

# Application of Machine Learning Algorithm for Osteoporosis Disease Prediction System

Rajendra Artanto Wiryawan Sujana <sup>1\*</sup>, I Made Artha Agastya <sup>2\*</sup>

\* Informatika, Universitas Amikom Yogyakarta

[rantowiryana@amikom.ac.id](mailto:rantowiryana@amikom.ac.id) <sup>1</sup>, [artha.agastya@amikom.ac.id](mailto:artha.agastya@amikom.ac.id) <sup>2</sup>

## Article Info

### Article history:

Received 2024-09-06

Revised 2024-10-09

Accepted 2024-10-14

### Keyword:

*Gradient Boosting,  
Machine Learning Algorithms,  
Osteoporosis,  
Random Forest,  
Support Vector Machine.*

## ABSTRACT

Osteoporosis is a condition characterized by decreased bone density, leading to fragile and easily fractured bones. This disease is a significant concern as it can cause disability, fractures, and death, particularly in the elderly population. Early detection of osteoporosis is crucial to prevent disease progression through timely interventions. This study aims to develop a machine learning-based prediction system capable of detecting osteoporosis using three different algorithms, Random Forest, Support Vector Machine (SVM), and Gradient Boosting. The study involves analyzing and comparing the performance of these algorithms based on evaluation metrics such as Confusion Matrix, Classification Report, AUC-ROC, and K-Fold Cross-Validation. The data used is processed in two formats, namely ordinal and one-hot encoding, to assess the impact of encoding techniques on model performance. The results show that the Gradient Boosting algorithm performs the best on both types of data, with the highest Accuracy of 91.07% on the one-hot encoded data. Meanwhile, SVM and Random Forest also demonstrate competitive performance but with slightly lower results. This study concludes that Gradient Boosting is the most effective algorithm for osteoporosis prediction in this research. These findings can serve as a foundation for further development in the early detection of osteoporosis and support more effective and efficient prevention and treatment efforts.



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

## I. PENDAHULUAN

Osteoporosis merupakan kondisi di mana berkurangnya kepadatan tulang, sehingga mengakibatkan tulang menjadi rapuh dan rentan terhadap patah tulang [1]. Osteoporosis disebabkan oleh ketidakseimbangan dalam proses metabolisme tulang, di mana penyerapan mineral seperti kalsium menjadi tidak optimal [1]. Faktor risiko osteoporosis mencakup usia, genetika, dan lingkungan. Usia merupakan faktor risiko signifikan untuk osteoporosis, dengan peningkatan risiko sebesar 1,4 hingga 1,8 kali untuk setiap dekade penambahan usia. Dari segi genetika, seseorang yang berasal dari ras Kaukasia, Polinesia, dan Negroid lebih rentan terhadap osteoporosis. Beberapa faktor lingkungan yang berperan meliputi defisiensi kalsium, minimnya aktivitas fisik, merokok, konsumsi alkohol, penggunaan obat-obatan seperti kortikosteroid, antikonvulsan, heparin, dan

siklosporin, serta peningkatan risiko jatuh akibat gangguan penglihatan dan keseimbangan [2].

Osteoporosis dipilih sebagai topik penelitian karena penyakit ini memerlukan perhatian serius, mengingat dampaknya yang bisa menyebabkan kecacatan, patah tulang, hingga kematian. Selain itu, pengobatan osteoporosis membutuhkan waktu yang lama, biaya yang sangat tinggi, dan menyebabkan penderitaan jangka panjang [3]. Menurut *World Health Organization* (WHO), sekitar 200 juta orang di dunia menderita osteoporosis. Menurut laporan dari *International Osteoporosis Foundation* (IOF), di Indonesia, satu dari empat perempuan berusia 50 hingga 80 tahun berisiko mengalami osteoporosis [3]. Rasio kejadian osteoporosis antara perempuan dan laki-laki di Indonesia adalah 4:1, dengan prevalensi yang lebih tinggi pada wanita setelah menopause. Menurut data dari Perhimpunan Osteoporosis Indonesia tahun 2007, prevalensi osteoporosis

pada penduduk di atas usia 50 tahun mencapai 32,3 persen pada perempuan dan 28,8 persen pada laki-laki [2].

Melakukan prediksi terhadap kemungkinan seseorang terkena osteoporosis memiliki peran krusial dalam upaya pencegahan dan penanganan penyakit ini secara efektif. Dengan kemampuan untuk melakukan prediksi dini, intervensi medis dapat dilakukan lebih cepat dan lebih efektif untuk mencegah atau mengurangi keparahan penyakit. Prediksi dini memungkinkan identifikasi individu yang berisiko tinggi, sehingga mereka dapat menjalani pemeriksaan dan perawatan yang lebih intensif [4]. Selain itu, dengan mengetahui risiko sejak awal, individu dapat mengadopsi gaya hidup yang lebih sehat. Oleh karena itu, pengembangan sistem prediksi berbasis algoritma *machine learning* untuk osteoporosis menjadi sangat penting dalam mendukung upaya pencegahan dan penanganan penyakit ini secara lebih efektif dan efisien.

Pada penelitian tahun 2021 yang dilakukan oleh Purbolaksono Mahendra Dwifabri dan rekan-rekannya, algoritma *Support Vector Machine* (SVM) dan *Modified Balanced Random Forest* (MBRF) diterapkan untuk analisis perbandingan algoritma dalam deteksi pasien penyakit diabetes. Dalam penelitian tersebut, data yang digunakan dilakukan proses *handling missing value* dengan teknik imputasi dan proses normalisasi data, selanjutnya dilakukan uji validasi pada kedua model algoritma menggunakan teknik *K-Fold Cross-Validation* dengan nilai  $K = 10$ . Kedua model memperoleh hasil akurasi maksimum sangat baik dengan nilai 91,48% untuk model SVM dan 97,80% untuk model MBRF. Hasil penelitian menunjukkan bahwa penggunaan algoritma SVM dan MBRF memiliki performa sangat baik dalam mengklasifikasikan data pasien penyakit diabetes [5].

Pada penelitian sebelumnya yang dilakukan oleh Patasik Eva Sapan dan Yulianto Sri pada tahun 2023, algoritma *gradient boost* dan *random forest* diterapkan untuk klasifikasi bahasa daerah. Dalam penelitian tersebut, data yang digunakan memuat kumpulan bahasa sehari-hari atau percakapan yang dilakukan yaitu bahasa yang diambil dari bahasa daerah Jawa, Nias, dan Toraja dengan total jumlah data sebanyak 9000 dan masing-masing bahasa memiliki data berjumlah 3000. Hasil penelitian menunjukkan bahwa kedua metode yang digunakan yaitu *Gradient Boosting* dan *Random Forest* memiliki performa cukup bagus dalam melakukan klasifikasi bahasa dengan tingkat akurasi 88.5% untuk model *Gradient Boost* dan 87.94% untuk model *Random Forest*. Dari hasil tersebut, dapat disimpulkan bahwa metode *Gradient Boost* memiliki performa yang lebih baik dalam melakukan klasifikasi bahasa [6].

Pada penelitian yang dilakukan oleh Jajang Jaya Purnama, dkk, melakukan klasifikasi untuk mengidentifikasi potensi risiko kesehatan ibu hamil menggunakan lima algoritma *machine learning* dan juga menerapkan perbandingan kelima algoritma yang menggunakan *hyperparameter tuning* dan tidak. Hasil penelitian menunjukkan bahwa kelima model prediksi yang tidak ditambah dengan metode *hyperparameter tuning* memiliki hasil akurasi yang tinggi, dimana skor

akurasi tertinggi diperoleh dengan algoritma *Random Forest* dengan skor 82,15% [7].

Berdasarkan uraian diatas, penelitian yang akan dilakukan untuk deteksi dini penyakit osteoporosis menggunakan metode analisis komparatif untuk membandingkan kinerja tiga algoritma *machine learning* yang berbeda, yaitu *Random Forest*, *Support Vector Machine*, dan *Gradient Boosting* dalam memprediksi penyakit osteoporosis. Ketiga algoritma tersebut dipilih dalam penelitian ini karena masing-masing algoritma memiliki keunggulan masing-masing. Algoritma *Gradient Boosting* dipilih karena kemampuannya yang sangat baik dalam meningkatkan akurasi prediksi dan menurunkan *error* [8]. *Random Forest* dipilih karena kemampuannya untuk menangani banyak fitur serta memberikan akurasi yang stabil dalam klasifikasi [9]. Di sisi lain, *Support Vector Machine* (SVM) dipilih karena kemampuannya dalam mengolah data *non-linear* dengan efektif [10]. Pemilihan algoritma-algoritma ini didasarkan pada karakteristik data klinis yang memiliki banyak fitur dan variasi dalam distribusi risiko osteoporosis.

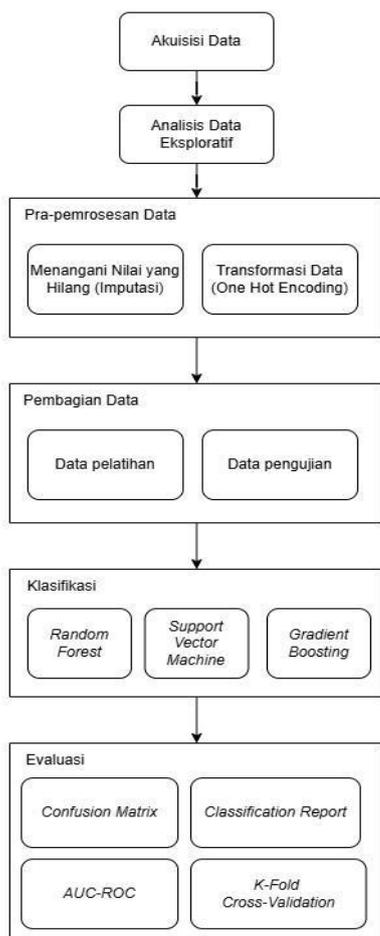
Penelitian ini bertujuan untuk mengevaluasi dan mengidentifikasi algoritma *machine learning* yang paling optimal dalam mendeteksi risiko osteoporosis, guna mendukung tenaga medis dalam membuat keputusan klinis yang lebih cepat, tepat, dan akurat. Dengan menggunakan berbagai metrik evaluasi seperti *Confusion Matrix*, *Classification Report*, *AUC-ROC*, dan *K-Fold Cross-Validation*, penelitian ini diharapkan dapat mengungkap algoritma terbaik serta faktor-faktor yang mempengaruhi prediksi tersebut.

## II. METODE

Penelitian ini menggunakan tiga algoritma *machine learning*, diantaranya adalah algoritma *Random Forest*, *Support Vector Machine* (SVM), dan *Gradient Boosting*. Dataset yang digunakan diperoleh melalui Kaggle yang berupa tabular CSV. Software yang digunakan adalah Google Colaboratory. Berikut merupakan alur proses penelitian yang akan dilakukan pada Gambar 1.

### A. Akuisisi Data

Pada tahap akuisisi data, proses dimulai dengan mengunduh dataset yang diperoleh dari situs Kaggle. Dataset dengan nama *osteoporosis.csv* berisi data pasien osteoporosis dengan total 1958 sampel dan 15 fitur yang digunakan dalam klasifikasi terkait informasi klinis dan demografis ini diunduh dalam format (.csv). Dataset ini disimpan di Google Drive dan kemudian diimpor ke dalam Google Colaboratory untuk digunakan dalam penelitian. Rincian lengkap mengenai fitur dan distribusi data dapat dilihat pada Tabel I dan II.



Gambar 1. Alur penelitian

TABEL I  
DESKRIPSI DATA NUMERIKAL

Variabel	Deskripsi	Mean	Range
Usia	Usia Individu dalam tahun	39.1011	18-90

TABEL II  
DESKRIPSI DATA KATEGORIKAL

Variabel	Deskripsi	Modus	Range
Jenis Kelamin	Jenis kelamin individu. "Male" atau "Female".	Laki-laki	Laki-laki dan Perempuan
Perubahan Hormonal	Menunjukkan apakah individu telah mengalami perubahan hormonal, terutama terkait menopause.	Normal	Normal dan Pasca-menopause
Riwayat Keluarga	Menunjukkan apakah ada riwayat keluarga osteoporosis atau patah tulang.	Tidak	Ya dan Tidak

Ras/Etnis	Ras atau etnis individu.	Afrika-Amerika	Kaukasia, Afrika-Amerika, dan Asia
Berat Badan	Status berat badan individu.	Normal	Normal dan Kurus
Asupan Kalsium	Tingkat asupan kalsium dalam diet individu.	Rendah	Kurang dan Tercukupi
Asupan Vitamin D	Tingkat asupan vitamin D dalam diet individu.	Cukup	Kurang dan Tercukupi
Aktivitas Fisik	Tingkat aktivitas fisik individu.	Aktif	Kurang Gerak dan Aktif
Merokok	Menunjukkan apakah individu adalah perokok.	Ya	Ya dan Tidak
Konsumsi Alkohol	Tingkat konsumsi alkohol oleh individu.	Tidak	Ya dan Tidak
Kondisi Medis	Kondisi medis yang ada pada individu.	Hipertiroidisme	Arthritis Reumatoid dan Hipertiroidisme
Konsumsi Obat-Obatan	Obat-obatan yang sedang dikonsumsi.	Tidak ada	Kortikosteroid dan Tidak ada
Riwayat Patah Tulang	Mengalami patah tulang sebelumnya.	Ya	Ya dan Tidak
Osteoporosis	Menunjukkan ada atau tidaknya osteoporosis.	Tidak	Ya dan Tidak

Dari hasil analisis, distribusi kelas pada kolom target "Osteoporosis", ditemukan bahwa dataset ini memiliki distribusi kelas yang seimbang, dengan masing-masing kelas (positif dan negatif osteoporosis) menyumbang 50% dari total data. Keseimbangan data ini mengindikasikan bahwa model machine learning yang akan diterapkan tidak akan mengalami bias terhadap salah satu kelas tertentu. Distribusi kelas osteoporosis dapat dilihat pada Tabel III.

TABEL III  
DISTRIBUSI KELAS OSTEOPOROSIS

Kategori	Persentase
Osteoporosis	50%
Non-osteoporosis	50%

B. Analisis Data Eksploratif

Proses Analisis Data Eksploratif (ADE) bertujuan untuk memahami karakteristik dan struktur dataset yang telah dikumpulkan. Pada tahap ini, teknik *boxplot* dan *cross-tabulation* digunakan untuk menganalisis data. Analisis ini membantu mengidentifikasi pola, tren, anomali, dan hubungan antar variabel dalam dataset. Selain itu, ADE juga digunakan untuk menentukan bagaimana cara terbaik mempersiapkan data untuk model *machine learning*,

termasuk menangani *missing values* dan *outliers* serta melakukan transformasi data jika diperlukan [11].

### C. Pra-pemrosesan Data

Tahap selanjutnya adalah menghapus nilai-nilai yang hilang atau anomali dari dataset, yang dapat berdampak pada kinerja model. Metode yang digunakan dalam menangani nilai yang hilang yaitu menggunakan metode imputasi. Metode imputasi adalah metode mengisi nilai yang hilang pada kolom dalam sebuah *dataframe* menggunakan nilai modus atau nilai yang paling sering muncul.

Diterapkan juga teknik *One-Hot Encoding* (OHE) pada dataset. *One-Hot Encoding* adalah salah satu teknik yang digunakan untuk mengonversi atau men-transformasi data kategorik menjadi bentuk numerik. Teknik ini bekerja dengan membuat *array* satu dimensi yang memiliki panjang sesuai dengan jumlah fitur, di mana setiap elemen *array* diisi dengan nilai biner 0 atau 1 [12].

Metode ini membuat representasi data kategorik menjadi lebih detail. Karena algoritma *machine learning* tidak dapat mengolah data kategorik secara langsung, data tersebut perlu diubah menjadi format numerik dengan nilai 0 dan 1. Setiap nilai dalam kolom kemudian diubah menjadi kolom baru yang diisi dengan angka 0 dan 1 berdasarkan fitur kategorik yang ada [12].

### D. Pembagian Data

Setelah data selesai diproses, langkah berikutnya adalah membagi data menjadi dua bagian, yaitu data pelatihan (*train*) sebesar 80% dan data pengujian (*test*) sebesar 20% dari total data. Pembagian ini penting karena memastikan model dapat dilatih dengan cukup data sekaligus menyediakan data yang belum pernah dilihat model untuk mengevaluasi kinerjanya secara objektif. Proses ini memastikan bahwa model yang dihasilkan tidak hanya sesuai dengan data latih, tetapi juga memiliki kemampuan generalisasi yang baik saat diterapkan pada data baru.

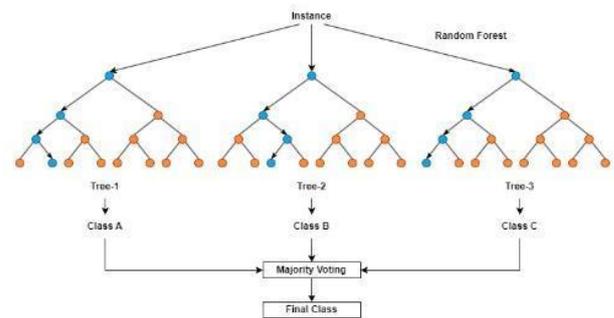
### E. Klasifikasi

Proses klasifikasi untuk prediksi penyakit osteoporosis dilakukan dengan menerapkan tiga algoritma *machine learning*, yaitu *Random Forest*, *Support Vector Machine* (SVM), dan *Gradient Boosting*. Masing-masing algoritma dipilih karena memiliki keunggulan tersendiri dalam menangani data yang kompleks dan bervariasi, serta mampu menghasilkan model prediksi yang akurat. Setelah pelatihan, kinerja masing-masing algoritma akan dievaluasi dan dibandingkan untuk menentukan algoritma yang paling efektif dalam memprediksi risiko osteoporosis.

#### 1) *Random Forest*

*Random Forest* (RF) merupakan algoritma terdiri dari serangkaian pohon keputusan [6]. Semakin banyak pohon yang digunakan, semakin tinggi akurasi yang dapat dicapai. RF menggunakan C4.5 atau J48 sebagai pengklasifikasi. Pada tahun 2001, Breiman memperkenalkan RF, yang mengintegrasikan metode *Bagging* dengan pemilihan fitur

secara acak untuk setiap pohon keputusan [9]. Berikut ini adalah gambaran yang digunakan pada metode *Random Forest*, seperti dalam Gambar 2.



Gambar 2. *Random Forest*

#### 2) *Support Vector Machine*

*Support Vector Machine* (SVM) adalah metode *supervised machine learning* yang dapat digunakan untuk prediksi dan klasifikasi. Metode ini dapat digunakan untuk data *linear* dan *non-linear* [5]. Model SVM mencoba menemukan *hyperplane* terbaik dengan memaksimalkan jarak marginal [5]. Metode ini dapat digunakan sebagai metode kernel untuk menemukan garis terbaik untuk pembagian sampel [13]. Persamaan (1) merupakan gambaran yang sesuai untuk model SVM, yang merupakan metode klasifikasi atau *Machine Learning*.

$$y(x_i) = w^T x_i + w_0 \quad (1)$$

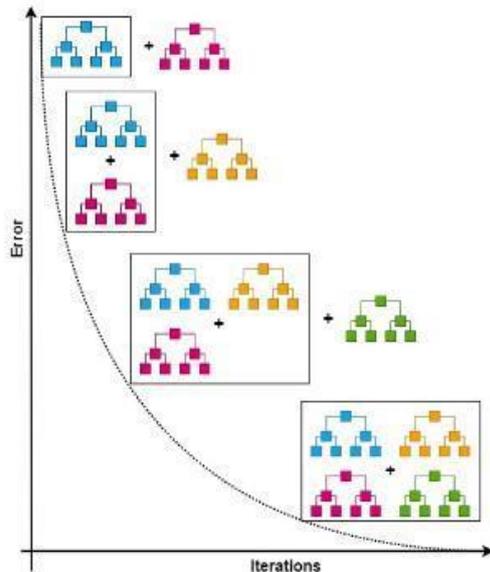
Dengan  $y(x_i)$  berfungsi sebagai prediksi dari  $t_i$ ,  $w$  adalah *vector weight* (parameter model),  $x_i$  adalah *variable data*, dan  $w_0$  adalah bias, dan untuk mengklasifikasi data dalam ruang tempatnya berada, model SVM bertujuan untuk memprediksi *hyperplane* dalam dimensi  $m$  [13]. Sub-ruang dari bidang dua dimensi biasanya disebut sebagai *hyperplane*.

#### 3) *Gradient Boosting*

*Gradient Boosting* adalah salah satu metode *Machine Learning* yang berfokus pada peningkatan kinerja model melalui pendekatan *boosting* [6]. Algoritma ini menggabungkan beberapa model kecil atau pembelajar lemah menjadi satu model yang lebih kuat dan akurat dalam memprediksi data [9]. *Gradient Boosting* beroperasi dengan mengukur kesalahan dari model sebelumnya dan menggunakan informasi tersebut untuk meningkatkan kinerja model berikutnya. Setiap model baru ditambahkan untuk mengurangi fungsi kerugian secara bertahap, mengikuti gradien dari fungsi kerugian keseluruhan. Visualisasi algoritma *Gradient Boosting* dapat dilihat pada Gambar 3.

*Gradient Boosting* terkenal akan kemampuannya menangani data yang kompleks dan memberikan prediksi yang akurat dengan menggabungkan kekuatan beberapa model lemah. *Gradient Boosting* fleksibel dan dapat digunakan untuk berbagai jenis tugas, baik regresi maupun klasifikasi. Namun, proses trainingnya cenderung lambat dan membutuhkan sumber daya komputasi yang tinggi, terutama pada dataset besar. Model ini juga rentan terhadap *overfitting*

jika tidak dikonfigurasi dengan baik, sehingga pemilihan parameter yang hati-hati dan validasi yang tepat sangat diperlukan untuk mencapai kinerja optimal [14].



Gambar 3. Gradient Boosting

## F. Evaluasi

Tahap evaluasi melibatkan beberapa langkah penting untuk menilai performa model yang telah dibangun. Beberapa metode yang digunakan untuk menilai performa model yang telah dibangun diantaranya, *Confusion Matrix*, *Classification Report*, dan *K-Fold Cross-Validation*.

### 1) Confusion Matrix

*Confusion Matrix* adalah tabel yang menunjukkan jumlah data uji yang diprediksi dengan benar dan salah oleh model klasifikasi yang digunakan. Tabel ini sangat penting untuk menentukan model klasifikasi dengan kinerja terbaik. Matriks konfusi berbentuk 2x2 dan merepresentasikan hasil klasifikasi biner dalam suatu kumpulan data [13].

Matriks ini memiliki empat kategori utama, yaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Setiap kategori menggambarkan hasil prediksi berdasarkan nilai sebenarnya dari data. *True Positive* (TP) terjadi ketika model memprediksi positif dan prediksinya sesuai dengan kenyataan. *True Negative* (TN) muncul saat model memprediksi negatif dan prediksi itu tepat. *False Positive* (FP) terjadi ketika model memprediksi positif, tetapi kenyataannya salah. Sedangkan *False Negative* (FN) terjadi ketika model memprediksi negatif, namun kenyataannya bertentangan [15]. Visualisasi mengenai analisis *confusion matrix* dapat dilihat pada Gambar 4.

Nilai Aktual Positif (1)	TN	FP
	FN	TP
	Negatif (0)	Positif (1)
	Nilai Prediksi	

Gambar 4. Confusion Matrix

### 2) AUC-ROC

AUC-ROC adalah metode evaluasi yang digunakan untuk mengukur kinerja model klasifikasi, khususnya untuk model biner. Kurva ROC menggambarkan hubungan antara *True Positive Rate* (TPR) di sumbu Y dan *False Positive Rate* (FPR) di sumbu X. *Area Under the ROC Curve* (AUC) menunjukkan seberapa baik model dapat membedakan antara kelas-kelas, dengan nilai AUC antara 0 hingga 1 [16]. Semakin tinggi AUC, semakin baik kinerja model dalam mengklasifikasikan data. Nilai AUC di atas 0,90 dianggap luar biasa, antara 0,80 dan 0,90 sangat baik, dan antara 0,70 hingga 0,80 dapat diterima [17]. AUC juga sering digunakan dalam skenario *multiclass* dengan metode *one-versus-rest* (OVR).

Dalam penelitian ini, AUC-ROC digunakan untuk membantu mengevaluasi dampak dari teknik transformasi data seperti *One-Hot Encoding*, yang dapat mempengaruhi kemampuan model dalam membedakan antara kelas positif (1) dan negatif (0) secara akurat [16]. Selain itu, karena dataset yang digunakan memiliki distribusi target (Y) yang seimbang, maka penggunaan AUC-ROC memungkinkan evaluasi performa model yang tidak bias terhadap distribusi kelas. Dengan demikian, AUC-ROC memberikan evaluasi yang lebih objektif dan komprehensif terhadap performa model, memungkinkan perbandingan yang jelas antara teknik transformasi data dan model klasifikasi yang digunakan dalam dataset.

### 3) Classification Report

Setelah dilakukan analisis menggunakan *Confusion Matrix*, ketiga model klasifikasi dievaluasi lebih lanjut menggunakan empat metrik utama, yaitu akurasi, presisi, *recall*, dan *f1-score*. Keempat metrik ini digunakan untuk mengukur seberapa baik model dalam melakukan klasifikasi. Beberapa rumus umum berikut dapat digunakan untuk menghitung kinerja klasifikasi, dan hasilnya dapat ditampilkan dalam bentuk persentase untuk nilai akurasi, presisi, dan *recall* [9].

Akurasi didefinisikan sebagai perbandingan antara jumlah data yang diklasifikasikan dengan benar (baik positif yang diprediksi sebagai positif maupun negatif yang diprediksi sebagai negatif dengan jumlah total data dalam dataset. Presisi adalah metrik evaluasi yang mengukur rasio antara jumlah sampel yang secara akurat diprediksi sebagai positif dibandingkan dengan total jumlah sampel yang diprediksi

sebagai positif. *Recall* adalah rasio antara jumlah sampel positif yang diprediksi benar dibandingkan dengan total jumlah sampel positif yang sebenarnya ada dalam dataset. *F1-Score* adalah sebuah metrik evaluasi yang mengkombinasikan presisi dan *recall* untuk memberikan nilai tunggal yang mencerminkan kualitas keseluruhan dari model [18]. Oleh karena itu Tabel IV di bawah dapat digunakan sebagai pedoman perhitungan pada penelitian.

TABEL IV  
RUMUS PERHITUNGAN CLASSIFICATION REPORT

Indikasi	Rumus
Akurasi	$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$
Presisi	$Precision = \frac{TP}{TP + FP}$
Recall	$Recall = \frac{TP}{TP + FN}$
F1-Score	$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall}$

4) *K-Fold Cross-Validation*

Setelah itu, dilakukan juga uji validasi menggunakan teknik *K-Fold Cross-Validation* untuk mengevaluasi kinerja model lebih mendalam. *K-Fold* merupakan suatu metode yang berguna untuk membagi data menjadi data pelatihan dan data pengujian guna mengevaluasi kinerja model pelatihan yang dibuat. Metode ini membagi dataset menjadi K partisi yang memiliki ukuran sama besar secara acak [19]. Setiap partisi akan digunakan sebagai data uji satu kali, sedangkan partisi lainnya digunakan sebagai data pelatihan. Proses ini diulang sebanyak K kali, dengan setiap percobaan menggunakan partisi yang berbeda sebagai data uji [20].

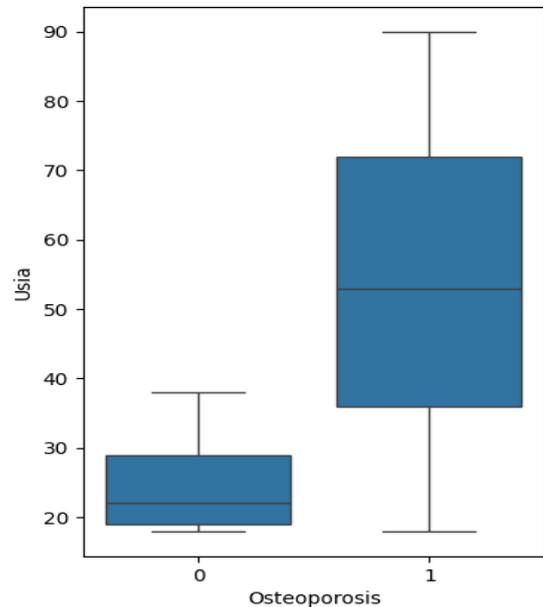
Ilustrasi kinerja dari proses *10-Fold Cross Validation* adalah sebagai berikut, dataset dibagi menjadi 10 bagian (subset) yang sama besar. Kemudian, satu subset dipilih sebagai data pengujian, sementara 9 subset lainnya digabungkan sebagai data pelatihan. Model dilatih menggunakan data pelatihan tersebut dan diuji dengan data pengujian yang dipilih. Proses ini diulangi sebanyak 10 kali, dengan setiap kali menggunakan subset yang berbeda sebagai data pengujian [21]. Hasil dari 10 percobaan tersebut kemudian dirata-ratakan untuk mendapatkan estimasi kinerja model yang lebih akurat dan *robust*.

III. HASIL DAN PEMBAHASAN

Pada bagian ini disajikan hasil dan pembahasan terkait penggunaan model *Random Forest*, *Support Vector Machine* (SVM), dan *Gradient Boosting* dalam memprediksi penyakit osteoporosis. Analisis data dilakukan menggunakan bahasa

pemrograman Python, serta hasilnya dibandingkan untuk menilai kinerja masing-masing model dalam memprediksi penyakit tersebut.

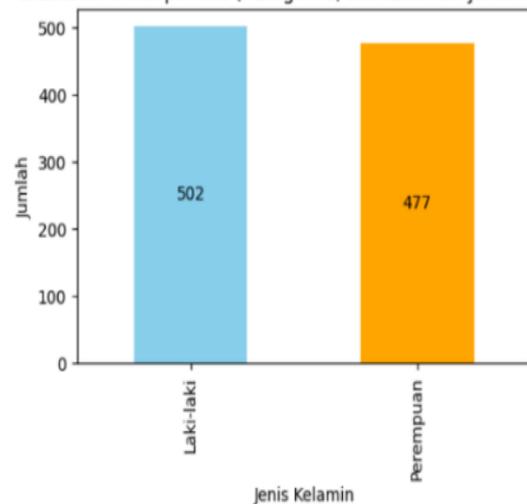
A. Analisis Data Eksploratif



Gambar 5. Boxplot Osteoporosis

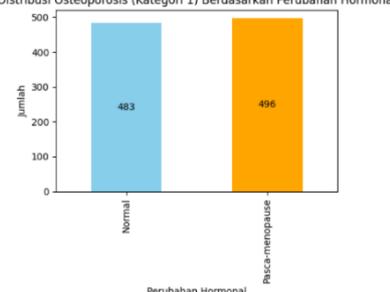
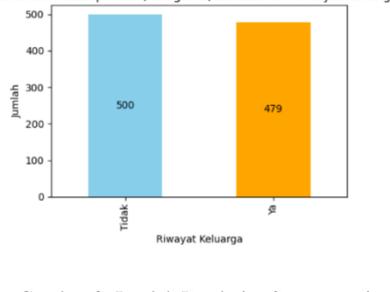
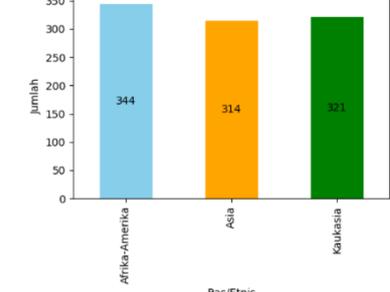
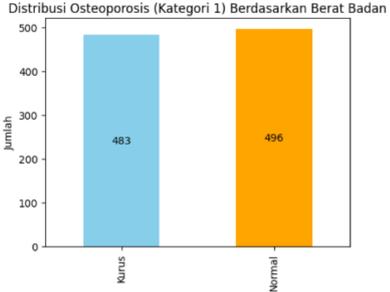
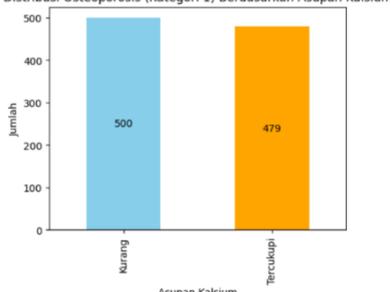
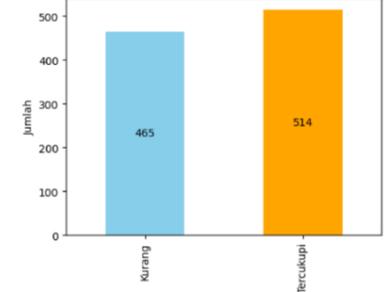
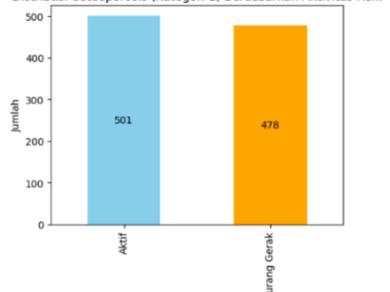
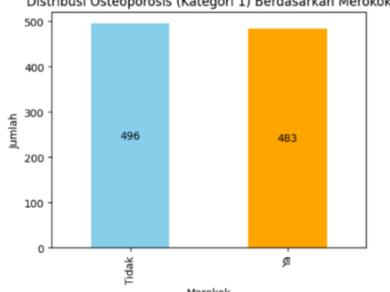
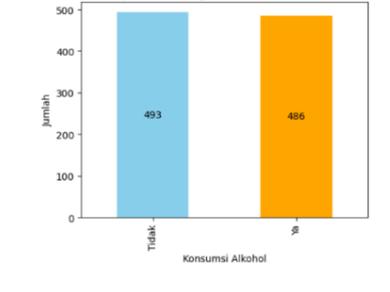
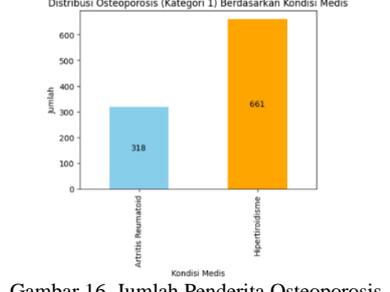
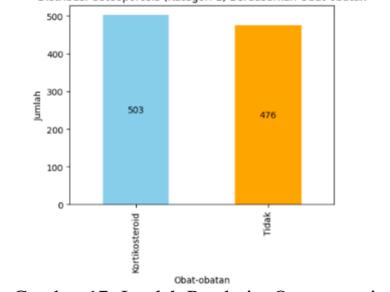
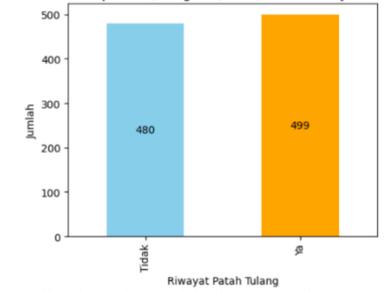
Seluruh Setelah dataset diakuisisi dari situs Kaggle, tahap selanjutnya dalam penelitian ini adalah melakukan analisis data eksploratif (ADE) menggunakan *boxplot* dan *cross-tabulation* untuk mengetahui atribut apa yang paling berpengaruh dalam prediksi. Visualisasi dapat dilihat pada Gambar 5. Pasien yang dikategorikan mengidap osteoporosis berada di usia 35 hingga 75 tahun, sedangkan pasien yang dikategorikan tidak mengalami osteoporosis berada di rentang usia 20 hingga 30 tahun.

Distribusi Osteoporosis (Kategori 1) Berdasarkan Jenis Kelamin



Gambar 6. Jumlah Penderita Osteoporosis berdasarkan Jenis Kelamin

TABEL V  
 DISTRIBUSI PASIEN OSTEOPOROSIS (KATEGORI 1)

<p>Distribusi Osteoporosis (Kategori 1) Berdasarkan Perubahan Hormonal</p>  <p><b>Gambar 7.</b> Jumlah Penderita Osteoporosis berdasarkan Perubahan Hormonal</p>	<p>Distribusi Osteoporosis (Kategori 1) Berdasarkan Riwayat Keluarga</p>  <p><b>Gambar 8.</b> Jumlah Penderita Osteoporosis berdasarkan Riwayat Keluarga</p>	<p>Distribusi Osteoporosis (Kategori 1) Berdasarkan Ras/Etnis</p>  <p><b>Gambar 9.</b> Jumlah Penderita Osteoporosis berdasarkan Ras/Etnis</p>
<p>Distribusi Osteoporosis (Kategori 1) Berdasarkan Berat Badan</p>  <p><b>Gambar 10.</b> Jumlah Penderita Osteoporosis berdasarkan Berat Badan</p>	<p>Distribusi Osteoporosis (Kategori 1) Berdasarkan Asupan Kalsium</p>  <p><b>Gambar 11.</b> Jumlah Penderita Osteoporosis berdasarkan Asupan Kalsium</p>	<p>Distribusi Osteoporosis (Kategori 1) Berdasarkan Asupan Vitamin D</p>  <p><b>Gambar 12.</b> Jumlah Penderita Osteoporosis berdasarkan Asupan Vitamin D</p>
<p>Distribusi Osteoporosis (Kategori 1) Berdasarkan Aktivitas Fisik</p>  <p><b>Gambar 13.</b> Jumlah Penderita Osteoporosis berdasarkan Aktivitas Fisik</p>	<p>Distribusi Osteoporosis (Kategori 1) Berdasarkan Merokok</p>  <p><b>Gambar 14.</b> Jumlah Penderita Osteoporosis berdasarkan Merokok</p>	<p>Distribusi Osteoporosis (Kategori 1) Berdasarkan Konsumsi Alkohol</p>  <p><b>Gambar 15.</b> Jumlah Penderita Osteoporosis berdasarkan Konsumsi Alkohol</p>
<p>Distribusi Osteoporosis (Kategori 1) Berdasarkan Kondisi Medis</p>  <p><b>Gambar 16.</b> Jumlah Penderita Osteoporosis berdasarkan Kondisi Medis</p>	<p>Distribusi Osteoporosis (Kategori 1) Berdasarkan Obat-obatan</p>  <p><b>Gambar 17.</b> Jumlah Penderita Osteoporosis berdasarkan Konsumsi Obat</p>	<p>Distribusi Osteoporosis (Kategori 1) Berdasarkan Riwayat Patah Tulang</p>  <p><b>Gambar 18.</b> Jumlah Penderita Osteoporosis berdasarkan Riwayat Patah Tulang</p>

Hasil visualisasi *cross-tabulation* pada gambar 6-18 menunjukkan bahwa distribusi penderita osteoporosis bervariasi berdasarkan beberapa faktor risiko. Dari segi jenis kelamin, lebih banyak laki-laki (502 orang) yang menderita osteoporosis dibandingkan perempuan (477 orang). Perubahan hormonal pasca-menopause (496 orang) juga sedikit lebih tinggi dibandingkan mereka yang memiliki hormon normal (483 orang). Dari segi ras/etnis, Afrika-Amerika memiliki jumlah kasus tertinggi (344 orang), diikuti oleh Kaukasia (321 orang) dan Asia (314 orang). Pasien dengan hipertiroidisme (661 orang) jauh lebih banyak dibandingkan dengan mereka yang menderita artritis reumatoid (318 orang). Selain itu, penggunaan kortikosteroid juga menunjukkan prevalensi yang lebih tinggi (503 orang) dibandingkan mereka yang tidak menggunakannya (476 orang). Faktor lainnya, seperti riwayat patah tulang (499 orang), asupan kalsium yang kurang (500 orang), dan kurangnya asupan vitamin D (465 orang), turut memengaruhi prevalensi osteoporosis. Kebiasaan merokok (483 orang) dan konsumsi alkohol (486 orang) juga terlibat, meskipun tidak terlalu signifikan jika dibandingkan dengan mereka yang tidak merokok (496 orang) dan tidak mengonsumsi alkohol (493 orang). Aktivitas fisik juga sedikit mempengaruhi, dengan jumlah penderita yang aktif (501 orang) sedikit lebih banyak dibandingkan yang kurang gerak (478 orang). Hasil ini menegaskan pentingnya berbagai faktor risiko dalam mempengaruhi prevalensi osteoporosis.

### B. Pra-pemrosesan Data

Sebelum melakukan klasifikasi, data perlu diolah melalui *preprocessing* atau pra-pemrosesan data terlebih dahulu agar dapat digunakan untuk membangun model. Dalam penelitian ini, dilakukan dua tahap pra-pemrosesan data, yaitu teknik imputasi dan transformasi data menggunakan teknik *One-Hot Encoding* (OHE).

TABEL VI  
PENGECEKAN ATRIBUT DATASET

No	Nama Atribut	Data Null
1	Usia	0
2	Jenis Kelamin	0
3	Perubahan Hormonal	0
4	Riwayat Keluarga	0
5	Ras/Etnis	0
6	Berat Badan	0
7	Asupan Kalsium	0
8	Asupan Vitamin D	0
9	Aktivitas Fisik	0
10	Merokok	0
11	Konsumsi Alkohol	988
12	Kondisi Medis	647
13	Konsumsi Obat	985
14	Riwayat Patah Tulang	0
15	Osteoporosis	0

Teknik imputasi diterapkan pada tiga atribut dataset, yaitu pada atribut *Konsumsi Alkohol*, *Medical Conditions*, dan *Konsumsi Obat*. Nilai null di ketiga atribut diimputasi atau diisi nilainya menggunakan nilai yang sering muncul di masing-masing atribut.

Setelah dilakukan proses menangani nilai yang hilang, selanjutnya masuk ke proses transformasi data menggunakan teknik *One-Hot Encoding*. Teknik *One-Hot Encoding* (OHE) diterapkan pada atribut yang memiliki nilai kategorikal terpisah, seperti kolom Jenis Kelamin. Proses OHE ini bertujuan sebagai komparator untuk menganalisis kinerja algoritma pada dataset tanpa proses OHE (ordinal) dengan dataset yang telah melalui proses OHE dalam memprediksi penyakit osteoporosis.

Atribut-atribut yang dipilih untuk dilakukan proses OHE yaitu, *Jenis Kelamin*, *Perubahan Hormonal*, *Riwayat Keluarga*, *Ras/Etnis*, *Berat Badan*, *Kondisi Medis*, dan *Konsumsi Obat*. Rincian atribut dataset yang telah dilakukan proses OHE dapat dilihat pada Tabel VI.

TABEL VII  
ATRIBUT DATASET SETELAH PROSES OHE

No	Nama Atribut
1	Usia
2	Jenis Kelamin_Laki-Laki
3	Jenis Kelamin_Perempuan
4	Perubahan Hormonal_Normal
5	Perubahan Hormonal_Pascamenopause
6	Riwayat Keluarga_Ya
7	Riwayat Keluarga_Tidak
8	Ras/Etnis_Afrika-Amerika
9	Ras/Etnis_Asia
10	Ras/Etnis_Kaukasian
11	Berat Badan_Normal
12	Berat Badan_Kurus
13	Asupan Kalsium
14	Asupan Vitamin D
15	Aktivitas Fisik
16	Merokok
17	Konsumsi Alkohol
18	Kondisi Medis_Hipertiroidisme
19	Kondisi Medis_Reumatoid Arthritis
20	Konsumsi Obat_Kortikosteroid
21	Konsumsi Obat_Tidak ada
22	Riwayat Patah Tulang
23	Osteoporosis

Setelah melalui proses *One-Hot Encoding* (OHE), atribut dataset bertambah dari yang semula berjumlah 15 kolom atribut, kini menjadi berjumlah 23 kolom atribut. Hasil proses OHE dapat dilihat pada Tabel VII.

### C. Pembagian Data

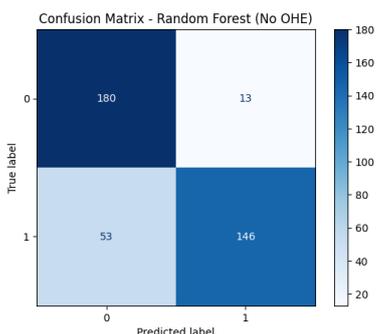
Setelah data selesai diproses pada tahap pra-pemrosesan, selanjutnya data dibagi menjadi dua bagian, yaitu (X) sebagai fitur dan (y) sebagai target. Bagian fitur (X) diambil dari semua kolom atribut kecuali kolom Osteoporosis, dan bagian

target (y) diambil dari kolom Osteoporosis. Selanjutnya, data dipisah menjadi data pelatihan sebesar 80% dan data pengujian sebesar 20% dari total data. Data pelatihan digunakan untuk membangun model, sementara data pengujian digunakan untuk menilai performa model.

**D. Evaluasi Confusion Matrix**

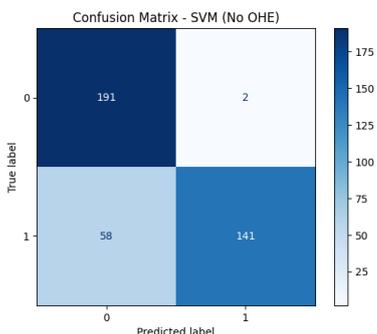
Proses selanjutnya setelah dataset dibagi yaitu proses klasifikasi menggunakan tiga algoritma *machine learning*, yaitu *Random Forest*, *Support Vector Machine (SVM)*, dan *Gradient Boosting*. Ketika model selesai dibangun, maka akan muncul hasil evaluasi, salah satunya *Confusion Matrix*.

1) *Confusion Matrix Random Forest Data Ordinal*



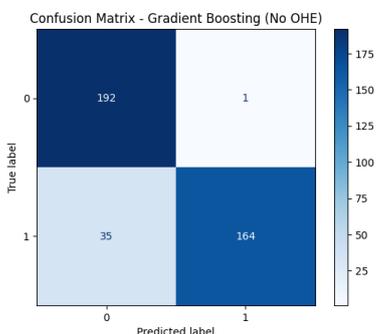
Gambar 19. Matriks Konfusi *Random Forest* Data Ordinal

2) *Confusion Matrix SVM tanpa Data Ordinal*



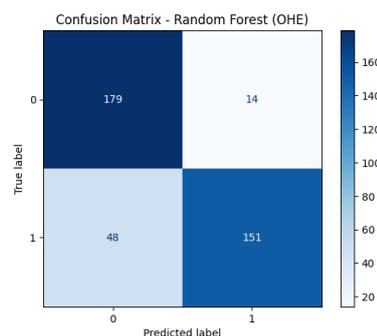
Gambar 20. Matriks Konfusi SVM Data Ordinal

3) *Confusion Matrix Gradient Boosting Data Ordinal*



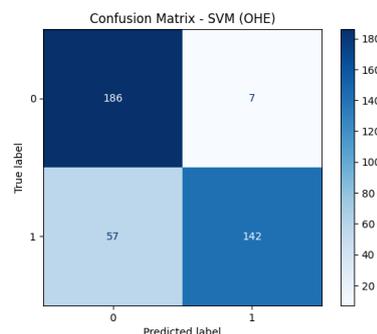
Gambar 21. Matriks Konfusi *Gradient Boosting* Data Ordinal

4) *Confusion Matrix Random Forest dengan OHE*



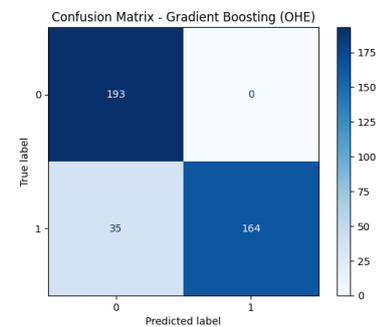
Gambar 22. Matriks Konfusi *Random Forest* dengan OHE

5) *Confusion Matrix SVM dengan OHE*



Gambar 23. Matriks Konfusi SVM dengan OHE

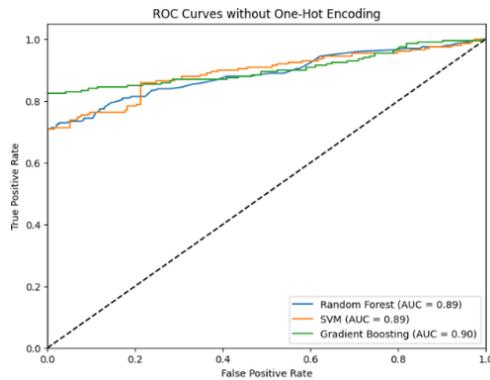
6) *Confusion Matrix Gradient Boosting dengan OHE*



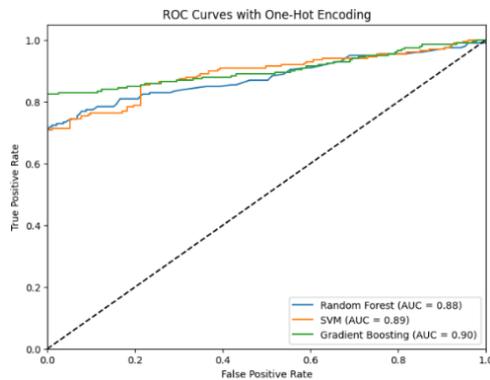
Gambar 24. Matriks Konfusi *Gradient Boosting* dengan OHE

Dari hasil *Confusion Matrix* pada gambar 19-24, terlihat bahwa algoritma *Gradient Boosting* dengan *One-Hot Encoding (OHE)* memberikan performa terbaik dengan jumlah true positive dan true negative yang tinggi serta false positive dan false negative yang rendah dibandingkan algoritma lainnya. *Random Forest* dan SVM juga mengalami peningkatan performa saat menggunakan OHE, tetapi masih memiliki kesalahan prediksi yang lebih tinggi dibandingkan *Gradient Boosting*. Dengan data ordinal, ketiga algoritma cenderung memiliki *false positive* dan *false negative* yang lebih tinggi, terutama pada SVM. Oleh karena itu, *Gradient Boosting* dengan OHE dinilai sebagai algoritma terbaik dalam memprediksi penyakit osteoporosis pada dataset ini.

E. Evaluasi AUC-ROC



Gambar 25. Uji AUC-ROC Data Ordinal



Gambar 26. Uji AUC-ROC dengan OHE

Hasil uji AUC-ROC pada gambar 25 dan 26 menunjukkan bahwa *Gradient Boosting* secara konsisten memiliki performa terbaik dengan nilai AUC sebesar 0.90, baik dengan maupun tanpa *One-Hot Encoding* (OHE). SVM juga memiliki kinerja yang baik dengan skor AUC 0.89, yang tetap stabil terlepas dari penggunaan OHE. Sementara itu, *Random Forest* sedikit menurun performanya saat menggunakan OHE, dengan AUC turun dari 0.89 menjadi 0.88. Secara keseluruhan, penggunaan OHE tidak memberikan perubahan besar terhadap kinerja model, namun *Gradient Boosting* tetap unggul dalam mendeteksi penyakit osteoporosis berdasarkan hasil kurva ROC ini.

F. Evaluasi Classification Report

Setelah berhasil memperoleh hasil *confusion matrix* untuk setiap algoritma yang digunakan, langkah selanjutnya dalam penelitian ini adalah membandingkan performa dari tiga algoritma klasifikasi yang berbeda, yaitu *Random Forest*, *Support Vector Machine* (SVM), dan *Gradient Boosting* menggunakan *Classification Report*. Perbandingan ini dilakukan untuk menganalisis kinerja masing-masing algoritma dalam memprediksi penyakit osteoporosis. Pengujian pertama dilakukan dengan menggunakan dataset yang belum dilakukan proses OHE (ordinal), dimana dataset masih menggunakan 15 atribut dalam melakukan prediksi. Pengujian menggunakan dataset ordinal atau tanpa melalui proses OHE bertujuan sebagai komparator dengan pengujian

kedua. Hasil pengujian pertama dari ketiga algoritma tersebut disajikan dalam VIII.

TABEL VIII  
PERBANDINGAN ALGORITMA DATA ORDINAL (TANPA OHE)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
<i>Random Forest</i>	84.44	85.79	84.59	84.33
<i>Support Vector Machine</i>	84.69	87.65	84.91	84.44
<i>Gradient Boosting</i>	90.82	91.99	90.95	90.77

Selanjutnya, pengujian kedua dilakukan pada data yang telah melalui proses *One-Hot Encoding* (OHE) untuk mengetahui apakah terdapat peningkatan ataupun penurunan kinerja algoritma. Hasil classification report dapat dilihat pada Tabel IX.

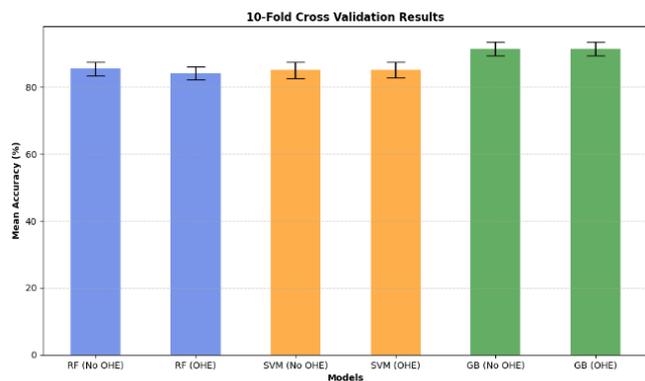
TABEL IX  
PERBANDINGAN ALGORITMA DATA OHE

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
<i>Random Forest</i>	84.95	84.79	85.07	84.89
<i>Support Vector Machine</i>	83.67	85.92	83.86	83.47
<i>Gradient Boosting</i>	91.07	92.32	91.21	91.02

Dari percobaan yang telah dilakukan, terlihat bahwa penerapan *One-Hot Encoding* memberikan dampak yang berbeda pada performa setiap algoritma. Algoritma *Gradient Boosting* menunjukkan peningkatan performa yang signifikan, dengan akurasi dan F1 Score yang lebih tinggi setelah menggunakan *One-Hot Encoding* yaitu sebesar 91.07%. Sebaliknya, *Random Forest* dan *Support Vector Machine* (SVM) mengalami penurunan performa dalam hal akurasi dan *F1-Score* ketika menggunakan *One-Hot Encoding* dibandingkan dengan data ordinal. Hal ini menunjukkan bahwa representasi data dapat mempengaruhi efektivitas algoritma tertentu, di mana *Gradient Boosting* tampaknya lebih mampu memanfaatkan fitur yang dihasilkan dari *One-Hot Encoding*.

G. Validasi Model

Untuk meningkatkan generalisasi model, dilakukan uji validasi model menggunakan metode *K-Fold Cross-Validation* dengan nilai K yang ditetapkan sebesar 10. *K-Fold Cross-Validation* merupakan salah satu teknik yang efektif untuk menguji model dan mengurangi risiko overfitting. Hasil uji perbandingan menggunakan metode *K-Fold Cross-Validation* dapat dilihat pada gambar 28.



Gambar 28. Grafik 10-Fold Cross Validation

Hasil uji *10-fold cross-validation* menunjukkan akurasi rata-rata setiap model dengan rentang kesalahan standar. Akurasi ini memberikan gambaran tentang seberapa baik model dapat memprediksi data baru atau data yang belum pernah dilihat. Akurasi tertinggi dicapai oleh model *Gradient Boosting* (GB) baik dengan data ordinal maupun dengan data OHE. Ini menunjukkan bahwa GB memiliki kemampuan generalisasi yang lebih baik dibandingkan model lainnya. Dari hasil ini, dapat disimpulkan bahwa *Gradient Boosting* merupakan model yang memiliki kemampuan generalisasi terbaik karena memiliki akurasi rata-rata yang tinggi dan rentang *error bar* yang kecil. *Random Forest* dan SVM meskipun juga memiliki kemampuan generalisasi yang cukup baik, namun terlihat memiliki variasi yang lebih tinggi, terutama pada RF dengan OHE.

#### H. Implikasi Klinis

Hasil penelitian tentang penerapan algoritma *machine learning* untuk prediksi osteoporosis memiliki relevansi langsung dengan aplikasi klinis dalam upaya deteksi dini dan pencegahan penyakit. Dengan menggunakan algoritma seperti *Random Forest*, *Support Vector Machine*, dan *Gradient Boosting* sistem prediksi ini dapat membantu tenaga medis mengidentifikasi pasien yang berisiko tinggi osteoporosis dengan lebih akurat dan cepat dibandingkan metode tradisional. Keunggulan *Gradient Boosting* yang terbukti memiliki akurasi terbaik dalam penelitian ini memungkinkan dokter untuk mengambil langkah preventif lebih dini, seperti penyesuaian gaya hidup atau pemberian terapi, sehingga mencegah komplikasi lebih lanjut seperti patah tulang. Implementasi sistem prediksi berbasis *machine learning* ini dapat mempercepat proses diagnostik di klinik, meningkatkan efisiensi pengambilan keputusan, dan mengurangi beban kesehatan jangka panjang pada pasien.

#### IV. KESIMPULAN

Penelitian ini berhasil mengembangkan model prediksi penyakit osteoporosis menggunakan tiga algoritma klasifikasi, yaitu *Random Forest*, *Support Vector Machine* (SVM), dan *Gradient Boosting*. Hasil penelitian menunjukkan bahwa model *Gradient Boosting* memiliki performa terbaik dengan akurasi sebesar 90.82%, *precision*

sebesar 91.99%, *recall* sebesar 90.95%, dan *F1-Score* sebesar 90.77% pada data ordinal. Ketika data diubah menggunakan *One-Hot Encoding*, akurasi model *Gradient Boosting* meningkat menjadi 91.07%, dengan *precision* sebesar 92.32%, *recall* sebesar 91.21%, dan *F1-Score* sebesar 91.02%. Ini menunjukkan bahwa *Gradient Boosting* lebih mampu menangani pola kompleks dalam data osteoporosis, dan mampu memberikan hasil prediksi yang lebih akurat dibandingkan dengan SVM dan *Random Forest*. Oleh karena itu, model ini dapat menjadi alat yang sangat berguna dalam membantu deteksi dini osteoporosis, memberikan manfaat bagi para praktisi medis dalam pengambilan keputusan.

Penelitian ini juga menghadapi keterbatasan dalam hal sensitivitas model terhadap perubahan kompleksitas data setelah transformasi menggunakan *One-Hot Encoding*, yang mempengaruhi akurasi model SVM dan *Random Forest*. Selain itu, *Gradient Boosting* memerlukan waktu pelatihan yang lebih lama dan sumber daya komputasi yang lebih besar dibandingkan algoritma lainnya. Untuk penelitian selanjutnya, disarankan untuk menggabungkan algoritma yang berbeda atau mencoba algoritma lain seperti *XGBoost* dan *LightGBM* yang mungkin lebih efisien dalam hal komputasi. Penelitian juga dapat diperluas dengan melakukan eksplorasi lebih dalam mengenai parameter tuning pada masing-masing algoritma untuk mendapatkan hasil yang optimal. Penelitian lebih lanjut juga perlu melakukan validasi model pada dataset yang lebih besar dan beragam untuk menguji generalisasi model. Penelitian ini menegaskan bahwa pemilihan algoritma yang tepat, seperti *Gradient Boosting*, sangat penting dalam aplikasi medis, terutama untuk deteksi dini penyakit yang kompleks seperti osteoporosis.

#### UCAPAN TERIMA KASIH

Ucapan terima kasih disampaikan kepada Program Studi Informatika Fakultas Ilmu Komputer Universitas Amikom Yogyakarta, dosen pembimbing, serta keluarga dan teman-teman yang telah memberikan dukungan dan bantuan sehingga penelitian ini dapat diselesaikan.

#### DAFTAR PUSTAKA

- [1] X. Wu and S. Park, "A Prediction Model for Osteoporosis Risk Using a Machine-Learning Approach and Its Validation in a Large Cohort," *J Korean Med Sci*, vol. 38, no. 21, 2023, doi: 10.3346/jkms.2023.38.e162.
- [2] D. S. Wicaksono and R. Y. Maulana, "Manfaat Ekstrak Dandelion Dalam Mencegah Osteoporosis," *Jurnal Penelitian Perawat Profesional*, vol. 2, no. 2, 2020, doi: 10.37287/jppp.v2i2.87.
- [3] N. Sani, Y. Yuniastini, A. Putra, and Y. Yuliyana, "Tingkat Pengetahuan Osteoporosis Sekunder dan Perilaku Pencegahan Mahasiswa Universitas Malahayati," *Jurnal Ilmiah Kesehatan Sandi Husada*, vol. 11, no. 1, 2020, doi: 10.35816/jiskh.v11i1.236.
- [4] V. V. Khanna et al., "A decision support system for osteoporosis risk prediction using machine learning and explainable artificial intelligence," *Heliyon*, vol. 9, no. 12, Dec. 2023, doi: 10.1016/j.heliyon.2023.e22456.
- [5] M. D. Purbolaksono, M. Irvan Tantowi, A. Imam Hidayat, and A. Adiwijaya, "Perbandingan Support Vector Machine dan Modified Balanced Random Forest dalam Deteksi Pasien Penyakit Diabetes," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 2, 2021, doi: 10.29207/resti.v5i2.3008.

- [6] E. S. Patasik and S. Yulianto, "Classification Of Regional Languages Using Methods Gradient Boots And Random Forest," *Jurnal Teknik Informatika (Jutif)*, vol. 4, no. 5, 2023, doi: 10.52436/1.jutif.2023.4.5.1459.
- [7] A. Algoritma et al., "Analisis Algoritma Klasifikasi Untuk Mengidentifikasi Potensi Risiko Kesehatan Ibu Hamil," *Journal of Applied Computer Science and Technology*, vol. 5, no. 1, pp. 120–127, Jun. 2024, doi: 10.52158/JACOST.V5I1.809.
- [8] S. P. Nainggolan and A. Sinaga, "Comparative Analysis Of Accuracy Of Random Forest And Gradient Boosting Classifier Algorithm For Diabetes Classification," *Sebatik*, vol. 27, no. 1, pp. 97–102, Jun. 2023, doi: 10.46984/sebatik.v27i1.2157.
- [9] C. Z. V. Junus, T. Tarno, and P. Kartikasari, "Klasifikasi Menggunakan Metode Support Vector Machine Dan Random Forest Untuk Deteksi Awal Risiko Diabetes Melitus," *Jurnal Gaussian*, vol. 11, no. 3, pp. 386–396, Jan. 2023, doi: 10.14710/j.gauss.11.3.386-396.
- [10] S. D. Wahyuni and R. H. Kusumodestoni, "Optimalisasi Algoritma Support Vector Machine (SVM) Dalam Klasifikasi Kejadian Data Stunting," *Bulletin of Information Technology (BIT)*, vol. 5, no. 2, pp. 56–64, 2024, doi: 10.47065/bit.v5i2.1247.
- [11] Kairos Abinaya Susanto et al., "Implementasi Bahasa Python Dalam Menganalisis Pengaruh Rokok Terhadap Risiko Pasien Terkena Penyakit Stroke," *Jurnal Publikasi Teknik Informatika*, vol. 2, no. 2, pp. 48–58, May 2023, doi: 10.55606/jupti.v2i2.1722.
- [12] S. Ana, R. Kurniawan, and A. Nazir, "Pengklasteran Risiko Covid-19 Di Riau Menggunakan Teknik One Hot Encoding Dan Algoritma K-Means Clustering," *Jurnal Informasi dan Komputer*, vol. 10, no. 1, 2022, doi: 10.35959/jik.v10i1.291.
- [13] I. Lestari, M. Akbar, and B. Intan, "Perbandingan Algoritma Machine Learning Untuk klasifikasi Amenorrhoea," *Journal of Computer and Information Systems Ampera*, vol. 4, no. 1, pp. 32–43, Jan. 2023, doi: 10.51519/JOURNALSISA.V4I1.371.
- [14] A. Febriansyah Istanto, A. Id Hadiana, and F. Rakhmat Umbara, "Prediksi Curah Hujan Menggunakan Metode Categorical Boosting (Catboost)," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 4, 2024, doi: 10.36040/jati.v7i4.7304.
- [15] R. Safdari, A. Deghatipour, M. Gholamzadeh, and K. Maghooli, "Applying data mining techniques to classify patients with suspected hepatitis C virus infection," *Intelligent Medicine*, vol. 2, no. 4, 2022, doi: 10.1016/j.imed.2021.12.003.
- [16] S. Dewi, H. A. Al Kautsar, and D. Y. Utami, "Prediksi Keberhasilan Pemasaran Layanan Jasa Perbankan Menggunakan Algoritma Logistic Regreesion," *Computer Science (CO-SCIENCE)*, vol. 3, no. 2, 2023, doi: 10.31294/coscience.v3i2.1931.
- [17] T. Tan, H. Sama, G. Wijaya, and O. E. Aboagye, "Studi Perbandingan Deteksi Intrusi Jaringan Menggunakan Machine Learning: (Metode SVM dan ANN)," *Jurnal Teknologi dan Informasi*, vol. 13, no. 2, pp. 152–164, Aug. 2023, doi: 10.34010/jati.v13i2.10484.
- [18] L. Syafaâ€TMah, Z. Zulfatman, I. Pakaya, and M. Lestandy, "Comparison of Machine Learning Classification Methods in Hepatitis C Virus," *Jurnal Online Informatika*, vol. 6, no. 1, pp. 73–78, Jun. 2021, doi: 10.15575/JOIN.V6I1.719.
- [19] K. Auliyatuz Zahroh et al., "Perbandingan Ekstraksi Fitur Untuk Klasifikasi COVID-19, MERS, dan SARS Menggunakan Algoritma Extreme Learning Machine," *Jurnal Fourier*, vol. 13, no. 1, pp. 30–41, Apr. 2024, doi: 10.14421/FOURIER.2024.131.30-41.
- [20] J. Homepage et al., "Implementasi Algoritma Naïve Bayes Classifier dan K-Nearest Neighbor untuk Klasifikasi Penyakit Ginjal Kronik," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 2, pp. 710–718, Apr. 2024, doi: 10.57152/MALCOM.V4I2.1229.
- [21] D. Nurlailly, Y. P. Irfandi, N. Santoso, S. Qomariyah, and D. Wibowo, "Classification of Hepatitis Patients Using Logistic Regression and Support Vector Machines Methods," *Jurnal Pendidikan Matematika (Kudus)*, vol. 5, no. 2, p. 237, Dec. 2022, doi: 10.21043/jpmk.v5i2.17052.