

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



TUGAS AKHIR

**Disusun Sebagai Salah Satu Syarat
untuk Memperoleh Gelar Sarjana Komputer
pada Departemen Ilmu Komputer / Informatika**

**Disusun Oleh:
IKHSAN WISNUADJI G
24010313130108**

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Saya yang bertanda tangan dibawah ini:

Nama : Ikhsan Wisnuadji Gamadarenda

NIM : 24010313130108

Judul : Implementasi Data Mining untuk Identifikasi Penyakit Ginjal Kronis (PGK)

Menggunakan *K-Nearest Neighbor* (kNN) dengan *Backward Elimination*

Dengan ini menyatakan bahwa dalam tugas akhir/skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan di dalam daftar pustaka.

Semarang, xx November 2018

Ikhsan Wisnuadji Gamadarenda

NIM. 24010313130108

HALAMAN PENGESAHAN

Judul : Implementasi Data Mining untuk Identifikasi Penyakit Ginjal Kronis (PGK) Menggunakan K-Nearest Neighbor (kNN) dengan *Backward Elimination*.

Nama : Ikhsan Wisnuadji Gamadarenda

NIM : 24010313130108

Telah diujikan pada sidang tugas akhir pada tanggal xx November 2018 dan dinyatakan lulus pada tanggal xx November 2018

Semarang, xx November 2018

Mengetahui,
Ketua Departemen Ilmu
Komputer/Informatika

Panitia Penguji Skripsi,
Ketua,

Dr. Retno Kusumaningrum, S.Si, M.Kom
NIP. 198104202005012001

XXXXXXXXXXXX
NIP. XXXXXXXXXXXXXXXX

HALAMAN PENGESAHAN

Judul : Implementasi Data Mining untuk Identifikasi Penyakit Ginjal Kronis (PGK)
Menggunakan K-Nearest Neighbor dengan Backward Elimination
Nama : Ikhsan Wisnuadji Gamadarenda
NIM : 24010313130108

Telah diujikan pada sidang tugas akhir pada tanggal xx November 2018

Semarang, xx November 2018

Dosen Pembimbing

Indra Waspada, ST, MTI

NIP. 197902122008121002

ABSTRAK

ABSTRACT

KATA PENGANTAR

DAFTAR ISI

HALAMAN PERNYATAAN KEASLIAN SKRIPSI	ii
HALAMAN PENGESAHAN	iii
HALAMAN PENGESAHAN	iv
ABSTRAK.....	v
ABSTRACT.....	vi
KATA PENGANTARDAFTAR ISI	vii
DAFTAR TABEL.....	xi
DAFTAR GAMBAR	xii
DAFTAR PERSAMAAN	xiii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah.....	3
1.3 Tujuan dan Manfaat	3
1.4 Ruang Lingkup	3
1.5 Sistematika Penulisan	4
BAB II TINJAUAN PUSTAKA	5
2.1 Penelitian Terkait	5
2.2 Penyakit Ginjal Kronis.....	6
2.3 Pemodelan Data Mining	7
2.4 k-Nearest Neighbor	8
2.5 Feature Selection.....	12
2.6 Algoritma Backward Elimination	13
2.7 k-Fold Cross Validation.....	15

2.8	Evaluasi Sistem.....	15
BAB III METODE PENELITIAN		17
3.1	Metode Penelitian	17
3.2	Lokasi Penelitian.....	17
3.3	Garis Besar Penyelesaian Masalah	17
3.3.1	Integrasi dan Pemahaman Data	18
3.3.2	Pembuatan Dataset dan Target.....	20
3.3.3	<i>Data Cleaning</i> dan <i>Pre-Processing</i>	20
3.3.4	Transformasi Data	21
3.3.5	<i>Data Mining</i> : Pemilihan Jenis Tugas <i>Data Mining</i>	21
3.3.6	<i>Data Mining</i> : Penentuan Algoritma dan Metode	21
3.3.7	<i>Data Mining</i> : Implementasi Algoritma Data Mining.....	21
3.3.8	Evaluasi dan Interpretasi	23
3.3.9	Penggunaan Informasi yang Didapatkan.....	23
3.4	Analisis dan Desain Sistem.....	23
3.4.1	Deskripsi Sistem.....	23
3.4.2	Deskripsi Umum Sistem.....	24
3.4.3	Kebutuhan Fungsional Sistem.....	25
3.4.4	Pemodelan Fungsi	25
BAB IV HASIL DAN PEMBAHASAN		28
4.1	Hasil Pengembangan Sistem.....	28
4.1.1	Lingkungan Implementasi.....	28
4.1.2	Implementasi Antarmuka	28
4.2	Skenario Pengujian Sistem	38
4.2.1	Skenario Pengujian Fungsional Sistem	38
4.2.2	Skenario Pengujian.....	39
4.3	Analisis Hasil Pengujian	41

4.3.1	Analisis Hasil Pengujian Fungsional Sistem.....	41
4.3.2	Analisis Hasil Skenario 1	41
4.3.3	Analisis Hasil Skenario 2	42
BAB V KESIMPULAN DAN SARAN		45
5.1	Kesimpulan	45
5.2	Saran	45
DAFTAR PUSTAKA		46

DAFTAR TABEL

Tabel 3.3.1-1 Penelitian Terkait	5
Tabel 3.3.1-1 Data Latih Kasus Algoritma kNN	10
Tabel 3.3.1-2 Data Uji Kasus Algoritma kNN	11
Tabel 3.3.1-3 Perhitungan Selisih Nilai Data Latih dengan Data Uji.....	11
Tabel 3.3.1-4 Hasil Perhitungan Euclidean Distance	11
Tabel 3.3.1-1 <i>Confusion Matrix</i>	16
Tabel 3.3.1-1 Data Atribut dan Tipe Data	19
Tabel 3.4.3-1 Kebutuhan Fungsional	25
Tabel 4.3.2-1 Hasil Pengujian Skenario 1	41
Tabel 4.3.3-1 Hasil Pengujian Skenario 2	43

DAFTAR GAMBAR

Gambar 3.3.1-1 Framework <i>Knowledge Data Discovery</i>	7
Gambar 3.3.1-1 <i>Flowchart</i> Algoritma kNN	10
Gambar 3.3.1-1 Teknik Seleksi Atribut	13
Gambar 3.3.1-1 <i>Logic Function</i> dari Teknik <i>Wrapper Approach</i> pada Operator <i>Backward Elimination</i> dengan RapidMiner	14
Gambar 3.3.1-2 <i>Flowchart</i> proses <i>Backward Elimination</i>	14
Gambar 3.3.1-1 Ilustrasi <i>k-Fold Cross Validation</i>	15
Gambar 3.3.1-1 Kerangka Kerja Penelitian	18
Gambar 3.3.7-1 <i>Flowchart</i> Penentuan Nilai <i>k-</i> pada kNN.....	22
Gambar 3.4.2-1 Arsitektur Sistem Identifikasi Penyakit Ginjal Kronis	24
Gambar 3.4.4-1 DFD level 0.....	26
Gambar 3.4.4-2 DFD level 1.....	27
Gambar 4.1.2-1 Antarmuka halaman <i>Homepage</i>	29
Gambar 4.1.2-2 Perbandingan antarmuka <i>Pemodelan Langsung</i> (atas) dan <i>Pemodelan Manual</i> (bawah).....	30
Gambar 4.1.2-3 Antarmuka halaman <i>Dataset</i> pada proses Pemodelan Manual.....	31
Gambar 4.1.2-4 Antarmuka Pembersihan Outlier pada proses Pemodelan Manual..	32
Gambar 4.1.2-5 Antarmuka halaman Transformasi Data Nominal pada proses Pemodelan Manual	33
Gambar 4.1.2-6 Antarmuka halaman Penanganan <i>Missing Value</i> pada proses Pemodelan Manual	34
Gambar 4.1.2-7 Antarmuka halaman Normalisasi pada proses <i>Demo</i>	35
Gambar 4.1.2-8 Antarmuka halaman hasil proses Pemodelan	36
Gambar 4.1.2-9 Antarmuka halaman <i>Input</i> proses Identifikasi	37
Gambar 4.1.2-10 Antarmuka halaman Hasil Identifikasi	38
Gambar 4.2.2-1 Jumlah Data yang Digunakan Sistem	40
Gambar 4.3.2-1 Grafik Performa kNN Skenario 1	42
Gambar 4.3.3-1 Perbandingan Performa Akurasi	43
Gambar 4.3.3-3 Perbandingan Performa Sensitivity	44
Gambar 4.3.3-4 Perbandingan Performa Specificity	44

DAFTAR PERSAMAAN

Persamaan 2.4-1 Rumus Perhitungan Jarak.....	9
Persamaan 2.8-1 Perhitungan <i>Sensitivity</i>	16
Persamaan 2.8-2 Perhitungan <i>Specifity</i>	16

BAB I

PENDAHULUAN

Bab ini menyajikan latar belakang, rumusan masalah, tujuan dan manfaat, ruang lingkup, dan sistematika penulisan dalam pembuatan Implementasi *Data Mining* untuk Identifikasi Penyakit Ginjal Kronis (PGK) Menggunakan Algoritma *k-Nearest Neighbor* (kNN) dengan *Backward Elimination*.

1.1 Latar Belakang

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. Kelainan tersebut dapat mempengaruhi struktur dan fungsi ginjal dengan tingkat keparahan yang berbeda-beda. Didefinisikan sebagai Penyakit Ginjal Kronis (PGK) jika pernah didiagnosis menderita penyakit gagal ginjal kronis (minimal sakit selama 3 bulan berturut-turut) oleh dokter (Riset Kesehatan Dasar, 2013). Penyakit tersebut pada awalnya tidak menunjukkan tanda dan gejala namun dapat berjalan progresif menjadi gagal ginjal (Kementrian Kesehatan, 2017).

PGK merupakan masalah kesehatan publik diseluruh dunia dengan insiden yang terus meningkat. Diperkirakan 2,5-11,2% populasi penduduk dewasa dari Eropa, Asia, Amerika Utara dan Australia dilaporkan mengalami PGK (Zhang dan Rothenbacher, 2008). Lebih dari 27 juta individu di Amerika Serikat mengalami PGK (Baumgarten dan Gehr, 2011). Sedangkan prevalensi penduduk Indonesia menderita PGK adalah 0,2% (Riskesdas, 2013).

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 (Locatelli, et al., 2002). Perawatan PGK merupakan ranking kedua pembiayaan terbesar dari BPJS kesehatan setelah penyakit jantung (Kementrian Kesehatan, 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 algoritma 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, et al., 2013).

Tujuan yang ingin dicapai dalam penulisan tugas akhir ini adalah menghasilkan sistem dengan salah satu algoritma *data mining* untuk membantu pendeteksian PGK. Sehingga pasien yang terdiagnosis dapat dilakukan tindakan lanjut secara cepat dan tepat untuk menanggulangi tingkat kerusakan dan biaya pengobatan yang lebih besar.

Data mining yang pernah diaplikasikan dalam bidang kesehatan misalnya diagnosis penyakit Diabetes Mellitus dengan menggunakan algoritma C4.5 76,10% dan *k-Nearest Neighbor* 79,14% (Karyono, 2016). Ada pula penelitian yang pernah membandingkan algoritma yang digunakan dalam pendeteksian PGK dari dataset *UC Irvine Machine Learning Repository* (UCI) menunjukkan *k-Nearest Neighbor* (kNN) dengan akurasi 78,75% dibandingkan dengan Support Vector Machine (SVM) 73,75% (Sinha, 2015). Berdasarkan penelusuran penelitian terkait yang pernah dilakukan dalam mendiagnosis PGK, algoritma kNN memiliki tingkat akurasi paling tinggi dalam diagnosis penyakit.

Teknik *k-Nearest Neighbour* atau kNN merupakan model klasifikasi non parametrik, dimana memiliki beberapa kelebihan, penerapannya yang sederhana namun efektif dalam banyak kasus. 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 algoritma 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 atau bernilai kosong serta mempunyai nilai atribut yang berbeda dengan atribut sejenis lainnya, sehingga dapat menyebabkan masalah pada perhitungan jaraknya. Hal ini mengakibatkan menurunnya akurasi dalam proses klasifikasi pada algoritma kNN. *Backward Elimination* pada tahap

preprocessing bertujuan untuk menghilangkan atribut-atribut yang tidak relevan tersebut sehingga diharapkan akurasi yang didapatkan meningkat. Penelitian sebelumnya pernah dilakukan dengan tujuan peningkatan akurasi pada *k-Nearest Neighbor* dengan *Backward Elimination* untuk mendiagnosis penyakit jantung. Hasil menunjukkan adanya peningkatan dengan metode ini hasil akurasi 88,62% menjadi 89,55% (Hermawanti dan Rabiha, 2014).

Berdasarkan masalah dan uraian yang telah dikemukakan, maka dibuatlah topik Tugas Akhir (TA) dengan judul “Implementasi *Data Mining* untuk Identifikasi Penyakit Ginjal Kronis (PGK) menggunakan Algoritma *k-Nearest Neighbor* (kNN) dengan *Backward Elimination*”.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dituliskan, disusun rumusan masalah yaitu:

1. Bagaimana penerapan dan perbandingan kinerja algoritma kNN dengan *Backward Elimination* dalam identifikasi PGK?
2. Apa saja atribut terbaik dari data rekam medis dalam mendiagnosis PGK.

1.3 Tujuan dan Manfaat

Tujuan dari penelitian ini adalah menghasilkan sistem yang dapat mendiagnosis penyakit diabetes dengan menggunakan algoritma kNN dengan *Backward Elimination*.

Adapun manfaat dilakukan penelitian Tugas Akhir ini adalah:

1. Hasil sistem dapat digunakan oleh masyarakat umum dan penyedia pelayanan kesehatan untuk identifikasi PGK.
2. Melakukan pendeteksian PGK secepat mungkin sehingga dapat menanggulangi kerusakan dan menekan biaya pengobatan.

1.4 Ruang Lingkup

Diberikan ruang lingkup agar pembahasan lebih jelas, terarah dan tidak menyimpang dari tujuan penelitian. Adapun ruang lingkup dalam penelitian ini adalah sebagai berikut:

1. Bahasa pemrograman yang digunakan dalam pengembangan sistem adalah Python.
2. Data set identifikasi 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.
3. *Output* dari system ini klasifikasi berupa 2 golongan, “PGK” atau “Normal”.

1.5 Sistematika Penulisan

Sistematika penulisan yang digunakan dalam tugas akhir ini terbagi dalam beberapa pokok bahasan, yaitu:

BAB I PENDAHULUAN

Bab ini berisi latar belakang masalah, rumusan masalah, tujuan dan manfaat, ruang lingkup, dan sistematika penulisan dalam pembuatan tugas akhir

BAB II TINJAUAN PUSTAKA

Bab ini menyajikan tinjauan pustaka yang berhubungan dengan topik tugas akhir. Dasar teori digunakan dalam penyusunan tugas akhir ini hingga selesai terciptanya sistem tersebut sehingga dapat diimplementasikan.

BAB III ANALISIS DAN PERANCANGAN

Bab ini membahas tahap analisis kebutuhan perancangan sistem serta hasil yang didapat pada tahap ini

BAB IV IMPLEMENTASI DAN PENGUJIAN

Bab ini membahas proses pengembangan sistem dan hasil yang didapat pada tahap implementasi. Bab ini berisi rincian pengujian sistem yang dibangun dengan metode *blackbox*.

BAB V PENUTUP

Bab ini berisi kesimpulan yang diambil berkaitan dengan sistem yang dikembangkan dan saran untuk pengembangan penelitian lebih lanjut.

BAB II

TINJAUAN PUSTAKA

Bab ini membahas tinjauan pustaka yang diambil dari literatur mengenai Aplikasi *Data Mining* untuk Implementasi *Data Mining* untuk Identifikasi Penyakit Ginjal Kronis (PGK) menggunakan Algoritma *k-Nearest Neighbor* (kNN) dengan *Backward Elimination*.

2.1 Penelitian Terkait

Dalam penelitian Tugas Akhir ini, penulis mereferensi dari penelitian – penelitian sebelumnya yang berkaitan dengan latar belakang masalah. Penelitian terkait dapat dilihat pada tabel 2.1-1.

Tabel 3.3.1-1 Penelitian Terkait

No	Peneliti dan Tahun	Dataset	Diagnosa Penyakit	Hasil
1	Giat Karyono, 2016	UCI: Pima Indians Diabetes	Diabetes Mellitus	Algoritma kNN menghasilkan akurasi lebih tinggi (79,14%) dibandingkan C4.5 (76,10%.)
2	Parul Sinha, 2015	UCI: Chronic Kidney Disease	Ginjal Kronis	kNN menghasilkan akurasi lebih tinggi (78,75%) dibandingkan SVM (73,75%)
3	Achmad Nuruddin Safriandono, 2016	UCI: Heart Disease	Jantung Koroner	kNN menghasilkan akurasi lebih tinggi jika menggunakan <i>Forward Selection</i> (95,29% → 96,08%)
4	Salekin dan Stankovic, 2016	UCI: Chronic Kidney Disease	Ginjal Kronis	kNN menghasilkan akurasi paling tinggi jika dilakukan penanganan missing value (99,3%) dibandingkan Random Forest (99%) dan Neural Network (98,5%) Ditambahkan tahap <i>feature selection</i> menghasilkan hasil akhir 11 atribut penting

Pada tabel 2.1-1 menunjukkan diagnosis Penyakit Ginjal Kronis (PGK) pernah dilakukan dengan membandingkan algoritma *k-Nearest Neighbors* (kNN) dan Support Vector Machine, algoritma kNN mendapatkan hasil yang lebih baik dibandingkan dengan

SVM. Pada penelitian kesehatan yang lain, menunjukkan bahwa kNN memiliki tingkat akurasi lebih tinggi dibandingkan algoritma *Decision Tree C4.5* pada diagnosis penyakit Diabetes Melitus. Penelitian Giat Karyono pada PGK tidak menggunakan *Feature Selection* pada tahap *preprocessing*, sedangkan pada penelitian kesehatan lain dengan kasus Jantung Koroner, dicontohkan penelitian Achmad Nuruddin pada kasus Jantung Koroner menunjukkan algoritma kNN akan memiliki akurasi yang lebih baik jika digabungkan dengan *Feature Selection* algoritma *Backward Elimination*.

Salekin dan Stankovic melakukan penelitian lain dengan menghasilkan algoritma kNN tertinggi ketika melakukan penanganan pada missing value sebesar 99,3% dibandingkan *Random Forest* dan *Neural Network*, penelitian ini juga melakukan *feature selection* dengan metode *wrapper approach* algoritma *Best First Search* yang mendapatkan 11 atribut terbaik. Hasil akhir algoritma ini mirip dengan *Forward Selection*, dimana algoritma ini hanya merepresentasikan kemampuan prediktif pada setiap atribut secara individu, sehingga pada penelitian ini akan membandingkan hasil atribut dari penelitian tersebut dengan metode *Wrapper Approach* dengan algoritma *Backward elimination*. Algoritma *Backward Elimination* dipilih karena kemampuan algoritma ini yang memungkinkan untuk mendapatkan beberapa atribut yang awalnya memiliki kemampuan prediktif rendah secara individu namun jika digabungkan dengan atribut lainnya akan memiliki akurasi yang tinggi (Gerard, 2012), sehingga dibuatlah topik Tugas Akhir ini dengan judul “Implementasi *Data Mining* untuk Identifikasi Penyakit Ginjal Kronis (PGK) Menggunakan Algoritma *k-Nearest Neighbor* (kNN) dengan *Backward Elimination*”.

2.2 Penyakit Ginjal Kronis

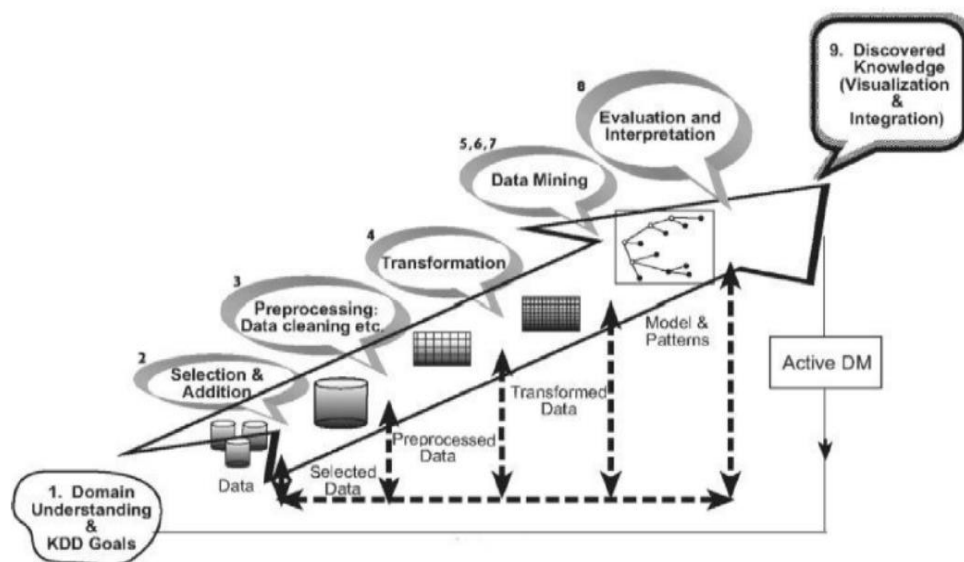
Penyakit Ginjal Kronis (PGK) adalah suatu proses patofisiologis dengan etiologi yang beragam, mengakibatkan penurunan fungsi ginjal yang progresif, penurunan fungsi ini bersifat kronis dan irreversible (Fakhrudin, 2013). Mengingat sifat penyakit ini yang *irreversible*, maka penyakit gagal ginjal lebih baik untuk dicegah, dan melakukan penganggulan lebih awal, sehingga pasien yang terkena PGK dapat mendapatkan terapi yang efektif. 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. Perawatan PGK merupakan ranking kedua pembiayaan terbesar dari BPJS kesehatan setelah penyakit jantung (Kemenkes, 2017).

2.3 Pemodelan Data Mining

Data mining adalah serangkaian proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui manusia dari suatu basis data. Informasi yang dihasilkan diperoleh dengan cara mengekstraksi dan mengenali pola penting atau menarik dari data yang terdapat dalam suatu basis data.

Terdapat tiga buah pemodelan yang cukup populer dalam data mining: 1) KDD; 2) CRISP-DM; dan 3) SEMMA. Dalam penelitian (Shafique dan Qaiser, 2014) yang membandingkan pemodelan tersebut, menyatakan bahwa semua pemodelan dapat digunakan dalam skenario apapun, CRISP-DM dan SEMMA merupakan pemodelan *enterprise* yang sering digunakan oleh perusahaan. Sedangkan pemodelan KDD lebih sering digunakan oleh peneliti dalam data mining karena dianggap lebih lengkap dan akurat.

Knowledge Data Discovery (KDD), adalah proses mengekstraksi pengetahuan tersembunyi dari sebuah database. KDD membutuhkan pengetahuan sebelumnya yang relevan dan pemahaman tentang domain dan tujuan aplikasi (Fayyad, Piatetsky-Shapiro dan Smyth, 1996). Adapun sembilan tahapan yang harus dilalui dalam proses data mining menurut (Shafique dan Qaiser, 2014) ditunjukkan pada gambar 2.3-1:



Gambar 3.3.1-1 Framework *Knowledge Data Discovery*

Dengan penjelasan tahapan *Knowledge Data Discovery* (KDD) sebagai berikut:

1. *Developing and Understanding of The Application Domain*, bertujuan untuk menentukan sudut pandang customer dan digunakan untuk mengembangkan dan memahami tentang domain dari aplikasi dan pengetahuan sebelumnya.

2. *Creating a Target Data Set*, fokus kepada pembuatan target data set dan subset dari data sampel atau variabel. Merupakan tahap yang penting dikarenakan penemuan pengetahuan dilakukan pada tahap ini.
3. *Data Cleaning and Pre-processing*, berfokus pada strategi pembersihan data target dan melengkapi *pre-processing* sehingga data konsisten dan tanpa *noise*.
4. *Data Transformation*, fokus pada transformasi data dari satu bentuk ke bentuk lainnya sehingga algoritma data mining dapat diimplementasikan dengan mudah.
5. *Choosing the Suitable Data Mining Task*, tugas *data mining* yang sesuai dipilih berdasarkan tujuan tertentu yang didefinisikan dalam tahap pertama. Contoh – contoh metode atau tugas *data mining* antara lain: *Classification* (Klasifikasi), *Clustering* (Pengelompokan), *Regression* (Regresi), *Summerization* (Peringkasan), dll.
6. *Choosing the Suitable Data Mining Algorithm*, satu atau lebih algoritma *data mining* yang cocok akan dipilih untuk mencari pola berbeda dari data. Ada sejumlah algoritma yang hadir saat ini untuk *data mining* tetapi algoritma yang sesuai dipilih berdasarkan pencocokan kriteria keseluruhan untuk *data mining*.
7. *Employing Data Mining Algorithm*, merupakan tahap implementasi algoritma *data mining* yang dipilih.
8. *Interpreting Mined Patterns*, fokus pada interpretasi dan evaluasi pola dari hasil yang didapat. Pada tahap ini mungkin melibatkan visualisasi dari pola yang telah diekstraksi.
9. *Using Discovered Knowledge*, merupakan tahap akhir dimana pengetahuan yang diperoleh digunakan dalam tujuan tertentu. Penemuan pengetahuan juga dapat digunakan pada pihak yang tertarik atau dapat mengintegrasikannya pada sebuah sistem untuk mendapatkan tindak lanjut.

2.4 k-Nearest Neighbor

Algoritma *k*-Nearest Neighbor (kNN) merupakan metode yang sangat populer dalam *data mining* dikarenakan implementasinya yang mudah. kNN adalah sebuah metode untuk melakukan klasifikasi terhadap objek menurut sampel data yang memiliki jarak paling dekat dengan objek tersebut. kNN merupakan algoritma *supervised learning* yang berarti hasil dari *query instance* yang baru diklasifikasikan didasarkan kepada mayoritas dari kategori pada

algoritma kNN. Kelas yang paling banyak muncul nantinya akan menjadi kelas hasil dari klasifikasi yang baru.

Tujuan algoritma kNN adalah mengklasifikasikan objek berdasarkan atribut dan *training sample*. *Classifier* tidak menggunakan model apapun untuk dicocokkan dan hanya berdasarkan pada memori. Algoritma metode kNN sangatlah sederhana, bekerja berdasarkan jarak terpendek dari query instance ke training sample untuk menentukan KNN-nya. Diberikan titik *query*, akan ditemukan sejumlah *k* objek (titik training) yang paling dekat dengan titik *query*. Kelebihan *k*-Nearest Neighbor:

1. Tangguh terhadap *training data* yang memiliki banyak *noise*.
2. Efektif apabila training datanya cukup besar.

Terdapat beberapa rumus perhitungan jarak dalam algoritma kNN, diantaranya yang akan dipakai adalah rumus *Euclidean Distance*. Rumus tersebut cocok untuk tipe data numerik. Berikut adalah rumus *Euclidean Distance* yang digunakan untuk menghitung jarak terdekat dari data uji ke data latih (2.4-1):

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

Keterangan:

$D(a, b)$ = Jarak antara *a* dan *b* dari matrik berdimensi *d*

a = data training

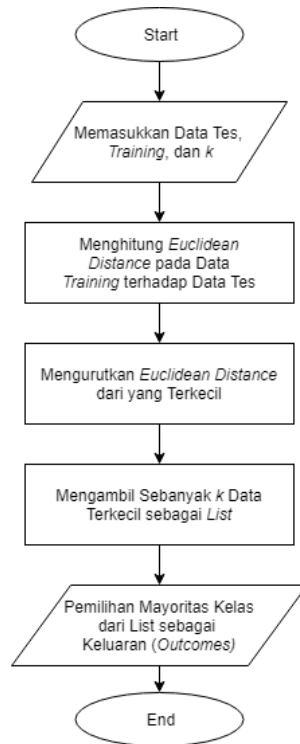
b = data uji

Persamaan 2.4-1 Rumus Perhitungan Jarak

Langkah-langkah untuk dalam membuat algoritma kNN, adalah:

1. Membuat fungsi untuk menentukan kuadrat jarak (*Euclidean Distance*) pada 2 buah objek.
2. Membuat fungsi pemilihan mayoritas (*Majority Votes*) dari sebuah *list*.
3. Mencari kuadrat jarak dengan fungsi *Euclidean distance* dari data terhadap *query instance* kemudian mengurutkannya, kemudian mengambil *k* titik terdekat (*Finding Nearest Neighbor*) sebagai sebuah *list*.
4. Melakukan pemilihan mayoritas keluaran (*outcomes*) dari *k* titik yang terpilih sebagai *list* dengan fungsi *Majority Votes* sebagai hasil prediksi.

Alur proses pelatihan kNN bentuk *FlowChart* ditunjukkan pada gambar 2.4-1



Gambar 3.3.1-1 Flowchart Algoritma kNN

Berikut contoh perhitungan sederhana pada Algoritma K-Nearest Neighbor diketahui 10 buah data yang terbagi kedalam 3 kelompok nilai yang dapat dilihat pada tabel 2.4.1:

Tabel 3.3.1-1 Data Latih Kasus Algoritma kNN

Data	Matkul A	Matkul B	Matkul C	Matkul D	Matkul E	Kelulusan
1	60	80	70	80	90	Tepat
2	70	90	50	70	70	Tepat
3	50	60	80	60	80	Terlambat
4	80	40	90	90	60	Terlambat
5	90	89	76	66	89	Tepat
6	75	75	60	50	99	Tepat
7	94	69	71	40	78	Tepat
8	71	70	94	99	96	Tepat
9	85	50	50	79	77	Terlambat
10	79	99	66	69	75	Tepat

Akan dicari untuk ketepatan waktu kelulusan mahasiswa dimana sebagai data uji, dengan menetapkan nilai $k = 5$. Data uji yang dimasukkan dapat dilihat pada tabel 2.4-2:

Tabel 3.3.1-2 Data Uji Kasus Algoritma kNN

Data	Matkul A	Matkul B	Matkul C	Matkul D	Matkul E	Kelulusan
11	30	10	90	60	80	?

Melakukan perhitungan jarak Euclidean (Query Instance) data uji dengan menggunakan persamaan 2.4-1:

1. Menghitung selisih nilai dari data uji terhadap setiap data latih yang ada. Hasil perhitungan pada tabel 2.4-3.

Tabel 3.3.1-3 Perhitungan Selisih Nilai Data Latih dengan Data Uji

Data	Matkul A	Matkul B	Matkul C	Matkul D	Matkul E	Kelulusan
1	60-30=30	80-10=70	70-90=-20	80-60=20	90-80=10	Tepat
2	70-30=40	90-10=80	50-90=-40	70-60=10	70-80=-10	Tepat
3	50-30=20	60-10=50	80-90=-10	60-60=0	80-80=0	Terlambat
4	80-30=50	40-10=30	90-90=0	90-60=30	60-80=-20	Terlambat
5	90-30=60	89-10=79	76-90=-14	66-60=6	89-80=9	Tepat
6	75-30=45	75-10=65	60-90=-30	50-60=-10	99-80=19	Tepat
7	94-30=64	69-10=59	71-90=-19	40-60=-20	78-80=-2	Tepat
8	71-30=41	70-10=60	94-90=-4	99-60=29	96-80=16	Tepat
9	85-30=55	50-10=40	50-90=-40	79-60=19	77-80=-3	Terlambat
10	79-30=49	99-10=89	66-90=-24	69-60=9	75-80=-5	Tepat

2. Menghitung jarak Euclidean:

Tabel 3.3.1-4 Hasil Perhitungan Euclidean Distance

Data	Perhitungan	Hasil
1	$\sqrt{30^2 + 70^2 + (-20)^2 + 20^2 + 10^2}$	81.85353
2	$\sqrt{40^2 + 80^2 + (-40)^2 + 10^2 + (-10)^2}$	98.99495
3	$\sqrt{20^2 + 50^2 + (-10)^2 + 0^2 + 0^2}$	54.77226
4	$\sqrt{50^2 + 30^2 + 0^2 + 30^2 + (-20)^2}$	68.55655
5	$\sqrt{60^2 + 79^2 + (-14)^2 + 6^2 + 9^2}$	100.7671
6	$\sqrt{45^2 + 65^2 + (-30)^2 + (-10)^2 + 19^2}$	87.24105
7	$\sqrt{64^2 + 59^2 + (-19)^2 + (-20)^2 + (-2)^2}$	91.33455
8	$\sqrt{41^2 + 60^2 + 4^2 + 29^2 + 16^2}$	84.10707
9	$\sqrt{55^2 + 40^2 + (-40)^2 + 19^2 + (-3)^2}$	81.20961

10	$\sqrt{49^2 + 89^2 + (-24)^2 + 9^2 + (-5)^2}$	104.9
----	---	-------

- Melakukan pengurutan data berdasarkan hasil perhitungan *Euclidean Distance* dari yang terkecil dengan $k=5$ pada tabel 2.4-5.

Tabel 3.3.1-1 Hasil Pengurutan berdasarkan Jarak Terdekat *k-Nearest Neighbor*

Urutan	Data	Jarak	Kelulusan
1	3	54.772	Terlambat
2	4	68.556	Terlambat
3	9	81.107	Terlambat
4	1	81.853	Tepat
5	8	84.107	Tepat
6	6	87.241	Tepat
7	7	91.334	Tepat
8	2	98.994	Tepat
9	5	100.767	Tepat
10	10	104.9	Tepat

- Dengan menggunakan $k=5$ (5 data teratas) pada tabel 2.4-5, akan diambil nilai kelulusan dengan frekuensi yang paling tinggi. Dalam hasil pengurutan berdasarkan jarak terdekat *Euclidean Distance* algoritma kNN didapatkan nilai “Terlambat” sebanyak 3 kali, nilai “Tepat” sebanyak 2 kali. Sehingga hasil untuk kasus diatas pada Data Uji menghasilkan nilai Kelulusan = “Terlambat”.

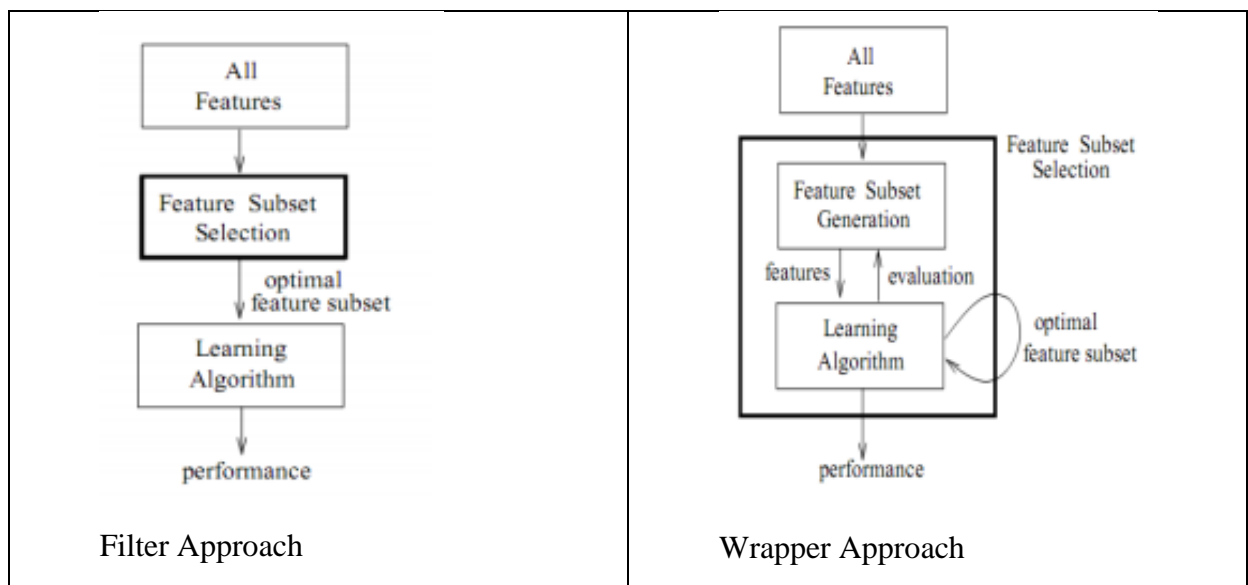
2.5 Feature Selection

Dataset PGK memiliki 24 buah atribut, sehingga dalam *dataset* tersebut dimungkinkan terdapat atribut yang memiliki korelasi kuat terhadap proses identifikasi, adapula atribut yang memiliki korelasi kuat dengan atribut prediktif lainnya, dan ada juga atribut yang tidak memiliki pengaruh sehingga mengurangi akurasi dalam identifikasi PGK. Sehingga seleksi atribut (*Feature Selection*) perlu dilakukan pada tahap *preprocessing* sebelum dilakukan tahap *data mining*.

Seleksi atribut merupakan proses untuk menghasilkan atribut dengan korelasi yang paling berpengaruh yang memudahkan dalam menganalisa dan menginterpretasikan hasil pemodelan *data mining* (Elkan, 2010). Dibandingkan dengan pendekatan *Brute Force* yang mengambil seluruh atribut yang ada, seleksi atribut juga dapat mempercepat proses

komputasi dalam pemodelan *data mining*, dikarenakan telah mengurangi jumlah atribut yang harus dikomputasi.

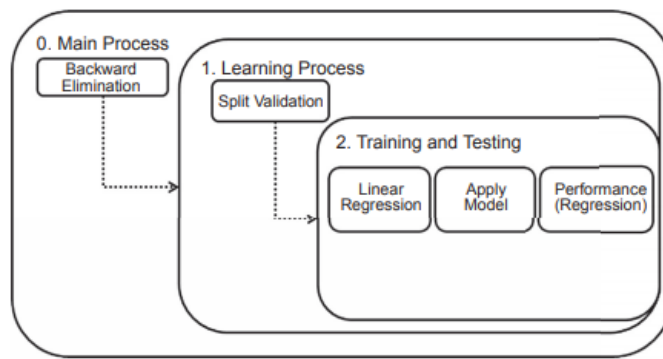
Terdapat dua teknik pendekatan dalam seleksi atribut: 1) *Filter Approach*, dan 2) *Wrapper Approach*. *Filter Approach* menilai relevansi dengan melihat sifat – sifat intrinsik data. Semua atribut diberi skor dan peringkat berdasarkan kriteria tertentu, beberapa atribut dengan peringkat tertinggi dipilih, dan atribut dengan skor rendah akan dihapus. *Wrapper Approach* merupakan teknik dengan mengevaluasi dan menguji model klasifikasi. Teknik ini secara iteratif menambah atau mengurangi atribut dari atribut sebelumnya untuk meningkatkan akurasi (Elkan, 2010).



Gambar 3.3.1-1 Teknik Seleksi Atribut

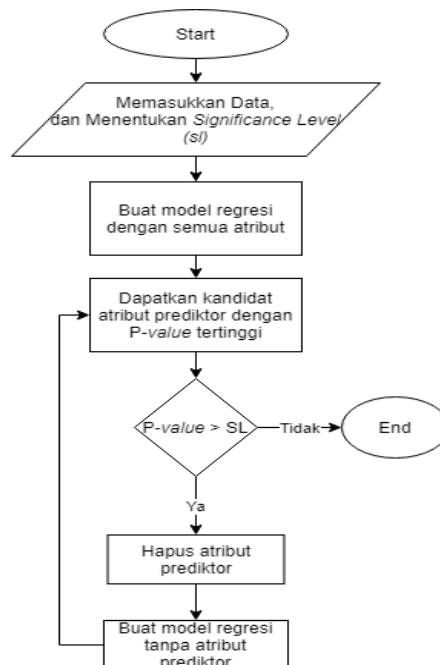
2.6 Algoritma Backward Elimination

Algoritma *Backward Elimination* merupakan salah satu algoritma dalam metode seleksi atribut untuk mengurangi ukuran *dataset*. Algoritma *Backward Elimination* menggunakan teknik *wrapper approach* yang didasarkan pada model regresi linear (Noori *et al.*, 2011). Kelebihan *wrapper approach* adalah memiliki interaksi dengan kelas(target), sehingga menghasilkan akurasi klasifikasi yang baik (Kumari dan Swarnkar, 2011). Pada pengerjaannya, algoritma *backward elimination* dibantu dengan tools datamining bernama *RapidMiner*, dan bagaimana operator tersebut mengikuti logic *wrapper approach* ditunjukkan pada gambar 2.6-1.



Gambar 3.3.1-1 *Logic Function* dari Teknik *Wrapper Approach* pada Operator *Backward Elimination* dengan RapidMiner

Backward Elimination dimulai oleh model yang memiliki semua potensial X atribut prediktif yang dicek model regresinya, kemudian diidentifikasi salah satu atribut prediktif yang memiliki nilai P -value terbesar, jika P -value terbesar tersebut lebih besar dari batas derajat (*significance level / predetermined limit*) yang telah ditentukan sebelumnya, maka atribut X tersebut dihilangkan. Kemudian gunakan model dengan sisa $P-2X$ atribut prediktif yang tersisa untuk kembali dicek model regresinya, dan kandidat atribut selanjutnya yang akan dihilangkan sesuai dengan yang sudah dijelaskan sebelumnya. Proses ini berlanjut hingga tidak ada lagi atribut X yang dapat dihilangkan. (Kutner *et al.*, 2004).

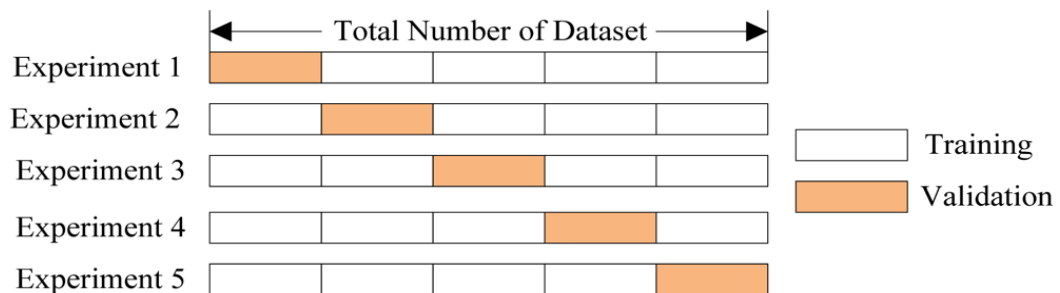


Gambar 3.3.1-2 *Flowchart* proses *Backward Elimination*

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.

2.7 k-Fold Cross Validation

k-Fold Cross Validation adalah teknik validasi dengan membagi data secara acak ke beberapa bagian, dan masing – masing bagian akan dilakukan proses klasifikasi. *k-Fold Cross Validation* melakukan iterasi sebanyak k kali untuk data pelatihan dan pengujian. Metode *k-Fold Cross Validation* berguna untuk memvalidasi akurasi sebuah prediksi atau klasifikasi terhadap data yang belum muncul dalam *dataset*. *Dataset* tersebut dibagi menjadi *k-subset* secara acak yang masing-masing *subset* memiliki jumlah *instance* pada proses iterasi klasifikasi (Han dan Kamber, 2011).



Gambar 3.3.1-1 Ilustrasi *k-Fold Cross Validation*

Kelebihan dari metode ini adalah tidak adanya masalah dalam pembagian data. Setiap data akan menjadi *test set* sebanyak satu kali dan akan menjadi *training set* sebanyak $k-1$ kali. Namun metode ini membuat pembelajaran yang dilakukan sebanyak k kali. Dimana menggunakan k kali waktu komputasi. Nilai k yang paling baik digunakan dalam penelitian menurut Kohavi adalah 10 jika dilihat dari variasi data dan bias yang dimiliki (Kohavi, 1995).

2.8 Evaluasi Sistem

Evaluasi sistem dilakukan dengan pengecekan hasil dari metode dengan menggunakan *confusion matrix*. *Confusion matrix* merupakan sebuah metode yang digunakan untuk mengetahui seberapa baik sebuah metode klasifikasi mengenali tuple dari kelas yang berbeda (Han dan Kamber, 2011). *Confusion matrix* merupakan perhitungan *Predicted Class* dan *Actual Class* pada gambar 2.6-1, dengan ketentuan tertentu, perhitungan yang dimaksud dapat meliputi *recall*, *precision*, *accuracy*, dan *error rate*.

Tabel 3.3.1-1 *Confusion Matrix*

		Predicted class	
		C_1	C_2
Actual class	C_1	true positives	false negatives
	C_2	false positives	true negatives

Dalam penelitian ini digunakan dua keluaran yaitu *sensitivity* dan *specificity* (proporsi kasus negatif yang diidentifikasi dengan benar) yang merupakan dasar dari perhitungan akurasi pada bidang kesehatan (Zhang *et al.*, 2008). Berikut perhitungan *sensitivity* (persamaan 2.8-1) dan *specificity* (persamaan 2.8-2) jika telah didapatkan *confusion matrix*:

$$\textbf{Sensitivity} = \frac{\textbf{True Positive}}{\textbf{True Positive} + \textbf{False Negative}}$$

Persamaan 2.8-1 Perhitungan *Sensitivity*

$$\textbf{Specificity} = \frac{\textbf{True Negative}}{\textbf{True Negative} + \textbf{False Positive}}$$

Persamaan 2.8-2 Perhitungan *Specifity*

BAB III

METODE PENELITIAN

Bab ini menjelaskan mengenai metode yang digunakan dalam pengambilan data, lokasi penelitian, arsitektur sistem, dan garis besar penyelesaian masalah dalam penyusunan Tugas Akhir.

3.1 Metode Penelitian

Metode penelitian yang digunakan dalam pengerjaan Tugas Akhir ini adalah studi pustaka dan eksperimental.

1. Studi Pustaka

Studi pustaka merupakan metodologi yang digunakan untuk menyusun Tugas Akhir ini. Penulis melakukan pengumpulan literature dan pembelajaran literature yang terkait dalam penyusunan tugas akhir ini seperti buku, jurnal, maupun artikel yang dapat digunakan untuk mengatasi permasalahan yang dihadapi dalam Tugas Akhir ini.

2. Eksperimental

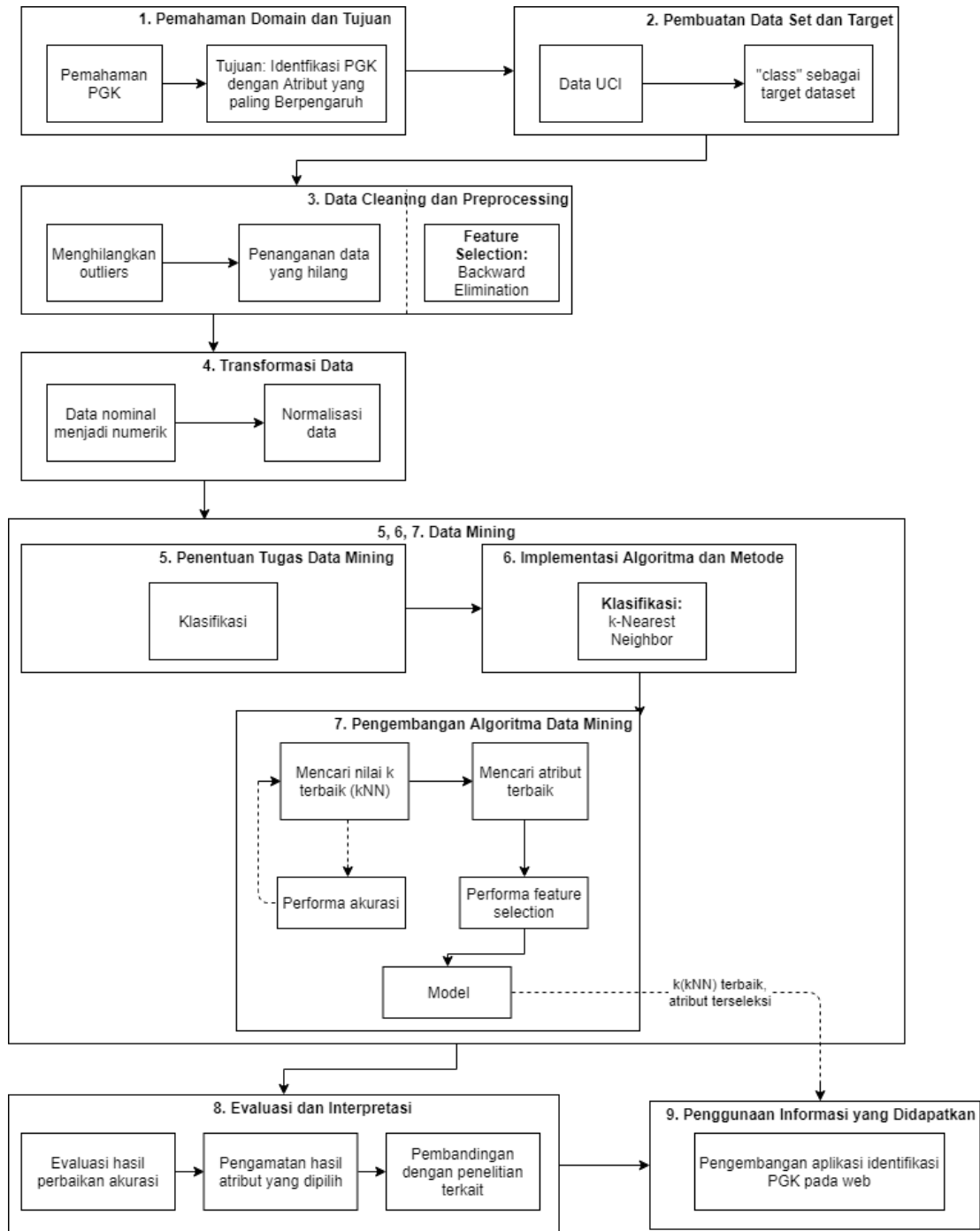
Eksperimental dilakukan untuk mengungkapkan hubungan sebab akibat dua variabel atau lebih dengan memperhatikan pengaruh variabel lainnya. Metode ini dilaksanakan dengan memberikan variabel bebas secara sengaja kepada objek penelitian untuk diketahui akibatnya di dalam variabel terikat. Dalam tugas akhir ini variabel bebas yang digunakan adalah data training, sedangkan variabel terikat yang digunakan adalah akurasi dari hasil penelitian.

3.2 Lokasi Penelitian

Lokasi penelitian yang digunakan dalam penyusunan tugas akhir ini bertempat di Departemen Ilmu Komputer/ Informatika Fakultas Sains dan Matematika Universitas Diponegoro Semarang.

3.3 Garis Besar Penyelesaian Masalah

Pada bagian ini menjelaskan mengenai bagaimana kerangka kerja penelitian Tugas Akhir ini berjalan dari awal hingga akhir. Kerangka kerja ini digambarkan pada proses gambar 3.3-1 dan disesuaikan dengan metode KDD pada pemodelan *Data Mining*.



Gambar 3.3.1-1 Kerangka Kerja Penelitian

3.3.1 Integrasi dan Pemahaman Data

Tahap pemahaman terhadap data yang diteliti. Setelah data didapatkan, maka akan dilakukan pembelajaran mengenai data yang digunakan. Dengan harapan dari pembelajaran tersebut, data dapat dikenali lebih lanjut. Tahap ini bertujuan untuk membiasakan diri dengan data yang dikerjakan dan menemukan wawasan awal mengenai informasi apa saja yang bisa didapatkan didalamnya.

Data didapatkan dari UCI *Machine Learning Repository*. Data memiliki 25 atribut, terdiri dari 1 kelas target dan 24 atribut seperti yang disajikan pada tabel 3.3.1-1.

Tabel 3.3.1-1 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 Gravity	Berat Jenis	Nominal (1.005, 1.010, 1.015, 1.020, 1.025)
4	al	Albumin	Albumin	Nominal (0, 1, 2, 3, 4, 5)
5	su	Sugar	Gula	Nominal (0, 1, 2, 3, 4, 5)
6	rbc	Red Blood Cells	Sel Darah Merah	Nominal (normal, abnormal)
7	pc	Pus Cell	Sel Darah Putih	Nominal (normal, abnormal)
8	pcc	Pus Cell Clumps	Gumpalan Sel Nanah	Nominal (present, notpresent)
9	ba	Bacteria	Bakteri	Nominal (present, notpresent)
10	bgr	Blood Glucose Random	Gula Darah Acak	Numerik (mgs/dl)
11	Bu	Blood Urea	Urea Darah	Numerik (mgs/dl)
12	Sc	Serum Creatinine	Kreatinin Serum	Numerik (mgs/dl)
13	sod	Sodium	Sodium	Numerik (mEq/L)
14	pot	Potassium	Potassium	Numerik (mEq/L)
15	hemo	Hemoglobin	Hemoglobin	Numerik (gms)
16	Pcv	Packed Cell Volume / Hematocrit	Hematokrit	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	Hipertensi	Nominal (yes, no)

20	Dm	Diabetes Mellitus	Diabetes Mellitus	Nominal (yes, no)
21	cad	Coronary Artery Disease	Penyakit Jantung Koroner	Nominal (yes, no)
22	appet	Appetite	Selera Makan	Nominal (good, poor)
23	Pe	Pedal Edema	Pembengkakan pada Kaki	Nominal (yes, no)
24	ane	Anemia	Anemia	Nominal (yes, no)
25	class	Class	Kelas (Variabel Terikat)	Nominal (ckd, notckd)

Terdapat beberapa informasi yang didapat dari pendalaman pemahaman atribut yang ada (Salekin dan Stankovic, 2016), antara lain:

- Diabetes Mellitus (dm):
Berdasarkan *National Kidney Foundation* (National Kidney Foundation, 2015), 1 dari 3 orang yang terkena diabetes memiliki kemungkinan teridentifikasi PGK.
- Sodium (sod) dan Potassium (pot):
Merupakan zat yang penting bagi tubuh namun berbahaya jika berlebihan, penderita PGK tidak dapat menghilangkan kelebihan Sodium, Potassium, dan cairan lainnya dalam tubuh.
- Pembengkakan pada kaki (pe):
Edema adalah istilah kedokteran dari pembengkakan. Edema terjadi ketika pembuluh darah kecil pecah dan melepaskan cairan ke jaringan di dekatnya. Cairan yang terakumulasi tersebut menjadikan pembengkakan.
- Sel darah putih (pc):
Jika terdeteksi sel darah putih pada urin merupakan indikasi infeksi PGK.
- Serum Kreatinin (sc):
Atribut ini berpengaruh terhadap filtrasi glomerulus pada ginjal.

3.3.2 Pembuatan Dataset dan Target

Target pada atribut “class”, telah dikelompokkan menjadi 2 output, PGK (ckd) atau Normal (notckd).

3.3.3 Data Cleaning dan Pre-Processing

Kegiatan yang ada pada tahap ini antara lain membersihkan dan memperbaiki data yang rusak; menghilangkan noise, yaitu menghapus data atau atribut yang tidak diperlukan, serta menyeragamkan data yang dianggap sama namun memiliki nilai yang berbeda atau membuatnya menjadi konsisten.

Seleksi atribut dilakukan untuk menghilangkan atribut yang dianggap tidak relevan dengan menggunakan Algoritma *Backward Elimination*. Dimana algoritma *Backward Elimination* akan mengeluarkan satu per satu atribut *predictor* yang tidak signifikan dan akan dilakukan terus menerus hingga tidak ada atribut *predictor* yang tidak signifikan. Tahap seleksi atribut ini dilakukan secara iteratif, yang akan dijelaskan lebih lanjut pada bagian implementasi.

3.3.4 Transformasi Data

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 algoritma k-NN yang mengimplementasikan rumus *Euclidean Distance*, dapat dicontohkan dengan mengubah nilai data yang dimasukkan harus berupa numerik (bilangan riil), sehingga atribut nominal yang bernilai binomial atau seperti: “Ya”-“Tidak”, “Ada”-“Tidak Ada”, “Normal”-“Abnormal” sehingga dapat ditransformasi menjadi 1 dan 0.

3.3.5 Data Mining: Pemilihan Jenis Tugas Data Mining

Jenis tugas data mining yang dipilih adalah *Classification* (Klasifikasi), didefinisikan sebagai *supervised learning*, dimana telah terdapat informasi mengenai bagaimana data tersebut dikelompokkan dan tidak ada penambahan kelompok.

3.3.6 Data Mining: Penentuan Algoritma dan Metode

Algoritma yang digunakan pada dataset PGK menggunakan algoritma *k*-Nearest Neighbor (kNN) dengan *Backward Elimination*. kNN diimplementasikan dengan rumus *Euclidean Distance* dikarenakan *dataset* yang digunakan sebagian besar adalah numerik. Sedangkan algoritma *Backward Elimination* bertujuan agar hasil dari proses *data mining* ini dapat ditafsirkan secara lebih baik.

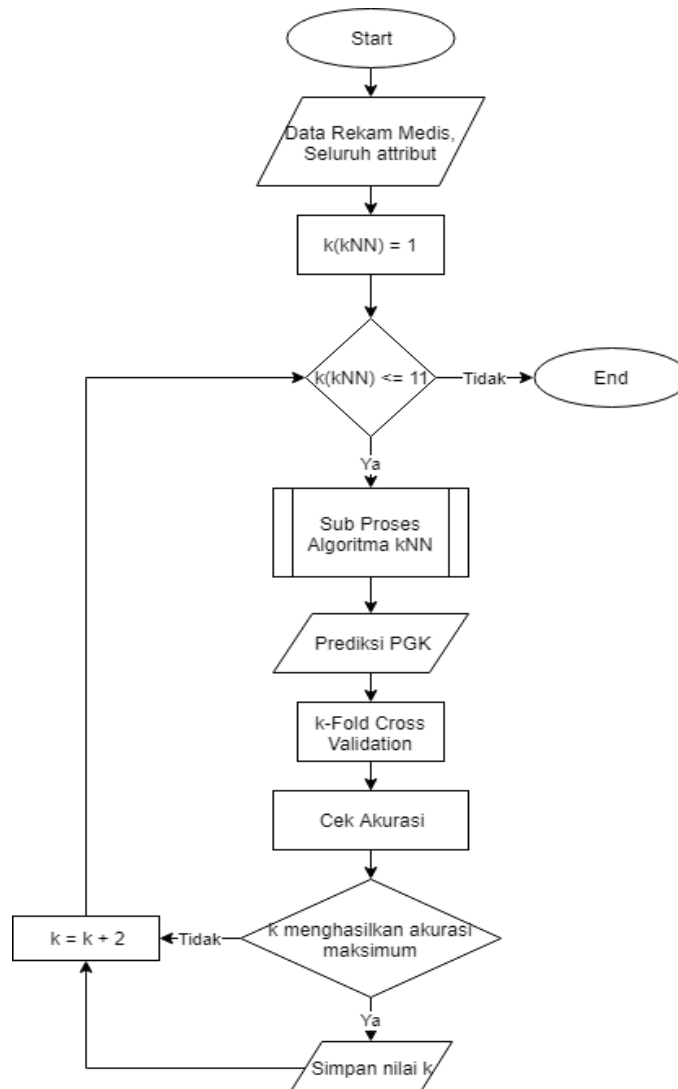
3.3.7 Data Mining: Implementasi Algoritma Data Mining

Tahap ini merupakan detail implementasi yang akan dilakukan dengan algoritma kNN dan *Backward Elimination*, pada tahap ini terdapat 2 proses utama:

1. Penentuan nilai k terbaik pada *k*-Nearest Neighbor

Dilakukan *training* dengan algoritma kNN kepada data yang sudah diberikan label kelas. Pada tahap ini dilakukan secara iteratif dan eksperimental untuk

mendapatkan nilai k yang memiliki akurasi terbaik. Flowchart iterasi yang bertujuan mendapatkan nilai k terbaik dari algoritma kNN ditunjukkan pada Gambar 3.3.7-1.



Gambar 3.3.7-1 Flowchart Penentuan Nilai k - pada kNN

Sub-proses algoritma kNN telah ditunjukkan pada gambar 2.4-1, diimplementasikan pada pada setiap iterasi dengan nilai k ganjil. Hal ini bertujuan agar hasil *Majority Votes* pada fungsi kNN mendapatkan nilai yang tidak berubah - ubah. Jika nilai k genap, maka penentuan kelas pada fungsi *Majority Votes* akan bersifat random, sehingga akurasi yang didapat setiap k genap akan tidak akurat dan selalu berubah – ubah.

2. Seleksi atribut algoritma *Backward Elimination*

Algoritma ini merupakan bagian dari metode *Stepwise Regression* dalam membangun model atribut terbaik, seperti yang telah di ilustrasikan pada gambar 2.5-1, metode ini melakukan evaluasi secara iteratif dengan model awal seluruh atribut dan dikurangi satu per satu secara iteratif hingga mendapatkan beberapa atribut yang paling berpengaruh (Kirill Eremenko, 2017).

Dalam iterasi yang dilakukan, hasil atribut yang didapatkan akan dikembalikan pada tahap *pre-processing*. Sehingga ketika kondisi iterasi terpenuhi dan berhenti, atribut yang tersisa pada iterasi terakhir merupakan atribut terbaik beserta mendapatkan nilai performa dari atribut yang terpilih.

3.3.8 Evaluasi dan Interpretasi

Dilakukannya evaluasi dan interpretasi terhadap hasil *data mining* yang telah dilakukan. Evaluasi sistem untuk memastikan apakah hasil yang didapat tidak terdapat kesalahan pada pengerjaan dan telah sesuai dengan tujuan awal yang telah dibuat dan membandingkan hasilnya dengan penelitian terkait. Setelah melakukan proses evaluasi, Interpretasi yang dilakukan adalah mendapatkan nilai k terbaik pada algoritma kNN dalam identifikasi PGK dan atribut – atribut terseleksi yang dianggap berpengaruh kuat pada identifikasi PGK.

3.3.9 Penggunaan Informasi yang Didapatkan

Merupakan tahap terakhir dimana pemodelan yang didapat berupa k terbaik pada algoritma kNN dan atribut terseleksi pada algoritma *Backward Elimination* dikembangkan sebagai sebuah aplikasi berbasis web dalam mengidentifikasi PGK dengan bahasa pemrograman utama Python.

3.4 Analisis dan Desain Sistem

Analisis dan desain sistem menjelaskan tentang deskripsi dari sistem, pemodelan analisis, dan perancangan sistem dari implementasi algoritma k-*Nearest Neighbor* (kNN) dengan *Backward Elimination* untuk identifikasi Penyakit Ginjal Kronis.

3.4.1 Deskripsi Sistem

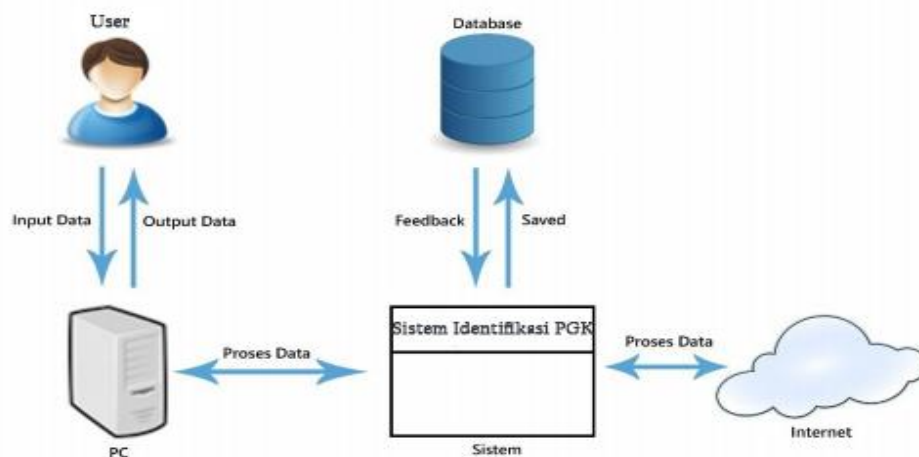
Deskripsi sistem menjelaskan mengenai gambaran Implementasi Algoritma kNN dengan *Backward Elimination* untuk identifikasi Penyakit Ginjal Kronis (PGK) yang

meliputi deskripsi umum sistem, kebutuhan fungsional sistem, dan kebutuhan non fungsional sistem.

3.4.2 Deskripsi Umum Sistem

Implementasi Algoritma *kNN* dengan *Backward Elimination* untuk identifikasi Penyakit Ginjal Kronis bekerja dengan membandingkan kinerja *kNN* dengan seluruh atribut sebagai model dan *kNN* dengan atribut yang telah diseleksi oleh *Backward Elimination*. Sistem ini nantinya dapat digunakan oleh tenaga medis dalam melakukan identifikasi PGK.

Sistem ini memiliki satu jenis pengguna (*user*). *User* dapat melakukan proses kelola data rekam medis PGK, melakukan identifikasi PGK dengan menggunakan algoritma *kNN* dan algoritma *kNN* dengan *Backward Elimination*, lalu melakukan pengujian dalam identifikasi PGK, dan seleksi atribut yang digunakan pada algoritma *Backward Elimination*. Arsitektur implementasi algoritma *kNN* dengan *Backward elimination* dapat dilihat pada gambar 3.4.2-1.



Gambar 3.4.2-1 Arsitektur Sistem Identifikasi Penyakit Ginjal Kronis

Sistem ini memiliki satu jenis pengguna (*user*). *User* dapat melakukan proses kelola data rekam medis PGK, melakukan identifikasi PGK dengan menggunakan algoritma *kNN* dan algoritma *kNN* dengan *Backward Elimination*, lalu melakukan pengujian dalam identifikasi PGK, dan seleksi atribut yang digunakan pada algoritma *Backward Elimination*. Arsitektur implementasi algoritma *kNN* dengan *Backward elimination* dapat dilihat pada gambar 3.4-1.

Adapun penjelasan dari gambar 3.4-1 adalah sebagai berikut:

- *User* : Pengguna sistem ini adalah dokter maupun tenaga medis penyakit ginjal kronis. User dapat melakukan *input* berupa data pasien baru. Dan menghasilkan *output* berupa diagnosis penyakit diabetes.
- *Database* : Data yang disimpan merupakan data awal repository dan dapat ditambahkan dengan data pasien yang telah terdiagnosis PGK.
- *PC* : *Personal Computer* sebagai media interaksi proses data antara user terhadap sistem identifikasi PGK.
- *Sistem* : Sistem identifikasi PGK yang dibuat menggunakan kNN dan kNN dengan *Backward Elimination*.
- *Internet* : Sistem identifikasi PGK terhubung dengan koneksi internet sehingga dapat diakses kapan saja dan dimana saja.
- *Input* : *Input* berisi data yang akan dimasukkan oleh *User* ke dalam sistem
- *Output* : Data yang ditampilkan setelah *User* melakukan *input* pada sistem.
- *Feedback* : Merupakan proses pengambilan data dari *database* yang telah ada untuk dilakukan pemrosesan.
- *Saved* : *Saved* merupakan proses menyimpan data jika telah terjadi perubahan atau ada data baru yang akan disimpan.

3.4.3 Kebutuhan Fungsional Sistem

Kebutuhan fungsional didefinisikan melalui spesifikasi *Software Requirement Specification* (SRS) aplikasi *data mining* dalam identifikasi PGK dilihat pada tabel 3.4.3-1

Tabel 3.4.3-1 Kebutuhan Fungsional

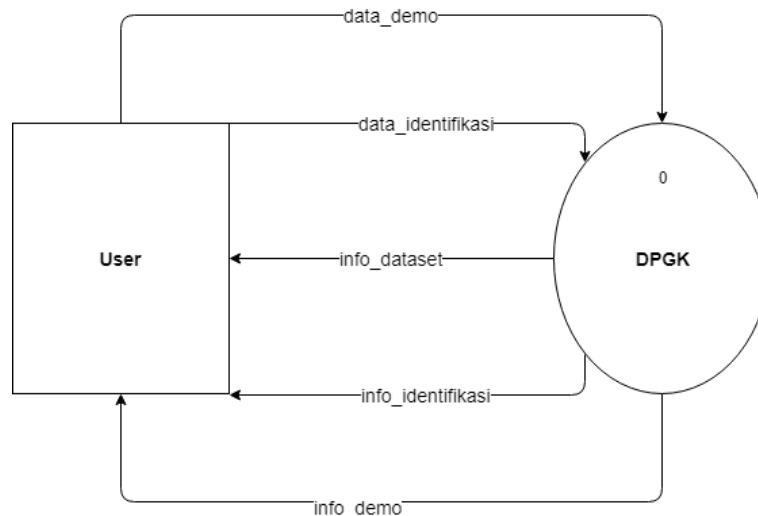
No	SRS ID	Deskripsi
1	SRSF-PGK-01	Sistem dapat menampilkan proses data mining pada tiap tahapan proses
2	SRSF-PGK-02	Sistem dapat melakukan identifikasi PGK
3	SRSF-PGK-03	Sistem dapat menampilkan performa pemodelan data mining yang dibuat

3.4.4 Pemodelan Fungsi

Pemodelan fungsional berguna untuk menggambarkan aspek dari sistem yang berhubungan dengan transformasi dari nilai, fungsi, pemetaan, dan batasan. Model fungsional Implementasi Algoritma *k*-Nearest Neighbor dengan *Backward Elimination* dalam mendiagnosis Penyakit Ginjal Kronis (PGK) digambarkan dalam *Data Flow Diagram*

(DFD). DFD menunjukkan apa yang dapat dikerjakan oleh sistem. Pada identifikasi Penyakit Ginjal Kronis terdapat 2 level DFD, yang meliputi DFD level 0 dan DFD level 1.

DFD level 0 disebut juga sebagai *Data Context Diagram* (DCD), berfungsi untuk menampilkan gambaran umum dari informasi yang terkandung dalam sebuah sistem. DCD digambarkan dengan sebuah symbol proses yang mirip seperti *black box* yang berinteraksi dengan entitas luar melalui *data flow*. Pada sistem identifikasi PGK, memiliki satu entitas yaitu *user* seperti yang ditunjukkan pada gambar 3.4.4-1.



Gambar 3.4.4-1 DFD level 0

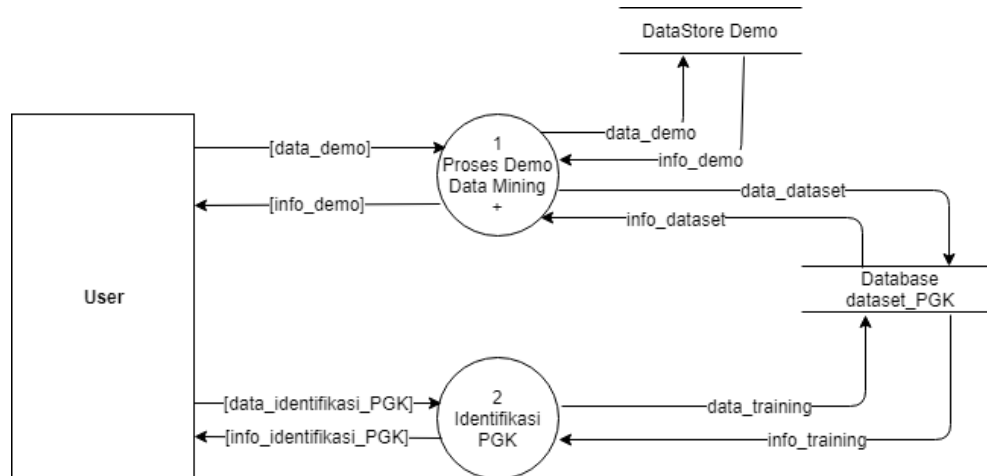
Sedangkan untuk DFD level 1 sistem ditunjukkan seperti pada Gambar 3.4.4-2. Pada DFD level 1 tersebut memiliki 2 proses, yaitu:

1. *Demo Data Mining*

Proses ini menerima inputan data yang akan dilakukan identifikasi beserta menampilkan pilihan tahap-tahap kelola proses *data mining* secara manual. *User* memasukkan data uji dengan pilihan atribut yang akan digunakan, kemudian digunakan data latih dari *database*, lalu memunculkan tahapan proses *data mining* yang akan digunakan, dengan beberapa pilihan yang dapat digunakan dalam identifikasi maupun pilihan validasi dalam proses identifikasi tersebut. Hasil akhir proses demo *data mining* yaitu menampilkan bentuk akhir dari data yang telah dimasukkan oleh *user*, informasi hasil identifikasi, serta informasi validasi performa dari tahapan *data mining* yang telah *user* pilih selama melakukan demo *data mining*.

2. Identifikasi PGK

Pada sub proses ini, pengguna dapat melakukan identifikasi status penyakit ginjal kronis secara otomatis dengan memasukkan data-data pasien sesuai atribut terbaik dan dengan metode terbaik menurut hasil penelitian ini.



Gambar 3.4.4-2 DFD level 1

BAB IV

HASIL DAN PEMBAHASAN

Bab hasil dan pembahasan menjelaskan tentang hasil pengembangan sistem, skenario pengujian sistem, dan hasil dan analisa sistem yang akan dilakukan pada Implementasi Data Mining untuk Identifikasi Penyakit Ginjal Kronis (PGK) Menggunakan K-Nearest Neighbor (KNN) dengan Backward Elimination.

4.1 Hasil Pengembangan Sistem

Hasil pengembangan sistem menyajikan mengenai lingkungan implementasi, implementasi data, implementasi fungsi, dan implementasi antarmuka.

4.1.1 Lingkungan Implementasi

Pada penelitian penerapan data mining untuk identifikasi penyakit ginjal kronis dengan menggunakan kNN dan Backward Elimination dikembangkan pada perangkat keras dengan spesifikasi berikut:

1. *Processor* : Intel® Core-i5-2450M @2.5GHz
2. *RAM* : 6 GB
3. *Harddisk* : 500 GB

dan lingkungan perangkat lunak sebagai berikut:

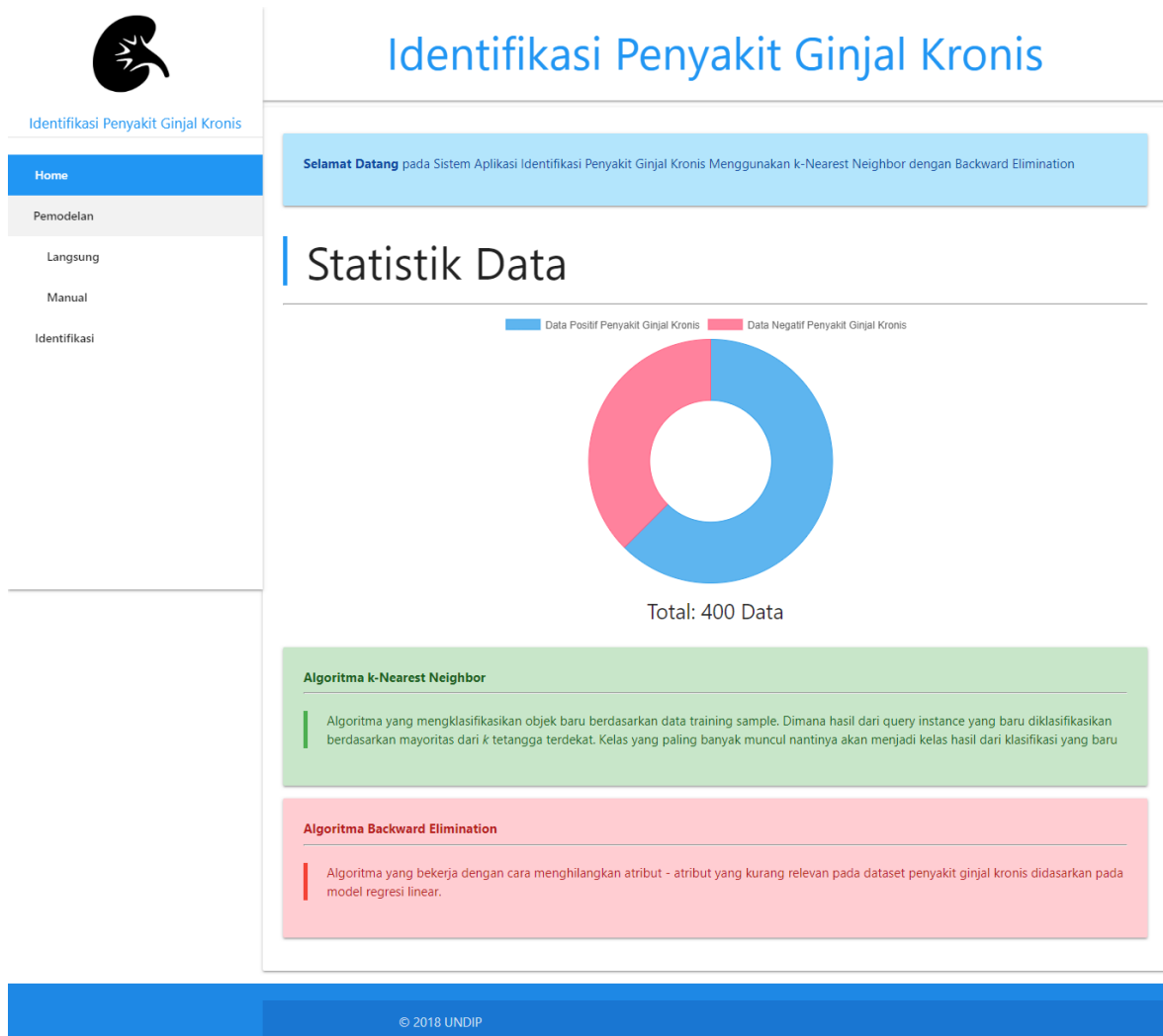
1. OS Windows 10
2. IDE PyCharm
3. Python versi 3
4. DBMS MySQL
5. *Browser* Google Chrome

4.1.2 Implementasi Antarmuka

Implementasi antarmuka adalah hasil perancangan desain yang telah disesuaikan dengan fitur-fitur yang dituliskan pada SRS.

1. Antarmuka *Homepage*

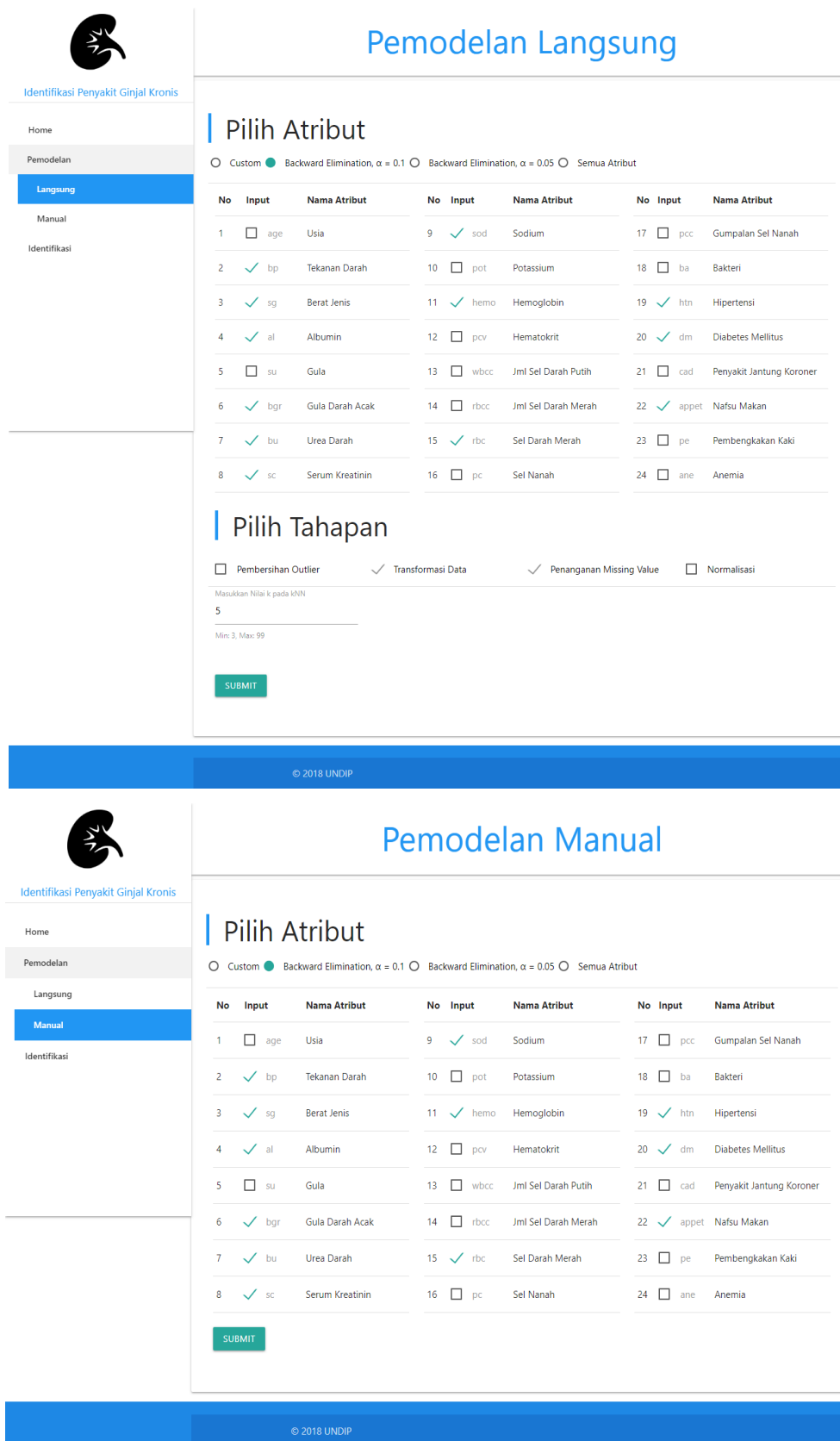
Antarmuka *Homepage* adalah halaman yang pertama kali ditampilkan ketika *user* mengakses sistem. Antarmuka halaman ini menampilkan statistik data, penjelasan mengenai gambaran umum sistem dan algoritma yang digunakan. Antarmuka halaman *Homepage* dilihat pada Gambar 4.1.2-1.



Gambar 4.1.2-1 Antarmuka halaman *Homepage*

2. Antarmuka halaman *Pemodelan*

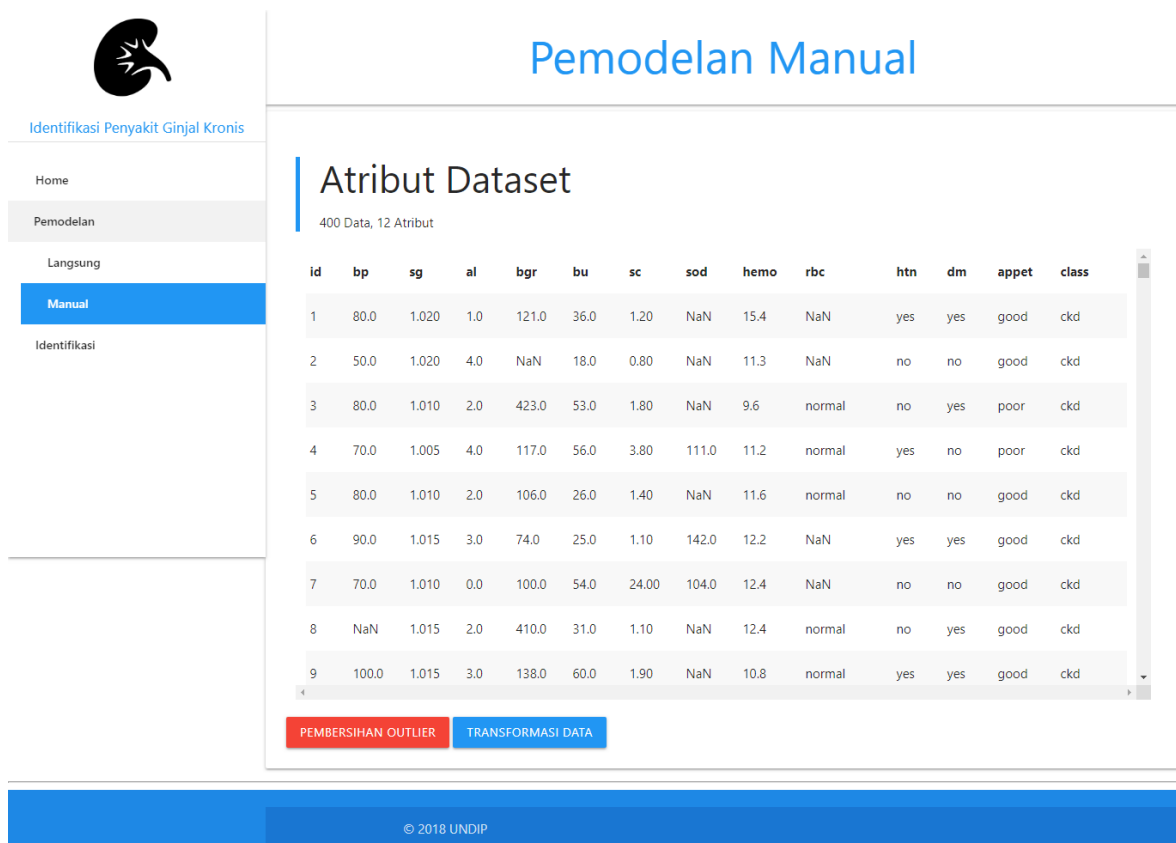
Antarmuka halaman *pemodelan* merupakan sebuah halaman untuk melakukan pengaturan model tahapan *data mining* yang akan digunakan dalam identifikasi PGK. Inputan memiliki 2 pilihan, 1) *Pemodelan Langsung*, yaitu memasukkan semua atribut yang akan digunakan dan tahapan data secara langsung; 2) *Pemodelan Manual*, menampilkan hasil terlebih dahulu dari tiap tahapan yang dipilih dari proses *data mining* yang dipilih hingga proses validasi. Perbandingan antarmuka halaman *Pemodelan Langsung*, dan *Pemodelan Manual* dapat dilihat pada Gambar 4.1.2-2.



Gambar 4.1.2-2 Perbandingan antarmuka *Pemodelan Langsung* (atas) dan *Pemodelan Manual* (bawah)

3. Antarmuka halaman *Dataset* pada proses *Pemodelan Manual*

Antarmuka halaman *dataset* dimunculkan ketika *user* telah memasukkan atribut yang akan digunakan. Halaman ini memunculkan *dataset* dengan atribut yang sesuai dengan masukkan *user* secara dinamis. Halaman ini memiliki tombol pilihan tahapan selanjutnya menggunakan “penanganan outlier” atau “transformasi data”. Antarmuka halaman *Dataset* pada *Pemodelan Manual* dapat dilihat pada Gambar 4.1.2-3




The screenshot shows the 'Pemodelan Manual' web application. The sidebar on the left has a logo and navigation links: 'Identifikasi Penyakit Ginjal Kronis', 'Home', 'Pemodelan', 'Langsung', 'Manual' (highlighted), and 'Identifikasi'. The main content area is titled 'Pemodelan Manual' and 'Atribut Dataset'. Below the title, it says '400 Data, 12 Atribut'. A table displays the dataset attributes and values for 9 rows. At the bottom of the table, there are two buttons: 'PEMBERSIHAN OUTLIER' (red) and 'TRANSFORMASI DATA' (blue). The footer of the application shows '© 2018 UNDIP'.

id	bp	sg	al	bgr	bu	sc	sod	hemo	rbc	htn	dm	appet	class
1	80.0	1.020	1.0	121.0	36.0	1.20	NaN	15.4	NaN	yes	yes	good	ckd
2	50.0	1.020	4.0	NaN	18.0	0.80	NaN	11.3	NaN	no	no	good	ckd
3	80.0	1.010	2.0	423.0	53.0	1.80	NaN	9.6	normal	no	yes	poor	ckd
4	70.0	1.005	4.0	117.0	56.0	3.80	111.0	11.2	normal	yes	no	poor	ckd
5	80.0	1.010	2.0	106.0	26.0	1.40	NaN	11.6	normal	no	no	good	ckd
6	90.0	1.015	3.0	74.0	25.0	1.10	142.0	12.2	NaN	yes	yes	good	ckd
7	70.0	1.010	0.0	100.0	54.0	24.00	104.0	12.4	NaN	no	no	good	ckd
8	NaN	1.015	2.0	410.0	31.0	1.10	NaN	12.4	normal	no	yes	good	ckd
9	100.0	1.015	3.0	138.0	60.0	1.90	NaN	10.8	normal	yes	yes	good	ckd

Gambar 4.1.2-3 Antarmuka halaman *Dataset* pada proses *Pemodelan Manual*

4. Antarmuka halaman *Pembersihan Outlier* pada proses *Pemodelan Manual*

Antarmuka halaman *Pembersihan Outlier* dimunculkan ketika *user* menekan tombol pilihan “pembersihan outlier” pada halaman *Dataset*. Halaman ini memunculkan kembali *Dataset* yang telah diproses untuk menghilangkan data-data yang dianggap memiliki *outlier*. Halaman ini memiliki sebuah tombol untuk melanjutkan dataset ke tahap transformasi data. Antarmuka halaman *Penanganan Outlier* pada *Pemodelan* dapat dilihat pada Gambar 4.1.2-4



Identifikasi Penyakit Ginjal Kronis

- Home
- Pemodelan
- Langsung
- Manual**
- Identifikasi

Pemodelan Manual

Pembersihan Outlier

285 Data, 12 Atribut

id	bp	sg	al	bgr	bu	sc	sod	hemo	rbc	htn	dm	appet	class
1	80.0	1.020	1.0	121.0	36.0	1.20	NaN	15.4	NaN	yes	yes	good	ckd
5	80.0	1.010	2.0	106.0	26.0	1.40	NaN	11.6	normal	no	no	good	ckd
6	90.0	1.015	3.0	74.0	25.0	1.10	142.0	12.2	NaN	yes	yes	good	ckd
13	70.0	1.015	3.0	208.0	72.0	2.10	138.0	9.7	NaN	yes	yes	poor	ckd
14	70.0	NaN	NaN	98.0	86.0	4.60	135.0	9.8	NaN	yes	yes	poor	ckd
15	80.0	1.010	3.0	157.0	90.0	4.10	130.0	5.6	normal	yes	yes	poor	ckd
17	70.0	1.015	2.0	99.0	46.0	2.20	138.0	12.6	NaN	no	no	good	ckd
18	80.0	NaN	NaN	114.0	87.0	5.20	139.0	12.1	NaN	yes	no	poor	ckd
20	60.0	1.015	1.0	100.0	31.0	1.60	NaN	10.3	NaN	yes	no	good	ckd

TRANSFORMASI DATA

© 2018 UNIDIP

Gambar 4.1.2-4 Antarmuka Pembersihan Outlier pada proses Pemodelan Manual

- Antarmuka halaman Transformasi Data Nominal pada proses Pemodelan Manual
Antarmuka halaman Transformasi Data dimunculkan ketika *user* menekan tombol pilihan “Transformasi Data” pada halaman *dataset* atau pada halaman pembersihan *outlier*. Halaman ini memunculkan kembali *dataset* yang telah diproses untuk mengubah atribut-atribut yang memiliki nilai nominal untuk ditransformasi menjadi data numerik. Halaman ini memiliki sebuah tombol untuk melanjutkan dataset ke tahap penanganan *missing value*. Antarmuka halaman transformasi data pada proses Pemodelan Manual dapat dilihat pada Gambar 4.1.2-5

Pemodelan Manual

Identifikasi Penyakit Ginjal Kronis

Home

Pemodelan

Langsung

Manual

Identifikasi

Transformasi Data

285 Data, 12 Atribut

id	bp	sg	al	bgr	bu	sc	sod	hemo	rbc	htn	dm	appet	class
1	80.0	1.020	1.0	121.0	36.0	1.20	NaN	15.4	NaN	0.0	0.0	0.0	ckd
5	80.0	1.010	2.0	106.0	26.0	1.40	NaN	11.6	0.0	1.0	1.0	0.0	ckd
6	90.0	1.015	3.0	74.0	25.0	1.10	142.0	12.2	NaN	0.0	0.0	0.0	ckd
13	70.0	1.015	3.0	208.0	72.0	2.10	138.0	9.7	NaN	0.0	0.0	1.0	ckd
14	70.0	NaN	NaN	98.0	86.0	4.60	135.0	9.8	NaN	0.0	0.0	1.0	ckd
15	80.0	1.010	3.0	157.0	90.0	4.10	130.0	5.6	0.0	0.0	0.0	1.0	ckd
17	70.0	1.015	2.0	99.0	46.0	2.20	138.0	12.6	NaN	1.0	1.0	0.0	ckd
18	80.0	NaN	NaN	114.0	87.0	5.20	139.0	12.1	NaN	0.0	1.0	1.0	ckd
20	60.0	1.015	1.0	100.0	31.0	1.60	NaN	10.3	NaN	0.0	1.0	0.0	ckd

PENANGANAN MISSING VALUE

© 2018 UNIDIP

Gambar 4.1.2-5 Antarmuka halaman Transformasi Data Nominal pada proses Pemodelan Manual

6. Antarmuka halaman Penanganan *Missing Value* pada Pemodelan Manual

Antarmuka halaman Penanganan *Missing Value* dimunculkan ketika *user* menekan tombol pilihan “penanganan *missing value*” pada halaman transformasi data nominal. Halaman ini memunculkan kembali *dataset* yang telah diproses dalam mengubah *missing value* yang semula dinotasikan dalam “NaN” menjadi rata-rata dari tiap bagian data atribut. Halaman ini memiliki dua buah pilihan tombol tahap lanjutan “Tanpa Normalisasi” dan “Normalisasi”. Antarmuka halaman penanganan *missing value* pada proses *demo* dapat dilihat pada Gambar 4.1.2-6

Pemodelan Manual

Identifikasi Penyakit Ginjal Kronis

Home

Pemodelan

Langsung

Manual

Identifikasi

Penanganan Missing Value

285 Data, 12 Atribut

id	bp	sg	al	bgr	bu	sc	sod	hemo	rbc	htn	dm	appet	cla
1	80.000000	1.020000	1.000000	121.000000	36.000000	1.200000	140.158371	15.400000	0.104396	0.000000	0.000000	0.000000	ckc
5	80.000000	1.010000	2.000000	106.000000	26.000000	1.400000	140.158371	11.600000	0.000000	1.000000	1.000000	0.000000	ckc
6	90.000000	1.015000	3.000000	74.000000	25.000000	1.100000	142.000000	12.200000	0.104396	0.000000	0.000000	0.000000	ckc
13	70.000000	1.015000	3.000000	208.000000	72.000000	2.100000	138.000000	9.700000	0.104396	0.000000	0.000000	1.000000	ckc
14	70.000000	1.018736	0.610687	98.000000	86.000000	4.600000	135.000000	9.800000	0.104396	0.000000	0.000000	1.000000	ckc
15	80.000000	1.010000	3.000000	157.000000	90.000000	4.100000	130.000000	5.600000	0.000000	0.000000	0.000000	1.000000	ckc
17	70.000000	1.015000	2.000000	99.000000	46.000000	2.200000	138.000000	12.600000	0.104396	1.000000	1.000000	0.000000	ckc
18	80.000000	1.018736	0.610687	114.000000	87.000000	5.200000	139.000000	12.100000	0.104396	0.000000	1.000000	1.000000	ckc
20	60.000000	1.015000	1.000000	100.000000	31.000000	1.600000	140.158371	10.300000	0.104396	0.000000	1.000000	0.000000	ckc

TANPA NORMALISASI NORMALISASI

© 2018 UNDIP

Gambar 4.1.2-6 Antarmuka halaman Penanganan *Missing Value* pada proses Pemodelan Manual

7. Antarmuka halaman Normalisasi pada Pemodelan Manual

Antarmuka halaman normalisasi dimunculkan ketika *user* menekan tombol pilihan “Normalisasi” pada halaman penanganan *missing value*. Halaman ini memunculkan kembali *dataset* yang masing-masing telah diproses untuk diubah ke skala $\max = 1$ dan $\min = 0$. Halaman ini memerlukan 2 input, nilai k untuk kNN dan nilai k untuk kFold dan tombol submit “Prediksi kNN” sebagai tahapan akhir proses *demo*. Antarmuka halaman normalisasi data pada proses *demo* dapat dilihat pada Gambar 4.1.2-7

Pemodelan Manual

Identifikasi Penyakit Ginjal Kronis

Home

Pemodelan

Langsung

Manual

Identifikasi

Normalisasi

285 Data, 12 Atribut

id	bp	sg	al	bgr	bu	sc	sod	hemo	rbc	htn	dm	appet	class
1	0.666667	0.750000	0.250000	0.274194	0.250000	0.166667	0.606335	0.803279	0.104396	0.000000	0.000000	0.000000	ckd
5	0.666667	0.250000	0.500000	0.193548	0.153846	0.208333	0.606335	0.491803	0.000000	1.000000	1.000000	0.000000	ckd
6	1.000000	0.500000	0.750000	0.021505	0.144231	0.145833	0.680000	0.540984	0.104396	0.000000	0.000000	0.000000	ckd
13	0.333333	0.500000	0.750000	0.741935	0.596154	0.354167	0.520000	0.336066	0.104396	0.000000	0.000000	1.000000	ckd
14	0.333333	0.686782	0.152672	0.150538	0.730769	0.875000	0.400000	0.344262	0.104396	0.000000	0.000000	1.000000	ckd
15	0.666667	0.250000	0.750000	0.467742	0.769231	0.770833	0.200000	0.000000	0.000000	0.000000	0.000000	1.000000	ckd
17	0.333333	0.500000	0.500000	0.155914	0.346154	0.375000	0.520000	0.573770	0.104396	1.000000	1.000000	0.000000	ckd
18	0.666667	0.686782	0.152672	0.236559	0.740385	1.000000	0.560000	0.532787	0.104396	0.000000	1.000000	1.000000	ckd
20	0.000000	0.500000	0.250000	0.161290	0.201923	0.250000	0.606335	0.385246	0.104396	0.000000	1.000000	0.000000	ckd

Masukkan Nilai k pada kNN

5

PREDIKSI KNN


Min: 3, Max: 99

© 2018 UNDIP

Gambar 4.1.2-7 Antarmuka halaman Normalisasi pada proses *Demo*

8. Antarmuka halaman Hasil Pemodelan

Antarmuka dimunculkan ketika *user* menekan tombol pilihan “Prediksi kNN” pada halaman normalisasi atau menekan tombol “Tanpa Normalisasi” pada halaman penanganan *missing value*. Halaman ini juga dapat dihasilkan setelah kita melakukan input form pada halaman pemodelan langsung. Halaman ini memunculkan validasi performa dari semua tahapan yang telah dipilih *user* selama proses pemodelan. Halaman ini memiliki tombol “kembali” untuk *redirect* ke halaman pemodelan manual dan juga “simpan pemodelan” yang dapat digunakan dalam identifikasi selanjutnya dan melakukan *redirect* ke halaman identifikasi. Antarmuka halaman hasil pemodelan dapat dilihat pada Gambar 4.1.2-8



Identifikasi Penyakit Ginjal Kronis

- Home
- Pemodelan
- Langsung
- Manual**
- Identifikasi

Pemodelan Manual

Hasil Pemodelan

Tahapan

- Atribut hasil Backward Elimination, $\alpha = 0.1$: bp, sg, al, bgr, bu, sc, sod, hemo, rbc, htn, dm, appet
- Pembersihan Outlier
- Transformasi Nominal Menjadi Numerik
- Penanganan Missing Value
- Normalisasi
- Prediksi kNN, k= 5

Validasi

(1.4551010131835938 s)

Fold	Akurasi (97.882%)	Sensitifity (95.685%)	Specificity (100.0%)
1	96.55172413793103	92.85714285714286	100.0
2	96.55172413793103	91.66666666666666	100.0
3	100.0	100.0	100.0
4	100.0	100.0	100.0
5	100.0	100.0	100.0
6	100.0	100.0	100.0
7	96.42857142857143	93.75	100.0
8	96.42857142857143	92.85714285714286	100.0
9	92.85714285714286	85.71428571428571	100.0
10	100.0	100.0	100.0

KEMBALI
SIMPAN PEMODELAN

© 2018 UNDIP

Gambar 4.1.2-8 Antarmuka halaman hasil proses Pemodelan

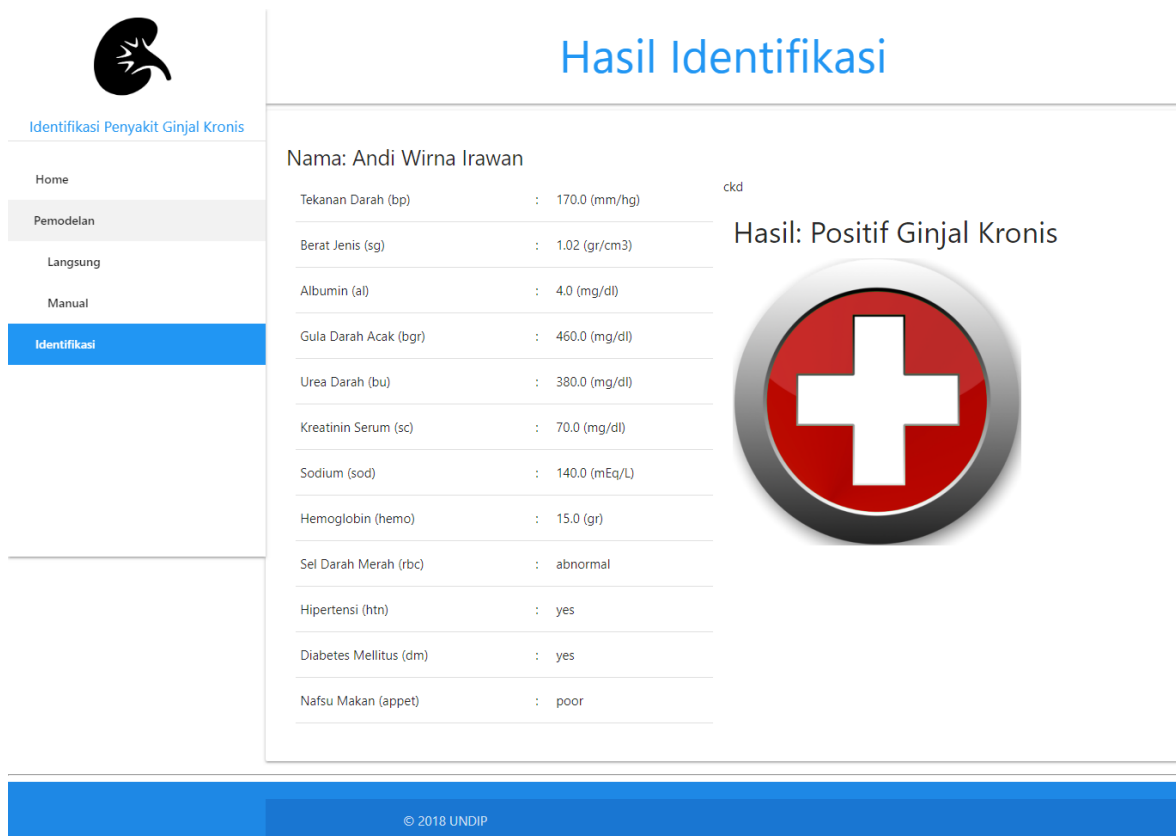
9. Antarmuka halaman *Input* pada proses Identifikasi

Antarmuka halaman *input* pada identifikasi merupakan sebuah halaman untuk melakukan proses pendeteksian PGK dengan otomatis menggunakan atribut dan tahapan yang telah direkomendasikan dari hasil penelitian ini atau hasil penyimpanan pemodelan yang dilakukan oleh *user*. Antarmuka halaman hasil *Input* proses Identifikasi dapat dilihat pada Gambar 4.1.2-9

Gambar 4.1.2-9 Antarmuka halaman *Input* proses Identifikasi

10. Antarmuka halaman Hasil proses Identifikasi

Antarmuka halaman hasil identifikasi merupakan sebuah halaman untuk menampilkan hasil proses pendeteksian PGK dari halaman *Input* proses Identifikasi. Antarmuka halaman Hasil pada proses Identifikasi dapat dilihat pada Gambar 4.1.2-10



Gambar 4.1.2-10 Antarmuka halaman Hasil Identifikasi

4.2 Skenario Pengujian Sistem

Skenario pengujian yang akan dilakukan pada Implementasi *Data Mining* untuk Identifikasi Penyakit Ginjal Kronis menggunakan *k-Nearest Neighbor* dengan *Backward Elimination* terdiri dari dua, yaitu pengujian fungsional sistem dan pengujian performa algoritma.

4.2.1 Skenario Pengujian Fungsional Sistem

Pengujian fungsional sistem dilakukan dengan metode pengujian *black box*. Pengujian *black box* adalah pengujian yang dilakukan dengan identifikasi kesalahan fungsionalitas perangkat lunak yang tampak dalam kesalahan output. Strategi pengujian ini tidak melihat mekanisme internal perangkat lunak dan hanya berfokus pada output yang dihasilkan dalam merespon input yang dipilih. Daftar rencana pengujian fungsional dapat dilihat pada tabel 4.2.1-1.

Table 4.2.1-1 Butir Pengujian Fungsional Sistem

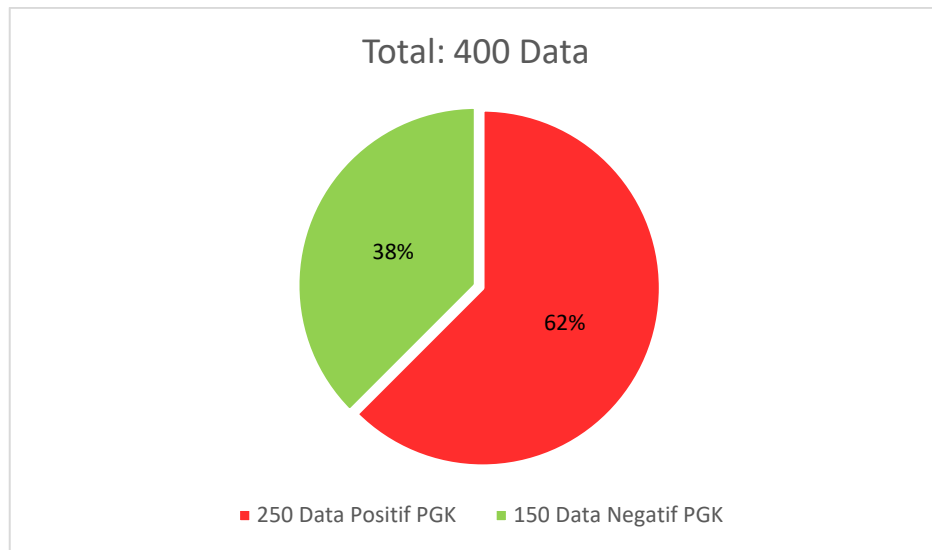
SRS-ID	Deskripsi	Butir Uji	Identifikasi
SRSF-PGK-01	Sistem dapat menampilkan proses data mining pada tiap tahapan proses	Melakukan pemodelan proses <i>data mining</i> dan menampilkan setiap tahapan proses	U-1-1
		Menyimpan pemodelan proses <i>data mining</i> yang dibuat dan dapat digunakan sebagai identifikasi	U-1-2
SRSF-PGK-02	Sistem dapat melakukan identifikasi PGK	Melakukan identifikasi status PGK menggunakan algoritma kNN	U-2-1
		Melakukan identifikasi status PGK menggunakan algoritma kNN dengan hasil seleksi fitur <i>Backward Elimination</i>	U-2-2
SRSF-PGK-03	Sistem dapat menampilkan performa pemodelan data mining yang dibuat	Menampilkan performa tahapan <i>data mining</i> ketika digunakan pada dataset di proses pemodelan manual dan pemodelan langsung	U-3-1

4.2.2 Skenario Pengujian

Skenario pengujian performa algoritma membahas mengenai data yang digunakan untuk pengujian performa algoritma dan skenario pengujian algoritma yang akan dilakukan.

4.2.2.1 Data Pengujian

Pengujian perangkat lunak ini menggunakan 400 data yang merupakan hasil proses *data mining* yang telah dilakukan. Dataset tersebut memiliki rincian berupa 250 dataset dengan klasifikasi positif teridentifikasi penyakit ginjal kronis (ckd) dan 150 data negatif teridentifikasi penyakit ginjal kronis (notckd). Data set identifikasi ginjal kronis diambil dari sebuah website UC Irvine Machine Learning Repository, merupakan survey dari Universitas Alagappa, India (https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease). Detail jumlah data yang digunakan dalam sistem dapat dilihat pada Gambar 4.2.2.1-1



Gambar 4.2.2-1 Jumlah Data yang Digunakan Sistem

4.2.2.2 Skenario Pengujian

Dalam pengujian diperlukan beberapa skenario pengujian. Digunakan skenario yang dilakukan pada proses pemodelan tahap data mining sebagai berikut:

a. Skenario 1

Skenario 1 dilakukan untuk mengetahui performa algoritma *k-Nearest Neighbors* terhadap dataset PGK seluruh atribut. Performa algoritma kNN dapat diketahui dengan menghitung akurasi, *specificity*, dan *sensitivity* dari algoritma tersebut.

Pada proses pengujian ini peneliti menggunakan model *k-Fold Cross Validation* dengan nilai $k = 10$, menggunakan jumlah tetangga terdekat sejumlah $k = 3$ hingga $k = 13$. Dengan hanya menggunakan k 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 algoritma kNN dengan dataset PGK seluruh atribut.

b. Skenario 2

Skenario 2 dilakukan untuk mengetahui performa algoritma *k-Nearest Neighbors* terhadap dataset PGK dengan atribut hasil algoritma *Backward Elimination* pada $\alpha = 0.1$ dan $\alpha = 0.05$. Performa algoritma kNN dapat

diketahui dengan menghitung akurasi, *specificity*, dan *sensitivity* dari algoritma tersebut.

Pada proses pengujian ini peneliti menggunakan model *k-Fold Cross Validation* dengan nilai $k = 10$, menggunakan jumlah tetangga terdekat sejumlah $k = 3$ hingga $k = 13$. Dengan hanya menggunakan k 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 algoritma kNN dengan dataset PGK dengan atribut hasil algoritma *Backward Elimination*.

4.3 Analisis Hasil Pengujian

Analisis hasil penelitian menjelaskan hasil dan analisis pengujian sistem dari rencana pengujian yang telah ditentukan sebelumnya untuk identifikasi PGK.

4.3.1 Analisis Hasil Pengujian Fungsional Sistem

Pengujian fungsional sistem dilakukan sesuai dengan butir pengujian sistem pada Tabel 4.2.1-1. Analisis hasil pengujian fungsional sistem dapat dilihat pada Lampiran 1.

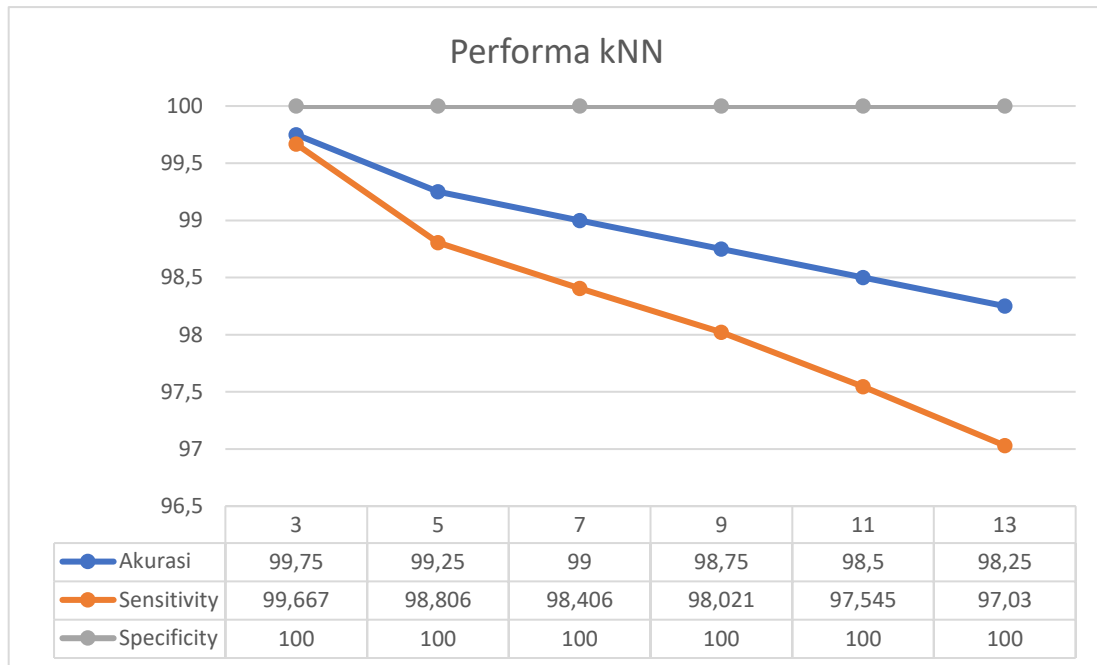
4.3.2 Analisis Hasil Skenario 1

Skenario 1 dilakukan untuk mengetahui performa *k-Nearest Neighbor* dengan menggunakan seluruh atribut yang ada pada dataset dengan menggunakan model *k-Fold Cross Validation* dengan nilai $k = 10$ dan tetangga terdekat sejumlah $k=3$ sampai dengan $k=13$ dengan nilai k (tetanga) pada kNN bernilai ganjil. Dalam bidang kesehatan, perhitungan yang dilakukan mempertimbangkan nilai rata-rata akurasi, *specificity*, dan *sensitivity*. Hasil pengujian skenario 1 dapat dilihat pada tabel 4.3.2-1

Tabel 4.3.2-1 Hasil Pengujian Skenario 1

Nilai k (kNN)	Akurasi (%)	Sensitivity (%)	Specificity (%)
3	99,75	99,667	100
5	99,25	98,806	100
7	99	98,406	100
9	98,75	98,021	100
11	98,5	97,545	100
13	98,25	97,03	100

Dari hasil skenario 1 pada tabel 4.3.2-1 dapat dilihat bahwa nilai k terbaik pada kNN adalah $k = 3$. Karena dapat memberikan nilai akurasi dan *sensitivity* tertinggi dengan nilai akurasi = 99.75% dan *sensitivity* = 99.667%, sedangkan *specificity* bernilai tetap 100% pada semua nilai $k = 3$ hingga $k = 13$.



Gambar 4.3.2-1 Grafik Performa kNN Skenario 1

Dari grafik performa kNN pada gambar 4.3.2-1 dapat disimpulkan bahwa semakin banyak nilai k pada kNN pada skenario 1 akan memiliki kecenderungan dalam menurunkan performa *akurasi* dan *sensitivity*.

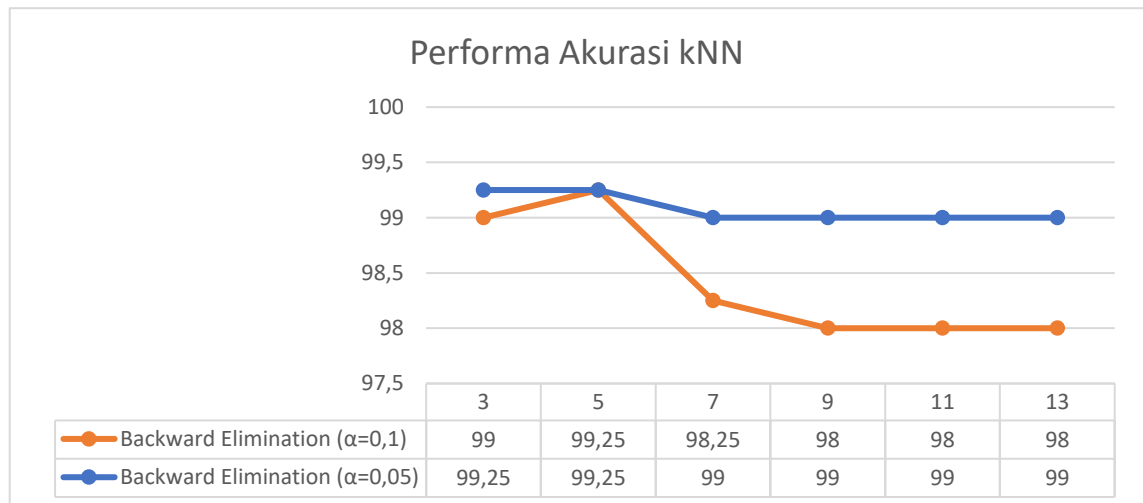
4.3.3 Analisis Hasil Skenario 2

Skenario 2 dilakukan untuk mengetahui performa *k-Nearest Neighbor* dengan menggunakan sebagian atribut hasil seleksi fitur *Backward Elimination* pada $\alpha = 0.1$ dan $\alpha = 0.05$, dengan menggunakan model *k-Fold Cross Validation* dengan nilai $k = 10$ dan tetangga terdekat sejumlah $k=3$ sampai dengan $k=13$ dengan nilai k (tetanga) pada kNN bernilai ganjil. Dalam bidang kesehatan, perhitungan yang dilakukan bertujuan untuk mempertimbangkan nilai rata-rata akurasi, *specificity*, dan *sensitivity*. Hasil pengujian skenario 2 dapat dilihat pada tabel 4.3.3-1.

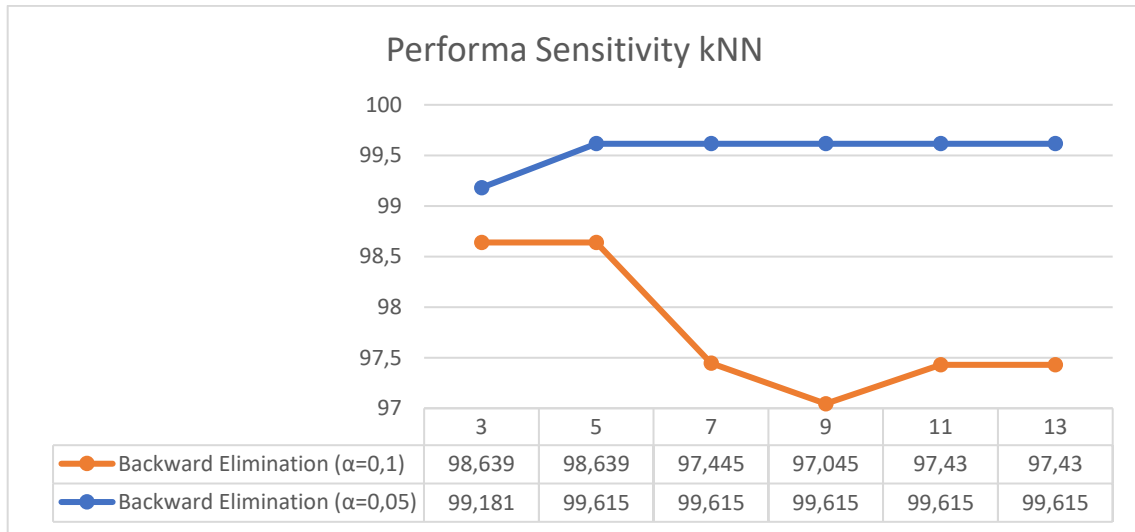
Tabel 4.3.3-1 Hasil Pengujian Skenario 2

α Backward Elimination	Atribut yang digunakan	Nilai k (kNN)	Akurasi (%)	Sensitivity (%)	Specificity (%)
0,1	bp, sg, al, bgr, bu, sc, sod, hemo, rbc, htn, dm, appet (12 Atribut)	3	99	98.639	99.412
		5	99.25	98.639	100
		7	98.25	97.445	99.412
		9	98	97.045	99.412
		11	98	97.43	98.745
		13	98	97.43	98.745
0,05	sg, al, bu, sc, sod, hemo, rbc, htn, dm, appet (10 Atribut)	3	99.25	99.181	99.412
		5	99.25	99.615	98.812
		7	99	99.615	98.245
		9	99	99.615	98.245
		11	99	99.615	98.245
		13	99	99.615	98.245

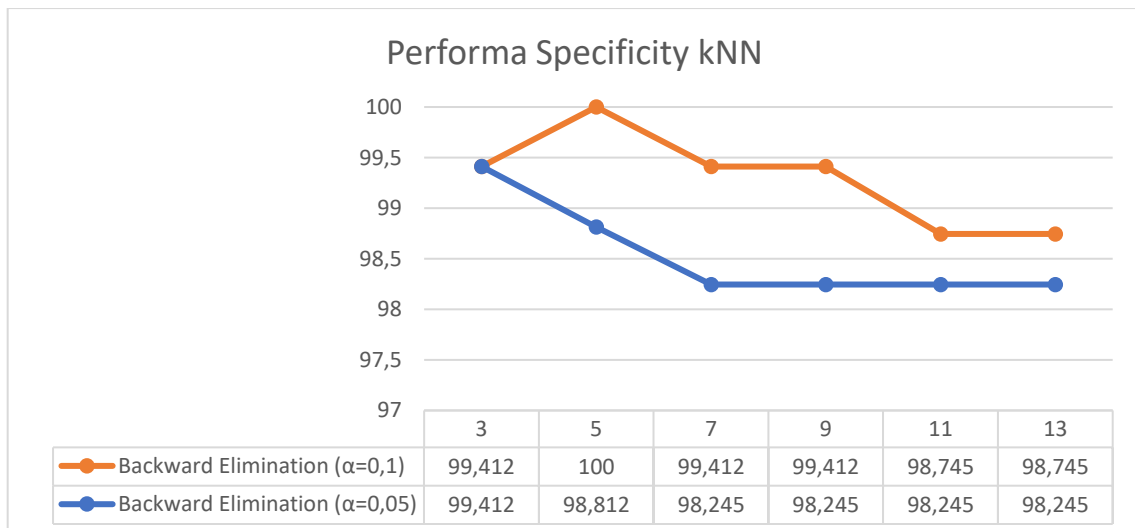
Dari hasil skenario 2 pada tabel 4.3.3-1 dan gambar grafik 4.3.3-1 dapat dilihat bahwa dengan *Backward Elimination*, kedua pengujian $\alpha = 0,1$ dan $\alpha = 0,05$ memiliki nilai akurasi terbaik kNN yang sama sebesar 99.25%.



Gambar 4.3.3-1 Perbandingan Performa Akurasi



Gambar 4.3.3-2 Perbandingan Performa Sensitivity



Gambar 4.3.3-3 Perbandingan Performa Specificity

Dapat dilihat pada gambar grafik 4.3.3-1, 4.3.3-2, dan 4.3.3-3, bahwa meskipun akurasi kNN dengan seleksi atribut *Backward Elimination* dengan $\alpha = 0,1$ dan $\alpha = 0,05$ memiliki akurasi tertinggi yang sama yaitu 99,25%, namun jika ditinjau dari hasil iterasi selama mencari nilai k kNN terbaik, *Backward Elimination* dengan $\alpha = 0,05$ memiliki performa akurasi dan *sensitivity* lebih baik dibandingkan $\alpha = 0,1$.

Seperti yang telah dijelaskan sebelumnya, meskipun performa *specificity* *Backward Elimination* dengan $\alpha = 0,1$ lebih baik daripada $\alpha = 0,05$, parameter performa yang menjadi prioritas pada kasus dalam bidang kesehatan dalam diagnosis penyakit adalah akurasi dan *sensitivity*.

BAB V

KESIMPULAN DAN SARAN

Bab ini menyajikan kesimpulan dari uraian yang telah dijabarkan pada bab-bab sebelumnya dan saran untuk pengembangan penelitian selanjutnya.

5.1 Kesimpulan

1. Perubahan nilai k pada *k-Nearest Neighbor* (kNN) mempengaruhi nilai rata-rata akurasi, sensitivity, dan specificity dalam identifikasi penyakit ginjal kronis (PGK).
2. Nilai k terbaik dalam identifikasi PGK menggunakan semua atribut adalah $k = 3$, dengan nilai akurasi = 99,75%, sensitivity = 99,667%, dan specificity = 100%
3. Dengan menggunakan seleksi atribut *Backward Elimination*, nilai α terbaik adalah 0,05 dengan menggunakan 10 dari 24 atribut awal, atribut yang terpilih 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).
4. Nilai k terbaik dalam identifikasi PGK menggunakan atribut hasil *Backward Elimination* ($\alpha = 0.05$) adalah $k = 5$, dengan nilai akurasi = 99,25%, sensitivity = 99,615%, dan specificity 98,812%.
5. Hasil akurasi yang dihasilkan tanpa seleksi atribut lebih tinggi dari penelitian terkait, sedangkan seleksi atribut membuat hasil akurasi lebih rendah, namun tidak menunjukkan penurunan yang berarti pada performa *sensitivitas*.
6. Rekomendasi sistem dari hasil penelitian ini adalah tetap menggunakan seleksi atribut dalam identifikasi PGK dengan tujuan dapat menurunkan harga tes pemeriksaan laboratorium dan menyederhanakan inputan yang perlu dimasukkan oleh pengguna.

5.2 Saran

Saran yang dapat diberikan dari tugas akhir ini untuk pengembangan penelitian selanjutnya adalah sebagai berikut:

1. Perlu penambahan data latih yang memungkinkan sistem untuk mengenali pola lebih banyak dalam identifikasi PGK sehingga akurasi menjadi lebih baik.
2. Perlunya pengembangan penelitian terhadap metode seleksi atribut (*feature selection*) lainnya.

DAFTAR PUSTAKA

- Elkan, C. (2010) *Predictive analytics and data mining*, Npl. Tersedia pada: <http://www.mendeley.com/research/data-mining-and-predictive-analysis/>.
- Fakhruddin, A. (2013) “Faktor-Faktor Penyebab Penyakit Ginjal Kronik Di Rsup Dr Kariadi Semarang Periode 2008-2012.”
- Fayyad, U., Piatetsky-Shapiro, G. dan Smyth, P. (1996) “From Data Mining to Knowledge Discovery in Databases,” *AI Magazine*, 17(3), hal. 37. doi: 10.1609/aimag.v17i3.1230.
- Gerard, E. D. (2012) “Simplifying a Multiple Regression Equation,” *The Little Handbook of Statistical Practice*, hal. 1–9.
- Han, J. dan Kamber, M. (2011) *Data Mining: Concepts and Techniques*, Elsevier. doi: 10.1007/978-3-642-19721-5.
- Hermawanti, L. dan Rabiha, S. G. (2014) “Penggabungan Algoritma Backward Elimination Dan K-Nearest Neighbor untuk Mendiagnosis Penyakit Jantung,” *Prosiding SNST*, hal. 7–12.
- 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)*, 14611(1), hal. 2319–7064. Tersedia pada: www.ijsr.net.
- Karyono, G. (2016) “Analisis Teknik Data Mining ‘Algoritma C4.5 dan K-Nereset Neighbor’ untuk Mendiagnosa Penyakit Diabetes Mellitus,” *Seminar Nasional Teknologi Informasi*, hal. 77–82. Tersedia pada: http://news.palcomtech.com/wp-content/uploads/downloads/2016/06/IT13_Giat-Karyono.pdf.
- Kemenkes (2017) *InfoDATIN*. Kementrian Kesehatan RI.
- Kirill Eremenko (2017) *Step by Step Building A Model*. Tersedia pada: <https://www.superdatascience.com/wp-content/uploads/2017/02/Step-by-step-Blueprints-For-Building-Models.pdf>.
- Kohavi, R. (1995) “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,” *Appears in the International Joint Conference on Artificial Intelligence (IJCAI)*, 5, hal. 1–7. doi: 10.1067/mod.2000.109031.
- Kumari, B. dan Swarnkar, T. (2011) “Filter versus Wrapper Feature Subset Selection in Large Dimensionality Microarray: A Review,” *International Journal of Computer*

Science and Information Technologies, 2(3), hal. 1048–1053.

Kutner, M. H. *et al.* (2004) *Applied Linear Statistical Models Fifth Edition*. 5 ed. McGraw-Hill/Irwin.

National Kidney Foundation (2015) *Diabetes - A Major Risk Factor for Kidney Disease*. Tersedia pada: <https://www.kidney.org/atoz/content/diabetes>.

Noori, R. *et al.* (2011) “Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction,” *Journal of Hydrology*. Elsevier B.V., 401(3–4), hal. 177–189. doi: 10.1016/j.jhydrol.2011.02.021.

Salekin, A. dan Stankovic, J. (2016) “Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes,” *Proceedings - 2016 IEEE International Conference on Healthcare Informatics, ICHI 2016*, hal. 262–270. doi: 10.1109/ICHI.2016.36.

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*, 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,” 4(12), hal. 608–612. doi: 10.17577/IJERTV4IS120622.

Zhang, S. *et al.* (2008) “Missing value imputation based on data clustering,” *Transactions on computational science I*, (60496327), hal. 128–138. doi: 10.1007/978-3-540-79299-4_7.