PERBANDINGAN KINERJA NAIVE BAYES DAN RANDOM FOREST DENGAN VARIASI SELEKSI FITUR DAN PENANGANAN *IMBALANCE DATA*

SKRIPSI

Program Studi Sarjana Informatika
Jurusan Informatika

Oleh:
FORTUNATUS ADHIETHERA TUAH PUTRA
NIM D1041211048

FAKULTAS TEKNIK
UNIVERSITAS TANJUNGPURA
PONTIANAK
2025

HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini:

Nama : Fortunatus Adhiethera Tuah Putra

NIM : D1041211048

menyatakan bahwa dalam skripsi yang berjudul "Perbandingan Kinerja Naive Bayes Dan Random Forest Dengan Variasi Seleksi Fitur Dan Penanganan *Imbalance Data*" tidak terdapat karya yang pernah diajukan untuk memperoleh gelar sarjana di suatu perguruan tinggi manapun. Sepanjang pengetahuan Saya, tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam Daftar Pustaka.

Demikian pernyataan ini dibuat dengan sebenar-benarnya. Saya sanggup menerima konsekuensi akademis dan hukum di kemudian hari apabila pernyataan yang dibuat ini tidak benar.

Pontianak, 31 Juli 2025

Fortunatus Adhiethera Tuah Putra
NIM D1041211048

KEMENTERIAN PENDIDIKAN TINGGI, SAINS, DAN TEKNOLOGI
UNIVERSITAS TANJUNGPURA
FAKULTAS TEKNIK
Jalan Prof. Dr. H. Hadari Nawawi Pontianak 78124
Telepon (0561) 740186, WA: +6282152280907
Email : ft@untan.ac.id Website : http://teknik.untan.ac.id

HALAMAN PENGESAHAN

PERBANDINGAN KINERJA NAIVE BAYES DAN RANDOM FOREST DENGAN VARIASI SELEKSI FITUR DAN PENANGANAN *IMBALANCE DATA*

Program Studi Sarjana Informatika
Jurusan Informatika

Oleh:

Fortunatus Adhiethera Tuah Putra
NIM D1041211048

Telah dipertahankan di depan Penguji Skripsi pada tanggal 31 Juli 2025
dan diterima sebagai salah satu persyaratan untuk memperoleh gelar sarjana

Susunan Penguji Skripsi:

Ketua,

Dr. Arif Bijaksana Putra Negara, S.T., M.T.
NIP. 197208081998021002

Penguji Utama,

Tursina, S.T., M.Cs.
NIP. 197801152002122003

Sekretaris,

Helen Sastypratiwi, S.T., M.Eng.
NIP. 198601172012122004

Penguji Pendamping,

Rina Septiriana, S.T., M.Cs.
NIP. 198709232020122001

Pontianak, 31 Juli 2025
Dekan,

Dr. Ing. Ir. Slamet Widodo, M.T., IPM
NIP. 196712231992031002

HALAMAN PERSEMBAHAN

Skripsi ini didedikasikan kepada kedua orang tuaku (Alm. Bapak Paulus Auk dan Ibu Birgita) dan saudara-saudaraku (Teras Rangga Binua Tuah Putra dan Alm. Rafael Tarigas Tuah Putra). Apa yang mereka telah berikan melebihi daripada apa yang pernah aku inginkan.

KATA PENGANTAR

Puji dan syukur kepada Tuhan Yang Maha Esa yang telah memberi berkat serta karunia-Nya kepada penulis sehingga dapat menyelesaikan laporan skripsi dengan judul "Perbandingan Algoritma Klasifikasi Naive Bayes dan Random Forest Terhadap Variasi Seleksi Fitur dan Penanganan *Imbalance* Data" sebagai syarat untuk menyelesaikan program sarjana (S1) pada Fakultas Teknik Universitas Tanjungpura.

Penulis menyadari bahwa tulisan ini tidak dapat terselesaikan tanpa dukungan dari berbagai pihak moril dan materil. Oleh karena itu, penulis ingin menyampaikan terima kasih kepada : Bapak Dr. Arif Bijaksana Putra Negara, S.T. M.T. selaku Dosen Pembimbing Utama dan Ibu Helen Sastypratiwi, S.T. M.Eng selaku Dosen Pembimbing Pendamping serta Ibu Tursina, S.T. M.Cs selaku Dosen Penguji Utama dan Ibu Rina Septiriana, S.T. M.Cs selaku Dosen Penguji Pendamping.

Penguji sadar bahwa skripsi ini tidak sempurna dan masih memilik banyak keterbatasan. Oleh karena itu, saran, kritik, dan masukan dari pihak-pihak yang berkepentingan sangat dihargai untuk perbaikan dan pengembangan lebih lanjut.

Dengan rendah hati, penulis berharap bahwa hasil penelitian ini dapat memberikan sumbangan kecil dalam pembelajaran dan pemahaman lebih mendalam mengenai pemilihan fitur dan penyeimbangan data dalam klasifikasi. Semoga apa yang telah penulis pelajari dan tuliskan dapat menjadi pijakan untuk mengembangkan iman dan pengetahuan yang lebih baik.

Pontianak, 31 Juli 2025
Penulis,

Fortunatus Adhiethera Tuah Putra

ABSTRAK

Data mining adalah proses penting untuk mengekstraksi informasi berharga dari kumpulan data besar dan kompleks. Salah satu teknik utamanya adalah klasifikasi, yang digunakan untuk memprediksi kategori data berdasarkan fitur tertentu. Penelitian ini membandingkan performa algoritma Naive Bayes dan Random Forest dalam mengatasi klasifikasi pada data tidak seimbang dengan berbagai skenario pemilihan fitur. Dataset yang digunakan adalah *Bank Marketing* dari UCI Machine Learning Repository yang memiliki karakteristik tidak seimbang, dengan perbandingan signifikan antara data berlabel "yes" dan "no". Penelitian ini mengevaluasi pengaruh teknik penyeimbangan data (SMOTE, undersampling) dan metode pemilihan fitur (Chi-square, Pearson's Correlation, Mutual Information), dengan evaluasi berdasarkan akurasi, presisi, *recall,* dan *F1-score*. Hasil penelitian menunjukkan bahwa Random Forest secara konsisten unggul dibandingkan Naive Bayes, terutama pada kombinasi SMOTE dan seluruh fitur yang menghasilkan akurasi tertinggi 93,08% (peningkatan 5,18% dari *baseline*). Sebaliknya, kinerja terendah diperoleh Naive Bayes pada kombinasi Mutual Information dan *undersampling* dengan akurasi 81,9% (penurunan 9,88%). Teknik SMOTE dan *undersampling* meningkatkan akurasi Random Forest masing-masing sebesar 5,18% dan 1,44%, namun menurunkan akurasi Naive Bayes sebesar 9,25% dan 8,42%. Pemilihan fitur tertentu, seperti Mutual Information dan Pearson's Correlation, umumnya berdampak positif pada akurasi Random Forest (hingga 4,99%), tetapi efeknya pada Naive Bayes bervariasi. Secara keseluruhan, Random Forest menunjukkan kinerja yang lebih stabil (87,51%–93,08%) dibandingkan Naive Bayes yang cenderung fluktuatif (82,47%–91,78%), membuktikan ketangguhannya dalam menghadapi ketidakseimbangan data dan variasi pemilihan fitur.

Kata kunci: *data mining*, klasifikasi, naive bayes, pemilihan fitur, random forest, SMOTE, *undersampling*

ABSTRACT

Data mining is a crucial process for extracting valuable information from large and complex datasets. One of its primary techniques is classification, which is used to predict data categories based on specific features. This study compares the performance of the Naive Bayes and Random Forest algorithms in handling classification tasks on imbalanced data under various feature selection scenarios. The dataset used is the Bank Marketing dataset from the UCI Machine Learning Repository, which is inherently imbalanced, with a significant disparity between the "yes" and "no" labels. This research evaluates the impact of data balancing techniques (SMOTE, undersampling) and feature selection methods (Chi-square, Pearson's Correlation, Mutual Information), with performance assessed using accuracy, precision, recall, and F1-score. The results indicate that Random Forest consistently outperforms Naive Bayes, particularly in the combination of SMOTE and all features, achieving the highest accuracy of 93.08% (an improvement of 5.18% from the baseline). Conversely, the lowest performance was observed in Naive Bayes with the combination of Mutual Information and undersampling, yielding an accuracy of 81.9% (a decrease of 9.88%). SMOTE and undersampling improved Random Forest accuracy by 5.18% and 1.44%, respectively, but reduced Naive Bayes accuracy by 9.25% and 8.42%. Certain feature selection methods, such as Mutual Information and Pearson's Correlation, generally had a positive effect on Random Forest accuracy (up to 4.99%), whereas their impact on Naive Bayes was more variable. Overall, Random Forest demonstrated more stable performance (87.51%–93.08%) compared to the more fluctuating results of Naive Bayes (82.47%–91.78%), underscoring its robustness in dealing with data imbalance and feature selection variations.

Keywords: data mining, classification, feature selection, naive bayes, random forest, SMOTE, undersampling

DAFTAR ISI

HALAMAN PERNYATAAN ii
HALAMAN PENGESAHAN iii
HALAMAN PERSEMBAHAN iv
KATA PENGANTAR v
ABSTRAK vi
ABSTRACT vii
DAFTAR ISI viii
DAFTAR TABEL x
DAFTAR GAMBAR xi
DAFTAR LAMPIRAN xiii
DAFTAR KODE PROGRAM xiv
BAB I PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Perumusan Masalah 4
1.3 Tujuan Penelitian 4
1.4 Pembatasan Masalah 4
1.5 Sistematika Penulisan 5
BAB II TINJAUAN PUSTAKA 6
2.1 Penelitian Terdahulu 6
2.2 Landasan Teori 10
2.2.1 Data Mining 10
2.2.2 Klasifikasi 10
2.2.3 Imbalance Data 11
2.2.4 Synthetic Minority Over-sampling Technique (SMOTE) 11
2.2.5 Undersampling 11
2.2.6 Chi-Square 11
2.2.7 Pearson's Correlation 12
2.2.8 Naive Bayes 12
2.2.9 Random Forest 13
2.2.10 Mutual Information 13
2.2.11 Python 14
2.2.12 Confusion Matrix 14
2.2.13 Subspace Outlier Detection 15
BAB III METODOLOGI PENELITIAN 16
3.1 Data Penelitian 16
3.2 Perangkat Penelitian 16
3.3 Metodologi Penelitian 16
3.3.1 Pengumpulan Data 19
3.3.2 Exploratory Data Analysis (EDA) 23
3.3.3 Data Cleaning 24

3.3.4 Penanganan Imbalanced Data 25
3.3.5 Pemilihan Fitur 26
3.3.6 Implementasi Model 29
3.3.7 Evaluasi 33
3.3.8 Analisis Hasil 36

BAB IV HASIL DAN ANALISIS 37

4.1 Exploratory Data Analysis (EDA) 37
4.2 Data Cleaning 39
4.3 Penanganan Imbalanced Data 40
4.4 Pemilihan Fitur 45
4.4.1 Chi-square 45
4.4.2 Pearson's Correlation 47
4.4.3 Mutual Information 48
4.5 Implementasi Model 50
4.6 Evaluasi 55
4.7 Analisis Hasil 79

BAB V KESIMPULAN DAN SARAN 92

5.1 Kesimpulan 92
5.2 Saran 92

DAFTAR PUSTAKA 94

DAFTAR TABEL

Tabel 2. 1 Penelitian Terdahulu 8
Tabel 3. 1 Skenario Penelitian 18
Tabel 3. 2 Skenario Penelitian 18
Tabel 3. 3 Penjelasan Dataset 21
Tabel 3. 4 Hyperparameter yang disetel 33
Tabel 3. 5 Penjelasan Skenario 35
Tabel 3. 6 Hasil yang diharapkan 35
Tabel 4. 1 Perubahan Data Setelah Cleaning 40
Tabel 4. 2 Hasil Penyeimbangan Data 44
Tabel 4. 3 Nilai P Setiap Fitur Kategorial 46
Tabel 4. 4 Nilai Korelasi Pearson 47
Tabel 4. 5 Mutual Information 49
Tabel 4. 6 Hyperparameter Naive Bayes 51
Tabel 4. 7 Hyperparameter Random Forest 52
Tabel 4. 8 Hasil Penelitian 76
Tabel 4. 9 Perbandingan Hasil Evaluasi Terhadap Skenario Default 79
Tabel 4. 10 Perbandingan Pemilihan Fitur NB + SMOTE 82
Tabel 4. 11 Perbandingan Pemilihan Fitur NB 83
Tabel 4. 12 Perbandingan Pemilihan Fitur NB + Undersampling 84
Tabel 4. 13 Pengaruh Pemilihan Ftur RF + SMOTE 86
Tabel 4. 14 Pengaruh Pemilihan Fitur RF 87
Tabel 4. 15 Pengaruh Pemilihan Fitur RF + Undersampling 88

DAFTAR GAMBAR

Gambar 3. 1 Metodologi Penelitian 17
Gambar 3. 2 Proses SOD 24
Gambar 3. 3 Proses SMOTE 25
Gambar 3. 4 Proses Undersampling 26
Gambar 3. 5 Proses Chi-square 27
Gambar 3. 6 Proses Pearson's correlation 28
Gambar 3. 7 Proses Mutual Information 29
Gambar 4. 1 Karakteristik Data 38
Gambar 4. 2 Hasil Pengecekan Outlier dan Duplikat 39
Gambar 4. 3 Hasil Pembersihan Data 40
Gambar 4. 4 Perbandingan data sebelum dan setelah penyeimbangan 45
Gambar 4. 5 Confusion Matrix Naive Bayes Skenario 1 56
Gambar 4. 6 Confusion Matrix Naive Bayes Skenario 2 56
Gambar 4. 7 Confusion Matrix Naive Bayes Skenario 3 57
Gambar 4. 8 Confusion Matrix Naive Bayes Skenario 4 58
Gambar 4. 9 Confusion Matrix Naive Bayes Skenario 5 58
Gambar 4. 10 Confusion Matrix Naive Bayes Skenario 6 59
Gambar 4. 11 Confusion Matrix Naive Bayes Skenario 7 60
Gambar 4. 12 Confusion Matrix Naive Bayes Skenario 8 60
Gambar 4. 13 Confusion Matrix Naive Bayes Skenario 9 61
Gambar 4. 14 Confusion Matrix Naive Bayes Skenario 10 62
Gambar 4. 15 Confusion Matrix Naive Bayes Skenario 11 62
Gambar 4. 16 Confusion Matrix Naive Bayes Skenario 12 63
Gambar 4. 17 Confusion Matrix Naive Bayes Skenario 13 64
Gambar 4. 18 Confusion Matrix Naive Bayes Skenario 14 64
Gambar 4. 19 Confusion Matrix Naive Bayes Skenario 15 65
Gambar 4. 20 Confusion Matrix Random Forest Skenario 1 66
Gambar 4. 21 Confusion Matrix Random Forest Skenario 2 66
Gambar 4. 22 Confusion Matrix Random Forest Skenario 3 67
Gambar 4. 23 Confusion Matrix Random Forest Skenario 4 68
Gambar 4. 24 Confusion Matrix Random Forest Skenario 5 69
Gambar 4. 25 Confusion Matrix Random Forest Skenario 6 69
Gambar 4. 26 Confusion Matrix Random Forest Skenario 7 70
Gambar 4. 27 Confusion Matrix Random Forest Skenario 8 71
Gambar 4. 28 Confusion Matrix Random Forest Skenario 9 71
Gambar 4. 29 Confusion Matrix Random Forest Skenario 10 72
Gambar 4. 30 Confusion Matrix Random Forest Skenario 11 73
Gambar 4. 31 Confusion Matrix Random Forest Skenario 12 73
Gambar 4. 32 Confusion Matrix Random Forest Skenario 13 74
Gambar 4. 33 Confusion Matrix Random Forest Skenario 14 75
Gambar 4. 34 Confusion Matrix Random Forest Skenario 15 75
Gambar 4. 35 Diagram Batang Perbandingan Akurasi 80
Gambar 4. 36 Perbandingan Pemilihan Fitur NB + SMOTE 83
Gambar 4. 37 Perbandingan Pemilihan Fitur NB 84
Gambar 4. 38 Pengaruh Pemilihan Fitur NB + Undersampling 85

Gambar 4. 39 Pengaruh Pemilihan Fitur RF + SMOTE 86
Gambar 4. 40 Pengaruh Pemilihan Fitur RF 87
Gambar 4. 41 Pengaruh Pemilihan Fitur RF + Undersampling 88

DAFTAR LAMPIRAN

Lampiran A DATA PENELITIAN A-1
Lampiran B REPOSITORY GITHUB.....B-1

DAFTAR KODE PROGRAM

Kode Program 4. 1 Melihat Karakteristik Data dan Missing Value 37
Kode Program 4. 2 Pengecekan Data Outlier dan Duplikat..... 39
Kode Program 4. 3 Pembersihan Data Duplikat 39
Kode Program 4. 4 Pembersihan Data Outlier..... 39
Kode Program 4. 5 Metode SMOTE..... 40
Kode Program 4. 6 Metode Undersampling 41
Kode Program 4. 7 Chi-Square 46
Kode Program 4. 8 Pearson's Correlation 47
Kode Program 4. 9 Mutual Information..... 49
Kode Program 4. 10 Tuning Hyperparameter Naive Bayes 51
Kode Program 4. 11 Tuning Hyperparameter Random Forest 52
Kode Program 4. 12 Implementasi Naive Bayes 53
Kode Program 4. 13 Implementasi Random Forest 54
Kode Program 4. 14 Implementasi Random Forest Mutual Information 54
Kode Program 4. 15 Implementasi Naive Bayes Mutual Information..... 55
Kode Program 4. 16 Evaluasi Model 55

BAB I
PENDAHULUAN

1.1 Latar Belakang

Data mining merupakan proses yang digunakan untuk mengekstrak informasi berharga dari kumpulan data yang besar dan kompleks. Proses ini melibarkan berbagai teknik dan metode dari beberapa disiplin ilmu, termasuk statistik, *machine learning,* dan kecerdasan buatan (Saputra et al., 2021). Teknik *data mining* dapat mengidentifikasi korelasi, pola, dan penemuan pengetahuan dari dataset. *Data mining* juga telah berhasil diterapkan di berbagai sektor seperti ritel, pemasaran, kesehatan dan lain sebagainya (Singh & Singh, 2021).

Salah satu teknik *data mining* adalah klasifikasi. Klasifikasi merupakan bentuk analisis data untuk memprediksi label atau kategori suatu data berdasarkan koombinasi fitur-fitur yang dimiliki. Algoritma klasifikasi yang baik mampu menghasilkan prediksi yang akurat meskipun dengan data yang tinggi.

Diantara berbagai algoritma klasifikasi, Naive Bayes dan Rando Forest merupakan dua algoritma yang umum digunakan dalam penelitian maupun praktik industri. Naive Bayes merupakan algoritma klasifikasi berbasis probabilitas yang menggunakan teorema Bayes untuk menghitung kemungkinan suatu data masuk dalam kelas tertentu. Algoritma ini dikenal karena kesederhanaan, kecepatan, dan kemampuannya untuk memberikan hasil yang baik meskipun dengan asumsi independensi antar fitur (Meidina & Abidin, 2023). Penelitian oleh Wibowo et al., menunjukkan bahwa meskipun Naive Bayes merupakan algoritma yang sederhana, algoritma ini dapat memberikan akurasi yang baik, terutama jika dikombinasikan dengan teknik pemilihan fitur yang tepat (Wibowo et al., 2023). Penelitian lain oleh Halasz et al., mengungkapkan bahwa Naive Bayes dipilih karena kemampuannya untuk menangani fitur kategorial sehingga memberikan prediksi yang andal dalam data kompleks (Halasz et al., 2021).

Di sisi lain, Random Forest merupakan algoritma berbasis pohon keputusan yang menggunakan pendekatan *ensemble learning*. Algoritma ini membangun banyak pohon keputusan dari sub set data yang berbeda dan menggabungkan prediksinya untuk meningkatkan akurasi dan mengurangi *overfitting*. Random

Forest terbukti efektif dalam menangani data yang tidak seimbang dan memiliki kemampuan untuk menangani *noise* dan *outlier* dalam data (Puspa et al., 2023). Penelitian Sari et al., menunjukkan bahwa Random Forest dapat menangani data dengan jumlah besar dan memiliki kecepatan tinggi serta kemudahan dalam implementasi (Sari et al., 2021).

Meskipun Naive Bayes dan Random Forest memiliki pendekatan yang berbeda. Naive Bayes berbasis probabilitas dan Random Forest berbasis pohon Keputusan. Perbandingan kedua algoritma ini akan memberikan perspektif yang komprehensif mengenai bagaimana kedua algoritma dengan karakteristik berbeda merespons tantangan umum dalam klasifikasi, yaitu pemilihan fitur dan ketidakseimbangan data. Naive Bayes sering digunakan sebagai algoritma karena ringan dan cepat, sedangkan Random Forest digunakan representasi algoritma kompleks yang lebih toleran terhadap *noise* dan *overfitting*. Oleh karena itu, perbandingan ini tidak dimaksudkan untuk menyamakan dua algoritma secara struktur, melainkan untuk mengevaluasi bagaimana setiap pendekatan bereaksi terhadap pra-pemrosesan data yang berbeda.

Penelitian ini juga akan melihat pengaruh ketidakseimbangan data yang akan dilakukan dengan beberapa cara, salah satunya adalah SMOTE (*Synthetic Minority Over-sampling Technique)* dan *undersampling* (mengurangi data mayoritas sehingga diperoleh data mayoritas dan minoritas yang lebih seimbang). SMOTE dapat meningkatkan akurasi yang cukup signifikan dibandingkan tanpa dilakukan SMOTE (Fitriani & Febrianto, 2021). Selain SMOTE juga dapat dilakukan *undersampling*. Teknik *undersampling* dapat membantu meningkatkan akurasi model dengan mengurangi bias terhadap kelas mayoritas (Kang et al., 2018).

Selain penyeimbangan data, pemilihan fitur juga menjadi aspek penting dalam meningkatkan performa algoritma klasifikasi. Beberapa skenario pemilihan fitur akan menggunakan metode Chi-square, Pearson’s Correlation, dan Mutual Information. Ketiga metode tersebut merupakan teknik yang sering digunakan untuk meningkatkan akurasi model klasifikasi. Ketiga metode ini memiliki karakteristik dan pengaplikasian yang berbeda, namun semuanya bertujuan untuk memilih fitur yang paling relevan dari dataset. Metode Chi-square terbukti efektif

dalam pemilihan fitur dengan data kategorial, sehingga meningkatkan performa model klasifikasi(Mahmud et al., 2024). Di sisi lain, Pearson's Correlation digunakan untuk mengukur hubungan antara variabel numerik dan telah terbukti efektif dalam konteks data numerik(Daru et al., 2021). Sementara itu, Mutual Information, yang dijelaskan oleh Pawening et al., dapat menangani fitur dengan berbagai jenis data baik kategorial maupun numerik (Pawening et al., 2016).

Pada penelitian yang dilakukan oleh Prakoso Indaryono, yang menunjukkan bahwa akurasi Naive Bayes mencapai 36,61% dan Random Forest 86,55% (Prakoso Indaryono, 2024). Namun, pada penelitian oleh (Momole, 2022) dalam klasifikasi bahasa daerah algoritma Naive Bayes mendapat akurasi 99% dan Random Forest 65%. Penelitian lain oleh (Leonardo et al., 2020) menunjukkan skor akurasi 85% untuk Naive Bayes dan 90% untuk Random Forest dengan data yang hampir sama.

Pada penelitian (Fitriani & Febrianto, 2021) algoritma Random Forest dengan metode SMOTE mencapai hasil 92,61% yang mana meningkat dari sebelumnya hanya 90, 38%, sedangkan untuk algoritma Naive Bayes dengan SMOTE mencapai hasil 82,17% yang mana hasil ini turun dibandingkan hasil tanpa SMOTE, yaitu 88%. Berdasarkan penelitian tersebut peneliti akan melanjutkan penelitian tersebut dengan menambahkan beberapa skenario termasuk melakukan *undersampling,* melakukan pemilihan fitur menggunakan Chi-Square, Pearson's Correlation, dan Mutual Information.

Penelitian ini menggunakan dataset *Bank Marketing* dari UCI Machine Learning Repository. Dataset *Bank Marketing* termasuk kategori data *imbalance* dikarenakan pada dataset *bank marketing* terdapat data dengan label 'yes' sebanyak 4640 baris sedangkan data dengan label 'no' sebanyak 36548 baris dan memiliki 20 fitur yang bervariasi baik numerik maupun kategorial. Hal ini memungkinkan dilakukan eksplorasi berbagai skenario penanganan data tidak seimbang (SMOTE dan *undersampling*) dan pemilihan fitur (Chi-square, Pearson's Correlation, dan Mutual Information). Selain itu, dataset ini sudah banyak digunakan dalam penelitian klasifikasi, sehingga memudahkan validasi dan perbandingan dengan penelitian sebelumnya.

Merujuk pada uraian di atas, maka penting untuk dilakukan penelitian lebih lanjut untuk membandingkan kinerja algoritma Naive Bayes dan Random Forest yang akan dievaluasi menggunakan *confusion matrix* dalam mengklasifikasikan data *bank marketing* dengan menerapkan tahap *preprocessing* seleksi fitur dengan chi-square, pearson's correlation, dan mutual information serta penanganan *imbalance data* dengan menggunakan SMOTE, tanpa penyeimbangan data, dan dengan *undersampling.*

1.2 Perumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, rumusan masalah pada penelitian ini adalah bagaimana perbandingan kinerja algoritma Naive Bayes dan Random Forest pada data *bank marketing* dengan menerapkan beberapa skenario pada tahap *preprocessing data*, yaitu penanganan data yang tidak seimbang dengan menggunakan metode SMOTE, tanpa penanganan data yang tidak seimbang, dan *undersampling* serta pemilihan fitur menggunakan metode Chi-square, Pearson's Correlation, dan Mutual Information.

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk menghasilkan analisis komparatif terhadap akurasi Algoritma Naive Bayes dan Random Forest pada data *bank marketing* dengan menerapkan beberapa skenario pada tahap *preprocessing data* yang mencakup penanganan *imbalance data* menggunakan metode SMOTE, tanpa penanganan data yang tidak seimbang, dan *undersampling* serta pemilihan fitur menggunakan metode Chi-square, Pearson's Correlation, Mutual Information .

1.4 Pembatasan Masalah

Agar pengerjaan penelitian menjadi lebih terarah, maka dibuat batasan masalah pada penelitian ini, diantaranya yaitu:

1. Fokus utama pada penelitian ini adalah untuk membandingkan kinerja algoritma Naive Bayes dan Random Forest pada data *bank marketing* dengan menerapkan beberapa skenario tahap *preprocessing data,* yaitu penanganan data yang tidak seimbang menggunakan metode SMOTE*,* tanpa penanganan data yang tidak seimbang*,* dan *undersampling* serta pemilihan fitur

menggunakan metode Chi-square, Pearson's Correlation, dan Mutual Information.

2. Data yang digunakan merupakan data dari UC Irvine Machine Learning Repository. Data memiliki 21 kolom dan 41188 baris. Data memiliki 10 kolom berisikan data kategorial, 10 kolom berisikan data numerik, dan 1 kolom label dengan tipe data kategorial.

1.5 Sistematika Penulisan

Dalam penelitian ini, pembahasan disajikan menjadi lima bab dengan sistematika pembahasan sebagai berikut.

BAB I : PENDAHULUAN menjelaskan latar belakang dan motivasi dari penelitian ini serta memberikan ulasan singkat terkait permasalahan yang dihadapi. Selain itu, tujuan dan manfaat yang diharapkan dari penelitian ini juga dijelaskan pada ini.

BAB II : TINJAUAN PUSTAKA menguraikan literatur terkait mengenai metode klasifikasi yang digunakan dalam perbandingan algoritma klasifikasi Naive Bayes dan Random Forest. Penulis menjelaskan teori-teori dan algoritma yang relevan dengan penelitian ini.

BAB III : METODOLOGI PENELITIAN merinci tahap-tahap penelitian, mulai dari data yang digunakan, cara mempersiapkan dataset, apa saja yang akan dilakukan sebelum dilakukan pengklasifikasian menggunakan algoritma Naive Bayes dan Random Forest, sampai dengan evaluasi model yang telah dibuat.

BAB IV : HASIL DAN ANALISIS menyajikan hasil pengujian metode klasifikasi, kelebihan dan kekurangan dari metode yang digunakan akan dibahas secara mendalam.

BAB V : PENUTUP menyimpulkan jawaban terhadap rumusan masalah yang diajukan di awal penelitian, mengambil kesimpulan berdasarkan hasil pengujian, dan analisis yang dilakukan. Selain itu, saran dan rekomendasi untuk penelitian lebih lanjut serta potensi peningkatan metode yang digunakan akan diuraikan.