

Sistem Prediksi Penyakit Kanker Serviks Menggunakan CART, Naive Bayes, dan k-NN

Cervical Cancer Disease Prediction System Using CART, Naive Bayes, and k-NN

Tutus Praningki*¹, Indra Budi²

^{1,2} Magister Teknik Informatika, Universitas AMIKOM Yogyakarta
E-mail: *praningki86@gmail.com, indra@cs.ui.ac.id

Abstrak

Tersedianya data histori rekam medis pasien kanker serviks pada institusi pelayanan kesehatan, tidak disertai dengan proses ekstraksi menjadi sebuah pengetahuan atau informasi. Penggunaan teknik data mining sangat berpotensi untuk diimplementasikan kedalam sistem yang dapat melakukan prediksi penyakit kanker serviks. Pada penelitian ini berfokus pada dataset diagnosa medis pasien yang akan melakukan tes Pap Smear. Algoritma yang digunakan untuk melakukan klasifikasi penyakit kanker serviks adalah Classification And Regression Trees (CART), Naive Bayes, dan k-Nearest Neighbor (k-NN). Pengujian yang dilakukan terhadap algoritma CART Decision Tree, Naive Bayes, dan k-NN, menggunakan formula Confusion Matrix, dengan menggunakan teknik pemecahan dataset Holdout. Hasil pengujian terhadap algoritma yang digunakan, menunjukkan algoritma Naive Bayes memiliki akurasi terbaik sebesar 94,44%, sedangkan tingkat akurasi yang dihasilkan algoritma CART dan k-NN adalah 88,89%, 85,04%. Performa yang didapatkan oleh masing-masing algoritma yang digunakan, memungkinkan penggunaan sistem prediksi penyakit kanker serviks untuk mendukung keputusan klinis pada pasien baru.

Kata Kunci — CART, Naive Bayes, k-NN, Klasifikasi

Abstract

The availability of historical data medical records of patients with cervical cancer at the health care facility, not accompanied by the extraction process into knowledge or information. The use of data mining techniques has the potential to be implemented into the system that can predict cervical cancer disease. In this research focuses on the patient's medical diagnosis dataset that will perform a Pap test. The algorithm used to classify a cervical cancer disease is the Classification and Regression Trees (CART), Naive Bayes and k-Nearest Neighbor (k-NN). Tests conducted on algorithm CART Decision Tree, Naive Bayes and k-NN, using Confusion Matrix formula, using the techniques of solving dataset holdout. The test results of the method used, showing Naive Bayes method has the best accuracy of 94,44%, while the level of accuracy resulted CART and k-NN algorithms is 88.89%, 85.04%. Performance obtained by each algorithms used, allows the use of cervical cancer disease prediction system for clinical decision support on new patients

Keywords — CART, Naive Bayes, k-NN, Classification

1. PENDAHULUAN

Saat ini penyakit kanker masih menjadi penyebab utama kematian diseluruh dunia, *International Agency for Research on Cancer* pada tahun 2012 merilis jenis penyakit kanker yang paling banyak diderita perempuan adalah kanker payudara, kolorektal, dan serviks. Pada tahun 2013 jumlah penderita kanker serviks di Indonesia lebih banyak dibandingkan kanker payudara, penderita kanker serviks paling banyak ditemukan di provinsi Jawa Timur dengan jumlah kasus sebanyak 21.313 [1]. Penyakit kanker serviks dapat dideteksi secara dini dengan melakukan tes *Pap Smear* [2]. Rumah Sakit Umum Daerah (RSUD) Kediri dan Yayasan Kanker Indonesia (YKI) kabupaten Kediri merupakan penyedia layanan untuk tes *Pap Smear*. Tersedianya data histori rekam medis pasien tidak disertai dengan proses ekstraksi data menjadi sebuah pengetahuan atau informasi yang dapat berguna untuk keputusan klinis. Penggunaan teknik *data mining* sangat potensial untuk diterapkan dalam pelayanan kesehatan, khususnya untuk prediksi penyakit kanker serviks.

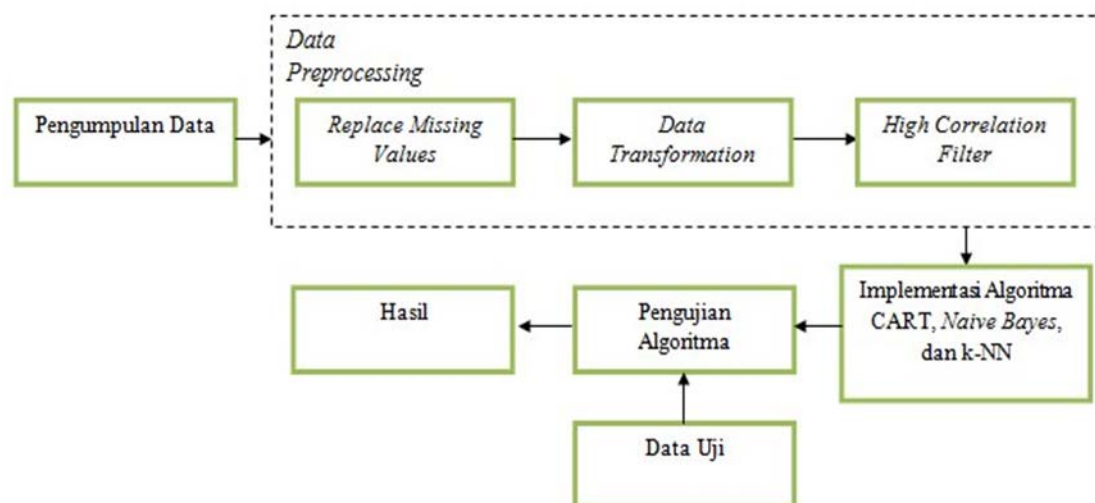
Pada kasus [3] teknik *data mining* digunakan untuk prediksi penyakit kanker serviks dengan menggunakan kombinasi nilai faktor risiko terhadap masing-masing nilai pada setiap atribut dataset. Dalam penelitian ini dataset yang digunakan adalah data rekam medis pasien ketika melakukan tes *Pap Smear* tanpa kombinasi nilai faktor risiko sehingga atribut yang digunakan sebagai target prediksi adalah hasil tes *Pap Smear* pasien.

Konsep klasifikasi merupakan bagian dari teknik data mining yang memiliki pekerjaan utama melakukan analisis prediksi [4]. Berbagai metode dalam klasifikasi banyak diterapkan untuk melakukan prediksi terhadap penyakit, seperti pada kasus [5]. Pada kasus [5], dilakukan perbandingan antara algoritma yang termasuk dalam kategori metode *Decision Tree*, dan hasilnya menunjukkan bahwa algoritma *Classification And Regression Trees* (CART) memiliki tingkat akurasi sebesar 83,2% dan lebih baik dibandingkan ID3 serta C4.5. Algoritma CART memiliki kemampuan klasifikasi yang baik, hasil prediksi pada penyakit kanker payudara menghasilkan tingkat akurasi yang signifikan sebesar 96% [6]. Pada kasus klasifikasi penyakit jantung oleh [7], metode k-Nearest Neighbor (k-NN) memiliki kinerja yang efisien dan hal tersebut sesuai dengan hasil [8], metode k-NN murni tanpa kombinasi memiliki tingkat akurasi sebesar 97,4%. Metode *Naive Bayes* dengan teknik probabilistik sederhana juga banyak digunakan untuk proses prediksi penyakit [9, 10, 11], pada kasus prediksi penyakit jantung, metode *Naive Bayes* menghasilkan tingkat akurasi sebesar 89% [10]. Menurut [12] ketiga algoritma yaitu CART, *Naive Bayes*, dan k-NN termasuk lima (5) algoritma klasifikasi yang masuk dalam sepuluh (10) algoritma terbaik dalam *data mining*.

Tujuan dari penelitian ini adalah membangun sistem yang dapat mendukung keputusan klinis untuk tenaga medis, potensi yang dimiliki oleh metode CART *Decision Tree*, *Naive Bayes*, dan k-NN pada penelitian sebelumnya memungkinkan pembuatan sistem. Beberapa penelitian menunjukkan bahwa dalam sebuah sistem dapat digunakan lebih dari satu algoritma, seperti [5] yang menggunakan algoritma CART, ID3, dan C4.5, sedangkan [11] menggunakan algoritma *Naive Bayes* dan k-NN. Algoritma yang digunakan dalam sistem akan dilakukan evaluasi dengan tujuan untuk menilai performa masing-masing algoritma yang digunakan dalam kasus prediksi hasil tes *Pap Smear*. Sangat penting untuk dapat mengetahui tingkat akurasi algoritma yang digunakan karena dengan tingkat akurasi yang signifikan maka sistem dapat digunakan untuk memprediksi hasil test *Pap Smear* pasien baru. Hasil dari evaluasi terhadap ketiga algoritma yang digunakan dapat digunakan untuk mengetahui algoritma yang memiliki kinerja paling baik.

2. METODE PENELITIAN

Tahapan penelitian mengikuti alur proses pada Gambar 1, setiap tahapan terdapat metode-metode yang digunakan untuk menyelesaikan penelitian.



Gambar 1. Alur proses penelitian

2.1. Pengumpulan Data

Data Metode pengumpulan data yang dilakukan adalah melakukan wawancara dengan tenaga medis yang melakukan penanganan terhadap pasien yang akan melakukan tes Pap Smear. Tahapan berikutnya adalah mengambil data sampel pasien yang pernah melakukan tes Pap Smear di RSUD Kediri dan YKI kabupaten Kediri. Preprocessing

2.2. Data Preprocessing

Data preprocessing merupakan tahapan penting dalam data mining [4], yang diperlukan untuk memperbaiki kualitas data yang dikumpulkan, sehingga dapat menghasilkan kualitas mining yang baik. Teknik-teknik yang digunakan pada tahap ini dapat berupa:

2.2.1. Replace Missing Values

Replace missing values difungsikan untuk estimasi data yang hilang atau kosong pada atribut-atribut atau fitur-fitur *dataset* dengan nilai rata-rata. Alat yang digunakan untuk teknik *replace missing values* adalah dengan menggunakan aplikasi *RapidMiner Studio* versi 7.3.

2.2.2. Data Transformation

Teknik *Discretization* merupakan bagian dari *data transformation* yang digunakan untuk merubah tipe data numerik atau kontinu kedalam interval [4], misalkan tinggi badan manusia (136-150, 151-165, 166-178) ditransformasi kedalam (Pendek, Sedang, Tinggi).

2.2.3. High Corelation Filter

High correlation filter merupakan bagian dari *fitur selection*, Formula *Pearson's product moment* [4], digunakan untuk mendapatkan fitur-fitur terbaik pada data rekam medis pasien dengan pengukuran tingkat korelasi pada fitur atau atribut dengan nilai numerik, berikut ini formula *Pearson's product moment*:

$$r_{xy} = \frac{N(\sum xy) - (\sum x)(\sum y)}{\sqrt{\{N(\sum x^2) - (\sum x)^2\}\{N(\sum y^2) - (\sum y)^2\}}} \quad (1)$$

Alat yang digunakan untuk melakukan fungsi *high correlation filter* adalah perangkat lunak SPSS 16.0. Sedangkan untuk atribut dengan tipe data nominal akan digunakan rumus *Pearson's chi-squared* [4], berikut formula dari *Pearson's chi-squared*:

$$\chi^2 = \sum_{j=1}^r \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \quad (2)$$

Dengan *fitur selection* dapat meningkatkan tingkat akurasi prediksi selain itu dengan menggunakan teknik *fitur selection* juga dapat mengidentifikasi fitur yang relevan [13].

2.3. Implementasi

Pada tahapan ini dilakukan implementasi algoritma CART, *Naive Bayes*, dan k-NN menggunakan *platform* visual basic 2010, sedangkan database dibangun menggunakan *Microsoft SQL Server R2*.

2.3.1. Algoritma CART

Algoritma CART merupakan teknik tanpa parametrik yang dikembangkan oleh Breiman, dkk (1984), dari algoritma CART menghasilkan antara klasifikasi dan pohon regresi, sehingga termasuk kedalam metode *Decision Tree*. Pemilahan pada metode algoritma CART dapat digunakan Gini Index [14], dengan formula:

$$i(t) = \sum_{i \neq j} p(i|t)p(j|t) \quad (3)$$

Untuk mendapatkan ukuran pohon yang tepat (*Goodness of splits*) [13], digunakan formula:

$$\Delta i(s, t) = i(t) - P_L i(t_L) - P_R i(t_R) \quad (4)$$

2.3.2. Algoritma Naïve Bayes

Klasifikasi *Bayesian* adalah pengklasifikasian statistik, bayesian mampu memprediksi kelas keanggotaan probabilistik [4]. Ide dasar dari *Naive Bayes* (NB) adalah menggunakan probabilitas gabungan dari kata-kata dan kategori untuk memperkirakan probabilitas dari kategori yang diberikan, secara umum persamaan yang digunakan untuk teorema *Bayes* adalah:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \quad (5)$$

Selama proses pelatihan harus dilakukan pembelajaran probabilitas akhir pada model untuk setiap kombinasi X dan Y berdasarkan data pelatihan. Dengan membangun model tersebut, suatu data uji X' dapat diklasifikasi dengan mencari nilai Y' dengan memaksimalkan nilai P(Y'|X') yang didapat, sehingga formulasi *Naive Bayes* untuk klasifikasi [15] adalah:

$$P(Y|X) = \frac{P(Y) \prod_{i=1}^q P(X_i|Y)}{P(X)} \quad (6)$$

2.3.3. Algoritma k-NN

Algoritma KNN selalu diawali dengan pencarian nilai k, kemudian dilakukan penghitungan pada data uji dan data latih [16]. Pada metode k-NN nilai k yang sering digunakan adalah ganjal [8, 15]. Penghitungan jarak terdekat antara data latih dan data uji secara umum digunakan formula *Euclidean Distance*, sehingga formulasi k-NN untuk klasifikasi adalah:

$$d_i = \sqrt{\sum_{i=1}^P (X_{1i} - X_{2i})^2} \quad (7)$$

X_1 merupakan data latih sedangkan X_2 merupakan data uji, i merupakan variabel data, P adalah dimensi data, dan d adalah jarak yang ingin diketahui, pada penelitian ini nilai k yang digunakan adalah k = 1, 3, 5, 7.

2.4. Analisis Data dan Pengujian

Pada metode k-NN diperlukan tipe data numerik agar dapat dilakukan perhitungan jarak *euclidian*, maka untuk tipe data kategorikal dilakukan perubahan ketipe data numerik seperti pada [11]. Pada metode *Naive Bayes* untuk tipe data kontinu digunakan formula *Gaussian Naive Bayes* [15]:

$$p(x = v|c) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}} \quad (8)$$

Pembagian dataset menggunakan metode *Holdout* [4], dimana 2/3 dari dataset dialokasikan sebagai data pelatihan dan sisanya 1/3 dialokasikan sebagai data uji. Setelah dilakukan pengujian terdapat data uji, maka langkah berikutnya adalah melakukan penghitungan terhadap tingkat performa metode *CART*, *Naive Bayes*, dan *k-NN*. Metode yang digunakan untuk penghitungan tingkat akurasi adalah *Confusion Matrix* [4]. Keluaran yang dihasilkan berupa nilai *accuracy*, *recall*, *precision*, *error rate*.

3. HASIL DAN PEMBAHASAN

3.1. Dataset

Hasil pengumpulan data dari Rumah Sakit Umum Daerah (RSUD) Kediri dan Yayasan Kanker Indonesia cabang Kabupaten Kediri, didapatkan jumlah data rekam medis pasien sebesar 702 *record*. Dari total 702 *record* terdapat 172 kasus positif dan 530 kasus negatif kanker serviks. Pada dataset tersebut terdapat 18 fitur dengan 1 fitur sebagai label kelas yaitu hasil test *Pap Smear*. Hasil deteksi menunjukkan terdapat *missing values* pada atribut atau fitur Berat Badan dan Tekanan Darah. Jumlah *missing values* pada fitur Berat Badan sebanyak 674 *record* atau 96%, sedangkan pada fitur Tekanan Darah sebanyak 32 *record* maka pada fitur Tekanan Darah akan dilakukan perbaikan dengan fungsi *replace missing values*. Pada fitur Berat Badan tidak akan digunakan untuk proses klasifikasi [4].

Fungsi *Discretization* [4] akan digunakan untuk fitur Tekanan Darah. Proses pembagian interval tekanan darah menggunakan standard dari [17], salah satu hasil dari komite bersama tersebut bahwa klasifikasi tekanan darah pada manusia dibagi seperti pada Tabel 1.

Tabel 1. Klasifikasi Tekanan Darah

Klasifikasi	Tekanan Darah (Sistolik/Diastolik)
Hipotensi	< 100/80
Normal	100/80 – 119/90
Pra Hipertensi	120/90 – 139/90
Hipertensi	≥ 140/90

Hasil dari *high correlation filter* didapatkan nilai korelasi antara atribut atau fitur sebagai *predictor* dengan atribut sebagai target. Hasil disajikan melalui Tabel 2.

Tabel 2. Hasil Uji Korelasi Dengan *Pearson's chi-squared*

Nama Atribut	Nilai <i>Asimp.Sig. (2-tailed)</i> terhadap atribut "Hasil"
Periode Menstruasi	0,000
Tekanan Darah	0,000
Bau Keputihan	0,000
Gatal-Gatal	0,013
Warna Keputihan	0,000
Jenis Kontrasepsi	0,105
Nyeri Pinggang	0,000
Nyeri Perut	0,000
Benjolan	0,000
Erosif (Lecet)	0,000
Berlendir (FL)	0,001
Pendarahan (FLX)	0,000

Hasil pada Tabel 2 merupakan nilai korelasi untuk atribut dengan tipe data nominal, sedangkan untuk tipe data numerik, hasil korelasi disajikan melalui Tabel 3.

Tabel 3. Hasil Uji Korelasi Dengan *Pearson's Product Moment*

Nama Atribut	Nilai <i>Asimp.Sig. (2-tailed)</i> terhadap atribut "Hasil"
Usia	0,000
Jumlah Anak	0,000
Lama Pemakaian Alat Kontrasepsi	0,248

Korelasi antara variabel X dan Y dikatakan mempunyai hubungan atau korelasi yang signifikan jika nilai *Sig. (2-tailed)* < 0,05, maka dari Tabel 2 dan 3 atribut dengan nilai korelasi rendah adalah Jenis Kontrasepsi dan Lama Pemakaian Alat Kontrasepsi. Atribut dengan korelasi yang rendah tidak akan digunakan untuk proses selanjutnya.

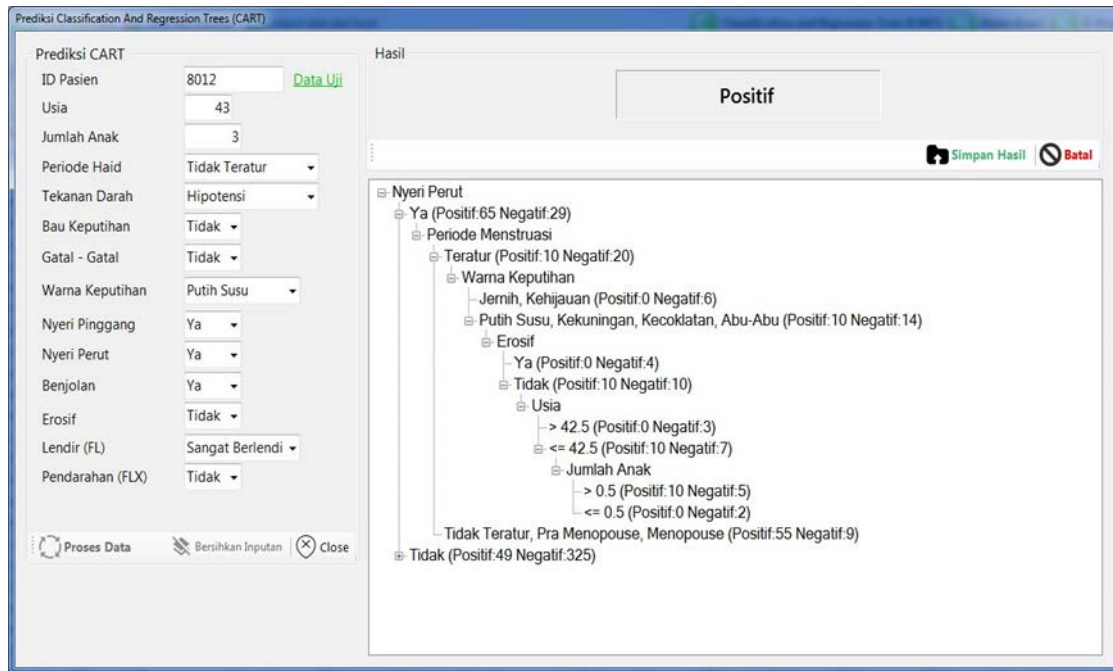
Hasil dari *data preprocessing* diperoleh atribut-atribut atau fitur-fitur rekam medis pasien yang akan digunakan untuk proses klasifikasi. Atribut-atribut tersebut dideskripsikan pada Tabel 4.

Tabel 4. Deskripsi dari Fitur-Fitur *Dataset* Penyakit Kanker Serviks

No.	Nama	Deskripsi
1	ID Pasien	Nomer ID Pasien
2	Usia	Usia dalam tahun
3	Jumlah Anak	Jumlah anak pasien
4	Periode Menstruasi	Teratur, Tidak Teratur, Pra Menopause, Menopause
5	Tekanan Darah	Hipotensi, Normal, Pra Hipertensi, Hipertensi
6	Bau Keputihan	Tidak, Ya
7	Gatal-Gatal	Tidak, Ya
8	Warna Keputihan	Jernih, Putih Susu, Kekuningan, Kecoklatan, Kehijauan, Abu-abu
9	Nyeri Pinggang	Tidak, Ya
10	Nyeri Perut	Tidak, Ya
11	Benjolan	Tidak, Ya
12	Erosif (Lecet)	Tidak, Ya
13	FL (Berlendir)	Tidak, Berlendir, Sangat Berlendir
14	FLX (Pendarahan)	Tidak, Ya
15	<i>Class</i> (Hasil Tes <i>Pap Smear</i>)	Negatif, Positif

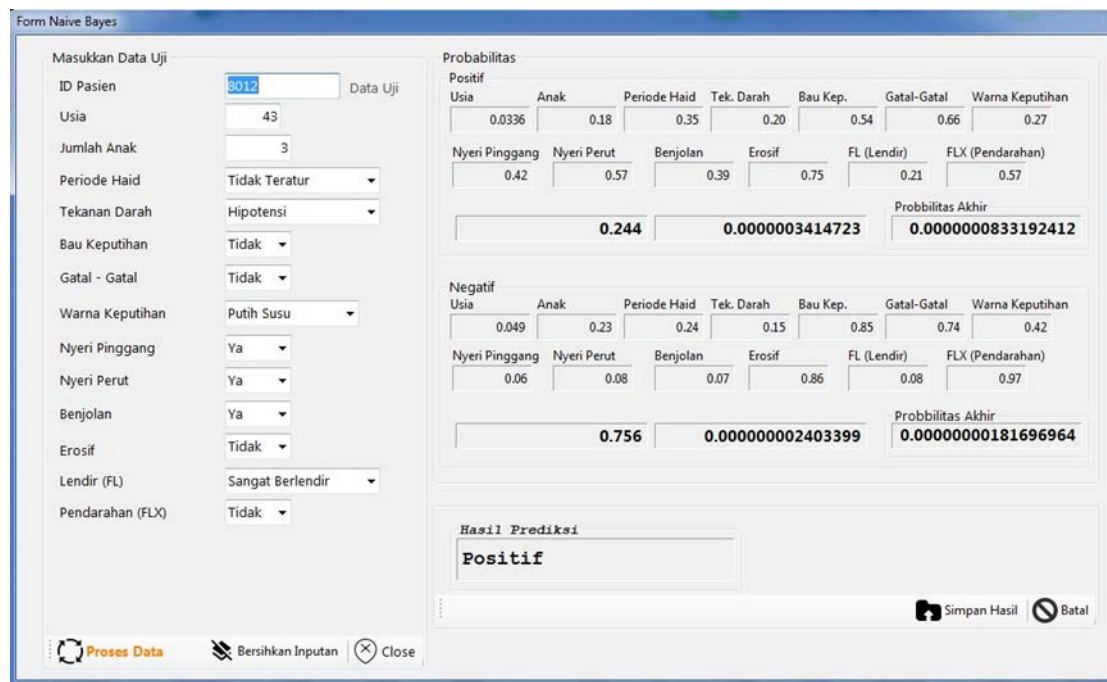
3.2. User Inteface

Pada Gambar 2 merupakan menu prediksi CART berfungsi untuk melakukan prediksi menggunakan algoritma CART Decision Tree, bentuk rule divisualisasikan melalui treeview.



Gambar 2. Tampilan menu prediksi CART

Menu prediksi Naive Bayes pada Gambar 3 difungsikan untuk melakukan prediksi dengan menggunakan algoritma Naive Bayes.



Gambar 3. Tampilan menu prediksi Naive Bayes

Menu prediksi k-NN pada Gambar 4 digunakan untuk melakukan prediksi menggunakan algoritma k-NN.

Gambar 4. Tampilan menu prediksi k-NN

3.3. Pengujian Algoritma

Pengujian algoritma CART, *Naive Bayes*, dan k-NN menggunakan data uji sebanyak 234 record dan pengukuran performa menggunakan formula *Confusion Matrix*. Pada algoritma CART didapatkan hasil seperti tertera pada Tabel 5.

Tabel 5. Hasil Pengujian Metode CART Decision Tree

Jumlah Data Uji	<i>Accuracy</i>	<i>Recall</i>	<i>Precision</i>	<i>Error Rate</i>
234	88,89%	91,38%	71,62%	11,11%

Algoritma CART yang menunjukkan tingkat akurasi dalam melakukan prediksi hasil tes *Pap Smear* sebesar 88,89 % sedangkan hasil pengujian yang didapatkan pada algoritma *Naive Bayes*, disajikan melalui Tabel 6.

Tabel 6. Hasil Pengujian Metode Naive Bayes

Jumlah Data Uji	<i>Accuracy</i>	<i>Recall</i>	<i>Precision</i>	<i>Error Rate</i>
234	94,44%	81,03%	95,92%	5,56%

Algoritma *Naive Bayes* menghasilkan performa tingkat akurasi sebesar 94,44%, sedangkan pada algoritma k-NN didapatkan hasil seperti pada Tabel 7.

Tabel 7. Hasil Pengujian Algoritma k-NN

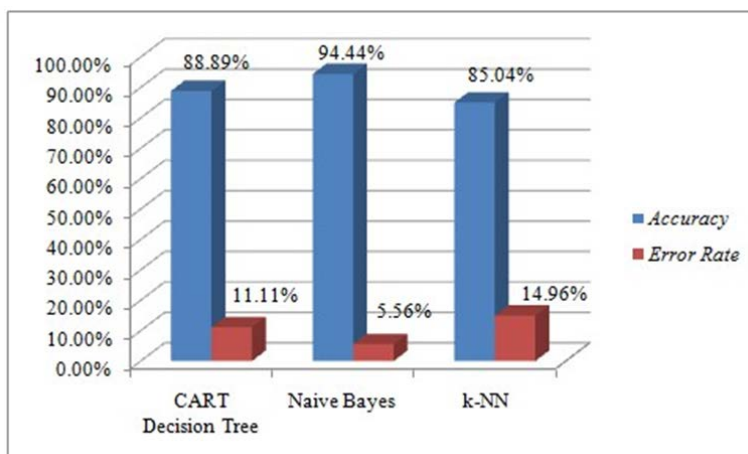
Jumlah Data Uji	Nilai k	Accuracy	Recall	Precision	Error Rate
234	1	85,04%	46,55%	87,10%	14,96%
234	3	85,04%	44,83%	89,66%	14,96%
234	5	85,04%	44,83%	89,66%	14,96%
234	7	83,33%	39,66%	85,19%	16,67%
Rata-rata		84,61%	43,97%	87,90%	15,39%

Performa algoritma k-NN didapatkan presentase tingkat akurasi yang sama sebesar 85,04% pada nilai k = 1, 3, dan 5, dan nilai rata-rata tingkat akurasi sebesar 84,61%. Perbandingan tingkat akurasi yang didapat algoritma CART, Naive Bayes, dan k-NN pada Tabel 8 dan gambar 5 menunjukkan tingkat akurasi terbaik dihasilkan algoritma Naive Bayes.

Tabel 8. Perbandingan Tingkat Akurasi

Jumlah Data Uji	CART Decision Tree	Naive Bayes	k-NN
234	88,89%	94,44%	85,04%

Performa algoritma k-NN dalam kasus ini tidak lebih baik jika dibandingkan dengan algoritma Naive Bayes, hal yang sama pada kasus [11, 19]. Pada kasus [20] performa algoritma k-NN juga tidak lebih baik dibandingkan algoritma Decision Tree, namun dalam kasus tersebut algoritma k-NN memiliki performa lebih baik dibanding dengan Naive Bayes.



Gambar 5. Grafik perbandingan tingkat akurasi dan kesalahan

4. KESIMPULAN

Teknik klasifikasi dalam data mining sangat efektif digunakan untuk melakukan pekerjaan prediksi dalam kasus penyakit kanker serviks dengan menggunakan dataset tes *Pap Smear*. Melalui hasil penelitian juga didapatkan beberapa poin kesimpulan sebagai berikut:

1. Proses implementasi dari tiga algoritma yang digunakan, algoritma CART Decision Tree memerlukan lebih banyak waktu dalam membangun pohon keputusan atau *classification rules*.
2. Sistem prediksi kanker serviks dapat digunakan untuk mendukung tenaga medis dalam memberikan keputusan klinis hasil tes *Pap Smear* pada pasien baru.

3. Algoritma *Naive Bayes* dengan teknik probabilistik mampu dengan baik melakukan klasifikasi terhadap kasus positif dan negatif kanker serviks, serta menghasilkan tingkat akurasi yang tinggi.
4. Algoritma k-NN dengan teknik klasifikasi sederhana pada kasus prediksi penyakit kanker serviks tidak dapat bekerja secara baik pada klasifikasi kasus positif.
5. Melalui hasil pengujian dapat diketahui bahwa algoritma yang termasuk dalam eager learner (*CART Decision Tree, Naive Bayes*) memiliki performa yang lebih baik dibandingkan lazy learner (k-NN)

5. SARAN

Pada penelitian masih terdapat beberapa kelemahan, sehingga dapat dilakukan perbaikan pada penelitian selanjutnya. Adapun saran untuk penelitian selanjutnya adalah sebagai berikut:

1. Pada penelitian ini tidak dilakukan deteksi dan *remove* data *outlier*, karena keterbatasan jumlah dataset. Dengan dilakukan *remove* data *outlier* dapat meningkatkan hasil klasifikasi.
2. Teknik pembagian dataset *Holdout* lebih efektif dengan jumlah dataset yang lebih besar, sehingga pada penelitian selanjutnya dapat dilakukan penambahan jumlah dataset.
3. Performa pada algoritma k-NN dapat ditingkatkan melalui teknik pembobotan pada setiap atribut, sehingga dapat diketahui tingkat kepentingan setiap atribut atau fitur.
4. Pada penelitian selanjutnya dapat dilakukan pengukuran dengan metode lainnya.
5. Pada penelitian selanjutnya dapat dilakukan penambahan atribut berupa faktor genetik dan riwayat keluarga.

DAFTAR PUSTAKA

- [1] Primadi, O., 2015, Situasi Penyakit Kanker – Kementerian Kesehatan, <http://www.depke.go.id/download.php?file=download/pusdatin/buletin/buletin-kanker.pdf>, diakses pada tanggal 31 Oktober 2016.
- [2] American Cancer Society, 2014, Cervical Cancer Prevention and Early Detection, <http://www.cancer.org/acs/groups/cid/documents/webcontent/003167-pdf.pdf>, diakses pada tanggal 4 November 2016.
- [3] Ramachandran, P., Girija, N., Bhuvaneswari, T., 2014, Early Detection and Prevention of Cancer using Data Mining Techniques, *International Journal of Computer Applications*, No. 13, Vol. 97.
- [4] Han, J., Kamber, M., Pei, J., 2012, *Data Mining Concepts and Techniques*, Ed. 3, Morgan Kaufmann, USA.
- [5] Kumar, D. S., Sathyadevi, G., Sivanesh, S., 2011, Decision Support System for Medical Diagnosis Using Data Mining, *International Journal of Computer Science Issue*, No. 3, Vol. 8
- [6] Venkatesan, E., Velmurugan, T., 2015, Performance Analysis of Decision Tree Algorithms for Breast Cancer Classification, *Indian Journal of Science and Technology*, No. 29, Vol. 8
- [7] Jabbar, M. A., Deekshatulu, B. L., Chandra, P., 2013, Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm, *International Conference on Computational Intelligence: Modeling Techniques and Applications*. Hal 85-94
- [8] Shouman, M., Turner, T., Stocker, R., 2012, Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients, *International Journal of Information and Education Technology*, No. 3, Vol 2.

-
- [9] Ratnam, D., Bindu, P. H., Sai, V. M., Devi, S. P. R., Rao, P. R., 2014, Computer-Based Clinical Decision Support System for Prediction of Heart Diseases Using Naive Bayes Algorithm, *International Journal of Computer Science and Information Technologies*, No.2, Vol. 5.
- [10] Medhekar, D. S., Bote, M. P., Deshmukh, S. D., 2013, Heart Disease Prediction System using Naive Bayes, *International Journal of Enhanced Research in Science Technology and Engineering*, No. 3, Vol. 2.
- [11] Swati, S., Priyadarshi, A., 2015, Decision Support System on Prediction of Heart Disease Using Data Mining Techniques, *International Journal of Engineering Research and General Science*, No. 3, Vol. 3.
- [12] Xindong, W., Kumar, V., 2009, *The Top Ten Algorithms in Data Mining*, Taylor & Francis Group, New York.
- [13] Lei, Y. C., Huan, L., 2003, Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution, *Proceedings of the Twentieth International Conference on Machine Learning*.
- [14] Timofeev, R., 2004, Classification and Regression Trees (CART) Theory and Applications, *Thesis*, Center of Applied Statistics and Economics, Univ. Humboldt, Berlin.
- [15] Prasetyo, E., 2012, *Data Mining Konsep dan Aplikasi Menggunakan Matlab*, Andi Offset, Yogyakarta.
- [16] Nugroho, A., Kusriani, Arief, M., R., 2014, Sistem Pendukung Keputusan Kredit Usaha Rakyat PT. Bank Rakyat Indonesia Unit Kaliangkrik Magelang, *Creative Information Technology Journal (CITEC)*, No.1, Vol. 2. <http://citec.amikom.ac.id/main/index.php/citec/article/viewFile/33/33.pdf>.
- [17] The Seventh Report of The Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure, <https://www.nhlbi.nih.gov/files/docs/guidelines/jnc7full.pdf>, diakses tgl 8 Desember 2016.
- [18] Meiyar, V. K. The Comparative Study for Diagnosing Heart Disease using KNN and Naive Bayes, *International Journal of Advance Research in Computer Science and Management Studies*, No. 8, Vol. 3.
- [19] Ani, R., Sasi, G., Sankar, U. R., Deepa, O. S., 2016, Decision Support System for Diagnosis and Prediction of Chronic Renal Failure using Random Subspace Classification, *Advances in Computing, Communications and Informatics, 2016 International Conference on*, Jaipur, 21-24 September 2016.
-