LAPORAN ANALISIS

MEMPREDIKSI WABAH DEMAM BERDARAH DI SAN JUAN DAN IQUITOS DENGAN METODE REGRESI BINOMIAL NEGATIF

Diajukan untuk Memenuhi Tugas Mata Kuliah *Capstone*

Oleh:

Nama:	**NIM:**
Amadea Franstella Tanugerah	01112180014
Eric Jahja	01112180012
Gabrielle Priscilla	01112180039
Veren Christi	01112180044

PROGRAM STUDI MATEMATIKA TERAPAN

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS PELITA HARAPAN

TANGERANG

2021

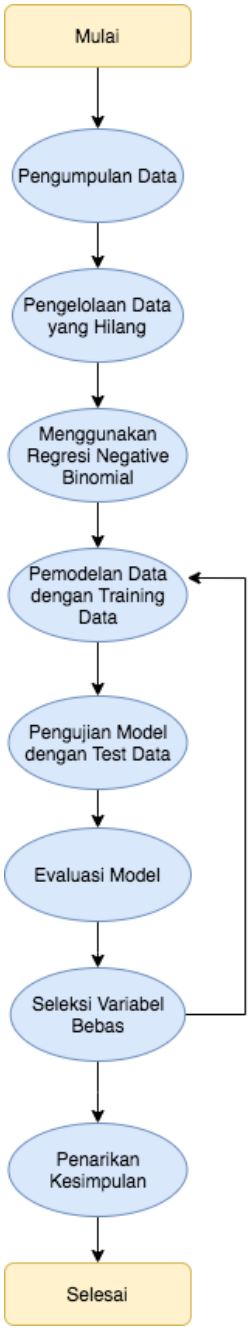
EXECUTIVE SUMMARY

Laporan ini membahas metodologi serta analisis hasil dari prediksi total kasus Demam Berdarah di kota San Juan dan Iquitos. Metode yang digunakan untuk mengolah data yang hilang adalah menggunakan data pada index waktu sebelumnya dan prediksi total kasus dilakukan dengan metode regresi Binomial Negatif. Semua pengolahan data dan regresi dilakukan dengan bahasa pemograman R. Hasil dari prediksi yang didapat memiliki *Mean Absolute Error* sebesar 26.1346.

Dalam laporan ini ditemukan bahwa rata-rata kelembaban spesifik, rata-rata *dew point temperature*, rata-rata temperatur, dan rata-rata temperatur minimum dapat memberikan *Mean Absolute Error* yang cukup rendah jika digunakan untuk memprediksi total kasus, dibandingkan dengan kombinasi variabel bebas lainnya. Selain itu pengisian data yang hilang dengan metode data pada index waktu sebelumnya memberikan dampak yang positif pada data dengan *seasonality*. Saran yang diberikan antara lain: pengolahan data yang hilang dan prediksi total kasus dapat dilakukan dengan metode lain dan; menggunakan metode seleksi variabel bebas yang lebih banyak, sehingga dapat meningkatkan kemungkinan perbaikan model.

1. METODOLOGI

Pada bagian ini akan dibahas langkah-langkah memprediksi total kasus Demam Berdarah (DB) di San Juan dan Iquitos yang terdiri dari metode pengolahan data yang hilang dan pemilihan model prediksi yang digunakan. Langkah-langkah yang akan dilakukan dalam laporan ini dapat dilihat pada Gambar 1.1.



Gambar 1.1: Flowchart Langkah-Langkah Pengerjaan

1.1 Pengumpulan Data

Data yang digunakan dalam laporan ini diunduh dari situs *DrivenData*. Set-set data yang ada meliputi *training* data yang dapat digunakan untuk pengembangan model dan *test* data yang akan digunakan untuk menguji seberapa akurat model yang didapatkan.

1.2 Pengolahan Data yang Hilang

Hal penting selanjutnya adalah menangani data yang hilang atau tidak ada. Data yang hilang dapat terjadi karena banyak faktor dan merupakan bagian yang tidak dapat dihindari. Pada laporan ini, data yang hilang akan diolah dengan metode *last observation carried forward* (*LOCF*). Dalam metode ini, data-data yang hilang akan digantikan dengan data yang sama dengan data pada indeks waktu sebelumnya. Hasil akhir dari perlakuan adalah data yang lengkap tanpa pengurangan data.

1.3 Pembuatan Model Regresi Binomial Negatif

Model regresi Binomial Negatif digunakan pada data ini karena model cocok untuk data hitung overdispersi, yaitu jenis data bilangan bulat non-negatif yang memiliki variansi lebih besar daripada rata-rata. *Training* data (Data Pelatih) akan digunakan untuk memodelkan regresi dan selanjutnya nilai-nilai variabel bebas *test* data (Data Tes) akan dimasukkan ke dalam model. Hasil akhir dari regresi Binomial Negatif adalah prediksi nilai total kasus DB dari *test* data. Setiap langkah *training* dan *testing* model akan dikerjakan menggunakan program R.

1.4 Seleksi Variabel

Seleksi variabel-variabel bebas merupakan salah satu cara untuk mendapatkan model yang bagus. Variabel bebas yang diinginkan adalah variabel-variabel yang berpengaruh secara signifikan dan tidak saling tumpang tindih. Seleksi dilakukan dengan cara mengikuti sugesti *benchmark* pada situs *DrivenData*.

1.5 Pengujian dan Evaluasi Model

Setelah mendapatkan model dari data pelatih, model akan diuji dengan data tes. Setiap *Mean Absolute Error* (MAE) dari model yang didapatkan dari membandingkan nilai prediksi dan nilai asli akan dicatat dan saling dibandingkan untuk evaluasi model terbaik. Rumus MAE sendiri adalah

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n},$$

di mana $y_i, x_i,$ dan n adalah nilai prediksi, nilai asli, dan total data secara berturut-turut.

1.6 Penarikan Kesimpulan

Hasil evaluasi model-model dapat digunakan untuk menarik kesimpulan pada laporan ini. Pada akhirnya akan dipilih model yang terbaik berdasarkan nilai evaluasi MAE yang terendah.

2. HASIL DAN ANALISIS

Pada bagian ini dijelaskan hasil dari penerapan data terhadap metode regresi Binomial Negatif serta analisisnya. Dimulai dengan pendeskripsian data-data yang digunakan pada laporan ini, dan dilanjutkan dengan hasil evaluasi metode pada model.

2.1 Data

Pada laporan ini digunakan dua set data yang diambil dari situs *drivendata.org* bagian kompetisi *DengAI: Predicting Disease Spread* di mana para kompetitor saling berusaha mendapatkan model yang terbaik untuk memprediksi total kasus DB di San Juan dan Iquitos. Rangkuman penjelasan data-data tersebut dijelaskan pada Tabel 2.1.

Tabel 2.1: Rangkuman Data Penelitian

No.	Nama Data	Jumlah Data	Jumlah Atribut
1.	Data Pelatih	1,456	25
2.	Data Tes	419	24

Terdapat beberapa atribut yang sama di dalam data pelatih dan data tes yang diperoleh dari stasiun-stasiun yang berbeda. Hal ini menyebabkan tumpang tindih pada data dan tidak baik jika dibiarkan untuk memprediksi menggunakan metode regresi.

2.1.1 Data Pelatih

Data Pelatih merupakan data jumlah total kasus DB per minggu di San Juan dan Iquitos. Set data ini berjumlah 1,456 buah dan memiliki 25 atribut yang dijelaskan pada Tabel 2.2.

Tabel 2.2: Penjelasan Atribut Data Pelatih

No.	Nama Atribut	Jenis Data	Rentang Nilai	Tipe Atribut
1.	Kota	Kategori	Sj, Iq	Lokasi
2.	Tahun	Numerik	[1990 – 2010]	Indeks Waktu
3.	Minggu Pada Tahun	Numerik	[1 – 53]	Indeks Waktu
4.	Tanggal Minggu Dimulai	Kategori	–	Indeks Waktu

5.	ndvi_ne	Numerik	[-0.406250 – 0.508357]	Indeks Vegetasi
6.	ndvi_nw	Numerik	[-0.456100 – 0.454429]	Indeks Vegetasi
7.	ndvi_se	Numerik	[-0.015533 – 0.538314]	Indeks Vegetasi
8.	ndvi_sw	Numerik	[-0.063457 – 0.546017]	Indeks Vegetasi
9.	precipitation_amt_mm	Numerik	[0 – 390.6]	Presipitasi
10.	reanalysis_air_temp_k	Numerik	[294.635714 – 302.2]	Suhu
11.	reanalysis_avg_temp_k	Numerik	[294.892857 – 302.928571]	Suhu
12.	reanalysis_dew_point_temp_k	Numerik	[289.642857 – 298.45]	Suhu
13.	reanalysis_max_air_temp_k	Numerik	[297.8 – 314]	Suhu
14.	reanalysis_min_air_temp_k	Numerik	[286.9 – 299.9]	Suhu
15.	reanalysis_precip_amt_kg_per_m2	Numerik	[0 – 570.5]	Presipitasi
16.	reanalysis_relative_humidity_percent	Numerik	[57.787143 – 98.61]	Kelembapan
17.	reanalysis_sat_precip_amt_mm	Numerik	[0 – 390.6]	Presipitasi
18.	reanalysis_specific_humidity_g_per_k	Numerik	[11.715714 – 20.461429]	Kelembapan
19.	reanalysis_tdtr_k	Numerik	[1.357143 – 16.028571]	Suhu
20.	station_avg_temp_c	Numerik	[21.4 – 30.8]	Suhu
21.	station_diur_temp_rng_c	Numerik	[4.528571 – 15.8]	Suhu
22.	station_max_temp_c	Numerik	[26.7 – 42.2]	Suhu
23.	station_min_temp_c	Numerik	[14.7 – 25.6]	Suhu
24.	station_precip_mm	Numerik	[0 – 543.3]	Presipitasi
25.	Total Kasus	Numerik	[0 – 461]	Jumlah Kasus

2.1.2 Data Tes

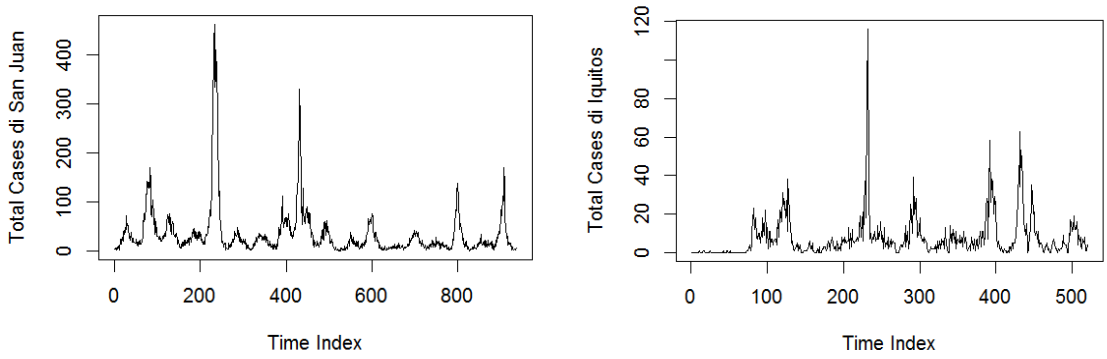
Data Tes merupakan data jumlah total kasus DB per minggu di San Juan dan Iquitos. Set data ini berjumlah 419 buah dan memiliki 24 atribut yang dijelaskan pada Tabel 2.3.

Tabel 2.3: Penjelasan Atribut Data Tes

No.	Nama Atribut	Jenis Data	Rentang Nilai	Tipe Atribut
1.	Kota	Kategori	Sj, Iq	Lokasi
2.	Tahun	Numerik	[2008 - 2013]	Indeks Waktu
3.	Minggu Pada Tahun	Numerik	[1 - 53]	Indeks Waktu
4.	Tanggal Minggu Dimulai	Kategori	–	Indeks Waktu
5.	ndvi_ne	Numerik	[-0.4643 – 0.5004]	Indeks Vegetasi
6.	ndvi_nw	Numerik	[-0.2118 – 0.649]	Indeks Vegetasi
7.	ndvi_se	Numerik	[0.0062 – 0.453043]	Indeks Vegetasi
8.	ndvi_sw	Numerik	[-0.014671 – 0.529043]	Indeks Vegetasi
9.	precipitation_amt_mm	Numerik	[0 – 169.34]	Presipitasi
10.	reanalysis_air_temp_k	Numerik	[294.554286 – 301.935714]	Suhu
11.	reanalysis_avg_temp_k	Numerik	[295.235714 – 303.328571]	Suhu
12.	reanalysis_dew_point_temp_k	Numerik	[290.818571 – 297.794286]	Suhu
13.	reanalysis_max_air_temp_k	Numerik	[298.2 – 314.1]	Suhu
14.	reanalysis_min_air_temp_k	Numerik	[286.2 – 299.7]	Suhu
15.	reanalysis_precip_amt_kg_per_m2	Numerik	[0 – 301.4]	Presipitasi
16.	reanalysis_relative_humidity_percent	Numerik	[64.92 – 97.982857]	Kelembapan
17.	reanalysis_sat_precip_amt_mm	Numerik	[0 – 169.34]	Presipitasi
18.	reanalysis_specific_humidity_g_per_kg	Numerik	[12.537143 – 19.598571]	Kelembapan
19.	reanalysis_tdtr_k	Numerik	[1.485714 – 14.485714]	Suhu
20.	station_avg_temp_c	Numerik	[24.157143 – 30.271429]	Suhu
21.	station_diur_temp_rng_c	Numerik	[4.042857 – 14.725]	Suhu
22.	station_max_temp_c	Numerik	[27.2 – 38.4]	Suhu
23.	station_min_temp_c	Numerik	[14.2 – 26.7]	Suhu
24.	station_precip_mm	Numerik	[0 – 212]	Presipitasi

2.2 Hasil Analisis Model Regresi Binomial Negatif

Pada bagian ini diuraikan hasil analisis dari prediksi total kasus demam berdarah (DB) di San Juan dan Iquitos dengan *Generalized Linear Model* (GLM) menggunakan Binomial Negatif sebagai *link function* dengan bantuan bahasa pemrograman R. Langkah pertama yang dilakukan untuk mengolah Data Pelatih dan Data Tes adalah mengolah data-data yang hilang. Metode pengolahan yang digunakan adalah dengan mengisi data yang hilang dengan data pada index waktu sebelumnya. Metode ini dipilih karena seperti yang terlihat pada Gambar 2.1, terdapat *seasonality* pada jumlah kasus yang terjadi dan data yang terkumpul diambil secara periodik. Maka, nilai data yang hilang cenderung tidak jauh berbeda dengan nilai data pada index waktu sebelumnya.



Gambar 2.1: Plot dari variabel *total cases* untuk kota San Juan dan Iquitos

Langkah berikutnya adalah memodelkan Data Pelatih dengan GLM menggunakan *link function* binomial negatif. *Link function* dipilih agar dapat mewakili variabel total kasus yang berdistiribusi binomial negatif serta memiliki variabilitas yang besar. Dari 24 variabel bebas hanya empat variabel bebas yang terpakai, di antaranya yaitu, rata-rata kelembapan spesifik, rata-rata suhu pengembunan, rata-rata suhu, dan suhu terendah. Keempat variabel ini dipilih berdasarkan informasi yang didapat dari *Centers for Disease Control and Prevention*, yang memaparkan bahwa siklus hidup nyamuk yang adalah pembawa virus DB erat kaitannya dengan suhu dan kelembapan. Dibentuk dua model – dengan penambahan faktor *exposure* – untuk mewakili dua kota di mana data diambil, yaitu San Juan dan Iquitos.

Berdasarkan hasil dari pemrograman R yang dapat dilihat pada Gambar 2.2, model untuk kota San Juan adalah

$$\begin{aligned} log\,(\mu_i) = 250.537 &+ 1.09636\, reanalysis\_specific\_humidity\_g\_per\_kg \\ &- 0.90001\, reanalysis\_dew\_point\_temp\_k \\ &+ 0.12338\, station\_avg\_temp\_c \\ &- 0.08584\, station\_min\_temp\_c\,, \end{aligned} \tag{2.1}$$

dengan asumsi bahwa variabel lain dianggap konstan, setiap peningkatan satu g/kg kelembapan di kota San Juan, total kasus akan meningakat 2.993 kali lipat. Untuk peningkatan satu derajat Kelvin rata-rata *dew point temperature*, jumlah kasus DB akan menjadi 0.4066 kali lebih banyak. Jika rata-rata suhu di San Juan meningkat satu derjat celcius, akan terjadi peningkatan jumlah kasus sebanyak 1.131 kali lipat. Satu derajat celcius peningkatan suhu terendah akan mengakibatkan total kasus menjadi 0.9177 kali lipat.

```
Call:
glm.nb(formula = form, data = combined, init.theta = 1.058685376,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7293  -1.0277  -0.4437   0.1456   4.4225

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     249.53700  106.59636   2.341  0.01923 *
reanalysis_specific_humidity_g_per_kg   1.09636    0.38543   2.845  0.00445 **
reanalysis_dew_point_temp_k            -0.90001    0.38276  -2.351  0.01870 *
station_avg_temp_c                      0.12338    0.05873   2.101  0.03566 *
station_min_temp_c                     -0.08584    0.05172  -1.660  0.09700 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.0587) family taken to be 1)

    Null deviance: 1165.6  on 935  degrees of freedom
Residual deviance: 1051.6  on 931  degrees of freedom
AIC: 8412.7

Number of Fisher Scoring iterations: 1


              Theta:  1.0587
          Std. Err.:  0.0458

 2 x log-likelihood:  -8400.7140
```

Gambar 2.2: Kesimpulan Model Data Pelatih untuk Kota San Juan

Berdasarkan hasil dari pemrograman R yang dapat dilihat pada Gambar 2.3, model untuk kota Iquitos adalah

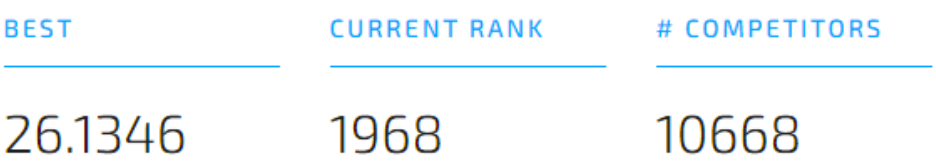
$$
\begin{aligned}
log\ (\mu_i\ ) = 498.70868 \\
&+ 1.96655\ reanalysis\_specific\_humidity\_g\_per\_kg \\
&- 1.80251\ reanalysis\_dew\_point\_temp\_k \\
&- 0.01955\ station\_avg\_temp\_c \\
&+ 0.17728\ station\_min\_temp\_c,
\end{aligned}
\tag{2.2}
$$

dengan asumsi bahwa variabel lain dianggap konstan, setiap peningkatan satu g/kg kelembapan di kota San Juan, total kasus akan meningakat 7.146 kali lipat. Untuk peningkatan satu derajat Kelvin rata-rata suhu kelembapan, jumlah kasus DB akan menjadi 0.16488 kali lipat. Jika rata-rata suhu di San Juan meningkat satu derjat celcius, akan terjadi peningkatan jumlah kasus sebanyak 0.9806 kali lipat. Satu derajat celcius peningkatan suhu terendah akan mengakibatkan total kasus menjadi 1.194 kali lipat.

```
Call:
glm.nb(formula = form, data = combined, init.theta = 0.7722358678,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2135  -1.1421  -0.3850   0.2827   3.3419

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                     497.70868  123.09516   4.043 5.27e-05 ***
reanalysis_specific_humidity_g_per_kg   1.96655    0.43325   4.539 5.65e-06 ***
reanalysis_dew_point_temp_k            -1.80251    0.44060  -4.091 4.30e-05 ***
station_avg_temp_c                     -0.01955    0.05251  -0.372     0.71
station_min_temp_c                      0.17728    0.04364   4.062 4.86e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.7722) family taken to be 1)

    Null deviance: 1176.4  on 919  degrees of freedom
Residual deviance: 1053.7  on 915  degrees of freedom
AIC: 5485.8

Number of Fisher Scoring iterations: 1


              Theta:  0.7722
          Std. Err.:  0.0422

 2 x log-likelihood:  -5473.8090
```

Gambar 2.3: Kesimpulan Model Data Pelatih untuk Kota Iquitos

Selanjutnya, Data Tes dimasukkan ke dalam persamaan (2.1) dan (2.2) untuk mendapatkan prediksi total kasus DB berdasarkan kotanya. Dapat dilihat pada Gambar 2.4, nilai MAE terendah yang didapatkan adalah 26.1346, dengan perolehan peringkat 1,968 dari 10,668 partisipan yang mengikuti kompetisi.



Gambar 2.4: Hasil MAE (bagian kiri) dan Peringkat (bagian tengah)

3. PENUTUP

3.1 Kesimpulan

Dalam laporan ini telah dilakukan regresi Binomial Negatif untuk memprediksi jumlah total kasus DB di San Juan dan Iquitos yang diukur berdasarkan nilai MAE. Berdasarkan analisis data yang telah dilakukan, dapat ditarik beberapa kesimpulan sebagai berikut.

1. Pengisian data yang hilang dengan data sebelum memberikan dampak yang positif pada pengolahan data yang memiliki *seasonality*.
2. Seleksi variabel bebas memberikan dampak yang cukup positif pada pengolahan data menggunakan regresi Binomial Negatif.
3. Hasil prediksi yang didapatkan memiliki MAE sebesar 26.1346 dan memberikan perolehan peringkat 1,968 dari 10,668 partisipan yang mengikuti kompetisi.

3.2 Saran

Metode yang digunakan dalam laporan ini masih dapat dikembangkan agar mendapatkan hasil yang lebih akurat dan memberikan gambaran yang lebih jelas mengenai variabel-variabel bebas yang menentukan prediksi total kasus. Terkait dengan hal – hal di atas, penulis menyarankan beberapa hal yang dapat diperhatikan seperti berikut.

1. Mengeksplorasi metode lain untuk mengatasi data hilang seperti *Expectation Maximization* (EM). Ini dikarenakan metode yang digunakan pada laporan ini bukan metode yang paling efektif.
2. Menggunakan metode lain untuk memprediksi kasus DB seperti *Time Series*.
3. Menggunakan metode seleksi variabel bebas yang lebih banyak, sehingga meningkatkan kemungkinan perbaikan model.

REFERENSI

[1] Kang H. (2013). *The prevention and handling of the missing data*. Korean journal of anesthesiology, *64*(5), 402–406. https://doi.org/10.4097/kjae.2013.64.5.402.

[2] "Dengue." *Centers for Disease Control and Prevention*, National Center for Emerging and Zoonotic Infectious Diseases (NCEZID), Division of Vector-Borne Diseases (DVBD), 14 Juli 2020, https://www.cdc.gov/dengue/.