UNIVERSITAS INDONESIA

FAKULTAS ILMU KOMPUTER

Evaluasi Model & *Performance Metric*

Siti Aminah*, Dinial Utami

CSGE603130: Kecerdasan Artifisial dan Sains Data Dasar

Genap 2022/2023

UNIVERSITAS INDONESIA

FAKULTAS ILMU KOMPUTER

Outline

1. Motivasi
2. Metodologi Evaluasi Model
3. Performance Metric

UNIVERSITAS INDONESIA

FAKULTAS ILMU KOMPUTER

Motivasi

Sumber:

- Slides Materi KASDD, “Evaluasi Model dan Performance Metric”, Semester Gasal 2022/2023
- Charu C. Aggarwal, “Data Mining: The Textbook”, Springer, 2015

Motivasi

Beberapa pertanyaan yang mungkin muncul terkait evaluasi model:

- Sebaiknya berapa besar data yang digunakan untuk membangun dan mengevaluasi model?
- Sebuah model prediksi sudah dibangun, bagaimana kita menilai kinerja model tersebut?
- Di antara beberapa model prediksi yang ada, bagaimana kita memilih model yang memiliki kinerja terbaik?
- Apa saja metrik yang dapat digunakan untuk menilai kinerja sebuah model prediksi?

Motivasi

Berdasarkan pertanyaan-pertanyaan tersebut, ada 2 isu atau tantangan dalam proses evaluasi:

1. Methodology Issue:
   - Bagaimana membagi dataset untuk training dan testing model dengan tepat
   - Pemilihan pendekatan yang tepat: holdout, bootstrap, cross-validation
2. Quantification Issue:
   - Memilih performance metrics yang tepat untuk evaluasi model berdasarkan tujuan

UNIVERSITAS INDONESIA

FAKULTAS ILMU KOMPUTER

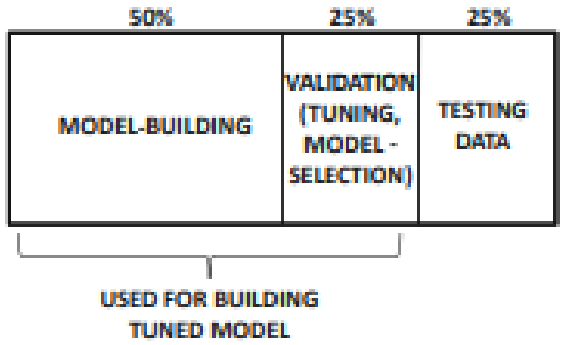
Metodologi Evaluasi Model

Sumber:

- Slides Materi KASDD, “Evaluasi Model dan Performance Metric”, Semester Gasal 2022/2023
- Charu C. Aggarwal, “Data Mining: The Textbook”, Springer, 2015
- Stuart Russel & Peter Norvig, “Artificial Intelligence: A Modern Approach”, 4th edition, Pearson, 2020
- Slides Materi Sains Data, “Learning to Classify”, Semester Genap 2020/2021

Training & Testing Models

- Ketika membangun dan mengevaluasi model prediksi, baik classification maupun regression, kita tetap membutuhkan data dengan pasangan variabel input (atribut) dan output (label kelas/nilai/*ground truth*).



Referensi lain mengatakan 80:20

Training: proses membangun model (*learning*)

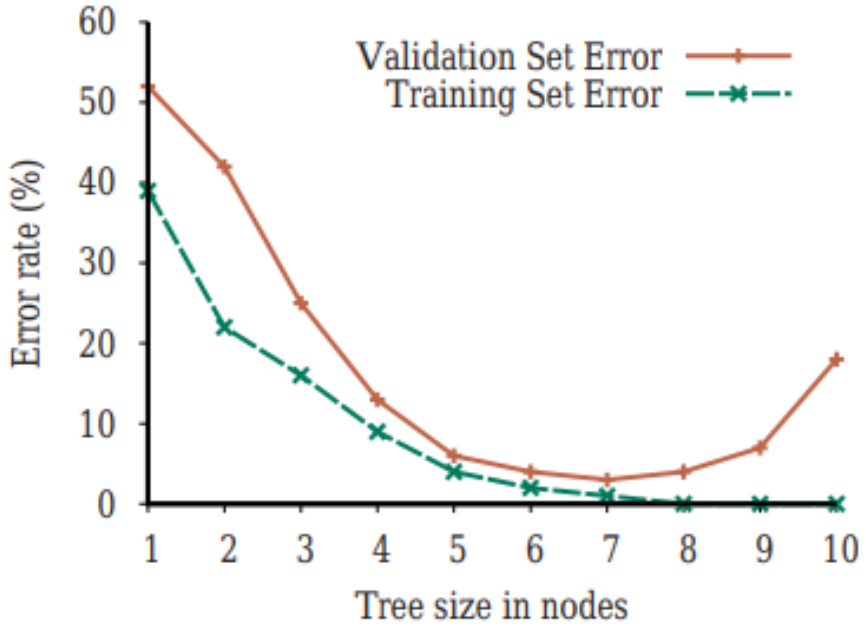
Validation: proses tambahan untuk melakukan tuning pada model, model selection

Testing: menguji kinerja model, bandingkan hasil prediksi dengan *ground truth*

Training & Testing Models

Validation

Final tuning pada model sebelum testing, misalkan menentukan hyperparameters atau untuk evaluasi kandidat-kandidat model dan pilih yang terbaik berdasarkan performance pada validation set untuk dilanjutkan ke testing.



Training & Testing Models

- No-nos ketika testing:
  - Menggunakan training set (sebagian atau semua) ketika melakukan testing pada model
  - Menggunakan testing set untuk parameter tuning atau hasil test dijadikan acuan untuk mengubah desain model
- Why? Akan *overestimate* kinerja model
  - Bayangkan jika soal ujian sama persis dengan soal-soal di kuis, PR, latihan di kelas, dsb ;)
  - estimasi error yang bias, tidak mencerminkan kinerja sebenarnya.

Model yang baik adalah model yang dapat memprediksi dengan tepat data yang tidak pernah dilihatnya!

Metode: Holdout

Data dibagi menjadi 2 set: (1) training set (2) testing set

Problem:

- Bila proporsi kelas tidak seimbang ketika membagi data, misal. ada kelas mayoritas di training set akan menjadi minoritas di testing set
- Bila data dari awal tidak mempunyai proporsi kelas yang seimbang (imbalanced classification)
- Menyebabkan estimasi error yang pessimistic

Solusi:

- Imbalanced classification: undersampling/oversampling, akan dibahas pada bab imbalanced classification :)

Metode: Cross-Validation

- estimasi kinerja model terhadap generalisasi (unbiased)
- dilakukan secara berulang
- membagi data (split) ke dalam training dan validation sets secara seragam dan acak, lalu merata-ratakan (averaging) error yang diperoleh dari seluruh bagian (splits).

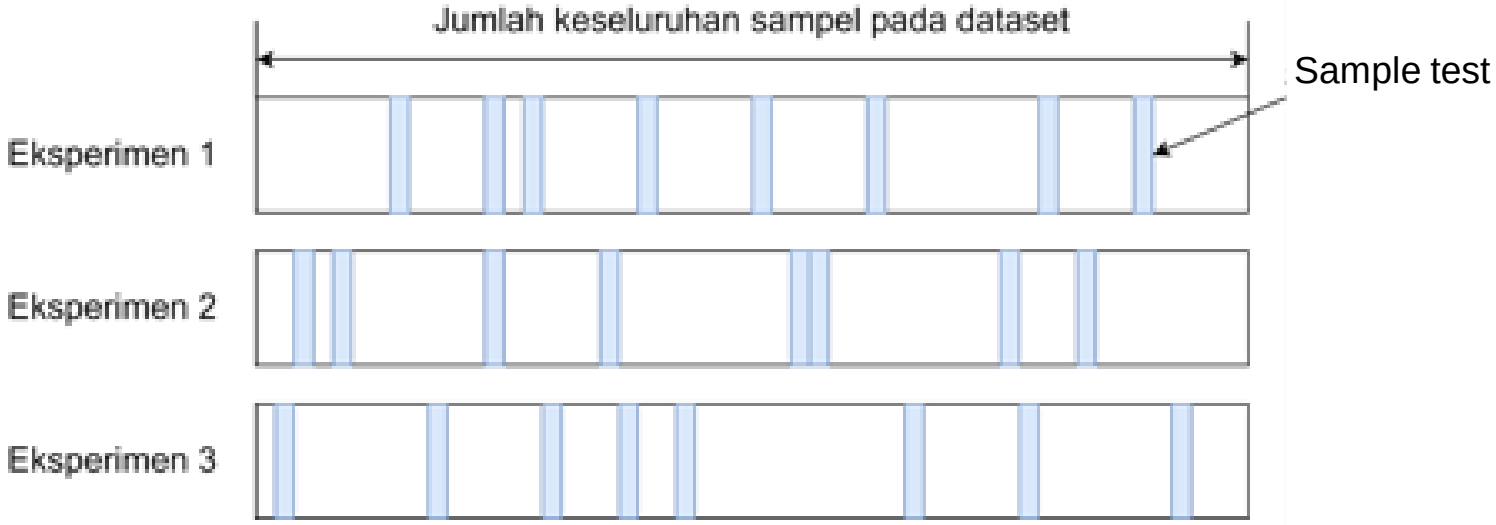
Teknik-teknik Cross-Validation

- Random Subsampling
- K-Fold Cross Validation
- Leave-one Out Cross Validation

Metode: Cross-Validation

Random Subsampling or Monte Carlo Cross-Validation

- buat K eksperimen (training-validation) terhadap seluruh dataset
- masing-masing eksperimen memilih secara acak sampel validation set (fixed size) tanpa pergantian/pengembalian
- Pada setiap eksperimen, latih ulang model dengan training set dan hitung estimasi error (E_i) pada validation set yang sesuai



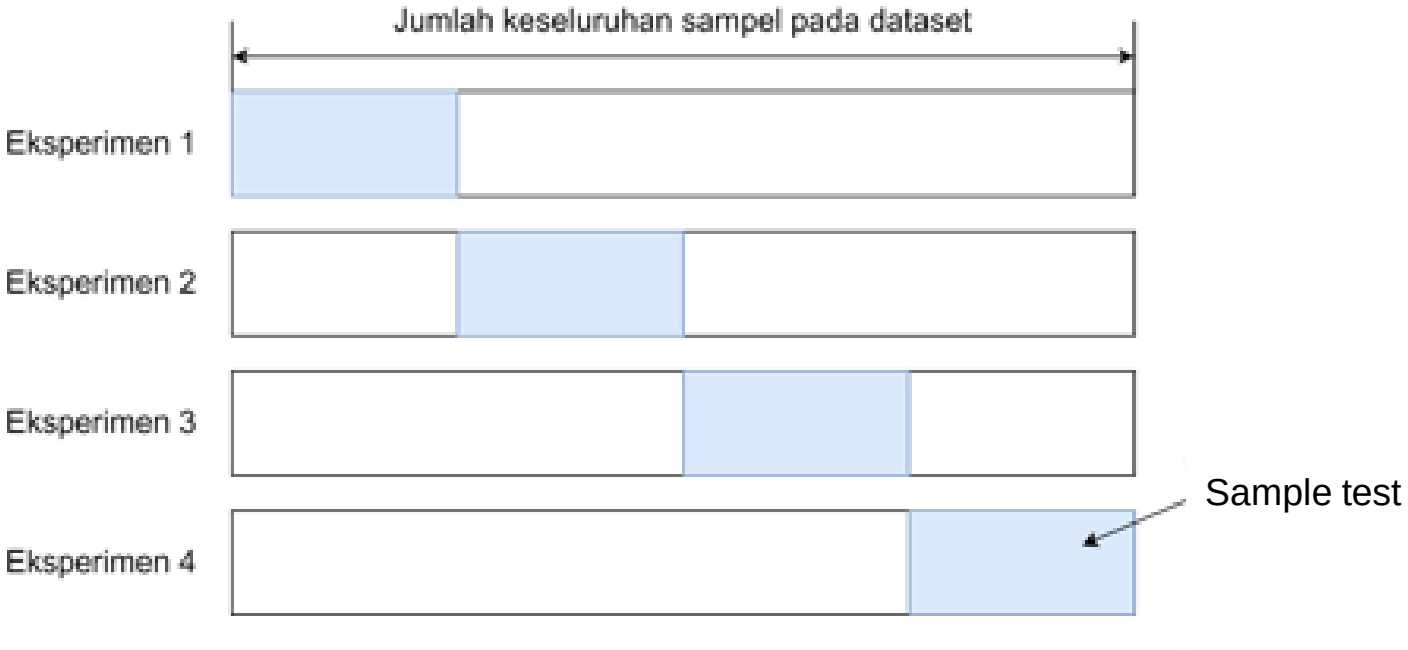
Estimasi error didapatkan dari rata-rata validation error pada setiap eksperimen

$$E = \frac{1}{K}\sum_{i=1}^{K} E_i$$

Metode: Cross Validation

K-Fold Cross Validation

- Buat partisi (fold) sebanyak K pada dataset.
- Untuk setiap eksperimen K, gunakan K-1 folds untuk training dan 1 fold sisanya untuk validation.



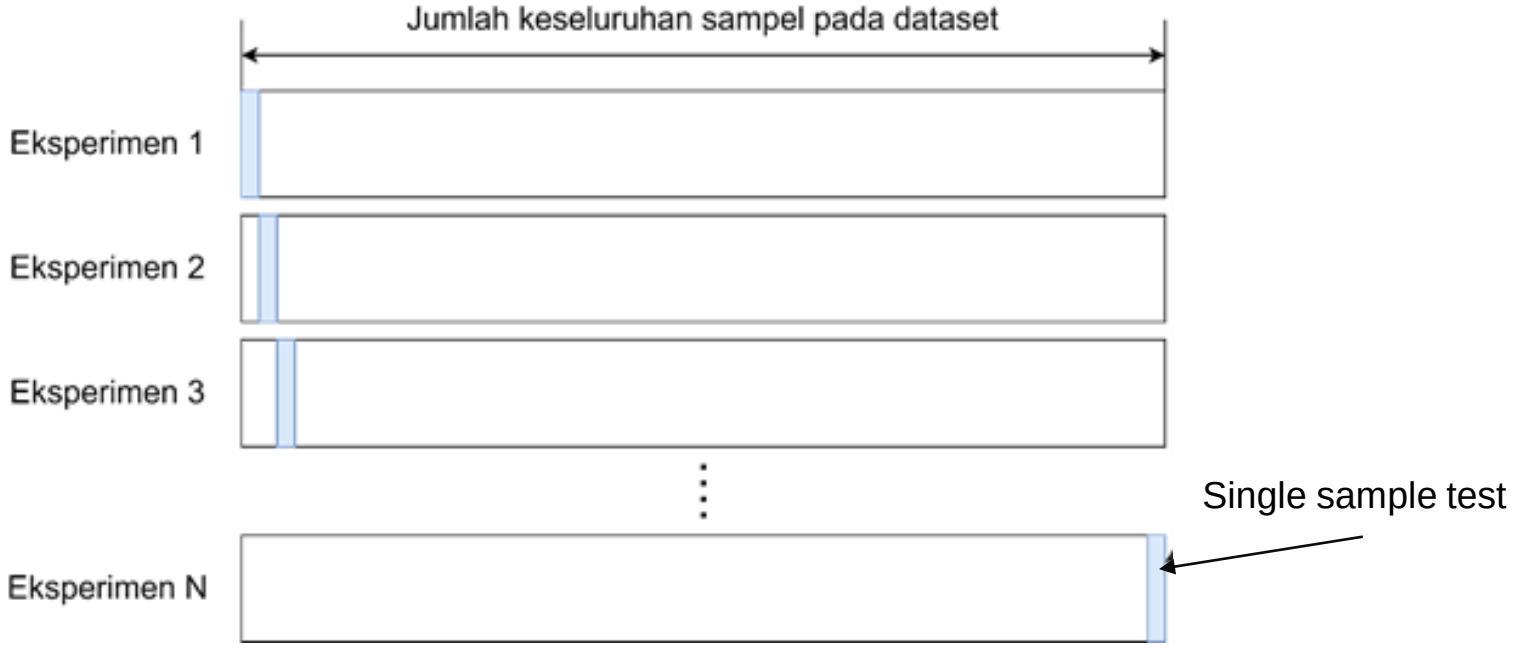
Estimasi error didapatkan dari rata-rata validation error pada setiap eksperimen

$$E = \frac{1}{K}\sum_{i=1}^{K} E_i$$

Metode: Cross-Validation

Leave-one-out Cross Validation (LOOCV)

- k-fold CV ketika K = N (ukuran dataset)
- lakukan eksperimen sebanyak N
- Pada masing-masing eksperimen, lakukan training pada N-1 sampel, dan sisanya untuk validation.



Estimasi error didapatkan dari rata-rata validation error pada setiap eksperimen

$$E = \frac{1}{N}\sum_{i=1}^{N} E_i$$

Metode: Cross-Validation

Jika K besar

- Bias dari estimasi error akan semakin kecil
- Variance dari estimasi error besar
- Waktu komputasi juga akan menjadi besar (karena eksperimen yang banyak)

Jika K kecil

- Waktu komputasi berkurang
- Variance dari estimasi error akan menjadi kecil
- Bias dari estimasi error menjadi semakin besar (lebih kecil atau sederhana dibandingkan dengan true error rate)

IRL, pemilihan jumlah folds bergantung pada ukuran dataset yang digunakan.

- Untuk dataset yang besar, 3-Fold CV dapat memberikan hasil yang cukup akurat.
- Untuk dataset yang kecil atau sparse, dapat menggunakan leave-one-out sehingga memungkinkan training dengan sampel yang banyak.

Umumnya, ambil K=10.

UNIVERSITAS INDONESIA

FAKULTAS ILMU KOMPUTER

Performance Metrics

Sumber:

- Slides Materi KASDD, “Evaluasi Model dan Performance Metric”, Semester Gasal 2022/2023
- Slides Materi Sains Data, “Regression”, Semester Genap 2020/2021
- D. Chahyati, et.al, Slides Materi Sains Data, “Metrik Evaluasi Klasifikasi”

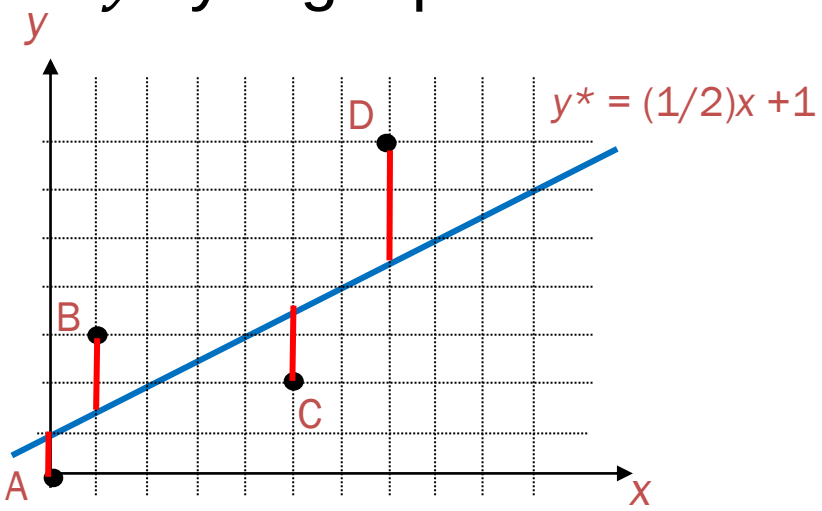
Performance Metric

- Setelah menentukan metode evaluasi model, tentukan metrik yang digunakan untuk menggambarkan kinerja model Anda

- Metrik classification berbeda dengan regression
  - classification adalah prediksi di mana variabel outputnya berjenis kategorikal
  - regression adalah prediksi di mana variabel outputnya adalah numerik

Performance Metric: Regression

- Bagaimana menilai apakah model regresi yang didapat sudah bagus atau belum?
- Terdapat setidaknya 4 metrik untuk mengetahui kinerja model regresi yaitu MAE, MSE, RMSE, dan R^2
- Semua metrik tersebut menghitung error antara nilai y data asli dan nilai y^* yang diprediksi oleh garis regresi (garis merah)



Titik	x	y	y^*= (1/2) x + 1
A	0	0	(1/2).0 +1 = 1
B	1	3	(1/2).1 +1 = 1.5
C	5	2	(1/2).5 +1 = 3.5
D	7	7	(1/2).7 +1 = 4.5

Performance Metric: Regression

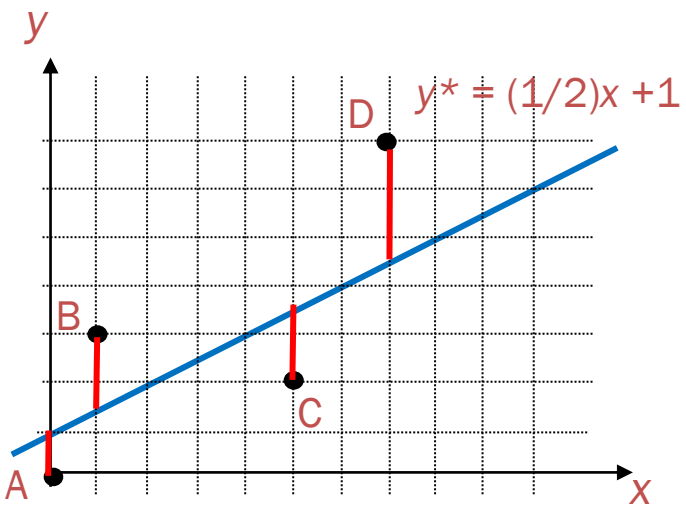
- **Mean Absolute Error (MAE)** $MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - y_i^*|$
- **Mean Squared Error (MSE)** $MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - y_i^*)^2$
- **Root Mean Squared Error (RMSE)** $RMSE = \sqrt{MSE}$

Keterangan:

- y_i : nilai output data ke-i (ground truth)
- y^*_i : nilai output data ke-i yang diprediksi
- N : ukuran dataset

Performance Metric: Regression

1. MAE (Mean Absolute Error) $= \frac{1}{n}\sum_{i=1}^{n}|y_i - y_i^*|$
2. MSE (Mean Squared Error) $= \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)^2$
3. RMSE (Root Mean Squared Error) $= \sqrt{MSE}$
4. R-squared $= R^2 = 1 - \frac{SS_{regression}}{SS_{total}}$ dimana SS = Sum of Squares



	x_i	y_i	y_i^*	$\lvert y_i - y_i^* \rvert$	$(y_i - y_i^*)^2$
A	0	0	1	1	1
B	1	3	1.5	1.5	2.25
C	5	2	3.5	1.5	2.25
D	7	7	4.5	2.5	6.25
			Sum	6.5	11.75
			Average	MAE = 1.625	MSE = 2.9375 RMSE = 1.7139

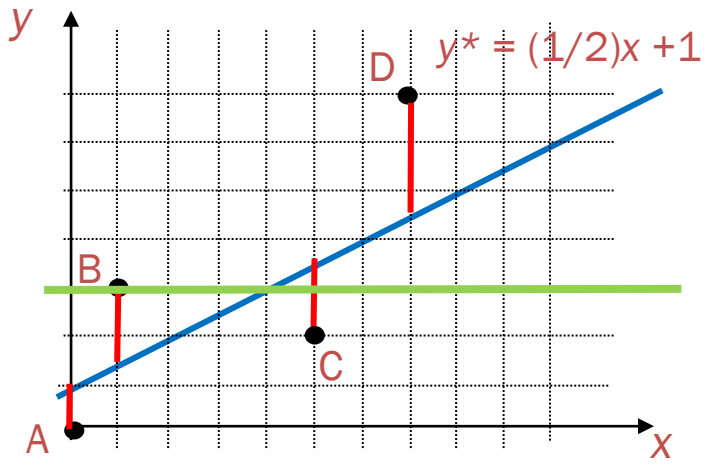
Performance Metric: Regression

$$R^2 = 1 - \frac{SS_{regression}}{SS_{total}} = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i^*)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Error terhadap garis regresi

Error terhadap garis y = (rerata y)

- Rerata y dalam contoh = (0+3+2+7)/4 =3
- $R^2 = 1 - \frac{SS_{regression}}{SS_{total}} = 1 - \frac{11.75}{26} = 1 - 0.45 = 0.55$
- Nilai R^2 berkisar dari 0 – 1
- Umumnya (tidak selalu)
  - Semakin **baik** regresi, semakin kecil $SS_{regression}$, semakin **besar** nilai R^2
  - Semakin **buruk** regresi, semakin besar $SS_{regression}$, semakin **kecil** nilai R^2



	x_i	y_i	y_i^*	$(y_i - \bar{y})^2$	$(y_i - y_i^*)^2$
A	0	0	1	9	1
B	1	3	1.5	0	2.25
C	5	2	3.5	1	2.25
D	7	7	4.5	16	6.25
			Sum	26	11.75

Performance Metric: Regression

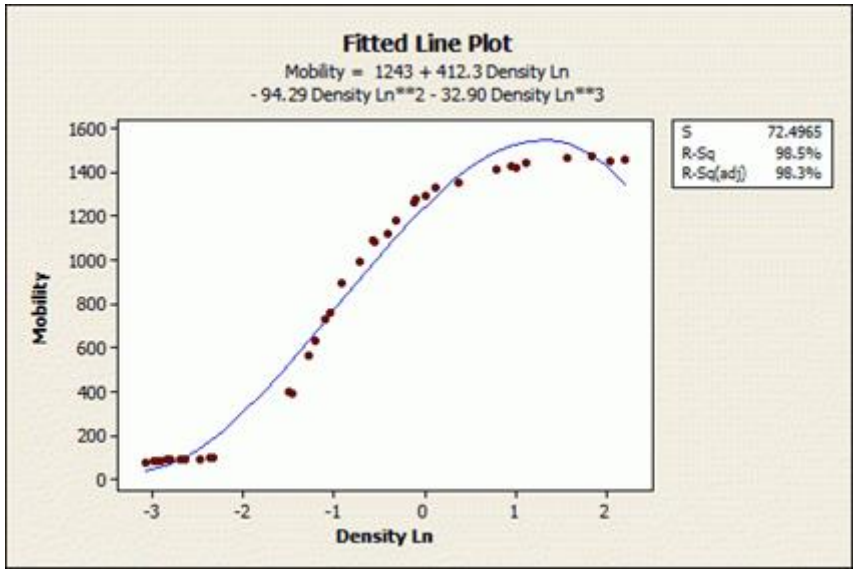
R-squared (R^2)

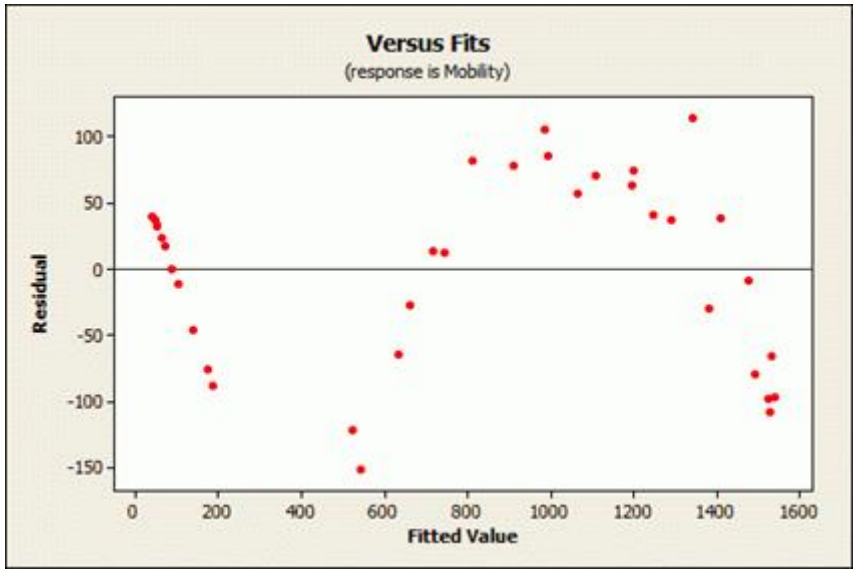
Nilai R^2 = 0.55

- 55% Variasi variabel yang diprediksi (y) dapat dijelaskan (explained) oleh variasi nilai input (x)
- 45% Variasi variabel yang diprediksi (y) diperoleh dari sumber lain yang tidak dapat dijelaskan.

Apakah nilai R^2 mendekati 1 selalu berarti bagus? Tidak! Pada gambar R^2 = 0.98 tapi jika kita lihat lagi, sebenarnya model regresi tersebut belum sempurna.

Jika dilihat dari analisis residunya, plotnya tidak acak melainkan membentuk pola tertentu. Pola ini masih bisa diekplorasi lebih lanjut.





https://statisticsbyjim.com/regression/interpret-r-squared-regression/

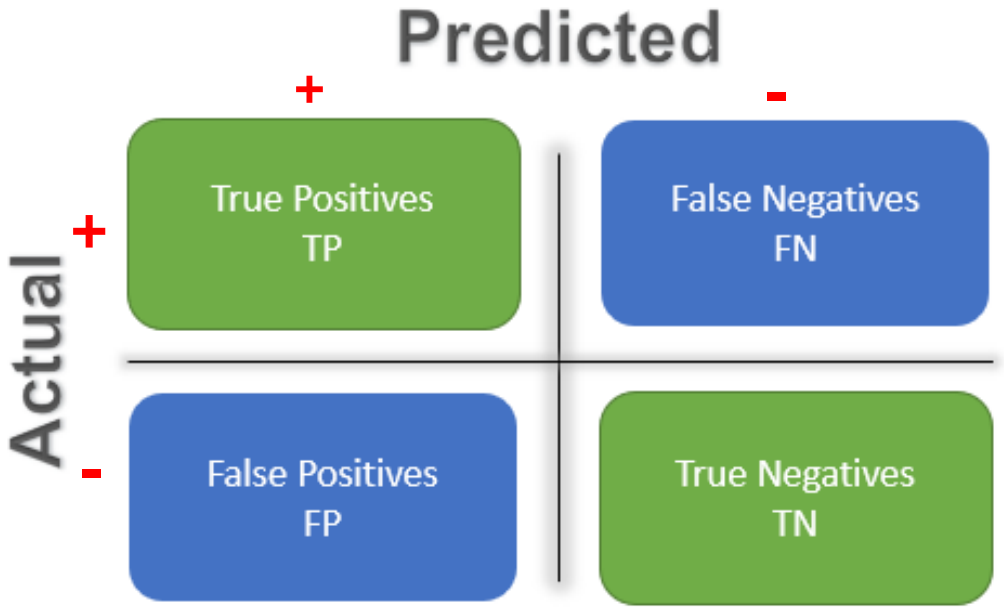
Performance Metric: Classification

- Bagus atau tidaknya suatu classifier (model yang digunakan untuk classfication) bergantung pada metrik evaluasi yang digunakan.
- Jika metrik yang digunakan tidak tepat, maka bisa jadi model yang dipilih tidak bagus, atau bahkan *misleading.*
- Untuk *imbalanced classification*, kita perlu lebih hati-hati karena bisa jadi metrik akan sangat dipengaruhi oleh kelas mayoritas.
- Pemilihan metrik bergantung pada aspek apa yang penting untuk diperhatikan oleh peneliti.

Performance Metric: Classification

Binary Classification

- ada kelas yang bisa dikategorikan sebagai kelas “Positif”, lainnya “Negatif”
- Ingat matriks di bawah ini?

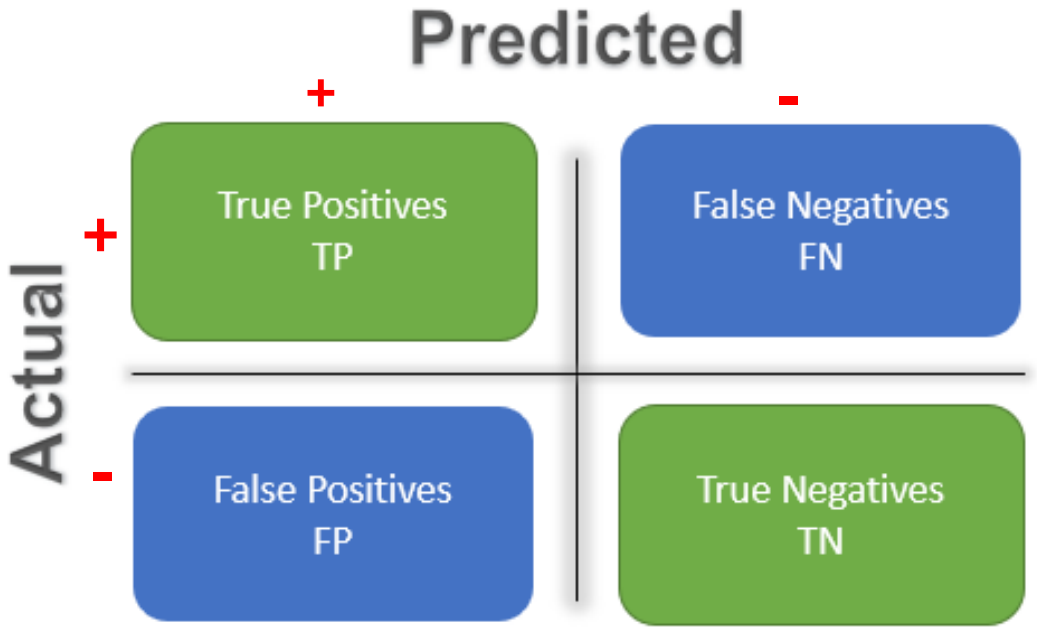


https://towardsdatascience.com/various-ways-to-evaluate-a-machine-learning-models-performance-230449055f15

Performance Metric: Classification

Confusion Matrix

- Matriks yang menyajikan jumlah TP, FN, FP, TN



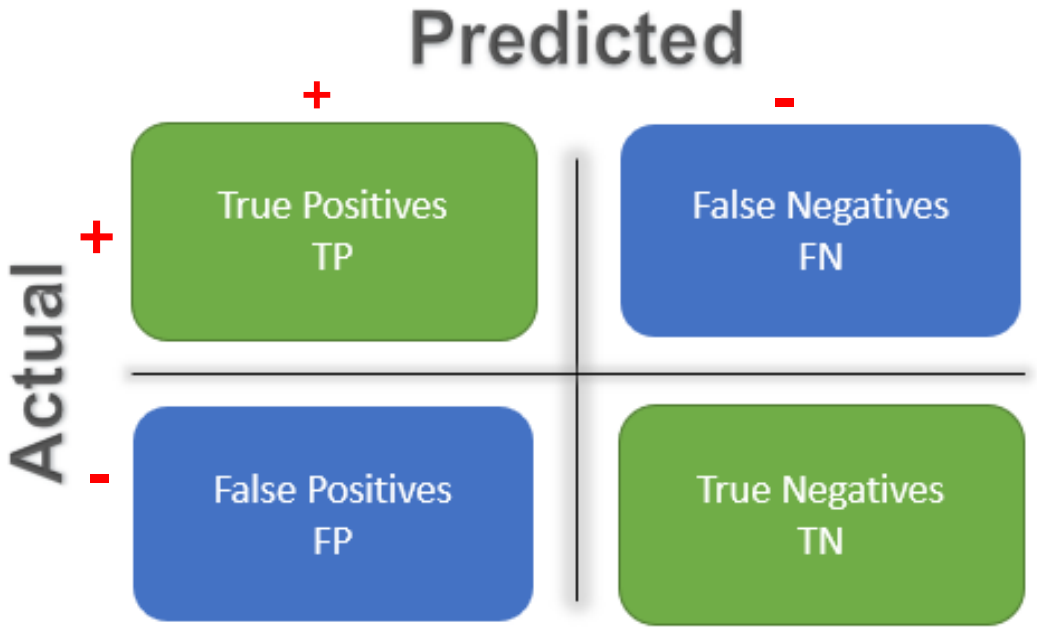
- **TP**: jumlah data yang berlabel kelas “positif” dan berhasil diprediksi sebagai kelas “positif”
- **FN**: jumlah data yang berlabel kelas “positif” tetapi diprediksi sebagai kelas “negatif”
- **FP**: jumlah data yang berlabel kelas “negatif” tetapi diprediksi sebagai kelas “positif”
- **TN**: jumlah data yang berlabel kelas “negatif” dan berhasil diprediksi sebagai kelas “negatif”

https://towardsdatascience.com/various-ways-to-evaluate-a-machine-learning-models-performance-230449055f15

Performance Metric: Classification

Accuracy (Akurasi)

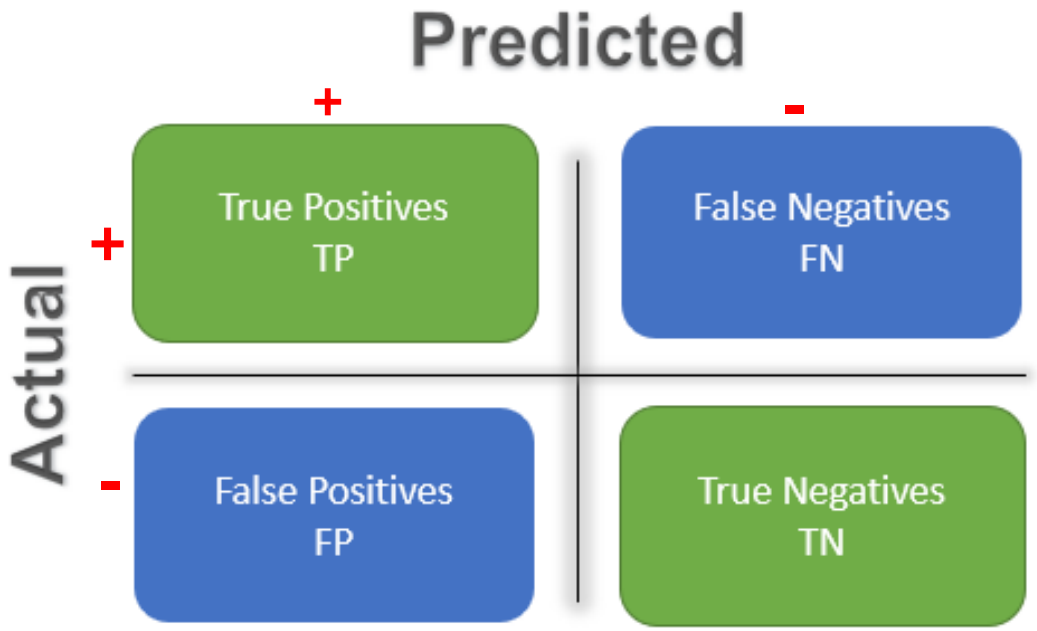
$$\frac{TP + TN}{TP + FP + TN + FN}$$



- Merupakan metrik yang paling mudah dipahami
- Dapat digunakan untuk binary classification maupun multiclass
- Baik digunakan jika tidak ada masalah imbalanced class

https://towardsdatascience.com/various-ways-to-evaluate-a-machine-learning-models-performance-230449055f15

Performance Metric: Classification



Precision

$$\frac{TP}{TP + FP}$$

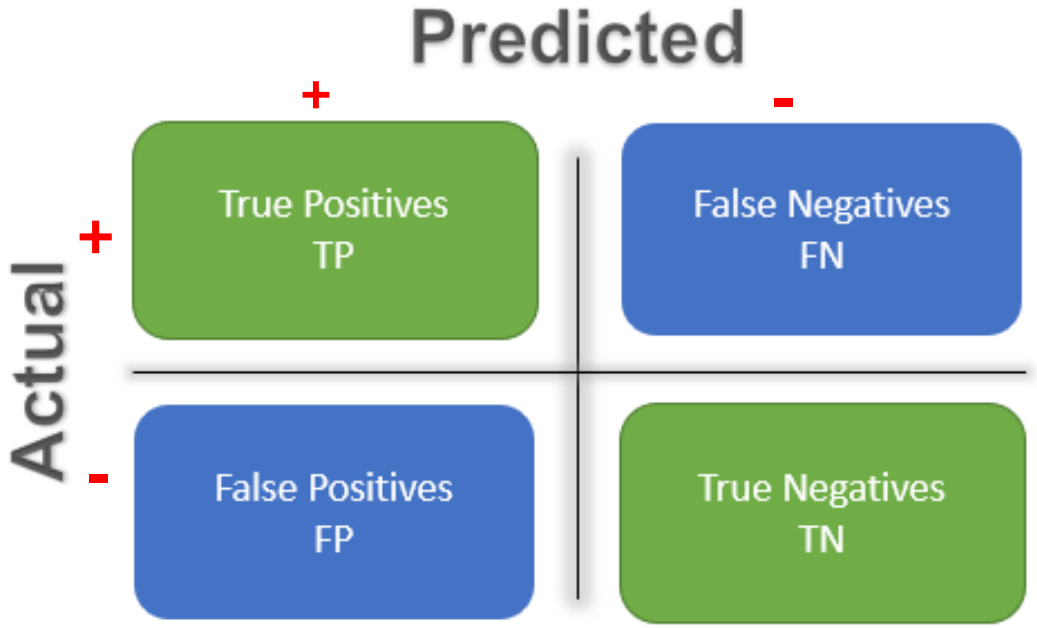
Recall/Sensitivity/True Positive Rate

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

Hati-hati: Jika classifier kita error dan menebak semua sample menjadi positif, maka hasil Recall = 1

https://towardsdatascience.com/various-ways-to-evaluate-a-machine-learning-models-performance-230449055f15

Performance Metric: Classification



Specificity

Mirip recall tapi dari sisi kelas “negatif”

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

Geometric Mean (G-Mean) of Sensitivity & Specificity

Mengukur keseimbangan antara kinerja model pada kelas “positif” dan “negatif”

$$G\text{-}mean = \sqrt{sensitivity * specificity}$$

https://towardsdatascience.com/various-ways-to-evaluate-a-machine-learning-models-performance-230449055f15

Sensitivity vs Specificity

- sensitivity: persentase true positives yang diklasifikasikan sebagai positive,
- Specificity: persentase true negatives diklasifikasikan sebagai negative.

Diskusi

Sensitivity vs Specificity, mana yang lebih perlu diperhatikan untuk mengukur

- Alat deteksi cancer
- Baggage detector di bandara

Performance Metric: Classification

F1-Score

- “harmonic mean” dari precision dan recall
- Range: [0, 1]
- bersama dengan G-Mean, berguna untuk mengukur kinerja pada imbalanced class
- Intuisi: F1-score baru akan besar kalau kedua nilai precision dan recall besar. Jika ada salah satu yang rendah, maka nilai F1 juga akan rendah)

- F1 = $\frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 * precision * recall}{precision + recall}$

- Mengapa tidak menggunakan **simple mean** dari precision dan recall saja?
- Kita akan lihat dua contoh kasus dimana terjadi *imbalanced class*

See: F-beta score

https://towardsdatascience.com/various-ways-to-evaluate-a-machine-learning-models-performance-230449055f15

F-Beta Score

- F1 score memberikan bobot yang sama untuk precision dan recall.
- Jika kita ingin mementingkan salah satunya, maka kita bisa menambahkan bobot β pada precision
- Rumusnya menjadi sebagai berikut

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}.$$

Contoh: F-1 Score

	Hasil Tes Positif	Hasil Tes Negatif
Orang sakit (Kelas +)	5	1
Orang sehat (Kelas -)	100	894

Kasus: Imbalanced Class

- Akurasi : 0.899
- Precision: 5/105 = 0.048
- Recall : 5/6 = 0.833
- Simple Mean = (0.048+0.833)/2= 0.4405
- Harmonic Mean = $F_1 = 2 * \frac{precision * recall}{precision + recall} = 2 * \frac{0.04}{0.881} = 0.09$

Contoh: F-1 Score

	Hasil Tes Positif	Hasil Tes Negatif
Orang sakit (Kelas +)	5	1
Orang sehat (Kelas -)	100	894

- Pada kasus ini, simple mean menghasilkan nilai yang masih tinggi (0.4405)
- Harmonic Mean (F1) menunjukkan angka yang mendekati 0, karena Precision juga mendekati 0
- Intuisi: Nilai F1 baru akan besar kalau **kedua nilai** precision dan recall besar. Jika ada salah satu yang rendah, maka nilai F1 juga akan rendah)
- Alat tes/classifier ini tidak ideal karena masih banyak orang sehat yang dinyatakan positif.
- Akurasi (0.899) tidak baik digunakan karena misleading

Latihan

Deteksi transaksi ilegal merupakan salah satu tugas lembaga perbankan nasional, untuk mencegah kejahatan keuangan seperti penyalahgunaan wewenang, money laundring, dan sebagainya. Lembaga perbankan ini menawarkan mahasiswa Fasilkom UI untuk mengembangkan sebuah *classifier* untuk mendeteksi transaksi perbankan yang ilegal berdasarkan data transaksi nasabah yang dimiliki oleh bank tersebut. Dari serangkaian uji coba pada 1500 subjek, didapatkan hasil sebagai berikut:

- *Classifier* berhasil mendeteksi transaksi legal sebagai transaksi legal sebanyak 1350 kali.
- *Classifier* salah mendeteksi transaksi legal sebagai transaksi ilegal 25 kali.
- *Classifier* salah mendeteksi transaksi ilegal sebagai legal sebanyak 75 kali.
- *Classifier* berhasil mendeteksi ilegal sebagai transaksi ilegal sebanyak 50 kali.

a) Buatlah *confusion matrix* berdasarkan narasi di atas
b) Hitunglah: *Sensitivity, Specificity, Total accuracy, Precision, F1-score*
c) Analisislah kinerja *classifier* yang dikembangkan berdasarkan nilai-nilai metrik di atas. Metrik mana yang menurut Anda paling penting pada kasus deteksi transaksi ilegal ini? Berikan penjelasan yang mendukung pilihan Anda.

		Predicted	
		+	-
Actual	+	1350	75
	-	25	50

Apa ini benar?

- Sensitivity = TP / (TP + FN) = 1350 / (1350 + 75) = 0.947 = Recall
- Specificity = TN / (TN + FP) = 50 / (50 + 25) = 0.667
- Accuracy = 1400 / 1500 = 0.933
- Precision = TP / (TP + FP) = 1350 / (1350 + 25) = 0.982
- F1 Score = 2 (precision* recall) / (precision + recall)
  = 2 (0.947 * 0.982) / (0.947 + 0.982) = **0.964**

		Predicted	
		+	-
Actual	+	50	75
	-	25	1350

Bagaimana dengan ini?

- Sensitivity = TP / (TP + FN) = 50 / (50+75) = 0.40 = Recall
- Specificity = TN / (TN + FP) = 1350 / (1350 + 25) = 0.98
- Accuracy = 1400 / 1500 = 0.93
- Precision = TP / (TP + FP) = 50 / (50 + 25) = 0.67
- F1 Score = 2 (precision* recall) / (precision + recall)
  = 2 (0.40 * 0.67) / (0.40 + 0.67) = **0.5**

Performance Metric: Classification

Classification dengan Threshold

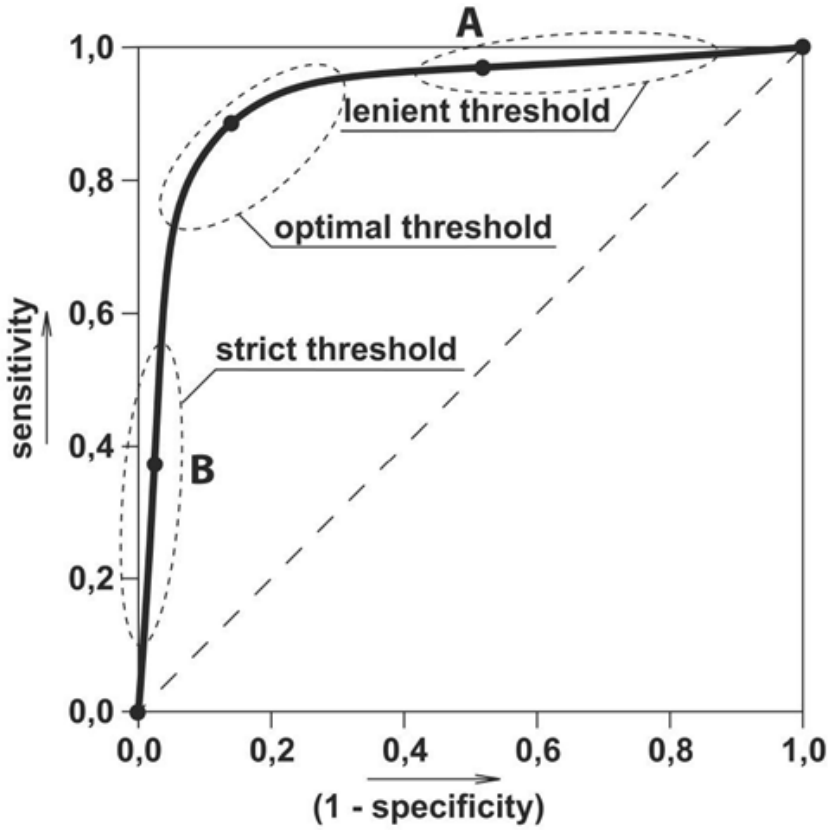
- Seringkali kita menggunakan thresholding untuk menentukan apakah suatu objek termasuk dalam suatu kelas tertentu
  - Misal, alat tes Covid-19 mengeluarkan angka dari 0 - 1
- Jika digunakan threshold 0.8, maka orang yang ditest dinyatakan positif jika hasil tes menunjukkan angka >= 0.8
- Cenderung akan sedikit kasus False Positive, tapi False Negative bisa jadi tinggi
- Jika digunakan threshold 0.4, maka orang yang ditest dinyatakan positif jika hasil tes menunjukkan angka >= 0.4
- Cenderung akan banyak kasus False Positive, tapi False Negative jadi rendah
- Menentukan threshold tidak mudah, maka dibuat metrik untuk melihat kinerja alat pada beberapa threshold

Performance Metric: Classification

Receiver Operating Characteristic (ROC) Curve

- kurva yang dibentuk dengan plotting nilai True Positive Rate (TPR) terhadap False Positive Rate (FPR) pada bermacam threshold.
- TPR = Sensitivity = Recall
- FPR = FP/(TN+FP) = 1 – Specificity

Threshold	TP	FN	FP	TN	Spec	1-Spec	Sens
1.0	0	50	0	50	1	0	0
0.9	30	20	3	47	0.94	0.06	0.6
...	...	...	...	...	...	...	...
0.6	40	10	5	45	0.9	0.1	0.9
0.2	45	5	30	20	0.4	0.6	0.95
0	50	0	50	0	0	1	1

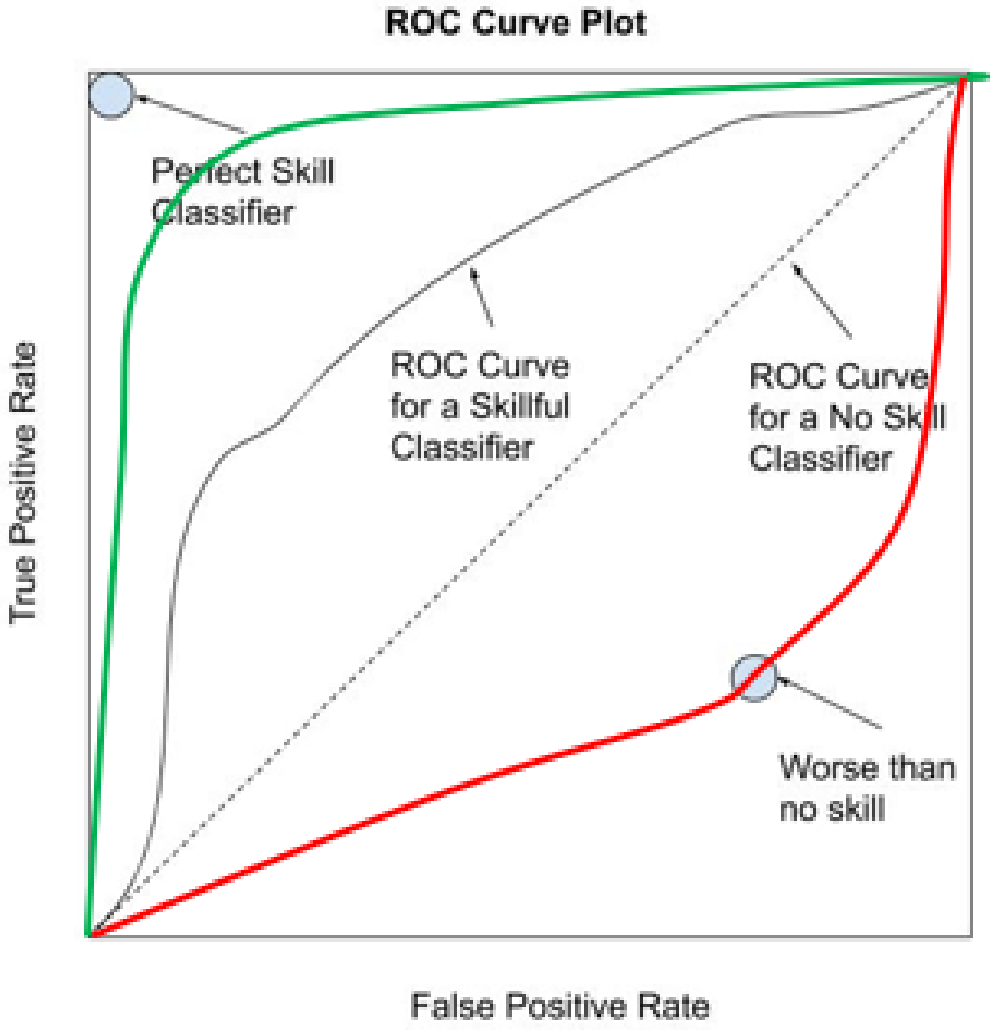


https://towardsdatascience.com/the-5-classification-evaluation-metrics-you-must-know-aa97784ff226
(image source)

Performance Metric: Classification

Receiver Operating Characteristic (ROC) Curve

- Kita mengharapkan classifier / alat tes yang kinerjanya mendekati kurva hijau
- Kita tidak menginginkan kurva yang kinerja mendekati kurva merah



https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/
(image source)

Performance Metric: Classification

Area Under ROC Curve (AUC)

- Merupakan luas dari daerah dibawah kurva ROC
- Menunjukkan probabilitas kelas positif dapat dipisahkan dari kelas negatif
- Contoh sebelumnya, luas daerah dibawah kurva hijau mendekati 1, artinya kinerjanya baik
- AUC bersifat scale-invariant
- AUC bersifat threshold-invariant, karena memperlihatkan kinerja classifier di berbagai nilai threshold, tidak seperti F1-score yang masih bergantung pada satu nilai threshold saja.
- PR-AUC: Area Under Precision vs Recall Curve, digunakan jika kelas positif lebih penting (Precision dan Recall sama-sama menghitung rasio TP terhadap sesuatu)

Performance Metric: Classification

Brier Score

- Pada beberapa metode klasifikasi, outputnya tidak langsung berupa kelas, namun berupa probabilitas suatu data masuk ke suatu kelas
- Binary Classification: hanya ada kelas 0 dan kelas 1

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$

Keterangan:

- f_t adalah nilai yang dikeluarkan oleh classifier
- o_t adalah kelas sesungguhnya dari sampel (0 atau 1)

Confusion Matrix pada Multiclass Classification

- False positive dan false negative rate dari *two-class classifier* dapat di-generalisasi untuk evaluasi dari *multi-class classifier*, membentuk **class confusion matrix**:
  - *cell ke-i, j berisi jumlah kasus (atau dapat berupa fraction) dimana true label adalah i dan predicted label adalah j.*

Class Confusion Matrix

Contoh **class confusion matrix** dari classifier untuk prediksi tingkat penyakit jantung berdasarkan pengukuran fisiologis dan fisik.

True	Predict					Class Error
	0	1	2	3	4	
0	151	7	2	3	1	**7,93%**
1	5	25	5	2	0	**32,43%**
2	1	8	150	2	3	**8,54%**
3	1	1	0	60	1	**4,76%**
4	1	2	3	2	100	**7,41%**

- Dataset tersebut memiliki 5 kelas (0 …. 4)
- Cell ke-*i, j* menampilkan jumlah data points dari true class *i* yang diklasifikasikan ke predicted class *j*.
- Pada setiap baris terdapat **class error rate**
- Perhatikan bagian diagonal dari table: jika nilai tertinggi terdapat di bagian tersebut, maka dapat dikatakan classifier bekerja dengan baik.
- Tabel menunjukkan prediksi yang buruk untuk kelas 1, dan cukup baik untuk kelas lainnya

F1 Score pada Multiclass Classification

Multi-class Classification

Jika jumlah kelas lebih dari dua, maka perhitungan nilai akurasi, precision, recall, F1, dst dapat menggunakan pendekatan sebagai berikut:

- micro: *Calculate metrics globally by counting the total number of times each class was correctly predicted and incorrectly predicted*.
- macro: *Calculate metrics for each "class" independently, and find their unweighted mean. This does not take label imbalance into account*.

F1 Score pada Multiclass Classification

Multi-class Classification: Macro

- Contoh: mendeteksi hewan apa yang ada di gambar

		True/Actual		
		Cat (🐱)	Fish (🐟)	Hen (🐓)
Predicted	Cat (🐱)	4	6	3
	Fish (🐟)	1	2	0
	Hen (🐓)	1	2	6

Class	Precision	Recall	F1-score
Cat	30.8%	66.7%	42.1%
Fish	66.7%	20.0%	30.8%
Hen	66.7%	66.7%	66.7%

- Ketika menghitung TP untuk Cat, anggap hanya ada 2 kelas yaitu “Cat” dan “Non-Cat”
  - Cat Precision = 4/13 = 30.8%
  - Cat Recall = 4/6 = 66.6%
  - Macro-F1 = (42.1% + 30.8% + 66.7%) / 3 = 46.5%
  - Macro-precision = (31% + 67% + 67%) / 3 = 54.7%
  - Macro-recall = (67% + 20% + 67%) / 3 = 51.1%

https://towardsdatascience.com/multi-class-metrics-made-simple-part-ii-the-f1-score-ebe8b2c2ca1

F1 Score pada Multiclass Classification

Multi-class Classification: Micro

Nilai FP, TP, FN, TN dihitung langsung dari tabel confusion matrix.

		True/Actual		
		Cat (🐱)	Fish (🐟)	Hen (🐓)
Predicted	Cat (🐱)	4	6	3
	Fish (🐟)	1	2	0
	Hen (🐓)	1	2	6

- TP = 4 + 2 + 6 = 12
- FP = 6 + 3 + 1 + 0 + 1 + 2 = 13
- False Negative: jika cat dianggap sebagai fish, maka itu dianggap false negative, sehingga nilainya sama dengan FP
- Micro Precision = 12/(12 + 13) = 48%
- Micro Recall = 12/(12+13) = 48%
- Karena nilai Micro precision = micro recall, maka nilai micro-F1 juga akan sama dengan kedua nilai ini.

https://towardsdatascience.com/multi-class-metrics-made-simple-part-ii-the-f1-score-ebe8b2c2ca1

UNIVERSITAS INDONESIA

FAKULTAS ILMU KOMPUTER

TERIMA KASIH

Disclaimer: Figures and content can be originated from other sources on the Web. The purpose of this slide set is educational only.