

**UNIVERSIDAD CATÓLICA BOLIVIANA “SAN PABLO”
CARRERA DE INGENIERÍA DE SISTEMAS
MAESTRÍA EN CIENCIA DE DATOS V1**

EXPLORATORY DATA ANALYSIS (EDA)

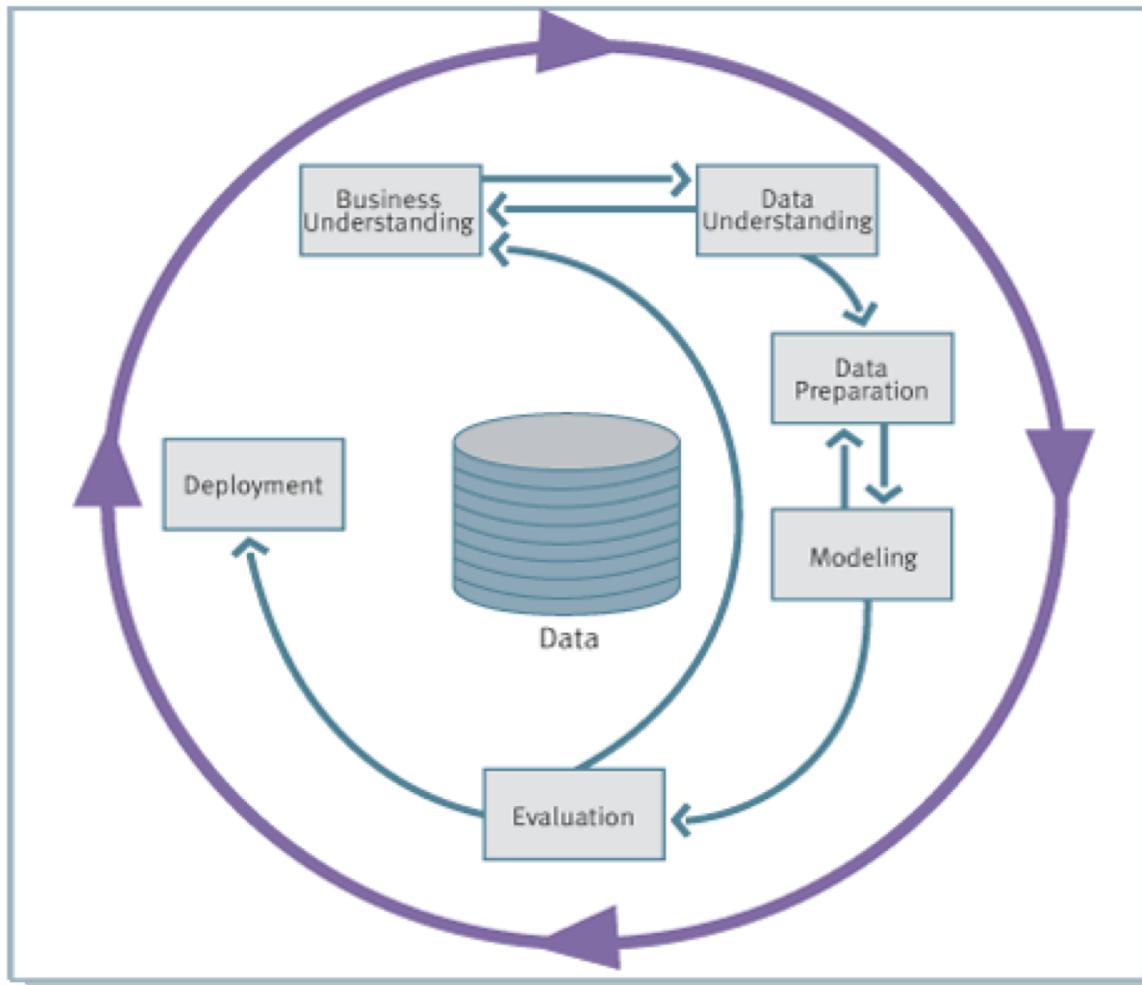
1. Exploración de Datos como proceso

- La exploración de datos se inicia con la identificación de una necesidad en la organización o el mercado.
- Se toma en cuenta las expectativas del tomador de decisiones respecto al retorno de la información descubierta.
- Lo que se necesita es un proceso que asegure la obtención del mayor provecho en el análisis de la información.

Data exploration project

	Time to complete (percent of total)	Importance to success (percent of total)
1. Exploring the problem	10	15
2. Exploring the solution	9	14
3. Implementation specification	1	51
4. Data mining		
a. Data preparation	60	15
b. Data surveying	15	3
c. Data modeling	5	2

Cross- Industry Standard Process for Data Mining - CRISP-DM



Ref.: www.crisp-dm.org

Fases en el proceso de CRISP-DM

- **Business Understanding**

Project objectives and requirements understanding, Data mining problem definition

- **Data Understanding**

Initial data collection and familiarization, Data quality problems identification

- **Data Preparation**

Table, record and attribute selection, Data transformation and cleaning

- **Modeling**

Modeling techniques selection and application, Parameters calibration

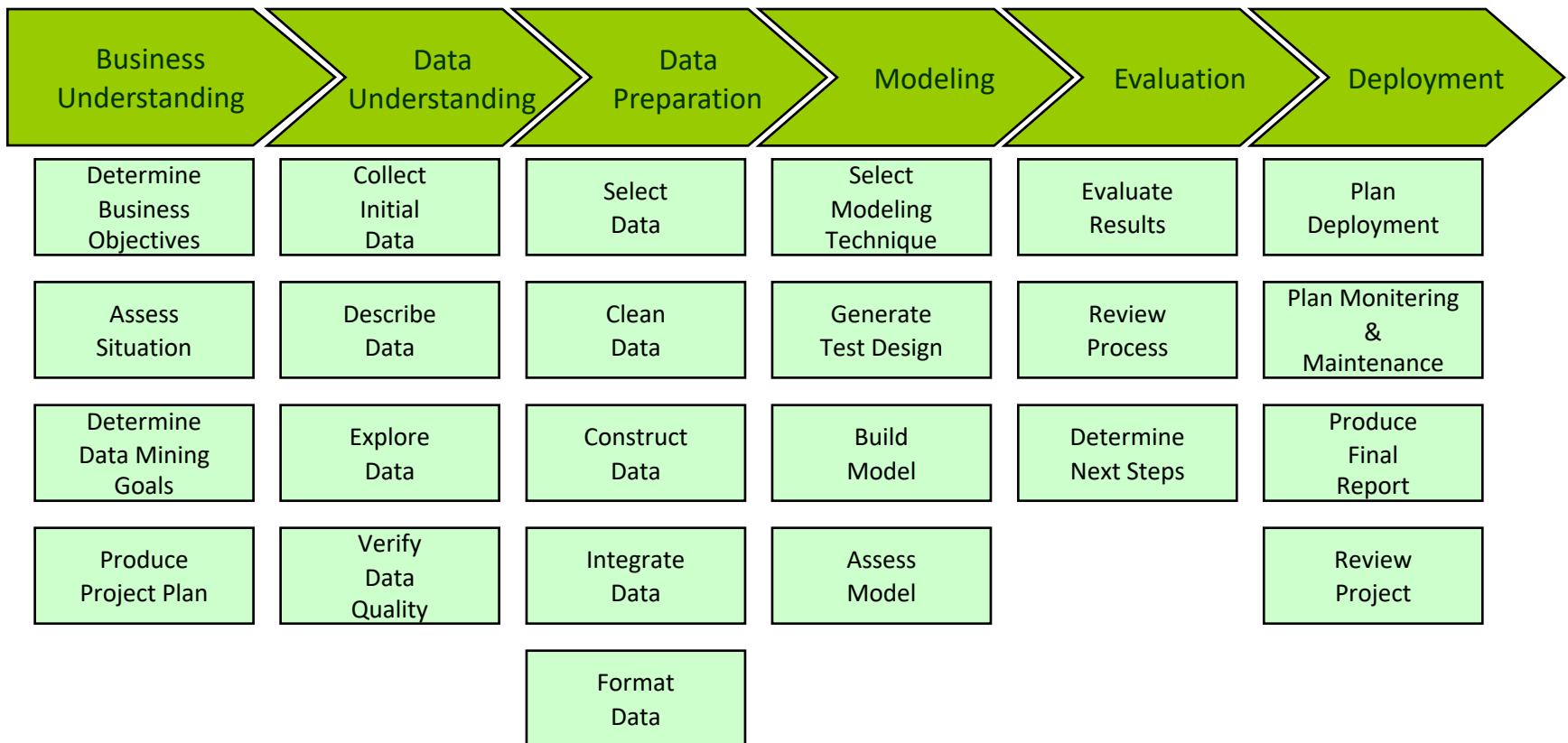
- **Evaluation**

Business objectives & issues achievement evaluation

- **Deployment**

Result model deployment, Repeatable data mining process implementation

Fases y Tareas en CRISP-DM



2. La naturaleza del mundo y su impacto en la Preparación de Datos

El mundo consiste de objetos que podemos identificar:
Autos, árboles, costo de vida, empleo, justicia, etc.

Objetos como colecciones de características de las cuales se pueden tomar medidas.

Capturando medidas

Ejemplo:

Un auto es azul, de dos puertas, cuatro cilíndros, cinco asientos.

Es decir, color, número de puertas, número de cilindros, etc.

Tipos de mediciones

- Todas las mediciones tienen alguna escala.
- Una variable recibe la medición que puede tomar un conjunto de valores

Algunas variables consisten de dos componentes:

- La escala
- La medición como tal, y otros que requieren más componentes

- *Variables escalares*: Variables que identifican la posición de un punto en alguna escala.

Mediciones escalares

Ejemplos

La temperatura del agua:

“hirviendo,” “Demasiado caliente,” “tibio”
“frio”, etc.

Qualitative (always discrete)

Nominal Scale Measurements (=, <>)

Categorías sin orden. Valores nominales llevan la menor cantidad de información. La idea es nombrar a los objetos

- Medidas categóricas nombran grupos de cosas, no entidades individuales.
- Los valores pueden ser agrupados en formas significativas

Mediciones categóricas denotan que existe una diferencia en tipo pero no es posible cuantificar tal diferencia.

Ejemplos:

- Número de ID, Estado civil, sexo, zona.
- Identificador de rubro de organización
- Standard industry classification (SIC) categorizan diferentes tipos de actividades de negocios.

Ordinal Scale Measurements (=, <>, <, >)

Categorías con orden significativo.

El ranking de las categorías es transitivo, si A es calificado más alto que B y B más alto que C, entonces A debe ser más alto que C.

Ejemplos:

- Temperatura
- Altura
- Peso
- Humedad

Quantitative (can be discrete or continuos) (=, <>, <, >, +, -)

Interval Scale Measurements

INTERVAL: No existe un cero verdadero, la división no tiene sentido.

Cuando exista información disponible no solamente acerca del orden para el ranking de valores pero también la diferencia en tamaño entre valores.

Ejemplo es la Temperatura, basada en la escala de intervalo utilizada, que la baja para dos días puede ser comparada. Es decir, 80 grados no es dos veces más caliente que 40 grados si se usa Fahrenheit frente a Celcius.

- **El punto zero** no está a la misma temperatura para las dos escalas.
- Las escalas tienen diferentes ratios a temperaturas equivalentes.
- El punto cero es fijado en forma arbitraria.

Ejemplos: Escalas de Temperaturas, Fechas de calendario, Scores.

Ratio Scale Measurements (=, <>, <, >, +, -, *, /)

Cero verdadero, La división tiene sentido

Ejemplos:

- Monto en Cuenta de Banco, comienza con un cero “verdadero”.
- Longitud
- Temperatura en la escala Kelvin

Nonscalar Measurements

Los valores escalares consisten de solamente dos componentes.
El valor de la medición y la escala.

Mediciones no escalares necesitan más componentes:

Ejemplo:

- Velocidad es el número de Km recorridos en una hora.
- Consumo de gasolina en Km por Litro.
- Fallas por hora

Ejemplos:

- # of phones in your house
(discrete, quantitative, ratio)
- Size of french fries
(discrete, qualitative, ordinal)
- Ownership of a cell phone
(binary, qualitative, nominal)
- # of local phone calls made in a month
(discrete, quantitative, ratio)
- Length of longest phone calls
(continuos, quantitative, ratio)
- Height
(continuos, quantitative, ratio)

- 
- Price of your text book
(discrete, quantitative, ratio)
 - Zipcode
(discrete, qualitative, nominal or ordinal all depends)
 - Temperature in degrees Farenheight
(continuos, quantitative, interval)
 - Temperature in degrees Celsius
(continuos, quantitative, interval)
 - Temperature in degrees Kelvin
(continuos, quantitative, ratio)

Monotonicity

- Una variable monotónica es la que se incrementa sin final.
- Monotonía también existe en la relación entre variables en las cuales una variable crece mientras la otra crece, decrece o se mantiene.

Ejemplos:

- Tiempo reflejado en fecha es una var. monotónica.
 - Números de Seguro Social
 - Números de Registros
 - Número de Factura
 - Códigos de Empleados
 - y otros.
-
- El problema es que estas variables, para entrar en un análisis, siempre deben ser transformadas en var. no monotónicas.
 - Por ejemplo una transformación es de tipo Datestamps a otra que identifique la estación del año.

Sparsity

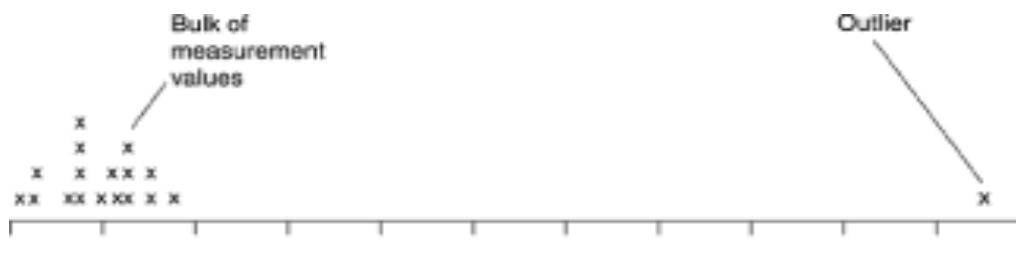
- Cuando variables individuales son sparse se debe decidir cuando quitarlas por que pueden o no tener valores insignificantes.
- En algunos casos las variables necesitan ser colapsadas a un número reducido de tal forma que cada una tenga información de varias de las demás variables.
- La reducción del número de variables se denomina *dimensionality reduction*.

Incrementando la Dimensionalidad

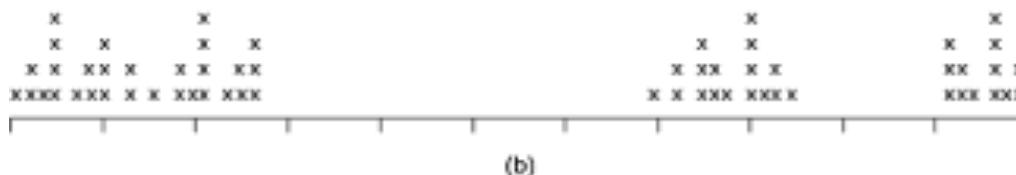
- Existen algunas circunstancias donde la dimensionalidad de una variable necesita ser incrementada, “one-hot encoding”.
- Un ejemplo son los ZIP codes, es beneficioso traducir el código ZIP de una lista categórica a una que refleje Latitud y Longitud.

Outliers

Un Outlier es la ocurrencia de un valor con frecuencia baja en una variable que está lejos del cúmulo del grupo de variables .



Examples of outliers: as an individual value (a) and as clumps of values (b).



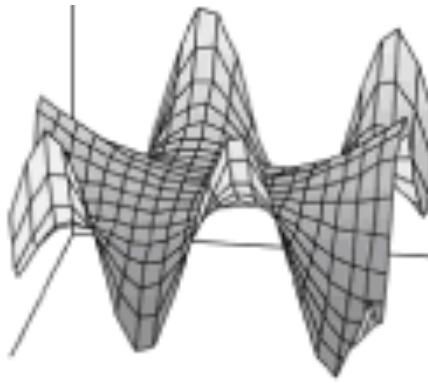
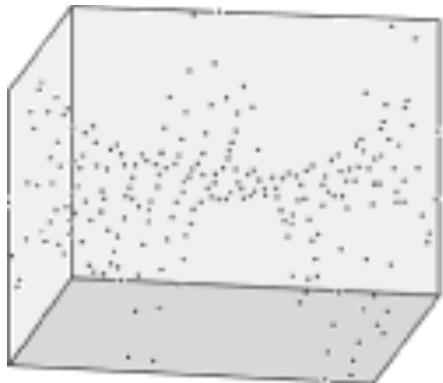
Numerando Valores categóricos

La experiencia muestra que técnicas de modelado que tratan bien con valores categóricos se benefician de una numeración válida de categorías

“One of the key principles in data preparation is to do as little damage as possible to the natural structure in a data set.”

The Shape of the Data Set

State space: Es llamado así a causa de las naturaleza de las instancias de los datos



Points plotted in a 3D phase space (left) can be represented by a manifold (right).

- El State Space puede ser extendido a la cantidad de dimensiones que número de variables existen.

Un nombre general para esta n-dimensional extensión de una linea o un plano se llama manifold. Es análogo a una hoja flexible como existe en tres dimensiones, pero puede ser propagado al número de dimensiones que se requiera.

La Preparación de Datos como Proceso

Etapas en la Preparación de Datos

- Las primeras son ***actividades no automatizadas*** que son procedurales-> Preparación básica
- El segundo conjunto de actividades ***actividades de preparación automatizadas*** -> *El código*
- Las etapas:
 1. Acceso a los datos
 2. Auditaje de datos
 3. Mejora y enriquecimiento de los datos
 4. Buscando el sesgo muestral
 5. Determinando la estructura de datos
 6. Construyendo el PIE
 7. Estudiando los datos (Survey)
 8. Modelando los datos

Stage 1: Accesando a los datos

- El punto de inicio de cualquier proyecto de preparación de datos es localizar los mismos.
- Los Warehouses tienen sus inconvenientes, uno significativo es que ellos a menudo son creados con una estructura particular que refleja alguna vista específica de la empresa.

Stage 2: Auditaje de datos

- Asumir que los datos adecuados están disponibles:
 - La fuente de datos
 - La cantidad de datos
 - La calidad de los datos

- El auditaje requiere el examen de pequeñas muestras de datos:
 - número de campos,
 - contenido de cada campo,
 - fuente de cada campo,
 - valores máximos y mínimos,
 - número de valores discretos,
 - y otros.

Existe alguna razón justificable para suponer que estos datos tienen el potencial para proveer la solución requerida al problema?

Stage 3: Mejora y enriquecimiento de los datos

Existen varias formas en las cuales los datos existentes puedan ser manipulados para extender su utilidad.

Stage 4: Buscando el sesgo muestral

El muestreo presenta algunos problemas. Existen algunos métodos automatizados para ayudar a detectar el sesgo muestral.

Stage 5: Determinando la estructura de datos

La estructura se refiere a la forma en la cual las variables en el Data Set se relacionan de acuerdo a sus niveles de agregación.

Stage 6: Construyendo datos de entrada

Data Issue: Muestras representativas

- Cuanta cantidad de datos es necesaria para el modelado.

“all of the data, all of the time” ???

- Estudio y el modelado todavía requieren al menos tres conjuntos: Training Set, Testing Set y el Validation Set.

Data Issue: Valores categóricos

- Datos categóricos son numerados o asignados con números apropiados.
- Cuando se construyen modelos predictivos o inferenciales es crítico que el orden natural de los valores categóricos se conserve.

Data Issue: Normalización

- La forma de normalización requiere cambiar los valores de instancia en formas especificadas y claramente definidas para exponer su contenido.

Data Issue: Missing and Empty Values

Data Set Issue: Reduciendo Width

- *Width* es descrita como el número de columnas y *depth* describe el número de filas.

Data Set Issue: Reducing Depth

- Sigue existiendo la necesidad de asegurar que el subconjunto de datos modelados en efecto refleje todas las interrelaciones que existen en el Data Set completo.

Data Set/Data Survey Issue:

- Este es realmente el primer paso del estudio (survey) como también el último paso de la Preparación de Datos.

Stage 7: Inspección de los datos (Survey)

- El estudio de los datos examina y reporta sobre las propiedades generales del manifold en el State Space.

Stage 8: Modelando los Datos

- El propósito principal de preparar y estudiar los datos es entenderlos.
- La mayoría de las herramientas son descritas como capaces de aprender las interrelaciones entre variables.
- El problema es lograr que las herramientas aprendan la verdadera interrelación antes de aprender el ruido.
- El propósito de la Preparación de Datos es transformar el Data set tal que la información contenida sea expuesta a la herramienta de minería.

Data Discovery

Data Access Issues

Aspectos legales. Pueden existir barreras legales para accesar algunos datos o partes de los Data Sets. Ej. Confidencialidad

Aspectos Gerenciales. Ej. Datos de producción pertenecen al secreto industrial de la empresa

Razones Políticas. Los datos y especialmente su propiedad es a menudo inconveniente.

Formato de los Datos. Los datos han sido generados y reunidos en muchos Formatos.

Conectividad. El acceso a los datos requiere que estén disponibles en línea y conectados al sistema que será utilizado para el minado.

Algunas medidas en variables numéricas

El análisis exploratorio de datos en el caso de valores numéricos
Medidas de tendencia Central, de dispersión y otros

Medidas de tendencia Central Medidas de dispersión

- | | |
|--|---|
| <ul style="list-style-type: none">• Media• Mediana• Moda | <ul style="list-style-type: none">• Varianza• Desviación Estándar• Rango de Variación |
|--|---|

Otros

- Cuenta
- Mínimo
- Máximo
- Cuartiles

Procesos generales para EDA

- Remove duplicate rows
- Drop constant features
- Get and remove duplicate columns
- Remove duplicate features
- Select numeric, non-numeric and object columns only
- Get numeric values from a categorical feature
- Print features levels

Preparación de datos

- Pasos en la exploración de datos
 - 1. Identificación de variables
 - 2. Análisis por variable y entre variables
 - 3. Tratamiento de valores faltantes
 - 4. Tratamiento de outliers
 - 5. Transformación de Variables
 - 6. Creación de Variables
- Pueden realizarse iteraciones en los pasos anteriores para mejorar el EDA

1. Identificación de variable

- En este paso se identifica la variable input y la variable Target o Label, luego el tipo de variable Categórica o numérica.
- Los datos pueden estar en DBMS
 - Protocolos ODBC, JDBC
- Datos en archivos planos
 - Formato de columna
 - Delimitadores de campo como tab, coma, etc.
 - Convertir los delimitadores de campo dentro de los strings
- Verificar el número de filas antes y después

- **Tipos de campos:**
 - binarios, nominal (categórico), ordinal, numérico, ...
 - Para campos nominlaes: Tablas que cuenten con la correspondencia de códigos a descripciones completas.
- **Rol del campo:**
 - Input: input para el modelado
 - Target: output
 - id/auxiliar : mantengalo, pero no utilizar para el modelado
 - Peso: peso de instancia
 - ...
- **Descripciones de los campos**

2. Análisis por variable y entre variables

- Se debe analizar variable por variable respecto a su distribución, principalmente en forma gráfica, de acuerdo a si es categórica o numérica.
- En el caso de numérica su sesgo
- En el caso de categórica su tabla de frecuencias
- Analizar la variable en cuestión respecto a su capacidad de predicción respecto al target o label

Remapeo de valores de variables

En general, el remapeo puede ser verdaderamente útil cuando una o más de estas circunstancias es verdad:

- La densidad de la información a ser remapeada en variables es baja.
- La dimensionalidad del modelo es modestamente incrementada
- El modelo requiere que las entradas categóricas sean representadas sin un orden implícito, lo cual se refleja cuando se mapea a una variable numérica.

Conversión: Nominal a Numérico

- Algunos métodos (redes neuronales, regresión, vecino más próximo) precisan que los atributos sean numéricos.
- Para usar atributos nominales hay que convertirlos a valores numéricos
 - P: ¿Por qué no simplemente ignorarlos?
 - R: Pueden contener información valiosa
- Hay diferentes estrategias según los atributos nominales sean
 - binarios
 - ordenados
 - valores múltiples

Conversión binario a numérico

- ¿Cómo convertirían atributos binarios a numéricos?
 - Ej. Sexo=M, F
- ¿Cómo convertiría atributos ordenados a numéricos?
 - Ej. calificaciones
 - sobresaliente, notable, aprobado,...
 - A+, A, A-, B+,...
- Atributos binarios
 - Ej. Deserta=S, N
- Convertir a Atributo_0_1 con valores 0, 1
 - Deserta= N → Deserta_0_0 = 0
 - Deserta= S → Deserta_0_1 = 1

Conversión: ordenados a numéricos

- Los atributos ordenados (ej. calificaciones) pueden convertirse a números conservando su orden natural
 - A → 4.0
 - A- → 3.7
 - B+ → 3.3
 - B → 3.0

Q: ¿Por qué es importante preservar el orden?

para permitir comparaciones significativas

ej. calificación > 3.5

Conversión: Nominal con pocos valores

- Atributos con un pequeño conjunto de valores posibles ($\sim < 32$) sin orden
 - ej. Color=Rojo, Naranja, Amarillo, ..., Violeta
 - para cada valor v crear una variable binaria C_v que es 1 si Color= v , 0 de otra forma.

The diagram illustrates the conversion of a nominal attribute 'Color' into binary variables. On the left, a table has columns 'ID', 'Color', and '...'. The first row has ID 371 and Color 'rojo'. The second row has ID 433 and Color 'amarillo'. An arrow points from this table to a second table on the right. The second table has columns 'ID', 'C_rojo', 'C_naranja', 'C_amarillo', and '...'. The first row corresponds to ID 371 with C_rojo = 1 and C_amarillo = 0. The second row corresponds to ID 433 with C_amarillo = 1 and C_rojo = 0.

ID	Color	...
371	rojo	
433	amarillo	

→

ID	C_rojo	C_naranja	C_amarillo	...
371	1	0	0	
433	0	0	1	

Conversión: Nominal con muchos valores

- Ejemplos:
 - Códigos de países (150+ valores)
 - Código de profesiones (7,000 valores, pero sólo algunas aparecen frecuentemente)

Q: ¿Qué hacer?

- Ignorar los que funcionan como IDs que tienen valores únicos para cada registro
- Agruparlos de forma natural
 - ej. países → 6-7 regiones
 - Profesión – seleccionar las más frecuentes, agrupar las demás con 0.
- Crear campos binarios para atributos seleccionados

Remapeo One-of- n

Una representación one-of- n *requiere la creación de pseudo variable con valor binario para cada valor categórico.*

Las desventajas ¿? serían:

- La dimensionalidad se puede incrementar considerablemente.
- La densidad de un estado particular es muy baja.
- Nuevamente, inclusive cuando una pseudo variable tiene una densidad razonable para el modelado, las variadas salidas estarán “prendidas” en algún grado si van a ser predichas.

Remapeo m-of-n

- Este remapeo m-of-n es una ventaja si cualquiera de las condiciones es satisfecha:

Primero: Si el número total de variables adicionales es menor que el número de etiquetas.

Segundo: Si el remapeo m-of-n adiciona información útil.

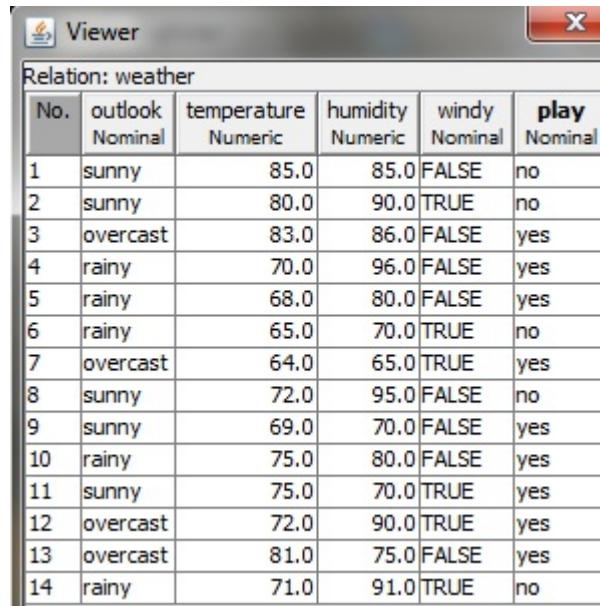
Remapeo para eliminar el ordenamiento

- Un otro elemento del remapeo es cuando es importante que no hayan implicaciones de orden entre las etiquetas.

7. Discretización

- Algunas técnicas requieren valores discretos como por ej. Naïve Bayes.
- La discretización es muy útil para generar resúmenes de datos
- La discretización es también denominada “binning”

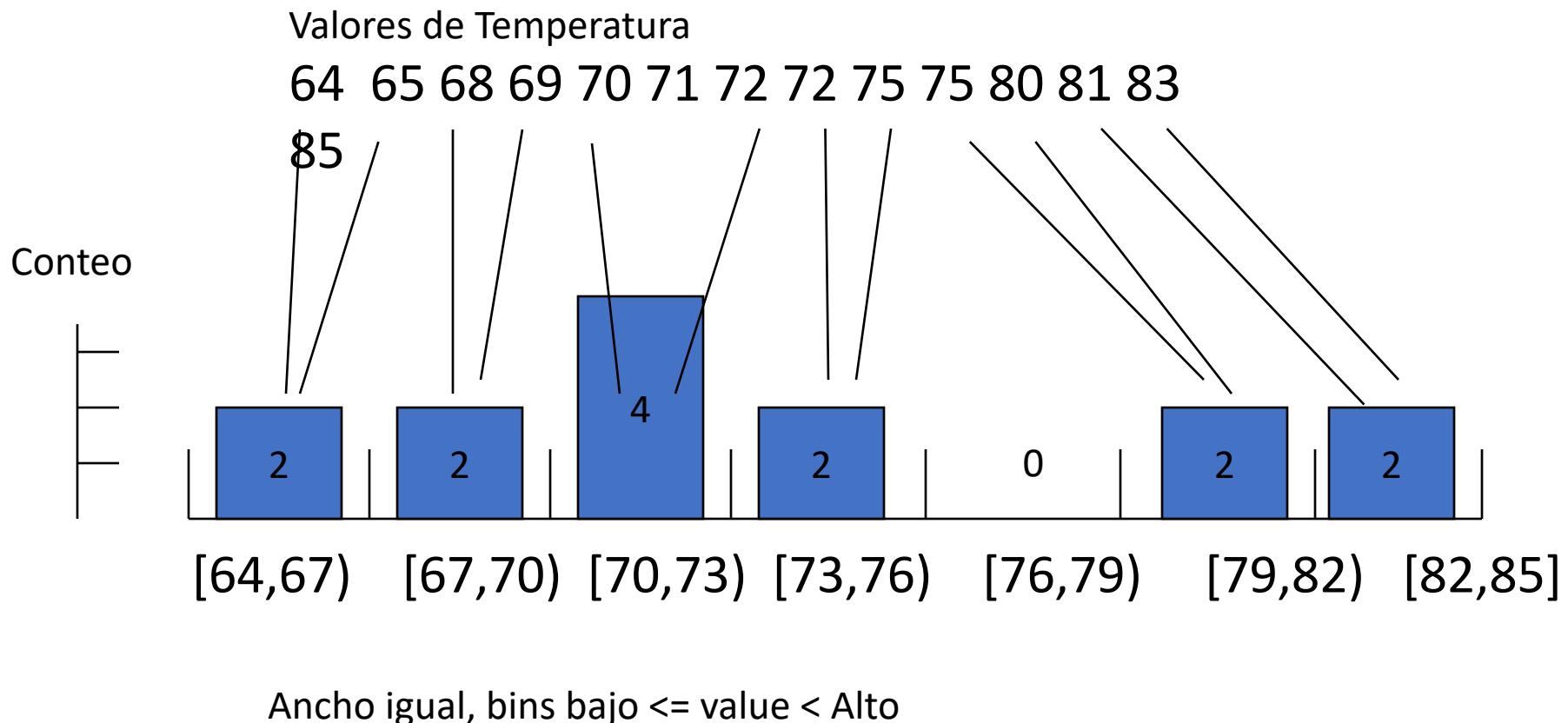
Ejemplo:



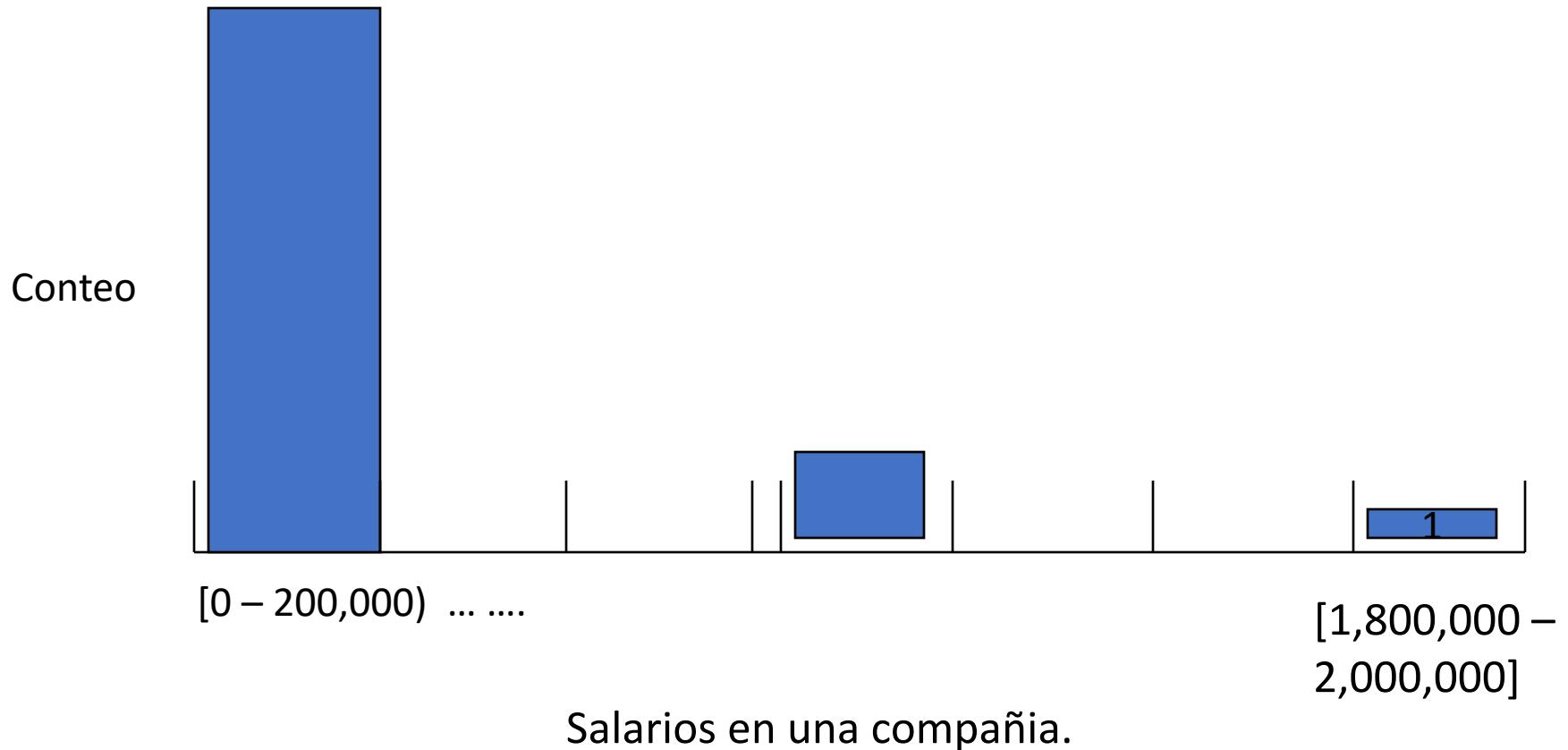
The screenshot shows a software window titled "Viewer" displaying a dataset named "weather". The table has the following structure:

No.	outlook Nominal	temperature Numeric	humidity Numeric	windy Nominal	play Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

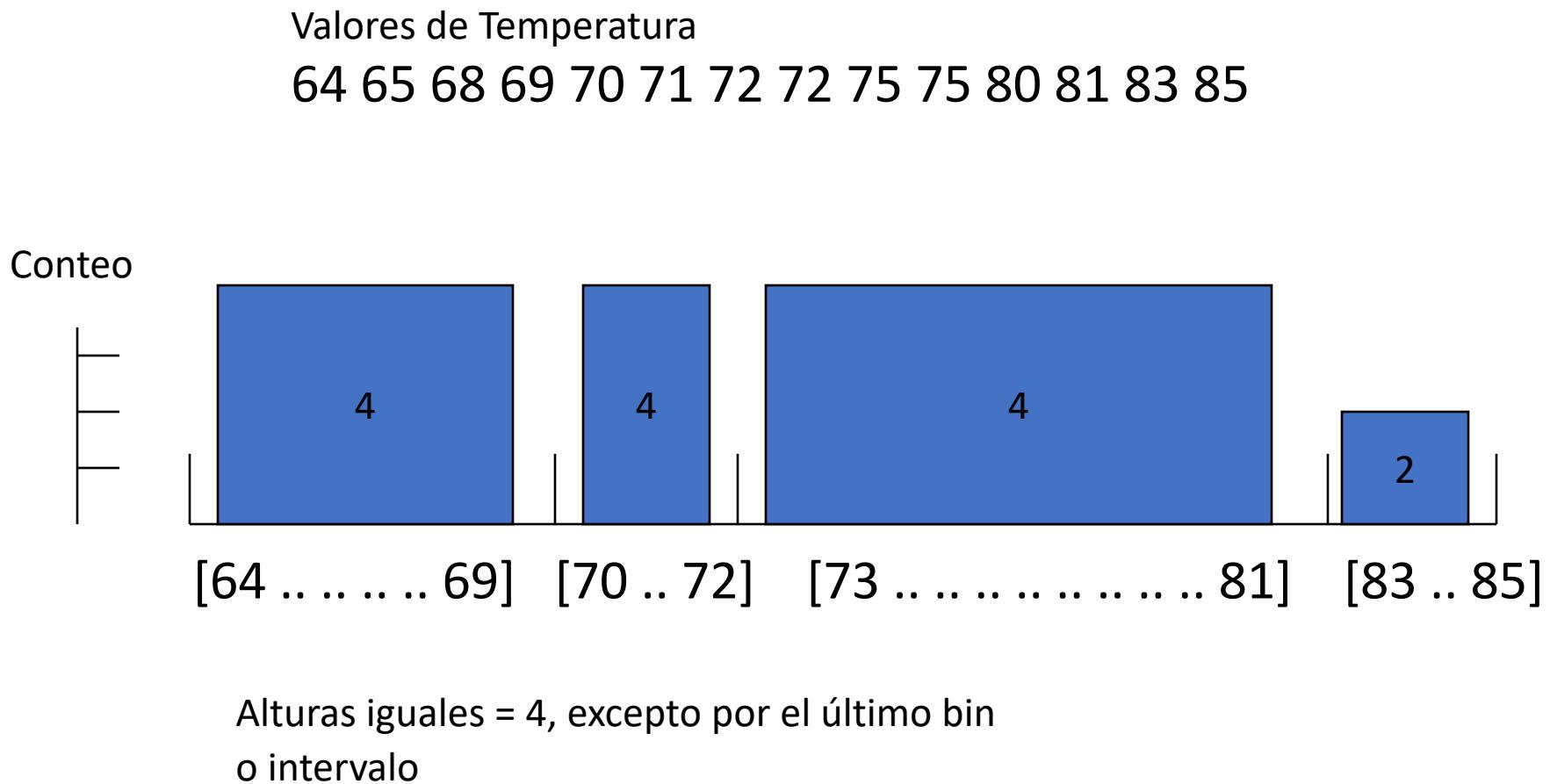
Discretización con ancho igual



Anchos iguales pueden producir aglutinamientos



Discretización con altura igual



Ventajas cuando la altura es la misma en los intervalos

- Se prefiere porque evita los aglutinamientos.
- Consideraciones adicionales:
 - No divide valores frecuentes a través de intervalos
 - Crea intervalos separados para valores especiales como el cero.
 - Puntos de corte legibles

Discretization: Consideraciones

- Ancho igual es más simple, adecuado para muchas clases.
 - Pero, puede fallar para distribuciones desiguales
- Altura igual da mejores resultados
- La dependencia de la clase puede ser mejor para la clasificación.
- Nota: Los árboles de decisión construyen la discretización “al vuelo”
- Naïve Bayes requiere una discretización inicial

Discretización no supervizada

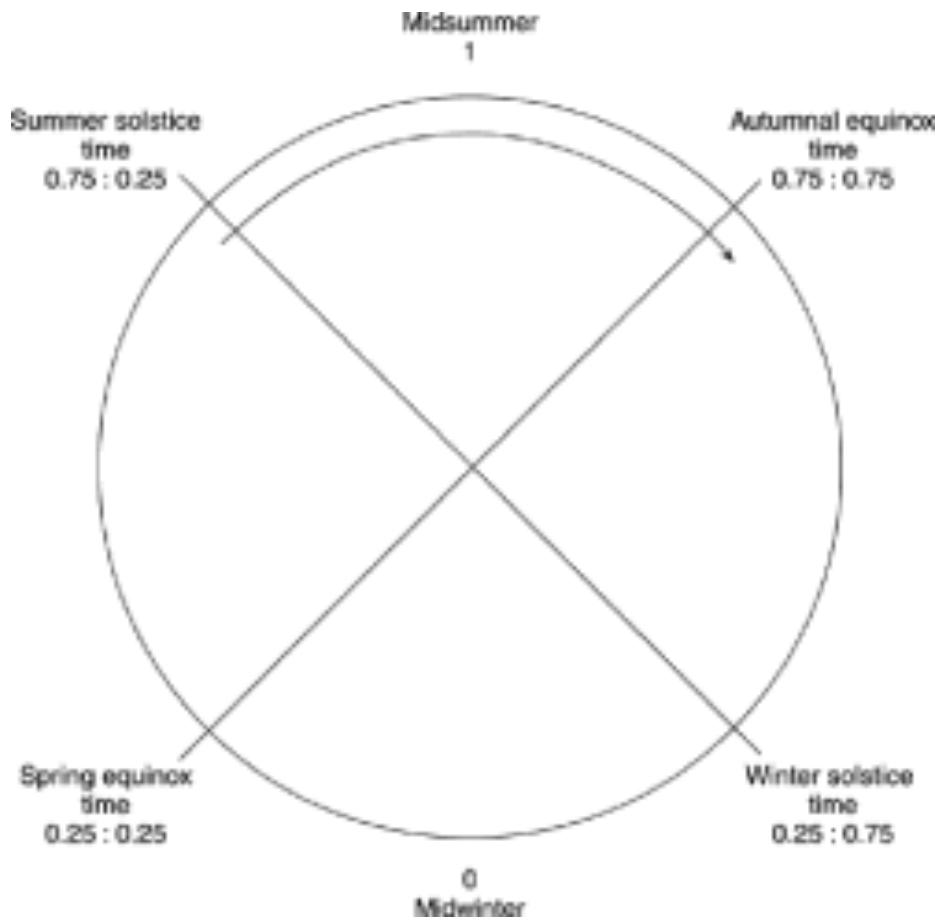
- Determina intervalos sin conocer la clase
 - Es el único medio posible cuando se utiliza clustering
- Dos estrategias:
 - *Equal-interval binning*
 - *Equal-frequency binning*
(también denominado *histogram equalization*)
 - Equal-frequency binning trabaja bien con Naïve Bayes si el número de intervalos es la raíz cuadrada del tamaño del dataset.

Discretización supervizada

- *Método Entropy-based*
- Construya un árbol de decisión con pre-poda en el atributo siendo discretizado
 - Utilice la entropía como criterio de división (ing. splitting)
 - Use el min description length como criterio para detenerse.
- Para aplicar min description length principle:
 - La “teoría” es
 - El splitting point ($\log_2[N - 1]$ bits)
 - Más la distribución de la clase en cada subconjunto

Remapeando la Discontinuidad Circular

Para los “mineros” de datos, el tiempo es circular. Las estaciones “dan vuelta”, es decir, luego de cada diciembre aparece enero.



Un reloj anual. El tiempo es representado por dos variables – una mostrando el tiempo ahora y una mostrando donde el tiempo fue un cuarto del año atrás.

Gráficos

- Histogram
- Scatter plot
- Boxplot
- Barplot
- Distributions
- Heatmap plot
- Facet grids
- Cluster map
- Joint plot - combines scatter + histograms
- Regression plots

State Space

- State space es un espacio como cualquier otro.
- Es diferente del espacio percibido normalmente en dos formas:
 - Primero, no está limitado a tres dimensiones.
 - Segundo, puede ser medido sin importar el número de dimensiones.
- Cuando se construyen state space para tratar con conjuntos de datos, el rango de valores dimensionales es limitado.

Unit State Space

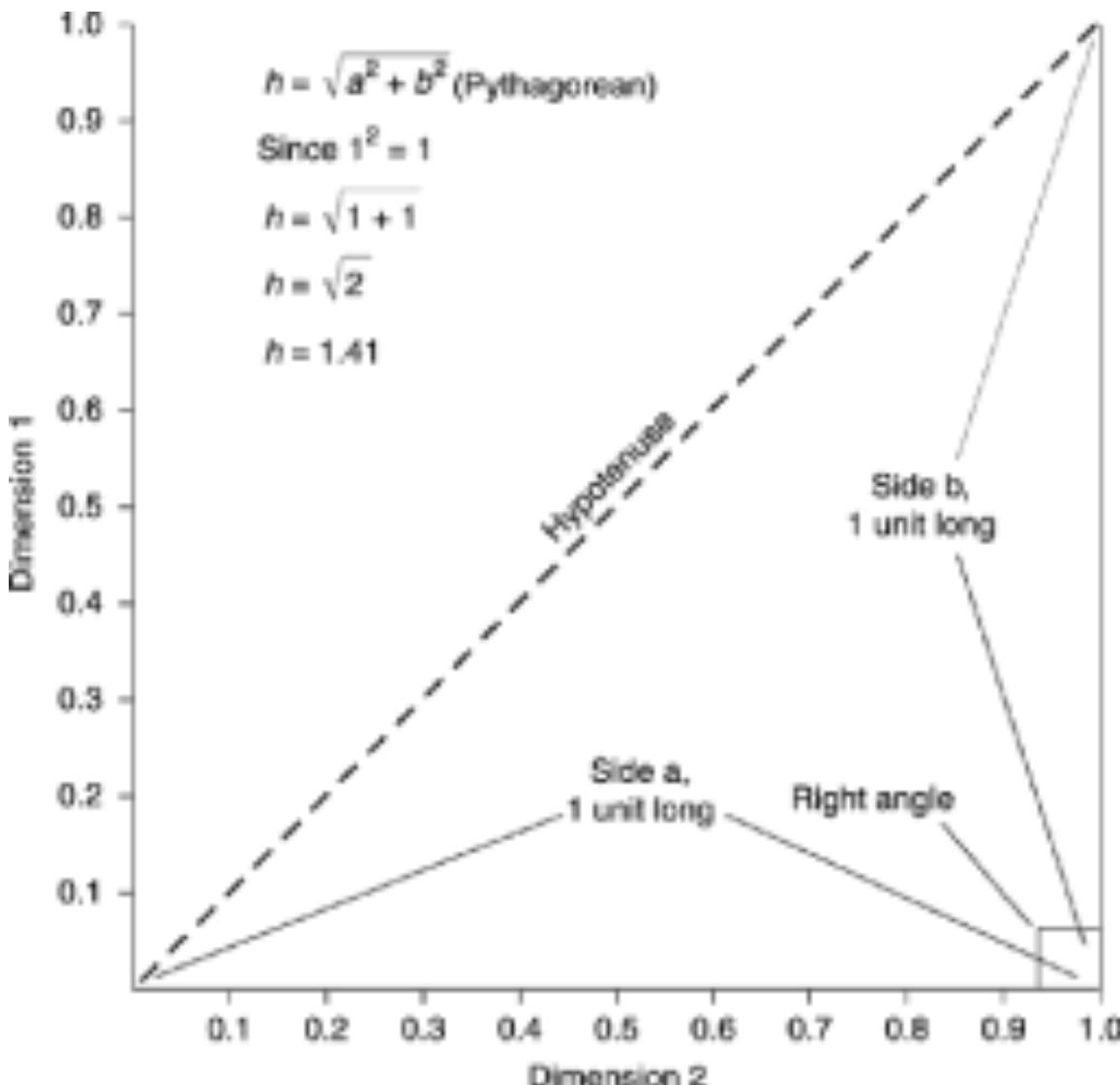
- Normalizar en este contexto significa que cada dimensión puede ser construida de tal forma que sus valores máximo y mínimo sean lo mismo.
- Cuando cada dimensión en el State Space es construida tal que los valores máximos y mínimos para cada rango sean 1 y 0, respectivamente el espacio se denomina Unit State Space.
- El State Space es construido tal que las dimensiones están todas en un ángulo recto entre ellas es así que un State Space de dos dimensiones es rectangular.

Pitagoras en State Space

- El teorema de Pitágoras puede ser extendido a un espacio de tres dimensiones, en el cual la línea diagonal más larga es de 1.73 unidades. ¿en caso de 4 dimensiones?
- El teorema de Pitágoras se cumple para cualquier dimensionalidad de State Space, sin importar el número de dimensiones.

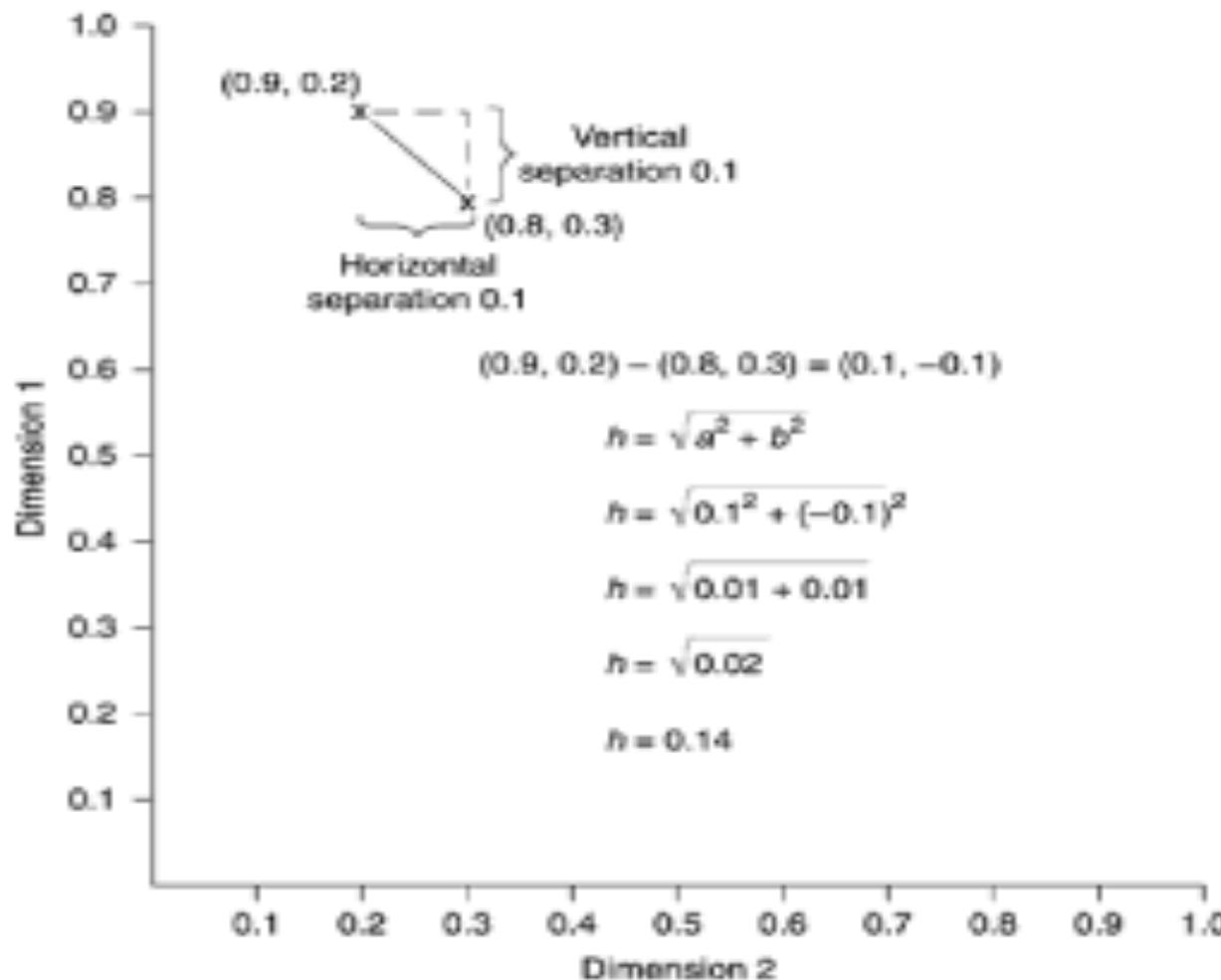
- La línea recta más larga que se puede dibujar en un State Space es siempre la raíz cuadrada de número de dimensiones.

La mayor separación posible en el State Space.



Posición en State Space

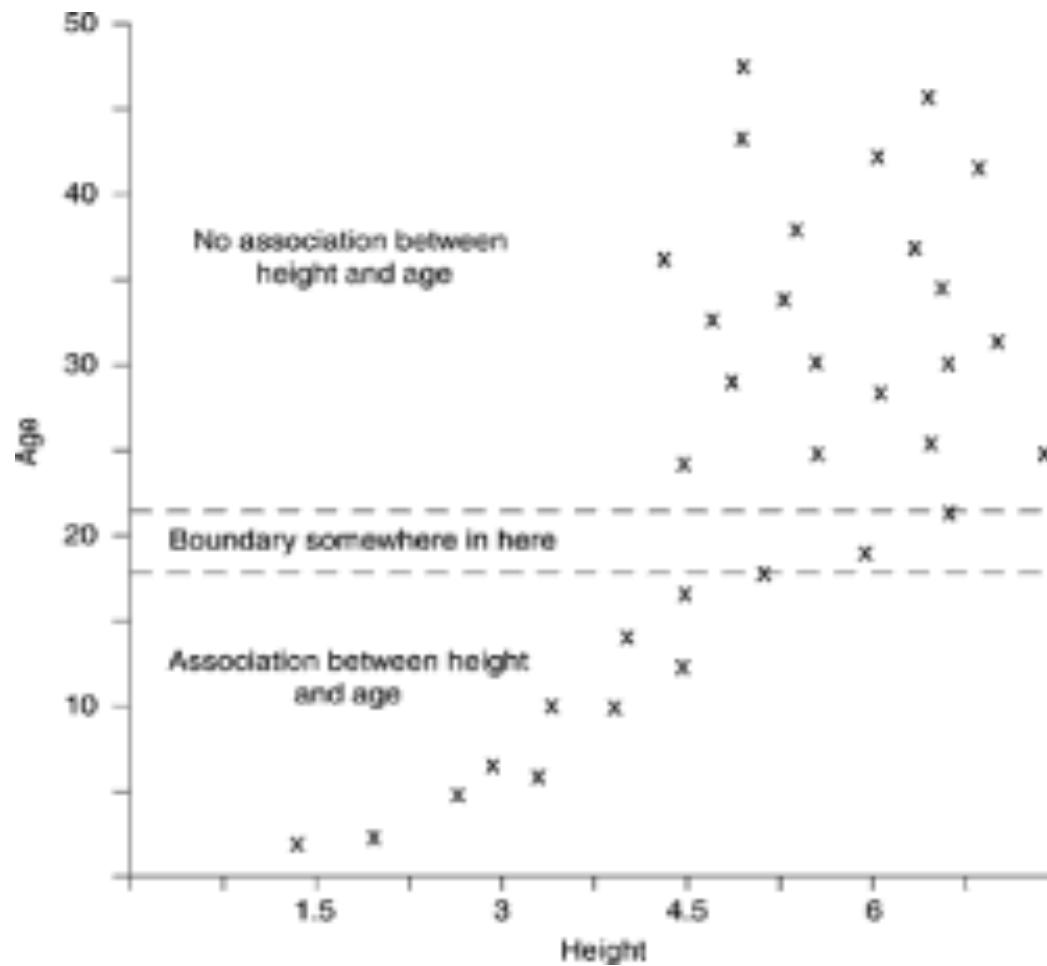
- En vez de encontrar la línea más larga en un State Space, el teorema de Pitágoras puede ser utilizado para encontrar la distancia entre cualquiera de dos puntos.



Encontrando la distancia entre dos puntos en un 2D State Space.

Vecinos

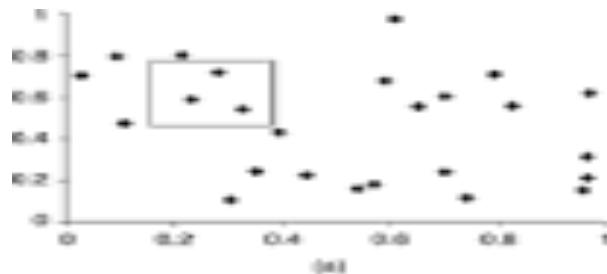
- Puntos en State Space que están cerca el uno del otro se denominan vecinos.



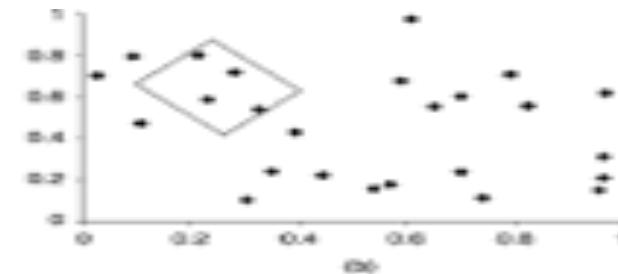
Mostrando la interrelación entre vecinos cuando existe y cuando no entre las variables.

Density and Sparsity

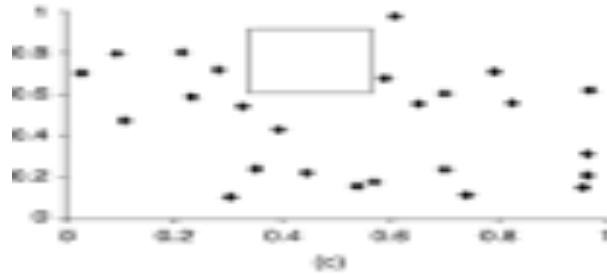
- Una forma de estimar la densidad es elegir un punto o posición y estimar la distancia de este a cada punto cerca en cada dimensión.



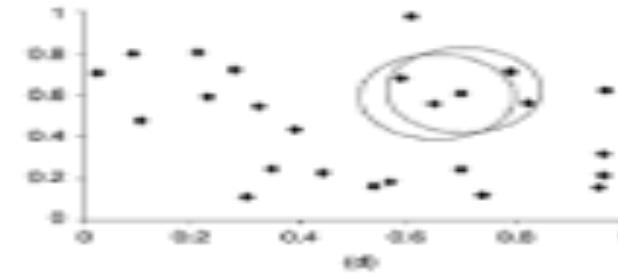
(a)



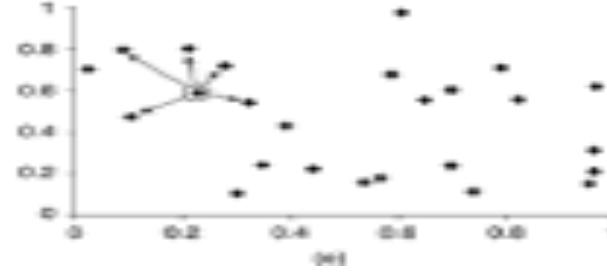
(b)



(c)



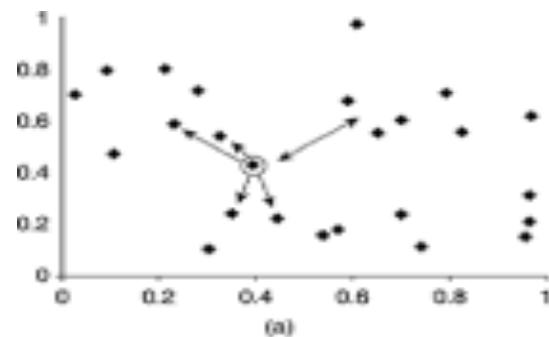
(d)



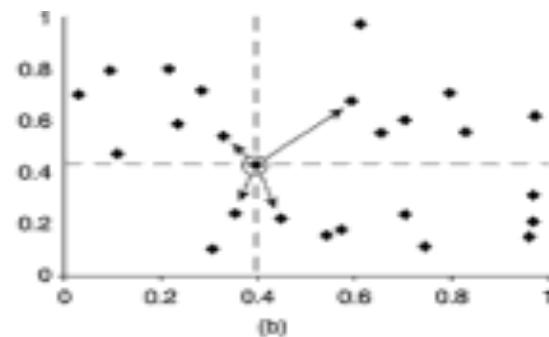
(e)

Vecindad y distancia

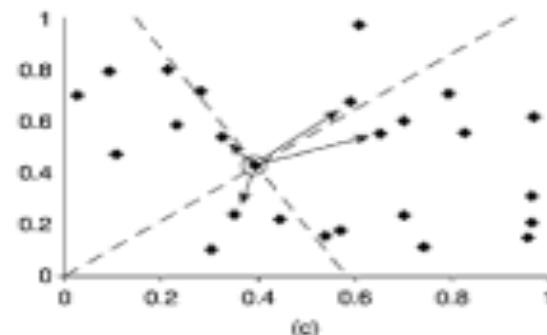
- Una forma más intuitiva de la densidad requiere encontrar los vecinos más próximos en todas las direcciones alrededor de un punto elegido.



(a)



(b)



(c)

Mapeando el State Space

- Es útil pensar que puntos en un State Space corresponden a un objeto geométrico de algún tipo inclusive cuando se piensa en más de tres dimensiones.

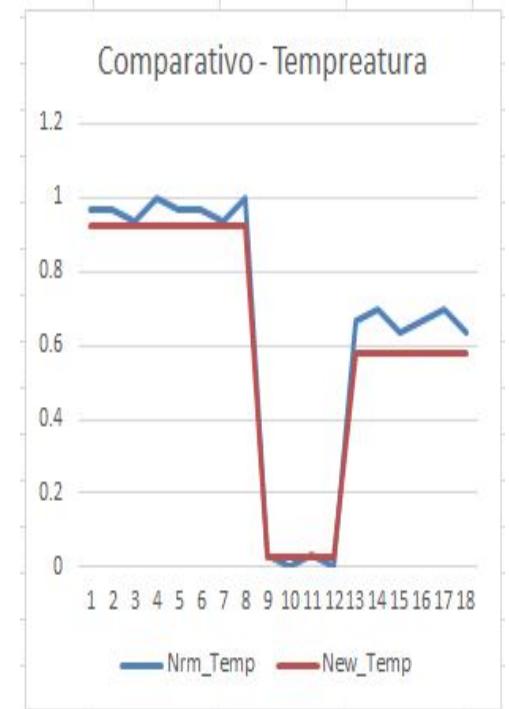
Mapeando valores Categóricos

- La discusión sobre State Space asume dimensiones que son numéricamente con escala y normalizadas en el rango de 0 a 1.
- Entre cualquier par de variables, sean categóricas o numéricas, existe algún tipo de relación.
- Es esta relación entre valores categóricos y el sistema de variables en conjunto que permiten una apropiada numeración.

Temp	Nrm_Temp	C_Temp	Hum	NorHum	New_Temp	Nrm_Temp - New_Temp	C_Temp	Promedios
38	0.96666667	a	40	0.9	0.92142857	0.045238095	a	0.92143
38	0.96666667	a	45	1.0	0.92142857	0.045238095	m	0.58095
37	0.93333333	a	42	0.9	0.92142857	0.011904762	b	0.02857
39	1	a	42	0.9	0.92142857	0.078571429		
38	0.96666667	a	40	0.9	0.92142857	0.045238095		
38	0.96666667	a	45	1.0	0.92142857	0.045238095		
37	0.93333333	a	42	0.9	0.92142857	0.011904762		
39	1	a	42	0.9	0.92142857	0.078571429		
10	0.03333333	b	10	0.0	0.02857143	0.004761905		
9	0	b	12	0.1	0.02857143	-0.028571429		
10	0.03333333	b	10	0.0	0.02857143	0.004761905		
9	0	b	12	0.1	0.02857143	-0.028571429		
29	0.66666667	m	30	0.6	0.58095238	0.085714286		
30	0.7	m	32	0.6	0.58095238	0.119047619		
28	0.63333333	m	29	0.5	0.58095238	0.052380952		
29	0.66666667	m	30	0.6	0.58095238	0.085714286		
30	0.7	m	32	0.6	0.58095238	0.119047619		
28	0.63333333	m	29	0.5	0.58095238	0.052380952		

ESCALAS

Temperatu	Humedad
a 37-39	a 40-45
m 28-31	m 29-32
b 9-10	b 10-12



Ejemplo: Es esta relación entre valores categóricos y el sistema de variables en conjunto que permiten una apropiada numeración.

Localización, Localización, Localización!

- En el espacio real, la localización lo es todo, de la misma forma en el mapeo de variables categóricas y por supuesto, numéricas.

Tablas de distribuciones conjuntas

- Un tipo de problema diferente surge si no existen variables numéricas.
- Cuando existe al menos una variable numérica presente, es utilizada para identificar el orden y espaciamiento de las variables categóricas.
- Sin una variable numérica presente, no existe nada para “calibrar” las variables categóricas.
- El problema es como encontrar algún tipo de ordenamiento lógico que revele las interrelaciones entre las variables categóricas.

La solución viene en pasos:

- Lo primero es descubrir como los valores categóricos de una variable se relacionan a los valores categóricos de otra variable.
- Una forma útil de comenzar es por medio del uso de una tabla de frecuencia conjunta

Normalizando Variables

Normalizar el rango de variables

- Varias herramientas de modelado requieren que el rango de la entrada sea normalizado.

Ej. las redes neuronales artificiales
las máquinas de vectores de sopote.

- Normalización requiere tomar valores de un rango y representarlo en otro.

Valores out-of-range en Training

- Una solución a los valores out-of-range es simplemente ignorarlos.

Existen dos problemas con esta posibilidad:

- El primero y menos significativo, es que reducir el número de instancias en la muestra reduce el nivel de confianza que la muestra sea representativa de la población.
- Un segundo problema, y más serio es la introducción de sesgo.



Transformación de los datos: Normalización

- Normalización mín-máx : a $[nuevo_min_A, nuevo_max_A]$

$$v' = \frac{v - min_A}{max_A - min_A} (nuevo_max_A - nuevo_min_A) + nuevo_min_A$$

- Ej. Normalizar el rango de ingresos \$12,000-\$98,000 a [0.0, 1.0].
Entonces a \$73,600 se le asigna el valor

$$\frac{73,600 - 12,000}{98,000 - 12,000} (1.0 - 0) + 0 = 0.716$$

- Estandarización Z-score (μ : media, σ :desviación típica):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

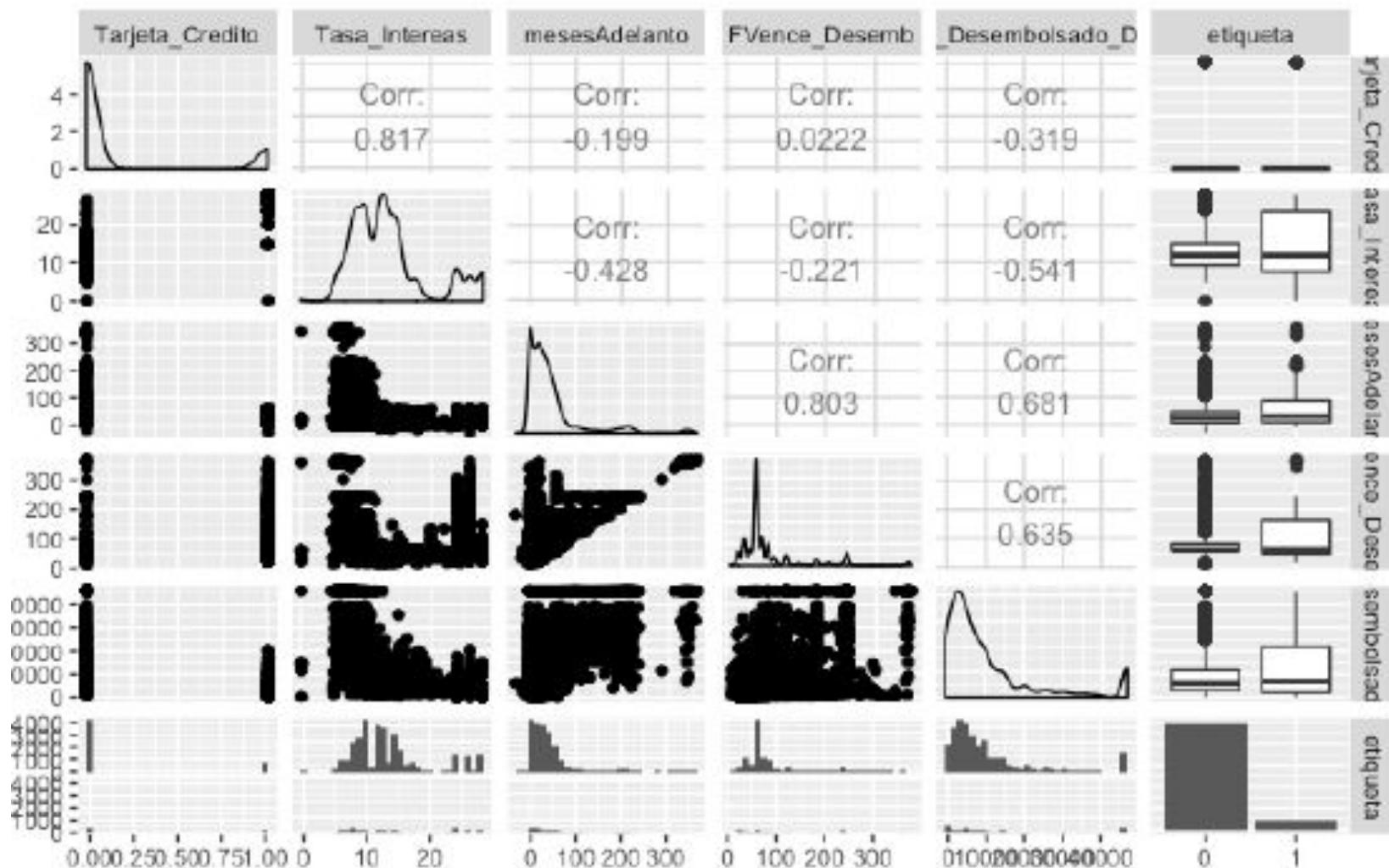
- Normalización por escalado decimal

$$v' = \frac{v}{10^j} \text{ donde } j \text{ es el menor entero tal que } \text{Max}(|v'|) < 1$$

Análisis entre Variables

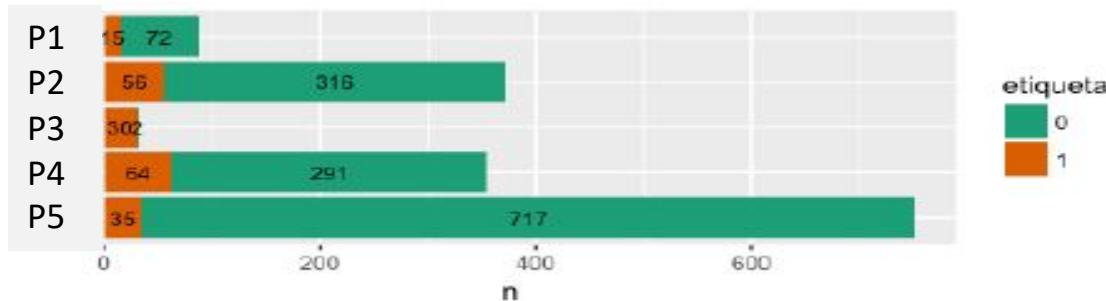
- El análisis entre variables pretende encontrar la interrelación entre las mismas, en términos de correlación de variables.
- Continua vs Continua
- Categórica vs Categórica
- Categórica vs Continua
- Analizar las variables en cuestión respecto a su capacidad de predicción respecto al target o label

Correlación y distribución de variables

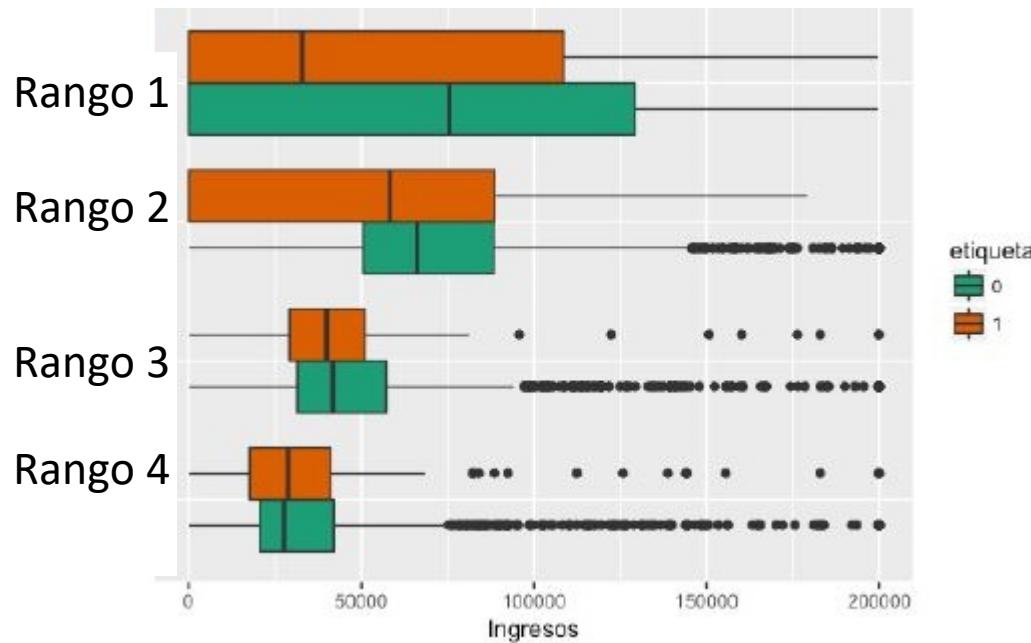


Continua vs Continua / Categórica

Fuga por producto

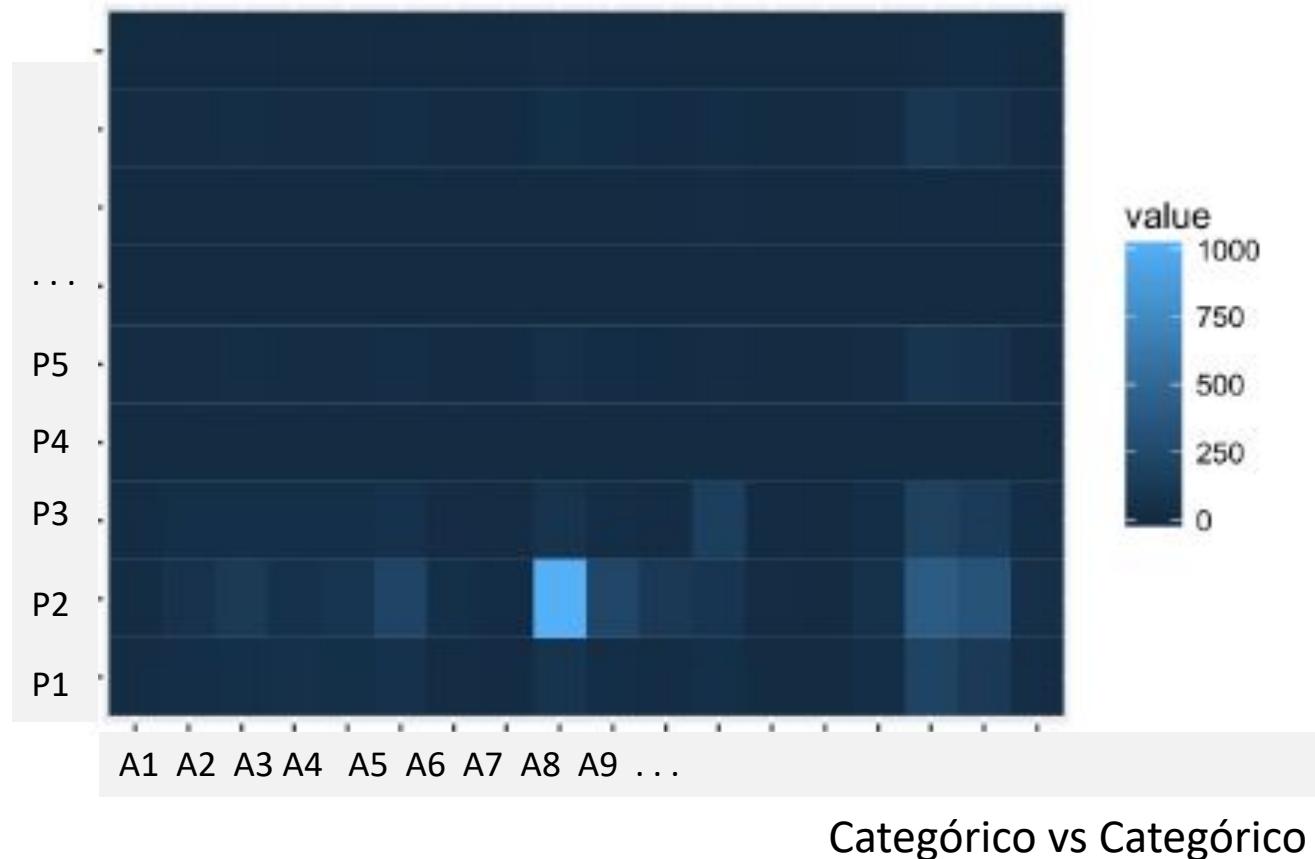


Rango de Crédito en relación a Ingresos



Continua vs Categórica

Producto vs Actividad Económica



Tratamiento de valores faltantes

- Forma de tratamiento
 - Valores faltantes pueden aparecer de varias formas
<empty field> “0” “.” “999” “NA” ...
- Eliminación
- Imputación de Media, Moda o Mediana
- Tratar a los valores faltantes como un valor separado
- Modelo Predictivo
- Imputación KNN

Reemplazo de valores faltantes y vacíos

- Pueden deberse a:
 - mal funcionamiento de equipos
 - datos que eran inconsistentes con otros y por tanto fueron borrados
 - datos no cargados por malos entendidos
 - el que cargó los datos no consideró que eran importantes
- Pueden ser significativos
 - un cliente sin número de teléfono porque quiere que no se le moleste
- Pueden tener que ser inferidos

¿Qué hacer con datos faltantes?

- Dejar pasar
 - algunos algoritmos son robustos ante datos faltantes
- Eliminar el atributo (columna)
 - cuando la proporción de nulos en ella es muy grande
- Ignorar el registro
 - suele hacerse cuando falta la etiqueta de la clase (en clasificación)
 - a veces sesga los datos porque las causas de que sea faltante pueden ser casos especiales.
- Llenarlo manualmente: tedioso + puede ser imposible por el volumen

¿Qué hacer con datos faltantes?

- Llenarlo automáticamente con
 - una constante global: ej. “desconocido”, una nueva clase?!
 - la mediana del atributo
 - la media del atributo para todas las instancias que pertenezcan a la misma clase: mejor.
 - La desviación estándar
 - el valor más probable: inferirlo utilizando métodos bayesianos, regresión lineal, K-NN o árboles de decisión.

Reteniendo información acerca de valores faltantes

Los valores faltantes deben ser reemplazados por varias razones:

- Primero, algunas técnicas de modelado no tratan con estos valores faltantes.
- Segundo, Las herramientas de modelado que usan métodos de reemplazo por default pueden introducir distorsión.
- Cuarto, La mayoría de los métodos de reemplazo descartan la información contenida en los patrones de valores faltantes.

Missing-Value Patterns

- Reemplazar valores faltantes difumina el hecho de ser faltantes.
- Puede suceder que el patron de valor faltante se convierta en el aspecto más importante durante el modelado.

Captura de patrones

- El missing-value pattern (MVP) es exactamente el patrón en el cual las variables tienen valores faltantes

Reemplazo de valores faltantes

- En la práctica, una de las partes que consume más tiempo de la preparación automática de datos es el reemplazo de valores faltantes.

Relación entre variables

- Como un sistema de variables, existe una relación que conecta los valores de variables entre ellos.
- Asignar un valor constante a una variable para todos sus valores faltantes ciertamente distorsiona la relación entre variables.
- Lo que es más, si los valores faltantes no son aleatoriamente faltantes usar un reemplazo tal que todos los valores faltantes tengan el mismo valor no solamente es inapropiado sino genera distorsión.
- Dados múltiples MVPs varias variables pueden tener simultáneamente valores faltantes, pero nunca todas las variables de una vez.

- Cualquier valor de variable que esté presente puede ser utilizado para estimar cual es el nivel apropiado de valor faltante que debería tener la variable.
- El propósito de reemplazar los valores faltantes no es usar los valores como tal, es hacer disponible a las herramientas de modelado la información contenida en otros valores de variables que están presentes.

ANALISIS DE INFORMACIÓN FALTANTE Y VALORES POR DEFECTO

- Información faltante es parte del mundo real.
- “null” **no es un valor**, es una marca o bandera.
- El concepto de “nulls” implica una lógica de tres valores: true, false, unknown.
- Las tablas de verdad en la lógica de tres valores:

AND	t	u	f
t	t	u	f
u	u	u	f
f	f	f	f

OR	t	u	f
t	t	t	t
u	t	u	u
f	t	u	f

NOT	
t	f
u	u
f	t

INFORMACIÓN FALTANTE Y VALORES POR DEFECTO

- Ejemplos de tipos de nulls:

1. **Valor no aplicable:** En emp(e#, ..., comisión, ...), la comisión solamente se aplica a los empleados en el Departamento de “Ventas”.
2. **Valor desconocido:** En emp (e#, ..., salario, ...) el salario de Juan es desconocido.
3. **Valor no existe:** No todas las personas cuentan con el número de seguro social.
4. **Valor indefinido:** en el caso de división por cero.
5. **Valor no provisto:** En un censo, “Se negó a contestar”, ‘Sin comentarios’.

INFORMACIÓN FALTANTE Y VALORES POR DEFECTO

7. Valor es el conjunto vacío: El OUTER JOIN natural de Departamento y empleado sobre el número de Departamento, asumiendo que por ejemplo el Departamento 'D5' no tiene empleados.

...

y ciertamente un número infinito de otras posibilidades de tipos de nulls.

- Los operadores del Álgebra Relacional (i.e. PRODUCT, JOIN, PROJECT, etc.) no se comportan de una forma correcta.

INFORMACIÓN FALTANTE Y VALORES POR DEFECTO

- Dada la variable lógica v del tipo “Valor de verdad”. Si el valor de v es unk, entonces se conoce que el valor de v es unk. Pero en el caso en que v sea UNK, entonces no se conoce si el valor de v es Verdad, Falso o unk, por lo tanto:

if v is unk then $v = v$ es Verdad

if v is UNK then $v = v$ es unk

- Un **dominio** no puede contener UNK, los dominios son conjuntos de valores:

→ Una relación que incluya UNK **no** es una relación.

Ejemplo: Dada una fecha F , la expresión

$(F < '03-03-2004') \text{ OR } (F = '03-03-2004') \text{ OR } (F > '03-03-2004')$

Sin importar el valor que tenga F esta expresión es Verdadera, pero si F contiene null, entonces la expresión es null.

INFORMACIÓN FALTANTE Y VALORES POR DEFECTO

Ejemplo:

DEPT	DEPT#	EMP	EMP#	DEPT#
	D2		E1	UNK

- Considere la expresión (como parte de un query):
 $\text{DEPT.DEPT\#} = \text{EMP.DEPT\#} \text{ AND } \text{EMP.DEPT\#} = \text{DEPT\#('D1')}$
- Un buen optimizador identificaría que este query tiene la forma $A=B \text{ AND } B=C$, lo cual resultaría en $A=C$
 $\text{DEPT.DEPT\#} = \text{EMP.DEPT\#} \text{ AND } \text{EMP.DEPT\#} = \text{DEPT\#('D1')}$
 $\text{AND DEPT.DEPT\#} = \text{DEPT\#('D1')}$
Esta expresión es siempre FALSE sea cual sea el valor real de UNK.

INFORMACIÓN FALTANTE Y VALORES POR DEFECTO

- Entonces la expresión siguiente:

```
EMP.EMP# WHERE EXISTS DEPT (  
NOT (DEPT.DEPT# =EMP.DEPT# AND EMP.DEPT# = DEPT#('D1')))
```

retornaría E1 si la “optimización” se hizo en el sentido considerado y UNK no fuera una marca sino un valor.

- Si el valor de un atributo dado dentro de una tupla dentro de una relvar es UNK, implica que el atributo no es un atributo, la tupla no es una tupla y la relvar no es una relvar.

→ **UNK y 3VL van en contra de todo el fundamento del Modelo Relacional**

INFORMACIÓN FALTANTE Y VALORES POR DEFECTO

- Identidades en 2VL, que **no** son identidades en 3VL.

p and not p	falso
p or not p	verdad
x = x	verdad
x <> x	falso
x < y or x = y or x > y	verdad
x = y and y = z	x = y and y = z and x = z
x < y and y < z	x < y and y < z and x < z
x - x	0
x / x	1
x + y > x	verdad
R JOIN R	R
R INTERSECT S	R JOIN S(*)

(*) INTERSECT ya no es un caso especial de JOIN.

INFORMACIÓN FALTANTE Y VALORES POR DEFECTO

- Las funciones agregadas en SQL(i.e. SUM, AVG, MAX y MIN) de un resultado de consulta vacío retorna null. Estos operadores se deberían evaluar de la sigte. forma:

sum (Φ) se evalúa a cero max (Φ) se evalúa a $-\infty$

min (Φ) se evalúa a $+\infty$ avg (Φ) se evalúa a error

INFORMACIÓN FALTANTE Y VALORES POR DEFECTO

Otro ejemplo: “Obtenga las partes en las cuales su peso no es igual al peso de las partes de ‘Tarija’”

1.

```
SELECT P.*  
      FROM P  
     WHERE P.PESO NOT IN  
           (SELECT P.PESO  
            FROM P  
           WHERE CIUDAD = 'Tarija');
```

2.

```
SELECT P.*  
      FROM P  
     WHERE NOT EXISTS  
           (SELECT Q.*  
            FROM P Q  
           WHERE CIUDAD = 'Tarija'  
             AND P.PESO = Q.PESO);
```

INFORMACIÓN FALTANTE Y VALORES POR DEFECTO

- Del ejemplo, suponga que hay solamente una parte en Tarija con un peso null, en la consulta 1, el subquery retorna un conjunto que contiene solo unk, por lo tanto ninguna parte es recuperada.
- En cambio en 2, el subquery retorna el conjunto vacío, el EXISTS retorna Falso, por lo tanto todas las partes son recuperadas.

INFORMACIÓN FALTANTE Y VALORES POR DEFECTO

Valores por defecto

- Valores “por defecto” para tratar con información faltante.
- Se quita completamente la idea de nulls, para utilizar **“valores especiales”** para representar la información faltante, con las ventajas siguientes:
 - Intuitivamente más sencillo de entender
 - Sencilla su implementación
 - Refleja la información faltante en el mundo real
 - Es extensible a otros tipos de información faltante, sin la necesidad de lógica de n valores.

VALORES FALTANTE Y POR DEFECTO

- Los valores por defecto no deberían ser codificados en programas. Lo que se necesita es una forma de referirse a estos valores simbólicamente.
- El valor UNK para una columna dada debe ser un valor del dominio en cuestión.

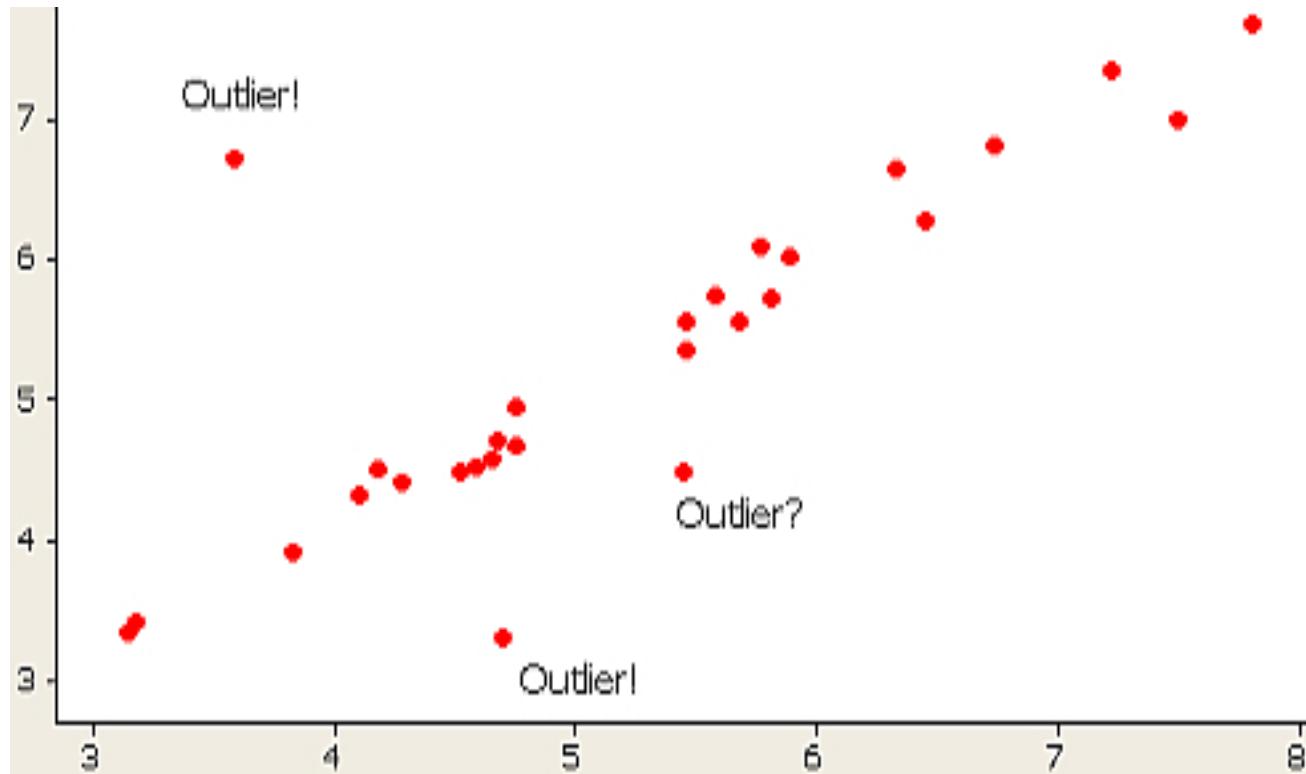
Tratamiento de outliers

- Principalmente graficación como medio de identificar outliers de tipo uni-variable, o multi-variable
- Impacto de los outliers en el dataset

Como quitarlos:

- Borrar las observaciones
- Transformar la variable discretizandolas
- Imputar outliers
- Tratarlos en forma separada

Outliers y posibles outliers



10. SELECCIÓN DE ATRIBUTOS

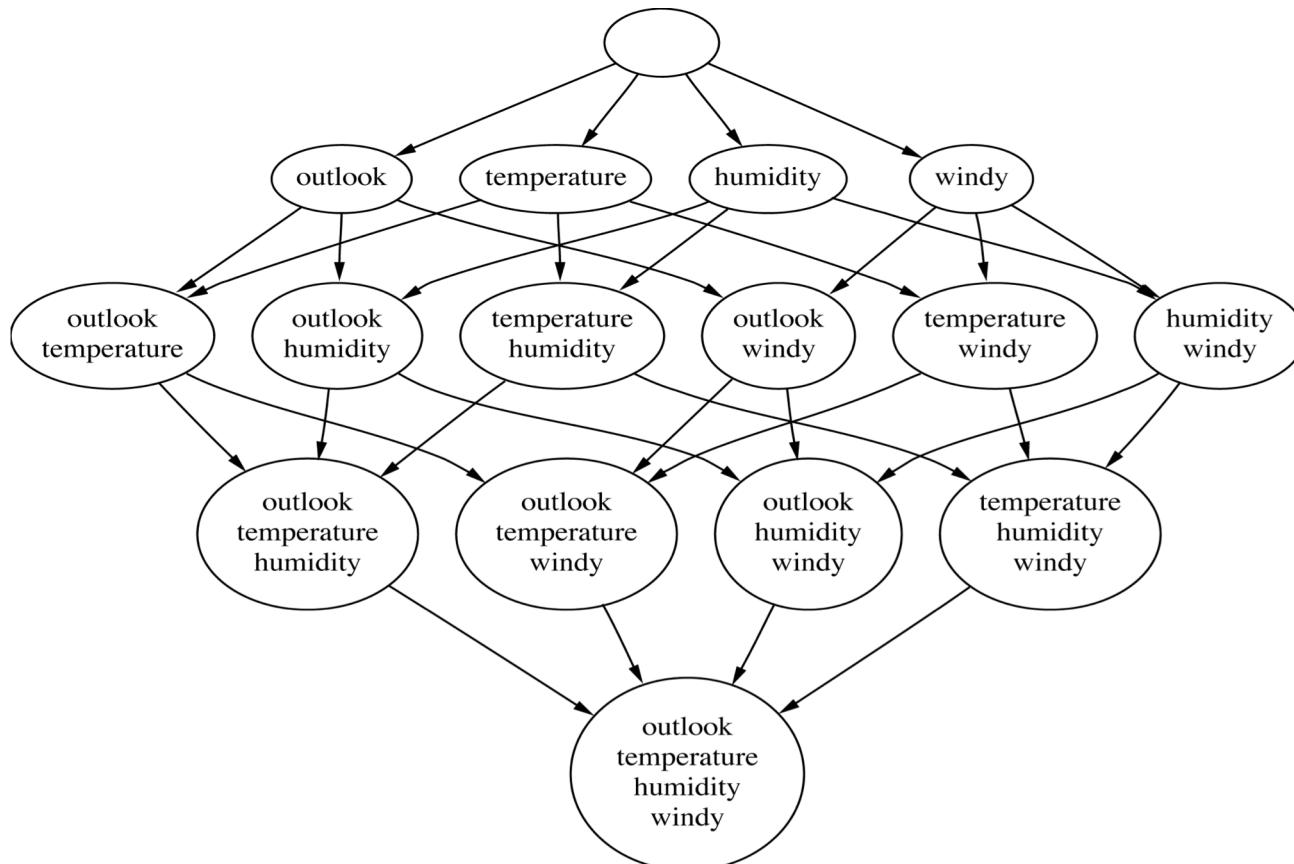
- Muchas de las técnicas de CD son utilizadas para identificar los atributos más apropiados para la clasificación, regresión, clustering, etc.
- Se han realizado experimentos con árboles de decisión (algoritmo C4.5), los cuales mostraron que la inclusión de atributos binarios generados aleatoriamente afectan el rendimiento en la clasificación en un 5% a 10%.
- Inclusive cuando los atributos son relevantes para el problema pueden también causar dificultades. Por ejemplo cuando en un dataset de dos atributos se incluye uno que tenga similares valores a la clase puede deteriorar la clasificación en un 1% a 5%.

- La mejor forma de elegir atributos para los procesos de CD es manualmente, basada en un entendimiento profundo del problema y el significado de los atributos.
- Cuando se selecciona un buen subconjunto de atributos, existen dos perspectivas:
 - Una es hacer una valoración independiente basada en características generales de los datos, en este caso se denominan métodos de filtrado **FILTERS**.
 - La otra es evaluar el subconjunto usando algoritmos de CD. en este caso se denominan métodos **WRAPPER**.

- Tiene sentido seleccionar el subconjunto más pequeño que distingue a todas las instancias en forma única.
- Algoritmos de DS pueden utilizarse para la selección de atributos, por ejemplo, se puede aplicar inicialmente un algoritmo de árboles de decisión al dataset para luego elegir solamente los atributos que fueron seleccionados en el árbol
- Los algoritmos de “nearest neighbor” son notoriamente susceptibles a atributos irrelevantes, generalmente se usa previamente algoritmos de árboles de decisión para seleccionar atributos con la mayor capacidad de predicción.

Buscando en el espacio de atributos

- La mayoría de los métodos de selección de atributos involucran la búsqueda del espacio de atributos que prediga la clase de mejor forma.



- Algunos métodos:
 - Aproximación por filtro: Evaluación basada en las características generales de los datos.
 - Encontrar el subconjunto más pequeño de atributos que separan los datos.
 - Utilizar los atributos seleccionados con C4.5, 1R
- El número de subconjuntos de atributos es exponencial en número de atributos
- Comunmente se utiliza:
 - *forward selection*
 - *backward elimination*

5. Transformación de variables (ing. Feature Engineering)

Transformación de variables

- Discretizar variables
- Utilizar el log, Raíz Cuadrada para la transformación
- Transformación de variables
 - ej. Fecha: convertir en mes , día, año, día de la semana, etc.
- Creación de variables dummy

6. Creación de variables (ing. Feature Engineering)

- La creación de variables que enrriquezcan la predicción
 - Ingreso / Crédito
 - Nro de cuartos / Precio
 - Ingresos * Categorización ASFI
- Adición de datos de Migración y de salubridad en Caso Malaria