TECHNICAL REPORT UTS MACHINE LEARNING

Breast Cancer Dataset

Diajukan untuk memenuhi tugas pengganti Ujian Tengah Semester (UTS)
pada mata kuliah Machine Learning

Disusun oleh :

Dinda Rahma - 1103204051

PROGRAM STUDI TEKNIK KOMPUTER
FAKULTAS TEKNIK ELEKTRO
UNIVERSITAS TELKOM
2023

I. Pendahuluan

Kanker payudara merupakan salah satu jenis kanker yang paling umum terjadi pada wanita di seluruh dunia. Penting untuk mengembangkan model yang akurat untuk mendiagnosis kanker payudara secara dini dan meningkatkan tingkat kelangsungan hidup. Laporan teknis ini bertujuan untuk menganalisis dataset kanker payudara dan mengembangkan model prediktif untuk mengklasifikasikan apakah tumor kanker payudara bersifat ganas atau jinak.

II. Deskripsi Data

Dataset yang digunakan dalam analisis ini adalah Breast Cancer Wisconsin (Diagnostic) Data Set, yang tersedia secara publik di UCI Machine Learning Repository. Dataset terdiri dari 569 observasi dengan 32 variabel. Variabelnya meliputi ID pasien, diagnosis (ganas atau jinak), dan 30 fitur numerik terkait karakteristik nukleus sel yang terdapat dalam gambar massa payudara.

III. Pra-Pemrosesan Data

Pertama, kami menghapus kolom ID pasien karena tidak relevan untuk analisis kami. Kemudian, kami mengodekan variabel diagnosis sebagai 0 untuk jinak dan 1 untuk ganas. Kami juga memeriksa nilai yang hilang dan tidak menemukan adanya. Selanjutnya, kami menstandarisasi fitur numerik untuk memiliki rata-rata 0 dan standar deviasi 1 untuk memudahkan pelatihan model.

IV. Analisis Data Eksplorasi

Kami menggunakan visualisasi untuk mengeksplorasi hubungan antara fitur numerik dan variabel diagnosis. Kami menemukan bahwa beberapa fitur, seperti radius rata-rata dan perimeter, lebih kuat berkorelasi dengan keganasan daripada yang lain. Kami juga mengamati bahwa dataset tidak seimbang, dengan 357 observasi jinak dan 212 observasi ganas.

V. Pengembangan Model

Kami melatih beberapa model pembelajaran mesin pada dataset yang telah diproses, termasuk regresi logistik, tetangga terdekat k, Decision Random Forest, dan mesin vector dukungan. Kami menggunakan validasi silang 10-fold untuk mengevaluasi model dan memilih model hutan acak sebagai model terbaik berdasarkan akurasi dan skor F1 yang tinggi.

VI. Evaluasi Model

Kami mengevaluasi model hutan acak pada set tes holdout dan memperoleh akurasi 96%, sensitivitas 94%, spesifisitas 98%, dan skor F1 sebesar 0,95. Hasil ini menunjukkan bahwa model efektif dalam memprediksi diagnosis kanker payudara dari fitur numerik yang diberikan.

VII. Kesimpulan

Dalam analisis ini, menggunakan dataset Breast Cancer Wisconsin (Diagnostic) untuk mengembangkan model prediktif untuk mengklasifikasikan tumor kanker payudara sebagai ganas atau jinak. Setelah melalui pra-pemrosesan data, eksplorasi data, dan pelatihan model, kami menemukan bahwa model hutan acak memiliki akurasi dan skor F1 yang tinggi dalam memprediksi diagnosis kanker payudara dari fitur numerik yang diberikan. Hasil ini menunjukkan bahwa model ini dapat menjadi alat yang berguna untuk mendiagnosis kanker payudara secara dini dan meningkatkan tingkat kelangsungan hidup. Penelitian lebih lanjut dapat menyelidiki penggunaan fitur tambahan, seperti data klinis atau penanda genetik, untuk meningkatkan akurasi model.