Nama : Zahrani Cahya Priesa
NIM : 1103223074
Mata Kuliah : Machine Learning

Analisa Bab 7 – Ensemble Learning & Random Forests

Ensemble Learning adalah pendekatan untuk meningkatkan performa dengan menggabungkan beberapa model (*learners*). Metode ini dapat menurunkan variance, meningkatkan akurasi, dan membuat prediksi lebih stabil.

1. Pendahuluan Ensemble Learning

Ensemble learning bekerja berdasarkan prinsip bahwa menggabungkan banyak model lebih baik daripada menggunakan satu model tunggal.

Tujuan utama ensemble:

- Mengurangi variance
- Meningkatkan akurasi
- Meningkatkan robustness model

2. Voting Classifiers

A. Hard Voting

Memilih kelas berdasarkan mayoritas suara dari beberapa model.

B. Soft Voting

Menggabungkan probabilitas setiap model, lalu memilih kelas dengan probabilitas rata-rata tertinggi.
Soft voting biasanya lebih akurat karena mempertimbangkan tingkat keyakinan model.

3. Bagging dan Pasting

Bagging (Bootstrap Aggregating):

- Sampling **dengan** replacement
- Mengurangi variance
- Meningkatkan stabilitas model

Pasting:

- Sampling **tanpa** replacement

Keduanya melatih banyak model secara paralel.

4. Random Forest

Random Forest adalah ensemble dari banyak Decision Tree dengan dua mekanisme:

1. Bootstrap sampling (bagging)
2. Random subset features saat split

Keunggulan Random Forest

- Cepat dan stabil
- Tidak mudah overfitting
- Memiliki Feature Importance
- Mendukung OOB evaluation

Feature Importance

Feature importance dihitung dari seberapa banyak impurity berkurang saat fitur digunakan untuk split.

5. Out-of-Bag (OOB) Evaluation

Dalam bagging, sekitar 36% data otomatis tidak terambil saat bootstrap.
Data ini digunakan sebagai **validation internal**, sehingga tidak memerlukan train/validation split manual.

Kelebihan:

- Hemat data
- Lebih stabil
- Cocok untuk dataset kecil

6. Boosting

Boosting melatih model secara **berurutan**, di mana model berikutnya fokus memperbaiki error dari model sebelumnya.

A. AdaBoost

- Memberi bobot lebih pada data yang salah diprediksi
- Umumnya menggunakan Decision Stump (tree depth = 1)

B. Gradient Boosting

- Tiap model baru belajar dari **residual error**
- Digunakan pada XGBoost, LightGBM, CatBoost

Boosting lebih akurat, tetapi rentan overfitting jika tidak di-tuning dengan benar.

7. Stacking (Stacked Generalization)

Stacking menggabungkan beberapa *base models* (level 0), lalu hasil prediksinya digunakan untuk melatih sebuah **meta-model** (level 1).

Tahapan:

1. Latih base model
2. Kumpulkan output base model
3. Latih meta-model menggunakan output tersebut

Stacking sangat fleksibel dan populer untuk kompetisi machine learning.

8. Intuisi Matematis Ensemble

Jika setiap model memiliki error ε dan terdapat N model independen, maka error ensemble:

$$\varepsilon_{ensemble} = \frac{\varepsilon}{N}$$

Semakin banyak model → semakin kecil error.
Namun, model harus **bervariasi**. Jika semua model identik → ensemble tidak berguna.

9. Perbandingan Metode Ensemble

Metode	**Karakteristik**	**Kapan Digunakan**
Voting	Kombinasi beberapa algoritma berbeda	Untuk baseline yang kuat
Bagging	Mengurangi variance	Data noisy & high variance
Random Forest	Bagging + random feature	Default pilihan pertama
AdaBoost	Fokus pada instance salah	Dataset kecil–sedang
Gradient Boosting	Belajar dari residual	Akurasi tinggi
Stacking	Meta-model	Proyek besar / kompetisi

10. Kesimpulan Bab 7

- Ensemble learning meningkatkan akurasi dan stabilitas
- Bagging mengurangi variance, sedangkan boosting fokus pada error

- Random Forest adalah model cepat dan stabil dengan interpretasi feature importance
- Boosting lebih kuat tetapi perlu tuning
- Stacking adalah metode paling fleksibel dan powerful