Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 2, No. 8, Agustus 2018, hlm. 2430-2434 http://j-ptiik.ub.ac.id

Implementasi Metode *Naïve Bayes* Dengan Perbaikan *Missing Value* Menggunakan Metode *Nearest Neighbor Imputation* Studi Kasus: Penyakit Malaria Di Kabupaten Malang

Riyant Fajar[1], Rizal Setya Perdana[2], Indriati[3]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]riyantfajar@yahoo.com, [2]rizalespe@ub.ac.id, [3]indriati.tif@ub.ac.id

Abstrak

Malaria adalah penyakit yang dapat ditularkan melalui gigitan nyamuk *Anopheles* betina. Ada empat tipe *Plasmodium* parasit yang sering ditemui pada kasus penyakit malaria di Indonesia yaitu *Plasmodium vivax (Tertiana*) dan *Plasmodium malariae (Quartana),* sedangkan malaria lainnya adalah *Plasmodium falcifarum (Tropica)* dan *Plasmodium ovale (Pernisiosia).* Gejala penyakit malaria selama ini hanya dikenal masyarakat awam melalui ciri-ciri yang diketahui tanpa oleh fakta dan pertimbangan medis lainnya. Sehingga masyarakat atau penderita mengalami kesulitan dalam membedakan penyakit malaria dengan penyakit demam, atau influenza. Akibatnya penyakit tersebut ditangani dengan cara yang salah. Gejala demam pada penyakit malaria tergantung pada jenis malaria. Sifat demam akut yang didahului oleh stadium dingin atau menggigil diikuti demam tinggi kemudian berkeringat banyak. Gejala klasik ini biasanya ditemukan pada penderita non imun (berasal dari daerah non *endemis*). Selain gejala klasik tersebut, dapat pula ditemukan gejala lain seperti nyeri kepala, mual, muntah, diare, dan nyeri otot. Penyakit malaria dapat menyebabkan anemia dan dapat mengakibatkan kematian. Pada wanita hamil dapat menyebabkan keguguran, lahir prematur dan berat badan lahir rendah, bahkan lahir mati. Oleh sebab itu dibutuhkan sistem komputer agar lebih cepat dalam mendeteksi gejala-gejala yang dialami. Sistem ini dibangun menggunakan metode *naïve bayes* dengan perbaikan *missing value* menggunakan metode *nearest neighbor imputation*. Hasil akurasi dari 2 skenario pengujian didapat akurasi terbaik sebesar 77,14% pada skenario pengujian 1 dan 64,70% pada skenario pengujian 2.

Kata kunci: *naïve bayes, nearest neighbor imputation, malaria, missing value*

Abstract

Malaria is an infectious disease that is transmitted among humans by the bites of female Anopheles mosquit. There are four types of Plasmodium that are frequently found in the case of malarial infection in Indonesia: Plasmodium vivax (Tertiana), Plasmodium malariae (Quartana), Plasmodium falcifarum (Tropica), and Plasmodium ovale (Pernisiosia). Thus far, people are having difficulty in differentiating the symptoms found in malaria and in another common cold or influenza as the laymen rely only on general knowledge without any medical facts and reviews. As a result, the patient of malaria is often mistreated. The symptoms of malaria depend on the types of malaria itself. Classic symptoms of malaria suffered by non-immune patients (patients who live in non-endemic area) are paroxysmal (sudden acute fever) preceded by chills and oversweating. On the other hand, classic symptoms of malaria suffered by immune patients are headache, nausea and vomitting, diarrhea , as well as muscle pain. Malaria is a life-threatening disease that can lead into death if not treated in an immediate manner. On that account, a computer system that can accelerate the detection is needed to help in diagnosing whether or not the patient is infected. The said system was designed using Naïve Bayes method and the improvement of missing value with the usage of nearest neighbor imputation method. The verdict of the system's accurateness from two testing scenarios has been acquired with the best accuracy point of 77.14% in the first testing scenario and 64.70% in the second testing scenario.

Keywords: *naïve bayes, nearest neighbor imputation, malaria, missing value*

1. PENDAHULUAN

Malaria adalah penyakit menular yang disebabkan oleh parasit (protozoa) dari genus plasmodium, yang dapat ditularkan melalui gigitan nyamuk Anopheles betina. Plasmodium merupakan genus protozoa parasit. Penyakit ini dikenal sebagai malaria. Penyakit ini juga mempunyai nama lain, seperti demam roma, demam rawa, demam tropik, demam pantai, demam charges, demam kura dan paludisme (Prabowo, 2004).

Ada empat tipe Plasmodium parasit yang sering ditemui pada kasus penyakit malaria di Indonesiayang dapat menginfeksi manusia yaitu Plasmodium vivax (Tertiana) dan Plasmodium malariae (Quartana), sedangkan malaria lainnya adalah Plasmodium falcifarum (Tropica) dan Plasmodium ovale (Pernisiosia). Di dunia ini hidup sekitar 400 spesies nyamuk Anopheles, tetapi hanya 60 spesies yang berperan sebagai vector malaria alami. Di Indonesia, ditemukan 80 spesies nyamuk Anopheles tetapi hanya 16 spesies sebagai vektor malaria (Prabowo, 2004).

Di Kabupaten Malang terdapat program eliminasi penyakit Malaria yang digalakkan oleh Dinas Kesehatan. Oleh sebab itu dibutuhkan sebuah sistem komputer agar dapat lebih cepat dalam mendeteksi gejala-gejala yang dialami, apakah terdiagnosa sebagai malaria atau tidak. Agar dapat segera diketahui dan ditangani secepatnya. Sistem aplikasi yang akan dibangin ini diharapkan dapat mengetahui tingkat akurasi dari sebuah metode agar dapat diketahui kualitasnya, dengan harapan bias berguna untuk masyarakat dalam mengetahui atau mendeteksi penyakit malaria dengan cepat.

2. KAJIAN PUSTAKA DAN DASAR TEORI

2.1. Kajian Pustaka

Penelitian sebelumnya yang berjudul "Sistem Pakar Diagnosa Penyakit Pada Sapi Potong Dengan Metode Naive Bayes (studi kasus di pos keswan Kab. Nganjuk)". Oleh Indriana Candra Dewi. Metode yang digunakan untuk menunjang keputusan pada diagnosis penyakit pada sapi potong adalah Naive Bayes. Sistem pakar ini mendiagnosa penyakit pada sapi potong dengan inputan memilih gejala (Dewi, 2015).

Penelitian sebelumnya yang berjudul Perbandingan Metode *Naïve Bayes* Dan *K-Nearest Neighbor* Untuk Klasifikasi Mutu Susu Sapi" oleh Suryandi. Metode yang digunakan pada penelitian ini adalah *Naïve Bayes* dan *K-Nearest Neighbor*. Penelitian ini membandingkan antara kedua metode tersebut, dimana kedua metode tersebut merupakan metode yang cukup baik dalam pengklasifikasian dan dapat menghasilkan tingkat akurasi yang tinggi (Suryadi, 2017).

Penelitian sebelumnya yang berjudul "Klasifikasi Jenis Kelamin Berdasarkan Nama Pengguna *Twitter* Menggunakan *Metode Naïve Bayes Classifier*" oleh Gagah Istaid Billah. Metode yang digunakan adalah *Naïve Bayes*. Pada sistem ini akan dilakukan pengelompokan jenis kelamin pengguna social media *twitter* berdasarkan nama, agar memudahkan penyebaran informasi untuk mendapatkan sasaran promosi yang tepat (Billah, 2016).

Penelitian sebelumnya yang berjudul "Penanganan *Missing Value* Dengan Algoritma *Weighted KNNI* Pada Data Kategori" oleh Akhmad Itsnaini Setyawan. Penelitian tersebut membahas tentang penyelesaian masalah nilai yang hilang (*missing value*) (Setyawan, 2013)

2.2. *Naïve Bayes*

Thomas Bayes menemukan suatu pendekatan untuk melakukan penalaran statistik yang jauh lebih maju dibandingkan dengan pola pikir sistematis tradisional pada waktu itu, fokus matematika pada waktu itu adalah pada suatu sampel dari populasi yang diketahui. Akan tetapi Bayes mengembangkan ide untuk menemukan properti dari populasi berdasarkan sampel tersebut dalam "*An essay towards the solving a problem in the doctrines of chance"* dia menyajikan tentang "*Proposition 9"* yang akhirnya dikenal dengan "*Teorema Bayes".* Selanjutnya teorema ini menjadi dasar dalam pengambilan keputusan. Rumus bayes secara umum dinyatakan dalam persamaan 1 (Prasetyo, 2012).

$$P(H|E) = \frac{P(E|H).P(H)}{P(E)} \quad (1)$$

Dimana:

P(H|E) = Probabilitas posterior bersyarat (*Conditional Probability*) suatu hipotesis H terjadi jika diberikan evidence/bukti E terjadi.

P(E|H) = Probabilitas sebuah evidence E terjadi akan mempengaruhi hipotesis H.

P(H) = Probabilitas awal (priori) hipotesis H terjadi tanpa memandang evidence

apapun.

P(E) = Probabilitas awal (priori) evidence E terjadi tanpa memandang hipotesis/evidence yang lain.

2.3. *Missing Value*

Menurut D. B. Rubin dan R. J. Little (2002), biasanya metode untuk mengatasi *missing value* dapat dibagi menjadi 3 kategori, yaitu, Parameter *Estimation*, *Case/Pairwise Deletion*, dan teknik Imputasi. Pada metode Parameter *Estimation*, dilakukan tahap-tahap *Maximum Likelihood* dan digunakan algoritma *Expetation – Maximization* dengan tujuan memperkirakan suatu nilai dari *missing value*. Pada metode *Case/Pairwise Deletion*, diperlakukan penghapusan terhadap *record* dataset yang terdapat *missing value* pada variabelnya. Pada metode Teknik Imputasi, *missing value* akan diganti dengan suatu nilai perkiraan yang berdasar pada informasi yang didapat dalam dataset.

2.4. *Nearest Neighbor Imputation*

Menurut Olivas, dkk (2010), *K-Nearest Neighbour Imputation* termasuk dalam *Machine Learning Solution* pada teknik imputasi. Pada metode ini, *missing value* pada suatu data akan ditangani dengan melakukan suatu imputasi dengan memperkirakan nilai yang didapatkan dari *record* yang paling serupa. Langkah-langkah *KNNI* adalah sebagai berikut:

1. *Dataset* dibagi menjadi dua bagian, yaitu *data complete* dan *data missing*. Apabila terdapat sedikitnya satu nilai yang kosong pada suatu *record*, data tersebut dikelompokkan ke dalam *data missing*.
2. Untuk tiap jarak dalam *data missing:*
   a. Menghitung jarak antara *record* data *missing* dengan setiap *record* data *complete* dengan mengabaikan atribut yang terdapat *missing value*. Langkah ini dilakukan pada semua *record* data *missing.*
   b. Untuk atribut kualitatif, digunakan K vector *record* yang paling dekat dan melakukan *voting* terhadap perkiraan nilai yang hilang. Sedangkan pada atribut numerik, *missing value* diganti dengan nilai rata-rata dari atribut pada K vector *record* yang terdekat.

3. METODOLOGI

Proses perancangan bertujuan untuk menjelaskan kerja sistem yang akan dibangun dan pengklasifikasian menggunakan metode *Naïve Bayes*. Proses awal adalah dengan membaca data *training,* kemudian akan dilanjutkan dengan *preprocessing.* Pada tahap *preprocessing* ini, nilai data yang kosong akan diisi dengan nilai yang didapat dari perhitungan menggunakan metode *Nearest Neighbor Imputation.* Setelah data yang dibutuhkan lengkap, dilakukan pengklasifikasian menggunakan metode *Naïve Bayes.* Contoh perhitungan manual juga disertakan guna untuk mengetahui bagaimana proses perhitungan yang ada pada metode *Nearest Neighbor Imputation* dan *Naïve Bayes*. Gambaran proses hingga akhir proses dapat dilihat pada Gambar 1.

Gambar 1 Gambaran Perancangan Algoritme

Pada tahap pengumpulan data, yang dibutuhkan adalah definisi penyakit, gejala – gejala yang timbul, dan diagnosis jenis penyakit malaria apa yang diderita oleh pasien yang menderita penyakit *malaria.* Sumber data diperoleh dari beberapa kali hasil wawancara yang dilakukan dengan Tenaga Medis dari Dinas Kesehatan Kabupaten Malang yaitu Fida Retno S.Kep. Dari hasil wawancara dengan beliau, didapatkan data pengetahuan tentang penyakit *Malaria* serta gejal-gejala dan jenis malaria pada tiap pasien guna perhitungan klasifikasi menggunakan metode *Naïve Bayes*.

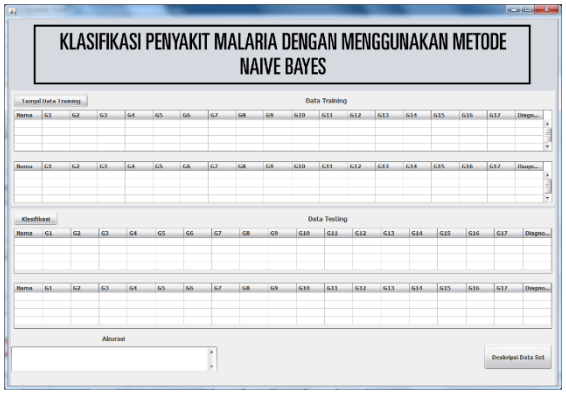
4. IMPLEMENTASI

Batasan Implementasi adalah batasan

proses yang dapat dilakukan oleh sistem berdasarkan perancangan awal. Beberapa batasan dalam implementasi aplikasi Diagnosis Penyakit Malaria Menggunakan Metode *Naïve Bayes* adalah sebagai berikut:

1. Aplikasi Diagnosis Penyakit Malaria dengan menggunakan metode *Naïve Bayes* dirancang dan dijalankan berbasis *Java*.
2. Masukan yang diterima oleh system adalah berupa data fakta gejala penyakit malaria pada pasien di Kabupaten Malang.
3. Keluaran yang diterima oleh pengguna berupa hasil akurasi data klasifikasi penyakit malaria.

Untuk tujuan pengujian, telah dibangun sebuah aplikasi yang bertujuan untuk menghitung hasil akurasi dari metode yang digunakan. Aplikasi tersebut mendapat inputan berupa data gejala-gejala yang diderita oleh pasien malaria dan menunjukkan hasil akurasi dari proses klasifikasi. Antarmuka aplikasi dapat dilihat pada gambar 2.



Gambar 2 Antarmuka Aplikasi

5. PENGUJIAN DAN HASIL

5.1. Perancangan Pengujian

Pada tahap ini, masing-masing dari skenario pengujian tersebut akan dilakukan lima skenario komposisi data. Dengan pembagian terhadap dataset menjadi dua bagian, yaitu data *training* dan data *testing*. Pada skenario komposisi data pertama, dataset akan dibagi dengan komposisi 20% data *training* dan 80% data *testing*. Pembagian komposisi data *training* kedua dengan komposisi 40% data *training* dan 60% pada data *testing*. Pembagian komposisi dengan data ketiga data training dan data testing masing-masing dibagi menjadi 50%. Pembagian data *training* sebesar 60% dan data *testing* sebesar 40% dilakukan pada komposisi keempat. Serta data *training* sebesar 80% dan data *testing* sebesar 20% pada komposisi terakhir. Pengambilan data pada tiap skenario akan dilakukan secara acak. Tujuan dari dilakukannya lima skenario tersebut adalah untuk mengetahui komposisi mana yang paling baik dalam penentuan akurasi pada proses klasifikasi. Dan dilakukan perbandingan skenario pengujian mana yang menghasilkan akurasi lebih baik.

5.2. Hasil Dan Analisis

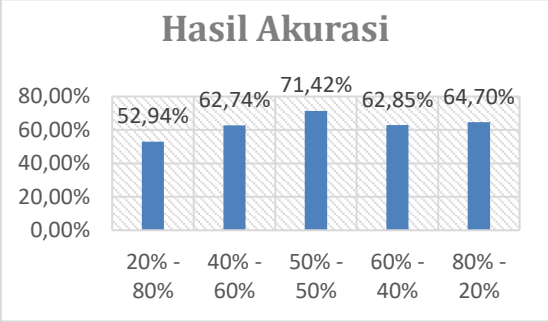
5.2.1. Hasil Pengujian Skenario 1

Dalam bab ini akan dibahas mengenai langkah-langkah pengujian ataupun analisis implementasi metode *Nearest Neighbor* untuk perbaikan *missing value* dan *Naïve Bayes* pada kasus penyakit malaria.

Pada pengujian pertama dilakukan perbaikan terhadap data yang kosong. Data yang kosong akan disimbolkan dengan tanda (-). Hasil pengujian skenario 1 dapat dilihat pada tabel 1. untuk mengetahui perbandingan akurasi pada masing-masing skenario komposisi data, dibuatlah grafik sebagaimana ditampilkan pada gambar 3.

Tabel 1 Hasil Pengujian Skenario 1

Komposisi Data		Hasil Akurasi
Data Training	Data Testing	
20%	80%	63.23%
40%	60%	70.58%
50%	50%	76.19%
60%	40%	77.14%
80%	20%	64.70%



Gambar 3 Grafik Hasil Pengujian Skenario 1

Berdasarkan hasil pengujian skenario 1 didapatkan hasil akurasi yang paling baik yaitu 77,14% pada komposisi data keempat dengan 60% data training dan 40% data testing. Didapatkan pula hasil akurasi yang paling rendah adalah 63,23% pada komposisi data pertama dengan 20% data *training* dan 80% data

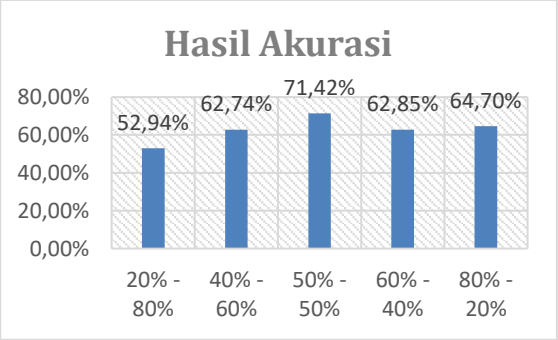
testing.

5.2.2. Hasil Pengujian Skenario 2

Tahap pengujian selanjutnya hampir sama dengan pengujian sebelumnya. Namun, yang membedakan adalah tidak dilakukan perbaikan terhadap data yang kosong atau *missing value*. Data yang kosong akan diganti dengan nilai baru, dalam hal ini adalah String "X". Hasil pengujian skenario 2 dapat dilihat pada tabel 2.

Tabel 2 Hasil Pengujian Skenario 2

Komposisi Data		Hasil Akurasi
Data Training	Data Testing	
20%	80%	52.94%
40%	60%	62.74%
50%	50%	71.42%
60%	40%	62.85%
80%	20%	64.70%



Gambar 4 Grafik Hasil Pengujian Skenario 2

Dari pengujian skenario 2 didapatkan hasil akurasi yang terbaik dengan nilai 71,42% pada komposisi data ketiga yaitu dengan pembagian masing-masing 50% untuk data training dan data testin, dan hasil akurai yang paling rendah atau buruk adalah pada komposisi data pertama dengan nilai 52,94%, komposisi data 20% untuk data *training* dan 80% untuk data *testing*.

6. KESIMPULAN

Berdasarkan hasil pengujian dari penelitian yang telah dilakukan, dapat ditarik beberapa kesimpulan yaitu metode klasifikasi *naïve bayes* dengan perbaikan *missing value* menggunakan metode *nearest neighbor imputation* telah diimolementasikan terhadap data penyakit mlaria. Metode perbaikan terhadap *missing value* yaitu *nearest neighbor imputation* digunakan pada tahap *preprocessing*, dan meode *naïve bayes* digunakan pada tahap selanjutnya untuk proses klasifikasi.

Rata-rata tingkat akurasi dari metode klasifikasi *naïve bayes* dengan dilakukan perbaikan terhadap *missing value* lebih baik dibanding dengan tanpa dilakukan perbaikan. Hasil akurasi proses klasifikasi dari proses klasifikasi menggunakan metode *naïve bayes* sangant dipengaruhi oleh proses perbaikan terhadap *missing value*.

DAFTAR PUSTAKA

Billah, Gagah, Istaid, 2016. *Klasifikasi Jenis Kelamin Berdasarkan Nama Pengguna Twitter Menggunakan Metode Naïve Bayes Classifier*. Universitas Brawijaya, Malang.

Dewi, Indriana, Candra. 2015. *Sistem Pakar Diagnosa Penyakit Pada Sapi Potong Dengan Metode Naïve Bayes (Studi Kasus Pos KESWAN Kab. Nganjuk*). Universitas Brawijaya, Malang.

Little, R. J. and Rubin, D. B. 2002. *Statistical Analysis with Missing Data Second Edition*. John Wiley and Sons, New York.

Olivas, Emilio Soria. 2010. *Handbook of Research on Machine Learning Applications and Trends*. IGI Global.

Prabowo, Arian. 2008. *Malaria: Mencegah Dan Mengatasinya*. Puspa Swara, Jakarta.

Prasetyo, Eko. 2012. *Data Mining – Konsep Dan Aplikasi Menggunakan MATLAB*. Yogyakarta: Andi.

Setyawan, Akhmad Itsnaini. 2013. *Penanganan Missing Values Dengan Algoritma Weighted KNNI Pada Data Kategori*. Universitas Brawijaya, Malang.

Suryadi, 2017. *Perbandingan Metode Naïve Bayes Dan K-Nearest Neighbor Untuk Klasifikasi Mutu Susu Sapi*. Universitas Brawijaya, Malang.