KLASIFIKASI PENYAKIT GINJAL KRONIS PADA MANUSIA MENGGUNAKAN ALGORITMA STOCHASTIC GRADIENT DESCENT (SGD), RANDOM FOREST, DAN SUPPORT VECTOR MACHINE (SVM)

Iqbal Dandy Lazuardi[1], Haris Saputra[2], Muhammad Faiq Ardyanto Putro[3]
[1,2,3]Prodi S1 Informatika, Fakultas Informatika, Universitas Telkom

Abstrak

Penyakit ginjal kronis merupakan masalah kesehatan yang sangat berbahaya yang telah menyebar secara global. Hal ini disebabkan oleh perubahan gaya hidup seperti kebiasaan makanan, perubahan atmosfer, dll. Jadi sangat penting untuk memutuskan obat apa pun yang harus dihindari dan untuk memprediksi penyakit pada tahap awal yang membantu menurunkan tingkat kematian. Dari hasil analisis yang telah dilakukan menunjukkan bahwa teknik *feature selection* sangat cocok untuk memprediksi penyakit ginjal kronis. *Principal component analysis* adalah salah satu teknik pemilihan fitur yang menyaring atribut yang kurang penting dan juga mengambil atribut penting dari dataset. Dalam penelitian ini dilakukannya perbandingan dari berbagai pendekatan klasifikasi data dalam hal seberapa akurat mereka memprediksi penyakit ginjal kronis. Dalam penelitian ini algoritma yang digunakan adalah *stochastic gradient descent* (SGD), *random forest,* dan *support vector machine* (SVM). Sedangkan untuk pengukuran akurasi yang digunakan untuk membandingkan pengklasifikasian adalah *recall*, *f1-score*, dan *precision*. Hasil yang didapat dari penelitian ini yaitu, perhitungan dengan *stochastic gradient descent* yang diterapkan pada *dataset* penyakit ginjal kronis menghasilkan nilai akurasi sebesar 94.16%. lalu dengan menggunakan *random forest* dengan *dataset* yang serupa menghasilkan nilai akurasi sebesar 98.33%. Terakhir dengan menggunakan algoritma *support vector machine* menghasilkan nilai akurasi yang serupa dengan *random forest* yaitu sebesar 98.33%. Sehingga dapat disimpulkan bahwa algoritma *random forest* dan *support vector machine* merupakan algortima terbaik untuk mendeteksi penyakit ginjal kronis.

Kata Kunci: Penyakit Ginjal Kronis, *Stochastic Gradient Descent, Random Forest, Support Vector Machine*

Abstract

Chronic kidney disease is a very dangerous health problem that has spread globally. This is caused by lifestyle changes such as food habits, atmospheric changes, etc. So it is very important to decide which drugs to avoid and to predict the disease at an early stage that helps reduce mortality. From the results of the analysis that has been done shows that the feature selection technique is very suitable for predicting chronic kidney disease. Principal component analysis is a feature selection technique that filters out less important attributes and also takes important attributes from the dataset. In this study, a comparison of various data classification approaches in terms of how accurately they predict chronic kidney disease. In this study the algorithm used is stochastic gradient descent (SGD), random forest, and support vector machine (SVM). As for the measurement of accuracy used to compare classifications are recall, f1-score, and precision. The results obtained from this study are, calculations with stochastic gradient descent applied to the chronic kidney disease dataset produce an accuracy value of 94.16%. then by using a random forest with a similar dataset produces an accuracy value of 98.33%. Finally, using the support vector machine algorithm produces an accuracy value similar to random forest, which is 98.33%. It can be concluded that the random forest algorithm and support vector machine are the best algorithms for detecting chronic kidney disease.

Keywords: Chronic Kidney Disease, Stochastic Gradient Descent, Random Forest, Support Vector Machine

1. Pendahuluan

Ginjal adalah salah satu organ yang sangat penting yang berfungsi untuk menjaga komposisi darah dengan mencegah menumpuknya limbah atau kotoran dan mengatur keseimbangan cairan di dalam tubuh, menjaga tingkatan elektrolit seperti sodium, potasium dan fosfat tetap stabil, serta memproduksi hormon dan enzim yang membantu

tubuh dalam mengendalikan tekanan darah, membuat sel darah merah dan menjaga tulang tetap kuat dan sehat (Kemenkes, 2017).

Penyakit ginjal kronis (PGK) adalah masalah kesehatan masyarakat dunia dengan prevalensi dan insiden gagal ginjal yang sering selalu meningkat setiap tahun. Prevalensi PGK selalu meningkat seiring meningkatnya jumlah penduduk usia lanjut dan penyakit diabetes melitus serta tekanan darah tinggi. Sekitar 1 dari 10 orang populasi dunia mengalami PGK pada stadium tertentu (Kemenkes, 2017). Nilai prevalensi di seluruh Indonesia untuk penyakit gagal ginjal memiliki nilai rata - rata berkisar kurang lebih 0.2 persen (Foundation, 2017). Dengan banyaknya jumlah penderita penyakit ginjal kronis dan perlunya diagnosis yang tepat maka perlu dilakukan penelitian untuk mencari metode yang memberikan nilai akurasi tertinggi.

Teknik *feature selection* sangat cocok untuk memprediksi penyakit ginjal kronis. *Principal component analysis* adalah salah satu teknik pemilihan fitur yang menyaring atribut yang kurang penting dan juga mengambil atribut penting dari dataset. Dalam penelitian ini dilakukannya perbandingan dari berbagai pendekatan klasifikasi data dalam hal seberapa akurat mereka memprediksi penyakit ginjal kronis. Algoritma yang digunakan pada penelitian ini adalah *stochastic gradient descent* (SGD), *random forest*, dan *support vector machine* (SVM). Sedangkan untuk pengukuran akurasi yang digunakan untuk membandingkan pengklasifikasian adalah recall, f1-score, dan precision.

2. Dasar Teori

2.1. Data

Data adalah sekumpulan keterangan atau fakta mentah berupa simbol, angka, kata-kata, atau citra, yang didapatkan melalui proses pengamatan atau pencarian ke sumber-sumber tertentu.

Data juga merupakan kumpulan keterangan-keterangan atau deskripsi dasar dari suatu hal (objek atau kejadian) yang diperoleh dari hasil pengamatan (observasi) dan dapat diolah menjadi bentuk yang lebih kompleks, seperti; informasi, *database*, atau solusi untuk masalah tertentu.

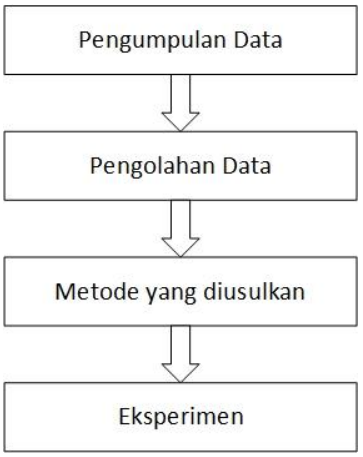
2.2. *Data Mining*

Data mining adalah suatu proses penambangan informasi penting dari suatu data. Dalam definisi yang lain, *data mining* dijelaskan sebagai serangkaian proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu basis data dengan melakukan penggalian pola-pola dari data dengan tujuan untuk memanipulasi data menjadi informasi yang lebih berharga yang diperoleh dengan cara mengekstraksi dan mengenali pola yang penting atau menarik dari data yang terdapat dalam *database*.

Data mining memiliki fungsi antara lain untuk menemukan pola dari data lalu menggunakan variabel untuk dapat memprediksi variabel lainnya, menemukan karakteristik penting dalam data yang sedang dipakai, menemukan model dan fungsi untuk menggambarkan kelas atau konsep data, dan menemukan suatu hubungan pada nilai atribut dari beberapa data.

3. Metodologi Penelitian

Dalam penelitian ini dilakukan beberapa tahapan penelitian, adapun tahapan penelitian yang dilakukan ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Penelitian

Pengumpulan Data

Pada tahap ini merupakan tahapan pencarian calon dataset yang akan diproses. Dalam penelitian ini, digunakan dataset yang didapatkan dari *website* The UCI Machine Learning Repository. Dataset dalam penelitian ini adalah 400 orang pasien yang terdiri dari 24 atribut dan 1 label. Hal ini ditunjukkan pada Tabel 1.

Tabel 1. Penjelasan Atribut

No.	Atribut	Tipe Atribut

1	age	Numerical
2	bp	Numerical
3	sg	Nominal
4	al	Nominal
5	su	Nominal
6	rbc	Nominal
7	pc	Nominal
8	pcc	Nominal
9	ba	Nominal
10	bgr	Numerical
11	bu	Numerical
12	se	Numerical
13	sod	Numerical
14	pot	Numerical
15	hemo	Numerical
16	pcv	Numerical
17	we	Numerical
18	re	Numerical
19	htn	Nominal
20	dm	Nominal
21	cad	Nominal
22	appet	Nominal
23	pe	Nominal
24	ane	Nominal
25	class	Nominal

Pengolahan Data

Pada tahap ini dilakukannya data *preparation* yaitu untuk membuat dataset menjadi normal sehingga data tersebut dapat diproses. Pada tahap ini ditemukannya beberapa *missing value* pada *dataset*, untuk memperbaiki *missing value* tersebut dilakukannya modifikasi data yang merupakan *missing value* menjadi angka nol. Lalu pada tahap ini juga dilakukan mengubah data dari *categorical* menjadi *numerical*.

Metode yang diusulkan

Pada tahap ini dilakukannya penentuan metode. Metode yang diusulkan untuk penelitian ini ialah algoritma *stochastic gradient descent* (SGD), *random forest*, serta *support vector machine* (SVM).

Eksperimen

Setelah menentukan metode yang akan digunakan, maka dilakukannya eksperimen untuk ketiga metode yang telah dipilih. Hasil yang ingin dicapai dari eksperimen tersebut ialah berupa data *precision*, *recall*, *f1-score*, *support*, serta akurasi.

4. Hasil dan Evaluasi

Tahapan terakhir dari penelitian yang dilakukan adalah tahapan evaluasi. Tahapan evaluasi dilakukan untuk menguji performansi dari sistem yang telah dibangun. Evaluasi dilakukan dengan cara membagi data latih yang telah diproses dengan data uji. Setelah dilakukan validasi maka dilakukan perhitungan performansi sistem menggunakan *recall*, *precision*, dan *F1-score*.

Dari dataset yang telah diolah kemudian kemudian dataset dibagi menjadi data training dan data testing. sebanyak 30% dari data keseluruhan digunakan untuk *data testing* dan sisanya menjadi *data training*. Dari data yang didapatkan tersebut kemudian dilakukan perhitungan dengan metode yang telah ditentukan yaitu, *Stochastic Gradient Descent (SGD)*, *Random Forest*, dan *Support Vector Machine (SVM).*

Metode pertama yang digunakan adalah *Stochastic Gradient Descent (SGD)*. Metode ini mendapatkan hasil *confussion matrix* sebagai berikut

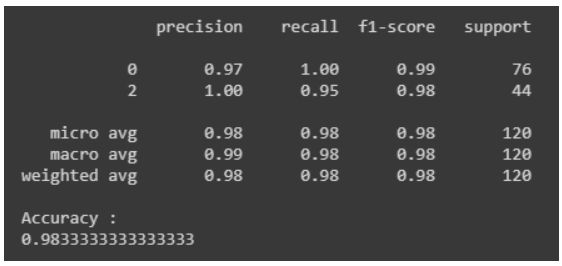
	precision	recall	f1-score	support
0	0.95	0.96	0.95	76
2	0.93	0.91	0.92	44
accuracy			0.94	120
macro avg	0.94	0.93	0.94	120
weighted avg	0.94	0.94	0.94	120

Accuracy :
0.9416666666666667

Gambar 1. Stochastic Gradient Descent (SGD)

Metode Berikutnya adalah metode *Random Forest.* Dengan menggunakan metode *Random Forest* didapatkan hasil yang sedikit lebih baik

dibandingkan metode *Stochastic Gradient Descent (SGD)* dan mendapatkan hasil sebagai berikut

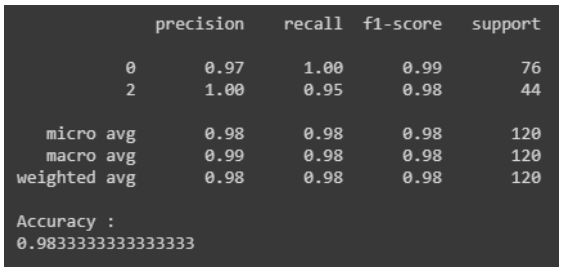


	precision	recall	f1-score	support
0	0.97	1.00	0.99	76
2	1.00	0.95	0.98	44
micro avg	0.98	0.98	0.98	120
macro avg	0.99	0.98	0.98	120
weighted avg	0.98	0.98	0.98	120

Accuracy :
0.9833333333333333

Gambar 2. Random Forest

Sementara itu dengan menggunakan metode algoritma *Support Vector Machine (SVM)* didapatkan hasil yang sama optimalnya dengan metode *Random Forest* yaitu sebagai berikut



	precision	recall	f1-score	support
0	0.97	1.00	0.99	76
2	1.00	0.95	0.98	44
micro avg	0.98	0.98	0.98	120
macro avg	0.99	0.98	0.98	120
weighted avg	0.98	0.98	0.98	120

Accuracy :
0.9833333333333333

Gambar 3. Support Vector Machine (SVM)

5. Kesimpulan dan Saran

Pada penelitian ini dilakukan permodelan menggunakan *stochastic gradient descent* (SGD), *random forest*, dan *support vector machine* (SVM) dengan menggunakan *dataset* penyakit ginjal kronis yang diambil dari UCI Repository. Hasil perhitungan dengan *stochastic gradient descent* yang diterapkan pada *dataset* penyakit ginjal kronis menghasilkan nilai akurasi sebesar 94.16%. lalu dengan menggunakan *random forest* dengan *dataset* yang serupa menghasilkan nilai akurasi sebesar 98.33%. Terakhir dengan menggunakan algoritma *support vector machine* menghasilkan nilai akurasi yang serupa dengan *random forest* yaitu sebesar 98.33%. Sehingga dapat disimpulkan bahwa algoritma *random forest* dan *support vector machine* merupakan algortima terbaik untuk mendeteksi penyakit ginjal kronis. Hal ini menjadi rekomendasi bahwa algoritma tersebut dapat digunakan oleh ahli patologi dalam membuat program untuk memprediksi penyakit ginjal kronis. Berdasarkan hasil dari penelitian yang telah dilakukan, maka peneliti mengajukan saran untuk melakukan eksperimen dengan menggunakan metode lainnya seperti *decision tree*, *naive bayes*. atau algoritma optimasi seperti *genetic algorithm*, *ant colony optimization*.

6. Daftar Pustaka

[1] UCI Machine Learning. *Chronic_Kidney_Disease_Data_Set.* https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease, 2015. Online; Accessed 9 April 2020.

[2] Maxmonroe. Pengertian Data: Definisi, Fungsi, dan Jenis-Jenis Data. https://www.maxmanroe.com/vid/teknologi/pengertian-data.html, 2020. Online; Accessed 29 April 2020.

[3] M Firman. Pengertian Data Mining dan Penerapannya. https://www.kompasiana.com/mfirman34/5c8fb0557a6d88244e001272/pengertian-data-mining-dan-penerapannya, 2019. Online; Accessed 29 April 2020.

[4] bootupacademyai. Data Mining Adalah? Pengertian Hingga Belajar Clustering Lengkap. https://bootup.ai/blog/data-mining-adalah/, 2019. Online; Accessed 29 April 2020.

[5] Sukma Arini. Apa yang dimaksud dengan data mining?. https://www.dictio.id/t/apa-yang-dimaksud-dengan-data-mining/13136, 2017. Online; Accessed 29 April 2020.

[6] Devishri P, Ragin O R, Anisha G S. (2019). *Comparative Study of Classification Algorithms in Chronic Kidney Disease. International Journal of Recent Technology and Engineering (IJRTE)*