



# MODUL DATA MINING

## Evaluation Classification



Pada modul ini dijelaskan mengenai jenis-jenis metric evaluasi klasifikasi dengan menggunakan bahasa pemrograman python.  
Diharapkan setelah mempelajari modul ini, mahasiswa mampu memahami jenis-jenis metric evaluasi dan penggunaannya dalam mengevaluasi performa model klasifikasi.

EPS  
5

## EVALUASI KLASIFIKASI

Untuk melakukan evaluasi performa model klasifikasi, dapat digunakan berbagai macam metric evaluasi. Beberapa metric evaluasi tersebut dikenalkan pada modul ini. Selain itu, pada modul ini juga dikenalkan beberapa jenis pembagian data.

### BAGIAN 1 – Jenis Pembagian Data Hold-out

Masih melanjutkan modul sebelumnya. Bagian ini dilaksanakan pada pembuatan model dan evaluasi model.

1. Import library untuk membagi data menggunakan metode hold-out

```
from sklearn.model_selection import train_test_split
```

2. Masukkan data sesuai atribut ke X -> atribut predictor dan y -> atribut target

```
X = data.drop('Survived', axis=1)  
y = data.Survived
```

3. Bagi data menjadi data training dan data testing.

Data training sebesar 70% dan sisanya adalah data testing.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=27)
```

4. Membuat model klasifikasi dengan algoritma KNN.

Import library KNN.

```
from sklearn.neighbors import KNeighborsClassifier
```

5. Buat model klasifikasi dengan algoritma KNN dengan data training.

```
# Train model  
KNN = KNeighborsClassifier().fit(X_train, y_train)
```

6. Prediksi data testing dengan model yang telah dibuat.

```
# Predict on training set  
KNN_pred = KNN.predict(X_test)
```

### Evaluasi Model – Confusion matrix (akurasi, precision, recall, F1 score)

7. Import library sklearn yang digunakan untuk metric evaluasi

```
from sklearn.metrics import classification_report, confusion_matrix
```

8. Hitung dan tampilkan confusion matrix.

**Cara 1:**

```
print(confusion_matrix(y_test, KNN_pred))
```

```
[[145  31]
 [ 30  61]]
```

Dengan kolom adalah predicted, dan baris adalah actual.  
Kolom [0] adalah kelas 0, dan kolom [1] adalah kelas 1. Begitupula untuk baris.

#### Cara 2:

```
pd.crosstab(y_test, KNN_pred, rownames=['Actual'], colnames=['Predicted'], margins=True)
```

Predicted	0	1	All
Actual			
0	145	31	176
1	30	61	91
All	175	92	267

#### CARI TAU:

Total data sejumlah 800 an, namun mengapa total data pada confusion matrik hanya 260 an?

9. Hitung nilai akurasi, precision, recall, dan F1 score

```
print(classification_report(y_test, KNN_pred))
```

	precision	recall	f1-score	support
0	0.83	0.82	0.83	176
1	0.66	0.67	0.67	91
accuracy			0.77	267
macro avg	0.75	0.75	0.75	267
weighted avg	0.77	0.77	0.77	267

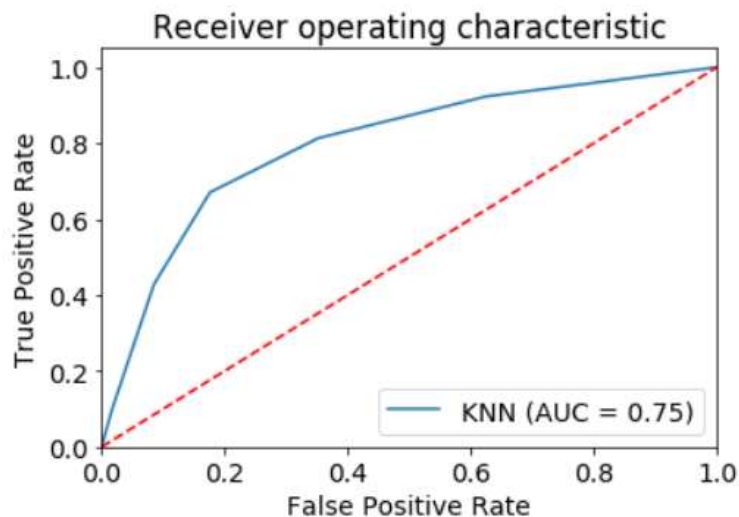
#### Evaluasi Model – ROC AUC

10. Import library.

```
from sklearn.metrics import roc_auc_score, roc_curve
import matplotlib.pyplot as plt
plt.rc("font", size=14)
```

11. Hitung FPR, TPR, dan nilai AUC

```
KNN_roc_auc = roc_auc_score(y_test, KNN.predict(X_test))
knn_fpr, knn_tpr, knn_thresholds = roc_curve(y_test, KNN.predict_proba(X_test)[: ,1])
plt.figure()
plt.plot(knn_fpr, knn_tpr, label='KNN (AUC = %0.2f)' % KNN_roc_auc)
plt.plot([0, 1], [0, 1], 'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.show()
```



### TUGAS MAHASISWA 1

1. Silahkan coba dengan 2 algoritma lain, kemudian lakukan evaluasi terhadap model tersebut (akurasi, precision, recall, dan F1 score serta ROC AUC)
2. Bandingkan evaluasi dari 2 model yang dihasilkan pada soal nomor 1. Jelaskan jawabanmu.

### BAGIAN 2 – Jenis Pembagian Data Cross-Val

Bagian ini melanjutkan bagian 1.

1. Import library

```
from sklearn.model_selection import cross_val_score
```

2. Definisikan algoritma yang digunakan

```
KNN = KNeighborsClassifier()
```

3. Pecah data menggunakan metode cross validation dengan k = 5, sekaligus hitung score akurasi.

```
scores_accuracy = cross_val_score(KNN, X, y, cv=5, scoring="accuracy")

print(scores_accuracy)
print("Rata-Rata Nilai Akurasi: %0.2f (+/- %0.2f)" % (scores_accuracy.mean(), scores_accuracy.std()))

[0.73595506 0.73033708 0.80898876 0.75706215 0.76271186]
Rata-Rata Nilai Akurasi: 0.76 (+/- 0.03)
```

#### 4. Hitung score Precision Macro

```
scores_precision = cross_val_score(KNN, X, y, cv=5, scoring="precision_macro")
print(scores_precision)
print("Rata-Rata Nilai Precision Macro: %0.2f (+/- %0.2f)" % (scores_precision.mean(), scores_precision.std()))

[0.7202501 0.71468927 0.79897436 0.74394319 0.75149573]
Rata-Rata Nilai Precision Macro: 0.75 (+/- 0.03)
```

#### 5. Hitung score Recall Macro

```
scores_recall = cross_val_score(KNN, X, y, cv=5, scoring="recall_macro")
print(scores_recall)
print("Rata-Rata Nilai Recall Macro: %0.2f (+/- %0.2f)" % (scores_recall.mean(), scores_recall.std()))

[0.71898396 0.70320856 0.81176471 0.73637345 0.73819482]
Rata-Rata Nilai Recall Macro: 0.74 (+/- 0.04)
```

#### 6. Hitung score F1 Macro

```
scores_f1 = cross_val_score(KNN, X, y, cv=5, scoring="f1_macro")
print(scores_f1)
print("Rata-Rata Nilai f1 Macro: %0.2f (+/- %0.2f)" % (scores_f1.mean(), scores_f1.std()))

[0.71959108 0.70723684 0.80260926 0.73946531 0.74301715]
Rata-Rata Nilai f1 Macro: 0.74 (+/- 0.03)
```

Catatan:

Cross-validation memiliki banyak variasi, variasinya dapat dibaca pada link berikut:

[https://scikit-learn.org/stable/modules/cross\\_validation.html](https://scikit-learn.org/stable/modules/cross_validation.html)

#### CARI TAHU:

Jika menggunakan cross validation, maka kita tidak dapat menampilkan confusion matrix dengan library SKLEARN seperti yang kita lakukan ketika membagi data dengan cara hold-out. Mengapa?

#### TUGAS MAHASISWA 2

1. Silahkan coba metode pembagian dengan crossval (nilai k menyesuaikan) dengan menggunakan algoritma-algoritma yang digunakan di TUGAS 1. Kemudian hitung nilai akurasi, precision, recall, dan F1 scorenya.
2. Bandingkan evaluasi dari 2 model yang dihasilkan pada soal nomor 1 TUGAS 2. Jelaskan jawabanmu.
3. Bandingkan evaluasi dari model-model yang dihasilkan pada soal nomor 1 TUGAS 2 dengan evaluasi model dari TUGAS 1. Jelaskan jawabanmu.

#### CATATAN TUGAS

1. Tugas dikumpulkan dalam bentuk laporan tertulis
2. Dikumpulkan paling lambat Selasa, 27 Oktober 2020, pukul 23.59 WIB via classroom