**LAPORAN TUGAS AKHIR  
MACHINE LEARNING**

**KLASIFIKASI KANKER PARU-PARU MENGGUNAKAN  
REGRESI (STOCHASTIC GRADIENT DESCENT)**

****

**Disusun Oleh :**

Andre 220411100028

Nico Dwi Putra 220411100101

Tofan Adi Nugroho 220411100131

Surya Eka Santoso 220411100149

**PROGRAM STUDI TEKNIK INFORMATIKA**

**FAKULTAS TEKNIK**

**UNIVERSITAS TRUNOJOYO MADURA**

**2024**

DAFTAR ISI

[DAFTAR ISI 1](#_Toc185619574)

[BAB I PENDAHULUAN 3](#_Toc185619575)

[1.1 Latar Belakang 3](#_Toc185619576)

[1.2 Rumusan Masalah 4](#_Toc185619577)

[1.3 Tujuan Penelitian 4](#_Toc185619578)

[1.4 Manfaat Penelitian 4](#_Toc185619579)

[BAB II LANDASAN TEORI 5](#_Toc185619580)

[2.1 Dataset 5](#_Toc185619581)

[2.2 Preprocessing 5](#_Toc185619582)

[2.2.1 Missing Value 5](#_Toc185619583)

[2.2.2 Normalisasi 5](#_Toc185619584)

[2.2.3 Transformasi 6](#_Toc185619585)

[2.2.4 Penyeimbangan Kelas/Target 6](#_Toc185619586)

[2.3 Stochastic Gradient Descent Classifier 7](#_Toc185619587)

[2.4 Model Score 8](#_Toc185619588)

[BAB III METODOLOGI PENELITIAN 10](#_Toc185619589)

[3.1 Desain Penelitian 10](#_Toc185619590)

[3.2 Pencarian Referensi 11](#_Toc185619591)

[3.3 Pengumpulan Data 11](#_Toc185619592)

[3.4 Preprocessing Data 11](#_Toc185619593)

[3.5 Pemisahan Data 11](#_Toc185619594)

[3.6 Modelling dan Training 11](#_Toc185619595)

[3.7 Evaluasi Model 11](#_Toc185619596)

[3.8 Testing 12](#_Toc185619597)

[3.9 Pelaporan 12](#_Toc185619598)

[BAB IV HASIL DAN PEMBAHASAN 13](#_Toc185619599)

[3.10 Preprocessing 13](#_Toc185619600)

[3.10.1 Normalisasi 13](#_Toc185619601)

[3.10.2 Transformasi 14](#_Toc185619602)

[3.11 Memisahkan Fitur dan Membagi Data 14](#_Toc185619603)

[3.12 Membangun Model Dengan Algoritma SGD 15](#_Toc185619604)

[3.12.1 Pengujian Dengan k-Fold Cross Validation 15](#_Toc185619605)

[3.12.2 Pengujian Dengan k-Fold Cross Validation Menggunakan SMOTE 16](#_Toc185619606)

[3.3.3 Perbandingan Performa Model dengan dan tanpa SMOTE 17](#_Toc185619607)

[3.4 Testing 17](#_Toc185619608)

[BAB IV KESIMPULAN 19](#_Toc185619609)

[DAFTAR PUSTAKA 20](#_Toc185619610)

# BAB I PENDAHULUAN

## Latar Belakang

Kanker, menurut definisi dari National Cancer Institute, adalah penyakit genetik yang yang disebabkan oleh perubahan gen yang mengontrol fungsi sel, terutama fungsi untuk tumbuh dan membelah. Kanker merupakan sel-sel baru yang tumbuh secara abnormal kemudian menyerang bagian tubuh kontralateral dan menyebar ke organ lain. Kanker merupakan penyebab kematian tertinggi kedua di dunia. Menurut The Global Cancer Burden memperkirakan kasus baru kanker meningkat menjadi 19,3 juta dan sekitar 10 juta kematian pada tahun 2020. Satu dari 5 orang di seluruh dunia mengidap kanker selama hidup mereka. Satu dari 8 pria dan satu dari 11 wanita meninggal karena penyakit ini[1].

Kanker paru adalah pertumbuhan sel kanker yang tidak terkendali dalam jaringan paru yang dapat disebabkan oleh sejumlah karsinogen, terutama asap rokok. Menurut World Health Organization (WHO),kanker paru merupakan penyebab kematian utama dalam kelompok kanker. Kanker paru memerlukan penanganan yang tepat. Buruknya diagnosis penyakit ini berkaitan dengan jarangnya penderita datang ke dokter karena penyakitnya masih berada pada stadium awal. hanya 15% kasus kanker paru yang ditemukan sejak stadium awal. Deteksi dini dan penanganan yang tepat pada pasien yang menderita kanker paru diharapkan mampu mengurangi angka kematian yang diakibatkan oleh kanker paru dan dapat meningkatkan angka harapan hidup[2].

Machine Learning adalah cabang ilmu komputer yang memungkinkan komputer untuk belajar dari data tanpa instruksi pemrograman secara eksplisit. Machine Learning (ML) merupakan salah satu aplikasi dari Kecerdasan Buatan (Artificial Intelligence, AI) yang berfokus pada pengembangan sistem yang mampu belajar secara mandiri tanpa perlu diprogram berulang kali. ML memerlukan data latih (training data) dalam proses pembelajarannya untuk menghasilkan suatu output. Dengan demikian, ML dapat digambarkan sebagai proses pemrograman komputer yang bertujuan mencapai performa atau kriteria tertentu menggunakan data latih atau pengalaman dari waktu sebelumnya (past experience). Beberapa penelitian menyimpulkan bahwa ML memiliki potensi besar di bidang medis, termasuk dalam memprediksi penyakit[3].

Salah satu algoritma populer dalam supervised machine learning adalah Stochastic Gradient Descent (SGD). SGD merupakan variasi dari metode optimasi gradient descent, di mana pembaruan parameter dilakukan setiap kali model memproses satu data dalam proses pelatihan. Dengan pembaruan parameter secara bertahap ini, SGD dapat bekerja lebih cepat pada dataset berukuran besar. Biasanya, nilai standar untuk learning rate pada SGD adalah 0,01[4]. SGD bekerja dengan memanfaatkan sejumlah sampel kecil data dalam setiap langkah iterasi, memungkinkan model untuk memperbarui parameter secara cepat. Teknik ini memungkinkan pemodelan pada data dengan dimensi tinggi dan volume besar tanpa membutuhkan memori yang besar, menjadikannya ideal untuk aplikasi yang memerlukan proses real-time dan batch[5].

Maka tujuan dari tugas ini adalah untuk melakukan klasifikasi kanker paru-paru menggunakan algoritma regresi Stochastic Gradient Descent (SGD). Pendekatan ini dipilih karena kemampuannya dalam menangani data berukuran besar dan berdimensi tinggi dengan efisiensi komputasi yang tinggi. Dengan memanfaatkan data latih yang relevan, model ini diharapkan mampu mengenali pola-pola penting dalam data pasien sehingga dapat memberikan prediksi yang akurat mengenai keberadaan kanker paru-paru. Melalui metode ini, diharapkan dapat mendukung upaya deteksi dini dan membantu meningkatkan efektivitas penanganan pasien, serta berkontribusi dalam mengurangi angka kematian akibat kanker paru-paru.

## Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah dalam penelitian ini adalah:

1. Bagaimana menerapkan metode regresi dengan optimasi Stochastic Gradient Descent untuk klasifikasi penyakit kanker paru-paru?
2. Seberapa efektif metode tersebut dalam mengidentifikasi penyakit paru-paru?

## Tujuan Penelitian

Penelitian ini bertujuan untuk:

1. Mengembangkan model klasifikasi diagnosis penyakit paru-paru dengan metode regresi menggunakan optimasi Stochastic Gradient Descent.
2. Mengukur akurasi dan efektivitas metode ini dalam mengidentifikasi penyakit paru-paru.

## Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini antara lain:

1. Memberikan kontribusi dalam deteksi dini kanker paru-paru menggunakan metode machine learning, khususnya algoritma Stochastic Gradient Descent.
2. Menyediakan acuan untuk pengembangan lebih lanjut bagi penelitian klasifikasi penyakit lainnya di bidang medis.
3. Memperlihatkan keefektifan metode Stochastic Gradient Descent dalam menangani kasus klasifikasi dengan dataset berukuran besar dan berdimensi tinggi pada aplikasi medis.

# BAB II LANDASAN TEORI

## Dataset

Data yang akan digunakan adalah dataset yang didapat dari website Kaggle yaitu Lung Cancer atau Kanker Paru Paru ( <https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer> ). Dataset ini berisi data yang dikumpulkan dari website sistem prediksi kanker paru paru online, Dataset berisi 309 data dengan 270 penderita kanker paru paru dan 39 yang bukan penderita kanker paru paru, pada dataset ini tidak memiliki nilai missing atau missing value. Dataset ini memiliki 16 atribut dimana 15 atribut sedangkan 1 atribut class atau target.

|  |  |
| --- | --- |
| Atribut | Deskripsi |
| Gender | Jenis kelamin pasien |
| Age | Usia pasien |
| Smoking | Status pasien merokok atau tidak |
| Yellow Fingers | Jari yang kuning akibat merokok |
| Anxiety | Indikasi kecemasan |
| Peer Pressure | Pengaruh tekanan sosial dalam kebiasaan merokok |
| Chronic Disease | Adanya penyakit kronis |
| Fatigue | Tingkat kelelahan |
| Allergy | Status alergi |
| Wheezing | Adanya gejala mengi |
| Alcohol Consuming | Kebiasaan konsumsi alkohol |
| Coughing | Intesitas batuk |
| Shortness of Breath | Gejala sesak nafas |
| Swallowing Difficulty | Kesulitan menelan |
| Chest Pain | Gejala nyeri dada |

Table 1. Fitur dalam Dataset

## Preprocessing

### Missing Value

Dalam proses preprocessing data, penanganan missing value dilakukan untuk memastikan kelengkapan data yang akan digunakan dalam pelatihan model. Missing value ditangani dengan metode imputasi, yaitu menggantinya menggunakan nilai rata-rata (mean) dari fitur terkait agar tidak mengurangi jumlah data yang tersedia dan menjaga distribusi data tetap konsisten.

### Normalisasi

Normalisasi data dilakukan menggunakan metode Min-Max Scaling untuk menyetarakan rentang nilai fitur-fitur dalam dataset. Normalisasi ini bertujuan untuk mengubah skala nilai menjadi lebih kecil tanpa melakukan modifikasi isi informasi didalamnya dengan melakukan operasi perubahan skala batas tepi atas 2 dan batas tepi bawah 1 (1,1), sehingga setiap fitur memiliki kontribusi yang seimbang dalam pelatihan model. Proses ini penting untuk meningkatkan performa model machine learning, terutama pada algoritma berbasis gradien seperti Stochastic Gradient Descent (SGD). Perhitungan Min-Max Scalaling dapat menggunakan persamaan sebagai berikut

Keterangan:

Xsc : Nilai Normalisasi

X : Nilai Value dalam Dataset

Xmin : Nilai minimal dalam suatu feature

Xmax : Nilai maksimal dalam suatu feature

### Transformasi

Proses mengubah data mentah menjadi format yang lebih sesuai untuk analisis atau pemodelan. Transformasi ini bertujuan untuk meningkatkan kualitas data, menyelaraskan nilai-nilai dalam data, dan memudahkan algoritma machine learning dalam memprosesnya. Proses transformasi data sangat penting untuk memastikan bahwa data dalam kondisi optimal sehingga algoritma machine learning dapat bekerja secara efisien dan menghasilkan model yang lebih akurat. Metode yang digunakan adalah Label Encoding dari library bawaan dalam scikit-learn, yaitu LabelEncoder dari modul sklearn.preprocessing.

Mengubah features kategori seperti GENDER (dengan nilai "M menjadi 1" dan "F menjadi 2") dan LUNG CANCER (dengan nilai "YES menjadi 2" dan "NO menjadi 1") diubah menjadi tipe numerik menggunakan label encoding.

### Penyeimbangan Kelas/Target

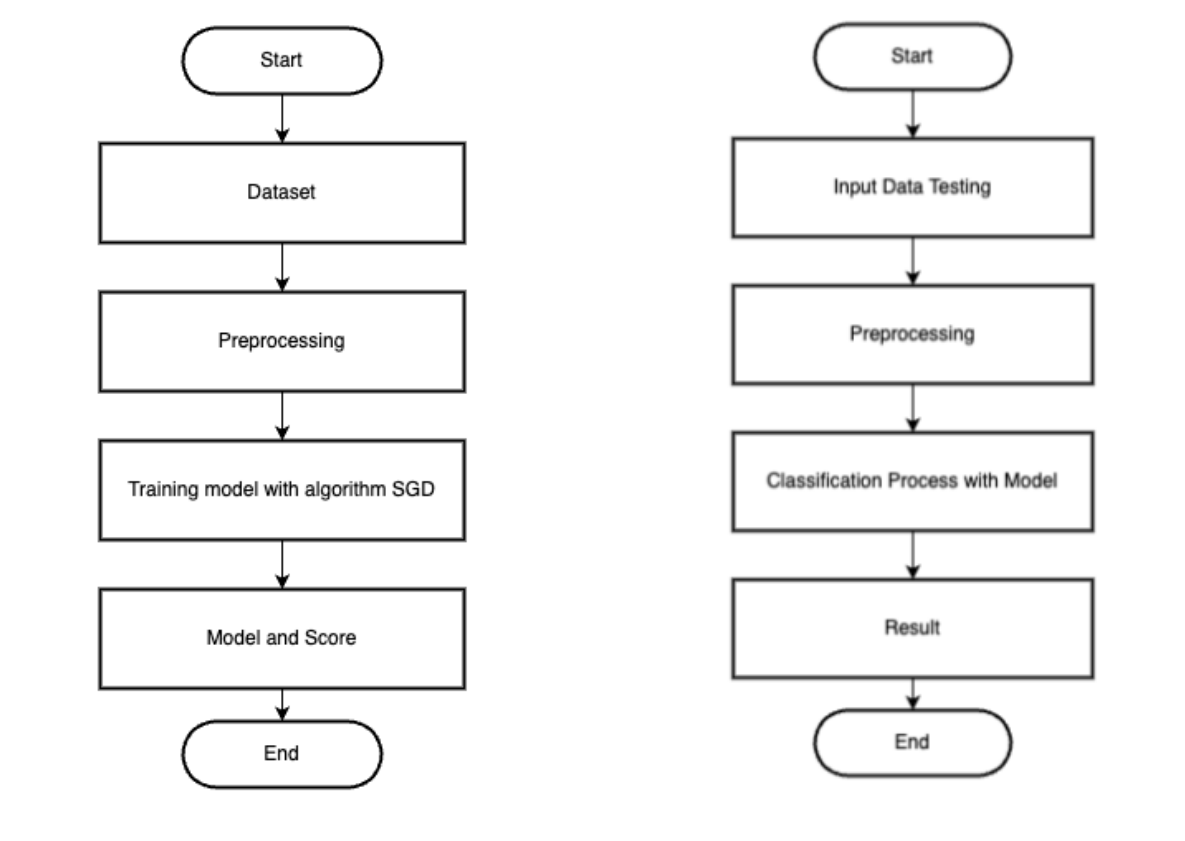
Penyeimbangan kelas adalah langkah penting dalam preprocessing data, terutama ketika menghadapi masalah kelas yang tidak seimbang (imbalanced class). Dalam masalah klasifikasi, kelas yang tidak seimbang terjadi ketika jumlah data untuk satu kelas jauh lebih banyak dibandingkan kelas lainnya. Misalnya, dalam dataset yang memprediksi keberadaan penyakit, jumlah data yang memiliki penyakit (positif) mungkin sangat sedikit dibandingkan dengan yang tidak (negatif). Jika hal ini tidak ditangani dengan baik, model machine learning cenderung memprediksi kelas mayoritas dengan lebih baik, sehingga mengabaikan kelas minoritas, yang dapat menyebabkan performa model yang buruk[6].

SMOTE merupakan teknik menyeimbangkan jumlah distribusi data sampel pada kelas minoritas dengan cara menyeleksi data sampel tersebut hingga jumlah data sampel menjadi seimbang dengan jumlah sampel pada kelas mayoritas. Penggunaan metode SMOTE memungkinkan akan adanya overfitting. Overfiting dapat terjadi karena data pada kelas minoritas diduplikasi sehingga memungkinkan adanya data latih yang sama. Tahapan dalam melakukan SMOTE dimulai dari menghitung jarak antar data pada data minoritas, selanjutnya menentukan nilai presentase SMOTE kemudian menetukkan jumlah k terdekat dan yang terakhir adalah menciptakan data sintesis dengan persamaan berikut[6]:

Dengan adalah data sintesis yang akan diciptakan data yang akan direplikasi, data yang memiliki jarak terdekat dari data yang akan direplikasi dan nilai random antara 0 dan 1.

## Stochastic Gradient Descent Classifier

Penelitian ini menggunakan Machine Learning dengan algoritma Stochastic Gradient Descent untuk mengklasifikasikan kanker paru-paru



Gambar 1. (a) Flowchart untuk membuat model and score (b) Flowchart proses klasifikasi

SGD merupakan sebuah pendekatan sederhana dan efisien dalam melakukan klasifikasi secara linier menggunakan pembelajaran diskriminatif. Metode ini berupa algoritma optimasi iteratif (ulang) yang berguna untuk mencari titik fungsi minimum yang dapat diturunkan. Pada awal proses algoritma dimulai dengan melakukan penebakan. Kesalahan penebakan diperbaiki selama terjadi pengulangan tebakan menggunakan aturan gradien (turunan) dari fungsi yang akan diminimalkan. SGD memiliki kemampuan belajar lebih cepat dalam melakukan pelatihan klasifikasi. Selain itu, berdasarkan ukuran dataset latih tidak terbatas waktu pelaksanaannya[7].

Proses algoritma SGD adalah dengan menemukan nilai θ yang dapat meminimalkan fungsi J(θ). Untuk menentukan nilai awal θ digunakan algoritme pencarian, kemudian pada setiap iterasi nilai θ agar terus diperbaharui sampai menemukan titik minimum atau nilai J yang paling minimum. Proses pembaharuan nilai θ pada setiap iterasi menggunakan Persamaan (2). Pembaharuan dilakukan secara bersamaan untuk semua nilai j = 0, …, n. Variable α merupakan learning rate yang mengatur seberapa besar pembaharuan nilai. Persamaan nilai J(θ) dapat dilihat pada Persamaan (3), di mana L merupakan loss function yang digunakan pada data pelatihan (𝑥1, 𝑦1), … (𝑥𝑛, 𝑦𝑥), dan R merupakan regularisasi atau penalty terhadap kompleksitas model[7].

: Parameter yang dioptimasi.

: Langkah pembaruan (learning rate).

: Arah perubahan berdasarkan turunan fungsi biaya.

: Nilai fungsi biaya yang akan diminimalkan.

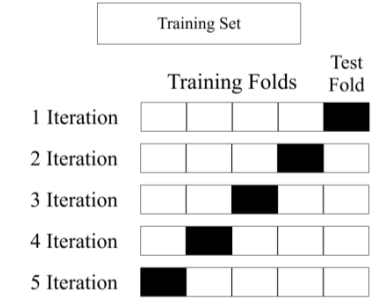
## Model Score

Tahap ini terdiri dari dua bagian utama yang digunakan untuk menilai kualitas model yang telah dibuat. Pertama, menggunakan cross-validation untuk mengevaluasi performa model dan mengukur seberapa baik model tersebut dalam menangani data. Kedua, menguji model dengan menghitung metrik evaluasi seperti akurasi, presisi, recall, dan F1-score. Metrik-metrik ini diperoleh dengan menganalisis hasil prediksi model, termasuk membandingkan jumlah prediksi true positive dengan false positive dalam klasifikasi. Untuk proses evaluasi dengan confusion matrix maka akan diperoleh nilai precision, recall, dan accuracy yang didapat dari rumus:

Keterangan:

|  |  |
| --- | --- |
| TP (True Positive) | : jumlah kasus kanker paru-paru yang berhasil diklasifikasikan dengan benar  sebagai positif (kanker paru-paru). |
| FP (False Positive) | : jumlah kasus non-kanker paru-paru yang salah diklasifikasikan sebagai positif  (kanker paru-paru). |
| TN (True Negative) | : jumlah kasus non-kanker paru-paru yang berhasil diklasifikasikan dengan  benar sebagai negatif (bukan kanker paru-paru). |
| FN (False Negative) | : jumlah kasus kanker paru-paru yang salah diklasifikasikan sebagai negatif  (bukan kanker paru-paru). |

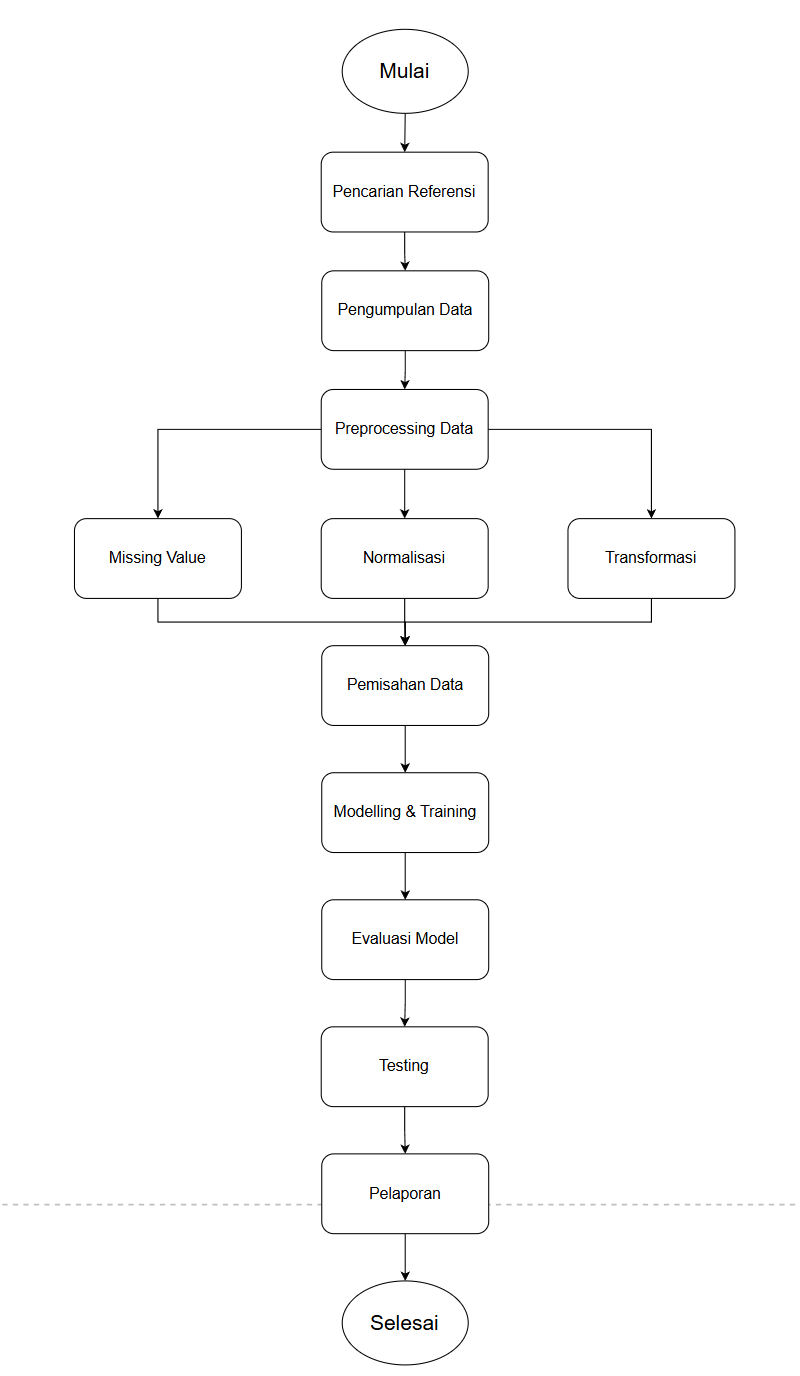
Pada tahap ini, juga diterapkan cross-validation, yaitu sebuah prosedur acak yang membagi data menjadi K bagian yang hampir sama besar. Setiap bagian (lipatan) digunakan secara bergantian untuk menguji model yang dilatih menggunakan K-1 bagian lainnya dengan algoritma klasifikasi tertentu. Pada Gambar 1 merupakan gambaran proses terjadinya cross validation



Gambar 2. Cross validation process

# BAB III METODOLOGI PENELITIAN

## Desain Penelitian



Gambar 3. Alur Penelitian

Penelitian ini dirancang dengan pendekatan kuantitatif menggunakan metode supervised machine learning. Fokus utama penelitian adalah pengembangan model klasifikasi kanker paru-paru dengan algoritma Stochastic Gradient Descent (SGD). Tahapan penelitian mencakup pengumpulan data, preprocessing, pemodelan, evaluasi, hingga pelaporan. Desain ini dirumuskan untuk memastikan proses berjalan secara sistematis, dimulai dari pengumpulan literatur sebagai dasar teori hingga implementasi dan pengujian model yang dapat memberikan hasil yang akurat dan terukur.

## Pencarian Referensi

Setelah desain penelitian dirumuskan, langkah berikutnya adalah mencari dan mengkaji referensi jurnal-jurnal terdahulu. Jurnal yang dipilih berkaitan dengan penerapan algoritma SGD, terutama pada kasus klasifikasi data medis. Kajian literatur ini bertujuan untuk memahami konsep, metode, dan performa SGD dalam berbagai konteks serta sebagai acuan untuk membandingkan hasil penelitian yang dilakukan.

## Pengumpulan Data

Dataset yang digunakan dalam penelitian ini adalah dataset kanker paru-paru yang diunduh dari platform Kaggle. Dataset ini terdiri dari 309 data dengan atribut-atribut pasien seperti usia, kebiasaan merokok, dan gejala tertentu yang relevan dengan deteksi kanker paru-paru. Dataset ini menjadi dasar untuk membangun dan menguji model klasifikasi.

## Preprocessing Data

Tahap preprocessing bertujuan untuk mempersiapkan dataset agar dapat diolah dengan optimal oleh algoritma machine learning. Langkah-langkah preprocessing meliputi:

* Penanganan missing values (jika ada) dengan Mean.
* Normalisasi data menggunakan metode Min-Max Scaling untuk menyetarakan rentang nilai fitur.
* Transformasi data kategori menjadi numerik dengan Label Encoding.

## Pemisahan Data

Dataset yang telah diproses dibagi menjadi dua bagian, yaitu data latih sebesar 80% dan data uji sebesar 20%. Data latih digunakan untuk membangun model, sedangkan data uji digunakan untuk mengevaluasi performa model setelah dilatih.

## Modelling dan Training

Pada tahap ini, algoritma Stochastic Gradient Descent (SGD) diterapkan untuk melatih model klasifikasi kanker paru-paru. Model dilatih menggunakan data latih yang telah diproses dengan menerapkan 1000 iterasi untuk memastikan parameter model dapat diperbarui secara optimal selama proses pelatihan. Teknik k-Fold Cross Validation digunakan untuk mengevaluasi performa model pada berbagai subset data, memastikan stabilitas, dan meningkatkan keakuratan hasil prediksi. Selain itu, teknik SMOTE (Synthetic Minority Oversampling Technique) diterapkan untuk menangani ketidakseimbangan data dengan menciptakan distribusi kelas yang lebih merata, sehingga model dapat mengenali kedua kelas dengan lebih baik dan menghindari bias terhadap kelas mayoritas.

## Evaluasi Model

Model yang telah dilatih dievaluasi dengan menghitung metrik seperti akurasi, presisi, recall, dan F1-score. Confusion matrix digunakan untuk menganalisis distribusi kesalahan klasifikasi, seperti false positives dan false negatives. Hasil evaluasi menunjukkan bahwa model dengan dataset yang seimbang memberikan performa yang lebih baik dibandingkan dengan dataset yang tidak seimbang.

## Testing

Pada tahap testing, model diuji menggunakan data uji untuk mengevaluasi performanya dalam mengklasifikasikan data baru. Hasil pengujian menunjukkan performa yang sangat baik dengan nilai akurasi, presisi, recall, dan F1-score yang tinggi. Hal ini menunjukkan bahwa algoritma SGD efektif dalam mendeteksi kanker paru-paru.

## Pelaporan

Tahap terakhir adalah menyusun laporan penelitian yang berisi penjelasan lengkap tentang proses dan hasil penelitian. Laporan ini terdiri dari beberapa bab utama, termasuk pendahuluan, landasan teori, metodologi penelitian, hasil dan pembahasan, serta kesimpulan. Semua hasil analisis, grafik, dan metrik evaluasi disajikan secara terstruktur. Referensi jurnal dan dataset yang digunakan juga disertakan untuk memastikan validitas dan transparansi penelitian. Laporan ini dirancang untuk memberikan panduan bagi penelitian lanjutan di bidang yang sama.

# BAB IV HASIL DAN PEMBAHASAN

Bagian ini membahas tentang pembuatan model, klasifikasi dan hasil pengujian dan model klasifikasi yang telah dibangun. Pada penelitian ini menggunakan dataset kanker paru-paru sebanyak 309 yang terbagi menjadi 270 penderita kanker paru paru dan 39 yang bukan penderita kanker paru paru. Contoh data ditunjukan pada Tabel 2.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| No. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| 0 | M | 69 | 1 | 2 | 2 | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | YES |
| 1 | M | 74 | 2 | 1 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 2 | 2 | 2 | YES |
| 2 | F | 59 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 2 | 1 | 2 | NO |
| … | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … |
| 306 | M | 58 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | YES |
| 307 | M | 67 | 2 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 2 | 2 | 2 | 1 | 2 | YES |
| 308 | M | 62 | 1 | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 1 | YES |

Table 2. Contoh Dataset

|  |  |
| --- | --- |
| Features | : 1. Gender |
|  | 2. Age |
|  | 3. Smoking |
|  | 4. Yellow Fingers |
|  | 5. Anxiety |
|  | 6. Peer Pressure |
|  | 7. Chronic Disease |
|  | 8. Fatigue |
|  | 9. Allergy |
|  | 10. Wheezing |
|  | 11. Alcohol Consuming |
|  | 12. Coughing |
|  | 13. Shortness of Breath |
|  | 14. Swallowing Difficulty |
|  | 15. Chest Pain |
| Target | : 16. Lung Cancer |
| Keterangan | : NO = 1 |
|  | YES = 2 |

## Preprocessing

### 4.1.1 Normalisasi

Pada tahap ini melakukan normalisasi pada features AGE menggunakan teknik MinMaxScaler dengan menyederhanakan data pada features menjadi antara 1 hingga 2 seperti terlihat pada Tabel 3 dibawah ini.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| No. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| 0 | M | 1.7272 | 1 | 2 | 2 | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | YES |
| 1 | M | 1.8030 | 2 | 1 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 2 | 2 | 2 | YES |
| 2 | F | 1.5757 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 2 | 1 | 2 | NO |
| … | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … |
| 306 | M | 1.5606 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | YES |
| 307 | M | 1.6969 | 2 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 2 | 2 | 2 | 1 | 2 | YES |
| 308 | M | 1.6212 | 1 | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 1 | YES |

Table 3. Dataset Setelah Normalisasi

### Transformasi

Mengubah features kategori seperti GENDER (dengan nilai "M menjadi 1" dan "F menjadi 2") dan LUNG CANCER (dengan nilai "YES menjadi 2" dan "NO menjadi 1") diubah menjadi tipe numerik menggunakan label encoding seperti terlihat pada Tabel 4 dibawah ini.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| No. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| 0 | 1 | 1.7272 | 1 | 2 | 2 | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 1 | 1 | 1.8030 | 2 | 1 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |
| 2 | 2 | 1.5757 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 2 | 1 | 2 | 1 |
| … | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … |
| 306 | 1 | 1.5606 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 2 |
| 307 | 1 | 1.6969 | 2 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 2 | 2 | 2 | 1 | 2 | 2 |
| 308 | 1 | 1.6212 | 1 | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 1 | 2 |

Table 4. Dataset Setelah Transformasi

## Memisahkan Fitur dan Membagi Data

Setelah melakukan preprocessing, selanjutnya membuat model dengan algoritma SGD. Sebelum itu dianjurkan untuk membuat sebuuah variabel X (features) dan variabel y (Target) seperti Tabel 5 dibawah ini.

|  |  |
| --- | --- |
| Features | 1. Gender |
| 2. Age |
| 3. Smoking |
| 4. Yellow Fingers |
| 5. Anxiety |
| 6. Peer Pressure |
| 7. Chronic Disease |
| 8. Fatigue |
| 9. Allergy |
| 10. Wheezing |
| 11. Alcohol Consuming |
| 12. Coughing |
| 13. Shortness of Breath |
| 14. Swallowing Difficulty |
| 15. Chest Pain |
| Target | 16. Lung Cancer |

Table 5. Definisi x dan y di Dataset

Langkah selanjutnya adalah membagi data menjadi dua yaitu data train dan data test, dengan pembagian data train sebesar 80% dan data test sebesar 20%.

## Membangun Model Dengan Algoritma SGD

Kami membangun model menggunakan algoritma Stochastic Gradient Descent (SGD) dengan bantuan pustaka pemrograman, seperti scikit-learn di Python, dan menetapkan jumlah iterasi sebanyak 1000 untuk melatih model. Dalam proses ini, model diperbarui secara bertahap menggunakan gradien fungsi kerugian dari subset data yang dipilih secara acak pada setiap iterasi, sehingga menghasilkan parameter model yang optimal. Parameter seperti learning rate juga disesuaikan untuk memastikan konvergensi menuju solusi terbaik. Setelah model dilatih, kami melakukan evaluasi menggunakan metode k-fold cross-validation untuk mengukur performa model secara menyeluruh dan memastikan generalisasi yang baik pada data yang tidak terlihat.

### Pengujian Dengan k-Fold Cross Validation

k-Fold Cross Validation adalah metode evaluasi model machine learning yang membagi dataset menjadi beberapa lipatan (fold) yang sama besar (di pengujian ini kami menggunakan 10 fold). Metode ini digunakan untuk memastikan bahwa performa model lebih stabil, akurat, dan representatif. Berikut adalah langkah-langkah bagaimana k-Fold Cross Validation dilakukan:

1. Pembagian Data

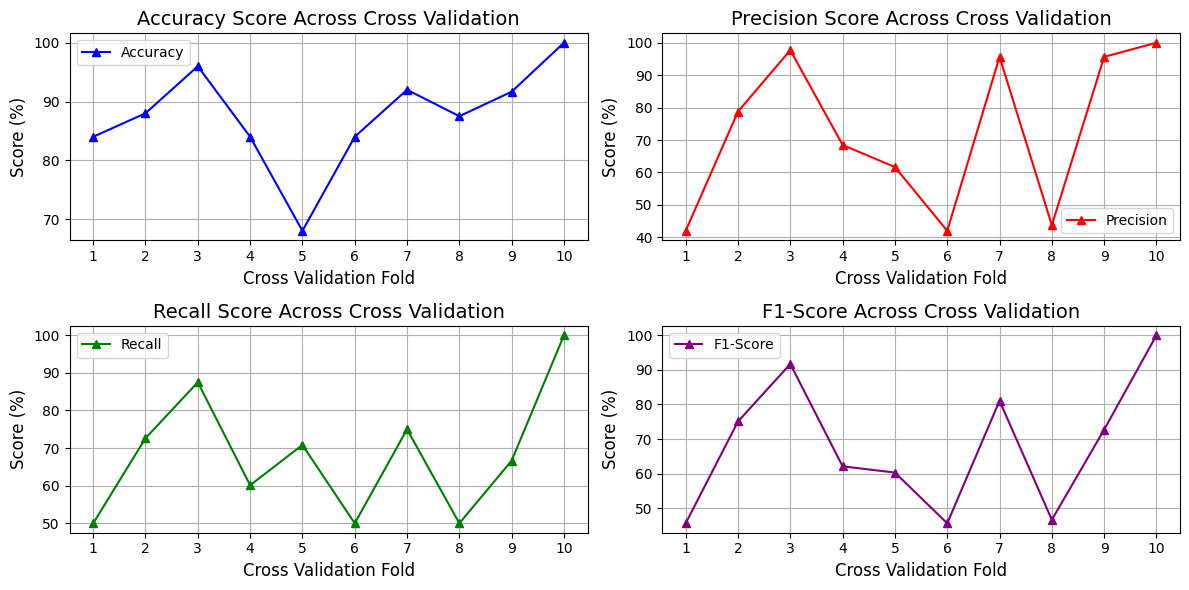
* Data dibagi menjadi 10 lipatan (fold), di mana setiap lipatan memiliki ukuran yang hampir sama.
* Pada setiap iterasi, satu lipatan digunakan sebagai data uji, sementara 9 lipatan lainnya digunakan sebagai data latih.

1. Proses Pengujian

* Model dilatih menggunakan data latih dari 9 lipatan.
* Model diuji pada 1 lipatan yang tidak digunakan untuk pelatihan.
* Proses ini diulang sebanyak 10 kali, sehingga setiap lipatan bergantian menjadi data uji dan data latih.

Tujuan utamanya adalah untuk mengevaluasi performa model, bukan untuk menghasilkan model akhir yang siap digunakan. Model tidak disimpan setelah setiap fold; evaluasi dilakukan hanya untuk menilai seberapa baik model pada berbagai subset data.

Setelah melakukan pengujian cross validation dengan nilai K yakni 10, maka diperoleh skor pada model dengan perhitungan accuracy, precision, recall, f1-score seperti pada Gambar 4 dibawah ini



Gambar 4. Hasil Pengujian Dengan k-Fold Cross Validation

Perolehan hasil diatas dapat dipastikan kembali dengan mencari nilai rata-rata dari masing-masing skor pada performa diatas yang dapat dilihat pada Tabel 6 dibawah ini.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Iteration | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| Accuracy | 0.840 | 0.880 | 0.960 | 0.840 | 0.680 | 0.840 | 0.920 | 0.875 | 0.916 | 1.000 | 0.875 |
| Precision | 0.420 | 0.787 | 0.977 | 0.684 | 0.616 | 0.420 | 0.956 | 0.437 | 0.956 | 1.000 | 0.725 |
| Recall | 0.500 | 0.726 | 0.875 | 0.601 | 0.708 | 0.500 | 0.750 | 0.500 | 0.666 | 1.000 | 0.682 |
| F1-Score | 0.456 | 0.750 | 0.916 | 0.621 | 0.603 | 0.456 | 0.810 | 0.466 | 0.727 | 1.000 | 0.680 |

Table 6. Rata-rata Performa

Pada Tabel 6 diperolelah hasil skor atau performa dari dataset kanker paru-paru yang berjumlah 309, dimana hasil akurasi 0.875, presisi 0.725, recall 0.682 dan f1-score 0.680.

### Pengujian Dengan k-Fold Cross Validation Menggunakan SMOTE

Dikarenakan jumlah skor atau performa dari pengujian k-Fold Validation terhitung rendah, maka kami akan meningkatkan akurasi dengan cara menyeimbangkan target kelas menggunakan metode SMOTE. Cara kerja SMOTE adalah pertama, SMOTE mengidentifikasi kelas yang memiliki jumlah lebih sedikit (kelas minoritas). Lalu untuk setiap contoh dalam kelas minoritas, SMOTE memilih sejumlah titik data terdekat (k-nearest neighbors) berdasarkan metrik jarak seperti Euclidean distance.

Distribusi kelas sebelum SMOTE pada kolom LUNG\_CANCER:

YES : 210

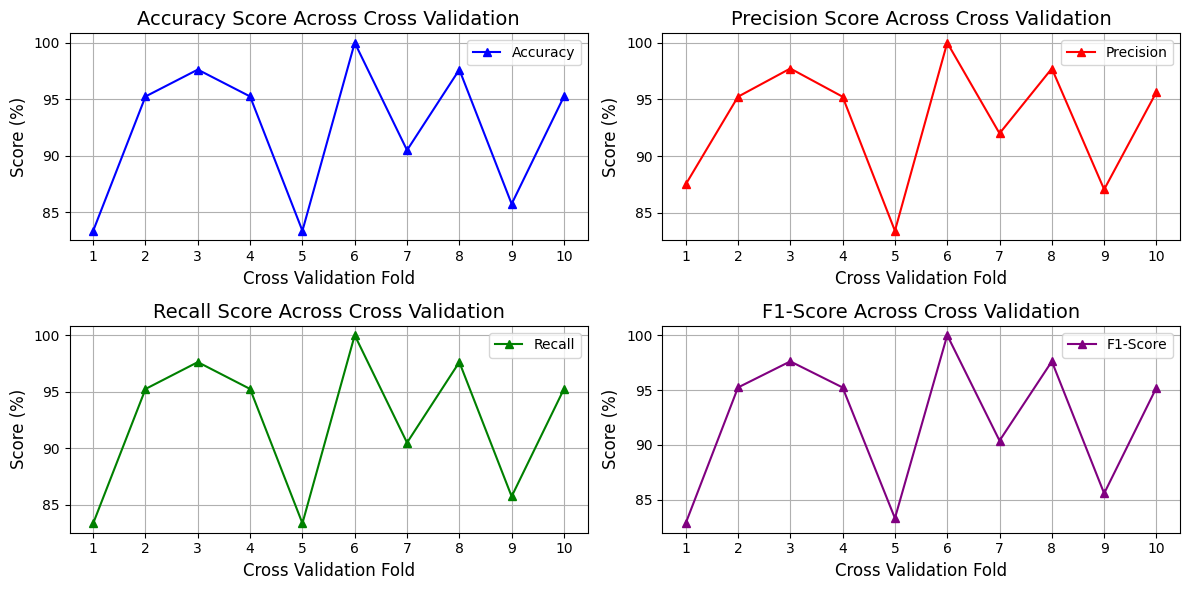
NO : 37

Distribusi kelas setelah SMOTE pada kolom LUNG\_CANCER:

YES : 210

NO : 210

Setelah dilakukan oversampling menggunakan SMOTE, selanjutnya melakukan pengujian cross validation dengan nilai K yakni 10, maka diperoleh skor pada model dengan perhitungan accuracy, precision, recall, f1-score seperti pada Gambar 5 dibawah ini



Gambar 5. Hasil Pengujian Dengan k-Fold Cross Validation setelah SMOTE

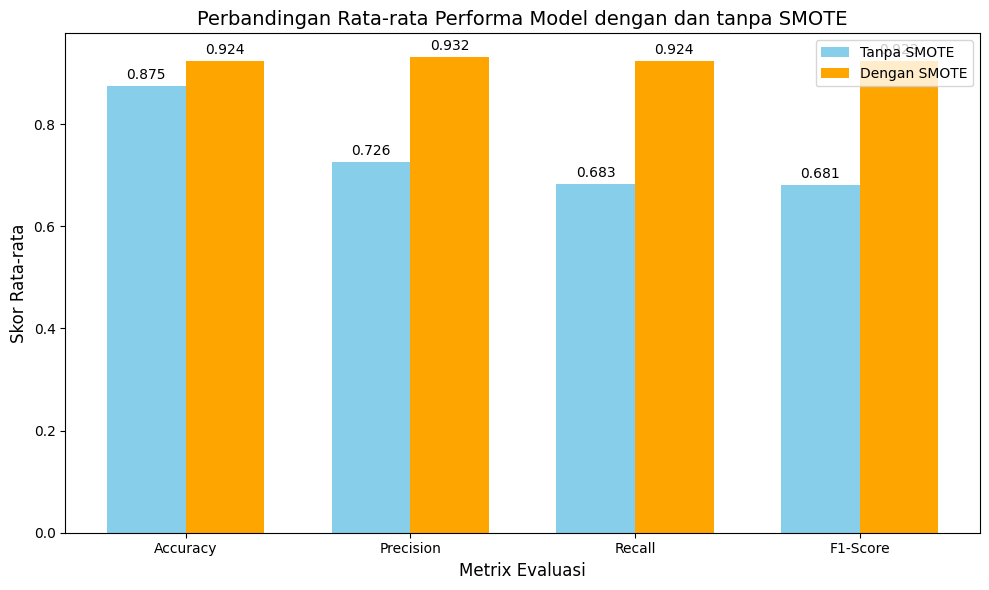
Perolehan hasil diatas dapat dipastikan kembali dengan mencari nilai rata-rata dari masing-masing skor pada performa diatas yang dapat dilihat pada Tabel 7 dibawah ini.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Iteration | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| Accuracy | 0.833 | 0.952 | 0.960 | 0.952 | 0.833 | 1.000 | 0.904 | 0.976 | 0.857 | 0.952 | 0.923 |
| Precision | 0.875 | 0.952 | 0.977 | 0.952 | 0.834 | 1.000 | 0.920 | 0.977 | 0.870 | 0.956 | 0.931 |
| Recall | 0.833 | 0.952 | 0.875 | 0.952 | 0.833 | 1.000 | 0.904 | 0.976 | 0.857 | 0.952 | 0.923 |
| F1-Score | 0.828 | 0.952 | 0.916 | 0.952 | 0.833 | 1.000 | 0.903 | 0.976 | 0.855 | 0.952 | 0.923 |

Table 7. Rata-rata Performa setelah SMOTE

Pada Tabel 7 diperolelah hasil skor atau performa meningkat, dimana hasil akurasi 0.924, presisi 0.932, recall 0.924 dan f1-score 0.923.

### Perbandingan Performa Model dengan dan tanpa SMOTE



Gambar 6. Perbandingan Performa

Berdasarkan hasil analisis performa model, model training yang menggunakan teknik SMOTE (Synthetic Minority Oversampling Technique) menunjukkan hasil yang lebih tinggi dan unggul dalam beberapa metrik evaluasi, termasuk Akurasi, Presisi, Recall, dan F1-Score, dibandingkan dengan model training tanpa SMOTE.

## Testing

Pengujian klasifikasi dilakukan dengan mengukur nilai akurasi, presisi, recall dan f1-score dari model yang telah dibuat sebelumnya berdasarkan algoritma SGD, dengan menggunakan data test. Hasil performansi ini menggunakan metode confusion matrix.

|  |  |
| --- | --- |
| Gambar 7. Classification Report | Gambar 8. Confusion Matrix |

Gambar 6 dan Gambar 7 menampilkan classification report dan confusion matrix berdasarkan hasil testing menggunakan data uji dari model yang telah dibuat, menghasilkan nilai performa akurasi sebesar 0.98, presisi 0.98, recall 0.98 dan f1-score 0.98.

# BAB IV KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan mengenai klasifikasi kanker paru-paru menggunakan algoritma Stochastic Gradient Descent (SGD), Algoritma Stochastic Gradient Descent (SGD) terbukti efektif dalam melakukan klasifikasi kanker paru-paru. Model yang dikembangkan mampu mengenali pola data dengan akurasi tinggi setelah diterapkan preprocessing, normalisasi, dan transformasi data. Dari pengujian yang dilakukan, model SGD menghasilkan nilai akurasi sebesar 0.98, precision 0.98, recall 0.98, dan F1-score 0.98. Hal ini menunjukkan bahwa model memiliki performa yang sangat baik dalam mengklasifikasikan data kanker paru-paru.

Tahap preprocessing, termasuk normalisasi data dengan Min-Max Scaling dan transformasi data menggunakan label encoding, membantu meningkatkan kualitas data sehingga lebih siap untuk diproses oleh model. Penerapan teknik SMOTE berhasil menangani masalah ketidakseimbangan kelas dalam dataset. Dengan menambah jumlah data sintetik untuk kelas minoritas, distribusi data menjadi lebih merata, yang berdampak pada peningkatan performa model.

Penelitian ini menunjukkan potensi machine learning, khususnya algoritma SGD, dalam mendukung deteksi dini kanker paru-paru. Implementasi metode ini diharapkan dapat berkontribusi dalam upaya pengurangan angka kematian akibat kanker paru-paru melalui sistem prediksi yang lebih akurat.

# DAFTAR PUSTAKA

[1] L. Sari, A. Romadloni, dan R. Listyaningrum, “Penerapan Data Mining dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest,” *Infotekmesin*, vol. 14, no. 1, hlm. 155–162, Jan 2023, doi: 10.35970/infotekmesin.v14i1.1751.

[2] J. Sinurat, “Bulletin of Information Technology (BIT) Jaringan Saraf Tiruan Diagnosa Penyakit Kanker Paru-Paru Menggunakan Metode Hebb Rule,” *Bulletin of Information Technology (BIT)*, vol. 2, no. 1, hlm. 20–27, 2021.

[3] F. T. Admojo dan Y. I. Sulistya, “Analisis performa algoritma Stochastic Gradient Descent (SGD) dalam mengklasifikasi tahu berformalin,” *Indonesian Journal of Data and Science (IJODAS)*, vol. 3, no. 1, 2022.

[4] G. Ben Arous, R. Gheissari, B. Arous, dan J. Gheissari, “Online stochastic gradient descent on non-convex losses from high-dimensional inference,” 2021.

[5] K. M. Arrazaq, D. Retno, S. Saputro, dan R. Setiyowati, “Penerapan Stochastic Gradient Descent Support Vector Regression pada Data Laju Pertumbuhan Produk Domestik Bruto di Indonesia.”

[6] S. Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti, A. Nikmatul Kasanah, U. Pujianto, T. Elektro, F. Teknik, dan U. Negeri Malang, “Terakreditasi SINTA Peringkat 2 Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN,” *Jurnal RESTI*, vol. 1, no. 3, hlm. 196–201, 2021.

[7] M. Raja dan M. Y. M. Parvees, “International Conference on IoT based Control Networks and Intelligent Systems (ICICNIS 2020) Stochastic Gradient Descent with Logistic Regression Technique for Medical Data Classification.” [Daring]. Tersedia pada: https://ssrn.com/abstract=3769853