MINERIA DE DATOS

PhD Maria Hallo

Mineria de datos

- La minería de datos suele describirse como *"el proceso de extraer información válida, auténtica y que se pueda procesar de las bases de datos de gran tamaño."*
- La minería de datos deriva patrones y tendencias que existen en los datos. Estos patrones y tendencias se pueden recopilar y definir como un modelo de minería de datos.

Ejemplos de aplicaciones

Los modelos de minería de datos se pueden aplicar a situaciones empresariales como las siguientes:

- Predecir ventas.
- Dirigir correo a clientes específicos.
- Determinar los productos que se pueden vender juntos.
- Buscar secuencias en el orden en que los clientes agregan productos a una cesta de compra.

Proceso de mineria de datos

- Consiste en los siguientes pasos
- Definir el problema
- Preparar los datos
- Explorar los datos
- Generar modelos
- Explorar y validar los modelos

Definir el problema

- Este paso incluye analizar los requisitos empresariales, definir el ámbito del problema, definir las métricas por las que se evaluará el modelo y definir el objetivo final del proyecto de minería de datos. Se aplican preguntas como las siguientes:
- ¿Qué está buscando?
- ¿Qué atributo del conjunto de datos desea intentar predecir?
- ¿Qué tipos de relaciones intenta buscar?
- ¿Desea realizar predicciones a partir del modelo de minería de datos o sólo buscar asociaciones y patrones interesantes?
- ¿Cómo se distribuyen los datos?

Preparar los datos

- Los datos pueden estar dispersos en la empresa y almacenados en distintos formatos; también pueden contener incoherencias como entradas que faltan o contienen errores.
- Se debe integrar los datos y realizar una limpieza
- Se puede usar herramientas como integration services para realizar esta tarea

Explorar los datos

- Debe comprender los datos para tomar las decisiones adecuadas al crear los modelos. Entre las técnicas de exploración se incluyen calcular los valores mínimos y máximos, calcular la media y las desviaciones estándar y examinar la distribución de los datos. Una vez explorados los datos, se puede decidir si el conjunto de datos contiene datos con errores y, a continuación, crear una estrategia para solucionar los problemas.
- El Diseñador de vistas de origen de datos de BI Development Studio contiene varias herramientas que se pueden utilizar para explorar los datos.

Generar modelos

- Antes de generar un modelo, se deben separar aleatoriamente los datos preparados en conjuntos de datos de entrenamiento y comprobación independientes. El conjunto de datos de entrenamiento se utiliza para generar el modelo y el conjunto de datos de comprobación para comprobar la precisión del modelo mediante la creación de consultas de predicción. Puede utilizar la Transformación Muestreo de porcentaje de Integration Services para dividir el conjunto de datos.

Generar modelos

- Normalmente, los modelos contienen columnas de entrada, una columna de identificación y una columna de predicción
- Una vez definida la estructura del modelo de minería de datos, se procesa rellenando la estructura vacía con los patrones que describen el modelo.
- Esto se conoce como *entrenar* el modelo.
- Los patrones se encuentran al pasar los datos originales por un algoritmo matemático

Generar modelos

- El modelo de minería de datos se define mediante un objeto de estructura de minería de datos, un objeto de modelo de minería de datos y un algoritmo de minería de datos.

Algoritmos incluidos en Sqlserver

- Algoritmo de árboles de decisión de Microsoft
- Algoritmo de clústeres de Microsoft
- Algoritmo Bayes naive de Microsoft
- Algoritmo de asociación de Microsoft
- Algoritmo de clústeres de secuencia de Microsoft
- Algoritmo de serie temporal de Microsoft
- Algoritmo de red neuronal de Microsoft (SSAS)
- Algoritmo de regresión logística de Microsoft
- Algoritmo de regresión lineal de Microsoft

Explorar y validar los modelos

- No se debe implementar un modelo en un entorno de producción sin comprobar primero si el modelo funciona correctamente. Además, puede que haya creado varios modelos y deba decidir cuál funciona mejor. Si ninguno de los modelos que ha creado en el paso Generar modelos funciona correctamente, puede que deba volver a un paso anterior del proceso y volver a definir el problema o volver a investigar los datos del conjunto de datos original.

Implementar los modelos

- Utilizar los modelos para crear predicciones que pueda utilizar para tomar decisiones empresariales. SQL
- Incrustar la funcionalidad de minería de datos en una aplicación.
- Crear un paquete en el que se utilice un modelo de minería de datos para separar de forma inteligente los datos entrantes en varias tablas.
- Crear un informe que permita a los usuarios realizar consultas directamente en un modelo de minería de datos existente.
- La actualización del modelo forma parte de la estrategia de implementación. A medida que la organización recibe más datos, debe volver a procesar los modelos para mejorar así su eficacia.