**Analisis Prediksi Penyakit Paru-Paru Menggunakan Algoritma Naive Bayes**

|  |  |  |
| --- | --- | --- |
| *Muhammad Shaadam Haidar Yuwono*  *241552010005*  *Department of Informatics Engineering*  *STMIK Tazkia*  *Bogor. Indonesia*  *241552010005.shaadam@student.stmik.tazkia.ac.id* |  | *Kresna risqi Ramadhan*  *241552010006*  *Department of Informatics Engineering*  *STMIK Tazkia*  *Bogor. Indonesia*  *241552010006.kresna@student.stmik.*  *tazkia.ac.id* |

## Abstrak

Penyakit paru-paru merupakan salah satu masalah kesehatan utama yang berdampak signifikan terhadap kualitas hidup masyarakat. Deteksi dini penyakit paru sangat penting untuk meningkatkan peluang keberhasilan pengobatan. Penelitian ini bertujuan mengembangkan sistem prediksi penyakit paru-paru berbasis machine learning dengan algoritma Naive Bayes. Dataset yang digunakan berisi data gejala pasien yang bersifat kategorikal, kemudian dilakukan proses preprocessing berupa penghapusan atribut yang tidak relevan, serta transformasi data menggunakan Ordinal Encoding. Data selanjutnya dibagi menjadi data latih (80%) dan data uji (20%). Model yang digunakan adalah Categorical Naive Bayes (CategoricalNB), yang sesuai untuk data kategorikal. Evaluasi kinerja model dilakukan dengan menghitung tingkat akurasi, classification report, dan analisis confusion matrix. Hasil penelitian menunjukkan bahwa model Naive Bayes mampu memprediksi kondisi penyakit paru-paru dengan akurasi yang cukup tinggi, serta memberikan distribusi prediksi yang seimbang antara kelas positif dan negatif. Kesimpulan dari penelitian ini adalah bahwa algoritma Naive Bayes dapat dijadikan pendekatan yang efektif dan efisien dalam membangun sistem pendukung keputusan untuk deteksi dini penyakit paru-paru, khususnya ketika data yang tersedia berbentuk kategorikal.

**Kata kunci**: Prediksi penyakit paru-paru, Naive Bayes, machine learning, klasifikasi, data kategorikal

**Abstract**

Lung disease is one of the major health issues that significantly impacts the quality of life of the community. Early detection of lung disease is crucial to increase the chances of successful treatment. This study aims to develop a machine learning-based lung disease prediction system using the Naive Bayes algorithm. The dataset used contains categorical patient symptom data, which underwent preprocessing in the form of removing irrelevant attributes and data transformation using Ordinal Encoding. The data was then divided into training data (80%) and test data (20%). The model used is Categorical Naive Bayes (CategoricalNB), which is suitable for categorical data. Model performance evaluation was carried out by calculating the accuracy rate, classification report, and confusion matrix analysis. The results showed that the Naive Bayes model was able to predict lung disease conditions with fairly high accuracy and provided a balanced prediction distribution between positive and negative classes. The conclusion of this study is that the Naive Bayes algorithm can be used as an effective and efficient approach in developing a decision support system for early detection of lung disease, especially when the available data is categorical.

**Keywords**: Lung disease prediction, Naive Bayes, machine learning, classification, categorical data

**1.Pendahuluan**

Penyakit paru-paru merupakan salah satu penyakit tidak menular yang berkontribusi besar terhadap angka kesakitan dan kematian di seluruh dunia. Organ paru-paru memiliki peran vital sebagai organ pernapasan yang bertugas menyalurkan oksigen ke dalam tubuh sekaligus mengeluarkan karbon dioksida. Gangguan pada organ ini dapat mengakibatkan berbagai masalah kesehatan yang serius, mulai dari kesulitan bernapas, penurunan kualitas hidup, hingga kematian. Beberapa penyakit paru-paru yang umum terjadi antara lain tuberkulosis (TBC), pneumonia, bronkitis kronis, asma, dan kanker paru-paru.

Di Indonesia, beban penyakit paru-paru masih tergolong tinggi. Data dari Kementerian Kesehatan menunjukkan bahwa tuberkulosis masih menjadi salah satu penyebab kematian utama, sementara penyakit paru obstruktif kronis (PPOK) juga terus meningkat seiring dengan tingginya prevalensi merokok. Selain itu, polusi udara, gaya hidup tidak sehat, serta keterbatasan akses pelayanan kesehatan turut memperparah kondisi ini. Fakta tersebut menunjukkan bahwa deteksi dini penyakit paru-paru sangatlah penting, karena semakin cepat suatu penyakit dikenali, maka semakin besar pula peluang keberhasilan pengobatan serta pencegahan komplikasi.

Seiring perkembangan teknologi, bidang kesehatan kini semakin banyak memanfaatkan pendekatan berbasis data dan kecerdasan buatan (Artificial Intelligence/AI). Salah satu cabang dari AI yang banyak digunakan adalah pembelajaran mesin (Machine Learning). Metode ini memungkinkan komputer untuk mempelajari pola dari sejumlah besar data dan membuat prediksi terhadap data baru. Dalam konteks penyakit paru-paru, machine learning dapat dimanfaatkan untuk mengolah data medis—baik berupa data klinis, rekam medis elektronik, maupun citra radiologi—untuk mengenali tanda-tanda adanya penyakit secara lebih cepat dan akurat.

Pemanfaatan machine learning dalam prediksi penyakit paru-paru diharapkan mampu menjadi solusi untuk mengatasi keterbatasan diagnosis konvensional yang memerlukan waktu lama dan bergantung pada keahlian dokter spesialis. Dengan adanya model prediksi berbasis komputer, proses analisis dapat dilakukan lebih efisien, sehingga membantu tenaga medis dalam pengambilan keputusan. Selain itu, sistem prediksi juga berpotensi digunakan sebagai alat bantu skrining awal bagi masyarakat, khususnya di daerah yang memiliki keterbatasan akses terhadap fasilitas kesehatan.

Dalam penelitian atau proyek ini, dibuat sebuah sistem prediksi penyakit paru-paru menggunakan pendekatan machine learning yang diimplementasikan dengan bahasa pemrograman Python. Sistem ini dirancang untuk menganalisis data masukan tertentu dan menghasilkan prediksi mengenai kemungkinan adanya penyakit paru-paru pada seseorang. Kehadiran sistem ini diharapkan tidak hanya membantu meningkatkan efektivitas deteksi dini, tetapi juga dapat mendukung upaya pemerintah dan tenaga kesehatan dalam menekan angka penderita penyakit paru-paru di Indonesia.

Untuk memodelkan hubungan antara fitur prediktor dengan status penyakit paru-paru, digunakan algoritma Categorical Naive Bayes yang sesuai untuk data kategorikal. Secara matematis, persamaan model dapat dinyatakan sebagai:

F(x)=x1​w1​+x2​w2​+⋯+xn​wn​+x0​w0​

**2. Metodologi penelitian**

## 2.2 Desain Penelitian

Penelitian ini menggunakan pendekatan **kuantitatif** dengan metode eksperimen komputasional berbasis **supervised learning**. Fokus penelitian adalah membangun model prediksi penyakit paru-paru menggunakan algoritma machine learning untuk mengidentifikasi kondisi kesehatan berdasarkan data pasien. Setiap tahapan penelitian, mulai dari pengumpulan data, preprocessing, pemodelan, hingga evaluasi, dilakukan secara sistematis agar diperoleh model dengan akurasi tinggi dan dapat diinterpretasikan.

**2.3 Pengumpulan Data**

Data yang digunakan dalam penelitian ini bersumber dari dataset kesehatan terkait penyakit paru-paru. Dataset memuat informasi pasien dengan atribut-atribut tertentu yang merepresentasikan kondisi fisik maupun kebiasaan yang berhubungan dengan kesehatan paru-paru.

Berikut adalah contoh fitur yang tersedia dalam dataset:

|  |  |
| --- | --- |
| **Nama Kolom** | **Deskripsi** |
| Name | Identitas responden (tidak digunakan sebagai fitur) |
| Age | Usia pasien |
| Gender | Jenis kelamin pasien |
| Smoking | Kebiasaan merokok (ya/tidak) |
| Coughing | Frekuensi batuk |
| Shortness\_of\_breath | Tingkat sesak napas |
| Chest\_pain | Gejala nyeri dada |
| Fatigue | Tingkat kelelahan |
| Family\_history | Riwayat penyakit paru-paru dalam keluarga |
| Disorder (target) | Status penyakit paru-paru (1 = sakit, 0 = normal) |

## 2.4 Preprocessing Data

Tahap preprocessing data bertujuan untuk menyiapkan dataset agar sesuai dengan kebutuhan algoritma machine learning. Tahapan yang dilakukan antara lain:

* **Pengecekan Missing Values**  
  Setiap kolom dicek menggunakan fungsi isnull() dan sum() dari library pandas. Hasil pemeriksaan menunjukkan jumlah nilai kosong pada setiap kolom. Jika ada missing value, maka dilakukan imputasi atau penghapusan data.
* **Encoding Variabel Kategorikal**  
  Variabel dengan tipe kategorikal, seperti Gender dan Smoking, diubah ke dalam bentuk numerik menggunakan LabelEncoder dari sklearn.preprocessing. Hal ini penting karena algoritma machine learning hanya dapat memproses data numerik.

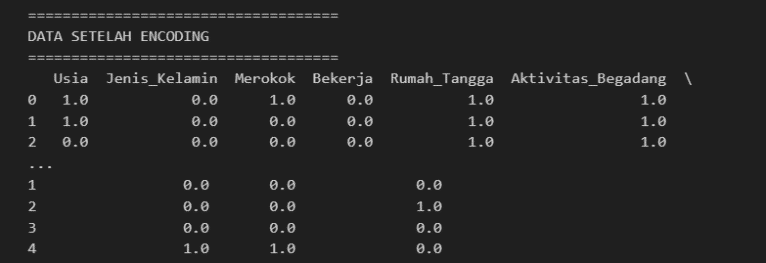


Image 2: encode kategorikal

* **Pemilihan Fitur dan Target**  
  Variabel Name dieliminasi karena tidak relevan terhadap prediksi. Variabel Disorder digunakan sebagai target (y), sedangkan kolom lainnya dijadikan fitur prediktor (X).

## 2.5 Pembagian Data

Dataset dibagi menjadi **training set** dan **testing set** dengan rasio **80:20**menggunakan fungsi *train\_test\_split()* dari *sklearn.model\_selection*. Tujuan utama dari pemisahan ini adalah memastikan bahwa model diuji pada data yang **belum pernah dilihat sebelumnya**, sehingga metrik evaluasi mencerminkan kemampuan generalisasi model. Pemilihan kedua algoritma ini didasarkan pada tujuan untuk membandingkan efektivitas model linear sederhana dengan model ansambel non-linear yang lebih kompleks.

**2.6 Categorical Naive Bayes**

Algoritma utama yang digunakan dalam penelitian ini adalah **Categorical Naive Bayes (CategoricalNB)**, yang sesuai untuk data dengan fitur kategorikal. Naive Bayes merupakan algoritma klasifikasi probabilistik berbasis Teorema Bayes dengan asumsi bahwa setiap fitur saling independen.

Proses klasifikasi dilakukan dengan menghitung probabilitas posterior untuk setiap kelas berdasarkan distribusi data pada masing-masing fitur. Kelebihan algoritma ini adalah:

* Sederhana dan efisien untuk data kategorikal.
* Cepat dilatih meskipun dataset cukup besar.
* Memberikan hasil yang kompetitif pada kasus klasifikasi sederhana.

Dengan karakteristik tersebut, Naive Bayes dipilih sebagai pendekatan yang efektif untuk membangun model prediksi penyakit paru-paru dalam penelitian ini.

#### **2.7 Evaluasi Model (baru)**

Evaluasi kinerja model dilakukan menggunakan beberapa metrik, antara lain:

* **Akurasi (Accuracy):** Mengukur seberapa banyak prediksi benar dibandingkan dengan total data.
* **Presisi (Precision):** Mengukur seberapa banyak prediksi positif yang benar dari semua prediksi positif.
* **Recall (Sensitivitas):** Mengukur seberapa banyak kasus positif yang berhasil dikenali dari seluruh kasus positif aktual.
* **F1-Score:** Rata-rata harmonis dari presisi dan recall untuk memberikan gambaran yang seimbang.
* **Confusion Matrix:** Memberikan ringkasan detail prediksi model, termasuk jumlah prediksi benar (true positive dan true negative) serta kesalahan (false positive dan false negative).

**3.Hasil dan pembahasan**

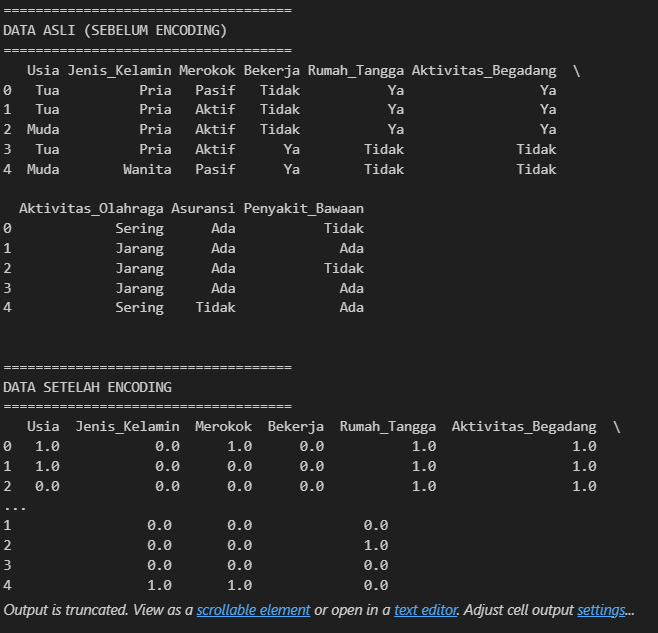
**3.1 Deskripsi Dataset dan Pra-pemrosesan Data**

Dataset yang digunakan untuk memprediksi penyakit paru-paru berisi **30.000 baris** dan **11 kolom**. Fitur-fitur yang tersedia adalah 'Usia', 'Jenis\_Kelamin', 'Merokok', 'Bekerja', 'Rumah\_Tangga', 'Aktivitas\_Begadang', 'Aktivitas\_Olahraga', 'Asuransi', dan 'Penyakit\_Bawaan', sedangkan kolom targetnya adalah 'Hasil'.

Karena semua fitur dan label target berupa data kategorikal (seperti 'Tua', 'Muda', 'Ya', 'Tidak'), data ini harus diubah menjadi format numerik agar bisa diproses oleh model *machine learning*. Teknik yang digunakan adalah **Ordinal Encoding**, yang mengonversi setiap kategori unik dalam suatu kolom menjadi nilai bilangan bulat. Misalnya, 'Tua' dan 'Muda' pada kolom 'Usia' diubah menjadi 1.0 dan 0.0, sedangkan 'Pria' dan 'Wanita' pada kolom 'Jenis\_Kelamin' diubah menjadi 0.0 dan 1.0.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| No | usia | Jenis kelamin | merokok | bekerja | Rumah tangga | begadang | olahraga | asuransi | Penyakit bawaan |
| 1 | tua | pria | pasif | tidak | ya | ya | sering | ada | ada |
| 2 | tua | pria | aktif | tidak | ya | ya | jarang | ada | tidak |
| 3 | muda | pria | aktif | tidak | ya | ya | jarang | ada | ada |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| No | usia | Jenis kelamin | merokok | bekerja | Rumah tangga | begadang | olahraga | asuransi | Penyakit bawaan |
| 1 | 1.0 | 0.0 | 1.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| 2 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 1.0 |
| 3 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 |



### 2. Pembagian Data dan Pelatihan Model

Setelah data di-*encode*, dataset dibagi menjadi **data latih (training data)** dan **data uji (testing data)** dengan perbandingan **80:20**. Ini berarti **24.000 baris** data digunakan untuk melatih model, sedangkan **6.000 baris** sisanya digunakan untuk menguji seberapa baik model dapat membuat prediksi pada data yang belum pernah dilihat sebelumnya.

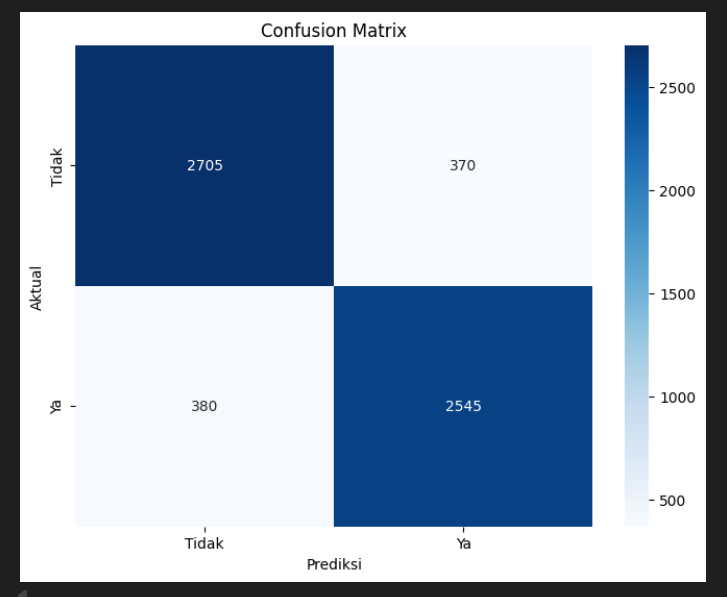
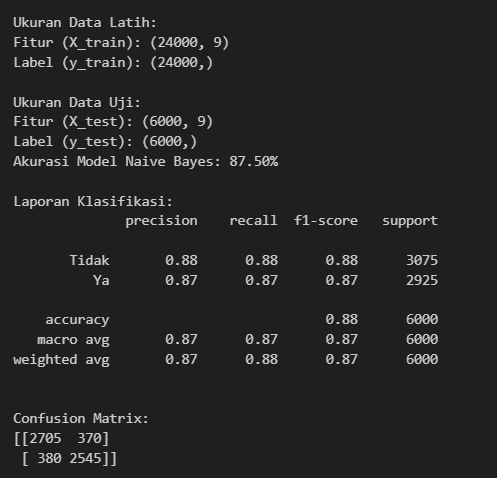
Model yang dipilih dalam klasifikasi ini adalah **Categorical Naive Bayes**. Model ini sangat cocok untuk data yang fitur-fiturnya bersifat kategorikal, seperti yang ada di dataset ini. Kelebihan utama dari model ini adalah kesederhanaan dan kecepatan komputasinya, menjadikannya pilihan yang efisien untuk *baseline* model.

### 3. Evaluasi Kinerja Model

Kinerja model dievaluasi menggunakan beberapa metrik, yaitu **akurasi**, **laporan klasifikasi**, dan **confusion matrix**.

* **Naive Bayes Classifier (NBC)** adalah algoritma klasifikasi probabilistik yang didasarkan pada **Teorema Bayes** dengan asumsi "naif" (sederhana) bahwa **semua fitur dalam dataset independen satu sama lain**. Meskipun asumsi ini jarang sepenuhnya benar di dunia nyata, Naive Bayes sering kali menunjukkan performa yang sangat baik dan efisien, terutama untuk data kategorikal.
* **Akurasi**: Model Naive Bayes mencapai akurasi sebesar **87,50%**. Angka ini menunjukkan bahwa model mampu memprediksi dengan benar apakah seseorang berpotensi memiliki penyakit paru-paru atau tidak sebanyak **87,5%** dari total 6.000 data uji.
* **Laporan Klasifikasi**:
  + **Presisi (Precision)**: Untuk kelas 'Tidak' (tidak memiliki penyakit), presisi adalah **0.88**, sedangkan untuk kelas 'Ya' (memiliki penyakit), presisi adalah **0.87**. Presisi mengukur seberapa banyak prediksi positif yang benar dari total prediksi positif yang dibuat oleh model.
  + **Recall**: Untuk kelas 'Tidak', recall adalah **0.88**, dan untuk kelas 'Ya', recall adalah **0.87**. Recall mengukur seberapa banyak kasus aktual positif yang berhasil diidentifikasi dengan benar oleh model.
  + **F1-Score**: F1-score untuk kedua kelas ('Ya' dan 'Tidak') adalah **0.87** dan **0.88**, menunjukkan keseimbangan yang baik antara presisi dan recall.
* **Confusion Matrix** adalah tabel yang merangkum kinerja model klasifikasi. Matriks ini sangat penting karena tidak hanya memberikan satu angka akurasi, tetapi juga menunjukkan di mana model melakukan kesalahan dan jenis kesalahan apa yang dibuatnya.
* **Confusion Matrix**: Matriks ini memberikan gambaran detail tentang hasil prediksi model.
  + **True Negatives (TN)**: Model dengan benar memprediksi **2.705 orang** tidak memiliki penyakit paru-paru.
  + **True Positives (TP)**: Model dengan benar memprediksi **2.545 orang** memiliki penyakit paru-paru.
  + **False Negatives (FN)**: Model salah memprediksi **380 orang** yang sebenarnya memiliki penyakit, sebagai orang yang tidak memiliki penyakit (tipe II error). Dalam konteks medis, ini adalah hasil yang paling kritis karena pasien yang sakit tidak terdeteksi dan tidak mendapatkan penanganan.
  + **False Positives (FP)**: Model salah memprediksi **370 orang** yang sebenarnya tidak memiliki penyakit, sebagai orang yang memiliki penyakit (tipe I error). Ini dapat menyebabkan kecemasan dan biaya tambahan untuk pemeriksaan medis yang tidak perlu.

Hasil ini menunjukkan bahwa model Naive Bayes memiliki kinerja yang solid dalam memprediksi penyakit paru-paru pada dataset ini, dengan tingkat akurasi yang tinggi dan kesalahan prediksi yang relatif rendah. Dalam kasus ini, **recall** untuk kelas 'Ya' (0.87) menjadi metrik yang paling penting, karena kita ingin memastikan sebanyak mungkin kasus penyakit terdeteksi dengan benar.



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No |  | Precision | Recall | F1-score | support |
| 1 | tidak | 0.88 | 0.88 | 0.88 | 3075 |
| 2 | ya | 0.87 | 0.87 | 0.87 | 2925 |
| 3 | accuracy |  |  | 0.88 | 6000 |
| 4 | Macro avg | 0.87 | 0.87 | 0.87 | 6000 |
| 5 | Weighted avg | 0.87 | 0.88 | 0.87 | 6000 |

**4.Kesimpulan**

Penelitian ini membangun sistem prediksi penyakit paru-paru menggunakan algoritma **Naive Bayes** dengan data gejala pasien yang bersifat kategorikal. Proses penelitian meliputi pengumpulan data, preprocessing (encoding data kategorikal, eliminasi atribut tidak relevan, dan pembagian data latih serta data uji), serta evaluasi model menggunakan akurasi, classification report, dan confusion matrix.

Hasil pengujian menunjukkan bahwa model **Categorical Naive Bayes** mampu mencapai akurasi sebesar **87,5%**, dengan nilai precision, recall, dan F1-score yang seimbang pada kedua kelas (memiliki penyakit dan tidak memiliki penyakit). Model ini terbukti cukup andal dalam mendeteksi kondisi paru-paru, meskipun masih terdapat kesalahan prediksi berupa false negative dan false positive.

Secara keseluruhan, penelitian ini menyimpulkan bahwa **Naive Bayes dapat dijadikan metode yang efektif, efisien, dan sederhana untuk mendukung deteksi dini penyakit paru-paru**, terutama ketika data yang tersedia berbentuk kategorikal. Model ini dapat membantu tenaga medis dalam pengambilan keputusan, sekaligus menjadi alat skrining awal bagi masyarakat untuk meningkatkan upaya pencegahan serta pengobatan lebih lanjut.

# **Ucapan Terima Kasih**

Penulis menyampaikan apresiasi yang sebesar-besarnya kepada dosen pengampu, Bapak Hendri **Karisma, S.Kom., M.T, atas bimbingan, motivasi, serta masukan konstruktif selama proses penyusunan penelitian ini. Ucapan terima kasih juga disampaikan kepada pihak yang telah membuka akses dataset secara publik melalui platform Kaggle, sehingga memungkinkan penelitian ini dapat terlaksana.**

**Daftar Pustaka**

1. Aulia, A. P., Adelia, Q., Mubarak, H. A., Fatan, M. A., & Sudarno. (2025). Lung Disease Risk Prediction Using Machine Learning Algorithms. *Public Research Journal of Engineering, Data Technology and Computer Science*, 3(1).
2. Hafiyyan, A., Seniwati, E., & Haryoko. (2023). Implementation of The Naïve Bayes Algorithm For Diagnosing Lung Disesase In An Expert System Concept.

*International Journal of Information System & Technology*, 7(1).

1. Ulumuddin, Rousyati, & Nurzuli, R. (2025). Lung Diseases Classification Using the Naïve Bayes Algorithm. *International Journal of Informatics and Computation (IJICOM)*, 7(2).
2. Fadlianda, D., Prananto, A., Eriska, C. A., Anjanira, S., Syadzwina, N., & Ula, M. (2025). DIAGNOSIS PENYAKIT JANTUNG MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE (SVM). *SENASTIKA* *Universitas Malikussaleh*, 744.
3. Haffandi, M. Y., Haerani, E., Syafria, F., & Oktavia, L. (2022). Klasifikasi Penyakit Paru-Paru dengan Menggunakan Metode Naïve Bayes Classifier. *Jurnal TEKINKOM*, 5(2), .