DECODE: JURNAL PENDIDIKAN TEKNOLOGI INFORMASI
P-ISSN: 2775-2984 | E-ISSN: 2775-1813

Vol. 4, No 3, November 2024, Hal: 1057-1070
Doi: http://dx.doi.org/10.51454/decode.v4i3.779
https://journal.umkendari.ac.id/index.php/decode

Prediksi Kanker Paru dengan Normalisasi menggunakan Perbandingan Algoritma *Random Forest, Decision Tree* dan *Naïve Bayes*

Banafshah Shafa[1*], Hanny Hikmayanti Handayani[1], Santi Arum Puspita Lestari[1], Yana Cahyana[1]

[1]Program Studi Teknik Informatika, Universitas Buana Perjuangan Karawang, Indonesia.

Artikel Info

Kata Kunci:

Confusion Matrix;
Decision Tree;
Naïve Bayes;
Penyakit Kanker Paru;
Random Forest.

Keywords:

Confusion Matrix;
Decision Tree;
Lung Cancer Disease;
Naïve Bayes;
Random Forest

Riwayat Artikel:

Submitted: 14 September 2024
Accepted: 7 November 2024
Published: 26 November 2024

Abstrak: Berdasarkan data Global Cancer Observatory Organisasi Kesehatan Dunia angka kematian kanker paru sebanyak 1.796.144 orang di seluruh dunia. Kematian akibat kanker paru di Indonesia sebanyak 30.843 pada tahun 2020. Penyakit yang dapat membunuh orang akibat keganasannya yang paling umum disebabkan oleh kanker paru mencapai 13% dari keseluruhan diagnosis kanker. Penyakit ini dapat disebabkan dari internal ataupun eksternal paru-paru. Membuat model prediksi dirasa perlu, guna mendeteksi penyakit ini lebih awal untuk menekan angka kematian yang diakibatkan oleh kanker paru. Menggunakan proses pemodelan menggunakan algoritma *Random Forest*, *Naïve Bayes* dan *Decision Tree* untuk memproses data tersebut. Tujuan penelitian melakukan perbandingan algoritma *Random Forest*, *Decision Tree* serta *Naïve Bayes* untuk memprediksi penyakit kanker paru dengan menggunakan data yang terdiri dari 26. 000 data. Data ini meliputi informasi tentang pasien, gaya hidup, dan kondisi medis, seperti umur, jenis kelamin, polusi udara, konsumsi alkohol, alergi debu, risiko genetik, penyakit paru kronis, diet seimbang, obesitas, kebiasaan merokok, dan riwayat penyakit lain. Tahapan pengolahan data terdiri dari, pembersihan Data, yaitu menghilangkan fitur yang tidak relevan, seperti Index dan Patient ID, dan mengubah fitur kategorikal "*Level*" menjadi bentuk numerik, lalu analisis Korelasi,yaitu Mengidentifikasi atribut yang memiliki korelasi tinggi, seperti "*Alcohol Use*", "*Occupational Hazards*", "*Genetic Risk*", dan "*Chronic Lung Disease*", selanjutya normalisasi data mengubah sebaran data dari empat atribut yang memiliki korelasi tinggi agar lebih mudah diproses, kemudian seleksi fitur yaitu memilih fitur penting dengan menggunakan metode chi-square, yang menunjukkan bahwa "*Coughing of Blood*", "*Passive Smoker*", dan "*Obesity*" memiliki score tertinggi dan dianggap sebagai fitur penting, dilanjutkan dengan pemisahan Data, yaitu membagi data menjadi 80% untuk data pelatihan dan 20% untuk data pengujian, selanjutnya pembuatan model dengan menggunakan tiga algoritma, yaitu *Random Forest, Decision Tree,* dan *Naïve Bayes*, untuk memprediksi kanker paru. *Random Forest* dan *Decision Tree* mencapai akurasi 100%, sementara *Naïve Bayes* mencapai akurasi 86%. Berdasarkan evaluasi penelitian yang telah dilakukan pada data penyakit kanker paru, algoritma *Random Forest* dan *Decision Tree* sangat cocok untuk prediksi data penyakit kanker paru karena mampu menghasilkan model prediksi yang baik dengan pengujian *Confusion Matrix* serta *Learning Curve*.

Abstract: *Based on data from the World Health Organization's Global Cancer Observatory, the number of lung cancer deaths is 1,796,144 people worldwide. Deaths from lung cancer in Indonesia amounted to 30,843 in 2020. Diseases that can kill people due to their malignancy are most commonly caused by lung cancer reaching 13% of all cancer diagnoses. This*

disease can be caused from internal or external lungs. Creating a prediction model is necessary, in order to detect this disease early to reduce the mortality rate caused by lung cancer. Using the modeling process using Random Forest, Naïve Bayes and Decision Tree algorithms to process the data. The purpose of the research is to compare the Random Forest, Decision Tree and Naïve Bayes algorithms to predict lung cancer using data consisting of 26,000 data. This data includes information about the patient, lifestyle, and medical conditions, such as age, gender, air pollution, alcohol consumption, dust allergies, genetic risk, chronic lung disease, balanced diet, obesity, smoking habits, and history of other diseases. Data processing stages consist of, Data cleaning, which removes irrelevant features, such as Index and Patient ID, and converts categorical features "Level" into numerical form, then Correlation analysis, which identifies attributes that have a high correlation, such as "Alcohol Use", "Occupational Hazards", "Genetic Risk", and "Chronic Lung Disease", then data normalization by changing the data distribution of the four attributes that have a high correlation to make it easier to process, then feature selection, namely selecting important features using the chi-square method, which shows that "Coughing of Blood", "Passive Smoker", and "Obesity" have the highest score and are considered important features, followed by data separation, which divides the data into 80% for training data and 20% for testing data, then model building using three algorithms, namely Random Forest, Decision Tree, and Naïve Bayes, to predict lung cancer. Random Forest and Decision Tree achieved 100% accuracy, while Naïve Bayes achieved 86% accuracy. Based on the evaluation of research that has been conducted on lung cancer data, the Random Forest and Decision Tree algorithms are very suitable for predicting lung cancer data because they are able to produce good prediction models with Conffusion Matrix and Learning Curve testing.

Corresponding Author:
Banafshah Shafa
Email: if20.banafshahshafa@mhs.ubpkarawang.ac.id

PENDAHULUAN

Organ terpenting pada tubuh manusia yang berfungsi untuk tempat pergantian oksigen yaitu paru-paru (Meiyanti & Komarudin, 2020). Kondisi tertentu pada kinerja sistem pernapasan dapat terganggu yang disebabkan oleh paru-paru, jika hal ini terjadi akan mengakibatkan suatu penyakit, salah satunya kanker paru (Zulaikhah Hariyanti Rukmana et al., 2022). Berdasarkan data *Global Cancer Observatory* Organisasi Kesehatan Dunia, kanker paru-paru merupakan bentuk kanker paling mematikan, dengan angka kematian sebanyak 1.796.144 orang di seluruh dunia. 30.843 kematian akibat kanker paru-paru dan 34.783 kasus baru dilaporkan di Indonesia pada tahun 2020 (Sari et al., 2023). Penyakit yang dapat membunuh orang akibat keganasannya yang paling umum disebabkan oleh kanker paru mencapai 13% dari keseluruhan diagnosis kanker. Ditemukan bahwa kanker paru banyak terjadi pada laki-laki, sementara pada perempuan kanker paru termasuk kasus keempat, pernyataan ini berdasarkan penelitian sebelumnya (Septhya et al., 2023). Penyakit ini dapat disebabkan dari internal ataupun eksternal paru- paru (Permana & Djamaluddin, 2023).

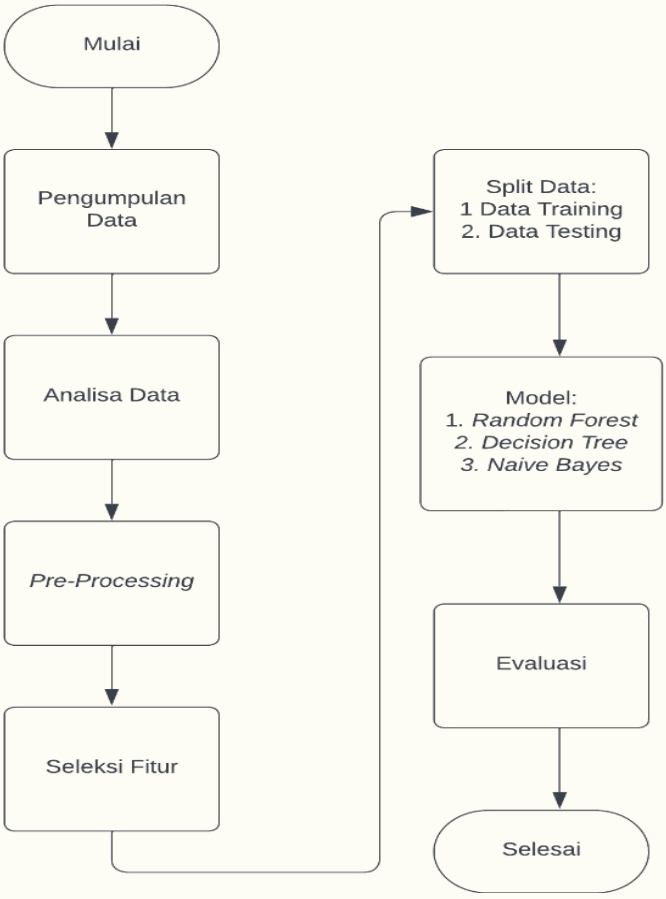
Berdasarkan data serta informasi terdahulu dapat digunakan untuk meramalkan kejadian masa depan yaitu tujuan dari prediksi (Rahman Wahid et al., 2023). Membuat model prediksi dirasa perlu, guna mendeteksi penyakit ini lebih awal untuk menekan angka kematian yang diakibatkan oleh kanker paru (Permana & Djamaluddin, 2023). Model dalam konteks prediksi penyakit dapat membantu dalam mengidentifikasi gejala-gejala penyakit paru telah menyerang, serta dapat membantu dalam hal pengambilan keputusan untuk langkah selanjutnya mengenai penanganan kanker paru (Rahman Wahid et al., 2023). Algoritma *machine learning* dengan menghasilkan banyak pohon keputusan yaitu

Random Forest. Algoritma ini terbukti menjadi salah satu algoritma *machine learning* terbaik untuk menyelesaikan kasus regresi serta klasifikasi dalam berbagai macam bidang (Apriliah et al., 2021) .Algoritma yang diuji dengan contoh data yang dimanfaatkan dalam memperoleh sebuah pohon yang menghasilkan nilai kebenaran yang teruji yaitu Algoritma *Decision TreeI* (Qisthiano et al., 2023). Metode klasifikasi menghitung nilai probabilitas dengan menjumlahkan beberapa nilai serta frekuensi dari data yang ada, dimana semua atribut saling berhubungan yang menghasilkan nilai di variabel, yaitu Algoritma *Naïve Bayes* (Kenang Candra Alivian Pratama et al., 2022). Penelitian kali ini bertujuan untuk melakukan perbandingan algoritma *Random Forest*, *Decision Tree* serta *Naïve Bayes* untuk memprediksi penyakit kanker paru sebagai saran dari penelitian sebelumnya (Meiyanti & Komarudin, 2020), hanya saja penelitian kali ini dilakukan proses normalisasi min-max terlebih dahulu.

Penelitian ini diharapkan dapat membantu dalam mengetahui kinerja dari setiap algoritma yang digunakan dalam memprediksi serta dapat mengetahui langkah selanjutnya yang akan diambil untuk menangani penyakit ini.

METODE

Berikut prosedur yang menggambarkan tahapan-tahapan penelitian ini:



Gambar 1. *Flowchart Penelitian*

Pengumpulan Data

Penelitian ini mengunakan dataset yang diperoleh dari *https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link/data* yang berupa data pasien yang mengidap penyakit kanker paru, data diakses pada tanggal 6 Desember 2023 pada pukul 20.00 WIB.

Analisa Data

Proses analisa data ini berfungsi untuk mengetahui dataset yang telah diperoleh, sebelum melakukan proses penelitian. Tahapan ini berguna untuk memperoleh informasi yang terdapat di dalam dataset, terdiri dari jumlah data, jumlah kolom, jumlah baris, jumlah *missing value*, jumlah data yang bukan numerik, jumlah duplikasi data, serta jumlah presentase data yang akan dihapus.

Pre-Processing

Tahap ini dilakukan untuk mengolah data, yang terdiri dari pembersihan data, proses ini dilakukan agar mengetahui atribut mana saja yang perlu dan tidak, pada dataset kali ini atribut yang akan dihilangkan yaitu *Index* dan *Patient Id*. Tahap selanjutnya merubah bentuk data yang masih non numerik menjadi numerik yang terdapat pada atribut level, 0 untuk *Low*, 1 untuk *Medium*, 2 untuk *High*. Berikutnya menghapus *Missing Value* serta duplikasi data, untuk dataset ini tidak ada *Missing Value* maupun duplikasi data. Selanjutnya menampilkan keseluruhan data setelah pembersihan.

Pemilihan Data, tahap ini dilakukan guna memilih data atau atribut yang akan dilakukan dalam proses penelitian ini. Data yang digunakan merupakan data yang relevan dengan tujuan penelitian, yaitu prediksi penyakit kanker paru. Dataset yang diperoleh tersebut dilakukan normalisasi dengan metode *min – max*, sebelumnya normalisasi ini telah digunakan pada kasus klasifikasi penyakit liver (Zulaikhah Hariyanti Rukmana et al., 2022).

Seleksi Fitur

Seleksi Fitur yang digunakan pada penelitian kali ini menggunakan pengujian *Chi-Square*. Seleksi fitur berfungsi untuk mengurangi jumlah fitur yang akan digunakan, serta mengoptimalkan nilai akurasi dengan pemilihan fitur yang sesuai (Septhya et al., 2023).

Split Data

Split Data, proses ini berguna untuk membagi data menjadi dua, yaitu *data training* dan *testing*. Data yang digunakan untuk pemodelan atau algoritma mengenali pola dalam suatu dataset ialah *Data training* sedangkan data yang belum digunakan pada proses *training* serta bertujuan untuk mengevaluasi kinerja suatu model atau algoritma disebut *Data Testing* . Data *Testing* dan data *Training* tidak boleh sama.

Penelitian kali ini menggunakan persentase 80 : 20, 80 *Data Training* dan 20 *Data Testing*. Persentase *Split Data* ini tidak memiliki aturan baku, pada penelitian sebelumnya oleh Agiel dkk, ketika analisis percobaan selesai dan hasilnya diperoleh, lebih banyak data latih akan menghasilkan akurasi yang lebih tinggi. Hal ini dikarenakan, dalam proses evaluasi yang dilakukan oleh *Confusion Matrix*, nilai *True Positive* dan *True Negative* akan lebih tinggi pada skenario dataset yang lebih besar. Hal ini akan berdampak negatif pada akurasi karena akurasi *True Positive* merupakan hasil dari prediksi positif yang akurat, dan akurasi *True Negative* merupakan hasil dari prediksi negatif yang akurat. Akibatnya, akurasi tertinggi juga dapat ditemukan dalam skenario dataset dengan tingkat kepercayaan tinggi (Fadillah Hermawan et al., 2022).

Model

Model yang digunakan pada penelitian ini ialah *Random Forest*, *Decision Tree* dan *Naïve Bayes*. Pada tahap ini akan dilakukan perbandingan dari ketiga model tersebut yang berfungsi untuk mengetahui model mana yang paling optimal untuk kasus ini. *Random Forest* mengelompokkan data berdasarkan pola-pola yang ada di dalam data (Agtira et al., 2023). *Random Forest* beroperasi dengan menghasilkan banyak pohon keputusan yang selanjutnya hasil keseluruhannya dilihat dari hasil terbanyak guna menentukan hasil prediksi akhir, sekaligus secara tidak langsung menangani permasalahan saat melakukan klasifikasi sering tidak optimal dengan menggunakan satu pohon keputusan (Sari et al., 2023). Klasifikasi dengan Random Forest telah digunakan untuk mendeteksi serangan DDoS di jaringan SDN dengan waktu pengambilan keputusan relative singkat sebesar 0.3 detik (Harto & Basuki, 2021). Penggabungan *Random Forest* dengan *Adaboost* pada kasus prediksi kanker paru memperoleh kinerja terbaik dalam hal akurasi, *presisi, recall* serta spesifikasi (Sinaga et al., 2022).

Penelitian yang berjudul Peningkatan Deteksi Kanker Paru Menggunakan *Random Walker Improved* dengan Jaringan Saraf Tiruan dan Klasifikasi *Random Forest* memperoleh hasil klasifikasi yang disarankan mempunyai potensi besar untuk mempengaruhi beban kanker paru secara global (Nair et al., 2024). Berikut rumus algoritma random forest :

$$f(x) = \sum_{i=1}^{m} w_i h_i(x) \qquad (1)$$

Keterangan: $f(x)$ ialah hasil dari pohon keputusan, m ialah jumlah node pohon keputusan, wi ialah nilai dari setiap node pohon keputusan, serta $hi(x)$ ialah fungsi yang hasilnya nilai 0 atau 1 dengan kondisi apakah x memenuhi ketentuan yang diberikan oleh node.

Decision Tree adalah jenis diagram alir yang terdiri dari node, dimana setiap node internal memiliki atribut dan node anaknya menampilkan hasil pengujian atau nilai atribut. Daun dapat digunakan untuk distribusi kelas atau emulasi kelas. Pohon keputusan biasanya digunakan untuk melakukan analisis statistik (Hidayanti et al., 2022). Langkah dalam *Decision Tree* salah satunya *Pruning* yang memiliki tujuan meminimalkan ukuran pohon keputusan dan menghindari *overfitting*. Pemangkasan dilakukan dengan memotong cabang-cabang pohon yang secara material tidak meningkatkan akurasi model. Secara rekursif menghilangkan cabang yang paling tidak signifikan adalah cara pemangkasan dilakukan hingga tingkat presisi yang dibutuhkan tercapai (Mia et al., 2024). Perbandingan performa algoritma *Decision Tree* dan *Support Vector Machine* dalam memprediksi gagal jantung memperoleh hasil pola yang berbeda dari data ditunjukkan oleh hasil pengelompokan, tetapi dengan variasi metodologis yang kecil dan keselarasan yang sempit dengan variabel objektif, temuan klasifikasi menunjukkan bahwa SVM mengungguli *Decision Tree* dalam hal akurasi, presisi, *Recall*, dan *F1-Score*, meskipun kedua teknik tersebut baik dalam memprediksi penyakit jantung (Arifuddin et al., 2024).

Naïve Bayes merupakan *Machine Learning* yang meringkas gagasan bahwa ada tidaknya sebuah fitur dalam sebuah kelas tidak selalu berkorelasi dengan ada atau tidaknya fitur lain dalam kelas yang sama, *Naïve Bayes* dapat dimanfaatkan dalam berbagai macam kasus klasifikasi, algoritma ini pernah dilakukan untuk penelitian klasifikasi kemungkinan bertahan hidup pada kasus kanker payudara yang menghasilkan nilai akurasi sebesar 86% (Arum & Triyono, 2021) (Alamsyah et al., 2023). *Naïve Bayes* disebut teorema bayes karena dapat memprediksi peluang masa depan berdasarkan dari pengalaman masa lalu (Aldiansyah Poetra et al., 2023). Perbandingan antara *Naïve Bayes* dan KNN dalam penelitian prediksi paru tahun 2023 menghasilkan nilai akurasi *Naïve Bayes* sebesar 98.8% (Naezer & Supriyanto, 2023). *Naïve Bayes* dengan Sistem Inkremental untuk Prediksi Kesehatan dengan waktu *Real-Time* menghasilkan tingkat akurasi tertinggi sebesar 66.6% pada dataset Agrawal, diikuti oleh Dialysis sebesar 61,6% serta akurasi liver sebesar 50% dan 58.3%. Ketepatan dipengaruhi oleh ukuran dataset dan distribusi data (Appasani et al., 2024). Rumus yang diperoleh berdasarkan *Teorema Bayes*, rumusnya sebagai berikut:

$$P(h|d) = \frac{P(d|h).P(h)}{P(d)} \qquad (2)$$

Keterangan: $P(h|d)$ ialah kemungkinan hipotesis h jika diberikan data d, $P(d|h)$ ialah kemungkinan data d jika hipotesis h benar, $P(h)$ ialah kemungkinan hipotesis h sebelum mengetahui data, serta $P(d)$ ialah kemungikinan data d.

Evaluasi

Mengukur evaluasi kinerja, dimana *Accuracy* berasal dari tabel yang terdiri dari evaluasi pemrosesan data untuk menunjukkan kesesuaian dengan akurasi yang terukur, *Precision* adalah

presentasi data diberi label positif yang disediakan oleh klasifikasi, *recall* mengacu pada evaluasi kumpulan data yang secara konsisten diprediksi oleh model sebagai positif, sedangkan *F1 Score* mengukur rata-rata harmonis dari akurasi dan recall merupakan pengertian *Confusion Matrix:*

$$Accuracy = \frac{True\ Positive\ +\ True\ Negative}{SUM\ The\ Number\ of\ Data}\ x\ 100\% \quad (3)$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}\ \text{x}\ 100 \quad (4)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}\ \text{x}\ 100 \quad (5)$$

$$F1\ Score = 2\ x\ \frac{Precision\ x\ Recall}{Precision\ x\ Recall} \quad (6)$$

HASIL DAN PEMBAHASAN

Berikut hasil dan pembahasan yang diperoleh dari pengolahan data prediksi kanker paru:

Pengumpulan Data

Data berbentuk file csv yang memiliki 26.000 data, yang terdiri dari jenis data numerik dan kategorikal. Data numerik ini meliputi *Index*, *Patient Id* (id pasien), *Age* (umur), gender, *Air Pollution* (polusi udara), *Alcohol Use* (penggunaan alkohol), *Dust Allergy* (alergi debu), *Occupational Hazards* (bahaya pekerjaan), *Genetic Risk* (resiko genetik), *Chronic Lung Disease* (penyakit paru kronis), *Balanced Diet* (diet seimbang), *Obesity* (obesitas), *Smoking* (merokok), *Passive Smoker* (perokok pasif), *Chest Pain* (nyeri dada), *Chouging Of Blood* (batuk berdarah), *Fatigue* (kelelahan), *Weight Loss* (kehilangan berat badan), *Shortness Of Breath* (sesak nafas), *Wheezing* (mengi), *Swallowing Difficulty* (sulit menelan), *Clubbing Of Finger Nails* (menggigiti kuku jari), *Frequent Cold* (sering kedinginan), *Dry Cough* (batuk kering), *Snoring* (mendengkur), dan kategorikal adalah *Level*.

Pre Processing

Pembersihan data untuk mengetahui fitur yang diperlukan atau tidak dengan penggalian jenis tipe data. Dataset ini memiliki 26 fitur, yang terdiri dari 2 jenis tipe data yaitu, integer dan *object*. Data numerik yang berisi bilangan bulat positif, nol dan negatif serta tidak memiliki desimal juga pecahan merupakan pengertian integer. Sementara penggambaran suatu entitas dunia nyata yang memiliki atribut atau metode tertentu adalah pengertian *object,* contohnya dalam dataset ini yang termasuk object ialah *Patient ID* mewakili identitas pasien satu dengan yang lain dan *Level* menggambarkan kategorikal penyakit kanker paru seorang pasien. *Integer* terdiri dari 24 fitur sementara *object* hanya 2 fitur, lebih detailnya dapat dilihat pada gambar dibawah.

```
Data columns (total 26 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   index                     1000 non-null   int64
 1   Patient Id                1000 non-null   object
 2   Age                       1000 non-null   int64
 3   Gender                    1000 non-null   int64
 4   Air Pollution             1000 non-null   int64
 5   Alcohol use               1000 non-null   int64
 6   Dust Allergy              1000 non-null   int64
 7   OccuPational Hazards      1000 non-null   int64
 8   Genetic Risk              1000 non-null   int64
 9   chronic Lung Disease      1000 non-null   int64
 10  Balanced Diet             1000 non-null   int64
 11  Obesity                   1000 non-null   int64
 12  Smoking                   1000 non-null   int64
 13  Passive Smoker            1000 non-null   int64
 14  Chest Pain                1000 non-null   int64
 15  Coughing of Blood         1000 non-null   int64
 16  Fatigue                   1000 non-null   int64
 17  Weight Loss               1000 non-null   int64
 18  Shortness of Breath       1000 non-null   int64
 19  Wheezing                  1000 non-null   int64
 20  Swallowing Difficulty     1000 non-null   int64
 21  Clubbing of Finger Nails  1000 non-null   int64
 22  Frequent Cold             1000 non-null   int64
 23  Dry Cough                 1000 non-null   int64
 24  Snoring                   1000 non-null   int64
 25  Level                     1000 non-null   object
dtypes: int64(24), object(2)
memory usage: 203.2+ KB
```

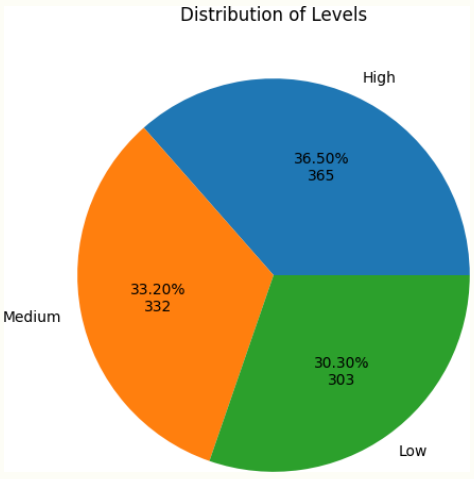
Gambar 2. *Tipe Data setiap Fitur*

Fitur yang akan dihilangkan yaitu *Index* dan *Patient ID* karena dua fitur tersebut tidak merepresentasikan data untuk pengolahan prediksi kanker paru. Selanjutnya, tersisa fitur level yang bertipe *object*. Atribut level ini harus diubah ke dalam bentuk numerik agar mudah dalam proses pengolahan data. Level *Low* diubah menjadi 0, *Medium* menjadi 1 dan *High* menjadi 2 seperti terlihat pada gambar 3.

```
Cancer Levels:  ['Low' 'Medium' 'High']
Cancer Levels:  [0 1 2]
```

Gambar 3. Hasil Perubahan Atribut Level

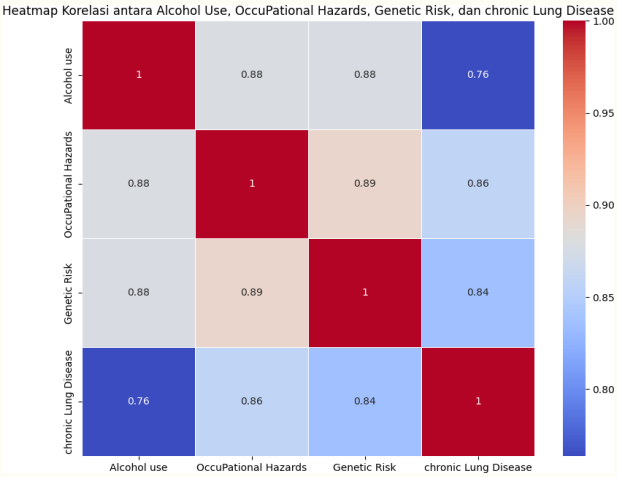
Langkah selanjutnya, memvisualisasikan atribut level.



Gambar 4. Visualisasi Data Atribut Level

Berdasarkan Gambar 4, level mengartikan bahwa seseorang yang diprediksi memiliki penyakit kanker paru, mempunyai 3 kategorik level, yaitu *Low* (kecil), *Medium* (sedang) dan *High* (tinggi). *Low* sendiri memiliki arti seorang pasien diprediksi dalam kategori kecil untuk dikatakan memiliki penyakit kanker paru, medium memiliki pengertian seorang pasien diprediksi dalam kategori sedang atau dengan kata lain kemungkinannya 50:50 untuk memiliki penyakit kanker paru. Sementara untuk kategori tinggi memiliki arti bahwa seorang pasien dapat dikatakan memiliki penyakit kanker paru.

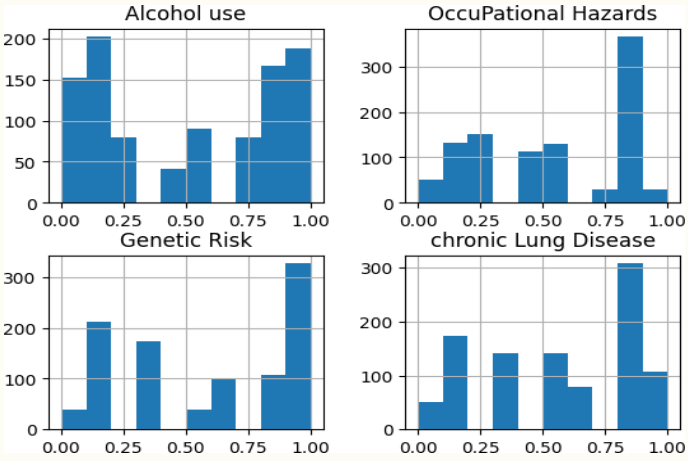
Gambar 4 menjelaskan bahwa pasien dengan kategori *High* sebanyak 36.50%, *Medium* sebanyak 33.20%, dan *Low* sebanyak 30.30% berdasarkan dataset. Proses berikutnya memunculkan korelasi data antar atribut, lebih jelasnya dapat dilihat pada gambar di bawah ini



Gambar 5. Korelasi Data

Gambar 5 adalah korelasi data yang menjelaskan bahwa atribut *Alcohol Use*, *Occupational Hazards*, *Genetic Risk* serta *Chronic Lung Disease* memiliki nilai korelasi mendekati 1. Seperti terlihat pada gambar 5, perbandingan *Alcohol Use* dengan *Occupational Hazards* memperoleh nilai 0.88 nilai ini sama seperti perbandingan *Alcohol Use* dengan *Genetic Risk*. Sementara perbandingan korelasi *Occupational Hazards* dengan *Genetic Risk* memperoleh nilai 0.89, *Occupational Hazards* dengan *Chronic Lung Disease* memperoleh nilai 0.86. Kemudian *Genetic Risk* dengan *Chronic Lung Disease* memperoleh nilai korelasi 0.84. Perbandingan korelasi atribut tersebut memiliki arti cukup berpengaruh satu sama lain, karena nilai korelasi antar atribut diatas mendekati nilai 1.

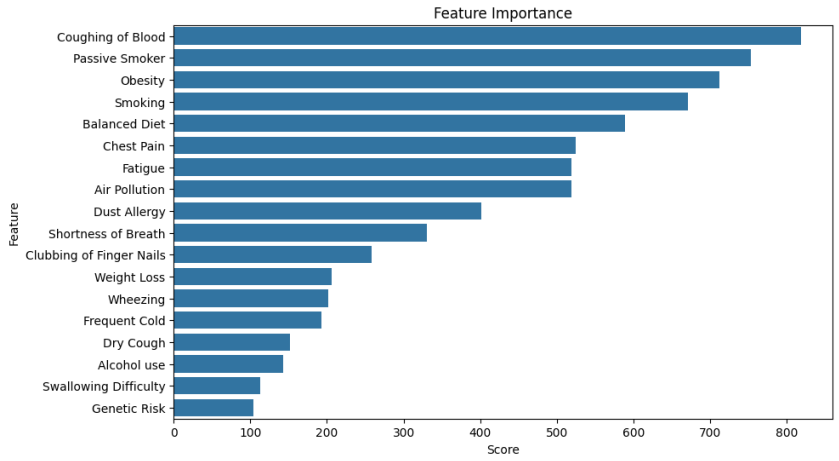
Proses berikutnya menormalisasikan empat atribut diatas. Gambar 6 memunculkan sebaran data dari empat fitur yang sebelumnya. Hasilnya ke empat fitur tersebut tidak ada yang memiliki sebaran data terdistribusi normal, hal ini bisa dilihat dari bentuk grafik dari masing-masing fitur, empat fitur tersebut grafiknya terlalu curam.



Gambar 6. Grafik *Normalisasi Min-Max*

Seleksi Fitur

Tahap selanjutnya yang dilakukan ialah pengujian *chi- square*. Proses ini berguna untuk memilih fitur terbaik dengan melihat hubungan atau asosiasi antara dua variabel kategori . Berikut grafik visualisasi dari pengujian chi-square



Gambar 7. *Feature Importance*

Gambar 7 memunculkan fitur yang memperoleh *score* 100 hingga 800. Tiga *score* tertinggi diperoleh oleh *coughing of blood* sebesar 800, *passive smoker* sebesar 750 dan *obesity* sebesar 725. Dari grafik di atas dapat diperoleh informasi fitur - fitur penting dalam data set.

Split Data

Split data penelitian ini menggunakan rasio 80 : 20, 80% untuk data training dan 20% untuk data testing. Rentang data *Training* pada dataset adalah 800 data teratas, sementara data *Testing* pada dataset 200 data setelah data Training. Berikut *Split Data* yang diproses

Tabel 1. *Split Data (Training)*

	Air Pollution	*Alcohol Use*	*Dust Allergy*	*…*	*Clubbing Of Finger Nails*	*Frequent Cold*	*Dry Cough*
0	1	0.71428571	7	…	8	7	7
1	6	1	7	...	2	1	7
2	1	0.71428571	7	…	8	7	7
…	…	…	…	…	…	…	…
797	2	0	5	…	1	3	2
798	6	0.57142857	6	…	4	6	7
799	6	0.57142857	6	…	4	6	7

Tabel 2. *Split Data (Testing)*

	Air Pollution	*Alcohol Use*	*Dust Allergy*	*…*	*Clubbing Of Finger Nails*	*Frequent Cold*	*Dry Cough*
800	1	0.71428571	7	…	3	1	2
801	6	1	7	...	9	3	4
802	6	1	7	…	9	3	4
…	…	…	…	…	…	…	…
997	6	0.85714286	7	…	2	4	5
998	2	0.28571429	2	…	2	3	2
999	1	0.71428571	7	…	8	7	7

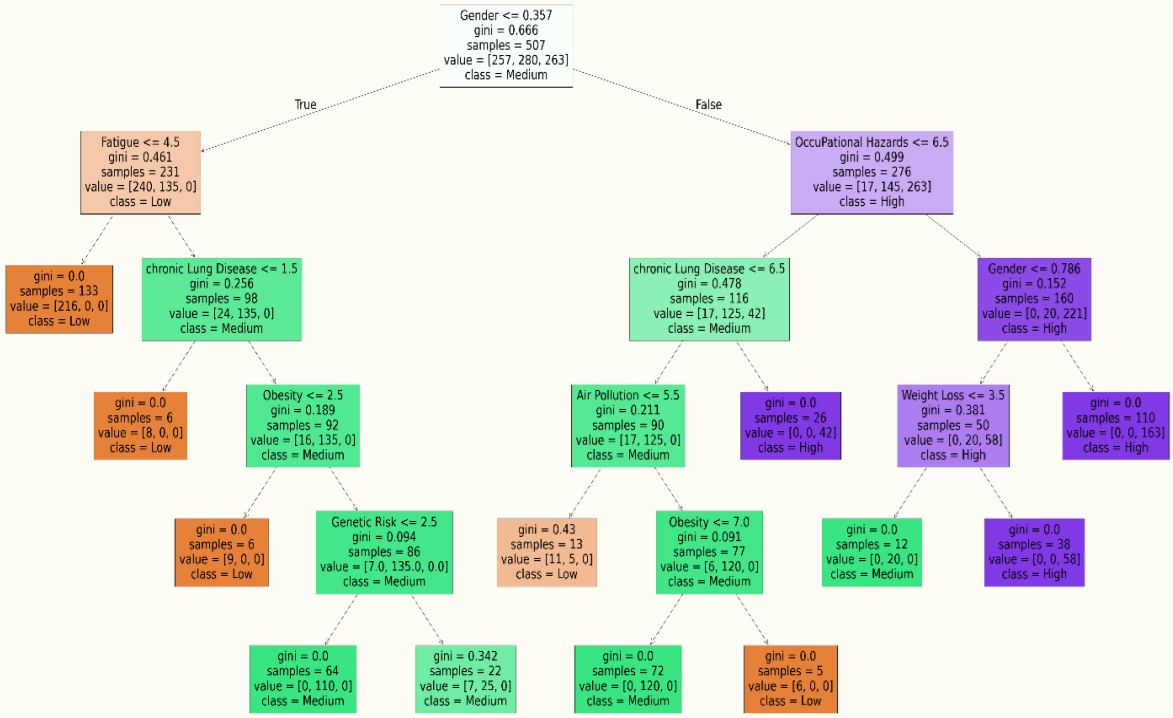
Model

Proses modelling penelitian ini menggunakan 3 algoritma, yaitu *Random Forest*, *Decision Tree* dan *Naïve Bayes*.

Tabel 3. Perbandingan Algoritma

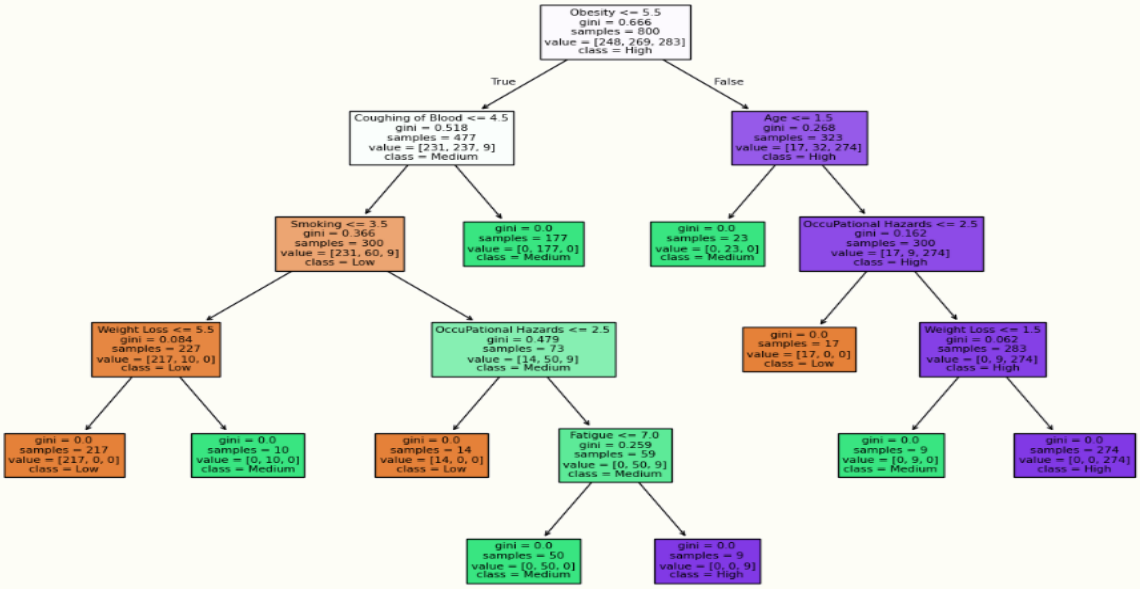
Algoritma	***Accuracy***	***Precision***	***Recall***
Random Forest	100%	100%	100%
Decision Tree	100%	100%	100%
Naïve Bayes	86%	84%	89%

Tabel 3 menunjukan hasil *Accuracy*, *Precision* dan *Recall Random Forest* dan *Decision Tree* memperoleh hasil 100% sementara *Naïve Bayes* 86% untuk *Accuracy*, 84% *Precision* dan 89% *Recall*. Salah satu dari beberapa output keputusan yang dihasilkan oleh model *Random Forest*. Gambar 8 *Random Forest* terdiri dari banyak pohon keputusan, yang masing-masing menggunakan kumpulan data yang berbeda dan kelompok fitur yang berbeda. *Random Forest* menggunakan beberapa prinsip untuk meningkatkan akurasi prediksi dan mengurangi *overfitting*.



Gambar 8. Pohon Keputusan Random Forest

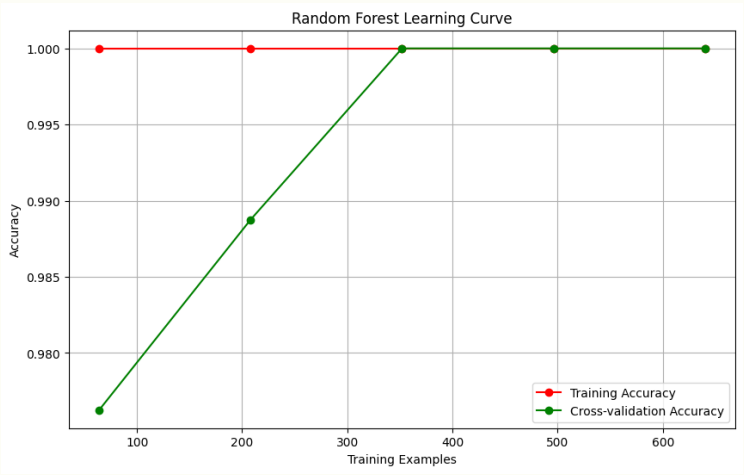
Gambar 9 Visualisasi komprehensif dari model *Decision Tree* yang dikembangkan menggunakan data Iris. Setiap *node* menampilkan keputusan berdasarkan fitur, dan pohon menghasilkan klasifikasi berdasarkan hasilnya. *Decision Tree* beroperasi secara independen dan menjadi lebih sederhana ketika terjadi *overfitting*.



Gambar 9. Pohon Keputusan *Decision Tree*

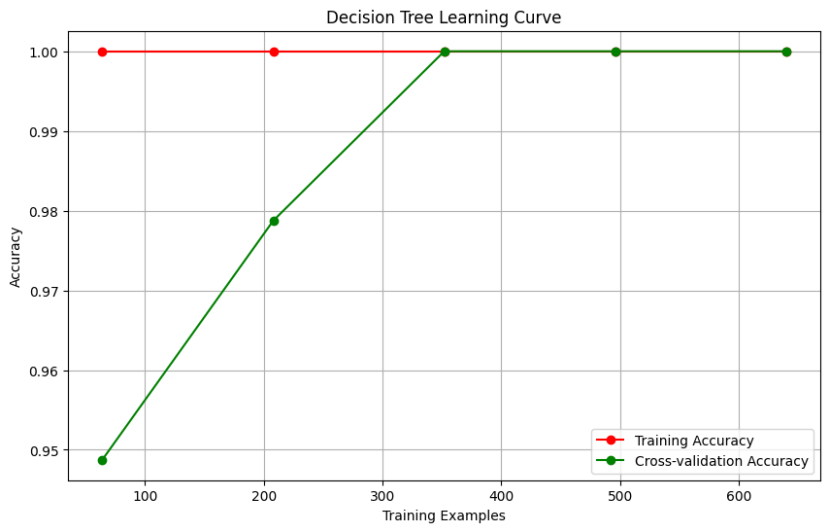
Evaluasi

Evaluasi penelitian kali ini menggunakan *learning curve*. *Learning curve* berfungsi untuk mengevaluasi sebuah model yang telah melakukan proses belajar, seiring berjalannya jumlah data yang tersedia dengan waktu proses. Hasil learning curve terdapat 3 jenis, yaitu *Goodfiting*, *Underfitting*, serta *Overfitting*. *Goodfitting* ialah hasil yang menandakan bahwa model yg digunakan memiliki performa cukup baik, sementara *Underfitting* menandakan bahwa model tidak dapat mengenali pola dalam data latih, sedangkan *Overfitting* menandakan bahwa model terlalu dapat mengenali data latih sehingga tidak dapat mengenali pola dalam data uji. Hasil *Learning Curve* untuk *Random Forest* dan *Decision Tree* dan *Naïve Bayes* dapat dilihat dari gambar di bawah ini.



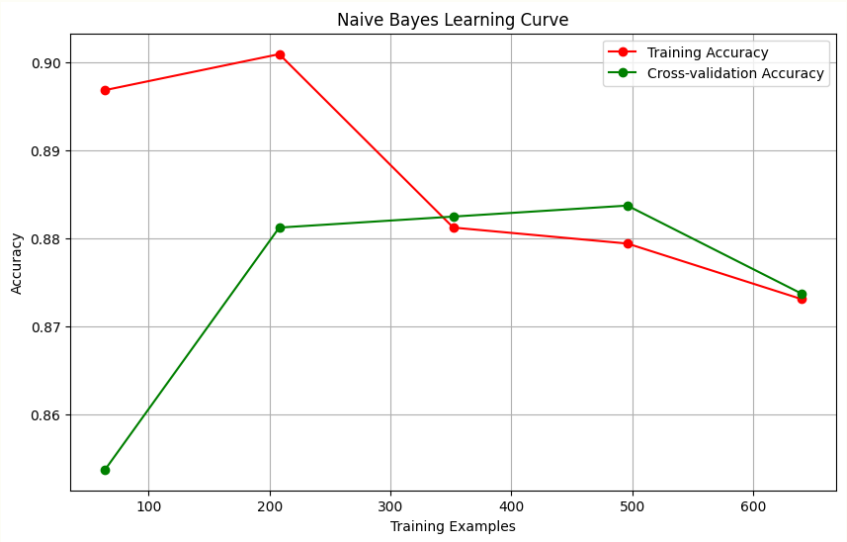
Gambar 10. *Random Forest Curve*

Gambar 10 visualisasi *Learning Curve* dari algoritma *Random Forest*, garis merah (sumbu x) sebagai *Data Training* dan garis hijau (sumbu y) sebagai nilai *Cross Validation Accuracy*. Dari grafik di atas garis hijau mendekati garis merah, hingga mecapai nilai 350 kedua garis bertemu dan setelahnya garis mendatar. Grafik ini dinamakan *Goodfitting* yang menggambarkan bahwa algoritma *Random Forest* menjadi algoritma terbaik untuk prediksi kanker paru.



Gambar 11. *Decision Tree Curve*

Gambar 11 visualisasi *Learning Curve* dari algoritma *Decision Tree*, garis merah (sumbu x) sebagai Data Training dan garis hijau (sumbu y) sebagai nilai *Cross Validation Accuracy*. Dari grafik di atas garis hijau mendekati garis merah, hingga mecapai nilai 350 kedua garis bertemu dan setelahnya garis mendatar. Grafik ini dinamakan *Goodfitting* yang menggambarkan bahwa algoritma *Decision Tree* menjadi algoritma terbaik untuk prediksi kanker paru.



Gambar 12. *Naive Bayes Curve*

Gambar 12 visualisasi *learning curve* dari algoritma *Naïve Bayes*, garis merah (sumbu x) sebagai Data Training dan garis hijau (sumbu y) sebagai nilai *Cross Validation Accuracy*. Dari grafik di atas garis hijau hanya naik hingga nilai 500 setelahnya grafik mengalami penurunan, sementara garis merah naik hingga titik 202 setelahnya grafik mengalami penurunan terus dan tidak sejajar dengan garis hijau, grafik ini dinamakan *Overfitting* yang menunjukkan algoritma *Naïve Bayes* kurang optimal dalam prediksi kanker paru. Dari evaluasi *learning curve* ketiga algoritma diatas menunjukan bahwa *Random Forest* dan *Decision Tree* menunjukkan bahwa algoritma tersebut memiliki kinerja yang baik, sementara Naïve Bayes kinerjanya kurang dikarenakan perbedaan metode kinerja dengan *Random Forest* juga *Decision Tree*.

KESIMPULAN

Nilai akurasi terbaik yang diperoleh oleh Random Forest dan Decision Tree sebesar 100%, Naïve Bayes menghasilkan nilai akurasi 86%. Hasil akurasi ini dipengaruhi oleh perbedaan metode kerja algoritma, Random Forest dan Decision Tree berbasis pohon keputusan sementara Naïve Bayes berbasis probabilitas teori bayes serta asumsi independen antar fitur. Terdapat beberapa ciri yang dapat dikategorikan seseorang diprediksi memiliki kanker paru, ciri-ciri tersebut antara lain batuk berdarah, perokok pasif, obesitas, dan perokok. Penelitian ini masih memiliki banyak kekurangan dan perlu dilakukan perbaikan. Saran untuk penelitian berikutnya bisa melakukan perbandingan Random Forest dengan algoritma lainnya seperti Gradient boosting, XGBoost, AdaBoost, sementara Naïve Bayes bisa dilakukan perbandingan dengan Bayesian Network, K-Nearest Neighbors dan lainnya

DAFTAR PUSTAKA

Agtira, B. H., Handayani, H. H., & Masruriyah, A. F. N. (2023). Perbandingan Algoritma NBC dan Decision Tree pada Sentimen Analisis Mengenai Vaksinasi Covid-19 Di Indonesia. *Remik*, *7*(1), 704–712. https://doi.org/10.33395/remik.v7i1.12151

Alamsyah, H., Cahyana, Y., & Pratama, A. R. (2023). Deteksi Fake Review Menggunakan Metode Support Vector Machine dan Naïve Bayes Di Tokopedia. *Jutisi: Jurnal Ilmiah Teknik Informatika Dan Sistem Informasi*, *12*, no.*2*, 585–598.

Aldiansyah Poetra, F., Rohana, T., & Elvira Awal, E. (2023). *Implementasi Algoritma Naïve Bayes Untuk Mendiagnosa Skizofrenia Berbasis Web*. *IV*(2), 146.

Appasani, Bokkisam, & Surendran. (2024). An Incremental Naive Bayes Learner for Real-Time Health Prediction. *Procedia Computer Science*, *235*, 2942–2954.

Apriliah, W., Kurniawan, I., Baydhowi, M., & Haryati, T. (2021). Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest. *Sistemasi*, *10*(1), 163. https://doi.org/10.32520/stmsi.v10i1.1129

Arifuddin, A., Buana, G. S., Vinarti, R. A., & Djunaidy, A. (2024). Performance Comparison of Decision Tree and Support Vector Machine Algorithms for Heart Failure Prediction. *Procedia Computer Science*, 628–636.

Arum, M. P., & Triyono. (2021). Genetic Algorithm For Feature Selection In Naive Bayes In Life Resistance Classification On Breast Cancer Patient. *Jurnal Ilmu KOmputer An Nuur*, *.1*, 32–37.

Fadillah Hermawan, A., Rakhmat Umbara, F., & Kasyidi, F. (2022). MIND (Multimedia Artificial Intelligent Networking Database Prediksi Awal Penyakit Stroke Berdasarkan Rekam Medis menggunakan Metode Algoritma CART(Classification and Regression Tree). *Journal MIND Journal | ISSN*, *7*(2), 151–164. https://doi.org/10.26760/mindjournal.v7i2.151-164

Harto, M. K., & Basuki, A. (2021). Deteksi Serangan DDoS Pada Jaringan Berbasis SDN Dengan Klasifikasi Random Forest. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *5*(4), 1329–1333. http://j-ptiik.ub.ac.id

Hidayanti, A., Siregar, A. M., Arum, S., Lestari, P., & Cahyana, Y. (2022). Model Analisis Kasus Covid-19 di Indonesia Menggunakan. *Jurnal Pengkajian Dan Penerapan Teknik Informatika*, *15*(1), 91–101.

Kenang Candra Alivian Pratama, H., Suharso, W., Kunci, K., Naïve Bayes, B., Naïve Bayes, G., & Naïve Bayes, M. (2022). Pengklasifikasian Kanker Payudara Dan Kanker Paru-Paru Dengan Metode Gaussian Naïve Bayes, Multinomial Naïve Bayes, Dan Bernoulli Naïve Bayes Classification Of Breast Cancer And Lung Cancer Using The Gaussian Naïve Bayes Multinomial Nave Bayes And Bernoul. *Jurnal Smart Teknologi*, *3*(4), 2774–1702.

http://jurnal.unmuhjember.ac.id/index.php/JST

Meiyanti, A., & Komarudin, R. (2020). Klasifikasi Diagnosa Penyakit Paru-Paru Pada Klinik Raditya Medical Center Dengan Metode Algoritma C4.5. *JSI: Jurnal Sistem Informasi (E-Journal)*, *12*(1), 1894–1905. https://doi.org/10.36706/jsi.v12i1.9456

Mia, Nur Masruriyah, A. F., & Pratama, A. R. (2024). Komparasi Model DecisionTree dan Random Forest untuk Memprediksi Penyakit Jantung. *Scientific Student Journal for Information, Technology and Science*, *V*(2), 123–130.

Naezer, M., & Supriyanto, R. (2023). Analisis Kinerja Algoritma Naïve Bayes dan k-NN untuk Memprediksi Penyakit Kanker Paru. *Jurnal Ilmiah KOMPUTASI, 22 No 2*(p-ISSN 1412-9434/e-ISSN 2549-7227).

Nair, Devi, & Bhasi. (2024). *Enhanced lung cancer detection: Integrating improved random walker segmentation with artificial neural network and random forest classifier*. Heliyon.

Permana, B. A. C., & Djamaluddin, M. (2023). Penerapan Python Dalam Data Mining Untuk Prediksi Kangker Paru. *Infotek : Jurnal Informatika Dan Teknologi*, *6*(2), 470–477. https://doi.org/10.29408/jit.v6i2.17816

Qisthiano, M. R., Prayesy, P. A., & Ruswita, I. (2023). Penerapan Algoritma Decision Tree dalam Klasifikasi Data Prediksi Kelulusan Mahasiswa. *G-Tech: Jurnal Teknologi Terapan*, *7*(1), 21–28. https://doi.org/10.33379/gtech.v7i1.1850

Rahman Wahid, M. A., Nugroho, A., & Halim Anshor, A. (2023). Prediksi Penyakit Kanker Paru-Paru Dengan Algoritma Regresi Linier. *Bulletin of Information Technology (BIT)*, *4*(1), 63–74. https://doi.org/10.47065/bit.v4i1.501

Sari, L., Romadloni, A., & Listyaningrum, R. (2023). Penerapan Data Mining dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest. *Infotekmesin*, *14*(1), 155–162. https://doi.org/10.35970/infotekmesin.v14i1.1751

Septhya, D., Rahayu, K., Rabbani, S., Fitria, V., Rahmaddeni, R., Irawan, Y., & Hayami, R. (2023). Implementasi Algoritma Decision Tree dan Support Vector Machine untuk Klasifikasi Penyakit Kanker Paru. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *3*(1), 15–19. https://doi.org/10.57152/malcom.v3i1.591

Sinaga, R. B., Widiyanto, D., & Wahyono, B. T. (2022). Deteksi Dini Penyakit Kanker Paru dengan Gabungan Algoritma Adaboost dan Random Forest. *Seminar Nasional Mahasiswa Ilmu Komputer Dan Aplikasinya (SENAMIKA)*, 1–10. https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer

Zulaikhah Hariyanti Rukmana, S., Aziz, A., & Harianto, W. (2022). Optimasi Algoritma K-Nearest Neighbor (Knn) Dengan Normalisasi Dan Seleksi Fitur Untuk Klasifikasi Penyakit Liver. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *6*(2), 439–445. https://doi.org/10.36040/jati.v6i2.4722