# KLASIFIKASI PENYAKIT GINJAL KRONIS DENGAN MENGGUNAKAN ALGORITMA MACHINE LEARNING

Muhammad Revaldo Adzim
*Informatics engineering*
*Faculty of Computer*
*Sriwijaya University*
South Sumatera, Indonesia
revaldoadzim@gmail.com

Annisa Darmawahyuni
*Inteligent System Research Group*
*Faculty of Computer*
*Sriwijaya University*
South Sumatera, Indonesia
riset.annisadarmawahyuni@gmail.com

***Abstract— Penelitian ini bertujuan mengembangkan model klasifikasi penyakit Ginjal Kronis dengan memanfaatkan Algoritma Machine Learning sebagai metode seleksi fitur sebelum proses klasifikasi. Dataset diperoleh dari platform Kaggle yang meliputi data klinis dan gaya hidup, proses preprocessing dilakukan melalui pembersihan data, penghapusan atribut non relevan, normalisasi menggunakan Min-Max Scaling, serta penanganan ketidakseimbangan data dengan tiga pendekatan, yaitu tanpa balancing (original), random undersampling, dan Synthetic Minority Oversampling Technique (SMOTE). Seleksi fitur dilakukan menggunakan metode SelectKBest untuk memilih atribut paling berpengaruh terhadap diagnosis CKD Tiga algoritma klasifikasi digunakan, yaitu K-Nearest Neighbor (KNN), Support Vector Machine (SVM), dan Random Forest (RF), dengan tuning parameter untuk memperoleh hasil optimal. Evaluasi dilakukan menggunakan metrik akurasi, sensitivitas, spesifisitas, presisi, dan f1-score. Hasil pengujian menunjukkan bahwa model RandomForest dengan n=200, pada skenario SMOTE mencapai performa terbaik dengan akurasi di atas 96%, precision di atas 99%, recall 93%. Hasil ini membuktikan bahwa seleksi fitur SelectKBest dengan algoritma klasifikasi serta strategi balancing data yang tepat mampu meningkatkan performa prediksi penyakit Ginjal Kronis secara signifikan.***

***Keywords—Chronic Kidney Disease, Seleksi Fitur, SelectKBest), Machine Learning, SMOTE, Random Forest***

## I. PENDAHULUAN

Penyakit Ginjal Kronis (Chronic Kidney Disease/CKD) merupakan salah satu masalah kesehatan serius yang menyerang Bagian Fungsi Ginjal dan dapat menurunkan kualitas hidup penderitanya[1]**.** Menurut World Health Organization (WHO), lebih dari 500 juta orang di seluruh dunia menderita penyakit Ginjal Kronis, dan jumlah ini terus meningkat setiap tahun. Gejala Ginjal kronis bervariasi namun tidak berspesifik, mulai dari lemas, gatal hingga kehilangan nafsu makan, penyakit Ginjal Kronis sangat berpotensi membahayakan nyawa. Oleh karena itu, deteksi dan diagnosis yang akurat menjadi kunci dalam penanganan Chronic Kidney Disease (CKD) secara efektif

Proses diagnosis penyakit Ginjal kronis memiliki banyak tahapan pemeriksaan, diawali dengan tes skrining (dipstick urin dan tes darah) diakhiri dengan tes keseluruhan yaitu Glomerular Filtratio Rate (GFR), namun dalam beberapa wilayah, keterbatasan alat serta minimnya pengalaman tenaga medis turut menjadi hambatan. Hal ini mendorong perlunya solusi alternatif berbasis teknologi untuk mendukung proses diagnosis, khususnya melalui pendekatan klasifikasi data medis secara otomatis.

Perkembangan teknologi informasi, khususnya pada bidang Kecerdasan buatan, telah membawa pengaruih signifikan dalam dunia kesehatan , mampu meningkatkan efektivitas kerja tenaga medis dan mempercepat pengambilan keputusan yang berbasis data. Penerapan teknologi digital berperan pentin dalm mendukung proses Klasifikasi maupun prediksi penyakit dengan hasil yang lebih akurat, serta dapat diterapkan pada sklala yang lebih luas.[2]

Perkembangan Machine Learning (ML) memberikan kontribusi penting dalam bidang kesehatan karena mampu mengelola data klinis dalam jumlah besar untuk menghasilkan informasi diagnostik[3]. ML bekerja dengan mempelajari pola dari data sebelumnya dan kemudian menggeneralisasikannya untuk melakukan prediksi terhadap data baru dengan tingkat ketepatan yang tinggi.. ML efektif dalam mengidentifikasi secara manual dari data pasien, sehingga mampu mempercepat proses diagnosis serta pengambilan keputusan medis. Berbagai algoritma seperti K-Nearest Neighbor (KNN), Support Vector Machine (SVM), dan Random Forest (RF) telah diaplikasikan secara luas dalam penelitian medis dan terbukti memiliki peforma yang baik [4][5]

Salah satu permasalah umum dalam penerapan algoritma klasifikasi pada data medis adalah ketidakseimbangan kelas (imbalanced data). Kondisi ini terjadi ketika jumlah data pada kelas dominan jauh lebih banyak dibandingkan kelas minoritas. Akibatnya, model yang dibangun sering kali lebih akurat dalam memprediksi kelas mayoritas, namun gagal mengenali kelas minoritas yang justru sering mewakili kasus penting pada domain medis. Beberapa penelitian menekankan bahwa isu ini harus diatasi melalui teknik balancing data, baik dengan menambah sampel kelas minoritas (oversampling) maupun mengurangi sampel kelas mayoritas (undersampling), agar performa model dapat lebih stabil.[6]

Selain penanganan ketidakseimbangan data, aspek lain yang berpengaruh terhadap akurasi klasifikasi adalah seleksi fitur (*feature selection*). Kehadiran atribut yang tidak relevan atau bersifat redundan dapat mengurangi kinerja model, baik dari segi akurasi maupun efesiensi komputasi. Oleh karena itu, diperlukan teknik untuk memilih fitur yang signifikan terhadap variable target. Salah satu metode umum adalah SeleckKBest dengan pendekatan ANOVA F-test

yang berfungsi menyeleksi nilai signifikansi statistiknya dan fitur yang paling berpengaruh terhadap diagnosis penyakit yang dipertahankan, sehingga performa algoritma Klasifikasi dapat ditingkatkan secara optimal .[7]

Penelitian ini bertujuan untuk menerapkan metode SelectKBest dalam proses seleksi fitur pada klasifikasi Ginjal kronis (CKD) serta menguji performa tiga algoritma *machine learning*, yaitu KNN, SVM, RF dan tiga skenario data, yaitu original, undersampling dan smote. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi dalam peningkatan akurasi klasifikasi data medis sekaligus mendukung pengembangan sistem pendukung keputusan dalam diagnosisi penyakit berbasis data.

## II. TINJAUAN PUSTAKA

### A. *Machine Learning dalam Klasifikasi Medis*

*Machine Learning (ML)* merupakan salah satu cabang kecerdasan buatan yang banyak digunakan dalam bidang kesehatan, khususnya untuk klasifikasi dan prediksi penyakit. ML mampu mengenali pola dari data klinis dan menghasilkan model prediktif yang membantu proses diagnosis maupun pemantauan kondisi pasien secara lebih akurat dan efisien. Sejumlah penelitian menunjukkan bahwa penerapan ML dalam data medis dapat mengungkap pola tersembunyi yang sulit dikenali secara manual, sehingga mempercepat proses klasifikasi penyakit dan mendukung pengambilan keputusan berbasis data. [3]

Dalam berbagai studi, algoritma seperti support Vector Machine (SVM), Random forest (RF) dan K-Nearest Neighbor (KNN) telah digunakan secara luas pada klasifikasi penyakit, termasuk ginjal kronis [5]. Penelitiam mengenai CKD menunjukan bahwa algoritma RF dam SVM cenderung memberikan akurasi lebih tinggi, sementara KNN tetao relevan pada dataset dengan karakteristik tertentu [4]. Selain itu, sejumlah literatur juga menekankan pentingnya tahapan *Preprocessing* dan *balancing data* dengan SMOTE maupun Undersampling dalam meningkatkan stabilitas model klasifikasi pada data medis yang umunya tidak seimbang [8]

### B. *Algoritma Klasifikasi*

Dalam bidang medis, berbagai algoritma *Machine learning* telah digunakan untuk menyelesaikan tugas klasifikasi, diantaranya **K-Nearest Neighbor (KNN), Support Vector Machine (SVM), dan Random forest (RF).**

Algoritma KNN bekerja dengan menghitung jarak antara data uji dan data latih, kemudian meenentukan kelas berdasarkan matyoritaas tetangga terdekat. Seju8mlah penelitian menunjukan bahwa KNN Mampu memberikan hasiil yang kompetitif dalam klasifikasi data medis, khususnya ketika nilai parameter K ditentukan secara optinmal.

Support Vector Machine (SVM) merupakan algoritma berbasis teori margin yang memisahkan data ke dalam dua kelas melalui *hyperplane* optimal. Kombinasi SVM dengan metode seleksi fitur terbukti dapat meingkatkan akurasi serta efesisensi komputasi, terutama pada dataset berskala tinggi[5]. Hal ini menjadikan SVM banyak digunakan dalam studi medis, termasuk klasifikasi penyakit Ginjal Kronis.

Random Forest (RF) adalah algoritma berbasis *ensemble* yang terdiri dari sejumlah pohon keputusan untuk menghasilkan prediksi yang lebih stabil. Penelitian sebelumnya menunjukan bahwa RF mampu mencapai akurasi tinggi pada klasifikasi penyakit kronis dengan dukungan tahapan preprocessing, normalisasi data, serta tuning parameter [5]**.** Keunggulan RF terletak pada kemampuannya menangani data dengan jumlah atribut yang besar serta menguirangi risiko *overfitting.*

### C. *Selection feature dengan SelectKBest*

Seleksi fitur merupakan salah satu tahap penting dalam pengolahan data untuk klasifikasi medis. Menurut El Touati Y dan J saidani T [7] Atribut yang tidak relevan atau redundan dapat menurunkan kinerja model karena meningkatkan komplektsitas perhitungan dan risiko overfitting. Oleh sebab itu, diperlukan metode seleksi fitur untuk memilih atribut yang paling berpengaruh terhadap variable target.

Selain itu, salah satu teknik seleksi fitur yang dilakukan juga oleh “Silimane J”[7] adalah SelectKBest, metode ini bekerja dengan menghitung skor statistik setiap fitur terhadap variable target, lalu memilih K fitur terbaik berdasarkan skor tersebut, Pada penelitian ini digunakan pendekatan **ANOVA F- test** untuk menghitung tingkat signifikansi fitur. ANOVA F-test mengevaluasi perbedaan rata rata antar kelompok, sehingga fitur dengan nilai signifikansi tinggi dianggap lebih relevan terhadap klasifikasi.[9]

Penelitian[7] menunjukan bahwa penerapan seleksi fitur manpu meningkatkan akurasi serta efisiensi model klasifikasi. Studi komparatif [5], [7] juga menegaskan bahwa kombinasi seleksi fitur dengan algoritma ML sepeti SVM dan Random forest menghasilkan model yang lebih stabil dan akurat dibandingkan tanpa seleksi fitur. Dengan demikian, penggunaan SelectKBest memperkuat performa model klasifikasi penyakit Ginjal Kronis pada penelitian ini.

### D. *Penanganan Ketidakseimbangan Data dengan Undersampling*

Ketidakseimbangan kelas merupakan salah satu tantangan yangsering muncul dalam analisis data medis, termasuk pada klasifikasi penyakit ginjal kronis. Kondisi ini dapat menyebabkan model klasifikasi cenderung lebih akurat dalam mengenali kelas mayoritas, tetapi gagal mendeteksi kelas minoritas yang justru sering mewakili kasus penting. Salah satu teknik mengatasi permasalahan ini adalah **Undersampling**, yaitu mengurangi jumlah sampel dari kelas mayoritas hingga seimbang dengan kelas minoritas.[6]

Menurut Mayuri.S[6] *Undersampling* dapat membantu menyeimbangkan distribusi kelas, sehingga model memiliki peliuang yang lebih besar untuk mengenali kelas minoritas. Studi sebelumnya menyebutkan bahwa Undersampling dapat meningkatkan kinerja model pada kondisi data yang sangat timpang, Namuin, Kelemahan utamanya adalah berpotensi hilangnya informasi penting akibat penghapusan data pada kelas mayoritas. Penelitian lain [8] juga melaporkan bahwa meskipun Undersampling dapat memperbaiki distribusi data, performanya cenderung lebih rendah dibandingkan oversampling seperti SMOTE, karena berkurangnya informasi

Dengan demikian, *undersampling* tetap dapat dipertimbangkan sebagai pendekatan awal untuk menangani ketidakseimbangan kelas, tetapi penerapannya perlu

dipertimbangkan secara hati-hati, terutama pada dataset medis yang berukuran terbatas.

*E. Penanganan Ketidakseimbangan Data dengan SMOTE*

Ketidakseimbangan distribusi kelas sering menjadi tantangan dalam klasifikasi medis, khususnya ketika jumlah sampel dari kelas minoritas sangat rendah. Untuk mengatasi hal ini, digunakan Synthetic Minority Over-sampling Technique (SMOTE). SMOTE bekerja dengan cara membuat sampel sintetis baru pada kelas minoritas berdasarkan interpolasi antar data yang berdekatan, sehingga distribusi data menjadi lebih seimbang tanpa mengurangi jumlah data pada kelas mayoritas.

SMOTE diperkenalkan oleh Chawla et al.(2002[8])dan hingga kini telah digunakan secara luas dalam berbagai penelitian klasifikasi medis.SMOTE mampu meningkatkan stabilitas dan akurasi model pada data yang sangat tidak seimbang, karena memberikan representasi yang lebih baik terhadap kelas minoritas

Meskipun demikian, SMOTE juga memiliki keterbatasan, terutama pada data berdimensi tinggi yang dapat menimbulkan risiko penciptaan sampel sintetis kurang representatif tetap dianggap sebagai salah satu metode balancing data yang paling efektif, terutama bila dikombinasikan dengan algoritma machine learning modern.

*F. Kesimpulan*

Berdasarkan tinjauan pustaka di atas, dapat disimpulkan bahwa kombinasi metode seleksi fitur SelectKBest dengan algoritma klasifikasi seperti KNN, SVM, dan RF memiliki potensi besar dalam meningkatkan akurasi klasifikasi penyakit. Penggunaan Undersampling dan SMOTE juga mendukung peningkatan performa model dalam situasi data yang tidak seimbang, menjadikan pendekatan ini sangat relevan untuk aplikasi klasifikasi Chronic Kidney Disease.

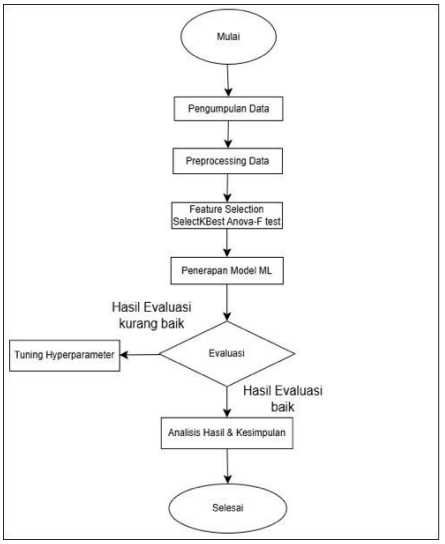
## III. METODOLOGI PENELITIAN

*A. DATASET*

Penelitian ini menggunakan dataset Chronic Kidney Disease (CKD) yang diperoleh dari situs Kaggle. Dataset ini berisi 1.659 sampel dengan 54 atribut yang mencakup informasi demografiis, gaya hidup, serta hasil pemeriksaan klinis pasien. Sebelum digunakan untuk proses pelatihan model, dataset dibagi tiga bagian, data latih, data validasi dan data uji

Atribut pada dataset ini meliputi variable demografis (misalnya usia, jenis kelamin), gaya hidup ( aktivitas fisik, pola makan), serta parameter klinis seperti tekanan darah, kreatin serum dan eGFR., targer kolom Diagnosis ada dua kelas yaitu =

- 0 untuk pasien tidak CKD
- 1 untuk pasien CKD




Gambar 1. Diagram Skema Metodologi Penelitian

*B. PRA-PEMROSESAN DATA*

Pra-pemrosesan data merupakan langkah penting sebelum membangun model klasifikasi agar data yang digunakan berada dalam kondisi yang konsisten dan optimal. Dalam penelitian ini, pra-pemrosesam dilakukan melalui beberapa tahap, yaitu pembersihan data, pembagian dataset, normalisasi serta penanganan ketidakseimbangan kelas, untuk penanganan ketidakseimbangan kelas, ada 3 seknario model, yaitu :

- Model 1 (Original): dataset digunakan tanpa penanganan ketidakseimbangan kelas (imbalanced)
- Model 2 Undersampling dilakukan random undersampling pada kelas mayoritas (kelas 0) agar distribusi kelas lebih seimbang
- Model 3 (SMOTE): dilakukan Synthetic Minority Over-sampling Technique (SMOTE) untuk menambah sampel sintetis pada kelas minoritas sehingga distribusi kelas menjadi seimbang

Ketiga skenario ini digunakan untuk membandingkan pengaruh metode balancing terhadap kinerja algoritma klasifikasi yang digunakan.

Tahapan pra-pemrosesan dilakukan dengan memeriksa adanya missing values dan data duplikat. Berdasarkan hasil pemeriksaan, dataset yang digunakan tidak mengandung nilai kosong, sehingga dapat langsung diproses ke tahap berikutnya. Dilanjuitkan dengan normalisasi pada fitur numerik agar klasifikasi dapat bekerja lebih optimal.

Langkah berikutnya adalah menangani ketidakseimbangan kelas. Pada skenario Model 2 (Undersampling), jumlah sampel dari kelas mayoritas dikurangi secara acak (*random undersampling)* sehingga proporsinya mendekati jumlah kelas minoritas. Sementara itu pada Model 3 (SMOTE) digunakan metode *Synthetic Minority Oversampling Technique,* yaitu dengan membangkitkan sampel sintetis baru pada kelas minoritas sebelum proses pembagian dataset. Strategi ini bertujuan agar data antara kelas mayoritas dan minoritas berada dalam kondisi yang seimbang. Proses pembangkitan data pada SMOTE dituliskan dirumus berikut :

antar kelas, sehingga fitur fitur tersebut dianggap

$$x_{new} = x_i + \delta \times (x_{nn} - x_i) \qquad (1)$$

Keterangan :

- $x_i$ = sampel minoritas,
- $x_{nn}$ =tetangga terdekat dari $x_i$,
- δ = bilangan acak antara 0 dan 1

Menurut penelitian[8] penerapan SMOTE terbukti mampu memberikan peningkatan signifikann terhadap perfoma klasifikasi pada dataset yang mengalami ketidakseimbangan, khususnya dibidang medis. Hal serupa juga ditegaskan oleh penelitian [10]yang menunjukan bahwa SMOTE sangat efektif dalam membantu model mendeteksi kelas minoritas yyang seringkali terabaikan oleh metode konvensional.

Setelah dilakukan balancing sesuai dengan masing-masing skenario (original, undersampling, dan SMOTE), dataset

Dataset kemudian dibagi menjadi tiga, yaitu 80% untuk data latih, 10% untuk data validasi dan 10% untuk data uji. Pembagian dilakukan dengan teknik stratified split, sehingga proporsi distribusi kelas pada setiap subset tetap terjaga dan representatif.

memungkinkan model untuk dilatih secara optimal dan dievaluasi pada data yang representatif terhadap distribusi populasi sebenarnya.

Langkah terakhir adalah melakukan normalisasi terhadap fitur numerik menggunakan metode *Min-Max Scaling*. Normalisasi ini penting untuk menyetarakan rentang nilai setiap fitur agar berada dalam interval 0, 1. Hal ini bertujuan untuk mencegah algoritma berbasis jarak seperti K-Nearest Neighbor (KNN) maupun algoritma berbasis kernel seperti Support Vector Machine (SVM) menjadi bias terhadap fitur yang memiliki skala lebih besar.

Proses Transformasi Min-Max dituliskan dengan rumus yaitu:

$$x^{\%} = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (2)$$

Keterangan :

- $X$ = nilai asli,
- $X^{\%}$= nilai hasil normalisasi,
- $X_{min}$ dan $X_{max}$ = nilai minimum dan maksimum

Metode ini membantu mencegah dominasi fitur berskala besar dan mempercepat proses pelatihan model, sebagaimana dijelaskan di dalam peneltiian [12]. Normalisasi diterapkan secara konsisten pada seluruh data latih, validasi, dan uji.

*C. SELEKSI FITUR DENGANSelectKBest*

SelectKBest merupakan salah satu metode seleksi fitur berbasis uji statistik yang digunakan untuk memilih sejumlah fitur terbaik dari dataset. SelectKBest bekerja dengan fungsi skor ANOVA F-test. Pendekatan ini bekerja dengan menghitung nikai F-statistic untuk setiap fitur terhadap label kelas, kemudian memilih K fitur dengan skor tertinggi. Nilai F-statitstic yang besar menunjukan perbedaan rata-rata yang signifikan

lebih relevan dalam proses klasifikasi.

Metode SelectKBest relatif sederhana dan efisien, karena tidak memerlukan parameter tambahan selain jumlah fitur (*K)* yang dipilih. Sejumlah penelitian sebelumnya juga menunjukan bahwa kombinasi seleksi fitur dengan algoritma klasifikasi seperti Support Vector Machine (SVM) dan Random Forest (RF) dapat meningkatkan akurasi sekaligus mengurangi waktu komputasi.

Dalam penelitian ini, setiap fitur dievaluasi berdasarkan nilai F-statistic yang dihitung menggunakan uji ANOVA F- test dengan nilai *K*= 20. Nilai F-statistic mencerminkan sejauh mana rata rata nilai suatu fitur berbeda secara signifikan antar kelas target.

Proses seleksi dilakukan dengan mengurutkan semua fitur berdasarkan skor F-test, kemudian dipilih dengan nilai tertinggi untuk gunakan tahap pelatihan model. Rumus umum perhitungan ANOVA F-test adalah:

$$x = \frac{MSantara}{MSdalam} \quad (3)$$

Fitur dengan nilai ***F*** lebih besar menunjukan adanya perbedaam lebih signifikan dan lebih informatif untuk proses klasifikasi.

- *M S*antara = variasi antar kelas seberapa besar perbedaan rata-rata kelas target
- *M S*dalam = variasi dalam kelas variabilitas dalam masing masing kelas

CKD : sehat vs sakit, nilai F besar berarti fitur tersebut memberikan perbedaan rata rata yang jelas antara kedua kelas.

*D. PEMODELAN KLASIFIKASI*

Setelah melalui tahap pra-pemrosesan data dan seleksi fitur menggunakan SelectKBest (ANOVA F-test), langkah berikutnya adalah membangun model klasifikasi. Pada penelitian ini digyuunakan tiga algoritma pembelajaran mesin, yaitu K-Nearest Neighbor (KNN), Support Vector Machine (SVM), dan Random Forest (RF).

- K-Nearest Neighbor (KNN)

KNN merupakan metode klasifikasi berbasis jarak, yang suatu data di uji berdasarkan mayoritas kelas dari label dari k tetangga terdekatnya. Penelitian [3] menyatakan bahwa KNN efektif dalam klasifikasi penyakit paru-paru dengan akurasi cukup tinggi. Dalam penelitian ini, model KNN diuji dengan beberapa parameter nilai k = 3, 5, 7, 9, dan 11 Pemilihan parameter terbaik dilakukan berdasarkan performa tertinggi pada data validasi dan uji.

- Support Vector Machine (SVM)

SVM adalah algoritma klasifikasi berbasis margin yang mencari *hyperplane* optimal untuk memisahkan dua kelas. Untuk meningkatkan performa, dilakukan *tuning* hyperparameter pada nilai **C** (0.1, 1, 10) dan parameter\. SVM dipilih karena terbukti efektif dalam klasifikasi data medis berskala besar dengan margin yang jelas.

- Random Forest (RF)

Random Forest adalah algoritma klasifikasi yang menggabungkan hasil dari sejumlah pohon keputusan. Algoritma ini bekerja dengan prinsip *bagging*, dimana setiap pohon dilatih pada subset data yang berbeda. Dalam penelitian ini, jumlah pohon yang digunakan adalah 100 dan 150. Random Forest dipilih karena robust terhadap noise dan mampu menangani data dengan banyak fitur.

### *E. EVALUASI MODEL*

Untuk mengukur performa model klasifikasi, digunakan lima metrik evaluasi utama, yaitu akurasi, presisi, recall, spesifisitas dan f1-score. Secara khusus, perhatian lebih difokuskan pada kelas minoritas (pasien CKD positif), karena ketidakseimbangan distribusi kelas dapat membuat akurasi keseluruhan menyesatkan jika hanya mengandalkan prediksi kelas mayoritas..Berikut adalah rumus-rumus dari masing-masing metrik evaluasi:

- Akurasi

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (4)$$

- Presisi

$$presisi = \frac{TP}{TP + FP} \times 100 \quad (5)$$

- Recall

$$Recall = \frac{TP}{TP + FN} \times 100 \quad (6)$$

- Spesifisitas

$$spesifisitas = \frac{TN}{TN + FP} \times 100 \quad (7)$$

- F1-Score

$$F1 = 2 \times \frac{Presisi \times Recall}{Presisi + Recall} \times 100 \quad (8)$$

Evaluasi performa model merupakan tahap penting dalam proses klasifikasi, terutama pada data medis yang umumnya memiliki distribusi kelas yang tidak seimbang. Oleh karena itu, penilaian performa tidak hanya didasarkan pada akurasi, tetapi juga menggunakan metrik lain seperti precision, recall, f1-score, sensivitas dan spensifisitas..

Menurut penelitian [17] penggunaan metrik evaluasi sangat krusial dalam menilai efektivitas algoritma machine learning pada domain kesehatan. Hal ini dikarenakan nilai akurasi yang tinggi belum tentu mencerminkan peforma yang optimal apabila model gagal mengenali kelas minoritas secara tepat. Pada kasus Klasifikasi penyakit, kemampuan model dalam mendeteksi pasien yang benar benar menderita penyakit (kelas positif) menjadi sangat penting, sehingga kombinasi recall dan precision yang direpresentasikan dalam f1-score dianggap sebagai indikator yang lebih representatif .

Metrik-metrik tersebut tidak hanya menggambarkan jumlah prediksi benar secara keseluruhan, tetapi juga mengukur sejauh mana model mampu mengenali pasien CKD secara akurat.

## IV. HASIL PENELITIAN

### *A. Hasil Penelitian*

Pada tahap ini dipaparkan hasil implementasi dari alur penelitian yang telah dirancang, meliputi pra-pemrosesan data, penanganan ketidakseimbangan kelas, pembagian dataset, serta seleksi fitur menggunakan SelectKBest (ANOVA-F test) dengan nilai *K*=20, Seluruh tahapan dilakukan untuk memastikan data berada dalam kondisi yang bersih, seimbang dan optimal seblum gigunakan dalam proses pelatihan model klasifikasi.

Tahap pra-pemrosesan dimulai dengan pemeriksaan kelengkapan data. Dataset Chronic Kidney Disease (CKD) tidak mengandung nilai kosong dan nilai duplikat, sehingga dpaat langsung diproses ketahap berikutnya. Selanjuitnya, dilakukan penanganan ketidakseimbangan kelas melalui tiga skenario berbeda, yaitu

- Model 1 (Original): dataset digunakan tanpa perubahan distribusi kelas
- Model 2 (Undersampling): sebagian data pada kelas mayoritas dihapus secara acak agar distribusi kelas lebih seimbang
- Model 3 (SMOTE): dilakukan penambahan sampel sintetis pada kelas minoritas menggunakan Synthetic Minority Over-sampling Technique (SMOTE) langsung pada dataset asli sebelum

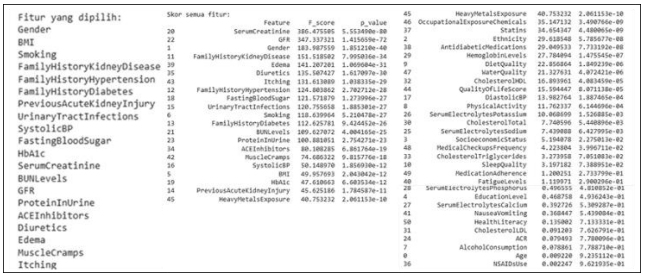
Pemisahan data

Setelah proses balancing, dataset kemudian dibagi menjadi tiga subset dengan perbandingan 80% data latih,

10% data validasi, dan 10% data uji. Proses pemisahan

dilakukan secara stratified sampling agar proporsi kelas tetap konsisten di setiap subset. Tahap berikutnya adalah normalisasi fitur numerik menggunakan metode Min–Max Scaling agar semua atribut berada dalam rentang [0,1]. Normalisasi diperlukan untuk mencegah bias pada algoritma berbasis jarak seperti KNN maupun algoritma dengan margin seperti SVM.

Gambar 2. Visualisasi fitur terpilih berdasarkan nilai f-score



```
Fitur yang dipilih:
Gender
BMI
Smoking
FamilyHistoryKidneyDisease
FamilyHistoryHypertension
FamilyHistoryDiabetes
PreviousAcuteKidneyInjury
UrinaryTractInfections
SystolicBP
FastingBloodSugar
HbA1c
SerumCreatinine
BUNLevels
GFR
ProteinInUrine
ACEInhibitors
Diuretics
Edema
MuscleCramps
Itching
```

Skor semua fitur:

| | Feature | F_score | p_value |
|---|---|---|---|
| 20 | SerumCreatinine | 386.475585 | 5.553490e-80 |
| 22 | GFR | 347.337321 | 1.415059e-72 |
| 1 | Gender | 183.987559 | 1.851210e-40 |
| 11 | FamilyHistoryKidneyDisease | 151.518582 | 7.995816e-34 |
| 39 | Edema | 141.207201 | 1.069004e-31 |
| 35 | Diuretics | 135.507427 | 1.617097e-30 |
| 43 | Itching | 131.613089 | 1.018315e-29 |
| 12 | FamilyHistoryHypertension | 124.803062 | 2.702712e-28 |
| 18 | FastingBloodSugar | 121.571879 | 1.273996e-27 |
| 15 | UrinaryTractInfections | 120.755658 | 1.885301e-27 |
| 6 | Smoking | 118.639964 | 5.210478e-27 |
| 13 | FamilyHistoryDiabetes | 112.625781 | 9.424612e-26 |
| 21 | BUNLevels | 109.627072 | 4.004165e-25 |
| 23 | ProteinInUrine | 100.881051 | 2.754271e-23 |
| 34 | ACEInhibitors | 80.108285 | 6.861764e-19 |
| 42 | MuscleCramps | 74.686322 | 9.815770e-18 |
| 16 | SystolicBP | 58.148970 | 1.856930e-12 |
| 5 | BMI | 49.957693 | 2.043042e-12 |
| 19 | HbA1c | 47.610663 | 6.603534e-12 |
| 14 | PreviousAcuteKidneyInjury | 45.625186 | 1.784587e-11 |
| 45 | HeavyMetalsExposure | 40.753232 | 2.061153e-10 |
| 45 | HeavyMetalsExposure | 40.753232 | 2.061153e-10 |
| 46 | OccupationalExposureChemicals | 35.147132 | 3.490766e-09 |
| 37 | Statins | 34.654147 | 4.480065e-09 |
| 2 | Ethnicity | 29.618548 | 5.785677e-08 |
| 38 | AntidiabeticMedications | 29.049533 | 7.733192e-08 |
| 29 | HemoglobinLevels | 27.784094 | 1.475545e-07 |
| 9 | DietQuality | 22.856864 | 1.649239e-06 |
| 47 | WaterQuality | 21.327631 | 4.072421e-06 |
| 32 | CholesterolHDL | 16.893961 | 4.883459e-05 |
| 44 | QualityOfLifeScore | 15.594447 | 8.071138e-05 |
| 17 | DiastolicBP | 13.982764 | 1.887465e-04 |
| 8 | PhysicalActivity | 11.762337 | 6.144696e-04 |
| 26 | SerumElectrolytesPotassium | 10.068169 | 1.526885e-03 |
| 30 | CholesterolTotal | 7.740596 | 5.448896e-03 |
| 25 | SerumElectrolytesSodium | 7.439088 | 6.427995e-03 |
| 3 | SocioeconomicStatus | 5.194078 | 2.275013e-02 |
| 48 | MedicalCheckupsFrequency | 4.223804 | 3.995711e-02 |
| 33 | CholesterolTriglycerides | 3.273958 | 7.051083e-02 |
| 10 | SleepQuality | 3.107182 | 7.388951e-02 |
| 49 | MedicationAdherence | 1.200251 | 2.733799e-01 |
| 40 | FatQualityLevels | 1.119971 | 2.900256e-01 |
| 28 | SerumElectrolytesPhosphorus | 0.696555 | 4.018852e-01 |
| 4 | EducationLevel | 0.468758 | 4.936243e-01 |
| 27 | SerumElectrolytesCalcium | 0.392726 | 5.309287e-01 |
| 41 | NauseaVomiting | 0.368447 | 5.439084e-01 |
| 50 | HealthLiteracy | 0.135082 | 7.133331e-01 |
| 31 | CholesterolLDL | 0.091203 | 7.626791e-01 |
| 24 | ACR | 0.079493 | 7.780096e-01 |
| 7 | AlcoholConsumption | 0.078861 | 7.788718e-01 |
| 0 | Age | 0.009220 | 9.235112e-01 |
| 36 | NSAIDsUse | 0.002247 | 9.621935e-01 |

Hasil seleksi menunjukan bahwa dari 54 atribut awal, hanya sebagian fitur yang terpilih sebagai fitur terbaik untuk setiap skenario balancing data. Beberapa fitur yang konsisten muncul antara lain hemogoblin, blood pressure, blood glucose dan serumcreatine. Fitur fitur tersebut memiliki relevansi klinis karena berhubungan dengan kondisi pasien

Dengan demikian, tahapan seleksi fitur memastikan bahwa dataset yang telah diproses, diseimbangkan dan direduksi fiturnya siap digunakan pada tahap pemodelan.

*B. Evaluasi Model*

Evaluasi model dilakukan untuk menilai kinerja algoritma klasifikasi pada ketiga skenario balancing data. Penilaian menggunakan metrik akurasi, precision, recall, spensifisitas dan F1 -score, dengan fokus utama pada kemampuan model dalam mengenali kelas positif (penderita CKD). Selain itu, confusion matrix digunakan untuk memvisualisasikan distribusi prediksi benar maupun salah pada setiap kelas.

Hasil evaluasi pada data uji terhadap seluruh model dan konfigurasi yang diuji disajikan dalam Tabel-tabel berikut:

- Model 1 (Original)

Pada skenario Model Original, dataset digunakan tanpa dilakukan penyeimbangan kelas sehingga distribusi antara kelas mayoritas dan minoritas tetap tidak seimbang. Kondisi ini merepresentasikan situasi nyata pada data medis, dimana jumlah pasien sehat biasanya lebih banyak dibandingkan pasien menderita penyakit.

Evaluasi pada skenario ini bertujuan untuk melihat bagaimana peforma masing masing algoritma klasifikasi KNN, SVM dan Random Forest setelah melalui tahap seleksi fitur dengan SelectKBest, model original ini dijadikan tolak ukur awal yang kemudian akan dibandingkan.

| Model | Parameter | Precision (0/1) | Recall (0/ | F1-score | acc |
|---|---|---|---|---|---|
| KNN | K=3 | 0.50/ 0.92 | 0.07/ 0.09 | 0.12/ 0.96 | 0.92 |
| | K=5 | 0.33/ 0.92 | 0.07 /0.99 | 0.12/ 0.95 | 0.91 |
| | K=7 | 0.50/ 0.92 | 0.07 /0.99 | 0.12/ 0.96 | 0.92 |
| | K=9 | 0.00/ 0.92 | 0.00/ 1.00 | 0.00/ 0.96 | 0.92 |
| | K=11 | 0.00/ 0.92 | 0.00/ 1.00 | 0.00/ 0.96 | 0.92 |
| RF | N=100 | 0.00/ 0.92 | 0.00/ 1.00 | 0.00/ 0.96 | 0.92 |
| | N=150 | 0.00/ 0.92 | 0.00/ 1.00 | 0.00/ 0.96 | 0.92 |
| SVM | rbf | 0.00/ 0.92 | 0.00/ 1.00 | 0.00/ 0.96 | 0.92 |
| | Rbf, c=10 | 0.29/ 0.92 | 0.14/ 0.97 | 0.19/ 0.95 | 0.90 |
| | Rbf, c=10,gamma=auto | 0.00/ 0.92 | 0.00/ 1.00 | 0.00/ 0.96 | 0.92 |

Tabel 1. Hasil Evaluasi Model 1

Berdasarkan hasil pengujian pada skenario Model Original, selurh algoritma klasifikasi KNN,SVM dan Random Forest masih menunjukan nilai akurasi yang relatif tinggi, yaitu berada pada kisaran 90% hingga 93%. Namun, ketika dianalisis lebih dalam, terlihat bahwa performa model dalam mengenali kelas minoritas (penderita CKD) masih sangat rendah.

Dengan demikian, meskipun Model Original memberikan hasil akurasi yang baik secara agregat, prediksi terhadap kelas minoritas (penderita CKD) tidak dapat diandalkan. Hal ini memperkuat alasan bahwa strategi penyeimbangan kelas seperti Undersampling dan SMOTE perlu diterapkan, agar model dapat menghasilkan yang lebih adil dan representatif pada kedua kelas.

- Model 2 (Random Undersampling)

Pada skenario Model 2 (Random Undersampling), penanganan ketidakseimbangan kelas dilakukan dengan cara mengurangi jumlah sampel pada kelas mayoritas (pasien tidak CKD) secara acak hingga jumlahnya seimbang dengan kelas minoritas (penderita (CKD). Proses ini menghasilkan dataset berukuran lebih kecil, yakni 248 sampel dengan distribusi masing-masing kelas sebanyak 124. Selanjutnya, data dibagi menjadi set latih, validasi, dan uji dengan proporsi 80:10:10 menggunakan stratified sampling untuk menjaga keseimbangan antar kelas. Strategi undersampling ini diharapkan mampu mengurangi bias model terhadap kelas mayoritas, meskipun dengan konsekuensi hilangnya sebagian informasi dari data asli. Evaluasi performa algoritma KNN, SVM, dan RF pada skenario ini disajikan dalam tabel berikut.

| Model | Parameter | Precision (0/1) | Recall (0/ | F1-score | acc |
|---|---|---|---|---|---|
| KNN | K=3 | 0.47/ 0.42 | 0.50/ 0.38 | 0.48/ 0.40 | 0.44 |
| | K=5 | 0.53/ 0.50 | 0.57 /0.46 | 0.55/ 0.48 | 0.52 |
| | K=7 | 0.47/ 0.42 | 0.50 /0.38 | 0.48/ 0.40 | 0.44 |
| | K=9 | 0.62/ 0.64 | 0.71/ 0.54 | 0.67/ 0.58 | 0.63 |
| | K=11 | 0.60/ 0.58 | 0.64/ 0.54 | 0.62/ 0.56 | 0.59 |
| RF | N=100 | 0.60/ 0.58 | 0.64/ 0.54 | 0.62/ 0.56 | 0.59 |
| | N=150 | 0.64/ 0.62 | 0.64/ 0.62 | 0.64/ 0.62 | 0.63 |
| SVM | rbf | 0.67/ 0.67 | 0.71/ 0.62 | 0.69/ 0.64 | 0.67 |
| | Rbf, c=10 | 0.56/ 0.55 | 0.64/ 0.46 | 0.60/ 0.50 | 0.56 |
| | Rbf, c=10,gamma=auto | 0.67/ 0.67 | 0.71/ 0.62 | 0.69/ 0.64 | 0.67 |

Tabel 1. Hasil Evaluasi Model 2

Berdasarkan Tabel di atas, performa model pada skenario Random Undersampling mengalami penurunan signifikan dibandingkan Model Original. Nilai akurasi tertinggi hanya

mencapai sekitar 60% pada KNN dengan k=9, sementara konfigurasi lain umumnya berada pada kisaran 48–60%. Random Forest dan SVM juga memperlihatkan pola serupa, dengan akurasi tidak stabil dan metrik precision serta recall pada kelas positif (penderita CKD) yang masih rendah.

Kondisi ini terjadi karena undersampling mengurangi jumlah data mayoritas secara drastis, sehingga informasi yang terkandung dalam dataset menjadi terbatas. Dengan dataset yang lebih kecil, kemampuan model dalam mempelajari pola menjadi berkurang yang berdampak pada menurunya kemampuan generalisasi..

Dengan demikian, Model 2 belum mampu memberikan performa yang optimal, sehingga diperlukan pendekatan alternatif seperti SMOTE yang dapat menjaga keseimbangan kelas tanpa mengorbankan jumlah data.

- Model 3 (SMOTE)

Pada skenario ketiga, penanganan ketidakseimbangan kelas dilakukan menggunakan Synthetic Minority Over-sampling Technique (SMOTE) yang diterapkan langsung pada dataset asli sebelum dilakukan proses pembagian data. Sehingga distribusi kelas menjadi lebih seimbang antara pasien CKD (1) dan non CKD (0).

Setelah dilakukan balancing dengan SMOTE, dataset. Kemudian dibagi menjadi 80% data latih, 10% data validasi, dan 10% data ui, dengan menggunakan *stratified split* agar proporsi kelas tetap terjaga disetiap subset data. [18]

Evaluasi performa algoritma KNN, SVM dan Random Forest pada skenario SMOTE disajikan dalam tabel berikut :

| Model | Parameter | Precision (0/1) | Recall (0/ | F1-score | acc |
|---|---|---|---|---|---|
| KNN | K=3 | 0.81/ 1.00 | 1.00/ 0.76 | 0.89/ 0.87 | 0.88 |
| | K=5 | 0.78/ 0.98 | 0.99 /0.83 | 0.87/ 0.83 | 0.85 |
| | K=7 | 0.81/ 1.00 | 1.00 /0.76 | 0.89/ 0.87 | 0.88 |
| | K=9 | 0.77/ 0.99 | 0.99/ 0.71 | 0.87/ 0.82 | 0.85 |
| | K=11 | 0.77/ 0.98 | 0.99/ 0.70 | 0.87/ 0.82 | 0.85 |
| RF | N=100 | 0.93/ 0.97 | 0.97/ 0.92 | 0.95/ 0.94 | 0.94 |
| | N=150 | 0.94/ 0.97 | 0.97/ 0.93 | 0.96/ 0.95 | 0.95 |
| SVM | rbf | 0.90/ 0.94 | 0.95/ 0.89 | 0.92/ 0.92 | 0.92 |
| | Rbf, c=10 | 0.90/ 0.99 | 0.99/ 0.89 | 0.95/ 0.94 | 0.94 |
| | Rbf, c=10,gamma=auto | 0.89/ 0.92 | 0.92/ 0.88 | 0.90/ 0.90 | 0.90 |

Tabel 3. Hasil Evaluasi Model 3

Berdasarkan Tabel di atas, penerapan SMOTE pada dataset asli memberikan dampak positif yang signifikan terhadap performa model. Pada algoritma KNN, meskiupun

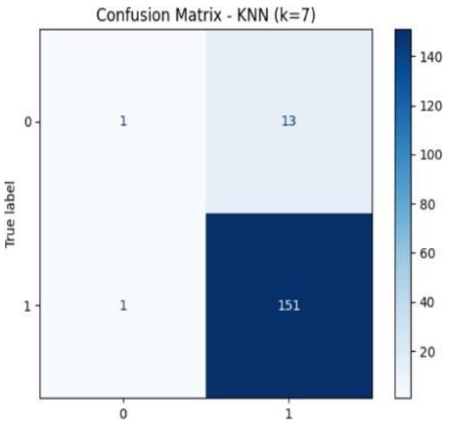
akurasi berada pada kisaran 77-81%, nilai precision dan recall untuk kelas minoritas menunjukan perbaikan yang jelas dibandingkan dengan Model Original maupun Undersampling, algoritma SVM juga menunjukan peningkatan yang stabil setelah diterapkanya SMOTE, dengan akurasi kisaran 68%-88%. Nilai precision, recall dan F11 score juga turun mengalami perbaikan dibanding sebelumnya.

Peningkatan paling mencolok terjadi pada Random Forest, yang berhasil mencapai akurasi tinggi dikisaran 95-96% dengan precision dan recall yang relatif seimbang. Hasil ini menunjukan bahwa Random Forest mampu memanfaatkan fitur seleksi secara konsisten[19]

Dengan demikian, Model 3 (SMOTE) dapat disimpulkan sebagai skenario terbaik dalam penelitian ini karena berhasil mengatasi bias terhadap kelas mayoritas sekaligus meningkatkan sensitivitas model dalam mengenali kelas minoritas (penderita CKD) [20]

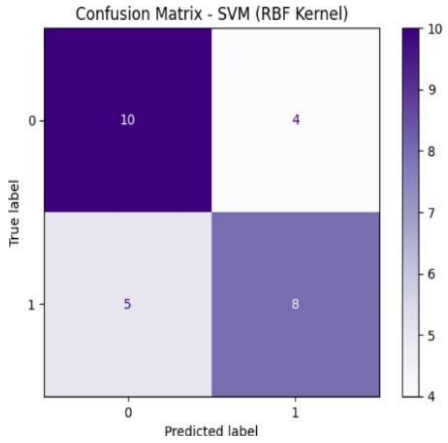
Untuk melengkapi hasil evaluasi yang telah disajikan dalam bentuk tabel, pada penelitian ini juga ditampilkan confusion matrix dari algoritma terbaik pada masing-masing skenario (Model 1, Model 2, dan Model 3). Visualisasi confusion matrix ini digunakan untuk menunjukkan secara lebih rinci distribusi prediksi benar dan salah pada kelas positif (penderita CKD) maupun kelas negatif (non CKD), sehingga dapat memperkuat analisis performa model yang diperoleh.

Gambar 3. *Confusion Matrix KNN pada Model 1*




Confusion matrix pada Model 1 (original tanpa penerapan undersampling maupun SMOTE) algoritma KNN dengan nilai K= 7 menunjukkan algoritma ini memperoleh nilai akurasi sebesar 92%. Namun, akurasi tinggi tersebut tidak diiukti dengan kemampuan yang baik dalam mengenali kedua kelas secara seimbang.

Kondisi ini merupakan bentuk accuracy paradox, dimana akurasi tampak tinggi namun tidak representatif untuk menggambarkan performa model secara keseluruhan. Oleh karena itu, diperlukan penyeimbangan data seperti Undersampling atau SMOTE, agar model lebih optimal dalam mengenali pasien CKD.

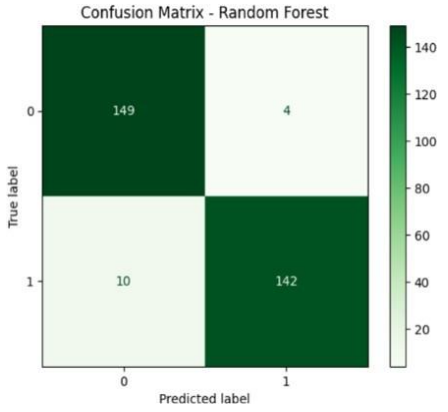



Gambar 4. *Confusion Matrix SVM (c =10 gamma auto) pada model 2*

Confusion matrix pada skenario Model 2 hanya menggunakan teknik Undersampling, dengan algoritma Support Vector Machine (SVM). Performa terbaik diperoleh pada konfigurasi C=10 gamma auto, dengan akurasi sekitar

67%, meskipun nilai akurasi tidak terlalu tinggi, SVM relatif lebih stabil dibandingkan algoritma lain pada skenario ini karena mampu menjaga keseimbangan antara precision dan recall pada kedua kelas.

Kondisi ini menunjukan bahwa teknik Undersampling memang dapat membantu mengurangi ketidakseimbangan data, namuin konsekuensinya adalah berkurangnya jumlah sampel pada kelas mayoritas,, dampaknya informasi untuk melatih model menjadi terbatas.




Gambar 5. *Confusion Matrix RF (n = 150, )pada model 3*

Berdasarkan confusion matrix pada Gambar, model Random Forest dengan parameter n=150 merupakan model terbaik dibandingkan dengan variasi parameter maupun algoritma lainnya. Hal ini terlihat dari hasil klasifikasi yang hampir sempurna, dengan akurasi yang mencapai 95%, precision dan recall (non CKD) – precision 0,94, recall 0,97, 1 (CKD).

Dibandingkan dengan model KNN maupun SVM yang masih menghasilkan kesalahan klasifikasi lebih besar, model ini mampu mencapai akurasi 95%, sensitivitas yang sangat tinggi, serta keseimbangan precision dan recall yang unggul. Dengan demikian, Random Forest dapat dianggap sebagai model paling optimal untuk klasifikasi penyakittlbo CKD dalam penelitian ini.

## V. KESIMPULAN

Penelitian klasifikasi Chronic Kidney Disease (CKD), membangun sistem dengan memanfaatkan tiga algoritma Machine Learning, yaitu K-Nearest Neighbor (KNN), Support vector Machine (SVM) dan Random Forest (RF). Tahapan pra pemrosesan yang dilakukan mencakup pembersihan data, normalisasi dengan Min-Max Scaling, serta penanganan ketidakseimbangan data menggunakan dua pendekatan, yaitu Random Undersampling dan Synthetic Minority Over-Sampling Technique (SMOTE).

Berdasarkan hasil eksperimen dengan tiga scenario model:

- Model 1 (original tanpa balancing) menghasillkan akurasi yang relatif tinggi, namun model gagal mengenali kelas minoritas (penderita CKD).
- Model 2 (undersampling) mampu menyeimbangkan distribusi kelas, tetapi karena jumlah data mayoritas berkurang drastis, performa model menurun dengan akurasi hanya sekitar 60%.
- Model 3 (SMOTE) terbukti sebagai skenario model terbaik karena meningkatkan representasi kelas minoritas tanpa mnegurangi data mayoritas. Pada skenario ini, algoritma Random Forest

(n=150) mencapai performa paling optimal dengan akurasi 95%, Algoritma KNN dan SVM juga menunjukan perbaikan dibandingkan skenario seblumnya namun masih dibawa hasil RF.

Secara keseluruhan, hasil penelitian ini menunjukkan bahwa kombinasi metode balancing SMOTE dan algoritma Random Forest n=150 memberikan hasil Klasifikasi terbaik dalam mendeteksi penyakit Ginjal Kronis. Hasil ini menegaskan bahwa pendekatan RF lebih optimal dan mampu menjaga keseimbangan performa antar kelas, sehingga dapat diandalkan untuk mendukung sistem prediksi berbasis data medis.

REFERENCES

[1] A. C. Webster, E. V. Nagler, R. L. Morton, and

P. Masson, “Chronic Kidney Disease,” Mar. 25, 2020, Lancet Publishing Group. doi: 10.1016/S0140-6736(16)32064-5.

[2] A. Sarce Joel, F. Abdussalaam, and Y. Yunengsih, “TATA KELOLA REKAM MEDIS BERBASIS TEKNOLOGI INFORMASI DALAM PENANGANAN KERAHASIAAN DAN KEAMANAN DATA PASIEN DENGAN

METODE KRIPTOGRAFI,” Jurnal Indonesia : Manajemen Informatika dan Komunikasi, vol. 4, no. 3, pp. 837–848, Sep. 2023, doi: 10.35870/jimik.v4i3.287.

[3] D. Bidang Kesehatan Fangatulo Dodo Telaumbanua, P. Hulu, T. Zulfiter Nadeak, R. Romeo Lumbantong, and A. Dharma, “Penggunaan Machine Learning,” Oct. 2020.

[4] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng, “Learning k for kNN Classification,” ACM Trans Intell Syst Technol, vol. 8, no. 3, Jan. 2020, doi: 10.1145/2990508.

[5] S. Y. I. S. Imaniar Ikko Mulya Rizky1a*, “Perbandingan Kinerja Algoritma Naive Bayes, Support Vector Machine dan Random forest untuk Prediksi Penyakit Ginjal Kronis Imaniar Ikko Mulya Rizky 1a* , Suhendro Yusuf Irianto 2b. Sriyanto 3c,” Aug. 2023.

[6] Dr. P. R. D. Prof. V. K. S. Miss. Mayuri S. Shelke1, “A Review on Imbalanced Data Handling Using Undersampling and Oversampling Technique,” International Journal of Recent Trends in Engineering and Research, vol. 3, no. 4, pp. 444–449, May 2021, doi: 10.23883/ijrter.2017.3168.0uwxm.

[7] Y. El Touati, J. Ben Slimane, and T. Saidani, “Adaptive Method for Feature Selection in the Machine Learning Context,” Engineering, Technology and Applied Science Research, vol. 14, no. 3, pp. 14295–14300, Jun. 2024, doi:

10.48084/etasr.7401.

[8] R. Blagus and L. Lusa, “SMOTE for high-dimensional class-imbalanced data,” 2020. [Online]. Available:

http://www.biomedcentral.com/1471- 2105/14/106

[9] A. Amato and V. Di Lecce, “Data preprocessing impact on machine learning algorithm performance,” Open Computer Science, vol. 13, no. 1, Jan. 2023, doi: 10.1515/comp-2022-0278.

[10] A. Fernández, S. García, F. Herrera, and N. V Chawla, “SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15- year Anniversary,” Apr. 2020.

[11] X. H. Cao, I. Stojkovic, and Z. Obradovic, “A robust data scaling algorithm to improve classification accuracies in biomedical data,” BMC Bioinformatics, vol. 17, no. 1, Sep. 2021,doi: 10.1186/s12859-016-1236-x.

[12] P. Palinggik Allorerung, A. Erna, M. Bagussahrir , and S. Alam, “Analisis Performa

Normalisasi Data untuk Klasifikasi K-Nearest Neighbor pada Dataset Penyakit," 2024.

[13] P. Budi Utomo, M. Faruqziddan, E. Herdika Septa Aulia, and S. Dini Azzahra, "Perbandingan Skenario Balancing Oversampling dan Undersampling dalam Klasifikasi Resiko Kambuh Kanker Tiroid menggunakan Algoritma SVM Linear," JAMI: Jurnal Ahli Muda Indonesia, vol. 5, no. 2, pp. 172–182, Dec. 2024, doi: 10.46510/jami.v5i2.320.

[14] M. Naseriparsa, A. Al-Shammari, M. Sheng,
Y. Zhang, and R. Zhou, "RSMOTE: improving classification performance over imbalanced medical datasets," Health Inf Sci Syst, vol. 8, no. 1, Dec. 2020, doi: 10.1007/s13755-020-00112-w.

[15] Narayanan and Jayashree, "Implementation of Efficient Machine Learning Techniques for Prediction of Cardiac Disease using SMOTE," in Procedia Computer Science, Elsevier B.V., 2024,

pp. 558–569. doi: 10.1016/j.procs.2024.03.245.

[16] I. Akbar, F. Supriadi, and D. I. Junaedi, "PEMANFAATAN MACHINE LEARNING DI BIDANG KESEHATAN," 2025.

[17] M. A. Al-Hashem, A. M. Alqudah, and Q. Qananwah, "Performance Evaluation of Different Machine Learning Classification Algorithms for Disease Diagnosis," International Journal of E-Health and Medical Communications, vol. 12, no. 6, 2021, doi: 10.4018/IJEHMC.20211101.oa5.

[18] E. Yunianto and S. Widya Pratama, "DIAGNOSTIK PENYAKIT GINJAL KRONIS MENGGUNAKAN MODEL KLASIFIKASI SUPPORT VECTOR MACHINE," vol. XIX, no.

1, 2024, [Online].
Available: http://ejournal.stmik-wp.ac.id

[19] L. Tugas et al., "Klasifikasi Penyakit Ginjal Kronis Menggunkan Metode Random Forest Dengan Hyperparameter Tuning," 2017.

[20] D. H. Jeong, S. E. Kim, W. H. Choi, and S. H. Ahn, "A Comparative Study on the Influence of Undersampling and Oversampling Techniques for the Classification of Physical Activities Using an Imbalanced Accelerometer Dataset," Healthcare (Switzerland), vol. 10, no. 7, Jul. 2022, doi: 10.3390/healthcare10071255.