Technical Report Machine Learning

"Visualization and Exploration of Breast Cancer Data from Sci-Kit Learn"

Oleh:

Nama : Juan Meta Sirgianto

NIM : 1103202092

PROGRAM STUDI TEKNIK KOMPUTER

FAKULTAS TEKNIK ELEKTRO

UNIVERSITAS TELKOM

2023

I. PENDAHULUAN

Kanker payudara adalah jenis kanker yang terjadi ketika sel-sel di dalam payudara mulai tumbuh secara tidak normal dan tidak terkendali. Kanker payudara dapat terjadi pada laki-laki dan perempuan, namun lebih umum terjadi pada perempuan. Kanker payudara adalah jenis kanker yang paling umum di antara wanita di seluruh dunia, dan dapat mempengaruhi kualitas hidup dan kesehatan perempuan secara signifikan. Pengobatan kanker payudara tergantung pada stadium kanker, jenis kanker, dan kondisi kesehatan pasien. Beberapa metode pengobatan yang umum digunakan termasuk operasi, radiasi, kemoterapi, terapi hormon, dan terapi target. Deteksi dini kanker payudara sangat penting untuk meningkatkan kemungkinan kesembuhan dan meminimalkan efek samping dari pengobatan. Oleh karena itu, wanita disarankan untuk melakukan pemeriksaan payudara sendiri dan menjalani pemeriksaan rutin dengan dokter.

II. DESKRIPSI DATASET

Dataset yang kita gunakan berisi tentang sampel tumor payudara pasien yang menderita kanker payudara. Tidak cuma itu, dataset yang kita gunakan juga berisi informasi mengenal ukuran dan bentuk tumornya. Dataset ini terdiri dari 569 observasi dengan 32 variabel. Variabelnya meliputi Diagnosis, ID pasien, Radius tumor, dan 29 fitur lainnya.

III. VISUALISASI DATA

Visualisasi data adalah proses mewakili data dalam bentuk grafik atau diagram, sehingga memudahkan kita untuk memahami pola, tren, hubungan, dan informasi penting lainnya yang terkandung dalam data. Tujuan utama dari visualisasi data adalah untuk memudahkan pemahaman dan analisis data dengan cara yang lebih intuitif dan efektif. Terdapat berbagai macam jenis visualisasi data, mulai dari grafik bar, grafik garis, scatter plot, decision tree, line plot, random forest dan masih banyak lagi. Pilihan jenis visualisasi yang tepat tergantung pada jenis data yang akan dianalisis dan tujuan dari analisis tersebut.

Scatter plot adalah jenis grafik yang digunakan untuk memvisualisasikan hubungan antara dua variabel numerik. Dalam scatter plot, data ditampilkan sebagai titik-titik pada bidang kartesian, di mana sumbu horizontal mewakili nilai dari satu variabel dan sumbu vertikal mewakili nilai dari variabel lainnya. Setiap titik pada scatter plot merepresentasikan sebuah pasangan nilai dari dua variabel. Jika terdapat banyak titik pada scatter plot, maka pola hubungan antara kedua variabel dapat dianalisis lebih lanjut. Scatter plot biasanya digunakan untuk mengidentifikasi korelasi antara dua variabel, yaitu seberapa kuat atau lemah hubungan antara keduanya. Scatter plot dapat memiliki bentuk yang berbeda-beda, seperti bentuk elips, lingkaran, atau segitiga, tergantung pada distribusi data. Scatter plot juga dapat diwarnai berdasarkan variabel kategorikal, sehingga memungkinkan kita untuk memvisualisasikan lebih dari dua variabel dalam satu grafik.

IV. DECISION TREE

Decision tree atau pohon keputusan adalah model prediksi yang menggunakan struktur pohon untuk memetakan serangkaian keputusan dan konsekuensi yang mungkin terjadi dari keputusan tersebut. Dalam pohon keputusan, setiap node pada pohon mewakili keputusan atau kondisi yang harus diperiksa, sedangkan cabang-cabang dari node tersebut merepresentasikan

kemungkinan hasil dari keputusan atau kondisi tersebut. Setiap daun pada pohon mewakili hasil akhir dari serangkaian keputusan dan kondisi yang diterapkan. Pohon keputusan dapat digunakan untuk memecahkan berbagai masalah klasifikasi atau regresi. Keuntungan dari penggunaan pohon keputusan adalah mudah dipahami dan diinterpretasi, serta dapat digunakan pada berbagai jenis data. Namun, pohon keputusan juga memiliki kekurangan, seperti kecenderungan untuk overfitting jika terlalu kompleks dan sulit untuk menangani data yang memiliki nilai yang hilang atau tidak lengkap.

V. RANDOM FOREST

Random forest adalah algoritma pembelajaran mesin yang menggabungkan beberapa pohon keputusan (decision tree) dalam satu model. Pada dasarnya, random forest menggunakan teknik ensemble learning atau penggabungan model untuk meningkatkan kinerja prediksi. Dalam random forest, setiap pohon keputusan dibuat dengan menggunakan subset data training yang dipilih secara acak dari keseluruhan data. Selain itu, dalam pembuatan setiap pohon keputusan, juga dilakukan pemilihan subset fitur secara acak dari keseluruhan fitur yang tersedia. Hal ini dilakukan untuk mengurangi kecenderungan overfitting dan meningkatkan kemampuan generalisasi model. Setelah terbentuk beberapa pohon keputusan, random forest menggabungkan hasil prediksi dari masing-masing pohon untuk menghasilkan prediksi akhir. Dalam klasifikasi, prediksi akhir diambil berdasarkan mayoritas suara, sedangkan dalam regresi, prediksi akhir diambil berdasarkan rata-rata hasil prediksi dari masing-masing pohon.

VI. KESIMPULAN

Dalam Technical Report ini kita mengetahui bahwa Machine Learning dapat digunakan untuk mengetahui diagnose kanker payudara melalui analisis dataset. Pada percobaan kali ini menggunakan dataset Breast Cancer Wisconsin (Diagnostic). Setelah melalui eksplorasi data, kita menemukan bahwa Random Forest adalah cara yang paling efektif untuk memprediksi diagnosis kanker payudara.