**Universidad Mariano Gálvez**

**Centro Universitario Retalhuleu**

**Algoritmo de Bayes Naive de Microsoft**

**Zelik Uriel Ramírez Reyes**

**2890-14-8668**

**Base de datos II**

**Ing.: Jorge Santos**

**Introducción**

La minería de datos, es una técnica en la que aplicamos a una gran base de datos, algoritmos de búsqueda de patrones de comparación. Para poder trabajar con la minería de datos necesitamos 2 cosas: una base de datos en una instancia de Analysis Services de SQL Server con un modelo de minería y el complemento de minería de datos de Office. Cuando trabajamos con grandes bases de datos, tipo datawarehouse, hay veces que analizando distintos datos es posible predecir algunos patrones. Por ejemplo, analizando las ventas de los últimos 5 años y analizando distintos factores de nuestros clientes, es posible por ejemplo, descubrir que los “clientes que van en bicicleta y tienen 2 hijos son más propensos a comprar un coche”, o que si “vendemos el producto en los meses de Febrero y Marzo las ventas de dicho producto subirá un 30%”. Dichos patrones es lo que denominamos **Minería de datos**. A continuación veremos información relacionada acerca del algoritmo de Bayes Naive de Microsoft.

**¿Qué es Algoritmo Bayes naive de Microsoft?**

El algoritmo Bayes naive de Microsoft es un algoritmo de clasificación basado en los teoremas de Bayes y se puede usar para el modelado de predicción y de exploración. La palabra naive (ingenuo en inglés) del término Bayes naive proviene del hecho que el algoritmo utiliza técnicas Bayesianas pero no tiene en cuenta las dependencias que puedan existir.

Desde el punto de vista computacional, el algoritmo es menos complejo que otros algoritmos de Microsoft y, por tanto, resulta útil para generar rápidamente modelos de minería de datos que detectan las relaciones entre las columnas de entrada y las columnas de predicción. Puede utilizar este algoritmo para realizar la exploración inicial de los datos y, más adelante, aplicar los resultados para crear modelos de minería de datos adicionales con otros algoritmos más complejos y precisos desde el punto de vista computacional.

**Ejemplo**

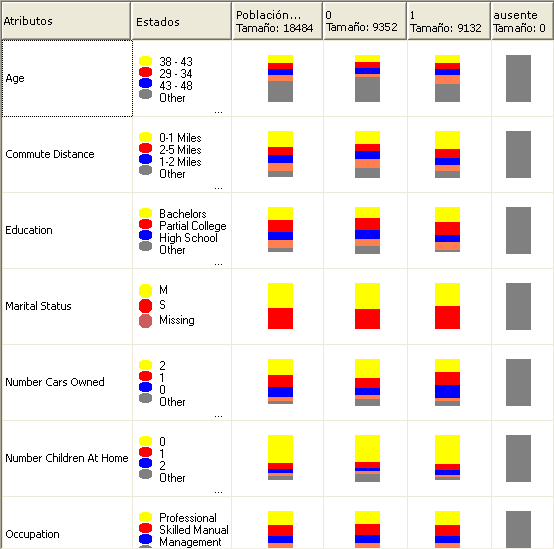
Como parte de su estrategia promocional, el departamento de comercialización de la empresa Adventure Works Cycles ha decidido atraer a posibles clientes realizando un envío por correo de folletos. Para reducir costos, desean enviar los folletos solo a los clientes de los que esperan recibir respuesta. La empresa almacena información en una base de datos sobre datos demográficos y respuestas a envíos de correo anteriores. Desean utilizar estos datos para ver el modo en que los datos demográficos como la edad o la ciudad pueden ayudarles a predecir la respuesta a una promoción, comparando los clientes potenciales con los que tienen características similares y con los que han adquirido productos de la empresa en el pasado. En concreto, lo que desean es ver las diferencias entre los clientes que adquirieron una bicicleta y los que no lo hicieron.

Mediante el algoritmo Bayes naive de Microsoft , el departamento de comercialización pude predecir rápidamente un resultado de un perfil de cliente concreto y, por tanto, puede determinar qué clientes responderán a los folletos con más probabilidad. Con el Visor Bayes naive de Microsoft de SQL Server Data Tools (SSDT), también pueden investigar visualmente qué columnas de entrada específicas contribuyen a conseguir respuestas positivas a los folletos.

**¿Cómo funciona el algoritmo?**

El algoritmo Bayes naive de Microsoft calcula la probabilidad de cada estado de cada columna de entrada, dado cada posible estado de la columna de predicción.

Para comprender cómo funciona, utilice el Visor Bayes naive de Microsoft de SQL Server Data Tools (SSDT) (como se muestra en el siguiente gráfico) para consultar una representación visual del modo en que el algoritmo distribuye los estados.



Aquí, el Visor Bayes naive de Microsoft muestra cada columna de entrada del conjunto de datos e indica cómo se distribuyen los estados de cada columna, dado cada estado de la columna de predicción.

Esta vista del modelo se utilizaría para identificar las columnas de entrada que son importantes para diferenciar los distintos estados de la columna de predicción.

Por ejemplo, en la fila Commute Distance que se muestra aquí, la distribución de valores de entrada es visiblemente diferente para los compradores en comparación con los no compradores. Esto indica que la entrada, Commute Distance = 0-1 miles, es un factor de predicción potencial.

El visor también proporciona valores para las distribuciones, de modo que pueda ver que para los clientes que viajan entre una y dos millas para ir a trabajar, la probabilidad de que compren una bicicleta es de 0,387, y la probabilidad que no la compren es de 0,287. En este ejemplo, el algoritmo utiliza la información numérica, derivada de un dato de cliente (como la distancia entre el domicilio y el lugar de trabajo), para predecir si un cliente compraría una bicicleta.

## Datos requeridos para los modelos Bayes naive

Al preparar los datos para su uso en un modelo de entrenamiento Bayes naive, conviene comprender qué requisitos son imprescindibles para el algoritmo, incluidos el volumen de datos necesario y la forma en que estos datos se utilizan.

Los requisitos para un modelo Bayes naive son los siguientes:

* **Una columna de una sola clave** : cada modelo debe contener una columna numérica o de texto que identifique cada registro de manera única. No están permitidas las claves compuestas.
* **Columnas de entrada**: en un modelo Bayes naive, todas las columnas deben ser discretas o se deben haber discretizado los valores. Para más información sobre cómo discretizar columnas.
* **Las variables deben ser independientes.** En un modelo Bayes naive, también es importante asegurarse de que los atributos de entrada sean independientes unos de otros. Esto es particularmente importante al utilizar el modelo para la predicción. Si usa dos columnas de datos que ya están estrechamente relacionadas, el efecto sería multiplicar la influencia de esas columnas, lo que puede ocultar otros factores que influyen en el resultado.

Al contrario, la capacidad del algoritmo de identificar las correlaciones entre las variables es útil cuando está explorando un modelo o conjunto de datos, para identificar las relaciones entre las entradas.

* **Al menos una columna de predicción** : el atributo de predicción debe contener valores discretos o discretizados.

## Ver el modelo

Para explorar el modelo, puede usar el **Visor Bayes naive de Microsoft**. El visor muestra cómo se relacionan los atributos de entrada con el atributo de predicción. El visor también proporciona un perfil detallado de cada clúster, una lista de los atributos que distinguen cada clúster de los demás, y las características del conjunto de datos de entrenamiento completo.

## Realizar predicciones

Una vez entrenado el modelo, los resultados se almacenan como un conjunto de patrones que se puede explorar o utilizar para realizar predicciones.

Puede crear consultas para devolver las predicciones sobre cómo se relacionan los nuevos datos con el atributo de predicción, o puede recuperar estadísticas que describan las correlaciones que ha hallado el modelo.

# Referencia técnica del algoritmo Bayes naive de Microsoft

El algoritmo Bayes naive de Microsoft es un algoritmo de clasificación que proporciona Microsoft SQL Server Analysis Services para el modelado predictivo. Este algoritmo calcula la probabilidad condicional entre columnas de entrada y de predicción y supone que las columnas son independientes. Esta suposición de independencia conduce al nombre Bayes naive.

## Implementación del algoritmo Bayes naive de Microsoft

Desde el punto de vista computacional, el algoritmo es menos complejo que otros algoritmos de Microsoft y, por tanto, resulta útil para generar rápidamente modelos de minería de datos que detectan las relaciones entre las columnas de entrada y las columnas de predicción. El algoritmo considera cada par de valores de atributos de entrada y de atributos de salida.

### **Selección de características**

El algoritmo Bayes naive de Microsoft realiza la selección automática de las características para limitar el número de valores que se consideran al generar el modelo.

El algoritmo está diseñado para reducir al mínimo el tiempo de proceso y seleccionar eficazmente los atributos que tienen la importancia máxima; sin embargo, puede controlar los datos que el algoritmo utiliza estableciendo los parámetros como se indica a continuación:

* Para limitar los valores que se utilizan como entradas, disminuya el valor de MAXIMUM\_INPUT\_ATTRIBUTES.
* Para limitar el número de atributos analizados por el modelo, disminuya el valor de MAXIMUM\_OUTPUT\_ATTRIBUTES.
* Para limitar el número de valores que pueden considerarse para cualquier un atributo, disminuya el valor de MINIMUM\_STATES.

## Personalizar el algoritmo Bayes naive

El algoritmo Bayes naive de Microsoft admite varios parámetros que influyen en el comportamiento, el rendimiento y la precisión del modelo de minería de datos resultante. También puede establecer marcas de modelado en las columnas de modelo para controlar cómo se procesan los datos, o establecer marcas en la estructura de minería de datos para especificar cómo se deberían administrar los valores nulos o que faltan.

### Establecer parámetros del algoritmo

El algoritmo Bayes naive de Microsoft admite varios parámetros que influyen en el rendimiento y la precisión del modelo de minería de datos resultante. Estos parámetros se describen en la tabla siguiente.

MAXIMUM\_INPUT\_ATTRIBUTES  
Especifica el número máximo de atributos de entrada que el algoritmo puede procesar antes de invocar la selección de características. La función de selección de atributos de entrada se deshabilita cuando este valor se establece en 0.

El valor predeterminado es 255.

MAXIMUM\_OUTPUT\_ATTRIBUTES  
Especifica el número máximo de atributos de salida que puede administrar el algoritmo antes de invocar la selección de características. La característica de selección de atributos de salida se deshabilita cuando este valor se establece en 0.

El valor predeterminado es 255.

MINIMUM\_DEPENDENCY\_PROBABILITY  
Especifica la probabilidad de dependencia mínima entre los atributos de entrada y salida. Este valor se utiliza para limitar el tamaño del contenido generado por el algoritmo. El valor de esta propiedad puede establecerse en un valor comprendido entre 0 y 1. Los valores mayores reducen el número de atributos en el contenido del modelo.

El valor predeterminado es 0,5.

MAXIMUM\_STATES  
Especifica el número máximo de estados de atributo que admite el algoritmo. Si el número de estados que tiene un atributo es mayor que el número máximo de estados, el algoritmo utiliza los estados más conocidos del atributo e interpreta que faltan los estados restantes.

El valor predeterminado es 100

## Requisitos

Un modelo de árbol de Bayes naive debe contener una columna de clave, al menos un atributo de predicción y al menos un atributo de entrada. Ningún atributo puede ser continuo; si los datos contienen datos numéricos continuos, se omitirán o se convertirán en discretos.

## Descripción de la estructura de un modelo Bayes naive

Un modelo Bayes naive tiene un nodo primario único que representa el modelo y sus metadatos, y debajo de dicho nodo, varios árboles independientes que representan los atributos de predicción seleccionados. Además de los árboles para los atributos, cada modelo contiene un nodo de estadísticas marginales (NODE\_TYPE = 26) que proporciona estadísticas descriptivas sobre el conjunto de casos de entrenamiento.

Para cada atributo de predicción y valor, el modelo genera un árbol que contiene información que describe cómo afectaron las columnas de entrada al resultado de ese atributo de predicción concreto. Cada árbol contiene el atributo de predicción y su valor (NODE\_TYPE = 9) y, a continuación, una serie de nodos que representan los atributos de entrada (NODE\_TYPE = 10). Dado que los atributos de entrada normalmente tienen varios valores, cada uno de dichos atributos (NODE\_TYPE = 10) puede tener varios nodos secundarios (NODE\_TYPE = 11), uno para cada estado específico del atributo.

**Contenido del modelo para un modelo Bayes naive**

En esta sección solo se proporcionan detalles y ejemplos de las columnas del contenido del modelo de minería de datos que tienen una relevancia especial para los modelos Bayes naive.

MODEL\_CATALOG  
Nombre de la base de datos en la que se almacena el modelo.

MODEL\_NAME  
Nombre del modelo.

ATTRIBUTE\_NAME  
Nombres de los atributos que corresponden a este nodo.

**Raíz del modelo** : nombre del atributo de predicción.

**Estadísticas marginales** : no aplicable.

**Atributo de predicción** : nombre del atributo de predicción.

**Atributo de entrada** : nombre del atributo de entrada.

**Estado de atributo de entrada** : solo el nombre del atributo de entrada. Para obtener el estado, use MSOLAP\_NODE\_SHORT\_CAPTION.

NODE\_NAME  
Nombre del nodo.

Esta columna contiene el mismo valor que NODE\_UNIQUE\_NAME.

NODE\_UNIQUE\_NAME  
Nombre único del nodo. Los nombres únicos se asignan según una convención que proporciona información sobre las relaciones entre los nodos.

NODE\_TYPE  
Un modelo Bayes naive genera los tipos de nodos siguientes:

| **Identificador del tipo de nodo** | **Description** |
| --- | --- |
| 26 (NaiveBayesMarginalStatNode) | Contiene estadísticas que describen el conjunto completo de casos de entrenamiento para el modelo. |
| 9 (atributo de predicción) | Contiene el nombre del atributo de predicción. |
| 10 (atributo de entrada) | Contiene el nombre de una columna de atributos de entrada, así como nodos secundarios que contienen los valores para el atributo. |
| 11 (estado de atributo de entrada) | Contiene los valores o los valores de datos discretos de todos los atributos de entrada que se emparejaron con un atributo de salida determinado. |

NODE\_CAPTION  
Etiqueta o título asociado al nodo. Esta propiedad se usa principalmente para la presentación.

**Raíz del modelo** : en blanco.

**Estadísticas marginales** : en blanco.

**Atributo de predicción** : nombre del atributo de predicción.

**Atributo de entrada** : nombre del atributo de predicción y del atributo de entrada actual. Ej.:

Bike Buyer -> Age

**Estado de atributo de entrada** : nombre del atributo de predicción y del atributo de entrada actual, más el valor de la entrada. Ej.:

Bike Buyer -> Age = Missing

CHILDREN\_CARDINALITY  
Número de elementos secundarios que tiene el nodo.

**Raíz del modelo** : recuento de los atributos de predicción del modelo, más 1 para el nodo de estadísticas marginales.

**Estadísticas marginales** : por definición, no tiene elementos secundarios.

**Atributo de predicción** : recuento de los atributos de entrada que estaban relacionados con el atributo de predicción actual.

**Atributo de entrada** : recuento de los valores discretos o discretizados para el atributo de entrada actual.

**Estado de atributo de entrada** : siempre es 0.

PARENT\_UNIQUE\_NAME  
Nombre único del nodo primario.

NODE\_DESCRIPTION  
Coincide con el título del nodo.

NODE\_RULE  
Representación XML del título del nodo.

MARGINAL\_RULE  
Coincide con la regla del nodo.

NODE\_PROBABILITY  
Probabilidad asociada a este nodo.

**Raíz del modelo** : siempre es 0.

**Estadísticas marginales** : siempre es 0.

**Atributo de predicción** : siempre es 1.

**Atributo de entrada** : siempre es 1.

**Estado de atributo de entrada** : número decimal que representa la probabilidad del valor actual. Los valores de todos los estados de los atributos de entrada bajo el nodo de atributo de entrada primario suman 1.

MARGINAL\_PROBABILITY  
Coincide con la probabilidad del nodo.

NODE\_DISTRIBUTION  
Tabla que contiene el histograma de probabilidad del nodo. Para obtener más información vea [Tabla NODE\_DISTRIBUTION](https://msdn.microsoft.com/es-es/library/cc645744.aspx#bkmk_nodedist).

NODE\_SUPPORT  
Número de casos que admiten este nodo.

**Raíz del modelo** : recuento de todos los casos de los datos de entrenamiento.

**Estadísticas marginales** : siempre es 0.

**Atributo de predicción** : recuento de todos los casos de los datos de entrenamiento.

**Atributo de entrada** : recuento de todos los casos de los datos de entrenamiento.

**Estado de atributo de entrada** : recuento de los casos de los datos de entrenamiento que solo contienen este valor concreto.

MSOLAP\_MODEL\_COLUMN  
Etiqueta que se utiliza para la visualización. Normalmente, coincide con ATTRIBUTE\_NAME.

MSOLAP\_NODE\_SCORE  
Representa la importancia del atributo o valor dentro del modelo.

**Raíz del modelo** : siempre es 0.

**Estadísticas marginales** : siempre es 0.

**Atributo de predicción** : siempre es 0.

**Atributo de entrada** : Puntuación interestingness para el atributo de entrada actual en relación con el atributo de predicción actual.

**Estado de atributo de entrada** : siempre es 0.

MSOLAP\_NODE\_SHORT\_CAPTION  
Cadena de texto que representa el nombre o el valor de una columna.

**Raíz del modelo** : en blanco.

**Estadísticas marginales** : en blanco.

**Atributo de predicción** : nombre del atributo de predicción.

**Atributo de entrada** : nombre del atributo de entrada.

**Estado de atributo de entrada** : valor o valor de datos discretos del atributo de entrada.

### Consulta de ejemplo 1: obtener metadatos del modelo usando DMX

Al consultar el conjunto de filas de esquema de minería de datos, puede buscar los metadatos del modelo. Esto podría incluir cuándo se creó, cuándo se procesó en último lugar, el nombre de la estructura de minería de datos en la que se basa el modelo y el nombre de las columnas que se usan como atributos de predicción. También se pueden devolver los parámetros que se utilizaron cuando se creó el modelo.

SELECT MODEL\_CATALOG, MODEL\_NAME, DATE\_CREATED, LAST\_PROCESSED,

SERVICE\_NAME, PREDICTION\_ENTITY, FILTER

FROM $system.DMSCHEMA\_MINING\_MODELS

WHERE MODEL\_NAME = 'TM\_NaiveBayes\_Filtered'

Resultados del ejemplo:

|  |  |
| --- | --- |
| MODEL\_CATALOG | AdventureWorks |
| MODEL\_NAME | TM\_NaiveBayes\_Filtered |
| DATE\_CREATED | 3/1/2008 19:15 |
| LAST\_PROCESSED | 3/2/2008 20:00 |
| SERVICE\_NAME | Microsoft\_Naive\_Bayes |
| PREDICTION\_ENTITY | Bike Buyer,Yearly Income |
| FILTER | [Region] = 'Europe' OR [Region] = 'North America' |

El modelo que se usa para este ejemplo está basado en el modelo Bayes naive que se crea en [Basic Data Mining Tutorial](https://msdn.microsoft.com/es-es/library/ms167167.aspx), pero se modificó agregando un segundo atributo de predicción y aplicando un filtro a los datos de entrenamiento.

### Consulta de ejemplo 3: buscar más información sobre atributos

Dado que un modelo Bayes naive a menudo contiene información compleja sobre las relaciones entre atributos diferentes, la manera más fácil de ver estas relaciones es utilizar el [Visor Bayes naive de Microsoft](https://msdn.microsoft.com/es-es/library/ms174534.aspx). Sin embargo, puede crear consultas DMX para devolver los datos.

En el ejemplo siguiente se muestra cómo devolver información del modelo sobre un atributo determinado, Region.

SELECT NODE\_TYPE, NODE\_CAPTION,

NODE\_PROBABILITY, NODE\_SUPPORT, MSOLAP\_NODE\_SCORE

FROM TM\_NaiveBayes.CONTENT

WHERE ATTRIBUTE\_NAME = 'Region'

Esta consulta devuelve dos tipos de nodos: el nodo que representa el atributo de entrada (NODE\_TYPE = 10) y nodos para cada valor del atributo (NODE\_TYPE = 11). El título del nodo se utiliza para identificarlo, en lugar del nombre, porque el título muestra tanto el nombre como el valor del atributo.

| **NODE\_TYPE** | **NODE\_CAPTION** | **NODE\_PROBABILITY** | **NODE\_SUPPORT** | **MSOLAP\_NODE\_SCORE** | **NODE\_TYPE** |
| --- | --- | --- | --- | --- | --- |
| 10 | Bike Buyer -> Region | 1 | 17484 | 84.51555875 | 10 |
| 11 | Bike Buyer -> Region = Missing | 0 | 0 | 0 | 11 |
| 11 | Bike Buyer -> Region = North America | 0.508236102 | 8886 | 0 | 11 |
| 11 | Bike Buyer -> Region = Pacific | 0.193891558 | 3390 | 0 | 11 |
| 11 | Bike Buyer -> Region = Europe | 0.29787234 | 5208 | 0 | 11 |

Algunas de las columnas almacenadas en los nodos son las mismas que se pueden obtener de los nodos de estadísticas marginales, como los valores de compatibilidad de los nodos y de puntuación de la probabilidad de los nodos. Sin embargo, MSOLAP\_NODE\_SCORE es un valor especial que solo se proporciona para los nodos de atributos de entrada e indica la importancia relativa de este atributo en el modelo. Puede ver casi toda esa misma información en el panel Red de dependencia del visor; sin embargo, el visor no proporciona puntuaciones.

**Conclusión**

* Naive Bayes es una técnica de clasificación y predicción que construye modelos que predicen la probabilidad de posibles resultados.

**Recomendación**

* Utilizar el algoritmo de Bayes Naves para calcular probabilidades, ya que este resulta ser menos complejo que otros, y por lo tanto resulta útil para generar rápidamente modelos de minería de datos.

Egrafía

https://msdn.microsoft.com/es-es/library/ms174806.aspx