

Penerapan Python Dalam Data Mining Untuk Prediksi Kanker Paru

Baiq Andrisca Candra Permana^{1*}, Muhammad Djamaluddin²

^{1,2}Program Studi Teknik Informatika, Universitas Hamzanwadi

andrisca.cp@hamzanwadi.ac.id

Abstrak

Kanker paru merupakan salah satu dari kelompok penyakit kanker yang paling banyak menyebabkan kematian termasuk di Indonesia. Banyak masyarakat penderita kanker paru tidak menyadari dirinya terinfeksi kanker paru yang mengakibatkan penanganan terhadap penyakit ini menjadi terlambat. Untuk itu perlu kiranya suatu metode yang memiliki tingkat akurasi yang baik dalam melakukan suatu prediksi sehingga nantinya dengan tingkat akurasi yang baik tersebut dapat menjadi acuan untuk dapat dikembangkannya suatu Artificial Intelligence (AI) dalam dunia kesehatan untuk mendeteksi dini kanker paru. Penelitian yang diusulkan menggunakan algoritma c4.5 untuk melakukan prediksi kemungkinan pasien penderita kanker paru dengan memberikan hasil akhir berupa tingkat akurasi prediksi dari algoritma yang diusulkan. Untuk melakukan implementasi datamining menggunakan bahasa pemrograman Python dengan memanfaatkan library yang telah disediakan untuk memudahkan melakukan implementasi machine learning. Dalam penelitian ini penggunaan c4.5 mampu memprediksi dengan tingkat akurasi yaitu sebesar 86%. Tingkat akurasi tersebut dapat dikatakan layak untuk dijadikan acuan untuk dapat memprediksi penderita kanker paru berdasarkan faktor-faktor gejala yang tampak pada pasien.

Kata Kunci: machine learning, c4.5, python, kanker paru

Abstract

Lung cancer is one of the groups of cancer that causes the most deaths, including in Indonesia. Many people with lung cancer do not realize that they are infected with lung cancer, which causes delays in treating this disease. For this reason, it is necessary to have a method that has a good level of accuracy in making a prediction so that later with a good level of accuracy it can be a reference for the development of an Artificial Intelligence (AI) in the world of health to detect lung cancer early. The proposed study uses the c4.5 algorithm to predict the likelihood of patients with lung cancer by providing the final result in the form of the prediction accuracy of the proposed algorithm. To carry out data mining implementation using the Python programming language by utilizing the library that has been provided to make it easier to implement machine learning. In this study the use of c4.5 was able to predict with an accuracy rate of 86%. This level of accuracy can be said to be worthy of being used as a reference to be able to predict lung cancer patients based on the symptoms that appear in the patient.

Keywords: machine learning, c4.5, python, lung cancer

1. Pendahuluan

Organisasi Kesehatan Dunia (WHO) menyatakan bahwa kanker merupakan kelompok penyakit yang berasal dari hampir seluruh organ tubuh dimana sel-sel yang terdapat pada organ tubuh tersebut tumbuh dan berkembang secara tidak

normal. Sel yang abnormal akan menyerang bagian tubuh yang ada disekitarnya. Menurut data WHO jumlah kematian yang disebabkan oleh penyakit kanker diseluruh dunia mencapai 1,8 kematian pada tahun 2022 dan kemungkinan akan terus meningkat pada tahun 2030 [1].

Salah satu kelompok penyakit kanker yang paling banyak mengakibatkan terjadinya kematian adalah kanker paru. Kanker paru merupakan penyakit keganasan yang dapat diakibatkan oleh penyakit pada paru itu sendiri atau dapat diakibatkan oleh penyakit diluar paru. Indonesia merupakan negara ke-4 dengan jumlah penderita kanker paru terbanyak di seluruh dunia dimana penderitanya sebagian besar pria usia > 40 tahun [2] [3].

Oleh sebab itu maka perlu kiranya suatu model prediksi untuk mendeteksi penyakit paru secara dini sehingga mencegah jumlah kematian yang diakibatkan oleh kanker paru. Penelitian sebelumnya telah dilakukan Eva dan Andreas, dimana penelitian dilakukan untuk memprediksi penyakit paru menggunakan algoritma Naïve bayes dan menghasilkan tingkat akurasi 73%.

Penelitian ini menggunakan algoritma c4.5 karena c4.5 dikenal memiliki akurasi yang baik dalam prediksi, dengan harapan dengan tingginya akurasi dalam melakukan prediksi pengembangan dapat dilakukan dalam dunia kesehatan sehingga deteksi dini kanker paru dapat dilakukan dan mendapat penanganan lebih awal.

2. Tinjauan Pustaka

2.1. Penelitian Terkait

Penelitian telah dilakukan oleh Ai dan Agus tahun (2017) yaitu melakukan prediksi kekambuhan

penyakit kanker payudara menggunakan algoritma c4.5. Berdasarkan penelitian yang telah dilakukan diperoleh hasil berupa nilai akurasi 75,5 % yang lebih baik jika dikomparasi dengan naïvebayes yang menghasilkan akurasi 72,7 % [4].

Penelitian yang dilakukan Noviandi pada (2018) yaitu membangun suatu model prediksi menggunakan c4.5 untuk melihat kemungkinan seorang pasien menderita penyakit diabetes dan menentukan tingkat akurasi dari algoritma. Berdasarkan penelitian diperoleh akurasi algoritma dalam prediksi penderita diabetes adalah sebesar 70.32% [5]

Penelitian selanjutnya dilakukan Andriska dan intan (2021) dimana melakukan komparasi algoritma naïve bayes dan decision tree untuk penyakit diabetes. Hasil penelitian yang telah dilakukan diperoleh bahwa akurasi yang dimiliki decision tree lebih tinggi jika dibandingkan dengan naïve bayes. Implementasi algoritma yang digunakan menggunakan Rapid Miner [6].

Penelitian selanjutnya dilakukan oleh Dwi dan Astried (2021), pada penelitian yang dilakukan untuk memprediksi penyakit jantung pasien berdasarkan kondisi medis yang dialami. Prediksi yang dilakukan menggunakan SVM dengan melakukan pengimplementasian dengan menggunakan python. Hasil yang diperoleh adalah nilai metrik akurasi sama dengan

90.11%, presisi 90.38% dan recall 92.15% dengan kernel linear [7].

Penelitian dilakukan oleh Yuri (2022) melakukan prediksi kelangsungan hidup pasien gagal ginjal dengan menggunakan seleksi fitur bestfirst dimana didapatkan terdapat 4 faktor paling berpengaruh terhadap kelangsungan hidup pasien yaitu *age*, *ejection fraction*, *serum creatinine* dan *time*. Dari penelitian ini didapatkan tingkat akurasi algoritma untuk memprediksi adalah sebesar 91.45%. [8]

2.2. Landasan Teori

1. Data mining

Data mining adalah suatu proses untuk mencari suatu pola atau mencari data yang menarik pada data yang sudah dipilih untuk dilakkan suatu pengolahan dengan menggunakan suatu teknik atau metode tertentu [9][10][11].

2. Algoritma C4.5

Algoritma c4.5 dipergunakan untuk melakukan prediksi keanggotaan suatu objek untuk kategori kelas yang berbeda dengan cara membandingkan nilai – nilai yang sesuai dengan atributnya [1][12]. Algoritma c4.5 berfungsi untuk membuat decision tree berdasarkan dataset yang telah dimiliki. Secara umum tahapan yang dilakukan c4.5 dalam membuat decision tree adalah [13]:

- a Memilih atribut sebagai akar
- b Membuat cabang untuk setiap nilai

- c Membagi kasus dalam cabang
- d Mengulangi proses untuk setiap nilai
- e Mengulangi proses utuk masing masing cabang hingga kasus selesai.

3. Model Prediksi/Supervised Learning

Prediksi merupakan suatu proses diman nantinya akan ditemukan suatu pola tertentu yang terdapat dalam sekumpulan data [14]. Nantinya pola-pola tersebut dapat diketahui berdasarkan variabel – variabel yang terdapat pada data tersebut. Jika pola sudah ditemukan maka pola tersebut dapat digunakan untuk memprediksi variable yang belum diketahui jenis dan polanya. Salah satu model supervised learning adalah decision Tree [15].

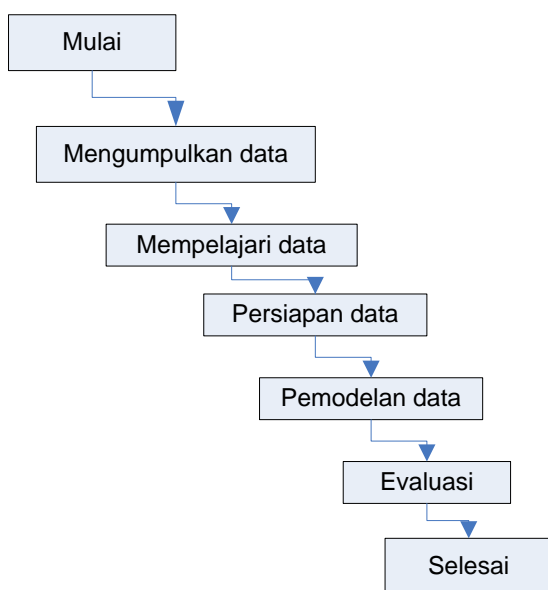
3. Metode Penelitian

Pada penelitian ini data yang digunakan adalah data public yang dapat diakses di halaman www.kaggle.com/datasets/imkrkannan/lung-cancer-dataset-by-staceyinrobert dimana data public yang disediakan adalah data para penderita penyakit kangker paru dengan jumlah 309 data yang digunakan pada penelitian ini.

Dalam penelitian ini data skunder yang digunakan berjumlah 309 data dan total 16 atribut dimana didalam nya termasuk 1 label yaitu Gender, Age, Smoking, Yellow Fingers, Anxiety, Peer Pressure, Chronic Disease, Fatigue, Allergy, Wheezing, Alcohol Consuming, Caughing, Shortness Of

Breath, Swallowing Difficulty, Chest Pain dan Class label nya adalah Lung Cancer.

Tahapan Yang dilakukan dalam penelitian ini adalah :



Gambar 1. Tahapan Penelitian

Tahapan yang dilakukan sesuai dengan gambar 1 adalah sebagai berikut :

- a. *Mengumpulkan data* : mendapatkan data/dataset para penderita kangker paru dimana data pada penelitian ini diperoleh dari data publik yang nantinya akan di oleh menggunakan model yang diusulkan.
- b. *Mempelajari data* : tahapan dimana mempelajari dataset yang dimiliki, mengecek data-data yang redundan.
- c. *Pemodelan data* : melakukan pengolahan data menggunakan algoritma yang diusulkan.
- d. *Evaluasi* : mengevaluasi hasil pengolahan data menggunakan algoritma yang sudah diusulkan.

4. Hasil dan Pembahasan

Adapun hasil dan pembahasan yang diperoleh dari pengolahan data untuk prediksi kangker paru yang terdiri dari beberapa proses diantaranya adalah :

1. Menyiapkan dataset

Data yang akan diolah adalah data dalam bentuk file CSV yang diperoleh dari data public yang di dapatkan di kaggle.com. Data yang diolah dalam penelitian ini terdiri atas 309 baris dan 16 kolom. Cuplikan data sebagaimana tampilan gambar berikut :

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	GENDER	AGE	SMOKR	YELLOW	ANXIET	PEERF	CHRON	FATIGU	ALLERG	WHEEZ	ALCOH	COUGH	SHORTN	SWALL	CHEST	LUNG_CANCER
2	M	63	1	2	2	1	1	2	1	2	2	2	2	2	2	2 YES
3	M	74	2	1	1	1	2	2	2	1	1	1	2	2	2	2 YES
4	F	59	1	1	1	2	1	2	1	2	1	2	2	1	2	2 NO
5	M	63	2	2	2	1	1	1	1	1	2	1	1	2	2	2 NO
6	F	63	1	2	1	1	1	1	1	2	1	2	2	1	2	2 NO
7	F	75	1	2	1	1	2	2	2	2	2	1	2	2	1	2 YES
8	M	52	2	1	1	1	1	2	1	2	2	2	2	2	1	2 YES
9	F	51	2	2	2	2	2	1	2	2	1	1	1	2	2	2 YES
10	F	68	2	1	2	1	1	2	1	1	1	1	1	1	1	2 NO
11	M	53	2	2	2	2	2	1	2	1	2	1	1	2	2	2 YES
12	F	61	2	2	2	2	2	2	1	2	1	2	2	2	2	2 YES
13	M	72	1	1	1	1	2	2	2	2	2	2	2	2	1	2 YES
14	F	60	2	1	1	1	1	2	1	1	1	1	1	2	1	2 NO
15	M	58	2	1	1	1	1	2	2	2	2	2	2	2	1	2 YES
16	M	69	2	1	1	1	1	1	2	2	2	2	2	1	1	2 NO
17	F	48	1	2	2	2	2	2	2	2	2	1	2	2	2	2 YES
18	M	75	2	1	1	1	2	1	2	2	2	2	2	2	1	2 YES
19	M	57	2	2	2	2	2	2	1	1	2	1	1	2	2	2 YES
20	F	68	2	2	2	2	2	2	1	1	1	2	2	1	1	2 YES
21	E															

Gambar 2. Cuplikan Dataset

2. Menyiapkan Data dan Library

Bila dataset yang dibutuhkan sudah tersedia, selanjutnya adalah menyiapkan library pada google colab yang diperlukan untuk menunjang pengolahan data.

```

import pandas as pd
import pydotplus
from sklearn.tree import DecisionTreeClassifier
import numpy as np
  
```

Gambar 3. Import Library Yang Digunakan Selanjutnya mengimport data dari kaggle tersebut di google colab.

```
[3] ! kaggle datasets download 'imkrkannan/lung-cancer-dataset-by-staceyinrobert'
```

```
Downloading lung-cancer-dataset-by-staceyinrobert.zip to /content
0% 0.00k/2.00k [00:00<?, ?B/s]
100% 2.00k/2.00k [00:00<00:00, 5.03MB/s]
```

Gambar 4 Download Dataset

Perintah diatas digunakan untuk mendownload dataset langsung dari kaggle untuk ditampilkan di google colab. Hal lain yang dapat dilakukan pada dataset adalah mendapatkan informasi terkait jumlah kolom, tipe data, jumlah seluruh data yang terdapat pada dataset.

```
df=pd.read_csv('survey_lung_cancer.csv')
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 309 entries, 0 to 308
Data columns (total 16 columns):
#   Column                Non-Null Count  Dtype
---  -
0   GENDER                 309 non-null   object
1   AGE                    309 non-null   int64
2   SMOKING                309 non-null   int64
3   YELLOW_FINGERS        309 non-null   int64
4   ANXIETY                309 non-null   int64
5   PEER_PRESSURE         309 non-null   int64
6   CHRONIC_DISEASE       309 non-null   int64
7   FATIGUE                309 non-null   int64
8   ALLERGY                309 non-null   int64
9   WHEEZING               309 non-null   int64
10  ALCOHOL_CONSUMING     309 non-null   int64
11  COUGHING               309 non-null   int64
12  SHORTNESS_OF_BREATH   309 non-null   int64
13  SWALLOWING_DIFFICULTY 309 non-null   int64
14  CHEST_PAIN            309 non-null   int64
15  LUNG_CANCER           309 non-null   object
dtypes: int64(14), object(2)
memory usage: 38.8+ KB
```

Gambar 5. Informasi Dataset

3. Melakukan Pembersihan Data

Tujuan pembersihan data adalah untuk menghapus data yang dianggap tidak seimbang, yang tidak lengkap atau data yang tidak sesuai.

```
lung = pd.read_csv("survey_lung_cancer.csv")
lung.isnull().sum()
```

```
GENDER      0
AGE          0
SMOKING      0
YELLOW_FINGERS 0
ANXIETY      0
PEER_PRESSURE 0
CHRONIC_DISEASE 0
FATIGUE      0
ALLERGY      0
WHEEZING     0
ALCOHOL_CONSUMING 0
COUGHING     0
SHORTNESS_OF_BREATH 0
SWALLOWING_DIFFICULTY 0
CHEST_PAIN   0
LUNG_CANCER  0
dtype: int64
```

Gambar 6. Pembersihan Data

Bila diperhatikan pada data dibawah ini, nilai dari gender tidak dalam bentuk numerik sebagaimana data pada atribut lainnya, Oleh sebab itu perlu dilakukan perubahan data agar nilai atribut Gender dapat diolah.

```
df.head()
```

index	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	C
0	M	69	1	2	2	1	
1	M	74	2	1	1	1	
2	F	59	1	1	1	2	
3	M	63	2	2	2	1	
4	F	63	1	2	1	1	

Gambar 7. Menampilkan Data Frame

Selanjutnya dilakukan perubahan nilai pada atribut gender sebagai berikut karena diperlukan keseragaman jenis data untuk dapat melakukan pengolahan data, karena jika data gender tidak dalam bentuk numerik tidak dapat dilakukan konversi data dari string ke jenis data float.

```
df['GENDER'].unique()
def change(col):
    if col == 'M':
        return 0
    else:
        return 1

df['GENDER'] = df['GENDER'].apply(change)
df.head()
```

C>

index	GENDER	AGE	SMOKING	YELLOW_FINGERS
0	1	69	1	2
1	1	74	2	1
2	1	59	1	1
3	1	63	2	2
4	1	63	1	2

Gambar 8. Merubah Data Menjadi Numerik

4. Menentukan variable x dan y

Selanjutnya menentukan data dan target menggunakan perintah berikut

```
X = lung.drop(columns=['LUNG_CANCER'],axis=1)
y = lung['LUNG_CANCER']
```

Gambar 9. Menentukan Variable x y

5. Melakukan split data training dan testing

```
[01] from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state = 1)# 70% training and 30% test

[02] from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.fit_transform(X_test)
```

Gambar 10. Split data training testing

Split data dilakukan untuk membagi data menjadi dua bagian, yaitu data training dan data testing dimana proporsi yang digunakan pada penelitian ini adalah 30:70. Rasio tersebut dimaksudkan untuk membagi data training sebesar 70% dan data testing sebesar 30%.

6. Evaluasi Model C4.5

Evaluasi dengan menggunakan model algoritma C4.5 menggunakan kode berikut

```
clf = DecisionTreeClassifier()
clf = clf.fit(X_train,y_train)
y_pred = clf.predict(X_test)
```

Gambar 11. Klasifikasi menggunakan decision tree

Dilakukan evaluasi terhadap performa dari algoritma c4.5 dalam melakukan prediksi yaitu dengan melakukan klasifikasi data.

Selanjutnya menentukan nilai akurasi algoritma C4.5 dengan menggunakan kode berikut :

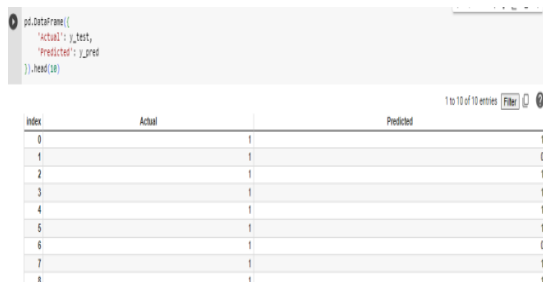
```
# Model Accuracy
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

Accuracy: 0.8709677419354839

Gambar 12. Pengecekan Akurasi

Dari model akurasi diperoleh sebesar 87%, artinya dari seluruh data yang diolah baik data training dan data testing, kemampuan algoritma c4.5 dalam melakukan prediksi adalah sebesar 87%. Nilai akurasi ini dapat dianggap sudah baik dalam melakukan prediksi dan dapat dijadikan acuan untuk dapat dikembangkan dalam dunia kesehatan untuk melakukan prediksi kemungkinan seseorang menderita penyakit kanker paru berdasarkan gejala yang tampak. Hal ini karena berdasarkan nilai aktual dan prediksi yang diberikan oleh algoritma c4.5 terlihat dalam memprediksi tidak banyak melakukan kesalahan sebagaimana tampak pada gambar

dibawah. Berikut adalah cuplikan perbedaan antara hasil prediksi dan data sebenarnya :



```

0    pd.DataFrame({
1      "Actual": y_test,
2      "Predicted": y_pred
3    }).head(10)

```

Index	Actual	Predicted
0	1	1
1	1	0
2	1	1
3	1	1
4	1	1
5	1	1
6	1	0
7	1	1
8	1	1

Gambar 13. Nilai Aktual Dan Prediksi

Dari data asli dan data prediksi, penggunaan algoritma c4.5 mengalami sedikit kesalahan dalam melakukan prediksi. Dari data yang tampak dari 10 data hanya ada 2 data yang meleset

5. Kesimpulan

Penggunaan google colab sangat membantu dalam melakukan penelitian terkait implementasi python dalam data mining untuk melakukan prediksi. Penggunaan model algoritma c4.5 berdasarkan penelitian yang dilakukan sudah layak untuk dikembangkan dalam dunia kesehatan sebagai model prediksi terhadap penyakit kanker paru karena akurasi yang baik yaitu 87%. Dengan tingkat akurasi tersebut besar kemungkinan prediksi algoritma terhadap suatu gejala dapat menentukan seseorang dinyatakan menderita kanker paru atau tidak. Butuh penelitian lebih lanjut untuk model prediksi lain yang mungkin memiliki tingkat akurasi yang lebih baik sehingga kecil kemungkinan prediksi dapat meleset.

6. Daftar Pustaka

- [1] B. A. C. P, "DOI : 10.29408/jit.v1i1. 892," *Baiq Andriksa Candra*, vol. 1, no. 1, pp. 32–39, 2018.
- [2] J. Joseph and L. W. A. Rotty, "Kanker Paru : Laporan Kasus," vol. 2, no. 1, pp. 17–25, 2020.
- [3] Y. Ernawati, S. Ermayanti, D. Herman, and R. Russilawati, "Faktor Risiko Kanker Paru pada Perempuan yang Dirawat di Bagian Paru RSUP Dr. M. Djamil Padang dan RSUD Solok: Penelitian Case Control," *J. Kesehat. Andalas*, vol. 8, no. 2S, p. 1, 2019, doi: 10.25077/jka.v8i2s.951.
- [4] C. Algoritma, "Prediksi Kekambuhan Kanker Payudara Dengan," vol. 15, no. 2, pp. 2017–2018, 2018.
- [5] Noviandi, "Implementasi Algoritma Decision Tree C4.5 Untuk Prediksi Penyakit Diabetes," *Inohim*, vol. 6, no. 1, pp. 1–5, 2018.
- [6] B. A. Candra Permana and I. K. Dewi Patwari, "Komparasi Metode Klasifikasi Data Mining Decision Tree dan Naïve Bayes Untuk Prediksi Penyakit Diabetes," *Infotek J. Inform. dan Teknol.*, vol. 4, no. 1, pp. 63–69, 2021, doi: 10.29408/jit.v4i1.2994.
- [7] D. S. Permana and A. Silvanie, "Prediksi Penyakit Jantung Menggunakan Support Vector Machine Dan Python Pada Basis Data Pasien," vol. 2, no. 1, pp. 29–34, 2021.
- [8] V. No and Y. Yuliani, "Algoritma Random Forest Untuk Prediksi Kelangsungan Hidup Pasien Gagal Jantung Menggunakan Seleksi Fitur Bestfirst," vol. 5, no. 2, pp. 298–306, 2022.
- [9] T. Informatika, S. Dharma, and W. Metro, "Klasifikasi Kanker Paru-Paru Menggunakan Metode Naive," vol. 6, no. 2, pp. 20–24, 2022.
- [10] A. U. Zailani and N. L. Hanun, "Penerapan Algoritma Klasifikasi Random Forest Untuk Penentuan Kelayakan Pemberian Kredit Di Koperasi Mitra Sejahtera,"

- Infotech J. Technol. Inf.*, vol. 6, no. 1, pp. 7–14, 2020, doi: 10.37365/jti.v6i1.61.
- [11] L. Sari, A. Romadloni, and R. Listyaningrum, “Penerapan Data Mining dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random,” vol. 14, no. 01, pp. 155–162, 2023, doi: 10.35970/infotekmesin.v14i1.1751.
- [12] M. R. Alfarabi, “Optimalisasi Algoritma C4.5 dalam Menganalisis Indikasi Penyebab Penyakit Feline Immunodeficiency Virus (FIV) pada Kucing,” *J. Sistim Inf. dan Teknol.*, vol. 4, pp. 6–9, 2022, doi: 10.37034/jsisfotek.v4i4.152.
- [13] C. Algoritma, “Rancang Bangun Aplikasi Pendeteksi Penyakit Ginjal Kronis dengan Menggunakan,” vol. IX, no. 1, 2017.
- [14] M. Kafil, “Penerapan Metode K-Nearest Neighbors,” *J. Mhs. Tek. Inform.*, vol. 3, no. 2, pp. 59–66, 2019.
- [15] A. Mujumdar and V. Vaidehi, “Diabetes Prediction using Machine Learning Algorithms,” *Procedia Comput. Sci.*, vol. 165, pp. 292–299, 2019, doi: 10.1016/j.procs.2020.01.047.