

Tema 1: Introducción al Data Mining



Universidad
Francisco de Vitoria
UFV Madrid

Alberto Nogales
alberto.nogales@ufv.es

Índice

- Motivación y orígenes
- Definición de Data Mining
- Tareas del Data Mining
- Proceso y fases de extracción del conocimiento
- Metodología CRISP-DM
- Casos de uso

Motivación y orígenes

- 1762
 - Se publica el artículo póstumo de Thomas Bayes en el que se habla del llamado teorema de Bayes.
- 1805
 - Legendre y Gauss aplican la regresión para estudiar las orbitas de los cuerpos alrededor del sol.

Motivación y orígenes

- 1936
 - Alan Turing formula en un artículo la idea de la maquina de Turing.
- 1943-1980's
 - Redes neuronales, bases de datos, algoritmos genéticos, concepto de Data Mining.

Motivación y orígenes

- 1989
 - Concepto de Knowledge Discovery in Databases (KDD) introducido por Gregory Piatetsky-Shapiro.
- 1990's
 - El concepto de Data Mining empieza a usarse en distintos ámbitos (Billy Bean).

Motivación y orígenes

- ¿Cuántas unidades se vendieron?
- ¿Cuántas unidades se vendieron en la comunidad de Madrid en el mes de Marzo?
- ¿Cuántas unidades se vendieron en la comunidad de Madrid en el mes de Marzo?
Dame un desglose de Alcorcón.
- ¿Cuánto voy a vender el próximo mes en Alcorcón? ¿Por qué?

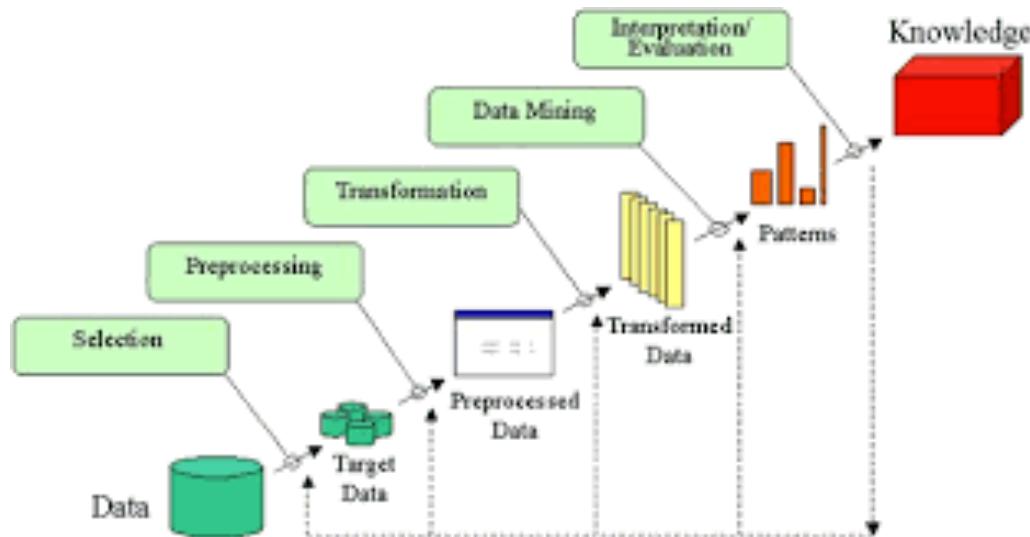
Motivación y orígenes

- Tipos de datos: numérico, textos, imágenes, etc.
- Cantidad de datos: WWW > 13.000.000 Petabytes.
- Interés por entender el porqué.
- El conocimiento es poder.

¿Qué es Data Mining?

Knowledge Discovery in Databases (KDD)

- (Fayyad, Piatetsky-Shapiro, & Smyth 1996): “El proceso no trivial consistente en descubrir patrones válidos, nuevos, potencialmente útiles y comprensibles dentro de un conjunto de datos.”



¿Qué es Data Mining?

Aspectos importantes de KDD:

- Patrones validos: conocimiento contrastable con la realidad.
- Potencialmente útiles: relación con el objetivo que nos proponemos.
- Comprensibles: relacionado con el usuario que maneja el conocimiento extraído de los datos.

¿Qué es Data Mining?

Data mining

- (Fayyad, Piatetsky-Shapiro, & Smyth, 1996): “El proceso de exploración y análisis, por medios automáticos o semiautomáticos, de grandes cantidades de datos para descubrir patrones y reglas útiles.”
- (Maimon & Rokach, 2010): “El proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos.”
- “El proceso de identificación de toda la información que es relevantes y es extraída de grandes cantidades de datos. El objetivo de esta extracción es descubrir patrones y tendencias estructurando la información que se ha obtenido de una manera que sea comprensible para su utilización.”

¿Qué es Data Mining?

Descubrimos conocimiento:

Palabra	Instancias	Porcentaje
Proceso	3	6%
Grandes	3	6%
Datos	3	6%
Descubrir	3	6%
Patrones	3	6%
Cantidades	2	4%
Información	2	4%
Exploración	1	2%
Análisis	1	2%

¿Qué es Data Mining?

No es Data Mining	Es Data Mining
Buscar un nombre en una agenda.	Descubrir según tu agenda los nombres más usados en cada ciudad.
Hacer una búsqueda en Google.	Descubrir que un perfil de persona busca noticias de un tipo.
Buscar en una base de datos lo que compra un usuario.	Encontrar las características de los clientes que compran videojuegos de coches.
Cuantos sofás se vendieron en los 3 primeros meses del año.	En que tiendas hubo unas ventas anómalas de sofás.

Proceso y fases de KDD

Definir que se va a hacer:

- Entender el dominio de aplicación.
- Entender el problema a resolver.
- Fijar los objetivos: asociación, clasificar, agrupar o predecir.

Proceso y fases de KDD

1.- Selección de los datos.

- Tipos de datos que se van a usar y sus fuentes.
- Se seleccionan los datos y se extraen.
- Costoso en tiempo y esfuerzo.
- Ej: data.gov, UCI Machine Learning, kaggle, European Data Portal, FAOStat, etc.

Proceso y fases de KDD

2.- Preprocesamiento.

Preparación y limpieza de los distintos datos de manera que se puedan manejar en el resto de fases. Puede llegar a ser el 70% del esfuerzo.

¿Tienen los datos suficiente calidad?

- Datos necesarios.
- Datos incompletos. Que falte algún valor. Se completan.

Proceso y fases de KDD

- Datos redundantes. Registros con la misma información o muy parecida.
- Datos incorrectos. Cuando el valor del dato se considera erróneo o no ha habido procesos de control.
- Errores de transcripción. Uso de mayúsculas y minúsculas.
- Datos envejecidos. Se han convertido en incorrectos al no ser actualizados.
- Variaciones de datos. Datos con distintas etiquetas.

Proceso y fases de KDD

3.- Transformación.

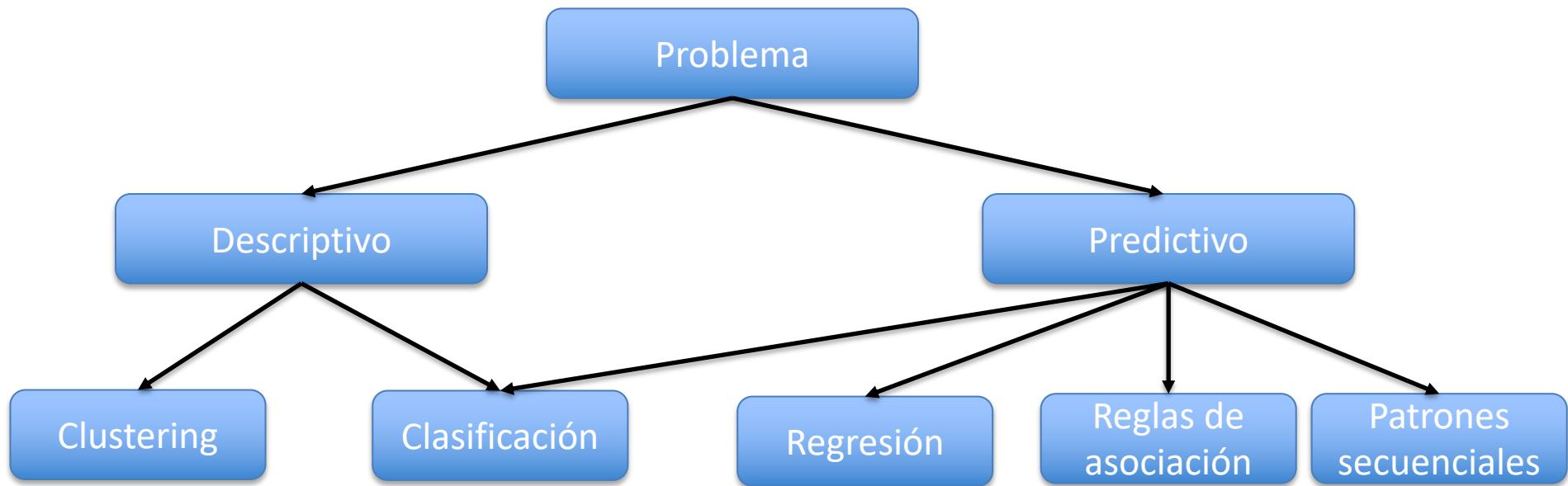
Tratamiento preliminar de datos, transformación y creación de nuevas variables. La fase 2 mejora la calidad, esta los adapta a un algoritmo.

- Convertir datos categóricos a numéricos y viceversa. Asignar a un rango numérico una etiqueta.
- Otras transformaciones: simplificar valores, agrupar valores, normalización de datos, etc.
- Reducción de la dimensionalidad: buscan trabajar con menos datos y mantener la calidad.

Proceso y fases de KDD

4.- Técnicas de Data Mining.

Elección de las técnicas que se van a usar para encontrar patrones o información que estaba oculta extrayendo nuevo conocimiento que dé valor extra a los datos.



Proceso y fases de KDD

- Predictivo: Usamos unas variables cuyos valores son conocidos para determinar el valor de otras variables en el futuro.
- Descriptivo: Encontrar patrones desconocidos interpretables para la persona.

Proceso y fases de KDD

4.1- Reglas de asociación.

- Dado un conjunto de registros, cada uno de los cuáles contiene un conjunto de elementos de una colección concreta.
- Encontrar reglas de dependencia que ayuden a predecir la ocurrencia de un elemento basándonos en la presencia de otros.
- Siguen el patrón: $X \rightarrow Y$

Proceso y fases de KDD

- Características importantes:
 - Cuando aparece un elemento/s por asociación aparece otro/s
 - Se trabaja con atributos categóricos.
 - Los atributos pueden expresarse en forma binaria (ausente o presente).
 - Los elementos de X no aparecen en Y.

Proceso y fases de KDD

- Ejemplos:

1. Gestión de stock en supermercados.
2. Prevención de enfermedades.
3. Gestión de inventarios en servicios técnicos.
4. Tipos de crímenes cometidos.

Proceso y fases de KDD

4.2- Patrones secuenciales.

- Es la extracción de patrones frecuentes relacionados con el tiempo u otro tipo de secuencia.
- Dado un conjunto de objetos, donde cada objeto se asocia a una serie de eventos temporales.
- Se quiere encontrar reglas fiables que predigan de forma correcta dependencias de secuenciamiento entre diferentes eventos.

Proceso y fases de KDD

- Características importantes:
- Son similares a las reglas de asociación pero los atributos tienen temporalidad o siguen una secuencia.

Proceso y fases de KDD

- Ejemplos:
 1. Ofertas temporales en tiendas online.
 2. Secuencias de caracteres como el ADN.
 3. Tendencias de exploración en un Web.
 4. Eventos generados por un sistema de sensores.

Proceso y fases de KDD

4.3- Clasificación.

- Dado una un conjunto de registros. Cada uno con un conjunto de atributos y donde uno de ellos se le denomina atributo de clase
- Encontrar un modelo para el atributo de clase en función de los valores del resto de atributos.
- Hay que validar los resultados sobre un conjunto de datos de validación no usados hasta entonces.

Proceso y fases de KDD

- Características importantes:
- Trabaja con datos etiquetados, supervisado.
- Utiliza un conjunto de datos de entrenamiento y clasifica nuevos datos.
- Encuentra los atributos que definen mejor una clase.

Proceso y fases de KDD

- Ejemplos:
 1. Encontrar un modelo de fraude para clientes de un portal online.
 2. Campañas de marketing en empresas de viajes.
 3. Prevención de abandono en empresas de telecomunicaciones.
 4. Clasificar imágenes.

Proceso y fases de KDD

4.4- Regresión.

- Predicción del valor de una variable basándose en el valor que reflejan otras y siguiendo modelos de dependencia que pueden ser lineales o no lineales (como los exponenciales).
- Cuidado con hacer hipótesis de causalidad.

Proceso y fases de KDD

- Características importantes:
 - Predicen valores numéricos.
 - La clasificación se puede entender como la predicción de una clase.

Proceso y fases de KDD

- Ejemplos:
 1. Lluvias con respecto a datos de temperatura y presión
 2. Ventas en función del gasto en publicidad.
 3. Riesgo de infarto en función del peso, edad y altura.
 4. Posibilidad de morir en Game of Thrones.

Proceso y fases de KDD

4.5- Clustering.

- Dado un conjunto de individuos con un conjunto de atributos individuales, y unas medidas de similitud entre ellos, encontrar agrupaciones similares.
- Las medidas de similitud pueden ser: la distancia euclídea para atributos continuos. En caso contrario: color, palabras frecuentes, etc.

Proceso y fases de KDD

- Características importantes:
- Trabaja con datos no etiquetados, no supervisado.
- Sugiere nuevos grupos con respecto a los atributos.

Proceso y fases de KDD

- Ejemplos:
 1. Agrupar clientes para campañas de marketing personalizadas.
 2. Documentos según su ámbito.
 3. Tipos de estudiantes según su personalidad.
 4. Reacciones de usuarios con móviles para GUI.

Proceso y fases de KDD

5.- Interpretación y evaluación.

Los modelos o patrones obtenidos deberían aceptar o contradecir la hipótesis inicial. Los patrones deben presentarse de forma que sean entendibles. Es por ello que las técnicas de visualización son muy útiles.

Proceso y fases de KDD

Integrar los resultados:

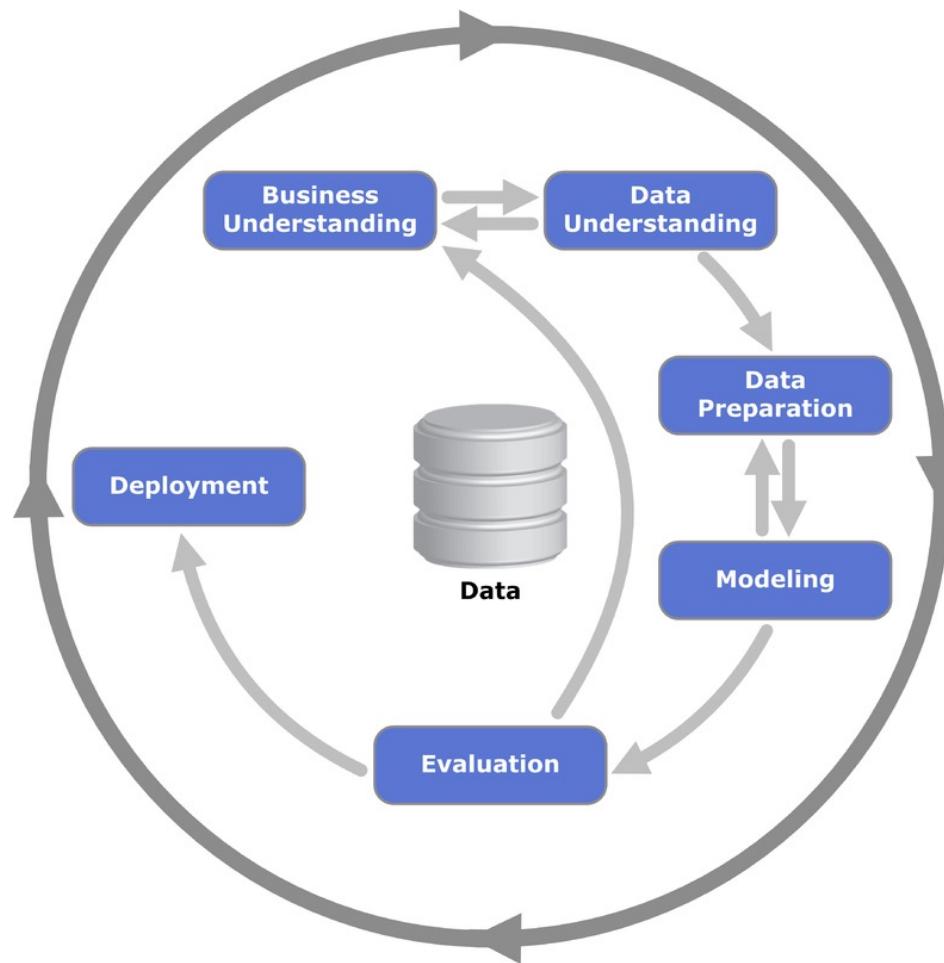
- Usar el modelo obtenido en el sistema de información que se está trabajando.
- Continuar con el proceso de KDD usando los nuevos datos que se obtienen.

Metodología CRISP-DM

CRoss-Industry Standard Process for Data Mining:

- Estándar abierto de Data Mining creado en 1996.
- En 1997 empezó como un proyecto financiado por la UE.
- Se estima que es usado en aproximadamente un 42% de los casos.

Metodología CRISP-DM



Metodología CRISP-DM

1.- Comprensión del negocio (Objetivos y requerimientos desde una perspectiva no técnica):

- Establecimiento de los objetivos del negocio (Contexto inicial, objetivos, criterios de éxito)
- Evaluación de la situación (Inventario de recursos, requerimientos, supuestos, terminologías propias del negocio)
- Establecimiento de los objetivos del Data Mining (objetivos y criterios de éxito)
- Generación del plan del proyecto (plan, herramientas, equipo y técnicas)

Metodología CRISP-DM

2.- Comprepción de los datos (Familiarizarse con los datos teniendo presente los objetivos del negocio):

- Recopilación inicial de datos.
- Descripción de los datos.
- Exploración de los datos.
- Verificación de calidad de datos

Metodología CRISP-DM

3.- Preparación de los datos (Obtener el formato final del dataset):

- Selección de los datos.
- Limpieza de datos.
- Construcción de datos.
- Integración de datos.
- Formateo de datos.

Metodología CRISP-DM

4.- Modelado (Aplicar las técnicas de minería de datos a los dataset):

- Selección de la técnica de modelado.
- Diseño de la evaluación.
- Construcción del modelo.
- Evaluación del modelo.

Metodología CRISP-DM

5.- Evaluación (Determinar si los resultados son útiles a las necesidades del negocio):

- Evaluación de resultados.
- Revisar el proceso.
- Establecimiento de los siguientes pasos o acciones.

Metodología CRISP-DM

6.- Despliegue (Explotar utilidad de los modelos, integrándolos en las tareas de toma de decisiones de la organización):

- Planificación de despliegue.
- Planificación de la monitorización y del mantenimiento.
- Generación de informe final.
- Revisión del proyecto.

Metodología CRISP-DM

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
<p>Determine Business Objectives <i>Background</i> <i>Business Objectives</i> <i>Business Success Criteria</i></p> <p>Assess Situation <i>Inventory of Resources Requirements, Assumptions, and Constraints</i> <i>Risks and Contingencies</i> <i>Terminology</i> <i>Costs and Benefits</i></p> <p>Determine Data Mining Goals <i>Data Mining Goals</i> <i>Data Mining Success Criteria</i></p> <p>Produce Project Plan <i>Project Plan</i> <i>Initial Assessment of Tools and Techniques</i></p>	<p>Collect Initial Data <i>Initial Data Collection Report</i></p> <p>Describe Data <i>Data Description Report</i></p> <p>Explore Data <i>Data Exploration Report</i></p> <p>Verify Data Quality <i>Data Quality Report</i></p>	<p>Select Data <i>Rationale for Inclusion/Exclusion</i></p> <p>Clean Data <i>Data Cleaning Report</i></p> <p>Construct Data <i>Derived Attributes</i> <i>Generated Records</i></p> <p>Integrate Data <i>Merged Data</i></p> <p>Format Data <i>Reformatted Data</i></p> <p>Dataset <i>Dataset Description</i></p>	<p>Select Modeling Techniques <i>Modeling Technique</i> <i>Modeling Assumptions</i></p> <p>Generate Test Design <i>Test Design</i></p> <p>Build Model <i>Parameter Settings</i> <i>Models</i> <i>Model Descriptions</i></p> <p>Assess Model <i>Model Assessment</i> <i>Revised Parameter Settings</i></p>	<p>Evaluate Results <i>Assessment of Data Mining Results w.r.t. Business Success Criteria</i> <i>Approved Models</i></p> <p>Review Process <i>Review of Process</i></p> <p>Determine Next Steps <i>List of Possible Actions</i> <i>Decision</i></p>	<p>Plan Deployment <i>Deployment Plan</i></p> <p>Plan Monitoring and Maintenance <i>Monitoring and Maintenance Plan</i></p> <p>Produce Final Report <i>Final Report</i> <i>Final Presentation</i></p> <p>Review Project <i>Experience Documentation</i></p>



Metodología CRISP-DM

KDD	SEMMA	CRISP-DM
Pasos previos	-	Comprensión del negocio
Selección	Muestreo	Comprensión de los datos
Preprocesamiento	Exploración	Preparación de los datos
Transformación	Manipulación	Preparación de los datos
Data Mining	Modelado	Modelado
Evaluación e interpretación	Valoración	Evaluación
Pasos posteriores	-	Despliegue

Casos de uso

Asociación: Encontrar ítems/características que se dan en común.

Romero, C., Romero, J., Luna, J. & Ventura, S. (2010). Mining Rare Association Rules from e-Learning Data. *EDM, RSJ de Baker, A. Merceron, and PIP Jr., Eds.* www.educationaldatamining.org, , 171--180.

Objetivo: Encontrar patrones en el uso de Moodle según la interacción de los estudiantes.

Casos de uso

230 estudiantes

Name	Description	Values
course	Identification number of the course.	C218, C94, C110, C111, C46
n_assignment	Number of assignments done.	ZERO, LOW, MEDIUM, HIGH
n_quiz	Number of quizzes taken.	ZERO, LOW, MEDIUM, HIGH
n_quiz_a	Number of quizzes passed.	ZERO, LOW, MEDIUM, HIGH
n_quiz_s	Number of quizzes failed.	ZERO, LOW, MEDIUM, HIGH
n_posts	Number of messages sent to the forum.	ZERO, LOW, MEDIUM, HIGH
n_read	Number of messages read on the forum.	ZERO, LOW, MEDIUM, HIGH
total_time_assignment	Total time spent on assignments.	ZERO, LOW, MEDIUM, HIGH
total_time_quiz	Total time spent on quizzes.	ZERO, LOW, MEDIUM, HIGH
total_time_forum	Total time spent on forum.	ZERO, LOW, MEDIUM, HIGH
mark	Final mark obtained by the student in the course.	ABSENT, FAIL, PASS, EXCELLENT

Casos de uso

Reglas obtenidas

Rule	Antecedent	Consequent
1	total_time_forum=HIGH	mark=PASS
2	n_posts=MEDIUM AND n_read=MEDIUM AND n_quiz_a=MEDIUM	mark=PASS
3	course=C110 AND n_assignment=HIGH	mark=PASS
4	total_time_quiz=LOW	mark=FAIL
5	n_assignment=LOW	mark=FAIL
6	n_quiz_a=LOW AND course=C218	mark=FAIL

Rule	Antecedent	Consequent
1	n_quiz=HIGH AND n_quiz_a=HIGH	mark=EXCELLENT
2	total_time_assignment=HIGH	mark=EXCELLENT
3	n_posts=HIGH AND course=C46	mark=EXCELLENT
4	total_time_assignment=ZERO AND total_time_forum=ZERO AND total_time_quiz=ZERO]	mark=ABSENT
5	n_posts=ZERO AND n_read=ZERO	mark=ABSENT
6	n_quiz=ZERO AND course=C111	mark=ABSENT

Casos de uso

Clasificación: Crear un clasificador con una serie de individuos. Luego usarlo para asignar una clase a los individuos nuevos.

Jabbar, M. A., Deekshatulu, B. L. & Chandra, P. (2015). Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm.. *CoRR*, abs/1508.02061.

Objetivo: Crear una herramienta de toma de decisiones para diagnosticar enfermedades del corazón.

Casos de uso

Probado con varios datasets. Se obtienen accuracies bastante altas en algunos casos.

Sl.no	Attribute	Data Type
1	Age	Numeric
2	Gender	Nominal
3	Diabetic	Nominal
4	BP Systolic	Numeric
5	BP Dialis	Numeric
6	Height	Numeric
7	Weight	Numeric
8	BMI	Numeric
9	Hypertension	Nominal
10	Rural	Nominal
11	Urban	Nominal
12	Disease	Nominal

Data set name	K=1	K=3	K=5
Weather data	85.71	85.71	85.71
Breast cancer	90	90	82.5
Heart stalog	100	90.74	83.3
Lympography	99.32	99.32	84.4
Hypothyroid	100	95.62	94.69
Primary tumor	75	65.48	61.35
Heart disease A.P	95	75	83.3

Casos de uso

Predicción: Dados unos datos históricos saber si un evento ocurrirá o no.

Kanyongo, G.Y., Certo, J.L., & Launcelot, B.I. (2006). Using Regression Analysis to Establish the Relationship between Home Environment and Reading Achievement: A Case of Zimbabwe.

Objetivo: Establecer que características propias del hogar están directamente relacionadas con el hábito de lectura.

Casos de uso

1329 chicos y 1368 chicas.

Variable	Scale	M	SD
Reading Score		25.60	10.05
Gender		1.51	0.50
Boys	1		
Girls	2		
Place the student stays during school week		1.25	0.43
With parents	1		
Other	2		
Reading at home		1.74	0.44
Never	1		
Sometimes	2		
SES (Combined the following)		1.24	0.35
<i>Possession of a TV</i>			
<i>Possession of a refrigerator</i>			
<i>Possession of piped water</i>			
<i>Possession of electricity</i>			
No	1		
Yes	2		
Meals (combined the following)		2.86	0.60
<i>Breakfast</i>			
<i>Lunch</i>			
<i>Supper</i>			
Not at all	1		
1 or 2 times per week	2		
3 or 4 days per week	3		
Everyday	4		
Home (combined the following)		1.26	0.44
<i>Someone makes sure you did homework</i>			
<i>Someone helps with homework</i>			
<i>Someone asks you to read to him/her</i>			
<i>Someone questions on what you read</i>			
<i>Someone looks at school work</i>			
Never	1		
Sometimes	2		
Most of the times	3		
Number of books at home		2.28	1.14

Casos de uso

Las propiedades que están directamente relacionadas son: el lugar donde se vive durante la semana, comodidades de la casa, número de comidas que hace a la semana, actividades en las que participa en casa y tiempo leyendo en casa.

	Estimate	S.E.	P
Reading score <-- Gender	0.52	0.34	0.13
Reading score <--Stay	-2.90	0.40	0.00
Reading score <-- SES	7.71	0.52	0.00
Reading score <-- Books	0.15	0.17	0.39
Reading score <-- Meals	1.88	0.30	0.00
Reading score <-- Home	1.50	0.34	0.00
Reading score <-- Home reading	3.74	0.44	0.00

Casos de uso

Clustering: Agrupar individuos de manera que los individuos cercanos se les considere de un mismo grupo.

Saylı, A., Ozturk, I., & Ustunel, M. (2016). Brand loyalty analysis system using K-Means algorithm.

Objetivo: Crear un Sistema que analice la lealtad de los clientes en las tiendas Migros Ticaret respecto a la marca en general, a los tipos de productos y a productos concretos.

Casos de uso

1628 registros

Account	Item	Brand Code	Total Distinct Day	Year Sum Quantity	Avg Sum Quantity	Category	Main	CGroup	Class	SubClass
CUSTOMER2	ITEM4	BRAND3	1	1,42	0,88	5	311	15	10	10
CUSTOMER2	ITEM5	BRAND3	1	1,15	0,91	5	311	15	10	10
CUSTOMER2	ITEM6	BRAND4	4	12	1,6	5	311	10	20	15

Casos de uso

Se obtienen 15 clusters en los que se puede ver que la lealtad a marcas y productos varía.

Cluster No	Sum Quantity Intervals	Number of Distinct Brands	Number of Items
Cluster 0	3,706.00 –	3	3
Cluster 1	1,620.00 –	8	9
Cluster 2	932.00 – 1,551.00	16	24
Cluster 3	578.00 - 892.00	27	42
Cluster 4	376.00 - 557.00	40	74
Cluster 5	261.00 - 373.00	66	113
Cluster 6	181.00 - 259.00	95	194
Cluster 7	126.00 - 180.00	128	290
Cluster 8	84.00 - 125.00	183	482
Cluster 9	52.00 - 83.00	247	732
Cluster 10	30.00 - 51.81	363	1159
Cluster 11	15.00 - 29.49	526	1953
Cluster 12	5.97 - 14.60	797	3713
Cluster 13	1.94 - 5.78	1162	7760
Cluster 14	0.00 - 1.90	888	3859

Casos de uso

[María Medina](#), data scientist en PiperLab y co-organizadora de la comunidad PyLadies Madrid.

<https://youtu.be/8I4XUYjon-c>