Technical Report: Klasifikasi Kanker Payudara dengan Model Decision Tree, Random Forest dan Self-Training.

Dalam dunia yang semakin terhubung dan canggih saat ini, komputer memiliki kemampuan untuk belajar dan memecahkan masalah dengan cara yang lebih cerdas. Salah satu konsep yang mendasari perkembangan ini adalah Machine Learning. Machine Learning adalah bidang dalam kecerdasan buatan yang bertujuan untuk mengembangkan algoritma dan model komputer yang dapat belajar dari data dan pengalaman.

Ada beberapa model machine learning yang sering digunakan, tergantung pada jenis masalah yang ingin dipecahkan dan jenis data yang tersedia. Berikut adalah beberapa model machine learning yang populer:

- Naive Bayes: Model ini berdasarkan pada teorema Bayes dan digunakan untuk klasifikasi. Ini bekerja dengan mengasumsikan bahwa setiap fitur independen satu sama lain. Model ini cocok untuk data yang memiliki banyak fitur dengan kelas target yang diketahui.

- Decision Tree: Model ini menggunakan struktur pohon keputusan yang terdiri dari node keputusan dan cabang berdasarkan fitur-fitur input. Ini memungkinkan pengambilan keputusan berdasarkan serangkaian pertanyaan ya/tidak. Decision tree dapat digunakan untuk klasifikasi dan regresi.

- Random Forest: Ini adalah kumpulan decision tree yang bekerja secara independen dan menggabungkan hasilnya. Setiap pohon dalam random forest dilatih pada subset data yang diambil secara acak. Model ini dapat memberikan prediksi yang lebih akurat dan mengurangi overfitting.

- Support Vector Machines (SVM): SVM adalah model yang digunakan untuk klasifikasi atau regresi. Ini mencari batas keputusan yang optimal untuk memisahkan kelas yang berbeda dalam ruang fitur. SVM juga dapat menggunakan fungsi kernel untuk menangani data yang tidak linier.

- Neural Networks: Model ini terinspirasi oleh struktur dan fungsi jaringan saraf dalam otak manusia. Ini terdiri dari beberapa lapisan neuron buatan yang saling terhubung. Neural networks dapat digunakan untuk klasifikasi, regresi, dan tugas-tugas lainnya. Model neural network yang dalam adalah Deep Learning.

Machine learning memanfaatkan algoritma dan model yang mempelajari pola dari data yang diberikan untuk memprediksi kelas atau kategori dari data yang belum diketahui. Hal tersebut dapat digunakan untuk mengklasifikasi berbagai dataset.

Terdapat banyak model machine learning yang bagus untuk klasifikasi, namun berikut ini adalah tiga model yang sering digunakan dan terkenal karena kinerjanya yang baik:

- Support Vector Machines (SVM): Model machine learning yang kuat untuk klasifikasi. SVM bekerja dengan mencari batas keputusan yang optimal untuk memisahkan dua kelas yang berbeda dalam ruang fitur. SVM mampu menangani data dengan dimensi yang tinggi dan efektif dalam mengatasi masalah klasifikasi linear maupun non-linear. Selain itu, SVM juga memiliki kemampuan untuk mengatasi masalah overfitting dan tahan terhadap noise dalam data.

- Random Forest: Model ensemble yang terdiri dari banyak pohon keputusan. Setiap pohon dalam Random Forest dilatih pada subset data yang diambil secara acak. Keputusan akhir diambil berdasarkan mayoritas suara dari semua pohon. Random Forest memiliki keunggulan dalam mengatasi masalah overfitting, tahan terhadap noise, dan mampu menghadapi data dengan fitur yang banyak dan kompleks. Model ini juga dapat memberikan estimasi pentingnya setiap fitur dalam klasifikasi.

- Gradient Boosting (misalnya, XGBoost atau LightGBM): Teknik ensemble learning yang membangun serangkaian model yang lemah secara berturut-turut untuk meningkatkan performa secara keseluruhan. Dalam klasifikasi, model gradient boosting seperti XGBoost atau LightGBM telah terbukti sangat efektif. Mereka bekerja dengan membuat model yang berturut-turut, di mana setiap model berusaha untuk memperbaiki kesalahan yang dibuat oleh model sebelumnya. Gradient Boosting biasanya memberikan hasil yang sangat akurat dan mampu menangani masalah dengan data yang besar dan kompleks.

Penting untuk dicatat bahwa setiap model memiliki kelebihan dan batasan sendiri, dan performa mereka tergantung pada karakteristik data yang digunakan. Pemilihan model yang tepat harus dipertimbangkan berdasarkan tujuan dan persyaratan spesifik dari masalah klasifikasi yang ingin dipecahkan.

Pada *Technical Report* ini akan dibahas langkah-langkah mengklasifikasikan kanker payudara menggunakan beberapa model diantaranya adalah *Decision Tree, Random Forest* dan *Self-Training*. Untuk dataset untuk kanker payudara sendiri sebenarnya banyak sekali seperti dari:

- Wisconsin Diagnostic Breast Cancer (WBCD) Dataset: Dataset ini berisi informasi mengenai fitur-fitur sel-sel payudara yang diambil dari citra mikroskopis. Termasuk di dalamnya adalah ukuran, bentuk, tekstur, dan fitur-fitur lainnya. Dataset ini digunakan untuk melakukan klasifikasi antara tumor ganas (malignant) dan tumor jinak (benign).

- Breast Cancer Wisconsin (Original) Dataset: Dataset ini mengandung informasi tentang fitur-fitur sel-sel payudara yang diambil dari citra mikroskopis. Dataset ini digunakan untuk melakukan klasifikasi antara tumor ganas (malignant) dan tumor jinak (benign). Fitur-fitur meliputi ukuran inti sel, tekstur, dan pengukuran lainnya.

- METABRIC (Molecular Taxonomy of Breast Cancer International Consortium): Dataset ini mencakup data ekspresi genetik, data klinis, dan informasi lainnya dari ribuan pasien kanker payudara. Dataset ini memberikan pemahaman lebih mendalam tentang aspek molekuler dari kanker payudara dan digunakan untuk berbagai analisis dan penelitian.

- UC Irvine Machine Learning Repository - Breast Cancer Wisconsin (Diagnostic) Dataset: Dataset ini berisi data medis yang dikumpulkan dari 569 pasien dengan kanker payudara. Termasuk di dalamnya adalah atribut-atribut seperti ukuran tumor, kekonsentrasian inti sel, dan pengukuran lainnya. Dataset ini digunakan untuk melakukan klasifikasi antara tumor ganas (malignant) dan tumor jinak (benign).

- The Cancer Imaging Archive (TCIA): TCIA menyediakan berbagai dataset citra medis yang mencakup kanker payudara. Dataset ini mencakup citra mammogram, citra MRI, dan citra lainnya yang dapat digunakan untuk analisis dan penelitian di bidang deteksi, diagnosis, dan pemantauan kanker payudara.

Dataset yang digunakan untuk laporan ini adalah sklearn.datasets.load_breast_cancer yang berisi 569 sampel tumor ganas dan jinak dengan 30 parameter untuk setiap sampelnya, dengan parameter sebagai berikut:

1. radius_mean: rata-rata jarak dari pusat ke tepi tumor
2. texture_mean: standar deviasi dari nilai-nilai skala abu-abu pada gambar
3. perimeter_mean: perimeter (panjang) tumor
4. area_mean: area tumor
5. smoothness_mean: rata-rata variasi panjang sel-sel di sekitar inti
6. compactness_mean: rata-rata perbandingan area^2 / perimeter - 1.0
7. concavity_mean: rata-rata tingkat keparahan cekungan pada kontur tumor
8. concave points_mean: rata-rata jumlah titik-titik cekung pada kontur tumor
9. symmetry_mean: rata-rata simetri sel-sel inti
10. fractal_dimension_mean: "coastline approximation" - 1

11. radius_se: standar error dari jarak pusat ke tepi tumor
12. texture_se: standar error dari nilai-nilai skala abu-abu pada gambar
13. perimeter_se: standar error dari perimeter (panjang) tumor
14. area_se: standar error dari area tumor
15. smoothness_se: standar error dari variasi panjang sel-sel di sekitar inti
16. compactness_se: standar error dari perbandingan area^2 / perimeter - 1.0
17. concavity_se: standar error dari tingkat keparahan cekungan pada kontur tumor
18. concave points_se: standar error dari jumlah titik-titik cekung pada kontur tumor
19. symmetry_se: standar error dari simetri sel-sel inti
20. fractal_dimension_se: standar error dari "coastline approximation" - 1
21. radius_worst: nilai terburuk (terbesar) dari jarak pusat ke tepi tumor
22. texture_worst: nilai terburuk (terbesar) dari standar deviasi nilai-nilai skala abu-abu pada gambar
23. perimeter_worst: nilai terburuk (terbesar) dari perimeter (panjang) tumor
24. area_worst: nilai terburuk (terbesar) dari area tumor
25. smoothness_worst: nilai terburuk (terbesar) dari variasi panjang sel-sel di sekitar inti
26. compactness_worst: nilai terburuk (terbesar) dari perbandingan area^2 / perimeter - 1.0
27. concavity_worst: nilai terburuk (terbesar) dari tingkat keparahan cekungan pada kontur tumor
28. concave points_worst: nilai terburuk (terbesar) dari jumlah titik-titik cekung pada kontur tumor
29. symmetry_worst: nilai terburuk (terbesar) dari simetri sel-sel inti
30. fractal_dimension_worst: nilai terburuk (terbesar) dari "coastline approximation" - 1