Technical Report

Machine Learning Using Python and Scikit-Learn

Oleh:

Nama : Alifia Mutiara Rahma

NIM : 1103200025

PROGRAM STUDI TEKNIK KOMPUTER

FAKULTAS TEKNIK ELEKTRO

UNIVERSITAS TELKOM

2023

I. Pendahuluan

Kanker payudara merupakan pertumbuhan sel-sel abnormal di jaringan payudara, dan dapat terjadi pada pria dan wanita meskipun lebih umum pada wanita. Gejalanya dapat berupa benjolan atau massa pada payudara, perubahan bentuk atau ukuran payudara, rasa sakit atau nyeri, perubahan pada puting payudara, serta keluarnya cairan dari puting payudara. Diagnosis kanker payudara dapat dilakukan melalui pemeriksaan fisik, mamografi, atau biopsi jaringan payudara. Deteksi sejak dini sangat penting karena semakin cepat dideteksi, semakin besar peluang kesembuhan. 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.

II. Pengumpulan dan Persiapan Data

Dalam membangun model machine learning kali ini, menggunakan dataset *Breast Cancer Winsconsin (Diagnostic)* yang sudah tersedia di perpustakaan Scikit-Learn. Sehingga dalam penggunaanya, cukup menggunakan *syntax* "breast_cancer = datasets.load_breast_cancer()" untuk memanggil datasetnya. Dataset ini berisi 569 sampel tumor ganas dan jinak. Selain itu, terdiri juga 568 baris dan 31 kolom dalam datasetnya. Dengan menggunakan dataset ini, akan dibagi datanya menjadi 80% data untuk pelatihan dan 20% sisanya untuk pengujian atau bisa juga perbandingannya 70:30.

III. Visualisasi Data

Visualisasi data dalam *machine learning* sangat penting karena dapat membantu kita memahami data, menemukan *outlier* dan data yang hilang, menentukan fitur yang paling penting, membuat model yang lebih baik, membuat laporan yang lebih baik, dan eksplorasi data, serta mempercepat proses pengambilan keputusan. Sehingga dalam teknisnya dilakukan beberapa visualisasi data. Visualisasi data yang dilakukan diantara dengan menggunakan histogram, pairplot, dan *heatmap*. Histogram ini memberikan visualisasi distribusi data dan membantu dalam mengidentifikasi nilai-nilai yang muncul paling sering atau *outliers* yang mungkin perlu diperiksa lebih lanjut. Selanjutnya, maksud dari variabel 'target' adalah variable yang akan diprediksi. Variabel 'target' ini berupa kolom terakhir dalam dataset yang bisa berupa nilai numerik atau kategori. Variable 'target' nantinya digunakan untuk melatih model dengan membandingkan hasil prediksi model dengan nilai sebenarnya dari target. Dalam visualisasi data menggunakan histogram variable 'target' ini akan menunjukkan jumlah pengamatan dari setiap nilai target pada dataset. Nilai 0 dalam histogram menujukkan hasil klasifikasi tumor payudara "jinak" sedangkan nilai 1 menunjukkan hasil klasifikasi tumor payudara "ganas". Selanjutnya visualisasi data menggunakan pairplot menunjukkan scaller plot dari setiap pasangan variabel pada dataset. Hal ini dapat membantu untuk memahami hubungan antara variabel-variabel pada dataset secara visual dan mengidentifikasi pola yang mungkin dapat terjadi antara variabel-variabel tersebut menggunakan parameter hue. Visualisasi data selanjutnya menggunakan *heatmap*, sesuai apa yang ada dalam gambar, menunjukan korelasi antara setiap pasangan variabel pada dataset secara virtual. Apabila warnanya semakin terang,

maka korelasinya semakin tinggi, dan berlaku sebaliknya. Warna-warna ini dapat membantu mengidentifikasi variabel yang paling berpengaruh pada target variabel.

IV. Eksplorasi Data

Digunakan tiga algoritma machine learning untuk mengklasifikasikan kanker payudara, antara lain *Decision Tree, Random Forest,* dan *Self-Training.* Setiap modelnya akan dievasulasi performa dengan menggunakan nilai akuransi, presisi, *recall*, dan F1-score. Selain itu, digunakan juga beberapa model visualisasi data seperti *heatmap*. Dalam *Decision Tree* menggunakan *code* untuk melakukan eksplorasi data pada dataset *Breast Cancer Wisconsin* untuk mengoptimalkan *hyperparameter* dari model *Decision Tree Classifier* dengan menggunakan *pruning complexity.* Hal ini dapat memungkinkan untuk menentukan nilai ccp_alpha yang optimal untuk meningkatkan akurasi model pada data test. Semakin tinggi nilai ccp_alpha, semakin banyak nodes yang akan dihapus, dan semakin kecil nilai ccp_alpha, semakin kompleks model *Decision Tree.* Selanjutnya dilakukan klasifikasi pada dataset dengan melakukan prediksi pada testing set, baru kemudian mengevaluasi model dengan menggunakan *accuracy score* dan diperoleh akurasi pada *testing set*. Kemudian ditampilkan dalam bentuk visualisasi untuk memvisualisasikan *Decision Tree* yang telah dilatih pada dataset. Dalam visualisasinya menghasilkan diagram yang menunjukkan bagaimana decision tree tersebut melakukan klasifikasi pada instance tertentu. Diagram tersebut akan menunjukkan node-node dalam decision tree, kondisi-kondisi yang digunakan untuk menentukan apakah sebuah instance termasuk dalam kelas 0 atau kelas 1, dan probabilitas kelas pada setiap node. Visualisasi ini dapat membantu dalam memahami cara D*ecision Tree* melakukan klasifikasi dan juga dapat membantu dalam mengidentifikasi node-node penting dalam decision tree tersebut.

Eksplorasi data selanjutnya menggunakan model *Random Forest.* Model ini bertujuan untuk melakukan eksplorasi data pada dataset yang diambil dari *library* sklearn.database. Pertama-tama, data dibagi menjadi data latih dan data uji dengan rasio 70:30 menggunakan train_test_split dari library sklearn. Selanjutnya, dilakukan training model menggunakan *Random Forest Classifier* dengan n_estimators = 100 dan random state = 42 pada data latih dan dilakukan prediksi pada data uji. Kemudian, dihitung akurasi model dengan menggunakan accuracy_score dari library sklearn. Setelah itu, dilakukan plot feature importances dengan menggunakan matplotlib dan seaborn. Visualisasi tersebut memberikan gambaran mengenai fitur-fitur (feature) apa saja yang paling berpengaruh dalam klasifikasi antara kanker payudara jinak atau ganas. Langkah terakhir, dilakukan evaluasi performa model menggunakan classification_report dan confusion_matrix dari library sklearn.metrics. Visualisasi confusion matrix ditampilkan menggunakan heatmap dari seaborn.

Eksplorasi data yang terakhir menggunakan model *Self-Training.* Dalam percobaan pertama dilakukan untuk melakukan eksplorasi data menggunakan self-training classifier pada dataset breast_cancer. Pertama, dilakukan split data menjadi training set dan testing set, kemudian dilakukan inisialisasi pada Random Forest Classifier untuk self-training. Selanjutnya, dilakukan self-training pada classifier dengan memasukkan threshold sebesar 0.9 dan jumlah iterasi maksimum sebesar 100. Hasil training digunakan untuk memprediksi data pada testing set, kemudian akurasi model dihitung dengan menggunakan accuracy_score. Selain itu, juga dilakukan visualisasi distribusi variabel target pada dataset menggunakan Seaborn. Hasil visualisasi menunjukkan bahwa target

variabel pada dataset relatif seimbang antara kelas *benign* dan *malignant*. Selanjutnya menggunakan algoritma *Support Vector Machine (SVM)* dilakukan *Self-Training* pada klasifikasi data set untuk menguji performa dari *Self-Training Classifier* dalam melakukan klasifikasi pada dataset Breast Cancer. Hasil performa dari classifier diukur dengan menggunakan *accuracy score*, dan hasilnya digambarkan dalam bentuk grafik. Grafik tersebut menunjukkan korelasi antara nilai *threshold* dengan akurasi, jumlah labeled samples, dan jumlah iterasi yang dibutuhkan oleh *classifier*.

V. Kesimpulan

Dalam *technical report* ini, dapat disimpulkan bahwa *machine learning* dapat digunakan untuk mendiagnosis kanker payudara melalui analisis gambar mamografi dan data klinis. Tiga algoritma machine learning, yaitu *Decision Tree, Random Forest*, dan *Self-Training*, digunakan untuk mengklasifikasikan kanker payudara, dan performa setiap model diuji menggunakan nilai akurasi dan ditampilkan dalam bentuk visualisasi data seperti *heatmap*. *Python* dan *Scikit-Learn* digunakan untuk membangun model machine learning, sedangkan *Seaborn* digunakan untuk memvisualisasikan data dan mendapatkan wawasan tentang hubungan antara fitur dan variabel target. Selain itu, beberapa model visualisasi data seperti histogram, pairplot, dan *heatmap* digunakan untuk melakukan eksplorasi data pada dataset. Mengingat pentingnya deteksi dini kanker payudara, penggunaan *machine learning* dalam diagnosis dapat membantu mempercepat proses dan meningkatkan akurasi diagnosis kanker payudara.

VI. Daftar Pustaka

Learn. scikit. (n.d.). Retrieved April 10, 2023, from https://scikit-learn.org/stable/

Sklearn.datasets.load_breast_cancer. scikit. (n.d.). Retrieved April 9, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html

Statistical Data Visualization#. seaborn. (n.d.). Retrieved April 10, 2023, from https://seaborn.pydata.org/

YouTube. (2021, December 11). EP01 - python - seaborn - getting started. YouTube. Retrieved April 10, 2023, from https://www.youtube.com/watch?v=EWThwGSIpuY&list=PLiHa1s-EL3vgyJdXmRQExosUuM5IhGBIi