Jurnal Komputer dan Informatika
Vol 17 No 1, April 2022: hlm 39-44

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

# KLASIFIKASI PENDERITA *MONKEYPOX* MENGGUNAKAN KNN, *NAIVE BAYES*, DAN *RANDOM FOREST*

**Noel**
Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Tarumanagara
Jl. Letjen S.Parman No.1, Jakarta Barat, DKI Jakarta, Indonesia 11410
e-mail: *noel.53521010@stu.untar.ac.id*

## ABSTRAK

Cacar monyet adalah infeksi virus zoonosis yang disebabkan oleh virus cacar monyet yang menyebabkan ruam yang mirip dengan cacar. Namun, penyebaran dari orang ke orang di luar kontak dekat langsung dan tingkat kematian jauh lebih rendah pada cacar monyet dibandingkan dengan infeksi cacar. Pada kasus ini menggunakan tiga metode klasifikasi, yaitu *K-Nearest Neighbors*, *Logistic Regression*, dan *Random Forest*, untuk memprediksi kasus *Monkeypox*. Pengumpulan dataset dari dataset pada *Kaggle* dengan terdiri dari 25.000 data, dengan 11 atribut dan dua kelas: negatif dan positif,. Dalam konteks ini, metode klasifikasi digunakan untuk mengetahui bagaimana agar metode yang digunakan dapat memprediksi dengan akurasi yang baik berdasarkan dataset yang digunakan untuk melatih dan menguji dibagi menjadi data latih 60% dan data uji 40% dengan menggunakan *test size* = 0.4 sebagai parameter saat membagi data. Hasil dari penelitian ini menunjukkan bahwa metode klasifikasi Random Forest menghasilkan nilai akurasi paling tinggi dengan menggunakan parameter akurasi, presisi, *recall*, dan *f1-score*. Nilai akurasi yang diperoleh sebesar 67%..

**Kata kunci:** Cacar Monyet , Klasifikasi, *K-Nearest Neighbors, Logistic Regression, Random Forest*

## *ABSTRACT*

*Monkey pox is a zoonotic viral infection caused by the monkey pox virus that causes a rash similar to smallpox. However, the person-to-person spread is beyond direct close contact and the mortality rate is much lower in monkey pox compared to smallpox infection. This case uses three classification methods, namely K-Nearest Neighbors, Logistic Regression, and Random Forest, to predict Monkeypox cases. The dataset was collected from a dataset on Kaggle consisting of 25,000 data, with 11 attributes and two classes: negative and positive, respectively. In this context, the classification method is used to find out how the method can predict with good accuracy based on the dataset used for training and testing divided into 60% training data and 40% test data by using test size = 0.4 as a parameter when dividing the data. The results of this study show that the Random Forest classification method produces the highest accuracy value using the accuracy, precision, recall, and f1-score parameters. The accuracy value obtained is 67%.*

***Keywords:*** *Monkeypox , Classification, K-Nearest Neighbors, Logistic Regression, Random Forest.*

## 1. PENDAHULUAN

Cacar monyet adalah infeksi virus zoonosis yang disebabkan oleh virus cacar monyet yang menyebabkan ruam yang mirip dengan cacar. Namun, penyebaran dari orang ke orang di luar kontak dekat langsung dan tingkat kematian jauh lebih rendah pada cacar monyet dibandingkan dengan infeksi cacar. Cacar monyet pertama kali dideskripsikan pada tahun 1958 pada monyet-monyet yang dikirim dari Singapura ke Denmark. Selama dekade berikutnya, wabah tambahan dilaporkan terjadi pada monyet-monyet yang dikurung di Amerika Serikat, Belanda, dan Perancis. Kasus pertama infeksi cacar monyet pada manusia dilaporkan pada tahun 1970 di Republik Demokratik Kongo pada seorang anak laki-laki berusia 9 bulan yang merupakan satu-satunya anggota keluarganya yang belum mendapatkan vaksinasi cacar. Menerima vaksinasi cacar sebelumnya diperkirakan 85% efektif dalam mencegah cacar monyet, meskipun kemanjuran jangka panjang vaksinasi cacar tidak jelas [1].

Pada tanggal 6 Mei 2022, kasus cacar monyet pertama dalam wabah multi-negara dikonfirmasi di Inggris pada seorang pria yang berasal dari Nigeria. Penelusuran kontak dilakukan dan

menemukan kasus-kasus lain di rumah tangganya. Namun, dalam beberapa hari berikutnya, kasus-kasus tanpa riwayat perjalanan yang terdokumentasi ke negara-negara dengan cacar monyet endemik dilaporkan di Inggris, menunjukkan penularan lokal yang tidak terdeteksi. Kasus-kasus baru juga terdeteksi di Portugal, Amerika Serikat, dan beberapa negara lainnya. Kasus pertama di Inggris awalnya dianggap kasus indeks, tetapi hipotesis ini ditolak karena tanggal timbulnya gejala yang lebih awal dilaporkan pada kasus-kasus di Portugal dan Inggris. Selain itu, adanya deteksi cacar monyet pada orang yang tidak berhubungan menunjukkan penyebaran tanpa gejala. Organisasi Kesehatan Dunia (WHO) dan lembaga kesehatan masyarakat lainnya telah meningkatkan kewaspadaan sejak 16 Mei 2022. Wabah saat ini disebabkan oleh virus Clade 3, yang berasal dari Afrika Barat. Pada tanggal 23 Juli 2022, WHO mengumumkan keadaan darurat kesehatan global [2]. Prevalensi HIV pada orang dengan cacar monyet mencapai 38%, dan 41% di antaranya telah didiagnosis dengan satu atau lebih infeksi menular seksual pada tahun sebelumnya. Sebanyak 94% orang dengan cacar monyet yang terinfeksi HIV telah mendapatkan perawatan HIV pada tahun sebelumnya, dan 82% memiliki viral load HIV yang rendah, menunjukkan bahwa virus HIV mereka terkendali. Proporsi orang dengan infeksi HIV yang terdaftar sebagai pasien rumah sakit lebih tinggi dibandingkan dengan mereka yang tidak terinfeksi HIV. Temuan ini menunjukkan pentingnya menggunakan sistem perawatan dan pencegahan HIV dan IMS untuk mengurangi kasus cacar monyet di kalangan populasi terkait. Memprioritaskan orang dengan infeksi HIV dan IMS untuk mendapatkan vaksin cacar monyet juga perlu dipertimbangkan [3] [4]. Dalam kasus ini menggunakan 3 metode untuk melakukan klasifikasi kategori. Tujuan dari klasifikasi adalah untuk melatih dan menguji model klasifikasi, yaitu *Logistic Regression*, *Random Forest*, dan *K-Nearest Neighbors*.

## 2. METODE PENELITIAN

### 3.1 Dataset

Data yang digunakan dalam eksperimen ini diambil dari Kaggle. Dataset memiliki 25000 data pasien dengan 11 atribut dan 2 kelas: *Positive* dan *Negative*. Fitur yang dipilih adalah *Systemic Illness, Rectal Pain, Sore Throat, Penile Oedema, Oral Lesions, Solitary Lesion, Swollen Tonsils, HIV Infection,* dan *Sexually Transmitted Infection*.

### 3.2 Algoritma Klasifikasi

Ketika perkiraan parametrik dari kepadatan probabilitas yang jelas tidak diketahui atau sulit untuk ditemukan, klasifikasi KNN dibuat untuk melakukan analisis karakteristik. Algoritma KNN adalah metode pembelajaran berbasis contoh yang digunakan untuk mengklasifikasikan objek berdasarkan contoh pelatihan terdekat dalam ruang fitur. Algoritma KNN menggunakan metrik jarak Euclidean untuk menemukan tetangga terdekat. Rumus dalam persamaan berikut ini digunakan untuk menghitung jarak geometris [6].

$$d(x_i, y_i) = \sqrt{\sum_{i=1}^{n}(x_i, y_i)^2},$$

dimana, $x$ adalah data uji dan $y$ adalah data latih. Sedangkan $i$ adalah iterasi data ke-$i$ dan $n$ merupakan jumlah data latih.

Untuk klasifikasi, regresi logistik digunakan. Tidak seperti regresi linier, data dalam regresi logistik tidak disusun dalam baris. Dengan menggunakan model ini, signifikansi statistik dari setiap variabel independen dibandingkan dengan probabilitas dihitung. Sangat efektif untuk memodelkan hasil binomial. Sebagai contoh, apakah individu mengalami WNV atau tidak dengan mengambil nilai 0 dan 1 dengan menggunakan satu atau lebih variabel penjelas? Contoh dibawah ini menggambarkan persamaan dari regresi logistik [7].

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

$$L_n\left(\frac{p}{1-p}\right) = B_0 + B_1X,$$

dimana, $B_0$ adalah konstanta dan $B_1$ adalah koefisien dari masing-masing variabel dengan nilai $p$ yang dapat diperoleh melalui persamaan berikut ini.

$$p = \frac{e^{B_0+B_1X}}{1+e^{B_0+B_1X}},$$

*Random Forest* adalah prosedur pembelajaran mesin yang populer yang dapat digunakan untuk mengembangkan model prediksi. *Random Forest* dibangun dengan mengkombinasikan hasil dari berbagai *Decision Tree*. Setiap *Decision Tree* menentukan prediktor label kelas untuk contoh baru (disebut sebagai suara). *Decision Tree* didapat dengan menghitung nilai entropi sebagai penentu nilai *information gain* dan tingkat ketidakmurnian atribut. Rumus pada persamaan dibawah ini digunakan untuk menghitung nilai entropi [8] [9].

$$Entropy\ (Y) = -\sum_i\ p(c|Y)\log^2 p(c|Y),$$

dimana, $Y$ merupakan himpunan kasus dan $p(c|Y)$ adalah proporsi nilai $Y$ terhadap kelas $c$. Sedangkan, nilai *information gain* dapat digambarkan melalui persamaan beriku ini.

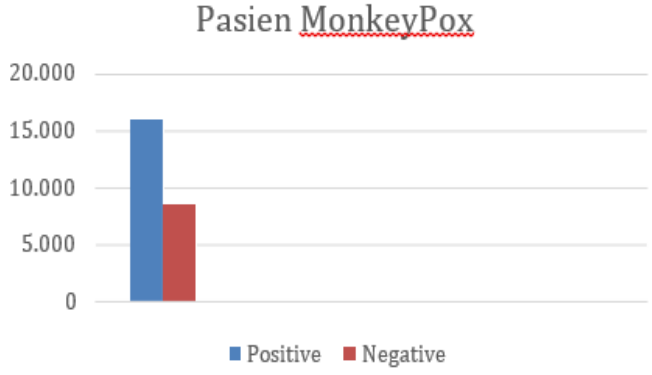
$$Information\ Gain\ (Y, a) = Entropy\ (Y) - \sum_{\cup}\ \varepsilon Values\ (a)\ \left(\frac{|Y_v|}{|Y_a|}\right) Entropy\ (Y_v)\ ,$$

dimana, $Values(a)$ menjadi nilai yang mungkin dalam himpunan kasus $a$ dengan $Y_v$ menjadi subkelas dari $Y$ dengan kelas $v$ yang berhubungan dengan kelas $a$ dan $Y_a$ adalah semua nilai yang sesuai dengan kelas $a$.

Metode Evaluasi yang akan dilakukan adalah membandingkan hasil percobaan. Dimulai dengan preprocessing data agar memudahkan data untuk dipahami komputer, selanjutnya melakukan *encode* untuk mengubah variabel menjadi angka, setelah itu membagi data menjadi data latih dan data uji (dengan perbandingan 60.40). Kesimpulan didapatkan dari hasil pengujian pada tiga model algoritma. Data hasil pengukuran tersebut kemudian digunakan sebagai panduan untuk menentukan hasil prediksi dalam penelitian ini.

## 3. HASIL DAN PEMBAHASAN

Hasil pengujian pasien *Monkeypox* ditunjukkan melalui grafik yang dapat dilihat seperti pada Gambar 1 berikut.



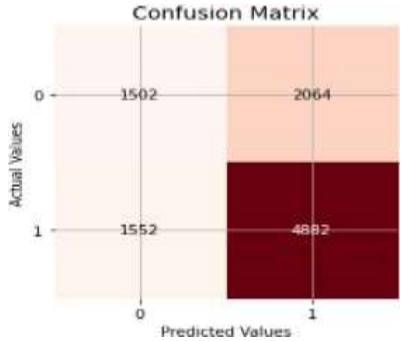
**Gambar 1.** Hasil Pengujian Pasien *Monkeypox*

Hasil nilai akurasi dari ketiga metode yang digunakan sangat baik, bisa dilihat pada Tabel 1. berikut.

**Tabel 1.** *Scatterplot* Temperatur Minimum

| Metode | Akurasi |
|---|---|
| *K-Nearest Neighbors* | 63% |
| *Logistic Regression* | 67% |
| *Random Forest* | 67% |

Hasil pengujian yang didapat menggunakan metode KNN dengan pembagian data latih dan data uji dengan perbandingan 6:4 dengan jumlah tetangga 5, memperoleh nilai akurasi terendah jika dibandingkan dengan dua metode lainnya, yaitu 63%. Hasil evaluasi dengan *confusion matrix* dan *classification report* masing-masing dapat dilihat pada Gambar 2 dan Gambar 3 berikut.



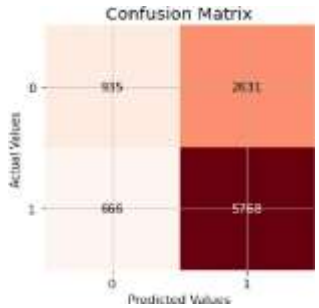

**Gambar 2.** Hasil Evaluasi KNN dengan *Confusion Matrix*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.49 | 0.42 | 0.45 | 3566 |
| 1 | 0.70 | 0.76 | 0.73 | 6434 |
| accuracy | | | 0.64 | 10000 |
| macro avg | 0.60 | 0.59 | 0.59 | 10000 |
| weighted avg | 0.63 | 0.64 | 0.63 | 10000 |

accuracy score=0.6384

**Gambar 3.** Hasil Evaluasi KNN dengan *Classification Report*

Hasil pengujian yang didapat menggunakan metode *logistic regression* dengan pembagian data latih dan data uji dengan perbandingan 6:4 diperoleh nilai akurasi yaitu 67,03%. Hasil evaluasi dari metode ini dapat dilihat masing-masing pada Gambar 4 dan Gambar 5.




**Gambar 4.** Hasil Evaluasi *Logistic Regression* dengan *Confusion Matirx*

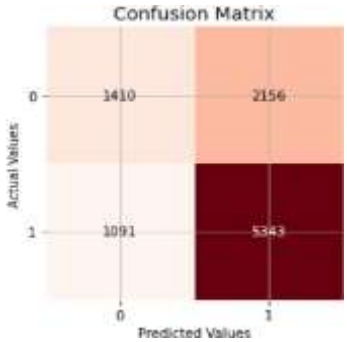
ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.58 | 0.26 | 0.36 | 3566 |
| 1 | 0.69 | 0.90 | 0.78 | 6434 |
| accuracy | | | 0.67 | 10000 |
| macro avg | 0.64 | 0.58 | 0.57 | 10000 |
| weighted avg | 0.65 | 0.67 | 0.63 | 10000 |

accuracy score=0.6703

**Gambar 5.** Hasil Evaluasi *Logistic Regression* dengan *Classification Report*

Hasil pengujian yang didapat menggunakan metode *random forest* dengan pembagian data latih dan data uji dengan perbandingan 6:4 diperoleh nilai akurasi yang hampir sama dengan *logistic regression* yaitu 67,53%. Kedua metode ini dapat dikatakan sebanding dan menghadapi masalah dengan baik, kedua metode mungkin mampu menggeneralisasi dengan baik dan menghasilkan prediksi yang serupa. Hasil evaluasi dari metode ini untuk *confusion matrix* dan *classification report* masing-masing dapat dilihat melalui Gambar 6 dan Gambar 7 dibawah.




**Gambar 6.** Hasil Evaluasi *Random Forest* dengan *Confusion Matrix*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.56 | 0.40 | 0.46 | 3566 |
| 1 | 0.71 | 0.83 | 0.77 | 6434 |
| accuracy | | | 0.68 | 10000 |
| macro avg | 0.64 | 0.61 | 0.62 | 10000 |
| weighted avg | 0.66 | 0.68 | 0.66 | 10000 |

accuracy score=0.6753

**Gambar 7.** Hasil Evaluasi *Random Forest* dengan *Classification Report*

## 4. KESIMPULAN

Dalam konteks ini, metode klasifikasi digunakan untuk mengetahui bagaimana agar metode yang digunakan dapat memprediksi dengan akurasi yang baik berdasarkan dataset yang digunakan untuk pelatihan dan pengujian dibagi menjadi data pelatihan (60% dari total data) dan data uji (40% dari total data) dengan menggunakan *test size* = 0.4 sebagai parameter saat membagi data, yaitu sekitar 67% untuk *Random Forest* dan *Logistic Regression* serta 63% untuk *K-Nearest Neighbors*. Studi ini diharapkan dapat memberikan pemahaman yang lebih baik tentang performa tiga metode klasifikasi yang berbeda (*Logistic Regression, Random Forest*, dan *K-Nearest Neighbors*) dalam memprediksi banyaknya kasus *Monkeypox*.

Kesimpulan yang diperoleh dari hasil klasifikasi banyaknya kasus *Monkeypox*, yang diklasifikasikan yaitu menentukan apakah hasil pengujian positif atau negatif dengan menggunakan tiga metode klasifikasi berbeda adalah metode klasifikasi *Random Forest* menghasilkan nilai akurasi

paling tinggi dengan menggunakan parameter. Nilai akurasi yang diperoleh sebesar 67.53%. Sedangkan nilai akurasi dari dua metode lainnya, yaitu 67% juga untuk *Logistic Regression* dan 63 % untuk metode KNN yang merupakan metode yang paling kecil. Jadi *Random Forest* merupakan metode yang paling cocok untuk digunakan dalam penelitian ini

## DAFTAR PUSAKA

[1] Curran, K. G., Eberly, K., Russell, O. O., Snyder, R. E., Phillips, E. K., Tang, E. C., ... & STI Team. (2022). HIV and sexually transmitted infections among persons with monkeypox— eight US jurisdictions, May 17–July 22, 2022. Morbidity and Mortality Weekly Report, 71(36), 1141.

[2] Gessain, A., Nakoune, E., & Yazdanpanah, Y. (2022). Monkeypox. New England Journal of Medicine, 387(19), 1783-1793.

[3] Rizk, J. G., Lippi, G., Henry, B. M., Forthal, D. N., & Rizk, Y. (2022). Prevention and treatment of monkeypox. Drugs, 82(9), 957-963.

[4] Sherwat, A., Brooks, J. T., Birnkrant, D., & Kim, P. (2022). Tecovirimat and the treatment of monkeypox—past, present, and future considerations. New England Journal of Medicine, 387(7), 579-581.

[5] Boateng, E. Y., Otoo, J., & Abaye, D. A. (2020). Basic tenets of classification algorithms K-nearest-neighbor, support vector machine, random forest and neural network: a review. Journal of Data Analysis and Information Processing, 8(4), 341-357.

[6] Zou, X., Hu, Y., Tian, Z., & Shen, K. (2019, October). Logistic regression model optimization and case analysis. In 2019 IEEE 7th international conference on computer science and network technology (ICCSNT) (pp. 135-139). IEEE.

[7] Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A comparative analysis of logistic regression, random forest and KNN models for the text classification. Augmented Human Research, 5, 1-16.

[8] Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. Expert systems with applications, 134, 93-101.