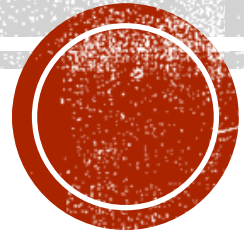
KNOWLEDGE DISCOVERY M14 – PREDICTIVE MINING

Nama : Ni Putu Devira Ayu Martini

Kelas : S2 Teknik Elektro 2020

NRP : 1120800012



DATASET "TRANSACTION"

InvoiceNo	StockCode	Qty	InvoiceDate	CustomerID	Country
537626	22725	830	12/7/10 14:57	12347	Iceland
537626	22729	948	12/7/10 14:57	12347	Iceland
537626	22195	695	12/7/10 14:57	12347	Iceland
542237	22725	636	1/26/11 14:30	12347	Iceland
542237	22729	536	1/26/11 14:30	12347	Iceland
542237	47559	919	1/26/11 14:30	12347	Iceland
542237	21154	803	1/26/11 14:30	12347	Iceland
542237	21035	532	1/26/11 14:30	12347	Iceland
…	…		…	…	…

SOAL!

1. dataset ← transaction.csv, and show it
2. data ← take the data in the dataset for feature of Qty, Country ("Germany"), month, year ("2011")
3. TotalQty ← take Month from the data and accumulated Qty in the same month, and show it
4. Visualize the movement of TotalQty values where the x axis = Month and the y axis = TotalQty
5. PredictedQty ← predict the total Qty of TotalQty in January 2012 with Linear Regression
6. Calculate the MAE, MSE and MAPE for within last 9 months

PERALATAN

1. Personal Computer
2. Dataset “Transaction”
3. Software Anaconda Jupyter (Bahasa Pemrograman Python)

LANGKAH PROGRAM

MENGUBAH FILE DATASET

InvoiceNo,StockCode,Qty,InvoiceDate,CustomerID,Country
537626,22725,830,12 7 2010 14:57,12347,Iceland
537626,22729,948,12 7 2010 14:57,12347,Iceland
537626,22195,695,12 7 2010 14:57,12347,Iceland
542237,22725,636,1 26 2011 14:30,12347,Iceland
542237,22729,536,1 26 2011 14:30,12347,Iceland
542237,47559,919,1 26 2011 14:30,12347,Iceland
542237,21154,803,1 26 2011 14:30,12347,Iceland
542237,21035,532,1 26 2011 14:30,12347,Iceland
549222,23076,383,4 7 2011 10:43,12347,Iceland
549222,21791,389,4 7 2011 10:43,12347,Iceland
549222,22550,500,4 7 2011 10:43,12347,Iceland
549222,22432,875,4 7 2011 10:43,12347,Iceland
549222,22195,434,4 7 2011 10:43,12347,Iceland
549222,21975,735,4 7 2011 10:43,12347,Iceland
556201,23171,135,6 9 2011 13:01,12347,Iceland
556201,23172,974,6 9 2011 13:01,12347,Iceland
556201,23175,82,6 9 2011 13:01,12347,Iceland
556201,51014,234,6 9 2011 13:01,12347,Iceland

Mengubah file dataset .csv pada kolom "Invoice Date" yang semula "bulan/tanggal/tahun" menjadi "bulan tanggal tahun" (tanda baca hilang diganti oleh spasi) agar memudahkan dalam memprogram/parsing date.

MENGINPUTKAN DATASET

```
import csv

contacts = []

with open('C:/Users/user/Downloads/transaction.csv') as csv_file:
    csv_reader = csv.reader(csv_file, delimiter=",")
    for row in csv_reader:
        contacts.append(row)

labels = contacts.pop(0)

print(f'{labels[0]} \t {labels[1]} \t\t {labels[2]} \t {labels[3]} \t\t {labels[4]} \t {labels[5]}')
print("-"*34)
for data in contacts:
    print(f'{data[0]} \t {data[1]} \t {data[2]} \t {data[3]} \t {data[4]} \t {data[5]}')
```

Memesan tempat array untuk variable "contacts"

Membaca dataset .csv pada directory

Menyimpan variable row (dataset) kedalam array "contacts"

Pada array "contacts" ke-0 dinamai dengan var "labels"

Print label dan data

Hasil :

```
InvoiceNo       StockCode               Qty     InvoiceDate             CustomerID      Country
----------------------------------
537626   22725   830     12 7 2010 14:57         12347   Iceland
537626   22729   948     12 7 2010 14:57         12347   Iceland
537626   22195   695     12 7 2010 14:57         12347   Iceland
```

MENGAMBIL DATA TESTING: 2011 & GERMANY (1)

```
import nltk
import re
from nltk.tokenize import word_tokenize
from nltk.tokenize import WordPunctTokenizer
import string

for data in contacts:
    data3=f'{data[3]}'
    data5=f'{data[5]}’
    data3=data3.translate(str.maketrans("","",string.punctuation))
    data3=word_tokenize(data3)
    data5=data5.translate(str.maketrans("","",string.punctuation))
    data5=word_tokenize(data5)
    #print(data3[2],data5[0])
    if (data5[0]=="Germany"):
        if(data3[2]=="2011"):
            print(f'{data[0]} \t {data[1]} \t {data[2]} \t {data[3]} \t {data[4]} \t {data[5]}')
```

Meng-Tokenize data3

Meng-Tokenize data5

Mengambil data5[0] yaitu Country

Mengambil data3[2] yaitu tahun

MENGAMBIL DATA TESTING: 2011 & GERMANY (2)

Hasil :

554985	21746	628	5 29 2011 12:26	12426	Germany
554985	21770	981	5 29 2011 12:26	12426	Germany
554985	22329	212	5 29 2011 12:26	12426	Germany
554985	22976	910	5 29 2011 12:26	12426	Germany
554985	22845	668	5 29 2011 12:26	12426	Germany
554985	16161	855	5 29 2011 12:26	12426	Germany
570452	22972	485	10 10 2011 15:15	12427	Germany
570452	23389	980	10 10 2011 15:15	12427	Germany
570452	22973	623	10 10 2011 15:15	12427	Germany
570452	22144	808	10 10 2011 15:15	12427	Germany
570452	22976	653	10 10 2011 15:15	12427	Germany
577135	22634	893	11 18 2011 8:56	12427	Germany
542371	22957	515	1 27 2011 13:29	12468	Germany
542371	21880	128	1 27 2011 13:29	12468	Germany
542371	21883	864	1 27 2011 13:29	12468	Germany
542371	22716	354	1 27 2011 13:29	12468	Germany
542371	21700	111	1 27 2011 13:29	12468	Germany
555523	22431	375	6 5 2011 11:36	12468	Germany
555523	22432	786	6 5 2011 11:36	12468	Germany

Terlihat bahwa Data disamping sudah difilter menjadi Country = Germany & Invoice Date (Year) = 2011

MENGAKUMULASI "BULAN" YANG SAMA (1)

```
import nltk
import re
from nltk.tokenize import word_tokenize
from nltk.tokenize import WordPunctTokenizer
import string

bulan1=1
bulan2=2
bulan3=3
bulan4=4
bulan5=5
bulan6=6
bulan7=7
bulan8=8
bulan9=9
bulan10=10
bulan11=11
bulan12=12
```

```
jumbln1=0
jumbln2=0
jumbln3=0
jumbln4=0
jumbln5=0
jumbln6=0
jumbln7=0
jumbln8=0
jumbln9=0
jumbln10=0
jumbln11=0
jumbln12=0
```

MENGAKUMULASI "BULAN" YANG SAMA (2)

```
for data in contacts:
    data3=f'{data[3]}'
    data5=f'{data[5]}'
    #WordPunctTokenizer().tokenize(data3)
    data3=data3.translate(str.maketrans("","",string.punctuation))
    data3=word_tokenize(data3)
    data5=data5.translate(str.maketrans("","",string.punctuation))
    data5=word_tokenize(data5)
    if (data5[0]=="Germany"):
        if(data3[2]=="2011"):
            if(data3[0]=="1"):
                jumbln1=jumbln1+1
            elif(data3[0]=="2"):
                jumbln2=jumbln2+1
            elif(data3[0]=="3"):
                jumbln3=jumbln3+1
            elif(data3[0]=="4"):
                jumbln4=jumbln4+1
            elif(data3[0]=="5"):
                jumbln5=jumbln5+1
            elif(data3[0]=="6"):
                jumbln6=jumbln6+1
            elif(data3[0]=="7"):
                jumbln7=jumbln7+1
            elif(data3[0]=="8"):
                jumbln8=jumbln8+1
            elif(data3[0]=="9"):
                jumbln9=jumbln9+1
            elif(data3[0]=="10"):
                jumbln10=jumbln10+1
            elif(data3[0]=="11"):
                jumbln11=jumbln11+1
            elif(data3[0]=="12"):
                jumbln12=jumbln12+1
```

MENGAKUMULASI "BULAN" YANG SAMA (3)

```
print("Jumlah bulan-1=",jumbln1)
print("Jumlah bulan-2=",jumbln2)
print("Jumlah bulan-3=",jumbln3)
print("Jumlah bulan-4=",jumbln4)
print("Jumlah bulan-5=",jumbln5)
print("Jumlah bulan-6=",jumbln6)
print("Jumlah bulan-7=",jumbln7)
print("Jumlah bulan-8=",jumbln8)
print("Jumlah bulan-9=",jumbln9)
print("Jumlah bulan-10=",jumbln10)
print("Jumlah bulan-11=",jumbln11)
print("Jumlah bulan-12=",jumbln12)
```

Hasil :

```
Jumlah bulan-1= 201
Jumlah bulan-2= 73
Jumlah bulan-3= 144
Jumlah bulan-4= 120
Jumlah bulan-5= 195
Jumlah bulan-6= 154
Jumlah bulan-7= 192
Jumlah bulan-8= 225
Jumlah bulan-9= 204
Jumlah bulan-10= 300
Jumlah bulan-11= 271
Jumlah bulan-12= 69
```

MEMPLOT GRAFIK AKUMULASI "BULAN" YANG SAMA

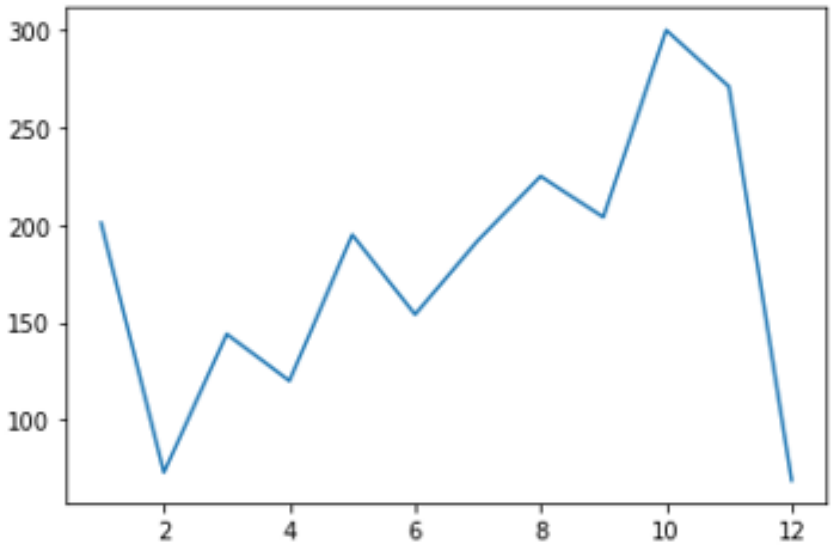
```
import matplotlib.pyplot as plt

x=[1,2,3,4,5,6,7,8,9,10,11,12]
y=[jumbln1,jumbln2,jumbln3,jumbln4,jumbln5,jumbln6,jumbln7,jumbln8,
jumbln9,jumbln10,jumbln11,jumbln12]
plt.plot(x,y)
plt.show()
```

Var array x menyatakan bulan

Var array y menyatakan akumulasi customer pada bulan 1-12

Memplot grafik x,y

Hasil :



MEMPREDIKSI JUMLAH CUSTOMER PADA JANUARY 2012

```
#LINEAR REGRESSION
jumx=0
xx=0
n=12
for i in range(1,13):
    jumx=jumx+i
    xx=xx+i**2

jumy=jumbln1+jumbln2+jumbln3+jumbln4+jumbln5+jumbln6+jumbln7+jumbln8+jumbln9+jumbln10+ju
mbln11+jumbln12

xy=1*jumbln1+2*jumbln2+3*jumbln3+4*jumbln4+5*jumbln5+6*jumbln6+7*jumbln7+8*jumbln8+9*jumb
ln9+10*jumbln10+11*jumbln11+12*jumbln12

a=(jumy*xx-jumx*xy)/(n*xx-jumx**2)
b=(n*xy-jumx*jumy)/(n*xx-jumx**2)
yy=a+b*13
print("Prediksi bulan ke Januari 2012 =",yy)
```

Mengakumulasi jumlah x yaitu bulan (1-12)

Mengakumulasi jumlah kuadrat dari x

Mengakumulasi jumlah y yaitu banyaknya customer perbulan

Menghitung x*y

$$a = \frac{(\Sigma Y)(\Sigma X^2) - (\Sigma X)(\Sigma XY)}{(n)(\Sigma X^2) - (\Sigma X)^2}$$

$$b = \frac{(n)(\Sigma XY) - (\Sigma X)(\Sigma Y)}{(n)(\Sigma X^2) - (\Sigma X)^2}$$

$$Y = a + b * X$$

MEMPREDIKSI JUMLAH CUSTOMER PADA JANUARY 2012

Hasil :

```
Prediksi bulan ke Januari 2012 = 223.77272727272725
```

MENGEVALUASI PREDIKSI

```
#EVALUATION

yyarray=[]
MAE=0
MSE=0
MAPE=0

for i in range(3,12):
    yy=a+b*i #Prediction
    MAE=MAE+abs(y[i]-yy)
    MSE=MSE+(y[i]-yy)**2
    MAPE=MAPE+abs((y[i]-yy)/y[i])
    yyarray.append(yy)
#print(yyarray)
MAE=MAE/9
MSE=MSE/9
MAPE=MAPE*100/9
print("MAE =",MAE)
print("MSE =",MSE)
print("MAPE =",MAPE)
```

Var yy menyatakan Prediction Value pada 9 bulan terakhir yaitu bulan ke-4 hingga ke-12

$$\frac{\sum_{t=1}^{N}\left|d_t - d_t'\right|}{N}$$

$$\frac{\sum_{t=1}^{N}\left(d_t - d_t'\right)^2}{N}$$

$$\frac{100}{N}\sum_{t=1}^{N}\left[\left|\frac{d_t - d_t'}{d_t}\right|\right]$$

Hasil :

```
MAE = 52.123931623931625
MSE = 4454.692455757141
MAPE = 39.36596439405867
```

ANALISA

- Pada praktikum kali ini membahas tentang Predictive Mining dengan menggunakan teknik Linear Regression yaitu teknik/fungsi bagaimana sebuah proses nantinya akan menemukan pola tertentu dari suatu data. Pola-pola tersebut dapat diketahui dari berbagai variabel yang ada pada data.
- Rumus yang digunakan dalam teknik Linear Regression adalah:

$$Y = a + b * X$$

$$a = \frac{(\Sigma Y)(\Sigma X^2) - (\Sigma X)(\Sigma XY)}{(n)(\Sigma X^2) - (\Sigma X)^2}$$

$$b = \frac{(n)(\Sigma XY) - (\Sigma X)(\Sigma Y)}{(n)(\Sigma X^2) - (\Sigma X)^2}$$

- Pada percobaan kali ini dicari prediksinya pada bulan Januari 2012 atau pola ke-13, didapat prediksi pada bulan tersebut adalah 223.77 yang mana hasilnya tidak keluar jauh/outlayer/masih dalam jangkauan bulan-bulan sebelumnya yaitu 69 hingga 300 customer.
- Kemudian untuk mengevaluasi prediksi digunakan perhitungan error MAE, MSE & MAPE, didapatkan nilai error MAE = 52.12, MSE = 4454.69 & MAPE = 39.36. Nilai yang tidak terlalu besar. Perlu diketahui bahwa teknik Linear Regression ini adalah teknik terbaik dalam memprediksi data, karena menggunakan pola dan menentukan prediksi berdasarkan korelasi antara variable-variable sebagaimana rumus diatas.