

Flu Shot Learning : Predict H1N1 and Seasonal Flu Vaccines

Elisa Satyawati, Irenne, dan Jeffrey Widjaya

Kelompok 1, Program Studi Matematika, Fakultas Sains dan Teknologi
Universitas Pelita Harapan

Virus H1N1 dan *seasonal flu* merupakan virus yang dapat menyebar sehingga vaksinasi merupakan salah satu upaya penyebaran yang cocok dilakukan. Dengan adanya vaksinasi, diharapkan timbulnya kekebalan kelompok untuk mengurangi penyebaran virus. Fokus utama dalam laporan ini adalah mencari probabilitas responden menerima vaksin H1N1 dan *seasonal flu* dengan membandingkan dua model, yaitu model Naive Bayes dan Regresi Logistik. Kedua model tersebut akan dibandingkan berdasarkan nilai AUC dan hasil prediksi yang diperoleh.

Kata Kunci: H1N1, Naive Bayes, Regresi Logistik, *seasonal flu*

1. Pendahuluan. Virus H1N1 atau influenza A pertama kali muncul pada musim semi tahun 2009 di Amerika Serikat dan menyebar hingga ke seluruh dunia. Virus ini mengandung kombinasi gen influenza unik yang sebelumnya belum teridentifikasi. Dalam kurun waktu satu tahun, peneliti memperkirakan virus ini menyebabkan kematian secara global sebanyak 151.000 - 575.000. Sehubungan dengan upaya untuk mengatasi penyebaran virus ini, vaksin untuk virus H1N1 mulai tersedia pada Oktober 2009¹. Menurut Biofarma, vaksin merupakan mikroorganisme yang dilemahkan dan akan diberikan kepada orang yang sehat sehingga dapat menimbulkan antibodi yang dapat membuat seseorang kebal terhadap virus ataupun bakteri tersebut². Dengan adanya vaksinasi pada setiap individu diharapkan adanya kekebalan kelompok sehingga dapat mengurangi penyebaran penyakit tersebut. Untuk melihat berapa banyak orang yang akan mau menerima vaksin berdasarkan data yang dapat diperoleh, dilakukan perhitungan probabilitas responden menerima vaksin.

Model Naive Bayes akan digunakan untuk mendapatkan hasil representasi dari model *probability distribution*. Hasil dari model Naive Bayes akan digunakan untuk memberikan nilai probabilitas responden menerima vaksin H1N1 dan *seasonal flu*. Selain model Naive Bayes, juga digunakan model Regresi Logistik untuk mencari probabilitas tersebut. Hasil dari kedua model yang digunakan ini akan dianalisis dan dilakukan perbandingan untuk melihat model mana yang memperoleh hasil lebih baik.

¹ Driven Data, “*Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines*” (<https://www.drivendata.org/competitions/66/flu-shot-learning/page/210/> diakses pada tanggal 30 Mei 2022 pukul 14.10)

² Biofarma, “*Vaksin*” (<https://www.biofarma.co.id/id/researcher/detail/vaksin> diakses pada tanggal 30 Mei 2022 pukul 14.21)

2. Rumusan Masalah. Dua model yang digunakan dalam pemodelan ini adalah model Naive Bayes dan Regresi Logistik. Regresi Logistik merupakan suatu model yang digunakan untuk menganalisis hubungan antara variabel dependen dan variabel independen dimana variabel dependen umumnya berupa data dikotomi[1]. Variabel dependen yang tersedia dalam data yang diberikan juga merupakan data dikotomi, sehingga hal tersebut menjadi alasan dalam pemilihan model Regresi Logistik. Model Naive Bayes menggunakan teorema Bayes dalam mengkombinasikan *prior probability* dengan probabilitas bersyarat untuk menghitung probabilitas klasifikasi yang mungkin. Model ini juga menggunakan asumsi independensi yang kuat dalam setiap kejadiannya[2]. Oleh karena hal tersebut, model Naive Bayes digunakan pada data ini.

Dalam pemodelan yang dilakukan, pada data terdapat 36 variabel yang diberikan. Akan tetapi, dalam pemilihan variabel terdapat beberapa variabel yang tidak digunakan dalam pemodelan. Hal tersebut dilakukan dengan alasan untuk menghindari terjadinya *overfit*. Ketika menggunakan model Regresi Logistik, akan lebih baik jika memilih variabel independen yang dapat menjelaskan hasil dengan baik dan juga memilih jumlah variabel independen yang tidak terlalu banyak[3]. Variabel independen yang digunakan pada model Naive Bayes akan disesuaikan dengan variabel independen yang digunakan dalam model Regresi Logistik dengan tujuan untuk membandingkan hasil probabilitas dari kedua model tersebut.

Model pertama yang digunakan untuk menghitung probabilitas responden mendapatkan vaksin H1N1 dan *seasonal flu* adalah model Naive Bayes. Formula umum yang digunakan dalam teorema Bayes[4].

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Dengan :

$P(A|B)$ = probabilitas akhir bersyarat hipotesis A terjadi jika diberikan B terjadi

$P(B|A)$ = probabilitas B mempengaruhi hipotesis A

$P(A)$ = probabilitas hipotesis A terjadi

$P(B)$ = probabilitas hipotesis B terjadi

Formula tersebut akan digunakan untuk memodelkan data, sehingga formula Naive Bayes yang digunakan adalah sebagai berikut.

$$= \frac{P(y_1 = k | x_1 x_2 x_3 x_4 x_5 x_6 x_7)}{P(x_1) \cdot P(x_2) \cdot P(x_3) \cdot P(x_4) \cdot P(x_5) \cdot P(x_6) \cdot P(x_7)}$$

Dengan nama variabel sebagai berikut :

y_1 = h1n1_vaccine

k = kelas dari h1n1_vaccine (dalam kasus ini keadaan seseorang menerima vaksin h1n1)

x_1 = h1n1_concern

x_2 = h1n1_knowledge

x_3 = opinion_h1n1_vacc_effective

x_4 = opinion_h1n1_risk

x_5 = opinion_h1n1_sick_from_vacc

x6 = age_group

x7 = sex

$$= \frac{P(y_2 = k | x_1 x_2 x_3 x_4 x_5 x_6 x_7)}{P(x_1) \cdot P(x_2) \cdot P(x_3) \cdot P(x_4) \cdot P(x_5) \cdot P(x_6) \cdot P(x_7)}$$

Dengan nama variabel sebagai berikut :

y₂ = seasonal_vaccine

k = kelas dari seasonal_vaccine (dalam kasus ini keadaan seseorang menerima vaksin *seasonal flu*)

x₁ = h1n1_concern

x₂ = h1n1_knowledge

x₃ = opinion_h1n1_vacc_effective

x₄ = opinion_h1n1_risk

x₅ = opinion_h1n1_sick_from_vacc

x₆ = age_group

x₇ = sex

Selanjutnya, model kedua yang digunakan untuk menghitung probabilitas serta sebagai perbandingan dengan model pertama adalah model Regresi Logistik. Model Regresi Logistik memiliki formula sebagai berikut[5].

$$y = \frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k)}}$$

Dengan :

y = variabel dependen

β_i = koefisien variabel independen ke-i

x_i = variabel independen ke-i

Formula tersebut akan digunakan untuk memodelkan data, sehingga formula yang digunakan adalah sebagai berikut.

$$y_1 = \frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7)}}$$

Dengan nama variabel sebagai berikut:

y₁ = h1n1_vaccine

x₁ = h1n1_concern

x₂ = h1n1_knowledge

x₃ = opinion_h1n1_vacc_effective

x₄ = opinion_h1n1_risk

x₅ = opinion_h1n1_sick_from_vacc

x₆ = age_group

x₇ = sex

$$y_2 = \frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7)}}$$

Dengan nama variabel sebagai berikut:

y_2 = seasonal_vaccine

x_1 = h1n1_concern

x_2 = h1n1_knowledge

x_3 = opinion_seas_vacc_effective

x_4 = opinion_seas_risk

x_5 = opinion_seas_sick_from_vacc

x_6 = age_group

x_7 = sex

Model Naive Bayes dan Regresi Logistik tersebut akan digunakan untuk menghitung probabilitas responden menerima vaksin H1N1 dan *seasonal flu*.

3. **Metodologi.** Tahapan-tahapan yang dilakukan untuk mencari probabilitas responden terhadap vaksin H1N1 dan *seasonal flu* baik dengan metode Naive Bayes dan Regresi Logistik adalah sebagai berikut :

1. Pemilihan Data

Tahap ini adalah tahapan yang bertujuan untuk memilih data dan variabel yang berkaitan atau memiliki hubungan dalam mencari probabilitas responden menerima vaksin H1N1 dan *seasonal flu*. Variabel independen yang digunakan untuk mencari probabilitas responden menerima vaksin H1N1 adalah h1n1_concern, h1n1_knowledge, opinion_h1n1_vacc_effective, opinion_h1n1_risk, opinion_h1n1_sick_from_vacc, age_group, dan sex. Sedangkan, variabel independen yang digunakan untuk mencari probabilitas responden menerima vaksin *seasonal flu* adalah h1n1_concern, h1n1_knowledge, opinion_seas_vacc_effective, opinion_seas_risk, opinion_seas_sick_from_vacc, age_group, dan sex.

Pada data yang diberikan, terdapat 36 variabel yang dapat digunakan sebagai variabel independen. Sehubungan dengan keterbatasan waktu, maka dalam proses pemilihan variabel yang digunakan hanya dilihat berdasarkan alasan kualitatifnya saja. Ketika dilihat dari sudut pandang data yang diperoleh melalui survei, terlihat sulit untuk bisa memperoleh seluruh 36 variabel tersebut. Oleh karena hal tersebut, dilakukan pemilihan variabel dengan pertimbangan variabel yang mudah diperoleh melalui survei. Selain itu, pada variabel terpilih dilakukan pengecekan korelasi untuk melihat korelasi antara variabel tersebut dengan variabel dependen/responnya. Pada model ini, dipilih 7 variabel independen untuk mencari probabilitas responden terhadap vaksin H1N1 dan 7 variabel independen untuk probabilitas responden terhadap vaksin *seasonal flu*. Walaupun demikian, akan lebih baik jika dapat dilakukan pemilihan data secara kuantitatif sehingga dapat memperoleh hasil yang lebih baik dan akurat.

Tabel 1. Tabel Korelasi Vaksin H1N1 dengan 7 Variabel Independen yang digunakan

	h1n1_vaccine
h1n1_concern	0.1215738
h1n1_knowledge	0.1177712
opinion_seas_vacc_effective	0.2673517
opinion_seas_risk	0.3205801
opinion_seas_sick_from_vacc	0.07458022

Tabel 2. Tabel Korelasi Vaksin *Seasonal Flu* dengan 7 Variabel Independen yang digunakan

	seasonal_vaccine
h1n1_concern	0.1544882
h1n1_knowledge	0.1197787
opinion_seas_vacc_effective	0.3588688
opinion_seas_risk	0.3869157
opinion_seas_sick_from_vacc	-0.06053783

2. Mempersiapkan Data / *Cleaning Data*

Tahap pengolahan data bertujuan untuk mengolah data mentah agar dapat digunakan untuk menghitung probabilitas. Data mentah tersebut akan dibersihkan jika terdapat *missing values*, melakukan pengolahan terhadap variabel yang memiliki nilai N/A dimana nilai N/A tersebut akan digantikan dengan nilai median dikarenakan model logistik tidak dapat bekerja dengan *missing values*. Variabel-variabel yang sudah terpilih dan diolah kemudian akan dimasukkan kedalam satu data frame yang baru dengan tujuan untuk memudahkan proses pengerjaan pemodelan.

3. Modeling

Tahap pemodelan data ini dilakukan dengan menggunakan algoritma Naive Bayes dan Regresi Logistik. Pada tahap pemodelan Naive Bayes akan dilakukan *K-Fold Cross Validation* terhadap data *training* dan data *testing*, sedangkan pemodelan Regresi Logistik

akan dilakukan dengan variabel yang sudah terpilih. Model yang telah dibangun akan digunakan untuk memprediksi nilai vaksin H1N1 dan *seasonal flu*.

4. Evaluasi

Tahap evaluasi bertujuan untuk mengetahui nilai probabilitas mana yang lebih baik dari responden penerima vaksin H1N1 dan *seasonal flu* apabila ditinjau menggunakan metode Naive Bayes dan Regresi Logistik. Melalui penelitian ini, modeling akan dibandingkan dengan menggunakan nilai *Receiver Operating Characteristic (ROC) Curve* dan *Area Under the Curve (AUC)* yang diperoleh.

Matriks Konfusi merupakan tabel yang digunakan untuk mengetahui ketepatan dalam pengklasifikasian antara nilai prediksi dan nilai observasi[6] seperti diberikan pada Tabel 3.

Tabel 3. Tabel Matriks Konfusi

Klasifikasi		Nilai Prediksi	
		Benar	Salah
Nilai Observasi	Benar	<i>True Positive (TP)</i>	<i>False Negative (FN)</i>
	Salah	<i>False Positive (FP)</i>	<i>True Negative (TN)</i>

Dengan nama variabel sebagai berikut:

TP : jumlah label prediksi bernilai benar dan label observasi juga bernilai benar

FN : jumlah label prediksi bernilai salah tetapi label observasi bernilai benar

FP : jumlah label prediksi bernilai benar tetapi label observasi bernilai salah

TN : jumlah label prediksi bernilai salah dan label observasi juga bernilai salah

Kurva ROC digunakan sebagai teknik untuk mengklasifikasi berdasarkan hasil prediksi. ROC akan dipetakan ke dalam bidang dua dimensi dengan nilai *true positif* (TP) pada sumbu Y dan *false positif* (FP) pada sumbu X. Sedangkan AUC merupakan luas daerah di bawah kurva yang dihasilkan oleh ROC dan nilainya selalu berada pada rentang 0 hingga 1. Dalam klasifikasi data mining, nilai AUC dapat dibagi ke dalam beberapa kelompok[7] seperti yang diberikan pada Tabel 4.

Tabel 4. Kategori Pengklasifikasian berdasarkan Nilai AUC dalam Data Mining

Nilai AUC	Klasifikasi
0.90 - 1.00	sangat baik
0.80 - 0.90	baik
0.70 - 0.80	cukup

0.60 - 0.70	buruk
0.50 - 0.60	salah

4. **Hasil dan Analisis.** Setelah melihat, memilih, mempersiapkan, serta melakukan modeling pada data yang telah tersedia, data akan diproses menggunakan R Studio untuk memperoleh hasil yang akan dianalisis. Hal yang pertama kali dilakukan adalah membangun model Naive Bayes untuk model y_1 dan y_2 . Kemudian dari model tersebut dicari nilai *fitted values* dan *predicted values* masing-masing. Nilai *fitted values* dan *predicted values* yang telah diperoleh akan digunakan untuk mencari nilai pada matriks konfusi. Matriks konfusi yang diperoleh dari variabel y_1 terhadap *fitted value* dengan nilai akurasi sebesar 1 tertera pada Tabel 5 dan untuk variabel y_1 terhadap *predicted values* dengan nilai akurasi sebesar 0.9994 tertera pada Tabel 6. Sama halnya dengan variabel y_1 , matriks konfusi yang diperoleh dari variabel y_2 terhadap *fitted values* dengan nilai akurasi sebesar 0.9997 tertera pada Tabel 7 dan untuk variabel y_2 terhadap *predicted values* dengan nilai akurasi sebesar 0.9998 tertera pada Tabel 8. Nilai akurasi yang diperoleh dapat digunakan untuk melihat tingkat akurasi model yang akan diprediksi.

Tabel 5. Tabel Matriks Konfusi Variabel y_1 terhadap *fitted values*

Klasifikasi		Nilai Prediksi	
		Benar	Salah
Nilai Observasi	Benar	16826	0
	Salah	1	4540

Tabel 6. Tabel Matriks Konfusi Variabel y_1 terhadap *predicted values*

Klasifikasi		Nilai Prediksi	
		Benar	Salah
Nilai Observasi	Benar	4203	0
	Salah	3	1134

Tabel 7. Tabel Matriks Konfusi Variabel y_2 terhadap *fitted values*

Klasifikasi		Nilai Prediksi	
		Benar	Salah
Nilai Observasi	Benar	11416	4
	Salah	2	9944

Tabel 8. Tabel Matriks Konfusi Variabel y_2 terhadap *predicted values*

Klasifikasi		Nilai Prediksi	
		Benar	Salah
Nilai Observasi	Benar	2854	1
	Salah	0	2486





Tahapan selanjutnya yang dilakukan adalah mencari nilai AUC dari model Naive Bayes yang digunakan. Dengan menggunakan fungsi `roc_auc` di R Studio, diperoleh nilai AUC dari variabel y_1 dan y_2 , yaitu sebesar 0,9997212. Berdasarkan kategori pengklasifikasian nilai AUC yang terdapat pada Tabel 4, nilai tersebut berada pada rentang 0,90 - 1,00 sehingga masuk dalam klasifikasi sangat baik. Hal tersebut menandakan bahwa model yang digunakan sangat baik dalam membedakan kelas benar dan salah.

Setelah membangun model Naive Bayes dan melakukan analisis terhadap hasil yang diperoleh, data akan dimodelkan kembali dengan model Regresi Logistik untuk dilakukan perbandingan. Selanjutnya dari data dan model yang telah dibangun, akan dicari nilai *predicted values* terhadap variabel y_1 dan y_2 . Dari nilai tersebut akan digunakan fungsi `roc_auc` di R Studio untuk mencari nilai AUC dari model Regresi Logistik. Nilai AUC yang diperoleh adalah 0,6819491. Berdasarkan kategori pengklasifikasian nilai AUC yang terdapat pada Tabel 4, nilai tersebut berada pada rentang 0,60 - 0,70 sehingga masuk dalam klasifikasi buruk.

Perbedaan nilai AUC yang diperoleh dari kedua model (Naive Bayes dan Regresi Logistik) tergolong cukup besar, yaitu 0,3177721. Selain dilihat dari perbedaan angka, dapat juga dilihat bahwa nilai AUC Naive Bayes masuk dalam klasifikasi sangat baik sedangkan nilai AUC Regresi Logistik masuk dalam klasifikasi buruk. Hal tersebut menandakan bahwa model yang lebih baik

dalam memprediksi variabel y_1 dan y_2 adalah model Naive Bayes. Selain itu, dapat juga dilihat dari *predicted values* yang telah diperoleh dari kedua model dimana *predicted values* dari variabel y_1 dan y_2 digabungkan dalam data frame yang baru untuk dikumpulkan dalam *competition Flu Shot Learning : Predict H1N1 and Seasonal Flu Vaccines*. *File submission* yang didapat dari menggunakan model Naive Bayes memperoleh nilai 0,8152 sedangkan *file submission* yang didapat dari menggunakan model Regresi Logistik adalah 0,6582. Terlihat bahwa hasil yang didapat dengan model Naive Bayes lebih baik dibandingkan model Regresi Logistik. Dengan menggunakan *file submission* dengan model Naive Bayes, diperoleh ranking 1053 Pada Gambar 1 terlampir *screenshot leaderboard* pada kompetisi tersebut.

Gambar 1. *Screenshot Leaderboard* Kelompok 1 dalam *Competition Flu Shot Learning : Predict H1N1 and Seasonal Flu Vaccines*

User or team		Best public l ₁ AUROC ⓘ	Timestamp ⓘ	# Entries
	Juane99_2	1051 0.8152	2022-01-12 10:57:45	3
	abcde1234	1052 0.8152	2022-06-03 07:18:31	3
	JeffreyElisalrenne	1053 0.8152	2022-06-03 09:18:31	2
	UGR_SA	1054 0.8149	2022-01-14 12:58:37	3

Pada Tabel 9 tertera cuplikan *predicted values* yang telah diperoleh dengan menggunakan model Naive Bayes dan pada Tabel 10 tertera cuplikan *predicted values* yang telah diperoleh dengan menggunakan model Regresi Logistik.

Tabel 9. Cuplikan *Predicted Values* yang diperoleh dari Model Naive Bayes

respondent_id	h1n1_vaccine	seasonal_vaccine
26707	0.0524305345	1.822444e-01
26708	0.0001477479	9.657739e-05
26709	0.6346189698	8.258033e-01
26710	0.3328776463	9.537287e-01

26711	0.6640647265	7.576616e-01
26712	0.9869571822	9.935735e-01

Tabel 10. Cuplikan *Predicted Values* yang diperoleh dari Model Regresi Logistik

respondent_id	h1n1_vaccine	seasonal_vaccine
26707	0.21697377	0.41832970
26708	0.07474886	0.09637859
26709	0.60606354	0.82286375
26710	0.14424452	0.75320794
26711	0.22613071	0.41773985
26712	0.38372564	0.89492689

5. **Kesimpulan.** Dalam mencari probabilitas responden menerima vaksin H1N1 dan *seasonal flu*, model yang digunakan adalah model Naive Bayes dan Regresi Logistik. Dari data yang tersedia dilakukan pemodelan sesuai dengan model yang digunakan untuk mencari *predicted values* dan juga nilai AUC untuk membandingkan kedua model yang digunakan. Pada model Naive Bayes diperoleh nilai AUC sebesar 0,9997212 yang masuk klasifikasi sangat baik sedangkan pada model Regresi Logistik diperoleh nilai AUC sebesar 0,6819491 yang masuk klasifikasi buruk. Berdasarkan nilai AUC, model Naive Bayes adalah model yang lebih baik untuk digunakan pada data ini.

Selain AUC, perbandingan kedua model juga dilihat dari *predicted values* yang dikumpulkan dalam *competition Flu Shot Learning : Predict H1N1 and Seasonal Flu Vaccines*. Berdasarkan nilai *submission, file submission* dengan model Naive Bayes memperoleh nilai yang lebih baik yaitu 0,8152 dan *file submission* dengan model Regresi Logistik memperoleh nilai 0,6582. Oleh karena itu, model yang lebih baik digunakan dalam mencari probabilitas responden menerima vaksin H1N1 dan *seasonal flu* adalah model Naive Bayes.

DAFTAR PUSTAKA

- [1] N. Rajagukguk, D. Ispriyanti dan Y. Wilandari, Perbandingan Metode Klasifikasi Regresi Logistik Biner dan Naive Bayes pada Status Pengguna KB di Kota Tegal Tahun 2014, *Jurnal Gaussian*, vol.4, no.2, pp.365-374, 2015
- [2] M. Bramer, *Principles of Data Mining*, Springer, London, 2007
- [3] J. C. Stoltzfus, Logistic Regression: A Brief Primer, *Society for Academic Emergency Medicine*, vol.18, pp.1099-1104, 2011
- [4] N. R. Indraswari dan Y. I. Kurniawan, Aplikasi Prediksi Usia Kelahiran dengan Metode Naive Bayes, *Jurnal SIMETRIS*, vol.9, no.1, pp.129-138, 2018
- [5] W. Mendenhall dan T. Sincich, *A Second Course in Statistics : Regression Analysis Seventh Edition*, Prentice Hall, Boston, 2011
- [6] N. Frastian, Implementasi Komparasi Algoritma Klasifikasi Menentukan Kelulusan Mata Kuliah Algoritma Universitas Budi Luhur, *Jurnal String*, vol.3, no.1, pp.4-5, 2018
- [7] L. Qadrini, A. Seppewali dan A. Aina, Decision Tree dan Adaboost pada Klasifikasi Penerima Program Bantuan Sosial, *Jurnal Inovasi Penelitian*, vol.2, no.7, pp.1963, 2021