Pengaruh Seleksi Fitur Pada Skema Klasifikasi *Naive Bayes* Berbasis *Gaussian* dan *Kernel Density*

Bagus Setya Rintyarna[1)]
[1)]*Prodi Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Jember*
Email: [1)]bagus.setya@unmuhjember.ac.id

Abstrak

Penyakait diabetes termasuk salah satu jenis penyakit yang perlu diwaspadai karena memiliki tingkat prevalensi yang cukup tinggi. Sebagai upaya deteksi dini penyakit diabetes, pada penelitian ini digunakan *Hidden Naïve Bayes* sebagai metode untuk klasifikasi penyakit diabetes. Hasil pengujian menunjukkan bahwa *Hidden Naïve Bayes* dapat digunakan untuk klasifikasi penyakit diabetes dengan kinerja yang lebih baik dibandingkan *Naïve Bayes Classifier.*

Kata kunci: *Hidden Naïve Bayes, Naïve Bayes Classifier.*

1. PENDAHULUAN

Dalam disiplin Kecerdasan Buatan (*Artificial Intelligence*) dikenal dua macam aktivitas/task penting, yaitu *supervised learning* dan *unsupervised learning*. Klasifikasi termasuk ke dalam *task supervised learning* sedangkan klastering termasuk dalam *task unsupervised learning*. Klasifikasi adalah proses menentukan kelas/kategori sebuah dataset berdasarkan pengetahuan awal yang dimiliki oleh metode yang digunakan. Proses mendapatkan pengetahuan awal itu disebut sebagai proses *learning*. Keberhasilan membangun sebuah metode Kecerdasan Buatan yang mampu mengklasifikasi dataset dengan kinerja yang baik mampu meningkatkan kecerdasan (*intelligence*) dari sistem komputer sehingga memiliki kemampuan mengklasifikasi tanpa campur tangan manusia.

Salah satu aspek yang berpengaruh besar terhadap hasil klasifikasi adalah fitur yang memiliki bobot signifikan terhadap klasifikasi. Dalam data *mining*, umumnya dataset yang dikumpulkan melalui proses *mining* memiliki dimensi data/jumlah yang besar yang sebenarnya tidak semua fitur memiliki dampak signifikan terhadap hasil klasifikasi yang benar. Oleh karena itulah diperlukan sebuah metode untuk menyeleksi fitur yang memiliki bobot sifnifikan terhadap klasifikasi. Proses klasifikasi dengan jumlah fitur yang sudah direduksi mampu meningkatkan kinerja metode klasifikasi dalam 2 aspek yaitu: akurasi klasifikasi dan kecepatan komputasi metode. John, Kohafi dan Pfleger menjelaskan konsep seleksi fitur dalam paper berjudul "I*rrelevant Features and The Subset Selection Problem*" (Yoon et al, 2006). Dalam paper tersebut dijelaskan ada dua macam fitur yaitu *relevant feature* dan *irrelevant feature*. *Relevant Feature* adalah fitur yang memiliki dampak signifikan terhadap hasil klasifikasi sedangkan *Irrelevant Feature* adalah fitur yang dampaknya kecil terhadap hasil klasifikasi. Dalam penelitian tersebut seleksi fitur dilakukan dengan metode *cross-validation* dalam skema klasifikasi berbasis ID3 dan C4.5.

Dalam penelitian yang lain, Jain dan Zonker melakukan seleksi fitur dengan menggunakan algoritma *Sequential Forward Floating Selection* (SFFS). Data set yang dipergunakan untuk menguji hasil seleksi fitur adalah dataset citra yang berasal dari SAR satelit yang menggunakan 4 model tekstur yang berbeda. Hasil penelitian mengindikasikan bahwa SFFS memiliki hasil yang lebih baik dibandingkan node pruning dan juga disimpulkan bahwa seleksi fitur baik dilakukan pada dataset yang memiliki jumlah fitur yang banyak (Tuomilehto et al., 2001).

Sementara itu Jimenez dan Landgrebe dalam Rismayanti (2005) melakukan investigasi

terhadap dataset *remote sensing* dari *Airborn Visible/Infrared Imaging Spectrometer* (AVIRIS) system. AVIRIS adalah sensor citra pertama yang digunakan untuk mengukur spectrum pantulan cahaya matahari dari kisaran panjang gelombang 400 nm sampai dengan interval 2500 nm pada interval 10 nm. Data yang dihasilkan oleh AVIRIS adalah data yang bersifat *multispectral* dan dalam penelitian Jimenez dihasilkan 220 band. Analisis dimensi tinggi dilakukan dengan menggunakan *Eucledian Distance* dan *Cartesian Geometry*. Penelitian mengembangkan metode seleksi fitur *Projection Pursuit* menjadi sebuah metode baru yaitu *Parametric Projection Pursuit* yang terdiri dari dua metode yaitu: *Parallel Parametric Projection Pursuit* dan *Sequential Parametric Projection Pursuit*.

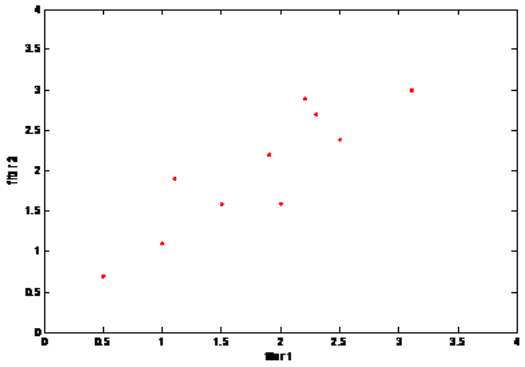
Heisele et. Al (1999) melakukan reduksi fitur pada skema klasifikasi citra wajah berbasis *support vector machine* dalam sebuah paper berjudul "*Hierarchichal Classification and Feature Reduction for Fast Face Detection With Support Vector Machines*". Dalam penelitian tersebut diusulkan sebuah metode untuk membangun sekaligus men-training sebuah hieararchy of classifier secara otomatis.Pada tahap berikutnya dilakukan reduksi fitur citra wajah untuk meningkatkan kinerja klasifier dalam mendeteksi citra wajah.

Pada penelitian ini diusulkan sebuah reduksi fitur berbasis *Principle Component Analysis* untuk meningkatkan kinerja klasifikasi berbasis Naïve Bayes Classifier. Untuk mengevaluasi Naïve Bayes akan dilakukan dua skema klasifikasi berbasis *Gaussian Distribution* dan *Kernel Density Estimation* dalam melakakukan perhitungan probabilitas yang diperlukan untuk mengetahui nilai posterior pada Naïve Bayes. Evaluasi akan dilakukan pada dataset Fisher Iris yang sudah tersedia di tool Matlab.

2. TINJAUAN PUSTAKA

2.1 *Principal Component Analysis*

Dalam implementasinya, PCA menggunakan prinsip-prinsip matematis untuk mentransformasikan sejumlah fitur yang berkorelasi menjadi sejumlah fitur lain yang ukurannya lebih kecil yang disebut sebagai *principle component*. PCA diimplementasikan dengan menggeser sumbu fitur ke arah fitur yang memiliki korelasi tinngi. Misalkan kita memiliki data dengan ukuran dua dimensi/fitur sebagai berikut: data = [2.5 2.4 ; 0.5 0.7 ; 2.2 2.9 ; 1.9 2.2 ; 3.1 3.0 ; 2.3 2.7 ; 2.0 1.6 ; 1.0 1.1 ; 1.5 1.6 ; 1.1 1.9]. Data-data tersebut apabila ditampilkan dalam bentuk plot didapatkan tampilan sebagai berikut:

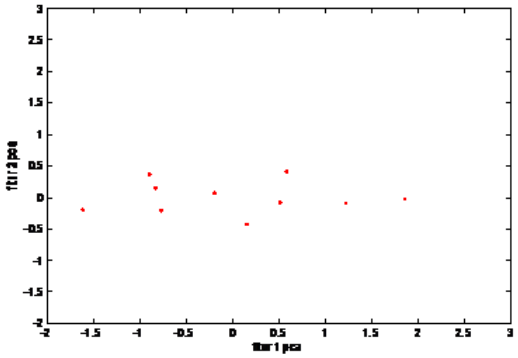


Gambar 1. Plot Grafik untuk data

Setelah data di-*adjust* terhadap nilai meannya, kemudian dihitung *covariance matrix*-nya, dihitung *eigenvector* dan *eigenvalue* serta *digenerate* data baru dengan menggunakan PCA maka didapatkan data baru dengan plot data tampak seperti Gambar 2. Dengan menggunakan matlab PCA bisa dieksekusi dengan *syntax* [COEFF, SCORE, latent] = *princomp* (data) di mana variabel SCORE menyimpan data baru yang di-*generate* dengan menggunakan PCA dan *latent* menyatakan nilai *eigenvalue* dari data yang menggambarkan urutan *principle component*-nya. Dari gambar 2 tampak bahwa fitur 2 bisa diabaikan/direduksi karena nilainya tidak se-signifikan fitur 1. Hal ini juga bisa dilihat dari nilai *eigenvalue* yang disimpan dalam variabel *latent* yang nilainya adalah [1.1410 0.0676], terlihat bahwa fitur 1 memiliki nilai eigenvalue lebih besar dari pada fitur 2. Dari gambar tampak pula bahwa PCA sebenarnya mentransformasi sumbu fitur ke arah *principle component*-nya.

Kontribusi penelitian ini adalah

penambahan *Principle Component Analysis* pada metode klasifikasi citra yang diusulkan oleh Chang yaitu "*A Bayesian Approach for Object Classification Based on Clusters of SIFT Local Features*". Shaw et al (2010) menyatakan bahwa *Principle Component Analysis* adalah sebuah metode statistik untuk mengurangi dimensi data dengan melakukan analisis kovarian antar faktor. Menurut Lindsay Smith, PCA diimplementasikan dalam beberapa tahap, yaitu (1) penyediaan dataset, (2) *adjust* dengan mengurangi tiap data dengan nilai meannya, (3) penghitungan *covariance-matrix* (4) penghitungan *eigenvalue* dan *eigenvector* dari *covariance* matrik yang didapatkan, (5) memilih *principle component* dan membentuk sebuah *feature vector* dan terakhir (6) menurunkan dataset baru.



Gambar 2. Plot Grafik untuk data setelah di-PCA

2.2 *Naive Bayes Classifier*

Naïve Bayes Classifier adalah salah satu metode klasifikasi berbasis penghitungan probability kemunculan fitur-fitur datasetnya terhadap fitur yang digunakan untuk training. Menurut Rish, Naïve Bayes mensimplifikasi proses learning dengan mengasumsikan bahwa fitur-fitur bersifat *independent* terhadap kelas. Meskipun asumsi bahwa fitur-fitur dataset bersifat independent adalah sebuah asumsi yang kurang baik tetapi kenyataannya hasil klasifikasi berbasis Naïve Bayes memiliki kinerja yang mampu berkompetisi dengan metode-metode lain yang lebih kompleks dalam aspek komputasinya.Keberhasilan Naïve Bayes tersebut setidaknya disebabkan oleh bahwa optimalnya hasil klasifikasi sebuah *classifier* tidak selalu berhubungan dengan kesesuaian kualitas distribusi probabilitas yang digunakan (Shaw, 2010).

Naïve Bayes Classifier memberikan kondisi sk dengan nilai attribute/fitur (A_1=v_1, A_2=v_2, …, A_m=v_m) terhadap kelas C_i dengan probabilitas maksimum (C_i |(v_1, v_2, …, v_m)) untuk semua *i*. Naïve Bayes mengasumsikan bahwa semua *attribute* adalah *independent*. Likelihood s_k terhadap kelas *i* (C_i) adalah sebagai berikut:

(1)

$$= \text{Prob}(C_i \mid (v_1, v_2, ..., v_m)) = \frac{P((v_1, v_2, ..., v_m) \mid C_i) P(C_i)}{P((v_1, v_2, ..., v_m))}$$

Sedangkan likelihood s_k terhadap kelas *j* (C_j) dapat dirumuskan dengan formula sebagai berikut :

(2)

$$= \text{Prob}(C_j \mid (v_1, v_2, ..., v_m)) = \frac{P((v_1, v_2, ..., v_m) \mid C_j) P(C_j)}{P((v_1, v_2, ..., v_m))}$$

Sehingga kelas data dapat ditentukan dengan membandingkan posterior data. Kelas ditentukan dengan nilai *likelihood/posterior* yang terbesar dan hanya perlu menghitung P((v_1, v_2, …, v_m)|C_i)P(C_i) dan P((v_1, v_2, …, v_m)|C_j)P(C_j).

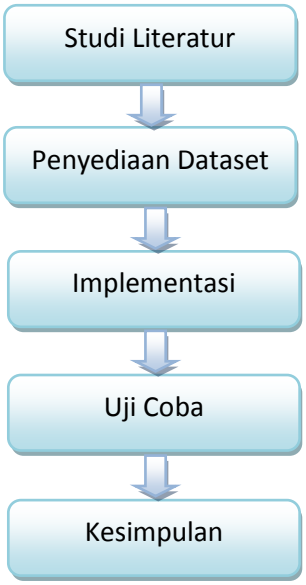
3. METODE PENELITIAN

Langkah-langkah atau tahap-tahap kegiatan penelitian dijelaskan dalam Blok Diagram seperti tampak pada gambar 3. Pada tahap pertama dilakukan studi literature yang bertujuan untuk menjelaskan kajian pustaka dari teori-teori penunjang yang mendukung konstruksi peneletian.Dengan studi literature diharapkan dapat dijelaskan secara lebih baik berkaitan dengan teori-teori penunjang yang membahas mengenai penyakit diabetes dan juga algoritma Hidden Naïve Bayes yang merupakan pengembangan dari Naïve Bayes Classifier. Juga dipelajari tentang beberapa parameter uji yang akan menjadi alat evaluasi kinerja algoritma yang digunakan dalam penelitian.

Pada tahap berikutnya dilakukan penyediaan data set. Pada penelitian kali

ini data set yang digunakan adalah data set Diabetes yang bisa diunduh secara gratis dari UCI *Machine Learning Repository.* Data set Diabetes yang disediakan di UCI *Machine Learning Repository* memiliki 8 atribute yaitu *preg, plas, pres, skin, insu, mass, pedi* dan *age* serta 2 kelas yaitu *tested positive* sebanyak 500 dataset dan tested negative sebanyak 270 dataset.

Selanjutnya implementasi metode dan pengujian dilakukan dengan tool Weka versi 3.6.9 untuk melakukan skema klasifikasi berbasis Hidden Naïve Bayes. Hasilnya juga akan dibandingkan dengan beberapa algoritma lain yang serumpun dengan Naïve Bayes yaitu Bayes Net, Naïve Bayes sendiri dan AODE. Parameter uji yang akan digunakan antara lain adalah: TP Rate, FP Rate, Precision, Recall dan F Measure. Sedangkan dua macam scenario uji yang akan digunakan adalah training tanpa *cross validasi* dan *training* dengan *cross validasi.*



Gambar 3. Langkah-langkah Kegiatan Penelitian

4. HASIL DAN PEMBAHASAN

4.1 Dataset

Dataset yang digunakan untuk pengujian dalam penelitian ini adalah dataset Diabetes yang diunduh dari UCI *Machine Learning Repository.* Owner dataset diabetes adalah *National Institutes of Diabetes and Digestive and Kidney Diseases* yang merupakan kontribusi dari *Vincent Sigillito Research Center*, RMI Group Leader Applied Phisics Laboratory The John Hopkins University. Jumlah fitur dataset Diabetes adalah 8 fitur yaitu: *preg, plas, pres, skin, insu, mass, pedi* dan *age*. Jumlah kelasanya ada dua yaitu *tested positive* dan *tested negative*. Jumlah keseluruhan dataset yang digunakan untuk pengujian metode Hidden Naïve Bayes (HNB) adalah sebanyak 668 data di mana jumlah kelas tested positive nya sebanyak 268 data dan jumlah kelas tested negativenya sebanyak 500 data.

4.2 Hasil Pengujian

Hasil klasifikasi dataset diabetes dengan algoritma Hidden Naïve Bayes dengan scenario pengujian dengan menggunakan training set dan menggunakan 10 *cross* validasi ditunjukkan dalam *confusion matrix* yang disajikan dalam tabel 1 berikut ini :

Tabel 1. Hasil Klasifikasi HNB dengan Training Set

Terklasifikasi sebagai	*tested positive*	*tested negative*
tested_positive	268	0
tested_negative	0	500

Jumlah data kelas *tested positive* dalam dataset adalah 268 dan terklasifikasi dengan HNB sebagai *tested positive* sebanyak 268, sedangkan jumlah data kelas *tested negative* dalam dataset adalah sejumlah 500 dan terklasifikasi sebagai *tested negative* sebanyak 500 sehingga sehingga dapat dihitung nilai parameter uji berupa TP Rate, FP *Rate, Precision, Recall* dan *F-Measure* seperti ditampilkan dalam tabel 2.

Tabel 2. Performance Klasifikasi HNB dengan Training Set

Kls/Par	TP Rate	FP Rate	Prec	Rec	F Meas
tested_positive	1	0	1	1	1
tested_negative	1	0	1	1	1
Rata-rata	1	0	1	1	1

Hasil tersebut lebih baik bila dibandingkan dengan klasifikasi dengan Naïve Bayes.

Tabel 3. Hasil Klasifikasi Naïve Bayes dengan Training Set

Terklasifikasi sebagai	*tested positive*	*tested negative*
tested positive	238	30
tested negative	25	475

Dari *convision matrix* di atas, bila dihitung rata-rata kinerja algoritma Naïve Bayes dapat ditampilkan dalam tabel 4.

Tabel 4. Performa Klasifikasi Naïve Bayes dengan Training Set

Kls/Par	TP Rate	FP Rate	Prec	Rec	F Meas
tested positive	0.888	0.050	0.905	0.888	0.896
tested negative	0.950	0.112	0.941	0.950	0.945
Rata-rata	0.919	0.081	0.923	0.919	0.921

5. KESIMPULAN

Dari hasil pengujian dapat disimpulkan beberapa hal sebagai berikut :

1. Algoritma Hidden Naïve Bayes adalah penyempurnaan Naïve Bayes Classifier.
2. Algoritma Hidden Naïve Bayes dapat dipergunakan untuk klasifikasi data Diabetes.
3. Hidden Naïve Bayes menunjukkan kinerja klasifikasi yang lebih baik dibandingkan dengan Naïve Bayes Classifier dalam pengujian dengan training set.

DAFTAR PUSTAKA

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. 2009. *The WEKA data mining software: an update*. ACM SIGKDD Explorations Newsletter, 11(1), 10-18.

Rismayanti C. 2005. *Terapi Insulin sebagai Alternatif Pengobatan Bagi Penderita Diabetes*. Repository. Universitas Negeri Yogyakarta.

Shaw J. E., Sicree R. A., Zimmet P. Z. 2010. *Global Estimates of The Prevalence of Diabetes,* Diabetes Research and Clinical Practice 87, 4-14.

Tuomilehto, J., et. al.2001. *Prevention of Type 2 Diabetes Mellitus by Changes in Lifestyle among Subjects with Impaired Glucose Tolerance*. The New England Journal of Medicine Vol. 344 No. 18.

Yoon, K. H., et. al. 2006. *Epidemic Obesity and Type 2 Diabetes in Asia, Lancet Division of Endocrinology and Metabolism*. College of Medicine, Catholic University of Korea.

www.mitrakeluarga.com diakses pada 11 Juli 2013.

Zhang, H., Jiang, L., & Su, J. 2005. *Hidden naive bayes*. *In Proceedings of the National Conference on Artificial Intelligence* (Vol. 20, No. 2, p. 919). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press.