SNESTIK

Seminar Nasional Teknik Elektro, Sistem Informasi, dan Teknik Informatika

https://ejurnal.itats.ac.id/snestik dan https://snestik.itats.ac.id

Informasi Pelaksanaan :
SNESTIK III - Surabaya, 11 Mareti 2023
Ruang Seminar Gedung A, Kampus Institut Teknologi Adhi Tama Surabaya

Informasi Artikel:
DOI : 10.31284/p.snestik.2023.4059

Prosiding ISSN 2775-5126

Fakultas Teknik Elektro dan Teknologi Informasi-Institut Teknologi Adhi Tama Surabaya
Gedung A-ITATS, Jl. Arief Rachman Hakim 100 Surabaya 60117 Telp. (031) 5945043
Email : snestik@itats.ac.id

Implementasi Metode Wrapper Sequential Feature Selection (WSFS) pada Dataset Stroke Menggunakan Metode Naïve Bayes Multinomial

Nadia Talidah S., Ester Yulitania T., Anindya Berlinani S., Chrisna Adrian Dwiputra H., Muchamad Kurniawan
Institut Teknologi Adhi Tama Surabaya
e-mail:adrianharyono77@gmail.com

ABSTRACT

There are many methods in the field of data mining to process data sets, one of which is Naive Bayes. The Naive Bayes algorithm is one of the classification algorithms with higher accuracy than decision tree algorithms and artificial neural networks. Multinomial Naive Bayes was able to reduce document misclassification by an average of 27%, while it reached 50% in the multivariate Bernoulli test. With less accuracy in searching so it requires this method. Results that can be carried out using feature selection with the Multinomial Naive Bayes model for better accuracy than Bernoulli's multivariate. Based on the results of the process carried out, the characteristic selection method can increase the influence of the test results on the model. Feature Selection Wrapper selects which features are important for objects/labels in the dataset and can select features that don't need to be used in the classification process using naive cells. The combination of using 6 functions can provide the highest accuracy compared to other combinations. The classification accuracy of Naive Bayes using a combination of six features increased to 81.575 from the accuracy value of the Naive Bayes algorithm, i. Based on the results of these experiments, the combination of feature selection and the Naive Bayes method shows

*better performance on medical datasets, especially in terms of hits, precision, recall and accuracy. This study examines the effectiveness of Feature Selection with the Naive Bayes classification algorithm. Based on the results of experimental testing and analysis of studies conducted, the WSFS method after testing with the 10-fold cross-validation method can provide recommendations to improve the performance of the classification algorithm.***Keywords:** *Feature Selection, Naïve Bayes*

ABSTRAK

Ada banyak metode di bidang data mining untuk mengolah kumpulan data, salah satunya adalah *Naive Bayes*. Algoritma *Naive Bayes* merupakan salah satu algoritma klasifikasi dengan akurasi yang lebih tinggi dibandingkan algoritma pohon keputusan dan jaringan syaraf tiruan. *Multinomial Naive Bayes* mampu mengurangi kesalahan klasifikasi dokumen rata-rata 27%, sementara itu mencapai 50% dalam pengujian Bernoulli multivariat. Dengan kurang keakuratan dalam mencari sehingga membutuhkan metode tersebut. Hasil yang dapat dilakukan dengan menggunakan seleksi fitur dengan model Multinomial Naive Bayes untuk akurasi yang lebih baik dibandingkan multivariat Bernoulli. Berdasarkan hasil proses yang dilakukan, metode pemilihan karakteristik dapat meningkatkan pengaruh hasil pengujian terhadap model. Pemilihan Fitur Wrapper memilih fitur mana yang penting untuk objek/label dalam dataset dan dapat memilih fitur yang tidak perlu digunakan dalam proses klasifikasi menggunakan sel naif. Kombinasi menggunakan 6 fungsi dapat memberikan akurasi tertinggi dibandingkan dengan kombinasi lainnya. Akurasi klasifikasi Naive Bayes menggunakan kombinasi enam fitur meningkat menjadi 81,575 dari nilai akurasi algoritma Naive Bayes, i. Berdasarkan hasil percobaan tersebut, kombinasi pemilihan fitur dan metode Naive Bayes menunjukkan kinerja yang lebih baik pada dataset medis, terutama dalam hal hit, presisi, recall dan akurasi. Penelitian ini menguji keefektifan Feature Selection dengan algoritma klasifikasi Naive Bayes. Berdasarkan hasil pengujian eksperimen dan analisis studi yang dilakukan, metode WSFS setelah dilakukan pengujian dengan metode 10-fold cross-validation dapat memberikan rekomendasi untuk meningkatkan kinerja algoritma klasifikasi.
Kata kunci: Feature Selection, Naïve Bayes.

PENDAHULUAN

Setiap dataset pasti memiliki karakteristik yang berbeda sehingga bentuk pengolahannya pun akan berbeda-beda. Pengolahan dataset akan disesuaikan juga dengan kebutuhan yang akan dituju sehingga dapat menghasilkan output yang sesuai dan nilai akurasi yang diinginkan. Setiap dataset juga memiliki jumlah yang sangat beragam karena sesuai dengan pengambilan dan kebutuhan pada setiap penelitiannya. Dataset sangat erat kaitannya dengan Data Mining. Data mining adalah proses menggali informasi atau sesuatu yang penting atau menarik dari data yang ada didalam database sehingga menghasilkan informasi yang sangat berharga.

Pada saat ini sudah banyak metode dalam pengolahan dataset pada bidang data *Mining*. Beberapa algoritma yang dapat digunakan adalah *Naïve Bayes, Decision Tree, Artificial Neural Network,* dan lain lain. Algoritma Naïve bayes adalah salah satu algoritma klasifikasi yang memiliki akurasi yang lebih tinggi dibandingkan oleh Algoritma Decision Tree dan *Artificial Neural Network*. Pengukuran kualitas dapat dihitung dengan akurasi. Tingkat akurasi dari masing-masing model memiliki perbedaan dari setiap model yang telah dilakukan pembelajaran. Tingkat akurasi yang baik terjadi jika tingkat akurasi mendekati 100% artinya model yang dihasilkan menunjukkan hasil yang tepat. Pada penelitian ini akan menggunakan Algoritma *Naïve Bayes*. Beberapa keuntungan menggunakan metode ini adalah sebagai metode *machine learning* yang menggunakan probabilitas, Jika ada nilai yang hilang, maka bisa diabaikan dalam perhitungan.

Pengklasifikasi *Naïve Bayes* dapat dibagi menjadi dua, yaitu *Multivariat Bernoulli* dan *Multinomial Naïve Bayes*. *Multinomial Naïve Bayes* mampu mengurangi kesalahan dalam klasifikasi dokumen dengan rata-rata 27% dia mencapai 50% Eksperimen dengan multivariat Bernoulli. Sebuah studi yang membandingkan kinerja dua model naif Pengklasifikasi Bayesian untuk pemfilteran anti-spam. Hasil yang dapat dilakukan oleh model Multinomial Naive Bayes untuk mencapai akurasi yang lebih baik dibandingkan Multivariat Bernoulli dengan pemilihan fitur selection. Begitu banyak peningkatan pada setiap metode sehingga sekarang mudah untuk pengolahan data tersebut.

Dalam penelitian tersebut dibahas banyak topik penting mulai dari pengumpulan data, hingga pemrosesan data dan akhirnya menggunakan data yang telah diproses tersebut untuk dilakukan tes secara efisien menggunakan algoritma *feature selection*. Penelitian tersebut menunjukkan beberapa peningkatan dramatis menggunakan hasil tersebut. Dan metode pemilihan feature harus diteliti lebih lanjut, pada data skala sangat besar. Jumlah pelatihan dan dokumen uji yang digunakan dalam penelitian tersebut sangat kecil dibandingkan apa yang tersebar luas di dunia maya. Selain itu, dalam penelitian tersebut, hanya ada 9 kategori. Di dunia nyata, ada ratusan kategori. Untuk memiliki pengkategorisasi berskala besar. Algoritma feature selection harus secara kuat dikembangkan. Dan topik ini dapat diteliti dan diuji lebih lanjut. Dengan melakukan penelitian lebih lanjut tentang topik yang disebutkan tersebut akan membantu, karena pada akhirnya dapat membantu mengkategorikan semua dokumen di dunia[1].

Pada penelitian sebelumnya yang berkaitan dengan feature selection serta mencakup Galavotti-Sebastiani-Simi Coefficient (GSS) di dalamnya yang berjudul Feature Selection for Text Categorisation oleh Øystein Løhre Garnes, 2009 membahas tentang langkah langkah untuk membangun pengklasifikasian menggunakan kumpulan file vektor (feature yang dipilih), mengevaluasi kinerjanya, dan menyimpan hasilnya di file. File berisi satu baris untuk setiap run, fold, dan dataset. Oleh karena itu dalam kasus penelitian tersebut, kalau hasil file dari percobaan Naïve Bayes berisi 600 baris hasil: 6 set ukuran feature (dari 500 hingga 10.000 feature) kali 10 kali fold 10 kali run (setiap kali lipat dijalankan), sedangkan hasil file dari percobaan Support Vector Machine berisi 150 baris hasil: 3 set ukuran feature (500, 1000, dan 2000 feature) kali 10 kali fold 5 kali run. Maka, dapat disimpulkan bahwa Naïve Bayes lebih cepat dibandingkan Support Vector Machine. Pada penelitian sebelumnya yang berkaitan dengan feature selection yang berjudul Categorical Proportional Difference: A Feature Selection Method for Text Categorization oleh Simeon, 2008 menyatakan bahwa sebuah feature dapat meningkatkan akurasi proses perhitungan. Metode feature selection adalah metode yang sangat populer dalam berbagai penelitian. feature selection digunakan untuk mengurangi dimensi dan mempercepat proses perhitungan. Selain itu feature selection juga mampu meningkatkan efisiensi dan akurasi dalam proses document extraction yang subset dengan pemilihan feature yang dianggap lebih relevan. zero, $\sum count(F, Ck) F \in V$ merupakan jumlah seluruh feature F pada kategori Ck, dan $|V|$ merupakan jumlah seluruh feature unik di seluruh kategori[1]

METODE

Naïve Bayes

Metode saat ini yang dipilih adalah Metode *Naïve Bayes*. *Naïve Bayes Classifier* merupakan sebuah metoda klasifikasi yang berakar pada *Teorema Bayes*.

Metode pengklasifikasian dengan menggunakan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai *Teorema Bayes.*

Keuntungan penggunaan adalah bahwa metoda ini hanya membutuhkan jumlah data pelatihan yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian. [engklasifikasian dokumen bisa dipersonalisasi, disesuaikan dengan kebutuhan setiap orang, Jika digunakan dalam bahasa pemrograman, *code*-nya sederhana, bisa digunakan untuk klasifikasi masalah *biner* ataupun *multiclass*

Berikut adalah rumus dari Metode *Naïve Bayes*:

$$P(H|X) = \frac{P(H|X).P(H)}{P(X)}$$

Wrapper Sequential Feature Selection (WSFS)

Selain menggunakan Naïve Bayes, penelitian ini menggunakan beberapa metode *Wrapper Sequential Feature Selection* untuk perbandingan hasil pengujian keakuratan. Penggunaan beberapa metode *Wrapper Sequential Feature Selection* ini bertujuan untuk melihat perbandingan nilai keakuratan yang lebih mendekati 100% pada metode *Naïve Bayes*. Feature Selection bertujuan mengurangi dimensi sehingga dapat meningkatkan keakuratan suatu dataset.

Multinomial Naïve Bayes

Multinomial Naïve Bayes digunakan untuk mengasumsikan independensi kemunculan kata dalam dokumen. Metode ini tidak memperhitungkan urutan kata dan *information context* dalam dokumen, namun memperhitungkan jumlah kata dalam dokumen (Destuardi dan Sumpeno, 2009). Dengan persamaan:

$$P(F, Ck) = \frac{count\ (F,Ck)+1}{(\sum F \in V\ count(F,Ck))+|V|}$$

$count\ (F,\ Ck\ )$ merupakan jumlah feature F yang muncul dalam suatu kategori Ck, penambahan nilai 1 untuk menghindari nilai zero, $\sum\ count(F,\ Ck\ )\ F \in V$ merupakan jumlah seluruh feature F pada kategori Ck, dan $|V|$ merupakan jumlah seluruh feature unik di seluruh kategori.

Metodologi yang digunakan dalam penelitian ini adalah sebagai berikut:

1. Studi Literatur, penulis akan mengumpulkan dasar teori serta mencari referensi yang diambil dari beberapa sumber seperti paper.
2. Pengumpulan Data, penulis akan melakukan pengumpulan data yang berasal dari Jurnal.
3. Pengolahan Data, penulis melakukan pengolahan yang telah dikumpulkan.
4. Implementasi Data, penulis melakukan implementasi data pada program yang sudah dibuat.
5. Penulisan Paper, penulis menulis paper sebagai bentuk hasil dari penelitian.

HASIL DAN PEMBAHASAN

A. Wrapper Sequential Feature Selection (WSFS)

Tabel 1. Hasil WSFS

Iterasi Ke-	**Atribut Index**	**Atribut *Scoring* terhadap target**
1	(11)	0,820402
2	(4, 11)	0,828448
3	(4, 7, 11)	0,858046
4	(3, 4, 7, 11)	0,862213
5	(3, 4, 7, 9, 11)	0,849569
6	(1, 3, 4, 7, 9, 11)	0,849713
7	(1, 3, 4, 7, 8, 9, 11)	0,849713
8	(1, 3, 4, 6, 7, 8, 9, 11)	0,845546
9	(0, 1, 3, 4, 6, 7, 8, 9, 11)	0,832902
10	(0, 1, 2, 3, 4, 6, 7, 8, 9, 11)	0,829023
11	(0, 1, 2, 3, 4, 6, 7, 8, 9, 10, 11)	0,816236
12	(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11)	0,832902

Tabel 1 menampilkan hasil dari proses klasifikasi menggunakan Naïve Bayes dan pengujian dengan 10-*Fold Cross Validation,* yang menunjukan bahwa akurasi yang dihasilkan adalah 75,241% dengan *recall sebesar* 43,56% dan presisi sebesar 67,66%.

Tabel 2. Hasil Performa Dari Klasifikasi *Naïve Bayes*

Naïve Bayes Tanpa Feature selection Wrapper			
Jumlah Fitur	***Accuracy***	***Recall***	***Precision***
11 Fitur	75,241	43,556	67,662

Penelitian ini menemukan bahwa fitur yang dapat digunakan untuk melakukan prediksi hanya *Ejection Fraction* dan serum *Creatinine* Kedua tersebut harus dipertahankan dan tidak boleh dihapus. Terdapat kombinasi 3-11 fitur yang dapat digunakan untuk melakukan prediksi.

Setelah mengimplementasikan metode *Naïve Bayes* dan melakukan pengujian model dengan menggunakan kombinasi 3 hingga 11 fitur, hasil klasifikasi dengan metode tersebut mengalami perubahan pada setiap nilai skoring setelah dilakukan proses *Wrapper Sequential Feature Selection (WSFS)*. Pengujian dengan menggunakan 10-*Fold Cross Validation* menghasilkan nilai akurasi tertinggi sebesar 81,575% dengan scoring 6 atribut. Nilai *recall* tertinggi adalah 54,889%, sedangkan performa precision 87,787% yang diperoleh dari kombinasi 3 fitur. Dari evaluasi tersebut, ditemukan bahwa model dengan menggunakan kombinasi 6 fitur memiliki akurasi tertinggi.

Tabel 3. Hasil Dari Proses *Wrapper Sequential Feature Selection (WSFS) & Naïve Bayes*

Jumlah Fitur	***Wrapper Sequential Feature Selection (WSFS)***		
	Accuracy	***Recall***	***Precision***
3 Fitur	81,241	53,889	**87,787**
4 Fitur	80.908	52.778	87.827

5 Fitur	80,575	51,889	87,454
6 Fitur	81.575	54.889	87.727
7 Fitur	77.230	46.778	77.475
8 Fitur	77,563	47,889	77,216
9 Fitur	78,908	52	83,209
10 Fitur	75,575	43,667	67,643
11 Fitur	75,575	42,667	68,337

Hasil akhir dari proses *Feature Selection dan Naïve Bayes* dapat dilihat pada Tabel 3. Dari hasil eksperimen, setiap fitur yang digunakan mempengaruhi hasil akhir dari proses klarifikasi pasien dengan stroke. Pengujian dilakukan dengan menggunakan metode *Naïve Bayes* dan 10-*Fold Cross Validation* pada model dengan 3 fitur, namun hasilnya kurang memuaskan. Oleh karena itu, dilakukan implementasi dengan menggunakan metode *Wrapper Sequential Feature Selection* (WSFS) untuk melakukan seleksi fitur yang lebih baik. Hasilnya menunjukan bahwa pengujian dataset dengan menggunakan metode WSFS dan klasifikasi *Naïve Bayes* dengan kombinasi atribut yang berbeda-beda akan menghasilkan tingkat akurasi yang berbeda juga.

Dalam proses ini, fitur-fitur yang didapatkan dengan menggunakan *Wrapper Sequential Feature Selection* (WSFS) adalah ('anaemia', 'diabetes', *'ejection_fraction'*, *'serum_creatinine*', *gender*', *'time*'). Model dengan menggunakan kombinasi 6 fitur ini, menghasilkan performa dengan nilai akurasi sebesar 81,575%, recall sebesar 54,889%, dan presisi sebesar 87,787%.

Metode Wrapper Sequential Feature Selection (WSFS) dapat meningkatkan performa dari hasil pengujian pada model, seperti terlihat pada Tabel 2 dan 3 yang menampilkan perbandingan hasil akurasi. Dengan menggunakan WSFS, fitur-fitur yang tidak diperlukan dapat dieliminasi sehingga dapat meningkatkan performa klasifikasi menggunakan *Naïve Bayes*.

Dari hasil eksperimen, diketahui bahwa penggunaan *Wrapper Sequential Feature Selection* (WSFS) dan naïve bayes dapat meningkatkan performa klasifikasi pada dataset medis, khususnya untuk kasus penyakit stroke. Terjadi peningkatan akurasi sebesar 6,334%, recall sebesar 11,333%, dan presisi sebesar 20,07%. Oleh karena itu, kombinasi metode WSFS dan naïve bayes sangat efektif dalam meningkatkan performa klasifikasi pada dataset medis.

KESIMPULAN

Penelitian ini telah menguji keefektivitasan *Wrapper Sequential Feature Selection (WSFS)* yang dikombinasikan dengan algoritma klasifikasi *Naïve Bayes*. Berdasarkan dari hasil eksperimen pengujian dan analisa dari penelitian yang telah dilakukan, metode WSFS dapat memberikan rekomendasi fitur yang dapat digunakan untuk meningkatkan performa algoritma klasifikasi setelah diuji dengan menggunakan metode validasi *10-Fold Cross validation*. Dari dataset medis yang dipilih sebagai data uji didapatkan bahwa kombinasi 6 (enam) fitur yaitu *('gender', 'age',' hypertension', 'avg_glucose_level', 'bmi', 'smoking_status')* dari total 11 (sebelas) fitur dapat meningkatkan performa algoritma klasifikasi *Naïve Bayes* dengan nilai akurasi sebesar 81.575%. Nilai *Recall* sebesar 54,889%. Nilai *Precision* sebesar 87,787%. Jika dibandingkan dengan hanya menerapkan algoritma klasifikasi *Naïve Bayes*, nilai

performa yang dihasilkan hanya 75,241% untuk performa akurasi, recall hanya 43,556% dan Precision yaitu 67,662%. Dari hasil tersebut dapat diketahui bahwa terjadi peningkatan sebesar 6,334%. untuk akurasi, 11,333% untuk nilai recall dan 20,07% untuk nilai performa precision-nya. Hal tersebut membuktikan bahwa metode seleksi fitur FS dapat secara signifikan meningkatkan performa algoritma klasifikasi khususnya *Naïve Bayes* dalam kasus yang diangkat ini.

DAFTAR PUSTAKA

[1] T. Marta Elisa Yuridis Butar Butar and M. Ali Fauzi, "Penentuan Rating Review Film Menggunakan Metode Multinomial Naïve Bayes Classifier dengan Feature Selection berbasis Chi-Square dan Galavotti-Sebastiani-Simi Coefficient," 2019. [Online]. Available: http://j-ptiik.ub.ac.id

[2] I. Destuardi and S. Sumpeno, "Klasifikasi Emosi Untuk Teks Bahasa Indonesia Menggunakan Metode Naive Bayes," 2009.