

PENERAPAN MACHINE LEARNING DALAM PREDIKSI METASTASIS TULANG PADA KARSINOMA PAYUDARA: STUDI DI RS KANKER DHARMAIS

PRADNYA SRI RAHAYU



**PROGRAM STUDI MAGISTER STATISTIKA DAN SAINS DATA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2025**

PENERAPAN MACHINE LEARNING DALAM PREDIKSI METASTASIS TULANG PADA KARSINOMA PAYUDARA: STUDI DI RS KANKER DHARMAIS

PRADNYA SRI RAHAYU



**PROGRAM STUDI MAGISTER STATISTIKA DAN SAINS DATA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2025**

- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



PERNYATAAN MENGENAI TESIS DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa tesis dengan judul “Penerapan *Machine Learning* dalam Prediksi Metastasis Tulang Pada Karsinoma Payudara: Studi di RS Kanker Dharmais” adalah karya saya dengan arahan dari dosen pembimbing dan belum diajukan dalam bentuk apapun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir tesis ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, 14 Agustus 2025
Pradnya Sri Rahayu
G1501211021

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

RINGKASAN

PRADNYA SRI RAHAYU. Penerapan *Machine Learning* dalam Prediksi Metastasis Tulang Pada Karsinoma Payudara: Studi di RS Kanker Dharmais. Dibimbing oleh UTAMI DYAH SYAFITRI dan AJI HAMIM WIGENA

Pembelajaran mesin (*Machine learning/ML*) sebagai bagian dari kecerdasan buatan, memungkinkan komputer belajar dari data tanpa pemrograman eksplisit dan telah berkembang sejak pertengahan abad ke-20 melalui berbagai algoritma seperti *support vector machines* dan *random forest* yang kini banyak digunakan dalam bidang medis, termasuk prediksi metastasis kanker payudara. ML terbagi menjadi dua pendekatan utama, yaitu *supervised* dan *unsupervised learning*, dengan *supervised learning* seperti regresi logistik, random forest, dan SVM banyak digunakan dalam penelitian klinis karena kemampuannya dalam klasifikasi dan interpretasi hasil. Salah satu penelitian klinis adalah pada penelitian kanker dimana salah satu karakteristiknya adalah kemampuannya bermetastasis sehingga perlu adanya suatu studi yang bertujuan mengevaluasi kinerja model *Logistic Regression*, *Random Forest*, dan *Support Vector Machine* dalam memprediksi metastasis tulang, menentukan model terbaik di antara ketiganya, serta mengidentifikasi peubah paling signifikan yang memengaruhi prediksi, guna mendukung praktik klinis yang lebih akurat dan personal di bidang onkologi

Dalam penelitian ini, variabel prediktor untuk memodelkan risiko metastasis tulang pada pasien kanker payudara dikelompokkan menjadi tiga kategori utama: subtype molekuler (ER, PR, HER2, Ki-67), karakteristik biologis tumor (morfologi, grade, invasi limfovaskular), dan faktor sosial (usia). Analisis bivariat menggunakan uji *Chi-square* menunjukkan bahwa hanya ekspresi ER dan PR yang memiliki hubungan signifikan dengan metastasis tulang ($p < 0,001$), sementara variabel lainnya tidak menunjukkan hubungan statistik yang signifikan, namun tetap dipertimbangkan dalam model prediksi karena potensi interaksi antar peubah dan relevansi klinis. Untuk membangun model prediksi yang lebih komprehensif, digunakan tiga algoritma ML: *Logistic Regression*, *Random Forest*, dan *Support Vector Machine* (SVM), dengan evaluasi dilakukan melalui 20 iterasi dan menggunakan berbagai metrik (akurasi, sensitivitas, spesifisitas, precision, F1 score).

Hasil evaluasi menunjukkan bahwa meskipun *Random Forest* unggul dalam *single run*, performanya menurun drastis saat diuji berulang, menandakan ketidakstabilan model. Sebaliknya, *Logistic Regression* tampil paling stabil dengan akurasi 66%, sensitivitas 60%, dan *F1-score* 68%, menjadikannya model terbaik secara keseluruhan. Analisis SHAP pada model ini menunjukkan bahwa ekspresi ER dan Ki-67 adalah prediktor terkuat, diikuti oleh morfologi, usia, dan HER2, sementara PR dan invasi limfovaskular memiliki kontribusi rendah. Konsistensi nilai SHAP antara data latih dan uji juga menunjukkan model memiliki generalisasi yang baik. Pendekatan ini menunjukkan bahwa integrasi data biologis dan sosial, didukung oleh teknik interpretabilitas seperti SHAP, mampu menghasilkan model prediktif yang akurat dan relevan secara klinis untuk deteksi dini metastasis tulang pada kanker payudara.

Kata kunci: Kanker payudara, Metastasis tulang, Pembelajaran mesin, Prediksi

SUMMARY

PRADNYA SRI RAHAYU. Application of Machine Learning for Bone Metastasis Prediction in Breast Carcinoma: A Study at Dharmas Cancer Hospital. Supervised by UTAMI DYAH SYAFITRI and AJI HAMIM WIGENA.

Machine learning (ML), as a part of artificial intelligence, enables computers to learn from data without explicit programming and has developed since the mid-20th century through various algorithms such as support vector machines and random forests, which are now widely used in the medical field, including predicting breast cancer metastasis. ML is divided into two main approaches: supervised and unsupervised learning. Supervised learning, such as logistic regression, random forest, and SVM, is widely used in clinical research due to its ability in classification and result interpretation. One example of clinical research is in cancer studies, where one characteristic is its ability to metastasize, thus requiring a study aimed at evaluating the performance of Logistic Regression, Random Forest, and Support Vector Machine models in predicting bone metastasis, determining the best model among the three, and identifying the most significant variables affecting predictions to support more accurate and personalized clinical practice in oncology.

In this study, predictor variables for modeling the risk of bone metastasis in breast cancer patients were grouped into three main categories: molecular subtypes (ER, PR, HER2, Ki-67), tumor biological characteristics (morphology, grade, lymphovascular invasion), and social factors (age). Bivariate analysis using the Chi-square test showed that only ER and PR expression had a significant relationship with bone metastasis ($p < 0,001$), while other variables did not show statistically significant relationships but were still considered in the prediction model due to potential variable interactions and clinical relevance. To build a more comprehensive predictive model, three ML algorithms were used: Logistic Regression, Random Forest, and Support Vector Machine (SVM), with evaluation conducted through 20 iterations and using various metrics (accuracy, sensitivity, specificity, precision, F1 score).

The evaluation results showed that although Random Forest excelled in a single run, its performance drastically decreased when tested repeatedly, indicating model instability. In contrast, Logistic Regression was the most stable, with an accuracy of 66%, sensitivity of 60%, and F1-score of 68%, making it the best overall model. SHAP analysis on this model showed that ER expression and Ki-67 were the strongest predictors, followed by morphology, age, and HER2, while PR and lymphovascular invasion had low contributions. The consistency of SHAP values between training and test data also indicated that the model had good generalization. This approach demonstrates that integrating biological and social data, supported by interpretability techniques such as SHAP, can produce an accurate and clinically relevant predictive model for early detection of bone metastasis in breast cancer.

Keyword: Bone metastasis, Breast cancer, Machine learning, Prediction

© Hak Cipta milik IPB, tahun 2025
Hak Cipta dilindungi Undang-Undang

Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan atau menyebutkan sumbernya. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik, atau tinjauan suatu masalah, dan pengutipan tersebut tidak merugikan kepentingan IPB.

Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apa pun tanpa izin IPB.

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



PENERAPAN MACHINE LEARNING DALAM PREDIKSI METASTASIS TULANG PADA KARSINOMA PAYUDARA: STUDI DI RS KANKER DHARMAIS

PRADNYA SRI RAHAYU

Tesis
sebagai salah satu syarat untuk memperoleh gelar
Magister Sains pada
Program Studi Statistika dan Sains Data

**PROGRAM STUDI MAGISTER STATISTIKA DAN SAINS DATA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2025**



@Hak cipta milik IPB University

IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Tim Penguji pada Ujian Tesis:

1. Prof. Dr. Ir. Anik Djuraidah M.S. _____

Judul Tesis : Penerapan *Machine Learning* dalam Prediksi Metastasis Tulang pada Karsinoma Payudara: Studi di RS Kanker Dharmais

Nama : Pradnya Sri Rahayu

NIM : G1501211021

Disetujui oleh

Pembimbing 1:

Dr. Utami Dyah Syafitri, S.Si., M.Si

Pembimbing 2:

Prof. Dr. Ir. Aji Hamim Wigena, M.Sc

Diketahui oleh

Ketua Program Studi:

Dr. Agus Mohamad Soleh S.Si., M.T
19750315 199903 1004

Dekan Sekolah Sains Data, Matematika, Dan Informatika

Prof. Dr. Ir. Agus Buono, M.Si, M.Kom
19660702 199302 1001

Tanggal Ujian: 11 Agustus 2025

Tanggal Lulus:

@Hak cipta milik IPB University

IPB University



PRAKATA

Puji syukur penulis panjatkan ke hadirat Allah SWT atas limpahan rahmat, karunia, dan kekuatan-Nya sehingga penelitian tesis berjudul “Penerapan Machine Learning dalam Prediksi Metastasis Tulang pada Karsinoma Payudara: Studi di RS Kanker Dharmais” dapat diselesaikan. Penelitian ini disusun sebagai bagian dari pemenuhan salah satu syarat untuk memperoleh gelar Magister Sains pada Program Studi Statistika dan Sains Data. Perjalanan penyusunan tesis ini bukanlah hal yang mudah. Setelah sempat tertunda sejak semester empat, penulis bersyukur akhirnya dapat melangkah kembali dan menyelesaikan tahap penting dalam proses akademik ini berkat doa, dukungan, dan bimbingan dari berbagai pihak.

Dengan penuh rasa hormat dan ketulusan hati, penulis menyampaikan terima kasih yang sebesar-besarnya kepada Dr. Utami Dyah Syafitri, S.Si., M.Si. dan Prof. Dr. Ir. Aji Hamim Wigena, M.Sc. selaku komisi pembimbing atas motivasi, arahan, dan kesempatan yang tiada henti, bahkan ketika penulis mengalami jeda yang cukup panjang. Ucapan terima kasih juga disampaikan kepada Dr. Mersi Kurniati, S.Si., M.Si. selaku moderator seminar, serta Prof. Dr. Ir. Anik Djuraidah, M.S. selaku penguji luar komisi pembimbing. Penghargaan mendalam penulis berikan kepada suami tercinta, almarhum Papa, Mama, kakak, dan adik, atas dukungan moral dan spiritual yang tidak pernah terputus. Terima kasih juga kepada Kepala Instalasi Pengendali Data Beban Kanker dan Jejaring Kanker Nasional RS Kanker Dharmais serta rekan-rekan sejawat di Subinstalasi Registrasi Kanker atas pengertian dan semangat di tengah kesibukan pekerjaan. Penulis juga berterima kasih kepada seluruh dosen di Departemen Statistika atas ilmu dan wawasan yang telah diberikan, serta kepada rekan-rekan mahasiswa Pascasarjana angkatan 2021, khususnya gelombang akhir, atas kebersamaan dan dukungan dalam menyelesaikan tesis ini.

Akhirnya, penulis berharap hasil penelitian ini dapat memberikan kontribusi nyata bagi pengembangan ilmu pengetahuan, khususnya di bidang onkologi dan penerapan machine learning dalam dunia medis. Semoga penelitian ini menjadi pijakan awal bagi langkah-langkah penelitian selanjutnya dan bermanfaat bagi berbagai pihak yang berkecimpung dalam bidang ini.

Untuk Mama, *Keep surviving* Ma...

Bogor, 12 Agustus 2025

Pradnya Sri Rahayu

DAFTAR ISI

DAFTAR TABEL	x
DAFTAR GAMBAR	xi
I PENDAHULUAN	1
1.1 Latar belakang.....	1
1.2 Perumusan Masalah	3
1.3 Tujuan	3
II TINJAUAN PUSTAKA.....	4
2.1 <i>Machine Learning</i>	4
2.1.1 Tipe <i>machine learning</i>	5
2.2 Model Klasifikasi dan Regresi pada <i>Machine Learning</i>	6
2.2.1 <i>Binary Logistic Regression</i>	6
2.2.2 <i>Support Vector Machine</i>	8
2.2.3 <i>Random Forest</i>	10
2.3. Pemilihan Data Latih dan Data Uji	12
2.4 Analisis Asosiasi antara Peubah Penjelasa dan Peubah Respon.....	13
2.5 Indikator Evaluasi Keabakan Model.....	13
2.6 Metode SHAP	17
2.6.1 SHAP pada Regresi Logistik	17
2.6.2 SHAP pada <i>Random Forest</i>	18
2.6.3 SHAP pada <i>Support Vector Machine</i>	18
2.7 Registrasi Kanker	18
2.7.1 Registrasi Kanker di RS Kanker Dharmais.....	18
2.8 Kanker Payudara	19
2.8.1 Karsinoma Payudara	19
2.8.2 Standar Laporan Patologi Karsinoma Payudara	19
2.9 Metastasis Tulang.....	19
III METODE	21
3.1 Data	21
3.2 Tahapan Penelitian	22
3.3 Analisis Data	23
3.3.1 Mengimport data dari File Excel	23
3.3.2 Melakukan eksplorasi data	23
3.3.3 Menguji asosiasi antara peubah penjelas dan peubah respon	23
3.3.4 Membagi data menjadi data latih data uji sebelum pemodelan	23
3.3.5 Melakukan pemodelan dengan regresi logistik biner.....	24
3.3.6 Melakukan pemodelan dengan <i>random forest</i>	24
3.3.7 Melakukan pemodelan dengan <i>support vector machine</i>	25
3.3.8 Membandingkan confusion matrix dari ketiga model	26
3.3.9 Menganalisa SHAP dan perbandingannya antara data latih dan data uji	26
IV HASIL DAN PEMBAHASAN	27
4.1 Hubungan antara peubah prediktor dan peubah hasil	27
4.2 Pemodelan Regresi Logistik	29
4.3 Pemodelan <i>Random Forest</i>	31

4.4	Pemodelan <i>Support Vector Machine</i>	34
4.5	Perbandingan evaluasi performa model	37
4.6.	Analisis SHAP	38
V	SIMPULAN DAN SARAN	41
5.1	Simpulan	41
5.2	Saran.....	41
	DAFTAR PUSTAKA	42



Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

DAFTAR TABEL

1 Klasifikasi Molekuler Karsinoma Payudara	19
2 Peubah respon dan peubah penjelas	22
3 Distribusi dan hubungan antara peubah penjelas dengan peubah respon.....	28
4 Perbandingan performa model regresi logistik pada berbagai <i>cutoff</i>	30
5 Analisis kinerja model regresi logistik pada nilai <i>cutoff</i> 0,4 dengan 20 kali iterasi	31
6 Hasil Tuning Hyperparameter pada Model Random Forest	32
7 Evaluasi performa model <i>random forest</i> pada parameter terbaik	33
8 Analisis kinerja model <i>random forest</i> pada <i>ntree</i> = 144 dan <i>mtry</i> = 2 dengan 20 kali iterasi	34
9 Perbandingan evaluasi performa model <i>support vector machine</i>	35
10 Perbandingan evaluasi performa model <i>support vector machine</i> tanpa pembobotan	35
11 Analisis kinerja model <i>support vector machine</i> pada <i>cost</i> = 0,01 dan terboboti dengan 20 kali iterasi	36
12 Ringkasan Metrik Evaluasi Model (<i>Single Run</i> vs 20 Iterasi)	38



DAFTAR GAMBAR

1 Fungsi logistik.....	7
2 <i>Hyperplane</i> (Garis pemisah kedua kelas)	8
3 Kemungkinan-kemungkinan <i>Hyperplane</i>	8
4 <i>Support vector, maximized margin, dan hyperplane</i>	9
5 <i>Linear SVM</i>	10
6 Diagram Arsitektur <i>Random Forest</i>	11
7 Diagram pembentukan <i>tree</i> pada <i>random forest</i>	12
8 Manajemen Data	21
9 Evaluasi performa model dengan <i>single run</i>	37
10 Evaluasi performa model dengan rata-rata dari 20 iterasi	38
11 Perbandingan Analisis <i>Feature Importance</i> menggunakan metode SHAP pada model SVM	39

