PROGRAM STUDI
TEKNIK INFORMATIKA – S1
FAKULTAS ILMU KOMPUTER
UNIVERSITAS DIAN NUSWANTORO

MATA KULIAH
DATA MINING

Technology vector created by sentavio - www.freepik.com

DATA MINING

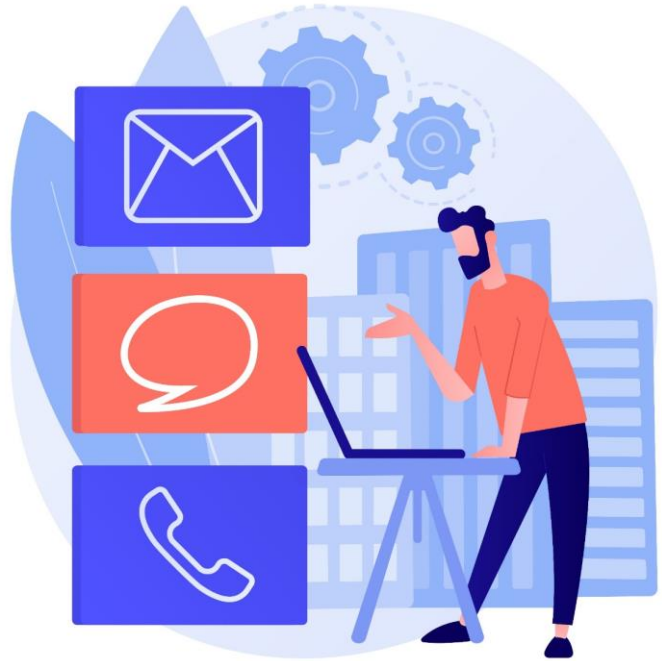
"Validasi dan Pengujian Model"

TIM PENGAMPU DOSEN DATA MINING

2023

Kontak Dosen

- Junta Zeniarja, M.Kom
- Email: junta@dsn.dinus.ac.id
- Youtube : https://www.youtube.com/JuntaZeniarja
- Scholar : http://bit.do/JuntaScholar



Pendahuluan [1]

- Evaluasi atau Validasi merupakan kunci dalam membuat aplikasi Data Mining.
- Ada berbagai macam cara dalam melakukan evaluasi.
- Jika kita memiliki data yang kita gunakan dalam proses pelatihan, maka tidak serta merta kita menjadikan data tersebut sebagai indikator keberhasilan aplikasi yang kita buat.
- Oleh karena itu, kita membutuhkan metode tertentu guna memprediksi performa berdasarkan eksperimen untuk berbagai macam data selain data training tersebut.
- Kualitas data merupakan faktor penting di dalam evaluasi baik untuk proses pelatihan maupun pengujian data.

Pendahuluan [2]

- Membandingkan performa antar beragam metode pada *Machine Learning* merupakan permasalahan yang tidak mudah.
- Performa suatu metode diukur dari sejauh mana kemampuan metode tersebut dalam melakukan klasifikasi secara akurat terhadap data tertentu yang diuji.
- Kebanyakan aplikasi Data Mining memerlukan biaya yang besar ketika melakukan pengujian karena melibatkan data yang berukuran besar.
- Hingga saat ini, tidak bisa dijumpai evaluasi yang pasti mengenai Data Mining, apalagi jika dipandang dari sisi filosofis.
- Satu – satunya teori yang benar adalah teori berdasarkan data (*theory of data*).

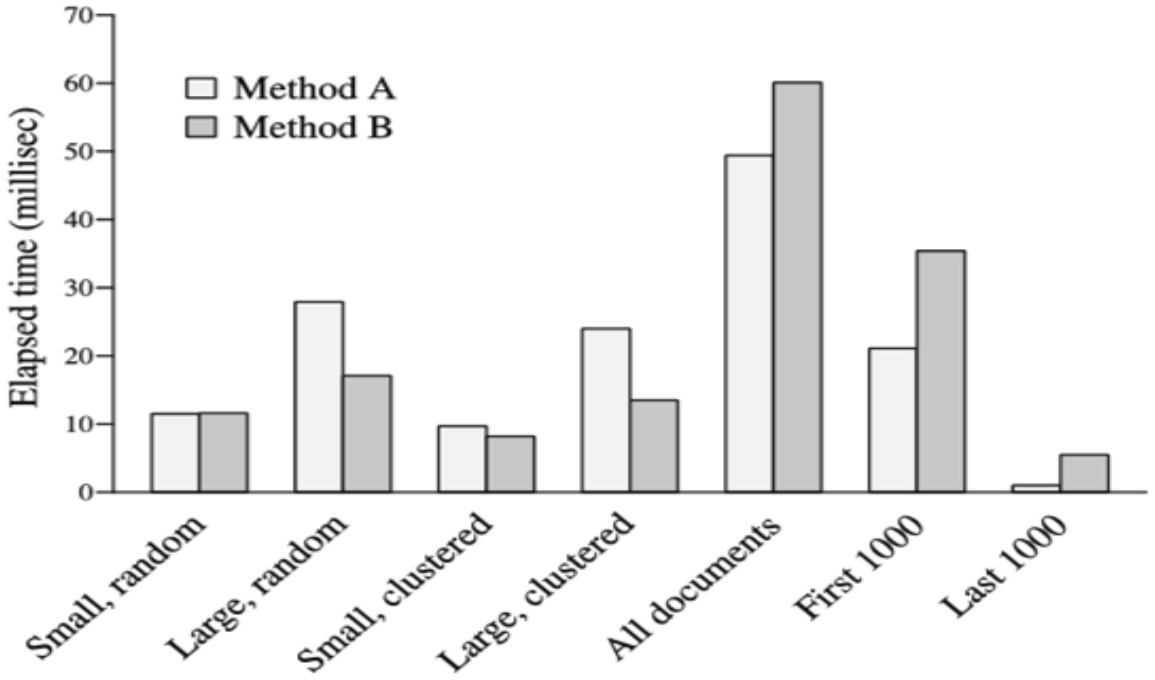
Performance

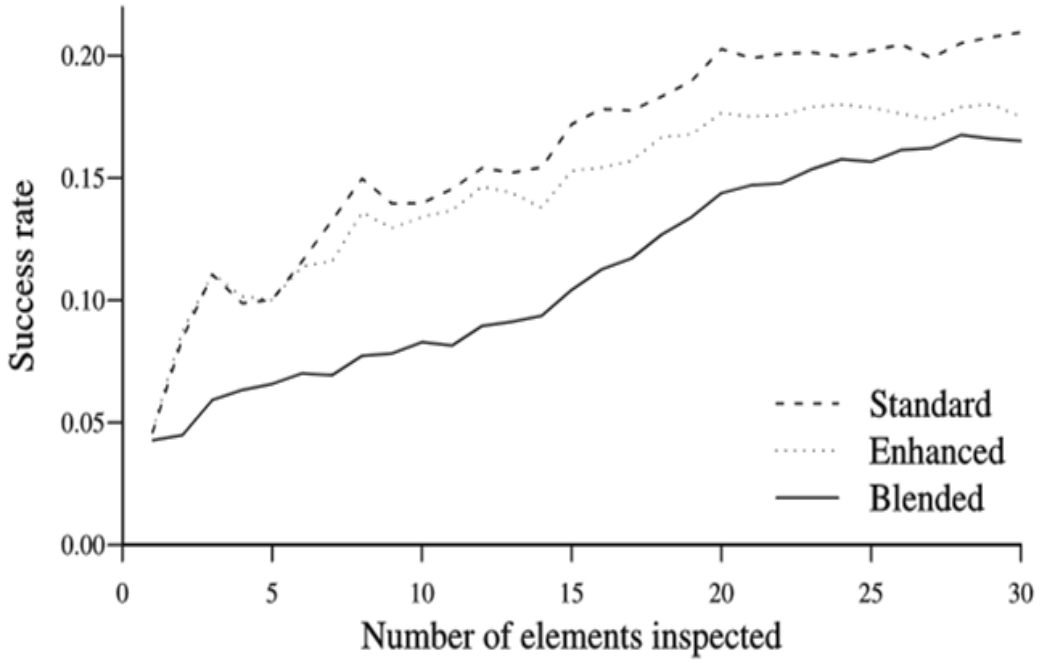
- Statistical performances are conducted for two or more methods.
- The methods are performed on many datasets or many parameter value.
- Performance measures:
  - Accuracy,
  - Computational time,
  - Complexity.
- Table, Bar chart, line graph.

Dataset	Method A	Method B
Breast cancer	78%	98%
Lung cancer	86%	80%
Iris	71%	68%
Zoo	55%	60%
mushroom	60%	55%
Average	70%	72.2%

The average acc of method B is higher.
Method B outperforms method A only on two datasets.
Is there a significant improvement?

Ploting the Results





Confusion Matrix

Actual

Predicted	Positive	Negative
Positive	TP	FP
Negative	FN	TN

TP = true positive

TN = true negative

FP = false positive

FN= false negative

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{F-measure} = \frac{2 \times P \times R}{P+R}$$

Highest value means that the algorithm achieves the best performance

Example

A dataset consists of 1000 fist and 400 non-fish. Naive bayes is used to classified them automatically and the results show that only 700 fish is correctly classified and 250 non-fish is correctly classified. How many % the accuracy of the experiment?

Confusion Matrix

Predicted \ Actual	Fish	Non-fish
Fish	700	150
Non-fish	300	250

$$Accuracy = \frac{700 + 250}{1400} = 68\%$$

$$Recall = \frac{700}{1000} = 70\%$$

$$Precision = \frac{700}{850} = 82\%$$

Memvalidasi Model

- Pada setiap topik di data mining seperti klasifikasi, klastering, asosiasi akan menghasilkan model.
- Model tersebut dapat berupa persamaan, diagram jaringan syarat tiruan, atau diagram pohon keputusan.
- Pertanyaannya adalah: seberapa anda yakin model tersebut siap digunakan untuk produksi dan apakah model tersebut memberikan hasil prediksi yang akurat terhadap data di masa depan?

Teknik Validasi

- *Split validation*: melakukan validasi sederhana dengan membagi dataset secara acak menjadi dua data terpisah — data latih & data uji.
- *Cross validation*: melakukan validasi berulang di mana dataset dibagi menjadi banyak subset (himpunan) data latih & validasi. Setiap iterasi memvalidasi (menguji) satu subset data dengan subset yang tersisa sebagai data latih. Pada c*ross validation, #* subset data adalah jumlah iterasi.

Cross validation

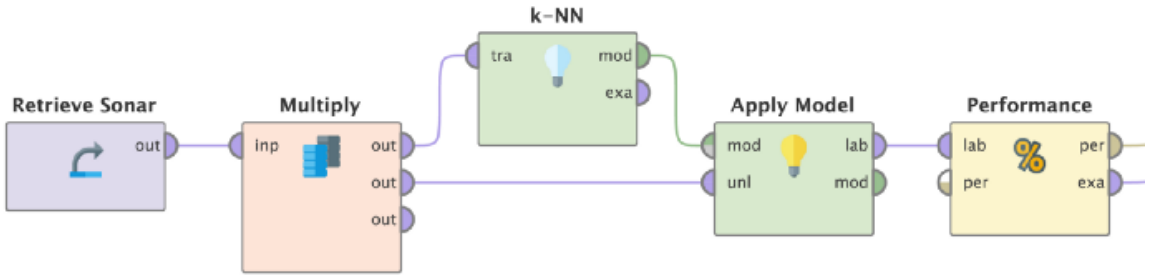
Dataset is divided into training and testing automatically

Using 5-fold, dataset is divided into 5 parts

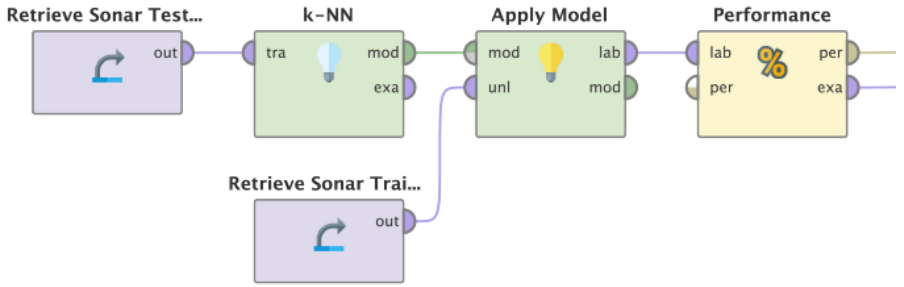
In each iteration, 4 parts become the training dataset and 1 part becomes the testing dataset

Split Validation

- *Training error* didapatkan dengan menghitung kesalahan klasifikasi model pada data yang sama dengan model yang dilatih.
- *Test error* didapatkan dengan menggunakan dua data yang sepenuhnya terpisah. Satu untuk melatih model (data latih) dan lainnya untuk menghitung kesalahan klasifikasi (data uji). Kedua dataset harus memiliki nilai label yang sama.



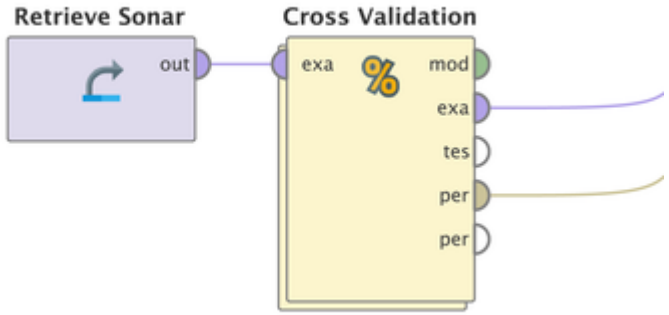
Training error

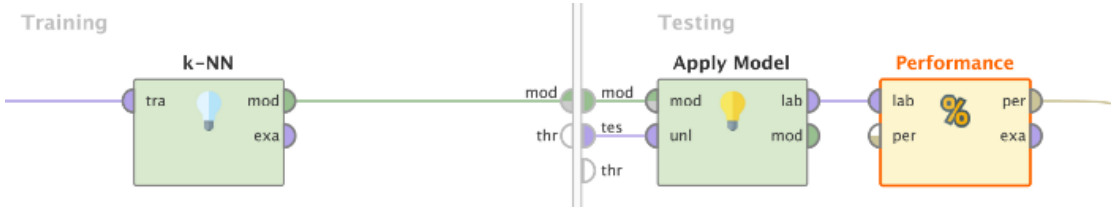


Test error

Cross Validation

- Bisa jadi sampel acak yang dipilih pada split validation tidak begitu acak, terutama jika hanya memiliki dataset yang sedikit. Dalam kasus tersebut, *test error* yang dihasilkan mungkin kurang mewakili akurasi model.
- Dari permasalahan tersebut, ide yang muncul adalah mengulangi *sampling* data uji beberapa kali dan menggunakan sampel yang berbeda untuk setiap kali pengujian. Cara ini disebut cross validation.

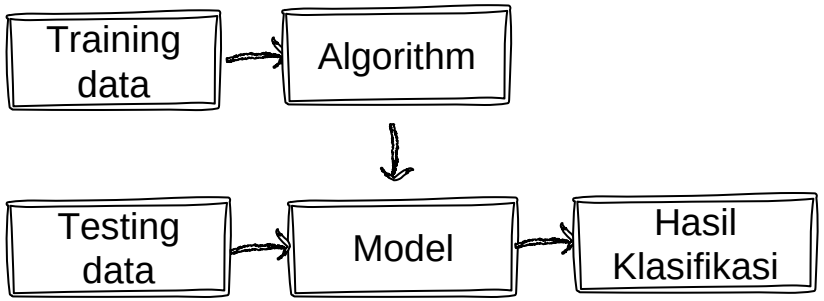




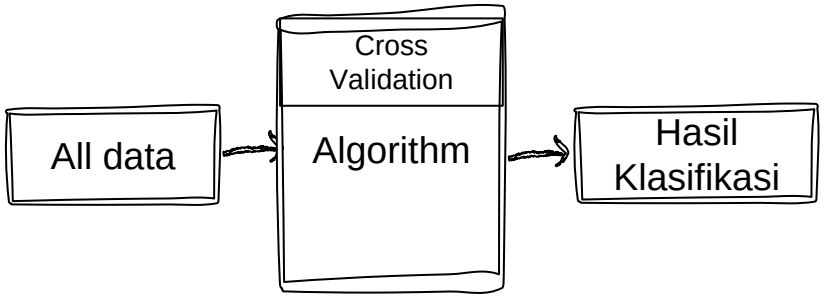
Subproses pada cross validation

Split Validation vs Cross Validation

- Split Validation



- With Cross Validation



Parametric vs Nonparametric

- Parametric
  - Based on assumption that data is normal distribution
  - T-test, ANOVA
- Non parametric
  - For small sample
  - Wilcoxon, Friedman

Implementasi Python

(Validasi dan Pengujian Model)

Pertemuan 14

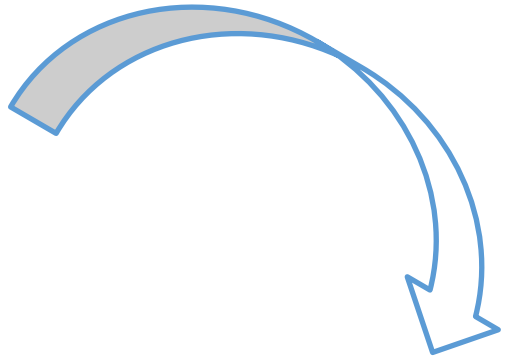
Studi Kasus Cross Validation dengan Python [Logistic Regression]

```python
# evaluate a logistic regression model using k-fold cross-validation
from numpy import mean
from numpy import std
from sklearn.datasets import make_classification
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
from sklearn.linear_model import LogisticRegression
# create dataset
X, y = make_classification(n_samples=1000, n_features=20, n_informative=15, n_redundant=5, random_state=1)
# prepare the cross-validation procedure
cv = KFold(n_splits=10, random_state=1, shuffle=True)
# create model
model = LogisticRegression()
# evaluate model
scores = cross_val_score(model, X, y, scoring='accuracy', cv=cv, n_jobs=-1)
# report performance
print('Accuracy: %.3f (%.3f)' % (mean(scores), std(scores)))
```

```
Accuracy: 0.868 (0.032)
```

Studi Kasus untuk Hitung Akurasi pada Python

```
# impor pustaka
import numpy as np
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
# data prediksi dan target
target=[1,1,1,1,1,1,1,0,0,0,0,0,0]
prediksi=[1,1,1,1,1,0,1,0,0,0,0,0,0]
# mengetahui akurasi
print(classification_report(target,prediksi))
akurasi=accuracy_score(target,prediksi)
print("Akurasi : ",round(akurasi,2))
```



```
              precision    recall  f1-score   support

           0       0.86      1.00      0.92         6
           1       1.00      0.86      0.92         7

    accuracy                           0.92        13
   macro avg       0.93      0.93      0.92        13
weighted avg       0.93      0.92      0.92        13

Akurasi :  0.92
```

Latihan Soal (Kuis)

- Carilah 3 paper tentang perkembangan aplikasi Data Mining dengan berbagai teknik Validasi dan Pengujian Model minimal 5 tahun terakhir (terbit antara thn 2016 – 2021), kemudian review paper tersebut, selanjutnya tuliskan kedalam paper A4 1 halaman penuh.

Referensi

1. Janez Demsar, Statistical Comparisons of Classifiers over Multiple Data Sets, Journal of Machine Learning Research, 2006.
2. Rahmadya Trias Handayanto, Herlawati, Data Mining dan Machine Learning menggunakan Matlab dan Python, Penerbit Informatika, 2020.
3. https://docs.biolab.si//3/data-mining-library/reference/evaluation.cd.html
4. https://medium.com/@ksnugroho/validasi-model-machine-learning-pada-rapidminer-50be0080df14
5. https://machinelearningmastery.com/repeated-k-fold-cross-validation-with-python/
6. Sumber gambar: www.freepik.com.

THANKS

ANY QUESTIONS?