Jurnal Sains dan Edukasi Sains
Vol.6, No.1, Februari 2023: 34-43
https://doi.org/10.24246/juses.v6i1p34-43

Perbandingan Kinerja *Regresi Decision Tree* dan Regresi Linear Berganda untuk Prediksi BMI pada Dataset *Asthma*

Alfida Tegar Nurani[1,*], Adi Setiawan[1], Bambang Susanto[1]

[1]Program Studi Matematika, Fakultas Sains dan Matematika, Universitas Kristen Satya Wacana, Jalan Diponegoro 52-60 Salatiga 50711, Jawa Tengah, Indonesia

*email korespondensi: alfidategarnurani25@gmail.com

Received: 26 Februari 2023; ***Revised***: 3 Maret 2023; ***Accepted***: 7 Mei 2023; ***Published***: 26 Mei 2023

ABSTRAK

Penelitian ini menerapkan dua metode regresi yaitu Decision Tree dan linear berganda untuk memprediksi Body Mass Index (BMI) berdasarkan variabel-variabel lainnya pada dataset Asthma. Metode Decision Tree merupakan salah satu cara data processing dalam memprediksi masa depan dengan cara membangun klasifikasi dan regresi model dalam bentuk struktur pohon. Analisis regresi linear berganda merupakan model regresi yang melibatkan lebih dari satu variabel independen. Analisis regresi linear berganda dilakukan untuk mengetahui arah dan seberapa besar pengaruh variabel independen terhadap variabel dependen. Untuk mengukur tingkat keakuratan model peramalan, digunakan suatu ukuran keakuratan MAPE (Mean Absolute Percentage Error). Hasil MAPE menunjukkan nilai rata-rata prosentase kesalahan mutlak dari nilai sebenarnya dengan nilai peramalan. Pada penelitian ini, metode regresi linear berganda menunjukkan hasil yang lebih baik dengan nilai MAPE berturut-turut sebesar 12,737%; 12,76%; 12,89%; dan 12,99% untuk proporsi data uji berturut-turut 10%, 20%, 30%, dan 40% sedangkan nilai MAPE dari metode regresi Decision Tree sebesar 12,758%; 12,79%; 12,92%; dan 13,13%. Apabila digunakan ukuran kebaikan yang lain seperti MAE dan RMSE akan memberikan hasil yang analog, sedangkan ukuran kebaikan R^2 berkebalikan. Penelitian ini dapat dilanjutkan dengan membandingkan hasil yang diperoleh dengan menggunakan metode lain dalam machine learning seperti Support Vector Machine (SVM), random forest, Artificial Neural Network (ANN), dan lain-lain.

Kata-kata kunci: asma; MAPE; regresi *Decision Tree*; regresi linear berganda; *Single Nucleotide Polymorphism*

PENDAHULUAN

Penelitian tentang *machine learning*, khususnya tentang *Decision Tree* banyak dilakukan seperti dapat dilihat pada penelitian berikut ini. Asma merupakan penyakit yang menyerang paru-paru manusia dengan gangguan inflamasi kronik yang menyebabkan sesak nafas, dada terasa berat serta batuk (terutama di malam dan dini hari). Asma dapat dipengaruhi oleh banyak faktor diantaranya jenis kelamin, umur, perokok aktif maupun pasif, genetik, serta *Body Mass Index* (BMI). Inflamasi kronik melibatkan sel-sel inflamasi diantaranya eosinofil, sel mast dan leukotrin. Asma sering diderita pada anak-anak, orang dewasa, maupun orang lanjut usia. Dengan menggunakan Algoritme Genetika, sistem pakar akan membantu dalam menemukan informasi jenis penyakit asma berdasarkan gejala klinis yang dirasakan sampai ditemukannya kesimpulan berdasarkan hasil diagnosis berupa informasi mengenai cara pengobatan asma. Kesimpulan dari hasil pengujian sistem diperoleh jawaban sangat menarik 47%, menarik 45%, dan tidak menarik 8%. Dan durasi waktu proses perhitungan pada algoritme genetika kurang efisien (Wijaya & Toyib, 2018).

Metode regresi merupakan sebuah metode analisis statistik untuk melakukan prediksi menggunakan pengembangan hubungan antara dua atau lebih variabel. Hubungan tersebut bersifat fungsional dan disajikan dalam bentuk matematis. Tujuan dari penelitian ini untuk mendapatkan prediksi penjualan properti berdasarkan data penjualan properti pada PT XYZ. Prediksi penjualan properti dengan menggunakan metode regresi linear ini dapat dikatakan tergolong dalam kategori sangat baik didasari hasil pengujian keakurasian terhadap keseluruhan tipe yang menampilkan nilai MSE, RMSE, dan MAPE yang memenuhi standar (Ayuni & Fitrianah, 2019).

Metode *Decision Tree* merupakan salah satu cara data *processing* dalam memprediksi masa depan dengan cara membangun klasifikasi dan regresi model dalam bentuk struktur pohon. *Decision Tree* adalah model visual untuk memungkinkan memahami proses pembuatan keputusan yang terstruktur, bertahap dan secara rasional. Pada diagram pohon keputusan proses pembuatan keputusan yang terstruktur dapat dilihat secara visual. Untuk mengembangkan usaha menjual mobil secara kredit PT. OTO Multiartha menggunakan

Visual Basic 6.0 dan *Microsoft Acces 2003* merupakan aplikasi yang digunakan untuk mengembangkan sistem dan sebagai media penyimpan data. Sistem yang dibuat adalah Sistem Pendukung keputusan dengan menggunakan metode *Decision Tree* untuk pemodelan sistemnya. Penelitian ini menghasilkan kesimpulan *error* pada pengolahan data menggunakan rule awal *Decision Tree* berkurang setelah dilakukan penyederhanaan rule melalui pengujian *Chi Square* yang menghasilkan rule akhir *Decision Tree,* yaitu dari 20% menjadi 15% (Junaedy et al., 2017). Namun demikian, masih relatif sedikit penelitian tentang pemanfaatan regresi *Decision Tree* dengan menggunakan data yang memperhitungkan faktor genetik khususnya dengan menggunakan marker SNP (*Single Nucleotide Polymorphism*). Penelitian ini bertujuan untuk memperoleh kinerja regresi *Decision Tree* yang digunakan untuk memodelkan target BMI (*Body Mass Index*) terhadap peubah prediktor lain seperti jenis kelamin, status merokok atau tidak, status terkena asma atau tidak dan faktor peubah genetik SNP pada dataset *Asthma*. Hasil yang diperoleh dibandingkan dengan kinerja model regresi linear ganda.

EKSPERIMEN

Teori yang dipelajari dalam penelitian ini meliputi model regresi linear ganda dan Regresi *Decision Tree*. Di samping, itu juga dijelaskan tentang ukuran kebaikan model dengan menggunakan statistik MAE, MAPE, RMSE, dan R^2. Selanjutnya diberikan gambaran tentang bagaimana menggunakan kedua metode tersebut pada data dengan ukuran kecil.

Metode regresi linier sederhana adalah metode statistik yang berfungsi untuk menguji sejauh mana hubungan sebab akibat antara variabel prediktor (independen) (*x*) dan variabel korektor (*y*) (Utami et al., 2022). Regresi linier berganda bertujuan untuk mempelajari lebih lanjut mengenai hubungan antara beberapa variabel independen (prediktor) dan variabel dependen (hasil) (Abdulhafedh, 2022).

Analisis regresi linear berganda adalah hubungan secara linear antara dua atau lebih variabel independen ($X_1, X_2, \dots X_p$) dengan variabel dependennya (Y). Biasanya data yang digunakan berskala interval atau rasio (Maharadja et al., 2021). Persamaan regresi linear berganda sebagai berikut (Basri, 2019):

$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_pX_p$$

Body Mass Index (BMI) merupakan indeks massa tubuh yang digunakan untuk menunjukkan berat badan. Dengan melalui BMI dapat mengetahui angka berat badan termasuk kategori normal, berlebih atau justru kurang. Cara menghitung BMI yaitu dengan membagi berat badan dalam satuan kilogram (Kg) dan tinggi badan dalam satuan meter kuadrat (m^2). Menurut Organisasi Kesehatan Dunia, perhitungan BMI terbagi menjadi 4 kategori diantaranya yaitu BMI dengan atau di atas 30 termasuk dalam kategori berat badan obesitas, BMI antara 25-29,9 termasuk dalam kategori berat badan berlebih, BMI antara 18,5-24,9 termasuk dalam kategori berat badan normal dan BMI di bawah 18,5 termasuk dalam kategori berat badan di bawah normal (Arini & Wijana, 2020).

MAE *(Mean Absolute Error)* adalah rata-rata selisih mutlak nilai sebenarnya (aktual) dengan nilai perkiraan (prediksi). MAE *(Mean Absolute Error)* metode yang digunakan untuk mengukur keakuratan suatu model statistik dalam melakukan perkiraan (prediksi). MAE dapat dihitung menggunakan rumus:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|f_t - y_t|$$

dengan

f_t : nilai hasil perkiraan (prediksi) ke-t $(t = 1, \dots, n)$,

y_t : nilai sebenarnya (aktual) ke-t $(t = 1, \dots, n)$,

n : banyaknya data yang diuji.

Komponen $|f_t - y_t|$ menunjukkan nilai kesalahan atau perbedaan nilai sebenarnya (aktual) dengan nilai perkiraan (prediksi). Semakin kecil nilai MAE *(Mean Absolute Error)*, maka semakin baik model tersebut dalam melakukan perkiraan (prediksi) (Suryanto, 2019).

MAPE adalah pengukuran statistik tentang akurasi perkiraan (prediksi) pada metode peramalan. MAPE digunakan untuk menghitung persentase kesalahan rata-rata secara mutlak (absolut) (Medya et al., 2022). Metode MAPE dapat dihitung menggunakan rumus :

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|Y_t - Y'_t|}{Y_t}$$

dengan

Y_t : nilai sebenarnya (aktual) ke-t $(t = 1, \dots, n)$,

Y'_t : nilai hasil perkiraan (prediksi) ke-t $(t = 1, \dots, n)$,
n : banyaknya data yang diuji.

Semakin kecil nilai persentase kesalahan (*percentage error*) pada MAPE maka, semakin akurat hasil peramalan tersebut. Dari **Tabel 1** dapat dilihat rentang nilai MAPE. Nilai MAPE kurang dari 10% memiliki tingkat akurasi prediksi yang sangat baik. Nilai MAPE antara 10% sampai 20% memiliki tingkat akurasi prediksi baik. Nilai MAPE antara 20% sampai 50% memiliki tingkat akurasi prediksi cukup sedangkan nilai MAPE lebih dari 50% memiliki tingkat akurasi prediksi buruk (Nurfadilah et al., 2022).

Tabel 1. Rentang Nilai MAPE

Nilai MAPE	Akurasi Prediksi
<10 %	Sangat baik
10% - 20%	Baik
20% - 50%	Cukup
>50%	Buruk

RMSE merupakan nilai rata-rata dari jumlah kuadrat kesalahan. RMSE juga dapat menyatakan ukuran besarnya kesalahan yang dihasilkan oleh suatu model perkiraan (prediksi). Rumus RMSE sebagai berikut :

$$RMSE = \frac{\sum_{t=1}^{n}(Y'_t - Y_t)^2}{n}$$

dengan
Y'_t : nilai data sebenarnya (aktual) ke-t $(t = 1, \dots, n)$,
Y_t : nilai sebenarnya (aktual) ke-t $(t = 1, \dots, n)$,
n : banyaknya data yang diuji.

Nilai hasil RMSE akan baik hasil perkiraan (prediksi) apabila nilai RMSE semakin rendah (Normah et al., 2022). R-Squared merupakan suatu nilai yang menampilkan seberapa besar variabel independen mempengaruhi variabel dependen. Angka yang berkisar sekitar 0 sampai 1 yang mengindikasikan besar kombinasi variabel independen secara bersama-sama mempengaruhi nilai variabel dependen. Jenis hubungan koefisien determinasi dinyatakan pada **Tabel 2**. Rumus R-Squared adalah (Mostafa & Amano, 2019):

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(y_t - \hat{y})^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2}$$

dengan
y_t : data yang diuji respon ke-t $(t = 1, \dots, n)$,
$\hat{y}$: ramalan respon ke-t $(t = 1, \dots, n)$,
$\bar{y}$: rata-rata,
n : banyaknya data yang diuji.

Tabel 2. Interpretasi Nilai R^2

Interval Koefisien	Hubungan
1 – 0,8	Sangat kuat
0,6 – 0,79	Kuat
0,4 – 0,59	Cukup kuat
0,2 – 0,39	Lemah
0 – 0,19	Sangat lemah

Contoh Kasus Data Iris

Ahli statistika dan biologis Inggris yang bernama Ronald Fisher, dalam *paper*-nya (1936) memperkenalkan dataset multivariat yang disebut Dataset Iris. Dataset Iris terdiri dari 3 spesies iris diantaranya Iris Setosa, Iris versicolor, dan Iris virginica. Setiap masing-masing spesies memiliki 50 sampel. Pada dataset ini diberikan 5 kolom yaitu Petal.Length, Petal.Width, Sepal.Length, Sepal.Width dan Species. Dalam contoh kasus ini, fitur Sepal.Length menjadi target T, sedangkan Sepal.Width (Sepal.Width dibuat menjadi 2 kelas yaitu Tinggi untuk yang lebih dari atau sama dengan 2,85 cm dan Rendah untuk yang lain) dan Species sebagai fitur/prediktor seperti yang terlihat pada **Tabel 3**.

Tabel 3. Data Iris

Sepal.Length	Sepal.Width	Species
5,1	Tinggi	Setosa
4,9	Tinggi	Setosa
4,7	Tinggi	Setosa
4,6	Tinggi	Setosa
5,0	Tinggi	Setosa
5,4	Tinggi	Setosa
4,6	Tinggi	Setosa
5,0	Tinggi	Setosa
4,4	Tinggi	Setosa
4,9	Tinggi	Setosa
7,0	Tinggi	Versicolor
6,4	Tinggi	Versicolor
6,9	Tinggi	Versicolor
5,5	Rendah	Versicolor
6,5	Rendah	Versicolor
5,7	Rendah	Versicolor
6,3	Tinggi	Versicolor
4,9	Rendah	Versicolor
6,6	Tinggi	Versicolor
5,2	Rendah	Versicolor
6,3	Tinggi	Virginica
5,8	Rendah	Virginica
7,1	Tinggi	Virginica
6,3	Tinggi	Virginica
6,5	Tinggi	Virginica
7,6	Tinggi	Virginica
4,9	Rendah	Virginica
7,3	Tinggi	Virginica
6,7	Rendah	Virginica
7,2	Tinggi	Virginica

Dalam contoh ini, hanya digunakan sebagian data untuk memberikan gambaran bagaimana metode *Decision Tree* digunakan. Dalam hal ini, standard deviasi untuk fitur *X* yaitu *S*(*X*) dihitung dengan rumus yaitu

$$S(X) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

sedangkan rumus standar deviasi *S*(*T*,*X*) untuk dua atribut yaitu target *T* dan fitur *X* adalah

$$\boldsymbol{S}(T,X) = \sum_{c \in X} P(c)S(c)$$

dengan,

c : kelas dalam fitur target *T* dan prediktor *X*,
P(*c*) : probabilitas kelas dari keseluruhan dalam fitur,
$S(c)$: standar deviasi kelas dari keseluruhan dalam fitur.

Dalam pembuatan Decision Tree untuk regresi digunakan langkah-langkah berikut ini:
Langkah 1. Dihitung standar deviasi dari Target yaitu Sepal.Length adalah 0,964073.
Langkah 2. Data dipisahkan dalam fitur Sepal.Width dan fitur Species serta masing-masing dihitung standard deviasi untuk dua atribut (yaitu target Sepal.Length dan fitur Sepal.Width). **Tabel 4** merupakan standar deviasi untuk tiap kelas dalam fitur Sepal.Width. Selanjutnya dihitung standard deviasi dua atribut yaitu target Sepal.Length dan Sepal.With yang dihitung dengan

$$\begin{aligned}\boldsymbol{S}(Sepal.Length, Sepal.Width) &= \boldsymbol{P}(Rendah) \times \boldsymbol{S}(Rendah) + \boldsymbol{P}(Tinggi) \times \boldsymbol{S}(Tinggi) \\ &= \left(\frac{8}{30}\right) \times 0{,}676 + \left(\frac{22}{30}\right) \times 1{,}054 \\ &= 0{,}953509\end{aligned}$$

Terlihat bahwa standard deviasi ini sedikit lebih kecil dibandingkan dengan standar deviasi target sehingga diperoleh pengurangan sebesar

SDR(Sepal.Width) = S(Sepal.Length) - S(Sepal.Length, Sepal.Width)
= 0,964073 – 0,953509
= 0,010564

Lebih lanjut dihitung standard deviasi dua atribut yaitu target Sepal.Length dan Species yang dihitung dengan

$$\boldsymbol{S}(Sepal.Length, Species) = \boldsymbol{P}(Setosa) \times \boldsymbol{S}(Setosa) + \boldsymbol{P}(Versicolor) \times \boldsymbol{S}(Versicolor) + \boldsymbol{P}(Virginica) \times \boldsymbol{S}(Virginica)$$
$$= \left(\frac{10}{30}\right) \times 0{,}291357 + \left(\frac{10}{30}\right) \times 0{,}727248 + \left(\frac{10}{30}\right) \times 0{,}804225$$
$$= 0{,}607610$$

Terlihat bahwa standard deviasi ini lebih kecil dibandingkan dengan standar deviasi target sehingga diperoleh pengurangan sebesar

SDR(Species) = *S*(Sepal.Length) - *S*(Sepal.Length, Species)
= 0,964073 – 0,607610
= 0,356464

Dalam hal ini, **Tabel 5** merupakan standar deviasi untuk tiap kelas dalam fitur Species.

Tabel 4. Standar Deviasi target untuk tiap kelas dalam fitur Sepal.Width

		Standar Deviasi	Cacah
Sepal.Width	Rendah	0,676	8
	Tinggi	1,054	22

Tabel 5. Standar Deviasi Target untuk tiap kelas dalam fitur Species

		Standar Deviasi	Cacah
Sepal.Width	Setosa	0,291357	10
	Versicolor	0,727248	10
	Virginica	0,804225	10

Langkah 3. Mengingat SDR(Species) lebih tinggi dibandingkan dengan SDR(Sepal.Width) maka dipilih Species sebagai *decision node* (lihat **Gambar 1**). Data dibagi menjadi dua bagian berdasarkan Species Setosa (karena hanya mempunyai kelas Sepal.Width Tinggi dan tidak ada kemungkinan lain, lihat **Tabel 6**) dan bukan Species Setosa (yaitu Species Versicolor dan Species Virginica, lihat **Tabel 7**) yang mungkin mempunyai kelas Sepal.Width Tinggi dan Rendah. Berdasarkan **Tabel 6**, terdapat 10 satuan yang mempunyai Spesies Setosa dan semuanya mempunyai Sepal.Width Tinggi dan mempunyai rata-rata Sepal.Length 4,68 cm (lihat **Gambar 1**).

Tabel 6. Tabel hasil percabangan pertama dari *Decision Tree* dengan Species Setosa

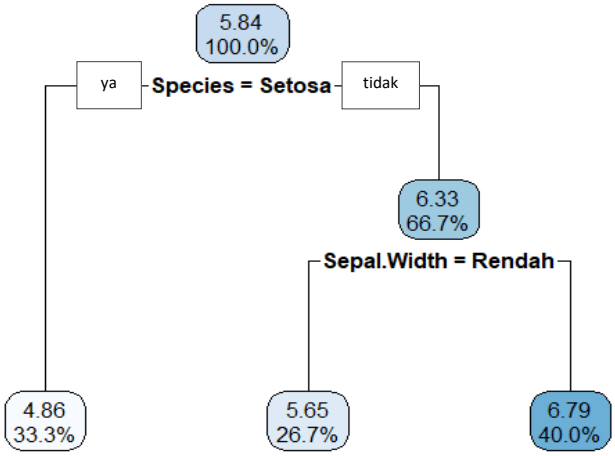
Species	Sepal.Width	Sepal.Length
Setosa	Tinggi	5,1
Setosa	Tinggi	4,9
Setosa	Tinggi	4,7
Setosa	Tinggi	4,6
Setosa	Tinggi	5,0
Setosa	Tinggi	5,4
Setosa	Tinggi	4,6
Setosa	Tinggi	5,0
Setosa	Tinggi	4,4
Setosa	Tinggi	4,9

Tabel 7. Tabel hasil percabangan pertama dari Decision Tree dengan Species bukan Setosa

Species	Sepal.Width	Sepal.Length
Versicolor	Tinggi	7,0
Versicolor	Tinggi	6,4
Versicolor	Tinggi	6,9
Versicolor	Rendah	5,5
Versicolor	Rendah	6,5
Versicolor	Rendah	5,7
Versicolor	Tinggi	6,3
Versicolor	Rendah	4,9
Versicolor	Tinggi	6,6
Versicolor	Rendah	5,2

Species	Sepal.Width	Sepal.Length
Virginica	Tinggi	6,3
Virginica	Rendah	5,8
Virginica	Tinggi	7,1
Virginica	Tinggi	6,3
Virginica	Tinggi	6,5
Virginica	Tinggi	7,6
Virginica	Rendah	4,9
Virginica	Tinggi	7,3
Virginica	Rendah	6,7
Virginica	Tinggi	7,2

Langkah 4. Mengingat **Tabel 7** mempunyai koefisien variasi (*coefficient of variation*) yaitu 0,12 yang lebih dari batas yang dipilih yaitu misalkan 0,1 maka dilakukan percabangan dalam *Decision Tree*. Fitur Sepal.Width cuma terdapat 2 kelas maka percabangan dilakukan berdasarkan kelas Rendah dan Tinggi sehingga kelas Rendah terdapat 8 satuan akan mempunyai rata-rata 5,65 cm dan kelas Tinggi terdapat 12 satuan mempunyai rata-rata 6,79 cm. Berdasarkan langkah-langkah di atas diperoleh *Decision Tree* seperti Gambar 1 dan gambar tersebut juga menyatakan hasil penggunaan program R dengan *package rpart* dan *rpart.plot*.



Gambar 1. *Decision Tree* dari Data Iris

Dalam penelitian ini juga digunakan *software* R untuk memprediksi dan melakukan regresi untuk memprediksi dataset *Asthma.* Metode pengambilan data yang digunakan dalam penelitian ini merupakan dataset *Asthma* diperoleh dari *software* R dengan nama paket "SNPassoc". Fitur-fitur yang digunakan sejumlah 57 fitur yaitu Country, Gender, BMI, Smoke, Case Control dan 51 data genetik SNP (Single Nucleotide Polymorphism) yang terdiri dari pasangan basa A (Adenine), T (Thymine), G (Guanine), dan C (Cytosine) *D*ata *Asthma* di *software* R kemudian dilakukan analisis dengan metode *Decision Tree* untuk regresi. Lalu dibandingkan dengan regresi linear berganda untuk mencapai hasil yang maksimal. Untuk mengukur tingkat keakuratan model peramalan, digunakan metode MAPE *(Mean Absolute Percentage Error*). Hasil MAPE menunjukkan nilai rata-rata kesalahan mutlak atau absolut dari nilai sebenarnya dengan nilai peramalan. Statistik yang lain yaitu RMSE, MAE, dan R-SQUARED digunakan sebagai pembanding terhadap MAPE.

HASIL DAN PEMBAHASAN

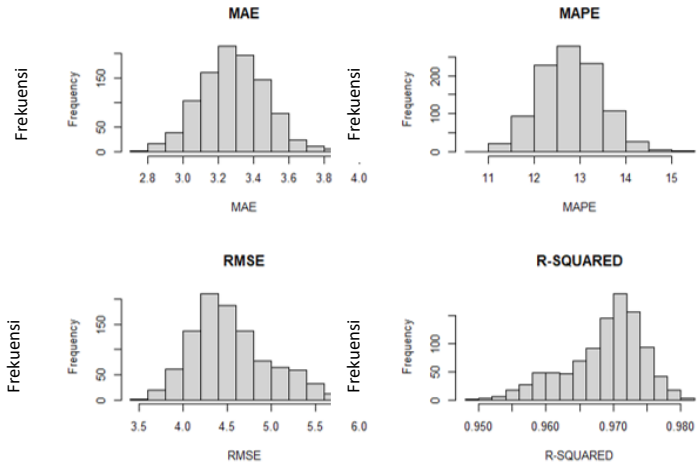
Regresi *Decision Tree*

Dalam data *Asthma* yang dimiliki akan dilakukan regresi menggunakan metode regresi *Decision Tree*. Data *Asthma* memiliki 1578 baris dengan 57 fitur. Setelah nilai NA dihilangkan, diperoleh baris memiliki banyak 1076 dengan kategori fitur. Dalam menentukan besar nilai presisi dan akurasi pada metode *Decision Tree* dilakukan 1000 kali simulasi model *Decision Tree* dengan membagi data menjadi 2 dimana 80% adalah data training dan 20% adalah data *test*. Hasil nilai-nilai MAE, MAPE, RMSE, dan R-SQUARED simulasi 1000 kali

tersebut dapat dilihat pada **Gambar 2** dan **Tabel 9**. Dari hasil **Tabel 9** tersebut dapat dilihat nilai *mean* pada MAPE *(Mean Absolute Percentage Error)* yaitu 12,79%. Dilakukan juga 1000 kali simulasi dengan proporsi data uji yang sama yaitu 10%, 20%, 30%, dan 40% menghasilkan nilai MAE *(Mean Absolute Error)*, MAPE *(Mean Absolute Percentage Error),* RMSE *(Root Mean Square Error),* dan R-Squared. Dari hasil simulasi 1000 kali, didapatkan hasil seperti pada **Tabel 10**.

Tabel 9. Ringkasan Statistik Hasil Simulasi 1000 kali *Decision Tree*

STATISTIK	MAE	MAPE	RMSE	R-SQUARED
Min	2,782	10,87%	3,560	0,9500
Kuartil 1	3,157	12,31%	4,244	0,9657
Median	3,281	12,79%	4,471	0,9700
Mean	3,285	12,79%	4,545	0,9688
Kuartil 3	3,408	13,21%	4,793	0,9727
Max	3,932	15,16%	5,941	0,9805



Gambar 2. Histogram nilai-nilai MAE, MAPE, RMSE dan R-SQUARED hasil simulasi 1000 kali

Tabel 10. Hasil Perhitungan *Decision Tree* Simulasi 1000 kali

Proporsi	MAE	MAPE	RMSE	R-SQUARED
10%	3,273	12,76%	4,492	0,9692
20%	3,285	12,79%	4,545	0,9688
30%	3,315	12,92%	4,603	0,9680
40%	3,370	13,13%	4,666	0,9671

Regresi Linear Berganda

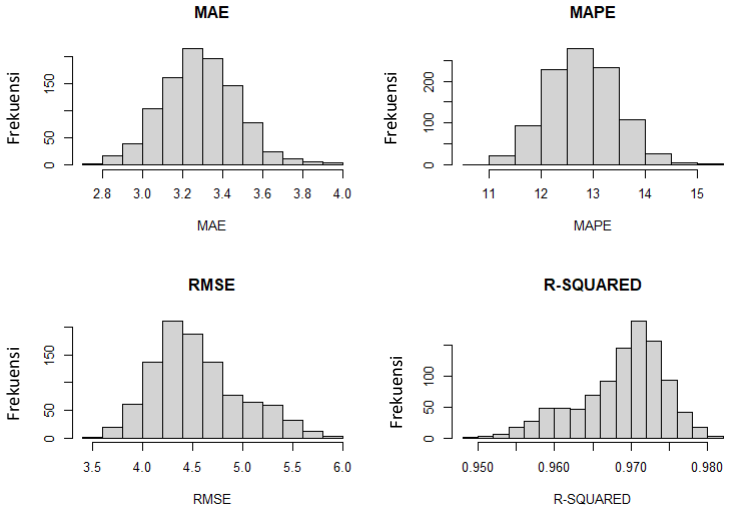
Metode yang digunakan untuk membandingkan dengan regresi *Decision Tree* adalah metode regresi linear berganda. Regresi linear berganda juga dilakukan dalam data *Asthma* sebanyak 1000 kali simulasi dengan 80% adalah *training dataset* dan 20 adalah *testing dataset.* Dari hasil 1000 kali simulasi pada regresi linear berganda, didapatkan hasil seperti pada **Tabel 11**. Disamping itu, untuk dapat mendapatkan hasil yang maksimal juga dilakukan simulasi 1000 kali dengan proporsi data uji sebesar 10%, 20%, 30%, dan 40%. Yang menghasilkan nilai MAE *(Mean Absolute Error)*, MAPE *(Mean Absolute Percentage Error),* RMSE *(Root Mean Square Error),* dan R-Squared. Dari hasil simulasi 1000 kali, didapatkan hasil seperti pada **Tabel 12**. Berdasarkan **Tabel 12** yang menunjukkan hasil perhitungan regresi linear berganda dengan simulasi 1000 kali dapat digambarkan histogram seperti terlihat pada **Gambar 3**.

Tabel 11. Hasil Simulasi 1000 kali Regresi Linear Berganda

	MAE	MAPE	RMSE	R-SQUARED
Min	2,689	10,55%	3,540	0,9507
Kuartil 1	3,137	12,33%	4,164	0,9663
Median	3,267	12,75%	4,407	0,9707
Mean	3,273	12,76%	4,502	0,9693
Kuartil 3	3,403	13,20%	4,766	0,9736
Max	3,866	14,72%	5,857	0,9805

Tabel 12. Hasil Perhitungan Regresi Linear Berganda Simulasi 1000 kali

Proporsi	MAE	MAPE	RMSE	R-SQUARED
10%	3,271	12,737%	4,482	0,9694
20%	3,273	12,76%	4,502	0,9693
30%	3,307	12,89%	4,560	0,9686
40%	3,331	12,99%	4,585	0,9683



Gambar 3. Histogram Regresi Linear Berganda Simulasi 1000 kali

Berdasarkan **Tabel 10** dan **Tabel 12** dapat dilakukan perbandingan *Decision Tree* dan regresi linear berganda berdasarkan simulasi yang telah dilakukan terhadap proporsi 10%, 20%, 30%, dan 40%. Hasil perbandingan menunjukkan bahwa nilai MAE *Decision Tree* dengan proporsi 20% sebesar 3,285 dimana angka tersebut lebih tinggi dibandingkan regresi linear berganda yakni sebesar 3,273. Nilai MAPE *Decision Tree* dengan proporsi 20% sebesar 12,79% dimana angka tersebut lebih tinggi dibandingkan regresi linear berganda yakni sebesar 12,76%. Nilai RMSE *Decision Tree* dengan proporsi 20% sebesar 4,545 dimana angka tersebut lebih tinggi dibandingkan regresi linear berganda yakni sebesar 4,502. Nilai R-Squared *Decision Tree* dengan proporsi 20% sebesar 0,9688 dimana angka tersebut lebih rendah dibandingkan regresi linear beranda yakni sebesar 0,9693. Hal itu berarti bahwa dalam kasus ini, *Decisien Tree* untuk Regresi cenderung memberikan hasil yang kurang baik dibandingkan dengan metode regresi linear ganda.

Penelitian lain terkini dan terkait dengan penelitian ini dapat dijelaskan berikut ini. Penelitian yang dilakukan oleh (Akbar et al., 2022) menyatakan bahwa algoritma *Decision Tree* C4.5 sangat bagus dalam memprediksi dataset pasien berpenyakit *stroke* dengan *splitting* data 70:30 terhadap 5110 dataset pasien penderita *stroke* yang diperoleh dari *Kaggle open datasets* karena nilai *Root Mean Square Error* (RMSE) yang dihasilkan sebesar 0.2350 dimana lebih kecil dibandingkan hasil dari *Support Vector Regression* yang menghasilkan *Root Mean Square Error* (RMSE) sebesar 0,3999. (Arista, 2022) menyatakan bahwa diagnosis Covid-19 dianggap sebagai topik penelitian yang sulit. Teknik pembelajaran mesin (*machine learning*) untuk diagnosis medis telah mendapatkan popularitas di komunitas medis karena kemampuannya untuk meramalkan penyakit sejak dini. Untuk menganalisis prediksi covid pada beberapa kriteria seperti akurasi, daya ingat, presisi, dan skor F1, teknik pohon keputusan dan regresi logistik digunakan dalam penelitian ini. Metode *Decision Tree* mengungguli algoritma linear berganda dalam percobaan ini. Dalam makalah ini hanya digunakan proporsi data uji 30 % sedangkan pada penelitian ini digunakan proporsi data uji yang lebih bervariasi dari 10%, 20%, 30%, dan 40%. Demikian juga ukuran kebaikannya menggunakan MAPE dan MAE disamping RMSE.

Menurut (Etode & Ree, 2021), meningkatnya jumlah penduduk di Indonesia beriringan dengan meningkatnya kesadaran serta tuntutan masyarakat atas mutu pelayanan kesehatan. Karena kurangnya tenaga medis, dokter spesialis, dan teknologi untuk mempermudah layanan kesehatan ini menyebabkan pasien hanya dirujuk ke dokter umum tanpa mempertimbangkan penyakit yang diderita. Digunakan sistem rekomendasi tujuan poli berbasis *Decision Tree* untuk memberikan rekomendasi tujuan poli pada pasien yang berkunjung ke Rumah Sakit Umum Daerah Bajawa. Dengan menggunakan metode *Decision Tree* untuk rekomendasi tujuan poli, *Decision Tree* memberikan rekomendasi tujuan poli berdasarkan waktu, kunjungan, keterangan penyakit, nama unit poli, dan jenis rawat dilakukan oleh admin bagian administrasi Rumah Sakit Umum Daerah Bajawa.

Regression tree atau pohon keputusan adalah suatu metode yang digunakan untuk membangun model prediksi dari data dimana model diperoleh dari hasil partisi secara rekursif ruang data dan menduga model prediksi sederhana di dalam setiap partisi. Dalam investasi saham seorang investor, sebaiknya memiliki prediksi untuk memprediksi harga saham yang ada pada waktu tertentu agar memiliki hasil yang memuaskan dengan mendapatkan keuntungan. Investor harus memprediksi harga saham di waktu mendatang dengan menggunakan *Regression Tree*. Berdasarkan hasil analisis terhadap prediksi harga saham serta pengambilan keputusan dengan metode *Regression Tree* dan *Self Organizing Map* (SOP), maka dihasilkan hasil prediksi menggunakan *Regression Tree* menghasilkan nilai MAPE pada data training sebesar 4,00% dan untuk data testing sebesar 3,77% yang artinya hasil prediksi sudah mendekati nilai aktualnya seperti penelitian oleh (Savitri et al., 2019).

KESIMPULAN DAN SARAN

Penelitian ini telah membandingkan akurasi (dalam bentuk MAPE) pada metode regresi *decision tree* dengan metode regresi linear ganda. Pada penelitian ini, metode regresi linear berganda menunjukkan hasil yang lebih baik dengan nilai MAPE berturut-turut sebesar 12,737%; 12,76%; 12,89%; dan 12,99% untuk 1000 kali percobaan dan proporsi data uji berturut-turut 10%, 20%, 30%, dan 40% sedangkan nilai MAPE dengan metode regresi *Decision Tree* sebesar 12,758%; 12,79%; 12,92%; dan 13,13% untuk 1000 kali percobaan dan proporsi data uji berturut-turut 10%, 20%, 30%, dan 40%. Nilai MAPE *(Mean Absolute Percentage Error)* antara metode regresi linear berganda dan *Decision Tree* memiliki perbedaan yang signifikan dengan menggunakan uji Kolmogorov-Smirnov dengan perolehan p-value lebih kecil dari 5%. Penelitian ini dapat dikembangkan dengan menggunakan metode lain dalam *machine learning* selain kedua metode di atas atau menggunakan *dataset* yang lain.

DAFTAR PUSTAKA

Abdulhafedh, A. (2022). Incorporating Multiple Linear Regression in Predicting the House Prices Using a Big Real Estate Dataset with 80 Independent Variables. *OALib*, *09*(01), 1–21. https://doi.org/10.4236/oalib.1108346

Akbar, F., Saputra, H. W., & Maulaya, A. K. (2022). Implementation of Decision Tree Algorithm C4 . 5 and Support Vector Regression for Stroke Disease Prediction Implementasi Algoritma Decision Tree C4 . 5 dan Support Vector Regression untuk Prediksi Penyakit Stroke. *Indonesian Journal of Machine Learning and Computer Science*, *2*(October), 61–67.

Arini, L. A., & Wijana, I. K. (2020). Korelasi Antara Body Mass Index (BMI) Dengan Blood Pressure (BP) Berdasarkan Ukuran Antropometri Pada Atlet. *JURNAL KESEHATAN PERINTIS (Perintis's Health Journal)*, *7*(1), 32–40. https://doi.org/10.33653/jkp.v7i1.390

Arista, A. (2022). Comparison Decision Tree and Logistic Regression Machine Learning Classification Algorithms to determine Covid-19. *Sinkron*, *7*(1), 59–65. https://doi.org/10.33395/sinkron.v7i1.11243

Ayuni, G. N., & Fitrianah, D. (2019). Penerapan metode Regresi Linear untuk prediksi penjualan properti pada PT XYZ. *Jurnal Telematika*, *14*(2), 79–86.

Basri, H. (2019). Pemodelan Regresi Berganda Untuk Data Dalam Studi Kecerdasan Emosional. *DIDAKTIKA : Jurnal Kependidikan*, *12*(2), 103–116. https://doi.org/10.30863/didaktika.v12i2.179

Etode, B. E. M., & Ree, D. E. T. (2021). Sistem Rekomendasi Tujuan Poli Pada Rumah Sakit Umum Daerah Bajawa Berdasarkan Metode Decision Tree. *Jurnal Ilmiah Teknik Informatika*, *15*(1), 62–74.

Junaedy, Musdar, I. A., & Angriani, H. (2017). Implementasi Metode Decision Tree Dalam Menentukan Pemberian Kredit Mobil Menggunakan Visual Basic (Studi Kasus UD PUTRA MAS Makassar). *Jtriste*, *4*(1), 55–67.

Maharadja, A. N., Maulana, I., & Dermawan, B. A. (2021). Penerapan Metode Regresi Linear Berganda untuk

Prediksi Kerugian Negara Berdasarkan Kasus Tindak Pidana Korupsi. *Journal of Applied Informatics and Computing*, *5*(1), 95–102. https://doi.org/10.30871/jaic.v5i1.3184

Medya, N., Ir, L., & Kamila, I. (2022). Metode Double Exponential Smoothing dalam Peramalan Jumlah Pemohon Paspor. *Jurnal Ilmiah Matematika*, *2*(1), 23–31.

Mostafa, S. M., & Amano, H. (2019). Effect of clustering data in improving machine learning model accuracy. *Journal of Theoretical and Applied Information Technology*, *97*(21), 2973–2981.

Normah, Rifai, B., Vambudi, S., & Maulana, R. (2022). Analisa Sentimen Perkembangan Vtuber Dengan Metode Support Vector Machine Berbasis SMOTE. *Jurnal Teknik Komputer AMIK BSI*, *8*(2), 174–180. https://doi.org/10.31294/jtk.v4i2

Nurfadilah, A., Budi, W., Kurniati, E., Suhaedi, D., Matematika, P. S., Bandung, U. I., & Statistik, B. P. (2022). *Penerapan Metode Moving Average untuk Prediksi Indeks Harga Konsumen*. *21*(1), 19–25.

Savitri, U. S., Saepudin, D., & Umbara, R. F. (2019). *Prediksi Harga Saham serta Pemberian Keputusan Jual Beli Menggunakan Metode Regression Tree dan Self Organizing Map (SOM)*. *6*(1), 2527–2538.

Suryanto, A. A. (2019). Penerapan Metode Mean Absolute Error (Mea) Dalam Algoritma Regresi Linear Untuk Prediksi Produksi Padi. *Saintekbu*, *11*(1), 78–83. https://doi.org/10.32764/saintekbu.v11i1.298

Utami, Y., Vinsensia, D., Nissa, A., & Sulastri, S. (2022). Forecasting the Sum of New College Students with Linear Regression Approach. *Jurnal Teknik Informatika C.I.T Medicom*, *14*(1), 10–15. https://doi.org/10.35335/cit.vol14.2022.231.pp10-15

Wijaya, A., & Toyib, R. (2018). Sistem Pakar Diagnosis Penyakit Asma Dengan Menggunakan Algoritme Genetik (Studi Kasus RSUD Kabupaten Kepahiang). *Pseudocode*, *5*(2), 1–11. https://doi.org/10.33369/pseudocode.5.2.1-11