Gaussian Naive Bayes

Adinda Putri - 13523071

Gaussian Naive Bayes merupakan teknik klasifikasi yang mengasumsikan bahwa setiap fitur independen dan memiliki distribusi normal. Teknik ini mengklasifikasikan data dengan cara mencari nilai maksimum dari posterior probability masing-masing class (target) dan menetapkan data tersebut ke class dengan nilai maksimum.

Cara Kerja

1. Pisahkan data point berdasarkan class.
2. Hitung mean dan variance untuk setiap fitur
3. Hitung probabilitas masing-masing class
4. Untuk data point dengan fitur ke-i adalah xi, hitung probabilitasnya dengan rumus berikut untuk masing-masing class

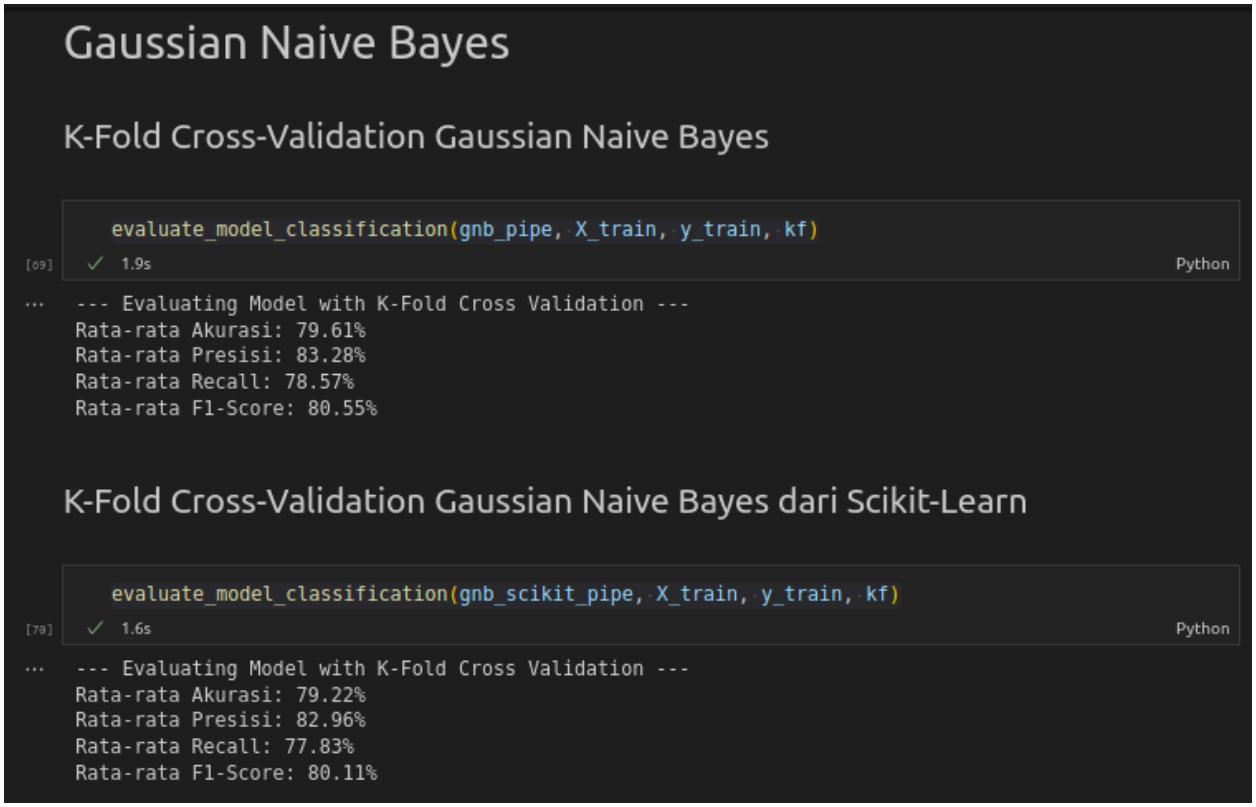
$$P(x_i|y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

5. Karena tiap fitur diasumsikan conditionally independent, maka posterior probability terjadinya class k untuk data point dengan fitur x1, x2, …, xn adalah sebagai berikut.

$$P(C = k|x_1, ..., x_n) \propto P(C = k) \prod_{i=1}^{n} P(x_i|C = k)$$

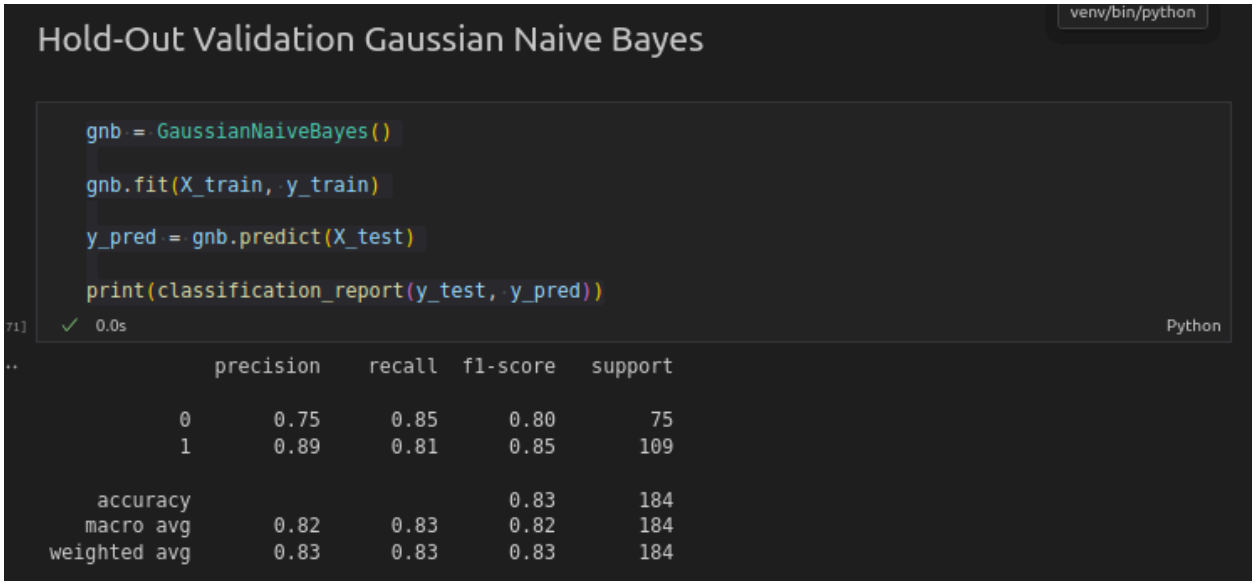
6. Lakukan prediksi dengan membandingkan posterior probability tiap class.
7. Klasifikasikan data point ke class dengan posterior probability paling tinggi.

Perbandingan model dari scratch dengan dari Scikit-Learn



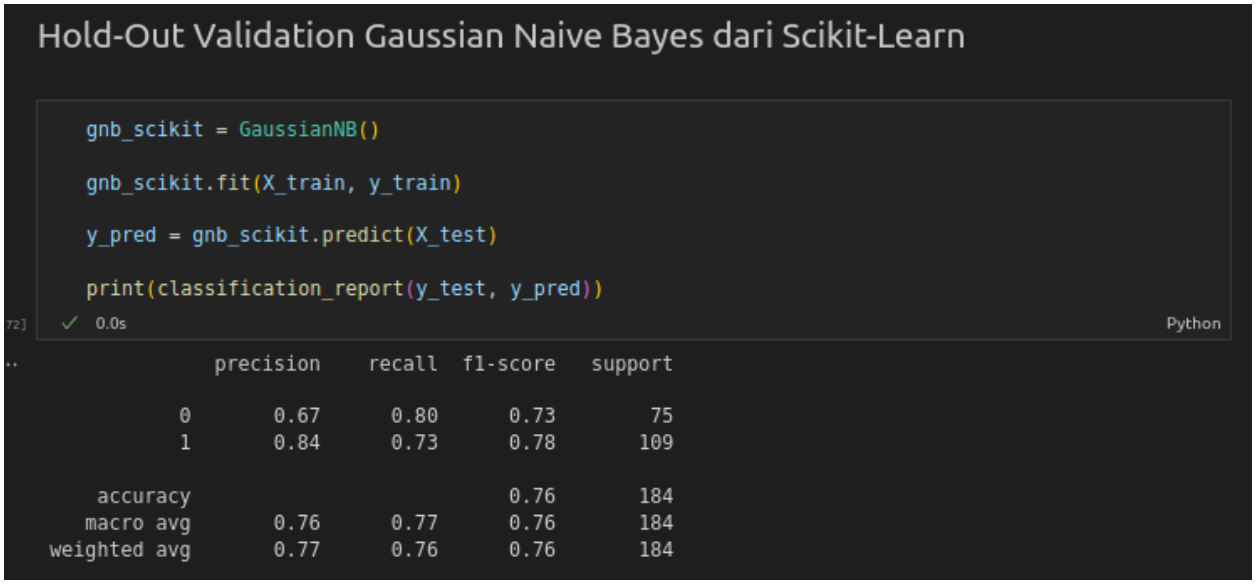
Gambar 1. K-Fold Cross Validation GNB dari Scratch dan GND dari Sklearn

Sumber: Penulis



Gambar 2. Hold-Out Validation GNB dari Scratch

Sumber: Penulis



Gambar 3. Hold-Out Validation GNB dari Sklearn

Sumber: Penulis

Hasil k-fold cross-validation antara model Gaussian Naive Bayes dari scratch dan Scikit-learn hampir identik. Selisih pada semua metrik (akurasi, presisi, recall, F1-score) sangat kecil, kurang dari 1%. Perbedaan minor ini dapat terjadi karena detail implementasi, yaitu cara penanganan fitur data kategorikal.

Namun, kesamaan hasil yang sangat tinggi ini membuktikan bahwa implementasi perhitungan probabilitas prior, likelihood (PDF Gaussian), dan posterior pada model dari scratch sudah tepat.

Sementara itu, berdasarkan hasil hold-out validation, model dari scratch memiliki performa yang lebih baik dengan selisih 7% akurasi. Kemungkinan alasannya sama, yaitu model dari scratch menangani data kategorikal secara khusus.

Ruang Improvement

- Mekanisme penangan data kategorikal masih bisa dioptimasi dengan mengombinasikan Gaussian Naive Bayes dengan Multinomial Naive Bayes
- Karena GNB bekerja baik untuk data dengan distribusi normal, maka handling outliers dapat meningkatkan kinerjanya. Dalam eksperimen ini, tidak dihandle data outliers secara khusus karena diasumsikan outliers tersebut merepresentasikan data ekstrem.

Referensi:

[1] *Gaussian Naive Bayes*, GeeksforGeeks, diperbarui 23 Juli 2025. [Daring]. Tersedia: https://www.geeksforgeeks.org/machine-learning/gaussian-naive-bayes/. [Diakses: 1 September 2025].

[2] *Gaussian Naive Bayes From Scratch*, GitHub (martian1231), 2025. [Daring]. Tersedia: https://github.com/martian1231/gaussianNaiveBayesFromScratch. [Diakses: 1 September 2025].