TEC | Tecnológico de Costa Rica

Aprendizaje Automático

Tecnológico de Costa Rica

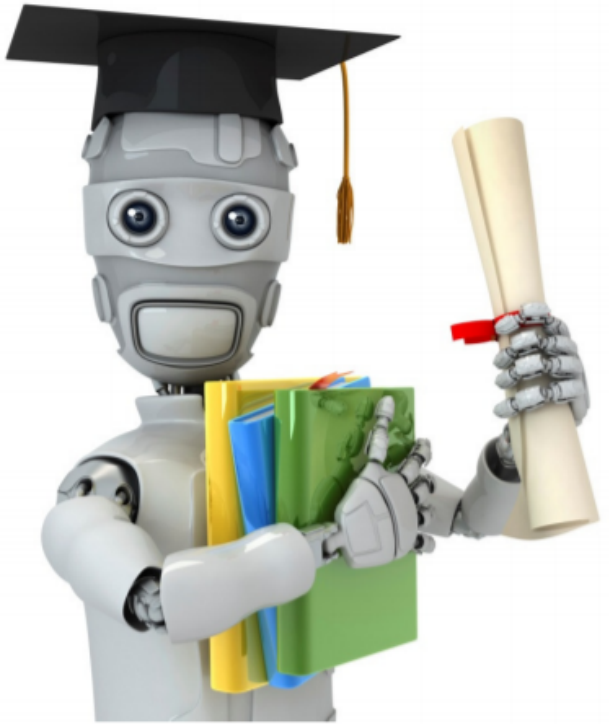
Programa de Ciencia de Datos

Frans van Dunné

Agenda

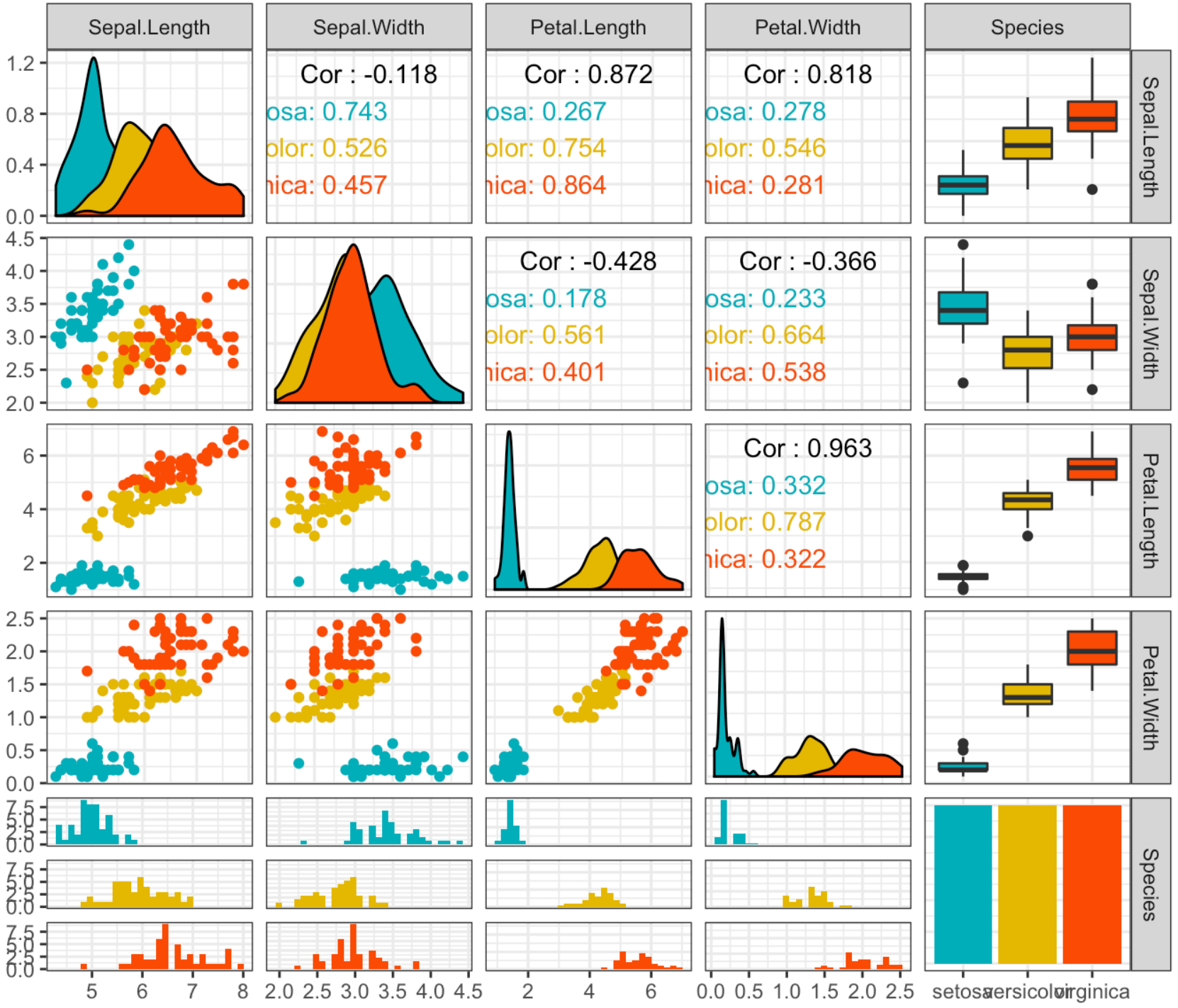
TEC | Tecnológico de Costa Rica

- **Aprendizaje Automático**
  - Etapa de preprocesamiento
  - Categorización y discretización de datos
  - Normalización
  - Eliminación de sesgos, redundancia y ruido



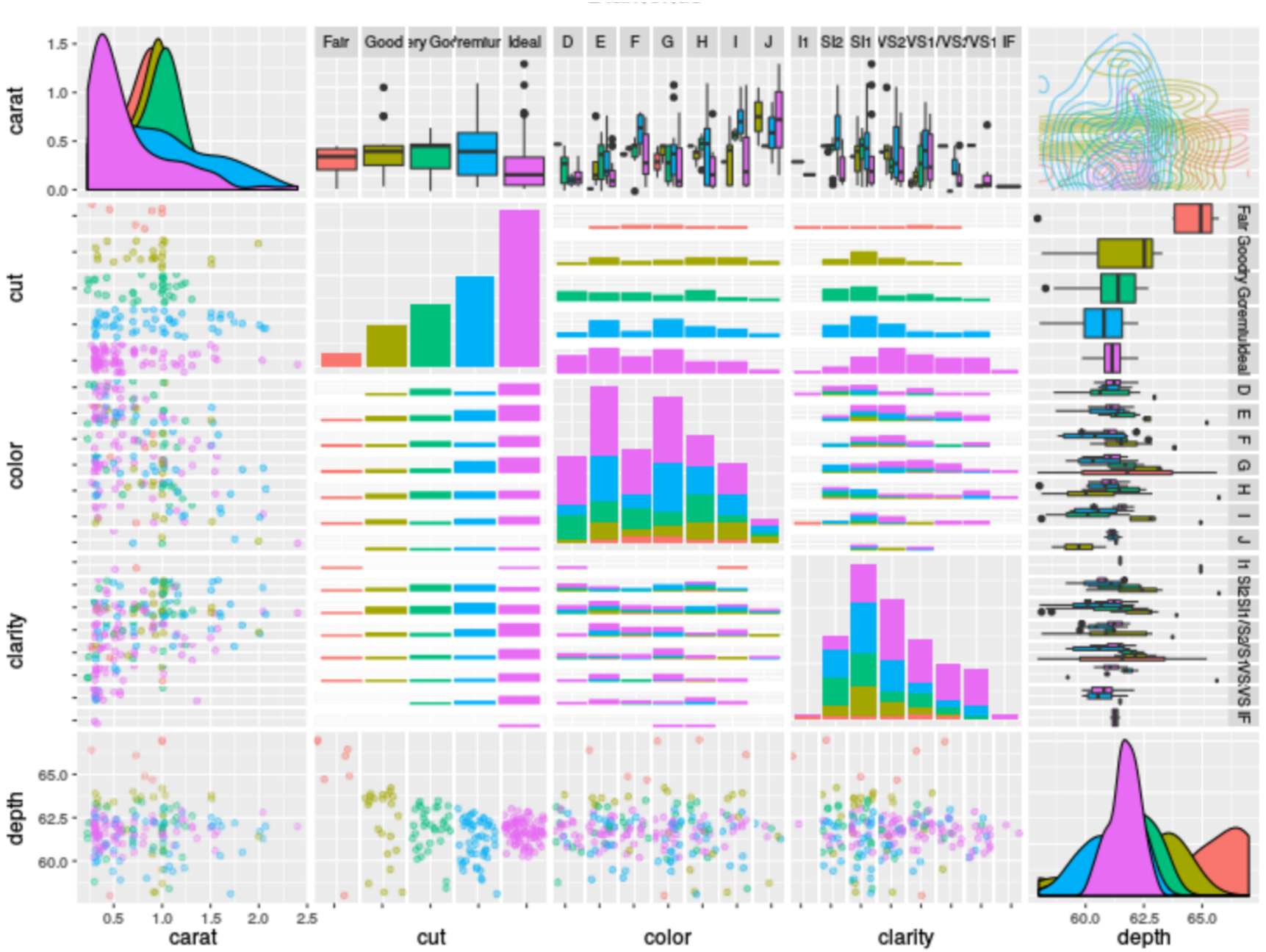
Ejemplos

http://www.sthda.com/english/sthda-upload/figures/r-graphics-essentials/008-plot-multivariate-continuous-data-r-graphics-cookbook-and-examples-for-great-data-visualization-scatter-plot-matrix-by-groups-ggpairs-1.png



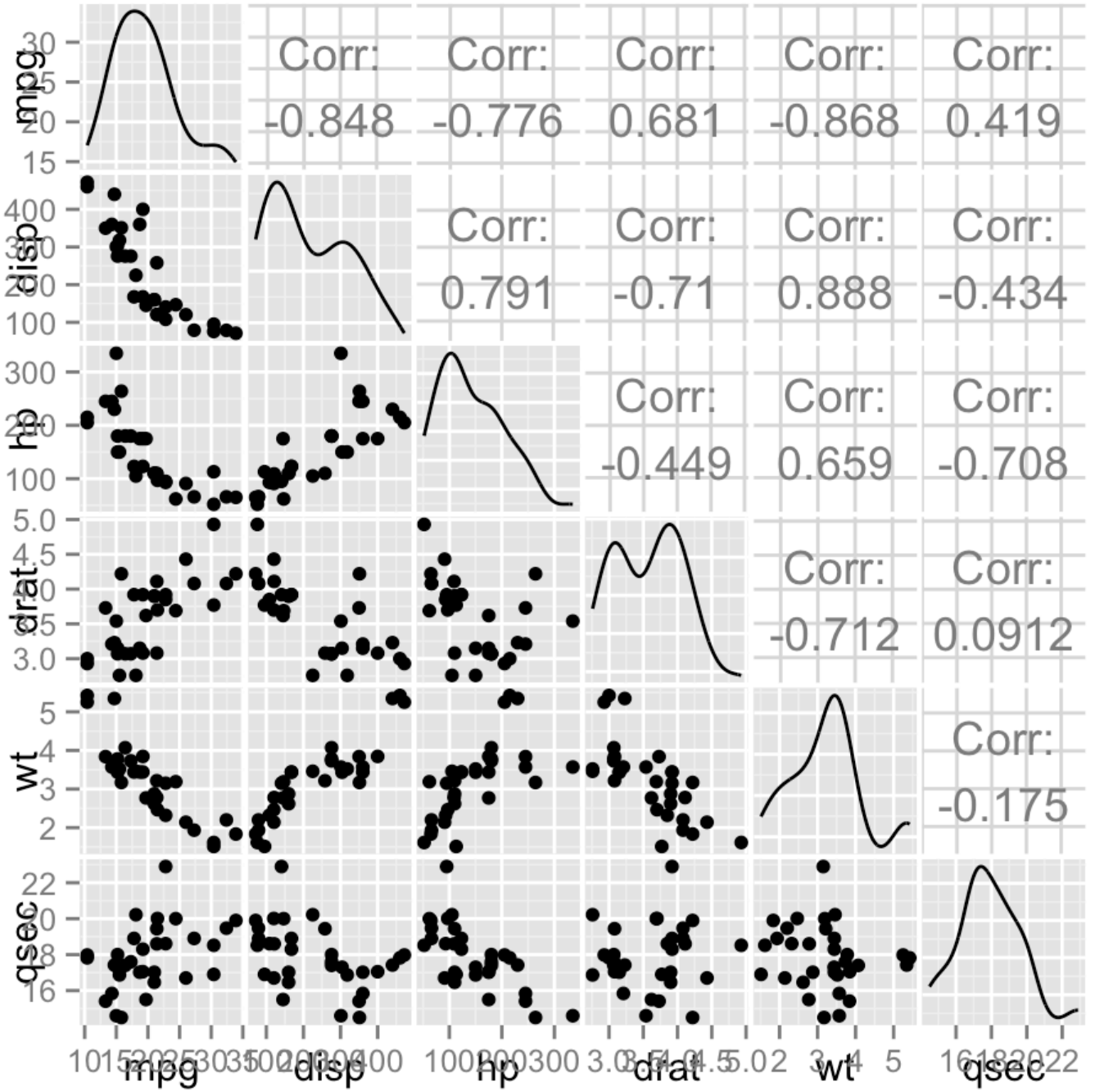
Ejemplos

https://stackoverflow.com/questions/34727408/edit-individual-ggplots-in-ggallyggpairs-how-do-i-have-the-density-plot-not-f/34853734



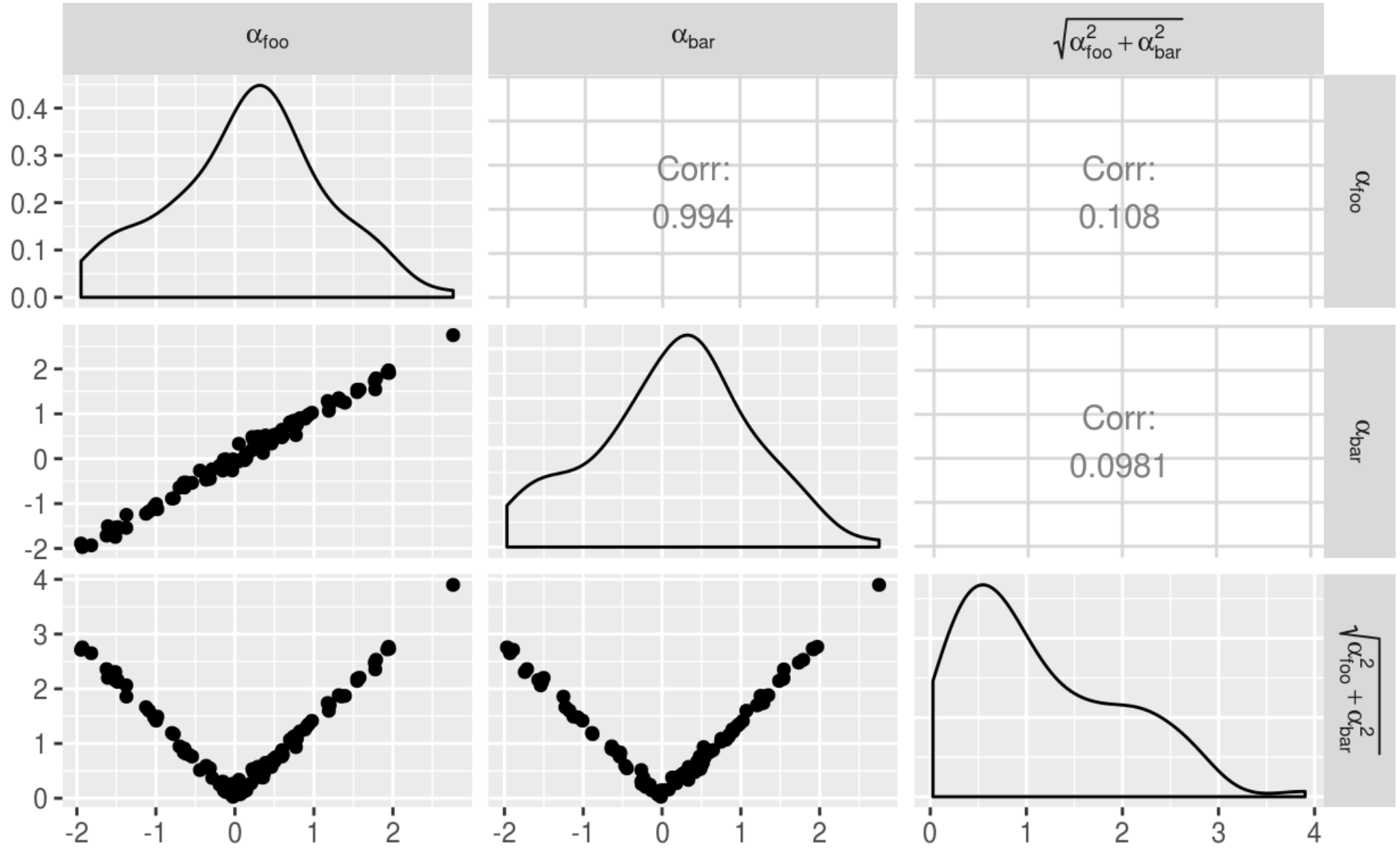
Ejemplos

http://www.sthda.com/english/wiki/ggally-r-package-extension-to-ggplot2-for-correlation-matrix-and-survival-plots-r-software-and-data-visualization



Ejemplos

https://ggobi.github.io/ggally/rd.html

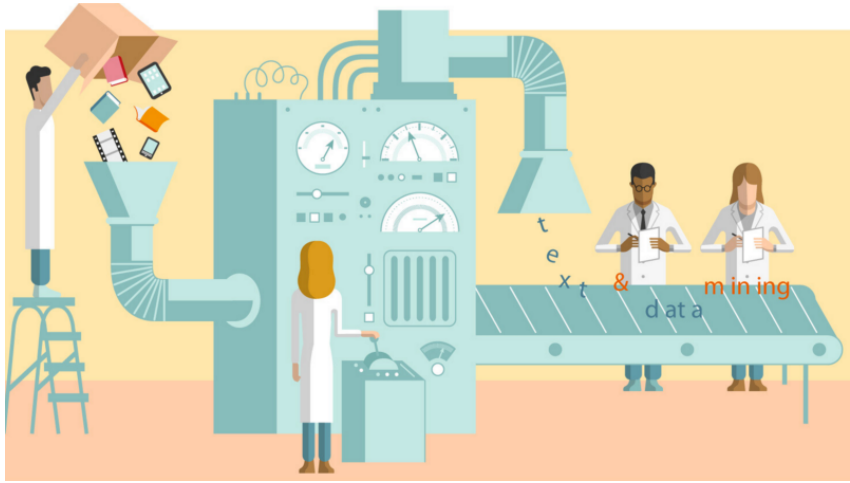


¿Por qué preparar los datos?

TEC | Tecnológico de Costa Rica

- Algún tipo de preparación de datos siempre es necesario para la mayoría de herramientas de ciencia de datos.

- El propósito de la preparación es transformar los conjuntos de datos de tal forma que la información que contienen esté mejor expuesta para la herramienta de ciencia de datos que se utilizará.



¿Por qué preparar los datos?

TEC | Tecnológico de Costa Rica

- Los errores de predicción deberían ser menores

- La preparación de datos también prepara al analista para producir mejores modelos y de manera más rápida.

- Muchas veces se trabaja con datos que tienen muchos años de antigüedad que contienen campos que ya no son válidos, relevantes o tienen datos perdidos.

¿Por qué preparar los datos?

TEC | Tecnológico de Costa Rica

- Los datos en la vida real están sucios:
  - **incompletos**: Falta de valores en los atributos, carecen de algunos atributos de interés, sólo contienen datos agregados:
    - ej., ocupación = “ ”
  - **anómalos**: errores y outliers
    - ej., Salario = -10
  - **inconsistentes**: contienen discrepancias en códigos y nombres
    - ej., Edad = 42 , Cumpleaños = 03/07/1997
    - ej., Rating previo 1,2,3, Rating actual A, B, C
    - ej., Discrepancia con registros duplicados
- Dependiendo del conjunto de datos, la etapa de preparación pueden tomar entre un 10% − 60% de todo el tiempo dedicado al proceso de ciencia de datos.

¿Por qué los datos están sucios?

TEC | Tecnológico de Costa Rica

- Los datos incompletos pueden venir de
  - Datos "No aplicables" al momento de ser colectados.
  - Diferentes consideraciones de tiempo cuando fueron recolectados y cuando son analizados
  - Problemas Humanos/hardware/software
- Datos anómalos (valores incorrectos) pueden venir de
  - Instrumentos de recolección de datos defectuoso
  - Errores humanos o de computadora en la entrada de los datos
  - Errores en la transmisión de datos
- Datos inconsistentes pueden venir de
  - Diferentes fuentes de datos
- Violación de dependencias funcionales (ej., modificación en algunos datos relacionados)
- Registros duplicados también necesitan ser limpiados

Medidas de Calidad

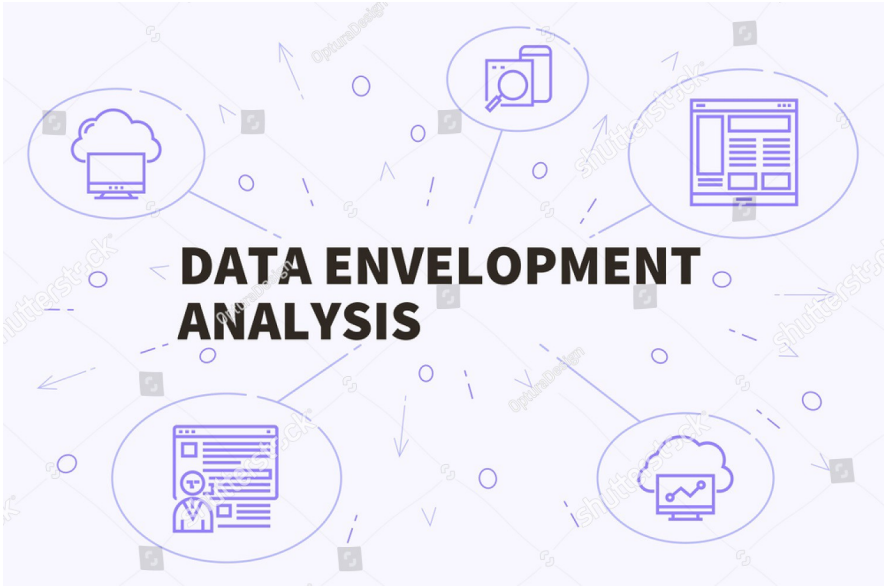
TEC | Tecnológico de Costa Rica

- **Exactitud**
- **Exhaustividad**
- **Consistencia**
- **Puntualidad**
- **Credibilidad**
- **Valor agregado**
- **Interpretabilidad**
- **Accesibilidad**

Etapa de preprocesamiento

TEC | Tecnológico de Costa Rica

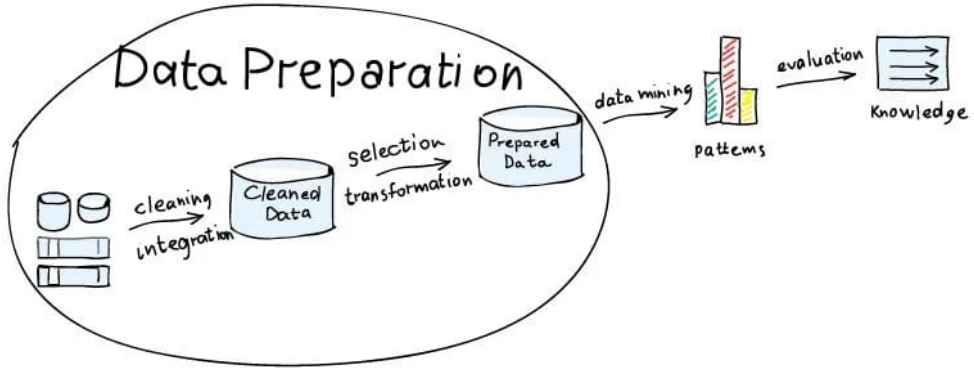
- **Limpieza de datos**
  - Completa valores faltantes, suavizar datos ruidosos, identificar o remover outliers y resolver inconsistencias.
- **Integración de datos**
  - Integración de múltiples bases de datos, cubos de datos, archivos.
- **Transformación de datos**
  - Normalización y agregación (totalización)
- **Reducción de datos**
  - Se obtiene una representación más reducida en volumen pero que produce los mismos o similares resultados analíticos.
- **Discretización de datos**
  - Parte de la reducción de datos pero con particular importancia, especialmente para datos numéricos



CRISP-DM: Entendimiento de los Datos

TEC | Tecnológico de Costa Rica

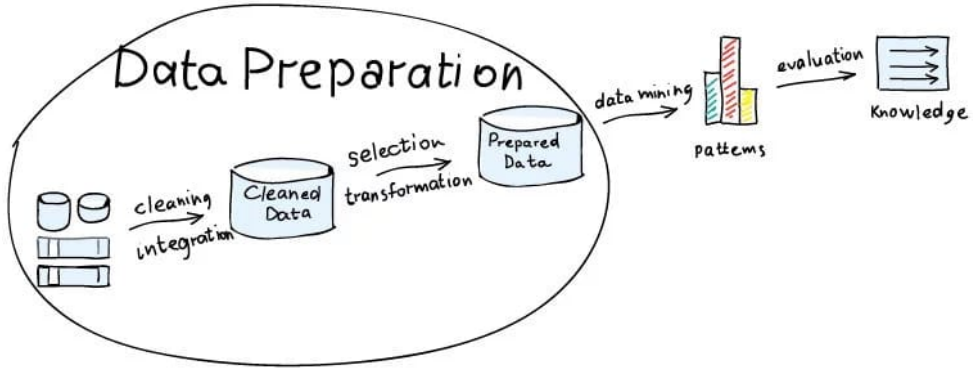
- **Recolección de datos**
  - Enumerar los conjuntos de datos adquiridos (lugares, los métodos utilizados para la adquisición, problemas encontrados y las soluciones alcanzadas).
- **Descripción de los datos**
  - Verificar el volumen de los datos y examinar sus propiedades.
  - Accesibilidad y disponibilidad de los atributos. Tipos de atributos, rango, correlaciones, identidades.
  - Comprender el significado de cada atributo y de los valores de los atributos en términos del negocio.
  - Para cada atributo, calcular estadísticos básicos (ej. promedio, máximo, mínimo, desviación estándar, varianza, moda, sesgo, etc.)



CRISP-DM: Entendimiento de los Datos

TEC | Tecnológico de Costa Rica

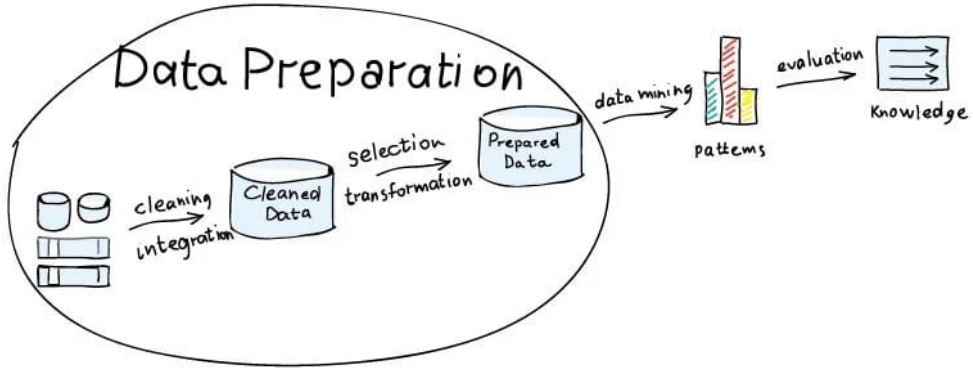
- **Exploración de datos**
  - Analizar en detalle las propiedades de los atributos de interés.
  - Distribuciones, relaciones entre pares de atributos, propiedades de subpoblaciones significativas, análisis estadístico simple.
- **Verificar la calidad de los datos**
  - Identificar valores especiales y catalogar su significado.
  - ¿Se cuenta con todos los casos requeridos? ¿Estos contienen errores y que tan comunes son?
  - Identificar atributos perdidos y en blanco. Significado de datos perdidos.
  - ¿El significado de los atributos y los valores que contienen encajan?
  - Verificar la escritura de los valores (ej. un mismo valor pero a veces empieza con letras mayúsculas, otras con letra minúscula)
  - Verificar la factibilidad de los valores, ej. todos los campos tienen los mismos o casi los mismos valores.



CRISP-DM: Entendimiento de los Datos

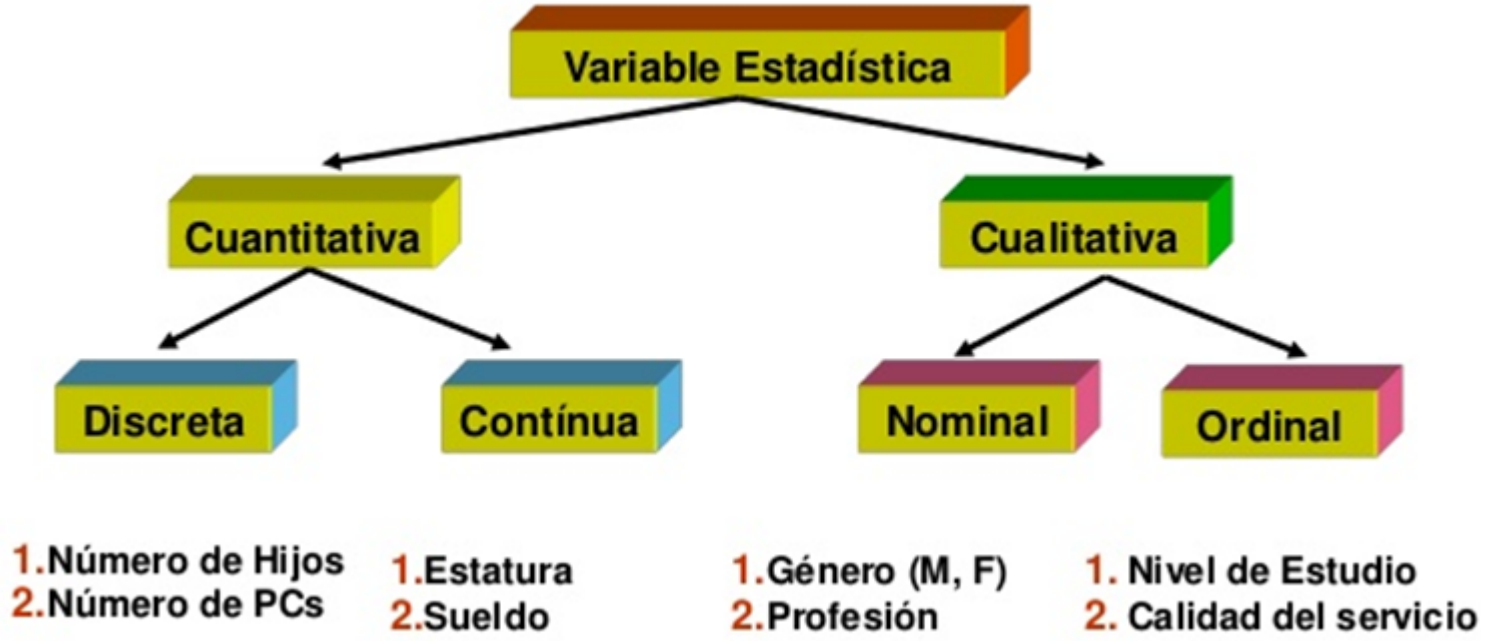
TEC | Tecnológico de Costa Rica

- **Selección de datos**
  - Reconsiderar el criterio de selección de los datos.
  - Decidir el conjunto de datos que será usado.
  - Recolectar data adicional que sea apropiada (interna o externa).
  - Considerar el uso de técnicas de muestreo.
  - Explicar por qué ciertos datos son incluidos o excluidos.
  - ¿Los datos perdidos pueden imputarse o reconstruirse?
- **Formato de Datos**
  - Reordenamiento de los atributos (Algunas herramientas tienen requerimientos en relación al orden de los atributos, ej. El primer campo debe ser un identificar único para cada registro o el último campo debe ser la variable respuesta a ser predicha).
  - Reordenamiento de registros (Puede ser que la herramienta de modelamiento requiera que los registros estén ordenados de acuerdo al valor de la variable respuesta)
  - Reformateo de valores (Cambios puramente sintácticos para satisfacer los requerimientos de una herramienta específica de modelamiento, ej. NA para datos perdidos en vez de 99, remover caracteres ilegales, letras mayúsculas o minúsculas, etc.)



Escalas de Medición

TEC | Tecnológico de Costa Rica



Escala Nominal

TEC | Tecnológico de Costa Rica

- Los valores de la variable clasifican a las unidades estadísticas en iguales o diferentes.
- Los valores de la variable funcionan simplemente como etiquetas que identifican a los distintos valores de las variables, por lo que incluso estos no necesitan ser números.
- Estas variables pueden transformarse preservando esta información mediante funciones 1-1.

- Ejemplos
  - Sexo: 1 = femenino; 2 = masculino.
  - Estado civil: 1 = casado; 2 = soltero; 3 = viudo; 4 = otro.
  - Especialidad de un alumno de Psicología: 0 = social; 1 =educacional; 2 = clínica.

Escala Ordinal

TEC | Tecnológico de Costa Rica

- Una escala ordinal es una escala nominal cuyos valores reflejan el orden existente entre los valores de la variable, según el mayor o menor grado en el que se encuentre presente la característica.
- Estas variables pueden transformarse preservando esta información mediante funciones monótonas crecientes.
- Ejemplos
  - Escala de pagos de un alumno: 1, 2, 3, 4, 5.
  - Grado de instrucción: 1 = primaria completa; 2 = secundaria completa; 3 = superior completa.
  - Grado de satisfacción de un cliente: 1 = muy insatisfecho; 2 =insatisfecho; 3 = satisfecho; 4 = muy satisfecho.

Escala Discreta

TEC | Tecnológico de Costa Rica

- Es una escala ordinal en la que, además, las diferencias (distancias) entre los valores asignados proporcionan información acerca de la diferencia en el grado en que se presenta la característica observada.
- Esta escala no tiene un cero real sino un cero relativo, definido arbitrariamente y que no indica ausencia de la característica medida.

- Ejemplos
  - Temperatura, en grados centígrados.
  - Animales en una granja (4, 5, 6, 7,.....)

Escala Continua

TEC | Tecnológico de Costa Rica

- Es una escala de intervalos en la que además los números asignados representan las cantidades de la característica que se mide.
- La proporción entre dos números corresponde a la misma proporción entre las cantidades de la característica medida.
- El cero aquí es real e indica ausencia total de la característica que mide la variable.

- Ejemplos
  - Sueldo bruto mensual en colones, de los empleados de una empresa.
  - Tiempo, en minutos, que tarda un alumno en terminar una prueba de agilidad mental.
  - Peso, en kilogramos, de una persona.

Datos Perdidos

TEC | Tecnológico de Costa Rica

- Impacto de los valores faltantes:
  - 1 % datos faltantes trivial.
  - 1-5 % manejable
  - 5-15 % requiere métodos sostificados
  - Más del 15 % interpretación perjudicial

Tratamiento de la no respuesta

TEC | Tecnológico de Costa Rica

- **Eliminar:**
  - Es la opción mas sencilla y consiste en eliminar las observaciones o variables que tengan los datos perdidos. Solamente debe realizarse si es poco el porcentaje de observaciones a eliminar y si es posible asumir que los valores faltante
- **Reemplazar (imputar):**
  - Reemplazar el valor perdido con un valor conocidos. Variedad de métodos, desde opciones sencillas (reemplazar por la media o mediana) hasta otras más complejas (modelos de regresión).
- **Mantener:**
  - No realizar imputación. A veces es factible analizar la información por separado.

Datos faltantes

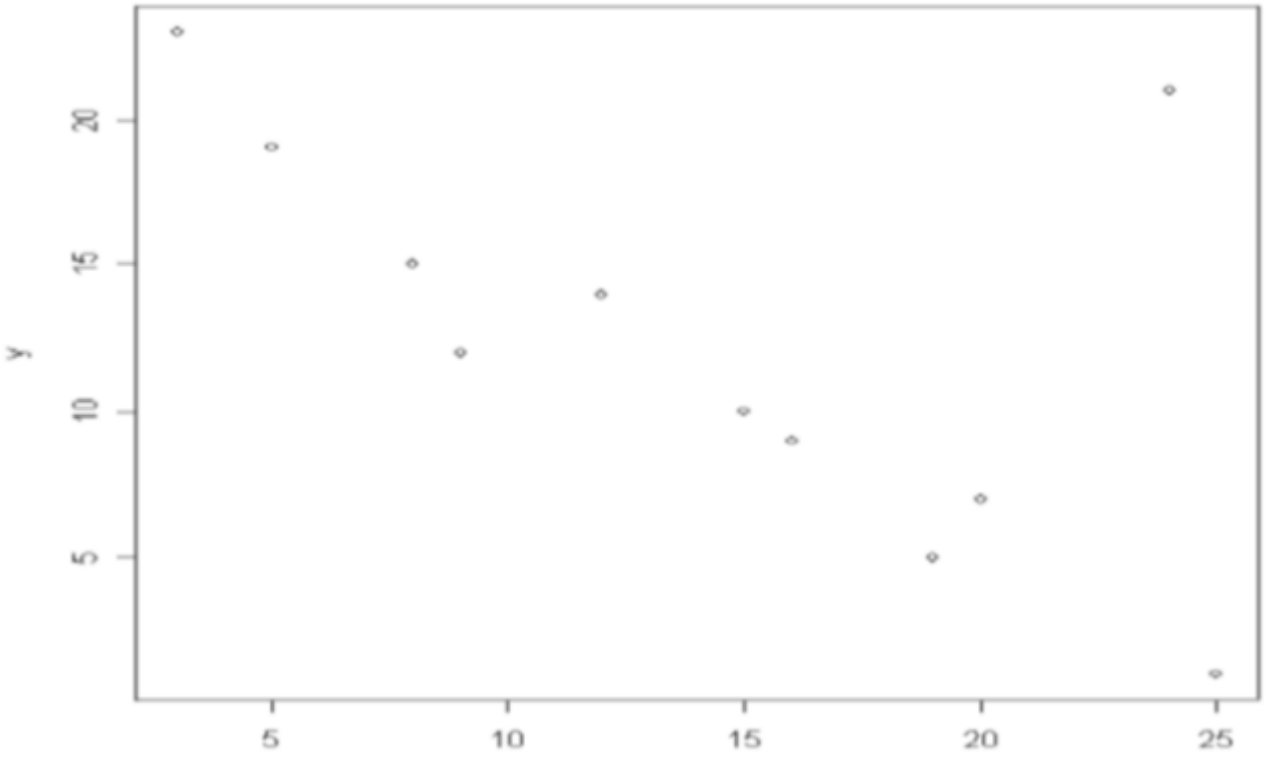
TEC | Tecnológico de Costa Rica

- **Imputación**: Los valores faltantes son reemplazados con valores estimados basados en la información disponible.
  - Imputación por la media
  - Imputación por la mediana
  - Imputación por la moda

Valores Outlier

TEC | Tecnológico de Costa Rica

- Un ”outlier” es una observación que se desvía tanto de las otras observaciones como para crear la sospecha de que fue generado por un mecanismo diferente.



Métodos para detectar Outliers

TEC | Tecnológico de Costa Rica

- Métodos basados en estadística robusta.
- Métodos basados en clustering.
- Métodos basados en distancia.
- Métodos basados en densidad local.

Transformación de Datos

TEC | Tecnológico de Costa Rica

- Suavizamiento: Remover datos ruidosos
- Agregación: resumen, construcción de cubos de datos
- Normalización
  - Normalización min-max
  - Normalización z-score
  - Normalización por escalamiento decimal
- Construcción de Atributos
  - Nuevos atributos construidos basados en los anteriormente especicados.

Normalización

TEC | Tecnológico de Costa Rica

- Consiste en reescalar los valores de los datos a un rango pre-especificado.
- Normalizar los datos de entrada ayudará a acelerar la fase de aprendizaje.
- Los atributos con rangos grandes de valores tendrán más peso que los atributos con rangos de valores más pequeños, y entonces dominarán la medida de distancia.
- También puede ser necesario aplicar algún tipo de normalización de datos para evitar problemas numéricos tales como pérdida de precisión y desbordamientos aritméticos(overows).

Normalización Z-score

TEC | Tecnológico de Costa Rica

- Este tipo de normalización funciona adecuadamente cuando:
  - No se conoce el mínimo ni el máximo de los datos originales.
  - Valores outlier pueden afectar el rango de los datos (pero no
  - los elimina).

Normalización por Escalamiento Decimal

TEC | Tecnológico de Costa Rica

- La normalización se realiza moviendo el punto decimal de los valores. El número de puntos decimales depende del máximo valor absoluto.
- Esta normalización transforma los datos al rango [-1,1]

Normalización softmax

TEC | Tecnológico de Costa Rica

- Es denominada de esta forma porque tiende suavemente hacia su valor máximo o mínimo sin llegar absolutamente. La transformación es mas o menos lineal en el rango medio, y tiene una ligera no linealidad a ambos extremos.
- Esta transformación lleva los valores al rango [0,1]

Logaritmo Natural

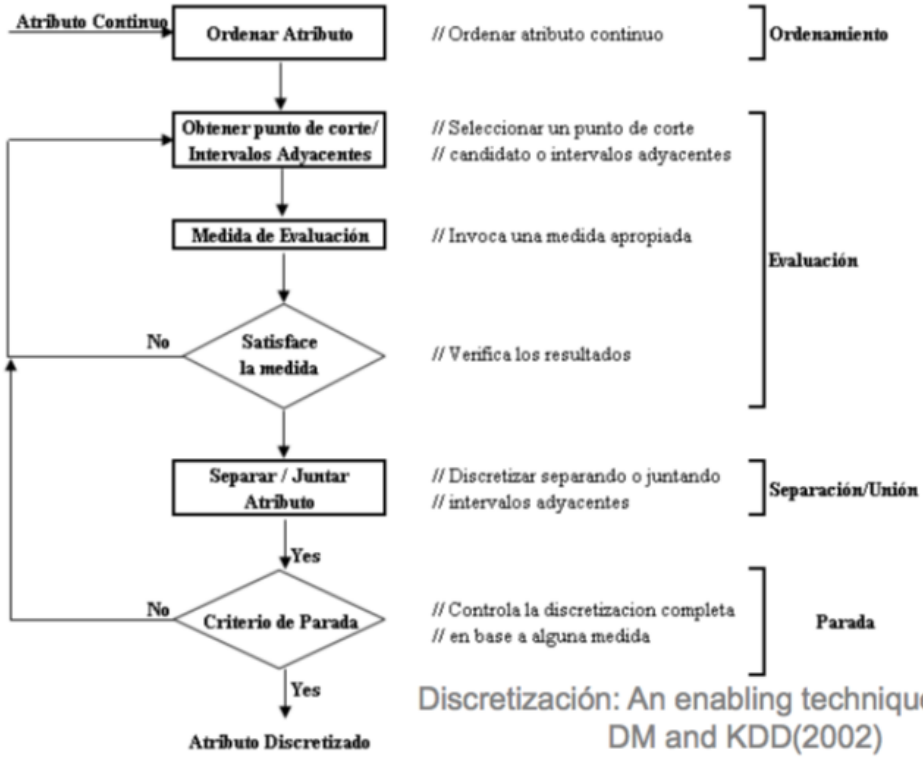
TEC | Tecnológico de Costa Rica

- Esta transformación reduce el rango de valores que puede tomar una variable positiva.
- Reduce la dispersión de los datos.
- Esta transformación es muy útil cuando se tiene outliers muy extremos.

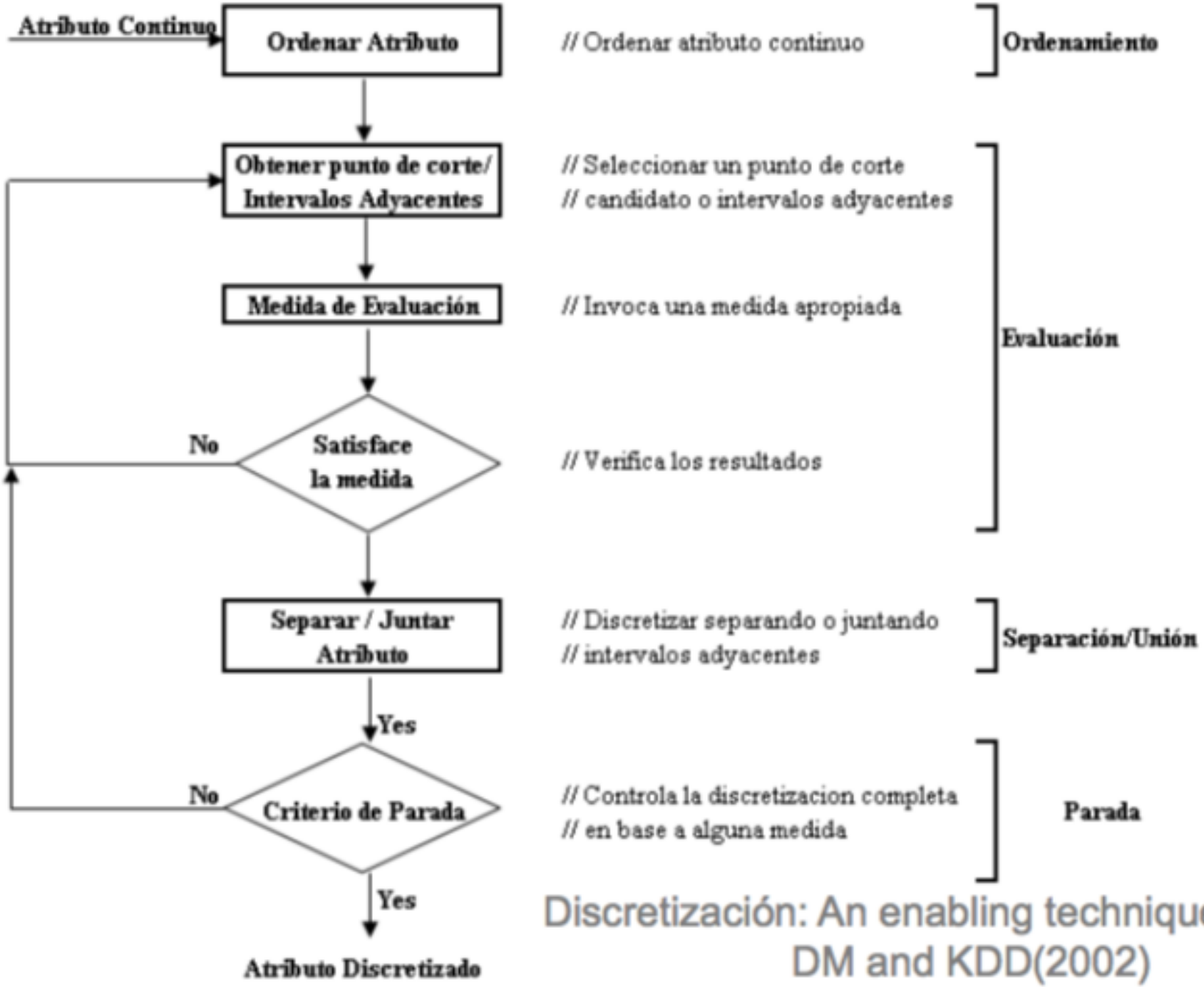
Discretización

TEC | Tecnológico de Costa Rica

- Es un método que transforma datos cuantitativos en cualitativos
- Algunas metodologías solo aceptan atributo categóricos.
  - Ejemplo : Naive Bayes.
- El proceso de aprendizaje es frecuentemente menos eficiente cuando las datos son solo cuantitativos



Discretización: An enabling technique. Liu et al DM and KDD(2002)



TEC | Tecnológico de Costa Rica

Discretización: An enabling technique. Liu et al DM and KDD(2002)

Discretización

TEC | Tecnológico de Costa Rica

- Métodos Top-Down: se inicia con una lista vacía de puntos de corte y se continúan agregando nuevos puntos a la lista separando los intervalos mientras la discretización progresa.

- Métodos Bottom-Up: se inicia con la lista completa de todos los valores continuos de la variable como puntos de corte y se eliminan algunos de ellos juntando los intervalos mientras la discretización progresa.

Discretización

TEC | Tecnológico de Costa Rica

- Discretización Dinámica: algunos algoritmos de clasificación tienen incorporados mecanismos para discretizar atributos continuos (por ejemplo, árboles de decisión). Los atributos continuos son discretizados durante el proceso de clasificación.

- Discretización Estática: Es un paso más en el preprocesamiento de datos. Los atributos continuos son previamente discretizados antes de la tarea de clasificación.

Discretización

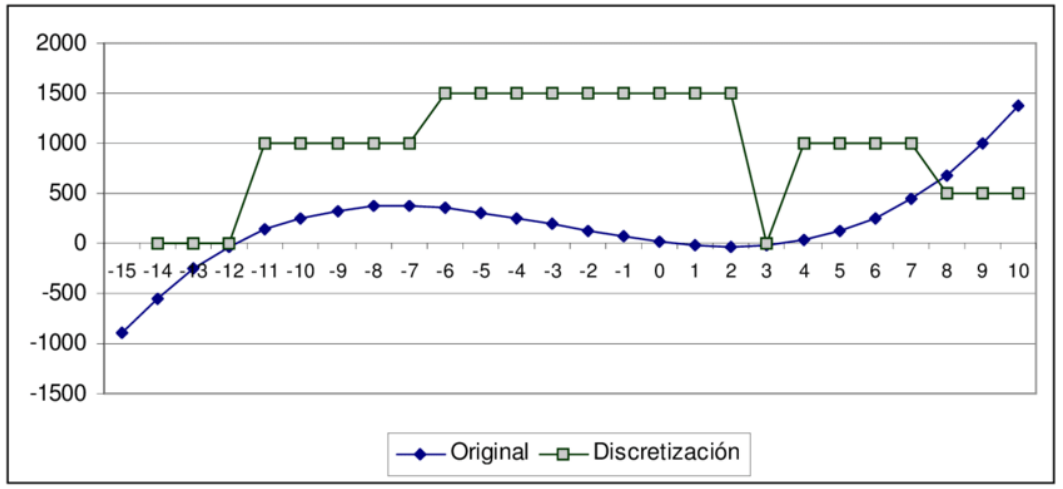
TEC | Tecnológico de Costa Rica

- Métodos Supervisados: Utilizan la información de la clase para la discretización.

- Métodos No Supervisados: No utilizan la información de la clase para la discretización.

Discretización mas popularesv

- Intervalo de igual amplitud
- Intervalos de igual frecuencia
- Discretización 1R
- Discretización por Entropía
- Discretización por chiMerge

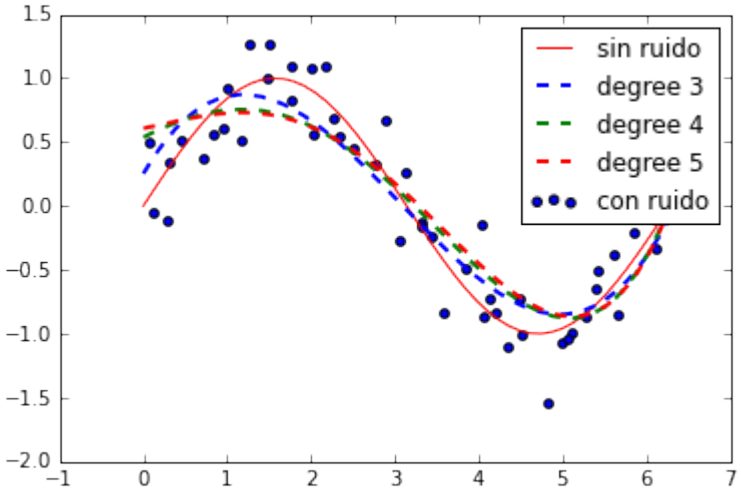


Ruido

TEC | Tecnológico de Costa Rica

- El ruido en la clase ocurre cuando un ejemplo pertenece a una clase incorrecta. El ruido en la clase puede ser atribuido a varias causas, incluyendo la subjetividad en el proceso de etiquetado, los errores en la entrada de datos o la ausencia de algunos atributos representativos.

- Dependiendo de la técnica empleada, se puede tratar el ruido utilizando las siguientes opciones:
  - Si es posible detectar que valores contienen ruido, el tratamiento puede ser similar al caso de los valores missing.
  - Algunas veces es posible conocer el error del instrumento de medida, (media y desviación estándar). Esto permite incorporar esta información en el conjunto de valores de un atributo con ruido mediante alguna transformación.

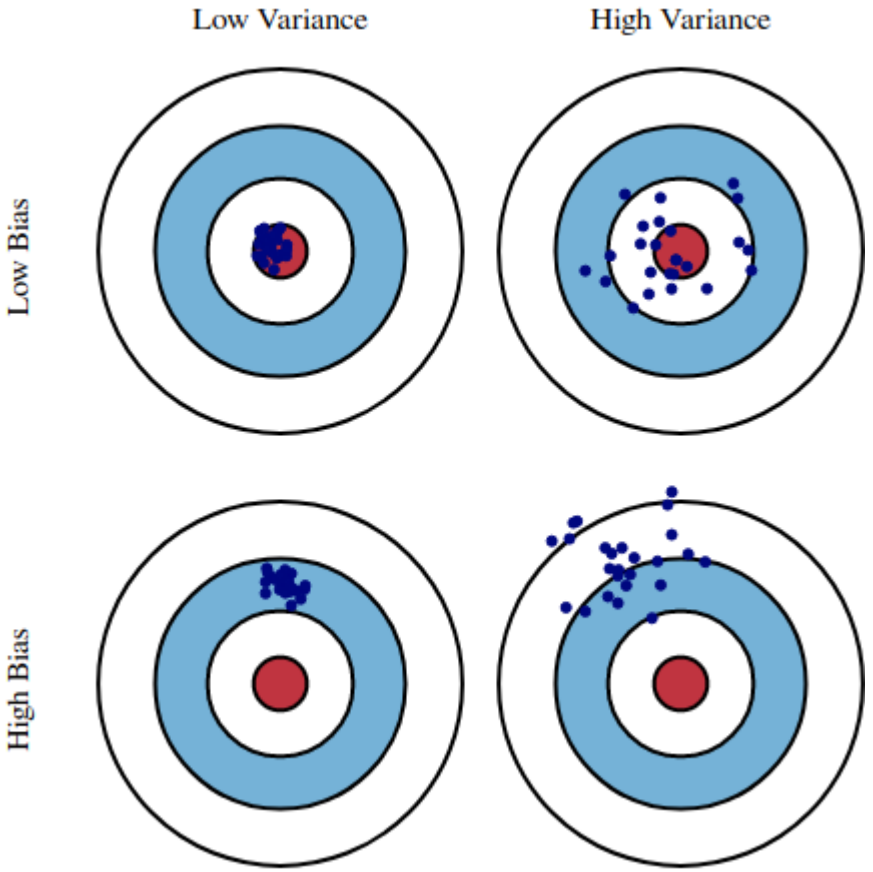


Sesgo

TEC | Tecnológico de Costa Rica

- El sesgo mide lo lejos que se encuentra el valor estimado respecto al real de la población completa. Por ejemplo, si se desea calcular la vida media de unas bombillas es necesario escoger una muestra.

- El tiempo de vida promedio de esta muestra es el que se le asocia a la población, pero no tiene porque se el de la población total. Este error es lo que se llama como sesgo.



Bibliografía

TEC | Tecnológico de Costa Rica

- Acosta-Mesa Héctor-Gabriel, Cruz-Ramírez Nicandro y García López Daniel Alejandro. Entropy Based Linear Approximation Algorithm for Time Series Discretization". Research in Computing Science. Instituto Politécnico Nacional, 2007(Publicación Pendiente).
- J. Dougherty, R. Kohavi and M. Sahami. "Supervised and unsupervised discretization of continuous features". Proceedings of the Twelfth International Conference on Machine Learning (pp. 194–202), Tahoe City, CA: Morgan Kaufmann, 1995.
- X. Hu, and N. Cercone. (1996). "Mining Knowledge Rules from Databases: A Rough Set Approach". In Proceedings of the Twelfth International Conference on Data Engineering
- H. Hua, and H. Zhao. (2009). "A Discretization Algorithm of Continuous Attributes Based on Supervised Clustering". In Chinese Conference on Pattern Recognition
- R. T. Ng. (1998). "Exploratory Mining and Pruning Optimizations of Constrained Associations Rules". In Proceedings of the ACM SIGMOD international conference on Management of data, (pp. 13-24).
- H. Sun, and S. Wang. (2011). "Measuring the component overlapping in the Gaussian mixture model". In Data Mining and Knowledge Discovery, 23, 479-502