by Azzah M

KLASIFIKASI KESEGARAN IKAN MENGGUNAKAN MACHINE LEARNING

Machine Learning Project
June 2025

Isi Konten

1 Project Overview

2 Tools dan Dataset

3 Preprocessing

4 Modeling

5 Result

6 Conclusion

Project Overview

Kualitas ikan segar merupakan faktor penting dalam industri perikanan karena berkaitan langsung dengan kepuasan konsumen dan nilai jual produk. Untuk membantu proses identifikasi kualitas ikan secara lebih efisien, **proyek ini bertujuan membangun model klasifikasi untuk mengidentifikasi kondisi ikan (segar atau busuk) berdasarkan data sensor aroma, serta membandingkan performa beberapa algoritma machine learning.**

Dataset yang digunakan berisi 80 data ikan dengan 24 fitur numerik yang diperoleh dari dosen sebagai bagian dari pembelajaran klasifikasi praktis. **Model dikembangkan menggunakan algoritma SVM, Random Forest, KNN, dan Decision Tree** dengan evaluasi performa menggunakan confusion matrix dan akurasi untuk menentukan algoritma terbaik.

Proyek ini diharapkan menjadi langkah awal automasi klasifikasi kualitas ikan sehingga pelaku usaha perikanan dapat menjaga kualitas produk dengan lebih cepat dan akurat.

Tools dan Dataset

TOOLS

- Jupyter Lab
- Python
- Library pandas, numpy, matplotlib, seaborn, dan scikit-learn.

DATA SOURCE

- **Jumlah data:** 80 entry ikan dengan label segar dan busuk.
- **Fitur:** 24 fitur numerik dari 8 sensor aroma.
- **Asal dataset:** Dataset bersifat internal yang diperoleh dari dosen untuk keperluan pembelajaran klasifikasi kondisi ikan pada mata kuliah Machine Learning.

Preprocessing

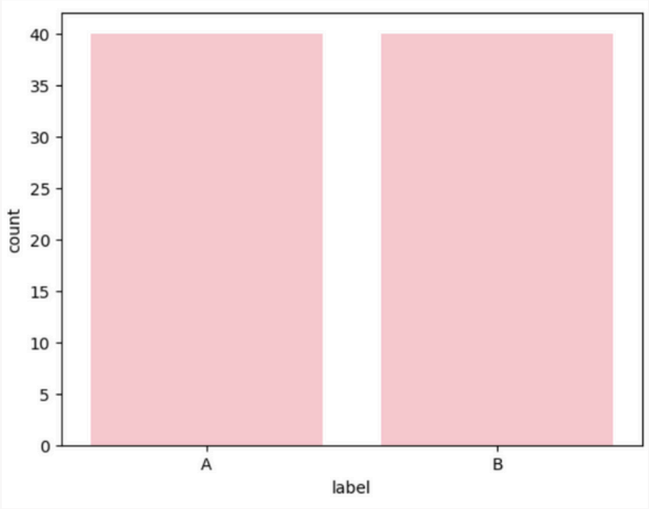
- Membuat dataframe dari file csv dengan `df = pd.read_csv('fitur_30s.csv')`
- Eksplorasi awal dengan `df.head()` untuk melihat isi dataset.
- EDA awal menggunakan `df.describe()` untuk melihat ringkasan statistik dataset.

  Tahap ini menghasilkan informasi:
  - Dataset berisi **80 data ikan dengan 24 fitur** aroma dari 8 sensor.
  - Setiap sensor memiliki 3 fitur:
    - **fn_mean**: Rata-rata respons sensor selama periode pengukuran yang memberikan informasi terkait kestabilan dan intensitas aroma selama waktu pengukuran.
    - **fn_max**: Nilai tertinggi respons sensor yang mendeteksi lonjakan aroma busuk selama periode pengukuran.
    - **fn_auc**: Luas area di bawah kurva respons sensor yang menggambarkan akumulasi intensitas aroma selama waktu pengukuran.

Preprocessing

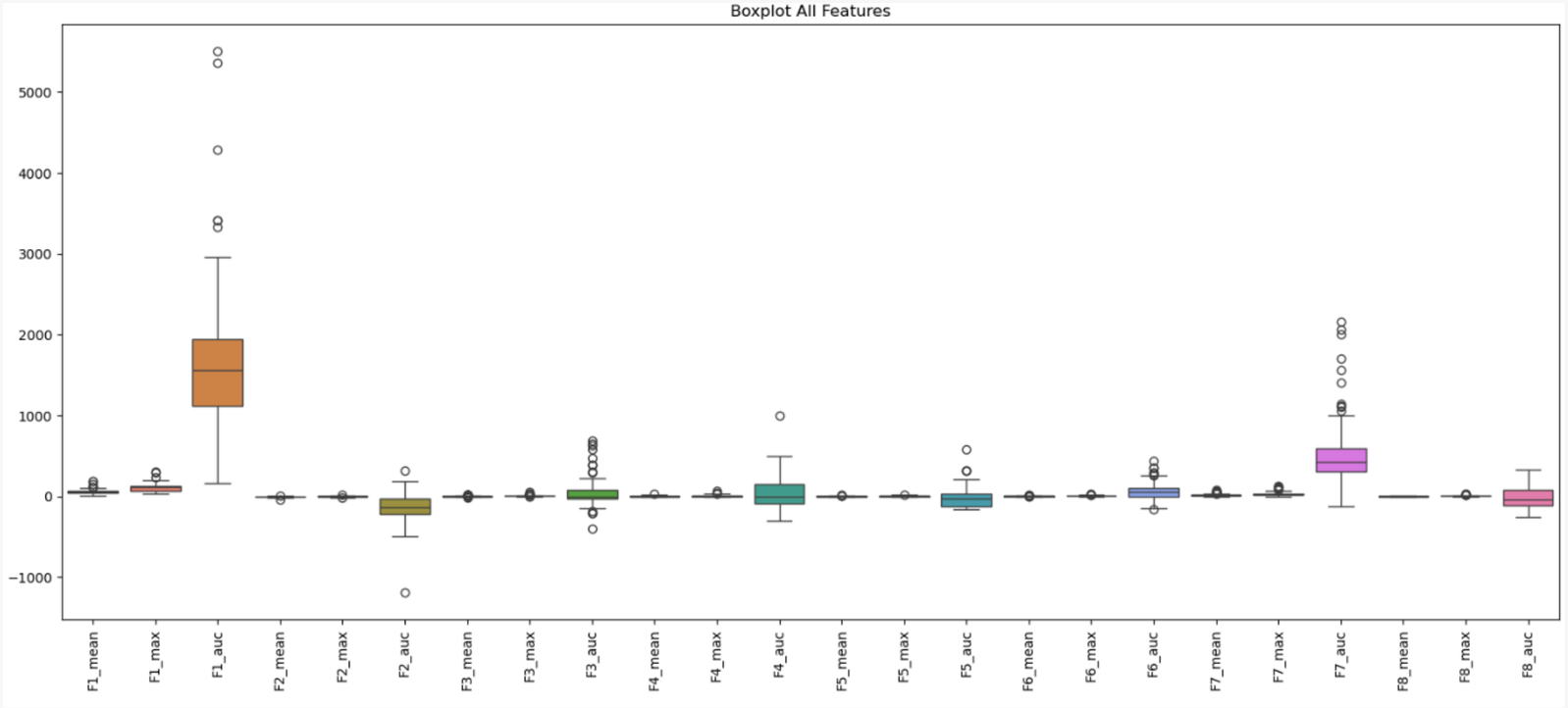
- Nilai antar fitur tidak seragam (rentang kecil hingga ribuan).
- Normalisasi diperlukan agar model tidak bias terhadap fitur berskala besar.
- Standar deviasi bervariasi yang menunjukkan persebaran data yang beragam.
- Terdapat indikasi outlier di beberapa fitur misalnya pada F7_auc yang memiliki mean 548 dan max 2152.
- Terdapat fitur dengan nilai negatif yang perlu diperhatikan.

- EDA lanjutan dengan:
  - Melihat distribusi kelas target. Dapat diketahui bahwa **distribusi kelas A dan B seimbang.**



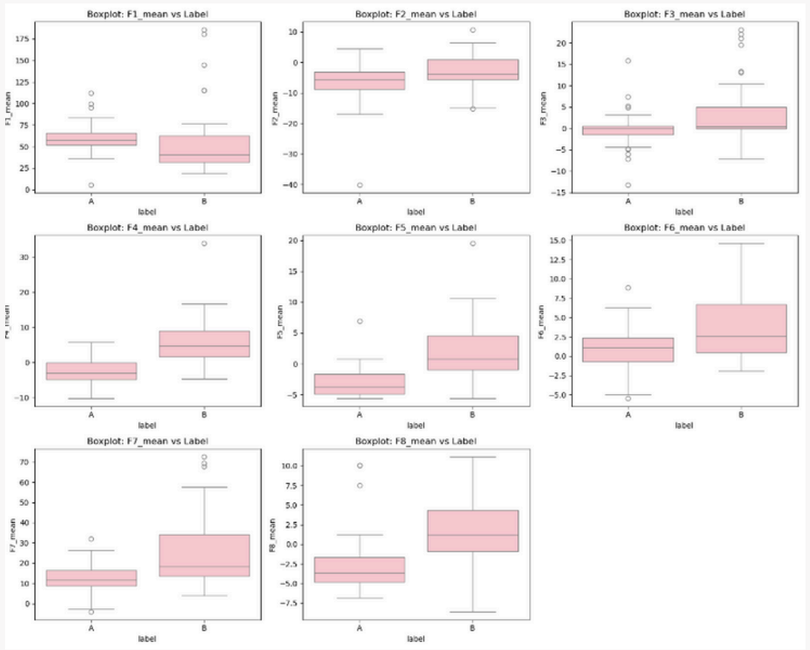
Preprocessing

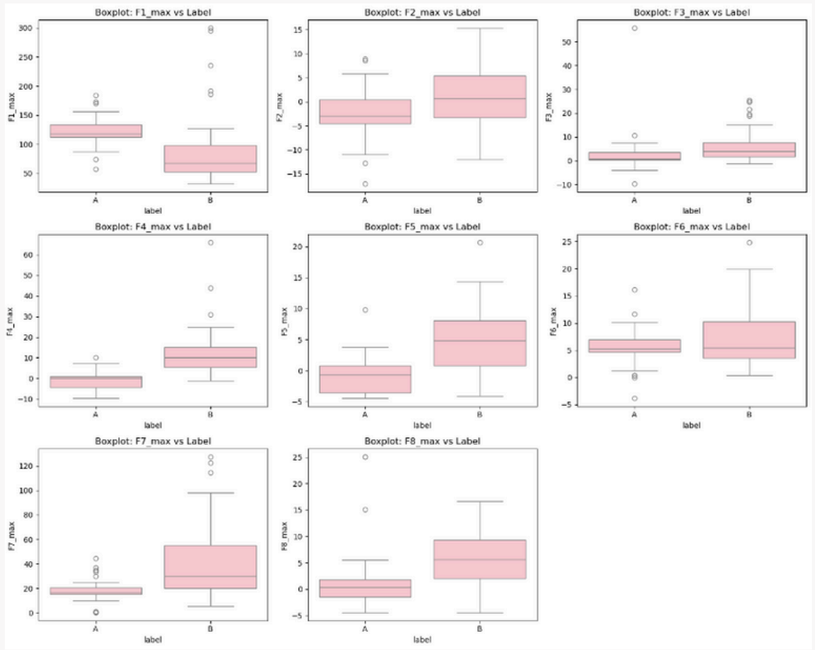
- Distribusi fitur

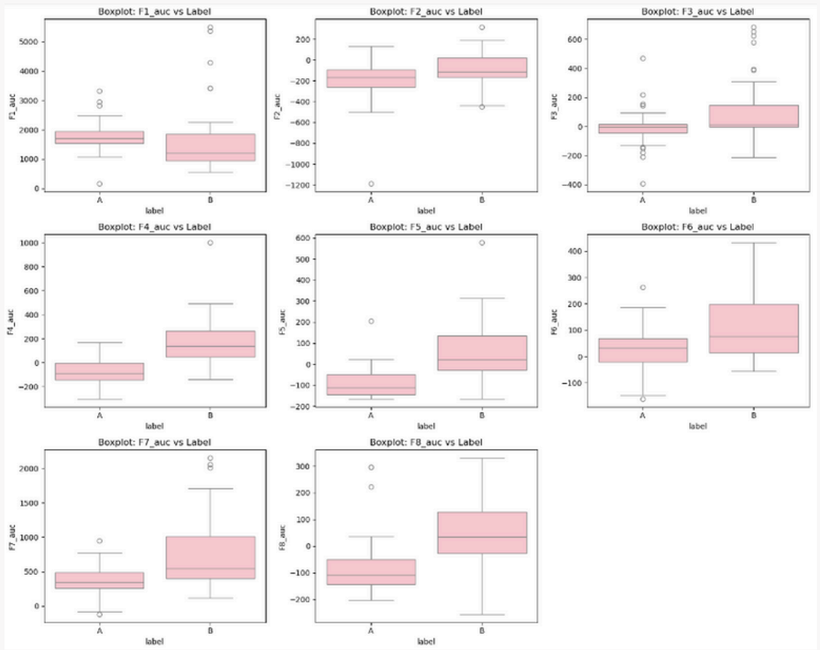


Preprocessing

- Distribusi fitur terhadap label



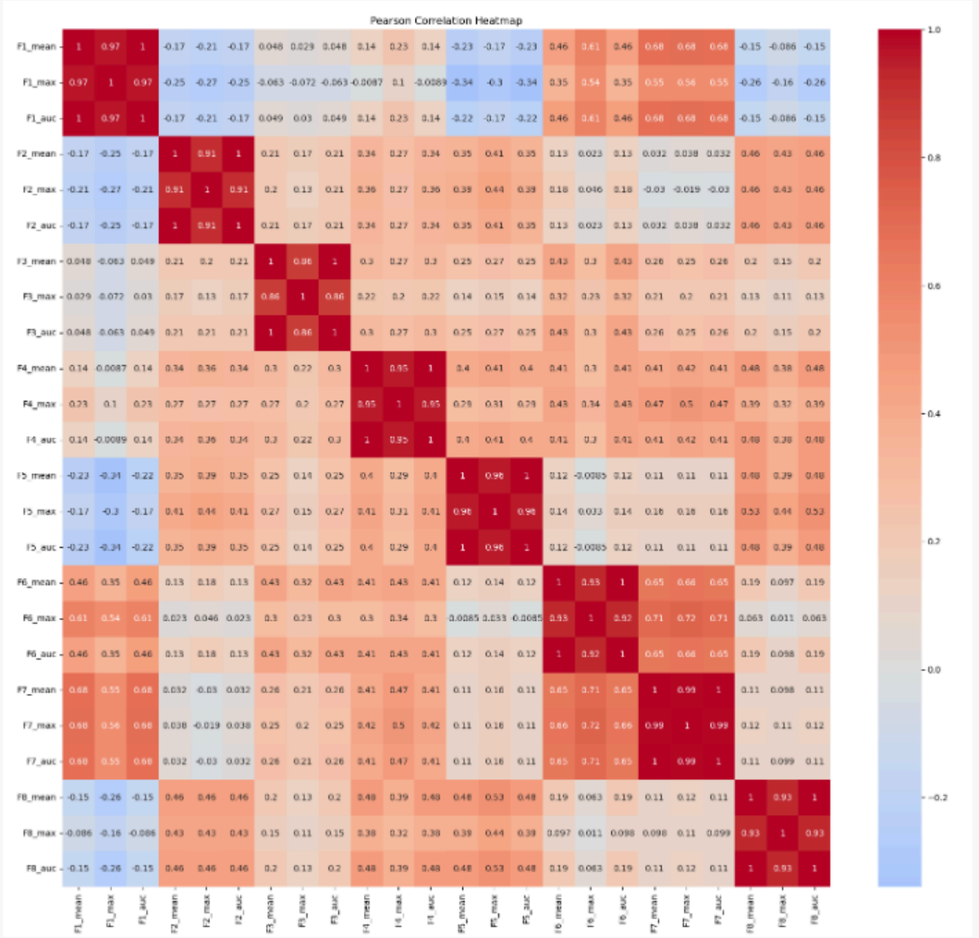




Beberapa fitur seperti *f2_mean, f6_max, dan f2_auc* **kurang informatif** untuk klasifikasi karena memiliki distribusi yang mirip antar kelas, median yang mirip, serta overlap yang tinggi, meskipun *f2_auc* memiliki outlier pada salah satu label yang bisa membantu klasifikasi pada nilai ekstrem. Fitur-fitur tersebut sebaiknya tidak digunakan atau dikombinasikan dengan fitur lain yang lebih informatif jika model memiliki mekanisme seleksi fitur internal seperti tree-based.

Preprocessing

- Korelasi fitur.
  Didapat bahwa korelasi antar fitur dalam masing-masing sensor memiliki nilai tinggi yang berarti data sensor stabil dan tidak noise.



Preprocessing

- Data cleaning, yaitu cek missing value dan duplicate value.

```
[13]:
df[df.duplicated()]
```

[13]:

	F1_mean	F1_max	F1_auc	F2_mean	F2_max	F2_auc

0 rows × 25 columns

dan *df.isnull().sum()*

- Normalisasi dilakukan di pipeline model untuk menghindari data leakege.
- Seleksi fitur juga dilakukan di pipeline karena beberapa model memiliki mekanisme internal secara otomatis untuk mengidentifikasi fitur yang paling relevan.