**Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)** DOI: 10.25126/jtiik.202071896
Vol. 7, No. 2, April 2020, hlm. 417-426 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 30/E/KPT/2018 e-ISSN: 2528-6579

# IMPLEMENTASI DATA MINING UNTUK DETEKSI PENYAKIT GINJAL KRONIS (PGK) MENGGUNAKAN K-NEAREST NEIGHBOR (KNN) DENGAN BACKWARD ELIMINATION

**Ikhsan Wisnuadji Gamadarenda[1], Indra Waspada*[2]**

[12]Universitas Diponegoro
Email: [1]ikhsangama@asia.com, [2]indrawaspada@undip.ac.id
*Penulis Korespondensi

(Naskah masuk: 21 Maret 2019, diterima untuk diterbitkan: 11 Februari 2020)

**Abstrak**

Penyakit ginjal kronis (PGK) merupakan masalah kesehatan publik di seluruh dunia dengan insiden yang terus meningkat. Berdasarkan sumber dari BPJS Kesehatan, perawatan PGK merupakan ranking kedua pembiayaan terbesar setelah penyakit jantung. Pendeteksian PGK juga memerlukan banyak atribut sehingga membutuhkan biaya yang cukup mahal. Oleh sebab itu dibuat sistem dengan tahapan data mining berbasis web yang memudahkan untuk melakukan deteksi PGK, sehingga PGK dapat dicegah, ditanggulangi, dan kemungkinan mendapatkan terapi yang efektif lebih besar jika diketahui lebih awal. Proses penelitian ini menggunakan sebuah rangka kerja *data mining Knowledge Data Discover*y (KDD). Dalam skenario rangka kerja yang digunakan, sistem ini menggunakan Algoritme *Backward Elimination* untuk mengurangi jumlah atribut yang dipakai dengan tujuan untuk mengurangi jenis pemeriksaan yang dilakukan, dan Algoritme k-*Nearest Neighbor* sebagai algoritme klasifikasi untuk mendeteksi penyakit. Hasil pemodelan terbaik *data mining* dari sistem yang dibuat menggunakan *Backward Elimination* ($\alpha$ = 0,05) dan kNN ($k$ = 3) dengan pertimbangan penurunan biaya pemeriksaan dan sensitivity tertinggi. Rekomendasi sistem menghasilkan 10 atribut yang terpilih dari 24 atribut awal yang digunakan, yaitu: berat jenis (*sg*), albumin (*al*), urea darah (*bu*), kreatinin serum (*sc*), sodium (*sod*), hemoglobin (*hemo*), sel darah merah (*rbc*), hipertensi (*htn*), diabetes mellitus (*dm*), dan nafsu makan (*appet*). Penggunaan atribut yang telah terseleksi tersebut, berhasil menekan biaya pemeriksaan hingga 73,36%. Selanjutnya dilakukan pendeteksian penyakit menggunakan Algoritme k-*Nearest Neighbor* menghasilkan nilai akurasi sebesar 99,25%, *sensitivity* sebesar 99,5%, dan *specificity* sebesar 98,745%.

**Kata kunci**: Deteksi Penyakit Ginjal Kronis, Algoritme k-*Nearest Neighbor*, Seleksi Atribut, Algoritme Backward Elimination.

# *DATA MINING IMPLEMENTATION FOR DETECTION OF CHRONIC KIDNEY (CKD) USING K-NEAREST NEIGHBOR (KNN) WITH BACKWARD ELIMINATION*

***Abstract***

*Chronic kidney disease (CKD) is a health problem for people around the world with increasing incidence. Based on sources from BPJS Kesehatan, CKD care is the second largest ranking of financing after heart disease. CKD detection also requires many attributes, so it requires quite expensive costs. Create a system with web-based data mining stages that makes it easy to detect CKD. Allowing CKD to be prevented, addressed, and advised to get effective therapy is greater if acknowledged earlier. The process of this research uses work methods of Data Mining Knowledge Data Discovery (KDD). In the framework of the framework used, this system uses the Backward Elimination Algorithm to reduce the number of attributes used to reduce the type of inspection performed, and the k-Nearest Neighbor Algorithm as an algorithm to update disease. The best data mining modeling results from the system are made using Backward Elimination ($\alpha$ = 0.05) and kNN ($k$ = 3) by calculating the increase in inspection costs and the highest sensitivity. System recommendations produce 10 attributes selected from the 24 initial attributes used, namely: specific gravity (sg), albumin (al), blood urea (bu), serum creatinine (sc), sodium (soil), hemoglobin (hemo), cell red blood (rbc), hypertension (htn), diabetes mellitus (dm), and appetite (appetite). The use of the selected attributes succeeded in achieving inspection costs of up to 73.36%. Furthermore, disease detection using the k-Nearest Neighbor Algorithm produces an accuracy value of 99.25%, sensitivity of 99.5%, and specificity of 98.745%.*

**Keywords**: *Detection of Chronic Kidney Disease, k-Nearest Neighbor Algorithm, Attribute Selection, Backward Elimination Algorithm*

## 1. PENDAHULUAN

Ginjal merupakan organ penting yang berfungsi menjaga komposisi darah dengan mencegah menumpuknya limbah dan mengendalikan keseimbangan cairan dalam tubuh. Penyakit ginjal adalah kelainan yang mengenai organ ginjal yang timbul akibat berbagai faktor, misalnya infeksi, tumor, kelainan bawaan, penyakit metabolik atau degeneratif, dan lain-lain. Seseorang didefinisikan sebagai Penyakit Ginjal Kronis (PGK) jika pernah didiagnosis menderita penyakit gagal ginjal kronis (minimal sakit selama 3 bulan berturut-turut) oleh dokter (Badan Penelitian dan Pengembangan Kesehatan, 2013). Penyakit tersebut pada awalnya tidak menunjukkan tanda dan gejala namun dapat berjalan progresif menjadi gagal ginjal (Kemenkes, 2017).

Penyakit gagal ginjal bisa dicegah, ditanggulangi, dan kemungkinan mendapatkan terapi yang efektif akan lebih besar jika diketahui lebih awal. Ketika PGK lambat terdeteksi maka memerlukan biaya yang lebih besar dalam pengobatannya serta membutuhkan tenaga medis yang lebih ahli dalam penanganannya dengan peluang penyembuhan yang semakin kecil (Jing et al., 2012). Perawatan PGK merupakan ranking kedua pembiayaan terbesar dari BPJS kesehatan setelah penyakit jantung (Kemenkes, 2017).

Menurut PERMENKES No: 269/MENKES/PER/III/2008 yang dimaksud rekam medis adalah berkas yang berisi catatan dan dokumen antara lain identitas pasien, hasil pemeriksaan, pengobatan yang telah diberikan, serta tindakan dan pelayanan lain yang telah diberikan kepada pasien. Melalui rekam medis ini dapat dilakukan proses *data mining*. *Data mining* adalah proses ekstraksi pengetahuan tertentu, dengan algoritme untuk mendeteksi pola spesifik, kecenderungan dalam data, dan aturan mekanis yaitu asosiasi antara data yang sebelumnya tidak terlihat berhubungan, sehingga mendapatkan pengetahuan baru yang menarik dan belum diketahui sebelumnya (Borges, Marques dan Bernardino, 2013).

Dalam penelitian ini didapatkan data yang digunakan untuk melakukan deteksi PGK menggunakan cukup banyak atribut sehingga membutuhkan biaya yang juga cukup mahal, sehingga dibuat tujuan yang ingin dicapai dalam penelitian ini adalah untuk membuat sebuah sistem berbasis web dari hasil suatu *framework data mining* yang memudahkan dalam melakukan deteksi PGK dengan mengurangi jumlah atribut yang harus dimasukkan, sehingga dapat menekan biaya tes laboratorium. Pasien yang terdiagnosis juga dapat dilakukan tindak lanjut secara cepat dan tepat untuk menanggulangi tingkat kerusakan dan mengurangi biaya pengobatan yang lebih mahal.

Teknik k-*Nearest Neighbour* atau kNN merupakan model klasifikasi yang memiliki beberapa kelebihan, penerapannya yang sederhana namun efektif dalam banyak kasus (Sinha, 2015). Data training pada kNN sangat cepat dan kuat meski pada noise data. kNN juga memiliki performa yang baik pada sistem dimana sebuah sample memiliki banyak label class (Jadhav dan Channe, 2013). Salah satu masalah dari algoritme kNN adalah semua atribut dalam record harus dihitung jaraknya satu sama lain. Dengan kata lain atribut pada record baru akan dihitung jaraknya dengan atribut pada record yang tersedia pada dataset training. Pada kenyataannya tidak semua atribut mempunyai nilai (*missing value*), serta mempunyai nilai atribut dengan *range value* berbeda dengan atribut sejenis lainnya, sehingga dapat menyebabkan masalah pada perhitungan jaraknya. Adapula kombinasi – kombinasi atribut yang memiliki rule tertentu yang dapat memperkuat klasifikasi. Maka diperlukan penanganan *preprocessing* sebelum melakukan deteksi dengan algoritme kNN (Han dan Kamber, 2011).

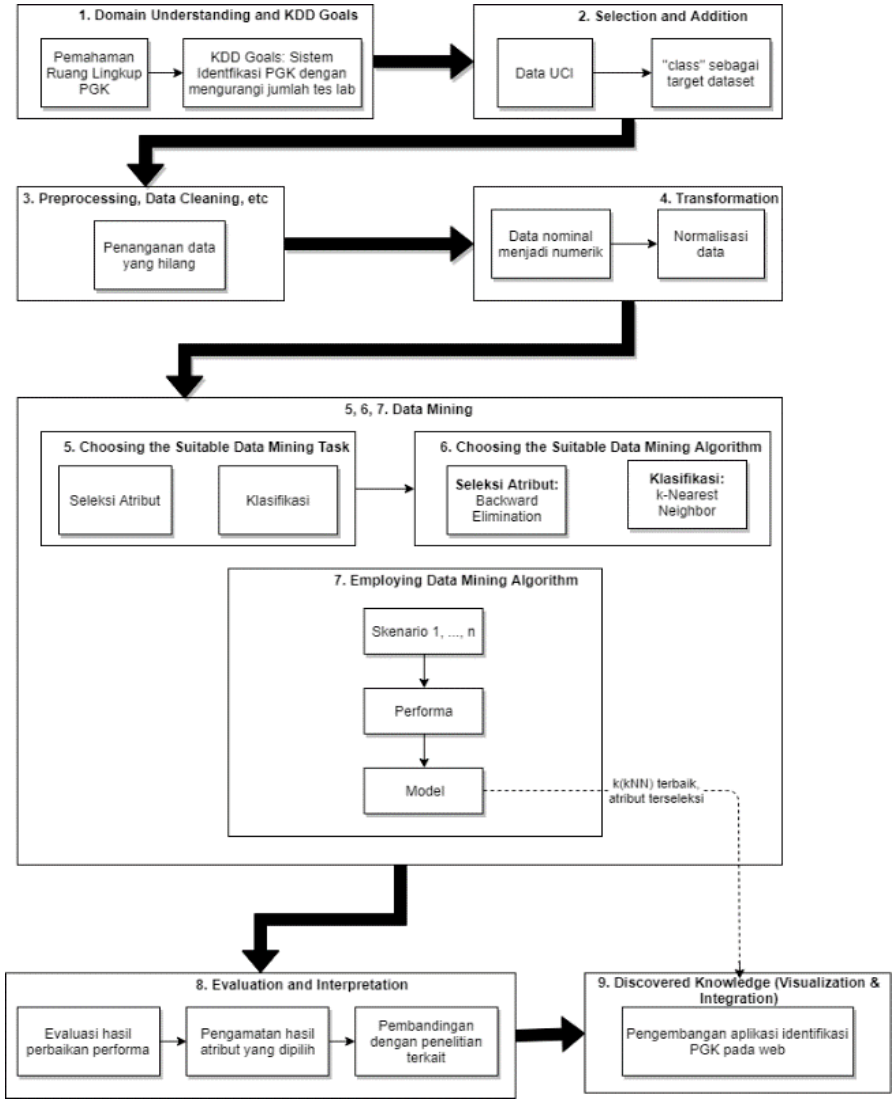
*Backward Elimination* merupakan salah satu tahap yang dapat dilakukan dalam melakukan seleksi atribut pada tahap *preprocessing*. Algoritme *Backward Elimination* dipilih karena kemampuan algoritme ini yang memungkinkan untuk mendapatkan beberapa atribut yang awalnya memiliki kemampuan klasifikasi rendah secara individu namun jika digabungkan dengan atribut lainnya akan memiliki akurasi yang tinggi (Gerard, 2012).

Dari permasalahan tersebut, solusi yang dapat dilakukan adalah mengimplementasikan Algoritme k-*Nearest Neighbor* untuk mendiagnosis Penyakit Ginjal Kronis (PGK) dengan menambahkan Algoritme Backward Elimination pada tahap preprocessing sehingga lebih optimal dan dapat menekan biaya pemeriksaan laboratorium.

Dalam pelaksanaan penelitian, menggunakan data set deteksi ginjal kronis diambil dari Universitas Alagappa (https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease) yaitu sejumlah 400 data dengan 25 atribut, dan 2 kelas pada atribut target. Dan hasil sistem dibangun menggunakan pahasa pemrograman Python, web *framework* Flask, dan CSS *framework* Materialize.

## 2. METODE PENELITIAN

Penelitian ini dilakukan dengan 9 tahapan yang telah disesuaikan dengan *framework data mining* KDD (*Knowledge Data Discovery*) (Shafique dan Qaiser, 2014).




Gambar 1 Rangka Kerja Penelitian

### 2.1. *Domain Understanding and KDD Goals*

Pada tahapan ini, penelitian melakukan studi pustaka mengenai hal – hal yang terkait permasalahan Penyakit Ginjal Kronis (PGK), hasil studi pustaka yang dilakukan dirangkum menjadi 5 poin utama yaitu: (1) Ginjal memiliki berbagai peranan penting dalam menjaga kesehatan tubuh manusia; (2) Penyakit Ginjal Kronis (PGK) merupakan masalah kesehatan publik di seluruh dunia. (3) PGK awalnya tidak menunjukkan gejala dan berjalan progresif semakin parah. (4) Pemeriksaan dan perawatan PGK yang mahal. (5) Penderita yang terdiagnosis PGK akan mendapatkan terapi yang efektif jika terdeteksi secara dini.

Tujuan dilakukannya rangka kerja *data mining* KDD ini adalah membuat sistem berbasis web yang dapat digunakan oleh masyarakat umum dan penyedia layanan kesehatan untuk melakukan deteksi PGK dengan penambahan fitur seleksi atribut untuk menekan biaya tes laboratorium.

### 2.2. *Selection and Addition*

Dalam proses selection, penelitian ini menggunakan dataset yang didapatkan dari UCI Machine Learning Repository. Dataset tersebut memiliki 400 responden, dengan masing – masing 250 data terdiagnosis PGK dan 150 data tidak terdiagnosis PGK. Data tersebut memiliki 25 kolom, yang terdiri dari 1 kelas target dan 24 atribut seperti pada tabel 3.3-1. Target pada atribut "class", telah dikelompokkan menjadi 2 output, PGK (ckd) atau Normal (notckd). Seluruh atribut ini dapat digunakan dalam deteksi PGK dan dapat dikurangi dengan proses feature selection yang akan dilakukan pada proses selanjutnya pada tahap ke-5 hingga ke-7 pada skenario *Data Mining*.

Tabel 1. Data Atribut dan Tipe Data

| No | Atribut | Keterangan | | Tipe Data |
|---|---|---|---|---|
| | | Inggris | Indonesia | |
| 1 | age | Age | Umur | Numerik (years) |
| 2 | bp | Blood Pressure | Tekanan Darah | Numerik (mm/hg) |
| 3 | sg | Specific Grafity | Berat jenis urin hasil tes urinalysis | Nominal (1.005, 1.010, 1.015, 1.020, 1.025) |
| 4 | al | Albumin | Albumin, protein hasil sintesis hati | Nominal (0, 1, 2, 3, 4, 5) |
| 5 | su | Sugar | Gula darah dalam tubuh | Nominal (0, 1, 2, 3, 4, 5) |
| 6 | rbc | Red Blood Cells | Bentuk sel darah merah | Nominal (normal, abnormal) |
| 7 | pc | Pus Cell | Bentuk sel darah putih | Nominal (normal, abnormal) |
| 8 | pcc | Pus Cell Clumps | Gumpalan sel nanah pada urin | Nominal (present, notpresent) |
| 9 | ba | Bacteria | Bakteri pada urin | Nominal (present, notpresent) |
| 10 | bgr | Blood Glucose Random | Gula darah setelah makan | Numerik (mgs/dl) |
| 11 | bu | Blood Urea | Kadar urea pada darah | Numerik (mgs/dl) |
| 12 | sc | Serum Creatinine | Kadar kreatinin serum pada darah | Numerik (mgs/dl) |
| 13 | sod | Sodium | Kadar sodium pada darah | Numerik (mEq/L) |
| 14 | pot | Potassium | Kadar potassium pada darah | Numerik (mEq/L) |
| 15 | hemo | Hemoglobin | Kadar hemoglobin pada darah | Numerik (gms) |
| 16 | pcv | Packed Cell Volume / Hematocrit | Jumlah hematocrit pada darah | Numerik (mEq/L) |
| 17 | wbcc | White Blood Cell Count | Jumlah sel darah putih | Numerik (cells/cumm) |
| 18 | rbcc | Red Blood Cell Count | Jumlah sel darah merah | Numerik (millions/cmm) |
| 19 | htn | Hypertension | Memiliki riwayat hipertensi | Nominal (yes, no) |
| 20 | dm | Diabetes Mellitus | Memiliki riwayat penyakit diabetes mellitus | Nominal (yes, no) |
| 21 | cad | Coronary Artery Disease | Memiliki riwayat penyakit jantung coroner | Nominal (yes, no) |
| 22 | appet | Appetite | Memiliki selera makan | Nominal (good, poor) |
| 23 | pe | Pedal Edema | Pembengkakan pada kaki | Nominal (yes, no) |
| 24 | Ane | Anemia | Memiliki riwayat anemia | Nominal (yes, no) |
| 25 | class | Class | Kelas (Variabel Terikat) | Nominal (ckd, notckd) |

### 2.3. *Pre-processing, Data Cleaning, etc*

Kegiatan yang ada pada tahap ini antara lain melakukan pemindahan data yang didapat kedalam database, membersihkan dan memperbaiki data yang rusak, melengkapi nilai yang hilang, menyeragamkan data sehingga menjadi konsisten. Penanganan missing value awal adalah saat mengambil data pada database adalah mengubah beberapa nilai atribut yang kosong dengan nilai "NaN" pada nilai. Contoh langkah hasil pengambilan data ini akan ditunjukkan 5 id awal dan 5 id akhir. Hasil awal saat diambil dari database ditunjukkan pada gambar 2.

| | id | age | bp | sg | al | su | bgr | bu | sc | sod | ... | pc | pcc | ba | htn | dm | cad | appet | pe | ane | class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 48.0 | 80.0 | 1.020 | 1.0 | 0.0 | 121.0 | 36.0 | 1.2 | NaN | ... | normal | notpresent | notpresent | yes | yes | no | good | no | no | ckd |
| 1 | 2 | 7.0 | 50.0 | 1.020 | 4.0 | 0.0 | NaN | 18.0 | 0.8 | NaN | ... | normal | notpresent | notpresent | no | no | no | good | no | no | ckd |
| 2 | 3 | 62.0 | 80.0 | 1.010 | 2.0 | 3.0 | 423.0 | 53.0 | 1.8 | NaN | ... | normal | notpresent | notpresent | no | yes | no | poor | no | yes | ckd |
| 3 | 4 | 48.0 | 70.0 | 1.005 | 4.0 | 0.0 | 117.0 | 56.0 | 3.8 | 111.0 | ... | abnormal | present | notpresent | yes | no | no | poor | yes | yes | ckd |
| 4 | 5 | 51.0 | 80.0 | 1.010 | 2.0 | 0.0 | 106.0 | 26.0 | 1.4 | NaN | ... | normal | notpresent | notpresent | no | no | no | good | no | no | ckd |

5 rows × 26 columns

| | id | age | bp | sg | al | su | bgr | bu | sc | sod | ... | pc | pcc | ba | htn | dm | cad | appet | pe | ane | class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 395 | 396 | 55.0 | 80.0 | 1.020 | 0.0 | 0.0 | 140.0 | 49.0 | 0.5 | 150.0 | ... | normal | notpresent | notpresent | no | no | no | good | no | no | notckd |
| 396 | 397 | 42.0 | 70.0 | 1.025 | 0.0 | 0.0 | 75.0 | 31.0 | 1.2 | 141.0 | ... | normal | notpresent | notpresent | no | no | no | good | no | no | notckd |
| 397 | 398 | 12.0 | 80.0 | 1.020 | 0.0 | 0.0 | 100.0 | 26.0 | 0.6 | 137.0 | ... | normal | notpresent | notpresent | no | no | no | good | no | no | notckd |
| 398 | 399 | 17.0 | 60.0 | 1.025 | 0.0 | 0.0 | 114.0 | 50.0 | 1.0 | 135.0 | ... | normal | notpresent | notpresent | no | no | no | good | no | no | notckd |
| 399 | 400 | 58.0 | 80.0 | 1.025 | 0.0 | 0.0 | 131.0 | 18.0 | 1.1 | 141.0 | ... | normal | notpresent | notpresent | no | no | no | good | no | no | notckd |

5 rows × 26 columns

Gambar 2. 5 ID Awal (atas) dan 5 ID Terakhir (bawah) Dataset

### 2.4. *Transformation*

Setelah data yang dipilih sudah diterapkan maka akan dilakukan tahapan untuk melakukan transformasi terhadap parameter tertentu. Transformasi akan dilakukan untuk memodifikasi sumber data ke format berbeda yang dapat diterima oleh proses *data mining* selanjutnya. Proses transformasi ini dilakukan jika diperlukan atau jika terdapat data yang dinilai perlu untuk dilakukan transformasi formatnya.

Dalam algoritme k-NN yang mengimplementasikan rumus *Euclidean Distance*, sehingga terdapat 3 langkah transformasi yang dilakukan:

1. Perubahan nilai dari beberapa atribut nominal dan kelas target yang bernilai binominal menjadi numerik (riil). Sistem perubahan nilai nominal menjadi numerik ini dikodekan secara dinamis dimana pada value nominal pada data id pertama menjadi 0, dan jika ada value yang berbeda pada data selanjutnya, maka akan diisi dengan nilai *increment*-nya, dalam konteks ini akan berubah menjadi nilai 1. Pengkodean tidak hanya dilakukan pada atribut, namun juga pada kelas target dikarenakan akan dilakukan regresi pada saat dilakukan *Backward Elimination*. Dapat dilihat hasil perubahan-perubahan nilai tersebut ditunjukkan pada tabel 3.3-2 seperti berikut:
2. Penanganan *missing value*, merupakan pengulangan pada bagian dari tahap pre-processing sehingga mengubah nilai yang hilang (NaN) menjadi suatu nilai agar dapat dilakukan perhitungan. Penelitian ini menggunakan imputasian yaitu nilai rata – rata (*mean*) sebagai pengganti pada data yang bernilai kosong.
3. Normalisasi data, langkah agar setiap atribut memiliki bobot yang sama. Proses normalisasi minimal – maksimal dengan batas bawah 0 dan batas atas 1 sehingga perhitungan jarak *Euclidean* dari algoritme kNN menjadi lebih akurat.

### 2.5. *Data Mining: Choosing the Suitable Data Mining Task*

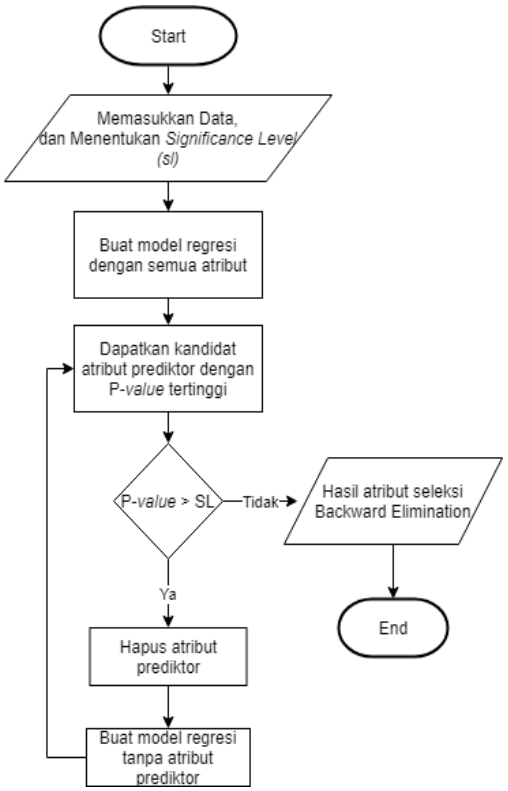
Jenis tugas *data mining* yang dipilih berdasarkan 2 tujuan utama penelitian ini, yaitu:

1. Seleksi atribut (*feature selection*), proses mengurangi atribut sehingga dapat untuk mengurangi biaya tes laboratorium.
2. Klasifikasi (*classification*) digunakan untuk deteksi penyakit ginjal kronis (PGK). Klasifikasi didefinisikan sebagai *supervised learning*, dimana telah terdapat informasi mengenai bagaimana data tersebut dikelompokkan dan tidak ada pertambahan kelompok.

### 2.6. *Data Mining: Choosing the Suitable Data Mining Algorithm*

Terdapat 2 algoritme yang akan digunakan untuk implementasi berdasarkan 2 tugas yang telah di jelaskan sebelumnya, yaitu:

1. ***Backward Elimination***, berfungsi sebagai seleksi atribut dimana memanfaatkan regresi statistik untuk mengetahui kedekatan setiap kombinasi atribut dengan target. Semakin kecil *significance level*, maka semakin ketat pemilihan atribut yang akan terpilih sehingga semakin sedikit atribut yang terpilih sebagai model. Pada berbagai riset dan penelitian, *significance level* yang digunakan adalah 0,05 atau 0,1 (Gerard, 2012), sehingga pada penelitian ini akan membandingkan kedua *significance level* tersebut dan menentukan yang terbaik.



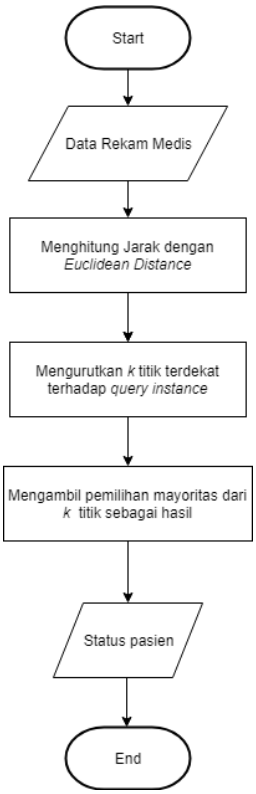

Gambar 3. *Flowchart Backward Elimination*

2. **k-*Nearest Neighbor* (kNN)**, digunakan dalam melakukan deteksi PGK. Algoritme ini memberikan output berdasarkan jarak terpendek dari query instance ke training. Diberikan titik *query*, akan ditemukan sejumlah *k* objek (titik training) yang paling dekat dengan titik *query*. Perhitungan jarak diimplementasikan dengan rumus *Euclidean Distance* pada Persamaan (1).

$$D(a,b) = \sqrt{\sum_{k=1}^{d} (a_k - b_k)^2} \quad (1)$$

Keterangan:
D(a, b) = Jarak antara a dan b dari matrik berdimensi d
a = data training
b = data uji

Sedangkan proses algoritme kNN dalam melakukan deteksi PGK ditunjukkan pada Gambar 4.



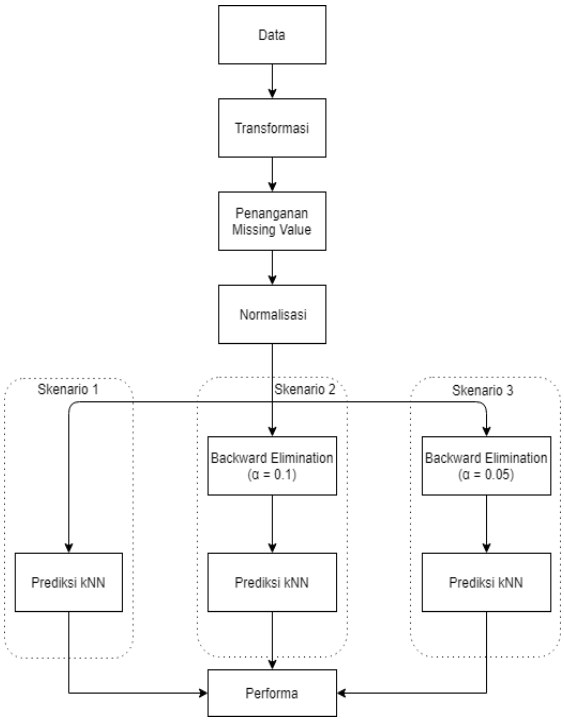

Gambar 4. *Flowchart* k-*Nearest Neighbor*(kNN)

### 2.7. *Data Mining: Employing Data Mining Algorithm*

Dalam mengimplementasikan algoritme *Backward Elimination* dan k-*Nearest Neighbor* (kNN), dilakukan 3 skenario pengujian untuk menentukan pemodelan *data mining* yang terbaik. Pada tiap pengujian ini menggunakan 10 *Cross Fold Validation* dan mencari nilai *k* terbaik dari algoritme kNN. Model terbaik didasarkan kepada rerata hasil akurasi, *sensitivity*, dan *specificity*. Skenario *data mining* yang akan diujikan dapat dilihat pada Gambar 3.

#### 2.7.1. Skenario Pengujian 1

Pada proses pengujian ini peneliti menggunakan model k-*Fold Cross Validation* dengan nilai KFold = 10, menggunakan jumlah tetangga terdekat sejumlah kNN = 3 hingga kNN = 23. Dengan hanya menggunakan kNN bernilai ganjil dimulai dari 3 dimaksudkan agar fungsi *majority votes* pada kNN menjadi konsisten dan memberikan nilai performa yang tetap. Perhitungan performa yang dilakukan akan mendapatkan nilai rata-rata dari akurasi, specificity, dan sensitivity pada setiap k tetangga terdekat, sehingga dapat diketahui jumlah tetangga (*k*) terbaik untuk algoritme kNN dengan dataset PGK seluruh atribut. Hasil pengujian dapat dilihat pada Tabel 2.




Gambar 5. Diagram Skenario Pengujian

Tabel 2. Hasil Pengujian Skenario 1

| Nilai *k* (kNN) | Akurasi (%) | Sensitivity (%) | Specificity (%) |
|---|---|---|---|
| **3** | **99.25** | **98.806** | **100** |
| 5 | 98.75 | 97.906 | 100 |
| 7 | 98 | 96.645 | 100 |
| 9 | 97.5 | 95.826 | 100 |
| 11 | 97.5 | 95.826 | 100 |
| 13 | 97.75 | 96.21 | 100 |
| 15 | 97.75 | 96.21 | 100 |
| 17 | 97.25 | 95.41 | 100 |
| 19 | 96.75 | 94.626 | 100 |
| 21 | 96.75 | 94.626 | 100 |
| 23 | 96 | 93.392 | 100 |

#### 2.7.2. Skenario Pengujian 2

Pada proses pengujian ini peneliti melakukan seleksi atribut terlebih dahulu dengan algoritme Backward Elimination α=0,1, tahapan seleksi atribut dapat dilihat pada Tabel 3.

Setelah dilakukan seleksi atribut, selanjutnya dilakukan pengujian pada sisa atribut hasil *Backward Elimination* sebagai atribut terpilih, peneliti menggunakan model k-*Fold Cross Validation* dengan nilai KFold= 10, menggunakan jumlah tetangga terdekat sejumlah kNN = 3 hingga kNN = 23. Dengan hanya menggunakan kNN bernilai ganjil dimulai dari 3 dimaksudkan agar fungsi *majority votes* pada kNN

menjadi konsisten dan memberikan nilai performa yang tetap. Perhitungan performa yang dilakukan akan mendapatkan nilai rata-rata dari akurasi, specificity, dan sensitivity pada setiap k tetangga terdekat, sehingga dapat diketahui jumlah tetangga (*k*) terbaik untuk algoritme kNN dengan dataset dataset PGK dengan atribut hasil algoritme *Backward Elimination* α=0,1. Hasil pengujian dapat dilihat pada Tabel 4.

Tabel 3. Proses Backward Elimination α=0,1

| | Tahap | Atribut | P-Value |
|---|---|---|---|
| **Atribut Terbuang (P-Val > 0,1)** | 1 | su | 0,977 |
| | 2 | pe | 0,891 |
| | 3 | age | 0,835 |
| | 4 | pc | 0,776 |
| | 5 | pot | 0,798 |
| | 6 | pcc | 0,508 |
| | 7 | pcv | 0,498 |
| | 8 | cad | 0,446 |
| | 9 | rbcc | 0,401 |
| | 10 | ane | 0,291 |
| | 11 | wbcc | 0,182 |
| | 12 | ba | 0,113 |
| **Sisa Atribut (P-Val ≤ 0,1)** | - | bp | 0,069 |
| | - | sg | 0,000 |
| | - | al | 0,000 |
| | - | bgr | 0,076 |
| | - | bu | 0,002 |
| | - | sc | 0,001 |
| | - | sod | 0,000 |
| | - | hemo | 0,000 |
| | - | rbc | 0,000 |
| | - | htn | 0,003 |
| | - | dm | 0,003 |
| | - | appet | 0,046 |

Tabel 4. Hasil Pengujian Skenario 2

| Atribut Terpilih | Nilai *k* (kNN) | Akurasi (%) | Sensitivity (%) | Specificity (%) |
|---|---|---|---|---|
| | 3 | 99 | 98.639 | 99.412 |
| | **5** | **99.25** | **98.639** | **100** |
| bp, sg, al, | 7 | 98.5 | 97.778 | 99.412 |
| bgr, bu, | 9 | 98.25 | 97.43 | 99.412 |
| sc, sod, | 11 | 98.25 | 97.43 | 99.412 |
| hemo, | 13 | 98.25 | 97.43 | 99.412 |
| rbc, htn, | 15 | 98.25 | 97.43 | 99.412 |
| dm, | 17 | 98.25 | 97.43 | 99.412 |
| appet | 19 | 98 | 97.045 | 99.412 |
| | 21 | 97.75 | 96.66 | 99.412 |
| | 23 | 97.75 | 96.66 | 99.412 |

### 2.7.3. Skenario Pengujian 3

Pada proses pengujian ini peneliti melakukan seleksi atribut terlebih dahulu dengan algoritme Backward Elimination α=0,05, tahapan seleksi atribut dapat dilihat pada Tabel 5.

Setelah dilakukan seleksi atribut, selanjutnya dilakukan pengujian pada sisa atribut hasil *Backward Elimination* sebagai atribut terpilih, peneliti menggunakan model k-*Fold Cross Validation* dengan nilai KFold= 10, menggunakan jumlah tetangga terdekat sejumlah kNN = 3 hingga kNN = 23. Dengan hanya menggunakan kNN bernilai ganjil dimulai dari 3 dimaksudkan agar fungsi *majority votes* pada kNN menjadi konsisten dan memberikan nilai performa yang tetap. Perhitungan performa yang dilakukan akan mendapatkan nilai rata-rata dari akurasi, specificity, dan sensitivity pada setiap k tetangga terdekat, sehingga dapat diketahui jumlah tetangga (*k*) terbaik untuk algoritme kNN dengan dataset dataset PGK dengan atribut hasil algoritme *Backward Elimination* α=0,05. Hasil pengujian dapat dilihat pada Tabel 6.

Tabel 5. Proses Backward Elimination α=0,05

| | Tahap | Atribut | P-Value |
|---|---|---|---|
| **Atribut Terbuang (P-Val > 0,05)** | 1 | su | 0,977 |
| | 2 | pe | 0,891 |
| | 3 | age | 0,835 |
| | 4 | pc | 0,776 |
| | 5 | pot | 0,798 |
| | 6 | pcc | 0,508 |
| | 7 | pcv | 0,498 |
| | 8 | cad | 0,446 |
| | 9 | rbcc | 0,401 |
| | 10 | ane | 0,291 |
| | 11 | wbcc | 0,182 |
| | 12 | ba | 0,113 |
| | 13 | bgr | 0,076 |
| | 14 | bp | 0,058 |
| **Sisa Atribut (P-Val ≤ 0,05)** | - | sg | 0,000 |
| | - | al | 0,000 |
| | - | bu | 0,001 |
| | - | sc | 0,000 |
| | - | sod | 0,000 |
| | - | hemo | 0,000 |
| | - | rbc | 0,000 |
| | - | htn | 0,001 |
| | - | dm | 0,000 |
| | - | appet | 0,038 |

Tabel 6. Hasil Pengujian Skenario 3

| Atribut Terpilih | Nilai *k* (kNN) | Akurasi (%) | Sensitivity (%) | Specificity (%) |
|---|---|---|---|---|
| | **3** | **99.25** | **99.5** | 98.745 |
| | 5 | 99.25 | 98.639 | **100** |
| | 7 | 98.25 | 97.445 | 99.412 |
| sg, al, bu, | 9 | 98.25 | 97.445 | 99.412 |
| sc, sod, | 11 | 98.25 | 97.83 | 98.745 |
| hemo, rbc, | 13 | 98 | 97.43 | 98.745 |
| htn, dm, | 15 | 98 | 97.43 | 98.745 |
| appet | 17 | 98.25 | 97.43 | 99.412 |
| | 19 | 98.26 | 97.43 | 99.412 |
| | 21 | 98.25 | 97.43 | 99.412 |
| | 23 | 98.25 | 97.43 | 99.412 |

### 2.8. *Evaluation and Interpretation*

Dilakukannya evaluasi dan interpretasi terhadap hasil *data mining* yang telah dilakukan. Evaluasi melibatkan perhitungan akurasi, *sensitivity*, dan *specificity* dari data tersebut. Pada penelitian ini dilakukan evaluasi menggunakan 10-*Fold Cross Validation* pada data yang berjumlah 400 dengan tetangga terdekat k=3 sampai dengan k=23. Selengkapnya tahapan ini akan dibahas lebih lanjut pada bab 3 bagian hasil dan pembahasan.

### 2.9. *Discovered Knowledge (Visualization and Integration)*

Merupakan tahap terakhir dimana pemodelan yang didapat berupa k terbaik pada algoritme kNN dan atribut terseleksi pada algoritme *Backward Elimination* dikembangkan sebagai sebuah aplikasi

berbasis web dalam mengdeteksi PGK dengan bahasa pemrograman utama Python. Selengkapnya tahapan ini akan dibahas lebih lanjut pada bab 4 bagian kesimpulan.

## 3. HASIL DAN PEMBAHASAN

Pada bab ini akan diuraikan hasil penelitian dan pembahasan mengenai hasil skenario pengujian yang bertujuan untuk mendapatkan pemodelan *data mining* terbaik dalam mendeteksi PGK.

### 3.1. Perbandingan Akurasi

Perbandingan akurasi pada skenario 1, 2 dan 3 dapat dilihat pada Gambar 4, hasil perbandingan tersebut menunjukkan performa akurasi skenario 1, 2 dan 3 memiliki akurasi tertinggi yang sama, yaitu 99.25%. Akurasi tertinggi skenario 1 dan 3 pada nilai *k*=3, sedangkan pada skenario 2 akurasi tertinggi ketika nilai *k*=5.

### 3.2. Perbandingan *Sensitivity*

Perbandingan *sensitivity* pada skenario 1, 2 dan 3 dapat dilihat di Gambar 5. Hasil perbandingan tersebut menunjukkan bahwa pada semua skenario, *sensitivity* tertinggi berada pada nilai *k*=3. Sedangkan nilai *sensitivity* tertinggi ada pada skenario 3 yang mencapai 99.5%. Dapat disimpulkan juga bahwa semakin besar nilai *k*, maka ada kecenderungan semakin menurunkan nilai *sensitivity*.

### 3.3. Perbandingan *Specificity*

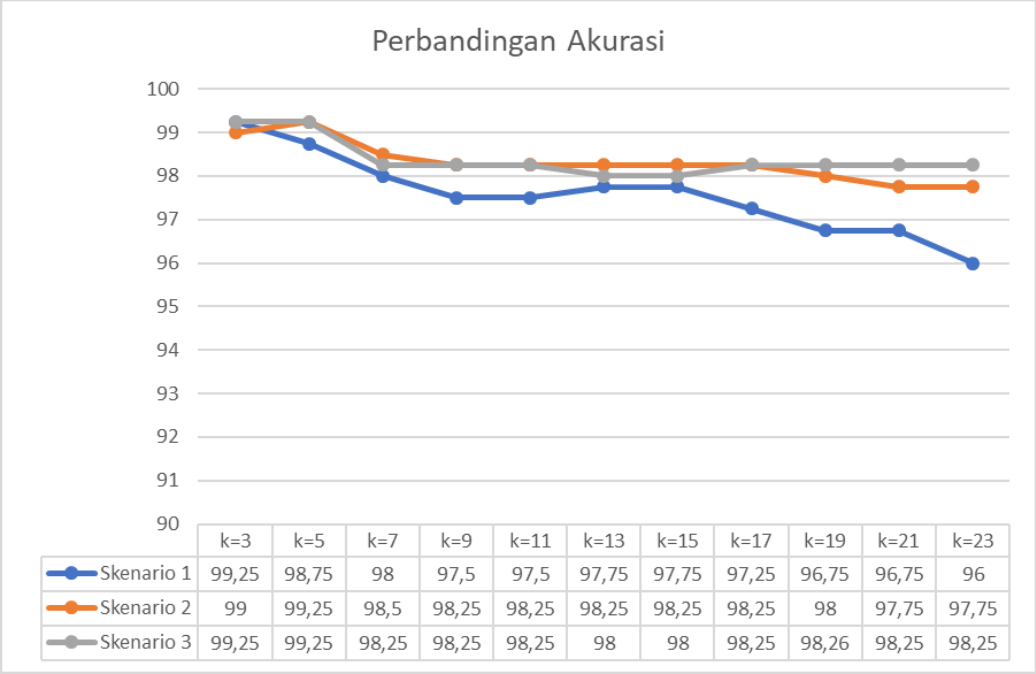
Perbandingan *specificity* pada skenario 1, 2 dan 3 dapat dilihat di Gambar 6. Performa *specificity* pada skenario 1 (pemakaian seluruh atribut) memiliki hasil yang paling baik dibandingkan skenario 2 dan 3 yang menggunakan seleksi atribut *Backward Elimination*. Hasil perbandingan tersebut menunjukkan performa *specificity* terbaik adalah skenario 1 dengan specificity tetap berada pada nilai 100%, sedangkan pada skenario 2 dan 3 yang menggunakan *Backward Elimination,* semakin besar nilai *k*, maka ada kecenderungan semakin menurunkan nilai *specificity*.

### 3.4. Analisa Biaya Tes Laboratorium

Dari tabel biaya tes laboratorium pada penelitian sebelumnya (Salekin & Stankovic, 2016) maka akan disimulasikan perbandingan biaya tes laboratorium seperti yang ditunjukkan pada Tabel 7.

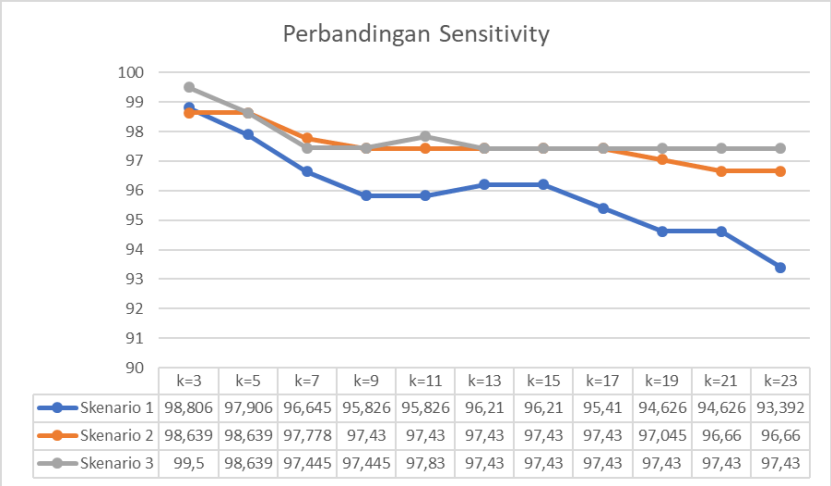
Seperti yang dilihat pada tabel 7, jika dilakukan simulasi penjumlahan tes laboratorium dengan penggunaan seluruh atribut maka membutuhkan biaya sebesar 391,36 USD. Sedangkan jika menggunakan hasil seleksi atribut *Backward Elimination* dengan α=0,1 dan α=0,05 berturut – turut adalah 133,1 USD dan 103,1 USD.

Melalui simulasi yang telah dilakukan, sehingga didapatkan hasil *Backward Elimination* (α=0,05) dapat menekan biaya pemeriksaan PGK paling efektif sebesar 73,36% dengan akurasi 99,25% dan sensitifitas tertinggi sebesar 99,5%.



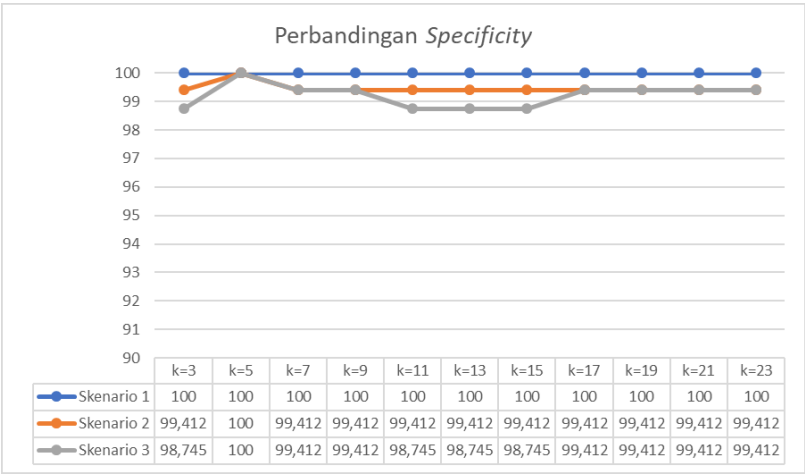

| | k=3 | k=5 | k=7 | k=9 | k=11 | k=13 | k=15 | k=17 | k=19 | k=21 | k=23 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Skenario 1 | 99,25 | 98,75 | 98 | 97,5 | 97,5 | 97,75 | 97,75 | 97,25 | 96,75 | 96,75 | 96 |
| Skenario 2 | 99 | 99,25 | 98,5 | 98,25 | 98,25 | 98,25 | 98,25 | 98,25 | 98 | 97,75 | 97,75 |
| Skenario 3 | 99,25 | 99,25 | 98,25 | 98,25 | 98,25 | 98 | 98 | 98,25 | 98,26 | 98,25 | 98,25 |

Gambar 6. Perbandingan Akurasi




| | k=3 | k=5 | k=7 | k=9 | k=11 | k=13 | k=15 | k=17 | k=19 | k=21 | k=23 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Skenario 1 | 98,806 | 97,906 | 96,645 | 95,826 | 95,826 | 96,21 | 96,21 | 95,41 | 94,626 | 94,626 | 93,392 |
| Skenario 2 | 98,639 | 98,639 | 97,778 | 97,43 | 97,43 | 97,43 | 97,43 | 97,43 | 97,045 | 96,66 | 96,66 |
| Skenario 3 | 99,5 | 98,639 | 97,445 | 97,445 | 97,83 | 97,43 | 97,43 | 97,43 | 97,43 | 97,43 | 97,43 |

Gambar 7. Perbandingan Sensitivity




| | k=3 | k=5 | k=7 | k=9 | k=11 | k=13 | k=15 | k=17 | k=19 | k=21 | k=23 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Skenario 1 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| Skenario 2 | 99,412 | 100 | 99,412 | 99,412 | 99,412 | 99,412 | 99,412 | 99,412 | 99,412 | 99,412 | 99,412 |
| Skenario 3 | 98,745 | 100 | 99,412 | 99,412 | 98,745 | 98,745 | 98,745 | 99,412 | 99,412 | 99,412 | 99,412 |

Gambar 8. Perbandingan *Specificity*

Tabel 7. Perkiraan Biaya Tes Laboratorium

| No | Atribut | Biaya (USD) | | |
|---|---|---|---|---|
| | | 24 Atribut | Backward Elimination | |
| | | | 12 Atribut (α=0,1) | 10 Atribut (α=0,1) |
| 1 | Usia (*age*) | Gratis | - | - |
| 2 | Tekanan Darah (*bp*) | Gratis | Gratis | - |
| 3 | Berat Jenis (*sg*) | Gratis | Gratis | Gratis |
| 4 | Albumin (*al*) | 25 | 25 | 25 |
| 5 | Gula (*su*) | 20 | - | - |
| 6 | Sel Darah Merah (*rbc*) | 39 | 39 | 39 |
| 7 | Sel Darah Putih (*pc*) | 30 | - | - |
| 8 | Gumpalan Sel Nanah (*pcc*) | | - | - |
| 9 | Bakteri (*ba*) | 50 | - | - |
| 10 | Gula Darah Acak (*bgr*) | 20 | 20 | - |
| 11 | Urea Darah (*bu*) | 11,85 | 11,85 | 11,85 |
| 12 | Serum Kreatinin (*sc*) | 14 | 14 | 14 |
| 13 | Sodium (*sod*) | 3,2 | 3,2 | 3,2 |
| 14 | Potasium (*pot*) | 49 | - | - |
| 15 | Hemoglobin (*hemo*) | 1,65 | 1,65 | 1,65 |
| 16 | Hematokrit (*pcv*) | 1,62 | - | - |
| 17 | Jumlah Sel Darah Putih (*wbcc*) | 30 | - | - |
| 18 | Jumlah Sel Darah Merah (*rbcc*) | | - | - |
| 19 | Hipertensi (*htn*) | Gratis | Gratis | Gratis |
| 20 | Diabetes Mellitus (d*m*) | 18,4 | 18,4 | 18,4 |
| 21 | Penyakit Jantung Koroner (*cad*) | 50 | - | - |
| 22 | Selera Makan (*appet*) | Gratis | Gratis | Gratis |
| 23 | Pembengkakan pada Kaki (p*e*) | Gratis | - | - |
| 24 | Anemia (*ane*) | 27,64 | - | - |
| **TOTAL** | | **391,36** | **133,1** | **103,1** |

## 4. KESIMPULAN

Dengan menggunakan seleksi fitur *Backward Elimination*, performa akurasi dan *sensitivity* yang dihasilkan lebih baik dibandingkan tanpa seleksi atribut. Perubahan nilai *k* pada k-*Nearest Neighbor* (kNN) mempengaruhi nilai rata-rata akurasi, sensitivity, dan specificity dalam deteksi penyakit ginjal kronis (PGK). Semakin besar nilai *k* memiliki kecenderungan penurunan performa akurasi dan *sensitivity*. Dikarenakan akurasi tertinggi semua skenario *data mining* sama, maka dilakukan pertimbangan selanjutnya pada *sensitivity*, dimana skenario ke-3 yaitu pemodelan *Backward Elimination* (α=0,05) dengan kNN nilai *k*=3 menghasilkan nilai *sensitivity* tertinggi sebesar 99,5% dibandingkan skenario pemodelan lainnya.

Skenario ke-3 *Backward Elimination* (α=0,05) menghasilkan seleksi atribut yaitu: berat jenis (*sg*), albumin (*al*), urea darah (*bu*), kreatinin serum (*sc*), sodium (*sod*), hemoglobin (*hemo*), sel darah merah (*rbc*), hipertensi (*htn*), diabetes mellitus (*dm*), nafsu makan (*appet*). Pemodelan skenario ini menurunkan biaya pemeriksaan laboratorium hingga 73,36% melalui penyederhanaan inputan yang perlu dimasukkan oleh *user*.

## DAFTAR PUSTAKA


BADAN PENELITIAN DAN PENGEMBANGAN KESEHATAN, 2013. Riset Kesehatan Dasar (RISKESDAS) 2013. *Laporan Nasional 2013*, hal.1–384.

BORGES, L.C., MARQUES, V.M. DAN BERNARDINO, J., 2013. Comparison of data mining techniques and tools for data classification. *Proceedings of the International C* Conference on Computer Science and Software Engineering - C3S2E '13*, [daring] (July), hal.113. Tersedia pada: <http://dl.acm.org/citation.cfm?doid=2494444.2494451>.

GERARD, E.D., 2012. *Simplifying a Multiple Regression Equation*. *The Little Handbook of Statistical Practice*, .

HAN, J. DAN KAMBER, M., 2011. *Data Mining: Concepts and Techniques*. [daring] *Elsevier*, Tersedia pada: <http://link.springer.com/10.1007/978-3-642-19721-5>.

JADHAV, S.D. DAN CHANNE, H.P., 2013. Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques. *International Journal of Science and Research (IJSR)*, [daring] 14611(1), hal.2319–7064. Tersedia pada: <www.ijsr.net>.

JING, Z., WEI-JIE, Y., NAN, Z., YI, Z. DAN LING, W., 2012. Hemoglobin Targets for Chronic Kidney Disease Patients with Anemia: A Systematic Review and Meta-analysis. *PLoS ONE*, 7(8).

KEMENKES, 2017. *InfoDATIN*. Kementrian Kesehatan RI.

SHAFIQUE, U. DAN QAISER, H., 2014. A Comparative Study of Data Mining Process Models ( KDD , CRISP-DM and SEMMA ). *International Journal of Innovation and Scientific Research*, [daring] 12(1), hal.217–222. Tersedia pada: <http://www.ijisr.issr-journals.org/>.

SINHA, P. SINHA; P., 2015. Comparative Study of Chronic Kidney Disease Prediction using KNN and SVM. [daring] 4(12), hal.608–612. Tersedia pada: <https://pdfs.semanticscholar.org/3ec0/5afd1eb4bb4d5ec17a9e0b3d09f5cbc30304.pdf>.