# Perbandingan Algoritma Klasifikasi Data Mining Untuk Prediksi Penyakit Kulit

**Dian Prajarini**
Sekolah Tinggi Seni Rupa Dan Desain Visi Indonesia
dianpraja@gmail.com

**ABSTRACT**
Skin diseases often thought as a insignificant problem compared to diseases with high mortality, such as HIV/AIDS, ischemic heart disease, stroke, tuberculosis, malaria and cancer. However, skin diseases are some of the most common diseases seen in the developing country. Skin diseases diagnosis and treatment is important because there are a lot of skin diseases that shows similar symptoms and screening for sign of skin disease is an important way for systemic diseases. Prediction of skin disease is diffucult because of a lot of skin disease shows similar symptoms. Data mining with clasification algorithm can be used to predict skin disease. Four data mining clasification algorithm C4.5, Naive Bayes, KNN and SVM was used for data analysis. Medical dataset used in this research is dataset download from UCI repository site. From the tes result by comparing accuration, precision dan recall, it is known that Naive Bayes has the highest accuracy and precision.

***Keyword:*** Clasification, C4.5, Naive Bayes, KNN, SVM, Data Mining, Skin Diseases

## 1. Introduction

Penyakit kulit sering dianggap sebagai masalah yang sepele bagi sebagian orang jika dibandingkan dengan penyakit yang menyebabkan kematian seperti HIV/AIDS, penyakit jantung iskemik, stroke, tuberkulosis, malaria dan kanker. Meskipun penyakit kulit dianggap sepele, penyakit kulit merupakan salah satu penyakit yang sering dijumpai di negara-negara berkembang. Diagnosis dan pengobatan penyakit kulit sangat penting pada negara berkembang, hal ini disebabkan karena banyaknya jumlah pasien penderita penyakit kulit, jumlah kasus yang sangat signifikan dalam hal kesalahan deteksi, ketidakmampuan serta gejala gatal yang tidak bisa dilacak dan skrining awal terhadap penyakit sistemik seperti leprosy dan HIV[5]. Meskipun penyakit kulit bisa merepresentasikan penyakit dalam yang serius, petugas kesehatan memiliki kekurangan tentang pengetahuan dasar ilmu penyakit kulit[4].

Sebuah diagnosis medis merupakan sebuah proses klasifikasi dimana dokter akan menganalisis banyak faktor sebelum menentukan diagnosis, umumnya menjadi masalah yang sulit [3]. Data mining telah terbukti menjadi pendekatan yang kuat dan efektif yang menyediakan proses penemuan pola dalam dataset besar[9]. Teknik klasifikasi yang disediakan oleh data mining bisa digunakan untuk membantu melakukan proses diagnosis atau prediksi penyakit kulit yang diderita oleh pasien dengan gejala tertentu. Berbagai macam teknik data mining digunakan untuk memprediksi penyakit [1]. Algoritma KNN digunakan untuk prediksi penyakit jantung [10].

Penelitian untuk deteksi penyakit kulit juga sudah dilakukan oleh beberapa peneliti. Penggabungan teknik Bayesian dan Best First Search digunakan untuk prediksi penyakit kulit dengan tingkat akurasi 99,31% [15]. Teknik Naive Bayes dan Decision Tree digunakan untuk prediksi penyakit kulit dengan hasil bahwa teknik Decision Tree lebih efisien dibandingkan teknik Naive Bayes dalam hal TP-rate- FP-rate, Precision, Recall dan ROC [12]. Teknik KNN juga diterapkan untuk deteksi penyakit kulit berdasarkan citra kulit [11][13]. Naive Bayes juga diterapkan untuk deteksi penyakit kulit berdasarkan citra kulit [5].

Pada penelitian ini akan dilakukan penerapan algoritma Decision Tree, Naive Bayes, KNN dan SVM untuk prediksi penyakit kulit dan diaplikasikan pada dataset penyakit kulit yang diunduh dari situs repository UCI[16]. Tujuan dari penelitian ini adalah untuk mengklasifikasikan algoritma klasifikasi yang akurat dan presisi untuk prediksi penyakit kulit.

## 2. Research Method

### 2.1. Data

Data yang digunakan dalam penelitian ini merupakan dataset yang diambil dari repositori basis data University of California at Irvine (UCI) [16]. Dataset penyakit kulit ini terdiri atas data diagnosis penyakit kulit sebanyak 366 data dan diklasifikasikan ke dalam 6 jenis penyakit. Dataset memiliki 34 atribut input berupa gejala dan 1 atribut target yaitu hasil diagnosis. Berdasarkan 366 dataset yang ada, 112 data didiagnosis psoriasis, 60 data didiagnosis

seboreic dermatitis, 72 data didiagnosa lichen planus, 49 data didiagnosis cronic dermatitis, 52 data didiagnosis pityriasis rosea dan 20 data didiagnosis pityriasis rubra pilaris[15][16].

### 2.2. Decision Tree (C4.5)

Algoritma C4.5 digunakan untuk melakukan klasifikasi atau segmentasi atau pengelompokan dan bersifat prediktif. Basis dari presikdi algoritma C4.5 ini adalah pohon keputusan. Cabang-cabang pohon keputusan merupakan pertanyaan klasifikasi dan daun-daunnya merupakan kelas-kelas atau segmen-segmennya [2].

### 2.3. Naive Bayes

Naive Bayes merupakan klasifikasi probabilitas sesuai dengan Teorema Bayes. Naive Bayes menganggap bahwa efek dari nilai atribut pada kelas tertentu independen dari nilai-nilai atribut lainnya, hal ini dilakukan untuk menyederhanakan perhitungan yang terlibat [6]. Rumus umum probalilitas bayes [6]:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

Keterangan:
$P(C_i|X)$ : Probabilitas $C_i$ jika diberi bukti X
$P(C_i)$ : Probabilitas $C_i$ tanpa memandang bukti apapun
$P(X|C_i)$ : Probabilitas X terjadi akan mempengaruhi $C_i$
$P(X)$ : Probabilitas X tanpa memandang bukti apapun

### 2.4. K-Nearest Neighbor (KNN)

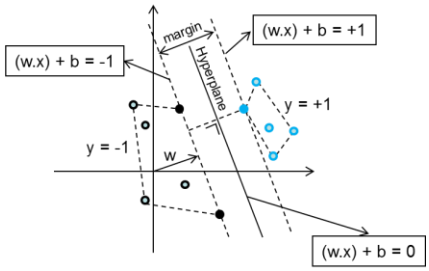
Algoritma KNN bertujuan untuk mengklasifikasikan data baru berdasarkan atribut dan data training [10]. Algoritma KNN menggunakan jarak kedekatan dengan tetangga sebagai nilai prediksi. Rumus algoritma KNN didesinifikan sebagai berikut [6]:

$$d(x_1,x_j) = \sqrt{\sum_{r=1}^{n}(a_r(x_i) - a_r(x_j))^2}$$

Keterangan:
$d(x_i,x_j)$ : Jarak Euclidean (*Euclidean Distance*)
$(x_i)$ : record ke-i
$(x_j)$ : record ke-j
$(a_r)$ : data ke-r
$i,j$ : 1,2,3,...n

### 2.5. Support Vector Machine (SVM)

SVM merupakan algoritma klasifikasi yang bisa digunakan untuk mengklasifikasikan data linier maupun data non liner [6]. Prinsip dasar SVM adalah menemukan *hyperlane* yang dapat mengklasifikasikan data menjadi 2 kelas.




**Gambar 1** SVM menemukan hyperlane terbaik

Rumus perhitungannya adalah sebagai berikut:

- Titik data : $x_i = \{x_1,x_2,\ldots,x_n\} \in R^n$
- Kelas data : $y_i \in \{-1,+1\}$
- Pasangan data dan kelas :
  $\{(x_i, y_i)\}_{i=1}^{N}$
- Maksimalkan fungsi:
  $Ld = \sum_{i=1}^{N}\alpha_i - \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i, x_j)$ *syarat* : $0 \le \alpha_i \le C$ *dan* $\sum_{i=1}^{N}\alpha_i y_i = 0$
- Menghitung nilai w dan b :

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i \quad b = -\frac{1}{2}\left(w.x^{+} + w.x^{-}\right)$$

- Fungsi keputusan klasifikasi sign(f(x)) :

$$f(x) = w.x + b \quad atau \quad f(x) = \sum_{i=1}^{m} \alpha_i y_i K(x, x_i) + b$$

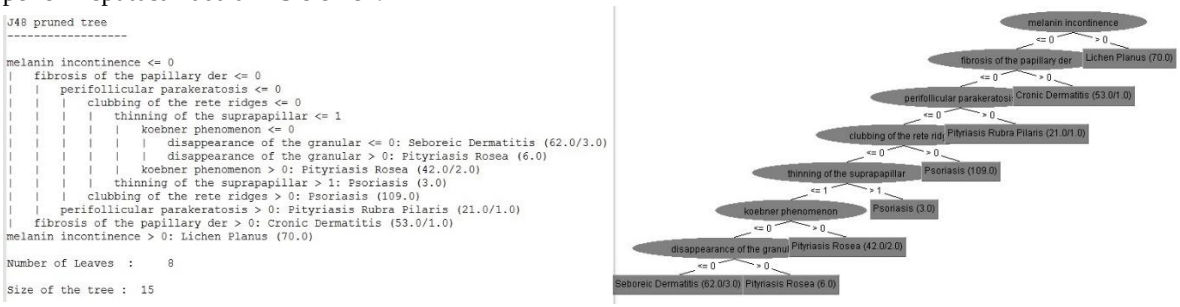
*Keterangan :*

N : banyaknya data
n : dimensi data atau banyaknya fitur
Ld : Dualitas Lagrange Multipier
$\alpha_i$ : nilai bobot setiap titik data
C : nilai konstanta
m : jumlah support vector/titik data yang memiliki $\alpha_i > 0$
$K(x,x_i)$ : fungsi kernel

## 3. Result and Analysis

### 3.1. Decision Tree (C4.5)

Algoritma klasifikasi decision tree menentukan nilai target sampel baru berdasarkan berbagai nilai atribut dari data yang tersedia. Pada gambar 2 menunjukkan bahwa pohon keputusan memiliki 8 daun dan ukuran total pohon keputusan adalah 15 elemen.




(a) (b)

**Gambar 2** Pohon Keputusan

Untuk mengevaluasi permorfa algoritma C4.5, dilakukan pembagian persentasi antara data training dan data tes. Tabel 1 menunjukkan hasil evaluasinya.

**Tabel 1** Hasil Pengujian C4.5

| Data Training (%) | Data Tes (%) | Keakuratan Hasil (%) | Presisi (%) | Recall (%) |
|---|---|---|---|---|
| 90 | 10 | 97,3 | 97,6 | 97,3 |
| 80 | 20 | 93,2 | 93,5 | 93,2 |
| 70 | 30 | 93,6 | 94 | 93,6 |
| 60 | 40 | 95,2 | 95,3 | 95,2 |
| 50 | 50 | 93,99 | 94,3 | 94 |

Nilai rata-rata perhitungan algoritma C4.5 diperoleh akurasi 94,7%, presisi 94,9% dan recall sebesar 94,7%.

### 3.2. Naive Bayes

Berdasarkan pengujian dengan WEKA, maka diperoleh akurasi, presisi dan recall sebagai berikut:

**Tabel 2** Hasil Pengujian Naive Bayes

| Data Training (%) | Data Tes (%) | Keakuratan Hasil (%) | Presisi (%) | Recall (%) |
|---|---|---|---|---|
| 90 | 10 | 100 | 100 | 100 |
| 80 | 20 | 97,3 | 97,5 | 97,3 |
| 70 | 30 | 98,2 | 98,3 | 98,2 |
| 60 | 40 | 97,9 | 98,1 | 97,9 |
| 50 | 50 | 97,3 | 97,5 | 97,3 |

Nilai rata-rata perhitungan algoritma j48 diperoleh akurasi 98,1%, presisi 98,3% dan recall sebesar 98,1%.

### 3.3. K-Nearest Neighbor

Berdasarkan pengujian dengan WEKA, maka diperoleh akurasi, presisi dan recall sebagai berikut:

**Tabel 3** Hasil Pengujian KNN

| Data Training (%) | Data Tes (%) | Keakuratan Hasil (%) | Presisi (%) | Recall (%) |
|---|---|---|---|---|
| 90 | 10 | 94,6 | 95,1 | 94,6 |
| 80 | 20 | 94,5 | 95,1 | 94,5 |
| 70 | 30 | 95,5 | 95,6 | 95,5 |
| 60 | 40 | 97,3 | 97,5 | 97,3 |
| 50 | 50 | 94,5 | 94,6 | 94,5 |

Nilai rata-rata perhitungan algoritma j48 diperoleh akurasi 95,3%, presisi 95,6% dan recall sebesar 95,3%.

### 3.4. Support Vector Machine (SVM)

Berdasarkan pengujian dengan WEKA, maka diperoleh akurasi, presisi dan recall sebagai berikut:

**Tabel 4** Hasil Pengujian SVM

| Data Training (%) | Data Tes (%) | Keakuratan Hasil (%) | Presisi (%) | Recall (%) |
|---|---|---|---|---|
| 90 | 10 | 100 | 100 | 100 |
| 80 | 20 | 95,9 | 96,1 | 95,9 |
| 70 | 30 | 98,2 | 98,2 | 98,2 |
| 60 | 40 | 97,9 | 98 | 97,9 |
| 50 | 50 | 98,4 | 98,4 | 98,4 |

Nilai rata-rata perhitungan algoritma SVM diperoleh akurasi 98,1%, presisi 98,2% dan recall sebesar 98,1%.

### 3.5. Perbandingan Hasil Uji

Berdasarkan hasil pengujian masing-masing algoritma, maka berikut adalah perbandingan hasil pengujian algoritma j48, Naive Bayes, KNN dan SVM.

**Tabel 5** Perbandingan Hasil Pengujian

| Data Training (%) | Data Tes (%) | Keakuratan Hasil (%) | | | | Presisi (%) | | | | Recall (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | C4.5 | Naive Bayes | KNN | SVM | C4.5 | Naive Bayes | KNN | SVM | C4.5 | Naive Bayes | KNN | SVM |
| 90 | 10 | 97,3 | 100 | 94,6 | 100 | 97,6 | 100 | 95,1 | 100 | 97,3 | 100 | 94,6 | 100 |
| 80 | 20 | 93,2 | 97,3 | 94,5 | 95,9 | 93,5 | 97,5 | 95,1 | 96,1 | 93,2 | 97,3 | 94,5 | 95,9 |
| 70 | 30 | 93,6 | 98,2 | 95,5 | 98,2 | 94 | 98,3 | 95,6 | 98,2 | 93,6 | 98,2 | 95,5 | 98,2 |
| 60 | 40 | 95,2 | 97,9 | 97,3 | 97,9 | 95,3 | 98,1 | 97,5 | 98 | 95,2 | 97,9 | 97,3 | 97,9 |
| 50 | 50 | 93,99 | 97,3 | 94,5 | 98,4 | 94,3 | 97,5 | 94,6 | 98,4 | 94 | 97,3 | 94,5 | 98,4 |
| Rata-rata | | 94,7 | 98,1 | 95,3 | 98,1 | 94,9 | 98,3 | 95,6 | 98,2 | 94,7 | 98,1 | 95,3 | 98,1 |

## 4. Conclusion (10 PT)

Hasil penggunaan algoritma klasifikasi data mining (C4.5, Naive Bayes, KNN, SVM) pada prediksi penyakit kulit dengan data set diambil dari UCI serta pengujiannya menunjukkan bahwa:

1. Algoritma C4.5, Naive Bayes, KNN, SVM bisa diterapkan untuk prediksi penyakit kulit dengan nilai akurasi, presisi dan recall di atas 94%.
2. Hasil pengujian menunjukkan algoritma Naive Bayes dan SVM memiliki nilai akurasi dan recall yang sama dan yang paling tinggi yaitu 98,1%, sedangkan nilai presisinya berbeda. Naive Bayes memberikan nilai presisi sebesar 98,3% dan SVM memiliki nilai presisi sebesar 98,2%, hanya selisih 0,1%.

3. Jika dibandingkan berdasarkan nilai, maka Naive Bayes dianggap sebagai algoritma klasifikasi yang lebih baik. Tetapi jika dibandingkan secara global, Naive Bayes dan SVM merupakan algoritma klasifikasi yang lebih baik dibandingkan C4.5 dan KNN untuk prediksi penyakit kuli

**References**


[1] Balakrishnan, Vimala; Shakouri, Mohammad R.; Huck Soo, Loo; , "Prediction Using Data Mining and Case-based Reasoning: A Case Study for Retinophaty", *International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering*, vol.6, no.3, pp.55-58, 2012.

[2] Bhargava, Neeraj; Sharma, Girja; Bhargava, Ritu; Mathuria, Manish;, "Decision Tree Analysis on J48 Algoritm for Data Mining", *International Journal of Advanced Research in Computer Science and Software Engineering*, vol.3, issue 6, June 2013.

[3] Dwiani, S.A.; Sam, A.; , "Diabetes Forecasting Using Supervised Learning Technique", *ACSIJ Advances in Computer Science: an International Journal*, vol.3, issue 5, no.11, Sept. 2014.

[4] Figueroa, J.; Fuller, L.; Abraha, A.; Hay, R.; , "The Prevalence of Skin Disease among Schoolchildren in Rural Ethiophia: A Preliminary Assessment of Dermatologic Needs", *Pediatric Dermatology* 13, pp. 378-81. 1996.

[5] Hafsah, Izzati Saila; Andono, Pulung Nurtantio; , "Deteksi Otomatis Penyakit Kulit Menggunakan Algoritma Naive Bayes", *Skripsi Fakultas Ilmu Komputer Universitas Dian Nuswantoro*, 2015, [on-line]: http://lppm.dinus.ac.id/index.php/home/TAView/16116/Deteksi-Otomatis-Penyakit-Kulit-Menggunakan-Ekstraksi-Fitur-Tekstur-pada-Citra-dan-Algoritma-Naive-Bayes.

[6] Han J.; Kamber, M; , "Data Mining:Concept and Techniques", New York:Morgan Kaufmann Publisher;2006

[7] Hay, R.; Bendeck, S.; Estrada, R..; Haddix A.; McLeod, T.; Mahe, A.; , "Skin Diseases", *Disease Control Priorities Project, The World Bank*. 2006

[8] Iswanto, Mukhamad Hasim; Permanasari, Adhistya Erna; Nugroho, Hanung Adi; , "Pemanfaatan Teknik Data Mining Untuk Diagnosis Penyakit Tuberculosis (TBS)", *Seminar Nasional Teknologi Informasi dan Multimedia 2015*, pp.121-126, Feb. 2015.

[9] Jaree, Thongkam; Guandong, Xu; Yanchun, Zhang; Fuchun Huang; , "Breast Cancer Survivabilityvia AdaBoost Algorithms", *Second Australian Workshop on Health Data and Knowledge Management*. 2008.

[10] Lestari, Mei; , "Penerapan Algoritma Klasifikasi Nearest Neighbor (k-NN) Untuk Mendeteksi Penyakit Jantung", *Factor Exacta*, vol.7, no.4, pp.366-371, 2014.

[11] Listyanto, Sebastian Rori; , "Implementasi K-Nearest Neighbor Untuk Mengenali Pola Citra Dalam Mendeteksi Penyakit Kulit", *Skripsi Fakultas Ilmu Komputer Universitas Dian Nuswantoro*, Apr. 2015, [on-line]: http://eprints.dinus.ac.id/id/eprint/15325.

[12] Manjusha, K.K.; Sankaranarayanan, K.; Seena, P.; ,"Prediction of Different Dermatological Condition Using Naive Bayesian Classification", *International Journal of Advance Research in Computer Science and Software Engineering* , vol.4, issue 1, pp.864-868 Jan. 2014.

[13] Manjusha, K.K.; Sankaranarayanan, K.; Seena, P.; ,"Data Mining in Dermatological Diagnosis : A Method for Severity Prediction", *International Journal of Computer Application* , vol.117, no.11, pp.11-14 May 2015.

[14] Othman, Mohd Fauzi bin; Yau, Thomas Moh Shan; , "Comparison of Different Classification Techniques Using WEKA for Breast Cancer", *IFMBE Proceedings*, vol.15, pp.520-523, 2007.

[15] Rambhajani, Madhura; Deepanker, Wyomesh; Pathak, Neelam; , "Classification of Dermatology Diseases through Bayes net and Best First Search", *International Journal of Advance Research in Computer Science and Software Engineering* , vol.4, issue 5, pp.116-119 May 2015.

[16] UCI, web source: https://archive.ics.uci.edu/ml/machine-learning-databases/dermatology/dermatology.data, accessed Oktober 2016.