



SOLID
QUALITY
MENTORS

Minería de Datos

Para el común de los mortales

ANTONIO SOTO

Director General

@antoniosql



<http://www.solidq.com>

Agenda

- Introducción a Minería de Datos
 - Problemas de Negocio
 - Tareas de Minería de Datos
 - Ciclo de un Proyecto de Minería de Datos
- Clasificación de Algoritmos de SQL Server Data Mining
- Minería de Datos aplicada utilizando Office

Introducción a Minería de Datos

- ¿Qué es la Minería de Datos?
- Problemas de Negocio
- Tareas de Minería de Datos
- Ciclo de un Proyecto de Minería de Datos

¿Qué es la Minería de Datos?

- «La **minería de datos** (**DM**, *Data Mining*) consiste en la extracción no trivial de información que reside de manera implícita en los datos. Dicha información era previamente desconocida y podrá resultar útil para algún proceso. En otras palabras, la minería de datos **prepara**, **sondea** y **explora** los datos para sacar la información oculta en ellos»

Fuente: Wikipedia

¿Qué es la Minería de Datos? (II)

- Deducir conocimiento examinando los datos y realizando predicciones
 - «examinar datos» examinar ejemplos de hechos conocidos sobre «casos» utilizando sus atributos – «variables»
 - «conocimiento»: Patrones, Clusters, Reglas, Árboles de Decisión, Redes Neuronales, Reglas de Asociación,....
- OLAP: Análisis orientado al modelo
- DM: Análisis orientado al dato
- Nombres alternativos: Análisis Predictivo

Problemas de Negocio

- Generación de Recomendaciones
 - ¿Qué productos o servicios deberíamos de ofrecer a nuestros clientes?
- Detección de anomalías
 - Detección de fraude
- Análisis de Rotación
 - ¿Qué clientes son más proclives de irse a la competencia?
- Gestión de Riesgos
 - ¿Debería de concederse el préstamo?
- Segmentación de clientes
 - Clasificación de nuestros clientes
- Anuncios Orientados
 - Personalización de anuncios, contenido,...
- Previsión
 - ¿Cuánto venderemos el próximo trimestre?

Tareas de Minería de Datos (I)

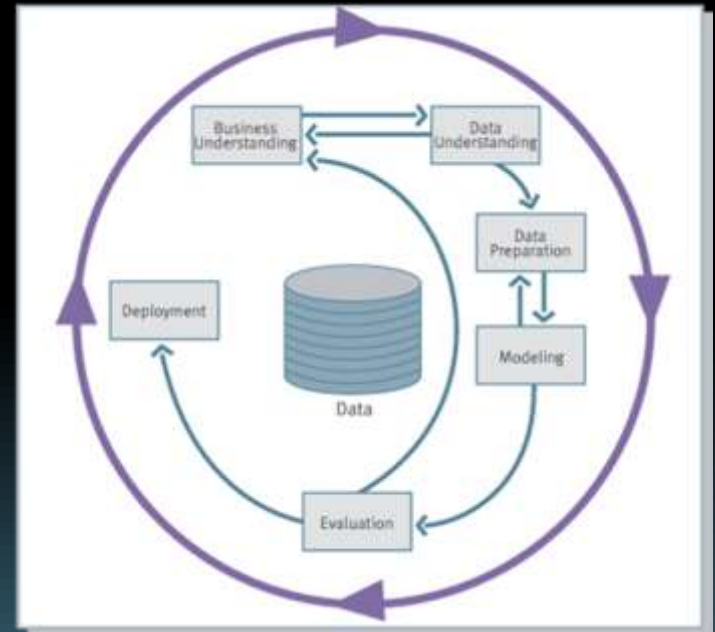
- **Clasificación**
 - Se asigna una categoría a cada caso. Cada caso tiene un conjunto de atributos uno de ellos es el atributo clase.
 - Se busca un modelo que describa el atributo clase como una función de los atributos de salida
- **Agrupación**
 - También conocido como segmentación
 - Identifica grupos naturales basándose en un conjunto de atributos
- **Asociación**
 - También conocido como análisis de cesta de la compra
- **Regresión**
 - Similar a clasificación pero con el objetivo de buscar patrones para determinar un valor numérico
 - Ej.: Predicción de la velocidad del viento basada en temperatura presión de aire y humedad

Tareas de Minería de Datos (II)

- Previsión
 - La entrada es un conjunto de valores a lo largo del tiempo de los que extrae valores futuros
- Análisis de secuencia
 - Busca patrones en una serie de eventos llamada secuencia
 - Ej. Secuencia de navegación en Web
- Análisis de desviaciones
 - Busca casos «raros» diferentes a los demás

Ciclo de un Proyecto de Minería (I)

- Formulación del Problema de Negocio
- Recolección de Datos
- Limpieza y Transformación de Datos
 - Transformación numérica
 - Agrupación
 - Agregación
 - Manejo de valores «perdidos»
 - Eliminar os «extremos»
- Creación del Modelo
 - Selección del Algoritmo
 - Prueba y Error en muchos casos



Ciclo de un Proyecto de Minería (II)

- Evaluación del Modelo
 - Evaluar la fiabilidad del modelo dentro de nuestro negocio
- Reporting y Predicción
- Integración en Aplicaciones
- Gestión del Modelo
 - Dependiendo del escenario puede ser muy volátil
 - Planificar «Entrenamiento»



Clasificación de Algoritmos

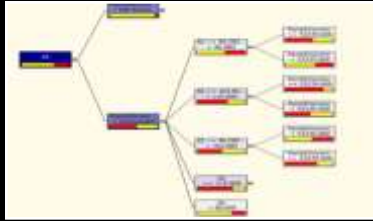
Clasificación de Algoritmos

- Tipos de Algoritmos
- Algoritmos de SQL Server Data Mining
- Clasificación de Algoritmos

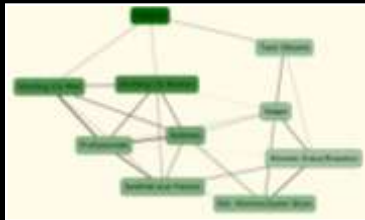
Tipos de Algoritmos

- De Forecasting.
 - Dada una tendencia ¿Cuál es la previsión?
- Supervisados.
 - Conocemos la respuesta ¿Qué está correlacionado?
- No Supervisados.
 - Desconocemos la respuesta ¿Cuáles son los grupos?

Algoritmos (I)



- Decision Trees
 - El más popular
 - Utilizado para clasificación

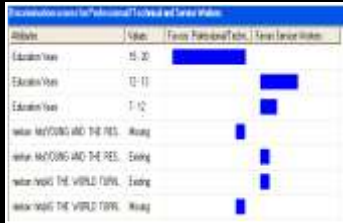


- Clustering
 - Encuentra agrupaciones naturales



- Sequence Clustering
 - Agrupo una secuencia de eventos discretos en grupos naturales basado en similaridad
 - Entender como los visitantes usan una web

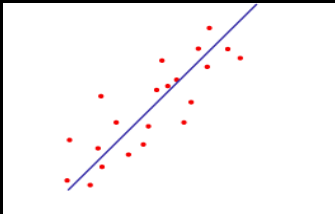
Algoritmos (II)



Attributes	Values	Feature Parameters/Values	Feature Parameter Values
EducationYears	15-20		
EducationYears	12-13		
EducationYears	7-12		
winner: HAZYUNG AND THE RES.	Winning		
winner: HAZYUNG AND THE RES.	Loosing		
winner: HAZYUNG AND THE RES.	Loosing		
winner: HAZYUNG AND THE RES.	Loosing		

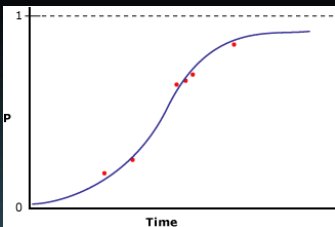
- Naïve Bayes

- Clasificación en escenarios similares a Decision Trees



- Linear Regression

- Encuentra la mejor línea recta posible a través de una serie de puntos
- Usado para análisis predictivo



- Logistic Regression

- Se adapta a un factor exponencial
- Usado para análisis predictivo

Algoritmos (III)



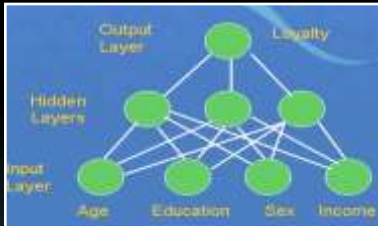
- Association Rules

- ## ■ Análisis de Cesta de la compra



- Time Series

- ▣ Algoritmo de previsión usado para previsión a corto y largo plazo
- ▣ Podemos usar varios escenarios para predecir escenarios “what if”



- Neural Network

- Tareas de Clasificación y Regresión
- Más sofisticado que Decision Trees y Naïve Bayes

Clasificación de Algoritmos

Tipo	Algoritmos
Forecasting	Time Series
Supervisados	Naive Bayes Linear Regression Decision Trees Neural Network Logistic Regression
No Supervisados	Clustering Sequence Clustering Association Rules Text Mining



Minería de Datos en Office

Minería de Datos en Office

- Tareas de Preparación de Datos
- Table Analysis Tools
- Data Modeling Tools
- Visores
- Testeo y Validación

Tareas de Preparación de Datos

- Explorar Datos
 - Información estadística básica
- Limpieza de datos
 - Outliers
 - Re-label
- Datos de Ejemplo





DEMO

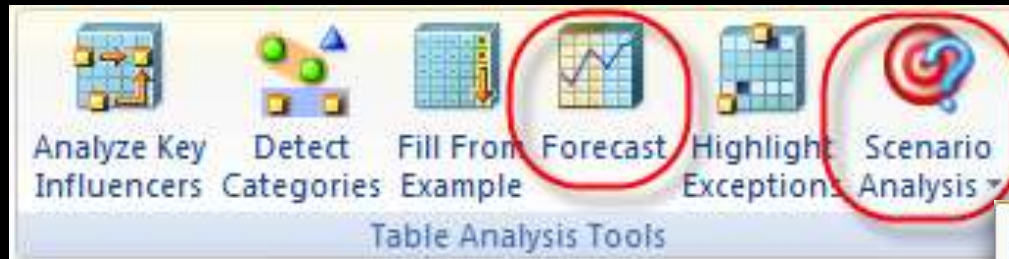
Preparando los datos

Table Analysis Tools (I)



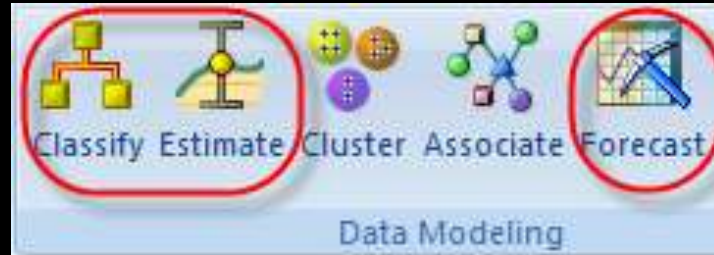
Herramienta	Muestra	Algoritmo
Analyze Key Influencers	Discriminación	Naïve Bayes
Detect Categories	Características clave de cada categoría	Naïve Bayes
Highlight Exceptions	Valores improbables	Microsoft Clustering

Table Analysis Tools(II)



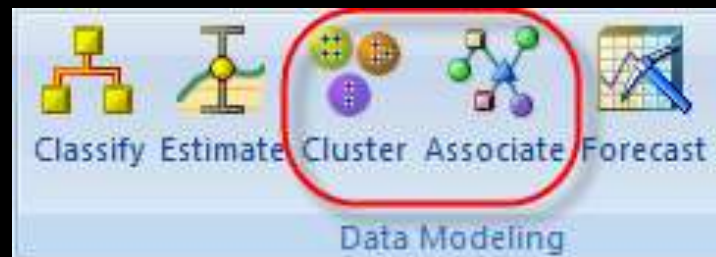
Herramienta	Algoritmo	Using This Algorithm
Forecast	Datos sobre tiempo	Microsoft Time Series
Goal Seek	Valores que deben de cambiar para cumplir un objetivo	Logistic Regression
What If	Cambio en valor predecido si cambian los valores de entrada	Logistic Regression

Data Modeling Tools



Herramienta	Predice	Algoritmo
Classify	Tendencia	Microsoft Decision Trees
Estimate	Factores que afectan a una salida	Microsoft Decision Trees
Forecast	Datos continuos sobre tiempo	Microsoft Time Series

Data Modeling Tools



Herramienta	Muestra	Algoritmo
Cluster	Grupos de filas con características comunes	Microsoft Clustering
Associate	Items encontrados juntos en múltiples transacciones	Microsoft Association

Visores de Modelo



Cluster Diagram



Cluster Profiles



Cluster Characteristics

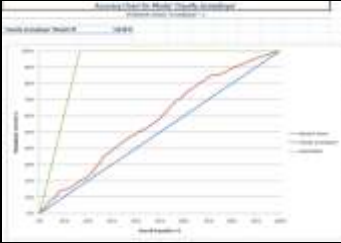


Cluster Discrimination

Otros Visores

- Decision tree
- Neural network
- Association rules
- Time series

Testeo y Validación



Accuracy Chart

Summary of Overall System Classification for model 1 based on average accuracy

Notes: Overall accuracy is calculated as follows:

Category	Actual	Predicted	Count
Total correct	83.25%	811.0	
Total misclassified	16.75%	162.0	

Results as Pre-conditions

Category	Actual	Predicted	Count
Correct	100.00%	100.00%	100
Misclassified	0.00%	0.00%	0

Results as Events

Category	Actual	Predicted	Count
Correct	83.25%	811.0	
Misclassified	16.75%	162.0	

Classification Matrix



Profit Chart

Resumen

- Problemas de Negocio que podemos resolver
- Que Tareas de Minería de Datos podemos aplicar
- Ciclo de un Proyecto de Minería de Datos
- Clasificación de Algoritmos de SQL Server Data Mining
- Minería de Datos aplicada utilizando Office



SOLID
QUALITY
MENTORS

Minería de Datos

Para el común de los mortales

ANTONIO SOTO

Director General

@antoniosql



<http://www.solidq.com>