 ISSN:2302-8149
e-ISSN:2540-9719

Klasifikasi Penyakit Kanker Paru-Paru dengan *Algoritma Extreme Gradient Boosting (XGBoost)* dan *Mutual Information* sebagai Metode *Feature Selection*

Lung Cancer Classification Using the Extreme Gradient Boosting (XGBoost) Algorithm and Mutual Information for Feature Selection

[1]Regitha Zizilia, [2]Yulison Herry Chrisnanto*, [3]Gunawan Abdillah
[1,2,3]Program Studi Informatika, Fakultas Sains dan Informatika, Universitas Jenderal Achmad Yani
[1,2,3]Jl. Terusan Jenderal Sudirman, Cimahi, Jawa Barat, Kota Cimahi, Jawa Barat 40525, Indonesia
*e-mail: yhc@if.unjani.ac.id

(***received***: 23 May 2025, ***revised***: 13 June 2025, ***accepted***: 14 June 2025)

Abstrak

Kanker paru-paru merupakan salah satu jenis kanker paling mematikan di dunia yang sering kali terlambat terdeteksi karena gejala yang tidak muncul pada tahap awal. Penelitian ini bertujuan untuk mengevaluasi pengaruh seleksi fitur menggunakan Mutual Information terhadap performa klasifikasi kanker paru-paru dengan algoritma XGBoost. Mutual Information digunakan untuk memilih fitur yang relevan, baik yang memiliki hubungan *linier* maupun *non-linier* dengan target, sementara XGBoost dipilih karena kemampuannya dalam mengolah *dataset* besar dan mengurangi *overfitting*. Penelitian dilakukan pada *dataset* berukuran 30000 baris data dengan skenario pembagian data yaitu 90:10, 80:20, 70:30, dan 60:40. Hasil menunjukkan bahwa akurasi *testing* sebelum penerapan Mutual Information sebesar 93.42% hingga 93.83% dan akurasi K-Fold Cross Validation sebesar 94.59% hingga 94.76%. Setelah seleksi fitur, akurasi *testing* tetap stabil pada skenario pembagian data 70:30 dan 60:40 sebesar 93.60% dan 93.42% dengan penurunan akurasi pada K-Fold Cross Validation menjadi 89.26% dan 90.88%. Namun, pada skenario pembagian data 90:10 dan 80:20 terjadi penurunan akurasi menjadi 88.63% dan 88.85% pada *testing*, serta 88.87% dan 90.24% pada akurasi K-Fold Cross Validation. Seleksi fitur dengan Mutual Information memberikan efisiensi waktu komputasi karena jumlah fitur yang lebih sedikit, dan dapat diterapkan secara efektif untuk menyederhanakan fitur tanpa mengurangi performa model pada beberapa skenario data dengan penyesuaian karakteristik data.

Kata kunci: klasifikasi, kanker paru-paru, mutual information, XGBoost, k-fold cross validation

Abstract

Lung cancer is one of the deadliest types of cancer worldwide and is often detected too late due to the absence of early symptoms. This study aims to evaluate the impact of feature selection using Mutual Information on the performance of lung cancer classification with the XGBoost algorithm. Mutual Information is employed to select relevant features, including those with linear and non-linear relationships with the target variable, while XGBoost is chosen for its ability to handle large datasets and reduce overfitting. The study was conducted on a dataset containing 30,000 data entries, with data split scenarios of 90:10, 80:20, 70:30, and 60:40. The results show that the testing accuracy before applying Mutual Information ranged from 93.42% to 93.83%, while K-Fold Cross-Validation accuracy ranged from 94.59% to 94.76%. After feature selection, testing accuracy remained stable for the 70:30 and 60:40 split scenarios, at 93.60% and 93.42% respectively. However, K-Fold Cross-Validation accuracy decreased to 89.26% and 90.88%. In contrast, for the 90:10 and 80:20 split scenarios, a decline in accuracy was observed — testing accuracy dropped to 88.63% and 88.85%, and K-Fold Cross-Validation accuracy fell to 88.87% and 90.24%. Feature selection using Mutual Information improves computational efficiency by reducing the number of features, and it can be effectively applied to simplify feature sets without significantly compromising model performance in certain data scenarios, depending on the characteristics of the dataset.

Keywords: *classification, lung cancer, mutual information, XGBoost, K-Fold cross validation*
http://sistemasi.ftik.unisi.ac.id

 ISSN:2302-8149
e-ISSN:2540-9719

1 Pendahuluan

Kanker paru-paru adalah salah satu kanker paling mematikan di dunia. Organisasi Kesehatan Dunia (WHO) melaporkan bahwa kanker paru-paru menjadi penyebab utama kematian terkait kanker, dengan banyak kasus yang tidak menunjukkan gejala pada tahap awal penyakit [1]. Hal ini menyebabkan sebagian besar pasien baru didiagnosa ketika penyakit sudah mencapai stadium lanjut, di mana pengobatan menjadi jauh lebih sulit dan seringkali kurang efektif. Kebiasaan merokok, paparan polusi udara, dan sejumlah faktor genetik dan lingkungan lainnya adalah faktor risiko kanker paru-paru. Dengan berbagai faktor risiko dan tingginya angka kematian menunjukkan perlunya metode deteksi yang lebih efektif dan efisien.

Untuk mengembangkan model klasifikasi berbasis algoritma yang dapat mendeteksi faktor risiko dari data terkait kanker paru-paru, penelitian ini mengaplikasikan teknik *machine learning*. Berbagai penelitian sebelumnya telah menggunakan algoritma seperti Support Vector Machine (SVM) [2] [3], Narrow Neural Network (NNN) [2], Random Forest [3] [4], Decision Tree [4][3], K-Nearest Neighbors (KNN) [2], dan Naïve Bayes [2][3][4] dengan hasil prediksi yang bervariasi. Meskipun algoritma-algoritma tersebut memberikan akurasi yang cukup baik, tetapi tetap ada tantangan dalam hal pemrosesan, penanganan dan variasi data yang perlu diperhatikan lebih lanjut [4]. dan kurangnya penggunaan atau ekplorasi teknik lain seperti *ensemble learning* seperti *boosting* secara mendalam [3]. Karena kanker paru-paru merupakan penyakit yang harus dikendalikan pada setiap tahap perkembangannya, diperlukan langkah-langkah yang memastikan diagnosis yang akurat. Dalam pengembangan algoritma machine learning untuk mengklasifikasi kanker paru-paru, penting untuk menemukan metode yang efektif agar klasifikasi penyakit ini bisa dilakukan dengan cara yang serupa dengan teknik analisis medis yang ada. Selain itu, pengolahan data juga sangat diperhatikan karena kualitas data yang digunakan akan sangat memengaruhi hasil klasifikasi. Oleh karena itu *feature selection* menjadi bagian penting dalam proses ini. *Feature selection* bertujuan untuk memilih fitur-fitur yang paling berpengaruh terhadap klasifikasi kanker paru-paru.

Berdasarkan penelitian terkait maupun latar belakang permasalahan yang telah diuraikan, penelitian ini memanfaatkan Mutual Information sebagai metode *feature selection* dalam meningkatkan kualitas data sebelum digunakan dalam model klasifikasi utama, yaitu XGBoost. Meskipun berbagai metode *feature selection* dapat digunakan dalam proses untuk memilih fitur-fitur yang lebih relevan[4]. Mutual Information dipilih dalam penelitian ini karena sangat efektif untuk mengukur korelasi antar variabel yang tidak terbatas pada hubungan linier saja tetapi juga dapat menangkap hubungan *non-linier* antar fitur dan target [5] dimana hubungan *non linier* sering kali ditemukan pada faktor resiko penyakit. Dengan memilih fitur-fitur yang paling relevan, penelitian ini menyaring variabel yang memiliki kontribusi signifikan terhadap klasifikasi sehingga model yang dibangun menjadi lebih fokus dan efektif.

Algoritma XGBoost digunakan untuk membangun model klasifikasi yang lebih akurat karena penyakit kanker paru-paru memiliki banyak faktor yang menjadi perhatian. Algoritma XGBoost dirancang dengan optimasi berbasis *gradient boosting*, dimana dapat melakukan penanganan *dataset* besar[6]. XGBoost juga memiliki kemampuan bawaan dalam generalisasi untuk mengurangi resiko *overfitting* dimana Gradient Boosting tidak memilikinya. Oleh karena itu, penelitian ini bertujuan untuk mengembangkan model klasifikasi kanker paru-paru yang akurat dan efisien dengan menggabungkan keunggulan metode Mutual Information dalam proses seleksi fitur yang dapat meningkatkan efisensi proses pelatihan model melalui pengurangan kompleksitas data dan kekuatan algoritma XGBoost dalam membangun model klasifikasi.

2 Tinjauan Literatur

Adapun penelitian terkait yang dilakukan beberapa peneliti dalam proses klasifikasi penyakit kanker paru-paru. Penelitian pertama [2] menggunakan KNN, SVM, Naïve Bayes, dan NNN untuk mengidentifikasi dan mengklasifikasikan kanker paru-paru dan membandingkan akurasi dari masing-masing algoritma untuk menemukan metode yang paling efektif. Hasil dari penelitian ini menunjukkan SVM memiliki akurasi tertinggi di antara algoritma lain yang diuji dengan akurasi 92.6% tetapi terdapat indikasi bahwa kinerja SVM dapat bervariasi tergantung pada parameter yang digunakan sehingga mungkin tidak efektif pada *dataset* yang lebih besar atau memiliki banyak *noise* sehingga peneliti menyarankan untuk lebih mengeksplor dataset yang lebih besar dan menggunakan

 ISSN:2302-8149
e-ISSN:2540-9719

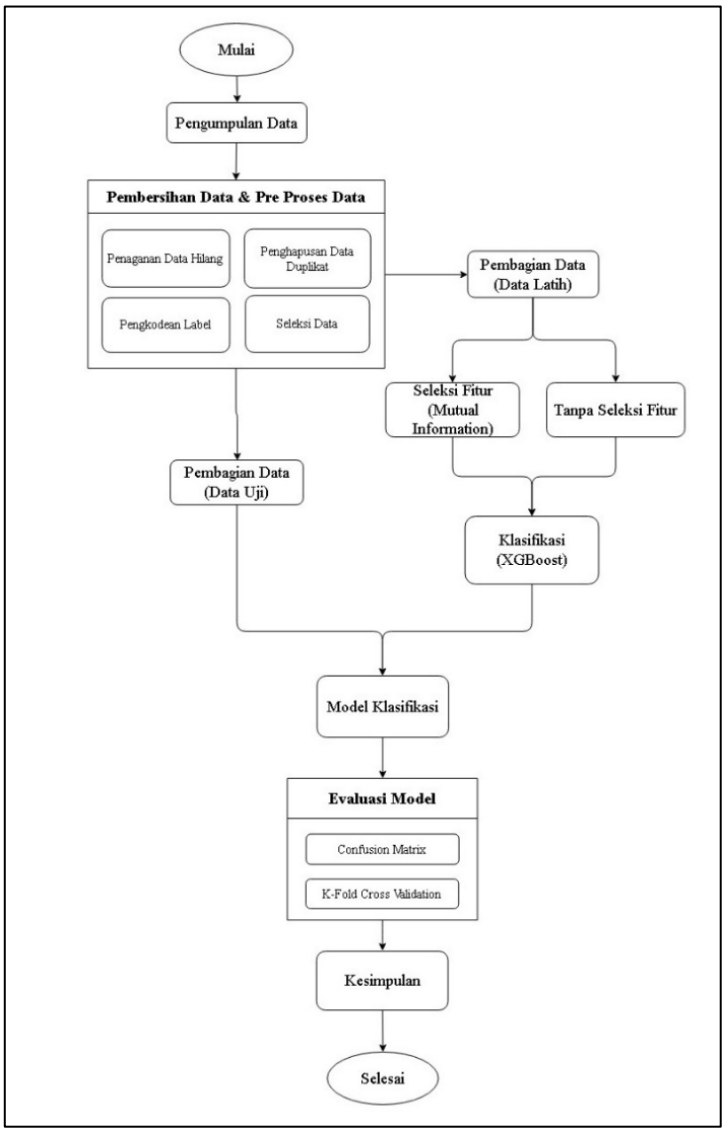
algoritma lain untuk meningkatkan akurasi dan mengurangi *overfitting* agar memberikan hasil yang lebih stabil.

Penelitian lainnya [3] menggunakan algoritma Support Vector Machines (SVM), Random Forest, K-Nearest Neighbors (KNN) , Naïve Bayes, Logistic Regression, Gradient Boosting dan Majority Voting untuk penerapan algoritma ensembles dan *machine learning* dalam deteksi kanker paru-paru karena dapat meningkatkan akurasi deteksi dengan menggabungkan kekuatan beberapa model klasifikasi. Hasilnya Gradient Boosting menjadi algoritma teratas dalam perbandingan berdasarkan nilai tertinggi yang diperoleh dan peningkatan keseluruhan dalam metrik. Penelitian ini menyarankan untuk mengeksplorasi teknik *ensemble* seperti *Stacking*, *Boosting* dan *Bagging* dalam meningkatkan kualitas model deteksi dan dapat digunakan pada berbagai jenis data.

Penelitian lainnya[4]melakukan klasifikasi data dengan menggunakan berbagai algoritma seperti Support Vector Machine (SVM), K-Neighbour, Decision Tree, Logistic Regression, Naïve Bayes, dan Random Forest. Hasil penelitian menunjukkan tingkat kinerja tertinggi adalah algoritma Random Forest dengan hasil 88.5%. Penelitian selanjutnya dapat melakukan menggunakan *dataset* yang lebih besar dan beragam, mengekplorasi algoritma lainnya dan penerapan teknik *feature selection* untuk mengedintifikasi fitur-fitur relevan.

3 Metode Penelitian

Berikut adalah alur metode penelitian yang dilakukan untuk mengklasifikasi penyakit kanker paru-paru yang dapat dilihat dalam Gambar 1.



Gambar 1. Metode penelitian

 ISSN:2302-8149
e-ISSN:2540-9719

3.1 Pengumpulan Data

Tahap pertama yang dilakukan adalah pengumpulan data yang nantinya data ini akan digunakan dalam *Training* dan *Testing*. Data diperoleh dari website Kaggle.com dengan nama "Dataset Predic Terkena Penyakit Paru-Paru" yang memiliki 11 kolom dan 30000 baris data.

3.2 Pembersihan Data dan Pra-proses Data

Pembersihan data bertujuan untuk memperbaiki dan menghapus data yang tidak konsisten dan persiapan data bertujuan untuk mengubah data menjadi format yang tersturktur.

3.2.1 Penanganan Data Hilang

Data dapat hilang karena berbagai alasan seperti kesalahan entri data, masalah pengumpulan data, atau kesalahan pemrosesan data. Penggantian nilai yang hilang merupakan salah satu masalah penting, banyak penelitian yang kinerja pengklasifikasiannya rendah karena tidak ada nilai yang sesuai untuk mengisi nilai yang hilang. Dalam proses pengklasifikasian, nilai yang hilang dalam kumpulan data asli akan mempengaruhi keakuratan dan kinerja dari proses pengklasifikasian tersebut sehingga akan berdampak pada hasil keluaran yang tidak dapat diandalkan dan salah.

Terdapat beberapa cara untuk menangani *missing* data diantaranya adalah metode imputasi dan penghapusan data. Imputasi dilakukan dengan mengisi nilai yang hilang dengan estimasi menggunakan nilai rata-rata (*mean*), *median*, modus, atau teknik lain. Sementara itu penghapusan dilakukan dengan membuang baris yang mengandung data hilang (*listwise deletion*).

3.2.2 Membuang Data Duplikat

Membuang data duplikat adalah proses untuk menghapus baris-baris yang memiliki nilai duplikat dalam dataset. Tujuan dari menghapus baris ini adalah agar mengurangi redudansi dan meningkatkan akurasi analisis.

3.2.3 Pengkodean Label

Pengkodean label adalah metode yang digunakan untuk mengubah data kategorikal atau ordinal menjadi data numerik dengan memberikan label atau kode angka pada setiap kategori atau tingkatan data. Pengkodean label bertujuan untuk mengubah data ke dalam format yang lebih sesuai agar dapat diproses oleh algoritma *machine learning* secara optimal [7].

3.2.4 Seleksi Data

Seleksi data adalah proses yang dilakukan dengan menghapus atribut-atribut yang tidak digunakan atau tidak ada kaitannya dalam proses klasifikasi.

3.3 Pembagian Data

Pembagian *dataset* ini bertujuan untuk melatih model pada data pelatihan dan menguji kinerja model pada data pengujian untuk mengevaluasi seberapa baik model dapat melakukan prediksi pada data baru yang belum pernah dilihat sebelumnya. Pada pembagian data, data dibagi menjadi dua yaitu *data training* dan *data testing* [8].

3.3.1 Data Latih

Data latih merupakan data yang telah diberi label untuk mengenali pola hubungan antar fitur dan label.

3.3.2 Data Uji

Data uji merupakan data yang digunakan untuk menguji performa akhir model setelah melalui proses pelatihan dan validasi, dengan tujuan memberikan estimasi objektif terhadap kemampuan model dalam memprediksi data baru yang belum pernah dilihat atau digunakan selama proses pelatihan [9].

3.4 Seleksi Fitur (Mutual Information)

Feature selection sangat penting dalam *machine learning* untuk meningkatkan kinerja model yang digunakan dengan mengidentifikasi atribut atau fitur yang relevan untuk dianalisis. Dengan memilih fitur yang tepat maka model akan menjadi lebih efisien dan akurat [10]. *Feature selection* yang dilakukan adalah mutual information dimana mutual information bekerja dengan memahami hubungan antara fitur-fitur dengan target prediksi. Dengan perhitungan pada persamaan (1) [11]:

 ISSN:2302-8149
e-ISSN:2540-9719

$$I(X;Y) = H(Y) - H(Y|X) = \sum_{x\epsilon X}\sum_{y\epsilon Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \quad (1)$$

Mutual Information akan menghitung selisih antara entropi awal H(Y) dan entropi kondisional H(Y|X) atau menghitung kontribusi mutual information untuk setiap pasangan (x,y) dan kemudian menjumlahkan semua kontribusi tersebut.
I(X;Y) : Mutual Information antara X dan Y
H(Y) : Entropi Y, yang mengukur ketidakpastian pada Y.
H(Y|X) :Entropi Y jika sudah mengetahui X.
p(x,y) : Fungsi probabilitas gabungan yang memberikan probabilitas terjadinya x dan y secara bersamaan.
p(x) dan p(y) : Distribusi marginal dari X dan Y

3.5 Klasifikasi dengan XGBoost

XGBoost adalah implementasi dari algoritma peningkatan gradien yang bekerja dengan membangun serangkaian pohon keputusan secara iteratif untuk memperbaiki kesalahan prediksi dari model sebelumnya [12]. Setiap pohon yang ditambahkan ke dalam model berusaha untuk mengurangi kesalahan prediksi dengan mempertimbangkan gradien dari fungsi kerugian yang sedang dioptimalkan. XGBoost terus memperbaiki akurasinya dengan melatih model menggunakan *data training* pada setiap iterasi. Sebagai metode pembelajaran kelompok, Gradient Boosting akan membuat model dipelajari secara berurutan. Hal ini membuat setiap model baru berkonsentrasi pada memperbaiki kesalahan prediksi dari model sebelumnya, yang menghasilkan prediktor yang kuat dan akurat. Terdapat beberapa rumus dalam tahapan kerja XGBoost [13] [14] dimana pada tahap pertama adalah melakukan asumsi model yang terdiri dari M pohon keputusan dengan tujuan optimasi yang terdapat pada persamaan (2)

$$\hat{y}_i = \sum_{m=1}^{M} f_m(x_i), f_m \in F \quad (2)$$

Dimana $\hat{y}_i$ merupakan nilai prediksi dari data ke-i. x_i merupakan fitur yang digunakan. $\sum_{m=1}^{M}$ merupakan penjumlahan dari semua pohon keputusan. f_m merupakan satu pohon keputusan yang berdiri sendiri dan F merepresentasikan semua kemungkinan pohon yang dapat digunakan dalam model. Persamaan (3) merupakan fungsi objektif dimana terdiri dari dua bagian yaitu Loss Function $(\sum l(\hat{y}_i, y_i))$ untuk mengukur kesalahan prediksi dan regularization term $(\sum \Omega(f_i))$ untuk mengontrol kompleksitas model dan mencegah *overfitting*.

$$L(f_i) = \sum l(\hat{y}_i, y_i) + \sum \Omega(f_i) \quad (3)$$

Persamaan (4) merupakan *update* prediksi dalam setiap iterasi dan persamaan (5) untuk regularization term pada keseluruhan model. $\hat{y}_i^{(t)}$ merupakan prediksi pada iterasi ke-t untuk data ke-i. $\hat{y}_i^{(t-1)}$ merupakan prediksi pada iterasi sebelumnya dan $f_{i(x_i)}$ merupakan fungsi yang dihasilkan pada iterasi ke-t untuk data ke-i.

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_{i(x_i)} \quad (4)$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (5)$$

Pada fungsi regularisasi, γ merupakan parameter regularisasi yang mengontrol seberapa besar penalty yang ditetapkan berdasarkan jumlah *leaf nodes*. T merupakan jumlah *leaf node*, λ merupakan parameter regularisasi, w merupakan bobot yang diberikan pada setiap *leaf node* dan j merupakan indeks *leaf node*. Persamaan (6) merupakan optimasi fungsi loss dengan pendekatan Taylor derajat ke-2.

 ISSN:2302-8149
e-ISSN:2540-9719

$$L^{(t)} \approx \sum_{i=1}^{n} \left[(g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t) \quad (6)$$

Dimana $L^{(t)}$ merupakan fungsi loss pada iterasi *boosting* ke-t, $f_t(x_i)$ merupakan *output* dari model baru terhadap *input* x. g_i dan h_i merupakan statistik gradien dari fungsi loss dan $\Omega(f_t)$ merupakan fungsi regularisasi untuk mengontrol kompleksitas pohon. Persamaan (7) merupakan optimasi bobot dari setiap *leaf node* dengan gradien dan hessian. Selain itu juga mengontrol *overfitting* dengan regularisasi.

$$L(f_i) \approx \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T \quad (7)$$

$L(f_i)$ merupakan fungsi loss untuk pohon keputusan ke-i. $\sum_{j=1}^{T}$ merupakan penjumlahan dari *leaf nodes* ke-1 sampai *leaf nodes* ke- T, dimana T adalah jumlah *leaf nodes*, j merupakan indeks *leaf* node, I_j merupakan himpunan data yang masuk ke *leaf node* ke-j. g_i dan h_i merupakan statistik gradien dari fungsi loss, W_j merupakan bobot *leaf node* ke j, λ merupakan parameter regularisasi dan i merupakan himpunan daun. Persamaan (8) merupakan perhitungan gain untuk memilih *split* terbaik dalam membangun pohon.

$$Gain = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} + \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma \quad (8)$$

Dimana G_L dan G_R merupakan jumlah gradien pada node kiri dan kanan. H_L dan H_R merupakan jumlah Hessian pada node kiri dan kanan. γ merupakan parameter regularisasi yang mengontrol minimum pengurangan loss untuk melakukan split. Persamaan (9) merupakan output bobot pada daun pohon.

$$w_j^* = -\frac{G_j}{H_j + \lambda} \quad (9)$$

w_j^* merupakan bobot optimal yang akan diberikan pada *leaf node* ke- j. G_j adalah jumlah gradien pada daun j dan H_j adalah jumlah Hessian pada daun j. λ merupakan parameter regularisasi.

Dalam model XGBoost diperlukan penentuan beberapa parameter pengukuran guna mengatur cara kerja model agar lebih efektif. Parameter yag digunakan ditunjukkan pada tabel 1.

Tabel 1. Parameter XGboost

Parameter	Deskripsi
Max_depth	Maksimal kedalaman pohon
Learning_rate	Mengatur besar langkah dalam proses pembelajaran setiap iterasi boosing
N_estimators	Jumlah total pohon yang akan dibangun dalam model
Subsample	Proporsi sampe data tiap iterasi untuk mengurangi overfitting
Colsample_bytree	Proporsi fitur tiap pohon
Reg_lambda	Regularisasi L2
Reg_alpha	Regularisasi L1
gamma	Mengontrol kompleksitas pohon keputusan
booster	Jenis model dasar yang akan digunakan dalam proses boosing

3.6 Evaluasi Model

Evaluasi perlu dilakukan untuk menganalisis hasil kinerja dari model seberapa baik model dalam melakukan klasifikasi.

 ISSN:2302-8149
e-ISSN:2540-9719

3.6.1 Confusion Matrix

Proses pengujian dengan Confusion Matrix yang dilakukan setelah proses klasifikasi selesai dilakukan. Dengan pengujian ini dapat memperoleh nilai presisi, *recall* dan keakuratan hasil tes. Perhitungan nilai dapat dilakukan dengan :
Accuracy digunakan untuk menghitung rasio dari prediksi yang benar, baik positif maupun negatif dengan keseluruhan data yang ada [15]. Dengan perhitungan pada persamaan (10) [16] .

$$Accuracy : A = \frac{TN+TP}{TP+FP+TN+FN} \quad (10)$$

Precision atau presisi digunakan untuk menghitung berapa keakuratan dari prediksi yang dilakukan oleh model itu benar [15]. Dengan perhitungan pada persamaan (11) [16].

$$Precision : P = \frac{TP}{TP+FP} \quad (11)$$

Recall merupakan kemampuan *classifier* untuk menemukan semua kasus positif [15]. Dengan perhitungan pada persamaan (12) [16].

$$Recall: R = \frac{TP}{TP+FN} \quad (12)$$

Dimana [1] :
True Positive (TP) merepresentasikan data uji positif yang berhasil dimasukkan ke dalam kelas positif oleh model.
True Negative (TN) merepresentasikan data uji negatif yang berhasil dimasukkan ke dalam kelas negatif oleh model.
False Positive (FP) merepresentasikan data uji negatif yang keliru dimasukkan ke dalam kelas positif oleh model.
False Negative (FN) merepresentasikan data uji yang positif yang keliru dimasukkan ke dalam kelas negatif oleh model.

3.6.2 K-Fold Cross Validation

Validasi model dengan menggunakan K-Fold Cross Validation adalah performa model dievaluasi berdasarkan nilai rata-rata dari kinerja model pada semua *subset* validasi. Metode ini mengevaluasi performa model lebih akurat dengan memaksimalkan penggunaan data dimana membagi *dataset* menjadi k *subset* atau *fold* dengan ukuran yang sama. Pada setiap iterasi, k-1 *subset* digunakan sebagai *training set*, sementara *subset* yang tersisa digunakan sebagai *validation set*. Proses ini akan diulang sebanyak k kali sesuai dengan inputan k. Hasil akhir dari K-Fold Cross-Validation adalah rata-rata performa model pada semua iterasi [17].

4 Hasil dan Pembahasan

Berdasarkan tahapan penelitian yang telah dijelaskan. Dilakukan percobaan-percobaan untuk menemukan hasil berdasarkan data yang dimiliki dan model yang digunakan.

4.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini diperoleh dari situs Kaggle.com dengan nama "Dataset Predic Terkena Penyakit Paru-Paru" dengan format *file* csv yang memiliki 11 kolom dan 30000 baris data. Struktur dataset ini disajikan pada Tabel 2, yang memuat informasi mengenai nama kolom beserta dengan deskripsi sebagai gambaran karakteristik data yang digunakan.

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 2. Deskripsi data

No	Nama Kolom	Tipe Data	Deskripsi
1	No	Numerik	Nomor urut atau indeks data
2	Usia	Kategorikal	Kategori usia pasien
3	Jenis_Kelamin	Kategorikal	Jenis kelamin pasien
4	Merokok	Kategorikal	Status kebiasaan pasien dalam merokok
5	Bekerja	Kategorikal	Status bekerja pasien
6	Rumah_Tangga	Kategorikal	Status rumah tangga pasien
7	Aktivitas_Begadang	Kategorikal	Kebiasaan begadang pasien
8	Aktivitas_Olahraga	Kategorikal	Frekuensi olahraga pasien
9	Asuransi	Kategorikal	Kepemilikan asuransi kesehatan
10	Penyakit_Bawaan	Kategorikal	Adanya penyakit bawaan
11	Hasil	Kategorikal	Label hasil apakah terindikasi penyakit paru-paru

4.2 Pembersihan Data & Pra Pemrosesan Data

Setelah data terkumpul, maka perlu membersihkan dan mempersiapkan data agar dapat digunakan oleh algoritma *machine learning* dengan baik dan memaksimalkan kinerja model.

4.2.1 Penanganan Data Hilang

Proses ini merupakan tahap pembersihan data dengan menghapus data yang hilang atau kosong. Berdasarkan hasil eksplorasi data yang terlihat pada Gambar 2, tidak ditemukan adanya nilai kosong atau hilang pada dataset yang digunakan sehingga tidak diperlukan proses penghapusan lebih lanjut.

Gambar 2. Penanganan data hilang

4.2.2 Menghapus Data yang Duplikat

Proses ini merupakan tahap menghapus data yang duplikat. Berdasarkan hasil eksplorasi data, tidak ditemukan adanya data duplikat pada dataset yang digunakan sehingga tidak diperlukan proses penghapusan lebih lanjut.

 ISSN:2302-8149
e-ISSN:2540-9719

4.2.3 Pengkodean Label

Dalam data ini terdapat beberapa atribut dengan tipe data kategorikal yang perlu diubah ke dalam format numerik agar dapat diproses oleh *machine learning*, yang umumnya hanya menerima input berupa nilai numerik. Proses ini dilakukan menggunakan fungsi replace() dari *library* pandas. Tabel 3 menyajikan repesentasi awal dari data sebelum dilakukan proses *encoding*, dimana nilai-nilai masih berupa kategorikal.

Tabel 3. Data awal

No	Usia	Jenis_Kelamin	Merokok	…	Penyakit_Bawaan	Hasil
1	Tua	Pria	Pasif	…	Tidak	Ya
2	Tua	Pria	Aktif	…	Ada	Tidak
3	Muda	Pria	Aktif	…	Tidak	Tidak
…	…	…	…	…	…	…
29999	Muda	Wanita	Pasif	…	Ada	Tidak
30000	Tua	Wanita	Pasif	…	Tidak	Ya

Pada Tabel 4, nilai-nilai pada atribut kategorikal telah dikonversi ke dalam bentuk numerik menggunakan teknik *label encoding* sehingga data tersebut sudah berada dalam format yang sesuai dan dapat diproses oleh algoritma *machine learning.*

Tabel 4. Data setelah dilakukan proses encoding

No	Usia	Jenis_Kelamin	Merokok	…	Penyakit_Bawaan	Hasil
1	1	0	0	…	0	1
2	1	0	1	…	1	0
3	0	0	1	…	0	0
…	…	…	…	…	…	…
29999	0	1	0	…	1	0
30000	1	1	0	…	0	1

Pada tabel 5 menunjukkan deskripsi lengkap mengenai data setelah melalui proses *encoding*, dimana setiap nilai numerik yang ditampilkan pada kolom-kolom tersebut merepresentasikan kategori asli dari masing-masing atribut dataset.

Tabel 5. Deskripsi data setelah di encoding

No	Nama Kolom	Deskripsi
1	Usia	0: Muda
		1: Tua
2	Jenis_Kelamin	0: Pria
		1: Wanita
3	Merokok	0: Pasif
		1: Aktif
4	Bekerja	0: Tidak
		1: Ya
5	Rumah_Tangga	0: Tidak
		1: Ya
6	Aktivitas_Begadang	0: Tidak
		1: Ya
7	Aktivitas_Olahraga	0: Jarang
		1: Sering
8	Asuransi	0: Tidak
		1: Ada
9	Penyakit_Bawaan	0: Tidak
		1: Ada
10	Hasil	0: Tidak
		1: Ya

 ISSN:2302-8149
e-ISSN:2540-9719

4.2.4 Seleksi Data

Dalam tahapan seleksi data dilakukan penghapusan atribut yang tidak relevan dalam proses klasifikasi yaitu atribut “No” . Atribut ini hanya berfungsi sebagai penomoran dan tidak memiliki kontribusi terhadap proses klasifikasi.

4.3 Pembagian Data

Pada tahap pembagian data, dilakukan percobaan dengan menggunakan beberapa nilai *test size* dari total data. Tujuan dari penggunaan beberapa *test size* ini adalah untuk menguji kinerja model terhadap proporsi data latih dan data uji yang berbeda serta untuk mengetahui proporsi pembagian data yang menghasilkan performa terbaik.

4.3.1 Data Latih

Data latih yang digunakan dalam penelitian ini dibagi ke dalam empat skenario yaitu, 90%, 80%, 70%, dan 60% dari total 30000 data. Jumlah data latih pada masing-masing skenario tersebut adalah 27000 data untuk skenario 90%, 24000 data untuk skenario 80%, 21000 data untuk skenario 70%, dan 18000 data untuk skenario 60%.

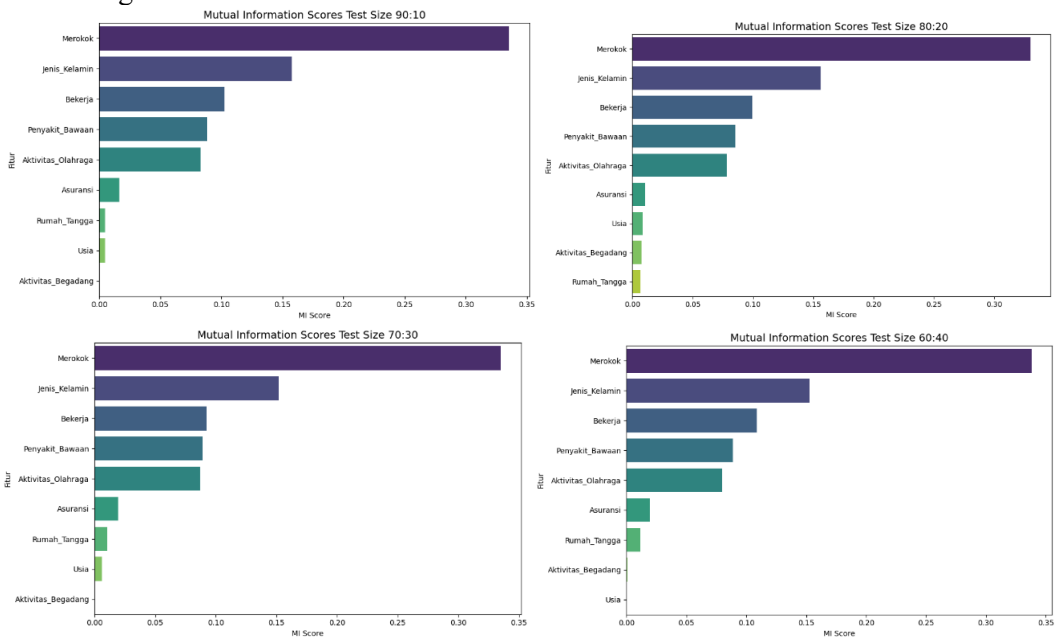
4.3.2 Data Uji

Data uji yang digunakan dalam penelitian ini dibagi ke dalam empat skenario yaitu, 10%, 20%, 30%, dan 40% dari total 30000 data. Jumlah data uji pada masing-masing skenario tersebut adalah 3000 data untuk skenario 10%, 6000 data untuk skenario 20%, 9000 data untuk skenario 30% dan 12000 data untuk skenario 40%.

4.4 Seleksi Fitur (Mutual Information)

Tahapan ini merupakan tahapan proses seleksi fitur dengan cara menghitung nilai Mutual Information untuk setiap atribut. Nilai Mutual Information dari setiap fitur diurutkan dari yang terbesar ke yang terkecil, fitur dengan nilai Mutual Information yang tinggi memiliki pengaruh yang lebih besar terhadap data.

Pada Gambar 3 menyajikan perbandingan hasil visualisasi nilai Mutual Information dari masing-masing fitur terhadap target pada berbagai skenario pembagian data, yaitu 90:10, 80:20, 70:30, dan 60:40. Visualisasi ini berbentuk diagram batang (*bar chart*) di mana setiap batang merepresentasikan satu fitur yang digunakan dalam model. Tinggi dari masing-masing batang menunjukkan besarnya nilai Mutual Information yang dimiliki fitur tersebut terhadap target. Semakin tinggi batangnya, maka semakin besar kontribusi informasi yang diberikan fitur tersebut dalam memprediksi target.



Gambar 3. Visualisasi nilai mutual information berdasarkan skenario pembagian data 90:10, 80:20, 70:30 dan 60:40

 ISSN:2302-8149
e-ISSN:2540-9719

Pada Tabel 6 menjelaskan hasil perhitungan dari nilai Mutual Information dari yang terbesar sampai terkecil. Nilai-nilai tersebut menunjukkan keterkaitan masing-masing fitur terhadap variabel target.

Tabel 6. Hasil nilai mutual information

Test Size	Fitur	Nilai Mutual Information
90:10	Merokok	0.335543
	Jenis_Kelamin	0.157484
	Bekerja	0.102564
	Penyakit_Bawaan	0.089739
	Aktivitas_Olahraga	0.082852
	Asuransi	0.018317
	Rumah_Tangga	0.004704
	Usia	0.004499
	Aktivitas_Begadang	0.000000
80:20	Merokok	0.330905
	Jenis_Kelamin	0.156350
	Bekerja	0.099778
	Penyakit_Bawaan	0.086092
	Aktivitas_Olahraga	0.078642
	Asuransi	0.012196
	Usia	0.008288
	Aktivitas_Begadang	0.007449
	Rumah_Tangga	0.006636
70:30	Merokok	0.334960
	Jenis_Kelamin	0.151711
	Bekerja	0.092437
	Penyakit_Bawaan	0.090074
	Aktivitas_Olahraga	0.087175
	Asuransi	0.021351
	Rumah_Tangga	0.010462
	Usia	0.005845
	Aktivitas_Begadang	0.000000
60:40	Merokok	0.339255
	Jenis_Kelamin	0.152737
	Bekerja	0.108888
	Penyakit_Bawaan	0.090641
	Aktivitas_Olahraga	0.079884
	Asuransi	0.021192
	Rumah_Tangga	0.011380
	Aktivitas_Begadang	0.000846
	Usia	0.000000

Untuk memilih fitur pada Mutual Information dengan menggunakan nilai mutual information score (MI Score) [5]. Pada penelitian ini menggunakan nilai MI-score > 0.01 sebagai *threshold* untuk mempertahankan fitur-fitur yang memiliki kontribusi lebih besar terhadap target, fitur dengan MI-score di bawah 0.01 akan dieleminasi. Berdasarkan kriteria tersebut, diperoleh hasil seleksi fitur seperti yang ditampilkan pada tabel 7.

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 7. Hasil seleksi fitur dengan mutual information

Test Size	Fitur
90:10	Merokok Jenis_Kelamin Bekerja Penyakit_Bawaan Aktivitas_Olahraga Asuransi
80:20	Merokok Jenis_Kelamin Bekerja Penyakit_Bawaan Aktivitas_Olahraga Asuransi
70:30	Merokok Jenis_Kelamin Bekerja Penyakit_Bawaan Aktivitas_Olahraga Asuransi Rumah_Tangga
60:40	Merokok Jenis_Kelamin Bekerja Penyakit_Bawaan Aktivitas_Olahraga Asuransi Rumah_Tangga

Berdasarkan hasil perhitungan MI-score dan terlihat pada tabel diatas, fitur yang secara konsisten memiliki kontribusi dengan informasi yang tinggi dan dipertahankan di semua skenario adalah Merokok, Jenis_Kelamin, Bekerja, Penyakit_Bawaan, Aktivitas_Olahraga, dan Asuransi. Sementara fitur seperti Aktivitas_Begadang dan Usia pada beberapa skenario menunjukkan skor MI yang sangat rendah sehingga dieliminasi dari model. Fitur Rumah_Tangga juga dieliminasi pada skenario 90:10 dan 80:20 karena memiliki nilai yang rendah.

4.5 Klasifikasi dengan XGBoost

Pada penelitian ini dilakukan proses klasifikasi menggunakan algoritma XGBoost dengan dua pendekatan, yaitu tanpa seleksi fitur dimana menggunakan seluruh fitur yang ada dan dengan seleksi fitur menggunakan Mutual Information. Tabel 8 menunjukkan hasil pemodelan klasifikasi dari pengujian dengan kedua pendekatan tersebut.

Tabel 8. Hasil pemodelan dengan mutual information dan tanpa mutual information

Test Size	Seleksi Fitur	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Komputasi (s)
90:10	Dengan Seleksi Fitur (6 Fitur)	88.63	88.64	88.64	88.63	612.62
	Tanpa Seleksi Fitur	93.83	94.44	93.92	93.82	671.20
80:20	Dengan Seleksi Fitur (6 Fitur)	88.85	88.85	88.87	88.85	535.33

 ISSN:2302-8149
e-ISSN:2540-9719

	Tanpa Seleksi Fitur	93.83	94.39	93.98	93.83	601.30
70:30	Dengan Seleksi Fitur (7 Fitur)	93.60	94.21	93.75	93.59	520.61
	Tanpa Seleksi Fitur	93.60	94.21	93.75	93.59	532.67
60:40	Dengan Seleksi Fitur (7 Fitur)	93.42	94.03	93.62	93.42	458.15
	Tanpa Seleksi Fitur	93.42	94.03	93.62	93.42	464.39

Hasil pemodelan menunjukkan bahwa akurasi model stabil pada skenario pembagian data 70:30 dan 60:40, baik ketika seluruh fitur digunakan maupun ketika hanya fitur terpilih yang digunakan. Namun, pada skenario 90:10 dan 80:20, penggunaan seleksi fitur menyebabkan penurunan akurasi yang menunjukkan bahwa pada proporsi data tertentu, proses seleksi fitur dapat menghilangkan informasi yang masih relevan[18].

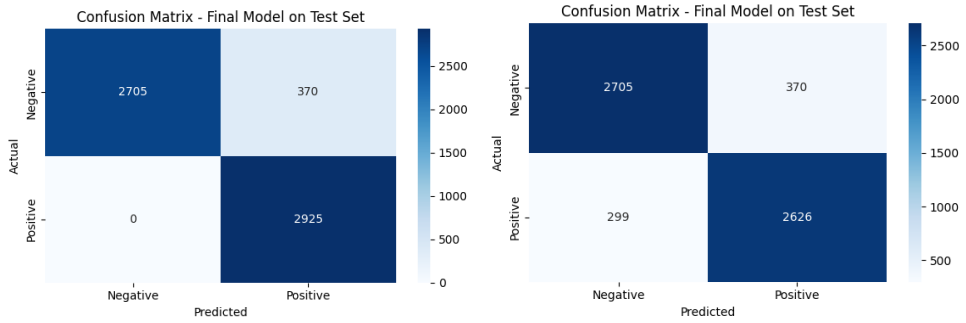
Meskipun proses seleksi fitur mengurangi jumlah fitur yang digunakan dalam pemodelan, fitur-fitur yang tersisa cukup mewakili informasi penting dalam dataset sehingga model masih dapat belajar secara efektif tanpa menurunkan performa model. Selain itu, penggunaan Mutual Information memberikan keuntungan dari segi efisiensi waktu komputasi. Dengan jumlah fitur yang lebih sedikit, proses pelatihan dan klasifikasi menjadi lebih cepat dibandingkan dengan model yang menggunakan seluruh fitur. Hal ini menunjukkan bahwa seleksi fitur dengan Mutual Information dapat meningkatkan kualitas data melalui penyederhanaan fitur tanpa mengurangi performa model secara signifikan.

4.6 Evaluasi Model

Untuk mengevaluasi performa model dalam melakukan klasifikasi yang dibangun menggunakan algoritma XGBoost dilakukan proses evaluasi model dengan beberapa pendekatan yaitu Confusion Matrix dan K-Fold Cross Validation. Kombinasi evaluasi dilakukan untuk menguji data secara keseluruhan bukan hanya pada *data test* saja.

4.6.1 Confusion Matrix

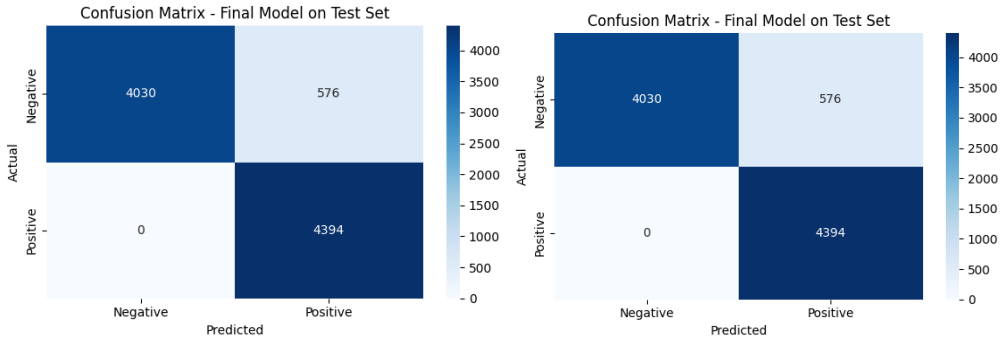
Untuk mengevaluasi performa model dalam melakukan klasifikasi, digunakan Confusion Matrix yang memberikan perbandingan antara nilai aktual dengan nilai prediksi. Hasil Confusion Matrix terdapat pada gambar 4 dan gambar 5.



Gambar 4. Hasil confuion matrix pada skenario 80:20 tanpa mutual information dan dengan mutual information

Gambar 4 menunjukkan hasil Confusion Matrix pada skenario 80:20 tanpa Mutual Information menunjukkan nilai *False Negative* yang 0. Hal ini menunjukkan bahwa model mampu mengklasifikasikan seluruh pasien yang benar-benar sakit (positif) secara tepat. Namun, ketika menggunakan Mutual Information FN meningkat menjadi 299.

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 5. Hasil confuion matrix pada skenario 70:30 tanpa mutual information dan dengan mutual information

Gambar 5 menunjukkan hasil Confusion Matrix pada skenario 70:30, di mana baik tanpa maupun dengan seleksi fitur Mutual Information, nilai FN tetap berada di angka 0

Berdasarkan hasil Confusion Matrix yang ditampilkan pada Gambar 4 dan Gambar 5, model menunjukkan performa yang sangat baik dalam mengenali data tanpa menggunakan seleksi fitur. Hal ini ditunjukkan dengan nilai *False Negative* yang didapatkan mencapai angka nol. Dalam diagnosis penyakit, nilai *False Negative* menjadi indikator krusial karena kesalahan model dalam mengklasifikasikan pasien yang sebenarnya positif (sakit) sebagai negatif (tidak sakit) dapat menyebabkan pasien tidak menerima perawatan yang diperlukan dengan tepat waktu. Oleh karena itu keberhasilan model dalam meminimalkan jumlah *False Negative* menunjukkan tingkat keandalan yang tinggi dalam proses klasifikasi. Hasil ini juga mewakili performa model pada pembagian data lainnya (90:10 dan 60:40). Namun, terdapat perbedaan pada hasil Confusion Matrix dengan seleksi fitur pada skenario 90:10 dan 80:20, di mana nilai FN meningkat mencapai angka 156 dan 299 karena proses seleksi fitur telah mengeliminasi fitur-fitur penting sehingga menurunkan model dalam mengklasifikasikan pasien.

4.6.2 K-Fold Cross Validation

Selain akurasi model yang didapat dari Confusion Matrix, penelitian ini juga menerapkan teknik K-Fold Cross Validation dengan nilai k = 10. Teknik ini membagi *dataset* menjadi sepuluh *subset* yang setara, kemudian proses petihan dan pengujian dilakukan secara bergantian sebanyak k kali atau sepuluh kali, dimana setiap satu *subset* digunakan sebagai data uji dan sembilan lainnya sebagai data latih. Dengan demikian, seluruh data digunakan sebagai data uji setidaknya satu kali. Hasil dari proses K-Fold Cross Validation ditunjukkan pada tabel 9.

Tabel 9. Hasil k-fold cross validation dengan mutual information dan tanpa mutual information

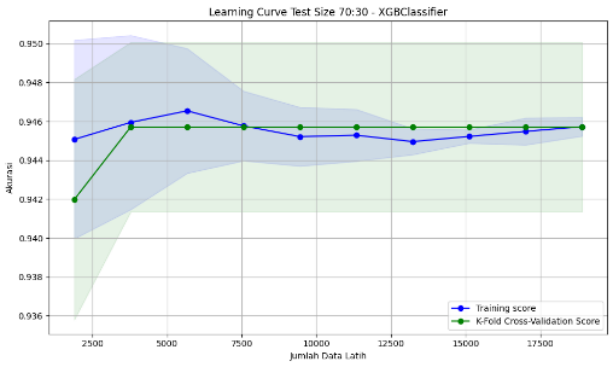
Test Size	Seleksi Fitur	Accuracy (%)
90:10	Dengan Seleksi Fitur (6 Fitur)	88.87
	Tanpa Seleksi Fitur	94.67
80:20	Dengan Seleksi Fitur (6 Fitur)	90.24
	Tanpa Seleksi Fitur	94.76
70:30	Dengan Seleksi Fitur (7 Fitur)	89.26

 ISSN:2302-8149
e-ISSN:2540-9719

	Tanpa Seleksi Fitur	94.63
60:40	Dengan Seleksi Fitur (7 Fitur)	90.88
	Tanpa Seleksi Fitur	94.59

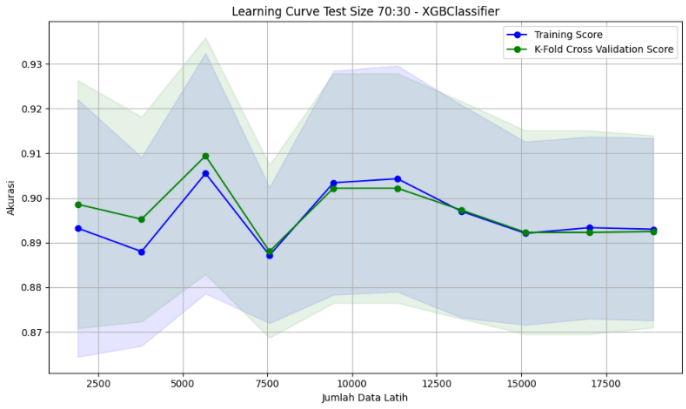
Dari tabel diatas menunjukkan hasil akurasi K-Fold Cross Validation pada semua skenario pembagian data memiliki akurasi yang lebih tinggi pada model tanpa seleksi fitur dibandingan dengan model yang menggunakan seleksi fitur. Namun selisih akurasi antara tanpa seleksi fitur dan dengan seleksi fitur tidak terlalu besar.

Setelah melalui evaluasi model, dilakukan perbandingan antara metrik perfoma pada data latih dan data validasi. Perbandingan ini dilakukan dengan menggunakan konsep dan visualisasi l*earning curve*. *Learning curve* merupakan alat yang digunakan untuk menganalisis kinerja generalisasi sebuah algoritma pembelajaran mesin terhadap jumlah data latih yang digunakan. Kurva ini menunjukkan bagaimaan skor validasi dan skor pelatihan berubah seiring bertambahnya jumlah sampel pelatihan [19]. Dengan membandingkannya maka dapat dilihat konsistensi performa model serta mendeteksi adanya indikasi *overfitting* atau *underfitting* selama proses pelatihan. Perbandingan performa ditunjukkan pada gambar 6 dan 7.



Gambar 6. Learning curve pada pembagian data 70:30 sebelum mutual infromation

Berdasarkan gambar 6 dengan skenario pembagian data 70:30 menggunakan model XGBoost Classifier, terlihat bahwa model menunjukkan performa yang stabil dan konsisten seiring bertambahnya jumlah data latih.



Gambar 7. Learning curve pada pembagian data 70:30 setelah mutual infromation

Berdasarkan gambar 7 setelah penerapan Mutual Information pada skenario pembagian data 70:30, kurva pelatihan dan validasi menunjukkan sedikit fluktuasi pada data 5000 hingga 7500 tetapi performa mulai stabil seiring bertambahnya jumlah data pelatihan. Nilai akurasi berkisar antara 88%

 ISSN:2302-8149
e-ISSN:2540-9719

hingga 90%, yang mencerminkan kinerja model yang konsisten dalam berbagai skenario ukuran data latih.

Berdasarkan gambar 6 dan 7, model XGBoost menunjukkan performa yang lebih optimal sebelum penerapan Mutual Information. Setelah diterapkan seleksi fitur, model tetap stabil namun performa model sedikit menurun, yang mengindikasikan kemungkinan tereliminasinya beberapa fitur penting. Meskipun demikian, akurasi model tetap berada pada tingkat yang baik, sehingga seleksi fitur tetap memberikan efisiensi tanpa mengorbankan performa secara drastis. Pola serupa juga tampak pada skenario pembagian data lainnya, seperti 90:10, 80:20 dan 60:40. Secara umum, Model XGBoost menunjukkan performa yang stabil, ditandai dengan semakin kecilnya selisih antara *training score* dan K-Fold Cross-Validation *score* seiring bertambahnya jumlah data latih. Hal ini mengindikasikan bahwa model tidak mengalami *overfitting* maupun *underfitting*.

5 Kesimpulan

Penelitian ini bertujuan untuk mengevaluasi pengaruh seleksi fitur menggunakan Mutual Information terhadap performa klasifikasi kanker paru-paru menggunakan algoritma XGBoost. Hasil menunjukkan bahwa akurasi model tanpa seleksi fitur berada di 93.42% hingga 93.83% pada semua skenario. Setelah penerapan seleksi fitur menggunakan Mutual Information akurasi tetap stabil di 93.42% hingga 93.60% dimana Mutual Information dapat mengurangi jumlah fitur tanpa menurunkan akurasi pada skenario pembagian data 70:30 dan 60:40. Hal ini menunjukkan bahwa Mutual Information mampu mengurangi jumlah fitur tanpa menurunkan akurasi pada skenario tersebut. Namun demikian, rata-rata akurasi pada K-Fold Cross Validation mengalami penurunan dari 94.63% dan 94.59% menjadi 89.26% dan 90.88%. Pada skenario pembagian data 90:10 dan 80:20, seleksi fitur menyebabkan peningkatan nilai *False Negative* yang berdampak pada penurunan akurasi testing menjadi 88.63% dan 88.85% serta penurunan akurasi rata-rata pada K-Fold Cross Validation yang sebelumnya 94.67% dan 94.76% menjadi 88.87% dan 90.24%. Di sisi lain, penerapan Mutual Information memberikan keuntungan dalam efisiensi komputasi dengan proses pelatihan dan klasifikasi yang lebih cepat karena jumlah fitur yang lebih sedikit. Dengan demikian, penerapan seleksi fitur dengan menggunakan Mutual Information dapat diterapkan secara efektif untuk meningkatkan efisiensi tanpa menurunkan kemampuan klasifikasi model, namun perlu dilakukan dengan pertimbangan dan analisis karakteristik data agar tidak menurunkan kemampuan klasifikasi model.

Penelitian ini memiliki batasan pada penggunaan satu metode seleksi fitur, yaitu Mutual Infromation, tanpa membandingkannya dengan metode seleksi fitur lainnya. Penelitian selanjutnya disarankan untuk melakukan pengujian pada *dataset* lain yang lebih besar dan memiliki jumlah fitur yang lebih banyak, membandingkan beberapa algoritma klasifikasi lainnya agar dapat mengetahui apakah penerapan Mutual Information memberikan dampak yang berbeda terhadap performa model lain dan melakukan eksplorasi serta perbadingan dengan berbagai metode seleksi fitur lainnya.

Referensi

[1] R. D. Marzuq, S. A. Wicaksono, and N. Y. Setiawan, "Prediksi Kanker Paru-Paru menggunakan *Algoritme Random Forest Decision Tree*," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, Vol. 7, No. 7, pp. 3448–3456, 2023.

[2] J. Al-Tawalbeh, B. Alshargawi, H. Alquran, W. Al-Azzawi, W. A. Mustafa, and A. Alkhayyat, "*Classification of Lung Cancer by Using Machine Learning Algorithms*," *IICETA 2022 - 5th Int. Conf. Eng. Technol. its Appl.*, pp. 528–531, 2022, doi: 10.1109/IICETA54559.2022.9888332.

[3] R. Bisht, N. Thapliyal, R. Bisht, and G. Wadhwa, "*Lung Cancer Detection using Ensemble Learning and Machine Learning Algorithms*," *2nd Int. Conf. Artif. Intell. Mach. Learn. Appl. Healthc. Internet Things, AIMLA 2024*, pp. 1–6, 2024, doi: 10.1109/AIMLA59606.2024.10531596.

[4] S. Bharathy, R. Pavithra, and B. Akshaya, "*Lung Cancer Detection using Machine Learning*," *Proc. - Int. Conf. Appl. Artif. Intell. Comput. ICAAIC 2022*, No. Icaaic, pp. 539–543, 2022, doi: 10.1109/ICAAIC53929.2022.9793061.

[5] A. Rahmadeyan and M. Mustakim, "Seleksi Fitur pada *Supervised Learning*: Klasifikasi

 ISSN:2302-8149
e-ISSN:2540-9719

Prestasi Belajar Mahasiswa Saat dan Pasca Pandemi COVID-19," *J. Nas. Teknol. dan Sist. Inf.*, Vol. 9, No. 1, pp. 21–32, 2023, doi: 10.25077/teknosi.v9i1.2023.21-32.

[6] H. I. H. Yusri, A. A. Ab Rahim, S. L. M. Hassan, I. S. A. Halim, and N. E. Abdullah, "*Water Quality Classification using SVM And XGBoost Method*," *2022 IEEE 13th Control Syst. Grad. Res. Colloquium, ICSGRC 2022 - Conf. Proc.*, No. July, pp. 231–236, 2022, doi: 10.1109/ICSGRC55096.2022.9845143.

[7] T. Gori, A. Sunyoto, and H. Al Fatta, "*Preprocessing* Data dan Klasifikasi untuk Prediksi Kinerja Akademik Siswa," *J. Teknol. Inf. dan Ilmu Komput.*, Vol. 11, No. 1, pp. 215–224, 2024, doi: 10.25126/jtiik.20241118074.

[8] S. Gadge and A. Karande, "*Study of Different Types of Evaluation Methods in Classification and Regression*," *2022 IEEE Reg. 10 Symp. TENSYMP 2022*, pp. 1–5, 2022, doi: 10.1109/TENSYMP54529.2022.9864426.

[9] I. W. Dharmana, I. G. A. Gunadi, and L. J. E. Dewi, "Deteksi Transaksi *Fraud* Kartu Kredit menggunakan *Oversampling* ADASYN dan Seleksi Fitur SVM-RFECV," *J. Teknol. Inf. dan Ilmu Komput.*, Vol. 11, No. 1, pp. 125–134, 2024, doi: 10.25126/jtiik.20241117640.

[10] W. Mostert and K. M. Malan, "*Comparative Analysis*," pp. 1–16, 2021.

[11] B. Yao, C. Li, and Y. Chen, "*Supervised Feature Selection based on Sparse Representation and Mutual Information*," *Proc. 2023 IEEE 5th Int. Conf. Civ. Aviat. Saf. Inf. Technol. ICCASIT 2023*, pp. 1354–1358, 2023, doi: 10.1109/ICCASIT58768.2023.10351667.

[12] A. Nageswari, U. Jyothi, G. Divya, T. Ammannamma, and V. Usha, "*Water Quality Classification using XGBoost method*," *ICCCMLA 2024 - 6th Int. Conf. Cybern. Cogn. Mach. Learn. Appl.*, pp. 302–306, 2024, doi: 10.1109/ICCCMLA63077.2024.10871422.

[13] V. S. Desdhanty and Z. Rustam, "*Liver Cancer Classification using Random Forest and Extreme Gradient Boosting (XGBoost) with Genetic Algorithm as Feature Selection*," *2021 Int. Conf. Decis. Aid SCI. Appl. DASA 2021*, pp. 716–719, 2021, doi: 10.1109/DASA53625.2021.9682311.

[14] V. Jagadeesh and P. Sivakumar, "*Enhanced Pipeline Safety: Cloud-based Leak Prediction Using XGBoost*," *Proc. - 2024 IEEE 16th Int. Conf. Commun. Syst. Netw. Technol. CICN 2024*, pp. 1087–1091, 2024, doi: 10.1109/CICN63059.2024.10847573.

[15] E. Helmud, E. Helmud, F. Fitriyani, and P. Romadiana, "*Classification Comparison Performance of Supervised Machine Learning Random Forest and Decision Tree Algorithms using Confusion Matrix*," *J. Sisfokom (Sistem Inf. dan Komputer)*, Vol. 13, No. 1, pp. 92–97, 2024, doi: 10.32736/sisfokom.v13i1.1985.

[16] R. Prasetya, "*Data Mining Application on Weather Prediction using Classification Tree, Naïve Bayes and K-Nearest Neighbor Algorithm with Model Testing of Supervised Learning Probabilistic Brier Score, Confusion Matrix and ROC*," *Jaict*, Vol. 4, No. 2, p. 25, 2020, doi: 10.32497/jaict.v4i2.1690.

[17] I. K. Nti, O. Nyarko-Boateng, and J. Aning, "*Performance of Machine Learning Algorithms with Different K Values in K-fold CrossValidation*," *Int. J. Inf. Technol. Comput. Sci.*, Vol. 13, No. 6, pp. 61–71, 2021, doi: 10.5815/ijitcs.2021.06.05.

[18] N. C. Ramadhan, H. H. H, T. Rohana, and A. M. Siregar, "Optimasi *Algoritma Machine Learning* menggunakan Seleksi Fitur *Xgboost* untuk Klasifikasi Kanker Payudara," *TIN Terap. Inform. Nusant.*, Vol. 5, No. 2, pp. 162–171, 2024, doi: 10.47065/tin.v5i2.5408.

[19] C. Giola, P. Danti, and S. Magnani, "*Learning Curves: A Novel Approach for Robustness Improvement of Load Forecasting* †," *Eng. Proc.*, Vol. 5, No. 1, 2021, doi: 10.3390/engproc2021005038.