


Nama : Rizki Ramadhan


NIM : 1103213091

Machine Learning

Classification Model

Dataset :

**Bank Marketing**
The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if...
Classification Multivariate 45.21K Instances 17 Features

**Bank Marketing**
Donated on 2/13/2012

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y).

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Business	Classification

Feature Type	# Instances	# Features
Categorical, Integer	45211	16

Dataset Information

Additional Information
The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. ...
SHOW MORE

Has Missing Values?
No

DOWNLOAD (999.8 KB)

IMPORT IN PYTHON

CITE

9 citations
289973 views

Creators
S. Moro
P. Rita
P. Cortez

DOI
10.24432/C5K306

License
This dataset is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

1. Importing Library

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split, cross_val_score, GridSearchCV
from sklearn.preprocessing import LabelEncoder, StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from xgboost import XGBClassifier
from sklearn.metrics import classification_report, accuracy_score
import numpy as np
```

✓ 0.0s

2. Membaca File CSV

	age	job	marital	education	default	balance	housing	loan	\
0	58	management	married	tertiary	no	2143	yes	no	
1	44	technician	single	secondary	no	29	yes	no	
2	33	entrepreneur	married	secondary	no	2	yes	yes	
3	47	blue-collar	married	unknown	no	1506	yes	no	
4	33	unknown	single	unknown	no	1	no	no	
	contact	day	month	duration	campaign	pdays	previous	poutcome	y
0	unknown	5	may	261	1	-1	0	unknown	no
1	unknown	5	may	151	1	-1	0	unknown	no
2	unknown	5	may	76	1	-1	0	unknown	no
3	unknown	5	may	92	1	-1	0	unknown	no
4	unknown	5	may	198	1	-1	0	unknown	no

Terdiri dari berbagai atribut seperti demografi (umur, pekerjaan, status pernikahan, pendidikan), informasi keuangan (saldo, pinjaman), serta data kampanye pemasaran (durasi, kontak, hasil sebelumnya). Beberapa fitur, seperti **education** dan **previous**, memiliki nilai "unknown," yang mengindikasikan adanya data yang hilang atau tidak diketahui. Target variabel y ("yes" atau "no") memungkinkan analisis klasifikasi untuk memprediksi keberhasilan kampanye berdasarkan fitur-fitur lainnya.

3. Explanatory Data informasi dataset

```
Informasi Dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 17 columns):
#   Column      Non-Null Count  Dtype
---  -
0   age         45211 non-null  int64
1   job         45211 non-null  object
2   marital     45211 non-null  object
3   education   45211 non-null  object
4   default     45211 non-null  object
5   balance     45211 non-null  int64
6   housing     45211 non-null  object
7   loan        45211 non-null  object
8   contact     45211 non-null  object
9   day         45211 non-null  int64
10  month       45211 non-null  object
11  duration    45211 non-null  int64
12  campaign    45211 non-null  int64
13  pdays       45211 non-null  int64
14  previous    45211 non-null  int64
15  poutcome    45211 non-null  object
16  y           45211 non-null  object
dtypes: int64(7), object(10)
memory usage: 5.9+ MB
None
```

Dataset ini memiliki 45.211 entri dengan 17 kolom, yang mencakup data demografi, keuangan, dan informasi kampanye pemasaran. Tidak ada nilai null di seluruh kolom, yang berarti data ini lengkap dan siap untuk dianalisis. Kolom-kolom seperti age, balance, duration, dan campaign bertipe numerik, sedangkan sebagian besar lainnya bertipe kategorikal (job, marital, education, dll.), sehingga preprocessing seperti encoding variabel kategorikal diperlukan sebelum modeling. Target variabel y bertipe kategorikal ("yes" atau "no"), memungkinkan analisis klasifikasi.

4. Statistik Deskriptif

```
Statistik Deskriptif:
      age      balance      day      duration      campaign \
count 45211.000000 45211.000000 45211.000000 45211.000000 45211.000000
mean   40.936210  1362.272058   15.806419   258.163080   2.763841
std    10.618762  3044.765829    8.322476  257.527812   3.098021
min    18.000000  -8019.000000    1.000000    0.000000    1.000000
25%    33.000000   72.000000    8.000000   103.000000    1.000000
50%    39.000000  448.000000   16.000000   180.000000    2.000000
75%    48.000000 1428.000000   21.000000   319.000000    3.000000
max    95.000000 102127.000000   31.000000  4918.000000   63.000000

      pdays      previous
count 45211.000000 45211.000000
mean   40.197828   0.580323
std    100.128746   2.303441
min    -1.000000   0.000000
25%    -1.000000   0.000000
50%    -1.000000   0.000000
75%    -1.000000   0.000000
max    871.000000  275.000000
```

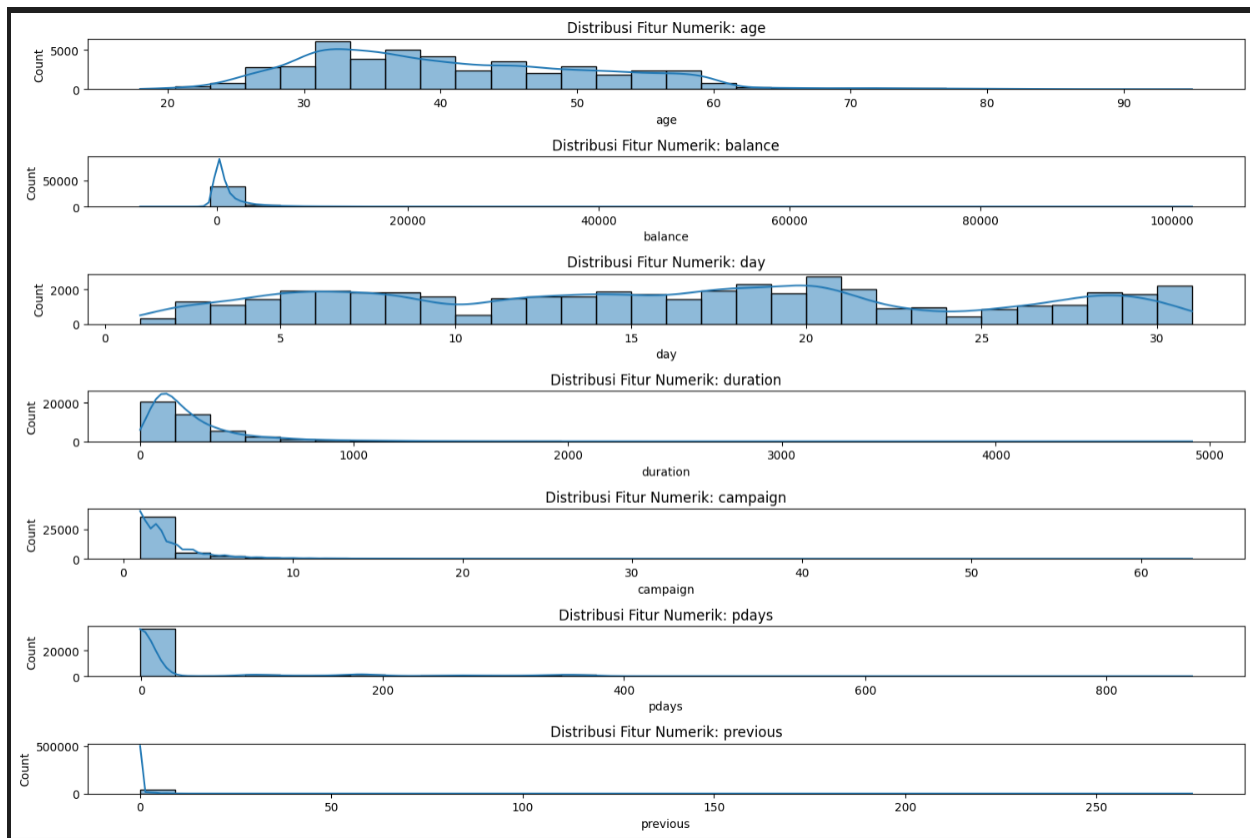
Dari statistik deskriptif, terlihat bahwa kolom *age* memiliki rentang nilai dari 18 hingga 95 tahun, dengan rata-rata 40,9 tahun, mencerminkan beragamnya kelompok usia dalam dataset ini. Kolom *balance* menunjukkan variasi besar, dengan nilai minimum -8019 dan maksimum 102127, mengindikasikan adanya outlier yang dapat memengaruhi analisis atau model. Kolom *duration*, yang mencerminkan durasi kontak, memiliki nilai maksimum 4918 detik, yang juga dapat dianggap *outlier*. Selain itu, kolom *pdays* dan *previous* memiliki banyak nilai -1 dan 0, yang kemungkinan menunjukkan ketidakhadiran atau ketidakaktifan dalam kontak sebelumnya.

5. Missing Values

```
Missing Values:
age           0
job           0
marital       0
education     0
default       0
balance       0
housing       0
loan          0
contact       0
day           0
month         0
duration      0
campaign      0
pdays        0
previous      0
poutcome     0
y             0
dtype: int64
```

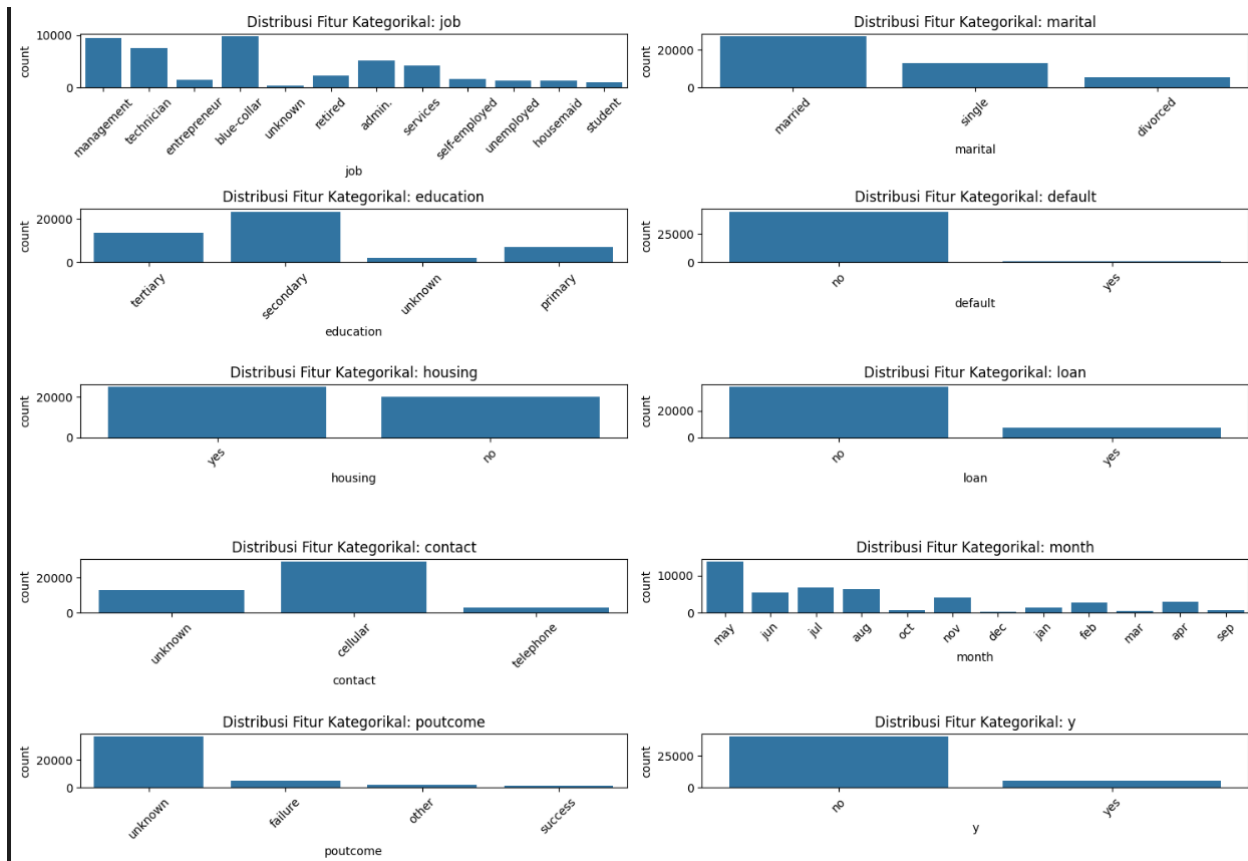
Dataset ini tidak memiliki nilai yang hilang pada seluruh kolom, sehingga data siap untuk analisis lebih lanjut. Hal ini memberikan keuntungan karena tidak ada risiko bias yang muncul akibat pengisian data kosong. Namun, meskipun tidak ada nilai kosong, beberapa kolom seperti *education*, *contact*, dan *poutcome* mungkin memiliki nilai kategori seperti "*unknown*" yang tetap perlu diperhatikan dalam *preprocessing*. Dengan data yang bersih ini, fokus dapat langsung diarahkan ke eksplorasi fitur, penanganan *outlier*, dan pembuatan model prediktif untuk menganalisis hubungan antara variabel dengan target *y*.

6. Visualisasi dasar untuk memahami distribusi data tipe numerikal



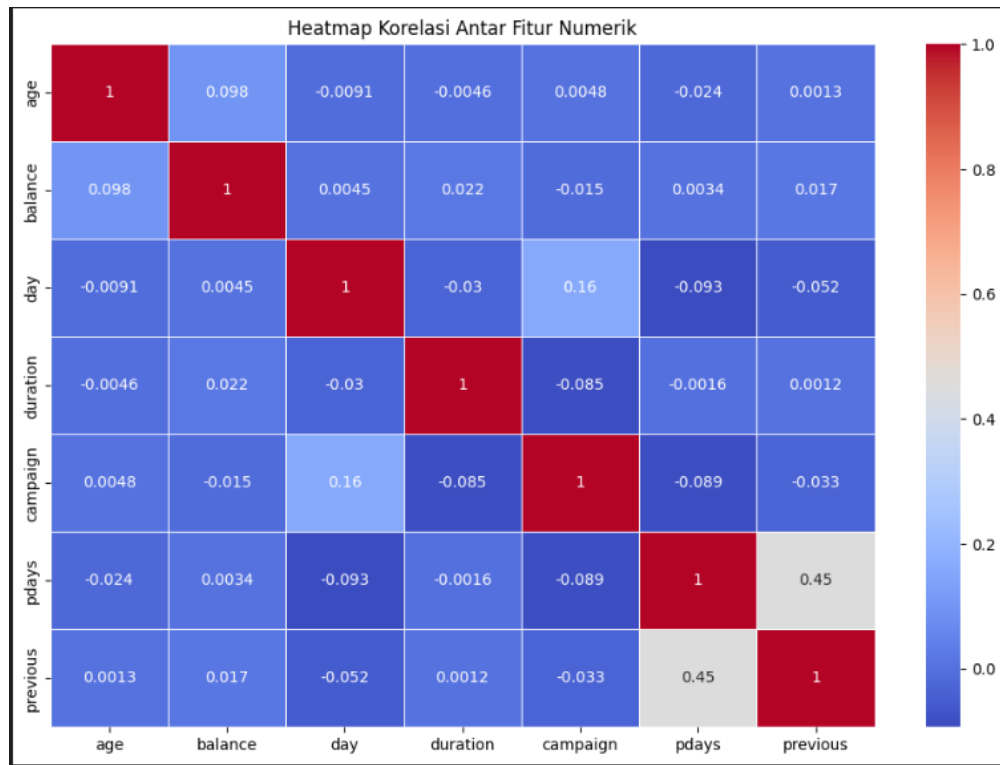
Distribusi fitur numerik dalam dataset menunjukkan pola yang beragam. Sebagian besar data usia terkonsentrasi pada rentang 30-50 tahun, sedangkan saldo (*balance*) memiliki distribusi yang sangat tidak merata dengan mayoritas di sekitar nol dan beberapa *outlier* ekstrem hingga mendekati 100.000. Durasi panggilan (*duration*) didominasi oleh nilai pendek dengan sedikit panggilan berdurasi sangat panjang, sementara jumlah kontak kampanye (*campaign*) sebagian besar berkisar antara 1 hingga 3 kontak, meskipun terdapat nilai hingga 63 yang menunjukkan *outlier*. Selain itu, fitur *pdays* dan *previous* mayoritas bernilai nol atau -1, menandakan banyak pelanggan yang tidak memiliki kontak sebelumnya.

7. Visualisasi dasar untuk memahami distribusi data tipe numerikal



Distribusi fitur kategorikal menunjukkan beberapa pola menarik. Fitur pekerjaan (*job*) didominasi oleh kategori seperti *blue-collar*, *management*, dan *technician*, sementara kategori seperti *student* dan *housemaid* jauh lebih sedikit. Sebagian besar pelanggan berstatus menikah (*marital*) dan memiliki tingkat pendidikan *secondary* (*education*). Mayoritas tidak memiliki pinjaman kredit (*loan*) dan pinjaman rumah (*housing*), serta tidak memiliki default pada catatan keuangan (*default*). Sebagian besar kontak dilakukan melalui seluler (*contact*), dan mayoritas terjadi pada bulan Mei. Fitur hasil kampanye sebelumnya (*poutcome*) didominasi oleh nilai *unknown*, yang dapat menunjukkan data yang hilang atau tidak relevan. Target *y* menunjukkan bahwa sebagian besar pelanggan tidak menyetujui penawaran *campaign*.

8. Visualisasi Heatmap



Heatmap korelasi antar fitur numerik menunjukkan bahwa sebagian besar variabel memiliki korelasi yang rendah satu sama lain, dengan nilai korelasi mendekati nol. Fitur *previous* dan *pdays* memiliki korelasi tertinggi di antara pasangan variabel, yaitu sekitar 0.45, yang menunjukkan hubungan moderat, kemungkinan karena keduanya mencerminkan riwayat kontak pelanggan sebelumnya. Variabel lain seperti *balance*, *duration*, dan *campaign* memiliki korelasi yang sangat rendah dengan fitur lain, menandakan bahwa mereka cenderung independen dan dapat memberikan informasi unik.

9. Pembagian data training dan juga data testing

```
Data pelatihan dan pengujian telah dibagi:  
Jumlah data pelatihan: 36168  
Jumlah data pengujian: 9043
```

Dataset telah dibagi menjadi **36.168 data pelatihan** dan **9.043 data pengujian**, yang sesuai dengan rasio umum 80:20. Pembagian ini memastikan model memiliki cukup data untuk belajar pada fase pelatihan dan menyediakan dataset pengujian yang memadai untuk mengevaluasi performa.

10. Pipeline untuk berbagai model klasifikasi

Model: Logistic Regression					
Akurasi: 0.8988167643481145					
	precision	recall	f1-score	support	
0	0.92	0.98	0.94	7952	
1	0.65	0.34	0.45	1091	
accuracy			0.90	9043	
macro avg	0.79	0.66	0.70	9043	
weighted avg	0.88	0.90	0.88	9043	
Model: Decision Tree					
Akurasi: 0.8727192303439124					
	precision	recall	f1-score	support	
0	0.93	0.93	0.93	7952	
1	0.47	0.49	0.48	1091	
accuracy			0.87	9043	
macro avg	0.70	0.71	0.70	9043	
weighted avg	0.87	0.87	0.87	9043	
Model: k-NN					
Akurasi: 0.8988167643481145					
	precision	recall	f1-score	support	
0	0.92	0.97	0.94	7952	
1	0.63	0.38	0.48	1091	
accuracy			0.90	9043	
macro avg	0.78	0.68	0.71	9043	
weighted avg	0.89	0.90	0.89	9043	

Model: XGBoost					
Akurasi: 0.9064469755612076					
	precision	recall	f1-score	support	
0	0.93	0.96	0.95	7952	
1	0.64	0.50	0.57	1091	
accuracy			0.91	9043	
macro avg	0.79	0.73	0.76	9043	
weighted avg	0.90	0.91	0.90	9043	

Preprocessing dan evaluasi model telah selesai dilakukan untuk berbagai model klasifikasi.

Berdasarkan hasil evaluasi model terhadap target y, model **Logistic Regression** dan **k-NN** memberikan akurasi tertinggi sebesar 89.88%, sementara model **Decision Tree** memiliki akurasi lebih rendah sebesar 87.27%. Namun, perbedaan signifikan terlihat pada nilai **recall** dan **F1-score** untuk kelas 1, di mana Logistic Regression hanya memiliki recall 34% dan F1-score 44%, menunjukkan bahwa model kesulitan mendeteksi pelanggan yang menyetujui produk. Kinerja k-NN sedikit lebih baik dengan recall 38% dan F1-score 48% untuk kelas 1, namun masih rendah dibandingkan kelas 0.

Model **XGBoost** memberikan performa terbaik dengan akurasi sebesar 90.64%. Selain itu, model ini menunjukkan peningkatan pada **recall** (50%) dan **F1-score** (57%) untuk kelas 1, yang lebih tinggi dibandingkan Logistic Regression, Decision Tree, maupun k-NN. Kelas 0 tetap memiliki metrik yang sangat baik, dengan precision, recall, dan F1-score di atas 93%. **Macro average** menunjukkan bahwa model mampu menangani kedua kelas dengan lebih seimbang dibandingkan model lain. Hasil ini menunjukkan bahwa XGBoost adalah pilihan optimal untuk dataset ini

11. Hyperparameter Tuning

```
Model: Logistic Regression (Setelah Hyperparameter Tuning)
Best Parameters: {'classifier__C': 1}
Akurasi: 0.8988167643481145
      precision    recall  f1-score   support

     0       0.92       0.98       0.94       7952
     1       0.65       0.34       0.45       1091

   accuracy          0.90       9043
  macro avg       0.79       0.66       0.70       9043
 weighted avg       0.88       0.90       0.88       9043

Model: Decision Tree (Setelah Hyperparameter Tuning)
Best Parameters: {'classifier__max_depth': 5, 'classifier__min_samples_split': 10}
Akurasi: 0.8972686055512551
      precision    recall  f1-score   support

     0       0.91       0.97       0.94       7952
     1       0.64       0.33       0.44       1091

   accuracy          0.90       9043
  macro avg       0.78       0.65       0.69       9043
 weighted avg       0.88       0.90       0.88       9043

Model: k-NN (Setelah Hyperparameter Tuning)
Best Parameters: {'classifier__n_neighbors': 3}
Akurasi: 0.8906336392790003
      precision    recall  f1-score   support

     0       0.92       0.96       0.94       7952
     1       0.57       0.39       0.46       1091

   accuracy          0.89       9043
  macro avg       0.74       0.67       0.70       9043
 weighted avg       0.88       0.89       0.88       9043

Model: XGBoost (Setelah Hyperparameter Tuning)
Best Parameters: {'classifier__max_depth': 3, 'classifier__n_estimators': 100}
Akurasi: 0.9060046444763906
      precision    recall  f1-score   support

     0       0.93       0.97       0.95       7952
     1       0.66       0.46       0.54       1091

   accuracy          0.91       9043
  macro avg       0.79       0.71       0.74       9043
 weighted avg       0.90       0.91       0.90       9043
```

Setelah dilakukan hyperparameter tuning, **Logistic Regression** dengan parameter terbaik {C: 1} dan **Decision Tree** dengan parameter terbaik {max_depth: 5, min_samples_split: 10} menunjukkan hasil yang hampir serupa dalam hal akurasi (Logistic Regression: 89.88%, Decision Tree: 89.72%). Namun, Logistic Regression memiliki **F1-score** yang sedikit lebih baik untuk kelas 1 (45%) dibandingkan Decision Tree (44%). Kedua model masih memiliki kesulitan dalam mendeteksi kelas 1, terlihat dari rendahnya nilai recall (34% untuk Logistic Regression dan 33% untuk Decision Tree).

Setelah dilakukan hyperparameter tuning, model **k-NN** dengan parameter terbaik {n_neighbors: 3} memiliki akurasi sebesar 89.06%, tetapi performa pada kelas minoritas (1) masih rendah dengan recall hanya 39% dan F1-score 46%. Sebaliknya, model **XGBoost** dengan parameter terbaik {max_depth: 3, n_estimators: 100} memberikan performa terbaik secara

keseluruhan, dengan akurasi sebesar 90.60%. XGBoost juga menunjukkan peningkatan pada kelas minoritas, dengan recall 46% dan F1-score 54%, yang lebih baik dibandingkan model lainnya.

Hyperparameter tuning meningkatkan stabilitas performa kedua model, dengan XGBoost tetap unggul dalam menangani ketidakseimbangan kelas sambil mempertahankan akurasi tinggi pada kelas mayoritas (0). Hal ini menjadikan XGBoost sebagai pilihan optimal untuk dataset ini.

KESIMPULAN

Dari evaluasi semua model, terlihat bahwa **XGBoost** adalah model dengan performa terbaik, baik sebelum maupun setelah hyperparameter tuning. Setelah tuning, XGBoost mencapai akurasi tertinggi sebesar **90.60%**, dengan peningkatan **F1-score** untuk kelas 1 sebesar **54%** dan recall sebesar **46%**, menjadikannya model yang lebih andal dalam menangani ketidakseimbangan kelas.

Model lainnya, seperti **Logistic Regression**, **Decision Tree**, dan **k-NN**, juga menunjukkan akurasi yang kompetitif, berkisar antara **89%-90%**, namun kesulitan dalam mendeteksi kelas 1. Setelah tuning, Logistic Regression dan Decision Tree masing-masing memiliki F1-score sebesar **45%** dan **44%**, sedangkan k-NN mencapai **46%**. Namun, XGBoost tetap unggul dalam keseimbangan antara kelas mayoritas dan minoritas.

Hyperparameter tuning memberikan peningkatan signifikan dalam stabilitas dan performa setiap model, terutama dengan mengurangi overfitting pada Decision Tree dan k-NN. Namun, untuk menangani dataset yang memiliki ketidakseimbangan kelas seperti ini, **XGBoost tetap menjadi model yang paling direkomendasikan** karena kemampuannya untuk menangkap pola kompleks dan memberikan hasil yang lebih seimbang di kedua kelas.