MATHunesa
Jurnal Ilmiah Matematika *Volume 13 No 02*
e-ISSN : 2716-506X | p-ISSN : 2301-9115 *Tahun 2025*

ANALISIS PERBANDINGAN METODE *DECISION TREE REGRESSION* DAN *RANDOM FOREST REGRESSION* PADA PREDIKSI KEPESERTAAN JAMINAN KESEHATAN NASIONAL

Aisha Rahma Putri Masita
Departemen Aktuaria, Fakultas Sains dan Analitika Data, Institut Teknologi Sepuluh Nopember
putrimasita03@gmail.com

Prilyandari Dina Saputri
Departemen Aktuaria, Fakultas Sains dan Analitika Data, Institut Teknologi Sepuluh Nopember
prilyandaridina @its.ac.id*

R. Mohamad Atok
Departemen Aktuaria, Fakultas Sains dan Analitika Data, Institut Teknologi Sepuluh Nopember
atok @its.ac.id

Abstrak

Kesehatan adalah salah satu aspek penting dalam mencapai kesejahteraan individu dan keberhasilan pembangunan negara. Untuk menjamin akses layanan kesehatan yang merata, pemerintah Indonesia meluncurkan program Jaminan Kesehatan Nasional (JKN) pada 1 Januari 2014. Namun, jumlah peserta aktif JKN hanya mencapai sekitar 80% pada tahun 2024, yang menyebabkan defisit keuangan pada BPJS Kesehatan. Penelitian ini bertujuan untuk memprediksi jumlah peserta aktif JKN menggunakan metode *Decision Tree Regression* dan *Random Forest Regression*, serta membandingkan performa kedua metode tersebut. Variabel respon adalah jumlah peserta aktif JKN, sedangkan variabel independen berupa jumlah fasilitas kesehatan seperti jumlah puskesmas, jumlah dokter praktik perorangan, jumlah klinik pratama termasuk klinik TNI/POLRI, jumlah FKRTL, jumlah penduduk miskin, tingkat pengangguran terbuka, pengeluaran perkapita disesuaikan, jumlah penduduk, angka harapan hidup (AHH), dan rata-rata konsumsi non makanan rumah tangga yang memiliki pengeluaran telekomunikasi, pada periode data dari 2016 – 2021. Hasil penelitian menunjukkan bahwa jumlah FKRTL, jumlah penduduk, jumlah dokter, dan jumlah puskesmas secara konsisten memberikan kontribusi besar dalam prediksi pada metode *Decision Tree Regression* maupun *Random Forest Regression*. Selain itu, *Decision Tree Regression* mengidentifikasi jumlah klinik pratama dan jumlah penduduk miskin sebagai variabel tambahan yang penting dalam memprediksi jumlah peserta aktif JKN. *Decision Tree Regression* menghasilkan nilai MAE sebesar 772.464,5, nilai MAPE sebesar 27,80%, dan nilai RMSE sebesar 1.324.906. *Random Forest Regression* memberikan hasil lebih baik dengan nilai MAE sebesar 518.909,8, nilai MAPE sebesar 14,79%, dan nilai RMSE sebesar 923.888,9. Secara keseluruhan, *Random Forest Regression* terbukti lebih baik dalam akurasi prediksi terlihat dari nilai MAE, MAPE, dan RMSE yang lebih kecil dibandingkan dengan *Decision Tree Regression*. Hal ini karena kemampuannya mengurangi kesalahan dengan menggabungkan prediksi dari banyak *tree*. Penelitian ini memberikan wawasan penting bagi BPJS Kesehatan untuk mengembangkan strategi berbasis data dan optimalisasi fasilitas kesehatan sebagai strategi utama dalam meningkatkan keikutsertaan masyarakat pada program JKN.
Kata Kunci: *Decision Tree Regression*, *Random Forest Regression*, Prediksi, Jaminan Kesehatan Nasional, Kepesertaan.

Abstract

Health is one of the important aspects in achieving individual welfare and the success of national development. To ensure equal access to health services, the Indonesian government launched the National Health Insurance (JKN) program on January 1, 2014. However, the number of active JKN participants will only reach around 80% in 2024, which will cause a financial deficit in BPJS Kesehatan. This study aims to predict the number of active JKN participants using the Decision Tree Regression and Random Forest Regression methods, and to compare the performance of the two methods. The response variable is the number of active JKN participants, while the independent variables are the number of health facilities such as the number of health centers, the number of individual practicing doctors, the number of primary clinics including TNI/POLRI clinics, the number of FKRTL, the number of poor people, the open unemployment rate, adjusted per capita expenditure, population, life expectancy (AHH), and the average non-food consumption of households that have telecommunications expenditure, in the data period from 2016 - 2021. The results of the study showed that the number of FKRTL, population, number of doctors, and number of health centers consistently made a large contribution

to the prediction of the Decision Tree Regression and Random Forest Regression methods. In addition, Decision Tree Regression identified the number of primary clinics and the number of poor people as additional influential variables. Decision Tree Regression produced a MAE value of 772,464.5, a MAPE value of 27.80%, and an RMSE value of 1,324,906. Random Forest Regression gives better results with MAE value of 518,909.8, MAPE value of 14.79%, and RMSE value of 923,888.9. Overall, Random Forest Regression is proven to be better in prediction accuracy as seen from the smaller MAE, MAPE, and RMSE values compared to Decision Tree Regression. This is due to its ability to reduce errors by combining predictions from many trees. This study provides important insights for BPJS Kesehatan to develop databased strategies and optimize health facilities as the main strategy in increasing community participation in the JKN program.

Keywords: *Decision Tree Regression, Random Forest Regression, Prediction, National Health Insurance, Participation.*

PENDAHULUAN

Di Indonesia, jaminan akan hak asasi warga atas kesehatan sudah dijamin dalam Undang-Undang No. 36 tahun 2009 tentang kesehatan. Salah satu upaya pemerintah Indonesia dalam meningkatkan pelayanan kesehatan yakni dengan meluncurkan program Jaminan Kesehatan Nasional (JKN) pada 1 Januari 2014 yang dikelola oleh Badan Penyelenggara Jaminan Kesehatan Sosial Kesehatan (BPJS Kesehatan) dan bersifat wajib berdasarkan UU No. 40 tahun 2004 tentang Sistem Jaminan Sosial Nasional (SJSN) (Kementerian Kesehatan Republik Indonesia, 2016). Sejak awal peluncuran hingga saat ini, jumlah peserta program JKN terus meningkat. Pada Agustus 2024, Indonesia berhasil mencapai target UHC yang ditetapkan, dimana pada tahun 2024 jumlah peserta JKN sudah melebihi angka 98% (RR/SK BPMI Setwapres, 2024). Meskipun begitu, terdapat kendala dimana dari tahun ke tahun jumlah peserta *non*-aktif semakin meningkat. Bahkan pada tahun 2024, jumlah peserta aktif hanya sekitar 80% (Al Ishaqi, 2024). Akibat jumlah peserta *non*-aktif yang semakin meningkat, setiap tahun BPJS Kesehatan selalu mengalami defisit dan semakin besar. Hal ini dikarenakan tingginya klaim kesehatan peserta tidak sebanding dengan premi peserta. Jika defisit keuangan BPJS Kesehatan terus berlanjut, maka akan berdampak besar pada perkembangan layanan kesehatan di Indonesia (Solikhin, 2019).

Metode *Decision Tree Regression* dan *Random Forest Regression* dipilih dikarenakan keduanya merupakan metode yang populer dan sudah terbukti memiliki akurasi yang baik dalam hal memprediksi dari penelitian-penelitian terdahulu. Penelitian yang dilakukan oleh Nopianti dkk. (2022) dengan membandingkan regresi linier berganda, *Support Vector Regression*, *Decision Tree Regression*, dan regresi K-Nearest menunjukkan bahwa *Decision Tree Regression* dan regresi K-Nearest menghasilkan koefisien korelasi yang tinggi. Sedangkan penelitian yang dilakukan oleh Gatera dkk. (2023) dengan membandingkan *Random Forest Regression* dan SVM untuk kasus kecelakaan menunjukkan bahwa *Random Forest Regression* lebih baik dari SVM. Namun terdapat perbedaan hasil antara penelitian-penelitian yang membandingkan *Decision Tree Regression* dan *Random Forest Regression*. Penelitian oleh (Reddy & Chandar (2023) yang menggunakan *Decision Tree* dan *Random Forest* dalam memprediksi harga rumah menunjukkan hasil bahwa *Decision Tree Regression* lebih baik dibandingkan dengan *Random Forest Regression*. Sedangkan penelitian oleh Putra dkk. (2023) yang memprediksi harga mobil mendapatkan hasil bahwa *Random Forest* lebih baik daripada *Decision Tree*.

Sehingga penelitian ini bertujuan untuk membandingkan dan melihat seberapa besar peningkatan akurasi dari kedua metode tersebut pada kasus kepesertaan JKN. Data yang digunakan pada penelitian ini memiliki rentang periode dari tahun 2016 – 2021, dengan parameter kontrol yang ditetapkan berupa *minsplit, maxdepth*, dan *ntree*. Dari penelitian ini pula ingin mengetahui variabel-variabel apa saja yang penting dalam memprediksi kepesertaan program JKN.

KAJIAN TEORI

DECISION TREE REGRESSION

Decision Tree merupakan model *non*-parametrik yang dibangun dengan membagi kumpulan data berdasarkan variabel prediktor secara rekursif dengan tujuan memprediksi variabel respon secara optimal (Alessi & Savona, 2021). *Decision Tree* membentuk sebuah *tree* yang terdiri *node* dan cabang yang menghubungkan *node* tersebut. *Node* yang terletak di bagian bawah *tree* disebut *leaf* atau daun dan menunjukkan kelas, sedangkan *node* teratas disebut *root* atau akar. *Root node* berisi data pelatihan yang akan dibagi menjadi beberapa kelas. Semua

node kecuali daun disebut *node* keputusan atau *decision node*s karena menentukan keputusan yang akan dilakukan pada *node* tersebut berdasarkan pada satu fitur (Cios dkk., 2007).

Decision Tree memiliki tiga pendekatan klasik yakni *Classification Tree* (*tree* klasifikasi), *Regression Tree* (*tree* regresi), dan *Classification* and *Regression Tree* (CART). *Classification Tree* digunakan jika hasil prediksi berupa kelas data atau kategorik. *Regression Tree* digunakan jika hasil prediksi berupa numerik seperti harga minyak, harga saham dan sebagainya. Sedangkan CART digunakan jika memperhitungkan kedua kasus klasifikasi dan regresi (Gorunescu, 2011).

Secara garis besar, terdapat dua langkah untuk membangun *Tree*. Pertama yakni dengan membagi ruang prediktor, yaitu himpunan nilai yang mungkin untuk $X_1, X_2, \dots, X_p$, menjadi J wilayah yang berbeda dan tidak tumpang tindih, $R_1, R_2, \dots, R_J$. Kedua, untuk setiap observasi yang termasuk dalam wilayah R_J akan dibuat prediksi yang sama, yang merupakan rata-rata nilai respon observasi pelatihan di R_J (James dkk., 2013).

Dalam memilih prediktor dan titik pisah untuk membangun *tree*, dipilih prediktor yang menghasilkan *Sum Square Error* (SSE) terendah. Tujuannya adalah mencari kotak $R_1, R_2, \dots, R_J$ sedemikian sehingga meminimalkan nilai SSE, yang dirumuskan sebagai berikut:

$$SSE = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 \quad (1)$$

dimana $\hat{y}_{R_j}$ merupakan rata-rata nilai untuk observasi pelatihan kotak ke- j (James dkk., 2013).

Untuk melakukan pemisahan cabang, pertama prediktor X_j dan titik potong s dipilih sedemikian sehingga membagi ruang prediktor menjadi wilayah $R_1(j, s) = \{X | X_j < s\}$ dan $R_2(j, s) = \{X | X_j \geq s\}$ dan menghasilkan kemungkinan terbesar pengurangan SSE. Artinya, semua prediktor $X_1, X_2, \dots, X_p$ dan semua kemungkinan nilai titik potong s untuk masing-masing prediktor dipertimbangkan, lalu pilih prediktor dan titik potong sedemikian sehingga *tree* yang dihasilkan memiliki SSE terendah. Selanjutnya proses ini diulang untuk mencari prediktor dan titik potong terbaik untuk membagi data lebih jauh sehingga meminimalkan SSE di setiap wilayahnya. Namun, kali ini bukan membagi seluruh ruang prediktor, melainkan membagi salah satu dari dua wilayah yang diidentifikasi sebelumnya sehingga sekarang menjadi tiga wilayah. Proses berlanjut hingga kriteria berhenti tercapai. Setelah wilayah $R_1, R_2, \dots, R_J$ terbentuk, selanjutnya memprediksi respon untuk observasi tes tertentu menggunakan rata-rata observasi pelatihan di wilayah tempat observasi tes berada (James dkk., 2013).

Decision Tree memiliki beberapa kelebihan diantaranya yakni metode ini mudah untuk dipahami dan diinterpretasikan. Metode ini juga murah untuk dibangun karena memerlukan sejumlah kecil data pelatihan dibandingkan dengan teknik klasifikasi lainnya. Kelebihan lainnya yakni dapat menggunakan data numerik dan kategorik tanpa batasan. Metode ini menggunakan teknik statistik klasik untuk memungkinkan validasi model. Dan keakuratan nya sebanding dengan teknik klasifikasi lain untuk banyak kumpulan data sederhana (Gorunescu, 2011).

RANDOM FOREST REGRESSION

Random Forest merupakan salah satu metode *ensemble learning* yang digunakan untuk klasifikasi. Metode ini menggunakan banyak model untuk memperoleh hasil prediksi yang lebih baik dibandingkan dengan model tunggal yang berdiri sendiri (Edwards & Gaber, 2014). *Random Forest* merupakan *supervised machine learning* yang berbasis *Decision Tree* atau pohon keputusan. Salah satu keuntungan utama dalam penggunaan *Random Forest* adalah performa generalisasi yang lebih baik untuk performa pelatihan serupa dibandingkan dengan *Decision Tree* (Montesinos López dkk., 2022).

Sama seperti *Decision Tree*, *Random Forest* juga dapat digunakan untuk klasifikasi maupun regresi. *Random Forest Classification* digunakan jika respon yang diprediksi berupa kategorik atau biner (kategorik dan numerik), dan menggunakan *Random Forest Regression* jika respon berupa kontinu (numerik) (Montesinos López dkk., 2022).

Pada *Random Forest Regression*, pemisahan untuk *root node* berbentuk $X \leq c$ dan $X > c$ untuk variabel X dan ambang batas c. Pemisahan terbaik adalah yang meminimalkan jumlah kesalahan kuadrat tertimbang atau *weighted Sum of Square Errors* (SSE) yang dirumuskan sebagai berikut:

$$SSE = SSE_L \Omega_L + SSE_R \Omega_R \quad (2)$$

$$SSE_L = \sum_{i=1}^{L} (y_i - \hat{y}_L)^2 \quad (3)$$

$$SSE_R = \sum_{i=1}^{R} (y_i - \hat{y}_R)^2 \quad (4)$$

dimana SSE_L mewakili jumlah kesalahan kuadrat untuk *node* kiri, L menunjukkan jumlah elemen yang memuat partisi kiri, $\hat{y}_L$ adalah *mean* dari variabel respon elemen di partisi kiri, dan $\Omega_L = \frac{n_L}{n}$ merupakan proporsi observasi pada *node* kiri. Sedangkan SSE_R mewakili jumlah kesalahan kuadrat untuk *node* kanan, R menunjukkan jumlah elemen yang memuat partisi kanan, $\hat{y}_R$ adalah *mean* dari variabel respon elemen di partisi kanan, dan $\Omega_R = \frac{n_R}{n}$ merupakan proporsi observasi pada *node* kanan (Montesinos López dkk., 2022).

Dalam *Random Forest* terdapat istilah *tuning parameters* (parameter penyetelan) yang penting dan berpengaruh dalam kinerja prediksi. Terdapat beberapa *tuning parameters* diantaranya $ntree$, $mtry$, dan *node size*. Parameter $ntree$ menunjukkan banyaknya jumlah *tree*. Parameter $mtry$ merupakan jumlah variabel independen yang dipilih secara acak untuk dipertimbangkan pada setiap pemisahan di *Random Forest*. Secara umum pada *Random Forest Regression* nilai $mtry = p/3$ dengan p merupakan jumlah total variabel *input*. Parameter *node size* merupakan parameter yang menentukan kedalaman *tree* keputusan (Montesinos López dkk., 2022).

VARIABLE IMPORTANCE MEASURE (VIM)

Perhitungan VIM pada *Decision Tree* dan *Random Forest* hampir sama. Jika pada *Decision Tree* VIM dihitung berdasarkan satu *tree* saja, maka pada *Random Forest* VIM dihitung dengan menggabungkan VIM sejumlah $ntree$ yang ada pada model (Wei dkk., 2015).

Salah satu pendekatan paling umum dalam menghitung VIM adalah menghitung penurunan akurasi prediksi dari kumpulan data pengujian. Untuk setiap *tree*, bagian dari set pengujian dari data dilewatkan melalui *tree* dan kesalahan prediksi (*prediction error* / PE) dicatat. Setiap variabel prediktor akan dipermutasi secara acak dan PE baru akan dihitung. Perbedaan antara keduanya kemudiaan di rata-ratakan di semua *tree*, dan dinormalisasi dengan standar deviasi perbedaannya. Variabel yang menunjukkan penurunan akurasi prediksi terbesar adalah variabel terpenting (Montesinos López dkk., 2022).

MEAN ABSOLUTE ERROR (MAE)

Indikator ini menghitung nilai yang diharapkan dari kesalahan absolut. Indikator ini dirumuskan sebagai berikut:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (5)$$

dengan $\hat{y}_i$ merupakan nilai prediksi sampel ke-i dan y_i merupakan nilai sebenarnya yang sesuai untuk total n sampel (Gatera dkk., 2023). Indikator MAE mudah ditafsirkan dan kurang sensitif terhadap *outlier* (Hoxha dkk., 2023).

MEAN ABSOLUTE PERCENTAGE ERROR (MAPE)

MAPE menghitung selisih rata-rata antara nilai prediksi dan nilai aktual variabel target. Hasilnya, kesalahan dinormalisasi sehingga serupa di berbagai titik data. Hal ini dapat membantu dalam situasi di mana variabel target berfluktuasi secara luas dan ketika ketidakakuratan dalam prediksi dapat mempengaruhi kinerja model secara keseluruhan (Hoxha dkk., 2023). Perhitungan MAPE dirumuskan sebagai berikut:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (6)$$

dimana $\hat{y}_i$ merupakan nilai prediksi sampel ke-i dan y_i merupakan nilai sebenarnya yang sesuai untuk total n sampel. MAPE mudah diartikan sebagai persentase, namun memiliki kelemahan yang signifikan ketika berhadapan dengan nilai aktual nol atau mendekati nol (Hoxha dkk., 2023).

Nilai MAPE diklasifikasikan menjadi 4 kriteria yang ditunjukkan dalam Tabel 1 (Vivas dkk., 2020).

Tabel 1. Kriteria MAPE

MAPE (%)	Keterangan
< 10	Hasil prediksi sangat akurat (*Highly accurate prediction*)
$10 - 20$	Hasil prediksi baik (*good prediction*)
$20 - 50$	Hasil prediksi wajar (*reasonable prediction*)
> 50	Hasil prediksi tidak akurat (*inaccurate prediction*)

ROOT MEAN SQUARED ERROR (RMSE)

RMSE memperhitungkan simpangan baku dari residu dan dirumuskan sebagai berikut:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (7)$$

dimana $\hat{y}_i$ merupakan nilai prediksi sampel ke-i dan y_i merupakan nilai sebenarnya yang sesuai untuk total n sampel (Gatera dkk., 2023). RMSE memiliki unit yang sama dengan variabel respon sehingga

lebih mudah diinterpretasikan. Indikator ini juga sensitif terhadap *outlier* (Hoxha dkk., 2023).

METODE

Metode yang digunakan pada penelitian ini adalah metode Decision Tree Regression dan Random Forest Regression. Analisis data dilakukan dengan menggunakan bantuan RStudio dan Microsoft Excel. Langkah analisis data dimulai dengan mengumpulkan dan mempersiapkan data yang relevan untuk kebutuhan penelitian. Data jumlah peserta JKN diperoleh melalui website Sistem Monitoring Terpadu miliki Dewan Jaminan Sosial Nasional. Data yang berkaitan dengan jumlah fasilitas kesehatan didapat dari Buku Statistik JKN terbitan Dewan Jaminan Sosial Nasional dan BPJS. Data lain yang berkaitan dengan demografi dan ekonomi didapat dari website miliki Badan Pusat Statistik. Seluruh data memiliki rentang waktu antara 2016 hingga 2021 dengan unit provinsi. Variabel yang digunakan pada penelitian ini dapat dilihat pada Tabel 2.

Setelah mengumpulkan data, langkah selanjutnya adalah menghitung statistika deskriptif dan analisis korelasi antar variabel untuk mengetahui karakteristik data. Pada penelitian ini, dilakukan penelitian dengan menggunakan data tanpa *preprocessing* dan data yang dilakukan *preprocessing* berupa *standardized.* Langkah selanjutnya yakni membagi data menjadi data training (pelatihan) dan data testing (pengujian) dengan proporsi 80:20, yakni 163 data training dan 41 data testing.

Setelah data dibagi, selanjutnya dilakukan pemodelan menggunakan metode *Decision Tree Regression* meliputi:

a. Memasukkan nilai parameter dengan kombinasi *minsplit* dan *maxdepth* masing-masing bernilai 1 hingga 20.
b. Melakukan pemodelan menggunakan *Decision Tree Regression*
c. Melakukan visualisasi struktur *Decision Tree Regression*
d. Membuat plot variable *importance* dari hasil pemodelan
e. Melakukan prediksi dengan data *testing*
f. Menghitung nilai MAE, MAPE, dan RMSE berdasarkan hasil prediksi
g. Memilih model *Decision Tree Regression* yang menghasilkan nilai MAE, MAPE, dan RMSE terkecil sebagai model terbaik

Tabel 2. Variabel Penelitian

No.	Variabel	Skala	Definisi
1	X_1	Rasio	Tingkat Penyelesaian Pendidikan Menurut Jenjang Pendidikan SD
2	X_2	Rasio	Tingkat Penyelesaian Pendidikan Menurut Jenjang Pendidikan SMP
3	X_3	Rasio	Tingkat Penyelesaian Pendidikan Menurut Jenjang Pendidikan SMA
4	X_4	Rasio	Jumlah Penduduk Miskin
5	X_5	Rasio	Tingkat Pengangguran Terbuka
6	X_6	Rasio	Jumlah Puskesmas yang Bekerja Sama dengan BPJS
7	X_7	Rasio	Jumlah Dokter Praktik Perorangan yang Bekerja Sama dengan BPJS
8	X_8	Rasio	Jumlah Klinik Pratama termasuk Klinik TNI/POLRI yang Bekerja Sama dengan BPJS
9	X_9	Rasio	Jumlah Fasilitas Kesehatan Rujukan Tingkat Lanjutan (FKRTL) yang Bekerja Sama dengan BPJS
10	X_{10}	Rasio	Pengeluaran Perkapita Disesuaikan
11	X_{11}	Rasio	Angka Harapan Hidup Laki-Laki
12	X_{12}	Rasio	Angka Harapan Hidup Perempuan
13	X_{13}	Rasio	Jumlah Penduduk
14	X_{14}	Rasio	Rata-Rata Konsumsi *Non* Makanan Rumah Tangga yang Memiliki Pengeluaran Telekomunikasi
15	Y	Rasio	Jumlah Peserta Aktif Program JKN

Selanjutnya dilakukan pemodelan menggunakan metode *Random Forest Regression* meliputi:

a. Memasukkan nilai parameter *ntree* sebesar 100 dan 500.
b. Melakukan pemodelan menggunakan *Random Forest Regression*
c. Membuat plot variable *importance* dari hasil pemodelan
d. Melakukan prediksi dengan data *testing*

e. Menghitung nilai MAE, MAPE, dan RMSE berdasarkan hasil prediksi
f. Memilih model *Random Forest Regression* yang menghasilkan nilai MAE, MAPE, dan RMSE terkecil sebagai model terbaik

Langkah selanjutnya yakni melakukan perbandingan hasil pemodelan menggunakan *Decision Tree Regression* dan *Random Forest Regression* berdasarkan kriteria evaluasi MAE, MAPE, dan RMSE. Langkah terakhir yakni menarik kesimpulan

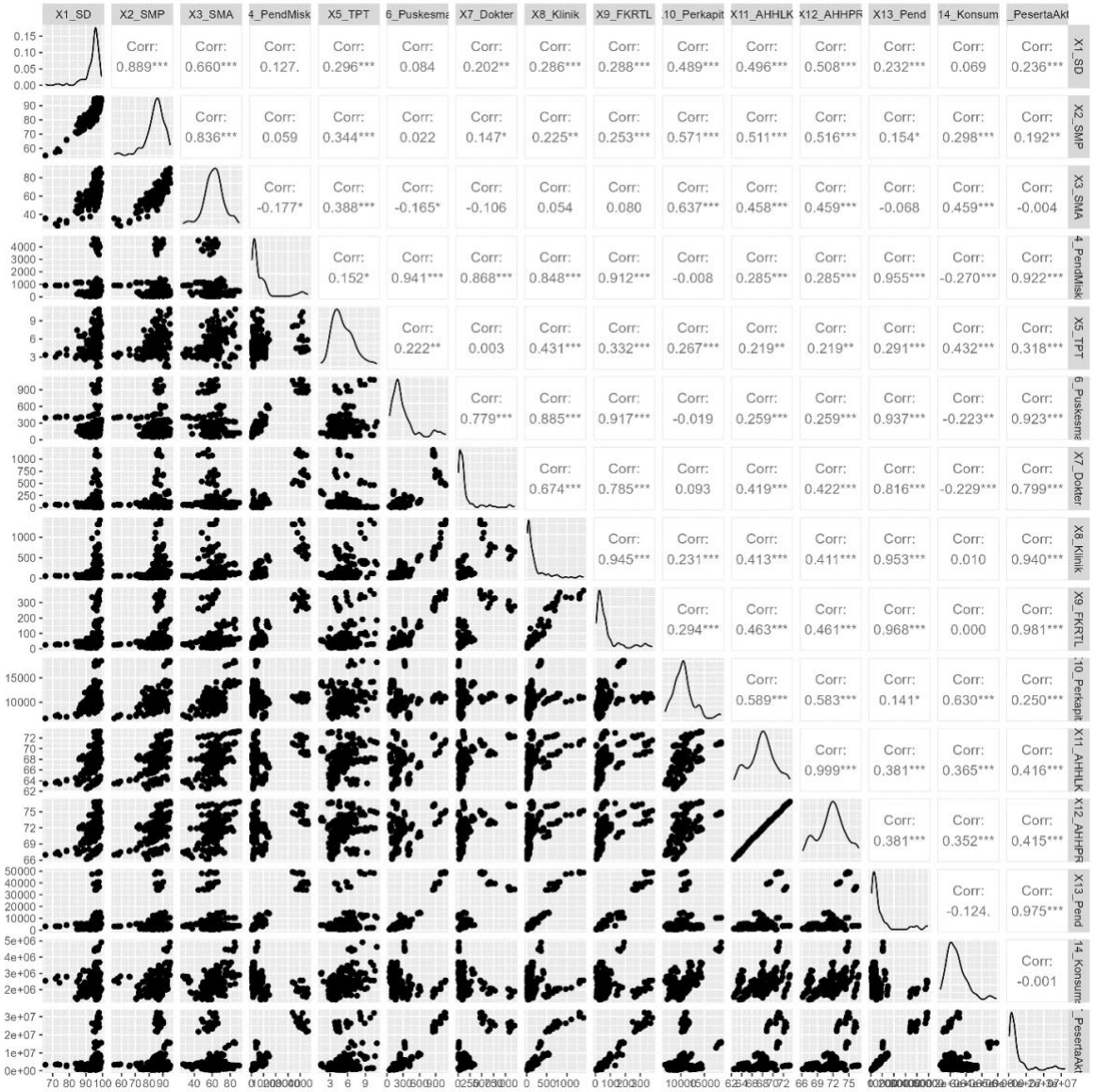
HASIL DAN PEMBAHASAN

EKSPLORASI DATA PESERTA JKN

Berikut merupakan hasil analisis data terhadap korelasi antar variabel yang dapat dilihat pada Gambar 1. Nilai korelasi antar variabel sendiri dapat dilihat pada Gambar 1 di bagian segitiga atas. Nilai korelasi 1 menunjukkan korelasi positif sempurna, korelasi sekitar -1 menunjukkan korelasi negatif sempurna, dan korelasi sekitar 0 menunjukkan tidak ada korelasi antar variabel. Contohnya, pada pasangan X_{11} dan X_{12} memiliki nilai korelasi 0,999. Nilai ini mendekati angka 1 sehingga dapat disimpulkan bahwa korelasi antar kedua variabel ini hampir positif sempurna. Lain halnya dengan pasangan X_9 dan X_{14} yang memiliki nilai korelasi 0,000

PEMODELAN *DECISION TREE REGRESSION*

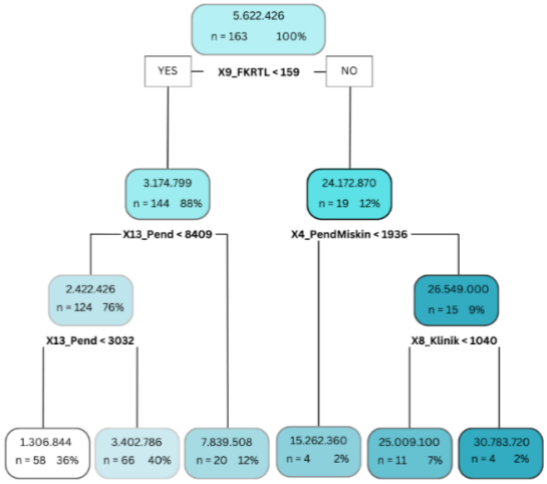
Pada penelitian ini, peneliti menguji beberapa kombinasi *minsplit* dan *maxdepth* dari nilai 1 hingga 20, sehingga total ada 400 pengujian yang dilakukan. Dari hasil pengujian, dipilih model *Decision Tree Regression* terbaik yakni model dengan



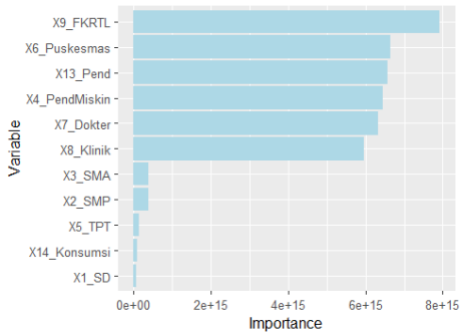
Gambar 1. Matriks Plot Variabel Penelitian

kombinasi *minsplit* dari nilai 2 hingga 13 dan *maxdepth* dengan nilai 3 hingga 20. Kombinasi-kombinasi ini menghasilkan model yang sama, baik dari bentuk *tree*, *variable importance*, dan juga nilai kriteria evaluasi.

Struktur *tree* dari model terpilih dapat dilihat pada Gambar 2. Terlihat *tree* yang dihasilkan hanya memiliki kedalaman sebesar 3 dengan variabel X_9, yakni jumlah FKRTL, menjadi pemisah pertama yang membagi data menjadi dua kelompok besar. Data yang memenuhi kondisi jumlah FKRTL lebih kecil dari 158,5 sebesar 144 data pengamatan atau sekitar 88% dengan prediksi rata-rata jumlah peserta aktif adalah 3.174.799 jiwa. Sedangkan data yang tidak memenuhi kondisi ini sebesar 19 data pengamatan atau sekitar 12% dengan prediksi rata-rata jumlah peserta aktif adalah 24.172.870 jiwa.



Gambar 2. Visualisasi Model *Decision Tree Regression* Terbaik



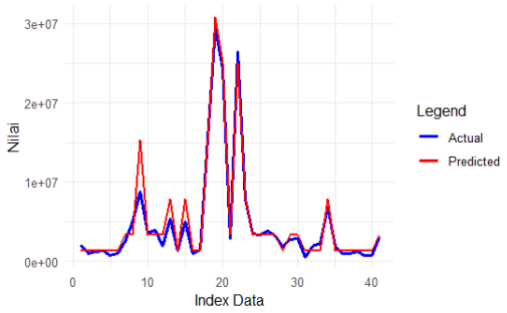
Gambar 3. Plot *Variable Importance* dari Model *Decision Tree Regression* Terbaik

Dari hasil analisis, model yang terpilih ini menemukan 11 variabel yang memiliki nilai VIM seperti yang terlihat pada Gambar 3. Namun, terlihat bahwa 6 di antaranya menunjukkan kontribusi yang jauh lebih besar terhadap prediksi jumlah peserta aktif JKN. Variabel-variabel tersebut secara berurutan adalah jumlah FKRTL, jumlah puskesmas, jumlah penduduk, jumlah penduduk miskin, jumlah dokter praktik perorangan, dan jumlah klinik pratama. Sebaliknya, 5 variabel lainnya, yaitu tingkat penyelesaian SMA, tingkat penyelesaian SMP, tingkat pengangguran terbuka, rata-rata konsumsi *non*-makanan, dan tingkat penyelesaian SD memiliki nilai VIM yang relatif kecil dan kontribusinya terhadap model dapat dianggap minimal. Perbedaan ini menunjukkan bahwa 6 variabel dengan VIM tertinggi memainkan peran utama dalam membangun prediksi, sementara 5 variabel dengan VIM rendah memiliki dampak yang kurang signifikan dalam model.

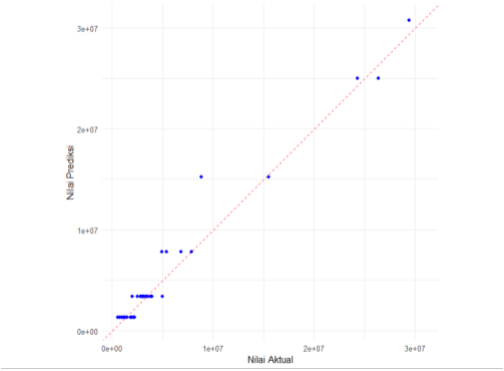
Nilai kriteria evaluasi yang dihasilkan dari model terpilih berupa nilai MAE sebesar 772.464,5 artinya secara rata-rata, model memprediksi jumlah peserta aktif dengan deviasi sekitar 772 ribu dari nilai sebenarnya. Selain itu nilai MAPE yang dihasilkan sebesar 27,80097%, artinya kesalahan prediksi rata-rata adalah sekitar 27,8% dari nilai aktual. Sedangkan nilai RMSE dihasilkan nilai sebesar 1.324.906, artinya model memiliki kesalahan prediksi rata-rata sekitar 1,32 juta dalam hal unit data.

Kinerja model regresi juga dievaluasi menggunakan *line plot* dan *scatter plot* nilai sebenarnya dan prediksi, sebagaimana ditunjukkan dalam Gambar 4 dan Gambar 5. Pada Gambar 4, terlihat bahwa garis yang terbentuk dari nilai prediksi hampir tumpang tindih dengan nilai sebenarnya. Hal ini menandakan bahwa nilai prediksi yang dihasilkan mendekati dengan nilai yang sebenarnya. Sedangkan pada Gambar 5, garis putus-putus merah merepresentasikan garis identitas di mana prediksi sempurna akan terletak, hal ini merupakan situasi di mana nilai prediksi tepat sama dengan nilai sebenarnya. Titik-titik yang dekat dengan garis ini menunjukkan prediksi yang akurat, sedangkan titik-titik yang jauh dari garis menunjukkan prediksi yang kurang akurat.

Pada penelitian ini, dilakukan pemodelan dengan menggunakan data tanpa *preprocessing* terlebih dahulu dan juga dengan data yang dilakukan *preprocessing* terlebih dahulu berupa *standardized*. Hasil kedua model menunjukkan hasil akhir yang sama.



Gambar 4. Line plot Perbandingan Nilai Aktual dengan Nilai Prediksi pada Model *Decision Tree Regression*

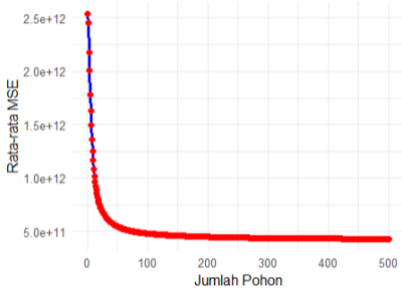


Gambar 5. *Scatterplot.* Perbandingan Nilai Aktual dengan Nilai Prediksi pada Model *Decision Tree Regression*

PEMODELAN *RANDOM FOREST REGRESSION*

Pada penelitian ini, peneliti menggunakan parameter awal berupa $ntree$ sebesar 100 dan 500, serta parameter $mtry$ dengan nilai default yakni $p/3$. Karena pada penelitian ini terdapat 14 variabel prediktor, maka nilai $mtry = \frac{14}{3} = 4{,}67$. Pada penelitian ini dilakukan replikasi sebanyak 1000 kali yang selanjutnya hasil pemodelan akan di rata-rata.

Pada Gambar 6, terlihat rata-rata MSE hasil replikasi 1000 kali menunjukkan penurunan seiring bertambahnya jumlah *tree*.



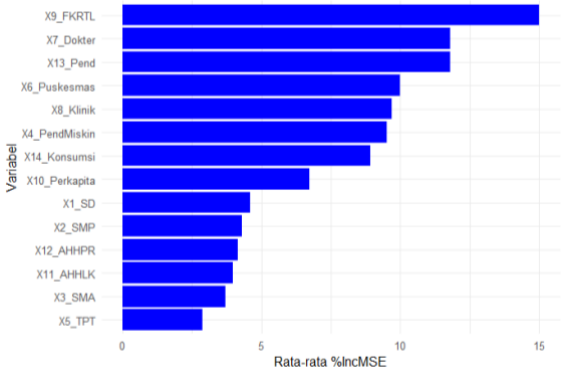
Gambar 6. Rata-Rata MSE per Jumlah *Tree (ntree)*

Tabel 3. Perbandingan Nilai Kriteria Evaluasi Model *Random Forest Regression* dengan *ntree* 100 dan 500

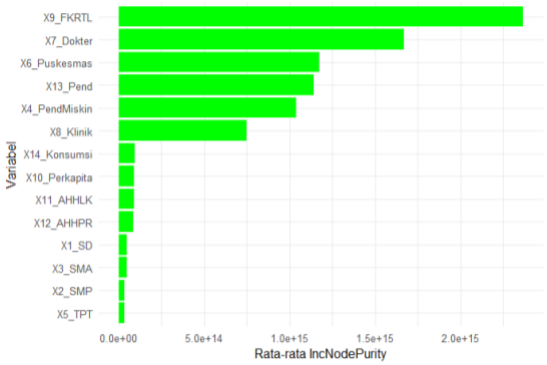
$ntree$	Rata-Rata MAE	Rata-Rata MAPE	Rata-Rata RMSE
100	520.152.3	14,89123 %	926.539,0
500	518.909.8	14,79462 %	923.888,9

Dari hasil pengujian menggunakan kriteria evaluasi, dipilih model terbaik untuk metode *Random Forest Regression* adalah model dengan $ntree$ sebesar 500. Perbandingan nilai kriteria evaluasi dari kedua model dapat dilihat pada Tabel 3. Nilai kriteria evaluasi yang dihasilkan oleh model terpilih berupa rata-rata nilai MAE sebesar 518.909,8 yang artinya secara rata-rata model memprediksi jumlah peserta aktif dengan deviasi sekitar 519 ribu dari nilai sebenarnya. Selain itu rata-rata nilai MAPE yang dihasilkan sebesar 14,79462% yang artinya kesalahan prediksi rata-rata adalah sekitar 15% dari nilai aktual. Sedangkan untuk rata-rata nilai RMSE dihasilkan nilai sebesar 923.888,9 yang artinya model memiliki kesalahan prediksi rata-rata sekitar 924 ribu dalam hal unit data.

Rata-rata *variabel importance* pada model terpilih dapat dilihat pada Gambar 7 dan Gambar 8. Dari Gambar 7, terlihat bahwa 5 variabel yang paling signifikan berdasarkan persentase kenaikan MSE berturut-turut adalah jumlah FKRTL, jumlah dokter, jumlah penduduk, jumlah puskesmas, dan jumlah klinik. Menghapus variabel-variabel ini akan meningkatkan kesalahan yang signifikan pada model. Sedangkan dari Gambar 8 terlihat bahwa 5 variabel yang paling signifikan berdasarkan peningkatan *purity* pada *node* adalah jumlah FKRTL, jumlah dokter, jumlah puskesmas, jumlah penduduk, dan jumlah penduduk miskin. Variabel-variabel tersebut memiliki kontribusi besar dalam mengklasifikasikan data ke dalam kelompok yang lebih homogen. Sehingga dapat disimpulkan bahwa variabel jumlah FKRTL, jumlah penduduk, jumlah dokter, dan jumlah puskesmas merupakan variabel yang paling penting baik dari segi kenaikan MSE maupun peningkatan *purity node*.



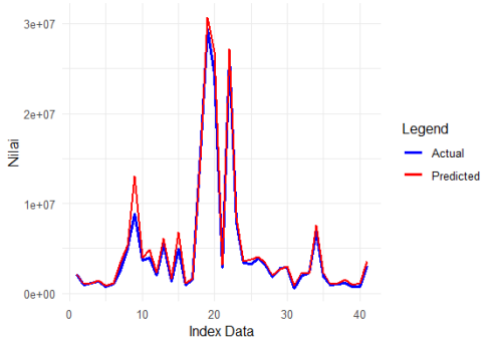
Gambar 7. Variable Importance Model *Random Forest Regression* Terbaik Berdasarkan *%IncMSE*



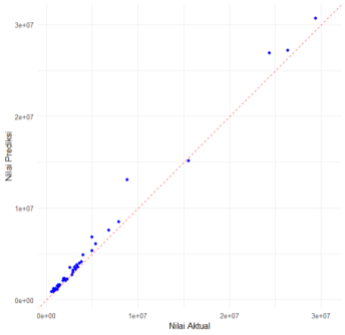
Gambar 8. *Variable Importance* Model *Random Forest Regression* Terbaik Berdasarkan *IncNodePurity*

Selain menggunakan kriteria evaluasi, kinerja model regresi juga dievaluasi menggunakan line plot dan *scatter plot* dari nilai sebenarnya dan nilai prediksi, sebagaimana ditunjukkan dalam Gambar 9 dan Gambar 10. Pada Gambar 9 terlihat bahwa garis yang terbentuk dari nilai prediksi hampir tumpang tindih dengan nilai sebenarnya. Hal ini menandakan bahwa nilai prediksi yang dihasilkan mendekati dengan nilai yang sebenarnya. Sedangkan pada Gambar 10, garis putus-putus merah merepresentasikan garis identitas di mana prediksi sempurna akan terletak, hal ini merupakan situasi di mana nilai prediksi tepat sama dengan nilai sebenarnya. Titik-titik yang dekat dengan garis ini menunjukkan prediksi yang akurat, sedangkan titik-titik yang jauh dari garis menunjukkan prediksi yang kurang akurat.

Pada penelitian ini, dilakukan pemodelan dengan menggunakan data tanpa *preprocessing* terlebih dahulu dan juga dengan data yang dilakukan *preprocessing* terlebih dahulu berupa *standardized*. Hasil kedua model menunjukkan hasil akhir yang sama.



Gambar 9. *Line plot* Perbandingan Nilai Aktual dengan Nilai Prediksi pada Model *Random Forest Regression* Terbaik



Gambar 10. Scatterplot Perbandingan Nilai Aktual dengan Nilai Prediksi pada Model *Random Forest Regression* Terbaik

PERBANDINGAN *DECISION TREE REGRESSION* DAN *RANDOM FOREST REGRESSION*

Dari kedua metode tersebut, dilakukan perbandingan untuk mengetahui metode mana yang menghasilkan prediksi lebih baik untuk memprediksi jumlah peserta aktif JKN yang dapat dilihat pada Tabel 4. Dari Tabel 4, terlihat bahwa metode *Random Forest Regression* menghasilkan nilai MAE, MAPE, dan RMSE lebih kecil dibandingkan dengan metode *Decision Tree Regression*. Dimana untuk data *testing*, nilai rata-rata MAE sebesar 518.909,8 yang artinya secara rata-rata model memprediksi jumlah peserta aktif dengan deviasi sekitar 519 ribu dari nilai sebenarnya. Untuk rata-rata MAPE dihasilkan nilai sebesar 14,79462% yang artinya kesalahan prediksi rata-rata adalah sekitar 15% dari nilai aktual. Sedangkan untuk rata-rata RMSE dihasilkan nilai sebesar 923.888,9 yang artinya model memiliki kesalahan prediksi rata-rata sekitar 924 ribu dalam hal unit data.

Sedangkan untuk data *training*, nilai rata-rata MAE sebesar 518.909,8 yang artinya secara rata-rata model memprediksi jumlah peserta aktif dengan deviasi sekitar 519 ribu dari nilai sebenarnya. Untuk rata-rata MAPE dihasilkan nilai sebesar 14,79462% yang artinya kesalahan prediksi rata-rata adalah sekitar 15% dari nilai aktual. Sedangkan untuk rata-rata RMSE dihasilkan nilai sebesar 923.888,9 yang artinya model memiliki kesalahan prediksi rata-rata sekitar 924 ribu dalam hal unit data.

Tabel 4. Perbandingan Nilai Kriteria Evaluasi Metode *Decision Tree Regression* dan *Random Forest Regression*

Model	Data	MAE	MAPE	RMSE
DTR	*Training*	770.557,9	26,6%	1.063.052
	Testing	772.464,5	27,8%	1.324.906
RFR	*Training*	162.353,9	4,3%	291.873
	Testing	518.909,8	14,8%	923.889

. Nilai MAE, MAPE, dan RMSE pada metode *Decision Tree Regression* yang tidak jauh berbeda antara data *training* dan *testing* menandakan tidak terjadinya *overfitting*. Namun pada nilai MAE, MAPE, dan RMSE metode *Random Forest Regression* menghasilkan nilai yang jauh berbeda antara data *training* dan *testing* sehingga menandakan indikasi terjadinya *overfitting*.

Hasil ini sejalan dengan penelitian yang dilakukan oleh Balogun & Tella (2022) yang melakukan perbandingan antara *Random Forest, Decision Tree Regression, Linear Regression*, dan *Support Vector Regression*, dimana *Random Forest* menghasilkan performa yang paling baik (Balogun & Tella, 2022). Penelitian lain yang membuktikan bahwa *Random Forest* menghasilkan performa yang lebih baik juga dilakukan oleh Ließ dkk. (2012) yang membandingkan performa *Random Forest Regression* dengan *Decision Tree Regression* (Ließ dkk., 2012). Penelitian yang dilakukan oleh Rodriguez-Galiano dkk. (2015) yang membandingkan *Random Forest*, *Regression Tree*, *Artificial Neural Networks*, dan *Support Vector Machine* juga menunjukkan bahwa *Random Forest* memiliki performa yang lebih baik (Rodriguez-Galiano dkk., 2015).

Performa *Random Forest* yang lebih baik dibandingkan dengan *Decision Tree* dikarenakan *Random Forest* memiliki performa generalisasi yang lebih baik untuk performa pelatihan yang serupa dibandingkan dengan *Decision Tree*. Peningkatan generalisasi ini dicapai dengan mengkompensasi kesalahan dalam prediksi *Decision Tree* yang berbeda. *Random Forest* mencari variabel independen terbaik di antara subset acak variabel independen. Hal ini menghasilkan heterogenitas luas yang secara umum meningkatkan kinerja model (Montesinos López dkk., 2022).

PENUTUP

SIMPULAN

Berdasarkan analisis perbandingan peramalan jumlah peserta aktif program JKN menggunakan metode *Decision Tree Regression* dan *Random Forest Regression*, ditemukan bahwa jumlah FKRTL, jumlah penduduk, jumlah dokter, dan jumlah puskesmas secara konsisten memberikan kontribusi besar dalam prediksi pada kedua metode. Selain itu, *Decision Tree Regression* juga mengidentifikasi jumlah klinik pratama dan jumlah penduduk miskin sebagai variabel tambahan yang berpengaruh, mencerminkan perbedaan pendekatan masing-masing metode dalam mengevaluasi kontribusi variabel. Model terbaik *Decision Tree Regression* diperoleh dengan kombinasi parameter *minsplit* sebesar 2 hingga 13 dan *maxdepth* sebesar 3 hingga 20, menghasilkan nilai MAE sebesar 772.464,5, nilai MAPE sebesar 27,80%, dan nilai RMSE sebesar 1.324.906. Sementara itu, model terbaik *Random Forest Regression* menggunakan parameter *ntree* sebesar 500, menghasilkan nilai rata-rata MAE sebesar 518.909,8, nilai rata-rata MAPE sebesar 14,79%, dan nilai rata-rata RMSE sebesar 923.888,9. Hasil perbandingan menunjukkan bahwa *Random Forest Regression* memiliki performa lebih unggul dibandingkan *Decision Tree Regression*, dengan nilai MAE, MAPE, dan RMSE yang lebih kecil. Keunggulan ini disebabkan oleh kemampuan *Random Forest* dalam mengurangi kesalahan model melalui penggabungan prediksi dari banyak pohon, sehingga lebih andal dibandingkan satu *Decision Tree* saja.

Secara keseluruhan, penelitian ini menunjukkan bahwa optimalisasi fasilitas kesehatan berdasarkan faktor-faktor yang memengaruhi jumlah peserta aktif dapat menjadi strategi utama bagi BPJS Kesehatan. Strategi ini tidak hanya diharapkan mampu meningkatkan jumlah peserta aktif program JKN, tetapi juga membantu BPJS Kesehatan menjaga keberlanjutan anggaran untuk pembayaran klaim serta menyediakan cadangan klaim yang cukup guna menghindari risiko tunggakan kepada mitra penyedia layanan kesehatan.

SARAN

Berdasarkan kesimpulan tersebut, terdapat beberapa saran yang dapat dipertimbangkan. Penelitian ini berfokus pada perbandingan antara *Decision Tree Regression* dan *Random Forest Regression*, sehingga untuk penelitian selanjutnya diharapkan dapat melakukan optimasi *hyperparameter* pada model *Random Forest* maupun *Decision Tree* dengan harapan dapat meningkatkan performa model. Selain itu, bagi instansi pemerintah disarankan untuk melakukan pengumpulan data yang lebih luas dan detail, misalnya hingga tingkat regional yang lebih kecil, dapat memberikan gambaran yang lebih spesifik dan membantu model menangkap pola yang lebih baik. Hasil dari model prediksi ini juga dapat

digunakan oleh pemerintah untuk merancang strategi peningkatan kepesertaan program JKN.

DAFTAR PUSTAKA

Al Ishaqi, A. M. (2024, Agustus 12). *BPJS Kesehatan Gaet 276,52 Juta Peserta JKN, Pengamat Keluhakan Jumlah Peserta Non Aktif*. Bisnis.com. https://finansial.bisnis.com/read/20240812/215/1790133/bpjs-kesehatan-gaet-27652-juta-peserta-jkn-pengamat-keluhkan-jumlah-peserta-non-aktif

Alessi, L., & Savona, R. (2021). Machine learning for financial stability. Dalam *Data Science for Economics and Finance: Methodologies and Applications* (hlm. 65–87). Springer International Publishing. https://doi.org/10.1007/978-3-030-66891-4_4

Balogun, A. L., & Tella, A. (2022). Modelling and investigating the impacts of climatic variables on ozone concentration in Malaysia using correlation analysis with random forest, decision tree regression, linear regression, and support vector regression. *Chemosphere*, *299*. https://doi.org/10.1016/j.chemosphere.2022.134250

Cios, K. j., Pedrycz, W., Swiniarski, R. W., & Kurgan, L. A. (2007). *Data Mining A Knowledge Discovery Approach*. Springer.

Edwards, K. J., & Gaber, M. M. (2014). *Astronomy and Big Data* (Vol. 6). Springer. http://www.springer.com/series/11970

Gatera, A., Kuradusenge, M., Bajpai, G., Mikeka, C., & Shrivastava, S. (2023). Comparison of random forest and support vector machine regression models for forecasting road accidents. *Scientific African*, *21*. https://doi.org/10.1016/j.sciaf.2023.e01739

Gorunescu, F. (2011). *Data Mining Concepts, Models and Techniques* (J. Kacprzyk & L. C. Jain, Ed.). Springer.

Hoxha, J., Çodur, M. Y., Mustafaraj, E., Kanj, H., & El Masri, A. (2023). Prediction of transportation energy demand in Türkiye using stacking ensemble models: Methodology and comparative analysis. *Applied Energy*, *350*. https://doi.org/10.1016/j.apenergy.2023.121765

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer. http://www.springer.com/series/417

Kementerian Kesehatan Republik Indonesia. (2016). *Buku Panduan Jaminan Kesehatan Nasional (JKN) Bagi Populasi Kunci*. Kementerian Kesehatan Republik Indonesia.

Ließ, M., Glaser, B., & Huwe, B. (2012). Uncertainty in the spatial prediction of soil texture. Comparison of regression tree and Random Forest models. *Geoderma*, *170*, 70–79. https://doi.org/10.1016/j.geoderma.2011.10.010

Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Random Forest for Genomic Prediction. Dalam *Multivariate Statistical Machine Learning Methods for Genomic Prediction* (hlm. 633–681). Springer International Publishing. https://doi.org/10.1007/978-3-030-89010-0_15

Nopianti, R., Tri Panudju, A., & Permana, A. (2022). *Prediksi Harga Saham Indonesia pada Masa Covid-19 Menggunakan Regresi Pohon Keputusan*. *6*(1). http://ejournal.bsi.ac.id/ejurnal/index.php/ecodemica

Putra, P. H., Azanuddin, Purba, B., & Dalimunthe, Y. A. (2023). Random forest and decision tree algorithms for car price prediction. *Jurnal Matematika Dan Ilmu Pengetahuan Alam LLDikti Wilayah*, *3*(2), 81–89.

Reddy, P. S. M., & Chandar, J. P. (2023). Decision Tree Regressor Compared with Random Forest Regressor for House Price Prediction in Mumbai. Dalam *Journal of Survey in Fisheries Sciences* (Vol. 10, Nomor 1S).

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, *71*, 804–818. https://doi.org/10.1016/j.oregeorev.2015.01.001

RR/SK BPMI Setwapres. (2024, Agustus 8). *Target 98 Persen UHC Tercapai di 2024, Wapres Apresiasi Sinergi BPJS Kesehatan dan Pemangku Kepentingan*. wapresri.go.id. https://www.wapresri.go.id/miliki-potensi-besar-wapres-minta-revitalisasi-jalur-rempah-bangkitkan-kejayaan-perdagangan-rempah-indonesia-2/

Solikhin, M. N. (2019, Maret 1). *Risiko Defisit BPJS bagi Industri Kesehatan*. sindonews.com. https://nasional.sindonews.com/berita/1382974/18/risiko-defisit-bpjs-bagi-industri-kesehatan

Vivas, E., Allende-Cid, H., & Salas, R. (2020). A systematic review of statistical and machine learning methods for electrical power forecasting with reported mape score. Dalam *Entropy* (Vol. 22, Nomor 12, hlm. 1–24). MDPI AG. https://doi.org/10.3390/e22121412

Wei, P., Lu, Z., & Song, J. (2015). Variable importance analysis: A comprehensive review. *Reliability Engineering and System Safety*, *142*, 399–432. https://doi.org/10.1016/j.ress.2015.05.018