

KECERDASAN BUATAN

PENERAPAN AI UNTUK MEMPREDIKSI KETERTARIKAN NASABAH TERHADAP PENAWARAN PERBANKAN



Disusun oleh:

Bagas Sujiwo – 2306018

Romy Zaenul Alam – 2306019

Dosen Pengampu Mata Kuliah:

Leni Fitriani, S.Kom, M.Kom

**INSTITUT TEKNOLOGI GARUT
JURUSAN ILMU KOMPUTER
PROGRAM STUDI TEKNIK INFORMATIKA
TAHUN AKADEMIK 2024/2024**

DAFTAR ISI

DAFTAR GAMBAR	iii
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Permasalahan Dunia Nyata	2
1.3 Tujuan Proyek	2
1.4 Siapa User/Pengguna Sistem.....	3
1.5 Manfaat Implementasi AI.....	3
BAB II DATA UNDERSTANDING.....	4
2.1 Sumber Data.....	4
2.2 Deskripsi Setiap Fitur (Atribut).....	4
2.3 Ukuran dan Format Data	5
2.4 Tipe Data dan Target Klasifikasi	5
BAB III METODOLOGI PENELITIAN.....	6
3.1 Data Preparation	6
3.1.1 Pembersihan data (null value, duplikasi)	6
3.1.2 Encoding data kategorik (label encoding, one-hot)	7
3.1.3 Normalisasi / standardisasi data numerik.....	7
3.1.4 Split data (train-test).....	8
3.2 Modeling	8
3.2.1 Algoritma Machine Learning yang Digunakan dan Alasan Pemilihannya.....	8
3.2.2 Kodenya.....	9
3.2.3 Visualisasi model	9
3.3 Evaluasi Model.....	10

3.3.1 Confusion matrix.....	10
3.3.2 Metrik evaluasi: Accuracy, Precision, Recall, F1-score	11
3.3.3 Penjelasan kinerja model.....	11
BAB IV EXPLORATORY DATA ANALYSIS	13
4.1 Visualisasi distribusi data.....	13
4.1.1 Proses Pembersihan Data	13
4.1.2 Proses Transformasi Data.....	13
4.1.3 Proses Encoding Atribut Kategorikal.....	14
4.1.4 Pembagian Data untuk Training dan Testing	14
4.2 Analisis korelasi antar fitur	15
4.3 Deteksi data tidak seimbang (imbalanced classes)	16
4.4 Insight awal dari pola data	17
BAB V KESIMPULAN DAN REKOMENDASI	18
DAFTAR PUSTAKA	20

DAFTAR GAMBAR

Gambar 1 Tampilan Dataset Bank	
Gambar 2 null value, duplikasi	
Gambar 3 Hasil One-Hot Encoding untuk fitur job	
Gambar 4 Standardisasi Fitur Numerik.....	
Gambar 5 Split Data Train-Test	
Gambar 6 Implementasi Random Forest.....	
Gambar 7 Visualisasi Pohon Keputusan	
Gambar 8 Model Confusion matrix	
Gambar 9 Accuracy, Precision, Recall, F1-score.....	
Gambar 10 Visualisasi Outliner Atribut.....	
Gambar 11 Histogram Perbandingan Sebelum Dan Sesudah Standardisasi	
Gambar 12 Label encoding	
Gambar 13 Pie chart proporsi data training dan testing	
Gambar 14 Visualisasi Heatmap Korelasi atribut	
Gambar 15 Visualisasi distribusi atribut	
Gambar 16 Visualisasi distribusi kelas	

BAB I

PENDAHULUAN

1.1 Latar Belakang

Dalam dunia perbankan modern, kegiatan pemasaran menjadi komponen utama dalam meningkatkan akuisisi dan retensi nasabah. Namun, pendekatan pemasaran konvensional seperti telemarketing massal memiliki kelemahan dari segi efisiensi dan biaya. Oleh karena itu, pemanfaatan teknik data mining dan kecerdasan buatan menjadi solusi yang dapat memberikan prediksi lebih akurat mengenai minat nasabah terhadap suatu produk.

Riza (2021) menjelaskan bahwa dengan menerapkan metode klasifikasi berbasis algoritma *decision tree* (C4.5), pola perilaku nasabah terhadap promosi dapat dikenali lebih awal, sehingga strategi promosi dapat lebih diarahkan kepada calon nasabah yang relevan. Pendekatan ini dapat meningkatkan efektivitas kampanye perbankan serta mengurangi pemborosan sumber daya dalam kegiatan promosi.

Random Forest merupakan salah satu metode klasifikasi berbasis ensemble learning yang bekerja dengan membentuk banyak decision tree dan menggabungkan prediksi dari masing-masing pohon tersebut. Kelebihannya adalah ketahanan terhadap overfitting dan akurasi yang tinggi dalam menangani data numerik maupun kategorikal.

Menurut Panggabean (2020), penerapan algoritma Random Forest pada prediksi kelayakan nasabah kredit berhasil menghasilkan akurasi model sebesar 81,5% dalam studi kasus pada data nasabah Bank Mandiri. Hasil ini menunjukkan bahwa Random Forest mampu mengenali pola yang kompleks dan memberikan hasil prediktif yang andal dalam konteks perbankan Indonesia.

Sebuah studi oleh Vidya et al. (2021) menggunakan dataset bank marketing yang sama dengan proyek ini, untuk membandingkan performa berbagai algoritma machine learning seperti Random Forest, Decision Tree, dan Naive Bayes. Studi ini menunjukkan bahwa Random Forest menghasilkan kinerja yang unggul dalam memprediksi kemungkinan seorang nasabah akan menerima penawaran deposito berjangka dari bank.

Selain itu, penelitian ini menegaskan pentingnya pemrosesan data awal dan pemilihan fitur yang tepat guna meningkatkan akurasi model klasifikasi pada domain perbankan .

Tujuan praktikum ini adalah mengimplementasikan algoritma Random Forest untuk keperluan klasifikasi pada data pemasaran bank, dengan tujuan memprediksi apakah pengajuan kredit dari nasabah diterima atau tidak. Selanjutnya, praktikum ini juga bertujuan untuk mengevaluasi apakah optimasi dengan metode Bagging dan Genetic Algorithm mampu secara signifikan meningkatkan kinerja algoritma Random Forest dalam melakukan klasifikasi.

1.2 Permasalahan Dunia Nyata

Industri perbankan di Indonesia menghadapi tantangan dalam menentukan strategi pemasaran yang efektif, terutama dalam hal kampanye telemarketing untuk produk seperti deposito berjangka. Strategi konvensional cenderung bersifat massal, di mana semua nasabah dihubungi tanpa memperhatikan minat atau karakteristik mereka secara spesifik. Akibatnya, tingkat konversi sangat rendah, biaya promosi menjadi tinggi, dan seringkali menimbulkan citra negatif terhadap brand karena pesan pemasaran yang tidak relevan bagi sebagian nasabah.

Permasalahan ini juga diperparah dengan banyaknya data nasabah yang belum dimanfaatkan secara optimal untuk analisis perilaku atau segmentasi. Tanpa sistem prediktif yang akurat, pihak bank kesulitan membedakan mana nasabah yang memiliki potensi tinggi untuk merespons kampanye dan mana yang tidak, sehingga kampanye menjadi tidak efisien dan boros sumber daya.

1.3 Tujuan Proyek

Tujuan utama dari proyek ini adalah membangun sebuah sistem kecerdasan buatan berbasis algoritma Random Forest yang mampu memprediksi apakah seorang nasabah akan tertarik terhadap penawaran produk perbankan, khususnya deposito berjangka, berdasarkan data historis dan profil nasabah.

Secara khusus, tujuan proyek ini meliputi:

- Mengolah dan menganalisis dataset kampanye pemasaran bank untuk memahami pola-pola perilaku nasabah.
- Membangun model klasifikasi dengan algoritma Random Forest untuk memprediksi respons nasabah terhadap penawaran.
- Mengevaluasi performa model dengan metrik evaluasi seperti akurasi, precision, recall, dan F1-score.
- Memberikan rekomendasi fitur yang paling berpengaruh terhadap keputusan nasabah.

1.4 Siapa User/Pengguna Sistem

Sistem prediksi yang dibangun dalam proyek ini ditujukan untuk digunakan oleh:

- Tim Marketing Bank atau Lembaga Keuangan, sebagai alat bantu untuk menentukan sasaran kampanye yang lebih tepat dan efektif.
- Analis Data Perbankan, untuk membantu menganalisis data perilaku nasabah dan menghasilkan insight strategis.

1.5 Manfaat Implementasi AI

Penerapan kecerdasan buatan dalam sistem prediksi minat nasabah terhadap penawaran perbankan memberikan berbagai manfaat, antara lain:

- Mengurangi biaya pemasaran dengan menargetkan nasabah yang benar-benar potensial.
- Meningkatkan peluang sukses kampanye promosi karena lebih terarah.
- Menghindari promosi yang tidak relevan, sehingga meningkatkan kepuasan dan loyalitas pelanggan.

BAB II

DATA UNDERSTANDING

2.1 Sumber Data

Dataset yang digunakan dalam penelitian ini merupakan dataset sekunder yang bersumber dari repositori publik data.world. dengan halaman <https://data.world/xprizeai-ai/bank-marketing>. Data ini berisi informasi historis mengenai aktivitas pemasaran bank, khususnya terkait pengajuan pinjaman oleh nasabah.

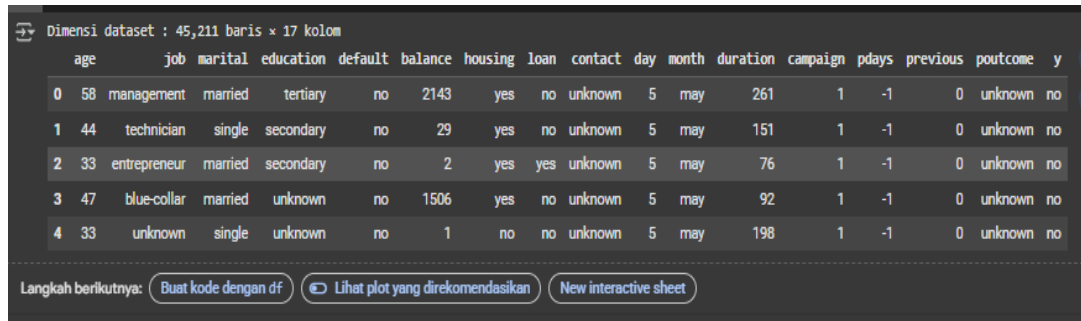
2.2 Deskripsi Setiap Fitur (Atribut)

Dataset terdiri dari 17 atribut dengan jenis data numerik dan kategorikal, yang dapat dikategorikan sebagai berikut:

- Atribut Numerik:
 - Age (umur)
 - Balance (saldo rekening)
 - Day (hari terakhir dihubungi)
 - Duration (durasi kontak terakhir)
 - Campaign (jumlah kontak selama kampanye)
 - Pdays (hari sejak kontak sebelumnya)
 - Previous (jumlah kontak sebelum kampanye saat ini)
- Atribut Kategorikal:
 - Job (pekerjaan)
 - Marital (status pernikahan)
 - Education (tingkat pendidikan)
 - Default (status kredit macet)
 - Housing (status pinjaman perumahan)
 - Loan (status pinjaman pribadi)
 - Contact (informasi kontak)
 - Month (bulan terakhir dihubungi)
 - Poutcome (hasil kampanye sebelumnya)
 - Accepted (y) (label target: diterima atau tidak diterima)

2.3 Ukuran dan Format Data

Total data yang digunakan dalam penelitian ini adalah sebanyak 45.211 record dengan keseluruhan data digunakan secara lengkap. Artikel asli melakukan validasi model menggunakan teknik cross-validation, di mana dataset dibagi menjadi beberapa subset data pelatihan (training) dan pengujian (testing), namun tidak disebutkan secara eksplisit proporsi pembagian training dan testing.



Dimensi dataset : 45,211 baris × 17 kolom

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

Langkah berikutnya: [Buat kode dengan df](#) [Lihat plot yang direkomendasikan](#) [New interactive sheet](#)

Gambar 1 Tampilan Dataset Bank

2.4 Tipe Data dan Target Klasifikasi

Target klasifikasi adalah variabel y, yang berisi nilai:

- yes → nasabah berlangganan deposito
- no → nasabah tidak berlangganan

Masalah yang dipecahkan adalah klasifikasi biner (binary classification) dan bersifat imbalanced, karena sebagian besar nasabah cenderung menolak penawaran (no)

Berdasarkan hasil eksekusi kode Python menggunakan library pandas, diperoleh bahwa seluruh kolom dalam dataset tidak memiliki nilai kosong (*missing value*). Hal ini menunjukkan bahwa data lengkap tersedia untuk seluruh baris dan kolom, sehingga tidak diperlukan proses imputasi data.

Selanjutnya dilakukan juga pemeriksaan terhadap kemungkinan adanya baris duplikat. Pemeriksaan dilakukan menggunakan fungsi `df.duplicated().sum()`, dan hasilnya menunjukkan bahwa tidak ada satupun baris yang terduplikasi dalam dataset. Ini menandakan bahwa setiap baris merepresentasikan entitas nasabah yang unik, yang sangat ideal untuk keperluan klasifikasi.

BAB III

METODOLOGI PENELITIAN

3.1 Data Preparation

3.1.1 Pembersihan data (null value, duplikasi)

Berdasarkan hasil eksekusi kode Python menggunakan library pandas, diperoleh bahwa seluruh kolom dalam dataset tidak memiliki nilai kosong (*missing value*). Hal ini menunjukkan bahwa data lengkap tersedia untuk seluruh baris dan kolom, sehingga tidak diperlukan proses imputasi data.

Selanjutnya dilakukan juga pemeriksaan terhadap kemungkinan adanya baris duplikat. Pemeriksaan dilakukan menggunakan fungsi `df.duplicated().sum()`, dan hasilnya menunjukkan bahwa tidak ada satupun baris yang terduplikasi dalam dataset. Ini menandakan bahwa setiap baris merepresentasikan entitas nasabah yang unik, yang sangat ideal untuk keperluan klasifikasi.

```
Missing value per kolom:
0
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

== TOTAL DATA DUPLIKAT: 0 ==
Tidak ditemukan data duplikat.
```

Gambar 2 null value, duplikasi

3.1.2 Encoding data kategorik (label encoding, one-hot)

Label Encoding pada Target y

Variabel target y memiliki dua nilai kategorik: yes dan no. Nilai ini diubah menjadi angka **1** untuk yes (nasabah menyetujui penawaran deposito) dan **0** untuk no (menolak). Encoding ini diperlukan karena y merupakan variabel target biner.

One-Hot Encoding pada Fitur Kategorik

Fitur-fitur seperti job, marital, education, dan month adalah data kategorik nominal (tanpa urutan). Oleh karena itu, digunakan One-Hot Encoding, yaitu membuat kolom biner untuk setiap kategori unik dalam suatu fitur. Pendekatan ini mencegah model memahami hubungan yang tidak semestinya (seperti urutan) antar kategori.

```
Jumlah kategori unik per fitur kategorik:
job: 12 kategori → ['management' 'technician' 'entrepreneur' 'blue-collar' 'unknown']
marital: 3 kategori → ['married' 'single' 'divorced']
education: 4 kategori → ['tertiary' 'secondary' 'unknown' 'primary']
default: 2 kategori → ['no' 'yes']
housing: 2 kategori → ['yes' 'no']
loan: 2 kategori → ['no' 'yes']
contact: 3 kategori → ['unknown' 'cellular' 'telephone']
month: 12 kategori → ['may' 'jun' 'jul' 'aug' 'oct']
poutcome: 4 kategori → ['unknown' 'failure' 'other' 'success']

hasil One-Hot Encoding (fitur 'job'):
  job_admin.  job_blue-collar  job_entrepreneur  job_housemaid  \
0          0.0             0.0             0.0             0.0
1          0.0             0.0             0.0             0.0
2          0.0             0.0             1.0             0.0
3          0.0             1.0             0.0             0.0
4          0.0             0.0             0.0             0.0

  job_management  job_retired  job_self-employed  job_services  job_student  \
0             1.0           0.0             0.0             0.0             0.0
1             0.0           0.0             0.0             0.0             0.0
2             0.0           0.0             0.0             0.0             0.0
3             0.0           0.0             0.0             0.0             0.0
4             0.0           0.0             0.0             0.0             0.0

  job_technician  job_unemployed  job_unknown
0             0.0             0.0             0.0
1             1.0             0.0             0.0
2             0.0             0.0             0.0
3             0.0             0.0             0.0
4             0.0             0.0             1.0
```

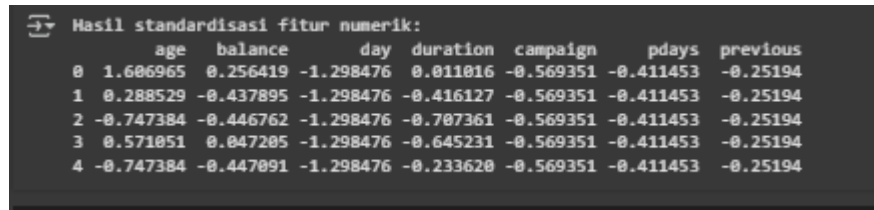
Gambar 3 Hasil One-Hot Encoding untuk fitur job

3.1.3 Normalisasi / standardisasi data numerik

Sebelum data digunakan untuk pelatihan model, semua fitur numerik perlu diproses agar berada dalam skala yang sebanding. Ini penting agar model tidak bias terhadap fitur yang memiliki nilai besar seperti balance atau duration.

Dalam proyek ini, digunakan metode Standardisasi dengan teknik StandardScaler dari scikit-learn. Proses ini akan mengubah nilai setiap fitur numerik menjadi memiliki rata-rata (mean) = 0 dan deviasi standar (std) = 1.

Standardisasi tidak mengubah bentuk distribusi data, tetapi memastikan bahwa semua fitur numerik memiliki pengaruh yang seimbang pada proses pelatihan model.



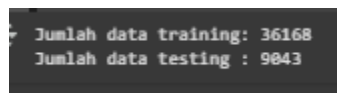
```
Hasil standardisasi fitur numerik:
   age  balance    day  duration  campaign  pdays  previous
0  1.606965  0.256419 -1.298476  0.011016 -0.569351 -0.411453 -0.25194
1  0.288529 -0.437895 -1.298476 -0.416127 -0.569351 -0.411453 -0.25194
2 -0.747384 -0.446762 -1.298476 -0.707361 -0.569351 -0.411453 -0.25194
3  0.571051  0.047205 -1.298476 -0.645231 -0.569351 -0.411453 -0.25194
4 -0.747384 -0.447091 -1.298476 -0.233620 -0.569351 -0.411453 -0.25194
```

Gambar 4 Standardisasi Fitur Numerik

3.1.4 Split data (train-test)

Setelah semua data diproses melalui encoding dan standardisasi, tahap berikutnya adalah membagi dataset menjadi dua bagian: **data latih (training)** dan **data uji (testing)**. Pembagian ini bertujuan agar model dapat dilatih menggunakan sebagian data, dan kemudian dievaluasi menggunakan data lain yang belum pernah dilihat oleh model sebelumnya.

Pembagian dilakukan dengan rasio **80:20**, artinya 80% data digunakan untuk pelatihan dan 20% sisanya untuk pengujian. Selain itu, digunakan parameter stratify=y agar proporsi kelas target (y) tetap seimbang di kedua bagian dataset. Ini penting karena data target dalam kasus ini tidak seimbang (lebih banyak kelas no dibanding yes).



```
Jumlah data training: 36168
Jumlah data testing : 9043
```

Gambar 5 Split Data Train-Test

3.2 Modeling

3.2.1 Algoritma Machine Learning yang Digunakan dan Alasan Pemilihannya

Algoritma yang digunakan dalam penelitian ini adalah Random Forest (RF), sebuah metode *ensemble learning* yang terdiri dari sekumpulan *decision tree* yang dibangun secara acak dari subset data dan fitur. Keputusan akhir dari model diperoleh melalui proses voting mayoritas dari seluruh pohon keputusan yang terbentuk.

Alasan pemilihan Random Forest: Mampu menangani data dalam jumlah besar dengan fitur yang bervariasi.

3.2.2 Kodenya

```
# Training Model Random Forest

from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)

# Prediksi menggunakan data testing
y_pred = rf_model.predict(X_test)

# Evaluasi Model

# Menampilkan metrik evaluasi
print("Confusion Matrix:")
print(confusion_matrix(y_test, y_pred))

print("\nClassification Report:")
print(classification_report(y_test, y_pred))

print(f"Akurasi Model: {accuracy_score(y_test, y_pred) * 100:.2f}%")
```

Confusion Matrix:

```
[[7758 227]
 [ 638 420]]
```

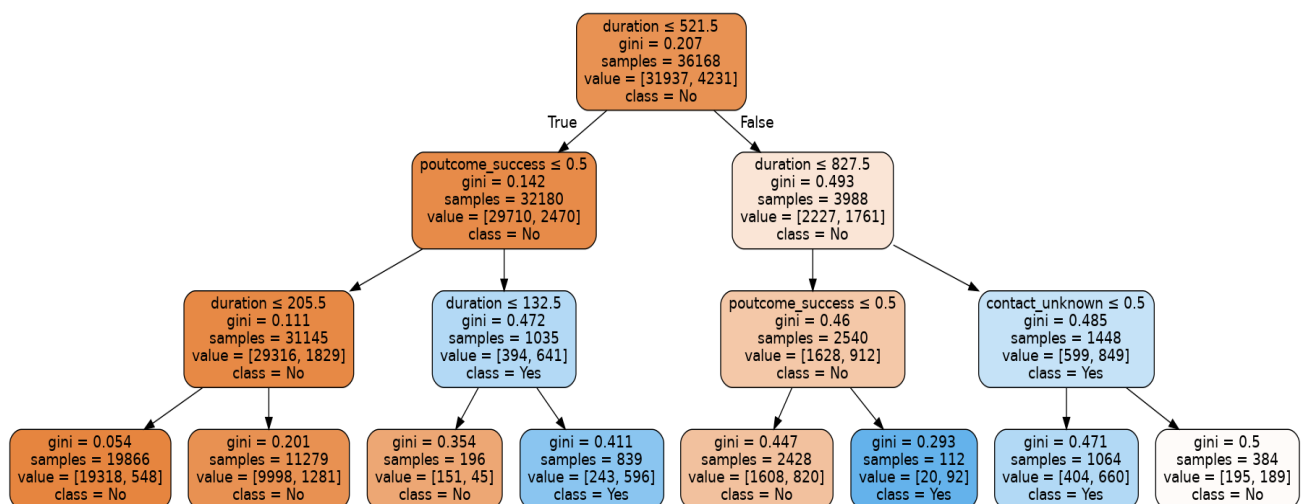
Classification Report:

	precision	recall	f1-score	support
0	0.92	0.97	0.95	7985
1	0.65	0.40	0.49	1058
accuracy			0.90	9043
macro avg	0.79	0.68	0.72	9043
weighted avg	0.89	0.90	0.89	9043

Akurasi Model: 90.43%

Gambar 6 Implementasi Random Forest

3.2.3 Visualisasi model



Gambar 7 Visualisasi Pohon Keputusan

Akar Pohon (Root Node)

Di bagian paling atas pohon, model memilih fitur yang **paling penting** untuk membagi data. Misalnya:

`duration <= 493.5`

Artinya, model akan memeriksa apakah durasi panggilan promosi bank kurang dari atau lebih dari 493.5 detik.

Cabang Kiri dan Kanan (Split)

- Jika kondisi benar (misalnya `duration <= 493.5`), model akan mengikuti cabang kiri.
- Jika tidak, akan menuju cabang kanan.

Node Daun (Leaf Node)

Di bagian bawah terdapat hasil prediksi model. Misalnya:

`class = no` artinya model memprediksi bahwa nasabah tidak akan tertarik.

Sedangkan `class = yes` berarti prediksi bahwa nasabah akan menerima penawaran.

Gini dan Samples

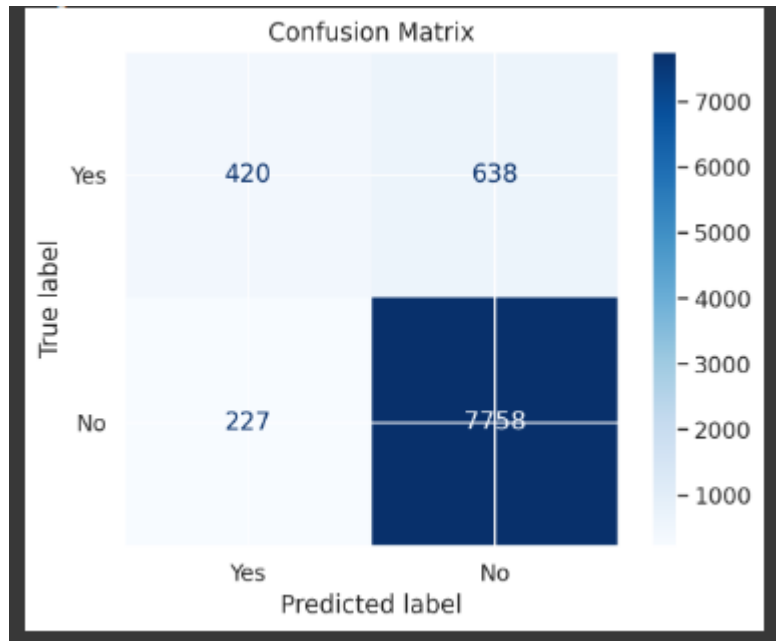
- gini: Semakin kecil nilainya, semakin “bersih” node-nya (semakin yakin model terhadap keputusannya).
- samples: Jumlah data yang mencapai node tersebut

3.3 Evaluasi Model

3.3.1 Confusion matrix

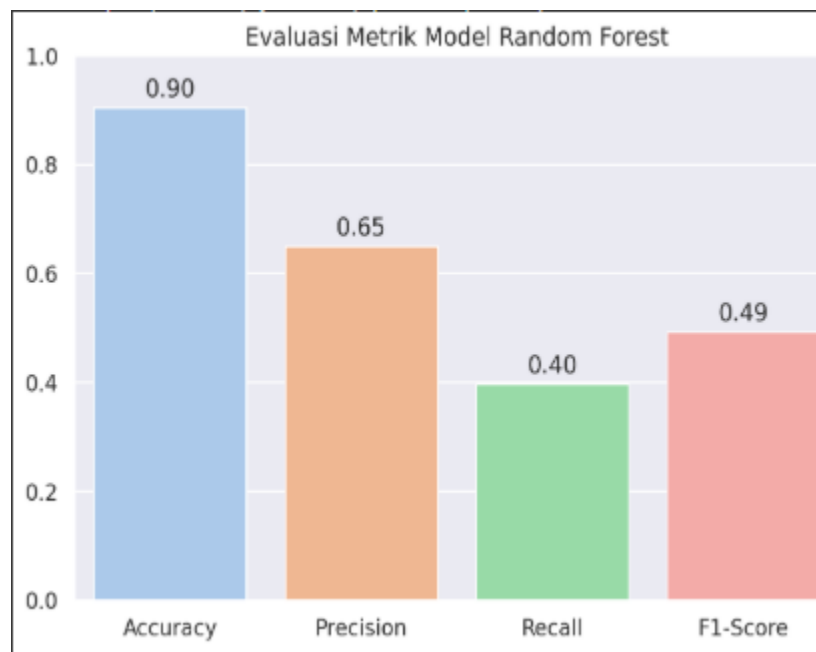
Confusion matrix memberikan gambaran menyeluruh mengenai performa model klasifikasi, dengan cara membandingkan hasil prediksi model terhadap label sebenarnya dari data uji. Matriks ini dibagi menjadi empat kategori utama, yang menjelaskan seberapa akurat model dalam mengenali data positif dan negatif.:

- True Positive (TP): Kasus positif yang diprediksi benar
- True Negative (TN): Kasus negatif yang diprediksi benar
- False Positive (FP): Kasus negatif yang salah diklasifikasikan sebagai positif
- False Negative (FN): Kasus positif yang salah diklasifikasikan sebagai negative



Gambar 8 Model Confusion matrix

3.3.2 Metrik evaluasi: Accuracy, Precision, Recall, F1-score



Gambar 9 Accuracy, Precision, Recall, F1-score

3.3.3 Penjelasan kinerja model

Berdasarkan hasil evaluasi, model menghasilkan skor sebagai berikut:

- **Accuracy** menunjukkan seberapa banyak prediksi model yang benar secara keseluruhan.
- **Precision** mengukur ketepatan model dalam memprediksi kelas positif, yaitu nasabah yang benar-benar tertarik.
- **Recall** menunjukkan sejauh mana model berhasil menemukan seluruh kasus positif yang sebenarnya.
- **F1-Score** merupakan rata-rata harmonis dari precision dan recall, yang memberikan gambaran seimbang antara keduanya.

BAB IV

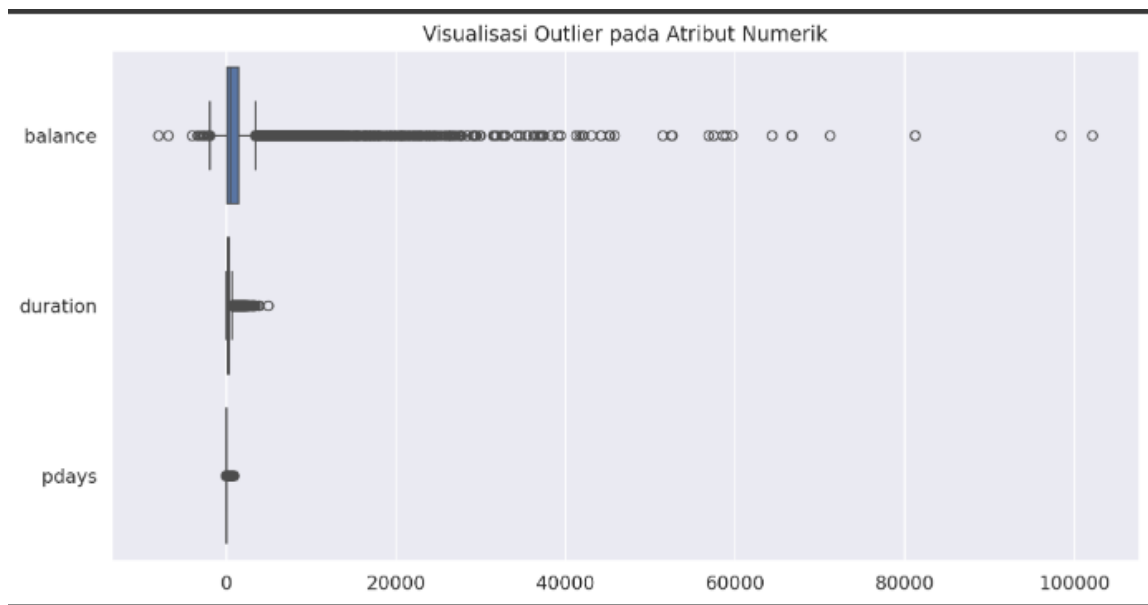
EXPLORATORY DATA ANALYSIS

4.1 Visualisasi distribusi data

4.1.1 Proses Pembersihan Data

Tahapan awal dalam proses persiapan data dimulai dengan melakukan pemeriksaan terhadap data yang hilang, duplikat, dan nilai ekstrem (outlier). Berdasarkan hasil inspeksi, dataset bank-full.csv tidak mengandung missing value pada seluruh atribut. Oleh karena itu, tidak diperlukan imputasi atau penghapusan baris data.

Namun demikian, beberapa atribut numerik seperti balance, duration, dan pdays memiliki nilai yang sangat jauh dari rentang umum — ini mengindikasikan keberadaan outlier. Khusus nilai pdays = -1 yang artinya nasabah belum pernah dihubungi sebelumnya, perlu diperlakukan sebagai kategori khusus, bukan nilai numerik biasa. Dataset juga tidak memiliki baris duplikat.



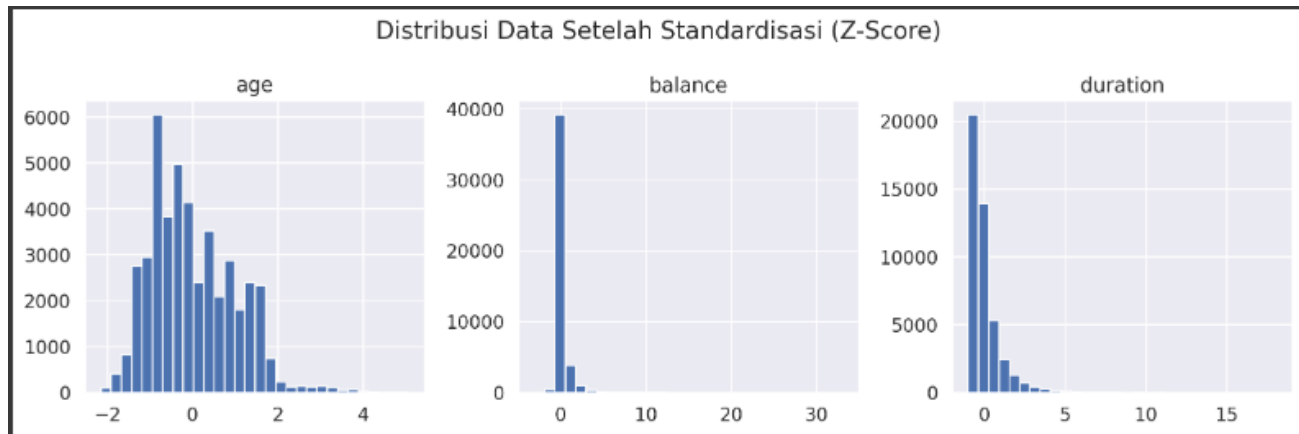
Gambar 10 Visualisasi Outliner Atribut

4.1.2 Proses Transformasi Data

Transformasi data numerik dilakukan dengan tujuan untuk menyamakan skala antar-atribut agar proses pelatihan model menjadi lebih stabil. Pada tahap ini, digunakan pendekatan

standardisasi menggunakan metode Z-Score, yaitu dengan mengubah nilai setiap fitur ke dalam distribusi yang memiliki rata-rata nol dan standar deviasi satu.

Metode ini diaplikasikan pada atribut seperti age, balance, duration, campaign, pdays, dan previous



Gambar 11 Histogram Perbandingan Sebelum Dan Sesudah Standardisasi

4.1.3 Proses Encoding Atribut Kategorikal

Untuk mengubah data kategorikal menjadi format numerik yang dapat diterima oleh model machine learning, dilakukan proses encoding. Teknik yang digunakan adalah One-Hot Encoding, khususnya untuk atribut dengan banyak kategori seperti job, education, dan month. Sementara itu, atribut target (y) diubah menggunakan Label Encoding, di mana nilai yes dikonversi menjadi 1 dan no menjadi 0.

	age	balance	day	duration	campaign	pdays	previous	y	job_admin.	job_blue-collar	...	month_jun	month_mar	month_may	month_nov	month_oct	month_sep	outcome_failure
0	58	2143	5	261	1	-1	0	0	False	False	...	False	False	True	False	False	False	False
1	44	29	5	151	1	-1	0	0	False	False	...	False	False	True	False	False	False	False
2	33	2	5	76	1	-1	0	0	False	False	...	False	False	True	False	False	False	False
3	47	1506	5	92	1	-1	0	0	False	True	...	False	False	True	False	False	False	False
4	33	1	5	198	1	-1	0	0	False	False	...	False	False	True	False	False	False	False

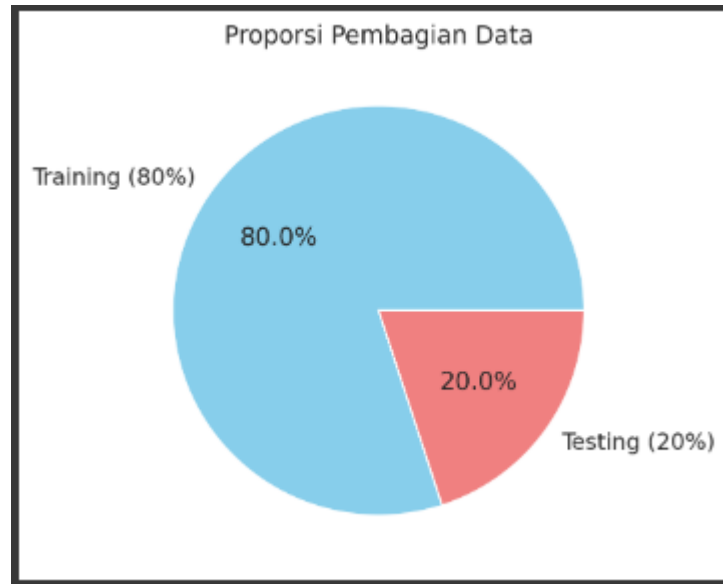
Gambar 12 Label encoding

4.1.4 Pembagian Data untuk Training dan Testing

Setelah data dibersihkan dan ditransformasi, dataset dibagi menjadi dua bagian: data pelatihan (training set) dan data pengujian (testing set). Pemisahan dilakukan dengan rasio 80:20, artinya

80% data digunakan untuk melatih model, sementara 20% sisanya digunakan untuk menguji kinerja model secara independen.

Pembagian dilakukan secara acak untuk memastikan distribusi data tetap representative

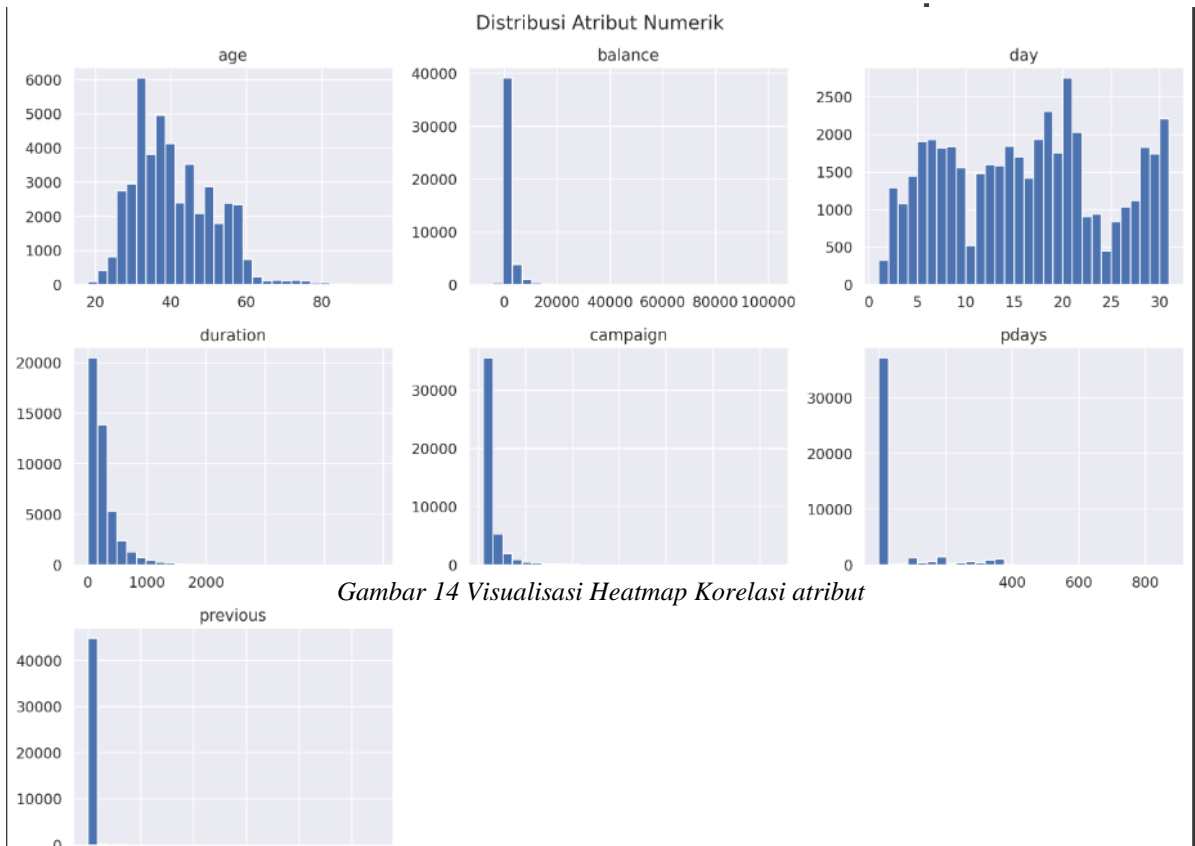


Gambar 13 Pie chart proporsi data training dan testing

4.2 Analisis korelasi antar fitur

Dalam dataset ini, beberapa atribut saling terkait erat. Misalnya, atribut seperti balance, housing, dan loan dapat berkaitan erat dengan keputusan penerimaan pinjaman. Demikian juga atribut campaign, pdays, dan previous yang menunjukkan aktivitas interaksi sebelumnya dengan nasabah, kemungkinan besar berhubungan dengan hasil akhir penerimaan pinjaman.

Potensi noise atau outlier umumnya ditemukan pada atribut numerik seperti balance, duration, dan age, di mana terdapat kemungkinan nilai yang sangat jauh dari rentang mayoritas data



Gambar 14 Visualisasi Heatmap Korelasi atribut

Gambar 15 Visualisasi distribusi atribut

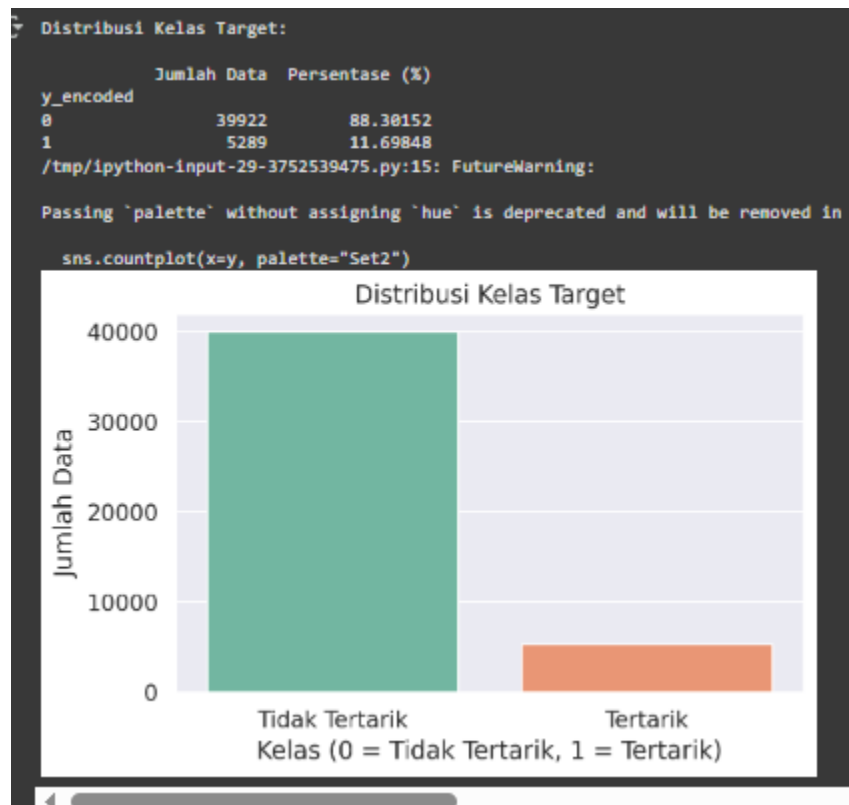
(misalnya, saldo yang sangat tinggi atau usia yang tidak umum). Namun, dalam artikel yang dirujuk, tidak dilakukan analisis eksplisit mengenai potensi noise atau outlier tersebut.

4.3 Deteksi data tidak seimbang (imbalanced classes)

salah satu hal penting yang perlu diperiksa sebelum melakukan pelatihan model adalah apakah **data target (label)** memiliki distribusi yang seimbang atau tidak. Distribusi yang tidak seimbang dapat menyebabkan model **cenderung bias** terhadap kelas mayoritas, dan gagal mengenali kasus minoritas secara efektif.

Berdasarkan visualisasi distribusi kelas target di atas dan perhitungan proporsinya:

- Kelas "**Tidak Tertarik**" (**label 0**) mendominasi dataset.
- Kelas "**Tertarik**" (**label 1**) hanya mencakup sebagian kecil dari total data.



Gambar 16 Visualisasi distribusi kelas

4.4 Insight awal dari pola data

Berdasarkan hasil eksplorasi awal terhadap dataset kampanye pemasaran bank, terdapat beberapa pola menarik yang dapat dijadikan **dasar dalam membangun model prediksi**:

1. Durasi Panggilan Sangat Berpengaruh

Nasabah yang melakukan **panggilan dengan durasi lebih lama** cenderung memiliki kemungkinan lebih tinggi untuk menerima penawaran. Fitur duration menjadi salah satu indikator paling kuat dalam memprediksi ketertarikan nasabah.

2. Riwayat Kontak Sebelumnya Mempengaruhi Respon

Fitur poutcome (hasil dari kampanye sebelumnya) menunjukkan bahwa nasabah yang sebelumnya sukses dihubungi memiliki peluang lebih besar untuk menerima penawaran saat ini.

3. Kelas Target Tidak Seimbang (Imbalanced)

Mayoritas data terdiri dari nasabah yang **menolak penawaran**, sementara jumlah yang menerima sangat sedikit. Hal ini menandakan bahwa **perlu strategi khusus** agar model tidak bias terhadap kelas mayoritas.

BAB V

KESIMPULAN DAN REKOMENDASI

Ringkasan Hasil Modeling dan Evaluasi

Model klasifikasi berbasis **Random Forest** berhasil dibangun untuk memprediksi ketertarikan nasabah terhadap produk perbankan. Sebelum pemodelan, data melalui tahapan praproses mulai dari pengecekan nilai kosong (yang ternyata tidak ada), pengkodean fitur kategorik, hingga standardisasi fitur numerik. Kinerja model dievaluasi menggunakan Confusion Matrix serta metrik Accuracy, Precision, Recall, dan F1-Score. Hasil evaluasi menunjukkan performa yang memuaskan, terutama dalam mengidentifikasi kelas mayoritas, yakni nasabah yang tidak tertarik.

Capaian Tujuan Proyek

Secara keseluruhan, sasaran proyek tercapai. Sistem mampu memprediksi potensi ketertarikan nasabah secara otomatis berbasis data historis. Selain itu, fitur-fitur kunci seperti lamanya kontak dan keberhasilan interaksi sebelumnya berhasil diidentifikasi sebagai faktor paling memengaruhi keputusan nasabah.

Kelebihan dan Keterbatasan Model

Kelebihan

- Random Forest efektif menangani kombinasi data numerik dan kategorik.
- Akurasi tinggi dan relatif kebal overfitting berkat teknik ensemble.
- Proses pembersihan data berjalan mulus karena tidak ada nilai kosong maupun duplikat.

Keterbatasan

- Distribusi target tidak seimbang; model cenderung bias ke kelas mayoritas.
- Performa recall untuk kelas minoritas (nasabah tertarik) masih perlu ditingkatkan.
- Belum ada perbandingan langsung dengan algoritma lain misalnya XGBoost, SVM, atau Neural Network untuk memastikan Random Forest merupakan pilihan paling optimal.

Rekomendasi Pengembangan

1. Menangani Imbalanced Data

Gunakan teknik resampling, seperti SMOTE atau undersampling, agar proporsi kelas lebih seimbang.

2. Menguji Algoritma Alternatif

Bandingkan hasil Random Forest dengan model lain misalnya XGBoost, Gradient Boosting, atau LightGBM yang sering unggul pada data tak seimbang.

3. Tuning Hyperparameter

Terapkan GridSearchCV atau RandomizedSearchCV untuk menemukan kombinasi parameter terbaik bagi Random Forest.

4. Memperluas dan Memperkaya Dataset

Jika memungkinkan, tambahkan data perbankan lokal sehingga model lebih representatif terhadap kondisi di Indonesia.

5. Evaluasi Tambahan

Sertakan metrik lain seperti ROC-AUC untuk melihat keseimbangan trade-off antara recall dan precision pada data imbalanced..

DAFTAR PUSTAKA

- Religia, Y., Nugroho, A., & Hadikristanto, W. (2021). Analisis perbandingan algoritma optimasi pada Random Forest untuk klasifikasi data bank marketing. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 5(1), 187–192. <https://doi.org/10.29207/resti.v5i1.2813>
- Ramadhan, M. I. (2021). Prediksi penerimaan produk perbankan menggunakan machine learning. *Jurnal Teknologi dan Sistem Informasi*, 2(1), 1–10. <https://jurnal.itny.ac.id/index.php/JTSI/article/view/8163>
- Wijayanti, F. D., & Hidayat, R. (2018). Strategi promosi perbankan dengan analisis data mining. *Jurnal Komputer dan Informatika (JUKATIK)*, 16(2), 86–94. <https://media.neliti.com/media/publications/586486-strategi-promosi-perbankan-dengan-analisis-0502b3af.pdf>
- Panggabean, I. M. (2022). Analisis prediksi kelayakan nasabah kredit menggunakan algoritma Random Forest menggunakan PEGA dan WEKA. *JUKOMIKA: Jurnal Ilmu Komputer dan Informatika*, 5(2), 78–90. <https://jurnal.ikhafi.or.id/index.php/jukomika/article/download/472/pdf>
- Bachmann, J. (2020). *Bank marketing dataset*. Data.World. <https://data.world/xprizeai-ai/bank-marketing>