sueldo	Edad	Sexo	IdD
1	Juan	2.100	45	H	Ge
2	Elena	2.400	40	M	Ma
3	María	?	53	M	Ge
4	Pedro	1.000	20	H	Ge
5	Lucía	3.55	35	M	Ma

IdD	DNombre	Director
Ge	Gestión	Rubio
Ve	Ventas	Burriel
Ma	Marketing	Torrubia

Listar la media de edad de los empleados de una empresa cuyo sueldo es mejor que € 2.000.-

```
SELECT D.IdD, D.DNombre, AVG(E.Edad)
FROM empleado E JOIN departamento D ON E.IdD=D.IdD
WHERE E.Sueldo >2000
GROUP BY D.IdD, D.DNombre
```

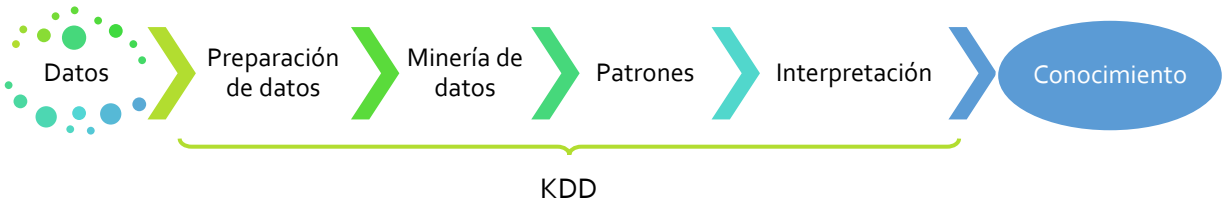
El resultado es el siguiente:

IdD	DNombre	AVG_edad
Ge	Gestión	45.0
Ma	Marketing	37.5

Proceso de descubrimiento del conocimiento

El proceso de descubrimiento del conocimiento en las bases de datos, o KDD (Knowledge Discovery from data), es un proceso complejo de transformación de datos brutos en información significativa y relevante para la toma de decisiones.

Graficando el proceso vemos:



"Es el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos"[1]

Pero analicemos el contenido de esta definición:

1. Válido: los patrones deben ser precisos para datos nuevos (con cierto grado de certidumbre).
2. Novedoso: debe aportar algo que por otros medios no era posible obtener de tal modo, que justifique su realización.
3. Potencialmente útil: debe ser información que genere un beneficio al usuario en relación a la toma de decisiones.
4. Comprensible: la extracción de patrones no comprensibles dificulta o imposibilita su interpretación, revisión, validación y uso en la toma de decisiones. De hecho, no proporciona conocimiento, sino que hasta puede llegar a confundir.

Por lo tanto, el proceso KDD nos permite la selección, limpieza, transformación y proyección de los datos; analizar los datos para extraer patrones y modelos adecuados; evaluar e interpretar los patrones para convertirlos en conocimiento; consolidar el conocimiento resolviendo posibles conflictos con conocimiento; hacer el conocimiento disponible para su uso.

[1] Fayyad (1996)

Tipos de tareas de Data Mining

Este proceso KDD es común a todos los tipos de minería, porque es justamente un proceso que engloba dicha industria. Es importante destacar que el uso de las tecnologías ha contribuido al avance de herramientas emergentes como son la minería de textos y la minería de sentimientos. Lo anterior debido a que el tratamiento de la información de forma manual sería casi imposible.

Las empresas que se dedican a la elaboración de programas cada vez más avanzados para el tratamiento de datos, textos y opiniones permiten a los dirigentes de las organizaciones una toma de decisiones.

La minería de sentimientos aún se encuentra en una etapa de maduración. Sin embargo, las diferentes combinaciones que surgen con diferentes técnicas y aplicaciones tecnológicas han generado la aparición de organizaciones cada vez más competitivas.

Por lo tanto, mencionaremos en esta instancia algunas de las aplicaciones más comúnmente realizadas en relación a la explotación:

Minería de datos

> *"Es el conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto."*[2]

Ampliaremos este enfoque de la minería durante el resto del curso.

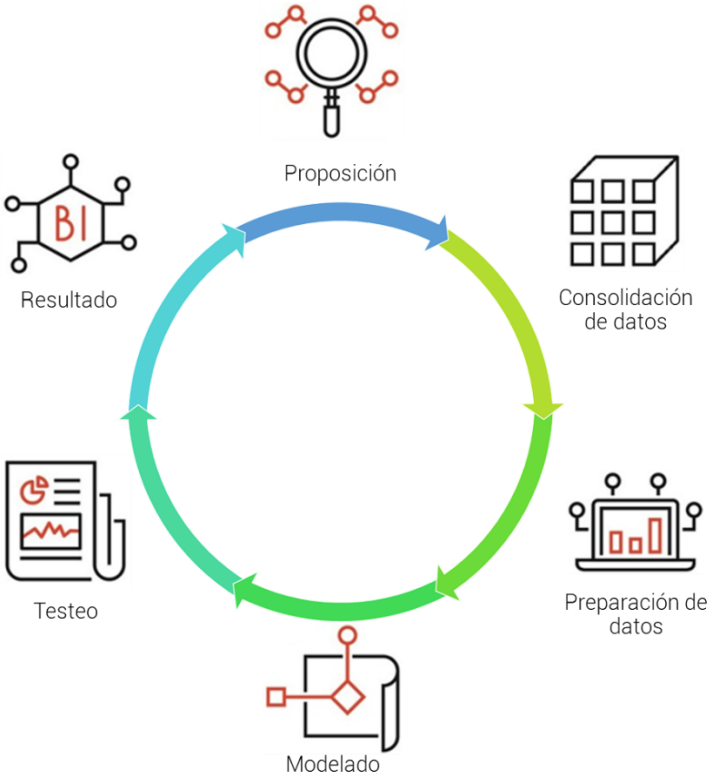
Minería de textos

> *"Es el proceso encargado del descubrimiento de información que no existía explícitamente en ningún texto de la colección, peor que surge de relacionar el contenido de varios de ellos"*[3]

El objetivo de la minería de texto es el descubrimiento de nueva información a partir de colecciones de documentos de texto no estructurado. Por no estructurado nos referimos a texto libre, generalmente en lenguaje natural, aunque también podría ser código fuente u otro tipo de información textual. La tarea de minería más habitual sobre estos datos es la categorización, la clasificación y el agrupamiento de textos. Podemos decir que la categorización es la tarea que identifica categorías, temas, materias o conceptos presentes en los textos, mientras que la clasificación es la tarea de asignar una clase o una categoría a cada documento.

[2] Sinnexus (2016)
[3] Rochina (2017)

Este tipo de minería puede usarse para proposiciones, y uno de los posibles procesos para realizarlas es el siguiente:



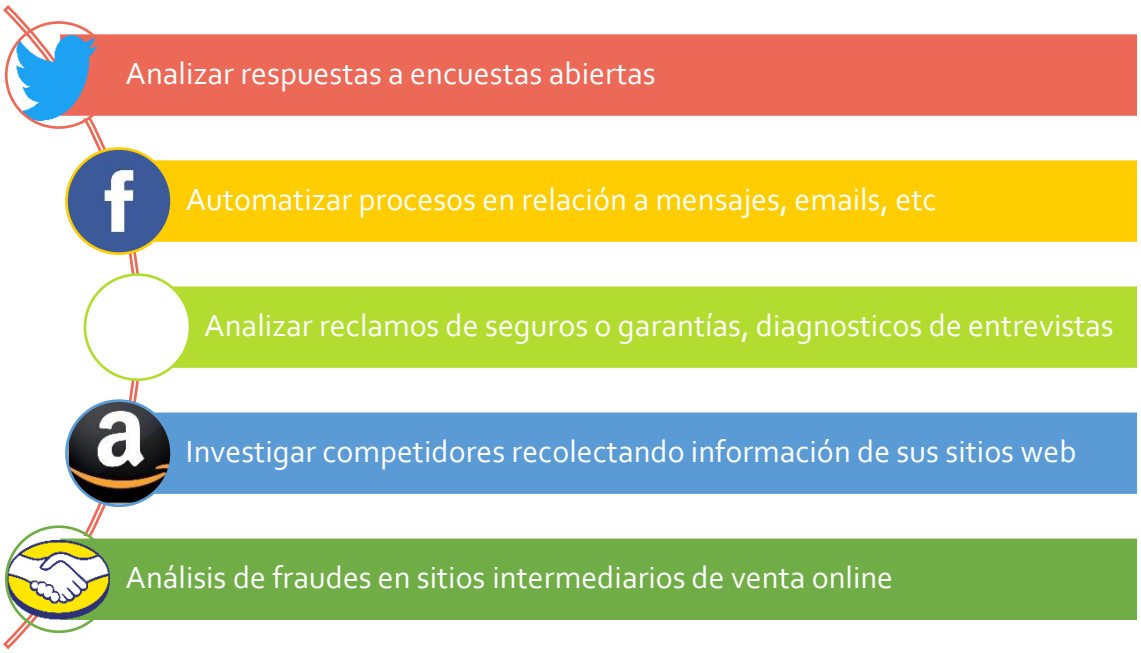
Retos de este tipo de minería

La minería de textos enfrenta algunos desafíos como pueden ser:

- Se debe conocer los **contextos** bajo los cuales se generan los contenidos. El poder conocer a los autores de la información, el momento en el que se está escribiendo, al igual que el lugar donde se escribe, permite tener un panorama más amplio del por qué se escribió alguna información.
- La cantidad de datos que se analizan es demasiado grande por lo que es necesaria la creación de algoritmos capaces de analizar grandes cantidades de datos utilizando al máximo los equipos de cómputo destinados a realizar la minería de datos.
- La información no siempre se convierte en conocimiento. Es necesario que la información obtenida ayude a las organizaciones a tomar importantes decisiones. Además de esto es necesario que la información sea transformada a un lenguaje que los ordenadores puedan manejar lo cual en muchas organizaciones es muy difícil de obtener.[4]

[4] LOZADA

Algunos ámbitos de aplicación son:



Minería de sentimientos

"Se refiere al uso de procesamiento de lenguaje natural, análisis de texto y lingüística computacional para identificar y extraer información subjetiva de los recursos"

Este tipo de minería surge, en general, como una necesidad de las organizaciones de "conocer" a sus clientes, las preferencias de los públicos que se manejan, más que nada cuando se trata de organizaciones que manejan su imagen pública por medios digitales. Para esto se generaron técnicas como el análisis de opiniones o las herramientas de monitoreo de medios sociales.

La minería de sentimientos trabaja en conjunto con la minería de textos a fin de encontrar el conocimiento en base a las opiniones.

- **Detección de la polaridad**: se realiza a fin de conocer si la opinión es positiva o negativa por lo cual se genera un rango para determinar dicha polaridad lo cual permita conocer las preferencias de las personas. Esto se realiza a través de diferentes algoritmos que realizan un análisis de las palabras utilizadas en un mensaje, eliminando las palabras ambiguas y dejando únicamente aquellas que representan una opinión de manera más concreta.
- **Análisis del sentimiento en base a las características**. De acuerdo a la opinión de los usuarios se puede conocer la aceptación o rechazo del producto. En base a esto se puede determinar si la aceptación es positiva o negativa.[5]

[5] CORTIZO, J. (s.f.). MINERIA DE OPINIONES. Obtenido de http://www.brainsins.com/es/blog/mineria-opiniones/3555

Algunas aplicaciones

Este tipo de técnicas son de gran utilización en diferentes ámbitos como son el empresarial y político.

- EMPRESARIAL: Permite conocer las opiniones de los empleados, cómo se siente en el trabajo y la aceptación o no de las estrategias o actividades que se están implementando.

Al mismo tiempo unas de las aplicaciones más comunes es que a través de la minería de sentimientos se puede conocer la aceptación de un producto o servicio por parte de las personas.

- POLÍTICO: A través de diversas herramientas se puede conocer la opinión de las personas en relación a un candidato, o partido político.

Ejemplo de Aplicaciones

Análisis de créditos bancarios

Un banco por Internet desea obtener reglas para predecir qué personas de las que solicitan un crédito no lo devuelven. La entidad bancaria cuenta con los datos correspondientes a los créditos concedidos con anterioridad a sus clientes (cuantía del crédito, duración en años, ...) y otros datos personales como el salario del cliente, si posee casa propia, etc. Algunos registros de clientes de esta base de datos son:

IDC	d.crédito (años)	c.crédito ($)	Salario ($)	Casa propia	Cuentas morosas	...	Devuelve el crédito
101	15	60.000.-	2.200.-	Sí	2	...	No
102	2	30.000.-	3.500.-	Sí	0	...	Sí
103	9	9.000.-	1.700.-	Sí	1	...	No
104	15	18.000.-	1.900.-	No	0	...	Sí
105	10	24.000.-	2.100.-	No	0	...	No
...	...	...	...	...	...	...	...

A partir de estos datos, las técnicas de minería de datos podrían sintetizar algunas reglas, como, por ejemplo:

SI [Cuentas morosas] > 0 **ENTONCES** [Devuelve el crédito] = No

SI [Cuentas morosas] = 0 **Y** ([Salario] >25.000.- **O** [d.crédito]>10)

ENTONCES [Devuelve el crédito] = Sí

El banco podría utilizar estas reglas para determinar las acciones a realizar en el trámite de los créditos: si se concede o no el crédito solicitado, si es necesario pedir avales especiales, etc.

Análisis de la cesta de la compra

Supongamos que un supermercado quiere obtener información sobre el comportamiento de compra de sus clientes; porque piensan que de esta manera pueden mejorar el servicio que les ofrece: reubicación de los productos que se suelen comprar juntos, localizar el empadronamiento idóneo para nuevos productos, etc.

Para ello dispone de la siguiente información de los productos que se recopila con cada una de las compras o cestas.

IDcesta	Huevos	Aceite	Pañales	Vino	Leche	Manteca	Salmón	Lechugas	…
1	Sí	No	No	Sí	No	Sí	Sí	Sí	…
2	No	Sí	No	No	Sí	No	No	Sí	…
3	No	No	Sí	No	Sí	No	No	No	…
4	No	Sí	Sí	No	Sí	No	No	No	…
5	Sí	Sí	No	No	No	Sí	No	Sí	…
6	Sí	No	No	Sí	Sí	Sí	Sí	No	…
7	No	No	No	No	No	No	No	No	…
8	Sí	Sí	Sí	Sí	Sí	Sí	Sí	No	…
…	…	…	…	…	…	…	…	…	…

Al analizar estos datos, el supermercado podría encontrar, por ejemplo, que el 100% de las veces que se compran pañales también se compra leche. Que el 50% de las veces que se compra vino y salmón, entonces se compran lechugas. También, se puede analizar cuáles de estas asociaciones son frecuentes, porque una asociación muy estrecha entre dos productos puede ser poco frecuente y, por lo tanto, poco útil.

Determinar las ventas de un producto

Una cadena de tiendas de electrodomésticos desea optimizar el funcionamiento de su almacén manteniendo un *stock* de cada producto suficiente para poder servir rápidamente el material adquirido por sus clientes. Para ello, la empresa dispone de las ventas efectuadas cada mes del último año de cada producto, tal y como se muestra a continuación:

Producto	Mes-12	…	Mes-4	Mes-3	Mes-2	Mes-1
Televisor plano 30'	20	…	52	14	139	74
Video-DVD recorder	11	…	43	32	26	59
Celular alta gama	50	…	61	14	5	28
Frigorífico no Frost	3	…	21	27	1	49
Microondas con grill	14	…	27	2	25	12
…	…	…	…	…	…	…

Con esta información, la empresa puede generar un modelo para predecir cuales van a ser las ventas de cada producto en el siguiente mes en función de las ventas realizadas en los meses anteriores, y efectuar así los pedidos necesarios a sus proveedores para disponer del *stock* necesario para hacer frente a esas ventas.

Determinar grupos diferenciados de empleados

El departamento de HR de una gran empresa desea categorizar a sus empleados en distintos grupos con el objetivo de entender mejor su comportamiento y tratarlos de manera adecuada. Para ello dispone en sus bases de datos de la siguiente información sobre los mismos:

Id	Sueldo	Casado	Auto	Hijos	Alq/prop	Sindicato	Bajas/año	Antigüedad	Sexo
1	1.000.-	Sí	No	0	Alquiler	No	7	15	H
2	2.000.-	No	Sí	1	Alquiler	Sí	3	3	M
3	1.500.-	Sí	Sí	2	Prop	Sí	5	10	H
4	3.000.-	Sí	Sí	1	Alquiler	No	15	7	M
5	1.000.-	Sí	Sí	0	Prop	Sí	1	6	H
6	4.000.-	No	Sí	0	Alquiler	Sí	3	16	M
7	2.500.-	No	No	0	Alquiler	Sí	0	8	H
8	2.000.-	No	Sí	0	Prop	Sí	2	6	M
9	2.000.-	Sí	Sí	3	Prop	No	7	5	H
10	3.000.-	Sí	Sí	2	Prop	No	1	20	H
11	5.000.-	No	No	0	Alquiler	No	2	12	M
12	800.-	Sí	Sí	2	Prop	No	3	1	H
13	2.000.-	No	No	0	Alquiler	No	27	5	M
14	1.000.-	No	Sí	0	Alquiler	Sí	0	7	H
15	800.-	No	Sí	0	Alquiler	No	3	2	H
...	...	...	...	...	...	...	...	...	...

Por medio de sistemas de minería de datos se podrían obtener tres grupos con las siguientes características:

Grupo 1

- Sueldo: $1.535,2
- Casado:
  - No 0,777
  - Sí 0,223
- Auto:
  - No 0,82
  - Sí 0,18
- Hijos 0,05
- Alq 0,99
- Prop 0,01
- Sindicato
  - No 0,8
  - Sí 0,2
- Bajas 8,3
- Antigüedad 8,7
- Sexo
  - H 0,61
  - M 0,39

Grupo 2

- Sueldo: $1.428,7
- Casado:
  - No 0,98
  - Sí 0,02
- Auto:
  - No 0,01
  - Sí 0,99
- Hijos 0,3
- Alq 0,75
- Prop 0,25
- Sindicato
  - Sí 1
- Bajas 2,3
- Antigüedad 8
- Sexo
  - H 0,25
  - M 0,75

Grupo 3

- Sueldo: $1.233,8
- Casado:
  - Sí 1
- Auto:
  - No 0,05
  - Sí 0,95
- Hijos 2,3
- Alq 0,17
- Prop 0,83
- Sindicato
  - No 0,67
  - Sí 0,33
- Bajas 5,1
- Antigüedad 8,1
- Sexo
  - H 0,83
  - M 0,17

Estos grupos podrían ser interpretados por el departamento de recursos humanos de la siguiente manera:

- Grupo 1: sin hijos y con vivienda de alquiler. Poco sindicados. Muchas bajas.
- Grupo 2: sin hijos y con auto. Muy sindicados. Pocas bajas. Normalmente son mujeres y viven en casas de alquiler.
- Grupo 3: con hijos, casados y con auto. Mayoritariamente hombres propietarios de su vivienda. Poco sindicados.

Otros ejemplos

- El área más conocida por todos probablemente sea RIESGOS.
- Detección de FRAUDES transaccionales es otra rama de Analytics, donde se hace foco en reducir las tasas de falsos positivos.
- Mejorar la relación con los CLIENTES, ofreciendo PRODUCTOS relevantes (hábitos, estilo de vida) a partir de información transaccional, comportamental, social.
- Modelos de ATTRITION para retención de clientes.
- Medición de la RENTABILIDAD prevista de los clientes y su LIFETIME VALUE.
- Análisis de SOCIAL MEDIA.
- Análisis de SENTIMIENTO.
- SEGMENTACIÓN de la cartera de clientes.
- Modelos de estimación de SALARIOS.
- Modelos de PRICING de Préstamos según plazos.
- Modelos de detección de REVOLVER en Tarjeta de Crédito.
- Análisis GENERACIONAL: Millenials y TC para estrategia de MKT y producto.
- Herramientas de reporting y VISUALIZACIÓN.
- Aplicación de técnicas colaborativas item-item y text-mining para RECOMENDACIÓN de mejor siguiente rubro de consumo.

Practica modulo 1

Para la siguiente actividad vamos a utilizar el Dataset de Titanic.

Este dataset cuenta con la siguiente información:

Variable	Definition	Key
Survival	Survival	0 = No, 1 = Yes
Pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
Sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg Q = Queenstown S = Southampton

Para realizar la práctica deberás entrar a GitHub y luego podes descargarte los archivos y trabajarlos en tu maquina o ir a https://mybinder.org y ejecutar Jupyter notebook desde ahí.

Test teórico

Ingresar al siguiente link e ingresar el código que aparece en pantalla: https://kahoot.it/

MODULO 2

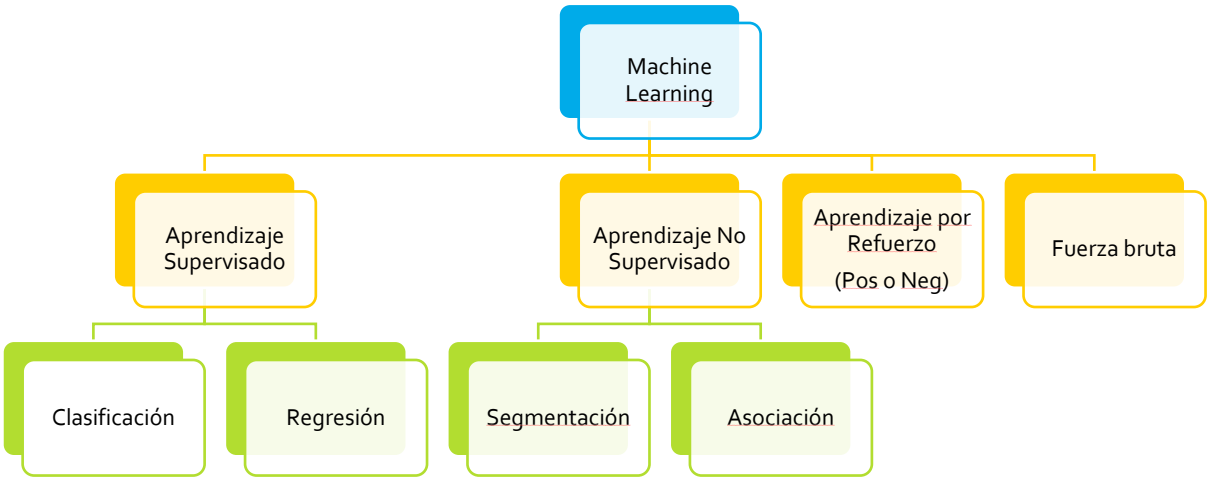
Definiciones iniciales

Antes de empezar el proceso de minería propiamente dicho, se deben tomar una serie de decisiones:

1- Determinar qué tipo de tarea de minería es el más apropiado. Por ejemplo, podríamos usar la clasificación para predecir en una determinada compañía quiénes dejarán de ser clientes.
2- Elegir el tipo de modelo. Por ejemplo, para una tarea de clasificación podríamos usar un árbol de decisión, porque queremos obtener un modelo en forma de reglas.
3- Elegir el algoritmo que resuelva la tarea y obtenga el tipo de modelo que estamos buscando. Esta elección es pertinente porque existen muchos métodos para construir los modelos. Por ejemplo, para crear árboles de decisión para clasificación podríamos usar CART o C5.0, entre otros.

Tipos de Machine Learning

Por lo tanto, para poder concretar las decisiones recientemente mencionadas, vamos a necesitar conocer las tareas y los modelos más utilizados en minería de datos.



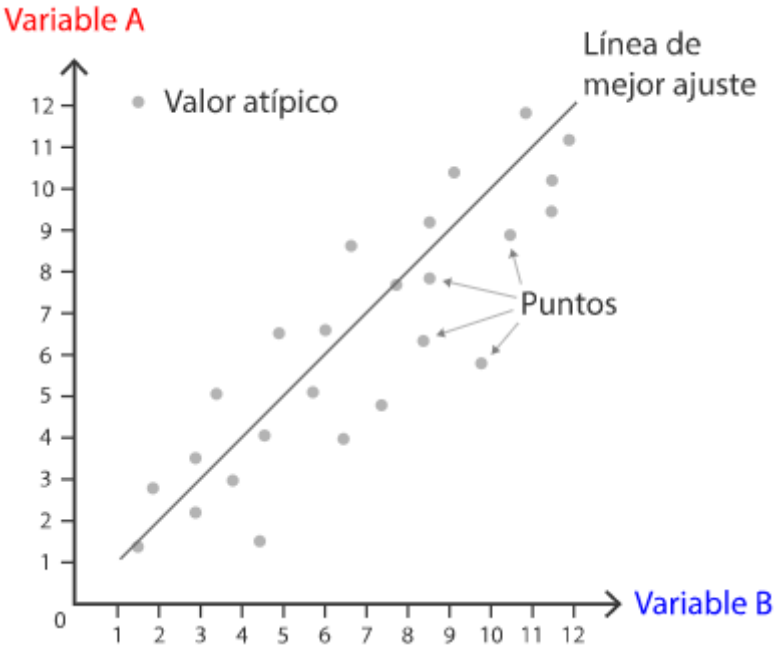
Estas tareas pueden considerarse como tipos de problemas diferentes a ser resueltos por un algoritmo de minería de datos. Por lo tanto, cada tarea tiene sus propios requisitos, y el tipo de información obtenida con una tarea puede diferir mucho de la obtenida con otra.

El conocimiento descubierto no sólo depende del algoritmo de minería utilizado, sino también de la calidad de los datos minados. Por lo tanto, después de la recopilación puede surgir, en la mayoría de los casos, que aparezcan problemas en relación a valores extremos, datos faltantes o erróneos.

En relación a su tratamiento, trabajaremos en los siguientes temas de este módulo.

Preprocesamiento de datos

Como vimos en el módulo anterior, el proceso de descubrimiento del conocimiento tiene como primera etapa el procesamiento de datos. En esta etapa está integrada en primer lugar por el preprocesamiento de datos. Es imprescindible conocer los datos para poder darles el mejor tratamiento, y así perfeccionar el descubrimiento de conocimiento, en primer lugar, porque las bases de datos no son perfectas y pueden aparecer valores extremos o faltantes.



Presencia de outliers

Como mencionamos, pueden aparecer problemas en relación a la calidad de los datos. Uno de estos problemas es encontrarnos con valores que no se ajustan al comportamiento general de los datos, y se conocen comúnmente como *outliers*. Estos valores anómalos pueden representar errores en los datos o pueden ser valores correctos que son simplemente diferentes a los demás.

Respecto a los algoritmos que utilizamos para la minería, algunos pueden ser robustos (indiferentes a la existencia de outliers), otros ignoran estos datos, otros los descartan por considerarlos ruido o excepciones, pero existen otros que son muy sensibles y el resultado se ve claramente perjudicado por ello.

Pero no siempre conviene eliminarlos. Veamos por ejemplo el análisis de consumo por tarjeta de crédito. Para realizar la detección de compras fraudulentas, los eventos raros pueden ser más interesantes que los regulares. Incluso, en general, cuando realizamos muchas compras en un mismo día y alguna de ellas de un monto más elevado que el usual suele suceder que nos bloquean la tarjeta.

Valores faltantes

La existencia de valores faltantes puede conducir a resultados poco precisos. Pero, es necesario primero reflexionar respecto al significado de esos valores faltantes antes de tomar una decisión sobre cómo tratarlos.

Sus causas pueden ser muy diversas: mal funcionamiento de un dispositivo, falta de motivación de un vendedor o problemas de integración, entre muchos otros.

Por lo tanto, es altamente probable que nos encontremos con valores faltantes en nuestro set de datos, que pueden ser reemplazados por varias razones. Primero, porque el método de minería que utilicemos puede no tratar bien a los campos faltantes, y hasta incluso puede no funcionar con este tipo de datos. Además, si necesitáramos realizar algún tipo de agregación, por ejemplo, realizar un promedio, dependiendo de la proporción de valores nulos la métrica no sería representativa. Finalmente, en caso de que, el método utilizado funcione con valores nulos, podría estar realizando internamente algún tipo de imputación y que quizás no sea la más adecuada porque no analiza el contexto del dato al momento de imputarlo.

Selección de atributos relevantes

Seleccionar los atributos relevantes es uno de los preprocesamientos más importantes, ya que es crucial que los atributos utilizados sean relevantes para la tarea de minería. En el mejor de los mundos podríamos construir un modelo utilizando todos los datos disponibles, pero esto implicaría mucho tiempo de procesamiento, así como un costo computacional importante. Por lo que, reducir la dimensionalidad es de gran ayuda a las restricciones en capacidades de procesamiento que pueden existir.

Por lo tanto, cuando pensamos en la selección de variables, tendríamos 3 etapas inicialmente:

Recopilar e integrar las fuentes de datos existentes

De todas las fuentes de datos que tiene la compañía, en relación al objeto de nuestro análisis, tenemos que realizar la integración de estos datos. Para ello, tenemos que llegar a tener un set de datos al cual aplicar nuestro algoritmo seleccionado. Pero que tengamos infinitas fuentes de datos, no implica que vayamos a utilizarlas todas.

Por lo tanto, a la hora de recopilar la información e integrarla necesitaremos de alguien del negocio con el criterio suficiente como para guiarnos en esta tarea de seleccionar las fuentes de datos relevantes para el análisis.

Identificar y seleccionar las variables relevantes en los datos

Una vez realizada la integración, puede suceder que la integración nos llevó a generar campos sin identidad propia. Pueden ser, por ejemplo, columnas que tienen como único valor un *null* o que el mismo valor se repite para todos los registros de nuestro set de datos. Si lo pensamos bien, una columna que tiene los valores antes mencionados es innecesaria ya que no muestra ningún tipo de variación respecto del target.

Limpieza de Datos

Esta tarea forma parte de la etapa de procesamiento de datos. La razón de por qué realizamos este proceso puede surgir de la base de datos que analizamos, como vimos en el punto anterior, o de la integración de diferentes fuentes de datos. De dicha integración pueden resultar:

- Valores atípicos o extremos
- Valores faltantes
- Valores erróneos

Dentro de las tareas de limpieza tendremos:

- Detectar y tratar la presencia de valores atípicos (outliers)
- Imputar la información faltante o valores perdidos (missing)
- Eliminar datos erróneos o irrelevantes

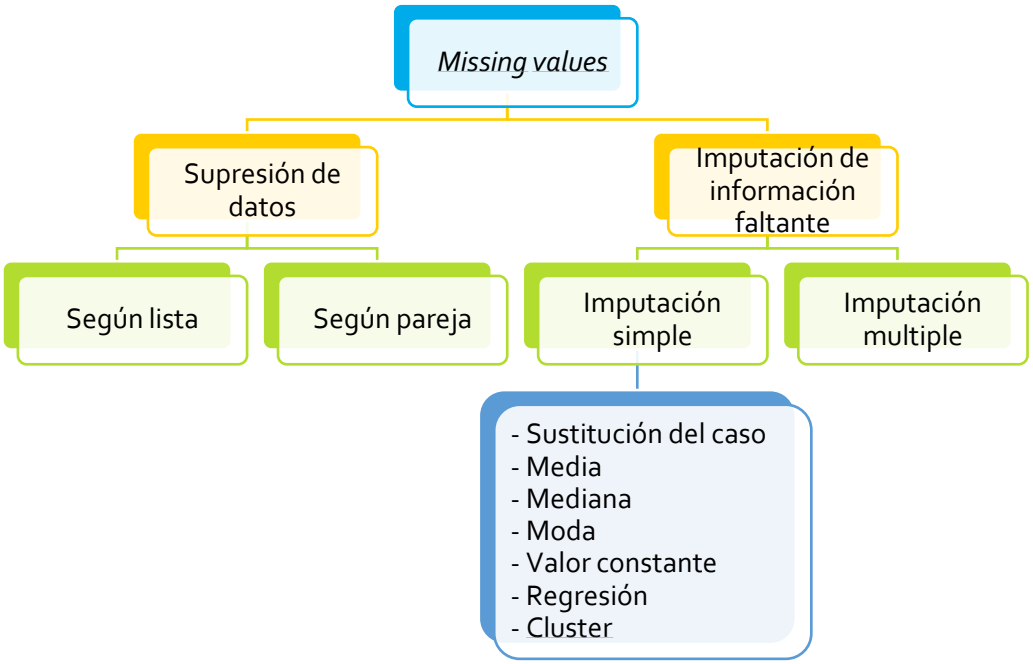
Tratamiento de outliers:

Como sabremos o podemos imaginarnos a esta altura, los valores extremos son justamente los que se destacan por no acompañar al comportamiento del conjunto de datos en general. Es por ello que, cuando se presentan en nuestro set de datos valores extremos, tenemos las siguientes posibilidades de tratamiento:

- Desecharlos no es una opción porque podemos perder información muy valiosa, valores extremos genuinos.
- Marcarlos es la opción más segura, para poder controlar si los valores extremos tienen algún efecto de correlación sobre el target.
- Aplicarles una transformación para cambiar la escala por medio de un logaritmo. No es aconsejable porque no tiene usualmente un gran efecto.

Tratamiento de Valores faltantes

Cuando aparecen en nuestro dataset valores que no tienen datos, los que conocemos como missing values, tenemos las siguientes posibilidades de comportamiento frente a ellos:



Teniendo estas opciones de tratamiento de outliers, y las que irán surgiendo a medida que pase el tiempo, lo importante es tener criterio. Nuestro criterio nos dirá si conviene suprimir el registro o suprimir la columna, y en caso de imputación hasta donde nos conviene llegar con las imputaciones.

Finalmente, ¿qué hacemos con los registros que no tienen valores para ciertos atributos/variables?

- Ignorar el registro
- Utilizar una constante global para rellenar el valor nulo
- Utilizar el valor de la media u otra medida de centralidad para rellenar el valor
- Utilizar la media u otra medida de centralidad de los objetos que pertenecen a la misma clase
- Utilizar alguna herramienta de DM para imputar el valor más probable

Integración y transformación de datos

Las integraciones de distintas fuentes deben realizarse a conciencia, porque pueden generar datos faltantes, para los cuales no existen soluciones fáciles, o duplicados que deben ser detectados durante la integración.

La integración es un proceso que se realiza generalmente durante la recopilación de datos, y si se genera un *data warehouse*, se realza en el proceso de ETL.

Esclarecimiento de identidad

Uno de los problemas que surge con la integración de diferentes fuentes de datos es **identificar objetos**, es decir, conseguir qué datos sobre el mismo objeto se unifiquen y qué datos de diferentes objetos permanezcan separados. Esto se conoce como el **problema de esclarecimiento de la identidad**.

Errores que ocurren a partir de esto:

- Se identifican dos o más objetos diferentes.
- Se multiplican dos o más objetos iguales.

Unificación de formatos

También puede suceder que a la hora de la integración nos encontremos con formatos diferentes o que los orígenes de datos hacer que sus medidas impliquen valores o interpretaciones diferentes. Para ayudar a la integración tenemos 2 posibilidades:

- Unificar los formatos
- Unificar las medidas

Reducción de dimensionalidad

La alta dimensionalidad puede causar que los patrones extraídos por los algoritmos a la hora de aprender sean caprichosos y poco robustos. Otro problema es en relación a lo que existe entre la eficiencia y la posibilidad de representar los datos, ya que en principio sólo podemos representar tres dimensiones.

Algunos métodos de reducción

Existen múltiples opciones para la reducción de dimensionalidad, para cada una de las temáticas siguientes enumeramos algunas posibilidades:

Valores redundantes:

Algunas soluciones:

- Eliminar algunas variables
- Crear variables compuestas a partir de las redundantes
- Crear combinaciones lineales ponderadas de las variables, considerando la mayor parte de la variabilidad de los datos.

Variables Irrelevantes:

- Son los que no tienen asociación con el target
- Cuando no hay una correlación lineal con el target
- Cuando el conjunto de datos tiene demasiadas dimensiones

Una buena estrategia es primero reducir la redundancia y luego la irrelevancia.

Selección de variables:

Existen abundantes métodos de selección de variables, varios de ellos son:

- Búsqueda de un subconjunto óptimo de atributos (Métodos Heurísticos)
    - Forward
    - Backward
    - Stepwise
    - Criterio de corte del árbol de decisión
- Criterio "experto"
- PCA

Muestreo

Son conjuntos de técnicas estadísticas que estudian la forma de seleccionar una muestra lo suficientemente representativa de una población cuya información permita inferir las propiedades o características de toda la población cometiendo un error medible y acotable.

Las estimaciones se realizan a través de funciones matemáticas de la muestra denominadas estimadores.

En este ámbito se utilizan técnicas de muestreo porque es muy costoso procesar los datos de toda la población. Y, por otro lado, porque a la hora de crear un algoritmo, en general partimos el set de datos de estudio en 2 porciones que deben tener las mismas características, resultando así ser 2 muestras, para poder generar el aprendizaje del algoritmo (training) y validar su funcionamiento (testing).

Los tipos de muestreo son:

Probabilístico, es una técnica en la cual las muestras son recogidas mediante un proceso que les brinda a todos los individuos de la población la misma oportunidad de ser seleccionados.

No probabilístico, es una técnica de muestreo donde las muestras se recogen por medio de un proceso que no les brinda a todos los individuos de la población las mismas oportunidades de ser seleccionados.

Estos métodos de muestro se utilizan para la **inferencia estadística**, que es la metodología que permite inferir resultados, predicciones y generalizaciones sobre la población estadística, basándose en la información contenida en las muestras representativas.

Para medir el grado de representatividad de la muestra es necesario utilizar **muestreo probabilístico**. El muestreo es probabilístico cuando puede establecerse la probabilidad de obtener cada una de las muestras que se posible seleccionar.

Tipos de muestreo probabilístico:

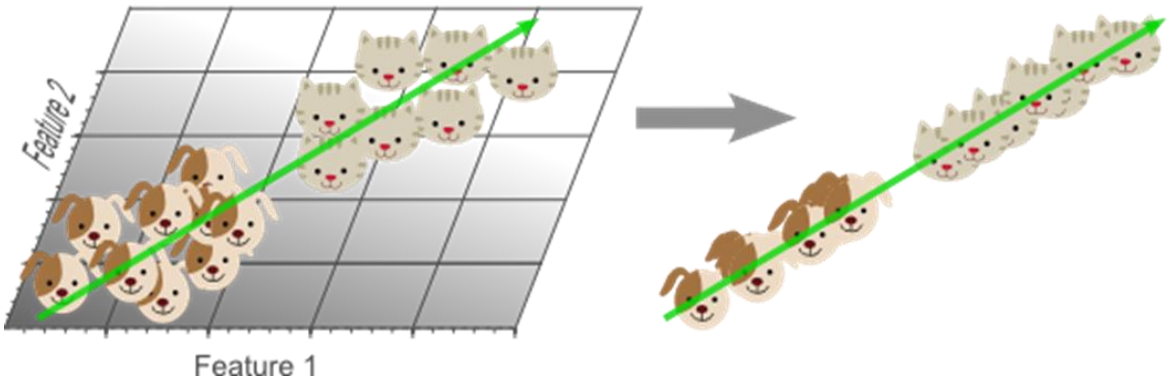
- Muestreo Aleatorio Simple: es un procedimiento de selección con probabilidades iguales que consiste en obtener la muestra unidad a unidad de forma aleatoria con reposición a la población de las unidades previamente seleccionadas.
- Muestreo Estratificado: se obtiene una muestra de tamaño *n* seleccionando *nh* elementos de cada uno de los L estratos en los que se subdivide la población
- Muestreo Sistemático: los elementos de la población se ponen en una lista y luego cada enésimo elemento de la lista se selecciona sistemáticamente para su inclusión en la muestra.
- Muestreo por Conglomerados: puede ser utilizado cuando es imposible o impráctico elaborar una lista exhaustiva de los elementos que constituyen a la población objetivo. Sin embargo, generalmente lo elementos de la población ya están agrupados en subpoblaciones y las listas de esas subpoblaciones ya existen o pueden ser creadas.

PCA

Es la técnica más tradicional, conocida y eficiente para reducir la dimensionalidad por transformación. Es también algo difícil de comprender en un curso inicial, pero por lo menos vamos a conocer de qué se trata.

De manera muy simplificada, se trata de tomar las variables originales y transformarlas de tal modo que se vea geográficamente como un cambio de ejes en la representación. Pero lo sorprendente es que los atributos que se generan resultan independientes entre sí, los nuevos atributos no pueden superar a los iniciales, y además se encuentran ordenados por relevancia. Todo esto nos asegura que, si ignoramos los últimos *k* atributos, estaremos descartando la información menos relevante.

A modo de ejemplo, supongamos que tenemos que usar PCA en el gráfico de 2 dimensiones y lo llevamos a 1 dimensión:



Ahí podemos ver que la misma diferenciación que hubiéramos hecho con 2 dimensiones podemos hacerla con 1 sola dimensión, gracias a PCA.

Usos de PCA

Este tipo de algoritmos tiene múltiples usos:

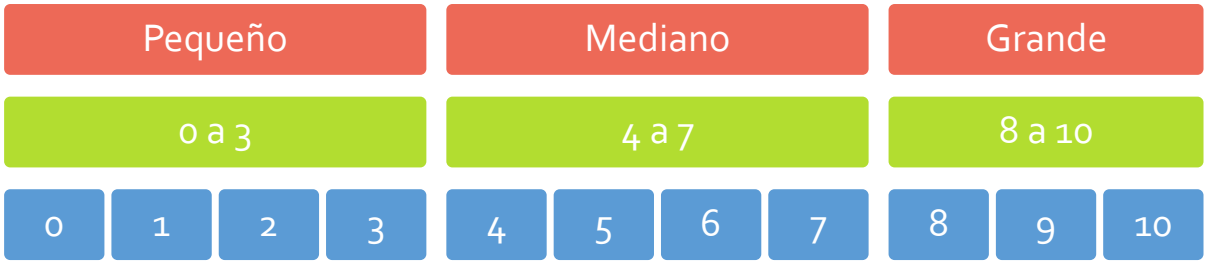
- Predicción
- Clasificación
- Detección de *outliers*
- Selección de variables

Discretización de variables

La discretización es un proceso que permite separar una serie de variables cualitativas o cuantitativas en clases. De modo que se simplifiquen las características de esta variable. Para ello destacamos 2 procesos de discretización:

Normalizar con Binning

Es conocido también como *binning*, que es la conversión de un valor numérico en un valor nominal ordenado. A modo de ejemplo, convertir en intervalo el atributo tamaño:



Puede, y en general suele suceder, que se mantiene el orden en que se generaron los intervalos. Pero puede suceder que se olvide el orden y se trate el atributo como un valor nominal sin orden.

La discretización se realiza cuando el error en la medida puede ser grande o existen ciertos umbrales significativos. Por ejemplo, que la diferencia mínima en una nota implique aprobar o reprobar un examen. Otra razón puede ser por escalas diferentes para un mismo concepto, siguiendo el ejemplo anterior, aprobado en El Salvador requiere un 6, en España un 5 y Argentina un 4. Para integrarlas sería preferible tener una escala común. Por último, cuando tenemos algunos atributos nominales y otros numéricos, y queremos que todos sean nominales para, por ejemplo, establecer reglas de asociación.

Binning por medio de clusters

Cuando no podemos generar una clase a simple vista, como en el punto anterior, podemos generar un cluster que nos genere una segmentación y con ella podamos asignar una clase que nos permita discretizarlas. A modo de ejemplo tenemos:

X	N_i	$N1_i$	$N0_i$	p_i	$\log(p_i/(1-p_i))$
J	5	4	1	0,80	0,60
I	12	6	6	0,50	0,00
B	970	432	538	0,45	-0,10
F	50	20	30	0,40	-0,18
G	23	8	15	0,35	-0,27
D	111	36	75	0,32	-0,32
H	17	5	12	0,29	-0,38
A	1564	441	1123	0,28	-0,41
E	85	23	62	0,27	-0,43
C	223	45	178	0,20	-0,60

X	N_i	$N1_i$	$N0_i$	p_i	$\log(p_i/(1-p_i))$
CL1	987	442	545	0,45	-0,09
CL2	184	64	120	0,35	-0,27
CL3	1666	469	1197	0,28	-0,41
CL4	223	45	178	0,20	-0,60

Practica modulo 2

Para realizar la práctica deberás entrar a GitHub y luego podes descargarte los archivos y trabajarlos en tu maquina o ir a https://mybinder.org y ejecutar Jupyter notebook desde ahí.

Test teórico

Ingresar al siguiente link e ingresar el código que aparece en pantalla: https://kahoot.it/

MODULO 3

Introducción estadística

Cuando hablamos del concepto de la Estadística nos referimos principalmente a una de las tantas ramas de la Matemática, pues, se encarga de analizar y estudiar datos, y también buscar las explicaciones de algunos fenómenos que alteran los resultados.

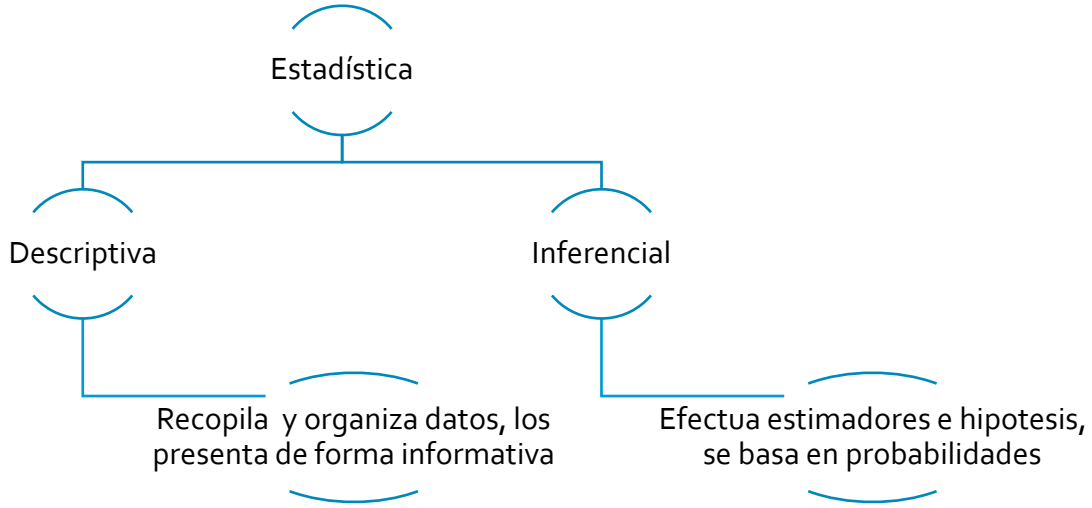
Por otro lado, también se considera una ciencia, ya que estudia a las poblaciones de forma específica, recopilando los diferentes datos para determinar algún problema o para darle solución.

Refrescamos algunos términos del módulo anterior:

- Población: La población en estadística es la recolección de un conjunto, elementos, artículos o sujetos que gozan de características comunes con el fin de estudiarlos y de esta forma se sacar conclusiones específicas para determinar sus resultados.
- Individuo: Cada uno de los elementos
- Muestra: Conjunto representativo, adecuado y válido de la población

Tipos de estadística

Dentro de los tipos de estadísticas existentes al momento, los que necesitamos conocer son:



Estadística Descriptiva o deductiva

Se encarga de recolectar, clasificar, ordenar, analizar y representar datos para obtener las características del grupo. Para ello se utiliza herramientas:

- Medidas de tendencia central:
  - Media, moda, mediana
- Medidas de dispersión:
  - Varianza, desvío estándar
- Medidas de forma:
  - Coeficiente de Pearson

Estadística Inferencial o inductiva

No se remite a la mera descripción, sino que va más allá. Trata de inferir características generales de una población a partir de prueba realizadas a una muestra de esta. Las principales características son:

- Infiere conclusiones generales
- Permite tomar previsiones
- Permite predecir el comportamiento de ciertos fenómenos
- Se apoya en la estadística descriptiva y en la probabilidad.

Entre las principales herramientas se encuentran:

- Contraste de hipótesis
- Intervalos de confianza
- Errores tipo 1 y tipo 2
- Teorema central de límite

Estás herramientas superan el contenido de este curso, pero los mencionaremos, así como a su implicancia para avanzar con nuestro objetivo de construir un árbol.

Herramientas estadísticas

Media aritmética o promedio:

Suma de todos los valores posibles de la variable, dividido por la cantidad total de casos.

Mediana:

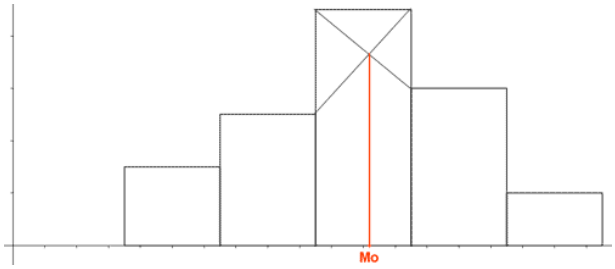
Es el punto medio de una distribución, es decir, es el número tal que la mitad de las observaciones son menores que ella y la otra mitad resultan mayores.

Moda:

Es el valor que tiene más frecuencia en un conjunto de datos. Si algunos valores tienen una frecuencia igual, cada uno representa una moda. Puede ser calculada para datos cualitativos. En caso de una variable continua será la clase modal o para hacerlo de una forma más exacta utilizamos la siguiente fórmula:

$$Mo = L_i + \frac{f_{Mo} - f_{Mo-1}}{(f_{Mo} - f_{Mo-1}) + (f_{Mo} - f_{Mo+1})} \cdot c$$

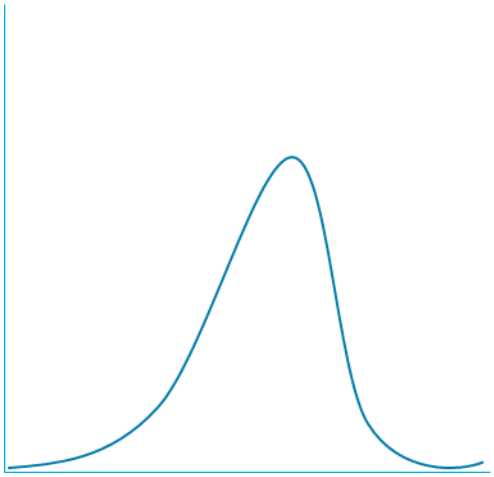
L_i : límite inferior de la clase modal
c: amplitud del intevalo de la variable estadística
f_{Mo}, f_{Mo+1}, f_{Mo-1} son, respectivamente, las frecuencias absolutas de la clase modal, la clase anterior y la posterior



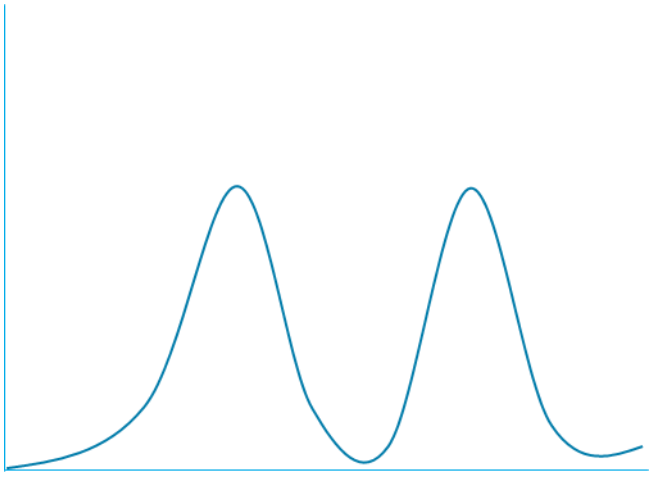
Diversidad modal

Es usual que una variable no tenga un único valor para la moda, sino dos o incluso más. A modo de ejemplo graficamos esta diversidad:

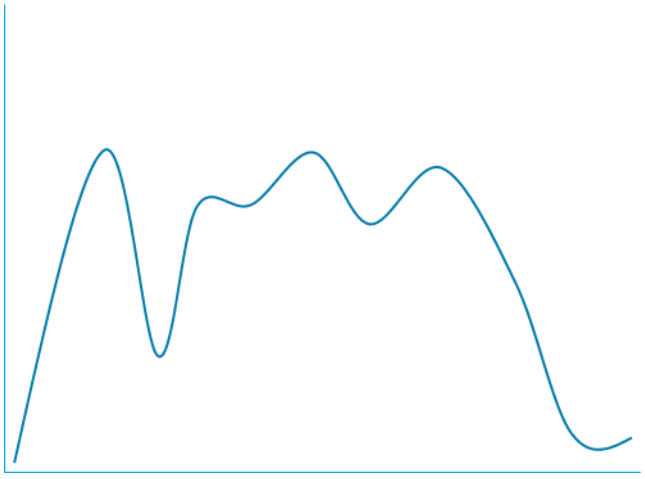
- Unimodal



- Bimodal



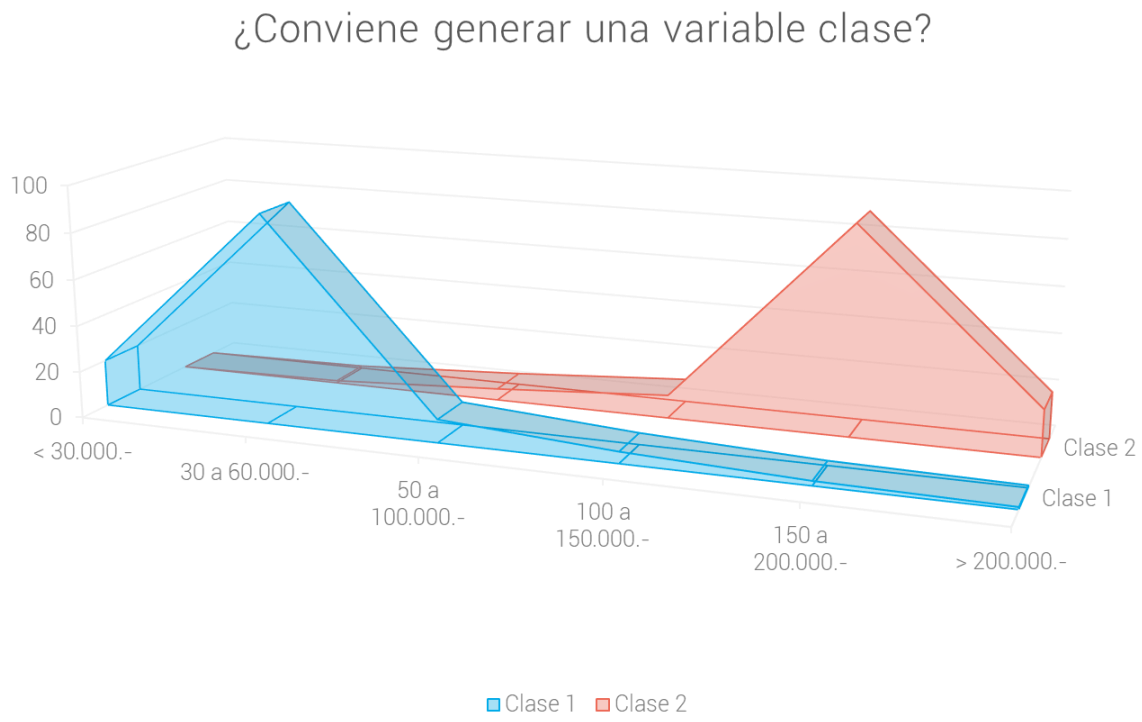
- Multimodal



Distribución Bimodal

En el caso de la distribución bimodal tendremos que analizar el comportamiento de dichas variables en relación con la variable target, porque puede suceder que se trate de dos tipos de comportamiento.

Por ejemplo, si estuviéramos prediciendo qué artículos se van a demandar más de un supermercado y nuestro set de datos incluye el salario de los individuos de la ciudad. Los rangos salariales son indicadores puros en relación al tipo de producto, de lujo o de necesidad, que pueda ofrecer mi supermercado.



Cuartiles:

Valores que dividen a los datos "ordenados" en cuartos.

- Q1: es el valor de la variable tal que el 25% de las observaciones son menores que él y el 75% resultan mayores.
- Q2: coincide con la Mediana.
- Q3: es el valor de la variable tal que el 75% de las observaciones son menores que él y el 25% resultan mayores.

Varianza y Desvío Estándar

Se conoce como varianza a la raíz cuadrada que se desprende de una desviación estándar, la cual permite que las industrias de manufactura encuentren precisión en el trabajo y producción y, al mismo tiempo, reduzcan el índice de errores.

Esto sucede ya que la varianza toma los datos dispersos de la media y luego de medirlos le da valor a las variaciones y a las desviaciones y también contabiliza y asume los errores cometidos previniendo posibles errores.

Cálculo de la varianza:

$$\sigma 2 = \frac{(X_1 - X_m)^2 + (X_2 - X_m)^2 + (X_3 - X_m)^2 + \ldots + (X_n - X_m)^2}{n}$$

La varianza es representada por "σ^2" (una letra griega sigma y elevada al cuadrado) y se hace el cálculo con la forma ya descrita.

El valor de Xm, es obtenido a través de la media aritmética o promedio de los valores a analizar. Mientras que Xn se obtiene a través del valor a analizar.

Desviación típica:

La desviación típica de una variable estadística es la raíz cuadrada positiva de la varianza.

- La desviación típica es el parámetro de dispersión más utilizado.
- Si se suma una constante a todos los valores de la variable, la desviación típica no varía.
- Si se multiplican todos los valores de la variable por un mismo número, la desviación típica queda multiplicada por el mismo número.

Medidas de dispersión:

- Rango:
  - Es la diferencia entre los valores más grandes y más pequeños en un conjunto de datos.
- Variancia (S^2):
  - Mide la dispersión de los datos con respecto a la media. Está medida en la misma unidad que la variable original pero elevada al cuadrado.
- Desvío Estándar (S):
  - Está medida en la misma unidad de medida que las variables originales. Este es un motivo por el cual se prefiere trabajar con S en lugar de S^2.
  - S = o solamente cuando no hay dispersión. Esto ocurre únicamente cuando todas las observaciones toman el mismo valor. En caso contrario S > o.
  - Igual que ocurre con la Media Aritmética, S no es robusta: fuertes asimetrías o unas pocas observaciones atípicas pueden hacer que aumente mucho S.
- Rango intercuartílico:
  - Mide la dispersión del 50% de los valores centrales. Se calcula como la diferencia entre el tercer y primer cuartil, denotado por RI=Q_3 -Q_1.
- MAD: Toma un conjunto de datos a partir de la mediana.

Análisis exploratorio

Cuando buscamos extraer conocimiento de una base de datos, las preguntas usuales que pueden surgir son:

- ¿Qué parte de los datos es pertinente analizar?
- ¿Qué tipo de conocimiento se desea extraer y cómo se debe presentar?
- ¿Qué conocimiento puede ser válido, novedoso e interesante?

- ¿Qué conocimiento previo me hace falta para realizar esta tarea?

Del mismo modo que nosotros necesitamos poder responder a estas preguntas para avanzar en la extracción de conocimiento, una herramienta de minería no puede digerir un conjunto de datos y producir algo razonable, si no se la *orienta*. En parte, porque las herramientas por ahora no tienen la capacidad de realizar algunas tareas de manera completamente automática.

Sucede, además, que incluso conociendo los datos y el dominio del que provienen, responder a algunas de las preguntas mencionadas no es algo sencillo. Es necesario, en muchos casos, *explorar* los datos, el contexto y los usuarios de la información.

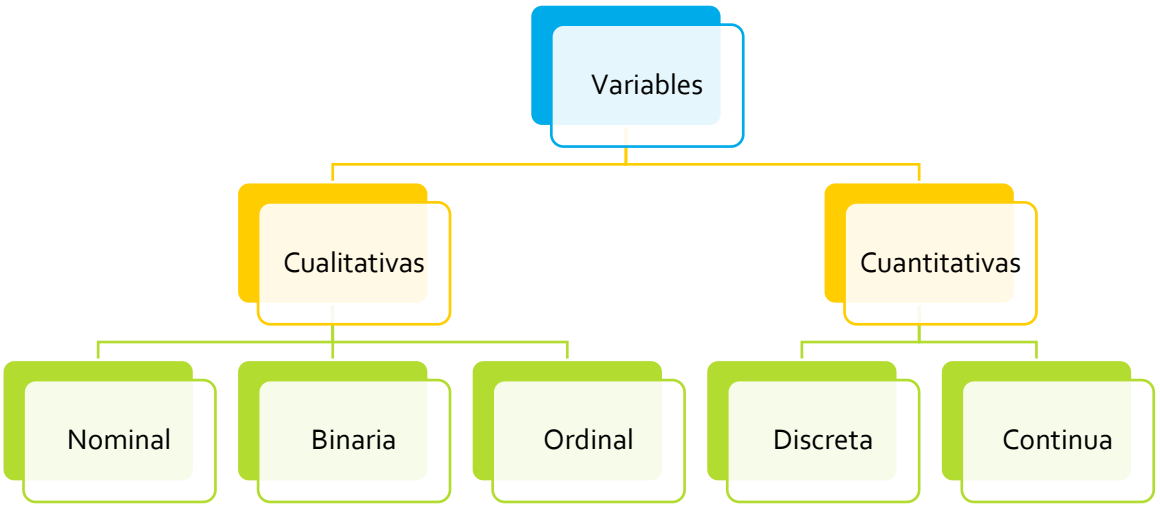
Por lo tanto, es necesario generar una vista minable, una tabla con todos los atributos relevantes para realizar la tares, acompañada de la tarea a realizar sobre ella y cómo evaluarla, así como la forma de presentar el resultado final, y en su caso, el conocimiento previo necesario.

Tipos de variables

La variable estadística se refiere a una característica o cualidad de un individuo que está propenso a adquirir diferentes valores. Estos valores se caracterizan por poder medirse.

Por ejemplo, el color de pelo de una persona, las notas de un examen, sexo, estatura de una persona, etc.

Los tipos de variable estadística se dividen de acuerdo a las características que la definan, entre ellas podemos encontrar los siguientes tipos.



Cualitativas

Las variables cualitativas son aquellas características o cualidades que no pueden ser calculadas con números, sino que lo hacen con palabras.

Este tipo de variable, a su vez se divide en las siguientes:

- **Cualitativa nominal**: Aquellas variables que no siguen ningún orden en específico.
  Por ejemplo: Colores (Negro, Naranja, Amarillo).
- **Cualitativa ordinal**: Aquellas que siguen un orden o jerarquía.
  Por ejemplo: Nivel socioeconómico (Alto, medio, bajo).
- **Cualitativa binaria**: En este caso, las variables son solamente dos.
  Por ejemplo: Si o No, Hombre o Mujer.

Cuantitativas

Las variables cuantitativas son aquellas características o cualidades que sí pueden expresarse y medirse a través de números.

Este tipo de variable a su vez se divide en:

- **Cuantitativa discreta**: Aquella variable que usa valores enteros y no finitos.
  Por ejemplo: La cantidad de familiares que tiene una persona (2, 3, 4 ó más).
- **Cuantitativa continua**: Aquella variable que utiliza valores finitos y objetivos. Suele caracterizarse por utilizar valores decimales.
  Por ejemplo: El peso de una persona (64.3 Kg, 72.3 Kg, etc).

Exploración visual

Son métodos que nos permiten a simple vista poder observar los datos. Estas gráficas, en general se centran en uno o dos atributos a lo sumo, y el objetivo principal es en esta esta etapa su colaboración con la limpieza y transformación de variables.

A modo de aclaración, se nos podrían presentar 2 tipos de visualización:

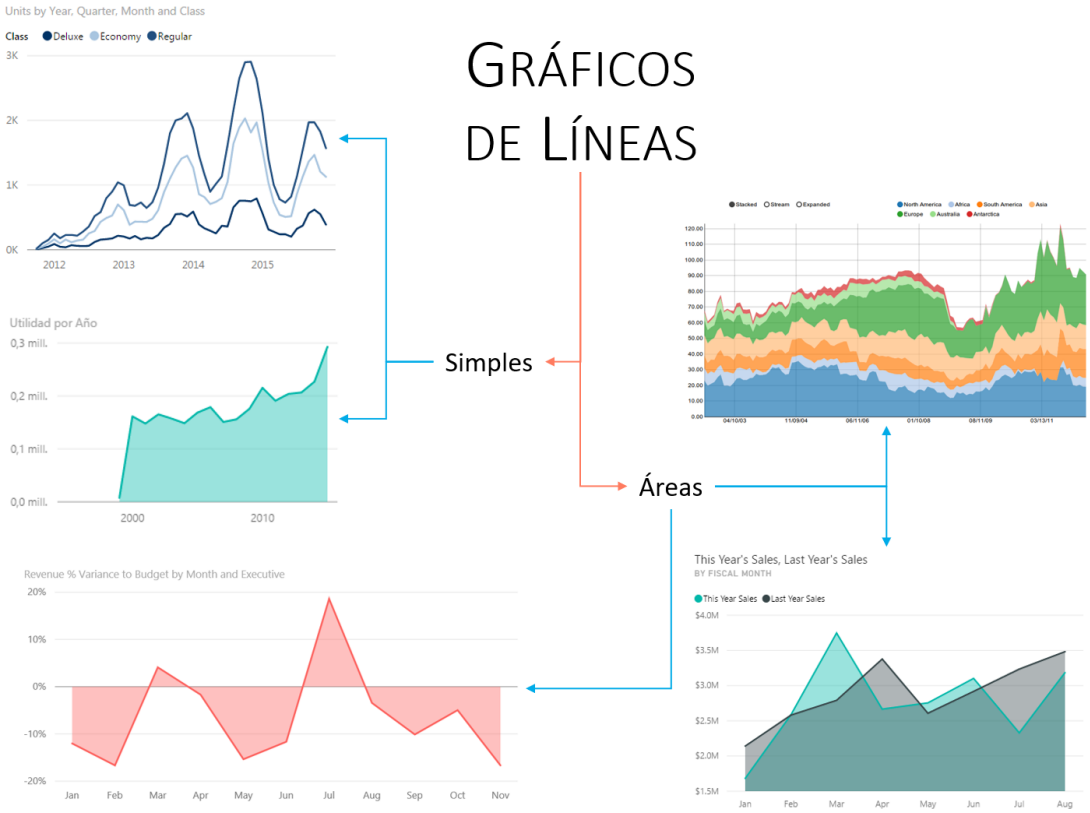
- Visualización previa:

  Se utiliza para entender mejor los datos y sugerir posibles patrones o tipos de herramientas de KDD a utilizar. La visualización previa se utiliza para ver tendencias y resúmenes de los datos, así como encontrar evidencia en relación al objetivo planteado.

- Visualización posterior:
  Se realiza después de la minería para mostrar los patrones y entenderlos mejor. Se utiliza frecuentemente para validar y mostrar a los expertos los resultados de la extracción de conocimiento.

En este apartado veremos, que los gráficos se utilizan de acuerdo al tipo de variable que estamos tratando de representar. Por el hecho de existir una abundancia importante de gráficos disponibles, así como mejores prácticas en relación a su utilización, vamos a dar un vistazo a algunos de ellos:

También, dentro de estos grupos existen variedades. Como puede ser en relación a gráficos de líneas o áreas, con los que se pueden representar variables continuas:

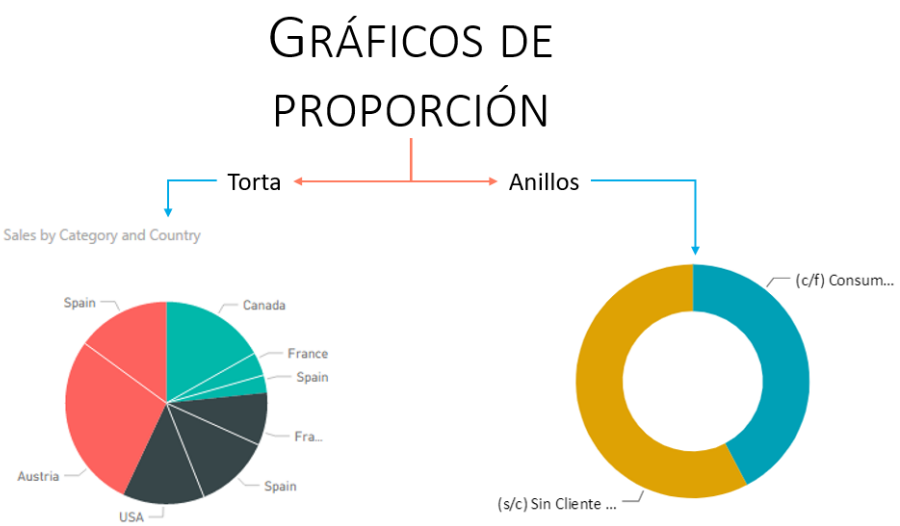


Diagramas de barras

Gráficos de barras para variables categóricas o discretas:

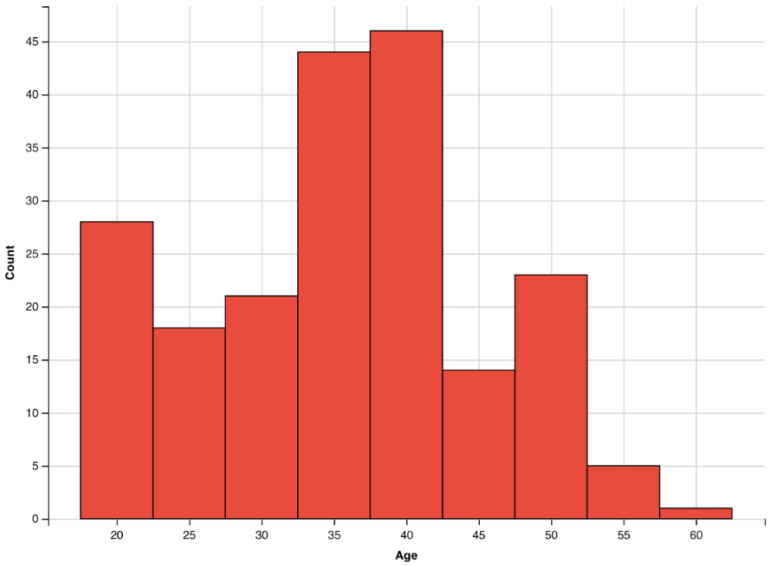
Gráficos combinados

Diagramas de sectores

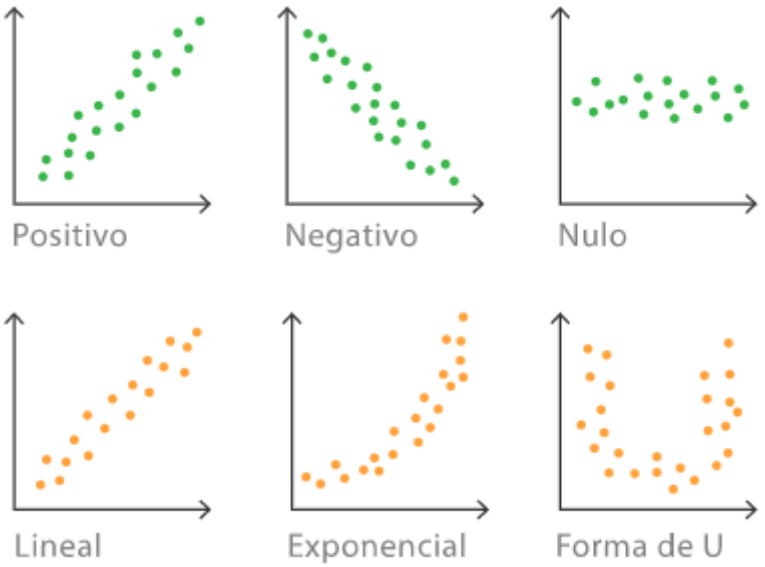


Estos gráficos muestran proporciones entre categorías.
Necesitan una variable dimensión y una métrica.

Histograma



Dispersión



Fuerza de Correlación

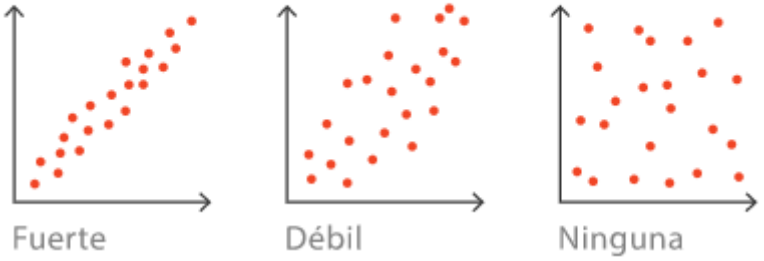
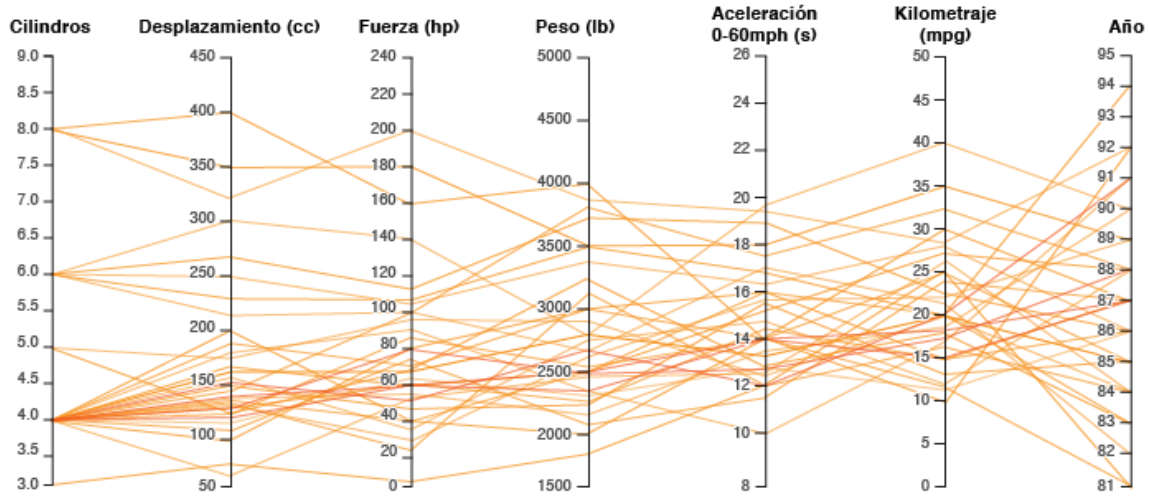
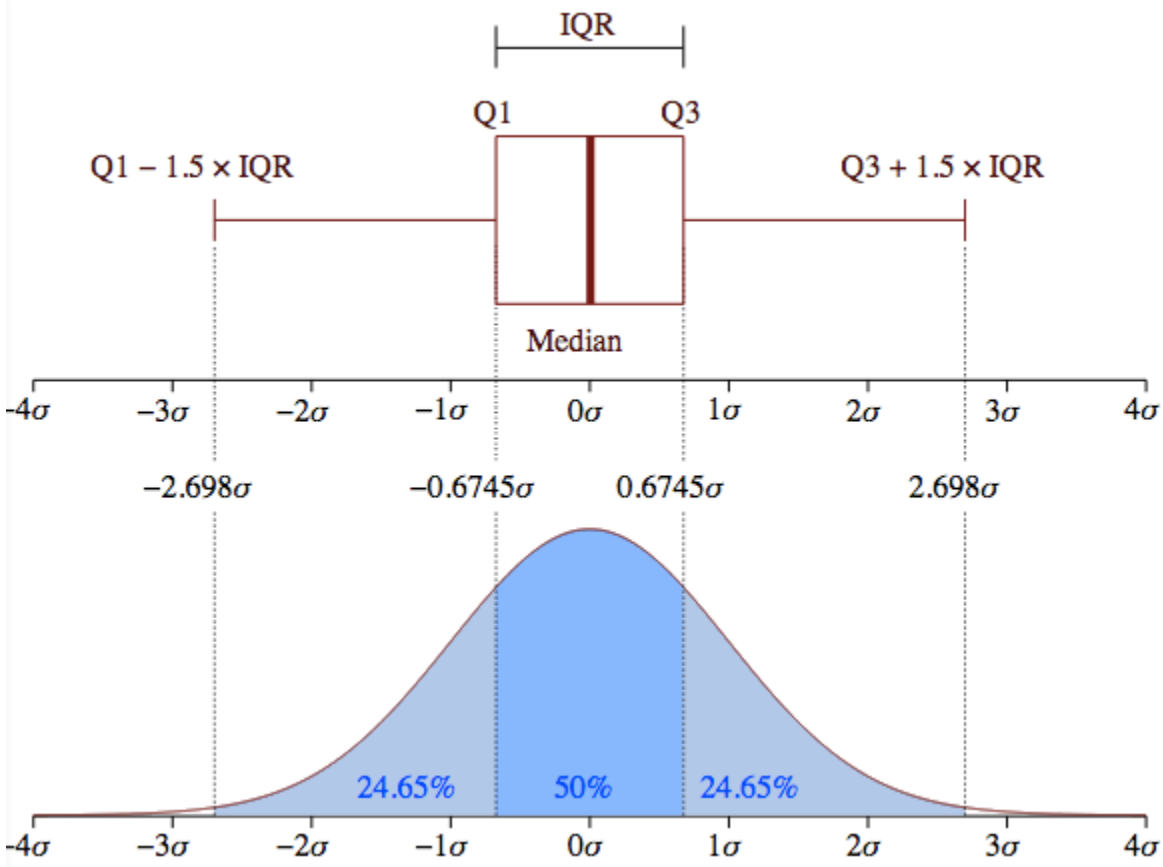


Gráfico de coordenadas paralelas



Diagramas de caja o Boxplot



Características de la distribución de las variables

Exploración Formal

Uso de estadísticos descriptivos como media o desvío en caso de distribución SIMÉTRICA Y UNIMODAL. Para casos más complejos se aconseja usar medidas más robustas. En general las distribuciones no suelen ser normales, ni simétricas. Pueden tener distintas modas o con problemas de curtosis. *Por lo que realizaremos transformaciones*.

Medidas de Forma

- Coeficiente de Simetría:
  Con los coeficientes de asimetría se trata de medir si las observaciones están dispuestas simétrica o asimétricamente respecto a un valor central (en general, la media aritmética) y el grado de esta asimetría.

- Coeficiente de Curtosis:
  Con el coeficiente de apuntamiento o curtosis se trata de medir el grado de apuntamiento de una distribución respecto a la distribución normal, que se toma como patrón, y cuyo coeficiente de curtosis es 0. La distribución normal es la más importante, tanto en la teoría de la probabilidad como en le práctica de los trabajos estadísticos. Se caracteriza por ser simétrica respecto al eje x. Su función de densidad, en la práctica histograma de frecuencias, tiene forma de campana.

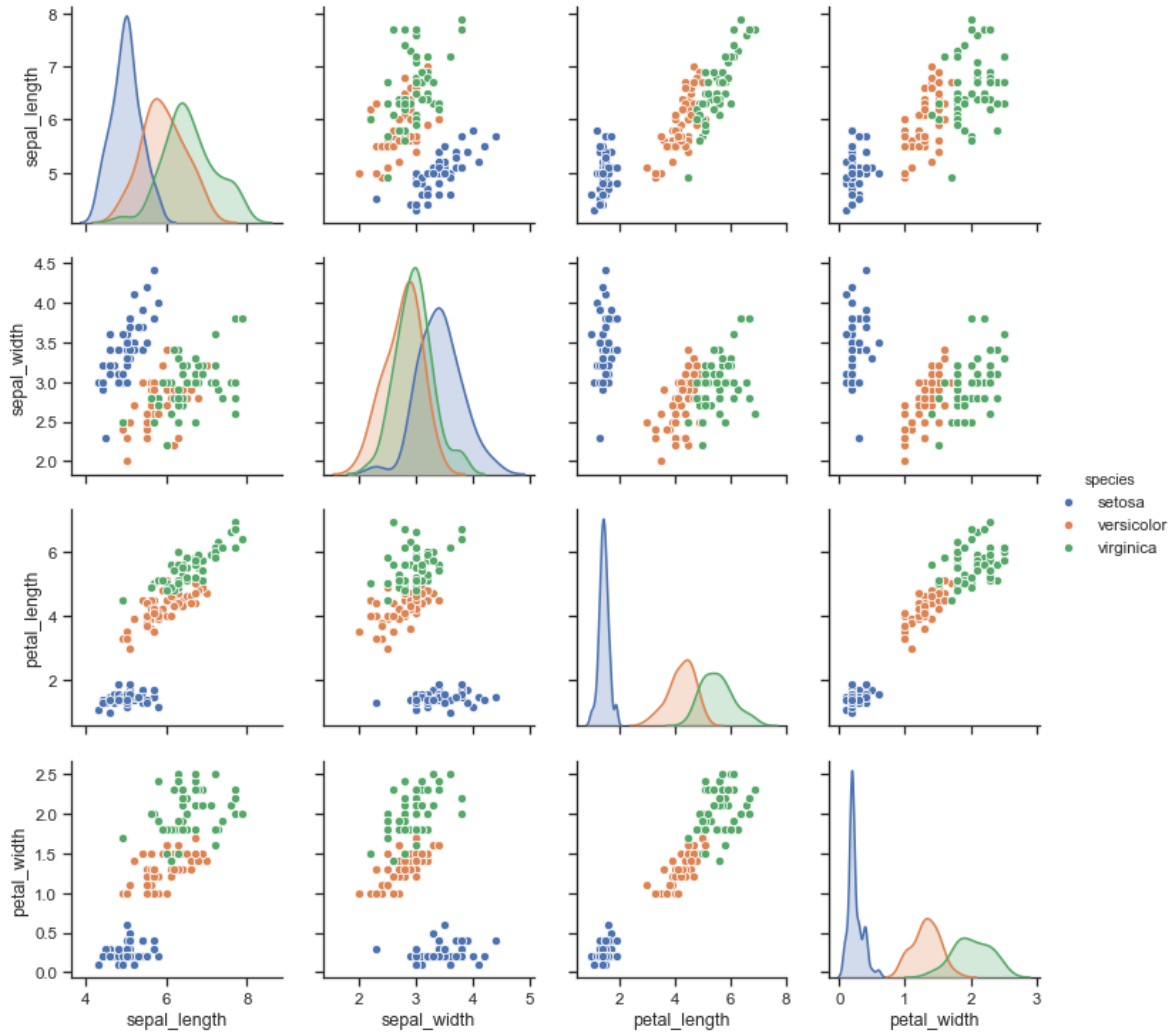
  La distribución es leptocúrtica (más apuntada que la normal), y platicúrtica (menos apuntada que la normal).

Correlación

El análisis de correlación es una tarea descriptiva que se puede usar para distinguir el grado de similitud entre los valores de dos variables numéricas. A modo de representar las relaciones entre las variables, se puede utilizar el gráfico de dispersión. Ahí pudimos observar que entre diferentes variables puede existir una relación, positiva o negativa, y en caso de no existir, ser indiferentes. Los valores de este coeficiente van a ir de 1 a -1, donde el cero muestra la independencia absoluta de las variables.

Este tipo de análisis de correlación, sobre todo las negativas, puede ser muy útil para establecer reglas de ítems correlacionados, como se puede observar en el siguiente ejemplo. *Un inspector de incendios que desea obtener información útil para la prevención de incendios probablemente está interesado en conocer correlaciones negativas entre el empleo de distintos grosores de protección del material eléctrico y la frecuencia de ocurrencia de incendios.*

Para poder avanzar en el análisis necesitamos crear una matriz de correlaciones que luego será lo que observamos en el gráfico de dispersión para múltiples variables o gráfico matriz, como el siguiente:



Transformación y análisis de las mejoras

Recodificación 1 a 1

- Enumeración aleatoria → Es la peor técnica, no recomendable!!!

- Codificación con variables indicadores (dummy).
  - Enfoque típico desde el punto de vista estadístico.
  - Se buscan niveles equidistantes que permitan ajustar la respuesta predicha del modelo para reflejar precisamente la diferencia de respuesta entre niveles.
  - Para una variable de n niveles se deben generar n o menos variables indicadoras.
  - El problema de una codificación con muchos dummies es uno de sobregeneralización.
  - En este caso primero conviene agrupar niveles antes de recodificar con variables indicadoras.
- Umbral: cuando por tener pocos valores se agrupa bajo el nombre "otros".
- Transformación basada en el target: En su forma más simple, se enumera los niveles según el promedio del target para el nivel.

Según target

Puede asignarse una categoría numérica por similitud entre los registros, y para ello podría utilizarse un cluster.

- Con un número moderado a grande de niveles, donde no se conoce la semántica de los mismos, pueden emplearse técnicas de clustering para agrupar los niveles en un número más o menos pequeño de categorías.
- Las técnicas de clustering son consideradas técnicas de reducción de datos (de variables), aunque en este contexto pueden usarse para reducir el número de niveles.

X	N_i	$N1_i$	$N0_i$	p_i	$\log(p_i/(1-p_i))$
J	5	4	1	0,80	0,60
I	12	6	6	0,50	0,00
B	970	432	538	0,45	-0,10
F	50	20	30	0,40	-0,18
G	23	8	15	0,35	-0,27
D	111	36	75	0,32	-0,32
H	17	5	12	0,29	-0,38
A	1564	441	1123	0,28	-0,41
E	85	23	62	0,27	-0,43
C	223	45	178	0,20	-0,60

X	N_i	$N1_i$	$N0_i$	p_i	$\log(p_i/(1-p_i))$
CL1	987	442	545	0,45	-0,09
CL2	184	64	120	0,35	-0,27
CL3	1666	469	1197	0,28	-0,41
CL4	223	45	178	0,20	-0,60

Documentación para una vida más tranquila

Característica de Y	Transformación	Sentencia de DATA Step
Sesgo Positivo Moderado	Raíz Cuadrada (Y)	Y_T = SQRT(Y);
Sesgo Negativo Moderado	Raíz Cuadrada (K-Y) (K=Máx(Y)+1)	Y_T = SQRT(K-Y);
Sesgo Positivo Grande	Log(Y) ó Ln(Y)	Y_T = LOG10(Y); ó Y_T = LOG(Y);
Sesgo Negativo Grande	Log(K-Y) ó Ln(K-Y)	Y_T = LOG10(K-Y); ó Y_T = LOG(K-Y);
Forma L extrema	Recíproco de Y	Y_T =1/Y;
Forma J extrema	Recíproco de (K-Y)	Y_T =1/(K-Y);
Si hay valores negativos o 0 en los datos, sumar una constante a Y antes de realizar las transformaciones recíproco o Log/Ln.		

Practica modulo 3

Para realizar la práctica deberás entrar a GitHub y luego podes descargarte los archivos y trabajarlos en tu maquina o ir a https://mybinder.org y ejecutar Jupyter notebook desde ahí.

Test teórico

Ingresar al siguiente link e ingresar el código que aparece en pantalla: https://kahoot.it/

MODULO 4

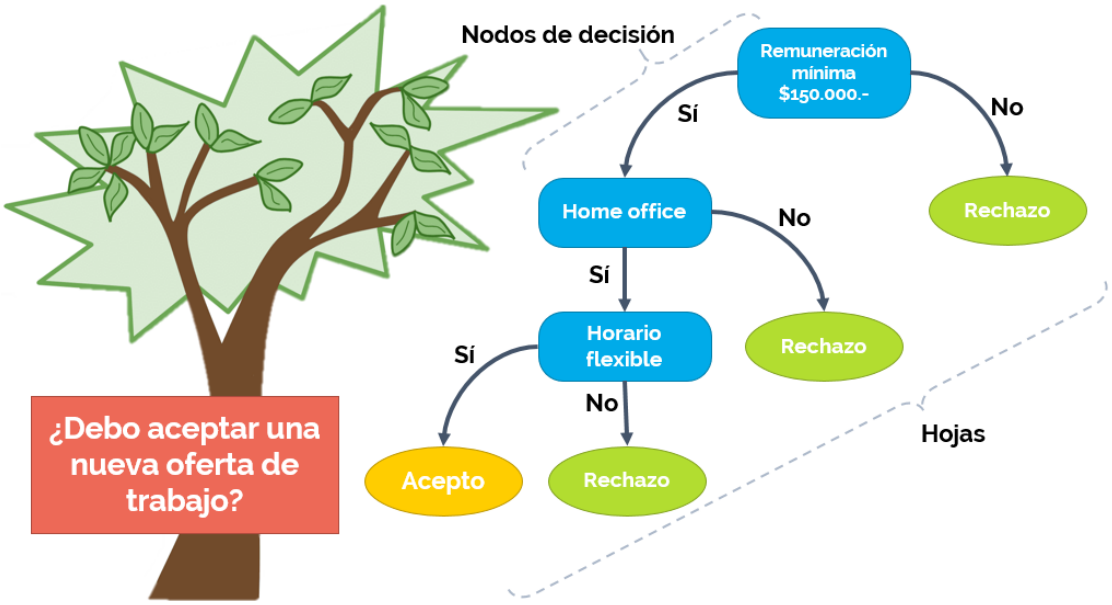
Dentro de las técnicas y métodos para el aprendizaje de *modelos comprensibles* y *proposicionales* se encuentran los árboles de decisión y los sistemas de reglas. Se los considera *comprensibles* porque se pueden expresar de una manera simbólica, en forma de conjunto de condiciones y, por lo tanto, pueden tener como resultado modelos inteligibles para los seres humanos. Son, además, *proposicionales* por ser métodos que se restringen a algoritmos que aprenden sobre una única tabla de datos y que no establecen relaciones entre más de una fila de la tabla a la vez, ni sobre más de un atributo a la vez.

Primer modelo, un árbol de decisión

Los sistemas de aprendizaje basados en árboles de decisión son quizás el método más fácil de utilizar y de entender. Un árbol de decisión es un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas.

Una de las grandes ventajas de los árboles de decisión es que, en su forma más general, las opciones posibles a partir de una determinada condición son excluyentes. Esto permite analizar una situación y, siguiendo el árbol de decisión apropiadamente, llegar a una sola acción o decisión a tomar.

Supongamos que recibimos una propuesta de trabajo, y que un modelo construido sobre un determinado set de arrojó el siguiente árbol de decisión. Como podemos observar, es sencillo aplicar el árbol de decisión a una nueva oferta de trabajo para recomendar si debe aceptarse o rechazarse la propuesta recibida. Basta con realizar las preguntas y seguir las respuestas hasta alguna de las hojas del árbol, catalogadas con un Acepto o Rechazo.



Cómo funciona

Lo que nos concierne respecto a los módulos en esta instancia es aprender a construir un árbol de decisión a partir de datos. En relación a los árboles de decisión, la característica más importante es que se asume que las clases son disjuntas. No pueden ser al mismo tiempo de clase *a* y clase *b*. En este sentido, la clasificación se diferencia de la categorización, donde se permiten más de una clase, etiqueta o categoría para cada instancia.

El espacio de instancias se iba partiendo de arriba abajo, utilizando cada vez una partición, es decir, un conjunto de condiciones excluyentes y exhaustivas. Por esta razón, estos son algoritmos de partición o de "divide y vencerás".

Tipos de árboles

Existen varios tipos de árboles, que se fueron creando desde 1983 hasta la fecha:

- CART
- ID3
- C4.5
- ASSISTANT
- C5.0
- CHAID
- QUEST, etc.

Lo que los diferencia es en relación a los puntos más importantes, que hacen que un algoritmo funcione mejor que otro:

1- Particiones que considerar
2- Criterio de selección de particiones.

A la hora de realizar splits, buscamos generar nodos hijos con la menor impureza posible. Para ello existen diferentes criterios, que mencionamos ahora, pero veremos en detalle más adelante:

- Índice de GINI
- Entropía (Ganancia de información)
- Test Chi-cuadrado
- Proporción de ganancia de información

CART

Es sinónimo de árboles de clasificación y regresión. Se caracteriza por el hecho de que construye árboles binarios, es decir, cada nodo interno tiene exactamente dos bordes salientes. Las divisiones se seleccionan utilizando los criterios de dosificación y el árbol obtenido se recorta mediante la reducción de costos y la complejidad. CART puede manejar variables numéricas y categóricas y puede manejar fácilmente valores atípicos.

Desventajas:

- Se puede dividir en una sola variable.
- Los árboles formados pueden ser inestables.

ID3

Construye un árbol de decisión para los datos dados de forma descendente, a partir de un conjunto de objetos y una especificación de propiedades Recursos e Información. En cada nodo del árbol, se prueba una propiedad en función de maximizar la ganancia de información y minimizar la entropía, y los resultados se utilizan para dividir el conjunto de objetos. Este proceso se realiza de forma recursiva hasta que el conjunto en un subárbol dado es homogéneo (es decir, contiene objetos que pertenecen a la misma categoría). El algoritmo ID3 utiliza una búsqueda codiciosa. Selecciona una prueba usando el criterio de ganancia de información, y luego nunca explora la posibilidad de opciones alternativas.

Desventajas:

- Los datos pueden ser sobre ajustados o sobre clasificados, si se analiza una pequeña muestra.
- Solo se prueba un atributo a la vez para tomar una decisión.
- No maneja atributos numéricos y valores perdidos.

C4.5

Versión mejorada del ID3. Las novedades son:

1- acepta características tanto continuas como discretas;
2- maneja puntos de datos incompletos;
3- resuelve el problema de ajuste excesivo mediante una técnica ascendente (muy inteligente) generalmente conocida como "poda";
4- se pueden aplicar diferentes pesos a las características que comprenden los datos de entrenamiento.

Desventajas:

- C4.5 construye ramas vacías con valores cero
- El ajuste excesivo ocurre cuando el modelo de algoritmo recoge datos con características poco comunes, especialmente cuando los datos son ruidosos.

Este evolucionó luego al c5.0

Resultados

Las diferencias de resultados son a causa de lo hiper-parámetros que utiliza cada uno de los algoritmos anteriores. Por ejemplo:

	Criterio Split	**Tipo de atributo**	**Nulos**	**Estrategia de poda**	**Detección de outliers**
ID3	Information gain	Sólo variables categóricas	No los soporta	No incluye poda	Susceptible a outliers
CART	Towing criterio	Categóricas y numéricas	Funciona ok	Usa costo de complejidad	Funciona ok
C4.5	Gain ratio	Categóricas y numéricas	Funciona ok	Usa costo de complejidad	Funciona ok

Aplicaremos estos conocimientos a nuestro set de datos que venimos manipulando a lo largo del curso y veremos qué efectos genera.

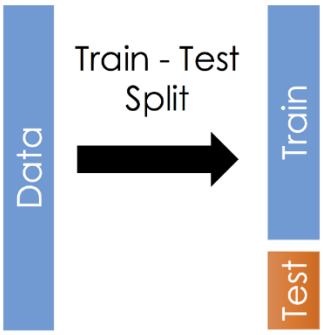
Mejoras de performance del modelo

Dado que los modelos aprenden, por lo tanto, pueden cometer errores, y por esa razón debemos conocer mejor el tipo de errores y el costo asociado del error. En los problemas de clasificación se usa una **matriz de confusión**, la cual muestra el recuento de casos de las clases predicas y sus valores actuales. Si se dispone de información sobre los costos de cada error/acierto en la clasificación, entonces las celdas de la matriz pueden asociarse con el costo de cometer un cierto error de clasificación o de efectuar una clasificación correcta. En este caso, la matriz suele denominarse **matriz de costos** de clasificación. Con estas dos matrices podemos evaluar los modelos con sus costos de error de clasificación y, por ejemplo, buscar un modelo que minimice el costo global.

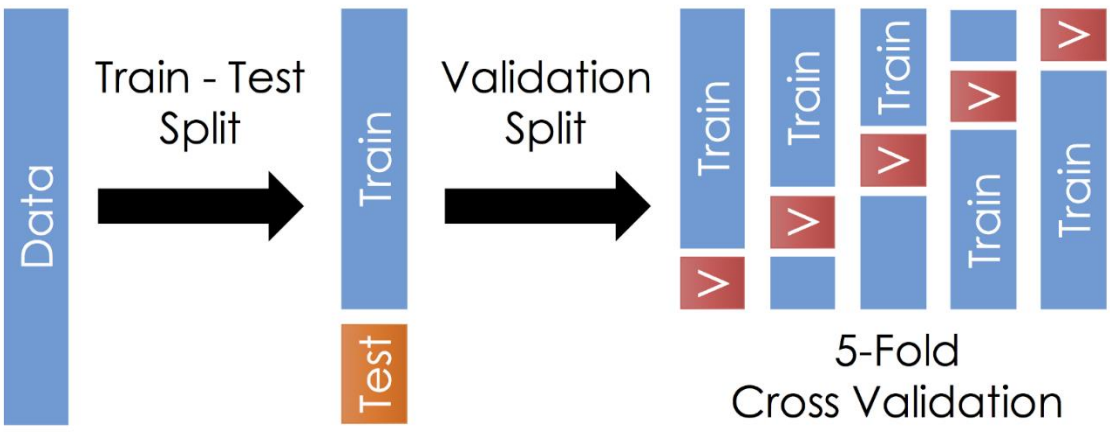
La consideración de que no todos los errores son iguales puede incluso tenerse en cuenta en situaciones donde los costos de error suelen ser difíciles de estimar o incluso desconocidos para muchas aplicaciones. En estos casos, se usan estrategias alternativas como la **curva ROC** (*Receiver Operating Characteristic*)

Cross Validation

A la hora de realizar una validación del modelo generado, tenemos que contar inicialmente con la cantidad de registros para que el algoritmo no sólo pueda aprender, sino también reservar una muestra representativa de dicho data set que nos permita corroborar su buen funcionamiento.

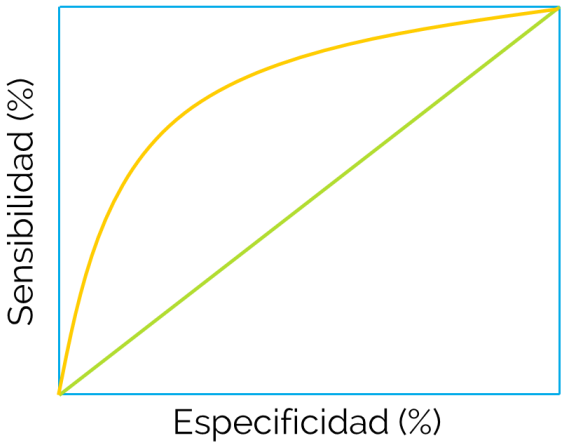


En relación al método de validación cruzada, que se usa normalmente, el data set de entrenamiento se divide aleatoriamente en *n* grupos. Donde se entrena con *n-1* grupos juntos y se guarda *1* para prueba, realizando de manera iterativa este proceso *n* veces y siempre seleccionando diferentes subsets para prueba, como observamos en la siguiente imagen:



Curva ROC

Compara la tasa de falsos positivos con la de verdaderos positivos, y para que sea mínimamente aceptable es necesario que supere el 0.5 que implica la línea recta que cruza el siguiente gráfico:

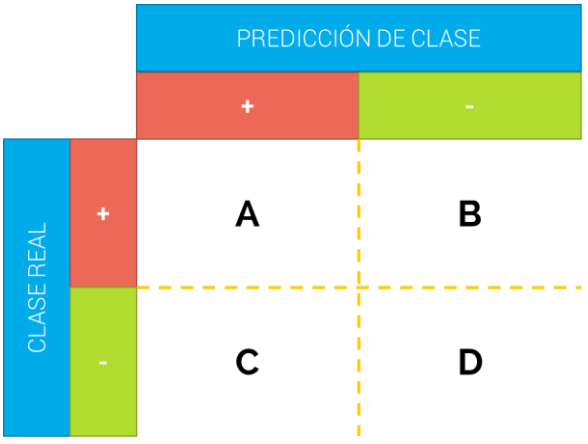


La curva ROC representa la especificidad frente a la sensibilidad para cada posible valor umbral o punto de corte en la escala de resultados de la prueba en estudio.

Este análisis se utiliza, porque lamentablemente no siempre tendremos la posibilidad de contar con una matriz de costos. Además, esta técnica provee herramientas que permiten seleccionar el subconjunto de clasificadores que tienen un comportamiento óptimo en general. Finalmente, el análisis ROC permite evaluar clasificadores de manera más independiente y completa a la clásica precisión.

Matriz de confusión

El análisis ROC utiliza la matriz de confusión siguiente:



Que genera las siguientes ratios:

- Verdaderos Positivos (VP) = A/(A+C)
- Falsos Negativos (FN) = C/(A+C)
- Falsos Positivos (FP) = B/(B+D)
- Verdaderos Negativos (VN) = D/(B+D)

Con estos ratios se genera la curva ROC, que muestra el límite de aceptación de los modelos clasificadores. De modo que el mejor sistema de aprendizaje será aquel que produzca el conjunto de clasificadores con mayor área bajo la superficie convexa o AUC (*Area Under the ROC Curve*)

Accuracy

Hace referencia a la conformidad de un valor medido con su valor verdadero, es decir, se refiere a cuán cerca del valor real se encuentra el valor medido. En términos estadísticos, la exactitud está relacionada con el sesgo de una estimación.

$$Accuracy = \frac{\mathrm{VP} + \mathrm{VN}}{\mathrm{VP} + \mathrm{FP} + \mathrm{FN} + \mathrm{VN}}$$

Precision

La precisión es el ratio entre las categorías asignadas correctamente y las que siendo de similar categoría se asignó con otra. De esta forma, cuanto más se acerque el valor de la precisión al valor nulo, es decir nulo el valor del denominador, mayor serán los desaciertos. Si por el contrario, el valor de la precisión es igual a uno, sé que se asignaron correctamente las categorías.

$$Precision = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FP}}$$

Recall

Este ratio viene a expresar la proporción de categorías asignadas correctamente, comparado con el total de las categorías existentes en la base de datos. Si el resultado de esta fórmula arroja como valor 1, se tendrá la exhaustividad máxima posible, y esto viene a indicar que se ha asignado correctamente la categoría a todos los valores predichos. Por el contrario, en el caso que el valor de la exhaustividad sea igual a cero, se tiene que los documentos obtenidos no poseen relevancia alguna.

$$Recall = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FN}}$$

F_1

Es la medida de precisión que tiene un test. Se emplea en la determinación de un valor único ponderado de la precisión y la exhaustividad (Recall).

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Análisis de parámetros del modelo

En esta sección haremos referencia a dos tipos de criterios que pueden utilizar este tipo de algoritmos:

- Criterios de partición
  - Entropía tiene preferencia por grupos más pequeños y puros
  - Gini tiene preferencia por grupos similares en tamaño
- Criterio de parada
  - Cuando un algoritmo utiliza iteraciones, para no caer en overfitting es importante saber cuándo frenar las iteraciones.

GINI

Como cada nodo del árbol define un subconjunto de los datos de entrenamientos.

Dado un nodo *t* del árbol, Gini(*t*) mide el grado de "pureza" de *t* con respecto a las clases:

- Mayor Gini implica menor pureza
- Gini = 1 – Probabilidad de sacar dos registros de la misma clase

Por lo tanto, el índice de Gini indica la probabilidad de NO sacar dos registros de la misma clase del nodo.

Empíricamente puede observarse:

- Tiende a seleccionar splits que aíslan una clase mayoritaria en un nodo y el resto en otros nodos.
- Tiende a crear splits desbalanceados.

Entropía

Entropía:

- Favorece splits balanceados en número de datos.
- Tiende a encontrar grupos de clases que suman más del 50% de los datos

Poda o pruning

La manera más frecuente de limitar el problema de especificación de los modelos, por lo tanto, que sean modelos más generales, es por medio de la eliminación de condiciones de las ramas de los árboles. Este procedimiento se conoce como poda. Cabe distinguir entre métodos de prepoda y porpoda.

- **Prepoda**: el proceso se realiza durante la construcción del árbol. Se trata en realidad de determinar el criterio de parada a la hora de seguir especializando una rama. En general, los criterios de prepoda pueden estar basados en el número de ejemplos por nodo, en el número de excepciones respecto a la clase mayoritaria (error esperado) o técnicas más sofisticadas, como el criterio MDL.
- **Pospoda**: el proceso se realiza después de la construcción del árbol o el conjunto de reglas. En el caso de los árboles de decisión se trata de eliminar nodos de abajo a arriba hasta cierto límite.

Aunque los criterios están basados en las mismas medidas, prepoda y pospoda, dado que la pospoda se realiza con una visión completa del modelo, suele obtener mejores resultados. Obviamente, la pospoda es menos eficiente que la prepoda, ya que ásta no genera nada que luego deba eliminarse.

Una consecuencia de utilizar prepoda o pospoda (o ambos) es que los nodos hoja ya no van a ser puros, es decir, es posible que tengan ejemplos de varias clases. Normalmente se elegirá la clase con más ejemplos para etiquetar el nodo hoja y hacer, por lo tanto, la predicción.

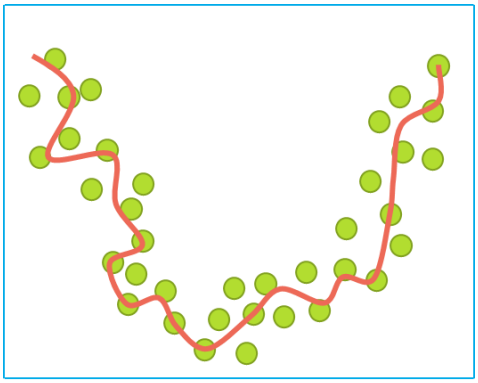
Lo que, si interesa conocer, a estas alturas, es el nivel de poda óptimo. Para ello la mayoría de los métodos de poda tienen uno o más parámetros que permiten decidir el grado de poda. Aunque por defecto en su mayoría tienen un valor asignado.

Probaremos esto con nuestro set de datos.

Overfitting

Overfitting es una de las principales causas para obtener malos resultados, ya que sucede cuando sin mucho criterio intentamos *encajar* los datos de entrada en nuestro modelo. De modo que generamos un modelo tan complejo que aprende de manera exhaustiva nuestro set de datos de entrenamiento, pero a la hora de testearlo con nuevos datos no tiene la capacidad de generalizar y flexibilizarse frente a datos nuevos. Así, la predicción que genera suele tener pésimos resultados, aun cuando en nuestro entrenamiento aparentaba funcionar a la perfección.

Gráficamente sería algo así:



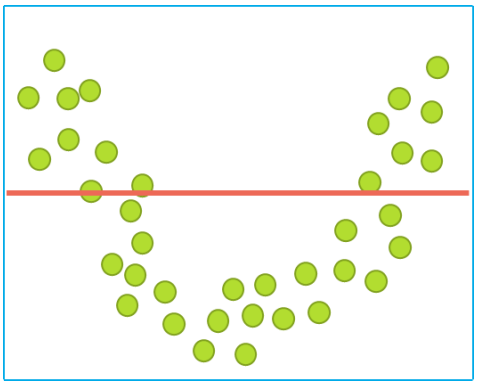
Para tratar el overfitting tenemos las siguientes opciones:

- Restringir el crecimiento del árbol
- Podar el árbol, de abajo por arriba
    - Se hace con datos nuevos, no con los que se creó el árbol
    - Se puede hacer mediante un algoritmo de IBM, MDL.

Underfitting

Puede suceder lo inverso con el subajuste, que, por miedo de caer en el overfitting, generemos un modelo tan pero tan general que no llegue a predecir nada.

Lo veremos gráficamente:



Practica modulo 4

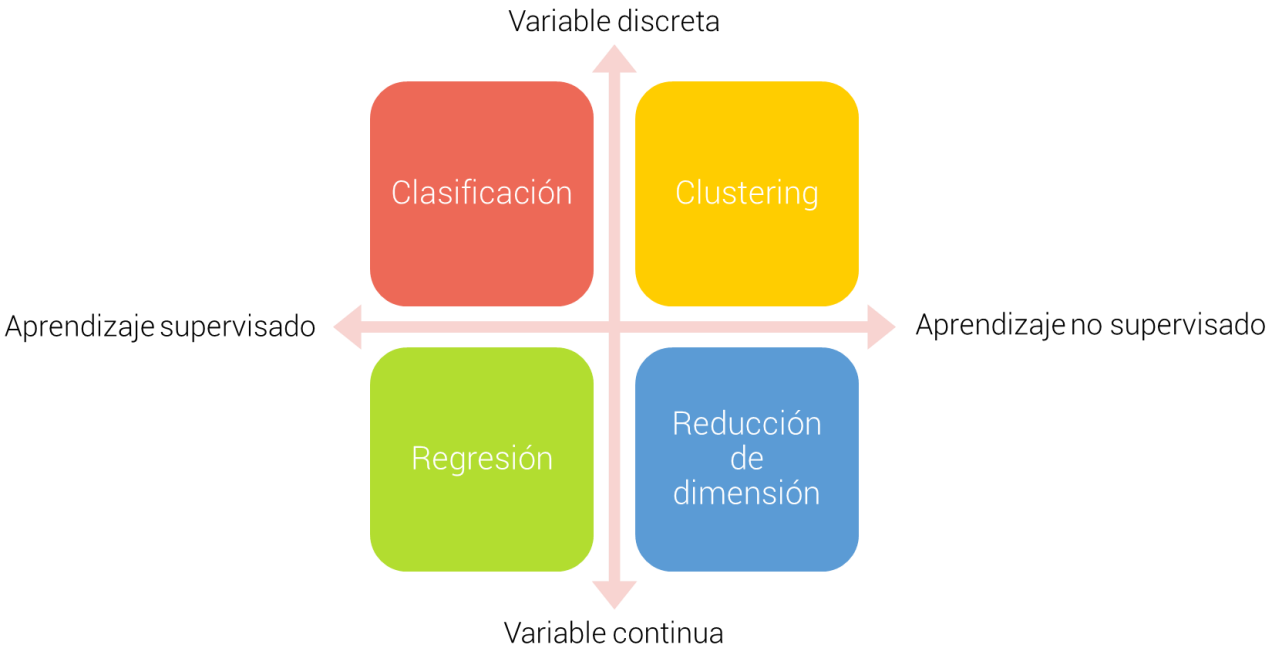
Para realizar la práctica deberás entrar a GitHub y luego podes descargarte los archivos y trabajarlos en tu maquina o ir a https://mybinder.org y ejecutar Jupyter notebook desde ahí.

Test teórico

Ingresar al siguiente link e ingresar el código que aparece en pantalla: https://kahoot.it/

Otros tipos de modelos, finalidad y características básicas

Podríamos diferenciar todo a partir de la siguiente imagen:



Por tipo de Aprendizaje

Supervisado

La primera modalidad de aprendizaje que tiene el machine learning es la de aprendizaje supervisado. Usándola, se entrena al algoritmo otorgándole las preguntas, denominadas características, y las respuestas, denominadas etiquetas. Esto se hace con la finalidad de que el algoritmo las combine y pueda hacer predicciones.

No supervisado

A diferencia del aprendizaje supervisado, en el no supervisado solo se le otorgan las características, sin proporcionarle al algoritmo ninguna etiqueta. Su función es la agrupación, por lo que el algoritmo debería catalogar por similitud y poder crear grupos, sin tener la capacidad de definir cómo es cada individualidad de cada uno de los integrantes del grupo.

Por tipo de Target

Existen, dos tipos de aprendizaje supervisado:

- Regresión: tiene como resultado un número específico. Si las etiquetas suelen ser un valor numérico, mediante las variables de las características, se pueden obtener dígitos como dato resultante.

- Clasificación: en este tipo, el algoritmo encuentra diferentes patrones y tiene por objetivo clasificar los elementos en diferentes grupos.

Parámetros que se definen

A modo de ejemplo para algunos de los mencionados, los parámetros que se definen son:

Random forest

- Número de árboles a realizar
- Número de variables que probar para cada división
- Fracción de observaciones usadas para construir un árbol

Support Vector Machine

- Grado polinomial
- Valor de penalización

Gradient Boosting

- Número de árboles
- Tasa de aprendizaje
- Tasa de muestreo
- Variables que considerar en cada división
- Parámetros de regularización L1 y L2

Redes Neuronales

- Cantidad de capas ocultas o neuronas
- Parámetros de regularización L1 y L2
- SGD

Cluster

El agrupamiento, también conocido como clustering o segmentación, es la tarea descriptiva por excelencia que tiene por objetivo obtener grupos "naturales" a partir de los datos. Se les llama grupos, no clases, porque a diferencia de la clasificación, en lugar de analizar los datos etiquetados en una clase, los analiza para generar esta etiqueta.

Los datos son agrupados basándose en el principio de maximizar la similitud entre los elementos de un grupo minimizando la similitud entre los distintos grupos.

Tipos de cluster

Algunos tipos de cluster, los que pueden ilustrarnos a esta altura del curso, son:

Particional:

1. Cluster de Grafos:

2. Búsqueda de Modas
3. Mezcla de densidades
4. Mínimo error cuadrático

Jerárquico:

1. Enlace simple
2. Enlace poderado
3. Enlace completo
4. Método de Ward

Regresión

Es una tarea predictiva que consiste en aprender una función real que asigna a cada instancia un valor real. Se diferencia de la clasificación porque el valor a predecir es numérico.

El objetivo en este caso es minimizar el error (generalmente el error cuadrático medio) entre el valor predicho y el valor real.

Tipos de Regresión

El objetivo principal de construir un modelo de regresión puede ser, por ejemplo, evaluar cómo afecta el cambio en unas características determinadas (variables independientes) sobre otra característica en concreto (variable dependiente), denominado modelo con fines explicativos; o también nuestro objetivo podría ser intentar estimar o aproximar el valor de una característica (variable dependiente)

en función de los valores que pueden tomar en conjunto otra serie de características (variables independientes), denominado entonces modelo con fines predictivos.

Existen varias opciones para estimar un modelo de regresión, de entre los que destacan por su facilidad de aplicación e interpretación, el modelo de regresión lineal y el modelo de regresión logística. Teniendo en cuenta el tipo de variable que deseemos estimar (variable dependiente o respuesta) aplicaremos un modelo de regresión u otro. Simplificando, cuando la variable dependiente es una variable continua, el modelo de regresión más frecuentemente utilizado es la regresión lineal, mientras que cuando la variable de interés es dicotómica (es decir, toma dos valores como sí/no, hombre/mujer) se utiliza la regresión logística

Test teórico

Ingresar al siguiente link e ingresar el código que aparece en pantalla: https://kahoot.it/