Universidad de los Andes

Facultad de Ciencias Económicas y Sociales

Área de Métodos Cuantitativos

Notas y Ejercicios de Computación I

Gerardo A. Colmenares L.

Colección y Procesamiento de Datos

1. Introducción.

Esta sección, que pareciera salida del contexto, es con el simple propósito de tratar de dejar claro algunos conceptos que forman parte del lenguaje coloquial que se emplea en los ambientes donde las herramientas computacionales están involucradas.
En efecto, se hará un comentario breve de algunos términos que generalmente se usan indistintamente ignorando las diferencias que en ellos existen. Además, se harán breves comentarios de la aplicación de estos conceptos en el campo de la estadística, estadística y economía aplicada y en la administración y finanzas.
Finalmente, cualquier intención de profundizar alguna de las áreas mencionadas o algunos de los conceptos o términos que en esta sección se mencionan, se incluye unas referencias que podrían ser útiles.

2. Definición de datos e información.

2.1 ¿Qué es dato?

Se comienza por comprender el concepto de dato. Supóngase que se desea investigar sobre las relaciones genéticas entre parientes y para ello, se prepara un estudio genético de los individuos y se realiza una encuesta en un sector específico de Mérida. El cuestionario aplicado fue diseñado para preguntarle a cada encuestado, la edad, el color del pelo, la altura, el peso, rasgos de los ojos, tipo de nariz, forma del rostro y estado de salud.

Al revisar las encuestas y elegir una al azar se observó que la respuesta a cada una de estas preguntas fueron, **17** para la edad, **negro** para el color del pelo, **1,65** metros para la altura, **74** kilogramos para el peso, **aguileña** para el tipo de nariz, y **ovalado** para la forma del rostro y **buena salud** para su aparente estado de salud. Estas respuestas o valores al observarlos aisladamente son abstractos, sin sentido, sin un mayor significado. Es decir, son respuestas representadas mediante valores cuantitativos o cualitativos que identifican a los datos observados para un encuestado y que han sido registrados en las encuestas.

Por tanto, *dato se puede entender como la descripción de un conocimiento cuantitativo o cualitativo de factores observados en un ente real.* En nuestro ejemplo anterior, entonces, se tiene que el ente es el individuo encuestado y los factores observados son los incluidos en el cuestionario: edad, peso,

altura, salud aparente etc. Los valores dados en las respuestas son las descripciones correspondientes a cada factor.

En el campo empresarial, todas las organizaciones necesitan datos y algunos sectores son totalmente dependientes de ellos. Bancos, compañías de seguros, agencias gubernamentales y la Seguridad Social, son algunos ejemplos. Pero en general, para la mayoría de las empresas tener muchos datos no siempre es bueno. Los datos no tienen significado en sí mismos.

2.1.1. Tipos de datos

De acuerdo con la sección anterior, la descripción de los datos puede ser cualitativa o cuantitativa. Tal como se puede observar en la encuesta o cuestionario, la pregunta que se refiere a forma del rostro produce una respuesta textual; pero, en cambio la referente al peso produce una respuesta cuantitativa, un valor. De ahí, se puede expresar que existen dos grandes grupos de datos: cuantitativos y cualitativos. Sin embargo, los datos cualitativos, por lo general, son convertidos a un grupo de categorías, encasillándolos a un código relativo que ha sido asignado a cada una de las respuestas. Algunos valores que podrían ser respuestas en las encuestas, son los siguientes:

En la *forma del rostro* se tendría 1 – ovalado, 2 – redondo, 3- largo.
En el *peso* se tendría 55,70; 58; 65; 87,5; etc.
En la *edad* por su lado, se tendría 17; 32; 26; 55; 15;
En el estado de salud, se tendría 0- mala, 1- regular, 2 – buena, 3- excelente, etc

Se observa que los datos cuantitativos pareciera que pueden estar descritos mediante valores discretos (14, 15, 5, 26) o valores continuos (15,25; 265,3221; 12,50). Mientras tanto, los datos cualitativos fueron transformados a números entre 1 y 4 o entre 0 y 3.. En este el primer caso representan datos cualitativos nominales, los cuales asocian cada característica evaluada en la forma del rostro a un número. Sin embargo, para el segundo caso, el estado de salud, se puede observar que la transformación de la referencia textual a un número conserva un cierto orden o categoría, calificándose como datos cualitativos ordinales, y de este modo estableciéndose una diferencia con el tipo de dato inmediato anterior.

En síntesis, se pueden encontrar datos:

- cualitativos ordinales y cualitativos nominales
- cuantitativos discretos y cuantitativos continuos.

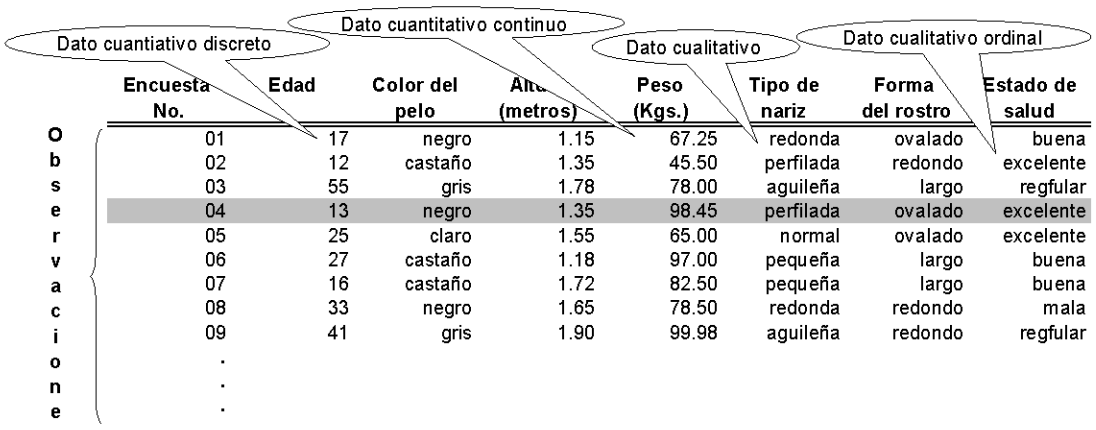
2.1.2. Presentación de los datos: Colección, Diseño y Recopilación.

Haciendo referencia a la encuesta anterior, se tiene que la presentación de los datos de acuerdo a su contenido descriptivo, se podría presentar organizadamente en filas para cada encuesta y las columnas para agrupar

las respuestas a cada pregunta formulada en el cuestionario. Estos datos organizados se pueden observar con mayor detalle en la tabla que se muestra a continuación.

Por supuesto que para poder tabular los datos mostrado en la tabla se realizaron unas actividades previas que forman parte de la recolección para su procesamiento y análisis. Estas actividades incluyen el diseño de la encuesta y la recopilación de datos. Hay que tener presente que la encuesta incluirá preguntas relacionadas a la investigación, con respuestas entendidas como datos cualitativos o cuantitativos.

Luego de la preparación de un formato especial para el vaciado de los datos de cada encuesta, los datos son recopilados directamente en el espacio geográfico definido como campo de estudio (hospital, mercado, industria, aula, etc), de manera individual.



	Encuesta No.	Edad	Color del pelo	Alt. (metros)	Peso (Kgs.)	Tipo de nariz	Forma del rostro	Estado de salud
Observaciones	01	17	negro	1.15	67.25	redonda	ovalado	buena
	02	12	castaño	1.35	45.50	perfilada	redondo	excelente
	03	55	gris	1.78	78.00	aguileña	largo	regfular
	04	13	negro	1.35	98.45	perfilada	ovalado	excelente
	05	25	claro	1.55	65.00	normal	ovalado	excelente
	06	27	castaño	1.18	97.00	pequeña	largo	buena
	07	16	castaño	1.72	82.50	pequeña	largo	buena
	08	33	negro	1.65	78.50	redonda	redondo	mala
	09	41	gris	1.90	99.98	aguileña	redondo	regfular
	.							
	.							
	.							

Ya obtenidos los formularios llenos con los datos requeridos, se preparan organizadamente (por ejemplo en forma tabular) incluyendo los cambios en los valores observados por unos nuevos que describen, por ejemplo numéricamente, las respuestas que sufran estos cambios. Este proceso se le conoce como *recodificación* y consiste básicamente en transformar el contenido original de la respuesta en algún valor cualitativo nominal u ordinal o algún valor cuantitativo discreto o continuo vinculado con la respuesta. De este modo quedan preparados y se pueden expresar tal como lo muestra la tabla anterior. Posterior a una recodificación, los datos quedarían expresados de la siguiente manera:

	Encuesta No.	Edad	Color del pelo	Altura (metros)	Peso (Kgs.)	Tipo de nariz	Forma del rostro	Estado de salud
Observacione	01	17	1	1.15	67.25	1	1	2
	02	12	2	1.35	45.50	2	2	3
	03	55	3	1.78	78.00	3	3	1
	04	13	7	1.35	98.45	2	1	3
	05	25	4	1.55	65.00	4	1	3
	06	27	2	1.18	97.00	5	3	2
	07	16	2	1.72	82.50	5	3	2
	08	33	1	1.65	78.50	1	2	0
	09	41	3	1.90	99.98	3	2	1
	.							
	.							
	.							

De este modo se puede apreciar que la numeración establecida para las variables cualitativas fueron:

Color del pelo	
negro	1
castaño	2
gris	3
claro	4

Tipo de nariz	
redonda	1
perfilada	2
guileña	3
normal	4
pequeña	5

Forma del rostro	
ovalado	1
redondo	2
largo	3

2.2. ¿Qué es información?

Después de realizada la fase de recolección y preparación de los datos en formas similares a la que se muestra en la tabla anterior, se estaría en disposición de un conjunto de datos preparados para algún fin. Se insiste en ellos separadamente no tienen ningún significado, y por esta razón se requiere de alguna transformación en o con ellos. Esta transformación produce información.

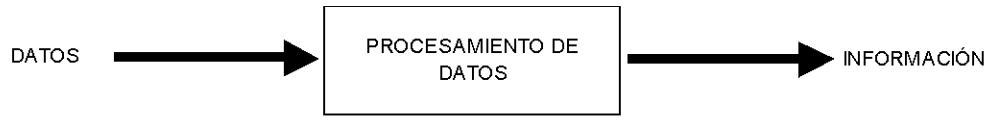
A diferencia de los datos, la información tiene significado (relevancia y propósito). No sólo puede formar potencialmente al que la recibe, sino que esta organizada para algún propósito. Los datos se convierten en información cuando su creador les añade significado. Transformamos datos en información añadiéndoles valor en varios sentidos. Hay varios métodos de transformación pero para los propósitos de este curso vale la pena mencionar:

- *Calculando:* los datos pueden haber sido analizados matemática o estadísticamente.
- *Corrigiendo*: los errores se han eliminado de los datos.
- *Condensando*: los datos se han podido resumir de forma más concisa

2.2.1. Procesamiento de datos. Esquema

De este modo, cualquier método que implique una transformación se le conoce como procesamiento de los datos. Estos procesamiento pueden involucrar cualquier técnica o herramienta de carácter científico que

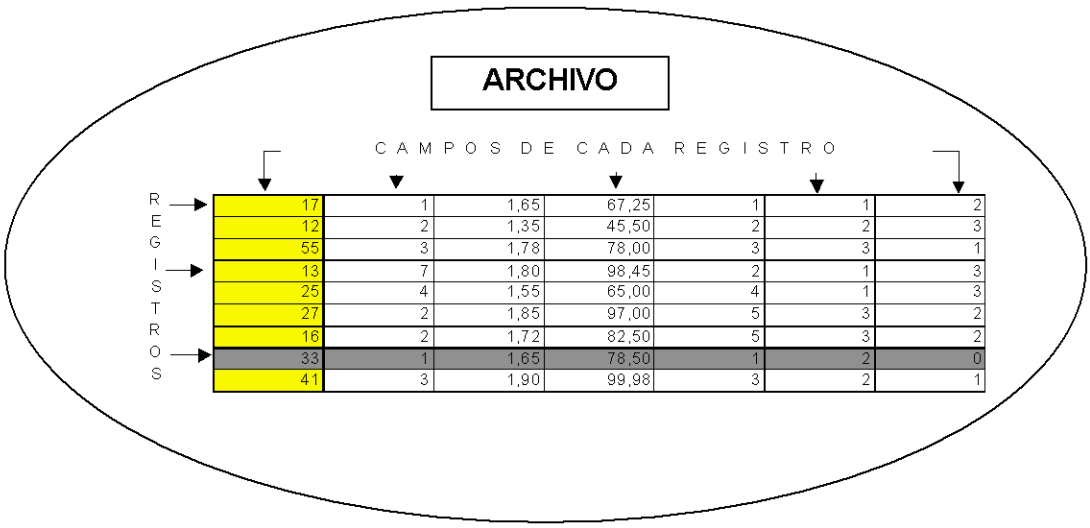
involucre la matemáticas y/o la estadística. En cualquier caso todas ellas tiene el siguiente esquema;



Este procesamiento a menudo es requerido para transformar en nueva información, datos que ya habían sido previamente procesados o información obtenida con anterioridad. De ahí que el procesamiento de datos sea de un significado tan amplio, pudiendo relacionarse tales procesos con la estadística y la matemáticas mediante herramientas y modelos preparados para tales fines.

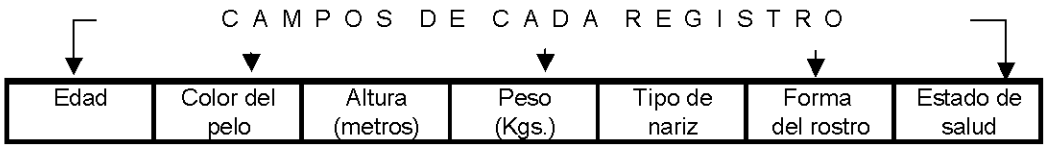
Un conjunto de observaciones como las mostradas en la tabla anterior, pueden ser manipulados para que puedan ser almacenados en dispositivos especiales, tales como los discos duros, discos flexibles o discos compactos, y posteriormente utilizados para ser procesados mediante herramientas especializadas (Ej.: Excel) y de este modo obtener información.

Esta nueva disposición de los datos almacenados, a pesar de representar las mismas características de las encuestas o cuestionarios originales, adquieren un nuevo significado en el ambiente computacional. Se estaría hablando de archivos en lugar de muestras o poblaciones, registros en lugar de cuestionarios u observaciones y campos en lugar de respuestas a las preguntas formuladas en el cuestionario. Este archivo puede ser transformado con mayor eficiencia si respeta estos mínimos criterios de estructura de datos. Observe en figura siguiente como fueron dispuestos los datos. Todos los campos son numéricos, no hay un identificador de la encuesta (fue eliminado), el usuario debe saber que el primer campo se corresponde con la edad, el segundo con el color del pelo, y así sucesivamente.



17	1	1,65	67,25	1	1	2
12	2	1,35	45,50	2	2	3
55	3	1,78	78,00	3	3	1
13	7	1,80	98,45	2	1	3
25	4	1,55	65,00	4	1	3
27	2	1,85	97,00	5	3	2
16	2	1,72	82,50	5	3	2
33	1	1,65	78,50	1	2	0
41	3	1,90	99,98	3	2	1

Este conocimiento previo de los campos y su significado se debe a que cuando se diseñó la estructura del registro, los datos se dispusieron de tal modo que ellos fueran preparado tal como se muestra:



2.2.2. Tipo de información

Aclarado el concepto de procesamiento de datos y la transformación inherente que sufren ellos para ser expresados mediante información, hay que precisar que se requiere para que esta información pueda ser obtenida. De ahí que se puede mencionar que la información puede venir expresada de diferentes formas. La más común es la numérica y algunos ejemplos de ella son los resultados al aplicar la estadística, resultados de los modelos numéricos y matemáticos, etc. La información no numérica puede ser expresada mediante símbolos, gráficos, mapas, planos, imágenes, etc. Des este tipo de información, el curso se concentrará en la numérica obtenida por el cálculo directo mediante expresiones algebraicas usando funciones o fórmulas. En el campo no numérico se hará énfasis en los gráficos y tablas descriptivas de síntesis de datos (frecuencias).

En cada caso, al convertir la fuente de datos en un medio más sencillo y expedito para procesarlos bien sea numérica o no numéricamente, se estaría hablando de análisis de datos.

3. Análisis de datos.

Al igual que en la sección anterior, analizar datos no es exclusividad de una especialidad científica. Se puede realizar análisis estadístico de datos, análisis numérico de datos, análisis financiero de datos, análisis exploratorio de datos, etc. En todos ellos se requiere una definición del esquema de análisis y por supuesto las variables y las observaciones involucradas. Igual que para los datos, las variables pueden contener valores cualitativos (nominales u ordinales) o cuantitativos (discretos o continuos), pueden ser variables respuestas (de salida o dependientes) y variables insumo (independientes o de entrada).

	VARIABLES DE CADA OBSERVACIÓN						
	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Observaciones	17	1	1.15	67.25	1	1	2
	12	2	1.35	45.50	2	2	3
	55	3	1.78	78.00	3	3	1
	13	7	1.35	98.45	2	1	3
	25	4	1.55	65.00	4	1	3
	27	2	1.18	97.00	5	3	2
	16	2	1.72	82.50	5	3	2
	33	1	1.65	78.50	1	2	0
	41	3	1.90	99.98	3	2	1
	.						
	.						
	.						

En la tabla anterior se puede observar que se definieron siete variables (X_1, X_2, X_3, X_4, X_5, X_6, y X_7). Estas variables son todas numéricas y además, la primera es discreta; la segunda, la quinta y la sexta son nominales; la tercera y la cuarta son continuas, y la séptima es ordinal.

Con este conjunto de variables se podría organizar un análisis estadístico de datos mediante la preparación de tablas de frecuencias para las variables nominales y ordinales, frecuencia por intervalos a las variables discretas y continuas, algún modelo de clasificación genética usando el conjunto completo de variables, debe existir una variable dependiente (Supóngase por ejemplo, la altura), tal que la respuesta de ser mediano, alto o bajo sea una de los efectos genéticos conseguidos con el uso del resto de variables preparadas posterior a la recolección. En fin, se puede desarrollar un análisis de datos del género que se corresponda con la necesidad de información.

3.1. Preprocesamiento.

Una actividad en la que se involucra un análisis preparatorio a los datos para el procesamiento definitivo, es lo que se le conoce como pre–procesamiento de datos. Cualquier cálculo, corrección o reducción que se le haga a los datos e implique una transformación de sus valores originales, se le puede considerar pre-procesamiento. Estos datos pre-procesados serán la nueva fuente de valores que permitirá producir la información que se esté buscando. Por ejemplo, en la encuesta anteriores y continuando con la formulación hipotética del modelo de clasificación genética, se podría pensar que la variable respuesta (X_3) sufra una modificación de sus valores advirtiendo que: todas las alturas inferiores a 1,20 cms. y sea mayor de 15 años son bajos (1); de igual modo son medianos (2) con altura entre 1,21 y 1,60 y mayores de 15 años; son altos (3) si su estatura es mayor que 1,60 y son mayores de 15 años: por último altura normales (0) para las otras condiciones. Otro cambio que pudiera sufrir los valores originales, es reduciendo la presencia de variables, mediante la eliminación de la variable estado de salud (X_7) , en vista de que pareciera estar muy relacionada con

el resto de ellas. De este modo los datos preprocesados quedan como se muestra en la tabla a continuación.

	VARIABLES DE CADA OBSERVACION					
	X_1	X_2	X_3	X_4	X_5	X_6
Observacione	17	1	1	67.25	1	1
	12	2	0	45.50	2	2
	55	3	3	78.00	3	3
	13	7	0	98.45	2	1
	25	4	2	65.00	4	1
	27	2	1	97.00	5	3
	16	2	3	82.50	5	3
	33	1	3	78.50	1	2
	41	3	3	99.98	3	2
	.					
	.					
	.					

Esta tabla con este nuevo conjunto de variables (transformadas o eliminadas) muestra
el conjunto de observaciones que serían utilizadas en un específico procesamiento

3.2. Procesamiento

Con estas variables transformadas o con las variables se podría organizar un análisis pormenorizado mediante algún tipo de cálculo estadístico (tablas de frecuencias, frecuencia por intervalos, modelo de clasificación, modelo de pronóstico, etc). A este cálculo se le conoce como procesamiento. Puede estar involucrado o no un preprocesamiento de datos. En fin, se puede desarrollar un análisis de datos del género que se corresponda con la necesidad de información. Excel es una herramienta que permite hacer procesamiento de datos. Algunas herramientas más especializadas como SAS, SPSS, Minitab, Stata, Stat Graphics, Matlab, etc, se dedican a procesar los datos. Saint, Lindo, Eviews son herramientas especializadas para procesar datos administrativos, de investigación de operaciones y de econometría respectivamente.

3.3. Análisis Exploratorio de Datos

Si el análisis de los datos es exhaustivo con la idea de identificar nuevas variables, nuevos patrones de identificación que pueden originar nuevas variables, variables cuyos datos están ausentes, variables sobrantes o redundantes, etc., se podría pensar que se hace una exploración con los datos. A esto comúnmente se le llama análisis exploratorio de datos. Para no caer en detalles, se podría comentar muy superficialmente que existen técnicas de exploración de datos, mayormente estadísticas, tales como análisis de regresión, análisis factorial, series temporales, análisis discriminante, de correspondencia, de componentes principales, y otros más. Fuera de este rango se mueven otros métodos exploratorios de datos

cuyo contexto no parte exactamente de la estadística, pero numéricamente son en algunos casos, mucho más exactos y efectivos. Tal es el caso de las técnicas heurísticas facilitadas por el cálculo numérico y la inteligencia artificial. Un par de ejemplo de ellas, es Data Mining (minería de datos) y las redes neuronales artificiales (RNA)

3.4. Algunos comentarios de DATA MINING y RNA

En las organizaciones la buena *gestión de los datos* es esencial para su funcionamiento, ya que operan con millones de transacciones diarias. Pero en general, para la mayoría de las empresas tener muchos datos no siempre es bueno. Las organizaciones almacenan datos sin sentido. Realmente esta actitud no tiene sentido por dos razones. La primera es que demasiados datos hacen más complicado identificar aquellos que son relevantes. Segundo, y todavía más importante, es que *los datos no tienen significado en sí mismos.*

La compilación de nuevos datos mediante algunas técnicas de recolección (ej.: encuestas, adquisición automática de datos, etc) ha menudo resulta muy costosa. Posiblemente resulta más confiable y económico re-usar la información. Es decir reciclarla. Se puede realizar inferencia desde archivos históricos de datos mediante métodos que permiten pronosticar resultados que se están requiriendo y de este modo evitando nuevas recolecciones.

Un nuevo estilo de empresa en el mercado es aquella cuyo principal objetivo comercial es el de recopilar datos de cualquier género en grandes cantidades, organizarlos, sintetizarlos y ponerlos a disposición como producto final a clientes que requieren algunos datos especializados para hacer sus propias cálculos e inferencias. Estas empresas de datos han surgido de la necesidad de darle un mejor provecho a la información histórica que fue captada en un momento dado a través de un conjunto de respuestas y que pierde su vigencia pasando a engrosar los grandes archivos del recuerdo.

Pues bien, técnicas emergentes del mundo de la inteligencia artificial y de la estadística aplicada permiten realizar exploraciones a esos datos históricos para reconocer datos que no fueron aprovechados en su momento, crear nuevos datos a partir de esos datos históricos, identificar patrones de comportamiento en los datos, extraer c\onjuntos reducidos de datos que son representativos para los objetivos actuales, etc. Estas técnicas novedosas, son las que se están empleando en las ciencias aplicadas y en la administración de los negocios. Ejemplo de ellos es en la geofísica para el pronóstico de movimientos símicos, en las finanzas para evaluar el comportamiento de las acciones, en la arquitectura para el desarrollo de nuevos urbanismos, en las finanzas para predecir crecimientos de las empresas, en la economía para pronosticar comportamientos de los ingresos económicos de un grupo social, en la geología para la evaluación y pronóstico de exploraciones petroleras, etc. Es decir, es de múltiple aplicación.

Las redes neuronales y la minería de datos (data mining) son algunas de estas herramientas novedosas. Las redes neuronales artificiales (RNA), emulando marginalmente las redes de neuronas biológicas, puede construir modelos RNA debidamente entrenados y probados para resolver problemas

de pronóstico o clasificación específicos. Ejemplo de ellos, son los modelos que pronostican el comportamiento del tiempo. Otro ejemplo es la autenticación de las tarjetas de créditos, mediante el reconocimiento de firmas, huellas digitales, e imágenes.

La minería de datos no es más que una versión de la estadística aplicada adaptada a la administración y negocios para realizara análisis exploratorio de datos. Sin embargo, su aplicación se a ha ampliado hacia el uso de algunos métodos numéricos emergentes como son los algoritmos genéticos y la lógica difusa para construir modelos que permitan la selección extracción y reconocimiento de patrones desde los datos históricos y crear nueva información para la toma de decisiones. Ejemplo de ellos son los datos requeridos para clasificar las acciones de la bolsa de valores entre buenas, regulares o malas. Otro ejemplo es la determinación de si un banco pasa las pruebas del ácido que lo certifican como un banco confiable para mantener depósitos en ellos a través de índices de confianza.

En conclusión, el mejor aprovechamiento de los datos garantiza resultados efectivos y confiables. El uso de los datos indicados y adecuados al problema, a través de el procesamiento, pre-procesamientos y análisis exploratorio de los datos, mediante técnicas convencionales o no, garantizan que la información que se obtenga sea la requerida a menores costos y con un alto porcentaje de consistencia y calidad.