Perbandingan Klasifikasi Kanker Payudara dengan Menggunakan Decision Tree dan Support Vector Machine

Disusun oleh Akbar Arta Putra , Ukasyah Rianza Alfarizi , Mikayla Almadevi

Pendahuluan

Payudara merupakan organ yang penting di tubuh manusia dan salah satu organ reproduksi pada wanita yang berfungsi untuk memproduksi ASI. Payudara terdiri dari kelenjar dan saraf, serta memiliki jaringan darah, seperti pembuluh darah dan pembuluh getah bening yang berguna untuk melawan infeksi. Dengan beberapa fungsi yang penting ini, menjaga payudara merupakan hal yang harus dilakukan supaya terhindar dari penyakit. Salah satu penyakit yang bisa menyerang organ tubuh ini adalah kanker payudara. Kanker payudara adalah sebuah penyakit di mana sel dalam payudara meningkat tidak terkendali, awalnya muncul di lobulus, kemudian di saluran susu, hingga akhirnya menyebar ke seluruh payudara. Sel yang tumbuh secara tidak normal tersebut dapat dirasakan sebagai benjolan yang disebut tumor.

Berdasarkan data yang dipublikasi oleh Globocan (Global Burden of Cancer) di tahun 2020, diketahui bahwa kanker payudara merepresentasikan 1 dari 4 kanker yang didiagnosis di kalangan wanita secara global. International Agency for Research on Cancer (IARC) memperkirakan bahwa pada tahun 2020 akan terdapat 2,2 juta kasus baru kanker payudara dan 684.996 kematian akibat kanker payudara di seluruh dunia [1]. Pada kenyataannya, menurut Data Globocan tahun 2020, kasus baru kanker payudara di Indonesia mencapai 68.858 kasus atau sekitar 16,6% dari total 396.914 kasus baru kanker di Indonesia. Sementara itu, jumlah kematian untuk kasus kanker payudara lebih dari 22 ribu jiwa kasus. Tingginya kasus kanker payudara membuat Kemenkes RI ingin melakukan pemerataan pelayanan kesehatan untuk kanker, terutama kanker payudara (Kementerian Kesehatan RI, 2022) [2].

Secara umum, penyebab terjadinya kanker payudara adalah karena terjadinya pertumbuhan yang tidak normal dari sel-sel pada payudara yang disebabkan oleh mutasi gen yang diturunkan secara genetik. Gen keturunan yang bermutasi diidentifikasi sebagai gen kanker payudara 1 (BRCA1) dan gen kanker payudara 2 (BRCA2). Selain gen keturunan, hal-hal yang menyebabkan terjadinya kanker payudara adalah gaya hidup, hormon, lingkungan tempat tinggal juga berpengaruh. Maka dari itu, dengan mengetahui kemungkinan apakah menderita kanker payudara sejak dini sangat penting agar dapat mengetahui langkah pencegahan atau pengobatan yang akan dilakukan untuk mengatasi kondisi yang terjadi. Untuk itu, perlu dilakukan algoritma klasifikasi yang sesuai dalam menentukan kemungkinan terkena kanker payudara. Pada penelitian ini, akan dilakukan perbandingan menggunakan algoritma Decision Tree dan Support Vector Machine untuk mengetahui algoritma yang terbaik mengklasifikasikan kanker payudara.

Metode dan Material

Dataset

Dataset yang digunakan dalam penelitian ini merupakan sampel yang didapatkan dari *website* Kaggle sejumlah 78.288 gambar untuk 7 kategori dengan setiap kategori memiliki 11.184 data. Dataset dibagi terlebih dahulu menjadi data training dan data testing dengan persentase 80% data training (8.947 data) dan 20% data testing (2.237 data).

Support Vector Machine

Support Vector Machine (SVM) diperkenalkan oleh Vapnik pada tahun 1992 sebagai suatu teknik klasifikasi yang efisien untuk masalah nonlinear. *Support Vector Machine* (SVM) juga dikenal sebagai teknik pembelajaran mesin (*machine learning*) paling mutakhir setelah pembelajaran mesin sebelumnya yang dikenal sebagai *Neural Network* (NN). Baik SVM maupun NN tersebut telah berhasil digunakan dalam pengenalan pola. Pembelajaran dilakukan dengan menggunakan pasangan data input dan data output berupa sasaran yang diinginkan. Konsep SVM dapat dijelaskan secara sederhana sebagai usaha mencari *hyperplane* terbaik yang berfungsi sebagai pemisah dua buah kelas pada input space. SVM berusaha menemukan fungsi pemisah (*hyperplane*) dengan memaksimalkan jarak antar kelas. Dengan cara ini, SVM dapat menjamin kemampuan generalisasi yang tinggi untuk data-data yang akan datang[3].

Decision Tree

Decision tree atau pohon keputusan adalah algoritma *machine learning* yang menggunakan seperangkat aturan untuk membangun klasifikasi dalam bentuk struktur pohon yang memodelkan kemungkinan hasil, biaya sumber daya, utilitas dan kemungkinan konsekuensi atau resiko. Konsepnya adalah dengan cara menyajikan algoritma dengan pernyataan bersyarat, yang meliputi cabang untuk mewakili langkah-langkah pengambilan keputusan yang dapat mengarah pada hasil yang menguntungkan. Dimana setiap cabang mewakili hasil untuk atribut, sedangkan jalur dari daun ke akar mewakili aturan untuk klasifikasi. Algoritma ini disebut dengan pohon keputusan karena pilihannya bercabang, membentuk struktur yang terlihat seperti pohon.

***Area Under Curve* (AUC)**

Area Under Curve (AUC) adalah suatu ukuran yang digunakan untuk membandingkan model klasifikasi berdasarkan kinerjanya. Range nilai dari AUC sendiri berkisar antara 0 sampai 1. Nilai AUC terbaik adalah nilai yang mendekati 1.

$$AUC = \frac{1+TP_{rate}-FP_{rate}}{2}$$

Nilai AUC dibagi menjadi beberapa kategori berdasarkan tingkat kebaikannya. Berikut tabel kategori AUC.

Tabel 1. Kategori nilai AUC

Nilai AUC	Kategori
0.90-1.00	*Excellent Classification*
0.80-0.90	*Good Classification*
0.70-0.80	*Fair Classification*
0.60-0.70	*Poor Classification*
0.50-0.60	*Failure*

Sintesis

Berdasarkan metode dari konsep penelitian di atas, maka urutan pengerjaannya sebagai berikut:

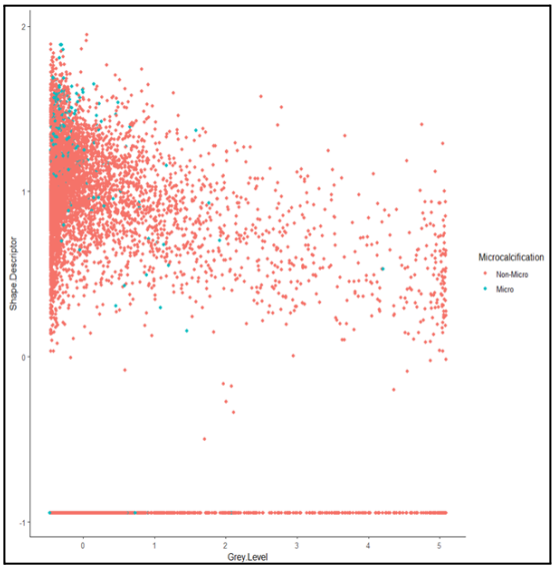
1. Tahap Pengumpulan Data
   Pada tahap pengumpulan data penelitian ini, yaitu data sekunder diambil dari *website* Kaggle.
2. Tahap Preprocessing
   Dalam tahap ini, data yang telah diperoleh akan diolah untuk menemukan data yang diperlukan dan membuang data-data yang tidak diperlukan untuk proses analisis. Teknik pengolahan data pada tahap pre-processing meliputi:
   a. Input data
   b. Cleaning Data
   c. Penyeimbangan Kelas
3. Pembagian Data
   Dalam penelitian, data yang telah diperoleh akan dibagi menjadi *data training* dan *data testing.*
   a. *Data training*, yaitu data yang akan diolah dengan metode yang akan digunakan. Hasilnya akan digunakan sebagai prediksi untuk data testing.
   b. *Data testing* merupakan data yang akan diuji dan diprediksi.
4. Metode *Decision Tree* dan SVM
5. Hasil dan Kesimpulan
   Dalam tahap ini dilakukan penalaran kesimpulan yang diperoleh berdasarkan hasil dari perbandingan klasifikasi dengan menggunakan algoritma *Support Vector Machine* dan *Decision Tree*.

Hasil dan Pembahasan

Pre-Processing Data

Data Pre-processing merupakan salah satu tahapan dalam melakukan mining data sebelum menuju ke tahap pemrosesan. Data mentah akan diolah terlebih dahulu. *Data Pre-processing* atau pra-proses data biasanya dilakukan melalui cara eliminasi data yang tidak sesuai. Selain itu dalam proses ini data akan diubah dalam bentuk yang akan lebih dipahami oleh sistem. Berikut *pre-processing* data pada penelitian ini.

1. **Visualisasi Data**



Gambar 1. Visualisasi data kanker payudara untuk variabel *microcalcification*

Penelitian ini akan dilakukan klasifikasi untuk membedakan klasifikasi *micro* dan *non-micro*. *Non-micro* adalah representasi dari seseorang yang tidak mengidap penyakit kanker payudara, sedangkan *micro* adalah representasi dari seseorang yang mengidap penyakit kanker. Gambar 1 menampilkan sebaran data untuk variabel microcalcification. Berdasarkan gambar 1 dapat dilihat bahwa sebaran data kelas antara *micro* dan *non-micro* tidak seimbang dengan proporsi 97.67% *non-micro* dan 2.33% *micro*, sehingga diperlukan suatu metode untuk menyeimbangkan kelas tersebut.

2. ***Cleaning Data* (Cek *Missing Value*)**

Missing Value merupakan hilangnya beberapa data yang telah diperoleh. Dalam dunia *data science*, *missing value* erat kaitannya dalam proses perselisihan data (*data wrangling*) sebelum nantinya akan dilakukan analisis dan prediksi data. Berikut tabel jumlah *missing value* pada data kanker payudara.

Tabel 2. Jumlah *missing value* pada masing-masing variabel

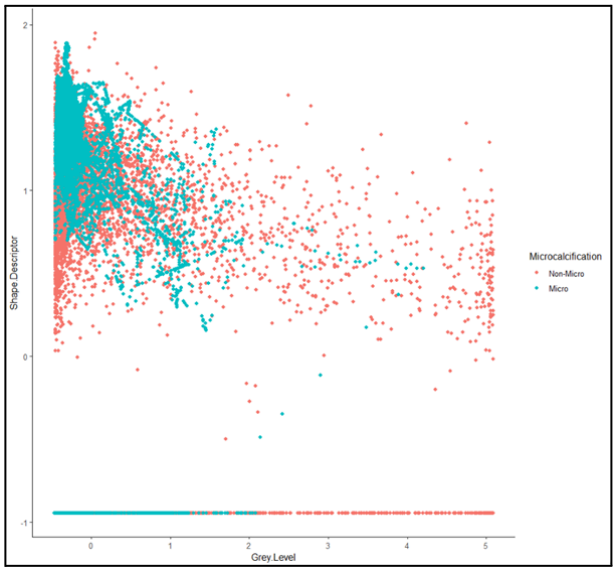
Variabel	Jumlah Missing Value
Area	0
Grey.Level	0
Gradient.Strength	0
Noise.Fluctuation	0
Contrast	0

Shape.Descriptor	0
Microcalcification	0

Berdasarkan tabel 1 dapat disimpulkan bahwa tidak terdapat missing value pada data penelitian. Sehingga dapat dilanjutkan ke tahap berikutnya.

3. Menyeimbangkan kelas

Ketidakseimbangan kelas pada klasifikasi merupakan suatu permasalahan yang sering terjadi. Ketidakseimbangan ini dapat menyebabkan algoritma pembelajaran mesin tidak berjalan dengan baik, hal tersebut dapat mempengaruhi hasil klasifikasi yang akan dilakukan. Untuk menangani permasalahan tersebut akan dilakukan penyeimbangan kelas dengan menggunakan metode smote. Berikut hasil penyeimbangan data dengan menggunakan metode smote.



Gambar 2. Hasil penyeimbangan data dengan menggunakan metode smote

Ketidakseimbangan kelas pada klasifikasi merupakan suatu permasalahan yang sering terjadi. Setelah dilakukan penyeimbangan kelas, kelas lebih seimbang dengan proporsi proporsi 54.55% *non-micro* dan 45.45% *micro*.

Pembagian Data

Dalam pemodelan, sampel data perlu dibagi menjadi dua subset, yaitu *data training* dan *data testing*, dimana jumlah *data training* harus lebih besar dari jumlah data testing untuk pengujian model. Pada data akan diambil sebanyak 20% untuk testing dan sebanyak 80% untuk training. Data training digunakan untuk membangun model dan data testing digunakan untuk mengetahui akurasi model.

Metode *Decision Tree* dan SVM

Metode yang digunakan untuk klasifikasi pada penelitian ini adalah metode *decision tree* dan *support vector machine*. Metode decision tree akan menggunakan beberapa algoritma, sedangkan untuk metode *support vector machine* akan menggunakan beberapa kernel. Untuk melihat kebaikan dari model yang dibentuk dari masing-masing metode akan digunakan beberapa parameter, yaitu :

1. *Accuracy*

   Seberapa mampu model menebak dengan benar variabel target Y.

2. *Recall*

   Dari semua data aktual yang positif, seberapa mampu proporsi model menebak dengan benar kelas

positif.

3. *Specificity*
   Dari semua data aktual yang negatif, seberapa mampu proporsi model menebak dengan benar kelas negatif.
4. *Precision*
   Dari semua hasil prediksi yang positif, seberapa mampu model menebak dengan benar kelas positif.

Berikut tabel hasil pengukuran kebaikan model dari masing-masing metode.

Tabel 3. Kebaikan model dengan metode *Decision Tree*

Metode	Algoritma	*Accuracy*	*Recall*	*Specificity*	*Precision*
Decision Tree	C5.0	0.961	0.964	0.957	0.964
	Random Forest	0.973	0.981	0.962	0.969

Tabel 3 menunjukan ukuran kebaikan model yang dihasilkan oleh metode *Decision Tree*. Metode *Decision Tree* menggunakan dua algoritma, yaitu algoritma C5.0 dan algoritma *random forest*. Algoritma C5.0 merupakan pohon keputusan non-biner di mana cabang pohon bisa lebih dari dua, sedangkan algoritma *random forest* merupakan algoritma *machine learning* yang menggabungkan keluaran dari beberapa decision tree untuk mencapai satu hasil. Model-model yang dibentuk dari metode *Decision Tree* memiliki nilai yang sangat baik, karena memiliki nilai di atas 90%.

Tabel 4. Kebaikan model dengan metode *Support Vector Machine*

Metode	Kernel	*Accuracy*	*Recall*	*Specificity*	*Precision*
Support Vector Machine	Linear	0.888	0.906	0.867	0.891
	Radial	0.934	0.955	0.908	0.927
	Polinomial	0.919	0.917	0.921	0.934
	Sigmoid	0.832	0.842	0.820	0.850

Tabel 4 menunjukan ukuran kebaikan model yang dihasilkan oleh metode *Support Vector Machine* Metode *Support Vector Machine* menggunakan 4 kernel algoritma, yaitu linear, radial, polinomial, dan sigmoid. Model-model yang dibentuk dari metode *Support Vector Machine* memiliki nilai yang sangat baik, karena memiliki nilai di atas 80%.

Setelah dilihat nilai kebaikan masing-masing model, langkah selanjutnya adalah memilih model terbaik. Model terbaik adalah model yang memiliki nilai *Area Under Curve* (AUC) terbesar. Berikut nilai AUC dari masing-masing model.

Tabel 5. Nilai AUC dari masing-masing model

Model	Nilai AUC
Decision Tree C5.0	0.9602405
Decision Tree Random Forest	0.9716186

SVM Linear	0.8864988
SVM Radial	0.9316374
SVM Polinomial	0.9192761
SVM Sigmoid	0.8308946

Setelah dilihat nilai kebaikan masing-masing model, langkah selanjutnya adalah memilih model terbaik. Model terbaik Pada tabel 5 nilai AUC terbesar adalah model *Decision Tree Random Forest* sebesar 0.9716186, tidak hanya itu model *Decision Tree Random Forest* juga mempunyai nilai kebaikan tertinggi. Oleh karena itu, model terbaik untuk klasifikasi data kanker payudara pada penelitian ini adalah *Decision Tree Random Forest*.

Kesimpulan

Pada penelitian ini, perbandingan klasifikasi kanker payudara dilakukan dengan menggunakan sampel acak yang didapatkan dari *website* Kaggle sejumlah 78.288 gambar untuk 7 kategori dengan setiap kategori memiliki 11.184 data. Klasifikasi kanker payudara menggunakan algoritma *Decision Tree* dan *Support Vector Machine* yang membentuk 6 model algoritma, yaitu *Decision Tree* C5.0, *Decision Tree* algoritma *Random Forest*, SVM Linear, SVM Radial, SVM Polinomial, dan SVM Sigmoid. Berdasarkan hasil klasifikasi dari 6 model yang terbentuk, didapatkan model yang terbaik untuk klasifikasi, yaitu *Decision Tree* algoritma *Random Forest* karena memiliki nilai AUC terbesar diantara model lainnya, yaitu senilai 0.9716186.

REFERENSI

1. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 71(3), 209-249.
2. Kemenkes, R. I. (2022). Kanker Payudara Paling Banyak di Indonesia, Kemenkes Targetkan Pemerataan Layanan Kesehatan. Diakses melalui website: https://www. kemkes.go.id/article/view/220204, 2.
3. Han. J., Kamber. M., & Pei. J., “Data Mining: Concepts and Techniques Third edition”, Waltham: Elsevier, 2011
4. Institutional Repository UIN Syarif Hidayatullah Jakarta: Invalid Identifier. (n.d.). https://repository.uinjkt.ac.id/dspace/bitstream/123456789/56253/1/ALAM+WAHYU+HUTOMO-FST.pdf