JASIE "Jurnal Aplikasi Sistem Informasi Dan Elektronika"
Vol. 1, No. 2, Desember 2019, Hal: 127 -134
p-ISSN: 2714-612X, e-ISSN: Dalam Proses

ANALISIS PENERAPAN ALGORITMA NAIVE BAYES UNTUK KLASIFIKASI PENYAKIT GINJAL KRONIS

Dudi Irawan[1], Hardian Oktavianto[2], Moh Khoirul Anam[3]
Teknik Informatika, Universitas Muhammadiyah Jember
Email: dudi.irawan77@gmail.com[1], hardian@unmuhjember.ac.id[2],
moh.khoirul.anaam@gmail.com[3]

ABSTRAK

Pada bidang medis, klasifikasi berbasis *machine learning* telah banyak digunakan untuk membantu dokter dan ahli kesehatan dalam diagnosis penyakit maupun penentuan tindakan perawatan dan pengobatan dengan tujuan meminimalisir salah diagnosis. Klasifikasi adalah salah satu bentuk *machine learning* berbasis pembelajaran yang bertujuan mengelompokkan atau mengkategorikan sesuatu berdasarkan atribut atau fitur yang ada. Naive bayes adalah algoritma *machine learning* berbasis pembelajaran yang paling sering digunakan karena kesederhanaan konsepnya. Pada penelitian ini diterapkan algoritma naive bayes terhadap dataset *chronic kidney disease* (CKD) untuk melakukan klasifikasi apakah pasien tergolong mengidap penyakit ginjal kronis atau tidak, dengan menggunakan dataset diambil dari UCI *Machine Learning Repository*. Algoritma naïve bayes bekerja dengan baik pada dataset yang digunakan pada penelitian ini yang ditunjukkan bahwa ketika menggunakan k dengan nilai 4 sampai 10, total 158 data dapat diklasifikasikan dengan tepat. Nilai akurasi tertinggi mencapai 100%, sedangkan nilai akurasi terendah hanya 99%. Nilai akurasi tertinggi diperoleh ketika menggunakan nilai k = 4 sampai dengan nilai k = 10. Nilai *precision* dan *recall* tertinggi adalah bernilai 1 yang menunjukkan bahwa untuk dataset yang digunakan, kualitas dan kuantitas klasifikasi sangat bagus.

Kata Kunci: deteksi, klasifikasi, penyakit ginjal kronis, naive bayes

1. PENDAHULUAN

Penyakit ginjal kronis adalah suatu keadaan di mana fungsi ginjal mengalami penurunan secara berkala. Penyakit kronis ini diawali dengan sakit pada ginjal yang tidak segera dilakukan perawatan dan pengobatan. Setelah bertahun – tahun maka penyakit ginjal akan mencapai kondisi kronis dan pada tahap akhir akan menjadi gagal ginjal (Ahmad, Tundjungsari, Widianti, Amalia, & Rachmawati, 2017). Penyakit ginjal kronis merupakan salah satu masalah penyakit yang sering terjadi di seluruh dunia, berdasarkan penelitian ditemukan bahwa penyakit ini lebih banyak menyerang para lansia, yaitu orang dengan usia diatas 65 tahun (Yildirim, 2017). Penyakit ginjal kronis jika tidak segera ditangani bisa menyebabkan gagal ginjal dan berujung kepada kematian, oleh sebab itu dokter dan ahli medis akan melakukan identifikasi berupa serangkaian tes melalui diagnosis, seperti tes darah, tes urin, dan sebagainya, untuk menentukan kondisi pasien (Ahmad, Tundjungsari, Widianti, Amalia, & Rachmawati, 2017) (Charleonnan, et al., 2016). Permasalahan klasik akan tetapi membawa dampak yang signifikan adalah terjadinya salah diagnosis. Misalnya, dalam suatu kasus diagnosis penyakit tertentu, fakta sebenarnya adalah si pasien mengidap penyakit akan tetapi dokter atau ahli medis menentukan hasil diagnosis tidak mengidap penyakit atau yang dikenal dengan istilah "*false positive*" (Avci, Karakus, Ozmen, & Avci, 2018) (Yildirim, 2017).

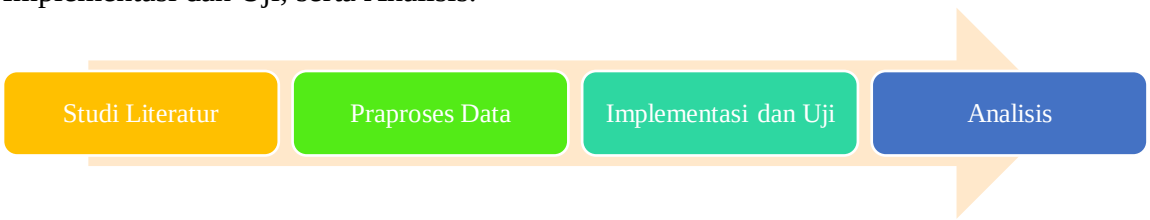
Pada bidang medis, klasifikasi berbasis *machine learning* telah banyak digunakan untuk membantu dokter dan ahli kesehatan dalam diagnosis penyakit maupun penentuan tindakan perawatan dan pengobatan, konsep umum klasifikasi

adalah mengelompokkan atau mengkategorikan sesuatu berdasarkan atribut atau fitur yang ada (Amrane, Oukid, Gagaoua, & Ensari, 2018). Algoritma – algoritma yang dapat digunakan dalam klasifikasi diantaranya adalah *Naive Bayes, Support Vector Machine, Artifial Neural Networking, Logistic Regression, K-Nearest Neigbor*, dan *Decission Tree* (Safuan, Wahono, & Supriyanto, 2015). Naive bayes adalah algoritma yang paling sering digunakan karena kesederhanan konsepnya. Naive bayes bekerja berdasarkan probabilitas, yang menghitung kemunculan fitur atau atribut terhadap suatu kelas, selain itu pada naive bayes setiap atribut atau fitur diasumsikan tidak bergantung satu sama lain (Chandra, Indrawan, & Sukajaya, 2016) (Rusland, Wahid, Kasim, & Hafit, 2017).

Pada penelitian ini akan diterapkan algoritma naive bayes terhadap dataset *chronic kidney disease* (CKD) untuk melakukan klasifikasi apakah pasien tergolong mengidap penyakit ginjal kronis atau tidak. Dataset diambil dari UCI *Machine Learning Repository* dan uji klasifikasi dilakukan dengan menggunakan *k-fold cross validation*.

2. METODE PENELITIAN

Metode penelitian dibagi menjadi bebrapa tahapan seperti yang dapat dilihat pada gambar 1. Terdapat 4 tahapan, yaitu : Studi Literatur, Praproses Data, Implementasi dan Uji, serta Analisis.



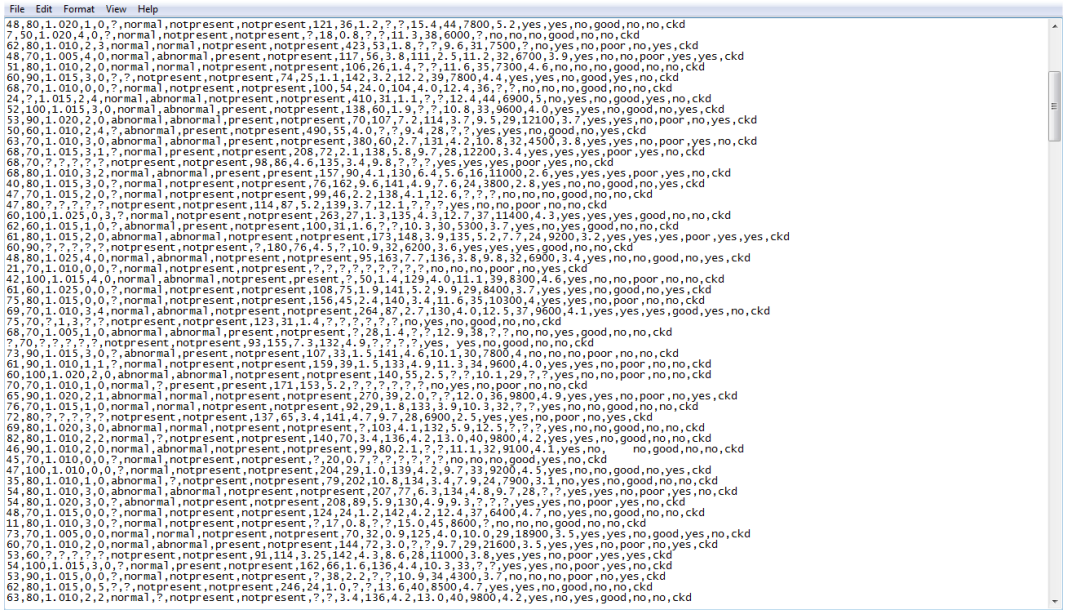
Gambar 1 Tahapan Penelitian

Studi literatur dilakukan sebagai tahapan persiapan untuk mencari dasar – dasar teori yang digunakan dalam penelitian, selain itu juga dilakukan pengumpulan data tentang penyakit ginjal kronis. Setelah studi literatur dilakukan maka tahapan selanjutnya adalah praproses data, yaitu melakukan analisis dan persiapan terhadap data penyakit ginjal kronis yang telah diperoleh agar siap untuk diolah pada proses selanjutnya. Pada tahap implementasi dan uji dilakukan dengan menggunakan WEKA. Tahapan terakhir yaitu melakukan analisis atau penarikan kesimpulan terhadap hasil uji yang diperoleh.

A. Praproses Data

Dataset yang digunakan adalah data penyakit ginjal kronis yang diakses dari https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease. Dataset ini mempunyai 400 buah *record* yang terdiri dari 25 buah atribut termasuk 1 atribut kelas, dengan perbandingan jumlah *record* diantara 2 kelas yaitu kelas kronis dan kelas tidak kronis adalah 250 dibanding 150. Rincian ke-25 atribut pada dataset ini adalah : *age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite,*

pedal edema, anemia, class. Gambar 2 merupakan *screenshot* dari dataset yang akan digunakan.



Gambar 2 *Screenshot* Dataset Spambase

Sebelum data digunakan untuk tahapan klasifikasi, dataset akan dipersiapkan terlebih dahulu. Adapun beberapa proses yang dilakukan nantinya meliputi: pembersihan data yang tidak lengkap (*missing value*), transformasi data, dan konversi data.

B. Implementasi dan Uji

Klasifikasi nantinya akan menerapkan algoritma Naive Bayes melalui *tools machine learning* yaitu WEKA, baik ketika melakukan pelatihan maupun ketika melakukan uji, dan skenario uji dataset menggunakan *k-fold cross validation*. Pada *k-fold cross validation*, total keseluruhan dataset dibagi menjadi k bagian, yang selanjutnya k-1 bagian digunakan sebagai data latih dan bagian sisa lainnya dijadikan data uji.

Tabel 1 Contoh 5-*fold cross validation*

Uji	Data Latih	Data Uji
1	F2, F3, F4, F5	F1
2	F1, F3, F4, F5	F2
3	F1, F2, F4, F5	F3
4	F1, F2, F3, F5	F4
5	F1, F2, F3, F4	F5

Sebagai contoh apabila digunakan k = 5 maka akan ada 5 bagian atau 5 kelompok data, kita misalkan 5 kelompok tersebut dengan F1, F2, F3, F4, dan F5. Uji juga akan dilakukan sebanyak 5 kali. Uji pertama F2, F3, F4 dan F5 digunakan sebagai data latih sedangkan F1 digunakan sebagai data uji. Pada uji kedua F1, F3, F4 dan F5 digunakan sebagai data latih sedangkan F2 digunakan sebagai data uji, dan

seterusnya. Sebagai ilustrasi penjelas proses *k-fold cross validation* ini dapat dilihat pada Tabel 1.

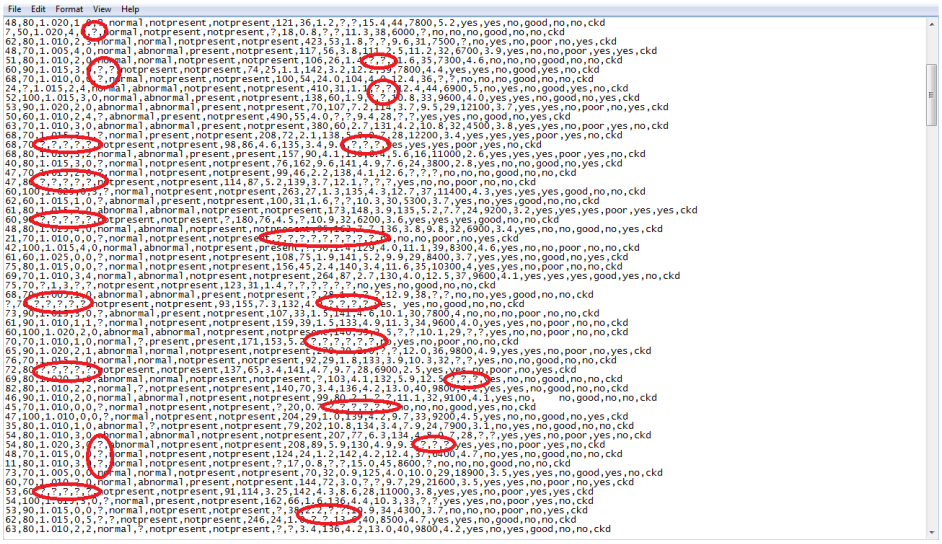
C. Analisis

Pada analisis akan dilakukan pengamatan dan penarikan kesimpulan terhadap hasil uji pada tahap sebelumnya. Analisis dilakukan agar mengetahui sejauh mana hasil akurasi, TP *Rate*, FP *Rate*, *Precision*, dan *Recall*, dari hasil klasifikasi berdasarkan skenario uji yang telah dirancang, yaitu membandingkan hasil – hasil yang diperoleh dengan menggunakan berbagai variasi nilai k pada *k-fold*.

3. HASIL PENELITIAN

A. Praproses Data

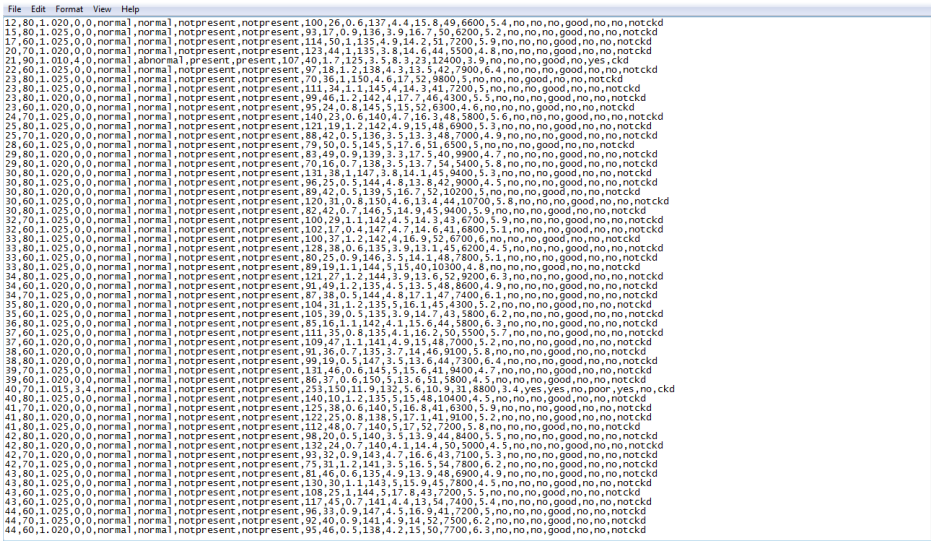
Dataset yang digunakan ternyata memiliki banyak nilai yang diwakili oleh tanda "?", hal ini pada umumnya disebut sebagai "*missing value*" atau nilai yang hilang, dan ini terjadi tidak hanya pada suatu atribut tertentu, akan tetapi pada banyak atribut. *Missing value* ini akan mempengaruhi hasil klasifikasi, sehingga pada dataset akan dilakukan pembersihan *missing value* ini. Pada gambar 3 dapat dilihat beberapa *missing value* yang terdapat pada dataset.



Gambar 3 Distribusi Jenis Dataset Berdasarkan Kelas

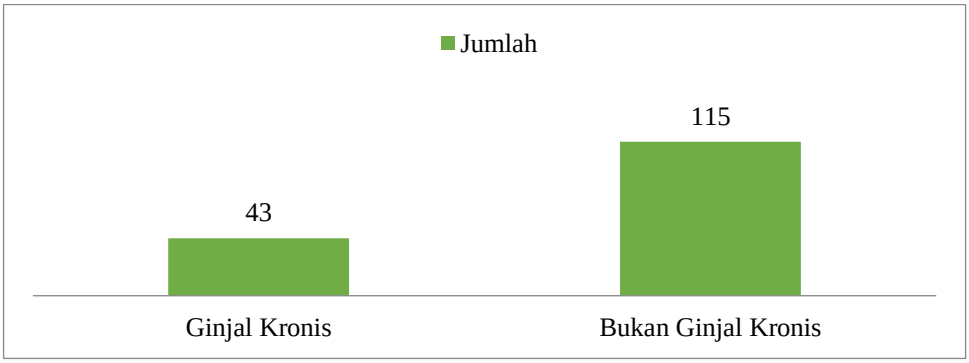
Pembersihan data dilakukan dengan bantuan perangkat lunak WEKA. Pembersihan data bertujuan untuk menghapus data mana saja yang mengandung nilai "?", apabila sebuah baris data mengandung "?" atau lebih, maka satu baris data tersebut dihapus.

Hasil dari pembersihan data dapat dilihat pada gambar 4. Dari total 400 data maka setelah dilakukan pembersihan data, total data menjadi 158 buah saja. 158 data inilah yang akan digunakan menjadi dataset bagi keseluruhan tahap penelitian.



Gambar 4 Hasil Praproses Data

Dari 158 data yang digunakan, distribusi masing – masing kelas ginjal kronis dan kelas bukan ginjal kronis adalah 43 buah termasuk ginjal kronis dan 115 termasuk kelas bukan ginjal kronis.



Gambar 5 Distribusi Kelas Data

B. Implementasi dan Uji

Klasifikasi dilakukan dengan bantuan perangkat lunak WEKA, dan klasifikasi menggunakan algoritma naïve bayes dengan skenario uji menggunakan 10-*fold cross validation*. Nilai k yang digunakan adalah mulai k = 2 sampai dengan k = 10. Uji penelitian digunakan untuk mengetahui performa dari algoritma naïve bayes terhadap dataset, yang meliputi sejauh mana algoritma mampu mengklasifikasikan data dengan benar, akurasi, serta *precision* dan *recall*. Rekapitulasi hasil uji dapat dilihat pada tabel 2.

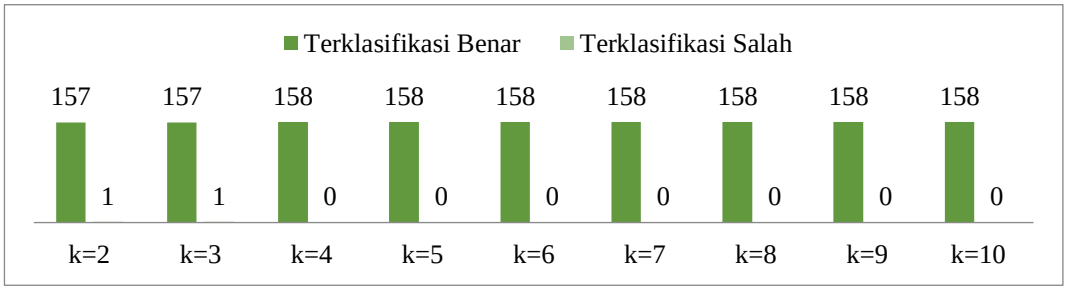
Berdasarkan tabel 2 dapat kita ketahui bahwa ketika nilai k yang digunakan adalah 2 dan 3, algoritma mampu mengklasifikasikan 157 data dengan benar dari keseluruhan 158 data dengan akurasi 99%, sedangkan untuk nilai k mulai 4 sampai dengan 10 algoritma mampu melakukan klasifikasi untuk semua 158 data dengan akurasi 100%. nilai terendah *precision* dan *recall* adalah 0.994 yang diperoleh ketika k = 2 dan k = 3, sedangkan nilai tertinggi adalah 1 yang diperoleh ketika nilai k = 4 sampai k = 10.

Tabel 2 Rekapitulasi Hasil Uji

K	Terklasifikasi Benar	Terklasifikasi Salah	% Akurasi	Precision	Recall
k=2	157	1	99%	0.994	0.994
k=3	157	1	99%	0.994	0.994
k=4	158	0	100%	1	1
k=5	158	0	100%	1	1
k=6	158	0	100%	1	1
k=7	158	0	100%	1	1
k=8	158	0	100%	1	1
k=9	158	0	100%	1	1
k=10	158	0	100%	1	1

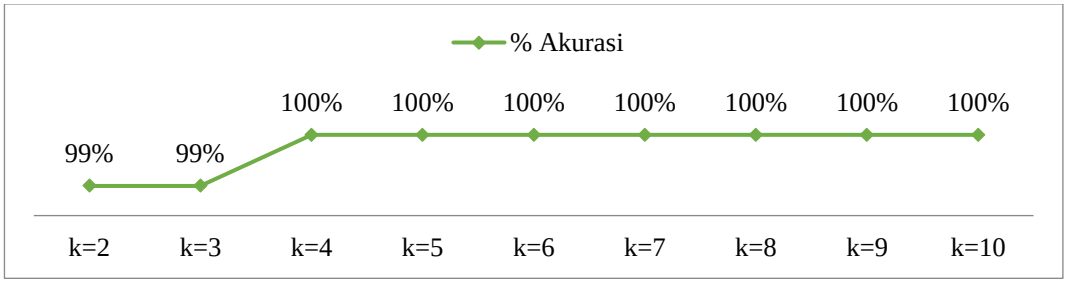
C. Analisis

Algoritma naïve bayes bekerja dengan baik pada dataset yang digunakan pada penelitian ini. Naïve bayes mampu melakukan klasifikasi data dengan benar, pada uji dengan menggunakan k bernilai 4 sampai 10, algoritma ini mampu mengklasifikasikan keseluruhan data yaitu 158 data dengan tepat tanpa kesalahan. Kecuali untuk nilai k =2 dan nilai k = 3 algoritma mampu mengklasifikasikan 157 data dari total 158 data.



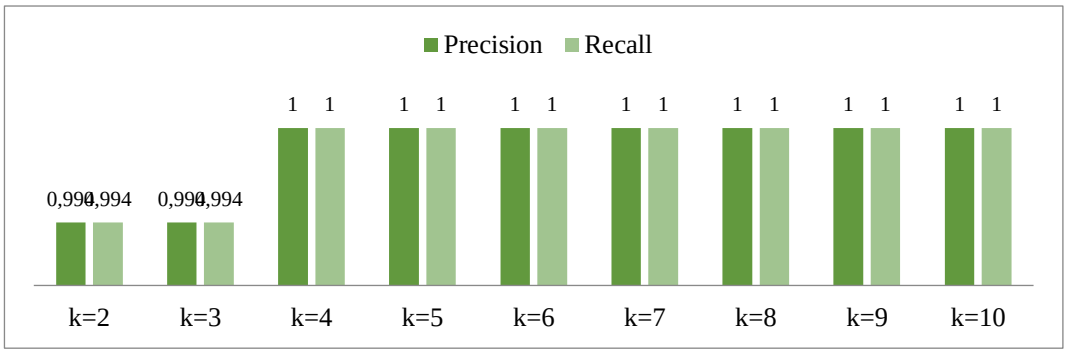
Gambar 6 Perbandingan Klasifikasi Benar dan Salah

Nilai akurasi tertinggi mencapai 100%, sedangkan nilai akurasi terendah hanya 99%. Akurasi ini diperoleh dari perhitungan berapa jumlah data yang terklasifikasi dengan tepat atau benar. Nilai akurasi tertinggi diperoleh ketika menggunakan nilai k = 4 sampai dengan nilai k = 10.



Gambar 7 Persentase Akurasi

Precision digunakan untuk mengukur kualitas klasifikasi dimana mengukur tingkat keberhasilan dari kelas ginjal kronis yang diklasifikasi dengan benar dari keseluruhan hasil klasifikasi kelas ginjal kronis. *Recall* digunakan untuk mengukur kuantitas klasifikasi dimana menunjukkan tingkat keberhasilan dari kelas ginjal kronis yang diklasifikasikan benar dari keseluruhan data kelas ginjal kronis. Nilai *precision* dan *recall* terletak diantara 0 sampai 1, apabila mendekati 1 maka menunjukkan hasil yang semakin bagus. Gambar 8 menyajikan grafik nilai *precision* dan *recall*.



Gambar 8 Perbandingan Precision dan Recall

4. KESIMPULAN

Kesimpulan pada penelitian ini adalah, algoritma naïve bayes bekerja dengan baik pada dataset yang digunakan pada penelitian ini, yang ditunjukkan dengan hasil ketika menggunakan k dengan nilai 4 sampai 10, 158 data dapat diklasifikasikan dengan tepat. Nilai akurasi tertinggi mencapai 100%, sedangkan nilai akurasi terendah hanya 99%. Nilai akurasi tertinggi pada uji klasifikasi diperoleh ketika menggunakan nilai k = 4 sampai dengan nilai k = 10. Nilai *precision* dan *recall* tertinggi adalah bernilai 1 yang menunjukkan bahwa untuk dataset yang digunakan, kualitas dan kuantitas klasifikasi sangat bagus.

Saran terhadap penelitian ini adalah mencoba algoritma naïve bayes pada model dataset yang lain, baik yang memiliki jumlah data lebih banyak ataupun atribut yang lebih banyak, juga terhadap data yang berjenis tidak seimbang atau *imbalanced* data.

5. DAFTAR PUSTAKA

1. Ahmad, M., Tundjungsari, V., Widianti, D., Amalia, P., & Rachmawati, U. A. (2017). Diagnostic Decision Support System of Chronic Kidney Disease Using Support Vector Machine. *Second International Conference on Informatics and Computing (ICIC)* (pp. 1 - 4). Jayapura, Indonesia: IEEE.
2. Amrane, M., Oukid, S., Gagaoua, I., & Ensari, T. (2018). Breast Cancer Classification Using Machine Learning. *2018 Electric Electronics, Computer Science, Biomedical Engineerings Meeting (EBBT)* (pp. 1 - 4). Istanbul, Turkey: IEEE.
3. Avci, E., Karakus, S., Ozmen, O., & Avci, D. (2018). Performance Comparison of Some Classifiers on Chronic Kidney Disease Data. *6th International Symposium on Digital Forensic and Security (ISDFS)* (pp. 1 - 4). Antalya, Turkey: IEEE.

4. Chandra, W. N., Indrawan, G., & Sukajaya, I. N. (2016, February). Spam Filtering Dengan Metode Pos Tagger Dan Klasifikasi Naïve Bayes. *Jurnal Ilmiah Teknologi dan Informasia ASIA, X*(1), 47-55.
5. Charleonnan, A., Fufaung, T., Niyomwong, T., Chokchueypattanakit, W., Suwannawach, S., & Ninchawee, N. (2016). Predictive Analytics for Chronic Kidney Disease Using Machine Learning Techniques. *The 2016 Management and Innovation Technology International Conference (MITiCON-2016)* (pp. 80 - 83). Bang-San, Thailand: IEEE.
6. Dulhare, U. N., & Ayesha, M. (2016). Extraction of action rules for chronic kidney disease using Naïve bayes classifier. *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)* (pp. 1 - 5). Chennai, india: IEEE.
7. Gunarathne, Perera, & Kahandawaarachchi. (2017). Performance Evaluation on Machine Learning Classification Techniques for Disease Classification and Forecasting through Data Analytics for Chronic Kidney Disease. *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)* (pp. 291 - 296). Washington DC, USA: IEEE.
8. Rusland, N. F., Wahid, N., Kasim, S., & Hafit, H. (2017). IOP Conference Series: Materials Science and Engineering. *International Research and Innovation Summit (IRIS2017). 226*, p. 012091. Melaka, Malaysia: IOP Publishing.
9. Safuan, Wahono, R. S., & Supriyanto, C. (2015, December). Penanganan Fitur Kontinyu dengan Feature Discretization Berbasis Expectation Maximization Clustering untuk Klasifikasi Spam Email Menggunakan Algoritma ID3. *Journal of Intelligent Systems, I*(2), 148-155. Retrieved from http://journal.ilmukomputer.org
10. Stimpson, A. J., & Cummings, M. L. (2014). Assessing Intervention Timing in Computer-Based Education Using Machine Learning Algorithms. *IEEE Access*, 78 - 87.
11. Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., . . . Wang, C. (2018). Machine Learning and Deep Learning Methods for Cybersecurity. *IEEE Access*, 35365 - 35381.
12. Yildirim, P. (2017). Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron. *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)* (pp. 193 - 198). Turin, Italy: IEEE.