# Ensemble Learning

## Dasar Teori

Dalam kategori Statistika dan Machine Learning, Teknik Ensemble Learning berupaya membuat kinerja model prediktif menjadi lebih baik dengan meningkatkan akurasinya. Ensemble Learning adalah proses yang menggunakan berbagai model machine learning (seperti pengklasifikasi, regresi, dan lain-lain) yang dibangun secara strategis untuk memecahkan masalah tertentu.

Sebagai contoh, ketika melatih sekelompok pengklasifikasi Decision Tree, masing-masing pada subset acak yang berbeda dari set pelatihan. Untuk membuat prediksi, hanya mendapatkan prediksi dari semua pohon individu, lalu memprediksi kelas yang mendapat nilai voting terbanyak. Suatu ensemble dari Decision Trees disebut sebagai Random Forest, dan meskipun sederhana, ini adalah salah satu algoritma Machine Learning paling kuat yang tersedia saat ini.

Pada modul ini, akan membahas metode Ensemble Learning yaitu *Voting Classifier, Bagging* dan *Boosting, dan akan* menerapkan model Random Forest.

### Voting Classifier

Misalkan melatih beberapa pengklasifikasi, masing-masing mencapai akurasi sekitar 80%. Ada kemungkinan memiliki classifier Regresi Logistik, classifier SVM, classifier Random forest, classifier K-Nearest Neighbors, dan ada model lain (Lihat pada Figure 6-1).

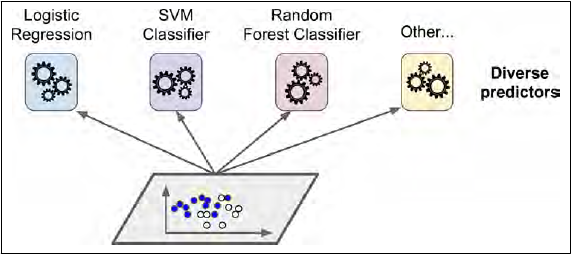


Figure 6‑1. Melatih aneka model klasifikasi

Cara yang sangat sederhana untuk membuat classifier yang lebih baik lagi adalah dengan mengumpulkan prediksi masing-masing classifier dan memprediksi kelas yang mendapat suara terbanyak. Pengelompokan suara terbanyak ini disebut pengelompokan *hard voting* (Lihat Figure 6-2).

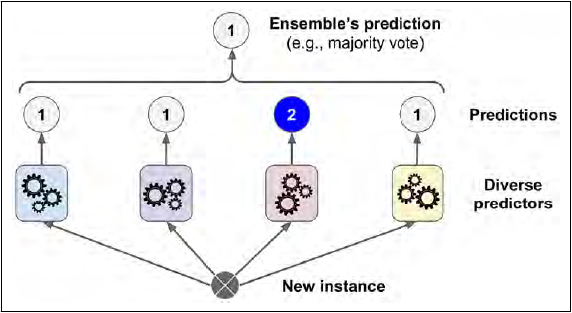


Figure 6‑2. *Hard Voting* Diagram

Sungguh menakjubkan, Voting Classifier ini sering mencapai akurasi yang lebih tinggi daripada penggolong terbaik dalam ensemble. Faktanya, bahkan jika masing-masing classifier adalah *weak-learner* (artinya hanya sedikit lebih baik daripada menebak secara random), ensemble masih dapat menjadi strong learner (mencapai akurasi tinggi), asalkan ada cukup banyak weak learner dan mereka cukup memadai berbeda.

### Bagging dan Pasting

Salah satu cara untuk mendapatkan beragam pengklasifikasi adalah dengan menggunakan algoritma training yang sangat berbeda, seperti yang baru saja dibahas sebelumnya. Metode Ensemble lain adalah dengan menggunakan algoritma training yang sama untuk setiap prediktor, tetapi untuk melatih mereka pada subset acak yang berbeda dari set pelatihan. Ketika pengambilan sampel dilakukan dengan penggantian, metode ini disebut *bagging* (Bootstrap Aggregating). Ketika pengambilan sampel dilakukan tanpa penggantian, maka akan disebut sebagai *pasting*.

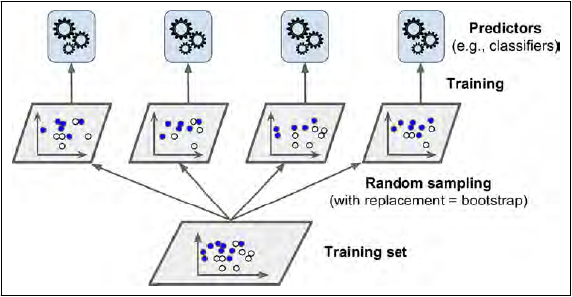


Figure 6‑3. Pasting/Bagging pada training set sampling dan training

### Boosting

Boosting (yang disebut sebagai hipotesis boosting) mengacu pada metode Ensemble apa pun yang dapat menggabungkan beberapa pelajar yang lemah menjadi pembelajar yang kuat. Gagasan umum dari sebagian besar metode peningkatan adalah untuk melatih prediktor secara berurutan, masing-masing mencoba untuk memperbaiki pendahulunya. Ada banyak metode peningkatan yang tersedia, tetapi sejauh ini yang paling populer adalah *AdaBoost*.

AdaBoost adalah iterative ensemble method. AdaBoost classifier membangun classifier yang kuat dengan menggabungkan beberapa pengklasifikasi berkinerja buruk sehingga Anda akan mendapatkan classifier kuat akurasi tinggi. Konsep dasar di balik Adaboost adalah mengatur bobot pengklasifikasi dan melatih sampel data dalam setiap iterasi sedemikian rupa sehingga memastikan prediksi akurat dari pengamatan yang tidak biasa. Algoritma pembelajaran mesin apa pun dapat digunakan sebagai klasifikasi dasar jika menerima bobot pada set pelatihan. Adaboost harus memenuhi dua syarat:

1. Pengklasifikasi harus dilatih secara interaktif pada berbagai contoh pelatihan yang diberi bobot.
2. Dalam setiap iterasi, ia mencoba untuk memberikan kesesuaian yang sangat baik untuk contoh-contoh ini dengan meminimalkan kesalahan pelatihan.

Dan untuk cara kerja AdaBoost sebagai berikut:

1. Awalnya, Adaboost memilih subset pelatihan secara acak.
2. Itu secara iteratif melatih model pembelajaran mesin AdaBoost dengan memilih set pelatihan berdasarkan prediksi akurat dari pelatihan terakhir.
3. Ini memberikan bobot yang lebih tinggi untuk pengamatan rahasia yang salah sehingga pada iterasi berikutnya pengamatan ini akan mendapatkan probabilitas tinggi untuk klasifikasi.
4. Juga, Ini memberikan bobot kepada classifier terlatih dalam setiap iterasi sesuai dengan akurasi classifier. Penggolong yang lebih akurat akan mendapatkan bobot tinggi.
5. Proses ini berulang hingga data pelatihan lengkap cocok tanpa kesalahan atau hingga mencapai jumlah maksimum penaksir.
6. Untuk mengklasifikasikan, lakukan "suara" di semua algoritma pembelajaran yang dibuat.

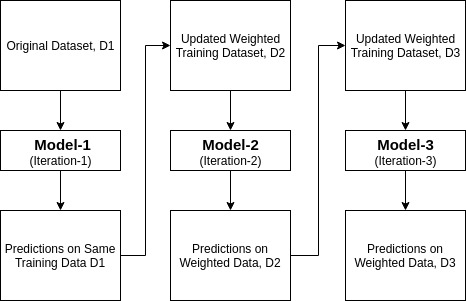


Figure 6‑4. Cara kerja AdaBoost

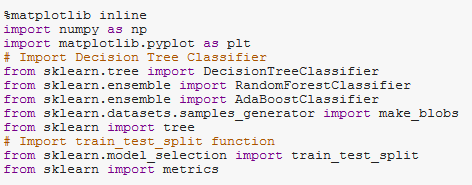
### Random Forest

Ensemble Learning di Decision Tree adalah Random Forest. Random forest melakukan metode *Bagging* secara internal. Random Forest menciptakan beberapa pohon, hingga ribuan, dan menghitung model terbaik untuk dataset yang diberikan. Tidak hanya mempertimbangkan semua fitur saat memisahkan node, algoritma Random Forest memilih fitur terbaik dari subset semua fitur. Sebagai contoh, terdapat dataset yang memiliki 16 fitur, saat mengimplementasikan model ini akan mengambil menggunakan rumus sqrt(jumlah\_fitur) sehingga menjadi 4 fitur yang akan dipakai secara random untuk memisah sebuah *node*. Teknik tersebut memperdagangkan bias yang lebih tinggi untuk varian yang lebih rendah, yang menghasilkan model yang jauh lebih baik. Random forest digunakan untuk Klasifikasi dan juga Regresi.

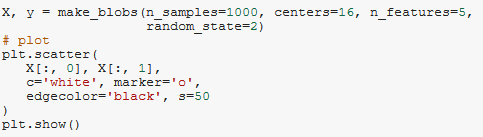
## Latihan

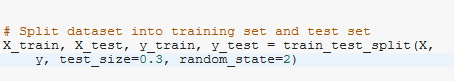
### Mempersiapkan Library dan datasets

##### Siapkan Library



##### Siapkan Dataset, dan split dataset



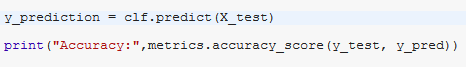


### Membangun Decision Tree dan Evaluasinya

##### Membangun Decision Tree



##### Mengecek akurasi

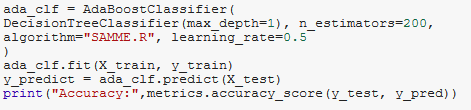


Dengan akurasi sekian, bagaimana hasil Akurasi Validasinya dengan import *cross\_validation\_score*

##### Cross Validation

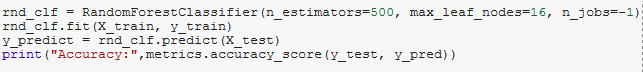


### Menerapkan AdaBoost



Selanjutnya terapkan cross\_validation dengan cv=5

### Menerapkan Random Forest



Selanjutnya terapkan cross\_validation dengan cv=5

## Tugas

1. Buat dataset artificial *binary class* dengan 20 baris.
2. Pilih satu algoritma ensemble untuk memprediksi dataset pada #1.
3. Buat penjelasan dalam bentuk video tutorial, step by step secara ilustratif.
4. Post video di 1 Youtube channel dan undang teman lain untuk view.
5. Deskripsi youtube video: nama Anggota dan NIM, url: elektro.um.ac.id