Perbandingan Seleksi Fitur Sequential, Chi-Square, dan Embedded Pada Klasifikasi Penyakit Kanker Payudara Menggunakan Algoritma Random Forest

Yudha Alif Auliya[1], **Muhammad 'Ariful Furqon**[2], **Nico Wibiyanto** [3]
[1,2,3]Fakultas Ilmu Komputer, Universitas Jember
Email: yudha.alif@unej.ac.id

Abstract. *Cancer is typically linked to malignant tumors that can metastasize to extensive body tissues. Breast cancer arises from the uncontrolled proliferation of breast cells, resulting in the formation of benign and malignant tumors. Breast cancer presents various indicators, including small, round, and soft lumps associated with benign breast conditions and non-cancerous growths. In contrast, malignant breast cancer presents as asymmetrical, irregular, painful, and various other manifestations. If untreated, the tumor may metastasize and present a fatal risk. This study intends to evaluate the efficacy of Sequential Feature Selection, Chi-Square, and Embedded methods in classifying breast cancer, alongside implementing hyperparameter optimization via grid search on the random forest algorithm. This study utilizes the Wisconsin Breast Cancer dataset from the UCI Machine Learning Repository, comprising 569 data entries, 30 attributes, and 1 class label. The performance of the model is assessed using a Confusion matrix, which quantifies accuracy, precision, recall, and F1-score. The test results were derived from twenty testing schemes employing a combination of data splitting, cross-validation, and hyperparameter tuning via grid search. The optimal performance outcomes were achieved using the random forest model, which was subjected to hyperparameter tuning alongside SFS feature selection. The integration of 20 features yielded an accuracy of 97.37%, precision of 95.83%, recall of 97.87%, and an F1 score of 96.84%. The employed prediction model demonstrates effective performance in identifying both positive and negative classes. The model accurately predicted the true negative class in 66 instances. The model accurately identified the true positive class in 46 instances. One instance involved the model predicting a false positive class, while another instance involved the model predicting a false negative class. These results demonstrate that the model exhibits a high degree of accuracy with negligible prediction errors.*
Keywords: *Breast Cancer, Random Forest, Selection Feature, Embedded*

Abstrak. *Kanker umumnya terkait dengan tumor ganas yang dapat menyebar ke jaringan tubuh yang lebih luas. Kanker payudara terjadi akibat pertumbuhan sel-sel payudara secara berlebih kemudian membentuk tumor yang bersifat jinak (benigna) dan ganas (maligna). Penyakit kanker payudara mempunyai beberapa tanda, seperti mempunyai benjolan kecil bulat, dan lembut untuk kanker payudara jinak sehingga memiliki keadaan dan pertumbuhan yang tidak bersifat kanker dan sebaliknya untuk kanker payudara ganas mempunyai bentuk yang tidak simetris, kasar, terasa nyeri, dan lainnya. Jika dibiarkan, tumor dapat menyebar ke seluruh tubuh dan memiliki resiko kematian. Penelitian ini bertujuan membandingkan kinerja Seleksi Fitur Sequential, Chi-Square, dan Embedded pada klasfikasi penyakit kanker payudara serta menerapkan tuning hyperparameter menggunakan grid search pada algoritma random forest. Dataset yang digunakan dalam penelitian ini adalah data Wisconsin Breast Cancer yang berasal dari UCI Machine Learning Repository dengan total 569 baris data, 30 fitur, dan 1 label kelas. Metode evaluasi performa model yang digunakan yakni Confusion matrix dengan mengukur tingkat akurasi, presisi, recall, dan f1-score yang dihasilkan oleh model. Berdasarkan hasil pengujian yang dilakukan dengan dua puluh skema pengujian menggunakan kombinasi splitting, cross validation dan hyperparameter tuning grid search. Hasil performa terbaik diperoleh model random forest yang dilakukan hyperparameter tuning dengan seleksi fitur SFS kombinasi 20 fitur menghasilkan akurasi 97,37%, presisi 95,83%, recall 97,87% dan F1 score 96,84%. Model prediksi yang digunakan menunjukkan performa yang baik dalam mendeteksi kelas positif dan negatif. Terdapat 66 kasus di mana model dengan benar memprediksi kelas true negatif. Model berhasil memprediksi kelas true positif sebanyak 46 kasus. 1 kasus model memprediksi kelas false positif dan 1 kasus model memprediksi kelas false negatif. Secara keseluruhan, hasil ini*

 ISSN: 2579-566X (Online)
ISSN: 2477-5274 (Print)

menunjukkan bahwa model memiliki tingkat akurasi yang tinggi dengan kesalahan prediksi yang minim.
Kata Kunci: *Kanker PayudaraRandom Forest, seleksi fitur, Hyperparameter Tuning, Grid Search, Sequential Fitur seleksi, Chi-Square, Embedded*

1. Pendahuluan

Kanker payudara merupakan salah satu penyakit yang paling mengancam jiwa di seluruh dunia. Untuk menurunkan tingkat kematian akibat kanker payudara, modernisasi sistem klasifikasi dengan menerapkan kecerdasan buatan telah menjadi suatu kebutuhan. Pemanfaatan machine learning, yang merupakan bagian dari kecerdasan buatan, dalam bidang kesehatan mencakup berbagai tujuan, termasuk pengidentifikasian penyakit kanker payudara. Machine learning memungkinkan sistem untuk menyelesaikan tugas dengan tepat, termasuk membantu klasifikasi dengan lebih cepat dan akurat. Sistem berbasis machine learning dapat memberikan dukungan pengambilan keputusan bagi dokter, meningkatkan ketepatan, kualitas, dan kecepatan dalam klasifikasi penyakit kanker payudara.

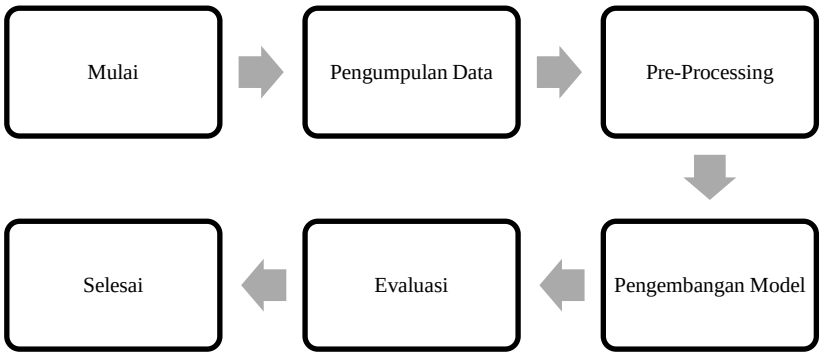
Data menunjukkan bahwa kanker payudara menyebabkan kematian sekitar 685.000 orang dan diperkirakan bisa mencapai 1 juta kasus pada tahun 2040 secara global [1]. Di Indonesia, menurut The Global Cancer Observatory (GLOBOCAN), kanker payudara menempati posisi pertama dengan 16,6% dari seluruh kasus kanker, atau sekitar 65.858 kasus [2]. Penyakit ini memiliki berbagai tanda, seperti benjolan kecil bulat dan lembut untuk kanker payudara jinak, serta bentuk yang tidak simetris, kasar, dan terasa nyeri untuk kanker payudara ganas [3]. Jika dibiarkan, tumor dapat menyebar ke seluruh tubuh dan memiliki risiko kematian yang tinggi [4].

Penelitian sebelumnya terdapat perbandingan antar Algoritma klasifikasi antara lain Random Forest (RF), k-NN, gradientB, Gaussian Naïve Bayes, MLP, linear regression, XGboost, MLP regression, Decision Tree dan Support Vector Machine (SVM). Dimana, hasil performa akurasi terbaik dicapai oleh SVM sebesar 90% [5]. Pada penelitian lainya yakni "A Linear Discriminant Analysis and Classification Model for Breast Cancer Diagnosis" menyatakan Algoritma RF dan SVM yang ditambahkan seleksi fitur yaitu Linear Discriminant Analysis (LDA) bisa meningkatkan akurasi. Dimana, hasil performa akurasi terbaik dicapai oleh SVM sebesar 96,4% dan RF sebesar 95.6% [6]. Sedangkan pada penelitian yang lain yakni "Exploring the performance of feature selection method using breast cancer dataset", terdapat hasil performa Algoritma RF dengan seleksi fitur yaitu Sequential, chi-square dan embedded belum dibandingkan dengan Algoritma klasifikasi SVM [7] [8].

Berdasarkan analisis tersebut, terdapat urgensi untuk melakukan perbandingan performa ketiga metode seleksi fitur yaitu: Sequential, Chi-square, dan Embedded. Penelitian ini akan menggunakan dataset Wisconsin Breast Cancer Dataset (WBCD) [9]. Penggunaan dataset ini memungkinkan hasil penelitian dapat dibandingkan dengan studi-studi sebelumnya dan memberikan kontribusi yang signifikan dalam pengembangan sistem klasifikasi kanker payudara yang lebih akurat [10]. Selain itu, penelitian ini akan mengevaluasi performa model menggunakan cross validation dengan nilai 5 K fold dan splitting 80:20 dengan confusion matrix, serta melakukan hyperparameter tuning dengan metode grid search untuk mengoptimalkan model setelah proses seleksi fitur [11]. Hasil dari penelitian ini diharapkan dapat meningkatkan performa model sehingga dapat membantu meningkatkan deteksi dini dan penanganan penyakit kanker payudara.

2. Metode Penelitian

Terdapat tahapan penelitian yang dilakukan oleh peneliti disajikan pada Gambar 1.



Gambar 1. Tahapan Penelitian

Pengumpulan Data

Dataset yang digunakan adalah Wisconsin Diagnostic Breast Cancer (WDBC) (UCI ML Repository), berisi 569 entri dan 30 atribut fitur (ekstraksi FNA: mean, se, worst untuk 10 fitur) serta 1 label (B/M). Distribusi kelas pada WDBC: 357 benign (B) dan 212 malignant (M).

Preprocessing

Tahap penelitian ini dilakukan drop fitur yang tidak diperlukan yaitu id. Hal ini bertujuan membuang atribut yang tidak berfungsi atau tidak digunakan. Tahap ini juga menggunakan library MinMaxScaler untuk mengstandarisasi nilai. Dengan demikian, fitur-fitur dalam data yang telah dikumpulkan akan memiliki rentang 0 hingga 1. Tujuannya adalah untuk menormalisasi nilai-nilai pada dataset agar memiliki skala yang seragam. Setelah itu, melakukan seleksi fitur yang bertujuan untuk mendapatkan fitur terbaik. Seleksi fitur yang digunakan terdiri dari SFS, Chi-Square, dan Embedded [12].

Metode seleksi fitur SFS beroperasi dengan direction forward, yang berarti proses dimulai dengan himpunan fitur yang hanya berisi satu fitur yang memiliki kontribusi signifikan terhadap performa model. Dalam langkah-langkah selanjutnya, fitur-fitur lain secara bertahap ditambahkan ke dalam himpunan berdasarkan kriteria tertentu hingga mencapai himpunan fitur yang optimal [13].

Metode Chi-Square dengan library chi2. Merupakan seleksi fitur yang digunakan untuk memilih fitur-fitur yang paling relevan. Algoritma ini menghitung nilai chi-square untuk setiap fitur dan memilih fitur-fitur dengan nilai chi-square tertinggi sebagai fitur-fitur yang paling optimal terkait dengan variabel target [14]. Metode Embedded pada RF dengan library RandomForestClassifier. Dalam seleksi fitur menggunakan RF, setiap fitur akan dinilai berdasarkan importance score yang sering digunakan untuk pemisahan dan yang secara signifikan mengurangi impurity akan dianggap penting [15]. Terakhir, Fitur-fitur tersebut kemudian diurutkan berdasarkan importance score, sehingga dapat dipilih sejumlah fitur teratas sesuai dengan fitur optimal. Untuk melakukan itu Sklearn menyediakan tools feature importance [16].

Pembangunan Model

Algoritma klasifikasi yang dihgunakan adalah RF. Pada tahap pembangan model ini, setelah melakukan proses pre-prosesing terdapat tiga data hasil seleksi fitur yang berasal dari SFS, Chi-Square, dan Embedded nantinya masing masing hasil seleksi fitur tersebut akan dilakukan pembangunan model menggunakan RF [17]. Peneliti juga akan meninjau penggunaan hyperparameter dengan hyperparameter tuning. Hal ini bertujuan hyperparameter yang telah digunakan setelah seleksi fitur, agar model memiliki performa dengan baik [18]. Search space atau nilai hyperparameter yang akan digunakan dalam metode grid search untuk melakukan tuning hyperparameter pada algoritma RF disajikan pada Tabel. 1.

Tabel 1. Penyetelan Hyperparameter RF.

Hyperparameter Criterion	Nilai [gini, entropy]
estimators	[100, 200, 300, 400, 500]
Max depth	[2, 4, 6, 8, 10]
Min samples leaf	[1, 5, 10]
min_samples_split	[2,4,6]
Hyperparameter	Nilai

Evaluasi Hasil

Peneliti akan melakukan tahapan evaluasi kinerja hasil dari pembangunan model. Pada penelitian ini pembagian dataset untuk data latih dan data uji mengacu pada penelitian sebelumnya dengan nilai sebesar 80:20 [19] dan peneliti juga akan melakukan pengujian dengan K-Fold Cross Validation dengan nilai 5-Fold [20]. Penentuan jumlah fitur dalam proses penerapan metode seleksi fitur mengacu pada penelitian sebelumnya yaitu 8 kombinasi fitur 12 kombinasi fitur dan 20 kombinasi fitur [21]. Pengujian menggunakan SVM dan RF dengan Seleksi Fitur Sequential, Chi-Square, dan Embedded ini menggunakan confusion matrix untuk mengetahui performa dari model disertai dengan penerapan hyperparameter tuning dengan algoritma grid search.

3. Hasil dan Pembahasan

Pengumpulan Data

Adapun fitur-fitur yang digunakan pada dataset ini sebagai penentu seseorang dapat terdiagnosa penyakit kanker payudara adalah rata-rata, standard error dan pengukuran terburuk dari sepuluh fitur, dan disertai dengan 1 label kelas yang berisikan jinak (Banign) atau ganas (Maligna). Deskripsi data penyakit kanker payudara dapat dilihat pada Tabel 2.

Tabel 2. Dataset Penyakit Kanker Payudara

 ISSN: 2579-566X (Online)
ISSN: 2477-5274 (Print)

id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	…	fractal_dimension_worst
1	M	17.99	10.38	122.8	1001	…	0.1189
2	M	20.57	17.77	132.9	1326	…	0.08902
3	M	19.69	21.25	130	1203	…	0.08758
4	M	11.42	20.38	77.58	386.1	…	0.173
5	M	20.29	14.34	135.1	1297	…	0.07678
6	M	12.45	15.7	82.57	477.1	…	0.1244
7	M	18.25	19.98	119.6	1040	…	0.08368
8	M	13.71	20.83	90.2	577.9	…	0.1151
9	M	13	21.82	87.5	519.8	…	0.1072
10	M	12.46	24.04	83.97	475.9	…	0.2075
11	M	16.02	23.24	102.7	797.8	…	0.08452
	…	…	…	…	…	…	…
569	B	7.76	24.54	47.92	181	…	0.07039

Preprocessing

Dilakukan pengecekan data yang memiliki missing values. Jika terdapat missing values, hal tersebut mempengaruhi akurasi model yang akan diimplementasikan dan pada data tersebut tidak terdapat missing values dan melakukan penghapusan atribut yang tidak perlu. Dilakukan transformasi data yaitu mentransfomasikan kolom target atau kelas yang awalnya *benign* (B) menjadi 0 dan *malign* (M) menjadi 1. Kemudian, melakukan standarisasi nilai. Standarisasi nilai dilakukan dengan *MinMaxScaler().* Dengan demikian, fitur dalam data yang telah dikumpulkan akan memiliki rentang 0 hingga 1. Tujuan dari standarisasi nilai-nilai dataset adalah untuk memastikan bahwa nilai-nilai tersebut memiliki skala yang seragam. Hasil tranformasi data disajikan pada Tabel 3.

Tabel 3. Hasil Transformasi Data

diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	…	fractal_dimension_worst
1	0,065527205	0,257693608	0,077322521	0,034368832	…	0,529501526
1	0,656202556	0,570172472	0,674206857	0,489401871	…	0,621905731
1	0,07257946	0,140344944	0,08023901	0,038831202	…	0,253136656
1	0,144914051	0,524518093	0,142907953	0,075774479	…	0,224059003
1	0,121406533	0,174839364	0,118295632	0,060713979	…	0,213292642
1	0,33542289	0,602975989	0,347204439	0,20955977	…	0,670227196
1	0,329056271	0,33885695	0,342225068	0,206212992	…	0,279331977
…	…	…	…	…	…	…
1	0,223272442	0,278322624	0,215891307	0,124388569	…	0,222787386

Seleksi Fitur

Pengimplementasian algoritma klasifikasi dengan metode seleksi fitur SFS, chi-square, dan Embedded. Algoritma klasifikasi yang digunakan yakni *Random Forest*. Parameter nilai fitur yang digunakan dalam teknik seleksi fitur yaitu 8,12 dan 20.

Hyperparameter tunning

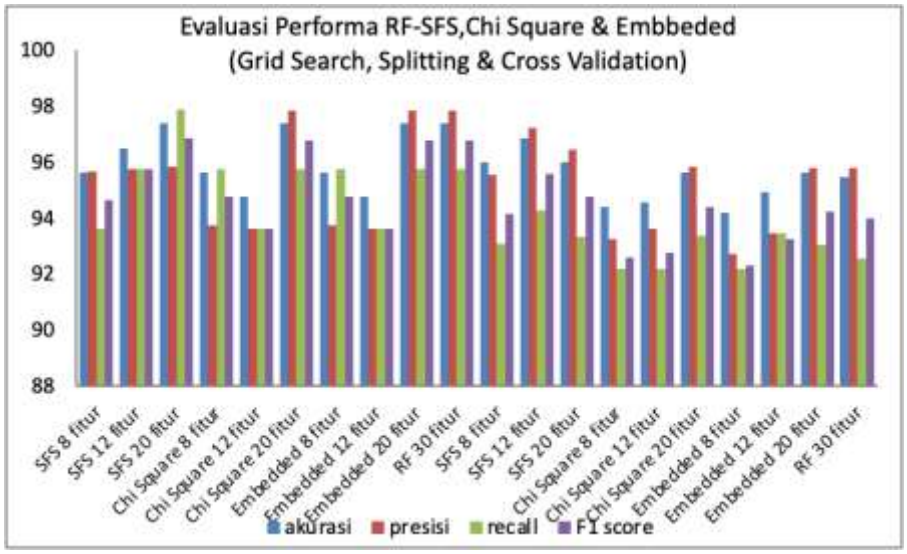
Tahapan hyperparameter tuning pada algoritma klasifikasi yakni RF. Proses ini bertujuan untuk mendapatkan hyperparameter untuk digunakan setelah kombinasi fitur didapatkan, sehingga kombinasi hyperparameter terbaik dapat menghasilkan performa model yang baik. Proses grid search dimulai dengan memilih model atau algoritma klasifikasi. Kemudian, menentukan rentang nilai hyperparameter yang ingin diproses. Pada RF, mempertimbangkan criterion, estimators, Max depth, Min samples leaf, dan min_samples_split. Setelah mengidentifikasi kombinasi nilai hyperparameter, algoritma grid search membentuk sebuah grid yang mencakup semua kemungkinan kombinasi tersebut. Selanjutnya, validasi silang digunakan untuk mengevaluasi kinerja model dari setiap kombinasi hyperparameter yang diuji. Evaluasi dilakukan dengan stratified 5-fold nested cross-validation (outer fold untuk estimasi performa, inner fold untuk hyperparameter tuning) dan juga hold-out stratified 80:20 yang diulang 10 kali. Semua pembagian data bersifat stratified untuk menjaga proporsi kelas. Hasil hyperparameter tunning random forest dengan grid search dapat dilihat pada Tabel 4.

Tabel 4. Hasil hyperparameter tuning RF-grid search

Seleksi fitur	**SFS (Sequential Feature Selection)**	**Chi Square**	**Embedded**
Jumlah Fitur	Tuned Hyperparameter	Tuned Hyperparameter	Tuned Hyperparameter
8 Splitting	criterion': 'entropy', 'max_depth': 8, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100	criterion': 'gini', 'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200	criterion': 'gini', 'max_depth': 6, 'min_samples_leaf': 5, 'min_samples_split': 2, 'n_estimators': 400
12 Splitting	criterion': 'gini', 'max_depth': 8, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300	criterion': 'gini', 'max_depth': 6, 'min_samples_leaf': 1, 'min_samples_split': 4, 'n_estimators': 300	criterion': 'entropy', 'max_depth': 8, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 400
20 Splitting	criterion': 'gini', 'max_depth': 8, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100	criterion': 'gini', 'max_depth': 6, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 500	criterion': 'entropy', 'max_depth': 8, 'min_samples_leaf': 1, 'min_samples_split': 4, 'n_estimators': 300
8 Cross Validation	'criterion': 'gini', 'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100	criterion': 'gini', 'max_depth': 6, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300	criterion': 'gini', 'max_depth': 8, 'min_samples_leaf': 1, 'min_samples_split': 6, 'n_estimators': 100
12 Cross Validation	criterion': 'gini', 'max_depth': 8, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100	criterion': 'entropy', 'max_depth': 4, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 400	criterion': 'gini', 'max_depth': 6, 'min_samples_leaf': 1, 'min_samples_split': 6, 'n_estimators': 200
20 Cross Validation	criterion': 'gini', 'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100	criterion': 'entropy', 'max_depth': 6, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300	criterion': 'entropy', 'max_depth': 6, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 500

Evaluasi Hasil

Pembangunan model yang disertai hyperparameter tuning menggunakan grid search mempengaruhi hasil evaluasi performa model. Performa model yang dibangun dengan grid search yakni RF tanpa seleksi fitur dengan splitting memiliki akurasi 97,37%, presisi 97,83%, recall 95,74% dan F1 score 96,77% dengan nilai hyperparameter criterion entropy, max depth sebesar 8, min samples leaf sebesar 1, min samples split sebesar 2, dan n estimators sebesar 300 sedangkan cross validation memiliki akurasi 95.43%, presisi 95.78%, recall 92.55% dan F1 score 93.97% dengan nilai hyperparameter criterion entropy, max_depth 8, min_samples_leaf 1, min_samples_split 6, dan n_estimators 300. Secara detail hasil pengujian dapat dilihat pada Gambar 2.

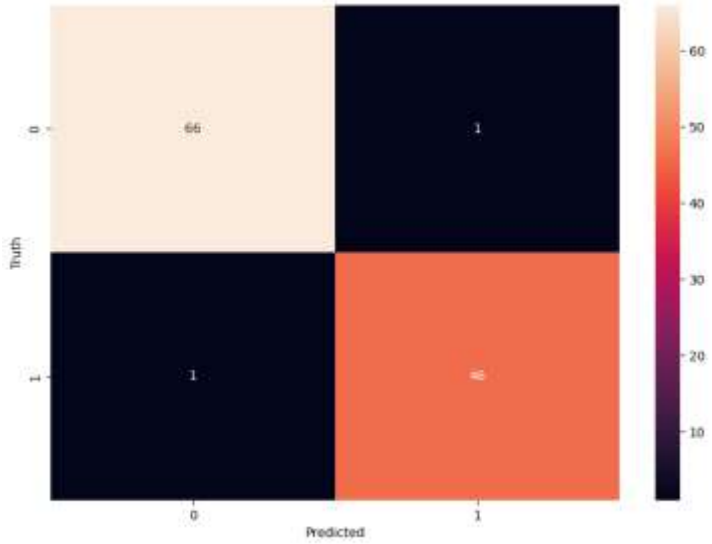


Gambar 2. Hasil Evaluasi Performa RF-SFS, Chi Square & Embbeded

Dilakukan pengukuran performa model dengan tiga metode seleksi fitur dari SFS, Chi square dan Embedded. Pengujian model menggunakan hyperparameter tuning dengan seleksi fitur SFS kombinasi 20 fitur memiliki hasil yaitu akurasi 97,37%, presisi 95,83%, recall 97,87% dan F1 score 96,84%. Pengujian model dengan

 ISSN: 2579-566X (Online)
ISSN: 2477-5274 (Print)

seleksi fitur chi square menghasilkan peningkatan performa tertinggi ketika menggunakan kombinasi 20 fitur. Cross validation seleksi fitur chi square kombinasi 20 fitur memiliki hasil yaitu akurasi 96.84%, presisi 97.22%, recall 94.25% dan F1 Score 95.58%. Pengujian model dengan seleksi fitur embedded menghasilkan peningkatan performa tertinggi ketika menggunakan kombinasi 20 fitur yaitu akurasi 95.61%, presisi 95.78%, recall 93.05% dan F1 Score 94.23%.



Gambar 3. Hasil confusion matrix random forest dengan Grid search, 20 fitur dan proporsi 80:20

Berdasarkan Gambar 3. model prediksi yang digunakan, bahwa menunjukkan performa yang cukup baik dalam mendeteksi kelas positif dan negatif. Terdapat 66 kasus di mana model dengan benar memprediksi kelas negatif, yang disebut sebagai True Negative. Model juga berhasil memprediksi kelas positif dengan benar pada 46 kasus, dikenal sebagai True Positive. Namun, terdapat 1 kasus di mana model memprediksi kelas positif padahal sebenarnya negatif, yang dikenal sebagai False Positive. Selain itu, ada 1 kasus di mana model memprediksi kelas negatif padahal sebenarnya positif, yang disebut sebagai False Negative. Secara keseluruhan, hasil ini menunjukkan bahwa model memiliki tingkat akurasi yang tinggi dengan kesalahan prediksi yang minim.

4. Kesimpulan

Berdasarkan hasil pengujian yang dilakukan dengan dua puluh skema pengujian menggunakan kombinasi splitting, cross validation dan hyperparameter tuning grid search. Hasil performa terbaik diperoleh model random forest yang dilakukan hyperparameter tuning dengan seleksi fitur SFS kombinasi 20 fitur menghasilkan akurasi 97,37%, presisi 95,83%, recall 97,87% dan F1 score 96,84%. Model prediksi yang digunakan menunjukkan performa yang baik dalam mendeteksi kelas positif dan negatif. Terdapat 66 kasus di mana model dengan benar memprediksi kelas negatif, yang disebut sebagai True Negative. Model juga berhasil memprediksi kelas true positif dengan benar pada 46 kasus. Namun, terdapat 1 kasus model memprediksi kelas false positif. Selain itu terdapat 1 kasus model memprediksi kelas false negatif. Secara keseluruhan, hasil ini menunjukkan bahwa model memiliki tingkat akurasi yang tinggi dengan kesalahan prediksi yang minim

Referensi

[1] Adebiyi, M. O., Arowolo, M. O., Mshelia, M. D., & Olugbara, O. O. (2022). A Linear Discriminant Analysis and Classification Model for Breast Cancer Diagnosis. *Applied Sciences*, *12*(22), 11455. https://doi.org/10.3390/app122211455

[2] Ali, N. M., Besar, R., & Aziz, N. A. A. (2023). A case study of microarray breast cancer classification using machine learning algorithms with grid search cross validation. *Bulletin of Electrical Engineering and Informatics*, *12*(2), 1047–1054. https://doi.org/10.11591/eei.v12i2.4838

[3] Anggoro, D. A., & Afdallah, N. A. (2022). Grid Search CV Implementation in Random Forest Algorithm to Improve Accuracy of Breast Cancer Data. *International Journal on Advanced Science, Engineering and Information Technology*, *12*(2), 515–520. https://doi.org/10.18517/ijaseit.12.2.15487

[5] Assegie, T. A., Tulasi, R. L., Elanangai, V., & Kumar, N. K. (2022). Exploring the performance of feature selection method using breast cancer dataset. *Indonesian Journal of Electrical Engineering and Computer Science*, *25*(1), 232. https://doi.org/10.11591/ijeecs.v25.i1.pp232-237

[6] Chaeikar, S. S., Manaf, A. A., Alarood, A. A., & Zamani, M. (2020). PFW: Polygonal fuzzy weighted—an SVM kernel for the classification of overlapping data groups. *Electronics (Switzerland)*, *9*(4). https://doi.org/10.3390/electronics9040615

[7] Das, L. N., Saini, S., Kataria, P., & Dipanshu. (2022). Breast cancer detection from histopathological images using machine learning models. *International Journal of Health Sciences*, 9542–9553. https://doi.org/10.53730/ijhs.v6nS3.8254

[8] Dasariraju, S., Huo, M., & McCalla, S. (2020). Detection and classification of immature leukocytes for diagnosis of acute myeloid leukemia using random forest algorithm. *Bioengineering*, *7*(4), 1–12. https://doi.org/10.3390/bioengineering7040120

[9] Deepa, B. G., & Senthil, S. (2020). Constructive Effect of Ranking Optimal Features Using Random Forest, Support Vector Machine and Naïve Bayes for Breast Cancer Diagnosis. *Big Data Analytics and Intelligence: A Perspective for Health Care*, *September 2020*, 189–202. https://doi.org/10.1108/978-1-83909-099-820201014

[10] Fajri, M., & Primajaya, A. (2023). Komparasi Teknik Hyperparameter Optimization pada SVM untuk Permasalahan Klasifikasi dengan Menggunakan Grid Search dan Random Search. *Journal of Applied Informatics and Computing*, *7*(1), 14–19. https://doi.org/10.30871/jaic.v7i1.5004

[11] Hafid, H. (2023). Penerapan K-Fold Cross Validation untuk Menganalisis Kinerja Algoritma K-Nearest Neighbor pada Data Kasus Covid-19 di Indonesia. *Journal of Mathematics*, *6*(2), 161–168. http://www.ojs.unm.ac.id/jmathcos

[12] Jasim, A. A., Jalal, A. A., Abdulateef, N. M., & Talib, N. A. (2022). Effectiveness evaluation of machine learning algorithms for breast cancer prediction. *Bulletin of Electrical Engineering and Informatics*, *11*(3), 1516–1525. https://doi.org/10.11591/eei.v11i3.3621

[13] Jonathan, M., Rostianingsih, S., Palit, H. N., & Surabaya, J. S. (n.d.). *Pengaruh Feature Selection terhadap Kinerja C5 . 0 , XGBoost , dan Random Forest dalam Mengklasifikasikan Website Phishing*.

[14] Julianto, Y., Setiabudi, D. H., & Rostianingsih, S. (2022). Analisis Sentimen Ulasan Restoran Menggunakan Metode SVM. *Jurnal Infra*, *10*(1).

[15] Kamelia, M., & Agus, S. (2021). Fine Needle Aspiration Biopsy (FNAB) Massa Intraabdomen dipandu Ultrasonografi. *Health and Medical Journal*, *4*(1), 55–61. https://doi.org/10.33854/heme.v4i1.819

[16] Kusumarini, A. I., Hogantara, P. A., Fadhlurohman, M., & Nurul Chamidah, S. K. . M. K. (2021). Perbandingan Algoritma Random Forest, Naive Bayes, Dan Decision Tree Dengan Oversampling Untuk Klasifikasi Bakteri E.Coli. *Prosiding Seminar Nasional Mahasiswa Bidang Ilmu Komputer Dan Aplikasinya*, *2*(1), 792–799.

[17] Mahdi, A. N., & Mohsin, A. A. (2021). Machine learning classification based on Radom Forest algorithm: a review. *International Journal of Science and Business*, *5*(2), 128–142. https://doi.org/10.5281/zenodo.4471118

[18] Maseno, E. M., & Wang, Z. (2024). Hybrid wrapper feature selection method based on genetic algorithm and extreme learning machine for intrusion detection. *Journal of Big Data*. https://doi.org/10.1186/s40537-024-00887-9

[19] Praghakusma, A. Z., & Charibaldi, N. (2021). Komparasi Fungsi Kernel Metode Support Vector Machine untuk Analisis Sentimen Instagram dan Twitter (Studi Kasus : Komisi Pemberantasan Korupsi). *JSTIE (Jurnal Sarjana Teknik Informatika) (E-Journal)*, *9*(2), 88. https://doi.org/10.12928/jstie.v9i2.20181

[20] Rusmalina, S. (2019). Pena medika. *Jurnal Kesehatan Pena Medika*, *9*(2), 48–54.Sandag, G. A. (2020). Prediksi Rating Aplikasi App Store Menggunakan Algoritma Random Forest. *CogITo Smart Journal*, *6*(2), 167–178. https://doi.org/10.31154/cogito.v6i2.270.167-178

[21] Spencer, R., Thabtah, F., Abdelhamid, N., & Thompson, M. (2020). Exploring feature selection and classification methods for predicting heart disease. *Digital Health*, *6*, 1–10. https://doi.org/10.1177/2055207620914777

ISSN: 2579-566X (Online)
ISSN: 2477-5274 (Print)

[Halaman ini sengaja dikosongkan]