Jurnal Teknik Elektro Vol. 11 No. 2
Juli - Desember 2019

P-ISSN 1411 - 0059
E-ISSN 2549 - 1571

Kombinasi Metode *Correlated Naive Bayes* dan Metode Seleksi Fitur *Wrapper* untuk Klasifikasi Data Kesehatan

Hairani[1] dan Muhammad Innuddin[2]

[1]*Program Studi Ilmu Komputer, Fakultas Teknik dan Desain, Universitas Bumigora*
[2]*Program Studi Sistem Informasi, Fakultas Teknik dan Desain, Universitas Bumigora*
Jl. Ismail Marzuki No.22, Mataram, 83127, Indonesia
Hairani@universitasbumigora.ac.id[1], Inn@universitasbumigora.ac.id[2]

Abstract*— Most features of health data that have many irrelevant features can reduce the performance of classification method. One health data that has many attributes is the Pima Indian Diabetes dataset and Thyroid. Diabetes is a deadly disease caused by the increasing of blood sugar because of the body's inability to produce enough insulin and its complications can lead to heart attacks and strokes. The purpose of this research is to do a combination of Correlated Naïve Bayes method and Wrapper-based feature selection to classification of health data. The stages of this research consist of several stages, namely; (1) the collection of Pima Indian Diabetes and Thyroid dataset from UCI Machine Learning Repository, (2) pre-processing data such as transformation, Scaling, and Wrapper-based feature selection, (3) classification using the Correlated Naive Bayes and Naive Bayes methods, and (4) performance test based on its accuracy using the 10-fold cross validation method. Based on the results, the combination of Correlated Naive Bayes method and Wrapper-based feature selection get the best accuracy for both datasets used. For Pima Indian Diabetes dataset, the accuracy is 71,4% and the Thyroid dataset accuracy is 79,38%. Thus, the combination of Correlated Naïve Bayes method and Wrapper-based feature selection result in better accuracy without feature selection with an increase of 4,1% for Pima Indian Diabetes dataset and 0,48% for the Thyroid dataset.

Keywords— Correlated Naive Bayes, Wrapper feature selection, Pima Indian Diabetes dataset, Thyroid dataset, health data

Abstrak—Kebanyakan fitur pada data kesehatan terdapat fitur tidak relevan sehingga dapat menurunkan kinerja metode klasifikasi. Salah satu data kesehatan yang memiliki atribut banyak adalah Pima Indian Diabetes dan Thyroid. Penyakit diabetes merupakan salah satu penyakit mematikan yang disebabkan meningkatnya gula darah yang diakibatkan oleh ketidakmampuan tubuh menghasilkan insulin yang cukup dan komplikasinya dapat mengakibatkan serangan jantung dan stroke. Tujuan dari penelitian ini adalah melakukan kombinasi metode *Correlated Naive Bayes* dan seleksi fitur berbasis *Wrapper* untuk klasifikasi data kesehatan. Tahapan penelitian ini terdiri dari beberapa tahapan yaitu (1) pengumpulan *dataset* Pima Indian Diabetes dan Thyroid dari UCI Machine Learning Repository, (2) data *pre-processing* seperti transformasi, *scaling*, dan seleksi fitur berbasis *Wrapper*, (3) klasifikasi menggunakan metode *Correlated Naive Bayes* dan *Naive Bayes*, dan (4) pengujian kinerja berdasarkan akurasinya menggunakan metode validasi *10-fold cross validation*. Berdasarkan hasil pengujian yang telah dilakukan, kombinasi metode *Correlated Naive Bayes* dengan seleksi fitur berbasis *Wrapper* mendapatkan akurasi terbaik kedua *dataset* yang digunakan. Untuk *dataset* Pima Indan Diabetes akurasinya sebesar 71,4% dan akurasi *dataset* Thyroid sebesar 79,38%. Dengan demikian, kombinasi metode *Correlated Naive Bayes* dan seleksi fitur berbasis *Wrapper* menghasilkan akurasi lebih baik tanpa seleksi fitur dengan kenaikan sebesar 4,1% untuk *dataset* Pima Indian Diabetes dan 0,48% *dataset* Thyroid.

Kata kunci— ***Correlated Naive Bayes,*** **seleksi fitur** ***Wrapper, dataset*** **Pima Indian Diabetes,** ***dataset*** **Thyroid, data kesehatan**

I. PENDAHULUAN

Untuk meningkatkan ketepatan klasifikasi pada data kesehatan membutuhkan metode klasifikasi dengan kinerja yang baik. Bagaimanapun, jumlah data kesehatan yang diperoleh dari mesin digital memiliki fitur yang banyak dan tidak semua atributnya relevan digunakan untuk klasifikasi penyakit [1]. Kebanyakan fitur pada data kesehatan terdapat fitur tidak relevan sehingga dapat menurunkan kinerja metode klasifikasi. Salah satu data kesehatan yang memiliki banyak atribut adalah Pima Indian Diabetes dan Thyroid. Penyakit diabetes merupakan salah satu penyakit mematikan yang disebabkan meningkatnya gula darah dalam tubuh [2]. Penyakit diabetes disebabkan ketidakmampuan tubuh memproduksi insulin yang cukup. Komplikasi penyakit diabetes dapat menyebabkan serangan jantung dan stroke. Salah satu cara untuk meningkatkan akurasi metode klasifikasi adalah penggunaan pemilihan fitur. Pemilihan fitur

merupakan teknik pra-pengolahan sangat penting untuk memilih fitur-fitur yang berpengaruh pada sebuah *dataset* [3]. Pemilihan fitur digunakan untuk memilih fitur-fitur yang berpengaruh, menghapus fitur tidak relevan pada atribut *dataset*, waktu komputasi menjadi cepat, dan dapat meningkatkan kinerja dari metode klasifikasi [4], [5]. Teknik pemilihan fitur dibagi menjadi 3 kelompok yaitu *Filter*, *Wrapper*, dan *Embedded* [4]. Penelitian pemilihan fitur berbasis *Filter* dilakukan oleh [6]–[8], berbasis *Wrapper* oleh [3], [5], dan *embedded* [9]. Penelitian ini menggunakan teknik pemilihan fitur berbasis *Wrapper* dikarenakan memiliki kinerja lebih baik dari pemilihan fitur berbasis *Filter* [10] dan *Embedded* [11].

Metode klasifikasi yang digunakan penelitian ini adalah algoritma *Correlated Naive Bayes*. Algoritma *Correlated Naive Bayes* adalah sebuah algoritma hasil pengembangan *Naive Bayes*. Parameter-parameter yang ditambahkan pada algoritma *Correlated Naive Bayes* adalah nilai korelasi antar fitur X dengan kelasnya dan bilangan laplacian. Perhitungan korelasi (*R-Square*) dilakukan untuk menunjukkan hubungan antar fitur dengan kelasnya pada algoritma *Correlated Naive Bayes* [12]. Bilangan laplacian digunakan untuk menghindari terjadinya *zero probability*. Untuk meningkatkan akurasi algoritma *Correlated Naive Bayes* dapat menggunakan teknik pemilihan fitur.

Penelitian dibidang *data mining* untuk klasifikasi penyakit sudah banyak dilakukan diantaranya adalah penelitian [13] melakukan klasifikasi penyakit diabetes menggunakan metode *Correlated Naive Bayes* dan *Naive Bayes*. Kelemahan penelitian ini tidak menggunakan seleksi fitur untuk memilih atribut yang relevan. Penelitian [14] mengatasi permasalahan ketidakseimbangan kelas pada *dataset* Pima Indian Diabetes menggunakan metode K-Means-Smote. Kelemahan penelitian tersebut tidak menggunakan proses seleksi fitur sebelum melakukan klasifikasi menggunakan metode C4.5, *Support Vektor Machine* (SVM), dan *Naive Bayes*. Penelitian [15] melakukan diagnosis penyakit rematik menggunakan penalaran maju dan metode faktor kepastian.

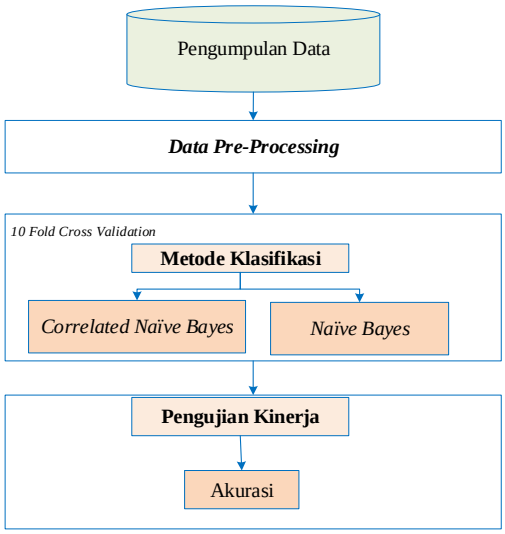
Penelitian [16] menggunakan teknik pemilihan fitur *Information Gain* untuk memilih fitur yang berpengaruh untuk deteksi penyakit jantung. Penelitian [17] mengimplementasikan teknik pemilihan fitur *Hybrid* untuk prediksi diabetes melitus. Teknik pemilihan fitur berbasis *Filter* menggunakan *Information Gain*, sedangkan berbasis *Wrapper* menggunakan SVM sebagai algoritma pembelajarannya dan *Sequential Backward Search* (SBS). Penelitian [18] melakukan perbandingan teknik pemilihan fitur *Information Gain* dan *Relief* untuk deteksi penyakit gagal ginjal kronis. Berdasarkan hasil penelitiannya, metode seleksi fitur *Relief* memiliki kinerja lebih baik dari *Information Gain*.

Berdasarkan uraian di atas, terdapat *gap* penelitian ini dengan penelitian sebelumnya yaitu belum ada penelitian yang mengkombinasikan metode *Correlated Naive Bayes* dan seleksi fitur berbasis *Wrapper* untuk klasifikasi data kesehatan yang memiliki banyak fitur. Sebagai perbandingan penelitian [13] menggunakan algoritma *Correlated Naive Bayes* untuk klasifikasi penyakit diabetes tanpa menggunakan seleksi fitur.

Oleh karena itu, penelitian ini mengkombinasikan algoritma *Correlated Naive Bayes* dan seleksi fitur berbasis *Wrapper* untuk klasifikasi data kesehatan untuk mendapatkan akurasi optimal.

II. METODE

Untuk menyelesaikan penelitian ini, aliran penelitian yang digunakan ditunjukkan pada Gambar 1. Tahapan pertama melakukan pengumpulan *dataset* kesehatan. *Dataset* dibidang kesehatan yang digunakan adalah *dataset* pima indians diabetes dan *dataset* Thyroid diperoleh dari *UCI Machine Learning Repository*. Detail *dataset* penelitian ini ditunjukkan pada Tabel I. Adapun detail atribut pada *dataset* yang digunakan ditunjukkan pada Tabel II dan Tabel III.



Gambar 1. Tahapan penelitian

TABEL I. DETAIL *DATASET*

No	*Dataset*	Data	Atribut	Kelas
1.	Pima Indian Diabetes	768	9	2
2.	Thyroid	215	6	3

TABEL II. ATRIBUT *DATASET* PIMA INDIAN DIABETES

No	Atribut	Label	Deskripsi
1.	*Number of times pregnant*	NP	Jumlah kehamilan
2.	*Plasma glucose*	GTT	Kadar glukosa
3.	*Blood pressure*	DBP	Tekanan darah
4.	*Skin thickness*	TSF	Ketebalan kulit
5.	*Insulin*	HSI	Insulin
6.	*BMI*	BMI	Berat masa tubuh
7.	*Diabetes pedigree function*	DPF	Riwayat penyakit keluarga
8.	*Age*	*Age*	Umur
9.	*Tested Negative and Tested Positive*	Kelas	

TABEL III. ATRIBUT *DATASET* THYROID

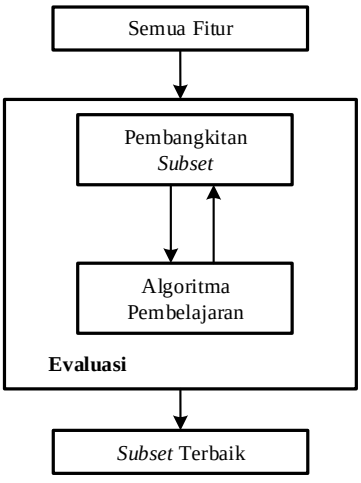
No.	Atribut	Label
1.	*T3-resin uptake test*	RUT
2.	*Total serum thyroxin*	TST
3.	*Total serum triiodothyronine*	TSTD
4.	*Basal thyroid-stimulating hormone (TSH*	TSH
5.	*Maximal absolute difference of TSH value*	MAD
6.	*Normal* *Hyper* *Hypo*	Kelas

Tahapan pra-pengolahan digunakan penelitian ini adalah transformasi, *scaling*, dan pemilihan fitur. Transformasi digunakan untuk merubah tipe atribut nominal menjadi angka. Perhitungan korelasi (*R-Square*) dilakukan untuk menunjukkan hubungan antar fitur dengan kelasnya pada algoritma *Correlated Naive Bayes*.

Scaling digunakan untuk mimalisir terjadinya dominasi fitur dengan jangkauan nilai terbesar (max_x) terhadap fitur dengan jangkauan nilai terkecil (min_x). Formulasi yang digunakan untuk melakukan *scaling* ditunjukkan pada (1).

Sedangkan pemilihan fitur digunakan untuk menghapus fitur tidak relevan dan memilih fitur yang berpengaruh pada *dataset* Pima Indian Diabetes dan Thyroid. Teknik pemilihan fitur *Wrapper* digunakan pada penelitian ini adalah SVM sebagai algoritma pembelajarannya dan S*equential Backward Search* (SBS). Teknik pemilihan fitur *Wrapper* dalam pemilihan atributnya melibatkan algoritma pembelajaran. Adapun proses kerja dari metode pemilihan fitur *Wrapper* ditunjukkan pada Gambar 2.

$$X' = \frac{x - \min_x}{\max_x - \min_x} \quad (1)$$



Gambar 2. Proses pemilihan fitur berbasis *Wrapper* [19]

Algoritma *Correlated Naive Bayes* adalah sebuah algoritma hasil pengembangan *Naive Bayes*. Parameter-parameter yang ditambahkan pada algoritma *Correlated Naive Bayes* adalah nilai korelasi antar fitur X dengan kelasnya dan bilangan laplacian. Perhitungan korelasi (*R-Square*) dilakukan untuk menunjukkan hubungan antar fitur dengan kelasnya pada algoritma *Correlated Naive Bayes* [12]. Bilangan laplacian digunakan untuk menghindari terjadi *zero probability*. Rumus algoritma *Correlated Naive Bayes* untuk klasifikasi ditunjukkan pada (2) [12], serta rumus perhitungan korelasinya ditunjukkan pada (3) dan (4).

$$P(Y \mid X) = \frac{P(Y)\prod_{i=1}^{q} P(X_i \mid Y)^{\ell} . R(X_i \mid Y)}{P(X)} \quad (2)$$

P(X|Y) merupakan probabilitas hipotesis Y berdasarkan kejadian X. P(Y) merupakan *prior probability* pada hipotesis Y. $\Pi_{i=1}^{q}(X_i \mid Y)$ merupakan probabilitas fitur X berdasarkan hipotesis Y. $R(X_i|Y)$ merupakan korelasi (*R-Square*) fitur X berdasarkan hipotesis Y. ℓ merupakan bilangan laplacian, sedangkan *P(X)* merupakan probabilitas X.

$$r = \frac{n.\left(\sum XY\right) - \left(\sum X\right).\left(\sum Y\right)}{\sqrt{\left(n.\sum X^2 - \left(\sum X\right)^2\right)}\sqrt{n.\sum Y^2 - \left(Y\right)^2}} \quad (3)$$

$$R = r^2 \quad (4)$$

R merupakan *R-Square* fitur antar kelasnya, sedangkan *r* merupakan nilai korelasi fitur antar kelasnya. n merupakan total data pada *dataset*. $\sum XY$ merupakan total perkalian fitur (X) dengan kelasnya (Y). $\sum X$ merupakan total dari fitur X, sedangkan $\sum Y$ merupakan total dari fitur Y. $\sum X^2$ merupakan total dari fitur X yang dikuadratkan, sedangkan $(\sum X)^2$ merupakan kuadrat total fitur X. $\sum Y^2$ merupakan total fitur Y yang dikuadratkan, sedangkan $(\sum Y)^2$ merupakan kuadrat total fitur Y.

III. HASIL DAN PEMBAHASAN

A. Pengumpulan Data

Pengumpulan *dataset* diperoleh dari *UCI Machine Learning Repository*. *Dataset* yang digunakan penelitian ini adalah *dataset* Thyroid dan Pima Indian Diabetes yang ditunjukkan pada Tabel IV dan Tabel V.

TABEL IV. CONTOH *DATASET* THYROID

No.	RUT	TST	TSTD	TSH	MAD	Kelas
1.	107	10,1	2,2	0,9	2,7	Normal
2.	113	9,9	3,1	2,0	5,9	Normal
3.	139	16,4	3,8	1,1	-0,2	Hyper
4.	125	2,3	0,9	16,5	9,5	Hypo
...	...	...	...	...	...	...
211.	104	6,1	1,8	0,5	0,8	Normal
212.	102	6,2	1,2	1,4	1,3	Normal
213.	102	5,3	1,4	1,3	6,7	Hypo
214.	79	19,0	5,5	0,9	0,3	Hyper
215.	92	11,1	1,2	0,7	-0,2	Hyper

TABEL V. CONTOH *DATASET* PIMA INDIAN DIABETES

No.	NP	GTT	DBP	TSF	HS1	BMI	DPF	Age	Kelas
1.	1	85	66	29	0	26,6	0,351	31	*Negative*
2.	1	89	66	23	94	28,1	0,167	21	*Negative*
3.	5	116	74	0	0	25,6	0,201	30	*Negative*
4.	10	115	0	0	0	25,2	0,134	29	*Negative*
...	...	...	...	...	...	...	...	...	...
764.	1	128	88	39	110	36,5	1,057	37	*Positive*
765.	0	123	72	0	0	36,3	0,258	52	*Positive*
766.	6	190	92	0	0	35,5	0,278	66	*Positive*
767.	9	170	74	31	0	44,0	0,403	43	*Positive*
768.	1	126	60	0	0	30,1	0,349	47	*Positive*

B. Pra-pengolahan Data

Tahapan pra-pengolahan digunakan penelitian ini adalah transformasi, *scaling*, seleksi fitur. Transformasi digunakan untuk merubah tipe atribut nominal menjadi angka. Perhitungan korelasi (*R-Square*) dilakukan untuk menunjukkan hubungan antar fitur dengan kelasnya pada algoritma *Correlated Naive Bayes*. Hasil transformasi atribut ditunjukkan pada Tabel VI dan Tabel VII.

Scaling digunakan untuk mimalisir terjadinya dominasi fitur dengan jangkauan nilai terbesar terhadap fitur dengan jangkauan nilai terkecil. Formulasi yang digunakan untuk melakukan *scaling* ditunjukkan pada (1). Contoh hasil *scaling* pada *dataset* Pima Indian Diabetes ditunjukkan pada Tabel VIII.

TABEL VI. TRANSFORMASI *DATASET* PIMA INDIAN DIABETES

No.	Data Nominal	Data Angka
1.	*Negative*	1
2.	*Positive*	2

TABEL VII. TRANSFORMASI *DATASET* THYROID

No.	Data Nominal	Data Angka
1.	*Normal*	1
2.	*Hyper*	2
3.	*Hypo*	3

Sebagai contoh hasil *scaling* atribut GTT pada *dataset* Pima Indian Diabetes untuk nilai 89 dimana nilai maksimum dan minimumnya adalah 139 dan 85 adalah sebagai berikut:

$$X' = \frac{x - \min_x}{\max_x - \min_x} = \frac{89 - 85}{139 - 85} = 0{,}074$$

Sedangkan pemilihan fitur digunakan untuk menghapus fitur tidak relevan dan memilih fitur yang berpengaruh pada *dataset* Pima Indian Diabetes dan Thyroid. Teknik pemilihan fitur *Wrapper* digunakan pada penelitian ini adalah SVM sebagai algoritma pembelajarannya berdasarkan (4) mengacu penelitian [20] dan S*equential Backward Search* (SBS). Adapun hasil fitur yang terpilih menggunakan teknik pemilihan fitur *Wrapper* ditunjukkan pada Tabel IX.

C. Klasifikasi

Metode klasifikasi penelitian ini menggunakan metode *Correlated Naive Bayes* berdasarkan (2).

TABEL VIII. HASIL *SCALING* ATRIBUT GTT PADA *DATASET* PIMA INDIAN DIABETES

No.	Tanpa *Scaling*	Hasil *Scaling*
1.	89 (Mg/dL)	0,074
2.	85 (Mg/dL)	0
3.	116 (Mg/dL)	0,574
4.	115 (Mg/dL)	0,556
5.	110 (Mg/dL)	0,463
6.	139 (Mg/dL)	1
7.	103 (Mg/dL)	0,333
8.	126 (Mg/dL)	0,759
9.	99 (Mg/dL)	0,259
10.	97 (Mg/dL)	0,222

TABEL IX. ATRIBUT TERPILIH HASIL SELEKSI FITUR

Dataset	Atribut Original	Atribut Terpilih
Pima Indian Diabetes	NP, GTT, DBP, TSF, HSI, BMI, DBF, Age	GTT, DBF
Thyroid	RUT, TST, TSTD, TSH, MAD	TST, TSTD, TSH, MAD

D. Pengujian Kinerja

Metode *10-fold cross validation* digunakan untuk memvalidasi hasil akurasi metode *Correlated Naive Bayes* tiap-tiap *fold*. Metode *10-fold cross validation* membagi data sebanyak 10 data tiap-tiap *fold*. Hasil akurasi tiap-tiap *fold* ditunjukkan pada Tabel X dan Tabel XI. Untuk mempermudah melihat nilai akurasi metode *Correlated Naive Bayes* ditunjukkan pada Tabel XII.

Berdasarkan pada Tabel XII, ditunjukkan akurasi terbaik diperoleh metode *Correlated Naive Bayes* dengan seleksi fitur pada *dataset* Pima Indian Diabetes sebesar 71,4%, sedangkan pada *dataset* Thyroid akurasinya sebesar 79,38%. Dengan demikian, penggunaan seleksi fitur *Wrapper* dapat meningkatkan akurasi metode *Correlated Naive Bayes* sebesar 4,1% untuk *dataset* Pima Indian Diabetes dan sebesar 0,48% untuk *dataset* Thyroid, dikarenakan hanya mengklasifikasikan fitur-fitur yang relevan [21]. Hal ini selaras dengan dengan penelitian [22]–[24] menggunakan teknik seleksi fitur untuk meningkatkan akurasi metode klasifikasi yang digunakan. Kebanyakan referensi seperti [13], [25], dan [26] hanya implementasi metode *Correlated Naive Bayes*, sehingga

penelitian ini menggunakan metode seleksi fitur untuk meningkatkan akurasi algoritma *Correlated Naive Bayes*.

TABEL X. AKURASI *DATASET* PIMA INDIAN DIABETES

Fold	*Correlated Naive Bayes*		*Naive Bayes*	
	Tanpa Pemilihan Fitur	**Pemilihan Fitur**	**Tanpa Pemilihan Fitur**	**Pemilihan Fitur**
1.	67,89%	**73,29%**	63,95%	70,39%
2.	**67,50%**	63,16%	63,95%	59,87%
3.	67,37%	**72,11%**	64,34%	71,40%
4.	66,71%	**73,03%**	63,82%	72,24%
5.	67,37%	**72,89%**	65,53%	70,13%
6.	66,97%	**72,63%**	64,47%	69,61%
7.	67,37%	**71,97%**	64,87%	71,84%
8.	67,24%	**72,24%**	63,29%	70,92%
9.	67,37%	**72,24%**	63,68%	70,39%
10.	67,24%	**72,37%**	64,74%	71,71%

TABEL XI. AKURASI *DATASET* THYROID

Fold	*Correlated Naive Bayes*		*Naive Bayes*	
	Tanpa Pemilihan Fitur	**Pemilihan Fitur**	**Tanpa Pemilihan Fitur**	**Pemilihan Fitur**
1.	80,00%	**80,48%**	61,43%	69,05%
2.	77,62%	**78,57%**	64,29%	69,05%
3.	80,95%	**81,43%**	61,43%	68,57%
4.	78,57%	**79,52%**	61,43%	68,09%
5.	79,05%	**79,05%**	62,38%	70,00%
6.	78,57%	**78,57%**	60,95%	70,00%
7.	77,62%	**78,09%**	62,38%	67,62%
8.	78,57%	**79,05%**	60,00%	70,48%
9.	77,14%	**78,57%**	63,33%	69,52%
10.	**80,95%**	80,48%	60,00%	67,62%

TABEL XII. HASIL AKURASI METODE KLASIFIKASI

Dataset	**Correlated *Naive Bayes***		***Naive Bayes***	
	Tanpa Pemilihan Fitur	**Pemilihan Fitur**	**Tanpa Pemilihan Fitur**	**Pemilihan Fitur**
Pima Indian Diabetes	67,30%	**71,40%**	64,26%	69,79%
Thyroid	78,90%	**79,38%**	61,76%	68,83%

IV. PENUTUP

Metode *Correlated Naive Bayes* dan *Wrapper* menghasilkan akurasi lebih baik dibandingkan tanpa seleksi fitur pada *dataset* Pima Indian Diabetes dan Thyroid. Hal ini disebabkan metode *Correlated Naive Bayes* hanya mengklasifikasikan fitur-fitur yang relevan pada *dataset* yang digunakan sehingga terjadi kenaikan akurasi Pima Indian Diabetes sebesar 4,1% dan Thyroid sebesar 0,48%. Penelitian selanjutnya dapat menggunakan kombinasi metode seleksi fitur berbasis *Filter* dan *Wrapper* untuk pemilihan atribut *dataset* kesehatan seperti Pima Indian Diabetes dan Thyroid.

UCAPAN TERIMA KASIH

Terima kasih kepada DRPM DIKTI untuk dana penelitian dalam skema Penelitian Dosen Pemula (PDP) tahun pelaksanaan 2020 yang sudah diberikan, sehingga penelitian ini dapat terlaksana.

REFERENSI

[1] J. D. Álvarez, J. A. Matias-Guiu, M. N. Cabrera-Martín, J. L. Risco-Martín, and J. L. Ayala, "An application of machine learning with feature selection to improve diagnosis and classification of neurodegenerative disorders," *BMC Bioinformatics*, vol. 20, no. 1, pp. 1–12, 2019, doi: 10.1186/s12859-019-3027-7.

[2] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 1578–1585, 2018, doi: 10.1016/j.procs.2018.05.122.

[3] M. A. Fahmiin and T. H. Lim, "Evaluating the Effectiveness of Wrapper Feature Selection Methods with Artificial Neural Network Classifier for Diabetes Prediction," in *Testbeds and Research Infrastructures for the Development of Networks and Communications*, 2020, pp. 3–17.

[4] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, "Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 13, no. 5, pp. 971–989, 2016, doi: 10.1109/TCBB.2015.2478454.

[5] N. K. Suchetha, A. Nikhil, and P. Hrudya, "Comparing the Wrapper Feature Selection Evaluators on Twitter Sentiment Classification," in *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*, 2019, pp. 1–6, doi: 10.1109/ICCIDS.2019.8862033.

[6] E. Hancer, B. Xue, and M. Zhang, "Differential evolution for filter feature selection based on information theory and feature ranking," *Knowledge-Based Syst.*, vol. 140, pp. 103–119, 2018, doi: 10.1016/j.knosys.2017.10.028.

[7] S. L. Shiva Darshan and C. D. Jaidhar, "Performance Evaluation of Filter-based Feature Selection Techniques in Classifying Portable Executable Files," *Procedia Comput. Sci.*, vol. 125, pp. 346–356, 2018, doi: 10.1016/j.procs.2017.12.046.

[8] M. Alirezanejad, R. Enayatifar, H. Motameni, and H. Nematzadeh, "Heuristic filter feature selection methods for medical datasets," *Genomics*, vol. 112, no. 2, pp. 1173–1181, 2020, doi: 10.1016/j.ygeno.2019.07.002.

[9] H. Zhou, X. Wang, and Y. Zhang, "Feature selection based on weighted conditional mutual information," *Appl. Comput. Informatics*, no. xxxx, 2020, doi: 10.1016/j.aci.2019.12.003.

[10] C. Liu, W. Wang, Q. Zhao, X. Shen, and M. Konan, "A new feature selection method based on a validity index of feature subset," *Pattern Recognit. Lett.*, vol. 92, pp. 1–8, 2017, doi: 10.1016/j.patrec.2017.03.018.

[11] S. S. Hameed, O. O. Petinrin, A. O. Hashi, and F. Saeed, "Filter-wrapper combination and embedded feature selection for gene expression data," *Int. J. Adv. Soft Comput. its Appl.*, vol. 10, no. 1, pp. 90–105, 2018.

[12] B. A. Muktamar, N. A. Setiawan, and T. B. Adji, "Pembobotan Korelasi pada Naive Bayes Classifier," *Semin. Nas. Teknol. Inf. dan Multimed. 2015 STMIK AMIKOM Yogyakarta, 6-8 Februari 2015*, no. 1, pp. 43–47, 2015.

[13] H. Hairani, G. Nugraha, M. Nurkholis Abdillah, and M. Innuddin, "Komparasi Akurasi Metode Correlated Naive Bayes Classifier dan Naive Bayes Classifier untuk Diagnosis Penyakit Diabetes," *InfoTekJar (Jurnal Nasional Informatika dan Teknologi Jaringan)*, vol. 3, no. 1, pp. 6–11, 2018, doi: 10.30743/infotekjar.v3i1.558

[14] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE untuk menangani ketidakseimbangan kelas dalam klasifikasi penyakit diabetes dengan C4.5, SVM, dan naive Bayes," *Jurnal Teknologi. dan Sistem Komputer*, vol. 8, no. 2, pp. 89–93, Apr. 2020, doi: https://doi.org/10.14710/jtsiskom.8.2.2020.89-93.

[15] Hairani, M. N. Abdillah, and M. Innuddin, "An Expert System for Diagnosis of Rheumatic Disease Types Using Forward Chaining Inference and Certainty Factor Method," in *2019 International Conference on Sustainable Information Engineering and Technology (SIET)*, 2019, pp. 104–109, doi: 10.1109/SIET48054.2019.8986035.

[16] S. H. A. Aini, Y. A. Sari, and A. Arwan, "Seleksi Fitur Information Gain untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naïve Bayes," *J. Pengemb. Teknol. Inf. dan Ilmu Komputer; Vol 2 No 9*, vol. 2, no. 9, pp. 2546–2554, Feb. 2018.

[17] H. Zheng, H. W. Park, D. Li, K. H. Park, and K. H. Ryu, "A Hybrid Feature Selection Approach for Applying to Patients with Diabetes Mellitus : KNHANES," in *2018 5th NAFOSTED Conference on Information and Computer Science (NICS)*, 2018, pp. 110–113.

[18] F. Kayaalp, M. S. Basarslan, and K. Polat, "A hybrid classification example in describing chronic kidney disease," in *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, 2018, pp. 1–4, doi: 10.1109/EBBT.2018.8391444.

[19] N. El Aboudi and L. Benhlima, "Review on wrapper feature selection approaches," in *2016 International Conference on Engineering & MIS (ICEMIS)*, 2016, pp. 1–5, doi: 10.1109/ICEMIS.2016.7745366.

[20] S. Manikandan, E. Susi, and S. Abirami, "Feature Selection on High Dimensional Data using Wrapper Based Subset Selection," in *2017 Second International Conference on Recent Trend and Challenges in Computational Models*, 2017, pp. 320–325, doi: 10.1109/ICRTCCM.2017.58.

[21] O. Somantri and M. Khambali, "Feature Selection Klasifikasi Kategori Cerita Pendek Menggunakan Naïve Bayes dan Algoritme Genetika," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 6, no. 3, pp. 301–306, 2017, doi: 10.22146/jnteti.v6i3.332.

[22] I. Santoso, W. Gata, and A. B. Paryanti, "Penggunaan Feature Selection di Algoritma Support Vector Machine untuk Sentimen Analisis Komisi Pemilihan Umum," *Rekayasa Sist. dan Teknol. Inf.*, vol. 3, no. 3, pp. 364–370, 2019.

[23] R. Nair and A. Bhagat, "Feature selection method to improve the accuracy of classification algorithm," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 6, pp. 124–127, 2019.

[24] H. Marcos and H. Utomo, "Perbandingan Kinerja Algoritme C.45 Dan Naive Bayes Mengklasifikasi Penyakit Diabetes," *J. Inform.*, vol. 15, no. 2, pp. 141–148, 2015, doi: 10.30873/ji.v15i2.596.

[25] B. A. Muktamar, N. A. Setiawan, and T. B. Adji, "Analisis Perbandingan Tingkat AKurasi Algoritma Naive Bayes Classifier dengan Correlated-Naive Bayes Classifier," *Semin. Nas. Teknol. Inf. dan Multimed. 2015 STMIK AMIKOM Yogyakarta, 6-8 Februari 2015*, pp. 49–54, 2015.

[26] Z. Ulhaq and T. B. Adji, "Technique (SMOTE) dengan Correlated Naïve Bayes pada Klasifikasi Siswa Berkesulitan Belajar," in *CITEE*, 2017, pp. 201–205.