SMARTLOCK: Jurnal Sains dan Teknologi
Volume 2 No 2, Tahun 2023: Hal. 32-38
e-ISSN: 2986-108X. Open Access at: https://e-jurnal.nobel.ac.id/index.php/smartlock
Penerbit: Program Studi Sistem dan Teknologi Informasi ITB Nobel Makassar, Sulsel.

Model Ensemble Algoritma Naive Bayes Dan Random Forest Dalam Klasifikasi Penyakit Paru-paru Untuk Meningkatkan Akurasi

Mursyid Ardiansyah[*1]*

[*1] *Ilmu Komputer, Institut Teknologi Sains dan Bisnis Muhammadiyah*

mursyidardiansyah@itsbm.ac.id[*1]

Abstract — This study explores an ensemble model combining Naive Bayes and Random Forest algorithms in the classification of lung diseases with the aim of improving accuracy. Involving a dataset of 30,000 instances, the research yields excellent performance in various aspects, including accuracy, precision, and recall. The combination of the Naïve Bayes model with Random Forest integrated using the VotingClassifier proves superior compared to using the Naïve Bayes model alone. Experimental results demonstrate that the ensemble model achieves an accuracy of 93%, with precision reaching 87.03%, and a recall of 100%. This superiority emphasizes that integrating the strengths of Naive Bayes and Random Forest in an ensemble approach can enhance the predictive capabilities of the classification system. The study significantly contributes to improving the diagnosis of lung diseases, opening opportunities for the development of more efficient classification systems in medical practice. Thus, this research not only enhances classification accuracy but also provides guidance for the development of artificial intelligence-based solutions in the healthcare domain.

Keyword — Naïve Bayes, Random Forest, Accuracy, Classification.

Abstrak — Penelitian ini membahas model ensemble yang menggabungkan algoritma Naive Bayes dan Random Forest dalam klasifikasi penyakit paru-paru dengan tujuan meningkatkan akurasi. Melibatkan dataset sebanyak 30.000 data, penelitian ini menghasilkan kinerja yang sangat baik dalam berbagai aspek, termasuk akurasi, presisi, dan recall. Penggabungan model Naïve Bayes dengan Random Forest yang terintegrasi menggunakan VotingClassifier mampu unggul dibandingkan dengan penggunaan model Naïve Bayes tunggal. Hasil eksperimen menunjukkan bahwa performa model ensemble mencapai akurasi sebesar 93%, dengan presisi mencapai 87.03%, dan recall mencapai 100%. Keunggulan ini menegaskan bahwa pengintegrasian kekuatan Naive Bayes dan Random Forest dalam pendekatan ensemble dapat meningkatkan kemampuan prediktif sistem klasifikasi. Penelitian ini memberikan kontribusi signifikan dalam meningkatkan diagnosis penyakit paru-paru, membuka peluang untuk pengembangan sistem klasifikasi yang lebih efisien dalam praktik medis. Dengan demikian, penelitian ini tidak hanya meningkatkan akurasi klasifikasi, tetapi juga memberikan panduan untuk pengembangan solusi berbasis kecerdasan buatan di bidang kesehatan.

Kata kunci — Naïve Bayes, Random Forest, Akurasi, Klasifikasi.

I. PENDAHULUAN

Penyakit paru-paru merupakan salah satu masalah kesehatan global yang memengaruhi kualitas hidup manusia secara signifikan. Dengan adanya kemajuan dalam teknologi informasi dan kecerdasan buatan, penerapan metode klasifikasi menjadi kritis dalam upaya diagnosis dini dan penanganan penyakit paru-paru. Dalam konteks ini, penelitian ini bertujuan untuk meningkatkan akurasi diagnosis penyakit paru-paru dengan menggabungkan kekuatan dua model algoritma, yaitu *Naive Bayes* dan *Random Forest,* melalui pendekatan *ensemble*.

Penyakit paru-paru, termasuk tetapi tidak terbatas pada penyakit seperti pneumonia, bronkitis, dan penyakit paru obstruktif kronis (PPOK), memiliki dampak yang signifikan pada kesehatan populasi. Diagnosis yang cepat dan akurat adalah kunci untuk memberikan perawatan yang tepat waktu, mengurangi tingkat mortalitas, dan meningkatkan kualitas hidup pasien. Oleh karena itu, pengembangan model klasifikasi yang handal menjadi suatu kebutuhan mendesak[1].

Algoritma *Naive Bayes* adalah metode klasifikasi yang berdasarkan pada teorema probabilitas Bayes. Metode ini mengasumsikan independensi antar fitur dan sangat efisien dalam mengolah data dengan dimensi tinggi[2]. Sementara itu, Random Forest merupakan metode ensemble yang memanfaatkan sejumlah besar pohon keputusan untuk meningkatkan akurasi dan mengatasi overfitting[3]. Kedua algoritma ini telah terbukti efektif dalam berbagai konteks klasifikasi, termasuk dalam diagnosis penyakit.

Model *ensemble* melibatkan penggabungan prediksi dari beberapa model untuk meningkatkan keakuratan dan kinerja[4]. Dalam penelitian ini, pendekatan ensemble akan mengintegrasikan kekuatan

algoritma *Naive Bayes* dan *Random Forest*. Gabungan ini diharapkan dapat mengoptimalkan kelebihan masing-masing model, sehingga menghasilkan sistem klasifikasi yang lebih andal dan responsif terhadap kompleksitas penyakit paru-paru.

Penelitian ini bertujuan untuk meningkatkan akurasi dalam klasifikasi penyakit paru-paru dengan menggabungkan model algoritma *Naive Bayes* dan *Random Forest* melalui pendekatan *ensemble*. Dengan memanfaatkan kombinasi dua metode klasifikasi yang berbeda, diharapkan penelitian ini dapat memberikan kontribusi signifikan dalam upaya diagnosis dini dan penanganan penyakit paru-paru. Peningkatan akurasi yang dihasilkan dapat menjadi landasan bagi pengembangan sistem klasifikasi yang lebih efisien dan dapat diandalkan dalam praktek medis.

II. PENELITIAN YANG TERKAIT

Mengimplementasikan algoritma Naïve Bayes untuk membuat "*Smart Heart Disease Prediction*" sehingga dapat membantu menyelesaikan masalah dalam mendiagnosa penyakit jantung. Dari penelitian tersebut dengan menggunakan model Naïve Bayes dapat menghasilkan akurasi 89.77%[5].

Klasifikasi harga dan spesifikasi *CPU* dan *eGPU* pada penilitan[6] menggunakan model Naïve Bayes untuk proses klasifikasinya. Kemudian untuk performanya memiliki akurasi 76.8%, presisi 100% Recall 76.4% dan F1-Score 86.6%

Penelitian[7] menggunakan *Random Forest* untuk membantu pemilihan variabel untuk model prediksi klasifikasi. Untuk hasil *Out of Bag Error Rate, Number of Variables, Computation Time* dan AUC memiliki hasil yang baik.

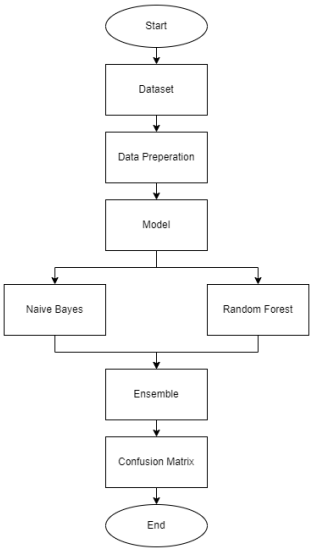
A robust technique of fake news detection using Ensemble Voting Classifer and comparison with other classifers penilitian yang dibuat oleh[8] adalah untuk mendeteksi berita palsu. *VotingClassifier* diintegrasikan dengan beberapa model termasuk model *Random Forest* dan memiliki rata-rata akurasi 94.5% untuk beberapa model klasifikasi

III. METODE PENELITIAN

A. Flowchart

Flowchart atau bagan alir adalah gambaran yang memvisualisasikan beberapa aspek sistem informasi atau proses-proses yang akan dilakukan[9]. Dimulai dari menyiakan *dataset*, mengolah *dataset* pada proses *data preperation*, pembuatan model menggunakan metode *Naïve Bayes* dikombinasikan dengan metode Random Forest lalu, diintegrasikan dengan model *Ensemble* metode *VotingClassifier*, kemudian hasil akhirnya adalah evaluasi menggunakan *Confusion Matrix*.

Gambar 1. *Flowchart*



B. Dataset

Dataset adalah kumpulan data yang terorganisir secara struktural, mencakup beragam informasi yang dapat diolah dan dianalisis. Dataset ini dapat berasal dari berbagai sumber, seperti hasil pengukuran, pengamatan, atau rekaman, dan digunakan sebagai dasar untuk pengembangan model, eksplorasi data, serta penelitian statistik dan machine learning[10]. Dengan menyimpan informasi yang bervariasi, dataset menjadi pondasi bagi pemahaman mendalam tentang pola, tren, dan hubungan yang mungkin terkandung di dalamnya.

C. *Data Preperation*

Data preparation merujuk pada serangkaian langkah dan proses yang dilakukan untuk membersihkan, mengorganisir, dan mengubah data mentah menjadi bentuk yang lebih sesuai untuk analisis atau pemodelan. Ini melibatkan tindakan seperti penanganan nilai yang hilang, normalisasi, transformasi fitur, dan pemilihan atribut, sehingga data menjadi siap untuk dieksplorasi atau digunakan dalam pembuatan model[11]. Proses data preparation mendukung pengambilan keputusan yang akurat dan memastikan bahwa data yang digunakan dalam analisis atau pembelajaran mesin memiliki kualitas yang memadai.

D. *Naïve Bayes*

Naive Bayes adalah sebuah metode klasifikasi dalam machine learning yang mendasarkan prediksinya pada asumsi bahwa setiap fitur dalam data bersifat independen satu sama lain, meskipun asumsi ini sederhana, metode ini efektif dalam memodelkan probabilitas kelas yang berbeda dalam sebuah dataset[12].

$$P(A \cap B) = P(A \mid B).P(B) = P(B \mid A).P(A) \quad (1)$$

$$P(A|B) = P(B|A)P(A)/P(B) \quad (2)$$

Keterangan:
P(B|A) adalah peluang kejadian B dengan syarat A terjadi
P(A|B) adalah peluang kejadian A dengan syarat B terjadi
P(A ∩B) adalah peluang kejadian A irisan B
P(A) Peluang kejadian A
P(B) Peluang kejadian B

E. *Random Forest*

Random Forest adalah suatu algoritma ensemble dalam machine learning yang menggabungkan hasil prediksi dari beberapa Decision Trees yang dibangun secara acak. Dengan memanfaatkan keberagaman dan kombinasi model-model ini, Random Forest dapat meningkatkan akurasi prediksi dan memiliki kemampuan yang baik untuk menangani overfitting pada data yang kompleks[13].

$$MSE = 1/N \sum_{i=1}^{N} (fi - yi)^2 \quad (3)$$

Keterangan:
N adalah jumlah data
Fi adalah nilai yang dikembalikan oleh model
Yi adalah nilai aktual untuk dala *i*

F. *Ensemble VotingClassifier*

Ensemble VotingClassifier adalah suatu pendekatan dalam machine learning di mana berbagai model yang berbeda, seperti Naive Bayes dan Random Forest, digabungkan untuk memberikan prediksi akhir berdasarkan mayoritas suara atau probabilitas. Dengan memanfaatkan kekuatan dan keunikannya masing-masing model, VotingClassifier bertujuan untuk meningkatkan kinerja prediksi secara keseluruhan dan meningkatkan ketahanan model terhadap variasi data[4].

G. *Confusion Matrix*

Confusion matrix adalah suatu metrik evaluasi dalam konteks klasifikasi pada machine learning yang menyajikan performa model dengan merinci hasil prediksi untuk setiap kelas target. Matrix ini memberikan informasi tentang jumlah true positive, true negative, false positive, dan false negative, memungkinkan analisis yang lebih mendalam terhadap sejauh mana model dapat membedakan kelas target dengan akurasi yang sesuai[14]. Confusion matrix menjadi instrumen penting dalam mengevaluasi keandalan model dan mengidentifikasi area di mana model dapat ditingkatkan untuk meningkatkan keakuratannya.

$$Akurasi = (TP+TN) / (TP+TN+FP+FN) \quad (4)$$
$$Presisi = TP / (TP+FP) \quad (5)$$
$$Recall = TP / (TP+FN) \quad (6)$$

Keterangan:
TP adalah *True Positive*
TN adalah *True Negative*
FP adalah *False Positive*
FN adalah *False Negative*

IV. HASIL DAN PEMBAHASAN

Dataset yang digunakan adalah "DATASET PREDIC TERKENA PENYAKIT PARU-PARU" yang diupload oleh Andot03 Bsrc di website kaggle(https://www.kaggle.com/datasets/andot03bsrc/dataset-predic-terkena-penyakit-paruparu) dengan jumlah 30000 data. Sedangkan jumlah kolom atau variabel yaitu 11.

Tabel 1.
Dataset

No	Usia	Jenis_Kelamin	Merokok	Bekerja	Rumah_Tangga	Aktivitas_Begadang	Aktivitas_Olahraga	Asuransi	Penyakit_Bawaan	Hasil
1	Tua	Pria	Pasif	Tidak	Ya	Ya	Sering	Ada	Tidak	Ya
2	Tua	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Ada	Tidak
3	Muda	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Tidak	Tidak
…	…	…	…	…	…	…	…	…	…	…
30000	Tua	Wanita	Pasif	Tidak	Ya	Tidak	Sering	Tidak	Tidak	Ya

Untuk *data preperation* semua data diubah ke *integer* 0 dan 1 sehingga lebih mudah dikelola. Untuk kolom yang digunakan yaitu 10 dari 11 kolom.

Gambar 2. *Data preperation*

```
[3]  dataset['Usia'] = dataset['Usia'].map({'Tua':1,'Muda':0})
     dataset['Hasil'] = dataset['Hasil'].map({'Ya':1,'Tidak':0})
     dataset['Jenis_Kelamin'] = dataset['Jenis_Kelamin'].map({'Pria':1,'Wanita':0})
     dataset['Merokok'] = dataset['Merokok'].map({'Aktif':1,'Pasif':0})
     dataset['Bekerja'] = dataset['Bekerja'].map({'Ya':1,'Tidak':0})
     dataset['Rumah_Tangga'] = dataset['Rumah_Tangga'].map({'Ya':1,'Tidak':0})
     dataset['Aktivitas_Begadang'] = dataset['Aktivitas_Begadang'].map({'Ya':1,'Tidak':0})
     dataset['Aktivitas_Olahraga'] = dataset['Aktivitas_Olahraga'].map({'Sering':1,'Jarang':0})
     dataset['Asuransi'] = dataset['Asuransi'].map({'Ada':1,'Tidak':0})
     dataset['Penyakit_Bawaan'] = dataset['Penyakit_Bawaan'].map({'Ada':1,'Tidak':0})
```

Model yang akan digunakan adalah *Naïve Bayes* dan *Random Forest* dimana model tersebut tersedia pada *library Google Colab*. Bahasa yang digunakan *Google Colab* adalah *python*. Berikut adalah *script* model *Naïve Bayes* dan *Random Forest:*

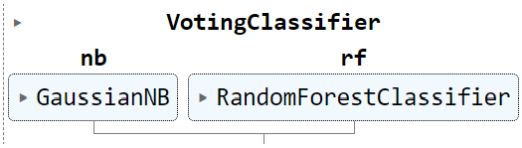
Gambar 3. Memanggil model *Naive Bayes* dan *Random Forest* di *Google Colab*

```
# Membuat model Naive Bayes
nb_model = GaussianNB()

# Membuat model Random Forest
rf_model = RandomForestClassifier
```

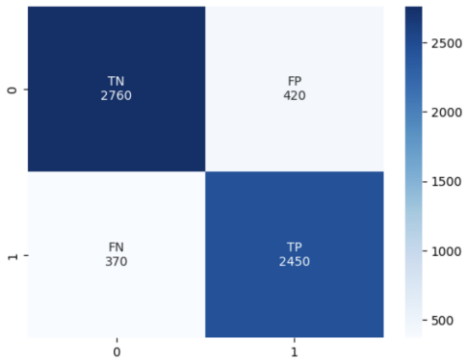
Ensebmle VotingClassifier digunakan untuk meningkatkan performa (akurasi, presisi dan recall) dari model yang dijalankan, sehingga hasilnya dapat lebih baik dibandingkan tanpa menggunakan *Ensemble VotingClassifier*. Model *Naïve Bayes* dan *Random Forest* yang telah diintegrasikan *VotingClassifier* dapat dilihat pada Gambar 4 berikut:

Gambar 4. Model *Naïve Bayes* dan *Random Forest* yang diintegrasikan dengan *VotingCLassifier*



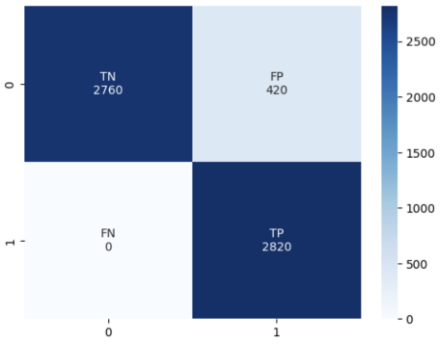
Sebelum menampilkan hasil performa dari model tersebut akan ditampilkan terlebih dahulu *confusion matrix* dari *Naïve Bayes* saja dan *Naïve Bayes* yang dikombinasikan dengan *Random Forest* yang telah diintegrasikan dengan *VotingClassifier*.

Gambar 5. *Confusion Matrix* model *Naïve Bayes*



Pada Gambar 5 menampilkan *Confusion Matrix* model *Naïve Bayes* dimana *True Positive* memiliki nilai 2450, *True Negative* memiliki nilai 2760, *False Positive* memiliki nilai 420 dan *False Negative* memiliki nilai 370.

Gambar 6. *Confusion Matrix* model *Naïve Bayes* dan *Random Forest* diintegrasikan *VotingClassifier*



Gambar 6 adalah *confusion matrix* model *Naïve Bayes* dan *Random Forest* yang diintegrasikan dengan *Ensemble VotingClassifier* dimana untuk *True Positive* 2820, *True Negative* 2760, *False Positive* 420 dan *False Negative* 0.

Tabel 2.
Akurasi, presisi dan Recall model *Naïve Bayes*

	Presentase (%)
Akurasi	86.83
Presisi	85.36
Recall	86.87

Pada Tabel 2 menampilkan performa dari model *Naïve Bayes* dimana untuk akurasinya memiliki nilai 86.83%, presisi 85.36% dan recall 86.87%.

Tabel 3.
Akurasi, presisi dan Recall model *Naïve Bayes* dikombinasikan dengan *Random Forest* yang terintegrasi *VotingClassifier*.

	Presentase (%)
Akurasi	93
Presisi	87.03
Recall	100

Tabel 3 menampilkan performa dari model *Naïve Bayes* dikombinasikan dengan *Random Forest yang terintegrasi VotingClassifier* dimana untuk akurasinya memiliki nilai 93%, presisi 87.03% dan recall 100%.

Hasil dari penelitian yang dilakukan sangat signifikan dimana dari setiap performa baik dari akurasi, presisi dan recall *Naïve Bayes* dikombinasikan dengan *Random Forest* yang terintegrasi *VotingClassifier* menghasilkan nilai performa lebih baik daripada hanya menggunakan model *Naïve Bayes* saja.

V. KESIMPULAN

Kesimpulan pada penelitian Model *Ensemble Algoritma Naive Bayes* Dan *Random Forest* Dalam Klasifikasi Penyakit Paru-paru Untuk Meningkatkan Akurasi yang telah dilakukan dengan jumlah dataset 30000 data menghasilkan performa yang sangat baik. Baik dari akurasi, presisi dan recall model *Naïve Bayes* dikombinasikan dengan *Random Forest* yang terintegrasi *VotingClassifier* unggul dibandingkan hanya menggunakan model *Naïve Bayes saja.* Performa model *Naïve Bayes* dikombinasikan dengan *Random Forest* yang terintegrasi *VotingClassifier* untuk akurasi 93%, presisi 87.03% dan recall 100%.

Saran untuk penelitian kedepannya, dapat dibuat aplikasi untuk mendeteksi Penyakit Paru-paru atau menggunakan algoritma *machine learning* atau pendekatan lainnya untuk mengetahui perbandingan performanya.

DAFTAR PUSTAKA

.

[1] D. Andita Kusuma, J. Teknik Informatika, I. Darmajaya, J. A. Pagar Alam, and B. Lampung, "Rancang Bangun Sistem Pakar Pendiagnosa Penyakit Paru-Paru Menggunakan Metode Case Based Reasoning," 2014.

[2] K. Vembandasamy, R. Sasipriya, and E. Deepa, "Heart Diseases Detection Using Naive Bayes Algorithm," *International Journal of Innovative Science, Engineering & Technology*, vol. 2, no. 9, pp. 441–444, 2015.

[3] W. Gata *et al.*, "Algorithm Implementations Naïve Bayes, Random Forest. C4.5 on Online Gaming for Learning Achievement Predictions," vol. 258, no. Icream 2018, 2019, doi: 10.2991/icream-18.2019.1.

[4] U. K. Kumar, M. B. S. Nikhil, and K. Sumangali, "Prediction of Breast Cancer using Voting Classifier Technique," 2014.

[5] A. N. Repaka, S. D. Ravikanti, and R. G. Franklin, "Design And Implementing Heart Disease Prediction Using Naives Bayesian," *Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019)*, 2019.

[6] M. Ardiansyah, W. Hidayat, E. Utami, and S. Raharjo, "CPU and eGPU Support System Based on Naive Bayes Classification," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 15, no. 2, p. 165, Apr. 2021, doi: 10.22146/ijccs.63689.

[7] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," *Expert Systems with Applications*, vol. 134. Elsevier Ltd, pp. 93–101, Nov. 15, 2019. doi: 10.1016/j.eswa.2019.05.028.

[8] A. Mahabub, "A robust technique of fake news detection using Ensemble Voting Classifier and comparison with other classifiers," *SN Appl Sci*, vol. 2, no. 4, Apr. 2020, doi: 10.1007/s42452-020-2326-y.

[9] S. Syamsiah, "PERANCANGAN FLOWCHART DAN PSEUDOCODE PEMBELAJARAN MENGENAL ANGKA DENGAN ANIMASI UNTUK ANAK PAUD RAMBUTAN," 2019.

[10] Sahar, "Analisis Perbandingan Metode K-Nearest Neighbor dan Naïve Bayes Classiffier pada Data Set Penyakit Jantung," *Indonesian Journal of Data and Science (IJODAS)*, vol. 1, no. 3, pp. 79–86, 2020.

[11] C. Accinelli, S. Minisi, and B. Catania, "Coverage-based Rewriting for Data Preparation," 2020.

[12] S. Anisah, A. S. Honggowibowo, and A. Pujiastuti, "Klasifikasi Teks Menggunakan Chi Square Feature Selection Untuk Menentukan Komik Berdasarkan Periode, Materi Dan Fisikdengan Algoritma Naivebayes," *Compiler*, vol. 5, no. 2, pp. 59–66, 2016, doi: 10.28989/compiler.v5i2.171.

[13] I. Ahmad, M. Basheri, M. J. Iqbal, and A. Rahim, "Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection," *IEEE Access*, vol. 6, pp. 33789–33795, 2018, doi: 10.1109/ACCESS.2018.2841987.

[14] I. Düntsch and G. Gediga, "Indices for rough set approximation and the application to confusion matrices," *International Journal of Approximate Reasoning*, vol. 118, pp. 155–172, 2020, doi: 10.1016/j.ijar.2019.12.008.