DIGITAL TALENT SCHOLARSHIP

THEMATIC ACADEMY

Evaluasi Model

Pertemuan #15 :

KOMINFO

#JADIJAGOANDIGITAL

Badan Penelitian dan Pengembangan Sumber Daya Manusia

Learning Objective

Rencana Pembelajaran		
1	**Pertemuan Ke**	15
2	**Topik**	**Membangun Model: Evaluasi** **Unit Kompetensi:** **1. J.62DMI00.014.1 - Mengevaluasi Hasil Pemodelan** **2. J.62DMI00.015.1 - Melakukan Proses Review Pemodelan**
3	**Deskripsi Topik**	**1. J.62DMI00.014.1 - Mengevaluasi Hasil Pemodelan** **a. Menggunakan model dengan data riil** **b. Menilai hasil pemodelan** **c. Membandingkan hasil pemodelan menggunakan hypothesis testing** **2. J.62DMI00.015.1 - Melakukan Proses Review Pemodelan** **a. Menilai kesesuaian proses pemodelan** **b. Menilai kualitas proses pemodelan**

Model yang diukur untuk:

Supervised Learning

- Klasifikasi
- Regresi

Unspervised Learning

- *Clustering*

Akurasi

Menunjukkan persentase klasifikasi yang bernilai valid terhadap total klasifikasi yang dilakukan.

$$a = \frac{t}{n} \times 100\%$$

dengan

a adalah akurasi dalam persen,

t adalah jumlah percobaan dengan prediksi valid, dan

n adalah jumlah percobaan

Contoh: Berapa akurasi dari percobaan di samping ini?

Data aktual	Output model (prediksi)	Kesimpulan
mangga	mangga	valid
jeruk	apel	invalid
apel	jeruk	invalid
mangga	apel	invalid
jeruk	jeruk	valid

Contoh

Data aktual	Output model (prediksi)	Kesimpulan
mangga	mangga	valid
jeruk	apel	invalid
apel	jeruk	invalid
mangga	apel	invalid
jeruk	jeruk	valid

Jumlah percobaan valid (t) = 2

Jumlah percobaan invalid = 3

Total Percobaan (a) = 5

$$a = \frac{t}{n} \times 100\%$$

$$= \frac{2}{5} \times 100\%$$

$$= 40\%$$

Akurasi dapat digunakan sebagai ukuran awal mengevaluasi model, namun tidak cukup dengan akurasi saja. Terkadang akurasi tiap kelas perlu diketahui juga.

Confusion matrix

*Bukan metric, namun bermanfaat melihat **sebaran validitas percobaan***

		Kelas Prediksi			
		mangga	apel	jambu	pear
Kelas Aktual	mangga	19	3	2	1
	apel	1	22	1	1
	Jambu	2	2	21	0
	Pear	0	1	1	23

Karakteristik:

- *Ada sumbu data aktual dan sumbu data prediksi (gunakan konvensi)*
- *Setiap kelas terpetakan satu sama lainnya*
- *Percobaan valid berada pada diagonal utama*
- *Matriks berbentuk bujur sangkar*

Confusion matrix

		Kelas Prediksi			
		mangga	apel	jambu	pear
Kelas Aktual	mangga	19	3	2	1
	apel	1	22	1	1
	Jambu	2	2	21	0
	Pear	0	1	1	23

Pada area kotak merah:

terdapat 25 mangga yang di uji, dengan 19 mangga dikenali sebagai mangga (valid), 3 mangga dikenali sebagai apel 9(invalid), 2 mangga dikenali sebagai jambu (invalid), dan 1 mangga dikenali sebagai pear (invalid)

Confusion matrix

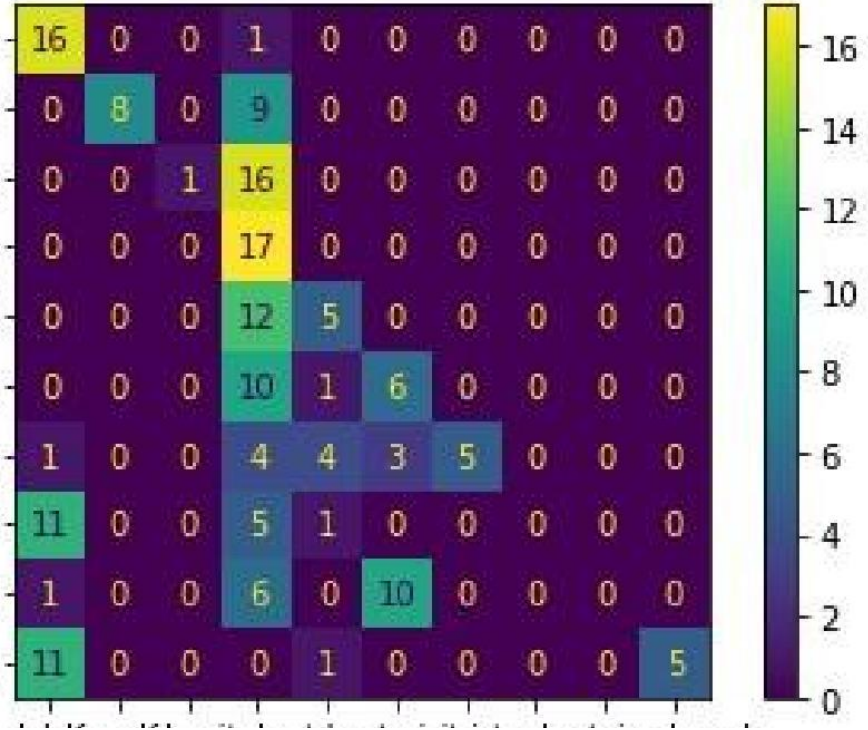
		Kelas Prediksi			
		mangga	apel	jambu	pear
Kelas Aktual	mangga	19	3	2	1
	apel	1	22	1	1
	Jambu	2	2	21	0
	Pear	0	1	1	23

Akurasi untuk pengujian kelas mangga adalah : 19/25 x 100% = 76%

Sedangkan akurasi total pengujian adalah : (19+22+21+23)/(25+25+25+25) x 100% = 85%

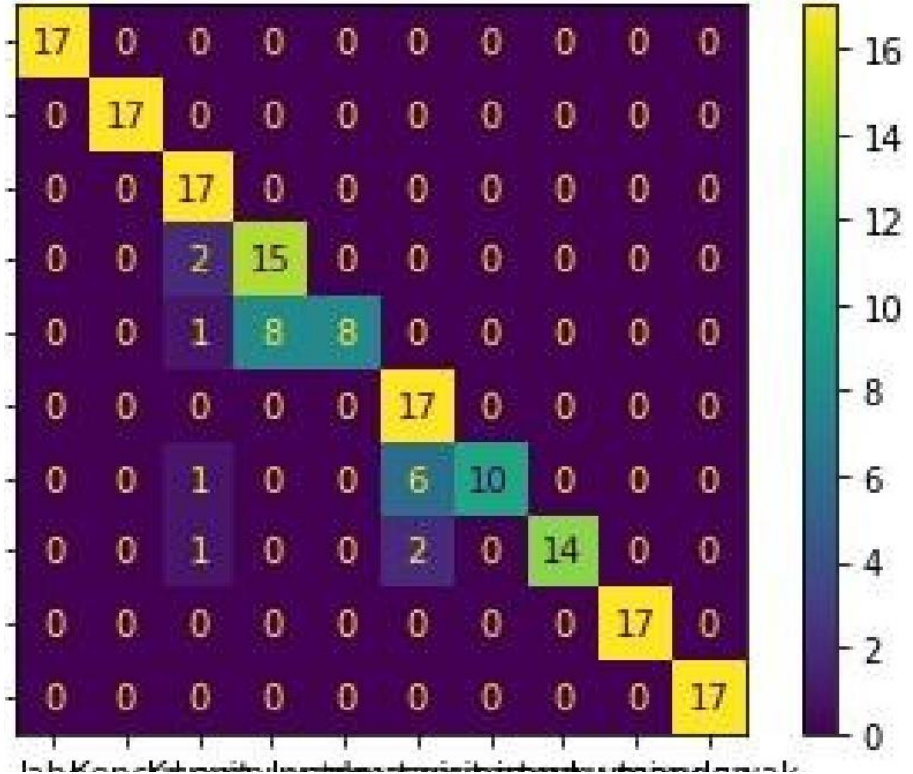
Visualisasi confusion matrix

- Representasi dengan heat map lebih baik.
- Contoh di samping adalah confusion matrix pada suatu percobaan di epoch 1.
- Berapa jumlah kelasnya?
- Jumlah data uji tiap kelas?



Visualisasi confusion matrix

- Contoh disamping hasil pada suatu percobaan pada epoch 4.
- Jumlah kelas = jumlah kolom = jumlah baris. Untuk case di samping, jumlah kelas adalah 10
- Jumlah data set **tiap kelas** sama untuk semua kelas = jumlah elemen tiap baris = 17
- Tanpa melihat nilainya, perbedaan heat map dapat diindera lebih cepat untuk membedakan hasil epoch 1 vs. epoch 4



Binary Classification

- *Hanya ada kelas: 0 atau 1, valid atau invalid, true atau false, positif atau negatif, bagus atau tidak bagus, cantik atau tidak cantik, rekomended atau tidak, lulus atau tidak, spam atau bukan spam, hoax atau bukan hoax, dsb.*

- *Bentuk yang paling umum: satu kelas dinyatakan sebagai kelas **positif** (menjadi fokus dalam klasifikasi), dan satu kelas lainnya dinyatakan sebagai kelas **negatif***

Contoh dalam dunia medis : sampel cairan mukus yang mengandung virus Covid-19 dinyatakan sebagai kelas positif dan sampel yang tidak mengandung virus dinyatakan sebagai kelas negatif

Contoh dalam kebencanaan : gempa yang mengakibatkan tsunami sebagai kelas positif dan yang tidak mengakibatkan tsunami sebagai kelas negatif

Keterbatasan akurasi

Dalam kasus deteksi pasien postifi Covid, sebuah detektor baru, sebut saja detektor X, diujikan pada 100 sampel. Sampel tersebut telah diuji dengan alat yang hasil deteksinya dijadikan acuan validitas (ground truth), yaitu PCR. Dari pengujian PCR sebelumnya diperoleh data bahwa 90 sampel adalah negatif dan 10 sampel adalah positif. Dengan menggunakan detektor X, ke 90 sampel negatif dideteksi negatif. Namun pada 10 sampel positif, diperoleh hasil bawah 5 sample dinyatakan sebagai positif dan 5 sampel sisanya negatif.

Berapa akurasi detektor X? (90+5)/(90+10) x 100% = 95% !

Apakah akurasi 95% merupkan hasil yang baik?

Kesalahan detektor X yang hanya pada 5% dapat berakibat fatal. Apakah ada metrik lain yang menjelaskan kasus semacam ini?

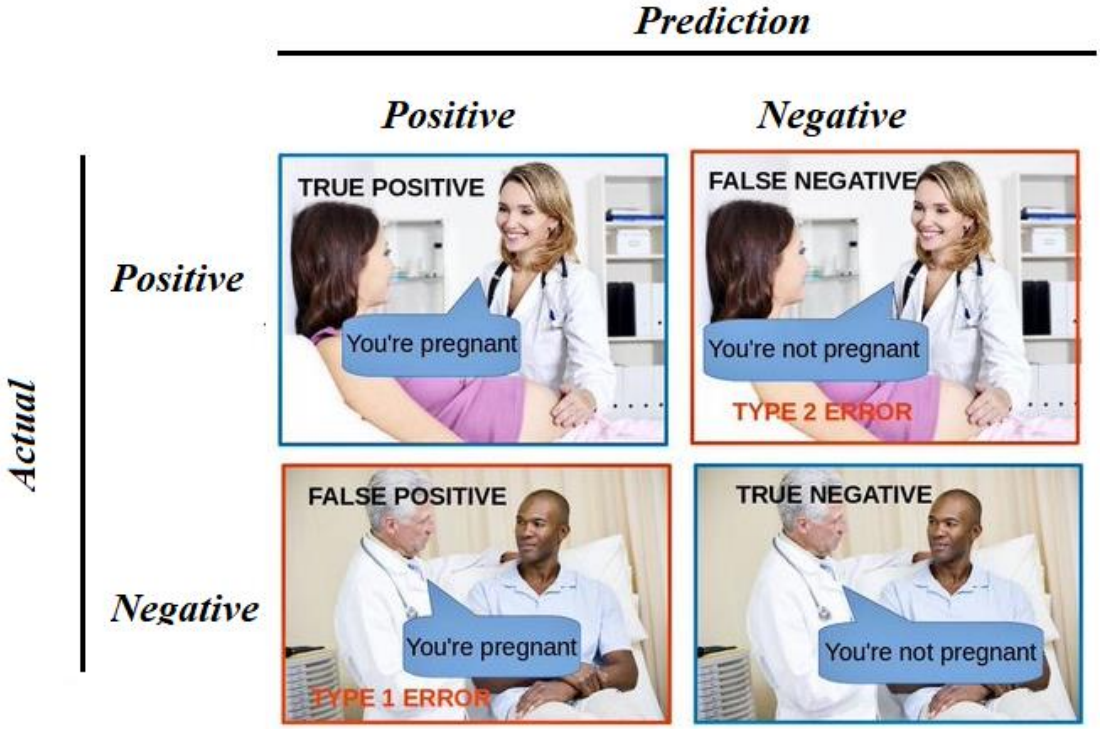
Confusion Matrix untuk Klasifikasi Biner

		Nilai Prediksi	
		Positive	Negative
Nilai Aktual	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

True Positive (TP): nilai sesungguhnya adalah positif dan diprediksi positif
False Positive (FP): nilai sesungguhnya adalah negatif namun diprediksi positif
True Negative (TN): nilai sesungguhnya adalah negatif dan diprediksi negatif, dan
False Negative (FN): nilai sesungguhnya adalah positif namun diprediksi negatif.

Klasifikasi yang bernilai **valid** adalah **TP dan TN**

Confusion Matrix untuk Klasifikasi Biner (Ilustrasi)



Source: Modifikasi dari https://skappal7.files.wordpress.com/2018/08/confusion-matrix.jpg

Metrik pada Klasifikasi Biner

		Nilai Prediksi: Positive	Nilai Prediksi: Negative	
Nilai Aktual	Positive	True Positive (TP)	False Negative (FN)	Recall, Sensitivity, True Positive Rate $\frac{TP}{TP + FN}$
	Negative	False Positive(FP)	True Negative (TN)	Specificity, True Negative Rate $\frac{TN}{FP + TN}$ False Positive Rate $\frac{FP}{FP + TN}$
		Precision $\frac{TP}{TP + FP}$	Negative Predictive Value $\frac{TN}{TN + FN}$	Accuracy $\frac{TP + TN}{TP + TN + FP + FN}$

Interpretasi metrik Recall - Precision

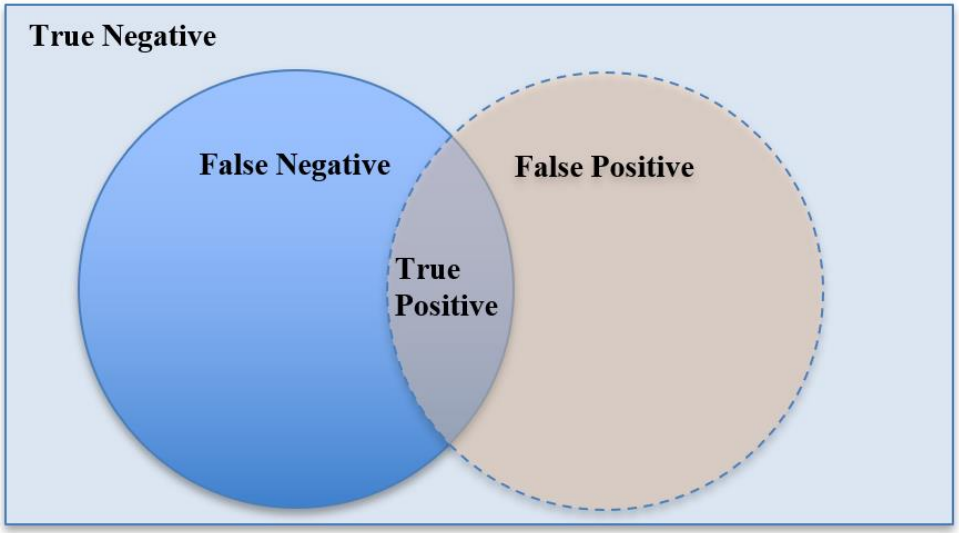
Recall dan Precision sering dihitung bersamaan untuk menggambarkan performansi model.

Kombinasi yang mungkin untuk keduanya:

- *Low recal low prescision*
- *High recall low precison*
- *Low recall high precision*
- *High recall high precision*

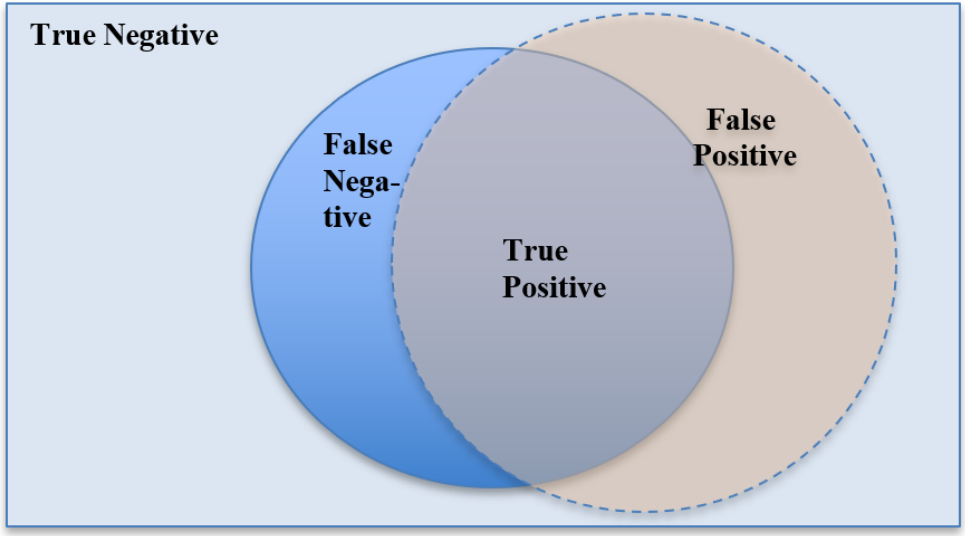
Low recall low precision

- *Model berkinerja kurang baik*
- *Baik False negatif dan maupun false positif bernilai besar pada model ini*

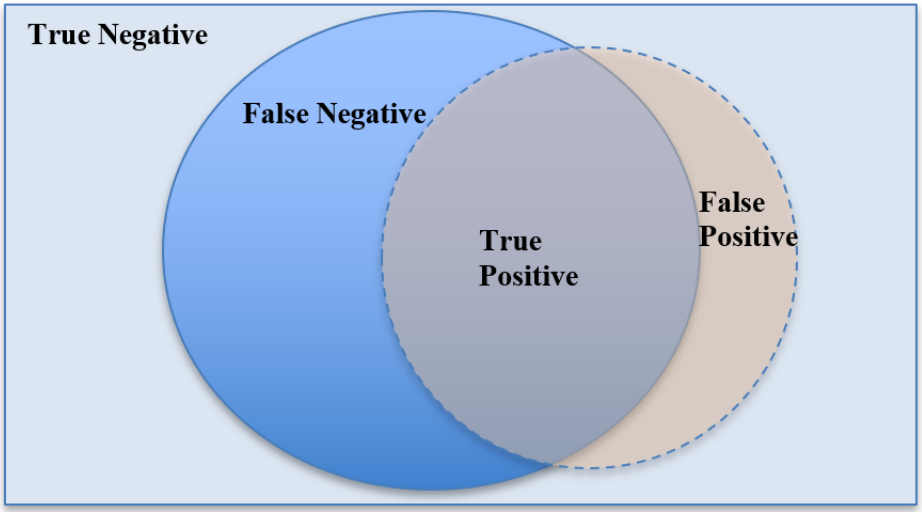


High recall low precision

- *Sebagian besar data positif dapat dikenali dengan baik (False Negative rendah)*
- *Tetapi banyak data negatif dikenali sebagai positif (False Positive tinggi*

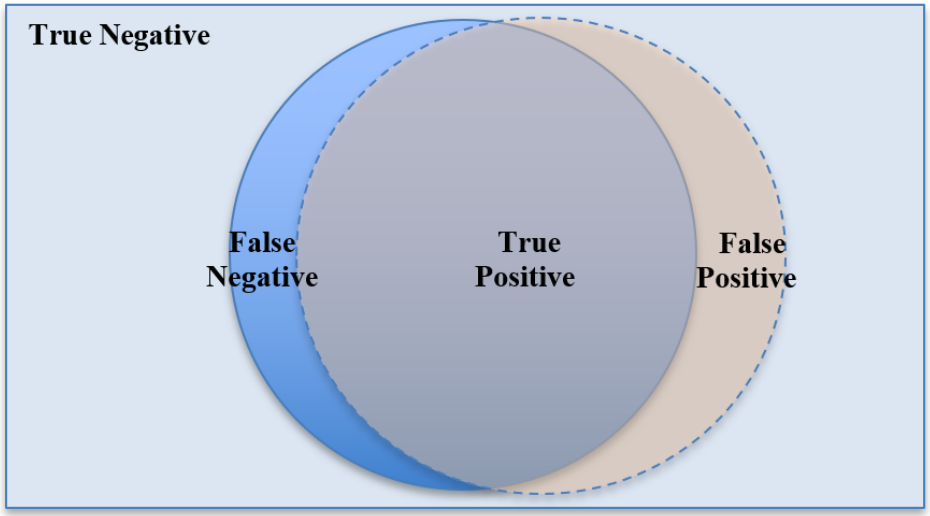


Low recall high precision



- *Banyak data positif yang teridentifikasi negatif (False Negative besar)*
- *Sebagian besar data yang teridentifikasi positif memang benar positif*

High recall high precision



Model memiliki kinerja baik

False Positive maupun False Negaative rendah

True Positve dan True Negative tinggi

F Score

- *Mengukur keseimbangan antara Precison – Recall*
- *Untuk model yang Precision dan Recall sama pentingnya digunakan F-1 Score, dinyatakan:*

$$F_1 = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}.$$

- *Untuk kasus recall relbih diutamakan dengan faktor β , maka formula diperluas menjadi:*

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

Evaluasi Model pada Probabilistik Model

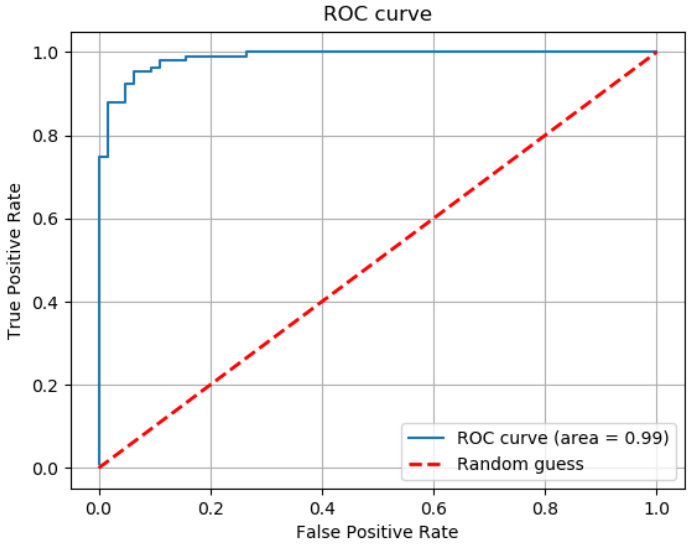
- *Model dalam memprediksi dengan menghasilkan nilai probabilitas [0,1] untuk suatu label*
- *Nilai Probabilitas menunujukkan seberapa yakin terhadap suatu kelas/label*
- *Label ditentukan dengan menetapkan suatu threshold*
- *Evaluasi dapat dilakukan beberapa metode, diantaranya:*

a. *Receiver Operator Characterisic (ROC) (dan pengembangan ke PR-AUC)*
b. *Logarithmic loss function*

https://www.analyticsvidhya.com/blog/2020/10/quick-guide-to-evaluation-metrics-for-supervised-and-unsupervised-machine-learning/

Evaluasi Model pada Probabilistik Model

- *Evaluasi dapat dilakukan metode Receiver Operating Characterisic (ROC) yaitu dengan mem-plot antara* ***Recall*** *(atau disebut juga* ***True Positive Rate****) sebagai sumbu –y dengan* ***False Positive Rate*** *sebagai sumbu-x untuk setiap threshold klasifikasi yang mungkin (threshold antara 0 hingga 1)*
- *Area yang diperoleh dari ROC dapat digunakan untuk analisis ROC – AUC (AUC: Area Under Curve),*

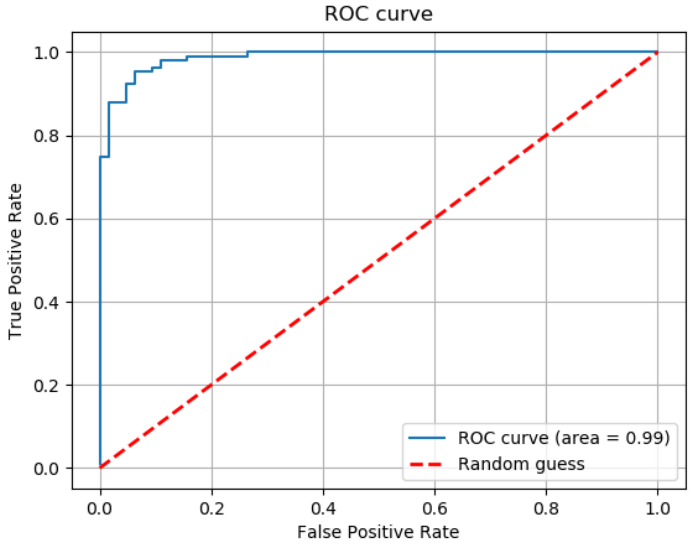


https://www.analyticsvidhya.com/blog/2020/10/quick-guide-to-evaluation-metrics-for-supervised-and-unsupervised-machine-learning/

ROC vs Akurasi

- *Mengapa perlu ROC-AUC dan tidak cukup menggunakan akurasi?*
- *ROC-AUC lebih menggambarkan secara lengkap visualisasi untuk semua threshold klasifikasi yang mungkin*

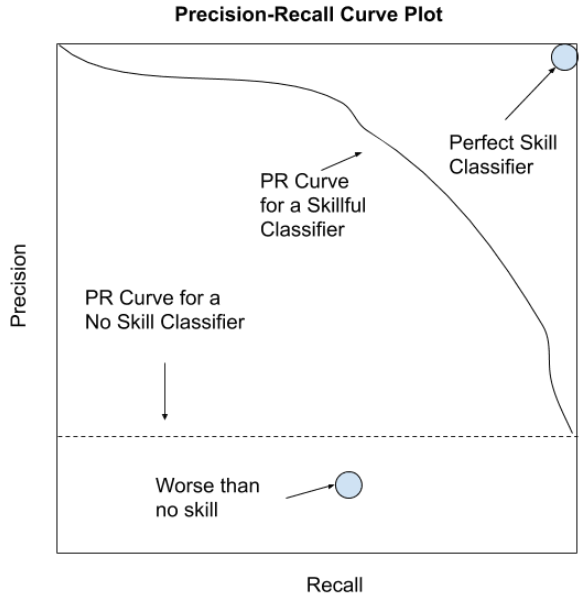
- *Akurasi hanya merepresentasikan performansi pada satu nilai threshold*
- ***Diskusi:** apa arti garus putus berwarna merah (random guess) pada diagram ROC-AUC ?*
- *Hasil paling optimal adalah yang menghasilkan AUC paling luas*



https://www.analyticsvidhya.com/blog/2020/10/quick-guide-to-evaluation-metrics-for-supervised-and-unsupervised-machine-learning/

Alternatif lain : PR - ROC

- *Plotting antara Precision (sumbu –y) dan Recall (sumbu –x)*
- *Digunakan untuk kelas yang minoritas yang menjadi perhatian utama cukup kecil*



https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/

Logarithmic loss (log loss)

- *Menunjukkan seberapa yakin pemberian label terhadap data yang diuji/diobeservasi*
- *Untuk setiap sample, perlu dihitung probabilitas untuk semua semua label yang mungkin*
- *Nilai log loss berada di [0,~) dan dinyatakan sebagai*

$$LogarithmicLoss = \frac{-1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} * \log(p_{ij})$$

  - *N: jumlah sampel ,*
  - *M : jumlah class ,*
  - y_{ij} *menunjukkan apakah sample i adalah kelas j atau bukan*
  - p_{ij} *menunjukkan probability sample i adalah kelas j*
- ***Diskusi:** semakin baik prediksi, nilai log loss semakin besar atau semakin kecil?*

Penerapan evaluasi model : Medis

Evaluasi model sangat bergantung kepada kasus dan distribusi data. Contoh model untuk medis:

- *deteksi penyakit menular (misalnya alat detektor G-Nose untuk penyakit Covid-19)*
- *deteksi kanker (misalnya kanker ganas / jinak pada kanker payudara , kanker serviks, dsb)*
- *deteksi kehamilan*
- *deteksi kekurangan gizi*

Ingat kembali detektor X untuk kasus Covid-19.

- *Dari parameter evaluasi : TP, TN, FP, FN, parameter manakah yang perlu untuk ditekan?*
- *Dari parameter FP dan FN parameter mana yang paling krusial, dimana kesalahannya berakibat fatal bagi manusia?*

Penerapan evaluasi model : Kebencanaan

Contoh model untuk kebencanaan:

- *Deteksi gunung meletus*
- *Deteksi kemarau panjang*
- *Deteksi banjir*
- *Deteksi gempa bumi*
- *Deteksi tsunami*

Penerapan evaluasi model : Telekomunikasi

Contoh model telekomunikasi:

- *Deteksi spam (dipakai sebagai spam filter)*
- *Deteksi hoax*
- *Deteksi fraud*
- *Deteksi pembajakan akun*

False Negative vs. False Postive?

- *Kesalahan prediksi berada di False Positive maupun false Negative.*

Pada Kebencanaan, misalnya deteksi dini tsunami di tepi pantai (tsunami : +):

- *False Negative : Diprediksi tidak ada tsunami, namun ternyata da tsunami*
- *False Positive : Diprediksi ada tsunami, namun ternyata tidak tsunami*

Pada case ini, Tsunami yang tidak terprediksi sebelumnya (FN) lebih membahayakan dan sangat merugikan, dibandingkan FP

Deteksi spam (dipakai sebagai spam filter, spam : +)

- *False Negative : Email Spam masuk inbox*
- *False positive : Email normal masuk folder spam*

Pada kasus ini, umumnya orang masih bisa menerima spam masuk inbox (FN) daripada email penting masuk spam (FP), bisa berakibat gagal kerja, gagal proyek, gagal sekolah, dsb.

Setiap permasalahan memiliki titik tekan parameter yang berbeda beda tergantung kasusnya!

Beberapa kasus aplikasi dengan Class Imbalance Problem

Isu dataset yang tidak seimbang muncul dalam berbagai persoalan dan menjadi bahasan tersendiri

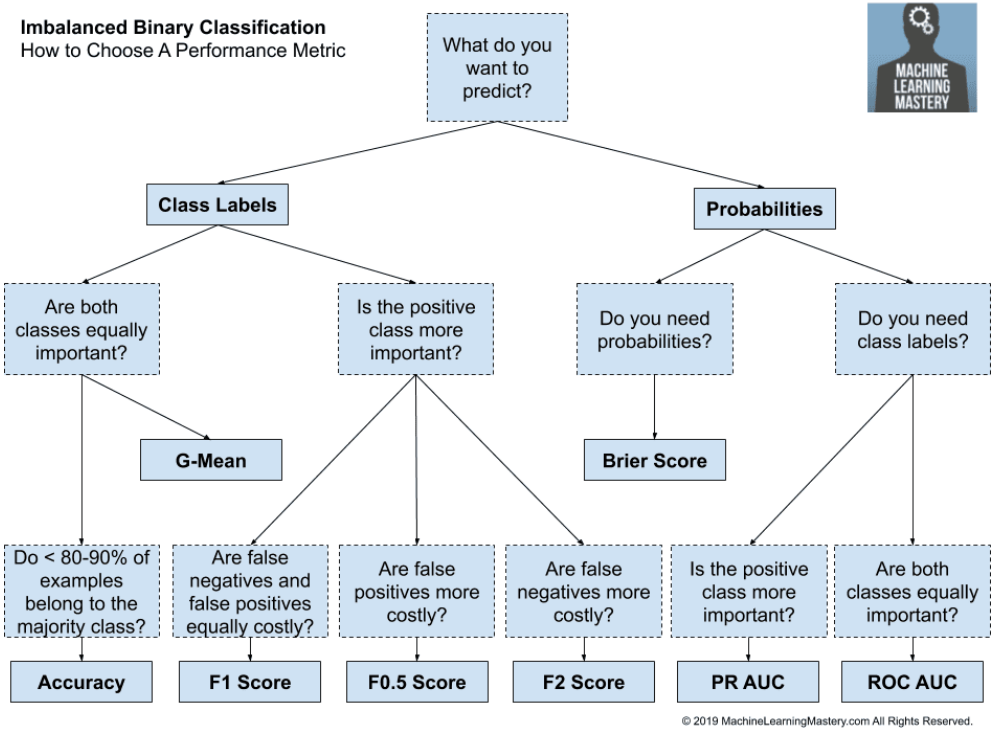
Berikut contoh aplikasi dengan data kelas tidak imbang (sumber: Learning from Imbalanced Data Sets, Alberto F. , et.al. Springer, 2018)

Applications of ML and DM where the class imbalance problem is present

Year	Domain	Subcategory	Application	Data-level	Internal	Cost-sensitive	Ensemble
1997	Engineering	Satellite radar images	Detection of oil spills in satellite radar images		×		
1997	Engineering	Satellite radar images	Detection of oil spills in satellite radar images	×			
1998	Engineering	Satellite radar images	Detection of oil spills in satellite radar images	×	×		
2012	Information technology	Software	Software defect prediction	×			×

Year	Domain	Subcategory	Application	Data-level	Internal	Cost-sensitive	Ensemble
2013	Bioinformatics	Protein identification	MicroRNA precursor classification	×			×
2014	Medicine	Quality control	Prediction of the post-operative life expectancy in lung cancer patients			×	×
2014	Bioinformatics	Protein identification	Five datasets that represent four different bioinformatics applications. These include miRNA identification, protein localization prediction, promoter identification from DNA sequences, kinase substrate prediction from protein phosphorylation profiling.	×			
2014	Information technology	Text mining	Text categorization	×			×
2014	Bioinformatics	Cell recognition	Mitotic cells recognition in Hep-2 images	×			×
2014	Medicine	Diagnosis	Lung nodule detection	×			×
2014	Information technology	Software	Software defect prediction	×		×	×
2014	Security	Video surveillance	Face re-identification	×			×
2014	Information technology	Network analysis	Botnet traffic detection	×		×	×
2014	Information technology	Network analysis	Network traffic classification				×
2015	Medicine	Quality control	Prediction of long stay patients in emergency department				×
2015	Bioinformatics	Protein identification	Protein data classification				×
2015	Medicine	Diagnosis	Diagnosis of diabetes mellitus	×			
2016	Business management	Customer relationship management	Customer churn prediction	×			
2016	Medicine	Diagnosis	Breast cancer malignancy classification				×
2016	Medicine	Diagnosis	Bleeding detection in endoscopic video	×			×
2016	Education	High school	Early dropout detection	×	×		
2016	Security	Video surveillance	Face re-identification		×		×
2016	Engineering	Semiconductors	Fault detection in semiconductors	×		×	×
2016	Medicine	Diagnosis	Thyroid nodule classification	×			
2016	Medicine	Diagnosis	Breast cancer classification from Magnetic Resonance Images (MRIs)				×
2016	Security	Biometric authentication	Multimodal biometric authentication				×
2017	Engineering	Energy	Short-term voltage stability assessment	×		×	
2017	Business management	Customer relationship management	Customer churn prediction	×		×	×
2017	Information technology	Network analysis	Mobile malware detection	×		×	
2017	Engineering	Semiconductors	Fault detection in semiconductors	×			×
2017	Medicine	Quality control	Prediction of the survival status of poly-trauma	×			

Alternatif Pilihan Metrik untuk Imbalanced Data Sets



https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/

Pengayaan: Evaluasi Model untuk Segmentasi Citra

Untuk operasi klasifikasi pixel citra (untuk keperluan segementasi), dimana citra akan dilabeli sebagai foregroud dan background, maka kualitas segemntasi dapat diukur dengan beberapa metode diantaranya Jaccard Index dan Similarity Index.

Misalnya A ada adalah hasil segmentasi dan B adalah ground truth, maka

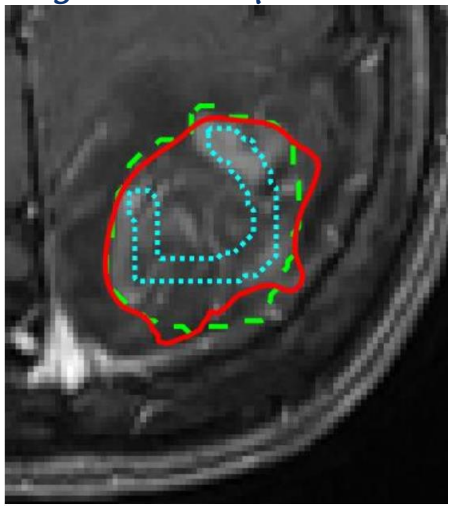
- *Jacard Index (JI)*

$$\mathrm{JI} = (A \cap B)/(A \cup B)$$

- *Similarity Index / Dice Coefficient (SI)*

$$\mathrm{SI} = (2|A \cap B|)/(|A| + |B|)$$

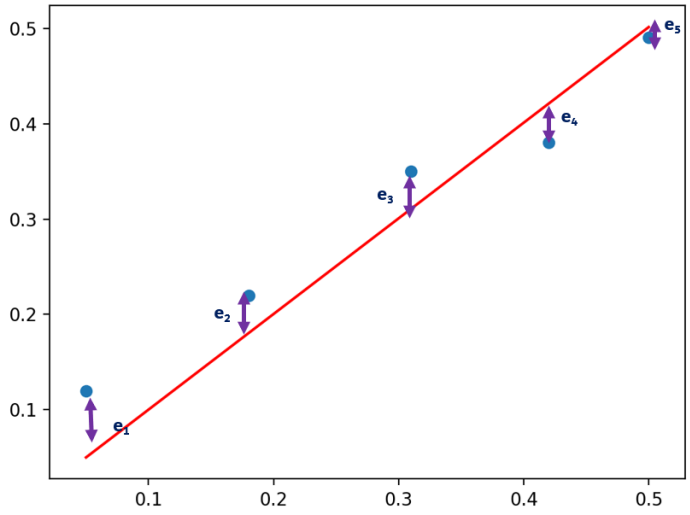
dimana $|.|$ menyatakan banyaknya elemen (dalam hal ini pixel)



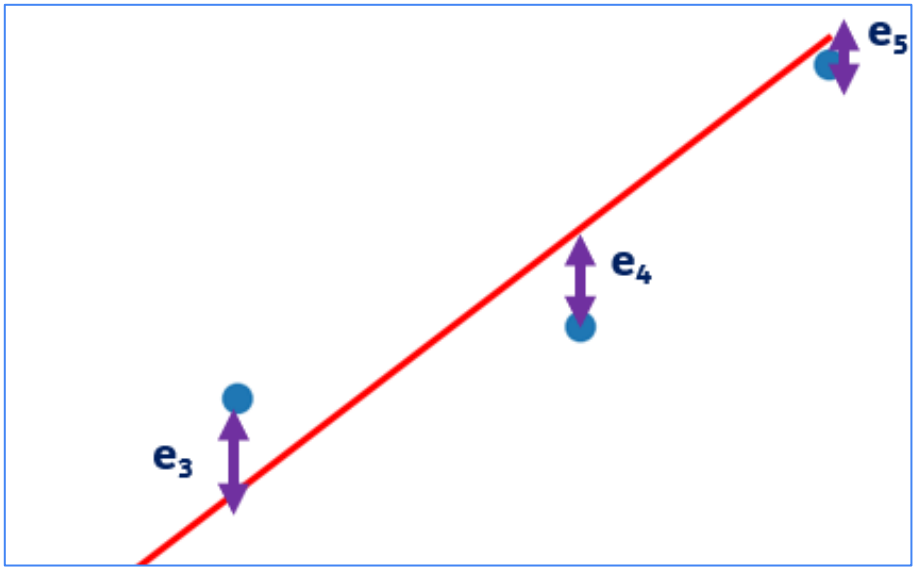
Biru : inisiasi, Merah : Hasil segemntasi, Hijau : Groundtruth

Evaluasi Model untuk Regresi

- *Model memprediksi suatu nilai konitnyu (bilangan real), bukan nilai diskrit (berupa kelas/label)*
- *Contoh: prediksi harga rumah, suhu maksimum, kekuatan gempa, harga saham*
- *Error merupakan selisi dari nilai aktual dengna nilai prediksi (real)*



zoom



Evaluasi Model untuk Regresi

Contoh pengukuran model untuk regresi:

a. *Mean Absolute Error (MAE)*
b. *Relative Absolute Error (RAE)*
c. *Mean Squared Error (MSE)*
d. *Relative Squared Error (RSE)*
e. *Root Mean Squared Error (RMSE)*
f. *Mean Absolute Percntage Error (MAPE)*
g. *Mean Percntage Error (MPE)*
h. *R- squared*

MAE dan RMSE akan diulas di slide berikut

Mean Absolute Error (MAE)

- *Ide : Setiap selisih error diambil nilai mutlaknya untuk selanjutnya dijumlahkan (**Diskusi:** mengapa nilai mutlak?)*
- *Jumlah nilai mutlak semua error di bagi rata dengan banyaknya sampel sehingga diperoleh nilai rata rata error, karenanya disebut Mean Absolute Error, dinyatakan sebagai:*

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n} |y_j - \hat{y}_j|$$

- *n : banyak sample*
- y_j *: nilai aktual untuk sample j*
- $\hat{y}_j$ *nilai prediksi untuk sample j*

Root Mean Squared Error (RMSE)

- *Ide : Setiap selisih error diambil nilai kuadratnya untuk selanjutnya dijumlahkan*
- *Jumlah nilai kuadrat setiap error di bagi rata dengan banyaknya sampel sehingga diperoleh nilai rata rata kuadrat error, untuk kemudian ditarik nilai akrnya, karenanya disebut Root Mean Squared Error, dinyatakan sebagai:*
  - *n : banyak sample*
  - y_j *: nilai aktual untuk sample j*
  - $\hat{y}_j$ *nilai prediksi untuk sample j*

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

- *RMSE memiliki fungsi kuadratik yang bersifat kontinu dan dapat diturunkan (dieferensiable) dan menguntungkan untuk optimasi (**Diskusi**: Mengapa?).*
- *Lebih sensitive terhadap pencilan (outlier), **Diskusi**: Mengapa?*

Perbandingan secara umum

Acroynm	Full Name	Residual Operation?	Robust To Outliers?
MAE	Mean Absolute Error	Absolute Value	Yes
MSE	Mean Squared Error	Square	No
RMSE	Root Mean Squared Error	Square	No
MAPE	Mean Absolute Percentage Error	Absolute Value	Yes
MPE	Mean Percentage Error	N/A	Yes

Pengukurkan Performansi pada Clustering

- *Termasuk unspervised learning, data tidak berlabel (tidak ada kelas)*
- *Tujuan : mengelompokkan data yang mirip sedekat mungkin dan memisahkan data yang tidak mirip sejauh mungkin*
- *Contoh pengukuran performansi untuk clustering :*
  - *Silhouette Coefficient*
  - *Rand Index*
  - *Mutual Information*
  - *Calinski-Harabasz Index (C-H Index)*
  - *Davies-Bouldin Index*
  - *Dunn Index*

Silhouette Coefficient

- *Silhouette Coefficient dinyatakan sebagai*

$$s = \frac{b - a}{max(a, b)}$$

s: silhouette Coefficient

a: rata-rata jarak sebuah sampel dengan sampel lainnya di cluster yang sama

b: rata-rata jarak sebuah sampel dengan sampel lainnya di cluster tetangga terdekat

- *Nilai berada diantara -1 dan +1*
- *Nilai -1 mengindikasikan clustering yang tidak tepat, disekitar 0 mengindikasikan adanya overlaping clustering, dan +1 mengindikasikan clustering yang padat dan terpisah dengan baik*

Membandingkan Model

- *Model yang telah dibangun diharapkan memiliki akurasi yang lebih baik.*
- *Contoh:*

Dua buah, yaitu model 1 dan model 2, dilakukan pengujian terhadap 10 data dengan hasil seperti di samping. Dengan menghitung ratio jumlah percobaan valid terhadap jumlah percobaan diperoleh:

Data Uji	Model 1	Model 2
1	valid	tidak valid
2	tidak valid	tidak valid
3	valid	valid
4	valid	valid
5	tidak valid	tidak valid
6	valid	valid
7	valid	tidak valid
8	tidak valid	tidak valid
9	valid	valid
10	tidak valid	valid

Akurasi model 1 : 6/10= 60%

Akurasi model 2 : 5/10 = 50%

Membandingkan Model

- *Akurasi model yang lebih tinggi belum cukup untuk dapat diklaim bahwa model tersebut **secara statisik** signifikan berbeda (dan lebih baik) dari dari model lainnya.*

- *Untuk mendukug klaim bahwa model 1 lebih baik dari model 2 perlu pengujian secara statistik dengan membuat dua hipotesa yang berlawanan:*

  H_0 : Kedua model memiliki akurasi yang sama

  H_1: Kedua model memiliki akurasi yang berbeda

- *Pengujian statistik yang sederhana dapat dilakukan dengan McNemar's Test*
- *Untuk pengujian lainnya yang lebih detil (5 cv test dsb.) silakan dilanjutkan ke pengayaan.*

McNemar's Test

- *Data pengujian di susun menjadi tabel contingency seagai berikut (perhatikan pasangan dalam model 1/model 2)*

	Model 2 valid	Model 2 tidak valid
Model 1 valid	valid/valid	valid/tidak valid
Model 1 tidak valid	tidak valid/valid	tidak valid/tidak valid

- *Sehingga dari tabel sebelumnya diperoleh :*

	Model 2 valid	Model 2 tidak valid
Model 1 valid	4	2
Model 1 tidak valid	1	3

McNemar's Test

- *McNemar's test statistic dihitung dengan*

- *S = **(valid/tidak valid - tidak valid/valid)^2 / (valid / tidak valid + tidak valid/valid)***

- *Hal penting dari S diatas adalah klaim statistik konsen kepada perbedaan valid dan tidak valid pda kedua model, bukan pada akurasi maupun tingkat error*

- *Melakui perhitungan statisik lebih lanjut, perlu memperhatikan masing masing nilai dalam tabel contingency. Distribusi χ^2 mengasumsikan nilai nilai lbesar untuk nilai elemen-elemen tabel contingency. Untuk nilai kecil, digunakan distribusi Binomial. Dalam praktikal, nilai S diatas dilakukan koreksi. Perhitungan detil statistik ini dapat dibaca di referensi.*

McNemar's Test

- *McNemar's test statistic dihitung dengan*

- *S = **(valid/tidak valid - tidak valid/valid)^2 / (valid / tidak valid + tidak valid/valid)***

- *Hal penting dari S diatas adalah klaim statistik konsen kepada perbedaan valid dan tidak valid pda kedua model, bukan pada akurasi maupun tingkat error*

- *Melakui perhitungan statisik lebih lanjut, perlu memperhatikan masing masing nilai dalam tabel contingency. Distribusi χ^2 mengasumsikan nilai nilai lbesar untuk nilai elemen-elemen tabel contingency. Untuk nilai kecil, digunakan distribusi Binomial. Dalam praktikal, nilai S diatas dilakukan koreksi. Perhitungan detil statistik ini dapat dibaca di referensi.*

Parameter penting dalam McNemar's Test

- *Parameter dalam McNemar's Test, selain s adalah p*
- *Dalam penggunaan praktikal, dapat digunakan perintah (dalam python) untuk mendapatkan dua nilai ini, dengan memperhatikan apakah nilai elemen tabel contingency besar atau kecil*
- *Contoh : dari table contingency sebelumnya, dapat dituliskan:*

  T = [[4,2],
  [1,3]]

- *Untuk case nilai-nilai kecil (misalnya tabel contingency T di atas), dapat digunakan perintah:*

```
s, p = mcnemar(T, exact=True)
```

- *Parameter lain adalah ambang batas p untuk threshold, yaitu α, misalnya $\alpha = 0.05$*

Penolakan / Penerimaan hipotesa

- *Perdasarkan nilai p dan ambang α dapat ditentukan:*

- *Jika p > α , hipotesa H_0 gagal untuk ditolak, kedua model scara statistik tidak ada perbedaan*
- *Jika p $\leq$ α , hipotesa H_0 ditolak, kedua model scara statistik secara signifikan ada perbedaan*

- *McNemar's adalah pengujian yang sederhana dan telah berkembang diantaranya 5xcv t-test beserta pengembangannya. Detil teori pengujian ini dapat dilihat di referensi.*

Source Code

```
# Contoh sederhana mcnemar test
from statsmodels.stats.contingency_tables import mcnemar
# Asumsi tabel contingency sudah tersedia
conti = [[4,2],
     [1, 3]]
# Perhitungan mcnemar test dilakukan dengan fungsi mcnemar
retVal = mcnemar(conti, exact=True)
# menampilkan nilai statistic dan p value
print('Nilai statistic =%.3f, \nNilai p-value =%.3f' % (retVal.statistic, retVal.pvalue))
# Pengecekan nilai p-value, dengan mengambil sebuah nilai alpha
alpha = 0.01
if retVal.pvalue > alpha:
  print('Hipotesis H0 gagal ditolak, kedua model memiliki peluang eror yang sama)')
else:
  print('Hipotesis H0 ditolak, kedua model memiliki peluang eror yang berbeda’)
```

Output:

```
Nilai statistic =1.000, Nilai p-value =1.000 Hipotesis H0 gagal ditolak, kedua model memiliki peluang eror yang sama)
```

Sources

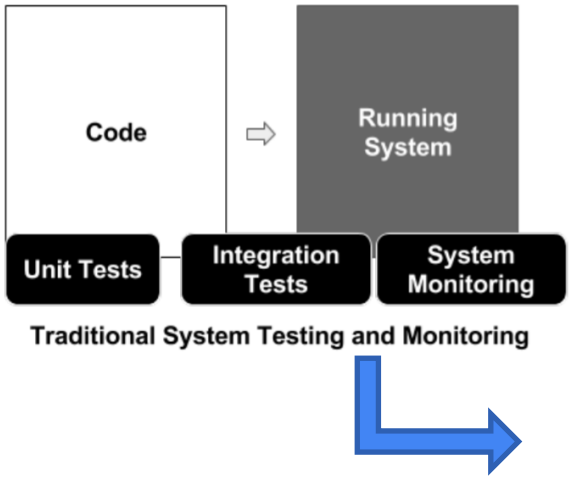
- *Klasifikasi*
- *Regresi:*
- https://www.dataquest.io/blog/understanding-regression-error-metrics/
- *Klasifikasi (ROC)*
- *https://www.youtube.com/watch?v=z5qA9qZMyw0*
- *https://www.youtube.com/watch?v=4jRBRDbJemM*
- *https://www.youtube.com/watch?v=4jRBRDbJemM&t=349s*
- *Regresi*
- *Clustering*
- *Evaluasi Model (Mc Nemar test dll.)*
- *https://sebastianraschka.com/pdf/lecture-notes/stat479fs18/11_eval-algo_slides.pdf*
- *https://www.youtube.com/watch?v=z5qA9qZMyw0*
- *https://machinelearningmastery.com/mcnemars-test-for-machine-learning/*
- *Thomas G. Dietterich; Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Comput 1998; 10 (7): 1895–1923. doi: https://doi.org/10.1162/089976698300017197*

Learning Objective

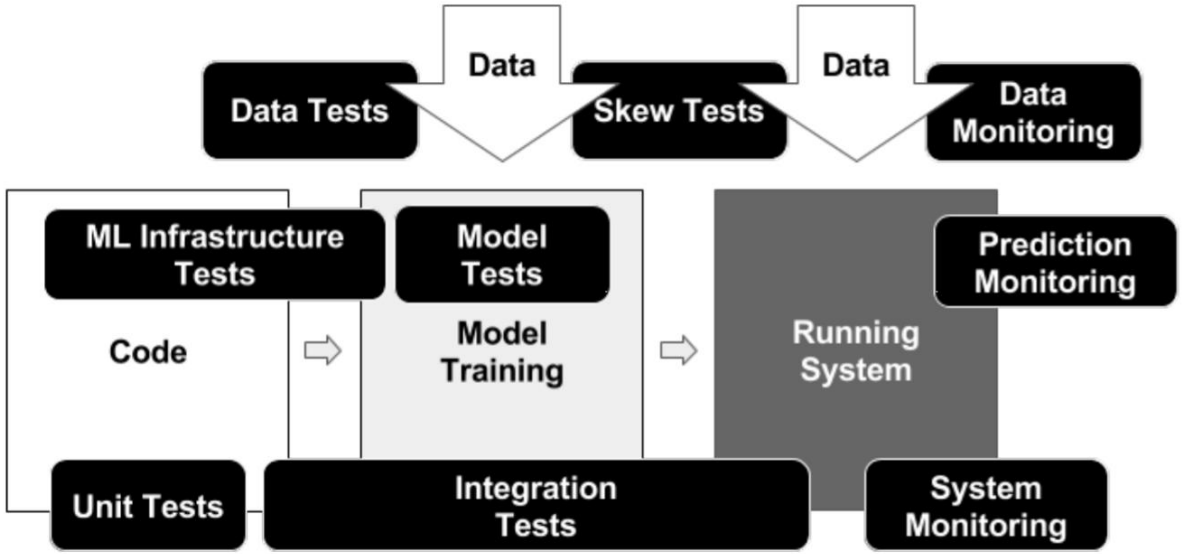
Rencana Pembelajaran		
1	**Pertemuan Ke**	15
2	**Topik**	**Membangun Model: Evaluasi** **Unit Kompetensi:** **1. J.62DMI00.014.1 - Mengevaluasi Hasil Pemodelan** **2. J.62DMI00.015.1 - Melakukan Proses Review Pemodelan**
3	**Deskripsi Topik**	**1. J.62DMI00.014.1 - Mengevaluasi Hasil Pemodelan** **a. Menggunakan model dengan data riil** **b. Menilai hasil pemodelan** **c. Membandingkan hasil pemodelan menggunakan hypothesis testing** **2. J.62DMI00.015.1 - Melakukan Proses Review Pemodelan** **a. Menilai kesesuaian proses pemodelan** **b. Menilai kualitas proses pemodelan**

Keunikan sistem berbasis ML

- *Dipengaruhi data yang dinamis*
- *Dipengaruhi konfigurasimodel*



Traditional System Testing and Monitoring



ML-Based System Testing and Monitoring

Google ML Test Score

Google ML Test Score menguji sistem berbasis ML melalui 28 kriteria yang secara umum dikelompokkan menjadi 4 :

- *Memelihara semua fitur dalam skema, hanya menyimpan fitur yang penting dan tidak terlalu rumit, dapat digunakan tanpa melanggar privasi atau peraturan yang berlaku.*
- *Membuat model dalam lingkungan yang tercatat perkembangannya, mengoptimalkan parameter model, dan melakukan pemeriksaan rutin terhadap model dasar*
- *Membangun pipeline ML terintegrasi yang dapat didebug dengan mudah dan diuji sebelum diimplementasikan ke sistem produksi (setiap penambahan disertai alternatif rollback).*
- *Memantau ketidaktersediaan atau perubahan data input, inkonsistensi antara sub-bagian training dan scoring, penurunan kualitas statistik model, atau kecepatan keseluruhan sistem.*

Data

1. Feature expectations are captured in a schema.
2. All features are beneficial.
3. No feature's cost is too much.
4. Features adhere to meta-level requirements.
5. The data pipeline has appropriate privacy controls.
6. New features can be added quickly.
7. All input feature code is tested.

Model

1. Every model specification undergoes a code review and is checked in to a repository.
2. Offline proxy metrics correlate with actual online impact metrics.
3. All hyperparameters have been tuned.
4. The impact of model staleness is known.
5. A simpler model is not better.
6. Model quality is sufficient on all important data slices.
7. The model has been tested for considerations of inclusion.

Infra

1. Training is reproducible.
2. Model specification code is unit tested.
3. The full ML pipeline is integration tested.
4. Model quality is validated before attempting to serve it.
5. The model allows debugging by observing the step-by-step computation of training or inference on a single example.
6. Models are tested via a canary process before they enter production serving environments.
7. Models can be quickly and safely rolled back to a previous serving version.

Monitor

1. Dependency changes result in notification.
2. Data invariants hold in training and serving inputs.
3. Training and serving features compute the same values.
4. Models are not too stale.
5. The model is numerically stable.
6. The model has not experienced a dramatic or slow-leak regressions in training speed, serving latency, throughput, or RAM usage.
7. The model has not experienced a regression in prediction quality on served data.

Sources

- https://static.googleusercontent.com/media/research.google.com/id//pubs/archive/aad9f93b86b7addfea4c419b9100c6cdd26cacea.pdf
- https://www.kaggle.com/discussion/217946
- https://medium.com/@rasmi/the-ml-production-readiness-of-teslas-autopilot-80acd03b3089
- https://ckaestne.github.io/seai/S2020/slides/13_infrastructurequality/infrastructurequality.pdf
- https://blog.dataiku.com/the-google-ml-test-score-measuring-your-sust-ai-nability

#JADIJAGOANDIGITAL

TERIMA KASIH

digitalent.kominfo

DTS_kominfo

digitalent.kominfo

digital talent scholarship