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 PENGESAHAN

Yang bertandatangan dibawah ini menyatakan bahwa Proposal Tugas Akhir yang berjudul:

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

Dipersiapkan dan disusun oleh:

Nama : Ikhsan Wisnuadji Gamadarenda

NIM : 24010313130108

Telah disahkan sebagai Proposal Tugas Akhir yang merupakan salah satu syarat untuk memperoleh gelar Sarjana Komputer.

Semarang, 7 Juni 2018

Mengetahui,	Menyetujui,
Ketua Departemen Ilmu Komputer/Informatika	Dosen Pembimbing,
Dr. Retno Kusumaningrum, S.Si, M.Kom	Indra Waspada, ST, MTI
NIP. 198104202005012001	NIP. 197902122008121002

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

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 Masalah

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 ekstrasi 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 aplikasi 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 aplikasi 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 Eliminiation* 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 Eliminatizon*"

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 aplikasi 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 aplikasi adalah Python.
2. Data set pasien 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 aplikasi tersebut sehingga dapat diimplementasikan.

BAB III ANALISIS DAN PERANCANGAN

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

BAB IV IMPLEMENTASI DAN PENGUJIAN

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

BAB V PENUTUP

Bab ini berisi kesimpulan yang diambil berkaitan dengan aplikasi 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 dengna latar belakang masalah. Penelitian terkait dapat dilihat pada tabel 2.1-1.

Tabel 2.1-1 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 *Forward Selection* (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 *feature selection* menghasilkan hasil akhir 11 attribut 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 mendapatakan 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 (Fakhruddin, 2013). Mengingat sifat penyakit ini yang

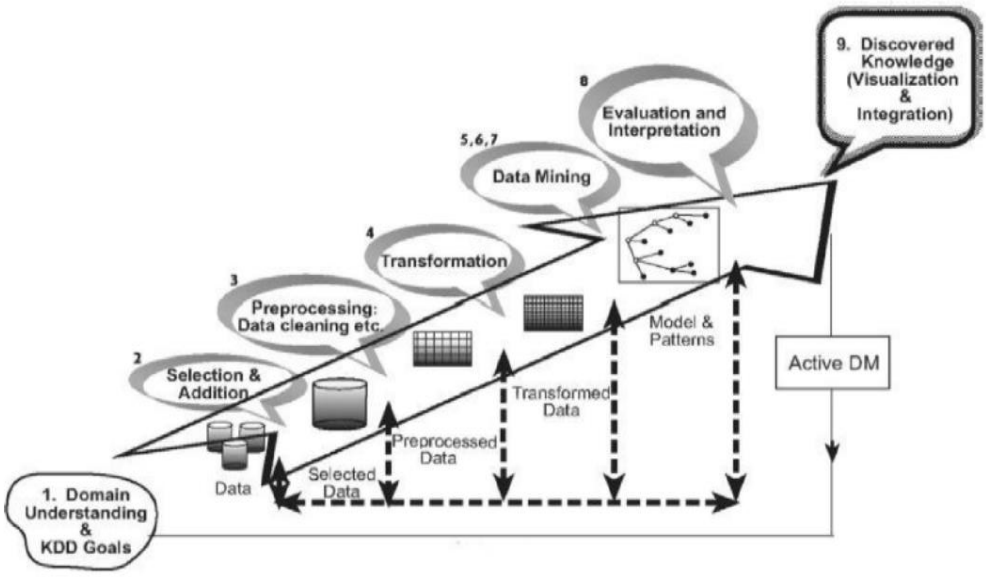
irreversible, maka penyakit gagal ginjal lebih baik untuk dicegah, dan melakukan penganggulangan lebih awal, sehingga pasien yang terkena PGK dapat mendapatkan terapi yang efektif. Ketika PGK lambat terdeteksi maka memerlukan biaya yang lebih besar dalam pengobatannya serta membutuhkan tenaga medis yang 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 manua 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 II-1 *Knowledge Data Discovery Framework*

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 interprestasi 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 popular 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):

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

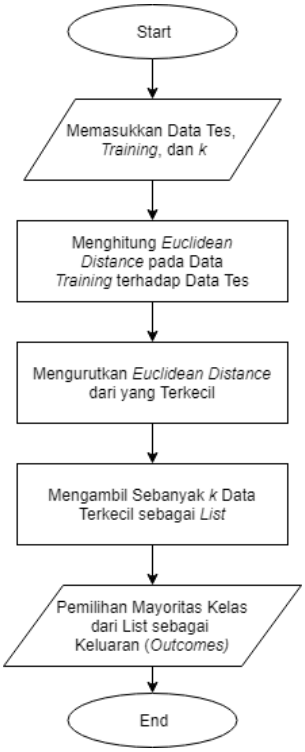
Keterangan:
D(a,b) = Jarak antara a dan b dari matrik berdimensi d
a = data training
b = data uji

Persamaan 2.4-1

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 II-2 *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 2.4-1 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 2.4-2 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 2.4-3 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 2.4-4 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

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

Tabel 2.4-5 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

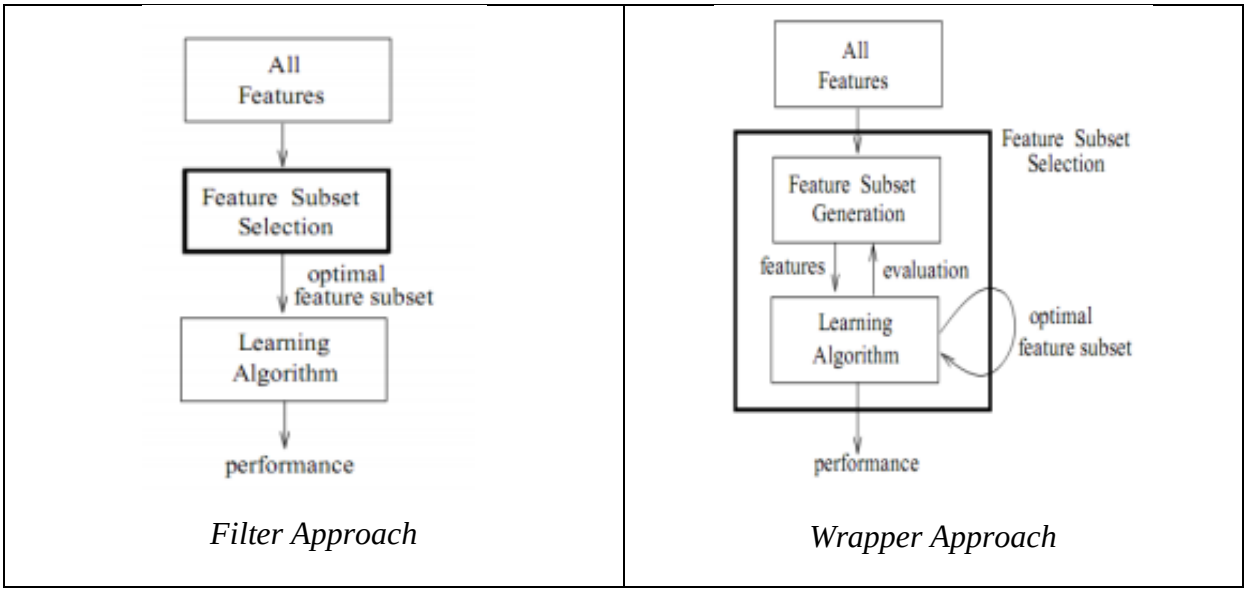
4) Dengan menggunakan *k*=5 (5 data teratas) pada tabel 2.4-5, akan diambil nilai kelulusan dengan frekuensi yang paling tinggi. Dalam hasil pengurutan berdasarkan jarak terdekat *Euclidean Distance* algoritma kNN didapatkan nilai "Terlambat" sebanyak 3 kali, nilai "Tepat" sebanyak 2 kali. Sehingga hasil 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.

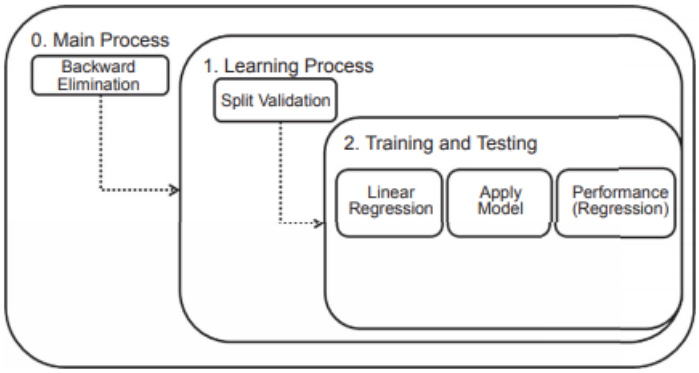
Terdapat dua teknik pendekatan dalam seleksi atribut: 1) *Filter Approach*, dan 2) *Wrapper Approach*. *Filter Approach* menilai relevansi dengan melihat sifat – sifat intrinsik data. Semua atribut diberi skor dan peringkat berdasarkan kriteria tertentu, beberapa atribut dengan peringkat tertinggi dipilih, 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 II-3 Teknik Seleksi Atribut

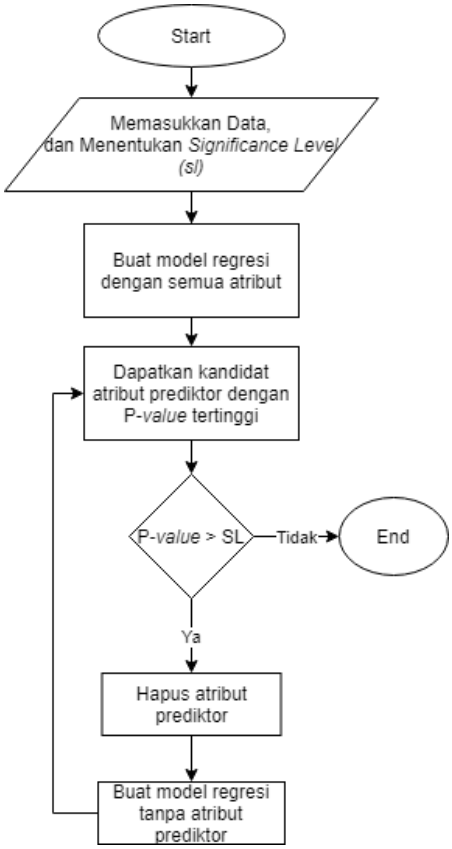
2.6. Algoritma Backward Elimination

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.



Gambar II-4 *Logic Function* dari Teknik *Wrapper Approach* pada Operator *Backward Elimination* dengan RapidMiner

Backward Elimination dimulai oleh 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*) yang telah ditentukan sebelumnya, maka atribut *X* tersebut dihilangkan. Kemudian gunakan model atribut prediktif yang tersisa kembali dicek model regresinya, dan ulangi kembali langkah tersebut untuk mencari kandidat atribut prediktif selanjutnya yang akan dihilangkan. Proses ini berlangsung hingga tidak ada lagi atribut prediktif X yang dapat dihilangkan, dengan kata lain tidak ada atribut prediktif yang memiliki P-*value* yang lebih besar dari batas derajat (*significance level*) yang telah ditentukan sebelumnya (Kutner *et al.*, 2004).

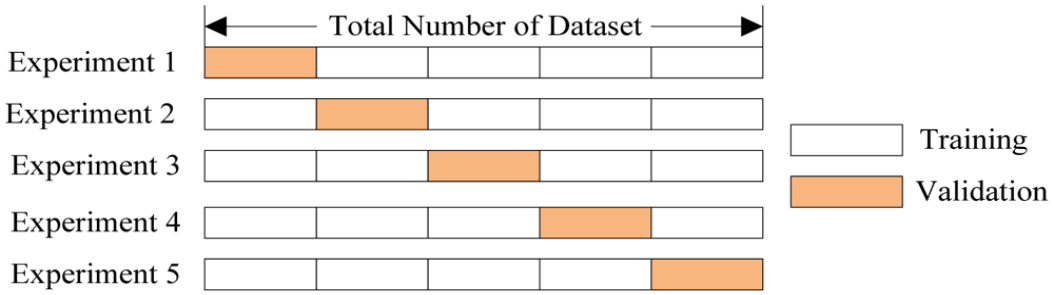


Gambar II-5 *Flowchart* proses *Backward Elimination*

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 II-6 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 2.8-1 *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*:

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

Persamaan 2.8-1

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

Persamaan 2.8-2

Dalam diagnosa medis, sensitivitas (*sensitivity*) menjelaskan tentang kemampuan dalam tes untuk mengidentifikasi secara benar seseorang yang mengidap penyakit (*true positive rate*), sedangkan kekhususan (*specificity*) menjelaskan kemampuan dalam tes untuk mengidentifikasi seseorang yang tidak mengidap penyakit (*true negative rate*).

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

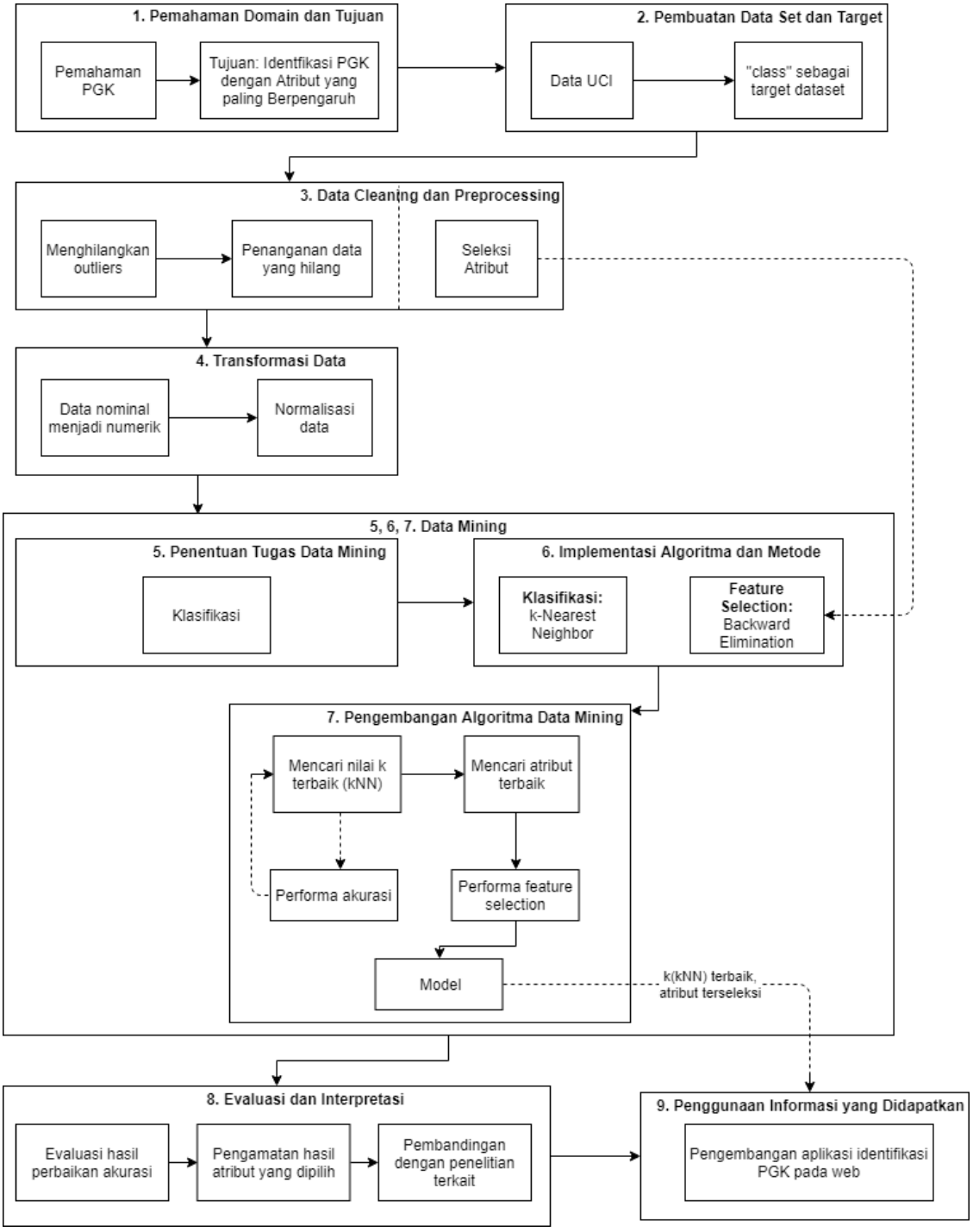
   Eksperimental dilakukan untuk mengungkapkan hubungan sebab akibat dua variabel atau lebih dengan memperhatikan pengaruh variabel lainnya. Metode ini dilaksanakan dengan memberikan variabel bebas secara sengaja kepada objek penelitian untuk diketahui akibatnya di dalam variabel terikat. Dalam tugas akhir 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 bertempatkan 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 III-1 Kerangka Kerja Penelitian

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.

Tabel 3.3-1 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	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 menghilakan 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.

2. Pembuatan Dataset dan Target

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

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.

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.

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 pertambahan kelompok.

6. *Data Mining*: Penentuan Algoritma dan Metode

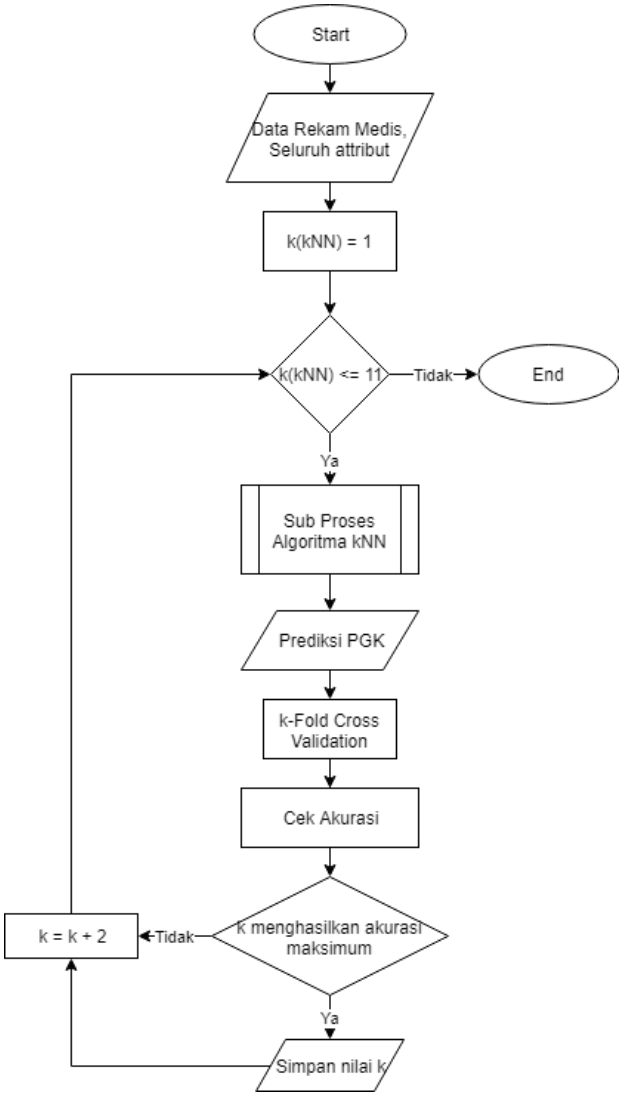
Algoritma yang digunakan pada dataset PGK menggunakan algoritma *k*-Nearest Neighbor (kNN) dengan *Backward Elimination*. kNN diimplementasikan dengan rumus *Euclidean Distance* dikarenakan *dataset* yang digunakan sebagian besar adalah numerik. Sedangkan algoritma *Backward Elimination* bertujuan agar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

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:

   - 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 k dengan akurasi terbaik. Sub-proses algoritma kNN telah ditunjukkan sebelumnya 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. Flowchart iterasi yang bertujuan mendapatkan nilai *k* terbaik dari algoritma kNN ditunjukkan pada gambar 3.3-2.



Gambar III-2 *Flowchart* Penentuan Nilai *k* pada kNN

- Seleksi atribut algoritma *Backward Elimination*

  Setelah mendapatkan nilai k terbaik pada algoritma kNN, akan dilanjutkan dengan seleksi atribut dengan 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.

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.

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.

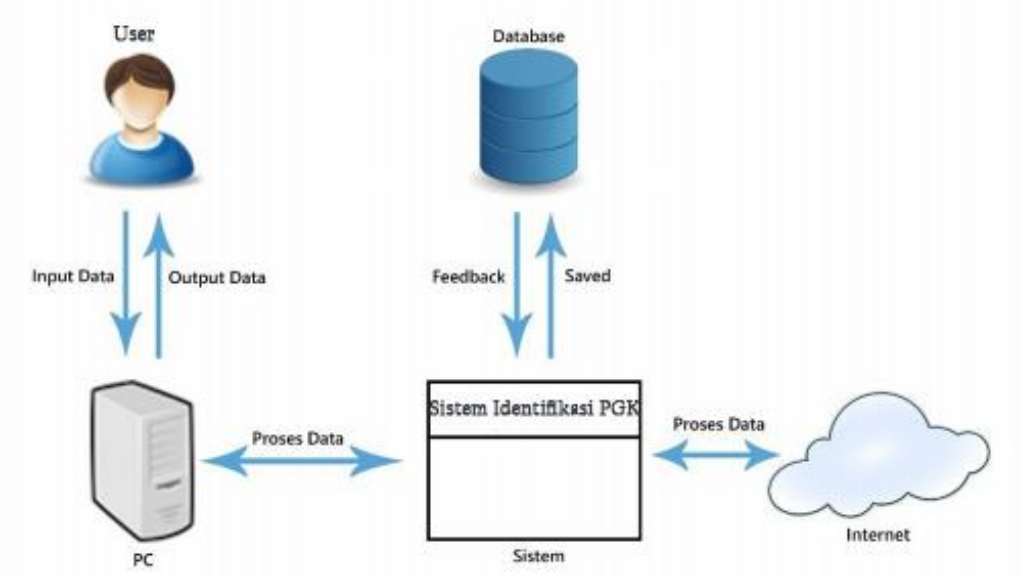
1. Deskripsi Sistem

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.

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 alrogitma *k*NN dan algoritma *k*NN 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.



Gambar III-3 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 alrogitma *k*NN dan algoritma *k*NN 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. Kebutuhan Fungsional Sistem

Kebutuhan fungsional akan didefinisikan melalui spesifikasi *Software Requirement Spesification* (SRS) aplikasi *data mining* dalam identifikasi PGK dapat dilihat pada tabel 3.4-1

Tabel 3.4-1 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 algoritma dari aplikasi

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 Perangkat Lunak

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

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:

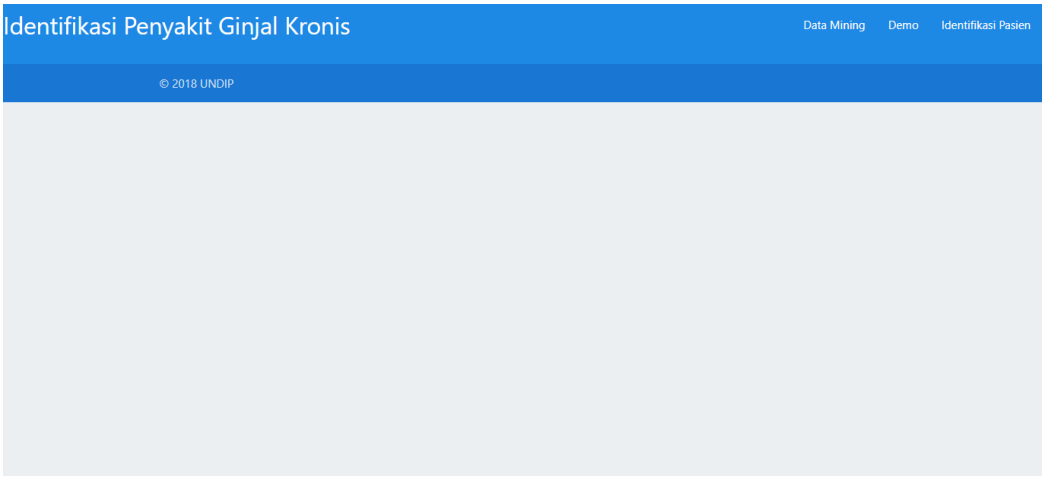
1. OS Windows 10
2. IDE PyCharm
3. Python versi 3
4. DBMS MySQL
5. *Browser* Google Chrome

Implementasi Antarmuka

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

1. Antarmuka *Welcome Page*

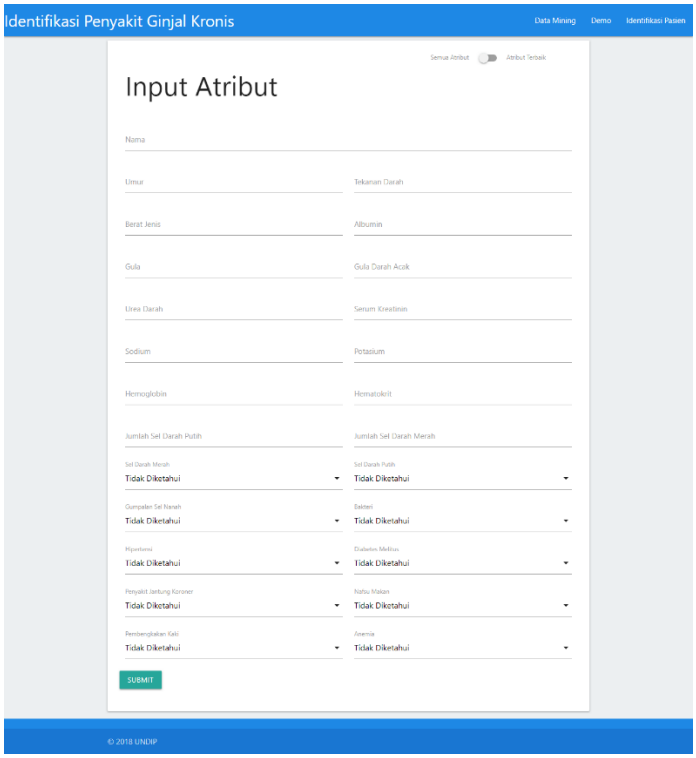
Merupakan halaman awal yang ditampilkan ketika pengguna akan mengakses perangkat lunak. Antarmuka *homepage* dapat dilihat pada gambar



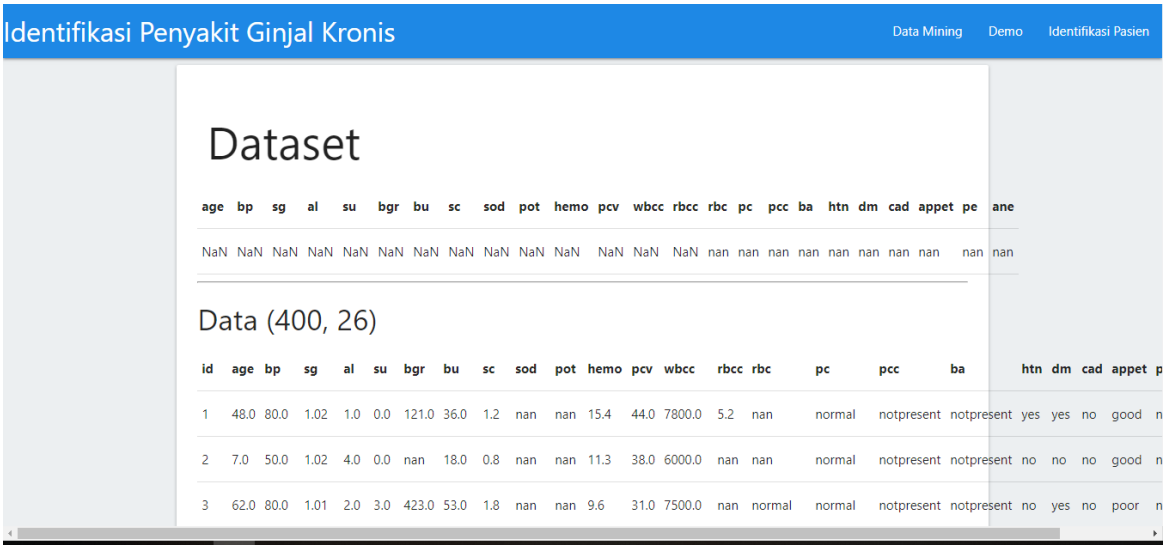
Gambar IV-1 Antarmuka *Welcome Page*

2. Antarmuka Demo

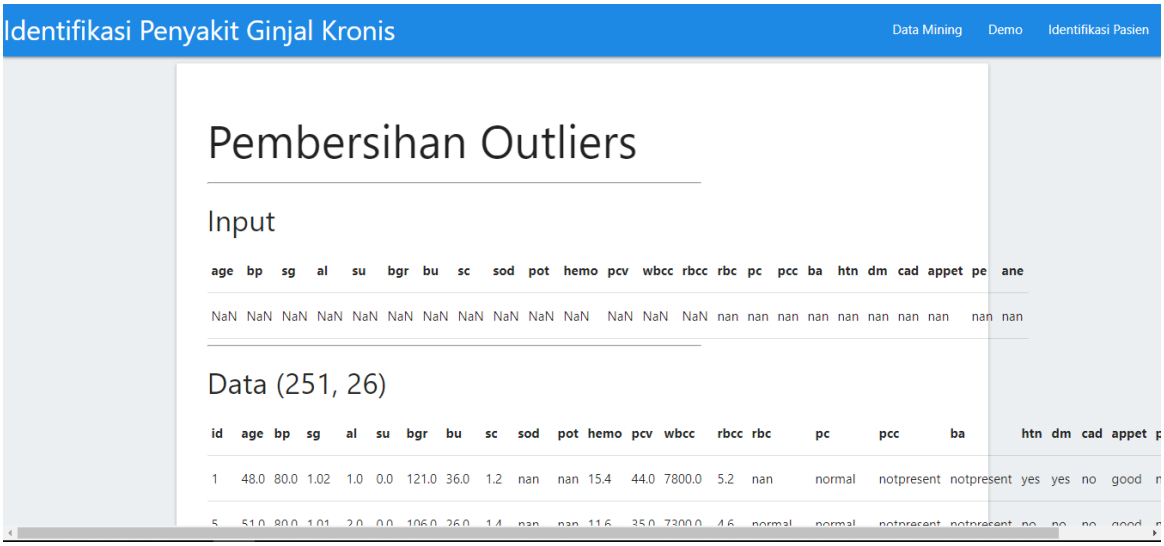
Antarmuka tahapan jika *User* ingin melakukan percobaan sendiri pada tahapan proses data mining untuk identifikasi PGK mulai dari pemilihan atribut, metode, hingga pengecekan hasil dan akurasi.



Gambar IV-2 Antarmuka Demo



Gambar IV-3 Antarmuka Dataset



Gambar IV-4 Antarmuka Pembersihan Outliers

Gambar IV-5 Antarmuka Perubahan Nominal Menjadi Numerik

3. Antarmuka Identifikasi

Merupakan implementasi terbaik dalam identifikasi PGK menurut penelitian Tugas Akhir.

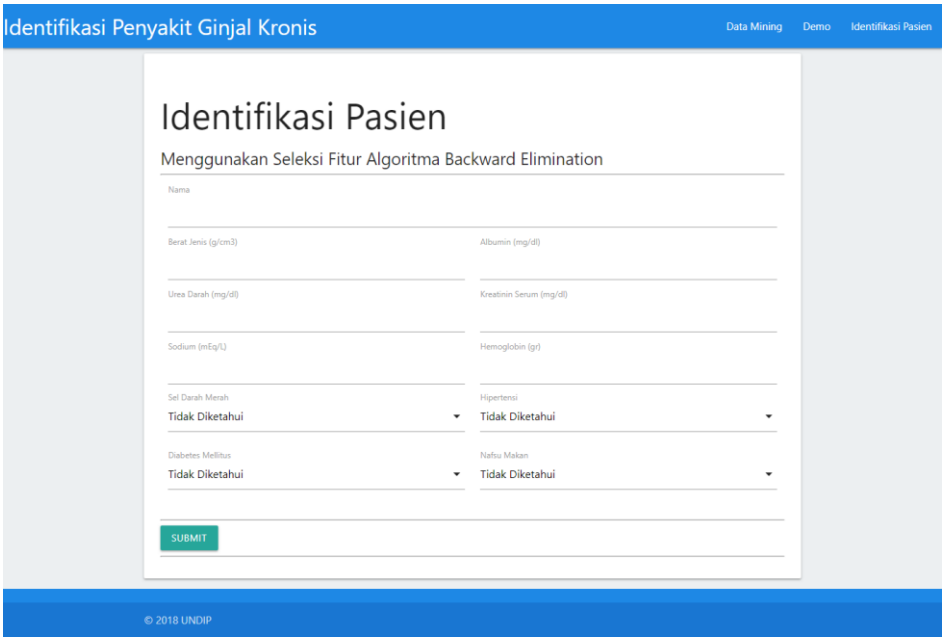


Figure IV-5 Antarmuka Identifikasi Pasien

4.2. Skenario Pengujian Perangkat Lunak

Skenario pengujian sistem dilakukan untuk identifikasi PGK menggunakan Algoritma kNN dan *Backward Elimination* terdiri dari dua, menggunakan pengujian fungsional dan kinerja sistem.

Pengujian Fungsional

Pengujian fungsional sistem dilakukan dengan metode *black box*, yaitu dilakukan dengan identifikasi kesalahan fungsionalitas perangkat lunak yang tampak pada kesalahan *output*. Strategi pengujian ini fokus pada *output* yang dihasilkan berdasarkan *input* yang dipilih, dan tidak melihat mekanisme internal perangkat lunak. Daftar rencana pengujian fungsional dapat dilihat pada tabel

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 Tektnik 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 Articial 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.