Instituto Politécnico Nacional
Escuela Superior de Cómputo
Secretaría Académica
Departamento de Ingeniería en Sistemas Computacionales

Minería de datos (*Data Mining*)

1

Profesora: *Dra. Fabiola Ocampo Botello*

¿Qué es la minería de datos?

La minería de datos es un campo multidisciplinario que integra trabajo de diversas áreas como tecnología de bases de datos, aprendizaje automático (*machine learning*), estadística, reconocimiento de patrones, recuperación de información, redes neuronales, sistemas basados en conocimiento, inteligencia artificial, cómputo de alto rendimiento y visualización de datos (Sahu, Shrma, & Gondhalakar, 2011).

La minería de datos ha sido comúnmente definida como encontrar información en una base de datos, ha sido llamada análisis de datos exploratorio, descubrimiento conducido por datos y aprendizaje deductivo (Dunham, M. H., 2002).

2

3

Características de la minería de datos:

- Surge como una tecnología para apoyar a comprender el contenido de una base de datos.
- Es una etapa de un proceso llamado extracción de conocimiento en bases de datos (*Knowledge Discovery in Databases, KDD*).
- Mediante la minería de datos se detectan relaciones entre los datos que no han sido identificadas, lo que permite conocer relaciones con sentido, patrones de comportamiento, secuencias, predicciones, agrupamiento que serán analizados para la toma de decisiones.
- La minería de datos generalmente implica el análisis de los datos almacenados en un almacén de datos (*datawarehouse*). Tres de las principales técnicas de la minería de datos son: la regresión, la clasificación y el agrupamiento (Sahu, Shrma, & Gondhalakar, 2011).

Extracción de conocimiento en bases de datos (*Knowledge Discovery in Databases, KDD*).

Según (Dunham, 2002), La minería de datos ha sido comúnmente definida como encontrar información en una base de datos, ha sido llamada análisis de datos exploratorio, descubrimiento conducido por datos y aprendizaje deductivo.

Este autor señala que la extracción de conocimiento en bases de datos (KDD) es el proceso de encontrar información útil y patrones en los datos. Mientras que la minería de datos (*data mining*) es el uso de algoritmos para extraer información y patrones de comportamiento de datos como parte del KDD.

4

5

Etapas de la Extracción de conocimiento en bases de datos (*Knowledge Discovery in Databases, KDD*).

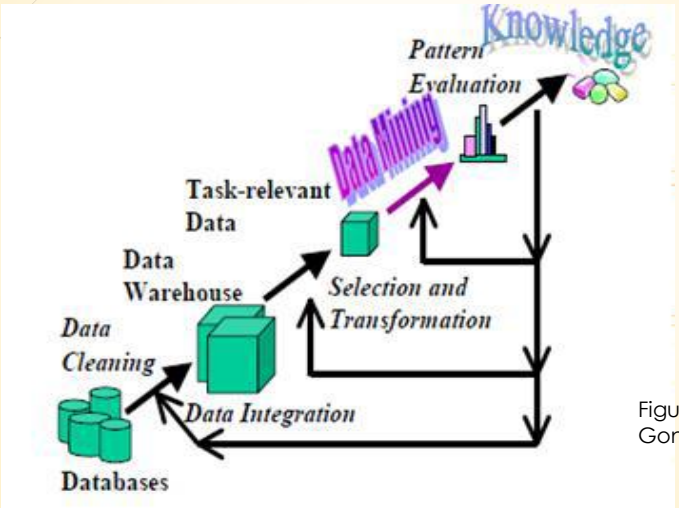


Figura tomada de Sahu, Shrma, & Gondhalakar (2011)

6

Etapas de la Extracción de conocimiento en bases de datos (*Knowledge Discovery in Databases, KDD*).

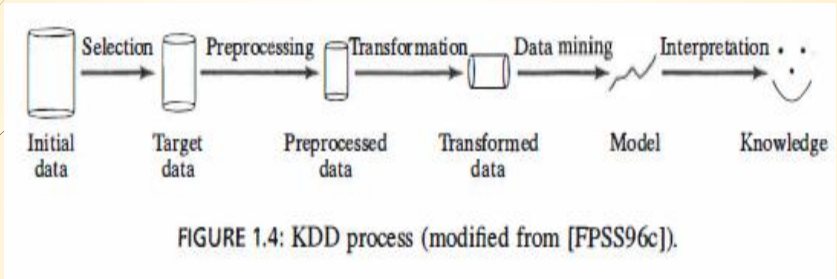


FIGURE 1.4: KDD process (modified from [FPSS96c]).

Figura tomada de Dunham (2002)

7

Aplicaciones de la minería de datos

La minería de datos se aplica para detectar patrones de comportamiento y extraer información valiosa, algunas de las aplicaciones son (Rodríguez y Díaz, 2009):

- La utilización de árboles de decisión en la construcción de modelos de clasificación de diferentes características del desarrollo de software.
- Aspectos climatológicos: predicción de tormentas, etc.

Medicina: encontrar la probabilidad de una respuesta satisfactoria a un tratamiento médico.

- Mercadotecnia: identificar clientes susceptibles de responder a ofertas de productos y servicios por correo, fidelidad de clientes, afinidad de productos.
- Inversión en casa de bolsa y banca: análisis de clientes, aprobación de préstamos, determinación de montos de crédito, etc.

8

Aplicaciones de la minería de datos

- Detección de fraudes y comportamientos inusuales: telefónicos, seguros, en tarjetas de crédito, de evasión fiscal, electricidad, etc.
- Análisis de canastas de mercado para mejorar la organización de tiendas, segmentación de mercado (*clustering*).
- Determinación de niveles de audiencia de programas televisivos.
- Industria y manufactura: diagnóstico de fallas.

9

Ventajas de la minería de datos

Las ventajas de la minería de datos son (Sahu, Shrma & Gondhalakar (2011):

- Mercadotecnia/Venta al menudeo. La minería de datos ayuda a las compañías a construir modelos fundamentados en datos históricos, con esta información los comercializadores pueden tener un enfoque adecuado para vender sus productos.
- Banca financiera. Considerando datos históricos de los clientes, las instituciones bancarias y financieras pueden estimar cuáles son los buenos y/o malos préstamos y su nivel de riesgo.
- Producción. Al aplicar la minería de datos con datos de ingeniería, los fabricantes pueden detectar equipos defectuosos y determinar patrones óptimos de control.
- Gobierno. La minería de datos ayuda a las agencias gubernamentales a excavar y analizar registros de transacciones financieras en la construcción de patrones que puedan detectar el lavado de dinero o la actividad criminal.

10

Desventajas de la minería de datos

Las desventajas de la minería de datos son (Sahu, Shrma & Gondhalakar (2011):

- Problemas de privacidad. Las empresas recopilan información acerca de sus clientes de muchas maneras para entender las tendencias de sus comportamientos de compras. El temor que tienen las personas es que la información personal de ellos se venda o se filtre.
- Problemas de seguridad. Ha habido muchos casos en los cuales lo piratas informáticos (hackers) han ingresado y robado grandes cantidades de datos de clientes de grandes corporaciones.
- Mal uso de la información/información inexacta. La información recopilada mediante la minería de datos con fines comerciales o éticos puede ser mal utilizada. Esta información es explotada por personas o empresas no éticas para tomar ventaja o abusar de la vulnerabilidad de la gente o discriminar a un grupo de personas.

Ejemplos de tipos de problemas que aborda la minería de datos

Ejemplos de tipos de problemas en los que se aplica la minería de datos (ejemplos adaptados de Dunham, M. H., 2002):

11

Ejemplo No. 1:

Las compañías de tarjetas de crédito deben determinar si autorizan las compras que se realizan con tarjeta de crédito. Supongamos que con base en información histórica sobre compras, cada compra se coloca en una de cuatro clases: (1) autorizar, (2) solicitar una identificación adicional antes de la autorización, (3) no autorizar, y (4) no autorizar y contactar a la policía.

Las funciones de minería de datos en este ejemplo son necesarias, debido dos aspectos:

- Primero se deben examinar los datos históricos para determinar cómo los datos encajan en las cuatro clases.
- Posteriormente el problema es aplicar este modelo a cada nueva compra.

Si bien la segunda parte puede expresarse como una simple consulta de base de datos, la primera parte no puede ser.

Ejemplo No. 2:

Se utiliza una estación de control de seguridad del aeropuerto para determinar si los pasajeros son terroristas o delincuentes potenciales. Para hacer esto, se escanea la cara de cada pasajero y se identifica su patrón básico (distancia entre los ojos, tamaño y forma de la boca, forma de la cabeza, etc.).

Este patrón se compara con los registros en una base de datos para ver si coincide con algún patrón asociado con delincuentes conocidos. *Clasificación (Reconocimiento de patrones).*

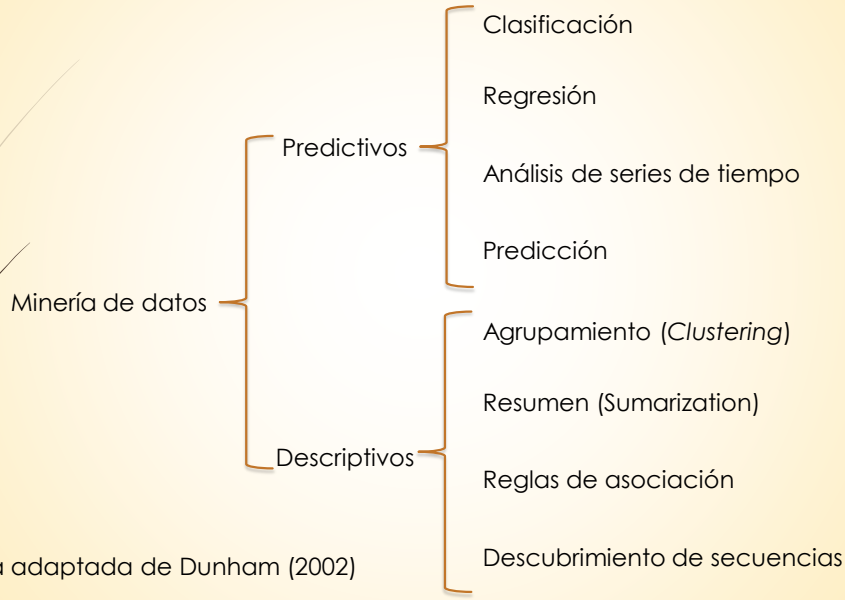
12

Ejemplo No. 3:

Un profesor desea alcanzar cierto nivel de ahorros antes de su jubilación. Periódicamente, ella predice cuáles serán sus ahorros según su valor actual y varios valores del pasado. Ella utiliza una fórmula de regresión lineal simple para predecir este valor ajustando el comportamiento pasado a una función lineal y entonces usa esta función para predecir los valores de esos puntos en el futuro.

Con base en estos valores, ella altera su cartera de inversiones. *Regresión lineal.*

Tipos de algoritmos



13

Figura adaptada de Dunham (2002)

Los tipos de datos que comúnmente se encuentran en las bases de datos los siguientes niveles de medición de datos:

Babbie (1988) y Hernández y otros (2003) establecen los siguientes niveles de medición:

- **Nominales**. Se utilizan para distinguir categorías comprendidas en una variable determinada, son mutuamente excluyentes entre sí. Existen dos o más categorías que no tiene orden ni jerarquía, por ejemplo: sexo (hombre o mujer), afiliación religiosa o política. Los números asignados a cada categoría son simplemente con fines de clasificación.

14

- **Ordinales**. Reflejan un orden de rango entre las categorías que forman una variable. Existen varias categorías que mantienen un orden y existe una jerarquía. Los números asignados reflejan tal jerarquía, reflejan intervalos que no necesariamente son iguales. Por ejemplo: clase social (alta, media, baja), categoría ocupacional en un empleo

- **Intervalo**. Es similar al anterior, pero en este tipo de dato los intervalos entre las categorías son iguales en la medición, también se conoce como intervalos iguales para resaltar la característica que la distingue de una escala ordinal. El cero es arbitrario. *Por ejemplo si desea expresar la temperatura ambiental en categorías de 5 en 5 grados, el cero no indica la ausencia de temperatura.*

- **Razón**. Tiene las mismas características que las medidas de intervalo, pero el cero no es arbitrario, es real. Por ejemplo: las horas a la semana que una persona ve la televisión, el número de hijos, las ventas de un producto en un periodo de tiempo, la edad en años. Una distancia de 10 km está al doble de una de 5 km.

15

Tipos de estadísticos

Bennet, Briggs y Triola (2011:2) establecen que **estadística** (en singular) es la ciencia que recolecta, organiza e interpreta datos y **estadísticas** (en plural) son los datos (números y otras partes de información) que describen o resumen algo.

Castillo Morales (2013) establece que un estadístico es un procedimiento de cálculo que usa datos y constantes conocidas.

Castillo Morales (2013) presenta diversos tipos de estadísticos:

- Estadísticos de localización.
- Estadísticos de dispersión.

16

Estadísticos de localización

Los estadísticos de localización son: mínimo, máximo, semirrango, mediana, percentil 25% o primer cuartil, percentil 75% o tercer cuartil, percentil 95%, media y moda.

Semirrango: es el valor intermedio entre el máximo y el mínimo.

17

Mason, Lind y Marchal (2000:121) establecen que la varianza y la desviación estándar son las medidas de dispersión más ampliamente usadas, pero existen otros medios para describir los valores que dividen un conjunto de datos en partes iguales. Estos son: cuartiles, deciles, centiles y percentiles.

La **mediana** indica el centro de los datos.

Cuartiles. Dividen el conjunto de observaciones en cuatro partes iguales.
Q_1 es el primer cuartil, es el valor abajo del cual se encuentra el 25% de las observaciones.
Q_2 es la mediana
Q_3 es el valor por abajo del cual se encuentra el 75% de las observaciones.

Deciles dividen el conjunto de datos en 10 partes iguales.

Centiles dividen el conjunto de datos en 100 partes iguales

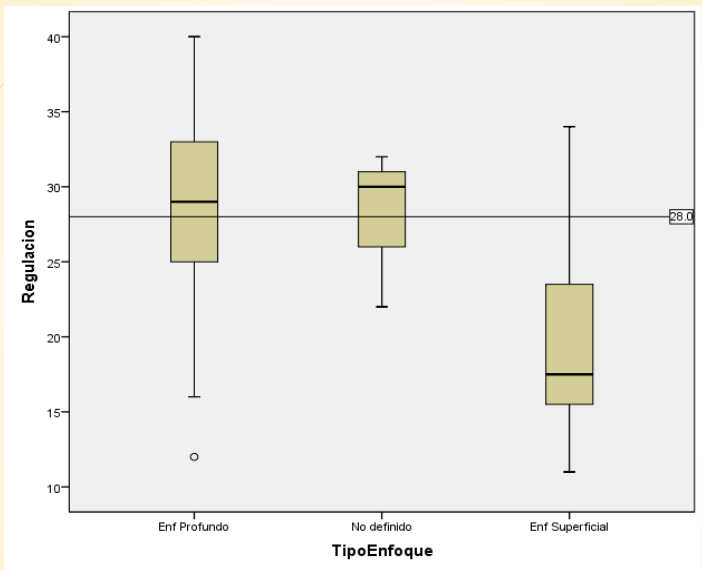
18

Los datos más representativos de una serie de observaciones son:

Valor mínimo
Cuartil 1
Mediana
Cuartil 3
Valor máximo

Los cuales son los expresados en una gráfica de caja y bigotes.

19



20

Estadísticos de dispersión

Se entiende por dispersión la separación que presentan los puntos entre sí o con respecto al centro de la gráfica. Si todos los datos tienen el mismo valor, no hay dispersión y esta vale cero (Castillo Morales, 2013).

Los estadísticos de dispersión son:

- Rango. se obtiene restando el valor mínimo al valor máximo y da la longitud del intervalo en donde se encuentra la muestra.
- Rango intercuartílico. Es la diferencia entre los percentiles 75% y 25%, entre éstos se encuentra el 50% de los valores intermedios de la muestra.
- Desviación estándar. Es la raíz cuadrada de la varianza.
- Varianza. En una muestra se refiere a la diferencia entre el dato y la media elevadas al cuadrado.

21

La medida de tendencia central más utilizada es la media.

La medida de variabilidad más utilizada es la varianza. También llamada cuadrado medio. Nos dice qué tan dispersos están los valores con respecto a la media.

Si se tiene un grupo muy heterogéneo en alguna puntuación, por ejemplo en el promedio, su varianza será muy grande con respecto a uno que es muy homogéneo.

La varianza es un estadístico muy utilizada en la comparación de grupos, en el análisis de hipótesis, entre otros más.

22

Correlación

Según Aguayo y Lora (2007), la correlación es una técnica matemática que evalúa el grado de asociación o relación entre dos variables cuantitativas, tanto en términos de direccionalidad como de fuerza o intensidad proporcionados por un coeficiente.

El coeficiente de correlación puede tener valores que oscilan entre -1 y +1, considerando el cero. Cuando el valor se acerca a +1 se considera que ambas variables se relacionan de manera muy estrecha, esto quiere decir que si se analiza la correlación entre dos variables X y Y, existe una correlación positiva si cuando se incrementa el valor de X también se incrementa el de Y o cuando hay un decremento en el valor de X también hay un decremento en el valor de Y.

23

Del mismo modo, cuando el valor se acerca a -1 refleja que existe una relación de forma inversa, esto es, cuando aumenta el valor de X existe un decremento en el valor de Y y cuando X obtiene puntajes bajos Y alcanza puntajes altos.

Representación de datos

La representación de datos se puede hacer mediante tablas o gráficas. Algunas de las utilizadas son:

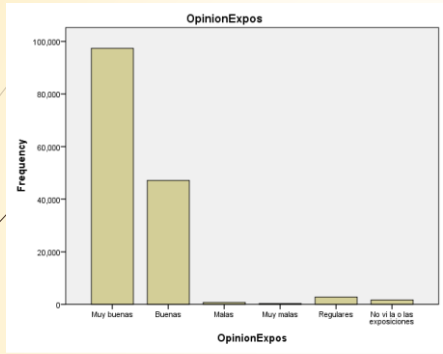
Tablas de frecuencias y de frecuencia acumulada

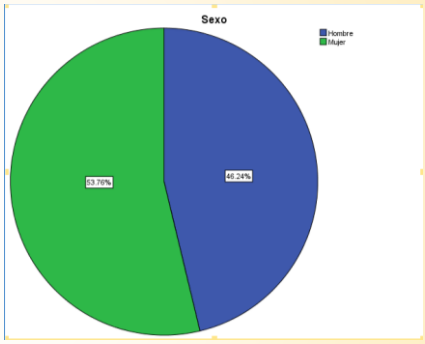
24

Eval_Gral

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	99	.1	.1	.1
	1	57	.0	.0	.1
	2	44	.0	.0	.1
	3	67	.0	.0	.2
	4	132	.1	.1	.3
	5	448	.3	.3	.6
	6	813	.5	.5	1.1
	7	2991	2.0	2.0	3.1
	8	14847	9.9	9.9	13.0
	9	33738	22.5	22.5	35.5
	10	96772	64.5	64.5	100.0
	Total	150008	100.0	100.0	

Gráficas de barras (datos cualitativos) y gráficas de pastel (datos cualitativos).

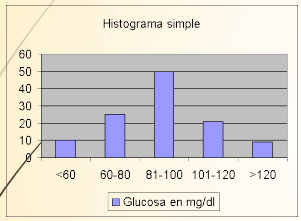


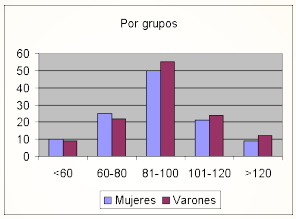


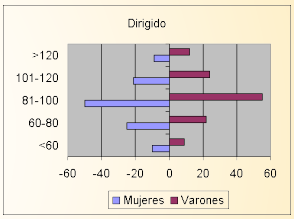
25

Histograma.

Un histograma es una gráfica de barras que muestra una distribución para datos cuantitativos (de intervalo o de razón de medida); las barras tienen un orden natural y las anchuras de las barras tienen un significado específico.







26

Figuras tomadas de: Ejemplos de tipos de representación gráfica. "s/f". http://www.hrc.es/bioest/Ejemplos_histo.html

27

Gráfica de líneas

Muestra una distribución de datos cuantitativos, conecta una serie de puntos. Los puntos van donde iría la parte superior de la barra en el histograma. La posición horizontal de los puntos corresponde al centro de la clase.

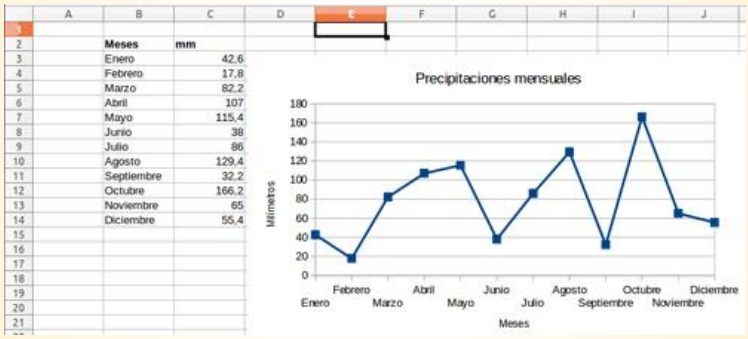


Figura tomada de: Crear un gráfico de líneas ("S/f").
En: https://ordenadorpractico.es/mod/assign/view.php?id=274

28

¿Qué hacemos con los datos sucios?

Han, Kamber & Pei (2012) presentan diversas rutinas de limpieza de datos para tratar los valores faltantes, suavizar el ruido al encontrar valores atípicos y corregir inconsistencias.

Valores faltantes:

1. Ignorar la tupla. Esto se hace por ejemplo cuando falta la etiqueta de la clase y lo que se realiza es la clasificación. Este método no es muy efectivo, a menos que a la tupla le falten varios valores.

2. Completar manualmente el valor faltante. Este enfoque consume mucho tiempo y puede no ser factible considerando un gran conjunto de datos con muchos datos faltantes.

3. Uso de una constante global para completar los valores faltantes. Se puede utilizar una etiqueta como "Desconocido". Pero, hay que tener cuidado por que el algoritmo de minería de datos puede detectar erróneamente que es un concepto interesante ya que existen muchos datos con ese valor.

29

4. Uso de una medida de tendencia central para el atributo. Puede ser la media o la mediana.

5. Usar la media o la mediana para todos las muestras que pertenecen a la misma clase. Por ejemplo, si se clasifica a los clientes de acuerdo al riesgo de crédito, se puede reemplazar el valor faltante con el promedio del valor de ingreso de los clientes de esa categoría.

6. Usar el valor más probable para completar el valor faltante. Este valor se puede determinar mediante una regresión.

Ejercicios de clase

30

Ejercicios adaptados de Bennet, Briggs & Triola (2011).

1. Una entrenadora de atletismo desea determinar un ritmo cardiaco apropiado para sus atletas, por lo que selecciona a 5 de sus mejores atletas y les coloca un monitor cardiaco durante un entrenamiento. A la mitad del entrenamiento lee el ritmo cardiaco y obtiene los siguientes datos: 130, 135, 140, 145, 325. ¿Cuál es la mejor medida del promedio para este caso: la media o la mediana? ¿Por qué?

2. ¿Cuál media?

Los 100 estudiantes de nuevo ingreso de una universidad pequeña toman tres cursos en el programa básico de estudios. Dos cursos se imparten en grupos con los 100 alumnos cada uno. El tercer curso se imparte en 10 grupos con 10 alumnos cada uno. Los alumnos y los administradores dicen que los grupos son muy grandes. Los alumnos dicen que el tamaño medio de los grupos es de 70. Los administradores dicen que es de 25. ¿Ambas partes son correctas? ¿Por qué?

31

3. ¿Quién jugó mejor?

Ejercicio adaptado de Bennet, Briggs & Triola (2011).

Se tiene el registro de dos jugadores de básquetbol: Sebastián y Javier.

Jugador	Primer tiempo			Segundo tiempo		
	Canastas	Intentos	Porcentajes	Canastas	Intentos	Porcentajes
Sebastián	4	10	40%	3	4	75%
Javier	1	4	25%	7	10	70%

El reporte final dice:
Primer tiempo: 40% (Sebastián) a 25% (Javier).
Segundo tiempo: 75% (Sebastián) a 70% (Javier).

¿Podemos asegurar que Sebastián jugo mejor?

La **paradoja de Simpson** o **efecto Yule-Simpson** es una paradoja en la cual una tendencia que aparece en varios grupos de datos desaparece cuando estos grupos se combinan y en su lugar aparece la tendencia contraria para los datos agregados.

32

4. La tabla siguiente muestra los registros de bateo de dos jugadores de beisbol.

	Primera mitad		
Jugador	Hits	Veces al bat	Promedio de bateo
Pedro	50	150	0.333
Beto	10	50	0.200
	Segunda mitad		
Jugador	Hits	Veces al bat	Promedio de bateo
Pedro	35	70	0.500
Beto	70	150	0.467

Ejercicio adaptado de Bennet, Briggs & Triola (2011).

¿Quién tuvo el promedio de bateo más alto en la primera mitad de la temporada?
¿Quién tuvo el promedio de bateo más alto en la segunda mitad de la temporada?
¿Quién tuvo el promedio de bateo más alto global de la temporada?
¿Se presenta la paradoja de Simpson?

33

5. Muertes por tuberculosis (TB) en la ciudad de Nueva York y en Richmond, Virginia en 1910.

Ejercicio adaptado de Bennet, Briggs & Triola (2011).

	Nueva York	
Raza	Población	Muerte por TB
Blanca	4 675 000	8 400
No Blanca	92 000	500
Total	4 767 000	8 900
	Richmond	
Raza	Población	Muerte por TB
Blanca	81 000	130
No Blanca	47 000	160
Total	128 000	290

a) Calcular la tasa de mortandad para blancos, no blancos y todos los residentes de la ciudad de Nueva York.
b) Calcular la tasa de mortandad para blancos, no blancos y todos los residentes de la ciudad de Richmond.
c) ¿Cómo se presenta la paradoja de Simpson?.

Referencias bibliográficas

Aguayo Canela, Mariano y Lora Monge, E. (2007). Cómo realizar "paso a paso" un contraste de hipótesis con SPSS para Windows: (III) Relación o asociación y análisis de la dependencia (o no) entre dos variables cuantitativas. Correlación y regresión lineal simple. *Documento de la Fundación Andaluza Beturia para la Investigación en Salud (fabis.org)*. Dot. Núm. 0702005. Disponible en: http://www.fabis.org/html/archivos/docuweb/contraste_hipotesis_3r.pdf

Babbie R. Earl. (1988). *Métodos de investigación por encuesta*. Fondo de Cultura Económica. México.

Bennet, Briggs & Triola (2011). *Razonamiento estadístico*. Pearson. México.

Castillo M., A. (2013). *Estadística aplicada*. México, ed. Trillas.

Crear un gráfico de líneas. "s/f". *Fundación esplai. Ordenador práctico*. Disponible en: https://ordenadorpractico.es/mod/assign/view.php?id=274

Dunham, M. H. (2002). *Data mining: introductory and advanced topics*. Prentice Hall.

Ejemplos de tipos de representación gráfica. "s/f". *Material docente de la unidad de bioestadística clínica del Hospital Universitario Ramón y Cajal*. Disponible en: http://www.hrc.es/bioest/Ejemplos_histo.html

Han, Jiawei; Kamber, Micheline & Pei, Jian. (2012). *Data Mining: concepts and techniques*. Third edition. Morgan Kaufman Series.

Hernández Sampieri, R.; Fernández Collado, C; Baptista Lucio, P. (2003). *Metodología de la Investigación*. Tercera Edición. Editorial Mc. Graw Hill. D. F. México.

Mason, Lind & Marchal. (2000). *Estadística para administración y Economía*. Alfaomega. México.

Rodríguez Suárez, Yuniet, Díaz Amador, Anolandy, Herramientas de Minería de Datos. *Revista Cubana de Ciencias Informáticas* [en linea] 2009, 3 (Julio-Diciembre) : [Fecha de consulta: 24 de julio de 2019] Disponible en: <http://google.redalyc.org/articulo.oa?id=378343637009> ISSN 1994-1536

Sahu, Hemlata; Shrma, Shalini; Gondhalakar, Seema. (2011). A Brief Overview on Data Mining Survey. *International Journal of Computer Technology and Electronics Engineering (IJCTEE)*. Vol.1.

34