UTS Komputasi Lanjut dan Big Data Option 1 : Data Visualization

Sayyid Abdullah 2206130800

May 6, 2023

Abstract

Kanker adalah kondisi medis yang terjadi ketika sel-sel abnormal tumbuh secara tidak terkendali dan merusak jaringan sekitarnya. Penelitian ini akan dilakukan analisis dan memodelkan data penyakit kanker payudara dengan menggunakan decision tree, random forest, dan self training. Data kanker payudara diambil dari breast cancer dataset. Tools yang digunakan adalah google collab.

1 Pendahuluan

Kanker adalah kondisi medis yang terjadi ketika sel-sel abnormal tumbuh secara tidak terkendali dan merusak jaringan sekitarnya. Kanker dapat terjadi di hampir semua bagian tubuh dan dapat menyebar ke bagian lain dari tubuh melalui proses yang disebut metastasis. Ada berbagai jenis kanker, seperti kanker payudara, kanker paru-paru, kanker kulit, kanker usus besar, dan banyak lagi. Menurut [AYGB+19] kanker payudara adalah kanker paling umum pada wanita di seluruh dunia, dengan hampir 2 juta diagnosis kanker payudara baru setiap tahun.1 Di Amerika Serikat, diperkirakan ada 268.670 kasus kanker invasif pada tahun 2018. Faktor-faktor seperti genetika, gaya hidup, paparan lingkungan, dan faktor lainnya dapat meningkatkan risiko seseorang untuk mengembangkan kanker. Pengobatan kanker tergantung pada jenis dan stadium kanker, tetapi dapat mencakup operasi, radioterapi, kemoterapi, imunoterapi, atau kombinasi dari beberapa metode pengobatan. Banyak metode yang bisa digunakan ntuk mendiagnosa seseorang menderita kanker. salah satu contohnya adalah decision tree, random forest, self-traininng, dll.

Pada penelitian ini akan dilakukan analisis dan memodelkan data penyakit kanker payudara dengan menggunakan decision tree, random forest, dan self training. Data kanker payudara diambil dari breast cancer dataset. Tools yang digunakan adalah google collab.

2 Metodologi

2.1 Decision Tree

Decision tree (pohon keputusan) adalah salah satu metode analisis data dan pembuatan keputusan yang digunakan dalam ilmu komputer, statistik, dan machine learning. Metode ini menggunakan struktur seperti pohon untuk menggambarkan keputusan dan konsekuensi yang terkait. Pada dasarnya, sebuah decision tree terdiri dari node-node yang merepresentasikan keputusan atau pernyataan, cabang-cabang yang merepresentasikan kemungkinan hasil atau konsekuensi dari keputusan tersebut, dan daun-daun yang merepresentasikan hasil akhir atau output. Decision tree biasanya digunakan untuk membuat keputusan pada masalah klasifikasi dan regresi, dengan mengambil beberapa variabel input dan menghasilkan prediksi atau klasifikasi pada variabel output. Metode ini juga sering digunakan dalam pembuatan keputusan bisnis dan manajemen, seperti dalam analisis risiko dan pengambilan keputusan investasi.

Algoritma dari decision tree dapat dijelaskan dalam beberapa langkah berikut:

1. Memilih variabel input: Langkah pertama adalah memilih variabel input yang akan digunakan dalam pembuatan decision tree. Variabel input ini harus memiliki pengaruh yang signifikan terhadap variabel output.

2. Membuat decision node: Setelah variabel input dipilih, kita membuat decision node pada level atas pohon keputusan. Decision node ini merepresentasikan pertanyaan atau keputusan yang akan diambil berdasarkan variabel input.

3. Membuat cabang-cabang: Berdasarkan nilai variabel input, kita membuat cabang-cabang yang merepresentasikan kemungkinan hasil atau konsekuensi dari keputusan tersebut.

4. Menentukan kondisi terminal: Kondisi terminal merupakan daun-daun pada pohon keputusan yang merepresentasikan hasil akhir atau output. Kondisi terminal ini dicapai ketika tidak ada lagi variabel input yang dapat ditambahkan pada cabang-cabang tersebut.

5. Membangun sub-tree: Langkah terakhir adalah membangun sub-tree dari decision node sampai kondisi terminal terpenuhi. Proses ini dilakukan secara rekursif hingga terbentuk decision tree yang lengkap.

2.2 Random Forest

Random forest adalah salah satu metode machine learning yang digunakan untuk melakukan klasifikasi, regresi, dan pemilihan variabel pada data. Metode ini menggabungkan banyak pohon keputusan (decision tree) yang dibangun secara acak, untuk menghasilkan prediksi yang lebih akurat dan stabil.

Pada dasarnya, random forest terdiri dari beberapa langkah berikut:

1. Mengambil sampel acak dari data: Langkah pertama adalah mengambil sampel acak dari data yang tersedia. Sampel ini digunakan untuk membangun setiap pohon keputusan.

2. Membangun pohon keputusan: Setelah sampel acak dipilih, kita membangun decision tree untuk setiap sampel tersebut. Pada setiap level pohon keputusan, kita memilih variabel input secara acak dari subset variabel input yang tersedia.

3. Menggabungkan pohon keputusan: Setelah semua pohon keputusan selesai dibangun, kita menggabungkan hasil prediksi dari setiap pohon keputusan untuk menghasilkan prediksi akhir. Prediksi akhir ini didapatkan dengan mengambil rata-rata dari hasil prediksi setiap pohon keputusan dalam kasus klasifikasi atau regresi.

Keuntungan dari random forest adalah mampu mengurangi overfitting dan meningkatkan akurasi prediksi. Hal ini disebabkan karena setiap pohon keputusan dibangun dari sampel acak dan variabel input yang dipilih secara acak, sehingga setiap pohon keputusan menjadi unik. Selain itu, random forest juga dapat mengevaluasi pentingnya variabel input dalam prediksi dengan menghitung nilai mean decrease impurity.

2.3 Self-Training

Self-training adalah salah satu metode pembelajaran mesin semi-terawasi (semi-supervised learning) yang melibatkan penggunaan model yang sudah dilatih sebelumnya untuk menghasilkan label pada data yang belum diberi label.

Secara umum, self-training melibatkan dua tahap:

1. Latihan model awal: Pada tahap pertama, model dilatih pada dataset yang diberi label. Setelah dilatih, model dapat digunakan untuk memprediksi label pada dataset yang belum diberi label.

2. Pemilihan sampel dan labeling ulang: Pada tahap kedua, model digunakan untuk memprediksi label pada dataset yang belum diberi label. Kemudian, beberapa data dengan label yang paling tinggi keyakinannya akan diambil dan ditambahkan ke dataset yang sudah diberi label. Data-data tersebut kemudian dilatih ulang dengan model yang sudah diperbarui dan proses ini diulang hingga tidak ada data yang belum diberi label atau sampai pencapaian tingkat akurasi yang memadai.

Dalam self-training, model awal yang digunakan biasanya adalah model yang sangat bagus dalam klasifikasi atau regresi. Dengan demikian, self-training dapat meningkatkan akurasi dan performa model dengan mengumpulkan lebih banyak data yang dilatih, terutama pada kasus-kasus ketika dataset terbatas dalam ukuran atau terdapat kekurangan pengumpulan data terlabel.

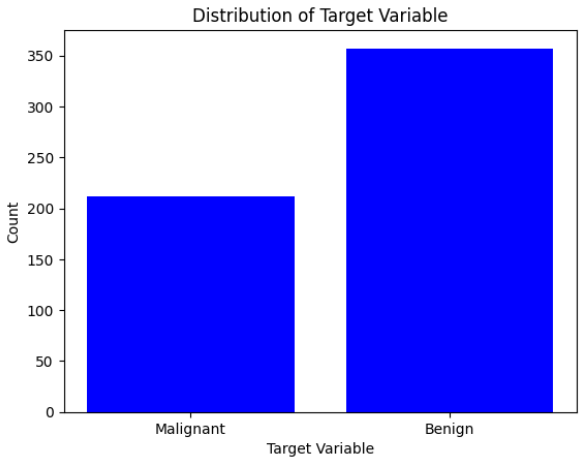

Figure 1: Visualisasi data target

```
[1 0 0 1 1 0 0 0 1 1 1 0 1 0 1 0 1 1 1 0 0 1 0 1 1 1 1 1 1 0 1 1 1 1 1 1 0
 1 0 1 1 0 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 0 0 1 1 0 0 1 1 1 0 0 1 1 0 0 1 0
 1 1 1 0 1 1 0 1 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 1 0 0 1 0 0 1 1 1 0 1 1 0
 1 1 0 1 0 1 1 1 0 1 1 1 0 1 0 0 1 1 0 0 0 1 1 1 0 1 1 1 0 1 0 1]
[1 0 0 1 1 0 0 0 1 1 1 0 1 0 1 0 1 1 1 0 1 1 0 1 1 1 1 1 1 0 1 1 1 1 1 1 0
 1 0 1 1 0 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 0 0 1 1 0 0 1 1 1 0 0 1 1 0 0 1 0
 1 1 1 0 1 1 0 1 1 0 0 0 0 0 0 0 1 1 1 1 1 0 1 1 1 0 0 1 0 0 1 0 0 1 1 1 0 1 1 0
 1 1 0 1 0 1 1 1 0 0 1 1 0 1 0 0 1 1 0 0 0 0 1 1 0 0 1 1 0 1 0 1]
Accuracy: 0.951
```

Figure 2: Hasil y prediksi

3 Hasil dan Pembahasan

Langkah pertama adalah memvisualisasikan data kanker payudara (breast cancer dataset). Dalam data tesebut terdapat sebanyak 30 variabel independen dan 1 variabel dependen yaitu target.Variabel yang divisualisasi berupa variabel target. Dari data target ada 2 nilai yaitu malignant(ganas) dan benign(jinak). Hasil visualiasi seperti Gambar 1. Pada Gambar 1, jumlah kasus kanker payudara ganas sebanyak 121 kasus dan untuk kasus kanker payudara jinak sebanyak 357 kasus.

Langkah selanjutnya adalah memodelkan dengan decision tree. Diperoleh hasil y prediksinya seperti pada Gambar 2. bagian array pertama adalah hasil y tesnya dan untuk array kedua adalah y prediksinya dengan akurasi 0,951. Langkah selanjutnya adalah melihat keterkaitan antara total impurity dengan α efektif tanpa memasukkan alfa efektif maksimum karena node trivial. Diperoleh seperti Gambar 3.

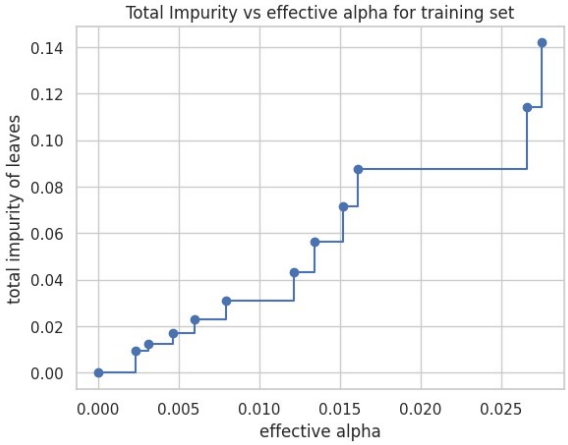

Figure 3: Total Impurity Vs α efektif untuk training set

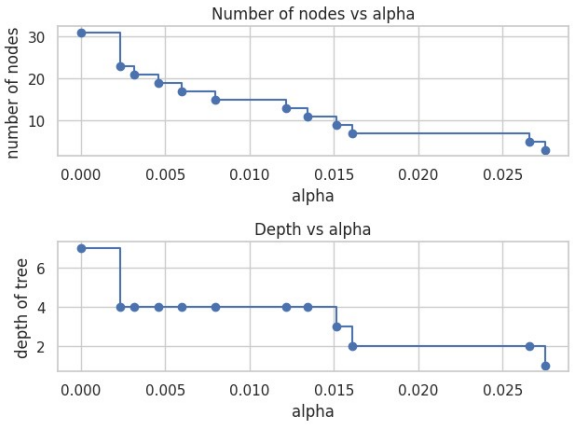

Figure 4: Jumlah node dan kedalaman pohon vs α

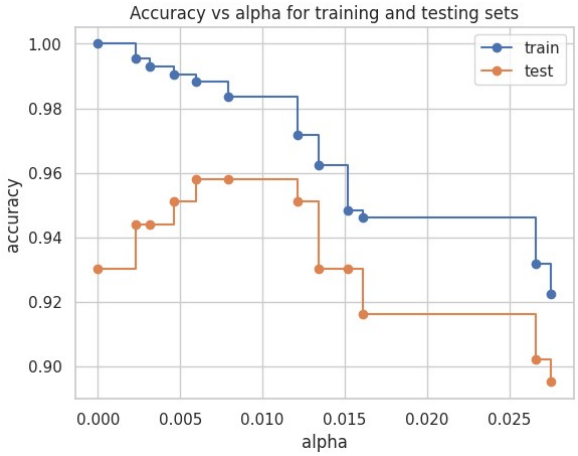

Figure 5: Akurasi Train dan Tes vs α

Model decision tree yang kompleks dan overfitting dapat menghasilkan model yang tidak dapat digeneralisasi dengan baik pada data baru. Untuk mengatasi masalah tersebut, digunakan teknik pruning untuk memotong cabang-cabang decision tree yang tidak signifikan sehingga model dapat lebih generalisasi. Dilakukan pruning menggunakan metode cost complexity pruning dengan mengoptimalkan nilai alpha. Selanjutnya adalah train model dengan α efektif dan diperoleh nilai α yang memangkas seluruh pohon pada simpul terakhir sebesar 0,3272984419327777.

Langkah selanjutnya adalah menghapus elemen terakhir karena itu adalah pohon trivial dengan hanya satu simpul. Setelah itu melihat keterkaitan jumlah node dan kedalaman pohon dengan α diperoleh seperti Gambar 4. Pada Gambar 4 disimpulkan bahwa semakin besar nilai α maka jumlah node dan kedalamannya akan semakin berkurang.

Selanjutnya adalah melihat akurasi train dan tes dengan membuat nilai α yang berbeda-beda diperoleh seperti pada Gambar 5. Pada gambar tersebut ketika α diatur ke nol dan mempertahankan parameter default lainnya dari data, pohon overfits, menghasilkan akurasi pelatihan 100% dan akurasi pengujian 88%. Saat α meningkat, lebih banyak pohon yang dipangkas, sehingga menciptakan pohon keputusan yang menggeneralisasi lebih baik. Dalam contoh ini, menyetel α 0,015 akan memaksimalkan akurasi pengujian.

Selanjutnya memodelkan dengan menggunakan random forest. Dari model diperoleh akurasi data sebesar 0.97 atau 97%. Karena dataset ini berisi variabel multikolinier, kepentingan permutasi akan menunjukkan bahwa tidak ada variabel yang penting. Salah satu pendekatan untuk menangani multikolinieritas adalah dengan melakukan pengelompokan hierarki pada Spearman rank-order correlations, memilih ambang batas, dan mempertahankan satu variabel dari setiap kluster.

Selanjutnya memplot kepentingan variabel berbasis pohon dan kepentingan permutasi seperti pada

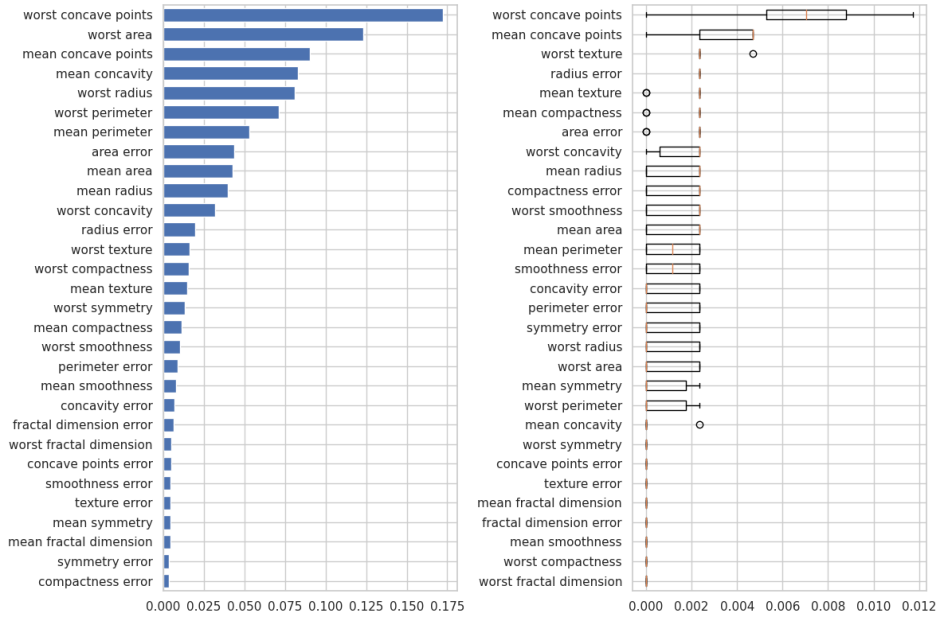

Figure 6: Kepentingan variabel berbasis pohon dan kepentingan permutasi

Gambar 6. Plot kepentingan permutasi menunjukkan bahwa mengubah variabel menurunkan akurasi paling banyak 0,012, yang menunjukkan bahwa tidak ada variabel yang penting. Ini bertentangan dengan akurasi tes yang dihitung di atas: beberapa variabel harus penting. Pentingnya permutasi dihitung pada set train untuk menunjukkan seberapa banyak model bergantung pada setiap variabel selama training.

Ketika variabel-variabelnya kolinear, permutasi satu variabel akan berdampak kecil pada kinerja model karena bisa mendapatkan informasi yang sama dari variabel yang dikorelasikan. Salah satu cara untuk menangani variabel multikolinier adalah dengan melakukan pengelompokan hierarkis pada korelasi urutan peringkat Spearman, memilih ambang batas, dan mempertahankan satu variabel dari setiap kluster. Pertama, memplot peta panas dari variabel yang dikorelasikan seperti Gambar 7. Selanjutnya, secara manual memilih ambang batas dengan inspeksi visual dendrogram untuk mengelompokkan variabel kami ke dalam kluster dan memilih variabel dari setiap kluster untuk dipertahankan, memilih variabel tersebut dari kumpulan data kami, dan memodelkan kembali. Keakuratan pengujian dari model baru tidak banyak berubah dibandingkan dengan model yang dilatih pada kumpulan data lengkap yaitu sama sebesar 0,97 atau 97%.

Selanjutnya menggunakan self training. Contoh ini mengilustrasikan efek dari berbagai ambang batas pada self training. Dataset breastcancer dimuat, dan label dihapus sehingga hanya 50 dari 569 sampel yang memiliki label. SelfTrainingClassifier dipasang pada dataset ini, dengan ambang batas yang bervariasi. Gambar 8 bagian atas menunjukkan jumlah sampel berlabel yang dimiliki pengklasifikasi pada akhir fit, dan akurasi pengklasifikasi. Gambar 8 bagian bawah menunjukkan iterasi terakhir di mana sampel diberi label. Semua nilai divalidasi silang dengan 3 lipatan. Pada rentang [0.4, 0.5], pengklasifikasi belajar dari sampel yang diberi label dengan kepercayaan rendah. Sampel berkeyakinan rendah ini kemungkinan besar memiliki label prediksi yang salah, dan akibatnya, pemasangan label yang salah ini menghasilkan akurasi yang buruk. Perhatikan bahwa classifier melabeli hampir semua sampel, dan hanya membutuhkan satu iterasi.

Untuk ambang batas yang sangat tinggi dalam rentang [0.9, 1] diperoleh bahwa pengklasifikasi tidak menambah kumpulan datanya (jumlah sampel yang diberi label sendiri adalah 0). Akibatnya, akurasi yang dicapai dengan ambang 0,9999 sama dengan yang dicapai oleh pengklasifikasi terbimbing biasa. Akurasi optimal terletak di antara kedua ekstrem ini pada ambang batas sekitar 0,7. Dari ketiga model, model dengan akurasi terbaik adalah model random forest dengan akurasi 0,97.

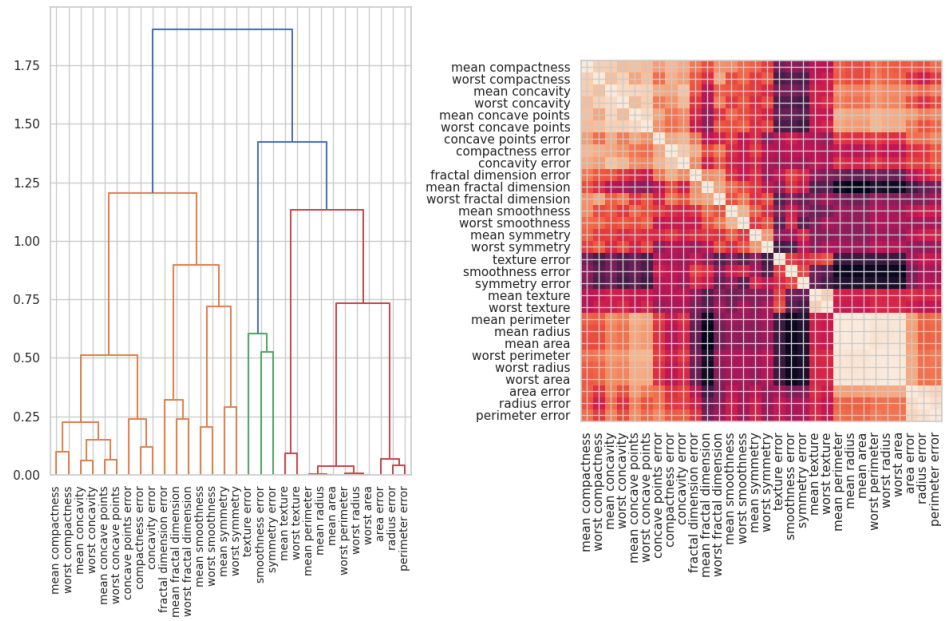

Figure 7: grafik panas dari variabel yg dikorelasikan

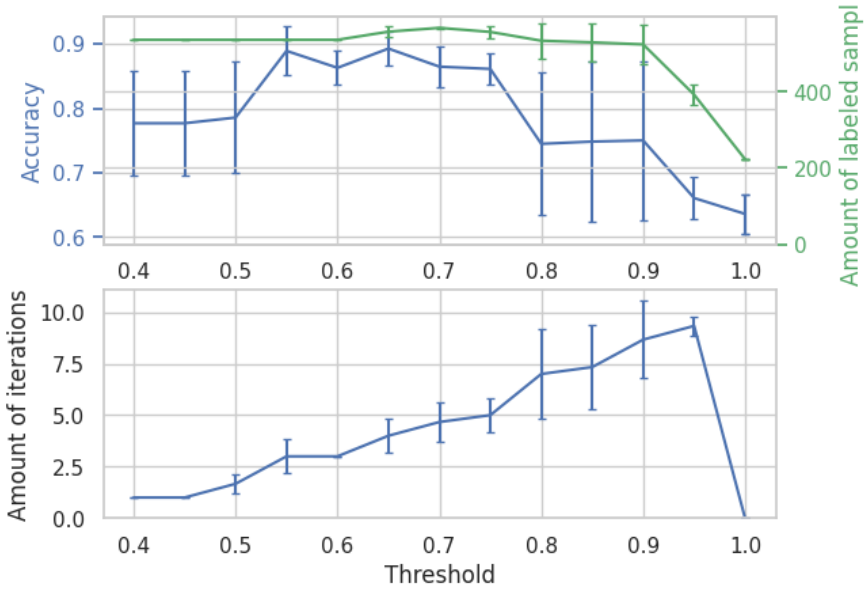

Figure 8: Ambang batas self-training

References

[AYGB+19] Arciero C. A., R. Jiang Y. Guo, M. Behera, R. O'regan, L. Peng, and X. Li. $ER^+/HER2^+$ Breast Cancer has Different Metastatic Patterns and Better Survival Than $ER^-/HER2^+$ Breast Cancer. *Clinical Breast Cancer*, 19(4):236–245, 2019.