

**ANALISIS PERFORMANSI *NAÏVE BAYES CLASSIFIER* DAN  
*RANDOM FOREST* TERHADAP SENTIMEN KEBIJAKAN  
KENAIKAN HARGA BBM DI INDONESIA**

**SKRIPSI**



Oleh :

**MUHAMMAD LUTFI PRATAMA**

**19081010049**

**PROGRAM STUDI INFORMATIKA**

**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS PEMBANGUNAN NASIONAL "VETERAN"**

**JAWA TIMUR**

**2023**

**LEMBAR PENGESAHAN  
SKRIPSI**

**Judul** : ANALISIS PERFORMANSI NAÏVE BAYES CLASSIFIER  
DAN RANDOM FOREST TERHADAP SENTIMEN  
KEBIJAKAN KENAIKAN HARGA BBM DI INDONESIA  
**Oleh** : MUHAMMAD LUTFI PRATAMA  
**NPM** : 19081010049

Telah Diseminarkan Dalam Ujian Skripsi Pada :  
Hari Jum'at, Tanggal 17 Maret 2023

**Mengetahui**

**Dosen Pembimbing**

1.

Yisti Vita Via, S.T., M.Kom

NIP : 19860425 2021212 001

2.

Eka Prakarsa Mandyartha, S.T., M.Kom

NIP : 19880525 2018031 001

**Dosen Penguji**

1.

Dr. Br. Ani Dijah Rahajoe, ST., M.Cs.

NIP : 19730512 200501 2003

2.

Agung Mustika Rizki, S.Kom., M.Kom.

NIP : 19930725 202203 1 008

**Menyetujui**

**Dekan**

**Fakultas Ilmu Komputer**

Dr. Ir. Noyirina Hendrasarie, MT.

NPT : 19681126 199403 2 001

**Koordinator Program Studi**

**Informatika**

Fetty Tri Anggrawan, S.Kom, M.Kom

NPT : 19820211 2021212 005

## **SURAT PERNYATAAN ANTI PLAGIAT**

Saya mahasiswa Program Studi Informatika UPN “Veteran” Jawa Timur,  
yang bertanda tangan di bawah ini:

Nama : Muhammad Lutfi Pratama

NPM : 19081010049

Menyatakan bahwa laporan skripsi yang saya ajukan dan kerjakan, dengan  
judul:

**“ANALISIS PERFORMANSI *NAÏVE BAYES CLASSIFIER* DAN  
*RANDOM FOREST* TERHADAP SENTIMEN KEBIJAKAN  
KENAIKAN HARGA BBM DI INDONESIA“**

Bukan merupakan plagiat dari skripsi/tugas akhir/penelitian orang lain dan  
juga bukan merupakan produk atau *software* yang dibeli dari pihak lain. Saya juga  
menyatakan bahwa skripsi ini adalah pekerjaan pribadi, kecuali yang dinyatakan  
pada daftar pustaka dan tidak pernah diajukan sebagai syarat memperoleh gelar di  
UPN “Veteran” Jawa Timur maupun di institusi pendidikan lain.

Jika ternyata di kemudian hari pernyataan ini terbukti tidak benar, maka  
saya siap mempertanggungjawabkan.

Surabaya, 18 Maret 2023

Hormat saya,



**MUHAMMAD LUTFI PRATAMA**

**NPM. 19081010049**

# **ANALISIS PERFORMANSI *NAÏVE BAYES CLASSIFIER* DAN *RANDOM FOREST* TERHADAP SENTIMEN KEBIJAKAN KENAIKAN HARGA BBM DI INDONESIA**

**Nama Mahasiswa : Muhammad Lutfi Pratama**

**NPM : 19081010049**

**Program Studi : Informatika**

**Dosen Pembimbing : Yisti Vita Via, S.T., M.Kom**

**Eka Prakarsa Mandyartha, S.T., M.Kom**

## **ABSTRAK**

Bahan Bakar Minyak (BBM) merupakan komoditas yang memegang peranan penting dalam aktifitas perekonomian masyarakat. Kebijakan kenaikan harga BBM dapat mendorong inflasi yang berpengaruh negatif terhadap pertumbuhan ekonomi. Namun pemerintah tidak tinggal diam, berbagai kebijakan baru dibuat, seperti Bantuan Langsung Tunai BBM, Bantuan Subsidi Upah, dan lain-lain. Fenomena ini menimbulkan beragam sentimen di masyarakat. Mengetahui sentimen masyarakat dapat menjadi tolak ukur pemerintah dalam menangani permasalahan ini.

Oleh karena itu, digunakan algoritma *Naïve Bayes Classifier* (NBC) dan *Random Forest* (RF) untuk klasifikasi sentimen masyarakat terhadap kebijakan kenaikan harga BBM. Tujuan utama adalah mengukur tingkat performansi algoritma NBC dan RF dalam mengklasifikasi data teks ke dalam tiga kelas, yaitu positif, netral, dan negatif. Model NBC dan RF akan dilakukan *tuning hyperparameter* untuk mendapatkan nilai parameter terbaik, lalu menggunakan nilai akurasi dan rata-rata skor AUC-ROC sebagai tolak ukur menentukan model terbaik. Dataset yang digunakan sekitar 250 ribu data *tweet* dari hasil *scraping* melalui platform Twitter.

Dari hasil pengujian didapatkan tingkat akurasi untuk NBC sebesar 79.74% dan nilai rata-rata AUC-ROC sebesar 89.83%. Untuk RF didapatkan akurasi sebesar 85.15% dan untuk nilai rata-rata AUC-ROC sebesar 94.62%. Diketahui bahwa algoritma *Random Forest* lebih unggul dalam menganalisis sentimen masyarakat terhadap kebijakan kenaikan harga BBM di Indonesia.

**Kata kunci:** *Naïve Bayes*, *Random Forest*, Analisis Sentimen, BBM, AUC-ROC

## KATA PENGANTAR

Alhamdulillah, puji syukur kehadiran Allah SWT yang telah memberikan rahmat dan inayah-Nya sehingga penulis dapat menyelesaikan laporan skripsi yang berjudul “Analisis Performansi *Naïve Bayes Classifier* dan *Random Forest* Terhadap Sentimen Kebijakan Kenaikan Harga BBM di Indonesia”.

Terima kasih saya ucapkan kepada seluruh pihak yang telah memberikan banyak dukungan kepada penulis sehingga penulis bisa menyelesaikan laporan skripsi ini dengan baik dan tepat waktu.

Penulis menyadari, bahwa laporan skripsi ini masih jauh dari kata sempurna dalam penyusunan, tata kelola bahasa, maupun aspek penulisan lainnya. Oleh karena itu, penulis sangat mengharapkan kritik dan saran yang membangun dari semua pembaca agar penulis bisa menjadi lebih baik lagi di masa mendatang.

Semoga laporan skripsi ini bisa menambah wawasan para pembaca dan bisa bermanfaat untuk perkembangan dan peningkatan ilmu pengetahuan.

Surabaya, Maret 2023

Penulis

## UCAPAN TERIMA KASIH

Laporan skripsi ini dapat diselesaikan dengan baik atas berkat dan rahmat Allah SWT yang telah memberikan kesehatan dan kelancaran untuk penulis selama proses menyelesaikan skripsi. Proses penyelesaian skripsi ini juga tak lepas dari bantuan dan dukungan berbagai pihak, untuk itu penulis ingin menyampaikan ucapan terima kasih kepada:

1. Bapak Prof. Dr. Ir. Akhmad Fauzi, M.MT selaku Rektor Universitas Pembangunan Nasional “Veteran” Jawa Timur.
2. Ibu Dr. Novirina Hendrasarie, S.T., M.T. selaku Dekan Fakultas Ilmu Komputer di Universitas Pembangunan Nasional “Veteran” Jawa Timur.
3. Ibu Fetty Tri Anggraeny S.Kom., M.Kom. selaku Koordinator Program Studi Informatika Fakultas Ilmu Komputer UPN “Veteran” Jawa Timur.
4. Ibu Yisti Vita Via, S.T, M.Kom. selaku dosen pembimbing 1 dan Bapak Eka Prakarsa Mandyartha, S.T, M.Kom. selaku dosen pembimbing 2 yang telah memberikan arahan selama proses penyelesaian skripsi.
5. Bapak Dr. Basuki Rahmat, S.Si, MT. selaku dosen wali penulis yang selalu memberikan arahan selama proses perkuliahan.
6. Staff dan seluruh bapak ibu dosen program studi Informatika UPN “Veteran” Jawa Timur yang telah mengajar dan memberikan ilmu serta pengalaman selama masa perkuliahan.
7. Kedua orang tua penulis, Bapak Joko Waluyo dan Ibu Evi Nurul Rahmawati, serta adik-adik yang selalu memberikan motivasi dan doa untuk penulis.
8. Teman-teman seperjuangan Informatika angkatan 2019, kakak dan adik tingkat, serta seluruh pihak yang telah membantu dan mendukung penulis dalam menyelesaikan skripsi dan perkuliahan ini.

Penulis hanya bisa berharap, semoga Allah SWT selalu memberikan perlindungan dan balasan yang lebih di kemudian hari. Amin.

## DAFTAR ISI

LEMBAR PENGESAHAN SKRIPSI .....	ii
SURAT PERNYATAAN ANTI PLAGIAT .....	iii
ABSTRAK .....	iv
KATA PENGANTAR .....	v
UCAPAN TERIMA KASIH.....	vi
DAFTAR ISI.....	vii
DAFTAR TABEL.....	xi
DAFTAR GAMBAR .....	xii
DAFTAR KODE PROGRAM.....	xiv
BAB I PENDAHULUAN .....	1
1.1. Latar Belakang .....	1
1.2. Rumusan Masalah .....	4
1.3. Tujuan .....	4
1.4. Manfaat .....	4
1.5. Batasan Masalah.....	5
BAB II TINJAUAN PUSTAKA.....	6
2.1. Penelitian Terdahulu .....	6
2.2. Kenaikan Harga BBM.....	8
2.3. Data Mining .....	9
2.4. Analisis Sentimen .....	10
2.5. Web Crawling dan Scraping .....	11
2.6. Data Preprocessing.....	11
2.6.1. Case Folding.....	12

2.6.2. Tokenizing.....	12
2.6.3. Spelling Normalization.....	12
2.6.4. Filtering .....	12
2.6.5. Stemming.....	13
2.7. Pelabelan Data.....	13
2.8. Visualisasi Data.....	17
2.9. Feature Extraction .....	17
2.10. Naïve Bayes Classifier .....	18
2.11. Random Forest .....	21
2.12. K-Fold Cross Validation .....	24
2.13. Evaluasi Hasil Klasifikasi .....	25
2.14. Twitter .....	27
<b>BAB III METODOLOGI.....</b>	<b>29</b>
3.1. Studi Literatur .....	30
3.2. Analisis Kebutuhan .....	30
3.3. Kebutuhan Data.....	30
3.4. Kebutuhan <i>Hardware</i> dan <i>Software</i> .....	30
3.5. Pengumpulan Data .....	31
3.5.1. Sumber Data .....	31
3.5.2. Rancangan Struktur Data.....	31
3.6. Pengolahan Data.....	32
3.7. Pembuatan Model Klasifikasi .....	34
3.8. Skema Pengujian.....	37
<b>BAB IV HASIL DAN PEMBAHASAN .....</b>	<b>41</b>
4.1. Pengumpulan Data .....	41



4.2.	Preprocessing Data.....	46
4.2.1.	Case Folding.....	47
4.2.2.	Tokenizing.....	48
4.2.3.	Spelling Normalization.....	49
4.2.4.	Filtering .....	50
4.2.5.	Stemming.....	52
4.3.	Labeling.....	53
4.4.	Visualisasi Dataset .....	55
4.5.	Pembagian Data .....	56
4.6.	Feature Extraction .....	57
4.7.	Klasifikasi Data.....	59
4.7.1.	Klasifikasi Naïve Bayes .....	59
4.7.2.	Klasifikasi Random Forest .....	62
4.8.	Pengujian Model .....	69
4.8.1.	Hasil Pengujian Tuning Hyperparameter Random Forest .....	73
4.8.2.	Hasil Pengujian Tuning Hyperparameter Naïve Bayes Classifier.....	79
4.8.3.	Hasil Pengujian K-Fold Cross Validation Model NBC dan R....	84
4.8.4.	Hasil Pengujian Menggunakan Data Asing Pada NBC dan RF...	85
4.9.	Interpretasi Model Terbaik.....	87
BAB V KESIMPULAN .....		91
5.1.	Kesimpulan .....	91
5.2.	Saran.....	92
DAFTAR PUSTAKA .....		93

LAMPIRAN.....	99
BIODATA PENULIS .....	102

## DAFTAR TABEL

Tabel 2.1 Penelitian Terdahulu .....	6
Tabel 2.2 Struktur <i>Document-term Matrix</i> .....	18
Tabel 2.3 Ilustrasi <i>k-fold cross-validation</i> .....	25
Tabel 2.4 Tabel <i>Confusion Matrix</i> .....	25
Tabel 3.1 Struktur data sebelum tahap <i>preprocessing</i> .....	32
Tabel 3.2 Struktur data <i>document-term</i> matriks.....	32
Tabel 3.3 Struktur tabel skenario pengujian <i>Naïve Bayes Classifier</i> .....	38
Tabel 3.4 Struktur tabel skenario pengujian <i>Random Forest</i> .....	38
Tabel 4.1 Banyak data <i>tweet</i> per bulan .....	46
Tabel 4.2 Contoh hasil <i>case folding</i> data <i>tweet</i> .....	48
Tabel 4.3 Contoh hasil filtering <i>tweet</i> Twitter .....	51
Tabel 4.4 Akumulasi jumlah data setiap label .....	54
Tabel 4.6 Hasil uji skenario <i>hyperparameter</i> RF.....	76
Tabel 4.7 Hasil uji skenario <i>hyperparameter</i> NBC .....	82

## DAFTAR GAMBAR

Gambar 2.1 <i>Data Processing Pipeline</i> .....	10
Gambar 2.2 Model arsitektur Transformer .....	14
Gambar 2.3 Hasil pelabelan data sample .....	16
Gambar 2.4 Visualisasi <i>Word Cloud</i> .....	17
Gambar 2.5 Struktur Decision Tree .....	22
Gambar 2.6 Struktur <i>Random Forest</i> .....	24
Gambar 2.7 Contoh kurva ROC.....	27
Gambar 2.8 Barchart data pengguna twitter menurut negara tahun 2022.....	28
Gambar 3.1 Langkah-langkah Penelitian .....	29
Gambar 3.2 Diagram alir metode pengolahan data.....	33
Gambar 3.3 Diagram alir metode klasifikasi .....	35
Gambar 3.4 Tahapan Skema Pengujian .....	37
Gambar 4.1 Dataframe output Crawling.....	42
Gambar 4.2 <i>Dataframe output scraping</i> .....	45
Gambar 4.3 Grafik jumlah tweet setiap bulan .....	46
Gambar 4.4 <i>Dataframe</i> hasil <i>tokenizing</i> .....	49
Gambar 4.5 <i>Dataframe</i> hasil normalisasi teks <i>tweet</i> .....	50
Gambar 4.6 <i>Dataframe</i> hasil <i>stemming tweet</i> .....	53
Gambar 4.7 <i>Dataframe</i> hasil <i>labeling</i> .....	54
Gambar 4.8 Grafik label sentimen terhadap jumlah data setiap bulan .....	55
Gambar 4.9 <i>Word Cloud</i> frekuensi fitur dataset .....	56
Gambar 4.10 <i>Dataframe</i> hasil ekstraksi fitur .....	58
Gambar 4.11 Data <i>test</i> baru.....	72
Gambar 4.12 <i>Dataframe</i> hasil pengujian <i>hyperparameter</i> NBC .....	74

Gambar 4.13 Grafik iterasi pengujian terhadap nilai akurasi RF.....	75
Gambar 4.14 Grafik pengaruh variabel parameter rasio splitting dan alpha terhadap nilai akurasi model RF .....	75
Gambar 4.15 Grafik iterasi pengujian terhadap nilai <i>recall</i> RF.....	77
Gambar 4.16 Nilai Confusion matrix model terbaik RF.....	77
Gambar 4.17 Grafik iterasi model terhadap kecepatan prediksi model RF .....	78
Gambar 4.18 Grafik korelasi waktu prediksi dan nilai akurasi model uji RF.....	78
Gambar 4.19 Dataframe hasil pengujian hyperparameter NBC .....	79
Gambar 4. 20 Grafik iterasi pengujian terhadap nilai akurasi NBC .....	80
Gambar 4.21 Grafik pengaruh variabel parameter rasio splitting dan alpha terhadap nilai akurasi model NBC.....	80
Gambar 4.22 Grafik iterasi model terhadap kecepatan prediksi model NBC.....	81
Gambar 4.23 Grafik korelasi waktu prediksi dan nilai akurasi model uji NBC .....	81
Gambar 4.24 Grafik iterasi pengujian terhadap nilai <i>recall</i> NBC .....	82
Gambar 4.25 Nilai <i>Confusion matrix</i> model terbaik NBC.....	83
Gambar 4.26 Hasil <i>k-fold cross validation</i> pada RF.....	84
Gambar 4.27 Hasil <i>k-fold cross validation</i> pada NBC.....	84
Gambar 4.28 Hasil <i>majority voting unknown data test</i> model RF dan NBC.....	86
Gambar 4.29 Hasil klasifikasi <i>unknown data</i> model <i>Random Forest</i> dan <i>Naive Bayes</i> .....	86
Gambar 4.30 Kurva ROC dan nilai AUC model <i>Naive Bayes</i> .....	88
Gambar 4.31 Kurva ROC dan nilai AUC model <i>Random Forest</i> .....	89

## DAFTAR KODE PROGRAM

Kode Program 4.1 <i>Crawling</i> data <i>tweet</i> .....	41
Kode Program 4.2 Akses token TwitterAPI .....	43
Kode Program 4.3 Fungsi scraping konten tweet Twitter.....	43
Kode Program 4.4 Membagi data frame <i>link tweet</i> menjadi baris kelipatan .....	44
Kode Program 4.5 Fungsi Case Folding .....	48
Kode Program 4.6 Fungsi <i>tokenize</i> data <i>tweet</i> .....	49
Kode Program 4.7 Normalisasi <i>tweet</i> .....	50
Kode Program 4.8 Menghapus stopword setiap kata.....	51
Kode Program 4.9 <i>Stemming</i> tweet.....	52
Kode Program 4.10 <i>Labeling</i> setiap data <i>tweet</i> .....	54
Kode Program 4.11 <i>Splitting dataset</i> menjadi data <i>training</i> dan data <i>testing</i> .....	57
Kode Program 4.12 Membuat <i>document-term matrix</i> menggunakan fungsi vectorizer.....	58
Kode Program 4.13 Kelas Multinomial Naive Bayes .....	61
Kode Program 4.14 Membuat model dan prediksi <i>data test</i> .....	62
Kode Program 4.15 Inisialisasi kelas Tree dan fungsi nilai prediksi.....	63
Kode Program 4.16 Inisialisasi model kelas <i>Random Forest</i> beserta <i>method</i> utamanya .....	66
Kode Program 4.17 Fungsi <i>splitting node (Gini Index)</i> dan fungsi prediksi .....	68
Kode Program 4.18 Membuat model dan prediksi model .....	68
Kode Program 4.19 Inisialisasi nilai parameter pengujian .....	69
Kode Program 4.20 Pengujian <i>hyperparameter</i> NBC .....	70
Kode Program 4.21 Pengujian <i>hyperparameter</i> RF.....	71
Kode Program 4.22 <i>K-fold cross-validation</i> .....	72
Kode Program 4.23 <i>Confusion Matrix</i> .....	73

Kode Program 4.24 *Classification Report* .....73