Technical Report Machine Learning

Testing Breast Cancer Dataset using Random Forest, Decision Tree, and Self-Training

Disusun oleh:

Galih Karya Gemilang

TK-44-03

1103202098

Program Studi S1 Teknik Komputer

Fakultas Teknik Elektro

Universitas Telkom

2023

Daftar Isi

1. Machine Leaning
2. Model-model umum yang digunakan dalam Machine Learning
3. Kumpulan dataset publik yang tersedia untuk kanker payudara
4. Decision Tree
5. Random Forest
6. Self-Training
7. Hasil
8. Kesimpulan

1. *Machine Learning*

Machine learning adalah cabang ilmu komputer yang memungkinkan komputer untuk belajar dari data dan pengalaman, serta membuat prediksi dan keputusan berdasarkan pemahaman yang didapat dari data.

Dalam *machine learning*, komputer belajar dengan mengidentifikasi pola dalam data dan memperoleh pemahaman tentang hubungan antara variabel yang berbeda. Hal ini dapat dicapai melalui pembuatan model statistik dan algoritma yang dapat mengekstrak informasi yang bermanfaat dari data yang diberikan.

Ada tiga jenis utama dari *machine learning*: *supervised learning*, *unsupervised learning*, dan *reinforcement learning*.

Supervised learning melibatkan pelatihan model pada data yang sudah diketahui outputnya, sedangkan *unsupervised learning* melibatkan identifikasi pola dalam data yang tidak memiliki output yang diketahui. *Reinforcement learning* melibatkan model yang belajar melalui interaksi dengan lingkungan sekitarnya, dan diberikan penguatan (*reward*) atau hukuman (*punishment*) berdasarkan tindakan yang diambil.

Machine learning digunakan dalam berbagai bidang seperti pengenalan suara dan gambar, analisis risiko kredit, pengenalan wajah, dan banyak lagi. Dalam beberapa tahun terakhir, *machine learning* telah menjadi sangat populer karena kemampuannya untuk memproses data yang sangat besar dengan cepat dan akurat, serta memberikan solusi untuk masalah yang sulit dipecahkan dengan metode tradisional.

2. Model-model umum yang digunakan dalam *Machine Learning*

Terdapat beberapa model umum dalam *machine learning* yang digunakan untuk mempelajari data, di antaranya:

a. Regresi: Model regresi digunakan untuk memprediksi nilai kontinu berdasarkan input. Contoh aplikasi dari regresi adalah prediksi harga rumah berdasarkan faktor-faktor seperti lokasi, ukuran, dan tahun pembangunan.
b. Klasifikasi: Model klasifikasi digunakan untuk memprediksi label kategori berdasarkan input. Contoh aplikasi dari klasifikasi adalah klasifikasi email sebagai spam atau bukan spam, atau klasifikasi gambar sebagai objek tertentu seperti kucing atau anjing.
c. *Clustering*: Model *clustering* digunakan untuk mengelompokkan data ke dalam kelompok-kelompok berdasarkan kesamaan karakteristik. Contoh aplikasi dari *clustering* adalah segmentasi pelanggan berdasarkan preferensi atau perilaku pembelian.
d. *Deep learning*: Model *deep learning* digunakan untuk memproses data yang sangat besar dan kompleks dengan menggunakan jaringan saraf tiruan (*artificial neural network*). *Deep learning* telah berhasil diterapkan dalam berbagai aplikasi, seperti pengenalan wajah, pengenalan suara, dan pengenalan teks.

e. *Reinforcement learning*: Model *reinforcement learning* digunakan untuk mempelajari bagaimana agen (*agent*) dapat memaksimalkan penguatan (*reward*) dalam interaksi dengan lingkungannya. Contoh aplikasi dari *reinforcement learning* adalah permainan komputer, robotika, dan optimisasi tugas-tugas berulang.
f. *Semi-supervised learning*: Model *semi-supervised learning* adalah kombinasi dari *supervised learning* dan *unsupervised learning*, di mana sebagian data memiliki label dan sebagian lagi tidak. Model ini digunakan ketika label data sulit atau mahal untuk dihasilkan, tetapi data tanpa label masih dapat memberikan informasi yang berguna.

Model-model tersebut dapat dikombinasikan dan disesuaikan dengan tujuan dan jenis data yang digunakan untuk mencapai hasil yang diinginkan.

3. Kumpulan dataset publik yang tersedia untuk kanker payudara

Terdapat beberapa kumpulan dataset publik yang tersedia untuk kanker payudara (*breast cancer*), di antaranya:

a. *Breast Cancer Wisconsin (Diagnostic) Dataset*: Kumpulan data ini berisi 569 sampel sel tumor yang dikumpulkan dari biopsi payudara pasien wanita di Wisconsin, AS. Data ini memiliki 30 fitur, termasuk karakteristik sel tumor seperti radius, tekstur, konsistensi, simetri, dan dimensi. Tujuan dari dataset ini adalah untuk memprediksi apakah tumor bersifat jinak (*benign*) atau ganas (*malignant*).
b. *Breast Cancer Wisconsin (Original) Dataset*: Kumpulan data ini juga dikumpulkan dari biopsi payudara pasien wanita di Wisconsin, AS, dan terdiri dari 699 sampel sel tumor. Dataset ini memiliki 9 fitur, termasuk ukuran sel, konsistensi, dan variasi ukuran inti. Tujuan dari dataset ini adalah untuk memprediksi apakah tumor bersifat jinak atau ganas.
c. *Breast Cancer Coimbra Dataset*: Kumpulan data ini berisi 116 sampel sel tumor yang dikumpulkan dari pasien wanita di Portugal. Dataset ini memiliki 10 fitur, termasuk karakteristik sel tumor seperti ukuran, bentuk, dan konsistensi. Tujuan dari dataset ini adalah untuk memprediksi apakah tumor bersifat jinak atau ganas.
d. *Mammographic Mass Dataset*: Kumpulan data ini berisi 961 sampel mammogram digital dari payudara wanita, termasuk 445 sampel dengan tumor dan 516 sampel tanpa tumor. Dataset ini memiliki 5 fitur, termasuk usia pasien dan karakteristik tumor seperti ukuran, bentuk, dan margin. Tujuan dari dataset ini adalah untuk memprediksi apakah ada tumor pada mammogram.
e. *INbreast Dataset*: Kumpulan data ini berisi 410 mammogram digital dari payudara wanita, termasuk 115 kasus kanker payudara dan 295 kasus tanpa kanker. Dataset ini mencakup informasi tentang usia pasien, jenis tumor, dan karakteristik tumor seperti ukuran dan bentuk. Tujuan dari dataset ini adalah untuk membantu dalam diagnosis dan deteksi kanker payudara.

Dataset tersebut dapat digunakan untuk pelatihan dan pengujian model machine learning untuk deteksi kanker payudara dan prediksi prognosis. Semua dataset tersebut tersedia secara publik dan dapat diunduh dari sumbernya.

4. *Decision Tree*

Decision tree atau pohon keputusan adalah salah satu model *machine learning* yang populer dalam pengambilan keputusan dan klasifikasi data. Model ini menghasilkan struktur seperti pohon yang terdiri dari serangkaian keputusan yang diambil berdasarkan fitur input dan hasil output yang diinginkan. Setiap node pada pohon keputusan mewakili suatu fitur, sedangkan cabang-cabang pada node tersebut mewakili kemungkinan nilai fitur tersebut. Setiap daun pada pohon keputusan mewakili hasil klasifikasi atau prediksi.

Ada beberapa metode yang digunakan dalam pembangunan pohon keputusan, di antaranya:

a. *ID3 (Iterative Dichotomiser 3)*: Metode ini menghitung entropy setiap fitur pada data dan memilih fitur dengan nilai informasi tertinggi sebagai node root (akar) pada pohon keputusan. Proses ini diulangi secara iteratif untuk membangun pohon yang lebih besar.
b. *C4.5*: Metode ini merupakan pengembangan dari ID3 dan memiliki kemampuan untuk menangani data dengan nilai yang hilang. Selain itu, C4.5 menggunakan algoritma pruning (pemangkasan) untuk mengurangi overfitting dan meningkatkan generalisasi model.
c. *CART (Classification and Regression Trees)*: Metode ini dapat digunakan untuk masalah klasifikasi dan regresi. Pada metode CART, setiap node pada pohon keputusan dibagi menjadi dua cabang yang saling eksklusif berdasarkan nilai ambang batas yang optimal. Proses ini diulangi secara rekursif sampai terbentuk pohon keputusan yang optimal.

Keuntungan dari penggunaan pohon keputusan antara lain mudah dipahami dan diinterpretasikan, dapat menangani data yang tidak terstruktur, serta efektif dalam memecahkan masalah klasifikasi dan regresi. Namun, pohon keputusan cenderung overfitting jika terlalu dalam atau kompleks, dan memerlukan pengujian dan optimisasi yang baik untuk mendapatkan hasil yang optimal.

5. *Random Forest*

Random forest adalah salah satu model *machine learning* yang memanfaatkan sekumpulan pohon keputusan (*decision tree*) untuk melakukan klasifikasi atau regresi. Model ini mampu meningkatkan akurasi dan mengatasi *overfitting* yang sering terjadi pada pohon keputusan tunggal. Random forest bekerja dengan cara membangun banyak pohon keputusan secara acak dan menggabungkan hasil prediksi dari setiap pohon untuk memperoleh prediksi akhir.

Proses pembangunan *random forest* melibatkan beberapa tahapan, di antaranya:

1. Pemilihan sampel: Sebuah *random forest* dibangun dengan menggunakan sampel data yang diambil secara acak dari data latih. Setiap pohon keputusan dibangun dengan menggunakan sebagian kecil data dari keseluruhan data latih. Proses ini dikenal dengan istilah bootstrap.
2. Pembangunan pohon: Setiap pohon keputusan pada *random forest* dibangun dengan menggunakan metode pembangunan pohon keputusan seperti ID3, C4.5, atau CART. Namun, pada setiap node di pohon keputusan, hanya sebagian dari fitur yang dipilih secara acak untuk digunakan dalam memilih nilai pemisah.

3. Kombinasi pohon: Setelah sejumlah pohon keputusan dibangun, prediksi dari setiap pohon digabungkan untuk menghasilkan prediksi akhir. Dalam klasifikasi, prediksi akhir dihasilkan melalui mayoritas suara, sedangkan pada regresi, prediksi akhir dihasilkan melalui rata-rata prediksi dari setiap pohon.

Keuntungan dari penggunaan *random forest* antara lain dapat mengatasi *overfitting* dan meningkatkan akurasi prediksi, mudah diimplementasikan, dan dapat digunakan untuk masalah klasifikasi dan regresi. Namun, *random forest* cenderung membutuhkan waktu yang lebih lama untuk proses pembangunan model karena memerlukan banyak pohon keputusan, serta membutuhkan pengaturan parameter yang tepat untuk menghasilkan hasil yang optimal.

6. *Self-Training*

Self-training adalah salah satu teknik *machine learning semi-supervised* yang memanfaatkan data tak berlabel untuk meningkatkan kinerja model dalam melakukan klasifikasi. Teknik ini berguna jika data latih yang terbatas atau mahal dalam pengumpulannya, sehingga membutuhkan penggunaan data tak berlabel untuk memperluas jangkauan data latih yang tersedia.

Pada *self-training*, model dilatih terlebih dahulu dengan menggunakan data latih yang ada. Setelah itu, model digunakan untuk melakukan prediksi pada data tak berlabel dan hasil prediksi tersebut kemudian ditambahkan ke dalam data latih sebagai data baru yang telah diberi label. Setelah itu, model dilatih kembali dengan data latih yang telah diperbarui dan proses prediksi dilakukan kembali pada data tak berlabel. Proses ini dilakukan secara berulang hingga model menghasilkan performa yang stabil atau mencapai batas iterasi tertentu.

Beberapa hal yang perlu diperhatikan dalam penggunaan *self-training*, antara lain:

a. Pemilihan data tak berlabel: Pemilihan data tak berlabel yang tepat sangat penting dalam proses *self-training*. Data yang dipilih sebaiknya sesuai dengan domain yang akan diprediksi dan memiliki keterkaitan yang erat dengan data latih.
b. Pemilihan model yang tepat: Pemilihan model yang tepat juga sangat penting dalam proses *self-training*. Model yang digunakan harus mampu mengatasi masalah klasifikasi yang kompleks dan dapat menghasilkan performa yang stabil.
c. Pengaturan batas iterasi: Proses *self-training* dilakukan secara berulang dengan menambahkan data tak berlabel ke dalam data latih hingga mencapai batas iterasi tertentu. Batas iterasi yang dipilih sebaiknya cukup untuk memperbarui data latih dan mencapai performa yang stabil tanpa mempengaruhi kinerja model.

Keuntungan dari penggunaan *self-training* antara lain dapat meningkatkan performa model dengan menggunakan data tak berlabel, memperluas jangkauan data latih, serta dapat digunakan dalam berbagai masalah klasifikasi yang kompleks. Namun, teknik *self-training* juga memiliki beberapa kelemahan, seperti memerlukan pemilihan data tak berlabel yang tepat, pemilihan model yang tepat, serta memerlukan pengaturan parameter yang tepat untuk menghasilkan hasil yang optimal.

7. Hasil

Laporan ini membahas hasil dari visualisasi dan eksplorasi data kanker payudara dengan menggunakan algoritma *tree decision*, *random forest*, dan *self-training*. Tujuan dari penelitian ini adalah untuk menganalisis dan memahami hubungan antara faktor risiko tertentu dan kemungkinan terjadinya kanker payudara. Data yang digunakan dalam penelitian ini diperoleh dari situs *kaggle.com* yang terkait dengan kanker payudara. Hasil dari analisis data menunjukkan bahwa algoritma *tree decision*, *random forest*, dan *self-training* dapat digunakan untuk memprediksi risiko kanker payudara dengan akurasi yang tinggi. Semoga laporan ini dapat memberikan pemahaman yang lebih baik terhadap faktor risiko kanker payudara dan kontribusi yang positif dalam upaya pencegahan dan pengobatan kanker payudara.

Tahap 1: Mencari dataset dan import library yang digunakan

Dataset yang digunakan sudah ditentukan, yaitu menggunakan dataset *breast cancer* menggunakan *API* dari *scikit-learn*. Hanya saja terdapat data kolom yang hilang, sehingga pada penugasan ini menggunakan dataset *breast cancer* dari *kaggle.com* yang isinya mirip-mirip dan lebih lengkap, sehingga memudahkan dalam melakukan visualisasi data. Tautannya adalah *https://www.kaggle.com/datasets/nancyalaswad90/breast-cancer-dataset*.

Library yang digunakan adalah *numpy* untuk memanipulasi array atau matriks secara efisien, *seaborn* untuk memvisualisasikan data secara statistik, *matplotlib* untuk memvisualisasikan data dalam bentuk grafik atau plot, *pandas* untuk melakukan analisis data dan memanipulasi data secara efisien dengan cara membentuk data tersebut menjadi sebuah *dataframe, dan beberapa dari sklearn* untuk membangun model machine learning. Yang paling banyak digunakan pada proyek ini adalah *pandas* karena dataset yang berupa *csv* ini disimpan secara lokal, kemudian diubah ke bentuk *dataframe* supaya mudah pengolahannya.

Tahap 2: Visualisasi data

Dataset yang sudah diubah ke bentuk *dataframe*, kemudian lakukan visualisasi. Dari yang paling mudah, mengelompokkan tumor yang ganas dengan yang jinak. Kemudian melakukan visualisasi data yang lebih mendetil lagi menggunakan *sns.swarmplot*. Dari grafik yang dihasilkan dapat terlihat jelas perbatasan antara tumor ganas dengan tumor jinaknya berdasarkan data lainnya. Selain itu juga, pada proyek ini terdapat grafik *heatmap* yang menampilkan korelasi data satu dengan yang lainnya.

Tahap 3: Eksplorasi data

Dari ketiga metode, untuk yang *Tree Decision* dan *Random Forest* sudah dibagi data tes dan data latihnya dengan perbandingan 1:4. Setelah itu program akan menampilkan akurasinya. Sedangkan untuk *self-training* menggunakan *gamma* sebesar 0,001, kemudian program akan menampilkan akurasi terhadap jumlah iterasinya dalam bentuk grafik.

8. Kesimpulan

Visualisasi dan eksplorasi data kanker payudara menggunakan algoritma *tree decision, random forest,* dan *self-training* dapat digunakan untuk menganalisis dan memahami hubungan antara faktor risiko tertentu dan kemungkinan terjadinya kanker payudara. Dalam penelitian ini, data yang digunakan berasal dari situs kaggle.com dan telah diolah menggunakan *library numpy, seaborn, matplotlib,* dan *pandas*. Visualisasi data dilakukan menggunakan *sns.swarmplot* dan *heatmap*, sedangkan eksplorasi data dilakukan dengan membagi data menjadi data test dan data latih pada algoritma *tree decision* dan *random forest*, serta *self-training*. Hasil analisis data menunjukkan bahwa algoritma *tree decision, random forest,* dan *self-training* dapat digunakan untuk memprediksi risiko kanker payudara dengan akurasi yang tinggi.