Jurnal Ilmu Komputer, Sistem Informasi, Teknik Informatika
E-ISSN: 2964-4763
Vol. 5, No. 1, Maret 2026

Perbandingan Performa Algoritma *Naive Bayes*, *Random Forest* dan *K-Nearest Neighbor* pada Prediksi Calon Jemaah Haji Indonesia yang Berpotensi Membatalkan Haji

Feri Setiadi[1], Handoyo Widi Nugroho[2], Suhendro Yusuf Irianto[3]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Institut Informatika dan Bisnis Darmajaya[1,2,3]

E-mail: verisetiadi@gmail.com[1], handoyo.wn@darmajaya.ac.id[2], suhendro@darmajaya.ac.id[3]

Abstract

This study aims to identify the most effective classification model in predicting prospective Hajj pilgrims who are likely to cancel their registration and to determine the most influential features in this decision. The study compares three classification models: Naïve Bayes, Random Forest, and K-Nearest Neighbor (k-NN), using a dataset of Hajj pilgrims from the Siskohat system of the Ministry of Religious Affairs Office in Pringsewu Regency. Additionally, the study applies the Recursive Feature Elimination Cross Validation (REFCV) feature selection method to identify the most relevant features influencing Hajj cancellations. The results show that the Random Forest model delivers the best performance, with higher accuracy, precision, and recall compared to the Naïve Bayes and k-NN models, both before and after feature selection. Features such as 'age,' 'occupation,' and 'address' were found to be the most significant attributes influencing Hajj cancellations. The application of the REFCV method proved to enhance model accuracy, particularly with the Random Forest model, which achieved 95% accuracy after feature selection. This study concludes that the Random Forest model with REF feature selection is the most effective combination for predicting Hajj registration cancellations, and it provides recommendations for Hajj administrators to improve prediction accuracy and data management efficiency.

Keywords*: Recursive Feature Elimination Cross Validation (REF), Naïve Bayes, Random Forest, K-Nearest Neighbor (k-NN), Hajj Cancellation.*

Abstrak

Penelitian ini bertujuan untuk mengidentifikasi model klasifikasi yang paling efektif dalam memprediksi calon jemaah haji yang berpotensi membatalkan pendaftaran hajinya serta menentukan fitur-fitur yang paling berpengaruh terhadap keputusan tersebut. Penelitian ini membandingkan tiga model klasifikasi, yaitu *Naïve Bayes, Random Forest, dan K-Nearest Neighbor (k-NN),* dengan menggunakan dataset jemaah haji dari Siskohat Kantor Kementerian Agama Kabupaten Pringsewu. Selain itu, penelitian ini juga menerapkan metode seleksi fitur *Recursive Feature Elimination Cross Validation (REFCV)* untuk mengidentifikasi fitur-fitur yang paling relevan dalam mempengaruhi pembatalan haji. Hasil penelitian menunjukkan bahwa model *Random Forest* memberikan performa terbaik dengan akurasi, presisi, dan recall yang lebih tinggi dibandingkan model *Naïve Bayes* dan *k-NN*, baik sebelum maupun setelah seleksi fitur. Fitur-fitur seperti 'usia', 'pekerjaan', dan 'alamat' ditemukan sebagai atribut yang paling signifikan dalam mempengaruhi pembatalan haji. Penerapan metode *REFCV* terbukti meningkatkan akurasi model, khususnya pada model *Random Forest* yang mencapai akurasi 95% setelah seleksi fitur. Penelitian ini menyimpulkan bahwa model *Random Forest* dengan seleksi fitur *REFCV* merupakan kombinasi yang paling efektif dalam memprediksi pembatalan pendaftaran haji, serta memberikan rekomendasi bagi pengelola haji dalam meningkatkan akurasi prediksi dan efisiensi pengelolaan data jemaah haji.

Kata Kunci: *Recursive Feature Elimination Cross Validation (REFCV), Naïve Bayes, Random Forest, K-Nearest Neighbor (k-NN),* Pembatalan Haji.

1. Pendahuluan

Ibadah haji merupakan salah satu rukun Islam yang wajib ditunaikan bagi setiap muslim yang telah mencukupi syarat-syaratnya [1]. Setiap tahun ribuan calon jemaah haji Indonesia diberangkatkan ke tanah suci untuk menunaikan ibadah haji[2]. Data pada Kementerian Agama RI dalam skala nasional melaporkan jumlah pendaftar jemaah haji mengalami tren yang meningkat. Pada tahun 2021 sebanyak 67.345 orang telah tercatat sebagai pendaftar baru Jemaah haji dan pada tahun 2022 meningkat 71% menjadi 398.622 orang pendaftar baru (satudata.kemenag.go.id) [3]. Namun, setiap tahun, sejumlah calon Jemaah haji membatalkan pendaftaran hajinya karena berbagai alasan, seperti kondisi kesehatan yang memburuk, usia lanjut, meninggal, atau kendala finansial.

Berdasarkan data dari Siskohat Kantor Kementerian Agama Kabupaten Pringsewu, alasan "lain-lain" menjadi penyebab terbesar dari pembatalan pendaftaran haji, mencakup 59,7% dari total kasus. Pembatalan akibat sakit mencapai 5,4%, sementara 34,9% disebabkan oleh wafatnya calon Jemaah haji. Secara keseluruhan, terdapat 149 kasus pembatalan pendaftaran haji dalam rentang tahun 2023-2024.

Tabel 1. Data Pembatalan Jemaah Haji

Sebab Batal	Tahun 2023	Tahun 2024	Jumlah	%
Lain - lain	56	33	89	59,7
Sakit	8	-	8	5,4
Wafat	39	13	52	34,9
Total	103	46	149	100

Membatalkan pendaftaran haji menjadi suatu kerugian bagi Jemaah tersebut, baik dari segi biaya, waktu, maupun kesempatan yang langka untuk menunaikan ibadah haji. Selain itu, pembatalan pendaftaran Jemaah haji juga berdampak pada kebijakan pemerintah dan pengelolaan kuota haji yang telah ditetapkan sebelumnya. Ahli hukum keuangan negara, Siswoyo Sujanto, dalam sidang uji materi terkait pengelolaan dana haji yang digelar oleh Mahkamah Konstitusi menyatakan, pembatalan pendaftaran Jemaah haji oleh calon Jemaah haji dapat menyebabkan kekacauan administrasi yang berdampak pada pneyelenggaraan haji secara keseluruhan [4]. Oleh karena itu, prediksi akurat terkait Jemaah haji yang berpotensi membatalkan pendaftaran Jemaah haji perlu dilakukan, sehingga dapat dilakukan tindakan pencegahan atau kebijakan yang tepat sebelum calon Jemaah haji membatalkan keberangkatan haji ke Tanah Suci.

Penelitian ini menggunakan algoritma klasifikasi dalam melakukan prediksi calon Jemaah haji Indonesia yang berpotensi membatalkan keberangkatan haji. Algoritma klasifikasi adalah suatu teknik dalam *data mining* yang digunakan untuk mengklasifikasikan suatu data menjadi kelompok atau kategori tertentu berdasarkan ciri-ciri atau atribut yang dimilikinya. Namun, dalam proses klasifikasi, pemilihan fitur yang relevan sangat penting untuk meningkatkan akurasi prediksi [5].

Beberapa studi terkait telah dilakukan mengenai algoritma klasifikasi dengan seleksi fitur. Penggunaan *Recursive Feature Elimination* dengan *Cross-Validation* (*REFCV*) dapat meningkatkan akurasi model pembelajaran mesin dengan mengidentifikasi dan memilih fitur-fitur yang paling relevan. Studi ini menunjukkan bahwa meskipun model pembelajaran mesin yang sederhana digunakan, hasil yang lebih baik dapat dicapai dengan seleksi fitur yang tepat. Dalam konteks deteksi diabetes tipe II, *REFCV* berhasil mengurangi dimensi data tanpa kehilangan akurasi prediksi yang signifikan. Dengan menerapkan *Cross Validation*, masalah *overfitting* dapat dihindari, dan hasil yang lebih akurat dicapai dibandingkan dengan menggunakan semua fitur yang tersedia [6].

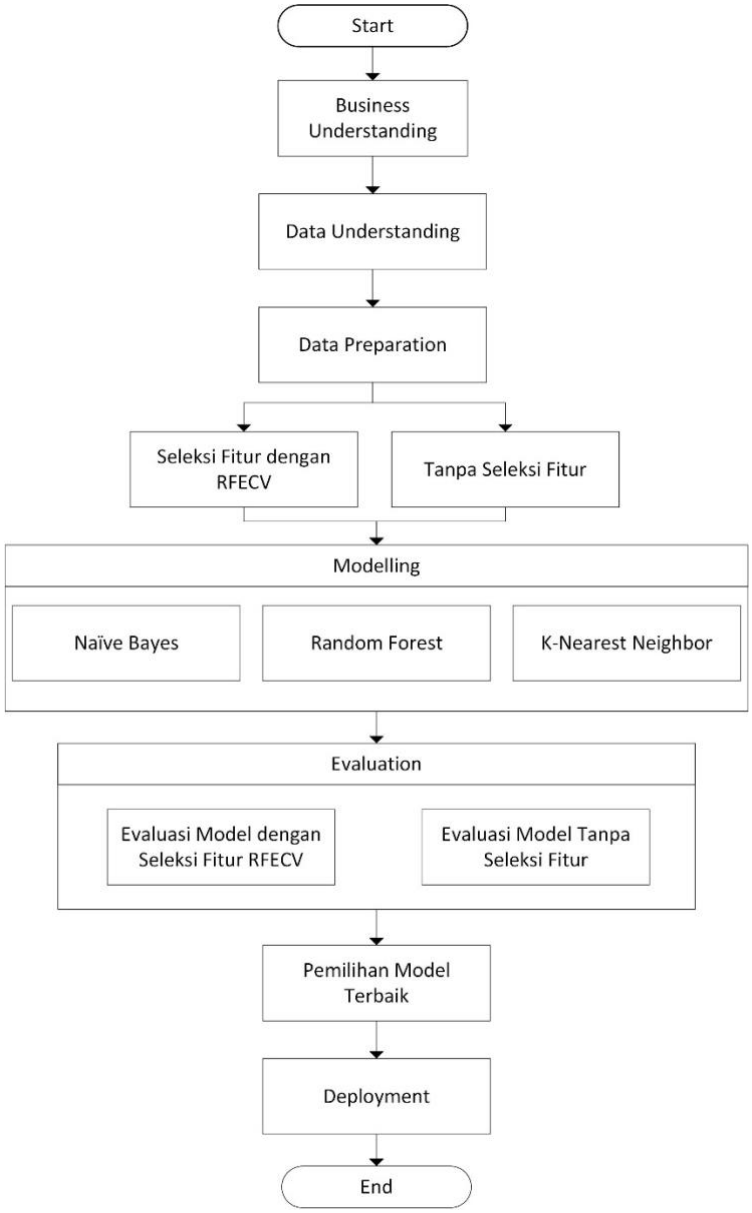
Dalam penelitian yang dilakukan oleh Ardea dan Fahrurozi pada tahun 2019, algoritma *Naive Bayes*, *K-Nearest Neighbor*, *Decision Tree*, dan *Random Forest* dibandingkan untuk mengklasifikasikan data terkait penyakit jantung koroner. Data yang digunakan adalah data penyakit jantung koroner yang berasal dari lokasi Cleveland Clinic Foundation yang telah diadopsi oleh instansi Hungarian Institute of Cardiology. Penelitian tersebut menemukan bahwa algoritma *Random Forest* memiliki nilai akurasi terbaik dibandingkan dengan algoritma lainnya untuk klasifikasi data [7].

Penelitian yang dilakukan oleh Agung Purwanto, Handoyo Widi Nugroho membandingkan Algoritma C. 45 dan Algoritma *K-Nearest Neighbor*s untuk klasifikasi penerima beasiswa. Dari penelitian tersebut, didapatkan bahwa *K-Nearest Neighbor*s memberikan nilai akurasi yang lebih baik [8].

Berdasarkan hasil akurasi yang relatif baik pada model studi literature tersebut, penelitian ini akan membandingkan antara model klasifikasi *data mining Naive Bayes*, *Random Forest* dan *K-Nearest Neighbor* dengan menambahkan tahap *Recursive Feature Elimination Cross-Validation*. Data yang digunakan dalam penelitian ini adalah data jemaah haji dari Kantor Kementerian Agama Kabupaten Pringsewu tahun 2019-2024. Penelitian ini dilakukan untuk mendapatkan model terbaik dalam memprediksi pembatalan haji sehingga dapat memberikan informasi yang berharga dalam upaya meningkatkan efektivitas pengelolaan kuota haji dan mencegah atau mengurangi jumlah pembatalan haji oleh masyarakat.

2. Metode

Metode yang digunakan dalam penelitian ini menggunakan standard Cross-Industry Standard Process for *Data mining* (*CRISP-DM*) yang terdiri dari beberapa tahapan yaitu *Business Understanding*, *Data Understanding*, *Data Preparation*, Modelling, *Evaluation*, dan *Deployment* [9].



Gambar 1. Langkah kerja implementasi *CRISP-DM*

Pada penelitian ini metode yang digunakan dalam pengumpulan data sebagai berikut : Wawancara kepada pegawai pada Seksi Penyelenggaraan Haji dan Umroh Kantor Kementerian Agama Kabupaten Pringsewu. Wawancara dilakukan untuk mengetahui masalah dan tujuan bisnis agar relevan dengan penelitian. Dalam tahap wawancara, akan diajukan pertanyaan-pertanyaan yang tepat dan terstruktur, serta mendengarkan dengan seksama jawaban yang diberikan oleh responden. Hasil dari tahap wawancara akan menjadi bagian penting dalam analisis data dan pembuatan model prediksi yang akan digunakan untuk mengidentifikasi jamaah haji yang berpotensi membatalkan haji.Studi pustaka dilakukan dengan mencari menggali informasi dan pengetahuan dari penelitian sebelumnya, serta publikasi dari buku, jurnal atau informasi lainnya dengan harapan pengetahuan tersebut dapat menjadi pendukung dalam penelitian ini terutama tentang teori, pendapat, serta pemikiran yang berkaitan dengan penelitian ini.

3. Hasil

3.1 Kerangka Penelitian

Penelitian ini menggunakan metodologi *CRISP-DM* dengan tahapan:

3.1.1 *Business Understanding*

Tujuan utama dari penelitian ini adalah untuk membantu Kementerian Agama khususnya Kantor Kementerian Agama Kabupaten Pringsewu dalam memprediksi calon jemaah haji yang berpotensi membatalkan haji dengan memanfaatkan Teknik *data mining* dengan target kriteria keberhasilan mencapai tingkat akurasi prediksi minimal 90%.

Langkah-langkah yang dilakukan meliputi pengumpulan data historis pendaftaran calon jemaah haji, pembersihan dan persiapan data untuk analisis, analisis data untuk mengidentifikasi pola dan tren yang relevan, pengembangan model prediksi menggunakan algoritma *Naive Bayes*, *Random Forest* dan *K-Nearest Neighbor* , serta pengujian dan validasi model dengan data uji untuk memastikan akurasi dan keandalan model prediksi. Setelah itu, performa antara *Naive Bayes*, *Random Forest* dan *K-Nearest Neighbor* akan dibandingkan berdasarkan akurasi, presisi, dan recall, sebelum menerapkan model terbaik dalam lingkungan produksi dan memantau kinerjanya secara berkala untuk perbaikan lebih lanjut.

3.1.2 Data Understanding

Data yang digunakan dalam penelitian ini adalah dataset yang diperoleh dari Sistem Komputerisasi Haji Terpadu (Siskohat) di wilayah Kantor Kementerian Agama Kabupaten Pringsewu dalam rentang tahun 2019 – 2024. Jumlah data didapatkan sebanyak 1.133 record dan 22 atribut.

Tabel 2. Dataset Keberangkatan dan Pembatalan Calon Jemaah Haji

No. Porsi	Tgl Daftar	Nama	Nama Ayah	Jenis Kelamin	Pekerjaan	Pendidikan	Tempat Lahir	Tgl Lahir	Alamat	Desa	Kecamatan	Kode Pos	Kab / Kota	Bank	Cabang	Jml Setoran	S. Aktif	S. Bayar	S. Haji	No. KTP	Status
0800089740	11/01/2012	ENDY SUSILO	SUBAGIO	L	Swasta	S1	TANJUNG	20/04/1972	PRINGSEW	PRINGSEW	PRINGSEWU	35373	KAB. PRIN	BSM	KCP PRINGSEWU	25000000	ACTIVE	CICIL	BELUM I	01200472	Berangkat
0800089742	11/01/2012	TITIN YENI HA	HADI SUMA	P	Swasta	S1	GISTING	13/09/1978	PRINGSEW	PRINGSEW	PRINGSEWU	35373	KAB. PRIN	BSM	KCP PRINGSEWU	25000000	ACTIVE	CICIL	BELUM I	01530978	Berangkat
0800091030	25/01/2012	SUHANTO SA	SANASMAD	L	Pegawai Negri Sip	SLTA	SIDOHARJ	22/08/1964	SIDOHARJ	SIDOHARJ	PRINGSEWU	35373	KAB. PRIN	BSM	KCP PRINGSEWU	25000000	ACTIVE	CICIL	BELUM I	01220864	Berangkat
0800092737	20/02/2012	MUHAMMAD	NUHAIRI	L	Dagang	SLTA	PRINGSEW	15/03/1970	LINGKUNG	PRINGSEW	PRINGSEWU	35373	KAB. PRIN	BSM	KCP PRINGSEWU	25000000	ACTIVE	CICIL	BELUM I	01150370	Berangkat
0800092738	20/02/2012	SUPRAPTI MI	MIYONO	P	Ibu Rumah Tangg	S1	TANJUNG	30/06/1972	LINGKUNG	PRINGSEW	PRINGSEWU	35373	KAB. PRIN	BSM	KCP PRINGSEWU	25000000	ACTIVE	CICIL	BELUM I	01700672	Berangkat
0800096135	27/04/2012	MUNFARIDAI	NASUHA	P	Swasta	SD	PONCOWA	10/06/1965	PRINGKUN	PRINGSEW	PRINGSEWU	35373	KAB. PRIN	BRIS	KCP LAMPUNG PR	25000000	ACTIVE	CICIL	BELUM I	01500665	Berangkat
0800096147	27/04/2012	BARUDIN HA	HASANUDI	L	Tani / Nelayan	SD	KUTOARJO	20/06/1967	PRINGKUN	PRINGSEW	PRINGSEWU	35373	KAB. PRIN	BSM	KCP PRINGSEWU	25000000	ACTIVE	CICIL	BELUM I	01200667	Berangkat
0800096681	10/05/2012	LILIS SUHAYA	WIHATMA	P	Pegawai Negri Sip	S1	CIKONENG	11/02/1966	JL.GOTON	PRINGSEW	PRINGSEWU	35373	KAB. PRIN	BRIS	KCP LAMPUNG PR	25000000	ACTIVE	CICIL	BELUM I	01510266	Berangkat
0800096682	10/05/2012	IMOP SUTOPI	SOEDARMA	L	Dagang	S1	KOTA BUN	23/03/1964	JL.GOTON	PRINGSEW	PRINGSEWU	35373	KAB. PRIN	BRIS	KCP LAMPUNG PR	25000000	ACTIVE	CICIL	BELUM I	01230364	Berangkat
0800096865	15/05/2012	SUWARTO SC	SOMO WAS	L	Pensiunan	D1 / D2	MAGELAN	05/07/1947	JL.PEMUD	PAGELARA	PAGELARAN	35375	KAB. PRIN	BSM	KCP PRINGSEWU	25000000	ACTIVE	CICIL	BELUM I	05050747	Berangkat
0800097010	21/05/2012	SUKINAH YOS	YOSO SUW	P	Ibu Rumah Tangg	SD	PODOMO	30/06/1962	PODOMO	PODOMO	PRINGSEWU	35373	KAB. PRIN	BRIS	KCP LAMPUNG PR	25000000	ACTIVE	CICIL	BELUM I	01700662	Berangkat
0800097063	22/05/2012	WASIMAN AS	ASMA DIM	L	Pegawai Negri Sip	S1	CILACAP	10/11/1963	JL.SATRIA	PRINGSEW	PRINGSEWU	35373	KAB. PRIN	BSM	KCP PRINGSEWU	25000000	ACTIVE	CICIL	SUDAH I	01101163	Berangkat
0800097064	22/05/2012	SITI MUFLIH	ABDUL	P	Pegawai Negri Sip	S1	PRINGSEW	20/05/1964	JL.SATRIA	PRINGSEW	PRINGSEWU	35373	KAB. PRIN	BSM	KCP PRINGSEWU	25000000	ACTIVE	CICIL	SUDAH I	01600056	Berangkat
0800097198	24/05/2012	SUNARTI MU	MUHAMMA	P	Ibu Rumah Tangg	SD	GISTING	07/12/1960	BANDUNG	BANDUNG	ADILUWIH	35374	KAB. PRIN	BRIS	KCP LAMPUNG PR	25000000	ACTIVE	CICIL	BELUM I	07471266	Berangkat
0800097205	24/05/2012	BADRUDIN BI	BUKHORI	L	Dagang	SLTP	WARGO M	02/01/1960	WARGO M	WARGO M	PARDASUKA	35382	KAB. PRIN	BSM	KCP PRINGSEWU	25000000	ACTIVE	CICIL	BELUM I	04020166	Berangkat
0800097206	24/05/2012	ALIYAH MUN	MUNAWIR	P	Ibu Rumah Tangg	SLTP	WARGO M	07/08/1968	WARGO M	WARGO M	PARDASUKA	35382	KAB. PRIN	BSM	KCP PRINGSEWU	25000000	ACTIVE	CICIL	BELUM I	04470868	Berangkat
0800097357	29/05/2012	SUKINAH AM	AMATDIME	P	Pegawai Negri Sip	D1 / D2	WARINGIN	29/06/1962	WARINGIN	WARINGIN	SUKOHARJO	35374	KAB. PRIN	BSM	KCP PRINGSEWU	25000000	ACTIVE	CICIL	BELUM I	08690662	Berangkat
0800097358	29/05/2012	TUMIRANUDI	KARTOREJO	L	Pegawai Negri Sip	D1 / D2	WARINGIN	17/09/1959	WARINGIN	WARINGIN	SUKOHARJO	35374	KAB. PRIN	BSM	KCP PRINGSEWU	25000000	ACTIVE	CICIL	BELUM I	08170959	Berangkat

Hasil pengumpulan data didapatkan 1.134 record data jemaah haji yang telah berangkat dan membatalkan porsi hajinya.

3.1.3 Data Preparation

Data yang diperoleh dari pengumpulan data, kemudian diperiksa untuk mendeteksi nilai yang hilang (*missing value*) dan duplikasi data.

```python
# Cek jumlah missing value per kolom
missing_values = dataset.isnull().sum()
print("Jumlah missing value per kolom:")
print(missing_values)

total_missing = missing_values.sum()
print(f"Total missing value di seluruh dataset: {total_missing}")
```

```
Jumlah missing value per kolom:
nomor_porsi       0
tanggal_daftar    0
usia_daftar       0
nama              0
nama_ayah         0
jenis_kelamin     0
pekerjaan         0
pendidikan        0
tempat_lahir      0
tanggal_lahir     0
alamat            0
desa              0
kecamatan         0
kode_pos          0
kab_kota          0
bank              0
cabang            0
jumlah_setoran    0
s_aktif           0
s_bayar           0
s_haji            0
nik               0
status            0
dtype: int64
Total missing value di seluruh dataset: 0
```

Gambar 2. Deteksi Nilai Hilang (*missing value*)

```python
import pandas as pd

# Mengecek jumlah baris yang duplikat
jumlah_duplikasi = dataset.duplicated().sum()
print(f"Jumlah baris yang duplikat: {jumlah_duplikasi}")

# Melihat baris yang duplikat
baris_duplikat = dataset[dataset.duplicated()]
print(baris_duplikat)
```

```
Jumlah baris yang duplikat: 0
Empty DataFrame
Columns: [nomor_porsi, tanggal_daftar, usia_daftar, nama, nama_ayah, jenis_kelamin,
Index: []

[0 rows x 23 columns]
```

Gambar 3. Deteksi Duplikasi Data

Dari hasil deteksi tidak ditemukan duplikasi data dan nilai yang hilang (*missing value*) pada dataset. Selanjutnya dilakukan transformasi data dengan mengubah data yang memiliki nilai kategorikal menjadi numerik. Sebagian besar algoritma *machine learning* dan statistik memerlukan data dalam bentuk numerik untuk dapat melakukan perhitungan yang relevan. Oleh karena itu, diperlukan langkah untuk mengubah data kategorikal ini menjadi format numerik. Metode yang akan digunakan pada penelitian ini menggunakan *label encoding*.

Seleksi fitur menggunakan teknik *Recursive Feature Elimination* diterapkan untuk mengidentifikasi atribut yang paling berpengaruh dan dapat meningkatkan performa model prediksi. Untuk memperoleh hasil pengujian dan performa yang lebih akurat, model prediksi akan dibandingkan antara dataset yang telah melalui proses seleksi fitur menggunakan *Recursive Feature Elimination* dan dataset yang tidak melalui seleksi fitur.

3.1.4 Modeling

3.1.4.1 Persiapan Model

Dataset calon jemaah haji dipersiapkan untuk proses modeling dengan menerapkan teknik 10-fold *Cross-Validation*. Teknik ini membagi dataset menjadi sepuluh subset (folds), di mana setiap fold digunakan secara bergantian sebagai data uji, sementara sembilan fold lainnya digunakan sebagai data latih. Proses ini diulang sepuluh kali sehingga setiap fold berperan sebagai data uji satu kali, memastikan bahwa setiap bagian dataset diuji dan dilatih dengan merata. Pendekatan ini diterapkan baik pada dataset yang telah melalui proses seleksi fitur menggunakan teknik *Recursive Feature Elimination* maupun pada dataset yang tidak mengalami seleksi fitur.

Dengan menggunakan teknik *Cross-Validation*, evaluasi model menjadi lebih baik dan hasil performa model lebih dapat diandalkan karena meminimalkan kemungkinan *overfitting* dan memberikan gambaran yang lebih akurat mengenai kemampuan generalisasi model pada data yang belum pernah dilihat sebelumnya.

3.1.4.2 Pemodelan Tanpa Seleksi Fitur

Pada tahap ini, dilakukan pembuatan model prediksi dengan algoritma *Naïve Bayes*, *Random Forest*, dan *K-Nearest Neighbor* tanpa seleksi fitur. Semua fitur yang ada dalam dataset akan digunakan secara langsung dalam proses pemodelan, tanpa ada pengurangan atau pemilihan fitur berdasarkan relevansi atau pentingnya terhadap target prediksi.

Pemodelan tanpa seleksi fitur memungkinkan model untuk memanfaatkan semua informasi yang tersedia dalam dataset, termasuk fitur yang mungkin dianggap kurang relevan atau memiliki korelasi rendah dengan target. Pendekatan ini memberikan keuntungan dalam beberapa kasus, terutama ketika tidak ada informasi yang jelas tentang mana fitur yang paling penting atau ketika dataset relatif kecil sehingga mengurangi risiko *overfitting*.

Namun, penggunaan semua fitur tanpa seleksi juga memiliki potensi kelemahan. Model bisa menjadi lebih kompleks dan memerlukan lebih banyak sumber daya komputasi, terutama untuk algoritma seperti *Random Forest* yang secara alami memiliki kompleksitas tinggi. Selain itu, fitur yang tidak relevan atau (noise) bisa mengurangi akurasi model dan menyebabkan *overfitting*, di mana model menjadi sangat sesuai dengan data training tetapi tidak mampu melakukan generalisasi dengan baik pada data baru.

Meskipun demikian, pendekatan tanpa seleksi fitur dapat memberikan gambaran awal tentang bagaimana model bekerja dengan menggunakan semua data yang tersedia. Hasil dari tahap ini bisa menjadi titik awal untuk kemudian melakukan analisis lebih lanjut, seperti seleksi fitur, untuk meningkatkan kinerja model dan membuatnya lebih efisien.

Pemodelan menggunakan algoritma *Naïve Bayes*, *Random Forest* dan *K-Nearest Neighbor* tanpa menggunakan seleksi fitur menghasilkan *Random Forest* sebagai model terbaik dengan rata-rata akurasi 94.97%, precision 95%, recall 94%.

```
Rata-rata Akurasi, Precision, dan Recall:
              Mean Accuracy Mean Precision Mean Recall  \
Naive Bayes        0.929444       0.932869    0.929444
k-NN               0.934715       0.935113    0.934715
Random Forest      0.949721       0.950539    0.949721

                                              Accuracy per Fold  \
Naive Bayes    [0.9298245614035088, 0.8859649122807017, 0.912...
k-NN           [0.9385964912280702, 0.8947368421052632, 0.938...
Random Forest  [0.9473684210526315, 0.9210526315789473, 0.947...

                                             Precision per Fold  \
Naive Bayes    [0.947685703620754, 0.8969923292160054, 0.9122...
k-NN           [0.94579383571117, 0.9015171858216972, 0.94066...
Random Forest  [0.9513963480128893, 0.9236221867800815, 0.945...

                                                Recall per Fold
Naive Bayes    [0.9298245614035088, 0.8859649122807017, 0.912...
k-NN           [0.9385964912280702, 0.8947368421052632, 0.938...
Random Forest  [0.9473684210526315, 0.9210526315789473, 0.947...
```

Gambar 4. Evaluasi Model Tanpa Seleksi Fitur

Hasil pemodelan tanpa menggunakan seleksi fitur tersebut disajikan pada Tabel 4.2 Hasil Evaluasi Model Tanpa Seleksi Fitur.

Tabel 3. Hasil Evaluasi Model Tanpa Seleksi Fitur

Model	Accuracy (%)	Precision (%)	Recall (%)
Naïve Bayes	92.94%	93.28%	92.94%
Random Forest	94.97%	95.05%	94.97%
K-Nearest Neighbor	93.47%	93.51%	93.47%

Berdasarkan hasil pengujian menggunakan *Cross-Validation* dengan 10 fold, didapatkan bahwa model *Random Forest* memiliki akurasi rata-rata sebesar 94.97%, merupakan nilai tertinggi di antara ketiga model. Hal ini menunjukkan bahwa *Random Forest* memiliki performa terbaik dalam memprediksi label secara keseluruhan. Selain itu, precision rata-rata sebesar 95.05% menunjukkan bahwa model ini sangat akurat dalam mengidentifikasi positif, melebihi *Naive Bayes* dan *k-NN*. Recall rata-rata sebesar 94.97% menunjukkan bahwa *Random Forest* juga sangat efektif dalam menangkap kasus positif, sebanding dengan precision-nya. Dengan demikian, *Random Forest* menunjukkan performa terbaik di semua metrik (akurasi, precision, dan recall), yang mencerminkan kemampuannya dalam menangani data dan membuat prediksi yang akurat.

3.1.4.3 Pemodelan Dengan Seleksi Fitur

Dalam penelitian ini, fitur-fitur yang paling relevan dalam memprediksi status keberangkatan calon jamaah haji akan dipilih menggunakan metode *Recursive Feature Elimination* dengan *Cross-Validation* (*REFCV*). Metode ini bertujuan untuk mengeliminasi fitur-fitur yang tidak signifikan dan memilih fitur-fitur yang memiliki pengaruh terbesar terhadap variabel target.

Dari seluruh fitur yang ada, *REFCV* telah memilih 14 fitur terbaik, yaitu: nomor_porsi, tanggal_daftar, usia_daftar, nama, nama_ayah, pekerjaan, tempat_lahir, tanggal_lahir, alamat, desa, kode_pos, bank, cabang, nik. Pada hasil *REFCV*, fitur seperti 'nomor_porsi', 'tanggal_daftar', 'nama', dan 'nama_ayah' terpilih sebagai fitur penting dalam pemodelan. Namun, fitur-fitur ini mungkin kurang relevan dalam konteks prediksi pembatalan haji. Hal ini bisa disebabkan oleh beberapa faktor, seperti bias dalam data atau korelasi yang tidak langsung tetapi signifikan secara statistik.

Fitur-fitur seperti 'nomor_porsi' dan 'tanggal_daftar' mungkin merefleksikan urutan pendaftaran atau kelompok tertentu yang memiliki kecenderungan berbeda dalam hal pembatalan. Sementara itu, 'nama' dan 'nama_ayah' mungkin terpilih karena variasi tertentu dalam data yang terkait dengan faktor-faktor lain, meskipun secara logika fitur-fitur ini tidak seharusnya mempengaruhi hasil prediksi.

Setelah fitur-fitur ini terpilih, model prediksi dibangun menggunakan tiga algoritma berbeda yaitu *Naive Bayes*, *K-Nearest Neighbor*s (*k-NN*), dan *Random Forest*. Evaluasi model dilakukan melalui metode *Cross-Validation* dengan 10 fold untuk mendapatkan estimasi akurasi yang lebih stabil dan reliabel.

Hasil evaluasi menunjukkan bahwa *Naive Bayes* memperoleh akurasi rata-rata sebesar 93% setelah menggunakan fitur-fitur yang dipilih oleh *REFCV*. Skor akurasi pada setiap fold berkisar antara 86,73% hingga 96,46%, dengan variasi yang menunjukkan konsistensi performa model dalam berbagai subset data.

k-NN juga menunjukkan performa yang sangat baik dengan akurasi rata-rata sebesar 93%. Skor per fold bervariasi antara 89,47% hingga 97,35%, yang menunjukkan bahwa *k-NN* dapat bekerja efektif dengan fitur-fitur yang dipilih.

Random Forest memberikan hasil yang paling optimal dengan akurasi rata-rata sebesar 95%. Skor per fold pada *Random Forest* berkisar antara 91,15% hingga 99,11%, menunjukkan bahwa model ini memiliki kemampuan yang kuat dalam menangani variabilitas data dan memberikan prediksi yang akurat.

Proses seleksi fitur menggunakan *REFCV* berhasil meningkatkan performa model, terutama pada *Random Forest* yang menunjukkan akurasi tertinggi di antara ketiga algoritma yang diuji. Pemilihan fitur yang tepat terbukti mampu menyederhanakan model tanpa mengorbankan akurasi, bahkan meningkatkan kemampuan model dalam memprediksi status keberangkatan calon jamaah haji.

Tabel 4. Hasil Evaluasi Model Dengan Seleksi Fitur

Model	Accuracy (%)	Precision (%)	Recall (%)
Naïve Bayes	93%	93%	90%
Random Forest	95%	96%	96%
K-Nearest Neighbor	93%	89%	90%

4. Hasil dan Pembahasan

4.1 Penentuan Model Terbaik

Dalam menentukan model terbaik, dilakukan perbandingan performa antara model yang menggunakan seluruh fitur dan model yang telah melalui seleksi fitur menggunakan *REFCV*.

Tabel 5. Perbandingan Performa dengan Seluruh Fitur dan Seleksi Fitur

Fitur/Atribut	Model	Accuracy (%)	Precision (%)	Recall (%)
Tanpa seleksi fitur (semua fitur)	*Naïve Bayes*	92.94%	93.28%	92.94%
	Random Forest	94.97%	95.05%	94.97%
	K-Nearest Neighbor	93.47%	93.51%	93.47%
Dengan seleksi fitur *REFCV*	*Naïve Bayes*	93%	93%	90%
	Random Forest	95%	96%	96%
	K-Nearest Neighbor	93%	89%	90%

Hasil perbandingan menunjukkan bahwa:

a. ***Naïve Bayes*:**
   - Dengan seluruh fitur, model ini mencapai akurasi 92.94%, precision 93.28%, dan recall 92.94%.
   - Setelah seleksi fitur dengan *REFCV*, akurasi meningkat sedikit menjadi 93%, namun precision dan recall menjadi lebih seimbang pada 93% dan 90%.

b. ***Random Forest*:**
   - Model ini menunjukkan performa terbaik baik dengan semua fitur maupun setelah seleksi fitur. Dengan semua fitur, akurasi mencapai 94.97%, precision 95.05%, dan recall 94.97%.
   - Setelah seleksi fitur, performa meningkat menjadi akurasi 95%, precision 96%, dan recall 96%. Hal ini menunjukkan bahwa seleksi fitur *REFCV* mampu meningkatkan kemampuan model dalam prediksi yang lebih tepat.

c. ***K-Nearest Neighbor* (*k-NN*):**
   - Dengan seluruh fitur, *k-NN* memiliki akurasi 93.47%, precision 93.51%, dan recall 93.47%.
   - Setelah seleksi fitur, akurasi tetap pada 93%, namun precision sedikit menurun menjadi 89%, sedangkan recall menjadi 90%.

Random Forest tetap menjadi model terbaik, baik dengan seluruh fitur maupun setelah seleksi fitur. Namun, seleksi fitur dengan *REFCV* terbukti mampu meningkatkan performa secara keseluruhan, terutama untuk *Random Forest* yang mengalami peningkatan signifikan pada precision dan recall.

5. Kesimpulan

Penelitian ini telah membandingkan tiga model pembelajaran, yaitu *Naive Bayes*, *Random Forest*, dan *K-Nearest Neighbor* , dalam memprediksi potensi pembatalan pendaftaran haji Indonesia oleh calon jemaah haji. Berdasarkan hasil analisis performa model menggunakan seluruh fitur dan seleksi fitur dengan *REFCV*, dapat disimpulkan bahwa pemodelan dengan seleksi fitur mampu memberikan hasil yang lebih optimal. Seleksi fitur *REFCV* terbukti meningkatkan akurasi, precision, dan recall terutama pada model *Random Forest*, yang menunjukkan peningkatan paling signifikan dibandingkan model lain.

Secara keseluruhan, *Random Forest* konsisten menjadi model dengan performa terbaik, baik sebelum maupun setelah seleksi fitur. Meskipun *Naïve Bayes* dan *k-NN* juga menunjukkan performa yang baik, *Random Forest* unggul dalam hal akurasi dan keseimbangan antara precision dan recall, terutama setelah dilakukan seleksi fitur.

Dalam konteks dataset ini, *Random Forest* dengan seleksi fitur *REFCV* direkomendasikan sebagai model terbaik untuk digunakan, karena mampu memberikan prediksi yang lebih tepat dan efisien dengan mengurangi fitur yang kurang relevan tanpa mengorbankan akurasi.

Berdasarkan hasil perangkingan atribut dan analisis performa model, berikut adalah fitur-fitur yang paling mempengaruhi pembatalan haji:

a. Usia Daftar
   Usia saat mendaftar dapat mempengaruhi peluang pembatalan, karena mungkin mencerminkan kesiapan fisik dan finansial calon haji.
b. Pekerjaan:
   Jenis pekerjaan mungkin berkaitan dengan stabilitas ekonomi atau fleksibilitas waktu yang mempengaruhi keputusan untuk melanjutkan atau membatalkan keberangkatan.
c. Tempat Lahir dan Tempat Tinggal
   Lokasi geografis dapat mempengaruhi aksesibilitas dan kesiapan logistik, sehingga memengaruhi keputusan pembatalan.
d. Tanggal Lahir
   Tanggal lahir bisa menjadi indikator usia yang juga berpengaruh dalam pengambilan keputusan.

e. Alamat
   Lokasi tempat tinggal juga bisa menunjukkan faktor ekonomi atau sosial yang berhubungan dengan pembatalan haji.
f. Bank dan Cabang
   Bank dan cabang yang digunakan untuk setoran haji bisa saja terkait dengan akses atau kebijakan finansial yang berbeda, mempengaruhi keputusan akhir calon haji.

5.2 Keterbatasan Penelitian

Berikut Adalah beberapa keterbatasan yang perlu diantisipasi dalam interpretasi hasil:

5.2.1 Keterbatasan Data

a. Ukuran Sampel:
   Dataset 1,133 records terbatas pada satu kabupaten (Pringsewu), belum merepresentasikan diversitas nasional
b. Class Imbalance:
   Rasio batal:berangkat = 13.1%:86.9% dapat menyebabkan bias prediksi
c. Temporal Coverage:
   Data 2019-2024 termasuk periode pandemi COVID-19 yang abnormal
d. *Missing Features*:
   Tidak tersedia data objektif seperti rekam medis, tes kesehatan, atau assessment finansial detail

5.2.2 Keterbatasan Metodologi

a. Hyperparameter Tuning:
   Belum melakukan exhaustive grid search untuk optimal hyperparameters
b. External Validation:
   Model belum divalidasi pada dataset dari wilayah lain

5.2.3 Keterbatasan Generalisasi

Model dilatih dengan data Lampung, mungkin tidak optimal untuk daerah dengan karakteristik demografis berbeda.

Daftar Pustaka

[1] MA. Dr.H.Johari and M. Dr.H. Johar Arifin, Lc, "TUNTUNAN HAJI UMROH.pdf," 2019.

[2] Z. Munawaroh, "Efektifitas Siskohat Dalam Penyelenggaraan Ibadah Haji," 2015.

[3] Kementerian Agama RI, Jumlah Pendaftar Baru Jemaah Haji Indonesia Menurut Jenis Kelamin, 03-Sep-2022. [Online]. Available: https://www.citationmachine.net/ieee/cite-a-website/custom. [Accessed: 01-Apr-2023].

[4] F. Fachrudin, *Kenapa Calon Jemaah Haji Diimbau Tidak Membatalkan Keberangkatannya?*, 10-Oct-2017. [Online]. Available: https://nasional.kompas.com/read/2017/10/18/15421821/kenapa-calon-jemaah-haji-diimbau-tidak-membatalkan-keberangkatannya?page=all. [Accessed: 01-Apr-2023].

[5] M. S. Pathan, A. Nag, M. M. Pathan, and S. Dev, "Analyzing the impact of feature selection on the accuracy of heart disease prediction," *Healthcare Analytics*, vol. 2, Nov. 2022, doi: 10.1016/j.health.2022.100060.

[6] P. Misra and A. Singh Yadav, "Improving the Classification Accuracy using *Recursive Feature Elimination* with *Cross-Validation*," *International Journal on Emerging Technologies*, vol. 11, no. 3, pp. 659–665, 2020, [Online]. Available: www.researchtrend.net

[7] A. B. Wibisono and A. Fahrurozi, "Perbandingan Algoritma Klasifikasi Dalam Pengklasifikasian Data Penyakit Jantung Koroner," *Jurnal Ilmiah Teknologi dan Rekayasa*, vol. 24, no. 3, pp. 161–170, 2019, doi: 10.35760/tr.2019.v24i3.2393.

[8] A. Purwanto *et al.*, "ANALISA PERBANDINGAN KINERJA ALGORITMA C4.5 DAN ALGORITMA *K-NEAREST NEIGHBOR*S UNTUK KLASIFIKASI PENERIMA BEASISWA," 2023. [Online]. Available: https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index.

[9] C. Schröerab, F. Kruseb, and J. M. Gómezb, "A Systematic Literature Review on Applying CRISP-DM Process Model," Elsevier, vol. 181, no. 2019, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199

[7] J. Han, M. Kamber, and J. Pei, "*Data mining*. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)," 2011.

[8] G. A. Marcoulides, *Discovering Knowledge in Data: an Introduction to Data mining*, vol. 100, no. 472. 2005. doi: 10.1198/jasa.2005.s61.

[9] F. Gorunescu, *Data mining. Intelligent Systems Reference Library*. 2011.

[10] M. S. Wibawa, K. Dwi, and P. Novianti, "Konferensi Nasional Sistem & Informatika," 2017.