TECHNICAL REPORT UTS MACHINE LEARNING

Breast Cancer Dataset

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

Disusun oleh :

Achmad Rionov Faddillah Ramadhan - 1103204030

PROGRAM STUDI TEKNIK KOMPUTER
FAKULTAS TEKNIK ELEKTRO
UNIVERSITAS TELKOM
2023

I. Pendahuluan

Proyek untuk Ujian Tengah Semester ini bertujuan untuk menganalisis dataset kanker payudara dan membangun model pembelajaran mesin yang dapat memprediksi apakah suatu pasien menderita kanker payudara jinak atau ganas. Analisis dan model ini dapat membantu dokter dalam diagnosis dan memberikan rekomendasi pengobatan yang tepat kepada pasien. Proyek ini menggunakan Python dan beberapa pustaka pembelajaran mesin seperti Scikit-learn.

II. Deskripsi Data

Dataset yang digunakan dalam proyek ini adalah dataset kanker payudara Wisconsin. Dataset ini terdiri dari 569 sampel dan 30 fitur, di mana 212 sampel adalah kanker payudara ganas dan 357 sampel lainnya adalah kanker payudara jinak. Fitur-fitur tersebut mencakup ukuran dan bentuk sel-sel di gambaran mikroskopik biopsi payudara, seperti jari-jari terbesar (radius), kehalusan tekstur, dan luas area.

III. Pra-Pemrosesan Data

Pra-pemrosesan data dilakukan menggunakan pandas. Pandas adalah salah satu *library* Python yang digunakan untuk analisis data. Pandas digunakan dalam tahap pra-pemrosesan data karena kemampuannya dalam membaca dan menulis berbagai format data, seperti CSV, Excel, SQL, dan lainnya. Dengan menggunakan Pandas, kita dapat membaca dataset yang belum terstruktur dan mengubahnya menjadi suatu format yang lebih mudah digunakan dalam pemrosesan data.

Sebelum melakukan analisis data, kami melakukan beberapa langkah pra-pemrosesan data. Langkah-langkah ini termasuk memeriksa dan menangani nilai-nilai yang hilang, mengubah label menjadi format numerik, serta membagi dataset menjadi set pelatihan dan pengujian. Pra-pemrosesan data dilakukan untuk membersihkan dataset dari nilai yang hilang dan fitur-fitur yang tidak relevan. Selain itu, dilakukan juga normalisasi data untuk memastikan bahwa semua fitur memiliki skala yang sama.

IV. Analisis Data Eksplorasi

Bagian ini berisi analisis eksplorasi data menggunakan tiga jenis model berbeda: Decision Tree, Random Forest, dan Self-Training Classifier.

1. Decision Tree

Decision Tree adalah algoritma pembelajaran mesin yang digunakan dalam masalah klasifikasi. Algoritma ini memecah dataset menjadi beberapa subset berdasarkan nilai-nilai tertentu dari fitur-fitur dalam dataset. Decision Tree membentuk struktur seperti pohon keputusan yang terdiri dari simpul dan cabang. Pada simpul, algoritma Decision Tree membuat keputusan berdasarkan fitur-fitur yang ada. Sedangkan pada cabang, algoritma Decision Tree memecah dataset menjadi subset yang lebih kecil.

Dalam model Decision Tree, kita menggunakan pohon keputusan untuk memprediksi apakah suatu kasus mengidap kanker payudara jinak atau ganas.

2. Random Forest

Decision Tree adalah algoritma pembelajaran mesin yang digunakan dalam masalah klasifikasi. Algoritma ini memecah dataset menjadi beberapa subset berdasarkan nilai-nilai tertentu dari fitur-fitur dalam dataset. Decision Tree

membentuk struktur seperti pohon keputusan yang terdiri dari simpul dan cabang. Pada simpul, algoritma Decision Tree membuat keputusan berdasarkan fitur-fitur yang ada. Sedangkan pada cabang, algoritma Decision Tree memecah dataset menjadi subset yang lebih kecil.

Model Random Forest juga menggunakan pohon keputusan, tetapi membangun banyak pohon dan menggabungkan hasil dari masing-masing pohon.

3. Self-Training

Self-Training adalah metode semi-supervised learning yang digunakan dalam klasifikasi data. Metode ini bekerja dengan menggunakan beberapa data yang terlabel dan beberapa data yang tidak terlabel untuk menghasilkan model klasifikasi. Dalam Self-Training, model klasifikasi pertama kali dilatih menggunakan data terlabel. Kemudian, model tersebut digunakan untuk mengklasifikasikan data yang tidak terlabel. Data yang dihasilkan oleh model klasifikasi tersebut kemudian ditambahkan ke data terlabel dan digunakan kembali untuk melatih model klasifikasi. Proses ini diulang beberapa kali hingga mencapai tingkat akurasi yang diinginkan.

Model Self-Training Classifier merupakan jenis semi-supervised learning yang melibatkan klasifikasi iteratif dan penambahan data baru ke set pelatihan setelah setiap iterasi. Tujuan dari analisis ini adalah untuk membandingkan performa ketiga model dalam memprediksi hasil diagnosis kanker payudara.

Model Self-Training Classifier merupakan jenis semi-supervised learning yang melibatkan klasifikasi iteratif dan penambahan data baru ke set pelatihan setelah setiap iterasi. Tujuan dari analisis ini adalah untuk membandingkan performa ketiga model dalam memprediksi hasil diagnosis kanker payudara.

V. Pengembangan Model

Dalam bagian ini, kita membangun model dengan menggunakan algoritma Decision Tree, Random Forest, dan Self-Training Classifier. Model-model ini ditraining pada dataset yang telah diproses sebelumnya untuk memprediksi diagnosis kanker payudara berdasarkan fitur-fitur tertentu dari pasien. Selanjutnya, kita menghasilkan probabilitas hasil prediksi dari masing-masing model.

VI. Evaluasi Model

Pada bagian ini, kita mengevaluasi performa model yang telah dibangun sebelumnya menggunakan beberapa metrik evaluasi seperti akurasi, presisi, recall, dan F1 score. Metrik evaluasi ini membantu kita untuk menilai performa model dalam memprediksi hasil diagnosis kanker payudara dan membandingkan performa model yang berbeda. Selain itu, kita juga melakukan visualisasi pada model Decision Tree dan Self-Training Classifier untuk membantu memahami bagaimana model bekerja.

VII. Kesimpulan

Berdasarkan hasil analisis dan evaluasi yang telah dilakukan, dapat disimpulkan bahwa model Random Forest memiliki performa yang lebih baik dibandingkan dengan model Decision Tree dan Self-Training Classifier dalam memprediksi hasil diagnosis kanker payudara. Hal ini menunjukkan bahwa penggunaan Random Forest sebagai model dapat membantu meningkatkan akurasi dan keandalan dalam diagnosis kanker payudara. Namun, kita perlu diingat bahwa setiap model memiliki kelebihan dan kekurangan tersendiri dan dapat memberikan hasil yang berbeda-beda tergantung pada dataset yang digunakan dan parameter yang diberikan.