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.

PERBAIKAN TECHNICAL REPORT

I. Pengertian Machine Learning

Machine learning (ML) adalah cabang dari ilmu komputer yang menggunakan algoritma untuk menghasilkan model statistik dari data, yang dapat digunakan untuk melakukan prediksi atau mengambil keputusan tanpa diprogram secara eksplisit. Dalam machine learning, mesin belajar dari data dengan cara menemukan pola dan hubungan dalam data, dan kemudian menggunakan pola dan hubungan tersebut untuk membuat prediksi atau keputusan. Bisa dibilang juga bahwa machine learning adalah cara bagi komputer untuk belajar dan membuat prediksi atau keputusan dengan menggunakan data yang telah dipelajari sebelumnya, mirip seperti manusia yang belajar dari pengalaman.

II. Model – Model Machine Learning

1. Regresi Linier: Model regresi linier digunakan untuk memodelkan hubungan antara variabel independen (input) dan variabel dependen (output) dengan mengasumsikan hubungan tersebut berupa garis lurus. Model ini digunakan untuk melakukan prediksi nilai output berdasarkan input yang diberikan.
2. K-Nearest Neighbor (KNN): Model KNN menghitung jarak antara data yang ingin diprediksi dengan data yang telah ada di dalam set pelatihan, kemudian memilih nilai output yang paling sering muncul pada tetangga terdekatnya. Model ini sering digunakan untuk klasifikasi data.
3. Decision Tree: Model decision tree digunakan untuk membuat model prediksi berdasarkan serangkaian keputusan yang dibuat berdasarkan input yang diberikan. Model ini sering digunakan dalam klasifikasi dan regresi.
4. Random Forest: Model random forest adalah model ensemble yang terdiri dari banyak decision tree yang dibuat dengan subset acak dari data pelatihan. Model ini dapat digunakan untuk klasifikasi atau regresi.
5. Neural Network: Model neural network adalah model yang terdiri dari beberapa lapisan neuron buatan. Model ini dapat digunakan untuk melakukan prediksi atau klasifikasi, dan sering digunakan untuk tugas-tugas seperti pengenalan wajah atau identifikasi citra.

III. 3 Model Machine Learning Terbaik untuk Klasifikasi

1. Naive Bayes Classifier: Model Naive Bayes Classifier adalah model probabilistik yang sering digunakan untuk klasifikasi teks dan dokumen. Model ini menghitung probabilitas setiap kelas berdasarkan input, dan kemudian memilih kelas dengan probabilitas tertinggi.

2. Decision Tree: Model Decision Tree adalah model klasifikasi yang mudah diinterpretasikan dan digunakan untuk memprediksi kelas target berdasarkan serangkaian keputusan yang diambil berdasarkan fitur-fitur input. Model ini membagi dataset menjadi subset yang lebih kecil berdasarkan fitur-fitur yang paling penting, dan kemudian membangun pohon keputusan berdasarkan subset tersebut.

3. Random Forest: Model Random Forest adalah model ensemble yang terdiri dari beberapa decision tree. Model ini bekerja dengan membuat banyak decision tree yang berbeda dengan menggunakan subset acak dari data pelatihan, dan kemudian menggabungkan hasil prediksi dari setiap tree untuk memutuskan kelas target yang paling mungkin.

Salah satu model machine learning yang digunakan pada eksplorasi dataset diatas adalah model Self-Training. Model Self-Training adalah salah satu teknik semi-supervised learning di mana model dipelajari pada data yang sebagian besar tidak bertanda (unsupervised) dan kemudian memperbarui model pada data yang bertanda (supervised) untuk meningkatkan akurasi prediksi.

Bila dibandingkan dengan 3 model diatas, model self-training mungkin tidak selalu berhasil pada semua jenis dataset dan tugas klasifikasi. Selain itu, model self-training memerlukan jumlah data yang tidak bertanda yang cukup besar dan berkualitas tinggi untuk bekerja secara efektif, yang mungkin tidak selalu tersedia. Pada akhirnya, ada banyak model klasifikasi yang lebih umum dan terbukti berhasil untuk berbagai jenis dataset dan tugas, seperti Decision Tree, Random Forest, dan Naive Bayes Classifier, yang mungkin lebih tepat untuk dipertimbangkan sebagai model teratas untuk klasifikasi data.

IV. Public Dataset Available for Breast Cancer

Dataset ini adalah salah satu dataset klasifikasi yang paling umum digunakan dalam machine learning dan digunakan untuk memprediksi apakah tumor yang terdeteksi di payudara adalah ganas atau jinak.

Dataset ini terdiri dari 569 sampel tumor yang dikumpulkan dari pasien wanita dengan tumor payudara, dan masing-masing sampel memiliki 30 fitur numerik yang dihasilkan dari citra digitized dari aspirasi jarum halus (FNA) dari tumor. Fitur-fitur ini menggambarkan properti sel-sel dalam gambar dan telah dinormalisasi sehingga nilai mereka berada dalam rentang yang sama.

Setiap sampel tumor memiliki label yang menunjukkan apakah tumor tersebut ganas (1) atau jinak (0). Tujuan dari dataset ini adalah untuk mengembangkan model klasifikasi yang dapat memprediksi apakah tumor yang baru terdeteksi adalah ganas atau jinak berdasarkan fitur-fitur numerik dari citra FNA.

V. Konten Dataset Breast Cancer

Dataset ini terdiri dari 569 sampel tumor yang diambil dari pasien wanita dengan tumor payudara. Setiap sampel tumor memiliki 30 fitur numerik yang dihasilkan dari citra digitized dari aspirasi jarum halus (FNA) dari tumor. Fitur-fitur ini menggambarkan properti sel-sel dalam gambar dan telah dinormalisasi sehingga nilai mereka berada dalam rentang yang sama.

Setiap sampel tumor memiliki label yang menunjukkan apakah tumor tersebut ganas (1) atau jinak (0). Label tersebut didasarkan pada hasil diagnosis klinis dan patologis yang dilakukan pada sampel tersebut.

Berikut adalah deskripsi dari 30 fitur numerik yang ada dalam dataset:

- mean radius: rata-rata jari-jari sel dalam gambar FNA
- mean texture: rata-rata tingkat variasi warna dalam gambar FNA
- mean perimeter: rata-rata keliling sel dalam gambar FNA
- mean area: rata-rata area daerah sel dalam gambar FNA
- mean smoothness: rata-rata perbedaan jarak antara titik-titik pada permukaan sel
- mean compactness: rata-rata perbandingan keliling^2/area-1.0
- mean concavity: rata-rata kedalaman celah di sekitar kontur
- mean concave points: rata-rata jumlah titik-titik konkaf dalam kontur
- mean symmetry: rata-rata simetri sel dalam gambar FNA
- mean fractal dimension: rata-rata "coastline approximation" pada gambar FNA

Selain itu, setiap fitur numerik di atas memiliki tiga varian: "mean", "standard error", dan "worst". Variasi "mean" adalah rata-rata dari fitur yang diukur pada setiap sel tumor. Variasi "standard error" adalah standar deviasi dari fitur tersebut, dan variasi "worst" adalah nilai terbesar dari fitur tersebut yang diukur pada setiap sel tumor.