Sesi 10 – Data Preprocessing

SI602 – Praktikum Kecerdasan Bisnis

Akhmad Rezki Purnajaya, SKom., Mkom.

Bahan

- Teknik Sampling Data
  - Combination of Over-undersampling (COUS)
- Performa Prediksi
  - Akurasi, Sensitifitas, Spesifisitas, AUC

Teknik Sampling Data

Inisiasi Data

- **Step 1:** Aktifkan Package Library Teknik Sampling Data (pada praktikum ini kita menggunakan teknik **Combination of Over-undersampling (COUS)**:
  - install.packages('ROSE')
  - library(ROSE)

- **Step 2:** Aktifkan Package Library Teknik Klasifikasi (pada praktikum ini kita menggunakan teknik **Support Vector Machine (SVM))**
  - install.packages('e1071')
  - library("e1071")

- **Step 3:** Bangkitkan 3000 data input secara random dengan jumlah variabel = 3 dan jumlah baris = 1000:
  - x<-data.frame(matrix(rnorm(3000), ncol=3))
  - summary(x[1:900,])
  - summary(x[901:1000,])

Inisiasi Data

- **Step 3:** Bangkitkan 3000 data input secara random dengan jumlah variabel = 3 dan jumlah baris = 1000:
  - x<-data.frame(matrix(rnorm(3000), ncol=3))
  - summary(x[1:900,])
  - summary(x[901:1000,])

```
       X1                  X2                  X3
 Min.   :-3.04697   Min.   :-3.82519   Min.   :-2.940898
 1st Qu.:-0.71440   1st Qu.:-0.69597   1st Qu.:-0.616591
 Median :-0.03963   Median : 0.01742   Median :-0.019201
 Mean   :-0.05009   Mean   :-0.00169   Mean   :-0.008764
 3rd Qu.: 0.63094   3rd Qu.: 0.63903   3rd Qu.: 0.648480
 Max.   : 3.30433   Max.   : 3.85176   Max.   : 2.757687
       X1                  X2                  X3
 Min.   :-2.27922   Min.   :-2.24487   Min.   :-2.0467
 1st Qu.:-0.87502   1st Qu.:-0.74405   1st Qu.:-0.3309
 Median :-0.04830   Median : 0.04271   Median : 0.2027
 Mean   :-0.08152   Mean   :-0.01854   Mean   : 0.1561
 3rd Qu.: 0.57434   3rd Qu.: 0.65532   3rd Qu.: 0.7754
 Max.   : 1.78311   Max.   : 1.78018   Max.   : 2.4523
```

Inisiasi Data

- **Step 4:** Modifikasi data dengan baris 1-900 dikurang 0,5, baris 901-1000 ditambah 0,5:
  - x[1:900,1:ncol(x)]<-x[1:900,1:ncol(x)]-0.5
  - x[901:1000,1:ncol(x)]<-x[901:1000,1:ncol(x)]+0.5
  - summary(x[1:900,])
  - summary(x[901:1000,])

```
       X1                  X2                  X3
 Min.   :-3.5470   Min.   :-4.3252   Min.   :-3.4409
 1st Qu.:-1.2144   1st Qu.:-1.1960   1st Qu.:-1.1166
 Median :-0.5396   Median :-0.4826   Median :-0.5192
 Mean   :-0.5501   Mean   :-0.5017   Mean   :-0.5088
 3rd Qu.: 0.1309   3rd Qu.: 0.1390   3rd Qu.: 0.1485
 Max.   : 2.8043   Max.   : 3.3518   Max.   : 2.2577
       X1                  X2                  X3
 Min.   :-1.7792   Min.   :-1.7449   Min.   :-1.5467
 1st Qu.:-0.3750   1st Qu.:-0.2440   1st Qu.: 0.1691
 Median : 0.4517   Median : 0.5427   Median : 0.7027
 Mean   : 0.4185   Mean   : 0.4815   Mean   : 0.6561
 3rd Qu.: 1.0745   3rd Qu.: 1.1555   3rd Qu.: 1.2754
 Max.   : 2.2831   Max.   : 2.2802   Max.   : 2.9523
```

Inisiasi Data

- **Step 5:** Bangkitkan kelas untuk data 1-900 menjadi kelas 0 dan data 901-1000 menjadi kelas 1, sehingga kelas 0 sebanyak = 900, kelas 1 sebanyak = 100:
  - kelasImbalanced <- c(rep(0, 900),rep(1,100))
  - kelasImbalanced <- factor(kelasImbalanced)
  - table(kelasImbalanced)

```
kelasImbalanced
  0   1
900 100
```

Inisiasi Data

- **Step 6:** Gabungkan data 3 variabel dengan kelasnya:
  - dataImbalanced <- cbind(x,kelasImbalanced)
  - head(dataImbalanced)
  - tail(dataImbalanced)

A data.frame: 6 × 4

	X1	X2	X3	kelasImbalanced
	<dbl>	<dbl>	<dbl>	<fct>
1	-1.07394550	-1.5696518	-1.74781057	0
2	-0.27945117	-2.4021770	-1.48481315	0
3	-1.33451573	0.5538667	-0.05877343	0
4	-0.63035416	-0.6397304	-1.12228934	0
5	0.06732944	-0.8805980	-2.80111173	0
6	0.70846682	0.9761975	0.03537822	0

A data.frame: 6 × 4

	X1	X2	X3	kelasImbalanced
	<dbl>	<dbl>	<dbl>	<fct>
995	-0.1038329	1.1218578	0.1979753	1
996	-0.4220310	-0.5131228	0.6017659	1
997	0.9799107	0.3368176	-1.0608056	1
998	0.6058816	1.4666371	-0.3775891	1
999	-0.6403780	0.8474853	-0.8254433	1
1000	-0.4714858	0.6663778	0.3770795	1

Sampling Data

- **Step 7:** Sampling Data Combination of Over-undersampling (COUS):
  - data.balanced <- ovun.sample(kelasImbalanced~., data=dataImbalanced, N=nrow(dataImbalanced), p=0.5, seed=1, method="both")$data
  - names(data.balanced)[4]<-"kelasBalanced"
  - summary(data.balanced)

```
       X1                  X2                  X3             kelasBalanced
 Min.   :-3.5470   Min.   :-4.32519   Min.   :-3.440898   0:520
 1st Qu.:-0.8005   1st Qu.:-0.82326   1st Qu.:-0.748869   1:480
 Median :-0.1411   Median :-0.10788   Median :-0.052356
 Mean   :-0.1036   Mean   :-0.07959   Mean   :-0.003791
 3rd Qu.: 0.6588   3rd Qu.: 0.72951   3rd Qu.: 0.785039
 Max.   : 2.2831   Max.   : 2.28018   Max.   : 2.952305
```

Klasifikasikan dengan Model SVM

- **Step 8:** Buat data uji dengan mengambil 500 baris secara acak pada data imbalanced yang bertotal 1000 baris data:
  - set.seed(123)
  - dataUji <- dataImbalanced[sample(nrow(dataImbalanced), 500),]
  - head(dataUji)
  - summary(dataUji)

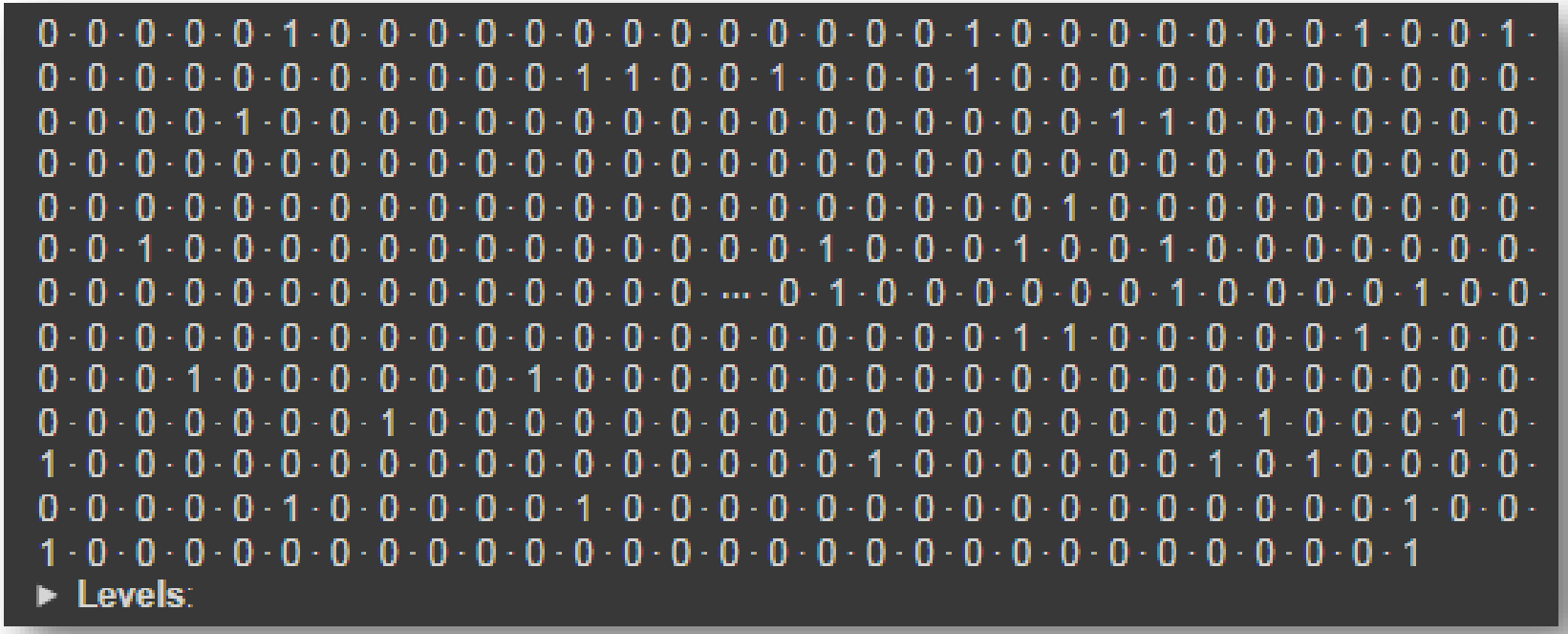
A data.frame: 6 × 4

	X1	X2	X3	kelasImbalanced
	<dbl>	<dbl>	<dbl>	<fct>
415	1.7572350	-0.7255436	-0.5050946	0
463	-1.4217221	1.1931888	-0.8238449	0
179	-0.3612473	1.1510123	0.2551887	0
526	-0.9849069	0.8820218	-0.5162069	0
195	-1.4422082	-1.5722952	0.4326354	0
938	2.2166491	0.6190363	0.4373757	1

```
       X1                 X2                 X3          kelasImbalanced
 Min.   :-3.5470   Min.   :-3.6455   Min.   :-3.4409   0:452
 1st Qu.:-1.1731   1st Qu.:-1.0644   1st Qu.:-1.0548   1: 48
 Median :-0.4472   Median :-0.3684   Median :-0.4803
 Mean   :-0.4592   Mean   :-0.4083   Mean   :-0.4179
 3rd Qu.: 0.2466   3rd Qu.: 0.2909   3rd Qu.: 0.3042
 Max.   : 2.8043   Max.   : 2.5176   Max.   : 2.9523
```

Klasifikasikan dengan Model SVM

- **Step 9:** Simpan kelas data uji untuk nanti dikomparasi dengan hasil klasifikasi SVM:
  - aktual <- dataUji[,4]
  - aktual



Klasifikasikan dengan Model SVM

- **Step 10:** Buat model klasifikasi dengan data imbalanced:
  - model.svm.imbalanced <- svm(kelasImbalanced~., data=dataImbalanced)
  - model.svm.imbalanced

```
Call:
svm(formula = kelasImbalanced ~ ., data = dataImbalanced)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Number of Support Vectors:  212
```

Klasifikasikan dengan Model SVM

- **Step 11:** Buat model klasifikasi dengan data balanced:
  - model.svm.balanced <-svm(kelasBalanced~., data=data.balanced)
  - model.svm.balanced

```
Call:
svm(formula = kelasBalanced ~ ., data = data.balanced)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Number of Support Vectors:  378
```

EVALUASI PERFORMA PREDIKSI

(Sonego *et al.* 2008)

Confusion Matrix

	AKTUAL: 0	AKTUAL: 1
PREDIKSI: 0	*True Positive* (TP)	*False Positive* (FP)
PREDIKSI: 1	*False Negative* (FN)	*True Negative* (TN)

EVALUASI PERFORMA PREDIKSI

(Sonego *et al*. 2008)

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensitivitas = \frac{TP}{TP + FN}$$

$$Spesifisitas = \frac{TN}{TN + FP}$$

$$AUC = \frac{Sensitivitas + Spesifisitas}{2}$$

Evaluasi Performa Prediksi

- **Step 12:** Prediksi model SVM data imbalanced dengan data uji serta hitung nilai akurasi dan AUC:
  - ypred <- predict(model.svm.imbalanced, dataUji)
  - (matcon <- table(prediksi=ypred,aktual))
  - (akurasiImbalanced <- sum(diag(matcon))/sum(matcon)*100)
  - (spesifisitasImbalanced <- matcon[1,1]/(matcon[1,1]+matcon[2,1]))
  - (sensitifitasImbalanced <- matcon[2,2]/(matcon[2,2]+matcon[1,2]))
  - (aucImbalanced <- (spesifisitasImbalanced+sensitifitasImbalanced)/2)

```
         aktual
prediksi   0   1
       0 440  31
       1  12  17
91.4
0.973451327433628
0.354166666666667
0.663808997050148
```

Evaluasi Performa Prediksi

- **Step 13:** Prediksi model SVM data balanced dengan data uji serta hitung nilai akurasi dan AUC:
  - ypred <- predict(model.svm.balanced, dataUji)
  - (matcon <- table(prediksi=ypred, aktual))
  - (akurasiBalanced <- sum(diag(matcon))/sum(matcon)*100)
  - (spesifisitasBalanced <- matcon[1,1]/(matcon[1,1]+matcon[2,1]))
  - (sensitifitasBalanced <- matcon[2,2]/(matcon[2,2]+matcon[1,2]))
  - (aucBalanced <- (spesifisitasBalanced+sensitifitasBalanced)/2)

```
         aktual
prediksi   0   1
       0 390   8
       1  62  40
86
0.86283185840708
0.833333333333333
0.848082595870207
```

Evaluasi Performa Prediksi

- **Step 14:** Bandingkan nilai akurasi dan AUC dengan Tabel Performa Prediksi :
  - data <- c('Imbalanced', 'Balanced')
  - akurasi <- c(akurasiImbalanced, akurasiBalanced)
  - spesifisitas <- c(spesifisitasImbalanced, spesifisitasBalanced)
  - sensitifitas <- c(sensitifitasImbalanced, sensitifitasBalanced)
  - auc <- c(aucImbalanced, aucBalanced)
  - tabelPerforma <- data.frame(data, akurasi, spesifisitas, sensitifitas, auc)
  - tabelPerforma

A data.frame: 2 × 5

data	akurasi	spesifisitas	sensitifitas	auc
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
Imbalanced	91.4	0.9734513	0.3541667	0.6638090
Balanced	86.0	0.8628319	0.8333333	0.8480826