

*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

### 1.1. Pengertian *Machine Learning*

*Machine learning* (pembelajaran mesin) adalah sebuah bidang ilmu komputer yang berkaitan dengan pengembangan algoritma dan teknik untuk memungkinkan komputer belajar dari data dan melakukan prediksi atau tindakan tertentu berdasarkan data yang telah diberikan.

Dalam *machine learning*, algoritma dan model matematis digunakan untuk mempelajari pola atau relasi dalam data, dan dengan demikian dapat membuat prediksi atau mengambil tindakan berdasarkan pola yang telah dipelajari. Algoritma *machine learning* dapat digunakan untuk berbagai tugas, seperti pengenalan gambar dan suara, klasifikasi data, prediksi, optimisasi, dan pengambilan keputusan.

### 1.2. Model *Machine Learning*

Dalam model *machine learning* kali ini menggunakan *Random Forest*. *Random Forest* adalah model ensemble dalam *machine learning* yang terdiri dari beberapa pohon keputusan atau *decision tree*. *Ensemble model* berarti bahwa model ini menggabungkan beberapa model yang berbeda untuk membuat prediksi yang lebih akurat dan stabil. Dalam kasus *Random Forest*, pohon keputusan yang digunakan dihasilkan secara acak dari sampel data yang berbeda.

Setiap pohon keputusan dalam *Random Forest* memutuskan kelas atau nilai target suatu data dengan cara membagi data menjadi subgrup yang semakin kecil, berdasarkan aturan-aturan yang didefinisikan, hingga tidak bisa dibagi lagi. Kemudian setiap subgrup akan diberi label atau nilai target. Dengan mengkombinasikan

prediksi dari beberapa pohon keputusan yang dihasilkan secara acak, *Random Forest* dapat mengurangi kecenderungan overfitting dan meningkatkan kinerja prediksi.

### 1.3. Dataset

Dataset yang digunakan adalah kumpulan data public untuk kanker payudara. Dataset ini adalah dataset yang tersedia secara publik dan dapat diakses oleh siapa saja untuk digunakan dalam penelitian atau pengembangan model *machine learning* terkait kanker payudara (breast cancer).

Beberapa contoh dataset public yang tersedia untuk kanker payudara adalah:

1. Breast Cancer Wisconsin (Diagnostic) Dataset - dataset ini berisi informasi mengenai fitur-fitur patologi dari sel-sel yang diambil dari aspirasi jarum halus pada kanker payudara. Dataset ini memiliki 569 sampel dan 30 fitur, serta terdiri dari dua kelas yaitu benign (jinak) dan malignant (ganas).
2. Breast Cancer Wisconsin (Original) Dataset - dataset ini merupakan dataset yang sama dengan dataset Breast Cancer Wisconsin (Diagnostic), namun dengan beberapa fitur tambahan. Dataset ini memiliki 699 sampel dan 10 fitur, serta terdiri dari dua kelas yaitu benign (j<|endoftext|>

### 1.4. Pengertian Kanker Payudara

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 penggunaannya, 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 menunjukkan hasil klasifikasi tumor payudara “jinak” sedangkan nilai 1 menunjukkan hasil klasifikasi tumor payudara “ganas”. Selanjutnya visualisasi data menggunakan pairplot menunjukkan scatter 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, menunjukkan 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 dievaluasi performa dengan menggunakan nilai akurasi, 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 *Decision 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/>
- Sklarn.datasets.load\_breast\_cancer. scikit. (n.d.). Retrieved April 9, 2023, from [https://scikit-learn.org/stable/modules/generated/sklarn.datasets.load\\_breast\\_cancer.html](https://scikit-learn.org/stable/modules/generated/sklarn.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>