LAPORAN

PREDIKSI PENYEBARAN PENYAKIT DEMAM BERDARAH

Ditulis untuk Memenuhi Tugas Mata Kuliah Kapita Selekta

Oleh:

Hanson Natalie	**01112170029**
Marcelina	**01112180011**
William Setiawan	**01112180001**

PROGRAM STUDI MATEMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS PELITA HARAPAN

JAKARTA

2021

BAB I

EXECUTIVE SUMMARY

Penyakit demam berdarah adalah suatu penyakit yang ditularkan oleh nyamuk yang terjadi pada daerah yang beriklim tropis dan sub-tropis. Untuk kasus demam berdarah yang ringan, gejala yang muncul memiliki kemiripan dengan penyakit flu, seperti demam tinggi, ruam, dan nyeri otot dan sendi. Sedangkan untuk kasus demam berdarah yang parah, dapat menyebabkan pendarahan, penurunan tekanan darah, dan bahkan bisa berujung pada kematian.

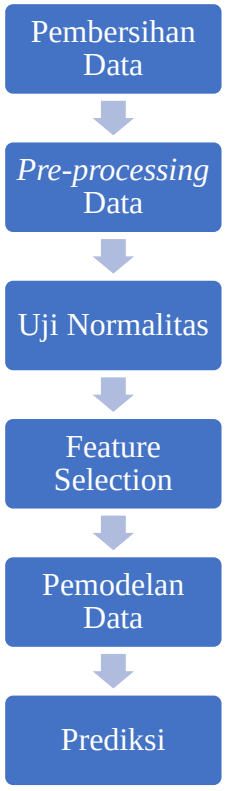
Berhubung penyakit demam berdarah ditularkan oleh nyamuk, transmisi dinamis dari penyakit demam berdarah memiliki korelasi kuat terhadap variabel yang mempengaruhi perubahan kepadatan nyamuk, seperti variabel iklim. Maka, para ilmuwan berargumen bahwa perubahan variabel iklim, seperti temperatur dan presipitasi (curah hujan), dapat menyebabkan perubahan infeksi penyakit demam berdarah yang cukup signifikan.

Pada tahun-tahun terakhir penyakit demam berdarah mulai menyebar. Tidak hanya pada daerah Asia Tenggara dan pulau-pulau Pasifik, penyakit demam berdarah juga mulai menyebar di daerah Amerika Latin. Demi meningkatkan inisiatif penelitian dan alokasi sumber daya untuk melawan penyakit demam berdarah, diperlukannya pemahaman terhadap hubungan antara penyakit demam berdarah dan variabel iklim. Pemahaman tersebut dapat berupa prediksi jumlah kasus demam berdarah berdasarkan data yang berisi variabel-variabel iklim.

Laporan ini bertujuan untuk melakukan prediksi banyaknya kasus demam berdarah dari kota San Juan dan Iquitos dengan menggunakan data dari perlombaan daring *DengAI: Predicting Diseases Spread.* Penulis menggunakan model regresi dengan *decision tree* tanpa *pruning* dan *decision tree* yang sudah di-*pruning.* Hasil prediksi penulis menunjukkan bahwa model *decision tree* tanpa *pruning* dapat memprediksi kasus demam berdarah dengan akurat dan memiliki nilai *Mean Absolute Error* sebesar 26.1971 dengan peringkat 2024 dari 10674 kompetitor.

BAB II
METODOLOGI

Pada bab II akan dijelaskan langkah-langkah yang digunakan dalam membuat model prediksi penyebaran demam berdarah. Langkah-langkah tersebut dapat dilihat pada Gambar 2.1.



Gambar 2.1: Langkah-langkah pengerjaan

2.1 Data Kasus Demam Berdarah

Data kasus demam berdarah yang disediakan berasal dari perlombaan *data science* daring berjudul *DengAI: Predicting Diseases Spread*. *Data* tersebut memuat catatan iklim dan total kasus demam berdarah mingguan untuk dua kota, yaitu kota *San Juan* selama 19 tahun dan kota *Iquitos* selama 11 tahun. Data iklim yang ada terdiri dari pengukuran stasiun cuaca data iklim harian GHCN NOAA, pengukuran curah hujan satelit PERSIANN dengan skala 0.25x0.25 derajat,

pengukuran analisis ulang sistem prakiraan iklim NCEP NOAA dengan skala 0.5x0.5 derajat dan pengukuran indeks vegetasi perbedaan ternormalisasi (NDVI).

2.2 Pembersihan Data

Data yang diunduh dari perlombaan dalam bentuk *spreadsheet* dalam format csv. Data tersebut tidak lengkap sebab ada beberapa nilai kosong (*Na*). Nilai-nilai kosong tersebut harus diatasi untuk mempermudah pengolahan data. Pertama, nilai-nilai kosong akan dihitung untuk setiap variabel. Variabel dengan jumlah nilai kosong lebih besar sama dengan suatu bilangan akan dihapus, sementara nilai-nilai kosong lainnya akan diisi. Salah satu cara untuk mengisi nilai-nilai kosong tersebut adalah metode interpolasi linier. Metode interpolasi linier adalah metode yang digunakan untuk mencari nilai dari sesuatu yang berada dalam suatu interval atau diantara dua buah titik yang segaris. Berikut adalah persamaan untuk metode interpolasi linier.

$$\frac{y - y_1}{y_2 - y_1} = \frac{x - x_1}{x_2 - x_1} \tag{2.1}$$

Dimana,

(x, y) : x adalah nilai dari suatu variabel yang akan dicari, sedangkan y adalah waktu dalam unit minggu.

(x_1, y_1) : nilai x dan y dari titik pertama.

(x_2, y_2) : nilai x dan y dari titik kedua.

2.3 *Pre-processing* Data

Pada tahap ini, data telah diisi lengkap dan beberapa variabel telah dihapus. Untuk meningkatkan akurasi prediksi, data akan dikoreksi terlebih dahulu. variabel-variabel dari data akan dianalisa untuk mencari nilai korelasi antar variabel dan data akan ditransformasi pula. Subbab 2.3.1 dan 2.3.2 akan menjelaskan prosesnya lebih lanjut.

2.3.1 Autokorelasi Antar Variabel

Variabel-variabel dari data akan diuji korelasi antar satu sama lain. Apabila terdapat nilai korelasi yang cukup tinggi, maka hal ini mengindikasikan adanya

autokorelasi. Autokorelasi dapat mengganggu akurasi dan presisi prediksi. Maka dari itu, variabel yang diduga autokorelasi akan dihapus.

2.3.2 Transformasi Data

Data yang memiliki banyak variabel dapat menimbulkan suatu masalah. Salah satu masalah yang dapat timbul adalah perbedaan skala antar variabel. Pemodelan prediksi data dapat lebih akurat apabila variabel-variabel yang dimasukkan memiliki skala atau jangkauan data yang sama. Contoh dari masalah adalah perbedaan skala temperatur terhadap curah hujan. Maka dari itu, data harus ditransformasi pula. Salah satu transformasi data adalah normalisasi data. Metode yang digunakan untuk normalisasi data pada penelitian ini adalah *z-score normalization*. Berikut adalah persamaannya untuk metode *z-score normalization*.

$$Z_{ij} = \frac{X_{ij} - \mu_j}{\sigma_j} \qquad (2.2)$$

Dimana,

X_{ij} : nilai data pada barisan ke-*i* dan pada variabel ke-*j*.

μ_j : nilai rata-rata dari variabel ke-*j*.

σ_j : nilai simpangan baku dari variabel ke-*j*.

2.4 Uji Normalitas

Penulis akan melakukan uji normalitas untuk masing-masing variabel dari data yang telah diunduh. Hal ini akan memperlihatkan apakah data yang digunakan pada laporan ini berdistribusi normal atau tidak normal. Uji normalitas akan dilakukan dengan metode *Shappiro Wilk test* dengan menggunakan *software Rstudio*.

2.5 *Feature Selection*

Data yang diunduh dari perlombaan dalam bentuk *spreadsheet* dalam format csv, ada banyak variabel dalam data ini prediksi akan lebih akurat apabila kita memilih variabel mana yang penting dan variabel mana yang tidak penting

dalam peramalan data ini, penulis menggunakan metode Boruta untuk menentukan seberapa pentingnya suatu variabel pada data ini.

2.6 Pemodelan Data dengan *Decision Tree*

Pemodelan data akan menggunakan data yang telah disiapkan. Penulis menggunakan metode *decision tree* untuk melakukan regresi. Berikut adalah langkah-langkah pembuatan model regresi dengan *decision tree.*

1. *Import* data yang telah diolah (pembersihan data, *pre-processing* data, transformasi data) pada program R.
2. Bagi data yang telah diolah menjadi data latih dan data uji.
3. Lakukan pembuatan model regresi dengan fungsi *rpart* menggunakan data latih.
4. Dengan model yang telah dibuat, prediksi *total_cases* pada data uji.
5. Evaluasi model dengan menghitung *mean absolute error* hasil prediksi dengan nilai sesungguhnya.
6. *Plot* nilai *complexity parameter* dan tentukan nilai *complexity parameter* optimal.
7. Lakukan *pruning* pada model yang telah dibuat pada langkah nomor 2.
8. Prediksi *total_cases* pada data uji menggunakan *decision tree* yang telah di *pruning*.
9. Evaluasi dengan model dengan menghitung *mean absolute error* hasil prediksi dengan nilai sesungguhnya.

2.9 Prediksi

Setelah menguji kemampuan model, maka akan dilakukan prediksi jumlah kasus demam berdarah dengan menggunakan data uji yang disediakan oleh *DengAi.*

BAB III

HASIL DAN ANALISIS

3.1 Data Kasus Demam Berdarah

Data kasus demam berdarah memiliki banyak variabel. Berikut adalah definisi dari variabel-variabel yang ada.

a. Indikator kota dan waktu
    - *city* - Kota
    - *year* - Tahun
    - *weekofyear* - urutan minggu dalam setahun
b. Pengukuran stasiun cuaca data iklim harian GHCN dari NOAA

    Jaringan Klimatologi Sejarah Global (GHCN) adalah database terintegrasi yang memuat ringkasan iklim dari statius permukaan tanah di seluruh dunia dan telah melewati tinjauan jaminan kualitas umum [1].
    - *station_avg_temp_c* - Suhu rata-rata (Celsius)
    - *station_diur_temp_rng_c* - Selisih suhu harian (Celsius)
    - *station_max_temp_c* - Suhu maksimum (Celsius)
    - *station_min_temp_c* - Suhu minimum (Celsius)
    - *station_precip_mm* - Total curah hujan (mm)
c. Pengukuran curah hujan satelit PERSIANN dengan skala 0.25x0.25 derajat

    Precipitation Estimation from Remotely Sensed Information using Artificial Neural Network – Climate Data Records (PERSIANN-CDR) adalah estimasi curah hujan harian pada resolusi spasial 0.25 derajat di lintang 60S-60N dari tahun 1983 hingga saat ini [2].
    - *precipitation_amt_mm* - Total curah hujan (mm)
d. Pengukuran analisis ulang sistem prakiraan iklim NCEP NOAA dengan skala 0.5x0.5 derajat

Analisis ulang sistem prakiraan iklim atau Climate Forecast System Reanalysis (CFSR) yang dilakukan oleh *National Centers for Environmental Prediction* (NCEP) merupakan analisis untuk mendapatkan perkiraan terbaik dari atmosfir, samudera, daratan, dan laut [3].

- *reanalysis_air_temp_k* - Mean suhu udara (Kelvin)
- *reanalysis_avg_temp_k* - Suhu udara rata-rata (Kelvin)
- *reanalysis_dew_point_temp_k* - Mean suhu titik embun (Kelvin)
- *reanalysis_max_air_temp_k* - Suhu udara maksimum (Kelvin)
- *reanalysis_min_air_temp_k* - Suhu udara minimum (Kelvin)
- *reanalysis_precip_amt_kg_per_m2* - Total curah hujan (kg/m2)
- *reanalysis_relative_humidity_percent* - Mean kelembapan relatif (%)
- *reanalysis_sat_precip_amt_mm* - Total curah hujan (mm)
- *reanalysis_specific_humidity_g_per_kg* - Mean kelembapan spesifik (g/kg)
- *reanalysis_tdtr_k* - Selisih suhu harian (Kelvin)

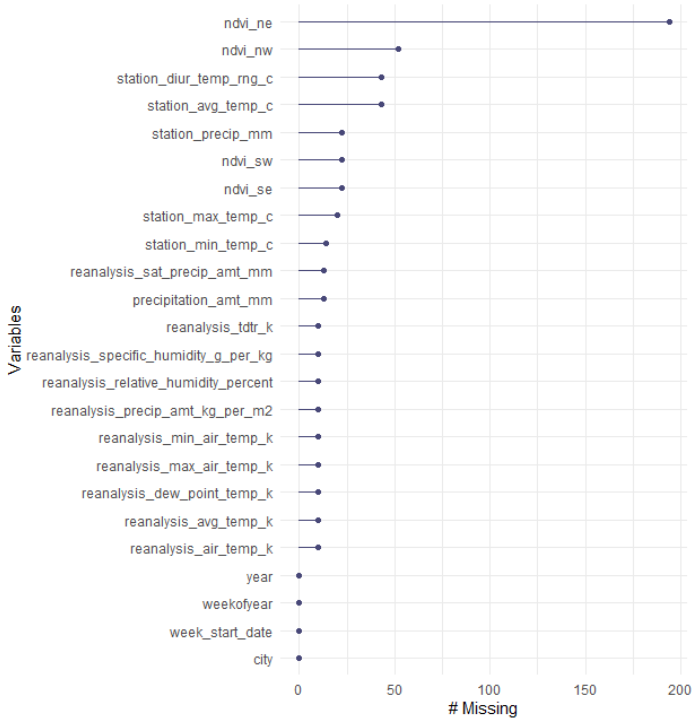
e. Pengukuran indeks vegetasi perbedaan ternormalisasi (NDVI)
Indeks vegetasi NDVI digunakan untuk mengukur aktivitas cakupan permukaan vegetasi [4]. Dengan dihitungnya indeks vegetasi, maka dapat diketahui potensi air tanah daerah tersebut.

- *ndvi_ne* - Piksel timur laut dari pusat kota
- *ndvi_nw* - Piksel barat laut dari pusat kota
- *ndvi_se* - Piksel tenggara dari pusat kota
- *ndvi_sw* - Piksel barat daya dari dari pusat kota

3.2 Proses Pembersihan Data

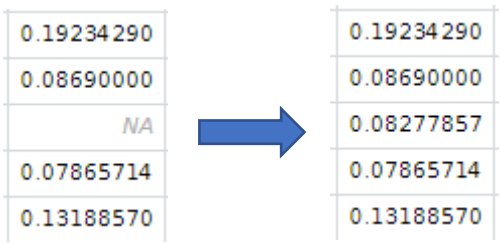
Data yang diunduh dan digunakan untuk pemodelan prediksi jumlah kasus demam berdarah tidaklah lengkap sebab ada banyak nilai kosong (*Na*). Penulis akan

menghitung jumlah nilai kosong pada masing-masing variabel terlebih dahulu sebelum menghapus variabel yang memiliki banyak nilai kosong. Untuk laporan ini, penulis memilih untuk menghapus variabel dengan jumlah nilai kosong lebih dari sama dengan 25. Dengan menggunakan *software Rstudio*, berikut adalah jumlah nilai kosong pada masing-masing variabel.



Gambar 3.1: Jumlah nilai kosong pada setiap variabel

Berdasarkan Gambar 3.1**:** Jumlah nilai kosong pada setiap variabel, ada empat variabel yang memiliki jumlah nilai kosong lebih dari 25, yaitu *ndvi_ne*, *ndvi_nw*, *station_diur_temp_mg_c*, dan *station_avg_temp_c*. Empat variabel tersebut akan dihapus dari data. Lalu, nilai-nilai kosong pada variabel lainnya akan diisi dengan metode interpolasi linier. Berikut adalah contoh pengisian nilai kosong pada suatu variabel.



Gambar 3.2: Pengisian nilai kosong dengan metode interpolasi linier

3.3 Proses *Pre-Processing* Data

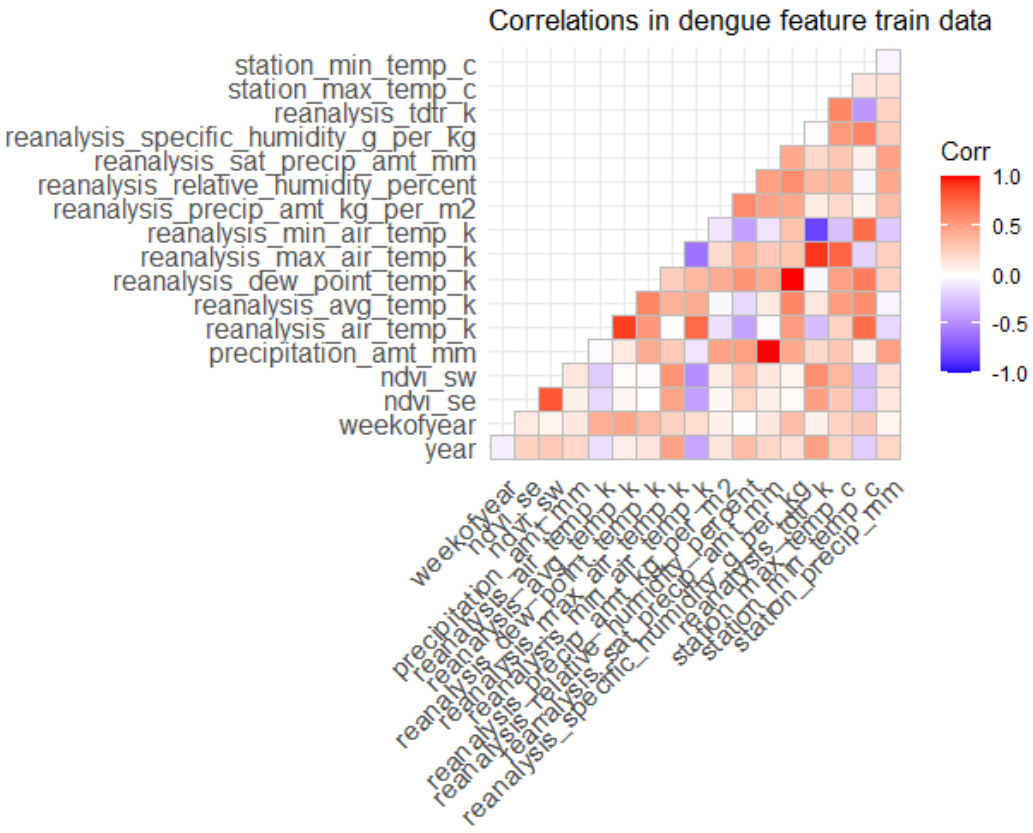
Dalam tahap *pre-processing* data, data akan dilakukan eksplorasi dan pengolahan data. Langkah ini bertujuan untuk mempersiapkan data dalam melakukan model prediksi. Pada Tabel 3.1 terlihat perbedaan rata-rata suatu variabel antara kota San Juan (*sj*) dan kota Iquitos (*iq*). Telihat bahwa nilai rata-rata *precipitation_amt_mm, reanalysis_precip_amt_kg_per_m2, reanalysis_tdtr_k,* dan *station_precip_mm* pada kota San Juan dan Iquitos cukup terpaut jauh.

city	iq	sj
ndvi_se	0,25	0,17709
ndvi_sw	0,26654	0,16609
precipitation_amt_mm	64,2723	35,3853
reanalysis_air_temp_k	297,871	299,157
reanalysis_avg_temp_k	299,134	299,271
reanalysis_dew_point_temp_k	295,498	295,104
reanalysis_max_air_temp_k	307,072	301,392
reanalysis_min_air_temp_k	292,877	297,297
reanalysis_precip_amt_kg_per_m2	57,5891	30,4236
reanalysis_relative_humidity_percent	88,6604	78,5689
reanalysis_sat_precip_amt_mm	64,2723	35,3853
reanalysis_specific_humidity_g_per_kg	17,1018	16,5463
reanalysis_tdtr_k	9,19613	2,51422
station_max_temp_c	33,977	31,5962
station_min_temp_c	21,2057	22,5938
station_precip_mm	62,0125	26,7788

Tabel 3.1: Perbedaan rata-rata nilai variabel bardasarkan kota

Penulis akan meningkatkan akurasi dan presisi dari model prediksi dengan cara membuang variabel-variabel yang berkorelasi (autokorelasi) tinggi dan

mentransformasi data. Berikut adalah *heat map correlation* dari semua variabel yang ada, kecuali variabel *city* dan *week_start_date*.



Gambar 3.3: Nilai korelasi antar variabel

Untuk laporan ini, penulis memilih untuk membuang variabel-variabel yang berkorelasi tinggi, yakni nilai korelasi yang lebih besar sama dengan 0,90. Berdasarkan hasil pengolahan data, ada empat variabel yang berkorelasi tinggi, yaitu *precipitation_amt_mm*, *reanalysis_air_temp_k*, *reanalysis _dew_point_temp_k*, dan *reanalysis_max_air_temp_k*. Empat variabel tersebut akan dihapus dari data. Pada tahap ini tersisa 16 variabel pada data.

Berikutnya data akan ditransformasi dengan menggunakan metode *z-score normalization*. Sebelumnya, variabel yang akan ditransformasi adalah data yang tidak berdistribusi normal. Maka, perlu dilakukan uji normalitas pada setiap

variabel, kecuali *city*, *year*, *weekofyear*, dan *week_start_date*. Uji normalitas yang akan digunakan adalah S*happiro Wilk test* dengan hipotesis berikut.

H_0 : Data berdistribusi normal

H_1 : Data tidak berdistribusi normal

Hasil uji normalitas untuk variabel-variabel tersebut adalah sebagai berikut.

Variabel	*p-value*
Ndvi_se	3.35e-13
Ndvi_sw	< 2.2e-16
reanalysis_avg_temp_k	6.71e-09
reanalysis_min_air_temp_k	< 2.2e-16
reanalysis_precip_amt_kg_per_m2	< 2.2e-16
reanalysis_relative_humidity_percent	< 2.2e-16
reanalysis_sat_precip_amt_mm	< 2.2e-16
reanalysis_specific_humidity_g_per_kg	< 2.2e-16
reanalysis_tdtr_k	< 2.2e-16
station_max_temp_c	1.81e-12
station_min_temp_c	1.81e-12
station_precip_mm	< 2.2e-16

Tabel 3.2: Hasil uji normalitas pada variabel data

Berdasarkan Tabel 3.2, dengan menggunakan $\alpha = 0{,}05$, terlihat bahwa tidak ada *p-value* bernilai lebih dari 0,05. Maka, semua variabel tidak berdistribusi normal sehingga semua variabel akan dinormalisasi. Sebelumnya, variabel-variabel yang dalam satu alat ukur yang sama, seperti suhu/temperatur, akan diubah jadi satuan yang sama, seperti Kelvin diubah menjadi Celcius. Berikut adalah hasil normalisasi data.

	city	year	weekofyear	week_start_date	ndvi_se	ndvi_sw	reanalysis_avg_temp_c
1	sj	1990	18	1990-04-30	0.19848330	0.1776167	24.59286
2	sj	1990	19	1990-05-07	0.16235710	0.1554857	25.29286
3	sj	1990	20	1990-05-14	0.15720000	0.1708429	25.72857
4	sj	1990	21	1990-05-21	0.22755710	0.2358857	26.07857
5	sj	1990	22	1990-05-28	0.25120000	0.2473400	26.51429
6	sj	1990	23	1990-06-04	0.25431430	0.1817429	26.61429
7	sj	1990	24	1990-06-11	0.20507140	0.2102714	26.07143
8	sj	1990	25	1990-06-18	0.15147140	0.1330286	26.37857

Gambar 3.4a: Nilai data sebelum dinormalisasi

	city	year	weekofyear	week_start_date	ndvi_se	ndvi_sw	reanalysis_avg_temp_c
1	sj	1990	18	1990-04-30	-0.06298129	-0.291601161	-1.1733454425
2	sj	1990	19	1990-05-07	-0.55261916	-0.556671857	-0.6179591306
3	sj	1990	20	1990-05-14	-0.62251612	-0.372733323	-0.2722594872
4	sj	1990	21	1990-05-21	0.33107159	0.406306950	0.0054336688
5	sj	1990	22	1990-05-28	0.65151657	0.543499091	0.3511333114
6	sj	1990	23	1990-06-04	0.69372635	-0.242180226	0.4304742132
7	sj	1990	24	1990-06-11	0.02631078	0.099515556	-0.0002335391
8	sj	1990	25	1990-06-18	-0.70015892	-0.825648367	0.2434563740

Gambar 3.4b: Nilai data setelah dinormalisasi

3.4 Proses *Feature Selection*

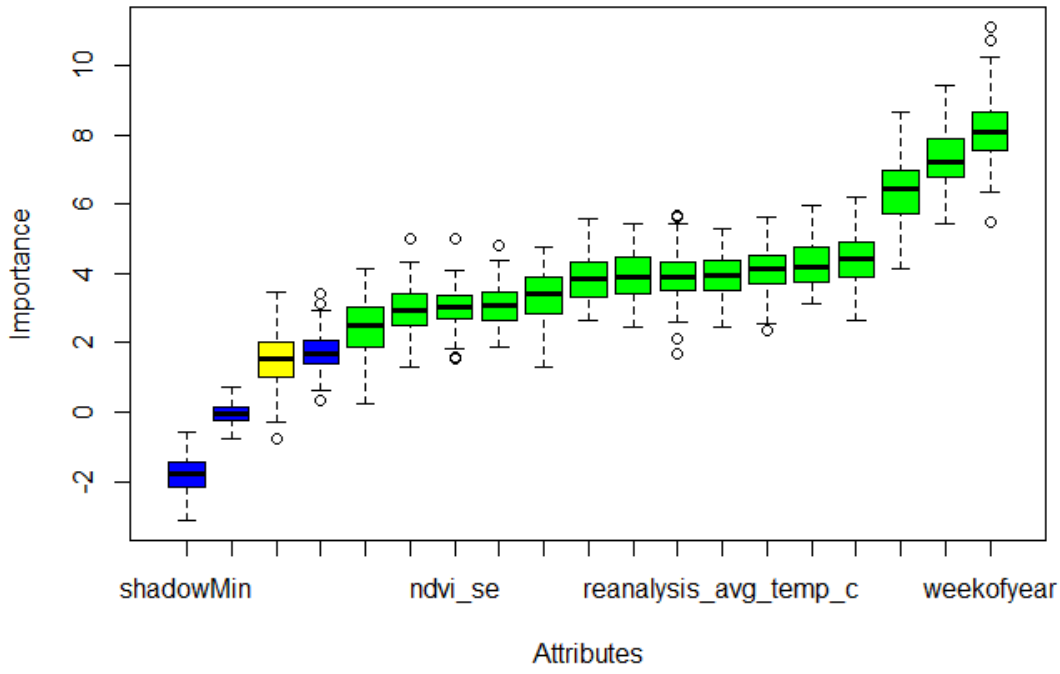
Data set yang diunduh memiliki banyak variabel. Untuk meningkatkan performa model prediksi, maka perlu diketahui variabel-variabel yang tidak berpengaruh terhadap penentuan jumlah kasus demam berdarah. Penulis melakukan dilakukan metode Boruta pada *software Rstudio*.

```
Boruta performed 99 iterations in 17.80431 secs.
 15 attributes confirmed important: city, ndvi_se, ndvi_sw,
reanalysis_avg_temp_c, reanalysis_min_air_temp_c and 10 more;
 No attributes deemed unimportant.
 1 tentative attributes left: reanalysis_sat_precip_amt_mm;
```

Gambar 3.5: Hasil pengolahan data dengan metode Boruta

Gambar 3.5 menunjukkan bahwa metode Boruta melakukan sebanyak 99 iterasi untuk memilih prediktor yang signifikan dalam mempengaruhi peramalan jumlah kasus demam berdarah. Hasil pengolahan data menunjukkan bahwa ada 15 variabel penting dan satu variabel yang tidak penting pada peramalan jumlah kasus demam berdarah, yaitu *reanalysis_sat_precip_amt_mm*. Pada Gambar 3.6, variabel

yang tidak penting merupakan *boxplot* berwarna kuning. Oleh karena itu, *reanalysis_sat_precip_amt_mm* tidak akan digunakan dalam pembuatan model prediksi.



Gambar 3.6: Plot variabel data terhadap tingkat penting

3.5 Model Prediksi dengan *Decision Tree*

Penulis membuat model regresi dengan *decision tree* untuk memprediksi kasus demam berdarah di kota San Juan dan kota Iquitos. Sebelum melakukan prediksi, penulis akan melakukan pembagian data menjadi dua macam data berdasarkan kota, yaitu kota San Juan (936 data) dan kota Iquitos (520 data). Lalu, kedua data akan dibagi menjadi 2 macam data, yaitu data latih (80%) dan data uji (20%). Pembagian data tidak akan dilakukan secara acak, karena data yang disediakan merupakan data *timeseries.*

Penulis akan membangun model prediksi dengan data latih terlebih dahulu dan memprediksi jumlah kasus demam berdarah dari data uji. Pengujian model prediksi akan dilakukan dengan menghitung *Mean Absolute Error* (MAE). Dalam pembuatan model, variabel *year* tidak digunakan karena nilai *year* pada data uji

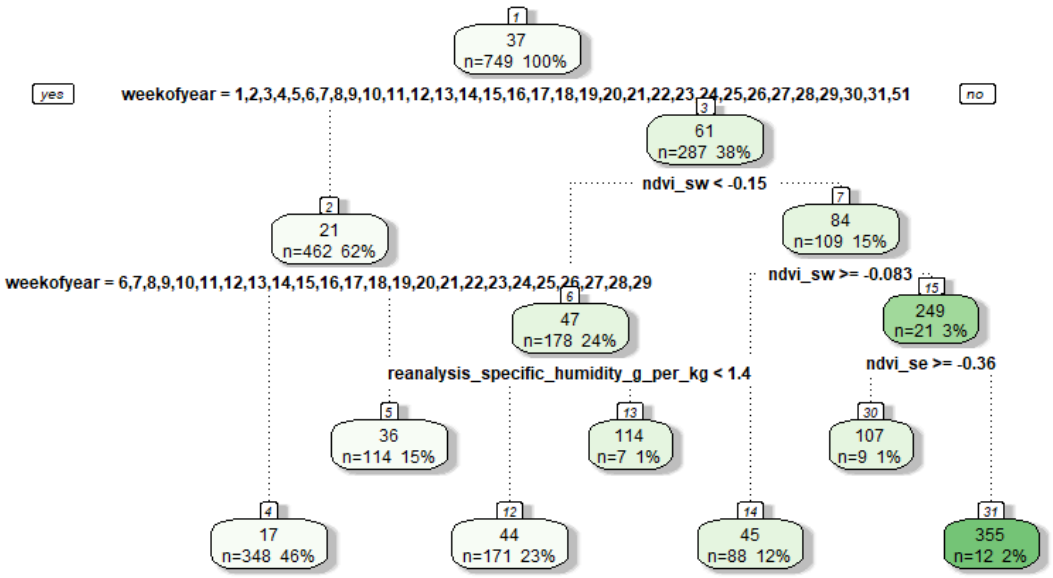
tidak berada pada data latih. Variabel *weekofyear* diubah menjadi variabel *categorical* karena tidak bersifat kontinu. Variabel yang digunakan sebagai *input* adalah:

- *weekofyear*,
- *ndvi_se*,
- *ndvi_sw*,
- *reanalysis_avg_temp_c*,
- *reanalysis_min_air_temp_c*,
- *reanalysis_precip_amt_kg_per_m*,
- *reanalysis_relative_humidity_percent*,
- *reanalysis_specific_humidity_g_per_kg*,
- *reanalysis_tdtr_k*,
- *station_max_temp_c*,
- *station_min_temp_c*,
- dan *station_precip_mm*.

3.5.1 Model Prediksi untuk Kota San Juan

```
n= 749

node), split, n, deviance, yval
      * denotes terminal node

 1) root 749 2265098.00  36.53538
   2) weekofyear=1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,
27,28,29,30,31,51 462  278418.90  21.32251
     4) weekofyear=6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29
 348  108901.30  16.66667 *
     5) weekofyear=1,2,3,4,5,30,31,51 114  138946.40  35.53509 *
   3) weekofyear=32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,52,53 287 17
07641.00  61.02439
     6) ndvi_sw< -0.1452499 178  334994.60  46.75843
      12) reanalysis_specific_humidity_g_per_kg< 1.358375 171  237178.00  44.01170 *
      13) reanalysis_specific_humidity_g_per_kg>=1.358375 7   65010.86 113.85710 *
     7) ndvi_sw>=-0.1452499 109 1277262.00  84.32110
      14) ndvi_sw>=-0.08344263 88  121811.30  45.09091 *
      15) ndvi_sw< -0.08344263 21  452490.30 248.71430
        30) ndvi_se>=-0.3641669 9   81522.00 107.00000 *
        31) ndvi_se< -0.3641669 12   54662.00 355.00000 *
```

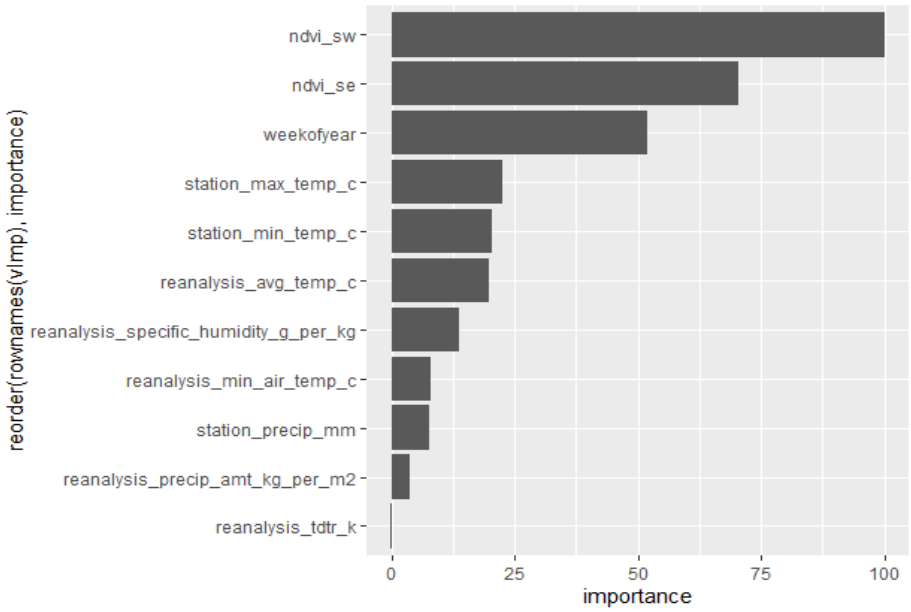
Gambar 3.7: Hasil *decision tree* kota San Juan



Gambar 3.8: Diagram *decision tree* kota San Juan

Dari Gambar 3.7 dan Gambar 3.8 terlihat bahwa *decision tree* kota San Juan menghasilkan 7 *terminal nodes*. *Node* pertama (*weekofyear* = 1, 2, 3, …, 29 ,30 ,31 ,51) menyatakan cabang sebelah kiri *tree* adalah data yang memiliki nilai *weekofyear* berkisar 1-31 dan 51, sedangkan cabang sebelah kanan adalah data-data yang memiliki nilai *weekofyear* 32-50 atau 52. Lalu pada *node* kedua (*weekofyear* = 6, 7, 8, 9, …, 27, 28, 29) menyatakan bahwa cabang sebelah kiri dari *node* tersebut memiliki *weekofyear* 6-29. Model regresi pada *decision tree* akan lebih mudah dibaca dari atas ke bawah [5]. Jika melihat dari *node* nomor 5, maka nilai rata-rata *total_cases* sebesar 36 akan muncul pada data dengan *weekofyear* yang tidak berada di rentang 6-29 namun berada di rentang 1-31 dan 51.

Penulis melakukan prediksi *total_cases* pada data uji kota San Juan yang telah dibuat sebelumnya. Dengan model *decision tree* yang telah dibuat, nilai MAE yang didapatkan adalah 21,92513. Hasil yang didapatkan cukup baik.

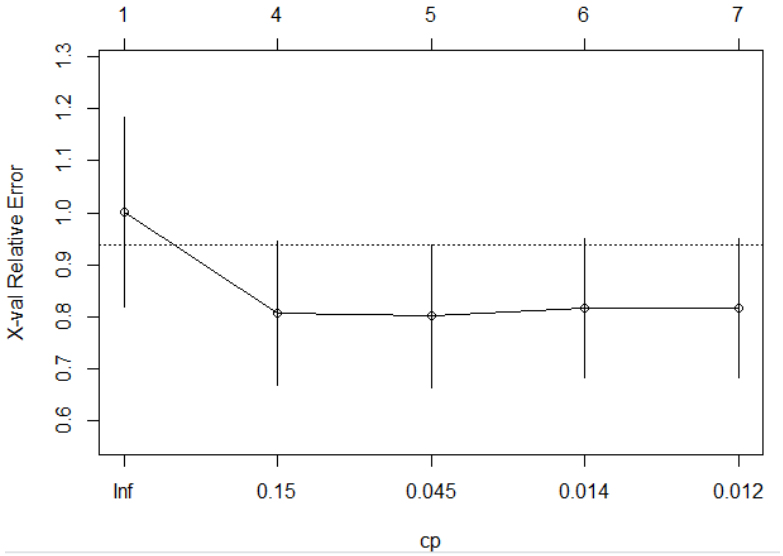


Gambar 3.9: Diagram kepentingan variable kota San Juan

Dari Gambar 3.9 terlihat bahwa indeks vegetasi, urutan minggu dalam tahun, suhu, kelembapan dan presipitasi berpengaruh dalam prediksi kasus demam berdarah di kota San Juan. Tiga variabel yang paling penting adalah *ndvi_sw*, *ndvi_se* dan *weekofyear*.

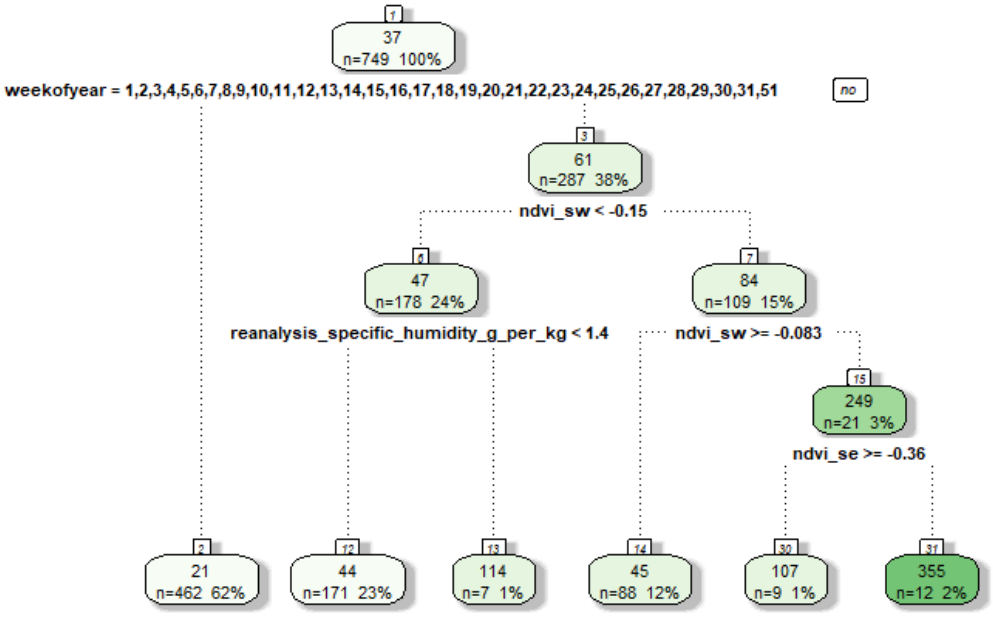
```
Regression tree:
rpart(formula = total_cases ~ ., data = subset(train_data_sj,
    select = -c(city, year, week_start_date)), method = "anova")

Variables actually used in tree construction:
[1] ndvi_se                              ndvi_sw
[3] reanalysis_specific_humidity_g_per_kg weekofyear

Root node error: 2265098/749 = 3024.2

n= 749

        CP nsplit rel error  xerror    xstd
1 0.158548      0   1.00000 1.00091 0.18250
2 0.139644      3   0.52435 0.80798 0.13895
3 0.014483      4   0.38471 0.80195 0.13689
4 0.013497      5   0.37023 0.81727 0.13305
5 0.010000      6   0.35673 0.81719 0.13270
```

Gambar 3.10: Hasil *complexity parameter*



Gambar 3.11: Grafik *complexity parameter*

Untuk meningkatkan performa model, maka perlu dipilih nilai *Complexity Parameter* (CP) dari *tree* terkecil yang memiliki *cross validation error (xerror)* terkecil. Dari Gambar 3.10 terlihat bahwa nilai CP terbaik adalah 0,014483. Nilai CP ini akan digunakan untuk melakukan *pruning.*



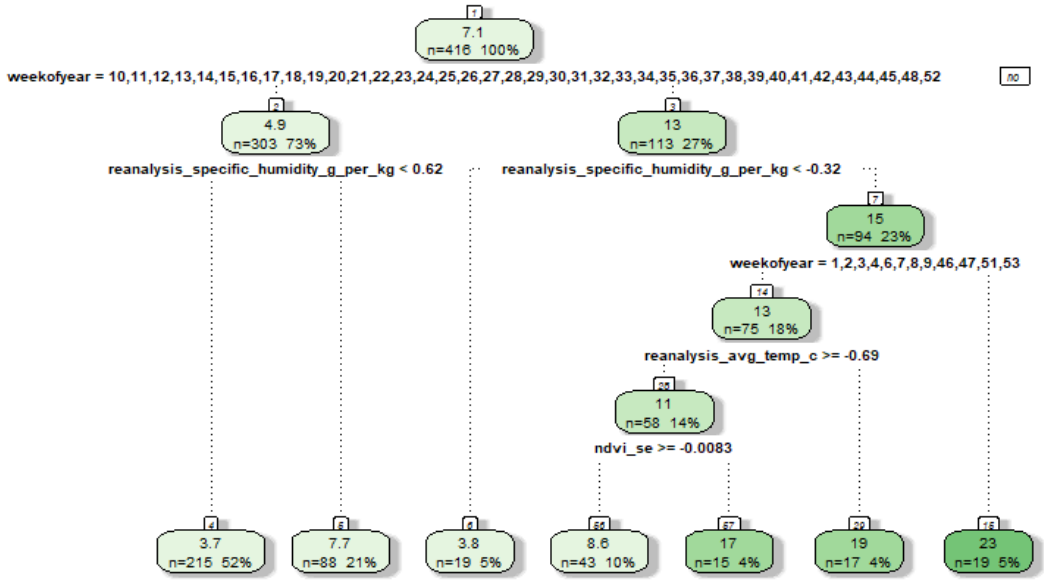
Gambar 3.12: Diagram *decision tree* setelah *pruning* kota San Juan

Pada Gambar 3.12 terlihat bahwa diagram *decision tree* terbaru memiliki lima *terminal nodes*. *Node* nomor 2 menunjukkan bahwa 462 data memiliki rata-rata kasus demam berdarah sebanyak 21. Lalu, 462 data tersebut memiliki nilai *weekofyear* dalam rentang 1-31 atau 51. Dengan model *decision tree* yang telah di-*pruning*, nilai MAE yang didapatkan setelah memprediksi data uji adalah 21,74332, lebih rendah 0,18181 dibandingkan MAE pada *decision tree* tanpa *pruning*.

3.5.2 Model Prediksi untuk Kota Iquitos

```
n= 416

node), split, n, deviance, yval
      * denotes terminal node

 1) root 416 46091.5400  7.076923
   2) weekofyear=10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,
32,33,34,35,36,37,38,39,40,41,42,43,44,45,48,52 303  9154.1850  4.887789
     4) reanalysis_specific_humidity_g_per_kg< 0.6222834 215  4906.9300  3.744186 *
     5) reanalysis_specific_humidity_g_per_kg>=0.6222834 88  3279.0910  7.681818 *
   3) weekofyear=1,2,3,4,5,6,7,8,9,46,47,49,50,51,53 113 31591.6800 12.946900
     6) reanalysis_specific_humidity_g_per_kg< -0.3172993 19   842.5263  3.842105 *
     7) reanalysis_specific_humidity_g_per_kg>=-0.3172993 94 28855.7400 14.787230
      14) weekofyear=1,2,3,4,6,7,8,9,46,47,51,53 75 10351.5500 12.626670
        28) reanalysis_avg_temp_c>=-0.6859656 58  5233.5860 10.724140
          56) ndvi_se>=-0.008346977 43  2576.2790  8.604651 *
          57) ndvi_se< -0.008346977 15  1910.4000 16.800000 *
        29) reanalysis_avg_temp_c< -0.6859656 17  4191.7650 19.117650 *
      15) weekofyear=5,49,50 19 16772.1100 23.315790 *
```

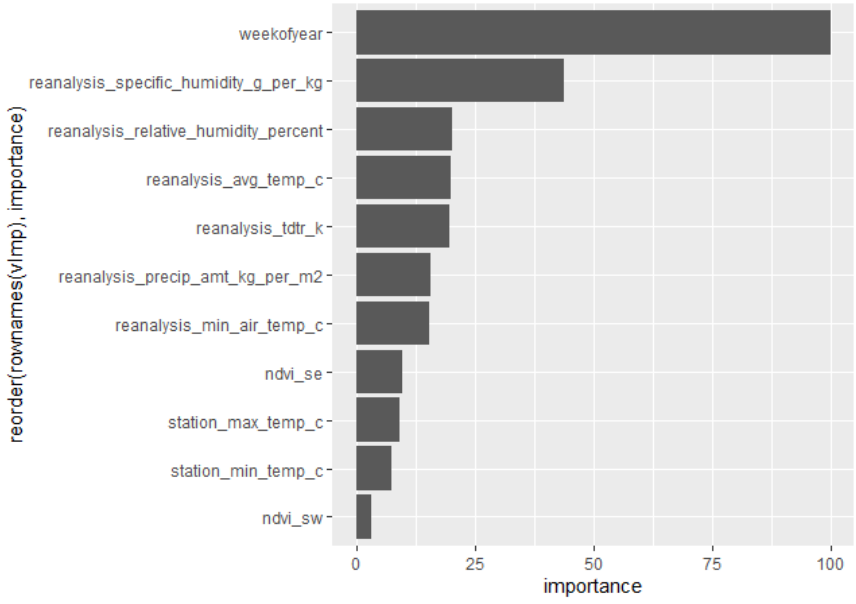
Gambar 3.13: Hasil *Decision Tree* kota Iquitos



Gambar 3.14: Diagram *Decision Tree* kota Iquitos

Dari Gambar 3.13 dan Gambar 3.14 terlihat bahwa *decision tree* kota Iquitos menghasilkan 7 *terminal nodes*. *Node* pertama (*weekofyear* = 10, 11, 12, …, 44, 45, 48, 52) menyatakan cabang sebelah kiri *tree* adalah data yang memiliki nilai *weekofyear* berkisar 10-45, 48, dan 52, sedangkan cabang sebelah kanan adalah data-data yang memiliki nilai *weekofyear* 1-9, 46-47, atau 49-51. Lalu pada *node* kedua (*reanalysis_specific_humidity_g_per_kg* $<$ 0,62) menyatakan bahwa cabang sebelah kiri dari *node* tersebut memiliki *reanalysis_specific_humidity_g_per_kg* yang bernilai lebih kecil dari 0.62. Jika melihat *node* nomor 6, maka nilai rata-rata *total_cases* sebesar 3.8 akan muncul pada data dengan *weekofyear* yang di rentang 1-9, 46-47, atau 49-51 dan memiliki nilai *reanalysis_specific_humidity_g_per_kg* $<$ $-$0,32.

Penulis melakukan pengujian model dengan memprediksi *total_cases* pada data uji kota Iquitos yang telah disiapkan. Hasil prediksi memiliki nilai MAE sebesar 6,817308.



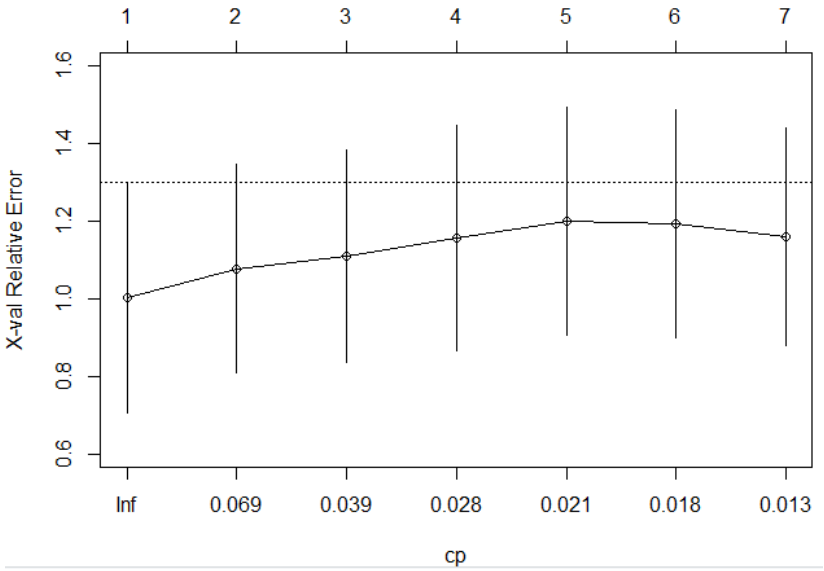
Gambar 3.15: Diagram kepentingan variabel kota Iquitos

Gambar 3.15 menunjukkan bahwa tiga variabel terpenting dalam prediksi kasus demam berdarah adalah *weekofyear, reanalysis_specific_humidity_g_per_kg,* dan *reanalysis_relative_humidity_precent*. Jika dilihat secara keseluruhan, urutan

minggu dalam tahun, kelembapan, suhu, presipitasi, dan indeks vegetasi merupakan kriteria penting dalam prediksi kasus demam berdarah di kota Iquitos

```
Regression tree:
rpart(formula = total_cases ~ ., data = subset(train_data_iq,
    select = -c(city, year, week_start_date)), method = "anova")

Variables actually used in tree construction:
[1] ndvi_se                               reanalysis_avg_temp_c
[3] reanalysis_specific_humidity_g_per_kg weekofyear

Root node error: 46092/416 = 110.8

n= 416

        CP nsplit rel error xerror    xstd
1 0.115979      0   1.00000 1.0035 0.29704
2 0.041079      1   0.88402 1.0778 0.26952
3 0.037579      2   0.84294 1.1107 0.27377
4 0.021005      3   0.80536 1.1574 0.29148
5 0.020095      4   0.78436 1.1998 0.29467
6 0.016205      5   0.76426 1.1927 0.29425
7 0.010000      6   0.74806 1.1605 0.28119
```

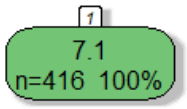
Gambar 3.16: Hasil *complexity parameter* kota Iquitos



Gambar 3.17: Grafik *complexity parameter* kota Iquitos

Untuk meningkatkan performa model, maka perlu dilakukan *pruning* pada model yang telah dibuat. Oleh karena itu, perlu dipilih nilai CP terbaik, yaitu CP dari *tree* terkecil yang memiliki *xerror* terkecil. Dari Gambar 3.16 terlihat bahwa

nilai CP terbaik adalah 0,115979 dengan *nsplit* bernilai nol. Hal ini berarti bahwa *tree* yang optimal hanya terdiri dari *root node* saja.



Gambar 3.18: Diagram *decision tree* setelah *pruning* kota Iquitos

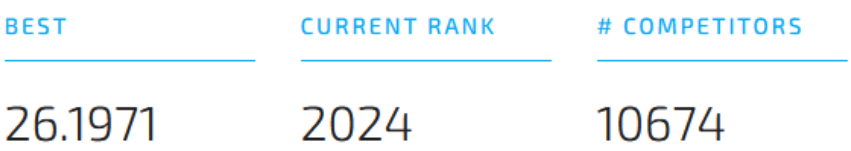
Gambar 3.18 menunjukkan bahwa rata-rata jumlah kasus demam berdarah untuk kota Iquitos adalah 7,1. Nilai MAE yang didapatkan dari model ini adalah 7,269231, lebih tinggi dibandingkan MAE pada *decision tree* sebelum *pruning.* Model *decision tree* setelah *pruning* menunjukkan adanya *overfitting* pada data latih.

3.6 Prediksi

Penulis menggunakan model regresi *decision tree* yang belum di-*pruning* dan yang sudah di-*pruning* untuk melakukan prediksi kasus demam berdarah pada data uji yang disediakan oleh *DengAI.* Tabel 3.3 menunjukkan nilai MAE yang didapatkan dengan menggunakan dua macam kombinasi model. Dari tiga percobaan yang dilakukan, didapatkan nilai MAE terendah sebesar 26,1971. Hal ini menandakan bahwa model regresi *decision tree* dapat memprediksi jumlah kasus demam berdarah dengan baik.

No.	Model prediksi kota San Juan	Model prediksi kota Iquitos	Hasil MAE
1.	*Decision tree* dengan *pruning*	*Decision tree* tanpa *pruning*	26,2260
2.	*Decision tree* dengan *pruning*	*Decision tree* dengan *pruning*	26,4159
3.	*Decision tree* tanpa *pruning*	*Decision tree* tanpa *pruning*	26,1971

Tabel 3.3: Hasil prediksi menggunakan model *decision tree*



Gambar 3.19: Nilai MAE dan peringkat penulis untuk model prediksi terbaik.

BAB IV

PENUTUP

5.1 Kesimpulan

Dari penelitian yang dilakukan, model regresi *decision tree* tanpa *pruning* adalah model terbaik untuk memprediksi kasus demam berdarah di kota San Juan dan kota Iquitos. Nilai MAE yang didapatkan penulis adalah 26,1971. Tiga variabel yang paling penting pada prediksi kasus demam berdarah di kota San Juan adalah *ndvi_sw, ndvi_se,* dan *weekofyear*. Hal ini menandakan bahwa peran nilai indeks vegetasi dan urutan minggu pada tahun memiliki dampak yang cukup besar dalam munculnya kasus demam berdarah. Pada kota Iquitos, tiga variabel yang paling berpengaruh adalah *weekofyear*, *reanalysis_specific_humidity_g_per_kg,* dan *reanalysis_relative_humidity_precent,* yang menandakan bahwa urutan minggu pada tahun dan kelembapan berpengaruh dalam mnculnya kasus demam berdarah.

5.2 Saran

Saran yang penulis berikan adalah sebagai berikut:

- Penelitian selanjutnya dapat menggunakan metode *pruning* lain terhadap *decision tree*¸ seperti *reduced error pruning*.
- Untuk mendapatan hasil yang lebih akurat, penelitian selanjutnya dapat melakukan pemodelan prediksi dengan menggunakan model regresi lain, seperti *Negative Binomial*, *Random Forest*, atau *XGBoost*.
- Bagi pihak pemerintah setempat dapat menggunakan hasil prediksi sebagai acuan untuk mempersiapkan diri dan merancang strategi penanganan pada saat minggu dimana kasus demam berdarah meningkat.

DAFTAR PUSTAKA

[1] *National Centers for Environmental Information. Global Historical Climatology Network (GHCN)*. Dapat diakses di https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/global-historical-climatology-network-ghcn. [Diakses pada 15 Juni 2021].

[2] NCAR. *PERSIAN-CDR: PRECIPITATION ESTIMATION FROM REMOTELY SENSED INFORMATION USING ARTIFICIAL NEURAL NETWORKS – CLIMATE DATA RECORD.* Dapat diakses di https://climatedataguide.ucar.edu/climate-data/persiann-cdr-precipitation-estimation-remotely-sensed-information-using-artificial. [Diakses pada 15 Juni 2021].

[3] *National Centers for Environmental Information*. *Climate Forecast System* (CFS). Dapat diakses di https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/climate-forecast-system-version2-cfsv2. [Diakses pada 15 Juni 2021].

[4] *National Centers for Environmental Information*. *Normalized Difference Vegetation Index.* Dapat diakses di https://www.ncdc.noaa.gov/cdr/terrestrial/normalized-difference-vegetation-index. [Diakses pada 15 Juni 2021].

[5] Wolfgang Jank. *Business Analytics for Managers.* Springer EBooks. 2011