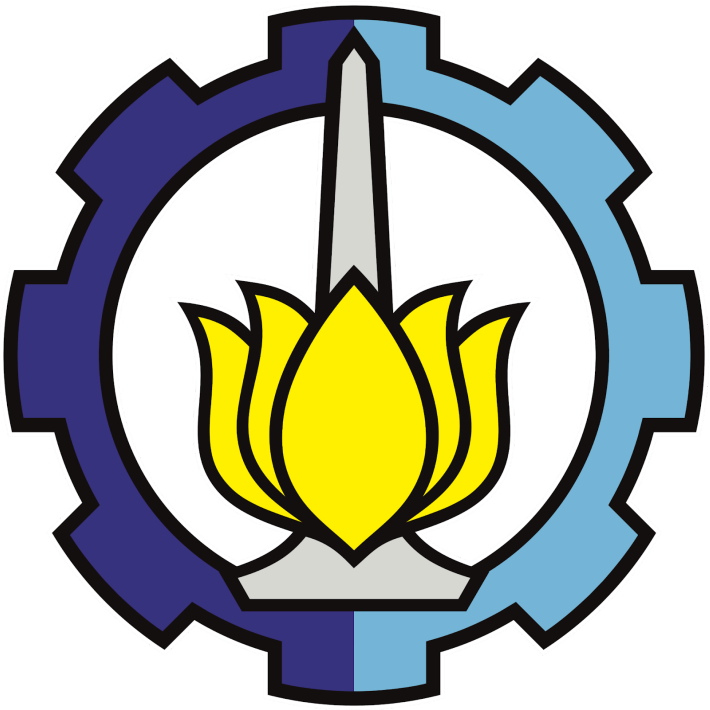
LAPORAN FINAL PROJECT MACHINE LEARNING 2025
KELOMPOK 01 MACHINE LEARING KELAS E
“Wine Quality Classification”



DOSEN PEMBIMBING:

Prof. Dr. Eng. Chastine Fatichah, S.Kom., M.Kom.

ANGGOTA:

Ananda Faris Ghazi Ramadhan	5025231280
Elmira Farah Azalia	5025211197
Izzudin Ali Akbari	5025231313

INSTITUT TEKNOLOGI SEPULUH NOPEMBER

FAKULTAS TEKNOLOGI ELEKTRO DAN INFORMATIKA CERDAS

DEPARTEMEN TEKNIK INFORMATIKA

2025

DAFTAR ISI

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Industri anggur merupakan salah satu pilar ekonomi utama bagi beberapa negara Barat, termasuk Prancis, Italia, dan Spanyol. Industri ini menghasilkan keuntungan yang besar dan terus berkembang untuk menarik pelanggan baru. Seiring pertumbuhan pasar, para produsen senantiasa mengembangkan rasa baru dan bereksperimen dengan teknik produksi inovatif. Namun, menganalisis produk‑produk baru tersebut hanya dengan keahlian manusia memakan banyak waktu dan biaya. Untuk mengatasi tantangan ini, industri dapat mengadopsi metode komputasi canggih melalui pembelajaran mesin. Ilmuwan data dapat menerapkan algoritma machine learning untuk menyederhanakan pengembangan produk, sehingga proses menjadi lebih efisien dan efektif.

1.2 Tujuan

Tujuan dari pengerjaan project ini adalah sebagai berikut:

1. Membuat produksi wine lebih efektif dan efisien
2. Mencegah kerugian dengan menentukan produk yang berkualitas rendah
3. Menguji dan membandingkan berbagai algoritma machine learning untuk menentukan model terbaik dalam mengklasifikasikan kualitas wine

1.3 Dataset dan Fitur

Dataset yang digunakan berasal dari platform **Kaggle** dan berisi informasi mengenai karakteristik kimia serta label kualitas wine. Terdapat empat fitur numerik sebagai variabel input, yaitu *fixed_acidity*, *residual_sugar*, *alcohol*, dan *density*, yang masing-masing merepresentasikan sifat fisik atau kimiawi dari wine. Sementara itu, *quality_label* merupakan variabel target yang mengklasifikasikan kualitas wine ke dalam tiga kategori: **low**, **medium**, dan **high**, dengan asumsi bahwa jarak antar kategori bersifat seimbang atau setara (*ordinal classification*). Fitur-fitur ini akan digunakan dalam proses pelatihan dan pengujian berbagai algoritma machine learning guna mengembangkan model prediktif yang mampu mengklasifikasikan kualitas wine secara otomatis dan akurat.

BAB 2
METODOLOGI

2.1 Preprocessing Data

Langkah-langkah *preprocessing* yang diterapkan meliputi:

- Pemeriksaan dan Penanganan Missing Values

  Dataset diperiksa untuk nilai kosong. Jika ditemukan, nilai kosong diisi dengan modus (nilai yang paling sering muncul).

```
# Mencari apakah ada data NULL di dalam masing-masing kolom (Attribut + Label)
for col in df.columns:
    if df[col].isnull().sum() > 0: # Jika NULL, ...
        most_frequent = df[col].mode()[0]
        df[col].fillna(most_frequent, inplace=True) # Isi kotak null dengan data yang paling sering muncul
        print(f"Filled missing values in column '{col}' with mode: {most_frequent}")
    else:
        print(f"No missing values in column '{col}'")
```

```
No missing values in column 'fixed_acidity'
No missing values in column 'residual_sugar'
No missing values in column 'alcohol'
No missing values in column 'density'
No missing values in column 'quality_label'
```

- Encoding Label Target

  Label `quality_label` yang semula berupa string `'low'`, `'medium'`, dan `'high'`, diubah menjadi angka `0`, `1`, dan `2` menggunakan dictionary mapping:

[5]:
```
# Mengubah data kategorical dalam kolom 'quality_label' menjadi numerik
mapping = {'low': 0, 'medium': 1, 'high': 2}
df['quality_label'] = df['quality_label'].map(mapping)
```

[6]:
```
print("\nMapping quality_label ke nilai numerik:")
for label, num in mapping.items():
    print(f"  {num} = {label}")
```

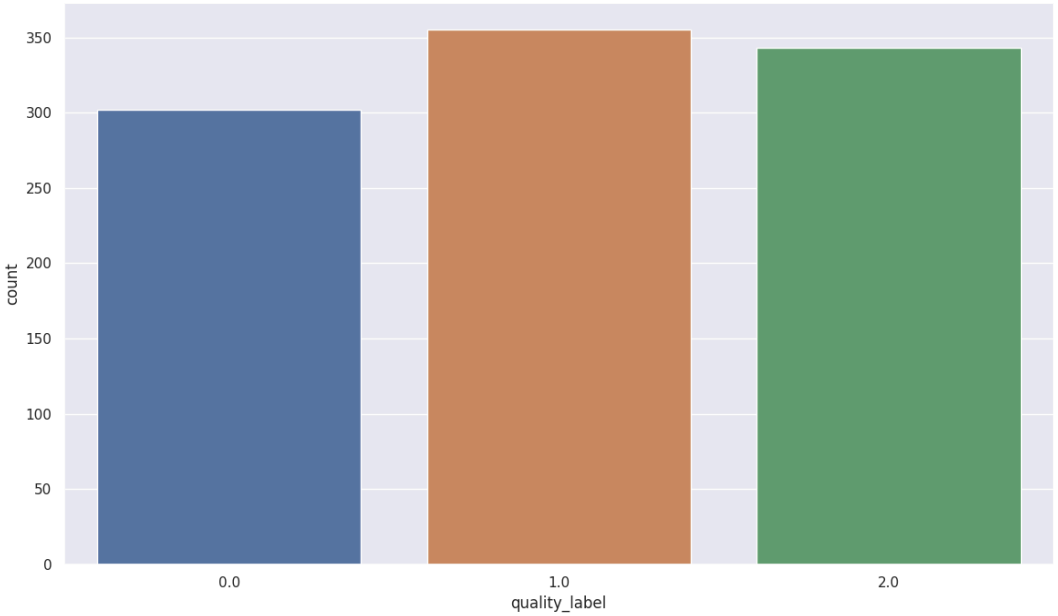
```
Mapping quality_label ke nilai numerik:
  0 = low
  1 = medium
  2 = high
```

2.2 Eksplorasi Data

Beberapa teknik visualisasi digunakan untuk memahami pola dan distribusi data:

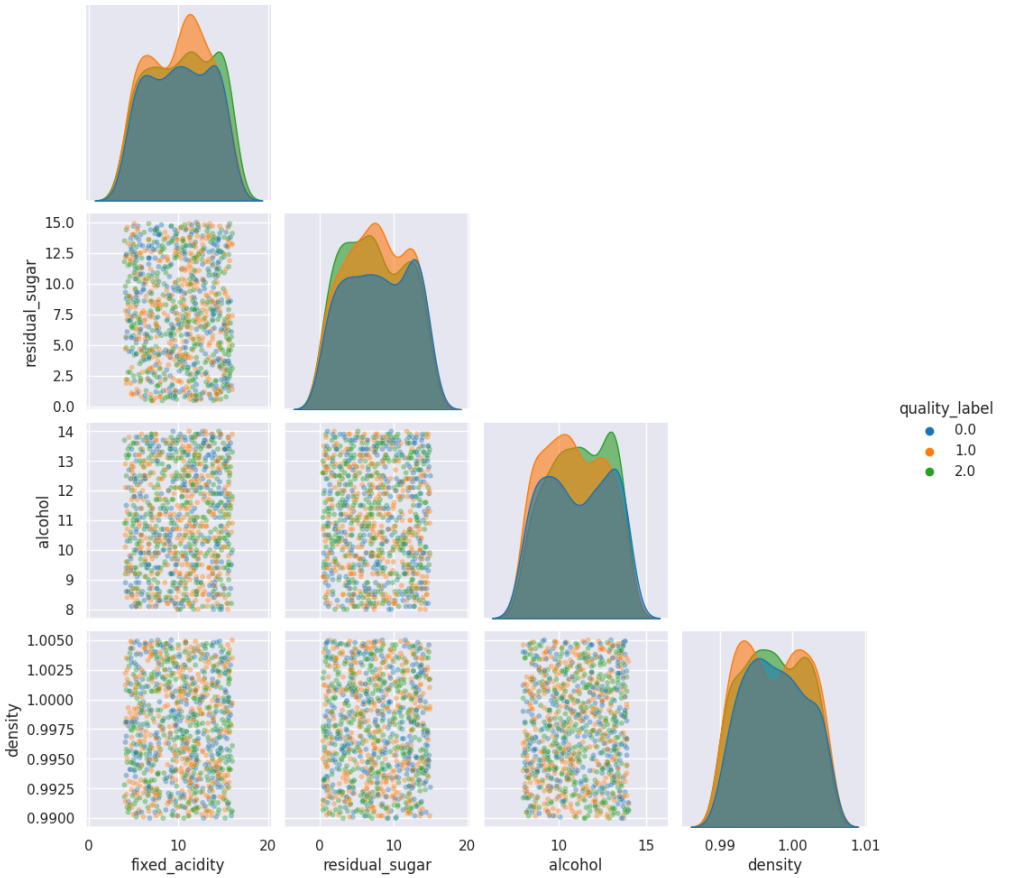
- Distribusi Label

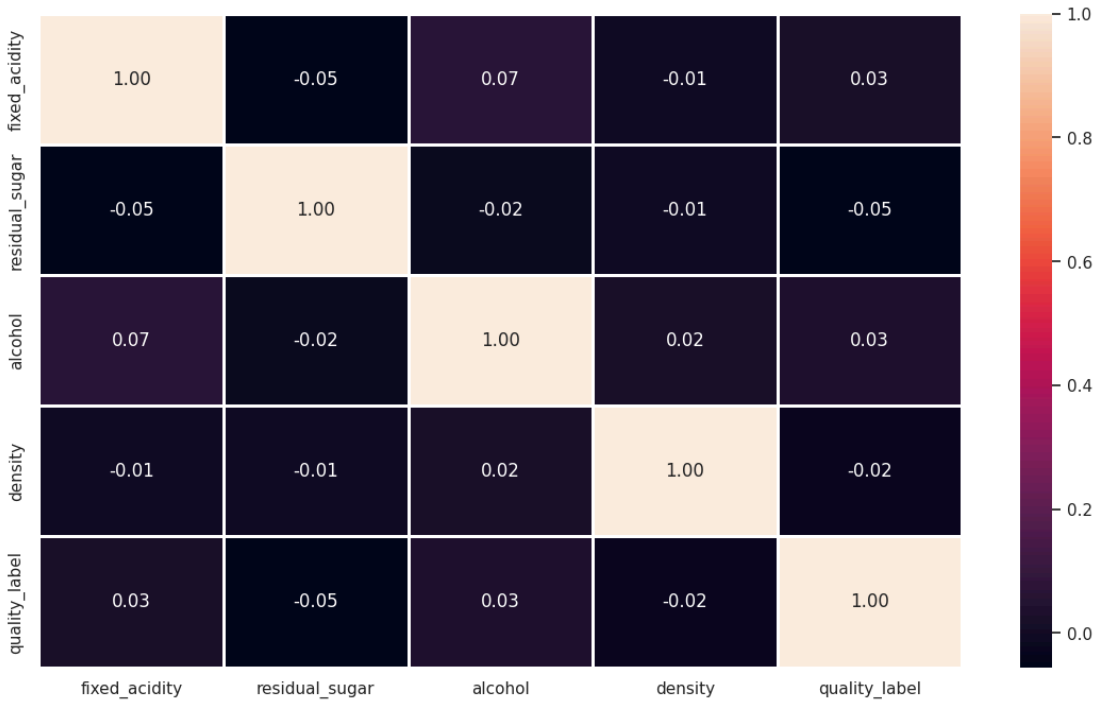
  Grafik count plot digunakan untuk menampilkan jumlah sampel untuk setiap kelas kualitas (`low`, `medium`, `high`).



- Pairplot dan Heatmap

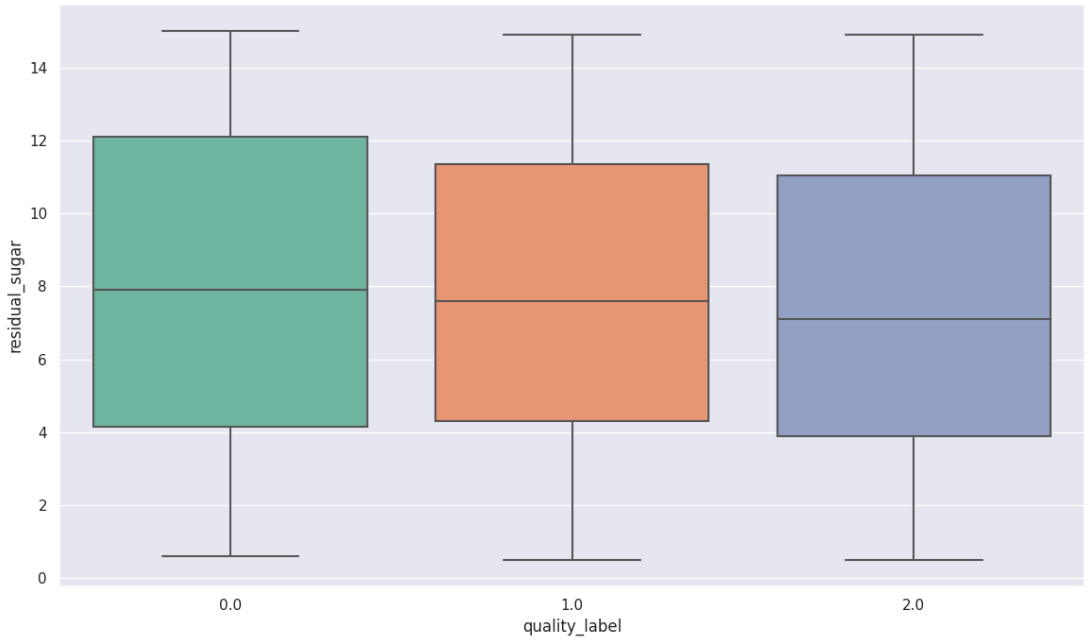
  Pairplot digunakan untuk melihat sebaran dan korelasi antar fitur berdasarkan label kualitas. Heatmap membantu mengidentifikasi korelasi antar fitur, di mana fitur seperti `alcohol` dan `density` menunjukkan korelasi yang mencolok dengan `quality_label`.





- Boxplot dan Statistik Deskriptif

  Boxplot fitur `residual_sugar` berdasarkan kualitas wine membantu memahami persebaran dan outlier. Deskripsi statistik seperti mean, min, dan max ditampilkan per kelas.



- Visualisasi 3D

  Visualisasi scatter plot 3D dilakukan dengan sumbu `fixed_acidity`, `alcohol`, dan `quality_label` untuk melihat pemisahan kelas dalam ruang fitur.

Tabel Low Quality (0):

	fixed_acidity	residual_sugar	alcohol	density
count	302.00	302.00	302.00	302.00
mean	10.13	7.95	11.07	1.00
std	3.42	4.32	1.86	0.00
min	4.10	0.60	8.00	0.99
25%	7.12	4.15	9.43	0.99
50%	10.10	7.90	11.10	1.00
75%	13.38	12.10	12.80	1.00
max	15.90	15.00	14.00	1.00

Tabel Medium Quality (1):

	fixed_acidity	residual_sugar	alcohol	density
count	355.00	355.00	355.00	355.00
mean	10.05	7.74	10.87	1.00
std	3.36	4.13	1.73	0.00
min	4.10	0.50	8.00	0.99
25%	7.10	4.30	9.40	0.99
50%	10.40	7.60	10.80	1.00
75%	12.70	11.35	12.40	1.00
max	16.00	14.90	14.00	1.00

Tabel High Quality (2):

	fixed_acidity	residual_sugar	alcohol	density
count	343.00	343.00	343.00	343.00
mean	10.34	7.41	11.19	1.00
std	3.58	4.19	1.69	0.00
min	4.00	0.50	8.00	0.99
25%	7.30	3.90	9.80	0.99
50%	10.50	7.10	11.20	1.00
75%	13.65	11.05	12.75	1.00
max	16.00	14.90	14.00	1.00

BAB 3
HASIL DAN PEMBAHASAN

3.1 Pembagian Data

Dataset dibagi menjadi data latih dan uji dengan rasio 80:20 menggunakan metode *stratified split* agar proporsi kelas tetap seimbang.

```
# Split dataframe menjadi x = atribut, y = target or label
# Lalu train_test_split x dan y
x = df.drop('quality_label', axis=1)   # Hanya drop kolom label, Attr = X
y = df['quality_label'] # Target = Y

x_train, x_test, y_train, y_test = train_test_split(
    x, y,
    test_size=0.20,
    random_state=42,
    stratify=y
)
```

3.2 Model

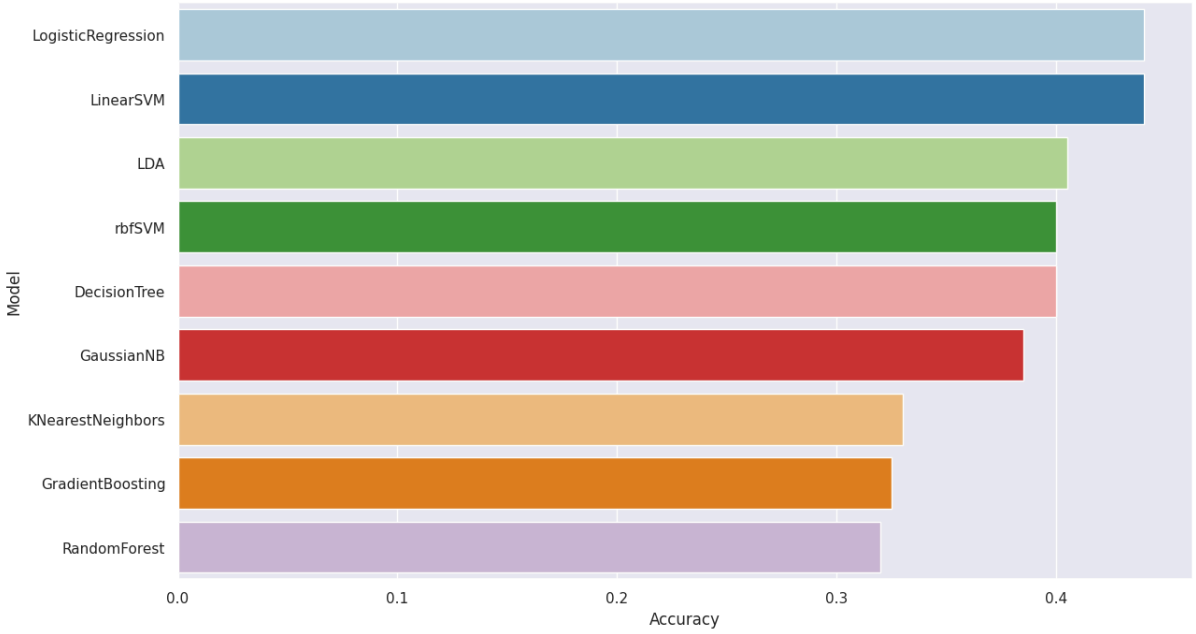
Sebanyak 9 model machine learning diuji untuk klasifikasi kualitas wine:

- Logistic Regression
- Support Vector Machine (Linear dan RBF)
- K-Nearest Neighbors
- Random Forest
- Decision Tree
- Gradient Boosting
- Gaussian Naive Bayes
- Linear Discriminant Analysis (LDA)

3.3 Evaluasi Awal

Setelah pelatihan awal, akurasi dari masing-masing model diuji terhadap data uji. Hasil akurasi awal divisualisasikan menggunakan grafik batang dan disusun dari yang tertinggi hingga terendah.

	Model	Accuracy
0	LogisticRegression	0.440
1	LinearSVM	0.440
2	LDA	0.405
3	rbfSVM	0.400
4	GaussianNB	0.385
5	DecisionTree	0.345
6	KNearestNeighbors	0.330
7	GradientBoosting	0.325
8	RandomForest	0.300



3.4 Evaluasi Lanjutan

Untuk memperoleh evaluasi yang lebih stabil, dilakukan Stratified K-Fold Cross Validation sebanyak 5 fold. Metrik yang dihitung, *Accuracy* (rata-rata dan standar deviasi), *Precision* (macro average), *Recall* (macro average) F1-Score (macro average). Model terbaik ditentukan berdasarkan rata-rata akurasi tertinggi, dengan mempertimbangkan stabilitas (standar deviasi rendah) dan performa metrik lain (*Precision*, *Recall*, F1).

```
=== Cross-Validation Results (5-Fold) ===
              Model  Accuracy Mean  Accuracy Std  Precision Mean  Precision Std  Recall Mean  Recall Std  F1-Score Mean  F1-Score Std
         GaussianNB          0.380      0.027749        0.387034       0.029653     0.366467    0.028116       0.336163      0.035694
          LinearSVM          0.379      0.035553        0.321449       0.072379     0.362750    0.034410       0.311709      0.038792
 LogisticRegression          0.378      0.038549        0.347023       0.062377     0.362289    0.037452       0.315958      0.041905
                LDA          0.378      0.038549        0.347023       0.062377     0.362289    0.037452       0.315958      0.041905
  KNearestNeighbors          0.357      0.013638        0.363010       0.012080     0.358882    0.014536       0.354231      0.014148
              rbfSVM          0.338      0.027129        0.336146       0.029158     0.330818    0.028044       0.321886      0.030323
        DecisionTree          0.322      0.033257        0.320606       0.033144     0.319667    0.034494       0.318975      0.033940
        RandomForest          0.317      0.022045        0.312619       0.024183     0.313890    0.024871       0.312088      0.025116
    GradientBoosting          0.290      0.040249        0.286191       0.042440     0.286068    0.041150       0.283626      0.041647

 quality_label
1.0    355
2.0    343
0.0    302
Name: count, dtype: int64
```

BAB 4
PENUTUP

4.1 Kesimpulan

Pada final project ini, kami telah menguji berbagai algoritma machine learning untuk mengklasifikasikan kualitas wine berdasarkan fitur-fitur kimia. Evaluasi dilakukan menggunakan metode Stratified K-Fold Cross-Validation sebanyak 5 lipatan untuk menjaga proporsi label dan memastikan kestabilan hasil. Dari sembilan model yang diuji, Gaussian Naive Bayes menunjukkan performa terbaik dengan akurasi rata-rata sebesar 38% dan skor F1 tertinggi di antara semua model, sedikit mengungguli Linear SVM dan Logistic Regression. Meskipun nilai akurasinya belum tinggi, model ini memiliki keseimbangan yang relatif baik antara presisi, recall, dan F1-score, serta hasil yang cukup konsisten di setiap lipatan.

4.2 Saran

Dengan akurasi tertinggi hanya sekitar 38%, performa model menunjukkan bahwa fitur dalam dataset belum cukup kuat untuk klasifikasi yang akurat. Untuk peningkatan, disarankan:

- Melakukan feature engineering dan menambah fitur relevan
- Memperluas dan menyeimbangkan dataset, misalnya dengan teknik SMOTE
- Menerapkan tuning hyperparameter dan model ensemble lanjutan seperti XGBoost
- Melakukan validasi silang dengan dataset serupa untuk menguji generalisasi

Sebagai pengembangan berikutnya, pendekatan deep learning seperti Multi-Layer Perceptron (MLP) dapat digunakan untuk menangani kompleksitas data lebih lanjut. Topik ini akan dibahas dalam perkuliahan lanjutan di bidang Deep Learning.