

TECHNICAL REPORT
MACHINE LEARNING MODEL USING PYTHON



Disusun Oleh :
Lutfiana Erniasari
1103204029

PROGRAM STUDI S1 TEKNIK KOMPUTER
FAKULTAS TEKNIK ELEKTRO
UNIVERSITAS TELKOM
2023

1. Pendahuluan

Kanker Payudara merupakan salah satu jenis kanker yang terjadi ketika sel-sel abnormal di payudara berkembang dan tumbuh dengan tidak terkendali. Kanker payudara bisa terjadi pada wanita dan pria, tetapi lebih umum pada wanita. Beberapa faktor risiko yang dapat meningkatkan kemungkinan seseorang mengalami kanker payudara adalah faktor genetik, usia, riwayat keluarga, serta gaya hidup yang tidak sehat. Tanda-tanda kanker payudara bisa berupa benjolan pada payudara atau ketiak, perubahan bentuk atau ukuran payudara, kulit yang terlihat seperti kulit jeruk, dan puting yang terlihat abnormal. Pemeriksaan payudara secara teratur dan tes medis dapat membantu mendeteksi kanker payudara lebih awal dan meningkatkan kemungkinan kesembuhan.

Sedangkan Machine Learning sendiri merupakan salah satu metode dalam bidang kecerdasan buatan (Artificial Intelligence) yang memungkinkan komputer belajar dari pengalaman sebelumnya tanpa harus di-program secara eksplisit. Dalam machine learning, komputer mempelajari pola dari data yang diberikan, sehingga mampu melakukan tugas tertentu seperti klasifikasi atau pengenalan pola. Lalu Machine Learning Model sendiri merupakan program komputer atau algoritma matematika yang mampu mempelajari pola dari data dan menghasilkan prediksi atau output berdasarkan input yang diberikan, tujuannya untuk dapat digunakan dalam berbagai tugas seperti klasifikasi, regresi, pengenalan wajah, atau prediksi harga saham, tergantung dari jenis model dan data yang digunakan. Sebelum digunakan, machine learning model harus melalui proses pelatihan dengan data yang telah dilabeli atau memiliki output yang diketahui, setelah proses pelatihan model machine learning akan menghasilkan parameter yang optimal untuk menghasilkan prediksi yang akurat. Model ini dapat digunakan untuk memprediksi output baru dari data input yang belum pernah dilihat sebelumnya.

Di era digitalisasi saat ini, deteksi dini kanker payudara dapat dilakukan dengan bantuan machine learning. Beberapa algoritma machine learning yang digunakan dalam technical report ini adalah Decision Tree, Random Forest, dan Self-Training, yang mampu menganalisis gambar mamografi dan data klinis untuk mengklasifikasikan kanker payudara. Python dan Scikit-Learn digunakan untuk membangun model machine learning yang akurat. Pembuatan machine learning menggunakan dataset breast cancer dari scikit merupakan salah satu contoh implementasi teknologi yang digunakan untuk mendukung pengambilan keputusan di

bidang kesehatan. Dataset ini adalah salah satu dataset kanker payudara yang paling populer digunakan untuk melatih model machine learning. Dataset breast cancer dari scikit dibangun oleh tim peneliti kesehatan yang terdiri dari Dr. William H. Wolberg, Dr. W. Nick Street, dan Olvi L. Mangasarian. Dataset ini berisi informasi tentang tumor payudara dari 569 pasien perempuan dengan 30 fitur yang diukur dari gambar citra digital. Pemanfaatan dataset breast cancer dari scikit dalam pembuatan model machine learning bertujuan untuk membantu memprediksi apakah tumor tersebut bersifat jinak atau ganas. Model machine learning dapat mempelajari pola dari data yang ada untuk membangun algoritma yang dapat digunakan untuk mengklasifikasikan tumor payudara menjadi kategori yang sesuai.

2. Pengolahan dan Visualisasi Data

Langkah pertama yang dilakukan adalah mengumpulkan dataset menggunakan fungsi `load_breast_cancer()` dari library `scikit-learn`. Dataset ini kemudian dikonversi menjadi `Pandas DataFrame` dengan menggunakan nama fitur sebagai nama kolom. Selanjutnya, kolom 'target' ditambahkan ke Data Frame dengan label 'malignant' dan 'benign' yang cocok dengan nilai target yang sebenarnya. Kemudian, data dipisahkan berdasarkan target ('malignant' atau 'benign') yang akan digunakan untuk analisis dan visualisasi lebih lanjut. Sejumlah visualisasi data dibuat menggunakan library `Matplotlib` dan `Seaborn`, yang menampilkan hubungan antara beberapa fitur dalam dataset dan memberikan pemahaman yang lebih baik tentang pola dalam data. Setelah semua langkah persiapan dataset selesai, proses pengolahan dataset dimulai.

Pertama, fitur dan target dipisahkan ke dalam variabel `x` dan `y`, masing-masing yang mewakili nilai fitur dan target. Kedua, dataset dibagi menjadi set pelatihan dan pengujian dengan proporsi 60% untuk data pelatihan dan 40% untuk data pengujian. Set data pelatihan digunakan untuk melatih berbagai model, seperti `Decision Tree Classifier`, `Random Forest Classifier`, dan `Self Training Classifier`, sedangkan set data pengujian digunakan untuk mengukur kinerja model tersebut. Setiap model dilatih dengan set pelatihan dan kemudian diuji dengan set pengujian untuk mengukur akurasi kinerja masing-masing model terhadap data yang belum dilihat sebelumnya.

Selain itu visualisasi data digunakan untuk membantu pemahaman dataset dengan lebih baik. Pertama, jumlah data masing-masing target ('malignant' atau 'benign') ditampilkan dalam diagram batang menggunakan library `Seaborn`. Selanjutnya, beberapa fitur yang penting ditampilkan dalam diagram `Scatterplot pairwise` menggunakan `PairGrid` juga dari library `Seaborn`. Setiap `Scatterplot pairwise`

menunjukkan hubungan antara dua fitur tertentu dalam dataset, dan diwarnai sesuai dengan nilai target. Selain itu, histogram untuk setiap fitur juga ditampilkan menggunakan Matplotlib. Histogram menunjukkan distribusi data untuk setiap fitur, berdasarkan jenis target ('malignant' atau 'benign'). Histogram juga digunakan untuk memvisualisasikan bagaimana setiap fitur mempengaruhi target. Histogram dengan warna yang berbeda digunakan untuk membedakan data masing-masing target ('malignant' atau 'benign').

Dalam kasus dataset ini, visualisasi data membantu untuk menarik kesimpulan tentang bagaimana setiap fitur mempengaruhi target. Misalnya, dari histogram pada fitur 'worst texture', kita dapat melihat bahwa pasien dengan target 'malignant' cenderung memiliki distribusi yang lebih tinggi pada setiap tingkat tekstur terburuk dibandingkan dengan pasien dengan target 'benign'. Ini menunjukkan bahwa fitur ini mungkin berguna dalam membedakan kelas pada kasus ini. Selain itu, Scatterplot pairwise dengan 'radius error' dan 'worst radius' menunjukkan bahwa kelas 'malignant' memiliki nilai error radius yang lebih besar dan 'worst radius' yang lebih besar daripada kelas 'benign'. Hal ini bertentangan dengan perbedaan yang ditunjukkan pada histogram 'mean radius', sehingga menunjukkan pentingnya untuk mempertimbangkan setiap fitur secara terpisah dalam analisis.

3. Kesimpulan

Program ini menggunakan beberapa model machine learning seperti Regresi Logistik (Self training), Random Forest, Pohon Keputusan dan pembagian data menjadi set pelatihan dan pengujian. Pada hasil running dari ketiga model tersebut, model random forest menempati akurasi tertinggi sebesar 92.98%, diikuti oleh model random forest sebesar 92.54% dan model self-training mempunyai akurasi yang sama seperti random forest sebesar 92.54%.

Model-model AI, seperti Random Forest, Decision Tree, dan Self Training, dapat membantu dokter dalam deteksi kanker payudara dengan tingkat akurasi yang tinggi. Model-model ini dapat memproses data dari berbagai faktor seperti usia, ukuran tumor, dan jumlah sel tumor untuk memberikan hasil yang dapat membantu dokter membuat keputusan yang lebih tepat. Model-model AI dapat memberikan manfaat yang signifikan bagi dokter dalam mendiagnosis kanker payudara. Dengan analisis dan pengolahan data yang cepat dan akurat, dokter dapat memperoleh informasi yang lebih akurat dan cepat dalam membuat keputusan diagnosa dan perawatan yang tepat untuk pasien.

Penggunaan teknologi AI juga dapat membantu dokter mendeteksi kanker payudara pada tahap awal, sehingga meningkatkan peluang kesembuhan pasien. Dengan kemampuan analisis data yang unggul, model-model AI memberikan alternatif baru bagi dokter untuk memperoleh informasi yang lebih tepat dan akurat dalam mengambil keputusan mengenai diagnosa dan perawatan pasien. Terutama dalam deteksi kanker payudara, model-model AI dapat membantu dokter dengan memberikan hasil yang cepat dan akurat sehingga tindakan selanjutnya dapat diambil dengan tepat dan waktu yang sesuai. Oleh karena itu, penggunaan teknologi AI dapat memberikan manfaat besar bagi dokter dan pasien dalam mengatasi masalah kesehatan yang krusial.