BAB III
METODE PENELITIAN

3.1. Objek Penelitian

Objek penelitian ini berfokus pada prediksi pendaftaran haji, dengan dataset berjumlah 27.000 dataset yang dibagi menjadi 9 Variabel yaitu no, no.porsi, tgl.daftar, nama, Jenis kelamin, bank, cabang, Kecamatan, status haji. Penelitian ini bertujuan untuk mengembangkan metode prediksi dalam menentukan tingkat akurasi prediksi pendafataran haji di Kabupaten Karawang. Data set dalam penelitian ini diperoleh dari Kantor Kementerian agama Kabupaten Karawang

3.2. Waktu Penelitian

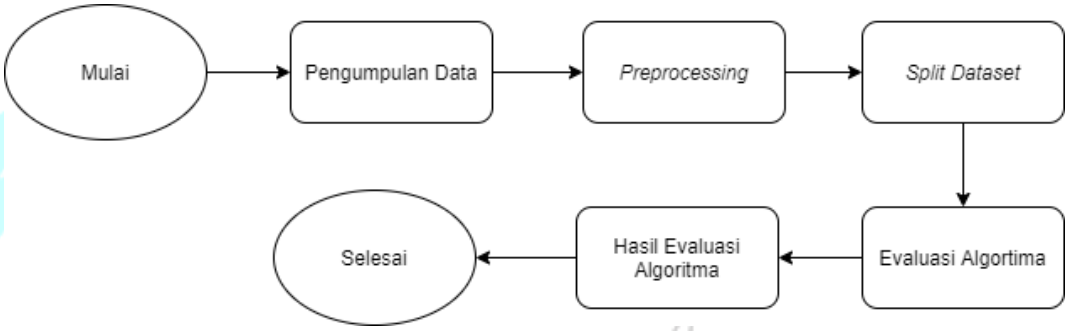
Penelitian dilakukan dengan memproses data yang telah dilakukan. Pengumpulan data berlangsung:

Tabel 3. 1 Waktu Penelitian

No	Tahap Penelitian	Oktober				November				Desember				Januari				Februari			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	Pengumpulan data																				
2	Preprocessing																				
3	Ekstrasi Fitur																				
4	Split dataset																				
5	Implementasi Algoritma																				
6	Hasil Evaluasi																				

3.3. Prosedur Penelitian

Prosedur penelitian ini dirancang untuk memberikan langkah-langkah sistematis dalam mengumpulkan dan menganalisis data, guna mencapai tujuan penelitian yang telah ditetapkan. Berikut gambar *flowchart* untuk prosedur penelitian.



Gambar 3. 1 Prosedur penelitian

3.3.1. Pengumpulan Data:

NO	NO. PORSI	TGL. DAFTAR	NAMA	JENIS	BANK	CABANG	KECAMATAN	STATUS HAJI
1	1001278149	03/02/2020	LEILY MARLINAWATY	P	BMI	CAPEM KARAWANG	TELUKJAMBE TIMUR	BELUM
2	1001278153	03/02/2020	SUNARTO	L	BMI	CAPEM KARAWANG	TELUKJAMBE TIMUR	BELUM
3	1001278177	03/02/2020	SITI ROMELAH	P	BSM	KCP CIKAMPEK	KOTA BARU	BELUM
4	1001278182	03/02/2020	DEDI DARMADI	L	BSM	KCP CIKAMPEK	KOTA BARU	BELUM
5	1001278201	03/02/2020	ENDANG JAYADI	L	BNIS	KCP KARAWANG	TELUKJAMBE BARAT	BELUM
6	1001278202	03/02/2020	WIDANINGSIH	P	BNIS	KCP KARAWANG	TELUKJAMBE BARAT	BELUM
7	1001278227	03/02/2020	CHAIRANNUR HUTABAR	P	BNIS	KCP KARAWANG	TELUKJAMBE TIMUR	BELUM
8	1001278240	03/02/2020	CARTIM HASANUDIN	L	BMI	CAPEM KARAWANG	PEDES	BELUM
9	1001278241	03/02/2020	RATNASARI	P	BMI	CAPEM KARAWANG	PEDES	BELUM
10	1001278252	03/02/2020	WARTI	P	MEGA SYARIAH	KCP KARAWANG	CILEBAR	BELUM
11	1001278262	03/02/2020	BHENI DAVID DWIANA	L	BNIS	KCP KARAWANG	TELUKJAMBE TIMUR	BELUM
12	1001278263	03/02/2020	MUHTAR	L	BANK SINARMAS	KCS Karawang	TIRTAJAYA	BELUM
13	1001278267	03/02/2020	ESEM	P	BANK SINARMAS	KCS Karawang	TIRTAJAYA	BELUM
14	1001278373	03/02/2020	YUS KUSHENDAR	L	BSM	KCP KARAWANG	KARAWANG TIMUR	BELUM
15	1001278377	03/02/2020	JUSUF ALI SURNAMA	L	BMI	CAPEM KARAWANG	TIRTAJAYA	BELUM
16	1001278384	03/02/2020	LAILA SARI	P	BSM	KCP KARAWANG	KARAWANG TIMUR	BELUM
17	1001278420	03/02/2020	JAJANG SUPRIATNA	L	BANK PERMATA SYARIAH	CAB KARAWANG TUPAREV	KARAWANG TIMUR	BELUM
18	1001278428	03/02/2020	TIN KARTIKA	P	BANK PERMATA SYARIAH	CAB KARAWANG TUPAREV	KARAWANG TIMUR	BELUM
19	1001278437	03/02/2020	CECEP MULYAWAN	L	BMI	CAPEM KARAWANG	KLARI	BELUM
20	1001278442	03/02/2020	IKHSANUL ARIFIN	L	BANK SINARMAS	KCS Karawang	KARAWANG BARAT	BELUM
21	1001278446	03/02/2020	ENENG YULYANI	P	BANK SINARMAS	KCS Karawang	KARAWANG BARAT	BELUM
22	1001278453	03/02/2020	ETI SUMIATI	P	BMI	CAPEM KARAWANG	KLARI	BELUM
23	1001278496	03/02/2020	NURAENI OKTAVIA	P	BSM	KCP KARAWANG	KLARI	BELUM
24	1001278498	03/02/2020	JUNAIDI	L	BSM	KCP KARAWANG	KLARI	BELUM
25	1001278503	03/02/2020	AGUS SALIM	L	BANK SINARMAS	KCS Karawang	TELUKJAMBE TIMUR	BELUM
26	1001278505	03/02/2020	DARSINAH	P	BMI	CAPEM KARAWANG	JAYAKERTA	BELUM
27	1001278506	03/02/2020	MUKSIN	L	BMI	CAPEM KARAWANG	JAYAKERTA	BELUM
28	1001278509	03/02/2020	TETI MARIA	P	BANK SINARMAS	KCS Karawang	TELUKJAMBE TIMUR	BELUM
29	1001278530	03/02/2020	NENDEN SUDARMINI	P	BANK SINARMAS	KCS Karawang	KARAWANG BARAT	BELUM
30	1001278534	03/02/2020	DIDIN BADRUDIN	L	BANK SINARMAS	KCS Karawang	KARAWANG BARAT	BELUM

Tabel 3.2. Pengumpulan data

Tahap pertama dalam penelitian ini adalah mengumpulkan data historis pendaftaran calon jamaah haji dari sumber resmi. Data ini diperoleh dari Kementerian Agama Kabupaten Karawang, yang memiliki sistem informasi resmi bernama SISKOHAT (Sistem Informasi dan Komputerisasi Haji Terpadu).

Data yang dikumpulkan harus komprehensif dan representatif, sehingga dapat digunakan untuk membangun model prediksi yang kuat. Oleh karena itu, data yang

dikumpulkan mencakup berbagai aspek yang dianggap berpengaruh terhadap jumlah pendaftar haji.

Beberapa variabel utama dalam dataset yang digunakan dalam penelitian ini, jenis data yang dikumpulkan meliputi:

1. Nama Jamaah: Identitas resmi calon jamaah haji.
2. Nomor Rekening: Rekening bank yang digunakan untuk pembayaran biaya haji.
3. NIK (Nomor Induk Kependudukan): Nomor identifikasi penduduk berdasarkan data kependudukan.
4. Tanggal Pendaftaran: Waktu ketika calon jamaah mendaftar dalam sistem.
5. Status Pendaftaran: Menunjukkan apakah calon jamaah sudah terverifikasi atau masih menunggu antrean.
6. Nomor Validasi: Nomor unik yang digunakan untuk memvalidasi pendaftaran calon jamaah.
7. Nomor Porsi: Nomor antrean calon jamaah dalam daftar tunggu keberangkatan haji.
8. Nama Bank: Nama bank tempat calon jamaah melakukan pembayaran biaya haji.
9. Nama Cabang: Cabang bank tempat pendaftaran dilakukan.
10. Jenis Kelamin: Apakah calon jamaah laki-laki atau perempuan.
11. Usia Pendaftar: Umur calon jamaah pada saat mendaftar.
12. Status Ekonomi: Informasi mengenai kondisi ekonomi calon jamaah berdasarkan penghasilan.

Variabel dari dataset tersebut digunakan pada penelitian. Tujuannya adalah sebagai berikut:

1. Menyiapkan dataset yang lengkap dan akurat agar dapat digunakan dalam proses analisis prediksi.
2. Memastikan data yang dikumpulkan relevan dengan tujuan penelitian, sehingga hasil prediksi dapat diandalkan.
3. Memperoleh informasi yang cukup untuk melatih dan menguji model regresi yang digunakan dalam penelitian.

3.3.2. Preprocessing Data:

Setelah data dikumpulkan, tahap selanjutnya adalah membersihkan dan mempersiapkan data agar siap digunakan dalam proses analisis prediksi. Data yang tidak bersih atau tidak lengkap dapat menyebabkan kesalahan dalam model prediksi, sehingga tahap ini sangat penting untuk meningkatkan akurasi hasil penelitian. Langkah-Langkah dalam *Preprocessing* Data:

1. Menghapus *Missing Values* (Data yang Hilang)

Missing *values* terjadi ketika suatu data tidak memiliki nilai dalam satu atau lebih atribut. Data yang hilang dapat menyebabkan bias dalam model prediksi dan mengurangi akurasi analisis. Penanganan diperlukan dalam menghadapi missing values, sebagai berikut :

1) Jika jumlah data yang memiliki *missing values* sedikit, maka baris tersebut perlu dihapus dengan menggunakan fungsi *dropna()* pada bahasa pemrograman python.
2) Jika jumlah data yang memiliki *missing values* cukup banyak, maka perlu dilakukan teknik imputasi dengan mengisi nilai yang hilang menggunakan rata-rata (*mean*), *median*, dan *modus*, serta *Fillna().*

Terdapat *pseudocode* dan contoh dari penanganan *Missing Values* Sebagai Berikut :

1. Missing Value

BEGIN

INPUT: *Dataset* yang digunakan Data Haji

// Langkah 1: Baca isi dataset

// Langkah 2: Cek total nilai kosong diseluruh dataframe

// Langkah 3: Simpan hasil data yang telah di cek total nilai kosong

// Langkah 4: Tampilkan hasil sebelum dan sesudah cek total nilai kosong *OUTPUT*: Tampilkan isi dataset Haji sebelum dan sesudah cek missing value

END

Contoh Hasil Sebelum dan Sesudah

Tabel 3. 2 Hasil dan Sesudah Penanganan *Missing Values*

Sebelum *Missing values*		Sesudah *Missing values*	
Nama	Kecamatan	Nama	Kecamatan
SUKARYA	TIRTAJAYA	SUKARYA	TIRTAJAYA
DEDEH YUNINGSIH	TIRTAJAYA	DEDEH YUNINGSIH	TIRTAJAYA
KHOIRUL UMAM	NaN	KHOIRUL UMAM	TIRTAJAYA
JEJEN ZAINI MIFTAH	TEGALWARU	JEJEN ZAINI MIFTAH	TEGALWARU

2. Menghapus Data Duplikat

Data duplikat terjadi ketika satu atau lebih entri dalam dataset memiliki nilai yang sama di beberapa atau semua atribut yang relevan. Keberadaan data duplikat dapat menyebabkan hasil analisis yang tidak valid, karena informasi yang berulang dapat memberikan bobot yang lebih besar dari seharusnya pada model prediksi. Penanganannya sebagai berikut :

1) Data duplikat dapat ditangani dengan melakukan identifikasi apakah terdapat data ganda pada atribut atau kolom unik seperti Nama Jamaah, Nomor Rekening, dan Nomor Porsi.
2) Penaganan data duplikat juga bisa dengan menggunakan fungsi *drop_duplicates()* pada bahasa pemrograman python.

2. Duplicate

BEGIN

INPUT: *Dataset* yang sudah dilakukan pengecekan data duplikat

// Langkah 1: Baca isi dataset haji

// Langkah 2: Cek jumlah data duplikat

// Langkah 3: Tampilkan hasil sebelum dan sesudah di cek nya data duplikat

kosong *OUTPUT*: Tampilkan isi dataset Haji sebelum dan sesudah cek *Duplicate*

END

Contoh Hasil Sebelum dan Sesudah

Tabel 3. 3 Hasil dan Sesudah Penanganan Data *Duplicate*

Sebelum *Duplicate*		Sesudah *Duplicate*	
Nama	No. Porsi	Nama	No. Porsi
SUKARYA	1001278149	SUKARYA	1001278149
DEDEH YUNINGSIH	1001493868	DEDEH YUNINGSIH	1001493868
KHOIRUL UMAM	1001493868	NaN	NaN
JEJEN ZAINI MIFTAH	1001493736	JEJEN ZAINI MIFTAH	1001493736

3. Transformasi Data Kategorikal ke Numerik

Secara umum, data numerik lebih mudah diolah dibandingkan data kategorikal. Variabel kategorikal seperti jenis kelamin seperti “P” dan “L” tidak memiliki urutan atau nilai kuantitatif yang bisa dihitung dengan metode matematis, maka perlu di konversi ke dalam format numerik.

1) Beberapa variabel dalam dataset, seperti Jenis Kelamin atau Nama Bank, bersifat kategorikal.
2) Data ini dikonversi ke dalam format numerik agar bisa diproses oleh algoritma regresi, menggunakan:
   A. *One-Hot Encoding* untuk data dengan banyak kategori.
   B. *Label Encoding* untuk data dengan dua kategori (misalnya Jenis Kelamin: 0 = Laki-laki, 1 = Perempuan).

3. Transformasi Data

BEGIN

INPUT: Dataset yang sudah dilakukan pengecekan dataset Haji

// Langkah 1: Baca dataset

// Langkah 2: Membuat pivot table

// Langkah 3: *Convert* ke *one-hot-encoding* (*binary*)

// Langkah 4: Tampilkan hasil setelah diubah menjadi *binary*

kosong *OUTPUT*: Tampilkan isi dataset Haji sebelum dan sesudah cek Transformasi Data

END

Contoh Hasil Sebelum dan Sesudah

Tabel 3. 4 Hasil dan Sesudah Penanganan Transformasi Data

Sebelum Transformasi Data		Sesudah Transformasi Data	
Nama	Kecamatan	Nama	Kecamatan
SUKARYA	TIRTAJAYA	SUKARYA	1
DEDEH YUNINGSIH	TIRTAJAYA	DEDEH YUNINGSIH	1
KHOIRUL UMAM	TIRTAJAYA	KHOIRUL UMAM	1
JEJEN ZAINI MIFTAH	TEGALWARU	JEJEN ZAINI MIFTAH	2

4. Normalisasi Data
    1) Data numerik seperti Usia Pendaftar yang memiliki rentang nilai yang berbeda-beda.
    2) Untuk memastikan bahwa semua variabel memiliki bobot yang sama dalam model, dilakukan normalisasi menggunakan metode *Min-Max Scaling.*

4. NORMALISASI DATA

BEGIN

INPUT: *Dataset* yang sudah dilakukan pengecekan dataset Haji

// Langkah 1: Baca *dataset*

// Langkah 2: Identifikasi kolom numerik untuk normalisasi

// Langkah 3: Tentukan metode normalisasi *Min-Max Scaling*

// Langkah 4: Lakukan normalisasi pada kolom yang dipilih

// Langkah 5: Tampilkan hasil sebelum dan sesudah normalisasi

OUTPUT: Tampilkan isi dataset Haji sebelum dan sesudah cek Normalisasi

END

Contoh Hasil Sebelum dan Sesudah

Tabel 3. 5 Hasil dan Sesudah Penanganan Normalisasi Data

Sebelum Normalisasi Data		Sesudah Normalisasi Data	
Nama	Umur	Nama	Umur
SUKARYA	88	SUKARYA	88
DEDEH YUNINGSIH	49	DEDEH YUNINGSIH	50

Sebelum Normalisasi Data		Sesudah Normalisasi Data	
Nama	Umur	Nama	Umur
KHOIRUL UMAM	32	KHOIRUL UMAM	30
JEJEN ZAINI MIFTAH	30	JEJEN ZAINI MIFTAH	30

5. Menghapus *Outlier*
    a. *Outlier* adalah data yang memiliki nilai ekstrem dan bisa mengganggu prediksi model.
    b. *Outlier* dideteksi menggunakan metode IQR (*Interquartile Range*).

Tujuan *Preprocessing* Data

1. Memastikan bahwa data dalam kondisi optimal untuk analisis prediksi.
2. Menghilangkan informasi yang tidak relevan atau salah.
3. Memastikan data dalam format yang sesuai untuk algoritma machine learning.

5. Outlier

BEGIN

INPUT: Dataset yang sudah dilakukan pengecekan dataset Haji

// Langkah 1: Baca *dataset*

// Langkah 2: Identifikasi kolom numerik untuk *Outlier*

// Langkah 3: Identifikasi outlier menggunakan metode IQR (*Interquartile Range*)

// 3.1: Hitung Q1 (kuartil pertama) dan Q3 (kuartil ketiga)

// 3.2: Hitung IQR = Q3 - Q1

// 3.3: Tentukan batas bawah = Q1 - 1.5 * IQR

// 3.4: Tentukan batas atas = Q3 + 1.5 * IQR

// 3.5: Tandai data yang berada di luar batas sebagai *outlier*

// Langkah 4: Tangani *outlier* dengan metode (hapus/ganti dengan median)

// Langkah 5: Lakukan normalisasi pada kolom numerik menggunakan *Min-Max Scaling*

// Langkah 6: Tampilkan hasil sebelum dan sesudah normalisasi

OUTPUT: Tampilkan isi dataset Haji sebelum dan sesudah cek *Outlier*

END

Contoh Hasil Sebelum dan Sesudah

Sebelum Normalisasi Data		Sesudah Normalisasi Data	
Nama	Umur	Nama	Umur
SUKARYA	88	SUKARYA	*Outlier*
DEDEH YUNINGSIH	49	DEDEH YUNINGSIH	50
KHOIRUL UMAM	32	KHOIRUL UMAM	30
JEJEN ZAINI MIFTAH	30	JEJEN ZAINI MIFTAH	30

3.3.3. *Split* Dataset:

Setelah data diproses, langkah selanjutnya adalah membagi dataset menjadi dua bagian:

1. *Training Data* (80%) → Digunakan untuk melatih model dan mengidentifikasi pola hubungan antara variabel prediktor.
2. *Testing Data* (20%) → Digunakan untuk menguji performa model setelah pelatihan dilakukan.

Pembagian dataset ini sangat penting karena memungkinkan model untuk diuji dengan data baru yang belum pernah dipelajari sebelumnya, sehingga dapat mengetahui seberapa baik model tersebut dalam melakukan prediksi.

3.3.4. Evaluasi Algoritma:

Pada tahapan ini, dilakukan penerapan algoritma *Random Forest* dan *XGBoost* untuk prediksi pendaftaran kuota haji di tahun 2025-2030.

1) *Pseudecode Random Forest*

Tabel 3.6 *Pseudecode Random Forest*

Algoritma: Prediksi Menggunakan *Random Forest*

Input: Dataset training $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$

x_i= fitur input (misalnya: usia, jenis kelamin, tanggal pendaftaran, jumlah pendaftar per tahun, dll)

y_i= label target (misalnya: jumlah pendaftar atau kategori prediksi)

Output: Model *Random Forest* berupa kumpulan pohon keputusan untuk prediksi pendaftaran calon jamaah haji

Algoritma: Prediksi Menggunakan *Random Forest*

1. Tentukan jumlah pohon keputusan yang ingin dibangun: n_estimators
2. Untuk setiap pohon t dari 1 sampai n_estimators:

1) Lakukan *bootstrap sampling* terhadap dataset training untuk membuat subset data secara acak dengan pengembalian

2) Bangun *decision tree* pada subset data tersebut dengan langkah:

a. Pada setiap node, pilih subset acak dari fitur (misalnya √jumlah total fitur)

b. Pilih fitur terbaik berdasarkan pengukuran impurity (misalnya Gini *Index* atau *Entropy*)

c. Lanjutkan pembentukan cabang hingga Maksimal kedalaman pohon tercapai, Jumlah minimum sampel dalam node dipenuhi, Impurity minimum tercapai

3) Simpan seluruh pohon dalam model *Random Forest*

4) Untuk memprediksi data baru x (misalnya prediksi jumlah pendaftar tahun mendatang):

a. Kirimkan x ke seluruh pohon

b. Setiap pohon menghasilkan prediksi (nilai numerik atau kelas)

c. Lakukan agregasi Untuk klasifikasi: gunakan mayoritas suara (voting), Untuk regresi: gunakan rata-rata hasil prediksi semua pohon

5) Hasil agregasi menjadi prediksi akhir jumlah atau kategori pendaftar haji

2) *Pseudecode XGBoost*

Tabel 3.7 *Pseudecode XGBoost*

Algoritma: Prediksi Menggunakan *XGBoost*

Input: Dataset pelatihan $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$

$\boldsymbol{x_i}$ = fitur input (misalnya: tahun pendaftaran, umur, jenis kelamin, lokasi kecamatan, dll)

$\boldsymbol{y_i}$ = label target (jumlah pendaftar atau kelas prediksi)

Output: Model prediksi jumlah pendaftar haji berbasis *XGBoost* (*boosted decision trees*)

1. Inisialisasi model prediksi awal:

$$\boldsymbol{f_0(x) = argmin_c \sum_{i=1}^{n} L(y_1, c)}$$

$\boldsymbol{L(y_1, c)}$=Fungsi loss (misalnya: *squared error* ($\boldsymbol{L(y_1, c)^2}$))

2. Untuk interasi t = 1 hingga T (Jumlah Boosting rounds):

- Hitung *residual/pseudo* residual untuk setiap data i:

$$\boldsymbol{r_i^{(t)} = \frac{\partial L(y_i,\ F^{(t-1)}(x_i))}{\partial F^{(t-1)}(x_i)}}$$

Untuk *Squared error loss*

$$\boldsymbol{r_i^{(t)} = y_i - F^{(t-1)}(x_i)}$$

- Buat tree model $\boldsymbol{h_t(x)}$

Gunakan $\{\boldsymbol{x_i}, \boldsymbol{r_i^{(t)}}\}$ sebagai data latih

- Hitung nilai leaf *weight* $\boldsymbol{\omega_j}$ untuk setiap node j di pohon, dengan meminimalkan loss

$$\boldsymbol{\omega_j = \frac{\sum_{i \in I_j} r_i^{(t)}}{\sum_{i \in I_j} 1}}$$

Algoritma: Prediksi Menggunakan *XGBoost*

Atau lebih umum (jika menggunakan regularisasi):

$$\omega_j = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}$$

Dimana:

g_i= *first derivative* (*gradient*)

h_i= *second derivative* (*hessian*)

I_j= indeks data yang jatuh ke dalam leaf j

λ= Regularisasi

- Tambahkan hasil pohon ke model sebelumnya:

$$F^{(t)}(x) = F^{(t-1)}(x) + \eta . h_t(x)$$

η= Learning rate

Model Akhir:

$$F(x) = F_0(x) \eta \sum_{t=1}^{T} h_t(x)$$

Prediksi Data Baru x :

1. Masukan x ke semua pohon $h_1, h_2, \ldots, h_t$

2. Hitung prediksi:

$$\hat{y} = F(x) = F_0(x) \eta \sum_{t=1}^{T} h_t(x)$$

Output hasil prediksi jumlah pendaftar haji

3.3.5. Hasil Evaluasi Algoritma:

Model yang sudah dibangun akan diuji menggunakan beberapa metrik evaluasi, yaitu:

1. *R-squared* (R^2) → Menunjukkan seberapa baik model dapat menjelaskan hubungan antar variabel.
2. *Mean Squared Error* (MSE)→ Metrik yang mengukur rata-rata kuadrat selisih antara nilai sebenarnya dan nilai prediksi dari model.
3. *Mean Absoulute Error* (MAE) → Mengukur rata-rata selisih absolut antara nilai sebenarnya dan nilai prediksi.

Pseucode prediksi dengan *Random Forest*

Algortima : *Random Forest*

Fungsi :

1. *Input*

Dalam prediksi jumlah pendaftar haji menggunakan *Random Forest*, data yang digunakan terbagi menjadi data latih (*X_Train, Y_Train*) dan data uji (*X_Test, Y_Test*). Data latih berisi variabel independen seperti tanggal daftar, umur, kecamatan, bank, dan jenis kelamin, serta variabel dependen berupa jumlah pendaftar, yang digunakan untuk membangun dan melatih model.

Data uji digunakan untuk mengukur performa model dengan membandingkan hasil prediksi terhadap nilai aktual. Selain itu, model ini dipengaruhi oleh beberapa hyperparameter seperti *learning rate*, jumlah iterasi, dan teknik normalisasi data agar performa prediksi lebih optimal.

2. *Output*

Hasil dari model *Random Forest* terdiri dari dua *output* utama. Pertama, prediksi jumlah pendaftar haji berdasarkan data *historis*, yang dapat digunakan oleh pemerintah dalam perencanaan kuota dan kebijakan haji. Kedua, evaluasi performa model menggunakan metrik seperti MSE, RMSE, MAE, dan R^2 untuk mengukur akurasi prediksi. Nilai MSE dan RMSE yang kecil menunjukkan kesalahan prediksi yang rendah, sementara R^2 yang mendekati 1 menandakan model mampu menjelaskan variasi data dengan baik.

3. *Langkah*-Langkah Prediksi Pendaftaran Haji dengan *Random Forest*

Proses prediksi jumlah pendaftar haji menggunakan *Random Forest* dilakukan melalui beberapa tahapan. Pertama, pengumpulan dan persiapan data dengan mengambil data historis dari sumber terpercaya, lalu membersihkannya dari data yang tidak lengkap, duplikat, dan mengonversi *variabel kategorikal* menjadi numerik. Kedua, pembagian data menjadi data latih (80%) dan data uji (20%) untuk menguji akurasi model. Ketiga, pembangunan model *Linear Regression* dengan metode *Ordinary Least Squares* (OLS) untuk menemukan hubungan terbaik antara *variabel independen* dan jumlah pendaftar.

Setelah itu, dilakukan pelatihan model menggunakan data latih, lalu pengujian model dengan data uji untuk mengukur akurasi prediksi. Model dievaluasi dengan metrik seperti MSE, RMSE, MAE, dan R^2 untuk mengukur tingkat kesalahan dan keakuratan prediksi. Jika performa kurang optimal, model dibandingkan dengan metode lain seperti SVR, *Decision Tree Regression*, atau *Random Forest Regression* untuk mendapatkan hasil terbaik. Model yang sudah dioptimalkan kemudian digunakan untuk prediksi dan pengambilan keputusan, membantu pemerintah dalam perencanaan kuota, distribusi pendaftar, serta pengelolaan logistik dan infrastruktur ibadah haji.

Pseucode prediksi dengan *Xgboost*

Algoritma : *XgBoost*

Fungsi :

1. *Input*

Dalam prediksi jumlah pendaftar haji menggunakan *Xgboost*, data terdiri dari data latih (*X_Train, Y_Train*) untuk membangun model dan data uji (*X_Test, Y_Test*) untuk mengukur performanya. *X_Train* mencakup variabel seperti tanggal daftar, umur, kecamatan, bank, dan jenis kelamin,, sementara *Y_Train* berisi jumlah pendaftar haji pada periode tertentu. Data uji digunakan untuk menguji akurasi model dengan membandingkan hasil prediksi terhadap Y_Test.

Selain itu, model ini dipengaruhi oleh *hyperparameter*, terutama nilai *alpha* dalam *XgBoost*, yang mengontrol kekuatan regularisasi untuk mencegah

overfitting dan memilih fitur yang paling relevan. Feature scaling seperti normalisasi atau standarisasi juga diterapkan agar model bekerja lebih optimal.

2. *Output*

Hasil dari *XgBoost* terdiri dari dua *output* utama. Pertama, prediksi jumlah pendaftar haji berdasarkan data historis, yang dapat digunakan oleh lembaga seperti Kementerian Agama dalam perencanaan kuota dan distribusi pendaftar. Kedua, evaluasi performa model menggunakan metrik seperti MSE, RMSE, MAE, dan R^2 untuk mengukur akurasi prediksi. MSE dan RMSE yang kecil menunjukkan kesalahan prediksi rendah, sedangkan R^2 yang mendekati 1 menandakan model dapat menjelaskan variasi data dengan baik.

3. Langkah-Langkah Prediksi Pendaftaran Haji dengan *XgBoost*

Pembangunan model *XgBoost* untuk prediksi jumlah pendaftar haji dilakukan melalui beberapa tahapan sistematis. Pertama, pengumpulan dan persiapan data, yang mencakup pembersihan data dari nilai kosong, duplikasi, serta konversi data kategorikal menjadi numerik, dengan normalisasi atau standarisasi untuk menyamakan skala variabel. Kedua, pembagian data menjadi data latih (80%) dan data uji (20%) untuk memastikan model dapat memprediksi data baru secara akurat. Ketiga, pembangunan model *XgBoost*, yang menggunakan regularisasi L1 untuk menyaring fitur yang paling relevan agar model lebih sederhana dan tidak *overfitting*. Keempat, pelatihan model dengan data latih untuk menemukan pola terbaik antara variabel independen dan jumlah pendaftar haji. Kelima, pengujian model dengan data uji untuk mengevaluasi akurasi prediksi. Keenam, evaluasi model menggunakan metrik seperti MSE, RMSE, MAE, dan R^2, di mana nilai MSE dan RMSE yang kecil menunjukkan kesalahan prediksi rendah, serta R^2 yang mendekati 1 menandakan model menjelaskan variasi data dengan baik. Ketujuh, perbandingan model dengan metode lain seperti *Ridge Regression, Elastic Net*, dan SVR untuk memilih model terbaik. Terakhir, implementasi hasil prediksi digunakan oleh Kementerian Agama untuk menentukan kuota haji, distribusi pendaftar, serta perencanaan logistik dan infrastruktur.