Kualifikasi Data Mining

SG15-1

1. **Frequent Pattern Analysis**

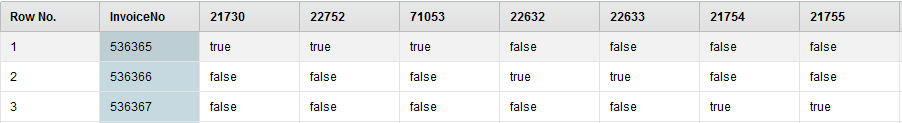
Bluejack Scientist Community sedang meneliti data mengenai **Online Retail** untuk menemukan pola pembelian barang dari data transaksi sudah terjadi sebelumnya, data transaksi di berikan dalam bentuk **Excel File**. Dimana isi dari **Excel File** tersebut adalah:

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Tipe** | **Deskripsi** |
| InvoiceNo | Polynominal | Kode identifikasi untuk suatu transaksi, terdiri dari 6 digit angka. Dengan ketentuan berikut:   * Setiap transaksi memiliki **InvoiceNo** yang unik * InvoiceNo yang di awali dengan huruf ‘**c**’ merupakan transaksi yang **dibatalkan** |
| StockCode | Polynominal | Kode identifikasi untuk suatu produk, terdiri dari 5 digit angka. Dengan ketentuan berikut:   * Setiap produk memiliki StockCode yang unik |
| Description | Polynominal | Deskripsi untuk suatu produk |
| Quantity | Integer | Jumlah produk yang dibeli |
| InvoiceDate | Date Time | Waktu kejadian transaksi |
| UnitPrice | Real | Harga satuan dari suatu produk |
| CustomerID | Integer | Kode identifikasi untuk pelanggan yang melakukan transaksi, terdiri dari 5 digit angka. Dengan ketentuan berikut:   * Setiap pelanggan memiliki **CustomerID** yang unik |
| Country | Polynominal | Negara tempat tinggal pelanggan |

**Table 1. Online Retail-1.xslx**

Sebagai data scientist, anda diminta untuk:

* + Meng-import data yang diberikan dalam **File Excel**: **Online Retail-1.xslx** ke dalam RapidMiner dan menyesuaikan tipe dari Attribute dengan data di **Table 1**
  + Melakukan **preprocessing** untuk data dari **Online Retail-1.xslx** menggunakan RapidMiner, dengan ketentuan sebagai berikut:
    1. Hapus setiap data transaksi yang **dibatalkan**
    2. Hapus setiap data transaksi yang **Quantity** nya bernilai **negative**
    3. Hapus setiap data transaksi yang **jumlah karakter** dari **StockCode** (kode produk)nya tidak sama dengan 5
    4. Pilah attribute yang akan digunakan:
       - InvoiceNo
       - StockCode
       - Quantity
    5. Ubah bentuk data agar siap digunakan, dengan ketentuan:
       - Putar setiap distinct value dari **StockCode** (kode produk) menjadi attribute yang dikelompokkan berdasarkan **InvoiceNo**
       - Ubah nama attribute dari hasil putaran menjadi masing-masing **StockCode** (kode produk), kecuali **InvoiceNo**
       - Ubah peran dari attribute **InvoiceNo** menjadi **identifier** untuk setiap example
       - Ubah setiap value dari masing-masing attribute **StockCode** (kode produk)menjadi:
         1. **True** untuk setiap **StockCode** (kode produk)yang muncul dalam suatu transaksi
         2. **False** untuk setiap **StockCode** (kode produk)yang tidak muncul dalam suatu transaksi

****

* + Mencari produk yang sering muncul dalam setiap transaksi menggunakan **FP-Growth** dengan **minimum support**: 0,03
  + Buat **association rules** berdasarkan data produk yang sering muncul, dengan **minimum confidence**: 0,5
  + Tampilkan **association rules** dan **frequent item sets** ke dalam result

**Pertanyaan**:

1. Dari hasil proses di atas, berapakah jumlah **association rules** yang memenuhi **minimum confidence**: 0,5? Berikan juga list **association rule** nya beserta **confidence** nya dalam bentuk table!
2. **Clustering and Classification**

Bluejack Scientist Community sedang meneliti data tentang **Online Retail** untuk melakukan segmentasi terhadap pelanggan berdasarkan rata-rata jumlah harga barang dan jumlah barang yang dibeli dari data transaksi yang sudah diberikan dalam bentuk **Excel File**. Lalu data pelanggan yang sudah memiliki **label** dari hasil segmentasi tersebut akan digunakan untuk mengklasifikasikan tipe dari pelanggan yang belum memiliki **label**. Dimana isi dari **Excel File** tersebut adalah:

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Tipe** | **Deskripsi** |
| InvoiceNo | Polynominal | Kode identifikasi untuk suatu transaksi, terdiri dari 6 digit angka. Dengan ketentuan berikut:   * Setiap transaksi memiliki **InvoiceNo** yang unik * InvoiceNo yang di awali dengan huruf ‘**c**’ merupakan transaksi yang **dibatalkan** |
| StockCode | Polynominal | Kode identifikasi untuk suatu produk, terdiri dari 5 digit angka. Dengan ketentuan berikut:   * Setiap produk memiliki StockCode yang unik |
| Description | Polynominal | Deskripsi untuk suatu produk |
| Quantity | Integer | Jumlah produk yang dibeli |
| InvoiceDate | Date Time | Waktu kejadian transaksi |
| UnitPrice | Real | Harga satuan dari suatu produk |
| CustomerID | Integer | Kode identifikasi untuk pelanggan yang melakukan transaksi, terdiri dari 5 digit angka. Dengan ketentuan berikut:   * Setiap pelanggan memiliki **CustomerID** yang unik |
| Country | Polynominal | Negara tempat tinggal pelanggan |

**Table 2. Online Retail-2.xslx**

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Tipe** | **Deskripsi** |
| CustomerID | Integer | Kode identifikasi untuk pelanggan yang melakukan transaksi, terdiri dari 5 digit angka. Dengan ketentuan berikut:   * Setiap pelanggan memiliki **CustomerID** yang unik |
| AverageOfQuantity | Real | Rata-rata dari jumlah barang yang pernah dibeli oleh pelanggan dari semua transaksi yang dilakukan |
| AverageOfTotalPrice | Real | Rata-rata dari jumlah harga barang yang pernah dibeli oleh pelanggan dari semua transaksi yang dilakukan |

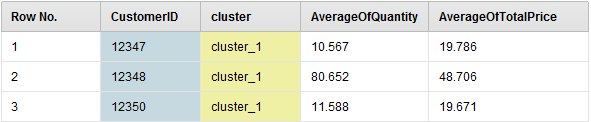
**Table 3. Online Retail-TestSet.xslx**

Sebagai data scientist, anda diminta untuk:

* + **Fase Clustering**
    1. Meng-import data yang diberikan dalam **File Excel**: **Online Retail-2.xslx** ke dalam RapidMiner dan menyesuaikan tipe dari Attribute dengan data di **Table 2**
    2. Melakukan **preprocessing** untuk data dari **Online Retail-2.xslx** menggunakan RapidMiner, dengan ketentuan sebagai berikut:
       - Hapus setiap data transaksi yang **dibatalkan**
       - Hapus setiap data transaksi yang tidak memiliki **CustomerID** (missing)
       - Tambahkan attribute dengan nama **TotalPrice**, dimana isi dari attribute itu adalah

|  |
| --- |
| **TotalPrice = Quantity \* UnitPrice** |

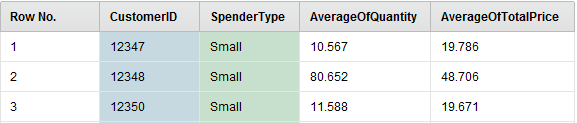
* + - * Pilah attribute yang akan digunakan:
        1. CustomerID
        2. Quantity
        3. TotalPrice
      * Gunakan fungsi aggregate untuk mendapatkan nilai **rata-rata** dari attribute **Quantity** dan **TotalPrice**
      * Ubah nama attribute, dengan ketentuan:
        1. Untuk attribute yang berisikan nilai **rata-rata** dari **Quantity**, ubah menjadi **AverageOfQuantity**
        2. Untuk attribute yang berisikan nilai **rata-rata** dari **TotalPrice**, ubah menjadi **AverageOfTotalPrice**
      * Ubah peran dari attribute **CustomerID** menjadi **identifier** untuk setiap example
    1. Melakukan **clustering** untuk membagi data pelanggan ke dalam **2 cluster** menggunakan **k-Means**, dengan **maximum run**: 1000 dan **local random seed**: 1992



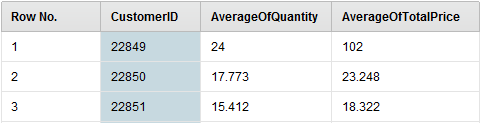
* + **Fase Classification**
    1. **Fase Training**
       - Memproses data yang sudah di cluster pada **Fase Clustering**, dengan ketentuan sebagai berikut:
         1. Ubah nama attribute yang merupakan hasil **clustering** menjadi **SpenderType**
         2. Ubah peran dari attribute **SpenderType** menjadi **label**
         3. Ubah setiap value dari attribute **SpenderType**, dengan ketentuan:

**Big** untuk pelanggan yang berada pada **cluster** yang memiliki **rata-rata** **AverageOfTotalPrice** dan **AverageOfQuantity** terbesar

**Small** untuk pelanggan yang berada pada **cluster** yang memiliki **rata-rata** **AverageOfTotalPrice** dan **AverageOfQuantity** terkecil



* + - * Membuat model untuk klasifikasi dari data yang sudah di proses, menggunakan **Naïve Bayes** dengan **laplace correction**: true
    1. **Fase Testing**
       - Meng-import data yang diberikan dalam **File Excel**: **OnlineRetail-TestSet.xslx** ke dalam RapidMiner dan menyesuaikan tipe dari Attribute dengan data di **Table 3**
       - Melakukan **preprocessing** untuk data dari **Online Retail-TestSet.xslx** menggunakan RapidMiner, dengan ketentuan sebagai berikut:
         1. Ubah peran dari attribute **CustomerID** menjadi **identifier** untuk setiap example



* + - * Melakukan klasifikasi untuk data dari **Online Retail-TestSet.xslx** yang sudah di **proprocessing** menggunakan **model** yang dibuat pada **Fase Training** untuk memprediksi **SpenderType** dari masing-masing data pelanggan
      * Tampilkan **data** **OnlineRetail-TestSet.xslx** yang sudah diberi label dari hasil klasifikasi, **model** yang digunakan untuk **klasifikasi**, dan **model** dari hasil **clustering** ke dalam result

**Pertanyaan**:

1. Berapakah nilai **mean** dari **AverageOfQuantity** dan **AverageOfTotalPrice** dari masing-masing **cluster**? Tentukan **cluster** mana yang memiliki **mean** terbesar dan terkecil!

Notes:

1. Untuk setiap **operator** yang memiliki **random seed**, gunakan **local random seed**: 1992
2. File yang harus dikumpulkan:
   * Semua file process (\*.rmp, \*.properties)
   * Semua file data (\*.ioo, \*.md, \*.properties)
   * File Doc untuk jawaban pertanyaan (\*.doc / \*.docx)

- Selamat mengerjakan ☺ -