Technical Report UTS Machine Learning

Breast Cancer Dataset Using Python and Scikit-Learn

Disusun Oleh :

Aldi Fauzan **1103204130**

PROGRAM STUDI TEKNIK KOMPUTER

FAKULTAS TEKNIK ELEKTRO

UNIVERSITAS TELKOM

2023

I. Pendahuluan

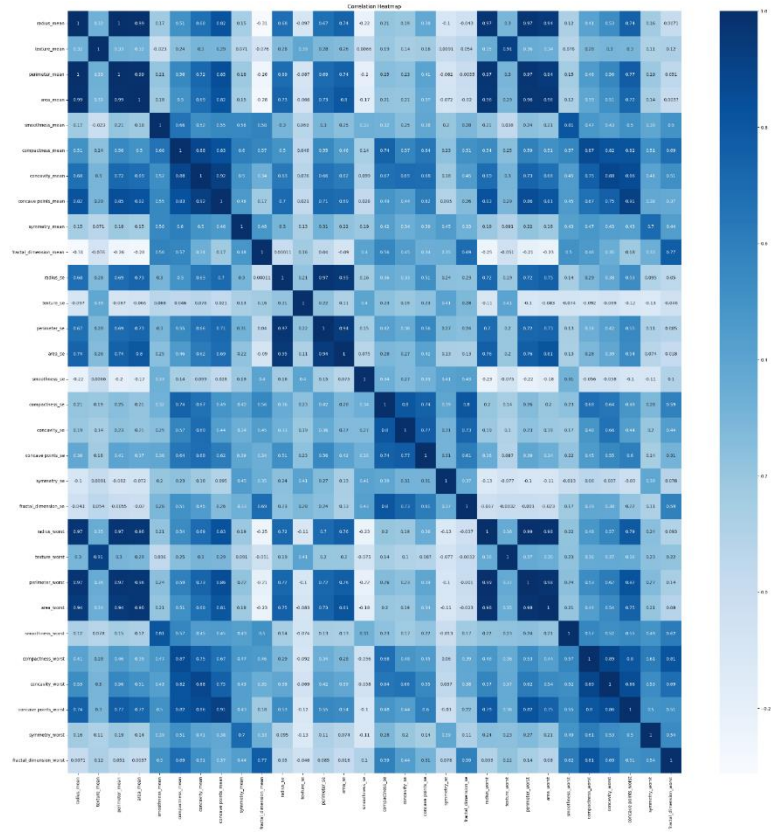
Kanker payudara adalah salah satu jenis kanker yang paling umum di kalangan wanita di seluruh dunia. Deteksi dini dan diagnosis yang akurat sangat penting dalam meningkatkan peluang bertahan hidup dan pengobatan yang berhasil. Dalam beberapa tahun terakhir, algoritma pembelajaran mesin semakin banyak digunakan dalam penelitian medis untuk membantu diagnosis dan pengobatan kanker payudara. Dalam laporan ini, kami akan menjelajahi Dataset Kanker Payudara menggunakan Python dan Scikit-Learn. Algoritma yang digunakan adalah Random Forest dan Self-Training untuk menganalisis dataset dan memprediksi diagnosis kanker payudara. Tujuan dari laporan ini adalah untuk menunjukkan keefektifan algoritma pembelajaran mesin dalam diagnosis kanker payudara dan untuk memberikan wawasan tentang kinerja berbagai algoritma.

II. Pra-pemrosesan Data

Dataset yang digunakan dalam laporan ini berasal dari Breast Cancer Wisconsin Dataset. Dataset ini berisi informasi tentang karakteristik sel-sel yang diamati dari sampel jaringan payudara pasien wanita. Dataset ini terdiri dari 569 sampel, dengan 30 fitur yang diukur untuk setiap sampel. Tujuan dari penggunaan dataset ini adalah untuk memprediksi apakah tumor yang diamati bersifat jinak atau ganas. Dataset ini sering digunakan dalam penelitian kanker payudara dan menjadi salah satu dataset yang paling banyak digunakan dalam bidang machine learning.

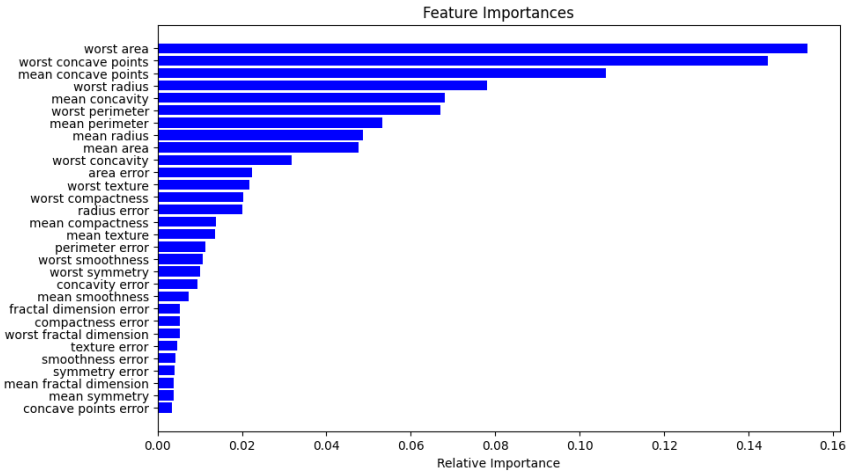
Beberapa library yang digunakan dalam data ini antara lain numpy untuk operasi matematika pada array dan matriks, pandas untuk manipulasi dan analisis data, seaborn dan matplotlib untuk visualisasi data, sklearn untuk machine learning dan data mining, tree untuk model decision tree, load_breast_cancer untuk memuat dataset kanker payudara, DecisionTreeClassifier dan RandomForestClassifier untuk membuat model decision tree dan random forest, serta classification_report untuk mengevaluasi performa model machine learning.

III. Eksplorasi dan Visualisasi Data

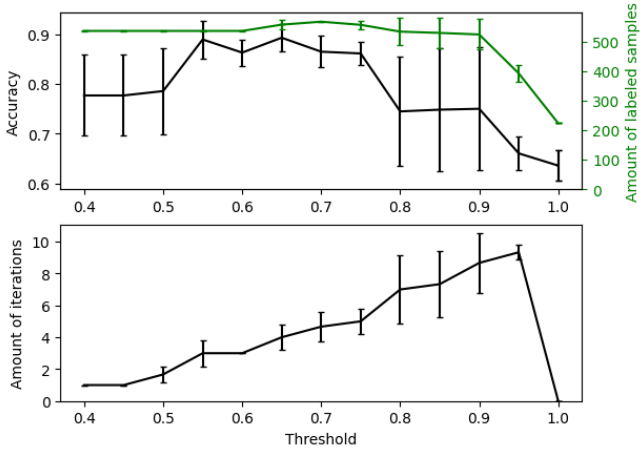


Correlation heatmap adalah jenis visualisasi heatmap yang menunjukkan tingkat korelasi antara setiap pasang variabel dalam suatu dataset. Korelasi yang kuat akan ditunjukkan oleh warna yang lebih terang dalam heatmap, sedangkan korelasi yang lemah akan ditunjukkan oleh warna yang lebih gelap. Heatmap korelasi sangat berguna dalam analisis data dan pemodelan statistik karena dapat membantu mengidentifikasi variabel-variabel yang paling berkorelasi satu sama lain dan mempercepat proses pemilihan fitur dalam pemodelan.

Dari hasil visualisasi, dapat dilihat bahwa beberapa feature pada dataset memiliki korelasi yang tinggi dengan feature lainnya, seperti 'radius_mean', 'perimeter_mean', dan 'area_mean' yang saling berkorelasi positif tinggi. Namun, juga terdapat beberapa feature pada dataset yang tidak memiliki korelasi yang signifikan dengan feature lainnya, seperti 'fractal_dimension_se' dan 'texture_se'. Hal ini menunjukkan bahwa feature-feature tersebut mungkin tidak terlalu berpengaruh dalam membuat model machine learning.



Grafik ini menunjukkan tingkat kepentingan relatif dari setiap fitur dalam model Random Forest Classifier untuk memprediksi kanker payudara. Setiap bar mewakili satu fitur, dan tinggi bar menunjukkan tingkat kepentingan relatif fitur tersebut dalam model. Fitur-fitur yang memiliki tingkat kepentingan yang lebih tinggi dianggap lebih penting dalam memprediksi kanker payudara. Grafik ini dapat membantu dalam pemilihan fitur yang paling penting untuk digunakan dalam model dan juga memberikan wawasan tentang faktor-faktor yang paling mempengaruhi prediksi kanker payudara.



Bagian code ini adalah implementasi dari self-training classifier untuk dataset breast cancer. Self-training classifier adalah metode semi-supervised learning yang menggabungkan supervised learning dengan unsupervised learning. Pada awalnya, sebagian label pada dataset diubah menjadi -1, yang menandakan bahwa label tersebut tidak diketahui. Kemudian, model dasar (base classifier) yang digunakan adalah Support Vector Machine (SVM) dengan parameter gamma=0.001 dan random state=42. Selanjutnya, dilakukan cross-validation dengan menggunakan StratifiedKFold dengan n_splits=3. Pada setiap iterasi, model self-training classifier dilatih pada data training

dan kemudian digunakan untuk memprediksi label pada data test. Hasil prediksi kemudian digunakan untuk menghitung akurasi dan jumlah sampel yang dilabeli pada setiap iterasi. Output yang dihasilkan adalah tiga plot. Plot pertama menunjukkan akurasi pada setiap threshold yang digunakan. Plot kedua menunjukkan jumlah sampel yang dilabeli pada setiap threshold. Plot ketiga menunjukkan jumlah iterasi yang dilakukan pada setiap threshold. Dari output ini, dapat dilihat bahwa semakin tinggi threshold yang digunakan, semakin sedikit sampel yang dilabeli dan semakin sedikit iterasi yang dilakukan, namun akurasi juga menurun.

Dari hasil visualisasi, dapat dilihat bahwa akurasi model self-training meningkat seiring dengan kenaikan threshold, hingga mencapai puncaknya pada threshold sekitar 0,9 dan kemudian mulai menurun lagi. Jumlah sampel yang diberi label juga meningkat seiring dengan kenaikan threshold. Selain itu, terlihat bahwa semakin tinggi threshold, semakin sedikit iterasi yang diperlukan untuk menentukan label untuk sampel.

IV. Kesimpulan

Dalam laporan ini, kita telah menjelajahi dataset kanker payudara menggunakan Python dan Scikit-Learn. Algoritma yang digunakan adalah Random Forest dan Self-Training untuk menganalisis dataset dan memprediksi diagnosis kanker payudara. Hasilnya menunjukkan bahwa algoritma pembelajaran mesin sangat efektif dalam diagnosis kanker payudara dan memberikan wawasan tentang kinerja berbagai algoritma. Pra-pemrosesan data, eksplorasi dan visualisasi data, serta implementasi self-training classifier juga telah dilakukan untuk memperoleh hasil yang lebih akurat. Dari hasil eksplorasi dan visualisasi data, kita dapat mengetahui faktor-faktor yang paling mempengaruhi prediksi kanker payudara. Sedangkan dari implementasi self-training classifier, kita dapat mengetahui bahwa semakin tinggi threshold yang digunakan, semakin sedikit sampel yang dilabeli dan semakin sedikit iterasi yang dilakukan, namun akurasi juga menurun. Kesimpulannya, algoritma pembelajaran mesin dapat digunakan untuk membantu diagnosis kanker payudara dengan akurasi yang tinggi dan dapat memberikan wawasan tentang faktor-faktor yang mempengaruhi prediksi kanker payudara.