Jurnal Teknik Informatika (JUTIF)
Vol. 4, No. 4, August 2023, pp. 693-701

DOI: https://doi.org/10.52436/1.jutif.2023.4.4.693
p-ISSN: 2723-3863
e-ISSN: 2723-3871

PCOS DISEASE CLASSIFICATION USING FEATURE SELECTION RFECV AND EDA WITH KNN ALGORITHM METHOD

Nadhira Triadha Pitaloka[1], Kusnawi[*2]

[1,2]Informatics, Faculty of Computer Science, Universitas Amikom Yogyakarta, Indonesia
Email: [1]nadhira.pitaloka@students.amikom.ac.id, [2]khusnawi@amikom.ac.id

(Article received: January 15, 2023; Revision: January 28, 2023; published: August 18, 2023)

Abstract

Polycystic ovary syndrome is an endocrine disorder of the ovaries that causes hormonal disturbances in women of reproductive age, where androgen secretion in the ovaries of women with Polycystic Ovary Syndrome (PCOS) is excessive compared to normal women. This usually occur in women with obesity which is characterized by irregular menstrual cycles, chronic anovulation, hyperandrogenism, and even infertility. Efforts are used to treat this disease in the form of hormone therapy, laparoscopic ovarian drilling, and in-vitro fertilization. However, these three therapies are focused on symptomatic therapy and are less effective in treating PCOS-related infertility. Detecting PCOS disease early is very necessary so that prevention and treatment can be carried out immediately. Therefore, a classification is carried out to detect PCOS disease by being able to analyze data that has a high degree of accuracy. The method used for the classification of PCOS disease is using the K Nearest Neighbor (KNN), method which previously carried out the feature selection process, namely the Exploratory Data Analysis (EDA), method which is used for the data analysis process by means of an analysis approach to data to find out the most accurate method and using the Recursive Feature Elimination and Cross-Validation (RFECV) selection method which ranks the features based on their level of importance to the prediction process. Further, the data classification process uses the K-Nearest Neighbors (KNN) algorithm. The results of the Exploratory Data Analysis (EDA) feature selection process produce 10 data attributes that are used and are continued by the Recursive Feature Elimination and Cross-Validation (RFECV) process by producing the 7 most important attributes used and finally the K-Nearest Neighbors (KNN) method has a high level high accuracy by producing an accuracy value of 93%, precision 82%, recall 100%, and F1 score 90%.

Keywords: *EDA, KNN, PCOS, RFECV.*

KLASIFIKASI PENYAKIT PCOS MENGGUNAKAN *FEATURE SELECTION* RFECV DAN EDA DENGAN METODE ALGORITMA KNN

Abstrak

Sindrom ovarium polikistik merupakan kelainan endokrin pada ovarium yang menyebabkan terganggunya hormonal pada wanita usia reproduksi dimana sekresi androgen pada ovarium wanita dengan *Polycystic Ovary Syndrome* (PCOS) berlebih dibandingkan dengan wanita normal. Hal ini biasanya terjadi pada wanita dengan obesitas yang ditandai oleh ketidak teraturan siklus menstruasi, anovulasi kronis, hiperandrogenisme bahkan adanya infertilitas. Upaya yang digunakan untuk mengobati penyakit ini berupa terapi hormon, *laparoscopic ovarian drilling*, dan *in-vitro fertilization*. Akan tetap ketiga terapi tersebut terfokus pada terapi simptomatis dan kurang efektif dalam mengatasi infertilitas terkait PCOS. Mendeteksi penyakit PCOS sejak dini sangat diperlukan agar pencegahan dan pengobatan dapat segera dilakukan. Oleh karena itu dilakukan klasifikasi untuk mendeteksi penyakit PCOS dengan dapat menganalisis data yang memiliki tingkat akurasi tinggi. Metode yang digunakan untuk klasifikasi penyakit PCOS ini yaitu menggunakan metode *K Nearest Neighbor* (KNN) yang sebelumnya dilakukan proses *selection feature* yaitu metode *Exploratory Data Analysis* (EDA) yang digunakan untuk proses analisis data dengan cara pendekatan analisis terhadap data untuk mengetahui metode yang paling akurat dan dengan metode seleksi *Recursive Feature Elimination and Cross-Validation* (RFECV) yang mengurutkan (*rangking*) fitur berdasarkan tingkat pentingnya terhadap proses prediksi. Selanjutnya proses klasifikasi data menggunakan algoritma *K-Nearest Neighbors* (KNN). Hasil dari proses *feature selection Exploratory Data Analysis* (EDA) dihasilkan 10 data atribut yang digunakan dan dilanjutkan proses *Recursive Feature Elimination and Cross-Validation* (RFECV) dengan dihasilkan 7 atribut terpenting yang digunakan dan terakhir proses metode *K-Nearest Neighbors* (KNN) memiliki tingkat akurasi tinggi dengan menghasilkan nilai akurasi 93%, *precision* 82%, *recall* 100%, dan *F1 Score* 90%.

Kata kunci: *EDA, KNN, PCOS, RFECV.*

1. PENDAHULUAN

PCOS (*Polycystic Ovary Syndrome*) merupakan kelainan endokrin pada ovarium yang menyebabkan terganggunya hormonal pada wanita usia reproduksi. Wanita dengan penyakit ini memiliki risiko infertilitas yang tinggi. Penderita PCOS sebanyak 50% diantaranya mengalami keterlambatan menstruasi lebih dari siklus normal yaitu di antara 21 sampai dengan 35 hari dan 20% belum atau tidak mendapatkan menstruasi pertamanya sampai dengan usia 15 tahun. Gejala kelainan ini biasanya ditandai dengan menstruasi yang tidak teratur, kelebihan hormon androgen (*hyperandrogenism*), dan polikistik ovarium yang dapat dilihat pada saat tes USG (*ultrasound*) [1] [2]. Wanita penderita PCOS sendiri memiliki dua kriteria dari tiga kriteria tersebut.

Penyebab PCOS diantaranya pola hidup tidak sehat, sering mengkonsumsi *junk food*, kelebihan insulin, dan kelainan genetik yang bisa berasal dari keturunan keluarga. Jika ibu atau kakak perempuan mengalami PCOS maka wanita tersebut juga akan memiliki risiko yang sama. Dalam jangka panjang PCOS dapat membuat komplikasi dan gangguan antara lain kanker pada ovarium, kanker pada endometrium yang disebabkan karena faktor-faktor seperti kelebihan berat badan, kelebihan hormon estrogen, dan infertilitas yang dimiliki wanita dengan penderita PCOS, kanker pada payudara penderita PCOS yang memiliki jumlah hormon estrogen yang tinggi hingga berlebihan dapat meningkatkan risiko terkenanya kanker payudara yang disebabkan oleh tingginya hormon estrogen pada wanita, kanker pada kardiovaskular (jantung dan pembuluh darah) dengan faktor penyebabnya adalah kelebihan berat badan, hipertensi, diabetes, usia, kolesterol, dan merokok [3][4].

Menurut WHO pada tahun 2015, sebanyak 5,8% dari 8.612 wanita usia 22-28 tahun mengalami penyakit PCOS. Upaya yang digunakan untuk mengobati penyakit ini berupa terapi hormon, *laparoscopic ovarian drilling*, dan *in-vitro fertilization*. Akan tetap ketiga terapi tersebut terfokus pada terapi simptomatis dan kurang efektif dalam mengatasi infertilitas terkait PCOS [1][5]. Mendeteksi penyakit PCOS sejak dini sangat diperlukan agar pencegahan dan pengobatan dapat segera dilakukan. Oleh karena itu, dibutuhkan teknologi untuk mendeteksi penyakit ini yang memiliki tingkat akurasi yang tinggi.

Penelitian ini menggunakan metode *K-Nearest Neighbor* (KNN). Dimana penelitian sebelumnya [5] membahas pendeteksi PCOS menggunakan klasifikasi ANN dengan tingkat akurasi 90,09%. Penelitian [3] membahas perbandingan klasifikasi SVM dan KNN untuk mendeteksi kanker dengan tingkat akurasi berturut-turut 85,60% dan 98,54%. Penelitian [6][7] membahas klasifikasi penyakit liver dan algoritma genetika menggunakan metode KNN dengan normalisasi dan seleksi fitur untuk meningkatkan nilai akurasi. Berdasarkan penelitian-penelitian tersebut, penelitian ini menggunakan metode KNN karena memiliki prinsip sederhana, sering digunakan untuk klasifikasi, dan memiliki nilai akurasi yang tinggi [3][7]. Selain itu, penelitian ini menggunakan *feature selection Exploratory Data Analysis* (EDA) dan *Recursive Feature Elimination and Cross-Validation* (RFECV). *Exploratory Data Analysis* (EDA) merupakan teknik pencarian heuristik untuk menemukan hubungan yang signifikan antara variabel secara keseluruhan dalam dataset atau kumpulan data yang besar [8]. RFECV merupakan proses rekursif yang mengurutkan (*rangking*) fitur berdasarkan tingkat pentingnya terhadap proses prediksi. Pada setiap iterasi, fitur *ranking* yang penting diukur dan fitur yang kurang relevan dihilangkan [9]. Kedua *feature* tersebut digunakan untuk mengefisiensi proses klasifikasi agar mendapatkan hasil klasifikasi yang lebih baik dari sebelumnya [8][9]. *Dataset* yang digunakan diambil dari *Kaggle Polycystic Ovary Syndrome* (PCOS) yang diperoleh dari 10 rumah sakit yang berbeda di Kerala, India.

Penelitian ini ditujukan untuk membantu para ahli medis untuk mendeteksi dan mengklasifikasi penyakit PCOS menggunakan metode algoritma KNN dengan *feature selection Recursive Feature Elimination and Cross-Validation* (RFECV) dan *Exploratory Data Analysis* (EDA) dengan menggunakan bahasa pemrograman *python* untuk mendapatkan hasil analisis data yang tingkat akurasi tinggi.

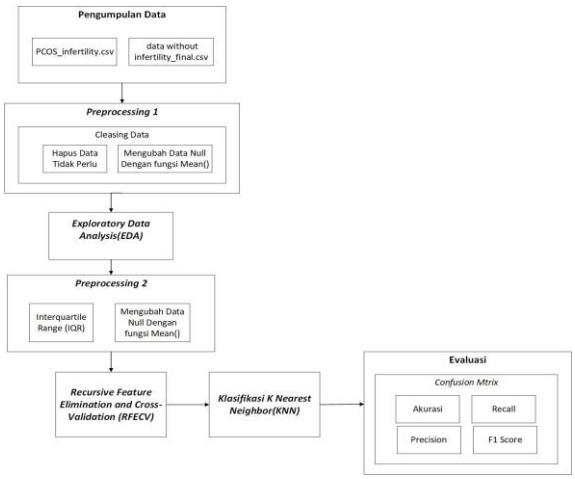
2. METODE PENELITIAN

Gambar 1 menunjukkan alur penelitian yang melalui beberapa tahap, yaitu dari *dataset* PCOS, *Preprocessing 1*, *Exploratory Data Analysis* (EDA), *Preprocessing 2*, *Recursive Feature Elimination and Cross-Validation* (RFECV), klasifikasi dengan *K-Nearest Neighbor* (KNN), dan Evaluasi.

2.1. Data PCOS

Pada tahap pertama penelitian dilakukan dengan pengumpulan data. Data yang digunakan berupa data publik dari PCOS: *Data Cleaning and Feature Importances* yang sudah banyak digunakan oleh peneliti. *Dataset* ini berisikan dua data csv, yaitu *PCOS_infertility.csv* yang ditunjukkan pada Tabel 1 dan data *without infertility_final.csv* yang ditunjukkan pada Tabel 2 dan untuk *dataset* dapat diunduh pada url:

https://www.kaggle.com/code/noelmat/pcos-data-cleaning-and-feature-importances/.



Gambar 1. Alur Penelitian

Tabel 1. *PCOS_infertility.csv*

Sl. No	Patient File No.	PCOS (Y/N)	I beta-HCG(mIU/mL)	II beta-HCG(mIU/mL)	AMH(ng/mL)
1	10001	0	1.99	1.99	2.07
2	10002	0	60.8	1.99	1.53
3	10003	1	494.08	494.08	6.63
4	10004	0	1.99	1.99	1.22
5	10005	0	801.45	801.45	2.26
6	10006	0	237.97	1.99	6.74
7	10007	0	1.99	1.99	3.05
8	10008	0	100.51	100.51	1.54
9	10009	0	1.99	1.99	1
10	10010	0	1.99	1.99	1.61
…..	…..	…..	…..	…..	…..
541	10541	1	1.99	1.99	20

Tabel 2. Data *without infertility_final.csv*

Sl. No	Patient File No.	PCOS (Y/N)	Age (yrs)	Weight (Kg)	…	Endometrium (mm)
1	10001	0	28	44.6	…..	8.5
2	10002	0	36	65	…..	3.7
3	10003	1	33	68.8	…..	10
4	10004	0	37	65	…..	7.5
5	10005	0	25	52	…..	7
6	10006	0	36	74.1	…..	8
7	10007	0	34	64	…..	6.8
8	10008	0	33	58.5	…..	7.1
9	10009	0	32	40	…..	4.2
10	10010	0	36	52	…..	2.5
….	…….	…….	…….	…….	…..	…….
541	10541	1	23	82	…..	6.9

2.2. *Preprocessing 1*

Preprocessing merupakan salah satu tahapan dalam melakukan *mining* data sebelum menuju ke tahap pemprosesan. *Preprocessing* dilakukan melalui cara eliminasi data yang tidak sesuai. Dalam proses ini data akan diubah dalam bentuk yang akan lebih dipahami oleh sistem. Melalui data *preprocessing* memungkinkan proses *mining* akan berjalan dengan lebih efektif dan efisien [10]. Data yang telah melalui Pra-pemrosesan merupakan data yang sudah melalui beberapa tahap pembersihan.

Tahap yang dilakukan pada *Preprocessing 1* yaitu melakukan data *cleaning* dengan dua langkah pertama mengganti data yang kosong atau *null* dengan mengisi data dengan rata-rata (*mean*) pada data csv *PCOS_infertility.csv* serta mengecek 43 atribut yang dilanjutkan dengan menggabungkan data *PCOS_infertility.csv* dengan data *without infertility _final.csv*.

2.3. *Exploratory Data Analysis* (EDA)

Exploratory Data Analysis (EDA) merupakan teknik pencarian heuristik untuk menemukan hubungan yang signifikan antara variabel secara keseluruhan dalam *dataset* atau kumpulan data yang besar. Pada tahap ketiga ini dilakukan proses *Exploratory Data Analysis (*EDA), yaitu proses untuk menentukan atribut apa yang akan digunakan. *Exploratory Data Analysis* (EDA) memiliki karakteristik fleksibel yang diperlukan untuk melakukan identifikasi dan investigasi suatu fenomena yang muncul pada saat melakukan penelitian empiris. Pengaplikasian *Exploratory Data Analysis* (EDA) tergantung pada konteks dan tergantung pada rincian analisis [8][11].

Tujuan *Exploratory Data Analysis* (EDA) adalah mencari pola data. *Exploratory Data Analysis* (EDA) juga digunakan diantaranya untuk mengoptimalkan pengetahuan mengenai data, menghasilkan variabel yang penting, mendeteksi *outlier*, anomali pada data, dan menguji asumsi awal [11][12]. Hal-hal inilah yang dapat digunakan dalam memperkaya analisis data dan membantu mengoptimalkan hasil klasifikasi.

2.4. Preprocessing 2

Tahap setelah dilakukan proses EDA dilanjutkan dengan Preprocessing ke 2 yaitu dengan melakukan proses *Interquartile Range* (IQR) dan mengisi data *null* dengan fungsi *mean*.

a. *Interquartile Range* (IQR)

IQR adalah *interquartile range* atau rentang antar kuartil dari sekumpulan data. IQR digunakan dalam analisis statistik untuk membantu menarik kesimpulan mengenai sekumpulan data [13]. IQR lebih sering digunakan daripada *range* karena tidak menyertakan data paling luar. Nilai IQR dapat diperoleh menggunakan rumus sebagai berikut.

$$IQR = Q_3 - Q_1 \quad (1)$$

Keterangan:
IQR = Jarak antar kuartil
Q_3 = Kuartil ke-3
Q_1 = Kuartil ke-1
Rumus untuk Q_1, yaitu:

$$Q_1 = data\ ke - \frac{n}{4} \quad (2)$$

Rumus untuk Q_3, yaitu:

$$Q_3 = data\ ke - \frac{3n}{4} \quad (3)$$

Keterangan:
n = Banyaknya data

Pada tahap keempat ini dilakukan perhitungan *outliers*. *Outliers* adalah data sampel yang memiliki karakteristik berbeda daripada mayoritas data sampel. Misalnya, nilainya terlalu tinggi atau terlalu rendah dibandingkan sebagian besar data sampel yang lainnya. Dalam analisis kelompok, kehadiran *outliers* perlu dideteksi sebab akan mengganggu analisis. Bila data telah tersedia untuk mendeteksi keberadaan *outliers*, maka dapat dilakukan dengan menggunakan dua cara, yaitu: menggunakan *z-score* dan *boxplot*.

b. Mengisi data null dengan fungsi mean

Setelah dilakukan pengecekan data Interquartile *Range* (IQR) dilanjutkan dengan pengecekan apakah terdapat data *null* dan jika ditemukan data *null* maka akan dilanjutkan dengan pengisian data *null* dengan nilai rata-rata (*mean*) pada atribut yang terdapat data *null*.

2.5. *Recursive Feature Elimination and Cross-Validation* (RFECV)

Metode seleksi RFECV pada dasarnya adalah proses rekursif yang mengurutkan (*rangking*) fitur berdasarkan tingkat pentingnya terhadap proses prediksi. *Ranking* tersebut dapat dihitung menggunakan metode *Support Vector Machine* (SVM) kernel linear [12][14]. Untuk klasifikasi biner, SVM membentuk fungsi linear, seperti yang dirumuskan pada Persamaan 4.

$$D(x) = sign(x.w) \quad (4)$$

Dimana x menandakan vektor *input* dan w menandakan vektor yang tegak lurus terhadap *hyperplane* yang terbentuk dari fungsi linear.

Pada tahap keenam ini dilakukan pengecekan kembali menggunakan *feature selection Recursive Feature Elimination and Cross-Validation* (RFECV). Metode RFECV menghasilkan *ranking* dengan variasi yang tinggi atau dengan kata lain metode ini sangat sensitif terhadap *training set*. RFECV mengatasi permasalahan tersebut dengan penggunaan *cross validation*. *Cross validation* memberi kesempatan pada seluruh untuk menjadi *testing set* sebanyak k-1, dimana *dataset* k adalah jumlah partisi pada *cross validation*. Dengan *cross validation*, metode RFECV akan lebih stabil dan lebih handal dalam pengurutan fitur [7][15].

2.6. Klasifikasi dengan KNN (*K- Nearest Neighbor*)

Algoritma *Nearest Neighbor Retrieval* (*K-Nearest Neighbor* atau K-NN) adalah sebuah algoritma untuk melakukan klasifikasi terhadap objek dengan data pembelajaran yang jaraknya paling dekat dengan objek tersebut (k = 1) [3][12]. Tahap ini menggunakan klasifikasi dengan algoritma *K-Nearest Neighbor (*KNN) dengan menentukan nilai k dengan pengecekan yang menghasilkan tingkat *error* yang paling rendah yang dilakukan pada nilai k yaitu 1 - 15. Rumus untuk menghitung bobot kemiripan (*similarity*) dengan *Nearest Neighbor* digunakan rumus *Euclidean*, seperti yang dirumuskan pada Persamaan 5.

$$d_{ij} = \sqrt{\sum_{j=1}^{m} (x_{ij} - c_{kj})^2} \quad (5)$$

Keterangan:
d_{ij} = Jarak
x_{ij} = Sampel Data
c_{kj} = Data Uji
j = Variabel Data
m = Dimensi Data

Cara kerja algoritma KNN diperlukan penentuan data latih, data uji, dan nilai k. Selanjutnya, data latih diurutkan berdasarkan hitungan jarak terdekat antara data uji dan data latih. Dan terakhir, diambil rata-rata data latih terkecil sesuai jumlah k untuk menentukan kelas regresi. Berikut proses alur perhitungan menggunakan metode *K-Nearest Neighbors* [12].

1. Menentukan parameter k
2. Menghitung jarak antara data yang akan dievaluasi dengan semua pelatihan
3. Mengurutkan jarak yang terbentuk
4. Menentukan jarak yang terdekat sampai nilai k
5. Memasangkan kelas yang sesuai
6. Mencari jumlah kelas yang terdekat dan menetapkan kelas tersebut untuk dievaluasi.

2.7. Evaluasi

Ini Pada tahap terakhir dilakukan evaluasi klasifikasi menggunakan *Confusion Matrix* dengan perhitungan *accuracy*, *precision*, *recall*, dan *f1 score*. *Confusion matrix* merupakan salah satu metode yang dapat digunakan untuk mengukur kinerja suatu metode klasifikasi. Pada dasarnya *confusion matrix* mengandung informasi yang membandingkan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi yang seharusnya [16]. Ada empat istilah yang merupakan representasi hasil proses klasifikasi pada *confusion matrix*, yaitu *True Positive*, *True Negative*, False *Positive*, dan False *Negative*. Nilai *True Negative* (TN) merupakan jumlah data negatif yang terdeteksi dengan benar, sedangkan *False Positive* (FP) merupakan data negatif namun terdeteksi sebagai data positif. Sementara itu, *True Positive* (TP) merupakan data positif yang terdeteksi benar. *False Negative* (FN) merupakan kebalikan dari *True Positive*, sehingga data posifit, namun terdeteksi sebagai data negatif.

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (6)$$

$$Presisi = \frac{TP}{FP + TP} \times 100\% \quad (7)$$

$$Recall = \frac{TP}{FN + TP} \times 100\% \quad (8)$$

Keterangan:

- TP adalah *True Positive*, yaitu jumlah data positif yang terklasifikasi dengan benar oleh sistem.
- TN adalah *True Negative*, yaitu jumlah data negatif yang terklasifikasi dengan benar oleh sistem.
- FN adalah *False Negative*, yaitu jumlah data negatif namun terklasifikasi salah oleh sistem.
- FP adalah *False Positive*, yaitu jumlah data positif namun terklasifikasi salah oleh sistem.

Nilai akurasi dapat diperoleh dengan Persamaan 6. Nilai presisi menggambarkan jumlah data kategori positif yang diklasifikasikan secara benar dibagi dengan total data yang diklasifikasi positif. Presisi dapat diperoleh dengan Persamaan 7. Sementara itu, *recall* menunjukkan berapa persen data kategori positif yang diklasifikasikan dengan benar oleh sistem. Nilai *recall* diperoleh dengan Persamaan 8.

3. HASIL DAN PEMBAHASAN

Tahap pertama yang dilakukan pada penelitian ini, yaitu menggunakan *dataset* dengan file csv yang diunduh dari *Kaggle* yaitu *PCOS_infertility.csv* dan data *without infertility _final.csv*. Proses pertama kali yang dilakukan adalah dengan menggunakan data yang ada pada Gambar 2 yaitu untuk *dataset PCOS_infertility.csv* dengan menggunakan 43 atribut yaitu dengan total data 541 data.

```
0   Sl. No                    541 non-null    int64
1   Patient File No.          541 non-null    int64
2   PCOS (Y/N)                541 non-null    int64
3    Age (yrs)                541 non-null    int64
4   Weight (Kg)               541 non-null    float64
5   Height(Cm)                541 non-null    float64
6   BMI                       541 non-null    float64
7   Blood Group               541 non-null    int64
8   Pulse rate(bpm)           541 non-null    int64
9   RR (breaths/min)          541 non-null    int64
10  Hb(g/dl)                  541 non-null    float64
11  Cycle(R/I)                541 non-null    int64
12  Cycle length(days)        541 non-null    int64
13  Marraige Status (Yrs)     540 non-null    float64
14  Pregnant(Y/N)             541 non-null    int64
15  No. of aborptions         541 non-null    int64
16  FSH(mIU/mL)               541 non-null    float64
17  LH(mIU/mL)                541 non-null    float64
18  FSH/LH                    541 non-null    float64
19  Hip(inch)                 541 non-null    int64
20  Waist(inch)               541 non-null    int64
21  Waist:Hip Ratio           541 non-null    float64
22  TSH (mIU/L)               541 non-null    float64
23  AMH(ng/mL)                540 non-null    float64
24  PRL(ng/mL)                541 non-null    float64
25  Vit D3 (ng/mL)            541 non-null    float64
26  PRG(ng/mL)                541 non-null    float64
27  RBS(mg/dl)                541 non-null    float64
28  Weight gain(Y/N)          541 non-null    int64
29  hair growth(Y/N)          541 non-null    int64
30  Skin darkening (Y/N)      541 non-null    int64
31  Hair loss(Y/N)            541 non-null    int64
32  Pimples(Y/N)              541 non-null    int64
33  Fast food (Y/N)           540 non-null    float64
34  Reg.Exercise(Y/N)         541 non-null    int64
35  BP _Systolic (mmHg)       541 non-null    int64
36  BP _Diastolic (mmHg)      541 non-null    int64
37  Follicle No. (L)          541 non-null    int64
38  Follicle No. (R)          541 non-null    int64
39  Avg. F size (L) (mm)      541 non-null    float64
40  Avg. F size (R) (mm)      541 non-null    float64
41  Endometrium (mm)          541 non-null    float64
42  Unnamed: 42               2 non-null      object
```

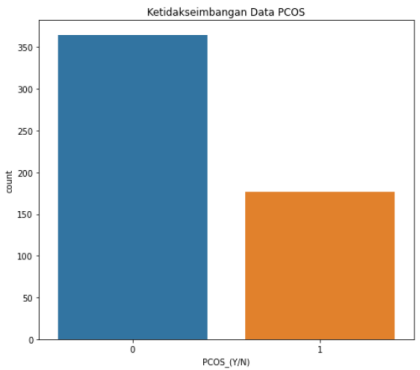
Gambar 2. Atribut *PCOS_infertility.csv*

Pada Gambar 2 dari data 541 PCOS_infertility.csv dilanjutkan dengan proses *Preprocessing* 1 yang dilakukan dengan pengecekan dataset *PCOS_infertility.csv* dan terdapat data *unamed :42*. Selanjutnya dilanjutkan proses *drop* atribut untuk menghilangkan data *unamed :42*, tidak hanya data *unamed :42* namun ditemukan juga data kosong atau *null* pada atribut *Marriage Status(Yrs)*, *Fast Food(Y/N)*, *AMH(ng/Ml)* yang terdapat total data 540 yang seharusnya untuk total data 541. Untuk mengatasi data kosong atau *null* dilakukan dengan mengisi data tersebut menggunakan nilai rata-rata (*mean*) agar data pada atribut bisa menjadi sama yaitu dari 540 menjadi 541. Kemudian langkah selanjutnya dilakukan proses menggabungkan dataset *PCOS_infertility.csv* dengan dataset *data without infertility _final.csv* dengan menambahkan 5 atribut yang ada pada data set *data without infertility _final.csv* pada Gambar 3, yaitu *Patient File No.*, *PCOS (Y/N)*, *I beta-HCG(mIU/mL)*,*II beta-HCG(mIU/mL)*, dan *AMH(ng/mL)*. Khusus untuk atribut *Sl.No*, *PCOS(Y/N)*, *AMH(ng/mL)* ditambahkan dibelakang atributnya *"_y"* sehingga untuk nama atribut menjadi *Sl.No_y*, *PCOS(Y/N)_y*, *AMH(ng/mL)_y*, hal ini dilakukan agar tidak menjadi data *double* antara dataset *PCOS_infertility.csv* dan dataset *data without infertility _final.csv*.

```
#   Column                Non-Null Count  Dtype
---  ------                --------------  -----
0   Sl. No                541 non-null    int64
1   Patient File No.      541 non-null    int64
2   PCOS (Y/N)            541 non-null    int64
3     I   beta-HCG(mIU/mL) 541 non-null    float64
4   II    beta-HCG(mIU/mL) 541 non-null    float64
5   AMH(ng/mL)            541 non-null    object
```

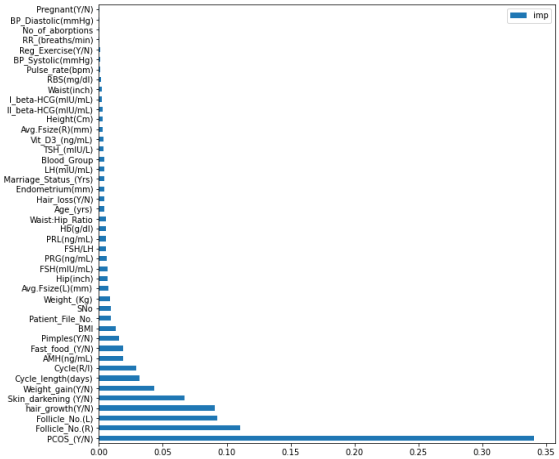
Gambar 3. Atribut data *without infertility _final.csv*

Setelah dilakukan penggabungan data dilanjutkan dengan proses *Exploratory Data Analysis* (EDA) yang dimulai dari proses pengecekan ketidakseimbangan data PCOS. Gambar 4 menunjukkan bahwa adanya ketidakseimbangan data PCOS. Hal ini dapat dilihat bahwa pada jumlah data kategori 0 (tidak penderita PCOS) memiliki jumlah 2 kali lipat dari jumlah data kategori 1 (penderita PCOS). *Imbalance data* atau ketidakseimbangan data yang terjadi dapat menyebabkan kasus dimana hasil klasifikasi mempunyai tingkat akurasi yang lebih tinggi tetapi hanya satu kategori data yang terklasifikasi secara tepat sehingga dapat disimpulkan bahwa model klasifikasi yang dihasilkan tidak bagus.



Gambar 4. Ketidakseimbangan Data PCOS

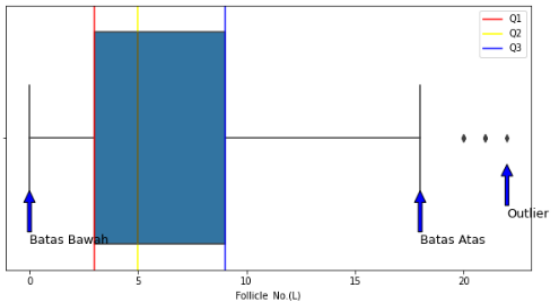
Dari data ketidakseimbangan, agar *dataset* pada data PCOS menjadi seimbang dilakukan *split dataset* yaitu dibagi menjadi data *testing* sebesar 15% dan data *training* 85%. Setelah *dataset* seimbang dilakukan proses *Exploratory Data Analysis* (EDA) yaitu menampilkan atribut yang sering digunakan. Pada penelitian ini hasil dari proses *Exploratory Data Analysis* (EDA) diambil 10 atribut yang digunakan, yaitu atribut *Follicle_No.(R)*, *Follicle_No.(L)*, *hair_growth(Y/N)*, *Skin_darkening (Y/N)*, *Cycle_length(days)*, *Cycle(R/I)*, *Weight_gain(Y/N)*, *AMH(ng/mL)*, *Fastfood(Y/N)*, dan *Pimples(Y/N)*.



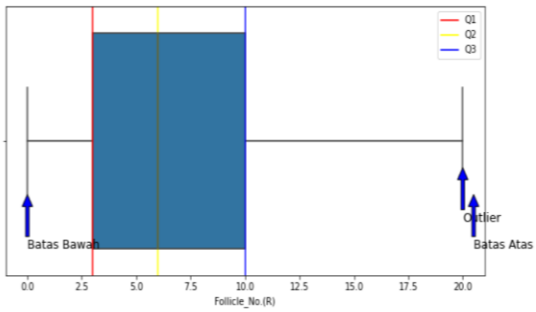
Gambar 5. Hasil proses EDA

Proses selanjutnya untuk menggunakan 10 atribut dari proses *Exploratory Data Analysis* (EDA) dilakukan untuk *drop* atribut yang tidak digunakan. Setelah dilakukan proses *drop* atribut dilanjutkan dengan proses *Interquartile Range* (IQR) untuk pengecekan data *outlier* dari 10 atribut hasil *Exploratory Data Analysis* (EDA) dan pada Gambar 6 menampilkan proses *Interquartile Range* (IQR) yang terdapat data batas bawah, batas atas, dan *outlier*. Jika data pada atribut ditemukan data *outlier* maka akan muncul bulatan pada *outlier* sesuai pada Gambar 6 sedangkan jika tidak ada data *outlier* maka tidak akan muncul bulatan pada *outlier* yang contohnya pada Gambar 7.Proses *Interquartile Range* (IQR) dilakukan hanya pada 5 atribut yaitu pada atribut *Cycle(R/I)*, *Cycle_length(days)*, *AMH(ng/mL)*, *Follicle_No.(R)*, dan *Follicle_No.(L)* karena untuk 5 atribut *Weight_gain(Y/N)*, *hair_growth(Y/N)*, *Skin_darkening (Y/N)*, dan *Fast_food_(Y/N)* untuk isi dari atribut tersebut adalah 1 dan 0 yang berarti 1 terindikasi penyakit PCOS dan 0 tidak terindikasi penyakit PCOS.

Dalam proses *Interquartile Range* (IQR) ditemukan data *outlier* pada atribut *Cycle(R/I)* sebanyak 0 data *outlier*, *Cycle_length(days)* sebanyak 77 data *outlier*, *AMH(ng/mL)* sebanyak 52 data *outlier*, *Follicle_No.(R)* sebanyak 6 data *outlier*, dan *Follicle_No.(L)* sebanyak 0 data *outlier*. Dari pengecekan *outlier* menghasilkan total data 541 menjadi 448 data.



Gambar 6. Proses IQR Follicle_No.(L)



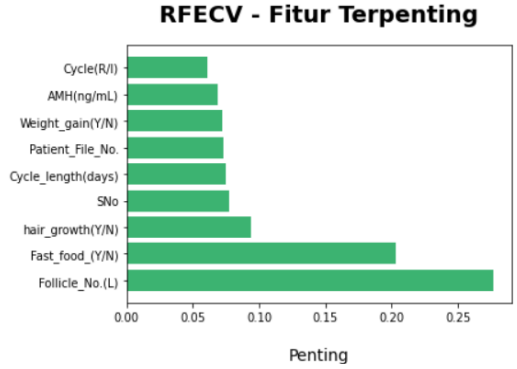
Gambar 7. Proses IQR Follicle_No.(R)

Setelah dilakukan pengecekan *Interquartile Range* (IQR) dilakukan *Preprocessing* 2 dengan mengecek nilai kosong atau nilai *null* pada 10 data atribut yang sudah dilakukan proses *Interquartile Range* (IQR) dan ditemukan nilai kosong atau nilai *null* pada *Cycle_length(days)*, *Follicle_No.(R)*, dan *Follicle_No.(R)* yang ada pada Gambar 8.

```
SNo                    0
Patient_File_No.       0
PCOS_(Y/N)             0
Cycle(R/I)             0
Cycle_length(days)     1
AMH(ng/mL)             0
Weight_gain(Y/N)       0
hair_growth(Y/N)       0
Skin_darkening (Y/N)   0
Pimples(Y/N)           0
Fast_food_(Y/N)        0
Follicle_No.(L)        9
Follicle_No.(R)       11
dtype: int64
```

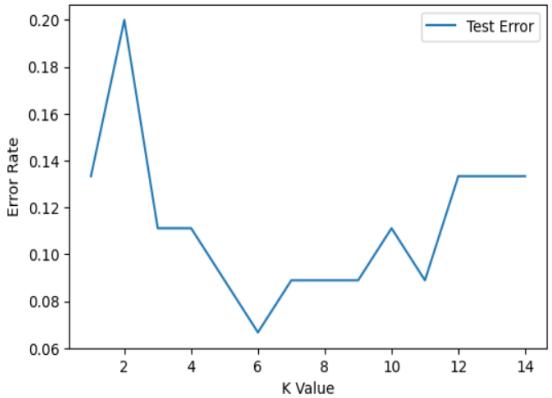
Gambar 8. Data RFECV

Untuk mengisi nilai kosong atau nilai *null* tersebut maka dilakukan proses pengisian nilai kosong atau nilai *null* dengan nilai rata-rata (*mean)*. Setelah diisi dengan nilai rata-rata (*mean)* dilanjutkan dengan dilakukan standarisasi data agar dapat dilanjutkan dengan proses *Recursive Feature Elimination and Cross-Validation* (RFECV).



Gambar 9. Data RFECV

Proses sebelum *Recursive Feature Elimination and Cross-Validation* (RFECV) dilakukan *split* data menjadi 10% data *testing* dan 90% data *training*. Dengan proses *Recursive Feature Elimination and Cross-Validation* (RFECV) dilakukan proses pengecekan atribut mana yang paling optimal yang akan digunakan pada proses klasifikasi *K Nearest Neighbor*(KNN) dengan menggunakan 10 atribut yang sebelumnya sudah dilakukan proses *Preprocessing* 2 dan menggunakan *Cross Validation* 10 k *fold* dan dari proses *Recursive Feature Elimination and Cross-Validation* (RFECV) menghasilkan 7 atribut terpenting atau optimal yaitu pada atribut *Follicle_No.(L)*, *Fastfood(Y/N)*, *hair_growth(Y/N)*, *Cycle_length(days)*, *Weight_gain(Y/N)*, *AMH(ng/mL)*, dan *Cycle(R/I)* yang ada pada Gambar 9. Setelah itu dilakukan proses perubahan data yang akan digunakan yaitu dengan menggunakan 7 data atribut dari proses *Recursive Feature Elimination and Cross-Validation* (RFECV) yang akan dilanjutkan dengan proses *K Nearest Neighbor (*KNN).



Gambar 10. *Hasil k terbaik*

Sebelum dilakukan klasifikasi dengan menggunakan *K Nearest Neighbor (*KNN), pada tahap pertama dilakukan proses penentuan nilai k dengan pengecekan nilai k dengan tingkat *error* yang rendah yang dilakukan pada nilai k yaitu 1 - 15. Pada pada Gambar 10 menghasilkan nilai k terbaik dengan tingkat *error* yang rendah yaitu pada nilai k = 6 yang akan digunakan untuk proses evaluasi dengan menggunakan *Confusion Matrix.*

Tabel 3. Hasil Pengujian

No	Pengujian	Hasil Pengujian
1	Akurasi	93%
2	Recall	82%
3	Precision	100%
4	F1 Score	90%

Pada Tabel 3 hasl pengujian adalah data dari hasil proses pengujian klasifikasi K-*Nearest Neighbor (*KNN) menggunakan *Confusion Matrix* dengan menghasilkan pengujian *accuracy* sebesar 93%, *recall* sebesar 82%, *precision* sebesar 100 %, dan *F1 Score sebesar* 90 %.

4. DISKUSI

Berbeda dengan penelitian sebelumnya, penelitian ini menggunakan *feature selection Exploratory Data Analysis* (EDA) yang menghasilkan 10 data yang digunakan dan dilakukan pengecekan kembali menggunakan *feature selection* Recursive *Feature Elimination and Cross-Validation* (RFECV) yang menghasilkan 7 atribut terpenting, kemudian dilanjutkan dengan proses klasifikasi *K-Nearest Neighbor (*KNN) dengan menghasilkan akurasi 93%.

Penelitian yang dilakukan oleh Hendrawan yang melakukan penelitian menggunakan algoritma *K-Nearest Neighbor (*KNN) dengan normalisasi dan seleksi fitur untuk mengklarifikasi penyakit liver. Algoritma ini digunakan karena memiliki prinsip sederhana dan mudah digunakan, tetapi memiliki nilai akurasi yang relatif rendah. Pada penelitian ini, nilai akurasi yang paling tinggi ketika menggunakan normalisasi *min-max* dengan seleksi fitur *Information Gain* dan *Gain Ratio Ratio*, nilai rata-rata yang digunakan untuk mengisi kekosongan data yaitu nilai k = 10.[2]

Penelitian yang dilakukan oleh Naufal dimana penelitian dilakukan pada jenis kanker. Penelitian ini menggunakan algoritma *K-Nearest Neighbor (*KNN) dan metode SVM yang mampu menghasilkan hasil akurasi tertinggi yaitu sebesar 98.54%. Namun SVM sangat sensitif dalam memilih nilai parameter yang digunakan sehingga perlu menggunakan kombinasi kernel dan parameter yang tepat guna menghasilkan akurasi yang maksimal.[3]

Berbeda dengan penelitian yang dilakukan Fajar pada penerapan metode *K-Nearest Neighbor* (KNN) untuk menentukan ikan cupang menggunakan deteksi tepi *canny* dan *invariant* menggunakan Matlab, menghasilkan nilai akurasi rata-rata sebesar 68,5714% dengan jumlah yang terdeteksi dengan baik sebanyak 48 data dan 22 data tidak terdeteksi dengan akurat dari total 70 data latih. Pengujian yang telah dilakukan menggunakan data uji terhadap 20 citra menghasilkan nilai akurasi rata-rata sebesar 70%, dengan jumlah data yang terdeteksi dengan baik adalah sebanyak 14 data dan 6 data tidak terdeteksi dengan baik, dan hasil akurasi yang kurang baik dapat disebabkan oleh bentuk sirip dan ekor ikan cupang yang tidak menentu.[17].

5. KESIMPULAN

Penelitian ini berhasil melakukan proses klasifikasi yang melalui 7 tahapan yaitu dari *download dataset* PCOS pada *Kaggle* dengan menggunakan dua *data set* file csv yaitu *PCOS_infertility.csv* dan data *without infertility _final.csv*, dilanjutkan *Preprocessing 1* dengan mengubah nilai kosong atau data *null* dan menghapus atribut yang tidak digunakan, selanjutnya dilanjutkan dengan proses *Exploratory Data Analysis* (EDA) dan dari proses tersebut digunakan 10 atribut kemudian dilanjutkan proses *Preprocessing 2* dengan proses IQR dan mengisi nilai kosong atau data *null* dengan menggunakan *mean*, selanjutnya dilakukan proses *Recursive Feature Elimination and Cross-Validation* (RFECV) yang menghasilkan 7 atribut terpenting atau optimal kemudian dilanjutkan klasifikasi dengan metode *K-Nearest Neighbor (*KNN), dan evaluasi. Kemudian dilanjutkan dengan proses *K-Nearest Neighbor (*KNN) dengan menghasilkan 7 data terpenting selanjutnya dilakukan klasifikasi dengan *K-Nearest Neighbor (*KNN) dengan menentukan nilai k terbaik yaitu nilai k = 6 dan tahap terakhir dilakukan proses evaluasi yang menghasilkan nilai *accuracy* sebesar 93%, nilai *precision* sebesar 100%, nilai *recall* sebesar 82%, dan nilai *F1 Scores* sebesar 90%.

UCAPAN TERIMA KASIH

Ucapan terima kasih ditujukan kepada Universitas Amikom Yogyakarta serta Program Studi S1 Informatika Fakultas Ilmu Komputer atas support dalam mengadakan riset kolaborasi penelitian.

DAFTAR PUSTAKA

[1] E. Maggyvin and M. I. Barliana, "*Literature Review*: Inovasi Terapi *Polycystic of Ovary Syndrome* (PCOS) Menggunakan *Targeted Drug Therapy Gen* CYP19 RS2414096," *Farmaka*, vol. 17, no. 1, pp. 107-118, 2019, doi: 10.24198/jf.v17i1.20829.g10054

[2] A. Hendrawan, L. M. Huizen, A. P. R. Pinem, and D. A. Wicaksana, "Implementasi Pemilihan Fitur Metode *Wrapper* dan *Embedded* dalam Prediksi Ketepatan Kelulusan Mahasiswa," *Pros. Sem. Nas. Pen. dan Peng. Kep. Masy. (SNPPKM 2021)*, 2021.

[3] S. A. Naufal, A. Adiwijaya, and W. Astuti, "Analisis Perbandingan Klasifikasi *Support Vector Machine* (SVM) dan *K-Nearest Neighbors* (KNN) untuk Deteksi Kanker dengan Data *Microarray*," *JURIKOM (Jur. Ris. Kom.)*, vol. 7, no. 1, pp. 162-168, 2020, doi: 10.30865/jurikom.v7i1.2014.

[4] M. S. Wibawa and K. D. P. Novianti, "Reduksi Fitur Untuk Optimalisasi Klasifikasi Tumor Payudara Berdasarkan data Citra FNA," *E-Proceedings KNS&I STIKOM Bali*, pp. 73-78, 2017.

[5] T. L. Basuki, J. Jondri, and U. N. Wisesty, "Deteksi *Polycystic Ovarian Syndrome* (PCOS) Menggunakan Klasifikasi *Microarray* Data dengan Algoritma *Artificial Neural Network* (ANN) *Backpropagation* dan *Principal Component Analysis*," *e-Proceeding of Engineering*, vol. 5, no. 3, 2018.

[6] S. Zulaikhah, A. Aziz, and W. Harianto, "Optimasi Algoritma *K-Nearest Neighbor* (KNN) dengan Normalisasi dan Seleksi Fitur untuk Klasifikasi Penyakit Liver," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 6, no. 2, pp. 439-445, 2022, doi: 10.36040/jati.v6i2.4722.

[7] R. T. Prasetio, "Seleksi Fitur dan Optimasi Parameter K-NN Berbasis Algoritma Genetik Pada *Dataset* Medis", *Jurnal Responsif: Riset Sains dan Informatika*, vol. 2, no. 2, pp. 213-221, 2020, doi: 10.51977/jti.v2i2.319.

[8] E. D. Wahyuni, A. A. Arifiyanti, and M. Kustyani, "*Exploratory Data Analysis* dalam Konteks Klasifikasi Data *Mining*," *ReTII (Rek. Tek. Indus. Infor XIV)*, pp. 263-269, 2019.

[9] I. Pratama, A. Y. Chandra, and P. T. Prasetyaningrum, "Seleksi Fitur dan Penanganan *Imbalanced* Data menggunakan RFECV dan ADASYN," *Jur. Eks. Inf.*, vol. 11, no. 1, pp. 38-49, 2021, doi: 10.30864/eksplora.v11i1.578.

[10] A. Ardiyansyah, P. A. Agustia, and R. Maulana, "Analisis Perbandingan Algoritma Klasifikasi Data *Mining* Untuk *Dataset Blogger* Dengan *Rapid Miner*," *Jurnal Khatulistiwa Informatika*, vol. 6, no. 1, pp. 20-28, 2018, doi: 10.31294/jki.v6i1.3799.g2437.

[11] F. V. P.Samosir, L. P. Mustamu, E. D. Anggara, A. I. Wiyogo, and A. Widjaja, "*Exploratory Data Analysis* terhadap Kepadatan Penumpang Kereta Rel Listrik," *Jurnal Teknik Informatika dan Sistem Informasi (JuTISI)*, vol. 7, no. 2, pp. 449-467, 2021, doi: 10.28932/jutisi.v7i2.3700.

[12] M. D. Nurmalasari, K. Kusrini, and S. Sudarman, "Komparasi Algoritma *Naive Bayes* dan *K-Nearest Neighbor* untuk Membangun Pengetahuan Diagnosa Penyakit Diabetes", *Jurnal Komtika*, vol. 5, no. 1, pp. 52-59, 2021.

[13] M. S. Faradisa, "Implementasi IQR-SMOTE Untuk Mengatasi Ketidakseimbangan Kelas Pada Klasifikasi Diabetes Menggunakan *K-Nearest Neighbors*," *JIK (Jurnal Ilmu Komputer)*, vol. 15, no. 1, pp. 48-60, 2022.

[14] R. Permatasari and A. Wibowo,

"*Implementation of Support Vector Machine - Recursive Feature Elimination for MicroRNA Selection in Breast Cancer Classification*," *Jurnal EECCIS*, vol. 14, no. 1, pp. 1-5, 2020, doi: 10.21776/jeeccis.v14i1.602.

[15] M. L. Huang, Y. H. Hung, W. M. Lee, R. K. Li, and B. R. Jiang, "*SVM-FRE Based Feature Selection and Taguchi Parameters Optimization for Multiclass SVM Classifier*," *Hindawi Publishing*, vol. 2014, pp. 1-10, 2014, doi: 10.1155/2014/795624.

[16] A. Amiruddin and R. Ishak, "Implementasi Seleksi Fitur Klasifikasi Waktu Kelulusan Mahasiswa Menggunakan *Correlation Matrix With Heatmap*," *Jambura Journal of Electrical and Electronics Engineering (JJEEE)*, vol. 4, no. 2, pp. 169-174, 2022.

[17] S. Fajar, E. W. Hidayat dan N. I. Kurniati, "Penerapan Metode *K-Nearest Neighboar* (KNN) untuk Menentukan Ikan Cupang Menggunakan Deteksi Tepi *Canny* dan *Invariant Moment," Jurnal Teknik Informatika (JUTIF)*, vol. 3, No 1, pp. 11-20, 2022, doi: 10.20884/1.jutif.2022.3.2.95.