Analisis Derajat Kesehatan di Provinsi Jawa Timur Menggunakan Model Regresi Multivariat, *Elastic Net*, dan *Multivariate Random Forest*

MUHAMMAD ADLANSYAH MUDA[1], TABITA YUNI SUSANTO[2], TIZA AYU VIRANIA[3], HASRI WIJI AQSARI[4], DEDY DWI PRASTYO[5], SANTI PUTERI RAHAYU[6]

Departemen Statistika, Fakultas Sains dan Analitika Data, Institut Teknologi Sepuluh Nopember (ITS)
e-mail: [1]adlansyahmuda@gmail.com, [2]tabitayuni56@gmail.com, [3]tiza.ayu99@gmail.com, [4]hasriwiji@gmail.com, [5]dedy.prastyo@gmail.com, [6]sprahayu@gmail.com

ABSTRAK

Angka harapan hidup dan persentase balita secara konsep memiliki hubungan. Sehingga pada penelitian ini kedua variabel tersebut digunakan sebagai variabel respon. Variabel lain yang diduga mempengaruhi variabel respon adalah persentase penduduk miskin, persentase KB pasca persalinan, persentase desa UCI, tingkat positif COVID-19 dan laju pertumbuhan penduduk. Data tersebut selanjutnya dianalisis menggunakan 3 metode yaitu metode regresi multivariat, regresi multivariat menggunakan *Elastic Net* dan *multivariate random forest*. Tiga variabel tersebut dibandingkan dengan menggunakan nilai RMSE dari masing-masing Y untuk mendapatkan model terbaik. Hasil dari perbandingan RMSE tersebut didapatkan bahma metode *Multivariate Random Forest* memiliki RMSE yang terkecil, sehingga metode *Multivariate Random Forest* adalah metode terbaik untuk data penelitian ini. Hasil dari analisis diketahui bahwa variabel persentase penduduk miskin merupakan variabel yang paling penting dalam memodelkan tingkat kesehatan di Jawa Timur menggunakan *Multivariate Random Forest* (MRF) diikuti dengan variabel persentase KB pasca persalinan, tingkat positif COVID-19, persentase desa UCI, dan laju pertumbuhan penduduk.

Kata Kunci: AHH, Elastic Net, Multivariat Random Forest, Regresi Multivariat, RMSE

1. PENDAHULUAN

Angka harapan hidup adalah rata-rata perkiraan banyak tahun yang dapat ditempuh oleh seseorang sejak lahir yang mencerminkan derajat kesehatan suatu masyarakat.Menurut BPS (2021), angka harapan hidup bergantung pada program pembangunan kesehatan, kesehatan lingkungan dan kecukupan gizi. Angka harapan hidup di Indonesia adalah 71.5 tahun dengan 73.3 tahun untuk Pria dan 69.4 tahun untuk Wanita. Karena diduga angka harapan hidup mempengaruhi pada kecukupan gizi, maka pada penelitian ini juga digunakan data persentase balita kurus.

Provinsi Jawa Timur dipilih sebagai objek penelitian dikarenakan nilai angka harapan hidup di Provinsi Jawa Timur tidak begitu jauh dengan angka harapan hidup di Indonesia, sehingga provinsi Jawa Timur dipilih diteliti. Nilai angka harapan hidup di Provinsi Jawa Timur adalah 73.27 untuk pria dan 69.42 untuk wanita.

Selanjutnya diduga angka harapan hidup dan persentase balita kurus dipengaruhi oleh beberapa variabel yaitu persentase penduduk miskin, persentase KB pasca persalinan, persentase desa UCI persentase tingkat positif COVID-19 dan laju pertumbuhan penduduk. Data yang digunakan yaitu berasal dari data profil Kesehatan Provinsi Jawa Timur pada tahun 2020 yang dipublikasikan oleh Dinas Kesehatan Provinsi Jawa Timur pada tahun 2021.

Untuk menguji dan memodelkan variabel-variabel tersebut, digunakan 3 pilihan metode yaitu model regresi multivariat, model regresi multivariat *Elastic Net* dan M*ultivariate Random Forest* (MRF). Ketiga metode tersebut diaplikasikan ke dalam data dan selanjutnya akan dipilih satu metode yang terbaik berdasarkan nilai RMSE yang didapatkan dari masing-masing model.

2. TINJAUAN PUSTAKA

2.1 Analisis Regresi Multivariat

Regresi multivariat adalah model regresi dimana variabel responnya lebih dari satu yang saling berkorelasi serta variabel prediktornya satu atau lebih (Rencher, 2002; Johnson, R. A. dan Wichern, 2007). Data multivariat dapat terdiri atas lebih dari satu variabel. Misalkan terdapat variabel respon berjumlah q yaitu $Y_1, Y_2, \ldots, Y_q$ dan p variabel prediktor yaitu $X_1, X_2, \ldots, X_p$, maka model regresi multivariat untuk pengamatan ke-i respon ke- q adalah:

$$\begin{aligned} y_{1i} &= \beta_{01} + \beta_{11}X_1 + \ldots + \beta_{p1}X_p + \varepsilon_{1i} \\ y_{2i} &= \beta_{02} + \beta_{12}X_1 + \ldots + \beta_{p2}X_p + \varepsilon_{2i} \\ &\vdots \\ y_{qi} &= \beta_{0q} + \beta_{1q}X_1 + \ldots + \beta_{pq}X_p + \varepsilon_{qi} \end{aligned} \quad \ldots\ldots (1)$$

Model regresi multivariat yang terdiri dari q model linier secara simultan dapat ditunjukkan bentuk matriks pada persamaan 2,

$$\boldsymbol{Y}_{(n\times q)} = \boldsymbol{X}_{n\times(p+1)}\boldsymbol{B}_{(p+1)\times q} + \boldsymbol{\varepsilon}_{(n\times q)} \quad \ldots\ldots (2)$$

$$\boldsymbol{Y}_{(n\times q)} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1q} \\ y_{21} & y_{22} & \cdots & y_{2q} \\ \cdots & \cdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nq} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1 & \boldsymbol{y}_2 & \cdots & \boldsymbol{y}_q \end{bmatrix};$$

$$\boldsymbol{X}_{n\times(p+1)} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & \cdots & \cdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}; \quad \boldsymbol{\varepsilon}_{(n\times q)} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1q} \\ \varepsilon_{21} & y_{22} & \cdots & \varepsilon_{2q} \\ \cdots & \cdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{nq} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\varepsilon}_1 & \boldsymbol{\varepsilon}_2 & \cdots & \boldsymbol{\varepsilon}_q \end{bmatrix};$$

$$\boldsymbol{B}_{(p+1)\times q} = \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0q} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1q} \\ \cdots & \cdots & \ddots & \vdots \\ \beta_{p1} & \beta_{p2} & \cdots & \beta_{pq} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 & \cdots & \boldsymbol{\beta}_q \end{bmatrix}$$

dengan $\boldsymbol{\varepsilon} \sim \boldsymbol{N}(\mu, \Sigma)$, $E(\varepsilon_i) = 0$, dan $\mathrm{cov}(\varepsilon_i, \varepsilon_i) = \sigma_{ii}\boldsymbol{I}$.

$\beta_{11}, \beta_{12}, \ldots, \beta_{pq}$ adalah parameter regresi yang nilainya belum diketahui, ε_{qi} adalah residual amatan ke-i untuk variabel respon ke- q dengan $i = 1, 2, \ldots, n$ dan n adalah banyaknya observasi.

Asumsi yang harus dipenuhi dalam pemodelan regresi multivariat adalah variabel respon berdistribusi normal multivariat dan antar variabel respon saling berhubungan. Pengujian distribusi normal multivariat dapat dilakukan dengan uji proporsi dengan hipotesis berikut.

H_0 : Data memenuhi asumsi distribusi normal bivariat.

H_1 : Data tidak memenuhi asumsi distribusi normal bivariat.

Statistik uji: $d_i^2 = (\boldsymbol{y}_i - \bar{\boldsymbol{y}})' \boldsymbol{S}^{-1} (\boldsymbol{y}_i - \bar{\boldsymbol{y}})$; $i = 1, 2, \ldots, n$ ..... (3)

$\boldsymbol{S}$ adalah matriks varians-kovarian dari variabel respon, $\bar{\boldsymbol{y}}$ adalah vektor rata-rata dari variabel respon, d_i^2 adalah square distance. H_0 diterima jika proporsi $d_i^2 \le \chi^2_{2;0,05}$ adalah tepat atau mendekati 50%.

Variabel $Y_1, Y_2, \ldots, Y_q$ dikatakan saling bebas (independen) jika matriks korelasi antar variabel membentuk matriks identitas. Untuk menguji independensi antar variabel ini dapat dilakukan dengan uji *Bartlett's Sphericity* berikut (Morrison, 2005).

H_0 : Antar variabel respon independent $\boldsymbol{\rho} = \mathbf{I}$.

H_1 : Antar variabel respon dependen $\boldsymbol{\rho} \neq \mathbf{I}$.

Statistik Uji:

$$\chi^2_{\text{hitung}} = -\left\{n - 1 - \frac{2q+5}{6}\right\} \ln|R| \quad \text{..... (4)}$$

$|R|$ merupakan determinan dari matrik korelasi. H_0 ditolak jika $\chi^2_{\text{hitung}} \geq \chi^2_{\alpha;\frac{1}{2}q(q-1)}$ yang berarti antar variabel respon bersifat dependen.

Estimasi parameter $\boldsymbol{\beta}$ pada model regresi multivariat dilakukan dengan estimasi kuadrat terkecil (Ordinary Least Square) yaitu meminimumkan jumlah kuadrat eror. Dalam model regresi multivariat pada persamaan (2), matrik parameter regresi berukuran $(p+1)\times q$, dengan estimasi $\hat{\boldsymbol{B}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{Y}\right)$ sedangkan $\boldsymbol{\varepsilon}$ yang merupakan matriks residual ditentunkan oleh estimasi $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \left(\boldsymbol{X}\hat{\boldsymbol{B}}\right)$ (Rencher, 2002).

$$\begin{aligned}
\hat{\boldsymbol{B}} &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{Y}\right) \\
&= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\left(y_1, y_2, ..., y_q\right) \\
&= \left[\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T y_1, \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T y_2, ..., \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T y_q\right] \\
\hat{\boldsymbol{B}} &= \left[\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, ..., \hat{\boldsymbol{\beta}}_q\right] \quad \text{..... (5)}
\end{aligned}$$

2.2 *Elastic Net*

Beberapa metode regularisasi yang digunakan untuk model fitting dan seleksi variabel antara lain yaitu LASSO, Ridge, dan *Elastic Net*. Regularisasi Ridge menginduksi penyusutan prediktor, dan dengan demikian membuat estimasi parameter lebih stabil, sedangkan regularisasi LASSO menyebabkan banyak koefisien regresi menjadi tepat nol dan karenanya memfasilitasi pemilihan variabel otomatis di mana hanya satu prediktor yang dipilih di antara prediktor yang berkorelasi (Cho *et al.*, 2010). *Elastic Net* merupakan salah satu metode regularisasi yang secara linier menggabungkan penalti ℓ_1-norm dan ℓ_2-norm dari metode LASSO (Least Absolute Shrinkage and Selection Operator) dan Ridge sehingga memberikan penyusutan (shrinkage) dan pemilihan variabel secara otomatis.

Menurut Zou dan Hastie, 2005 *Elastic Net* bekerja lebih baik daripada LASSO dalam representasi sparsity dan memperbaiki kekurangan LASSO. LASSO memiliki beberapa kekurangan yaitu sebagai berikut (Shen, Liu dan Wu, 2020).

(1) Dalam kasus p > n, di mana p adalah jumlah prediktor dan n adalah jumlah sampel, LASSO memilih paling banyak n variabel karena karakteristik dari masalah optimasi convex.

(2) Jika korelasi berpasangan sangat tinggi di antara sekelompok variabel, LASSO cenderung memilih hanya satu variabel, mengabaikan variabel lain, tidak peduli mana yang harus dipilih.

(3) Dalam kasus n > p, jika korelasi tinggi terjadi antara prediktor, telah diamati bahwa kinerja prediksi LASSO lebih buruk daripada regresi ridge.

Elastic Net menyelesaikan permasalahan berikut,

$$\min_{\boldsymbol{B}\in\mathbb{R}^{(p+1)\times q}}\left[\frac{1}{2}\left\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\right\|_F^2 + \lambda P_\alpha\left(\boldsymbol{B}\right)\right] \quad \text{..... (5)}$$

$$P_\alpha\left(\boldsymbol{B}\right) = \alpha\sum_{j=1}^{p}\left\|\beta_{j.}\right\|_2 + \frac{(1-\alpha)}{2}\left\|\boldsymbol{B}\right\|_F^2$$

dengan $P_\alpha\left(\boldsymbol{B}\right)$ adalah penalti *Elastic Net*, sehingga persamaan (5) menjadi persamaan (6).

$$\min_{\boldsymbol{B}\in\mathbb{R}^{(p+1)\times q}}\left[\frac{1}{2}\left\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}\right\|_F^2+\lambda\alpha\sum_{j=1}^{p}\left\|\beta_{j.}\right\|_2+\frac{\lambda(1-\alpha)}{2}\left\|\boldsymbol{B}\right\|_F^2\right] \quad \text{..... (6)}$$

$\beta_{j.}$ adalah baris ke-j dari matriks $\boldsymbol{B}$, $\lambda=\lambda_1+\lambda_2$, $\alpha=\dfrac{\lambda_1}{\lambda_1+\lambda_2}$ adalah parameter gabungan antara ridge dan LASSO. Jika $\alpha=0$, maka penalti penyusutan menjadi penalti penyusutan Ridge, untuk $\alpha=1$, maka penalti penyusutan menjadi penalti penyusutan LASSO. Persamaan 6 dapat diminimumkan dengan penurunan koordinat blockwise (blockwise coordinate descent) yaitu satu baris dari $\boldsymbol{B}$ pada satu waktu. Solusi penurunan koordinat blockwise ditunjukkan oleh persamaan 7.

$$\left\|X_{.j}\right\|_2^2\hat{\beta}_{j.}^T-X_{.j}^TR_{-j}+\lambda\alpha S\left(\hat{\beta}_{j.}\right)+\lambda(1-\alpha)\hat{\beta}_{j.}^T=0 \quad \text{..... (7)}$$

$R_{-j}=Y-\sum_{j\neq p}X_{.j}\beta_{j.}$ adalah residual parsial, $S\left(\hat{\beta}_{j.}\right)$ adalah sub differensial dengan,

$$S(\alpha)=\begin{cases}\dfrac{\alpha}{\|\alpha\|2} & ;\ \alpha\neq 0\\ \in\left\{u\ s.t.\ \|u\|_2\leq 1\right\} & ;\ \alpha=0\end{cases}.$$

Dari penyelesaian persamaan (7) didapat estimasi parameter $\hat{\beta}_{j.}$ sebagai berikut.

$$\hat{\beta}_{j.}=\frac{1}{\left\|X_{.j}\right\|_2^2+\lambda(1-\alpha)}\left(1-\frac{\lambda\alpha}{\left\|X_{.j}^TR_{-j}\right\|_2}\right)X_{.j}^TR_{-j} \quad \text{..... (8)}$$

Nilai parameter λ dapat bervariasi dan menyebabkan perbedaan penyusutan koefisien untuk nilai λ yang berbeda. Nilai λ yang semakin besar mengakibatkan jumlah koefisien yang menyusut menjadi nol meningkat, sehingga perlu dicari nilai λ yang optimum. Salah satu cara penentuan nilai λ optimum yaitu dengan menggunakan validasi silang (cross validation). Validasi silang merupakan metode untuk menduga kesalahan prediksi. Pada validasi silang, data dibagi menjadi data pemodelan dan data validasi secara acak dan setiap pengamatan memiliki peluang yang sama untuk menjadi data validasi. Salah satu jenis validasi silang adalah cross validation k-fold. Metode ini baik digunakan ketika jumlah observasi sedikit. Dalam cross validation k-fold, observasi dibagi ke dalam k gugus data secara acak dengan ukuran yang hampir sama. Dimana salah satu partisi sebagai data validasi dan k-1 partisi digunakan sebagai pembentuk model. *Cross Validation* dilakukan berulang sampai k kali, dengan masing-masing k partisi data digunakan satu kali sebagai validasi model. *Cross Validation* ini menghasilkan k MSE (Mean of Squares Error), $MSE_1, MSE_2, ..., MSE_k$. CVE (Cross Validation Error) k-fold diperoleh dengan cara berikut.

$$CVE(k)=\frac{1}{k}\sum_{i=1}^{k}MSE_i \quad \text{..... (9)}$$

Regresi *Elastic Net* Multivariat dapat dijalankan menggunakan packages R-joinet (Rauschenberger, 2021). Menurut Rauschenberger dan Glaab, 2021 joinet lebih fleksibel karena dapat memodelkan regresi multivariat dengan keluarga distribusi variabel respon yang berbeda.

2.3 *Multivariate random forest*

Kerangka pohon regresi, yang dikembangkan oleh (Breiman *et al.*, 1984) melibatkan empat komponen: (1) Satu set pertanyaan biner (ya/tidak), atau pemisahan, yang berfungsi untuk mempartisi ruang prediktor. Subsampel yang dibuat dengan menetapkan kasus menurut pemisahan ini disebut node (simpul). Sebuah node yang tidak memiliki keturunan node adalah node terminal atau leaf. (2) Sebuah ukuran node impurity, biasanya berkaitan dengan varians respon dalam konteks regresi. (3) Fungsi split, $\phi(s,t)$, yang dapat dievaluasi untuk setiap split s yang diperbolehkan, dari setiap node t. Pemisahan terbaik, yang mengoptimalkan ϕ, adalah

sedemikian rupa sehingga distribusi respons di node anak yang dihasilkan paling homogen di antara semua pemisahan yang bersaing, dengan homogenitas dinilai melalui ukuran impurity. (4) Sarana untuk menentukan ukuran pohon yang sesuai.

Pada variabel respon tunggal (univariat), misalkan y_i dan x_{ij}, $i = 1,...,n; j = 1,...,p$ menunjukkan variabel respon dan prediktor. Diberikan node t yang berisi sub-sampel kasus. Selanjutnya mempartisi t menjadi dua node anak, node 'kiri' t_L, dan node 'kanan' t_R. Misalnya j menjadi indeks dari prediktor kontinu atau kategori (ordinal). Kemudian pemisahan yang diizinkan adalah pemotongan biner yang mempertahankan urutan dalam bentuk $t_L = i \in t : x_{ij} \le c,\ t_R = i \in t : x_{ij} > c$ karena titik potong c berkisar pada semua nilai yang mungkin menghasilkan perbedaan t_L, t_R .s Untuk prediktor kategoris yang tidak berurutan, semua pembagian menjadi subset kategori yang terpisah diperbolehkan. Ukuran impurity node L_2 hanyalah jumlah kuadrat $SS(t) = \sum_{i \in t} \left(y_i - \mu(t)\right)^2$ di mana $\mu(t)$ adalah rata-rata dari y_i di node t. Maka fungsi split yang sesuai adalah sebagai berikut (Segal dan Xiao, 2011).

$$\phi(s,t) = SS(t) - SS(t_L) - SS(t_R) \quad \text{..... (10)}$$

Diberikan data multirespon (multivariat) $y_{iq} (i = 1,...,n;\ q = 1,...,q)$. Untuk kesederhanaan, diasumsi bahwa setiap individu memiliki jumlah respon yang sama (q) dan bahwa prediktornya adalah variabel 'dasar'; yaitu, tidak bervariasi dengan q. Semua yang diperlukan untuk memperluas pohon regresi ke beberapa respons adalah modifikasi dari fungsi split. Formulasi alami adalah mengganti ukuran pengotor node dengan analog berbobot (weighted analog) 'kovarians' sebagai berikut.

$$SS(t) = \sum_{i \in t} \left(y_i - \mu(t)\right)' V^{-1}(t,\eta) \left(y_i - \mu(t)\right) \quad \text{..... (11)}$$

Di sini η mewakili parameter yang mencirikan struktur kovarians yang ditentukan. Menggunakan persamaan 10 fungsi split untuk multirespon dibuat sesuai persamaan 9. Prediksi untuk setiap leaf dari pohon regresi multirespon (multivariat) hanyalah cara vektor variabel respon untuk kasus yang mencapai leaf (Segal dan Xiao, 2011).

Random forest adalah kumpulan prediktor pohon $h(\mathbf{x};\boldsymbol{\theta}_k), k = 1,...,K$ di mana $\mathbf{x}$ mewakili vektor input (prediktor) yang diamati dengan panjang p dengan vektor random terkait $\boldsymbol{X}$ dan $\boldsymbol{\theta}_k$ adalah vektor acak independen dan terdistribusi identik (L., 2001). Data yang diamati diasumsikan diambil secara independen dari distribusi gabungan $(\boldsymbol{X},\boldsymbol{Y})$ dan terdiri dari $n(p+1)$ tupel $(\mathbf{x}_1, y_1),...,(\mathbf{x}_n, y_n)$. Prediksi random forest adalah rata-rata tak tertimbang dari $h(\mathbf{x}) = (1/K)\sum_{k=1}^{K} h(\mathbf{x};\boldsymbol{\theta}_k)$.

3. METODOLOGI PENELITIAN

3.1 Sumber Data dan Variabel Penelitian

Data yang digunakan pada penelitian ini adalah data profil Kesehatan Provinsi Jawa Timur Tahun 2020 yang dipublikasikan oleh Dinas Kesehatan Provinsi Jawa Timur pada tahun 2021 (Dinas Kesehatan Provinsi Jawa Timur, 2021). Unit penelitian yang digunakan adalah kabupaten/kota di Jawa Timur yang terdiri dari 29 Kabupaten dan 9 Kota.

Variabel respon yang digunakan pada penelitian ini adalah Angka Harapan Hidup (AHH) sebagai Y_1 dan persentase balita kurus sebagai Y_2. Sedangkan untuk variabel prediktor yang digunakan adalah persentase penduduk miskin (X_1), persentase KB pasca persalinan (X_2), persentase desa UCI (X_3), tingkat positif COVID-19 (X_4)dan laju pertumbuhan penduduk (X_5). Adapun penjelasan dari masing-masing variabel adalah sebagai berikut.

1. Angka Harapan Hidup (Y_1) adalah didefinisikan sebagai rata-rata perkiraan banyak tahun yang dapat ditempuh oleh seseorang sejak lahir yang mencerminkan derajat kesehatan suatu masyarakat.
2. Persentase balita kurus (Y_2) merupakan persentase jumlah balita dengan status gizi yang didasarkan pada berat badan dan juga tinggi badan balita.
3. Persentase penduduk miskin (X_1) adalah persentase penduduk yang berada di bawah Garis Kemiskinan (GK).
4. Persentase KB pasca persalinan (X_2) merupakan persentase wanita yang melakukan KB pasca melahirkan.
5. Persentase desa UCI (X_3) merupakan persentase desa/kelurahan dimana ≥ 80% dari jumlah bayi yang ada di desa/kelurahan tersebut sudah mendapatkan imunisasi dasar lengkap dalam waktu satu tahun.
6. Tingkat positif COVID-19 (X_4) merupakan hasil perhitungan dari jumlah orang dengan hasil pemeriksaan positif dibagi dengan jumlah orang yang diperiksa pada suatu rentang waktu yang sama.
7. laju pertumbuhan penduduk (X_5) merupakan angka yang menunjukkan rata-rata tingkat pertambahan penduduk per tahun dalam jangka waktu tertentu. Angka ini dinyatakan sebagai persentase dari penduduk dasar.

3.2 Langkah Analisis

Langkah-langkah analisis yang dilakukan adalah:
1. Mengumpulkan data.
2. Melakukan eksplorasi data mengenai statistika deskriptif masing-masing variabel.
3. Melakukan pengujian variabel respon berditribusi normal multivariat.
4. Melakukan pengujian dependensi antar variabel respon dengan menggunakan uji Bartlett Test.
5. Melakukan deteksi multikolinieritas. Jika terdapat multikolinieritas maka diatasi dengan transformasi variabel atau dengan menghilangkan satu variabel.
6. Melakukan analisis regresi multivariat dengan langkah sebagai berikut.
   a. Melakukan estimasi parameter model regresi multivariat
   b. Melakukan pengujian signifikansi parameter model regresi multivariat secara serentak dan signifikan parsial secara multivariat dan univariat.
   c. Melakukan pengujian asumsi residual berdistribusi normal multivariat dan asumsi IIDN secara univariat.
   d. Menghitung nilai ukuran kebaikan model.
   e. Mendapatkan model persamaan faktor-faktor yang mempengaruhi derajat kesehatan Kabupaten/Kota di Provinsi Jawa Timur.
7. Melakukan analisis regresi multivariat dengan *Elastic Net* dengan langkah sebagai berikut.
   a. Menentukan rentang nilai α yaitu $0 \leq \alpha \leq 1$ dengan validasi silang.
   b. Memilih nilai pasangan nilai α dan λ optimal yang menghasilkan CVE minimum.
   c. Melakukan estimasi parameter model regresi multivariat *Elastic Net* dengan pasangan α dan λ optimal.
   d. Menghitung kebaikan model.
   e. Mendapatkan model persamaan faktor-faktor yang mempengaruhi derajat kesehatan Kabupaten/Kota di Provinsi Jawa Timur.
8. Melakukan pemodelan Multivariat Random Forest dengan langkah sebagai berikut.
   a. Menentukan jumlah pohon, jumlah daun dan maksimal nilai featurenya.
   b. Membagi data menjadi 5 fold.
   c. Melakukan pemodelan dengan menggunakan data training.
   d. Melakukan prediksi dengan menggunakan data testing.
   e. Menghitung kebaikan model dengan rata-rata RMSE dari ke-5 fold.
9. Membandingkan kebaikan model antara tiga metode tersebut.
10. Menarik kesimpulan.

4. HASIL DAN PEMBAHASAN

Pada bagian ini akan dilakukan statistika deskriptif untuk mengetahui karakteristik variabel-variabel yang digunakan dalam memodelkan tingkat kesehatan di Provinsi Jawa Timur tahun 2020, kemudian dilakukan pengujian dependensi dengan uji *Bartlett's Sphericity* dan pengujian asumsi distribusi normal multivariat, serta melakukan pemodelan tingkat kesehatan di Provinsi Jawa Timur dengan menggunakan regresi multivariat, regresi multivariat *Elastic Net*, dan *Multivariat Random Forest* (MRF). Terakhir dilakukan perbandingan performansi di antara ketiga model untuk mengetahui model terbaik.

4.1 Statistika Deskriptif

Hasil statistika deskriptif dari variabel-variabel yang digunakan dalam penelitian ini ditunjukkan pada Tabel 1 berikut ini:

Tabel 1. Statistika Deskriptif

Variabel	*Mean*	Variansi	Minimum	Maksimum
Y_1	71,64	3,97	66,74	74,18
Y_2	7,65	7,81	3,00	13,80
X_1	11,02	20,87	3,89	22,78
X_2	50,72	378,66	12,80	98,30
X_3	84,08	226,04	42,30	100,00
X_4	44,39	376,49	14,40	93,00
X_5	0,84	0,07	0,27	1,53

Tabel 1 menunjukkan bahwa rata-rata Angka Harapan Hidup (AHH) di Provinsi Jawa Timur tahun 2020 ialah sebesar 71,64 tahun dimana nilai tertinggi terdapat di Kota Surabaya dan Angka Harapan Hidup (AHH) terendah terdapat di Kabupaten Bondowoso. Kemudian rata-rata persentase balita kurus di Provinsi Jawa Timur tahun 2020 ialah sebesar 7,65% dimana nilai tertinggi terdapat di Kabupaten Pasuruan dan persentase balita kurus terendah terdapat di Kabupaten Banyuwangi. Selanjutnya rata-rata persentase penduduk miskin di Provinsi Jawa Timur tahun 2020 ialah sebesar 11,02% dengan persentase penduduk miskin tertinggi terdapat di Kabupaten Sampang dan persentase penduduk miskin terendah terdapat di Kota Batu. Kemudian rata-rata persentase KB pasca persalinan di Provinsi Jawa Timur tahun 2020 ialah sebesar 50,72% dengan persentase terendah terjadi di Kabupaten Pacitan dan persentase tertinggi terjadi di Kabupaten Situbondo. Selanjutnya rata-rata persentase desa UCI di Provinsi Jawa Timur tahun 2020 ialah sebesar 84,08 dimana persentase desa UCI tertinggi terjadi di Kabupaten Ngawi dan terendah terjadi di Kabupaten Bangkalan. Kemudian rata-rata tingkat positif COVID-19 di Provinsi Jawa Timur tahun 2020 ialah sebesar 44,39 dengan tingkat positif COVID-19 tertinggi terjadi di Kabupaten Lumajang dan terendah terjadi di Kabupaten Bangkalan. Selanjutnya rata-rata laju pertumbuhan penduduk di Provinsi Jawa Timur tahun 2020 ialah sebesar 0,84% dimana laju pertumbuhan penduduk tertinggi terdapat di Kabupaten Bangkalan dan terendah terdapat di Kota Malang. Nilai variansi untuk seluruh variabel yang digunakan kecuali variabel laju pertumbuhan penduduk memiliki nilai yang cukup besar, sehingga terdapat kecenderungan nilai yang heterogen. Selanjutnya dilakukan perhitungan nilai korelasi antar tiap variabel untuk mengetahui hubungan antar tiap variabel yang ditampilkan pada Tabel 2.

Tabel 2 menunjukkan bahwa nilai korelasi antara kedua variabel respon bernilai -0,404. Dikarenakan bernilai negatif maka dapat diartikan bahwa variabel Angka Harapan Hidup (AHH) dan persentase balita kurus memiliki hubungan yang berbanding terbalik, artinya semakin tinggi Angka Harapan Hidup (AHH) maka akan menurunkan persentase balita kurus di Provinsi Jawa Timur begitu juga sebaliknya. Berdasarkan Tabel 2 dapat diketahui bahwa korelasi antar variabel prediktor cenderung tidak terlalu tinggi, sehingga tidak terdapat kecenderungan terjadi kasus multikolinearitas.

Tabel 2. Nilai Korelasi

	Y_1	Y_2	X_1	X_2	X_3	X_4	X_5
Y_1	1	-0,404	-0,577	-0,616	0,448	-0,529	0,128
Y_2	-0,404	1	0,067	0,358	-0,073	0,146	-0,111
X_1	-0,577	0,067	1	0,270	-0,325	0,064	0,038
X_2	-0,616	0,358	0,270	1	-0,065	0,501	-0,314
X_3	0,448	-0,073	-0,325	-0,065	1	-0,030	-0,176
X_4	-0,529	0,146	0,064	0,501	-0,030	1	-0,149
X_5	0,128	-0,111	0,038	-0,314	-0,176	-0,149	1

4.2 UJI DEPENDENSI

Salah satu asumsi yang harus dipenuhi dalam melakukan analisis multivariat adalah kedua variabel respon yang diteliti harus saling dependen. Untuk mengetahui dependensi antar kedua variabel respon dilakukan dengan menggunakan uji *Bartlett's Sphericity* yang hasilnya ditampilkan pada Tabel 3 berikut:

Tabel 3. Uji *Bartlett's Sphericity*

χ^2_{hitung}	$\chi^2_{(1;0,05)}$	**Keputusan**
6,323	3,841	Tolak H_0

Berdasarkan Tabel 3 dapat dilihat bahwa keputusan tolak H_0, sehingga dapat disimpulkan bahwa terdapat hubungan antara variabel Angka Harapan Hidup (AHH) dan persentase balita kurus.

4.3 UJI MULTIKOLINEARITAS

Pemeriksaan untuk mendeteksi adanya multikolinieritas dapat dilakukan dengan melihat nilai VIF pada masing-masing variabel, dikatakan terjadi multikolinieritas apabila nilai VIF lebih dari 10, dan sebaliknya apabila nilai VIF kurang dari 10 maka tidak terjadi multikolinieritas. Berikut ini adalah hasil analisis untuk melihat adanya multikolinieritas.

Tabel 4. Uji *Multikolinearitas*

Variabel	**VIF**
X_1	1.165501
X_2	1.154468
X_3	1.346801
X_4	1.594642
X_5	1.218621

Berdasarkan Tabel 4 dapat dilihat bahwa nilai VIF pada variabel prediktor yang digunakan yaitu persentase penduduk miskin, persentase KB pasca persalinan, persentase desa UCI, tingkat positif COVID-19 dan laju pertumbuhan tidak terdapat multikolinearitas. Hal ini dapat dilihat dari nilai VIF yang bernilai < 10.

4.4 Uji Asumsi Distribusi Normal Multivariat

Asumsi selanjutnya yang harus dipenuhi dalam melakukan analisis multivariat adalah harus berdistribusi normal multivariat. Salah satu metode untuk melakukan pengujian asumsi distribusi normal multivariat adalah dengan menggunakan uji proporsi d_j^2 dengan hasil yang ditampilkan pada Tabel 5 berikut:

Tabel 5. Uji Asumsi Distribusi Normal Multivariat

$\chi^2_{(2;0,5)}$	**Proporsi** $d_j^2 < \chi^2_{(2;0,5)}$
1,386	0,526

Dikatakan mengikuti distribusi normal multivariat apabila proporsi d_j^2 yang kurang dan lebih dari 1,386 ialah tepat atau mendekati 50%. Tabel 5 menunjukkan bahwa proporsi d_j^2 < 1,386 ialah sebesar 52,6%, artinya dapat disimpulkan bahwa asumsi distribusi normal multivariat telah terpenuhi.

4.5 Model Regresi Multivariat

Setelah asumsi dependensi dan asumsi distribusi normal multivariat terpenuhi, maka selanjutnya dapat dilakukan pemodelan dengan menggunakan regresi multivariat. Hasil estimasi parameter dan pengujian signifikansi secara univariat untuk model regresi multivariat ditampilkan pada Tabel 6 berikut:

Tabel 6. Estimasi Parameter Model Regresi Multivariat

Respon	Prediktor	*Estimate*	*Std. Error*	*t*	*P-value*
Y_1	*Intercept*	72,880	1,704	42,770	0,000
	X_1	-0,163	0,044	-3,667	0,001
	X_2	-0,032	0,012	-2,694	0,011
	X_3	0,040	0,013	3,061	0,004
	X_4	-0,034	0,011	-3,085	0,004
	X_5	0,366	0,759	0,483	0,633
Y_2	*Intercept*	6,633	4,253	1,559	0,129
	X_1	-0,035	0,111	-0,318	0,753
	X_2	0,056	0,030	1,890	0,068
	X_3	-0,013	0,033	-0,391	0,698
	X_4	-0,007	0,027	-0,262	0,795
	X_5	-0,059	1,895	-0,031	0,976

Berdasarkan Tabel 6 dapat diketahui bahwa variabel laju pertumbuhan penduduk tidak berpengaruh signifikan secara univariat untuk kedua model, sehingga variabel tersebut dikeluarkan dari model dan hasil estimasi parameter model tanpa variabel laju pertumbuhan penduduk ditampilkan pada Tabel 7 berikut:

Tabel 7. Estimasi Parameter Model Regresi Multivariat Tanpa Variabel Laju Pertumbuhan Penduduk

Respon	Prediktor	*Estimate*	*Std. Error*	*t*	*P-value*
Y_1	*Intercept*	73,349	1,384	53,007	0,000
	X_1	-0,161	0,044	-3,684	0,001
	X_2	-0,034	0,011	-3,017	0,005
	X_3	0,039	0,013	3,059	0,004
	X_4	-0,034	0,011	-3,113	0,004
Y_2	*Intercept*	6,558	3,441	1,905	0,065
	X_1	-0,036	0,109	-0,326	0,746
	X_2	0,057	0,028	2,025	0,051
	X_3	-0,013	0,032	-0,398	0,693
	X_4	-0.007	0.027	-0.267	0.791

Tabel 7 menunjukkan bahwa secara univariat semua variabel prediktor berpengaruh signifikan pada model Angka Harapan Hidup (AHH), sedangkan pada model persentase balita kurus hanya variabel persentase KB pasca persalinan yang berpengaruh signifikan pada model. Selanjutnya dilakukan pengujian signifikansi parameter secara multivariat dengan menggunakan uji *Wilks' Lambda* yang hasilnya ditunjukkan pada Tabel 8 berikut:

Tabel 8. Uji *Wilks' Lambda*

Prediktor	*Wilks' Lambda*	*P-value*
Intercept	0,010	0,000
X_1	0,657	0,001
X_2	0,768	0.015
X_3	0,766	0.014
X_4	0,729	0.006

Berdasarkan Tabel 8 dapat dilihat bahwa secara multivariat seluruh variabel persentase penduduk miskin, persentase KB pasca persalinan, persentase desa UCI, dan tingkat positif COVID-19 berpengaruh signifikan dalam model. Sehingga model regresi multivariat untuk tingkat kesehatan di Provinsi Jawa Timur dapat dituliskan sebagai berikut:

$$\text{AHH} = 73,349 - 0,161X_1 - 0,034X_2 + 0,039X_3 - 0,034X_4$$

$$\text{BBK} = 6,558 - 0,036X_1 + 0,057X_2 - 0,013X_3 - 0,007X_4$$

Berdasarkan model tersebut dapat diketahui bahwa nilai konstanta pada tiap persamaan ialah 73,349 dan 6,558, artinya jika seluruh variabel konstan maka angka Angka Harapan Hidup (AHH) sebesar 73,349 tahun dan persentase balita kurus sebesar 6,558%. Koefisien variabel persentase penduduk miskin ialah -0,161 dan -0,036, artinya jika variabel prediktor lainnya konstan, maka setiap kenaikan 1% persentase penduduk miskin akan menurunkan Angka Harapan Hidup (AHH) sebesar 0,161 tahun dan persentase balita kurus sebesar 0,036%. Kemudian koefisien variabel persentase KB pasca persalinan ialah -0,034 dan 0,057, artinya jika variabel prediktor lainnya konstan, maka setiap kenaikan 1% persentase KB pasca persalinan akan menurunkan Angka Harapan Hidup (AHH) sebesar 0,034 tahun dan meningkatkan persentase balita kurus sebesar 0,057%. Selanjutnya koefisien variabel persentase desa UCI ialah 0,039 dan -0,013, artinya jika variabel prediktor lainnya konstan, maka setiap kenaikan 1% persentase desa UCI akan meningkatkan Angka Harapan Hidup (AHH) sebesar 0,039 tahun dan menurunkan persentase balita kurus sebesar 0,013%. Kemudian koefisien variabel tingkat positif COVID-19 ialah -0,034 dan -0,007, artinya jika variabel prediktor lainnya konstan, maka setiap kenaikan 1% tingkat positif COVID-19 akan menurunkan Angka Harapan Hidup (AHH) sebesar 0,034 tahun dan menurunkan persentase balita kurus sebesar 0,007%.

4.6 Model Regresi Multivariat *Elastic Net*

Elastic Net merupakan salah metode seleksi variabel dimana kelompok prediktor yang berkolerasi akan dipilih bersama-sama untuk dieliminasi. Pada penelitian ini nilai λ dan α yang dipilih merupakan nilai yang optimum. Hasil estimasi parameter untuk model regresi multivariat *Elastic Net* ditampilkan pada Tabel 9 berikut:

Tabel 9. Estimasi Parameter Model Regresi Multivariat *Elastic Net*

Respon	λ	α	Prediktor	*Estimate*
Y_1	0,084	1	*Intercept*	73.363
			X_1	-0.128
			X_2	-0.032
			X_3	0.029
			X_4	-0.026
			X_5	0.000
Y_2	0,012	1	*Intercept*	6.228
			X_1	0.034
			X_2	0.027
			X_3	-0.008
			X_4	0.007
			X_5	0.000

Elastic Net mengeliminasi variabel dengan cara menekan nilai koefisien menjadi 0. Pada Tabel 9 dapat dilihat bahwa variabel laju pertumbuhan penduduk dikeluarkan dari model. Sehingga model regresi multivariat *Elastic Net* dapat dituliskan sebagai berikut:

$$\text{AHH} = 73,363 - 0,128X_1 - 0,032X_2 + 0,029X_3 - 0,026X_4$$

$$\text{BBK} = 6,228 + 0,034X_1 + 0,027X_2 - 0,008X_3 + 0,007X_4$$

Model tersebut menjelaskan bahwa nilai konstanta pada tiap persamaan ialah 73,363 dan 6,228, artinya jika seluruh variabel konstan maka angka Angka Harapan Hidup (AHH) sebesar 73,363 tahun dan persentase balita kurus sebesar 6,2285. Koefisien variabel persentase penduduk miskin ialah -0,128 dan 0,034, artinya jika variabel prediktor lainnya konstan, maka setiap kenaikan 1% persentase penduduk miskin akan menurunkan Angka Harapan Hidup (AHH) sebesar 0,128 tahun dan meningkatkan persentase balita kurus sebesar 0,034%. Kemudian koefisien variabel persentase KB pasca persalinan ialah -0,032 dan 0,027, artinya jika variabel prediktor lainnya konstan, maka setiap kenaikan 1% persentase KB pasca persalinan akan menurunkan Angka Harapan Hidup (AHH) sebesar 0,032 tahun dan meningkatkan persentase balita kurus sebesar 0,027%. Selanjutnya koefisien variabel persentase desa UCI ialah 0,029 dan -0,008, artinya jika variabel prediktor lainnya konstan, maka setiap kenaikan 1% persentase desa UCI akan meningkatkan Angka Harapan Hidup (AHH) sebesar 0,029 tahun dan menurunkan persentase balita kurus sebesar 0,008%. Kemudian koefisien variabel tingkat positif COVID-19 ialah -0,026 dan 0,007, artinya jika variabel prediktor lainnya konstan, maka setiap kenaikan 1% tingkat positif COVID-19 akan menurunkan Angka Harapan Hidup (AHH) sebesar 0,026 tahun dan meningkatkan persentase balita kurus sebesar 0,007%.

4.6 Model *Multivariate Random Forest* (MRF)

Model *Multivariate Random Forest* (MRF) merupakan model yang terbentuk dari beberapa pohon (*tree*) dimana pada penelitian ini menggunakan 100 pohon. Pemodelan dengan menggunakan metode MRF ini mampu menghasilkan informasi mengenai variabel-variabel penting yang berkontribusi dalam model. Nilai frekuensi pada tiap variabel dalam menunjukkan tingkat kepentingan variabel dalam model ditunjukkan pada Tabel 10 berikut:

Tabel 10. Tingkat Kepentingan Variabel

Prediktor	Frekuensi	Peringkat
X_1	366	1
X_2	275	2
X_3	145	4
X_4	201	3
X_5	143	5

Tabel 10 menunjukkan bahwa berdasarkan nilai frekuensi, variabel persentase penduduk miskin merupakan variabel yang paling penting dalam memodelkan tingkat kesehatan di Jawa Timur menggunakan *Multivariate Random Forest* (MRF) diikuti dengan variabel persentase KB pasca persalinan, tingkat positif COVID-19, persentase desa UCI, dan terakhir adalah variabel laju pertumbuhan penduduk.

4.7 Memilih Model Terbaik

Setelah mendapatkan model regresi multivariat, regresi multivariat *Elastic Net*, dan *Multivariate Random Forest* (MRF), maka selanjutnya dilakukan pemilihan model terbaik dalam memodelkan tingkat kesehatan di Provinsi Jawa Timur. Salah satu penilaian yang dapat digunakan untuk memilih model terbaik ialah dengan menggunakan RMSE. Hasil nilai RMSE pada tiap model yang digunakan pada penelitian ini ditampilkan pada Tabel 11 berikut:

Tabel 11. Perbandingan Model

Model	Respon	RMSE	Peringkat
Regresi Multivariat	Y_1	1.031	2
	Y_2	2.564	
Regresi Multivariat *Elastic Net*	Y_1	1.075	3
	Y_2	2.612	
Multivariate Random Forest (MRF)	Y_1	0.569	1
	Y_2	0.891	

Berdasarkan nilai RMSE yang ditampilkan pada Tabel 11, dapat disimpulkan bahwa model *Multivariate Random Forest* (MRF) merupakan model terbaik dalam memodelkan tingkat kesehatan di Jawa Timur, diikuti dengan model regresi multivariat, dan model regresi multivariat *Elastic Net*.

5. KESIMPULAN

Model *Multivariate Random Forest* (MRF) merupakan model terbaik dalam memodelkan tingkat kesehatan di Provinsi Jawa Timur berdasarkan nilai RMSE dibandingkan dengan model regresi multivariat dan model regresi multivariat *Elastic Net*. Variabel persentase penduduk miskin merupakan variabel yang paling penting dalam memodelkan tingkat kesehatan di Jawa Timur menggunakan *Multivariate Random Forest* (MRF) diikuti dengan variabel persentase KB pasca persalinan, tingkat positif COVID-19, persentase desa UCI, dan laju pertumbuhan penduduk.

DAFTAR PUSTAKA

Breiman, L. *et al.* (1984) *Classification and regression trees*. CRC press.

Badan Pusat Statistik, 2021. *Angka Harapan Hidup (AHH) Menurut Provinsi dan Jenis Kelamin.* [Online]
Available at: https://www.bps.go.id/indicator/40/501/1/angka-harapan-hidup-ahh-menurut-provinsi-dan-jenis-kelamin.html
[Accessed 12 Desember 2021].

Cho, S. *et al.* (2010) 'Joint identification of multiple genetic variants via *Elastic Net* variable selection in a genome-wide association analysis', *Annals of human genetics*, 74(5), pp. 416–428.

Dinas Kesehatan Provinsi Jawa Timur, 2021. *Profil kesehatan 2020.* Surabaya: Dinas Kesehatan Provinsi Jawa Timur.

Johnson, R. A. dan Wichern, D. W. (2007) *Applied Multivariate Statistical Analysis*. 6th edn. New Jersey: Prentice Hal.

L., B. (2001) 'Random Forests. Machine Learning, 45', pp. 5–32.

Morrison, D. F. (2005) *Multivariat Statistical Methods*. Fourth. The Wharton School University of Pennsylvania.

Rauschenberger, A. (2021) 'Package "joinet" (Multivariate *Elastic Net* Regression)'. CRAN.

Rauschenberger, A. and Glaab, E. (2021) 'Predicting correlated outcomes from molecular data', *Bioinformatics*, 37(21), pp. 3889–3895.

Rencher, A. R. (2002) *Methods of Multivariate Analysis*. Second. New York: John Wiley and Sons Inc.

Segal, M. and Xiao, Y. (2011) 'Multivariate random forests', *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 1(1), pp. 80–87.

Shen, F., Liu, J. and Wu, K. (2020) 'Multivariate Time Series Forecasting based on *Elastic Net* and High-Order Fuzzy Cognitive Maps: A Case Study on Human Action Prediction through EEG Signals', *IEEE Transactions on Fuzzy Systems*.

Zou, H. and Hastie, T. (2005) 'Regularization and variable selection via the *Elastic Net*', *Journal of the royal statistical society: series B (statistical methodology)*, 67(2), pp. 301–320.