LAPORAN

ANALISIS PREDIKSI JUMLAH KASUS DEMAM BERDARAH *DENGUE* DI KOTA SAN JUAN DAN IQUITOS DENGAN METODE *GENERALIZED LINEAR MODELS*

Ditulis untuk memenuhi tugas analisa data
dan laporan dari mata kuliah Kapita Selekta

Oleh:

Monica Angelina Aprilia	**01112180042**
Vanessa Laurencia	**01112180026**
Veronica Cynthia	**01112180005**

PROGRAM STUDI MATEMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS PELITA HARAPAN
TANGERANG

2021

EXECUTIVE SUMMARY

Demam berdarah dengue (DBD) merupakan penyakit yang ditularkan oleh nyamuk *Aedes aegypti*. Nyamuk *Aedes aegypti* akan lebih mudah berkembang biak di lingkungan yang lembab dan memiliki suhu yang rendah. Oleh karena itu, pada dasarnya suatu daerah yang memiliki tingkat kelembapan yang tinggi dan suhu yang rendah memiliki jumlah kasus DBD lebih yang tinggi dibandingkan daerah lainnya.

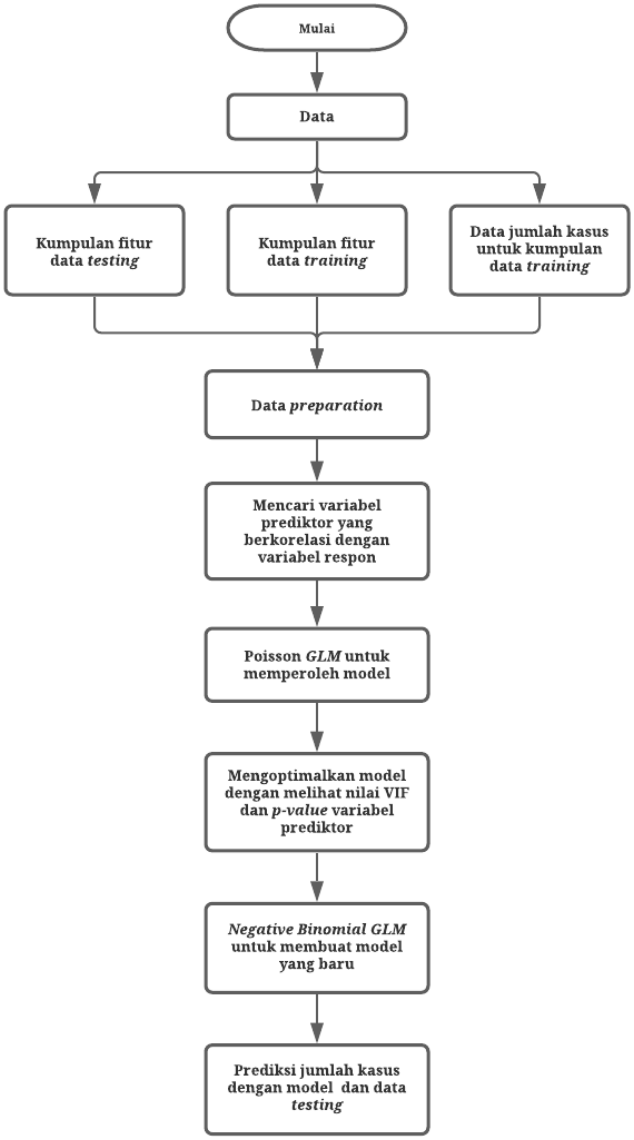
Dalam penelitian ini, akan digunakan data yang telah disediakan oleh situs *drivendata.org* pada salah satu perlombaan yang diselenggarakan dengan nama *DengAI: Predicting Disease Spread.* Perlombaan ini diselenggarakan untuk memprediksi jumlah kasus DBD pada kota San Juan dan Iquitos. Dalam data yang disediakan masih ada beberapa data yang kosong, sehingga perlu dilakukan proses *imputation* dengan *Multivariate Imputation by Chained Equations* (MICE). Beberapa metode yang digunakan dalam penelitian ini adalah regresi Poisson dan regresi binomial negatif. Kompetisi ini menggunakan metrik *Mean Absolute Error* (MAE) sehingga prediksi yang dilakukan bertujuan untuk meminimalkan nilai MAE.

Penelitian ini menggunakan 10 variabel pada masing-masing kota yang memiliki koefisien korelasi tertinggi dengan variabel *total_cases.* Variabel tersebut digunakan untuk membuat model yang menggunakan regresi Poisson, namun model yang dihasilkan mengalami *overdispersion.* Untuk mengoreksi *overdispersion* pada model tersebut, digunakan metode regresi binomial negatif. Hasil akhir dari pengembangan model pada penelitian ini mendapatkan nilai MAE sebesar 26,9159 dengan menempati posisi rank 2.631.

BAB I

METODOLOGI PENELITIAN

Pada bab ini akan dijelaskan data dan langkah-langkah yang digunakan dalam penelitian. Langkah-langkah penelitian dapat dilihat melalui Gambar 1.1.



Gambar 1.1: Diagram Alur Penelitian

1.1. Data

Dalam penelitian ini, akan digunakan data yang telah disediakan oleh situs *drivendata.org* pada salah satu perlombaan yang diselenggarakan dengan nama *DengAI: Predicting Disease Spread.* Perlombaan ini diselenggarakan untuk memprediksi jumlah kasus DBD pada kota San Juan dan Iquitos berdasarkan data yang disediakan. Data yang tersedia adalah *Training Data Features* (fitur data *training*), *Training Data Labels* (label data *training*), dan *Test Data Features* (fitur data *test*). Kurun waktu untuk data *training* pada kota San Juan adalah 18 tahun (1990 – 2008) dan pada kota Iquitos adalah 10 tahun (2000 – 2010). Sedangkan untuk data *testing*, kurun waktu pada kota San Juan adalah lima tahun (2008 – 2013) dan pada kota Iquitos adalah tiga tahun (2010 – 2013). Data terbagi menjadi dua bagian yaitu, fitur dan label. Penjelasan dari fitur dan label data dapat dilihat pada Tabel 1.1 dan Tabel 1.2.

Tabel 1.1: Tabel Penjelasan Fitur Data

Kelompok Fitur	Nama Fitur	Keterangan
Indikator kota dan tanggal	*city*	Singkatan nama kota, ‘sj’ untuk San Juan dan ‘iq’ untuk Iquitos.
	week_start_date	Tanggal dengan format yyyy-mm-dd.
Data iklim harian GHCN milik NOAA dengan pengukuran stasiun cuaca	*station_max_temp_c*	Temperatur maksimum.
	station_min_temp_c	Temperatur minimum.
	station_avg_temp_c	Rata-rata temperatur.
	station_precip_mm	Total pengendapan.
	station_diur_temp_rng_c	Rentang temperatur harian.
Pengukuran pengendapan satelit PERSIANN (0.25x0.25 skala derajat)	*precipitation_amt_mm*	Total pengendapan.
	reanalysis_sat_precip_amt_mm	Total pengendapan.
	reanalysis_dew_point_temp_k	Rata temperatur titik embun.
	reanalysis_air_temp_k	Rata temperatur udara.
	reanalysis_relative_humidity_percent	Rata kelembapan relatif.
	reanalysis_specific_humidity_g_per_kg	Rata kelembapan spesifik.

	reanalysis_precip_amt_kg_per_m 2	Total pengendapan.
	reanalysis_max_air_temp_k	Temperatur maksimum udara.
	reanalysis_min_air_temp_k	Temperatur minimum udara.
	reanalysis_avg_temp_k	Rara-rata temperatur udara.
	reanalysis_tdtr_k	Rentang temperatur harian.
Vegetasi satelit – *Normalized difference vegetation index (NDVI) – Pengukuruan CDR NDVI (0.5x0.5 skala derajat) milik NOAA*	*ndvi_se*	Piksel tenggara dari pusat kota.
	ndvi_sw	Piksel barat daya dari pusat kota.
	ndvi_ne	Piksel barat laut dari pusat kota.
	ndvi_nw	Piksel timur laut dari pusat kota.

Tabel 1.2: Tabel Penjelasan Label Data

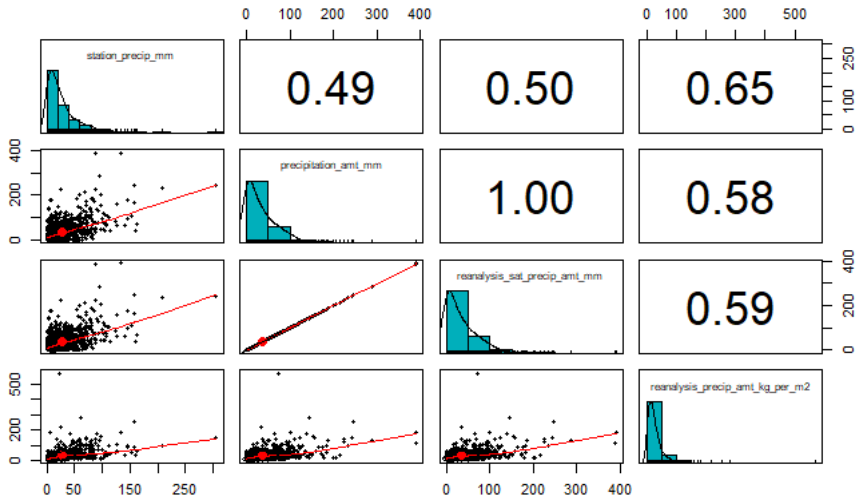
Nama Label	Keterangan
city	Singkatan nama kota, ‘sj’ untuk San Juan dan ‘iq’ untuk Iquitos.
year	Tahun.
weekofyear	Minggu dari tahun.
total_cases	Jumlah kasus DBD

1.2. *Data Preparation*

Data yang tersedia masih memiliki data yang kosong. Jumlah dari data yang hilang adalah 548 data yang tersebar pada 257 baris atau 17,6% dari data yang disediakan. Oleh karena itu, perlu dilakukan *data preparation* dengan langkah-langkah sebagai berikut.

1. Data akan dibagi berdasarkan variabel *city* yaitu, *sj* dan *iq.*
2. Data akan melewati proses *imputation* menggunakan MICE untuk mengisi data-data yang masih kosong pada fitur data *training* dan *testing*. Pengisian data-data yang kosong dilakukan dengan metode *predictive mean matching* (PMM) yang menggunakan nilai observasi yang terdekat dari model.
3. Setelah dilakukan proses *imputation*, masih ditemukan data yang kosong pada variabel *reanalysis_sat_precip_amt_mm.* Keterangan daripada variabel tersebut ternyata juga dimiliki oleh beberapa variabel lainnya yaitu,

reanalysis_precip_amt_kg_per_m2, *precipitation_amt_mm*, dan *station_precip_mm*. Oleh karena itu, uji korelasi dilakukan untuk melihat hubungan antara keempat variabel tersebut.



Gambar 1.2: Matriks Korelasi dan Histogram untuk Variabel Total Pengendapan

Melalui Gambar 1.2, ditemukan bahwa koefisien korelasi dari variabel *precipitation_amt_mm* dan *reanalysis_sat_precip_amt_mm* adalah 1,00. Koefisien korelasi antar dua variabel ini sangat tinggi sehingga variabel *reanalysis_sat_precip_amt_mm* tidak akan digunakan untuk penelitian.

4. Menggabungkan variabel *total_cases* dari label data *training* yang telah dibagi berdasarkan variabel *city* dengan fitur data *training* yang sudah melewati proses *imputation*.

1.3. Uji Korelasi Spearman

Uji korelasi Spearman digunakan untuk melihat korelasi koefisien dari variabel *total_cases* dengan variabel lainnya. Variabel *year* dan *weekofyear* tidak masuk ke dalam kelompok variabel yang akan diuji karena variabel tersebut merupakan skala waktu dari data yang disediakan. Fitur data terdiri dari 19 variabel yang berbeda.

Variabel ini terlalu banyak, sehingga akan dipilih 10 variabel dengan koefisien korelasi tertinggi.

1.4. Regresi Poisson

Model akan dibuat menggunakan metode *generalized linear model* (GLM) Poisson. Variabel respon yang digunakan adalah 10 variabel yang telah diperoleh sebelumnya melalui uji korelasi Spearman. Selanjutnya, optimisasi model dapat dilakukan dengan melakukan uji kolinearitas dan menghapus variabel-variabel dengan nilai *Variance Inflation Factors* (VIF) yang lebih besar dari 10. Formula dibentuk menggunakan sisa variabel yang ada. Model yang telah dibuat akan dievaluasi untuk melihat apakah model tersebut mengalami *overdispersion*, dimana *variance* lebih besar daripada *mean* (Utami, 2013). Hal ini dapat membuat hasil prediksi dari model menjadi kurang tepat karena nilai devians model yang besar.

1.5. Regresi Binomial Negatif

Model dibuat kembali menggunakan regresi binomial negatif dengan formula yang diperoleh pada akhir proses regresi Poisson untuk mengatasi *overdispersion*. Nilai *Akaike Information Criterion* (AIC) dari model akan dibandingkan dengan model regresi Poisson yang telah dibuat sebelumnya untuk melihat perkembangan dari penggunaan model regresi binomial negatif. Pengembangan model akan dilakukan sekali lagi dengan membuat model yang menggunakan variabel yang signifikan dengan *p-value* lebih kecil dari 0,05. Nilai MAE akan digunakan untuk melihat perkembangan dari kedua model tersebut.

1.6. Mencari Jumlah Kasus Demam Berdarah *Dengue*

Prediksi jumlah kasus demam berdarah akan menggunakan fitur data *testing* dan model regresi binomial negatif. Data uji yang digunakan merupakan data yang sudah melakukan proses *imputation* untuk mengisi data yang masih kosong. Hasil prediksi

jumlah kasus penyakit ini kemudian dimasukkan ke berkas *submission_format* dalam format *Comma Separated Values* (CSV). Berkas ini akan dikumpulkan ke situs perlombaan yang sudah disebutkan sebelumnya. Kompetisi ini menggunakan metrik MAE

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i|\,. \tag{1.1}$$

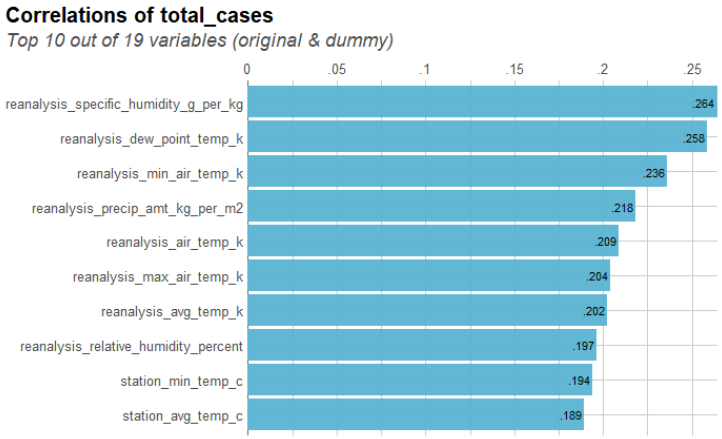
Tujuan dari prediksi ini adalah untuk meminimalkan nilai MAE. Jadi, prediksi yang baik adalah jika skor yang diperoleh serendah mungkin.

BAB II

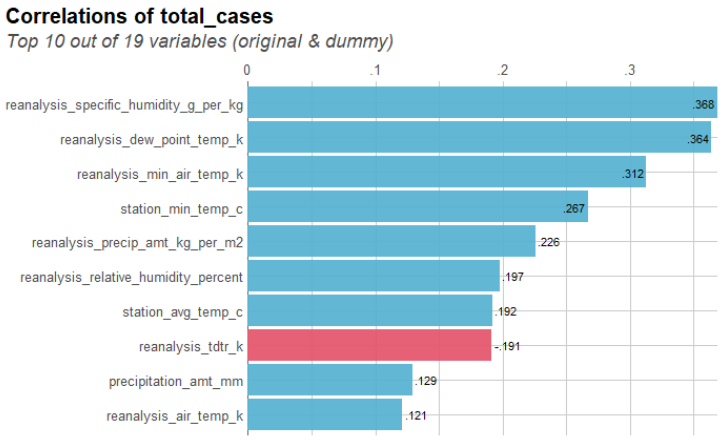
ANALISIS DAN PEMBAHASAN

2.1 Hasil Uji Korelasi Spearman

Pada Bab 1.3 telah dijelaskan bahwa peniliti akan memilih 10 variabel dengan koefisien korelasi tertinggi menggunakan uji korelasi Spearman dengan tujuan untuk mengurangi jumlah fitur data yang terlalu banyak. Variabel-variabel dengan korelasi tinggi untuk kota San Juan dan Iquitos ditunjukkan pada Gambar 2.1 dan Gambar 2.2.



Gambar 2.1: Grafik Korelasi Variabel *total_cases* dengan Variabel Lainnya pada kota San Juan.



Gambar 2.2: Grafik Korelasi Variabel *total_cases* dengan Variabel Lainnya pada kota Iquitos

Melalui Gambar 2.1 dan Gambar 2.2, ditemukan 10 variabel pada masing-masing kota yang memiliki koefisien korelasi tertinggi dengan variabel *total_cases*. Hal ini mungkin terjadi karena kesepuluh variabel tersebut merupakan ukuran kelembapan dan temperatur. Demam berdarah dengue merupakan penyakit yang dapat ditularkan oleh nyamuk *Aedes aegypti.* Nyamuk *Aedes aegypti* akan lebih mudah berkembang biak di lingkungan yang lembab dan memiliki suhu yang minimum. Oleh karena itu, peningkatan populasi nyamuk *Aedes aegypti* di daerah yang memiliki tingkat kelembapan yang tinggi dapat memperbanyak kasus deman berdarah (M. Reinhold et al., 2018: 7).

2.2 Hasil Regresi Poisson

Dalam penelitian ini, dilakukan empat kali regresi Poisson. Setelah dilakukan regresi sesuai metode yang telah dibahas pada Bab 1.4, diperoleh empat model. Model pertama adalah model yang menggunakan variabel prediktor untuk kota San Juan yang telah dibahas pada Bab 2.1, yaitu:

$$\begin{aligned} \textit{total\_cases} = 127 &+ 0{,}6423 * \textit{reanalysis\_specific\_humidity\_g\_per\_kg} - \\ &1{,}821 * \textit{reanalysis\_dew\_point\_temp\_k} + \\ &0{,}3732 * \textit{reanalysis\_min\_air\_temp\_k} + \\ &0{,}0006 * \textit{reanalysis\_precip\_amt\_kg\_per\_m2} + \\ &2{,}241 * \textit{reanalysis\_air\_temp\_k} + \\ &0{,}3881 * \textit{reanalysis\_max\_air\_temp\_k} - \\ &1{,}731 * \textit{reanalysis\_avg\_temp\_k} + \\ &0{,}2558 * \textit{reanalysis\_relative\_humidity\_percent} - \\ &0{,}0343 * \textit{station\_min\_temp\_c} + 0{,}1386 * \textit{station\_avg\_temp\_c}. \end{aligned} \tag{2.1}$$

Model kedua adalah model yang menggunakan variabel prediktor untuk kota Iquitos yang telah dibahas pada Bab 2.1, yaitu:

$$\begin{aligned} \textit{total\_cases} = 185{,}3 + 0{,}965 * \textit{reanalysis\_specific\_humidity\_g\_per\_kg} - \\ 0{,}8368 * \textit{reanalysis\_dew\_point\_temp\_k} + \\ 0{,}0049 * \textit{reanalysis\_min\_air\_temp\_k} + \\ 0{,}1082 * \textit{station\_min\_temp\_c} - \\ 0{,}0018 * \textit{reanalysis\_precip\_amt\_kg\_per\_m2} - \\ 0{,}0044 * \textit{reanalysis\_realtive\_humidity\_percent} - \\ 0{,}0649 * \textit{station\_avg\_temp\_c} - \\ 0{,}0949 * \textit{reanalysis\_tdtr\_k} - \\ 0{,}0003 * \textit{precipitation\_amt\_mm} + 0{,}1572 * \textit{reanalysis\_air\_temp\_k}. \end{aligned} \quad (2.2)$$

Setelah kedua model terbentuk, dilakukan uji kolinearitas antar variabel. Dari hasil uji kolinearitas yang dapat dilihat di Tabel 2.1, masih terdapat beberapa variabel yang memiliki nilai VIF lebih dari 10. Variabel tersebut tidak akan dipakai untuk model ketiga dan keempat untuk menghindari multikolinearitas.

Tabel 2.1: Tabel Hasil Uji Kolinearitas

Kota	Variabel	VIF
San Juan	*reanalysis_precip_amt_kg_per_m2*	1
	station_min_temp_c	1,07
	station_avg_temp_c	2
	reanalysis_specific_humidity_g_per_kg	211,07
	reanalysis_dew_point_temp_k	2.762,45
	reanalysis_min_air_temp_k	23,93
	reanalysis_air_temp_k	3.097,94
	reanalysis_max_air_temp_k	21,35
	reanalysis_avg_temp_k	159,92
Iquitos	*reanalysis_min_air_temp_k*	1,6

	station_min_temp_c	1,19
	reanalysis_precip_amt_kg_per_m2	1
	station_avg_temp_c	1,12
	reanalysis_tdtr_k	1,54
	precipitation_amt_mm	1
	reanalysis_specific_humidity_g_per_kg	404,89
	reanalysis_dew_point_temp_k	425,56
	reanalysis_relative_humidity_percent	34,6
	reanalysis_air_temp_k	230,94

Model ketiga merupakan pengembangan dari model 2.1 dengan menghapus variabel yang memiliki nilai VIF lebih dari 10, yaitu:

$$total\_cases = -2{,}5465 + 0{,}0026 * reanalysis\_precip\_amt\_kg\_per\_m2 - 0{,}0144 * station\_min\_temp\_c + 0{,}2322 * station\_avg\_temp\_c. \quad (2.3)$$

Model keempat merupakan pengembangan dari model 2.2 dengan menghapus variabel yang memiliki nilai VIF lebih dari 10, yaitu:

$$total\_cases = -41{,}94 + 0{,}1389 * reanalysis\_min\_air\_temp\_k + 0{,}132 * station\_min\_temp\_c - 0{,}0003 * reanalysis\_precip\_amt\_kg\_per\_m2 + 0{,}0271 * station\_avg\_temp\_c - 0{,}0337 * reanalysis\_tdtr\_k + 0{,}0001 * precipitaion\_amt\_mm. \quad (2.4)$$

Model ketiga dan keempat mengalami *overdispersion*. Hal ini dapat diketahui dengan nilai dari rasio *residual deviance* ke derajat kebebasan tidak sama dengan satu (Dormann, 2016) atau nilai dari *residual deviance* yang dibagi dengan *degrees of freedom* lebih besar dari satu (Utami, 2013). Hal ini dapat dilihat di dalam Tabel 2.2.

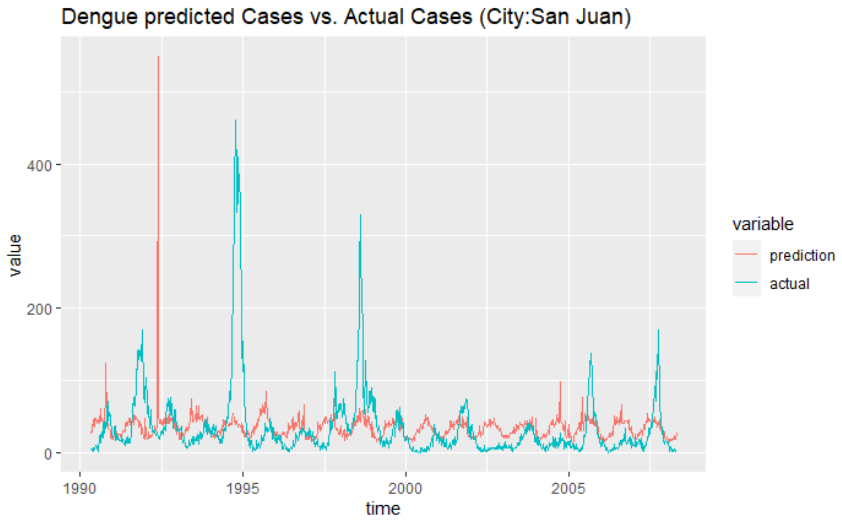
Tabel 2.2: Tabel Ringkasan Hasil *Fitting* Model Regresi Poisson

Model ke-	AIC	*Residual Deviance*	*Degrees of Freedom*	*Residual Deviance/Degrees of Freedom*
3	40.615	36.159	932	38,7972103
4	6.414,7	4865,4	513	9,484210526

2.3 Hasil Regresi Binomial Negatif

Regresi binomial negatif dilakukan sebanyak empat kali di dalam penelitian ini. Setelah dilakukan regresi sesuai metode yang telah dibahas pada Bab 1.5, diperoleh empat model. Model pertama adalah hasil pengembangan dari model 2.3 pada Bab 2.2, yaitu:

$$total\_cases = -2{,}5836 + 0{,}0049 * reanalysis\_precip\_amt\_kg\_per\_m2 - 0{,}0801 * station\_min\_temp\_c + 0{,}2857 * station\_avg\_temp\_c. \quad (2.5)$$

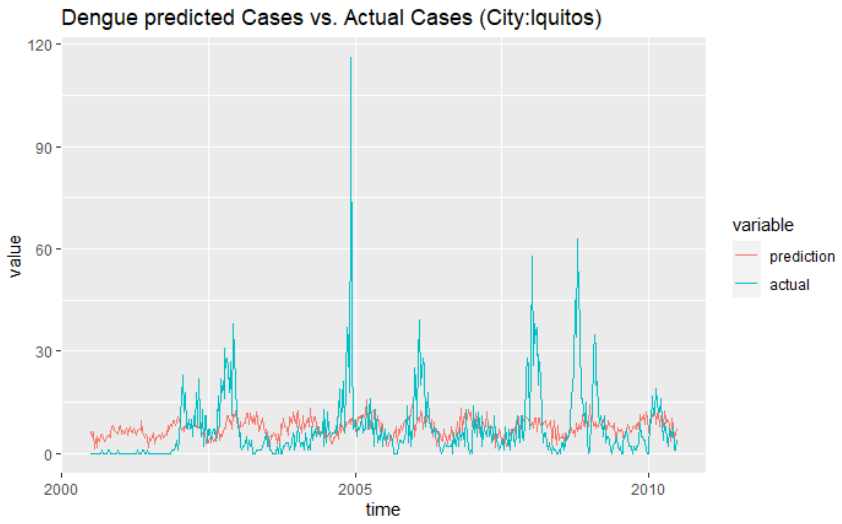


Gambar 2.3: Grafik Perbandingan Hasil Prediksi Model Pertama menggunakan Fitur Data *Training* dengan variabel *total_cases* kota San Juan pada Data *Training*

Untuk melihat performa model dalam memprediksi variabel *total_cases* (jumlah kasus), setiap model akan dilakukan percobaan untuk memprediksi dengan menggunakan data *training.* Hal ini dilakukan untuk melihat perbandingan hasil prediksi dengan data yang sudah diberikan. Melalui Gambar 2.3, dapat dilihat bahwa model pertama belum dapat memprediksi jumlah kasus yang tinggi dengan tepat. Namun, perbedaan hasil prediksi dengan data yang sebenarnya terlihat lebih sedikit untuk jumlah kasus yang kurang dari 100. Model kedua adalah hasil pengembangan dari model 2.4 pada Bab 2.4, yaitu:

$$\begin{aligned} total\_cases = {} & -35{,}47 + 0{,}1137 * reanalysis\_min\_air\_temp\_k + \\ & 0{,}1436 * station\_min\_temp\_c + \\ & 0{,}0007 * reanalysis\_precip\_amt\_kg\_per\_m2 + \\ & 0{,}0448 * station\_avg\_temp\_c - 0.0233 * reanalysis\_tdtr\_k + \\ & 0{,}0002 * precipitaion\_amt\_mm. \end{aligned} \tag{2.6}$$

Melalui Gambar 2.4, terlihat bahwa model 2.6 belum bisa memprediksi jumlah *total_cases* yang tinggi. Namun, hasil prediksi untuk jumlah kasus yang kurang dari 15 di tahun 2005-2010 tidak jauh berbeda dengan jumlah kasus yang sebenarnya.



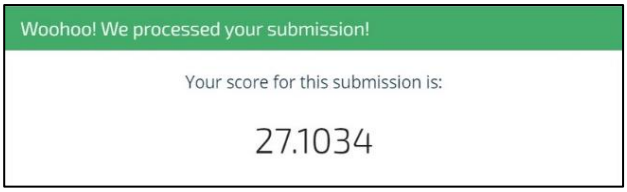
Gambar 2.4: Grafik Perbandingan Hasil Prediksi Model Kedua menggunakan Fitur Data *Training* dengan variabel *total_cases* kota Iquitos pada Data *Training*

Melalui Tabel 2.3, diperoleh nilai *residual deviance* yang mendekati *degrees of freedom* pada model 2.5 dan 2.6. Hal ini memperlihatkan bahwa model regresi binomial negatif telah mengoreksi *overdispersion* pada model regresi Poisson di Bab 2.2. Selain itu, kedua model ini memiliki nilai AIC yang lebih rendah jika dibandingkan dengan nilai AIC model regresi Poisson di Tabel 2.2. Nilai AIC yang lebih rendah ini menunjukkan bahwa model regresi binomial negatif merupakan model yang lebih baik untuk digunakan (Witten et al., 2013).

Tabel 2.3: Tabel Ringkasan Hasil *Fitting* Model Regresi Binomial Negatif Pertama dan Kedua

Model	**AIC**	***Residual Deviance***	***Degrees of Freedom***	***Residual Deviance/Degrees of Freedom***
2.5	8.412,2	1.051,8	932	1,128540773
2.6	3.149,3	596,41	513	1,162592593

Model 2.5 dan 2.6 digunakan untuk memprediksi variabel *total_cases* seperti yang telah dibahas pada Bab 1.6. MAE dari kedua model tersebut dapat dilihat pada Gambar 2.5.

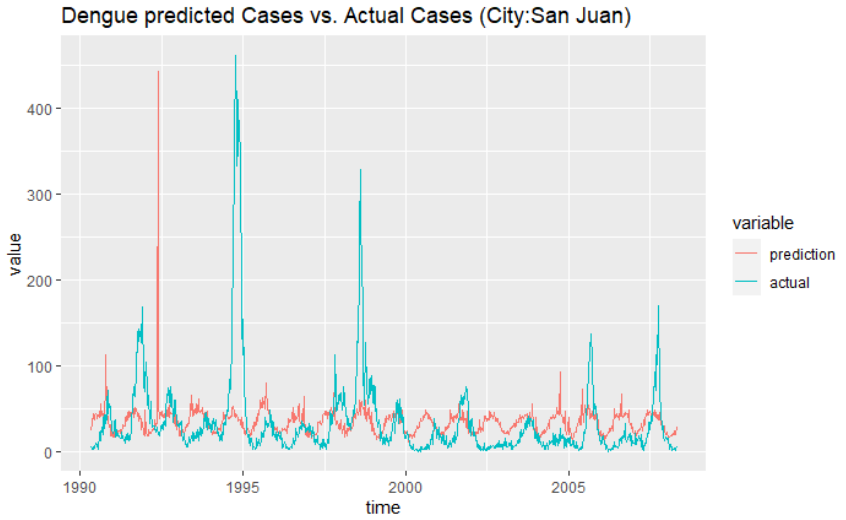


Gambar 2.5: MAE untuk Model Regresi Binomial Negatif Pertama dan Kedua.

Model ketiga adalah pengembangan dari model 2.5, yaitu:

$$total\_cases = -2{,}2494 + 0{,}0045 * reanalysis\_precip\_amt\_kg\_per\_m2 + 0{,}2068 * station\_min\_temp\_c. \quad (2.7)$$

Melalui Gambar 2.6, hasil prediksi model 2.7 tidak jauh berbeda dengan hasil prediksi model 2.5 pada Gambar 2.3. Hal ini dapat dilihat dengan ketidakmampuan model untuk memprediksi jumlah kasus yang lebih besar dari 100 dengan tepat.

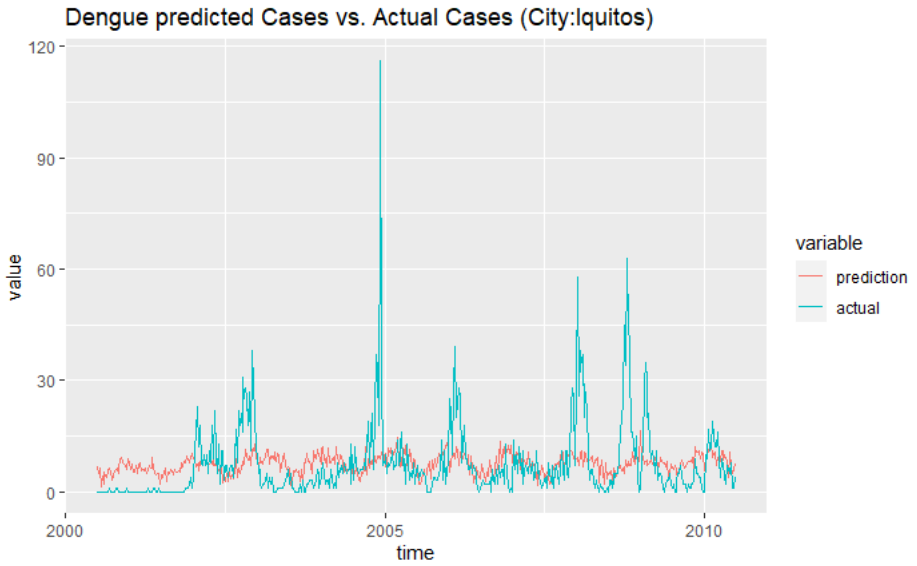


Gambar 2.6: Grafik Perbandingan Hasil Prediksi Model Ketiga menggunakan Fitur Data *Training* dengan variabel *total_cases* kota San Juan pada Data *Training*

Model keempat adalah pengembangan dari model 2.6, yaitu:

$$total\_cases = -43{,}1094 + 0{,}1425 * reanalysis\_min\_air\_temp\_k + 0{,}1569 * station\_min\_temp\_c. \quad (2.8)$$

Melalui Gambar 2.7, dapat dilihat bahwa hasil prediksi model keempat tidak jauh berbeda dengan hasil prediksi model kedua. Hal ini dapat dilihat dengan kemampuan model keempat untuk memprediksi jumlah kasus yang kurang dari 15 pada tahun 2005-2010 yang mendekati nilai sebenarnya.



Gambar 2.7: Grafik Perbandingan Hasil Prediksi Model Keempat menggunakan Fitur Data *Training* dengan variabel *total_cases* kota Iquitos pada Data *Training*

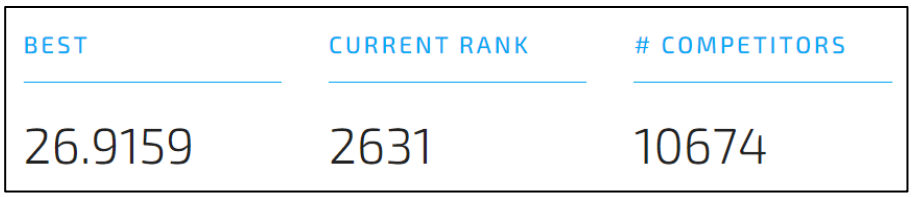
Nilai AIC yang diperoleh pada model 2.8 mengalami penurunan dan nilai rasio *residual deviance* terhadap derajat kebebasannya juga menurun. Hal ini menunjukkan model 2.8 yang kita peroleh mengalami perkembangan. Ringkasan nilai-nilai yang diperoleh dari hasil regresi disajikan pada Tabel 2.4.

Tabel 2.4: Tabel Ringkasan Hasil *Fitting* Model Regresi Binomial Negatif Ketiga dan Keempat

Model	**AIC**	***Residual Deviance***	***Degrees of Freedom***	***Residual Deviance/Degrees of Freedom***
2.7	8.412,7	1.052,2	933	1,127759914
2.8	3.142,7	596,51	517	1,153791103

Pengembangan yang berhasil dilakukan ini dapat dilihat juga dengan menurunnya nilai MAE pada prediksi dari model 2.7 dan 2.8. Nilai MAE turun sebanyak 0,21081 dari

nilai MAE sebelumnya. Hasil MAE dan *rank* yang didapatkan dapat dilihat pada Gambar 2.4.



Gambar 2.4: Nilai MAE dan *Rank* dari Hasil Prediksi Model Regresi Binomial Negatif yang Ketiga dan Keempat

BAB III

KESIMPULAN DAN SARAN

3.1 Kesimpulan

Dalam penelitian ini telah dilakukan regresi Poisson dan regresi binomial negatif pada data *training* yang telah disediakan oleh situs *drivendata.org.* Model yang dihasilkan telah digunakan untuk memprediksi variabel *total_cases* pada data *testing*. Performa model diukur dengan menggunakan AIC, *residual deviance*, dan MAE. Berdasarkan analisis data yang telah dilakukan pada Bab 2, dapat ditarik beberapa kesimpulan sebagai berikut.

1. Model *Poisson* yang dibuat mengalami *overdispersion.* Hal ini dapat diketahui dengan nilai dari rasio *residual deviance* ke derajat kebebasan tidak sama dengan satu dan nilai *residual deviance* yang jauh lebih besar daripada derajat kebebasan.
2. Model binomial negatif yang dibuat dapat mengoreksi *overdispersion* yang dimiliki model Poisson. Nilai dari rasio *residual deviance* ke derajat kebebasan sudah mendekati satu. Selain itu, model binomial negatif merupakan model yang lebih baik dengan nilai AIC yang lebih rendah daripada nilai AIC model Poisson.
3. Pengembangan model binomial negatif berhasil dilakukan dengan memilih variabel prediktor yang signifikan dengan selang kepercayaan sebesar 95%. Variabel prediktor tersebut adalah *reanalysis_precip_amt_kg_per_m2* dan *station_avg_temp_c* untuk kota San Juan dan *reanalysis_min_air_temp_k* dan *station_min_temp_c* untuk kota Iquitos. Hal ini dapat dilihat dengan penurunan nilai MAE sebesar 0,21081.

3.2 Saran

Berdasarkan kesimpulan, terdapat dua jenis saran untuk pengembangan penelitian ini sehingga hasil prediksi yang diperoleh lebih akurat. Saran untuk penelitian selanjutnya sebagai berikut.

3.2.1 Saran Teoretis

Pada penelitian ini, data yang diperoleh dari situs perlombaan masih memiliki data kosong. Akan lebih baik jika data yang digunakan lengkap sehingga model yang diperoleh akan lebih akurat untuk penelitian selanjutnya. Prediksi jumlah kasus untuk penelitian ini menggunakan metode *generalized linear model* (GLM), yaitu regresi Poisson dan regresi Binomial Negatif. Untuk penelitian selanjutnya, dapat digunakan metode lain, seperti *random forest* dan ARIMA untuk memperoleh model yang memiliki akurasi lebih baik. Model yang diperoleh juga dapat dikembangkan lagi menggunakan *Generalized Additive Model* untuk menghindari *overfitting.*

3.2.2 Saran Praktis

Berdasarkan model 2.7 dan 2.8 yang merupakan model terbaik yang ditemukan pada penelitian ini, terdapat empat variabel yang bersifat signifikan terhadap jumlah kasus DBD di kedua kota. Saran praktis yang diharapkan dapat bermanfaat untuk penduduk kota San Juan adalah untuk memperhatikan kemungkinan turunnya hujan. Sedangkan untuk kota Iquitos, perlu diperhatikan suhu pada daerah tersebut. Hal ini dapat dilakukan dengan memperhatikan perkiraan cuaca harian sehingga jumlah kasus DBD pada kedua daerah diharapkan dapat berkurang.

DAFTAR PUSTAKA

Dormann, Carstem F. (2016). Overdispersion, and how to deal with it in R and JAGS. APES.

Reinhold, M. Joanna, Claudio R. Lazzari, dan Cholé Lahondére. (2018). *Effects of the Environmental Temperature on Aedes aegypti and Aedes albopictus Mosquitoes: A Review. Jurnal MDPI insect.*

Utami, T. W. (2013). Analisis Regresi Binomial Negatif untuk Mengatasi Overdispersion Regresi Poisson pada Kasus Demam Berdarah Dengue. *Jurnal Statistika Universitas Muhammadiyah Semarang*, *1*(2).

Witten, D., James, G., Tibshirani, R., & Hastie, T. (2013). An introduction to statistical learning (Vol. 103, p. 226). New York: Springer.

LAMPIRAN A: Hasil Regresi Poisson

Regresi I: Sepuluh Variabel Prediktor dengan Korelasi Tertinggi

```
Call:
glm(formula = formula_10corr_sj, family = "poisson", data = imp_df_train_sj,
    contrasts = NULL)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-10.434   -4.896   -2.034    1.075   35.331

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                        1.270e+02  2.425e+01   5.237 1.63e-07 ***
reanalysis_specific_humidity_g_per_kg  6.423e-01  7.917e-02   8.113 4.93e-16 ***
reanalysis_dew_point_temp_k        -1.821e+00  2.909e-01  -6.262 3.81e-10 ***
reanalysis_min_air_temp_k           3.732e-01  1.741e-02  21.427  < 2e-16 ***
reanalysis_precip_amt_kg_per_m2     6.774e-04  1.807e-04   3.748 0.000178 ***
reanalysis_air_temp_k               2.241e+00  3.107e-01   7.213 5.49e-13 ***
reanalysis_max_air_temp_k           3.881e-01  1.562e-02  24.843  < 2e-16 ***
reanalysis_avg_temp_k              -1.731e+00  7.264e-02 -23.834  < 2e-16 ***
reanalysis_relative_humidity_percent  2.558e-01  6.397e-02   3.998 6.38e-05 ***
station_min_temp_c                 -3.436e-02  9.159e-03  -3.751 0.000176 ***
station_avg_temp_c                  1.386e-01  1.163e-02  11.911  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 39497  on 935  degrees of freedom
Residual deviance: 34654  on 925  degrees of freedom
AIC: 39133

Number of Fisher Scoring iterations: 6
```

Gambar A.1: Regresi Pertama untuk Kota San Juan

```
Call:
glm(formula = formula_10corr_iq, family = "poisson", data = imp_df_train_iq,
    contrasts = NULL)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.4129  -2.6423  -1.2046   0.6573  18.7347

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                        1.853e+02  6.740e+01   2.749 0.005974 **
reanalysis_specific_humidity_g_per_kg  9.650e-01  2.191e-01   4.405 1.06e-05 ***
reanalysis_dew_point_temp_k        -8.368e-01  2.482e-01  -3.371 0.000749 ***
reanalysis_min_air_temp_k           4.907e-03  2.172e-02   0.226 0.821218
station_min_temp_c                  1.082e-01  1.950e-02   5.550 2.85e-08 ***
reanalysis_precip_amt_kg_per_m2    -1.792e-03  4.193e-04  -4.273 1.93e-05 ***
reanalysis_relative_humidity_percent  -4.391e-03  4.033e-02  -0.109 0.913294
station_avg_temp_c                 -6.487e-02  2.524e-02  -2.571 0.010148 *
reanalysis_tdtr_k                  -9.487e-02  1.973e-02  -4.808 1.52e-06 ***
precipitation_amt_mm               -3.408e-04  5.143e-04  -0.663 0.507648
reanalysis_air_temp_k               1.572e-01  1.777e-01   0.885 0.376307
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 5354.0  on 519  degrees of freedom
Residual deviance: 4748.2  on 509  degrees of freedom
AIC: 6305.4

Number of Fisher Scoring iterations: 6
```

Gambar A.2: Regresi Pertama untuk Kota Iquitos

Regresi II: Menggunakan Variabel Prediktor dengan nilai VIF < 10

```
Call:
glm(formula = formula_vif1_sj, family = "poisson", data = imp_df_train_sj,
    contrasts = NULL)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-13.279   -4.962   -2.310    1.063   36.416

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -2.5465251  0.1204892 -21.135   <2e-16 ***
reanalysis_precip_amt_kg_per_m2  0.0025552  0.0001114  22.928   <2e-16 ***
station_min_temp_c           -0.0143982  0.0086990  -1.655   0.0979 .
station_avg_temp_c            0.2321881  0.0092474  25.108   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 39497  on 935  degrees of freedom
Residual deviance: 36150  on 932  degrees of freedom
AIC: 40615

Number of Fisher Scoring iterations: 6
```

Gambar A.3: Regresi Kedua untuk Kota San Juan

```
Call:
glm(formula = formula_vif1_iq, family = "poisson", data = imp_df_train_iq,
    contrasts = NULL)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.7885  -2.7005  -1.1960   0.7034  18.6546

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -4.194e+01  4.048e+00 -10.360  < 2e-16 ***
reanalysis_min_air_temp_k     1.389e-01  1.472e-02   9.434  < 2e-16 ***
station_min_temp_c            1.320e-01  1.888e-02   6.991 2.74e-12 ***
reanalysis_precip_amt_kg_per_m2 -3.156e-04  3.729e-04  -0.846 0.397342
station_avg_temp_c            2.711e-02  2.405e-02   1.127 0.259699
reanalysis_tdtr_k            -3.371e-02  9.400e-03  -3.586 0.000335 ***
precipitation_amt_mm          1.222e-04  4.992e-04   0.245 0.806656
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 5354.0  on 519  degrees of freedom
Residual deviance: 4865.4  on 513  degrees of freedom
AIC: 6414.7

Number of Fisher Scoring iterations: 6
```

Gambar A.4: Regresi Kedua untuk Kota Iquitos

LAMPIRAN B: Hasil Regresi Binomial Negatif

Regresi I: Menggunakan Varibel Prediktor dengan nilai VIF < 10

```
Call:
glm.nb(formula = formula_vif1_sj, data = imp_df_train_sj, contrasts = NULL,
    init.theta = 1.057312731, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7348  -1.0464  -0.4392   0.1948   3.9676

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -2.5835856  0.6399685  -4.037 5.41e-05 ***
reanalysis_precip_amt_kg_per_m2  0.0048861  0.0009252   5.281 1.28e-07 ***
station_min_temp_c         -0.0801179  0.0494799  -1.619    0.105
station_avg_temp_c          0.2857436  0.0520558   5.489 4.04e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.0573) family taken to be 1)

    Null deviance: 1164.2  on 935  degrees of freedom
Residual deviance: 1051.8  on 932  degrees of freedom
AIC: 8412.2

Number of Fisher Scoring iterations: 1


              Theta:  1.0573
          Std. Err.:  0.0457

 2 x log-likelihood:  -8402.1990
```

Gambar B.1: Regresi Pertama untuk kota San Juan

```
Call:
glm.nb(formula = formula_vif1_iq, data = imp_df_train_iq, contrasts = NULL,
    init.theta = 0.7635995045, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0503  -1.0852  -0.3696   0.2258   3.4163

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -3.547e+01  1.235e+01  -2.872  0.00407 **
reanalysis_min_air_temp_k    1.137e-01  4.524e-02   2.514  0.01194 *
station_min_temp_c           1.436e-01  5.839e-02   2.459  0.01394 *
reanalysis_precip_amt_kg_per_m2  6.761e-04  1.295e-03   0.522  0.60153
station_avg_temp_c           4.484e-02  7.470e-02   0.600  0.54831
reanalysis_tdtr_k           -2.330e-02  3.031e-02  -0.769  0.44213
precipitation_amt_mm         2.318e-04  1.683e-03   0.138  0.89046
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.7636) family taken to be 1)

    Null deviance: 645.74  on 519  degrees of freedom
Residual deviance: 596.41  on 513  degrees of freedom
AIC: 3149.3

Number of Fisher Scoring iterations: 1


              Theta:  0.7636
          Std. Err.:  0.0546

 2 x log-likelihood:  -3133.3080
```

Gambar B.2: Regresi Pertama untuk kota Iquitos

Regresi II: Menggunakan Varibel Prediktor dengan *p-value* < 0,05

```
Call:
glm.nb(formula = formula_pv_sj, data = imp_df_train_sj, contrasts = NULL,
    init.theta = 1.055033428, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7318  -1.0416  -0.4395   0.1899   4.1124

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -2.249358   0.623543  -3.607 0.000309 ***
reanalysis_precip_amt_kg_per_m2  0.004524   0.000911   4.966 6.82e-07 ***
station_avg_temp_c           0.206786   0.023148   8.933  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.055) family taken to be 1)

    Null deviance: 1161.9  on 935  degrees of freedom
Residual deviance: 1052.2  on 933  degrees of freedom
AIC: 8412.7

Number of Fisher Scoring iterations: 1


              Theta:  1.0550
          Std. Err.:  0.0456

 2 x log-likelihood:  -8404.7120
```

Gambar B.3: Regresi Kedua untuk kota San Juan

```
Call:
glm.nb(formula = formula_pv_iq, data = imp_df_train_iq, contrasts = NULL,
    init.theta = 0.7613113832, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0311  -1.0898  -0.3770   0.2316   3.4938

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -43.10938   11.16972  -3.859 0.000114 ***
reanalysis_min_air_temp_k   0.14252    0.04022   3.543 0.000395 ***
station_min_temp_c          0.15685    0.05331   2.942 0.003259 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.7613) family taken to be 1)

    Null deviance: 644.30  on 519  degrees of freedom
Residual deviance: 596.51  on 517  degrees of freedom
AIC: 3142.7

Number of Fisher Scoring iterations: 1


              Theta:  0.7613
          Std. Err.:  0.0544

 2 x log-likelihood:  -3134.7260
```

Gambar B.4: Regresi Kedua untuk kota Iquitos