

Nama : Muhammad Rafi Ramadhan

Program : Python for Data Science

Memahami Machine Learning – by Raka Ardhi

Memberikan Kemampuan Kepada Komputer untuk Belajar.

A. Prologue

Tahapan mempersiapkan data: mendapatkan data (*load, clean, inspect data*, dan merapihkan data), memilih algoritma, melatih model (*training*), serta uji akurasi (evaluasi model naïve bayes, peningkatan performa, *crossing validation*).

Apa itu machine learning? Sebuah aplikasi dari AI yang memprovide system kemampuan untuk belajar secara otomatis dan berkembang dari experience. Jadi akan terus berkembang tanpa kita mengoding. Contoh: saat buka email dibagian spam, pembelian user berdasarkan patren pencarian, autonomus driving.

Membangun model dari contoh input untuk membuat data-driven prediction.

Perbedaan *traditional programming* dengan *machine learning*: *traditional programming* butuh logic code (ex: *if, case, while/for*), sedangkan *machine learning* mengambil data yang bisa di proses dan memasukkannya ke algoritma untuk dianalisis sehingga akan menciptakan suatu model.

Tipe *machine learning*: *Supervised (Value Prediction*, perlu hasil prediksi, model prediksi data baru) dan *Unsupervised (Data clustering*, tidak perlu cluster, akses data cluster).

B. Mempersiapkan Data

Yang digunakan / instalasi: Jupyter notebook, python 3, anaconda

Workflow : Permasalahan > Persiapan data > Memilih algoritma > Melatih model > Uji model

Tidy data : Lebih mudah dimanipulasi, Variabelnya adalah kolom, dan Observasinya adalah baris.

Mendapatkan data bisa dari google, data pemerintahan, ataupun data perusahaan. Sebelum data digunakan perlu dilakukan *load, clean*, dan *inspect data*.

Load data `df = pd.read_csv("./data/kodeML.csv")`

Mengecek apakah ada data yang kosong `df.isnull().values.any()`

C. Memilih Algoritma

Faktor : Learning type, Result (Tipe: *Regression (continuous value)* dan *Classification (discret value)*), Complexity, Basic vs Enhanced.

Kandidat algoritma: *Naive bayes* (belajar dari data sebelumnya lalu memprediksi data yang terdapat celah), *logistic regression* (hasilnya biner 1 or 0), dan *Decision Tree* (seperti if/else).

Kenapa naive bayes? Probability based, bobot fitur sama, membutuhkan sedikit data.

D. Melatih Model

Model training yaitu membiarkan data yang spesifik untuk melatih algoritma machine learning (disini naive bayes) untuk menghasilkan model yang spesifik sesuai yang diinginkan.

Proses training: Split data > Train model > Evaluasi model.

```
From sklearn.model_selection import train_test_split
From sklearn.naive_bayes import GaussianNB
```

E. Uji Akurasi

Macam-macam uji akurasi:

1. Performance

```
From sklearn import metrics
Print("Accuracy: {0:4f}".format(metrics.accuracy_score(y_train, nb_predict_train)))
Print("{0}".format(metrics.confusion_matrix(y_test, nb_predict_test)))
Print(metrics.classification_report(y_test, nb_predict_test))
```

Confusion Matrix

```
[[118  33]
 [ 28  52]]
```

Classification Report

	precision	recall	f1-score	support
0	0.81	0.78	0.79	151
1	0.61	0.65	0.63	80
micro avg	0.74	0.74	0.74	231
macro avg	0.71	0.72	0.71	231
weighted avg	0.74	0.74	0.74	231

2. Peningkatan Performa

```
From sklearn.ensemble import RandomForestClassifier
From sklearn.linear_model import LogisticRegression
```

3. Cross Validation

```
From sklearn.linear_model import LogisticRegressionCV
```