JITE, 6 (2) January 2023 ISSN 2549-6247 (Print) ISSN 2549-6255 (Online)

JITE (Journal of Informatics and Telecommunication Engineering)

Available online http://ojs.uma.ac.id/index.php/jite **DOI : 10.31289/jite.v6i2.8373**

Received: 14 November 2022	Accepted: 17 January 2023	Published: 25 January 2023

Optimization of Support Vector Machine and XGBoost Methods Using Feature Selection to Improve Classification Performance

Doni Andriansyah1)*, Eka Wulansari Fridayanthie2)

1)Program Studi Informatika, Fakultas Teknologi Informasi, Universitas Nusa Mandiri, Indonesia
2) Program Studi Sistem Informasi, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika, Indonesia

*Coresponding Email: doni.dad@nusamandiri.ac.id

Abstrak

Kanker payudara masih menempati urutan pertama dalam jumlah kanker terbanyak di Indonesia dan menjadi salah satu penyumbang angka kematian yang diakibatkan oleh kanker. Umumnya terjadi pada wanita, namun tidak menutup kemungkinan dapat terjadi pada pria. Data Globocan pada tahun 2020 menunjukkan jumlah kasus kanker baru di Indonesia mencapai 65.858 atau 16,6% dari total 396.914 kasus kanker baru, dengan jumlah kematian mencapai 22.430 kasus. Pendeteksian dapat dilakukan secara dini sehingga memungkinkan pasien mendapatkan terapi yang tepat dan meningkatkan peluang untuk dapat bertahan hidup. Tujuan dari penelitian ini adalah mengimplementasikan algoritma pembelajaran mesin untuk mendeteksi kanker payudara pada wanita, algoritma yang akan digunakan adalah Support Vector Machine (SVM) dan XBGoost dengan menerapkan seleksi fitur untuk memperoleh akurasi yang lebih baik. Hasil klasifikasi dari kedua algoritma akan dibandingkan untuk mengetahui algoritma yang memiliki performa terbaik. Dataset yang digunakan berasal dari program SEER NCI pada November 2017 dengan melibatkan 4024 pasien. Penelitian menunjukkan dari 16 atribut yang terdapat pada dataset, ada 3 atribut (fitur) yang berpengaruh secara signifikan terhadap hasil klasifikasi yaitu 6th stage, reginol node positive, dan tumor size. XGBoost dengan seleksi fitur memiliki performa lebih baik sebesar 91,4 % dibandingkan dengan SVM yang hanya sebesar 89,8 %.

Kata Kunci: kanker payudara, pembelajaran mesin, SVM, XGBoost.

Abstract

Breast cancer still ranks first in the number of cancers in Indonesia and is one of the contributors to the death rate caused by cancer. Generally occurs in women, but does not rule out it can occur in men. Globocan data in 2020 shows the number of new cancer cases in Indonesia reached 65,858 or 16.6% of the total 396,914 new cancer cases, with the number of deaths reaching 22,430 cases. Early detection can allow patients to get the right therapy and increase their chances of survival. The purpose of this study is to implement a machine learning algorithm to detect breast cancer in women, the algorithms that will be used are Support Vector Machine (SVM) and XBGoost by implementing feature selection to obtain better accuracy. The classification results of the two algorithms will be compared to find out which algorithm has the best performance. The dataset used is from the SEER NCI program in November 2017 involving 4024 patients. The research shows that of the 16 attributes contained in the dataset, there are 3 attributes (features) that have a significant effect on the classification results, namely 6th stage, reginol node positive, and tumor size. XGBoost with feature selection has a better performance of 91.4% compared to SVM which is only 89.8%.

Keywords: breast cancer, machine learning, SVM, XGBoost.

How to Cite: Andriansyah, D., & Fridayanthie, E. W. (2023). Optimization of Support Vector Machine and XGBoost Methods Using Feature Selection to Improve Classification Performance. *JITE (Journal of Informatics and Telecommunication Engineering)*, 6(2), 484-493.

I. PENDAHULUAN

Dalam banyak penelitian telah disebutkan bahwa angka kematian telah meningkat pada wanita akibat kanker payudara (Abdulla et al., 2021). Kanker payudara adalah tumor ganas yang aktif di sel-sel payudara (Vaka et al., 2020). Penyakit ini mempengaruhi kedua jenis kelamin (Lukong, 2017), itu artinya baik pria maupun wanita dapat mengidap penyakit ini.

Menurut Kementerian Kesehatan Republik Indonesia, kanker payudara menempati urutan pertama dengan jumlah kanker terbanyak di Indonesia dan salah satu penyumbang kematian akibat kanker (Ministry of Health Republic of Indonesia, 2022). Data Globocan pada tahun 2020 menunjukkan jumlah kasus kanker baru di Indonesia mencapai 65.858 atau 16,6% dari total 396.914 kasus kanker baru, dengan jumlah kematian mencapai 22.430 kasus (Globocan, 2020).

Kanker payudara pada wanita usia subur terjadi pada usia 15-49 tahun, dan meningkat pada usia 35-54 tahun (Rahmayani et al., 2020). Deteksi dini sangat penting untuk kelangsungan hidup. Sekitar 70% kematian akibat kanker terjadi di negara berpenghasilan rendah dan menengah (Prastyo et al., 2020). Meskipun penyakit ini terjadi di seluruh dunia, insiden, mortalitas, dan tingkat kelangsungan hidup sangat bervariasi di berbagai belahan dunia, yang dapat disebabkan oleh banyak faktor seperti struktur populasi, gaya hidup, faktor genetik, dan lingkungan (Momenimovahed & Salehiniya, 2019).

Menurut sebuah laporan, sekitar 50% kanker payudara tidak terdeteksi pada pemeriksaan wanita dengan jaringan payudara yang sangat padat (Rabiei, 2022). Diagnosis penyakit pada stadium lanjut berkontribusi pada tingginya angka kematian pada wanita akibat kanker payudara (A. Gupta et al., 2015). Kanker payudara dapat dideteksi secara dini, memungkinkan pasien untuk mendapatkan terapi yang tepat dan dengan demikian meningkatkan peluang mereka untuk bertahan hidup (Arooj et al., 2022). Prediksi akurat dari perilaku kanker payudara sangat penting karena membantu dokter dalam proses pengambilan keputusan (Magboo & Magboo, 2021).

Machine learning merupakan kombinasi dari alat dan metode yang dapat digunakan untuk membuat algoritma yang dapat membantu dalam prediksi, klasifikasi, dan pengenalan pola (S. R. Gupta, 2022). Kesehatan adalah salah satu industri terbesar di dunia yang dapat memanfaatkan teknologi ini (Javaid et al., 2022). Ada banyak metode klasifikasi yang baik (Abdullah & Abdulazeez, 2021) yang dapat digunakan dalam penelitian antara lain algoritma Naive Baiyes, K-Nearest Neighbors, Decision Tree, dan Support Vector Machines (SVM). Namun pada penelitian ini, kami mencoba membandingkan algoritma SVM dan Extreme Gradient Boost (XGBoost) dengan seleksi fitur untuk mendapatkan nilai akurasi terbaik dari kedua metode tersebut.

II. STUDI PUSTAKA

Telah banyak penelitian yang menggunakan metode SVM, diantaranya adalah Comparative Analysis of SVM, XGBoost and Neural Network on Hate Speech Classification (Liang, 2021) dimana penelitian ini membahas tentang ujaran kebencian yang ditemukan di media sosial berupa teks, gambar, dan video. Akibatnya dapat memicu sejumlah pihak untuk melakukan hal-hal yang bertentangan dengan hukum dan merugikan pihak lain. Penelitian ini menganalisis tingkat akurasi, presisi, recall dan F1-Score dari 3 macam algoritma (SVM, XGBoost, dan Neural Network) dalam klasifikasi ujaran kebencian, dengan menggunakan dataset yang bersumber dari ujaran kebencian publik di Twitter dalam bahasa Indonesia. Hasil analisis menunjukkan bahwa algoritma SVM memiliki tingkat akurasi (83,2%), presisi (83%), recall (83%) dan F1-score (83%), SVM menempati level tertinggi dibandingkan XGBoost dan Jaringan Neural, sehingga algoritma SVM dapat dipertimbangkan untuk digunakan dalam klasifikasi ujaran kebencian.

Penelitian lainnya dengan judul Performance Evaluation of Machine Learning for Breast Cancer Diagnosis: A case study (Shanbehzadeh et al., 2022) menyatakan bahwa machine learning telah terbukti dapat mendiagnosis kanker payudara dengan cepat dan hemat biaya. Penelitian dilakukan dengan tujuan mengembangkan dan menguji model prediksi kanker payudara berdasarkan faktor gaya hidup wanita menggunakan beberapa pengklasifikasi machine learning dasar dan ensemble. Data diperoleh dari 1503 kasus yang dicurigai sebagai kanker payudara secara retrospektif diekstraksi dari basis data rumah sakit. Faktor risiko diidentifikasi menggunakan metode pembungkus-J48, pembungkus-SVM, pembungkus-NB, regresi logistik (LR), dan pemilihan fitur berbasis korelasi (CFS). Kemudian kinerja lima algoritma dasar ML, antara lain Naïve Bayes (NB), Bayesian network (BNeT), random forest (RF), multilayer perceptron (MLP), support vector machine (SVM), C4.5, eXtreme Gradient Boosting (XGBoost), pohon keputusan dan dua algoritma ensemble, termasuk Confidence weighted voting dan Voting dibandingkan untuk memprediksi BC sebelum dan sesudah melakukan feature section (FS). Hasil penelitian diperoleh Algoritma RF menunjukkan performa terbaik sebelum dan sesudah melakukan FS dengan AUC masing-masing sebesar 0,799 dan 0,798. Kombinasi model terbaik menggunakan metode Confidence weighted voting meningkatkan kinerja pengklasifikasi dan mencapai hasil terbaik dengan AUC 80%. Hasil penelitian

menunjukkan bahwa algoritma machine learning ensemble mewakili kemampuan yang lebih tinggi dari pada metode dasar. Model yang dikembangkan dapat secara akurat mengklasifikasikan individu yang berisiko tinggi untuk kanker payudara, dan dapat digunakan sebagai alat skrining untuk deteksi dini kanker payudara.

(Imaduddin & Hermansyah, 2021) menyebutkan bahwa salah satu kanker yang paling berbahaya didunia adalah kanker payudara yang umumnya banyak terjadi pada wanita, namun dalam beberapa kasus kanker ini dapat menyerang pria. Jenis kanker ini sangat berbahaya bagi manusia dan dapat menyebabkan kematian. Sehingga diperlukan pencegahan secara serius terhadap penyakit kanker ini. Salah satu pencegahan dapat dilakukan dengan pendeteksian secara dini. Tujuan dari penelitiannya adalah mengimplementasikan metode machine learning untuk mendeteksi kanker payudara pada wanita menggunakan algoritma Support Vector Machine (SVM) dan Decision Tree (DT). Setelah dilakukan klasifikasi terhadap data selanjutnya dilakukan perbandingan untuk mengetahui metode machine learning yang memiliki performa terbaik. Data sumber berasal dari Gynaecology Department of the University Hospital Centre of Coimbra (CHUC). Hasil dari penelitian ini menunjukan bahwa algoritma SVM dengan seleksi fitur memperoleh hasil klasifikasi terbaik dengan memperoleh akurasi sebesar 87,5%, sensitivitas 90%, dan spesitifitas 85%. Dengan demikian penelitian ini memperoleh hasil yang baik untuk dapat membantu memberikan solusi untuk mendeteksi penyakit kanker payudara.

A. Support Vector Machine (SVM)

Support Vector Machine (SVM) adalah metode pembelajaran mesin yang melakukan teknik untuk menemukan fungsi klasifikasi yang dapat membagi data menjadi dua kelas yang berbeda. Tujuan dari SVM adalah untuk mendapatkan *hyperplane* terbaik yang memisahkan dua kelas (Saikin et al., 2021) menjadi dua kelompok data dalam ruang dimensi yang lebih tinggi (Styawati & Mustofa, 2019). *Hyperplane* tersebut kemudian dapat digunakan untuk menentukan label yang paling mungkin untuk data yang tidak terlihat (Pisner & Schnyer, 2020). *Hyperplane* merupakan fungsi yang berperan sebagai batas dan membantu mengklasifikasikan titik data.

Algoritma SVM dapat digunakan untuk tugas analisis regresi, tetapi dalam praktiknya sering digunakan untuk aplikasi klasifikasi (B. M. Gupta et al., 2021). SVM telah digunakan sebagai alat yang ampuh untuk memecahkan masalah klasifikasi biner praktis sehingga telah terbukti bahwa SVM lebih unggul dari pada metode pembelajaran terawasi lainnya (Cervantes et al., 2020). Linear, Polynomial, *Radial Basis Function* (RBF), dan Sigmoid adalah kernel yang populer digunakan dalam klasifiksi SVM (Sarker, 2021). Pada dasarnya ada tiga tahap untuk analisis SVM: (1) seleksi fitur, (2) melatih dan menguji klasifikasi, dan (3) evaluasi kinerja (Pisner & Schnyer, 2020).

1. Seleksi Fitur, mencari korelasi yang berdampak tinggi pada hasil klasifikasi. Fitur yang tidak memiliki pengaruh signifikan pada hasil klasifikasi akan dibuang (Saikin et al., 2021). Sebagian besar metode seleksi fitur memberi peringkat terhadap fitur berdasarkan kriteria spesifik yang mencerminkan tingkat relevansinya.
2. Melatih dan menguji klasifikasi, SVM dilatih menggunakan observasi contoh di mana kita sudah mengetahui penetapan label dari contoh sebelumnya, konsekuensinya kita dapat mengawasi SVM untuk mengeksploitasi informasi utama untuk tujuan penetapan label baru (Pisner & Schnyer, 2020).
3. Evaluasi kinerja, kinerja klasifikasi dapat diukur menggunakan *Confusion Matrix*. *Confusion Matrix* merupakan alat yang digunakan untuk menganalisis seberapa baik *classifier* dapat mengenali *tuple* dari kelas yang berbeda dimana kelas yang diprediksi akan ditampilkan dibagian atas matrik dan kelas yang diobservasi akan ditampilkan dibagian kiri (Hutaminingsih, 2019). Pada dasarnya *Confusion Matrix* memberikan informasi perbandingan hasil klasifikasi yang dilakukan oleh model dengan hasil klasifikasi sebenarnya (Nugroho, 2020). Akurasi klasifikasi menunjukkan performa model klasifikasi secara keseluruhan, dimana semakin tinggi nilai akurasi klasifikasi yang dihasilkan maka semakin baik pula performa model klasifikasi.

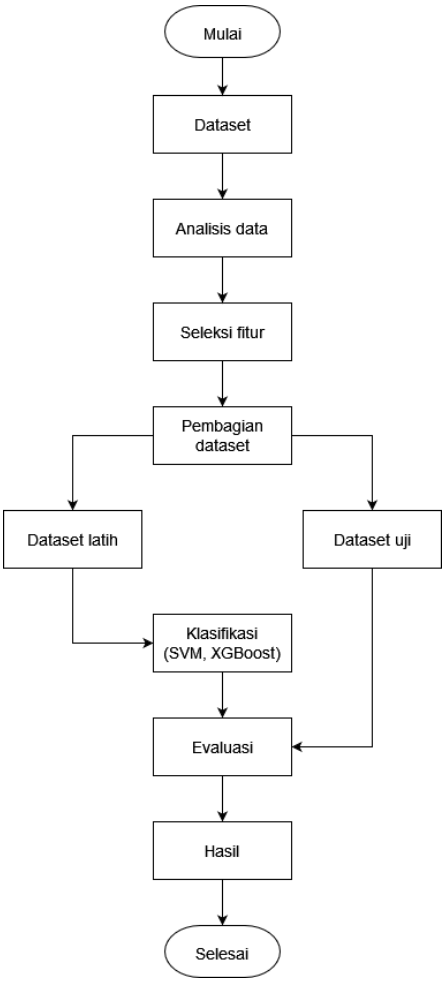
B. eXtreme Gradient Boost (XGBoost)

XGBoost merupakan pengembangan dari gradient boosting, algoritma yang dapat menemukan solusi optimal (Herni Yulianti et al., 2022) untuk mengatasi permasalahan regresi dan klasifikasi berdasarkan Gradient Bosting Decision Tree (Karo, 2020). Konsep dasar dari algoritma ini adalah menyesuaikan parameter pembelajaran secara berulang untuk menurunkan loss function (mekanisme

evaluasi atas model). XGBoost menggunakan model yang lebih teratur untuk membangun struktur pohon regresi, sehingga dapat memberikan kinerja yang lebih baik dan mampu mengurangi kompleksitas model untuk menghindari overfitting (Sunata, 2020). Boosting adalah algoritma pembelajaran ensemble yang memberikan bobot berbeda untuk distribusi data latih disetiap iterasi. Masing-masing iterasi peningkatan ditambahkan bobot untuk sample klasifikasi yang salah, dan mengurangi bobot untuk sample klasifikasi yang benar (Hanif, 2020).

III. METODE PENELITIAN

Penelitian menggunakan bahasa pemrograman Python dimana didalamnya telah terdapat library-library yang dibutuhkan untuk pemrosesan machine learning. Berikut adalah kerangka pemikiran penelitian:



Gambar 1. Alur Penelitian.

Alur penelitian dimulai dari pencarian dataset pada situs yang menyediakan data publik, kemudian melakukan analisis data terhadap dataset yang diperoleh untuk mengetahui kondisi data dilanjutkan dengan melakukan seleksi fitur untuk mengetahui fitur-fitur mana saja yang dapat mempengaruhi proses klasifikasi. Tahap selanjutnya adalah membagi dataset menjadi dua bagian yang berbeda yaitu data latih *(training data)* untuk proses pembuatan model dan data uji *(test data)* untuk proses evaluasi model. Tahap terakhir melakukan evaluasi menggunakan metode *Confusion Matrix* untuk mendapatkan tingkat akurasi, presisi, *recall*, dan *f1 score* yang diinginkan.

Dataset

Dataset yang digunakan dalam penelitian merupakan data publik kanker payudara yang diperoleh dari program SEER NCI November 2017 yang memberikan informasi tentang statistik kanker berbasis

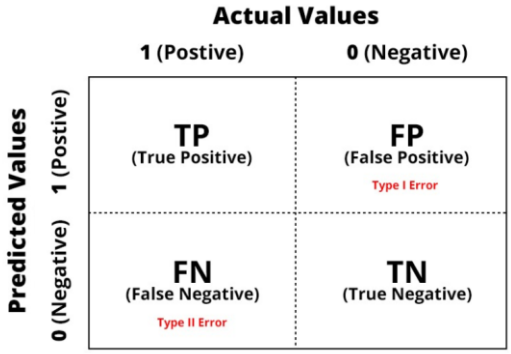
populasi yang dapat diakses secara gratis melalui situs Kaggle. Dataset melibatkan pasien wanita dengan duktus infiltrasi dan kanker payudara karsinoma lobular yang di diagnosa pada 2006-2010. Pasien dengan ukuran tumor tidak diketahui, LN regional yang diperiksa, LN regional positif, dan pasien dengan kelangsungan hidup kurang dari 1 bulan dikeluarkan, sehingga terdapat 4024 pasien yang disertakan. Data terdiri dari 16 atribut yaitu age, race, marital status, T stage, N stage, 6th stage, differentiate, grade, A stage, tumor size, estrogen status, progesterone status, regional node examined, reginol node positive, survival months, dan status.

Explaratory Data Analysis (EDA)

Explaratory Data Analysis merupakan proses kritis dalam melakukan investigasi awal pada data untuk menemukan pola, menemukan anomali, menguji hipotesis, dan untuk memeriksa asumsi dengan bantuan ringkasan statistik dan representasi grafis (Samosir et al., 2021). Dengan EDA, kondisi dataset dapat lebih mudah dipahami. Beberapa hal yang dapat dilakukan EDA antara lain adalah dapat mengetahui statistik dari dataset seperti standar deviasi, mean, min, dan max, dapat mengetahui ada atau tidaknya data yang hilang (missing value), serta dapat mengetahui ada atau tidaknya korelasi antar dua variabel.

Evaluasi

Evaluasi merupakan proses untuk mengukur kinerja model yang telah dibuat. Dalam penelitian ini menggunakan metode Confusion Matrix dimana metode tersebut menghasilkan nilai True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN).



Gambar 2. Confusion Matrix.

Selanjutnya adalah menghitung nilai akurasi, presisi, dan *recall* atau sensitifitas. Akurasi merupakan tingkat kedekatan niai prediksi dengan nilai *actual* (sebenarnya). Nilai akurasi dapat diperoleh dengan persamaan (1).

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Keterangan: TP = *True Positive*, TN = *True Negative*, FP = *False Positive*, FN = *False Negative*

Presisi menggambarkan tingkat keakuratan antara data yang diminta dengan hasil prediksi yang diberikan oleh model, dengan kata lain merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif. Nilai presisi dapat diperoleh dengan persamaan (2).

$$precision = \frac{TP}{TP + FP} \quad (2)$$

Keterangan: TP = *True Positive*, FP = *False Positive*

Recall atau sensitifitas menggambarkan keberhasilan model dalam menemukan kembali informasi, dengan kata lain *recall* merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif. Nilai *recall* dapat diperoleh dengan persamaan (3).

$$recall = \frac{TP}{TP + FN} \quad (3)$$

Keterangan: TP = *True Positive*, FN = *False Negative*

F1 score dapat diartikan sebagai rata-rata harmonik dari precision dan recall, dimana skor F1 mencapai nilai terbaiknya pada 1 dan skor terburuk pada 0. Kontribusi relatif dari precision dan recall pada skor F1 adalah sama. Skor F1 dapat diperoleh dengan persamaan (4).

$$F1\ score = 2 * \frac{Precision\ * Recall}{Precision + Recall} \quad (4)$$

Keterangan: Precision = nilai precision, Recall = nilai recall

IV. HASIL DAN PEMBAHASAN

Dataset yang telah diperoleh disimpan kedalam media penyimpanan Google Drive untuk selanjutnya diakses oleh Python menggunakan *library DataFrame*.

	Age	Race	Marital Status	T Stage	N Stage	6th Stage	differentiate	Grade	A Stage	Tumor Size	Estrogen Status	Progesterone Status	Regional Node Examined	Reginol Node Positive	Survival Months	Status
0	68	White	Married	T1	N1	IIA	Poorly differentiated	3	Regional	4	Positive	Positive	24	1	60	Alive
1	50	White	Married	T2	N2	IIIA	Moderately differentiated	2	Regional	35	Positive	Positive	14	5	62	Alive
2	58	White	Divorced	T3	N3	IIIC	Moderately differentiated	2	Regional	63	Positive	Positive	14	7	75	Alive
3	58	White	Married	T1	N1	IIA	Poorly differentiated	3	Regional	18	Positive	Positive	2	1	84	Alive
4	47	White	Married	T2	N1	IIB	Poorly differentiated	3	Regional	41	Positive	Positive	3	1	50	Alive

Gambar 3. Dataset Kanker Payudara 5 Teratas.

Sebelum masuk pada tahapan seleksi fitur, perlu dilakukan analisis data pada dataset kemudian dilanjutkan dengan menemukan fitur-fitur apa saja yang saling berkorelasi erat satu sama lain. Korelasi dilakukan menggunakan metode *Pearson Correlation* dengan memilih subset dari dataset yang hanya berisi fitur-fitur yang relevan saja. Hasilnya dapat dilihat pada gambar 4 dan 5.

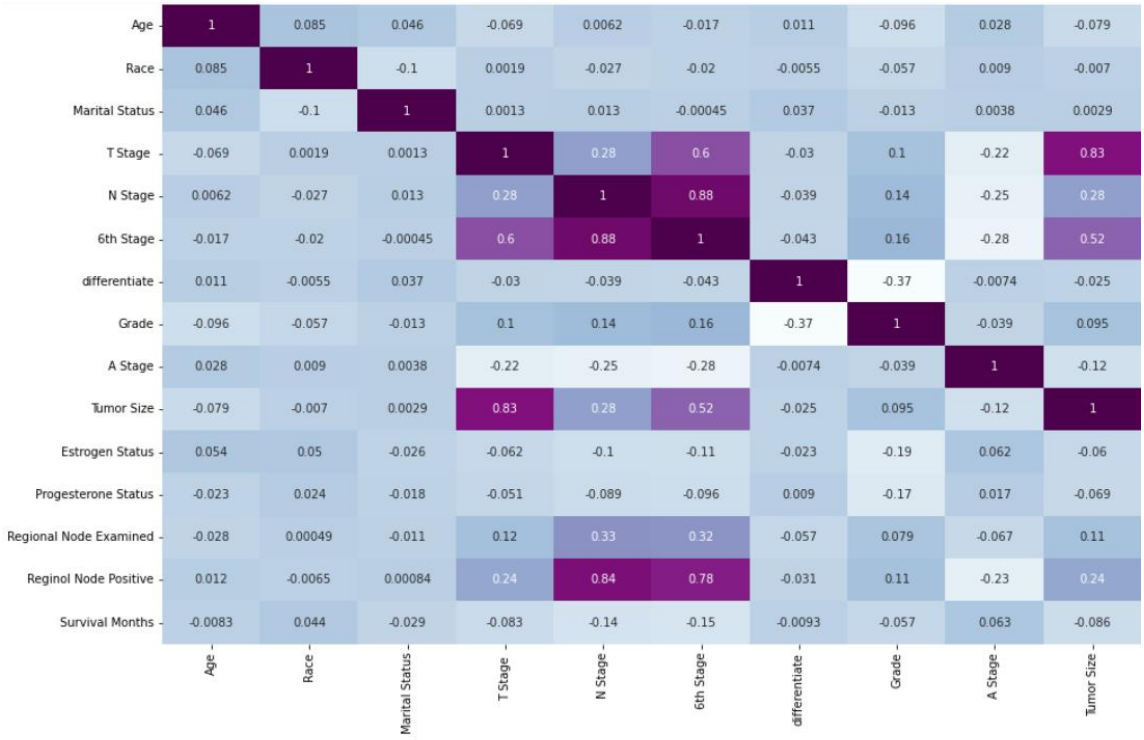
	Age	Tumor Size	Regional Node Examined	Reginol Node Positive	Survival Months
count	4024.000000	4024.000000	4024.000000	4024.000000	4024.000000
mean	53.972167	30.473658	14.357107	4.158052	71.297962
std	8.963134	21.119696	8.099675	5.109331	22.921430
min	30.000000	1.000000	1.000000	1.000000	1.000000
25%	47.000000	16.000000	9.000000	1.000000	56.000000
50%	54.000000	25.000000	14.000000	2.000000	73.000000
75%	61.000000	38.000000	19.000000	5.000000	90.000000
max	69.000000	140.000000	61.000000	46.000000	107.000000

```
# check column datatype

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4024 entries, 0 to 4023
Data columns (total 16 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Age                     4024 non-null   int64
 1   Race                    4024 non-null   object
 2   Marital Status          4024 non-null   object
 3   T Stage                 4024 non-null   object
 4   N Stage                 4024 non-null   object
 5   6th Stage               4024 non-null   object
 6   differentiate           4024 non-null   object
 7   Grade                   4024 non-null   object
 8   A Stage                 4024 non-null   object
 9   Tumor Size              4024 non-null   int64
 10  Estrogen Status         4024 non-null   object
 11  Progesterone Status     4024 non-null   object
 12  Regional Node Examined  4024 non-null   int64
 13  Reginol Node Positive   4024 non-null   int64
 14  Survival Months         4024 non-null   int64
 15  Status                  4024 non-null   object
dtypes: int64(5), object(11)
memory usage: 503.1+ KB
```

Gambar 4. Analisis data.



Gambar 5. Korelasi Fitur.

Dari hasil diatas, selanjutnya dilakukan seleksi fitur menggunakan data latih untuk mengetahui fitur-fitur apa saja yang memiliki pengaruh signifikan terhadap klasifikasi. Hasilnya, dari 16 fitur (atribut), yaitu age, race, marital status, T stage, N stage, 6th stage, differentiate, grade, A stage, tumor size, estrogen status, progesterone status, regional node examined, reginol node positive, survival months, dan status diperoleh 3 fitur yang berpengaruh secara signifikan yaitu 6th stage, reginol node positive, dan tumor size.

Tahap berikutnya adalah membagi dataset kedalam dua bagian yaitu data latih sebesar 80% dan data uji sebesar 20%, maka didapat data latih sebanyak 3.219 dan data uji sebanyak 805 dari total data keseluruhan sebanyak 4.024.

Dari hasil pembagian dataset pada tahap sebelumnya, maka akan dibuatkan model klasifikasi menggunakan algoritma SVM dengan kernel Radial Basis Function (RBF) dan XGBoost, dimana model akan menggunakan data uji dan seleksi fitur untuk memperoleh hasil akurasi yang terbaik.

Model klasifikasi SVM

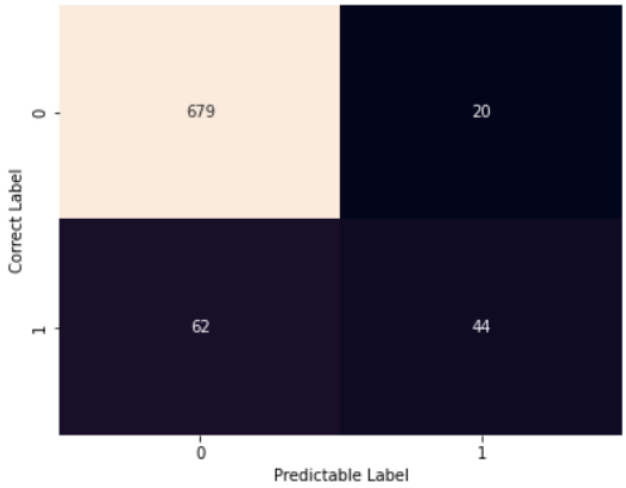
```
svcmodel = SVC(kernel = 'rbf', C = 10)
svcmodel.fit(X_train,y_train)
y_pred_scv = svcmodel.predict(X_test)
```

Model klasifikasi XGBoost

```
classifier = XGBClassifier()
classifier.fit(X_train, y_train)
y_pred_xgb = classifier.predict(X_test)
```

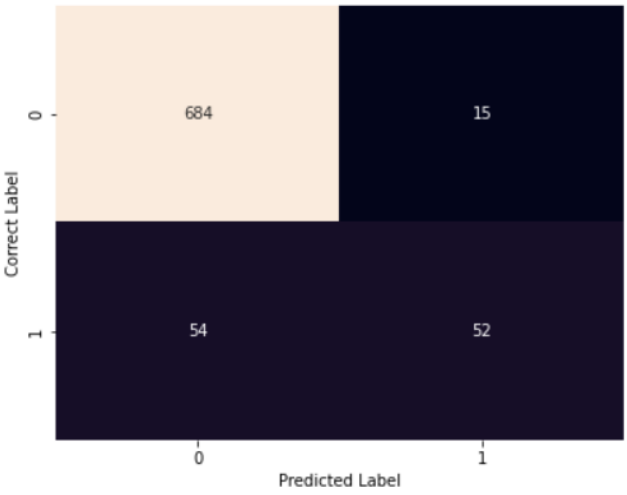
Evaluasi model klasifikasi dilakukan menggunakan metode *Confusion Matrix* untuk mengetahui kinerja model yang telah dibuat. Berikut hasil evaluasi yang diperoleh dari masing-masing algoritma.

```
accuracy_score:  0.8981366459627329
recall_score:  0.41509433962264153
f1_score:  0.5176470588235295
precision_score:  0.6875
```



Gambar 6. Hasil Evaluasi SVM.

```
accuracy_score:  0.9142857142857143
recall_score:  0.49056603773584906
f1_score:  0.6011560693641619
precision_score:  0.7761194029850746
```



Gambar 7. Hasil Evaluasi XGBoost.

Ringkasan hasil klasifikasi dapat dilihat pada tabel 2.

Tabel 2. Hasil Klasifikasi.

Algoritma	Akurasi	Presisi	Recall	F1 Score
SVM	89,8 %	68,7 %	41,5 %	51,7%
XGBoost	91,4 %	77,6 %	49,0 %	60,1%

V. SIMPULAN

Dari hasil penelitian yang diperoleh, dapat diketahui bahwa terdapat tiga fitur yang berpengaruh secara signifikan terhadap klasifikasi yaitu 6th stage, reginol node positive, dan tumor size. Hasil klasifikasi dengan seleksi fitur menunjukkan bahwa algoritma XGBoost memiliki performa yang lebih baik sebesar 91,4 % dibandingkan dengan algoritma SVM sebesar 89,8 % sehingga dapat digunakan untuk klasifikasi dan membantu dalam mendiagnosis penyakit kanker payudara.

Keterbatasan dalam penelitian ini adalah algoritma SVM hanya menggunakan satu jenis kernel yaitu RBF, tidak membandingkan dengan jenis kernel lainnya. Pada penelitian berikutnya perlu dilakukan perbandingan antar kernel sehingga dapat diketahui kernel mana yang memiliki tingkat akurasi terbaik dalam algoritma SVM. Hasil penelitian memiliki tingkat akurasi yang lebih tinggi dibandingkan dengan penelitian-penelitian sebelumnya dikarenakan adanya tahap seleksi fitur.

DAFTAR PUSTAKA

Abdulla, S. H., Sagheer, A. M., & Veisi, H. (2021). *Breast Cancer Classification Using Machine Learning Techniques: A Review*. 11.

Abdullah, D. M., & Abdulazeez, A. M. (2021). Machine Learning Applications based on SVM Classification A Review. *Qubahan Academic Journal*, *1*(2), Article 2. https://doi.org/10.48161/qaj.v1n2a50

Arooj, S., Atta-ur-Rahman, Zubair, M., Khan, M. F., Alissa, K., Khan, M. A., & Mosavi, A. (2022). Breast Cancer Detection and Classification Empowered With Transfer Learning. *Frontiers in Public Health*, *10*, 924432. https://doi.org/10.3389/fpubh.2022.924432

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189–215. https://doi.org/10.1016/j.neucom.2019.10.118

Globocan. (2020). *360-indonesia-fact-sheets*. The Global Cancer Observatory. https://gco.iarc.fr/today/data/factsheets/populations/360-indonesia-fact-sheets.pdf

Gupta, A., Shridhar, K., & Dhillon, P. K. (2015). A review of breast cancer awareness among women in India: Cancer literate or awareness deficit? *European Journal of Cancer*, *51*(14), 2058–2066. https://doi.org/10.1016/j.ejca.2015.07.008

Gupta, B. M., Dhawan, S. M., & Mamdapur, G. M. (2021). *INDIA: A SCIENTOMETRIC EVALUATION OF INDIA'S*. *57*, 14.

Gupta, S. R. (2022). Prediction time of breast cancer tumor recurrence using Machine Learning. *Cancer Treatment and Research Communications*, *32*, 100602. https://doi.org/10.1016/j.ctarc.2022.100602

Hanif, I. (2020). Implementing Extreme Gradient Boosting (XGBoost) Classifier to Improve Customer Churn Prediction. *Proceedings of the Proceedings of the 1st International Conference on Statistics and Analytics, ICSA 2019, 2-3 August 2019, Bogor, Indonesia*. Proceedings of the 1st International Conference on Statistics and Analytics, ICSA 2019, 2-3 August 2019, Bogor, Indonesia, Bogor, Indonesia. https://doi.org/10.4108/eai.2-8-2019.2290338

Herni Yulianti, S. E., Oni Soesanto, & Yuana Sukmawaty. (2022). Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit. *Journal of Mathematics Theory and Application*, *4*(1), 21–26. https://doi.org/10.31605/jomta.v4i1.1792

Hutaminingsih, N. W. (2019, July 9). Comparing Testing and Training Data SVM (Support Vector Machine) with R. *Medium*. https://medium.com/@nabilawrhtm/comparing-testing-and-training-data-svm-support-vector-machine-with-r-d69929a00708

Imaduddin, H., & Hermansyah, B. A. (2021). *ARISON OF SUPPORT VECTOR MACHINE AND DECISION TREE METHODS IN THE CLASSIFICATION OF BREAST CANCER*. *5*(1), 9.

Javaid, M., Haleem, A., Pratap Singh, R., Suman, R., & Rab, S. (2022). Significance of machine learning in healthcare: Features, pillars and applications. *International Journal of Intelligent Networks*, *3*, 58–73. https://doi.org/10.1016/j.ijin.2022.05.002

Karo, I. M. K. (2020). Implementasi Metode XGBoost dan Feature Importance untuk Klasifikasi pada Kebakaran Hutan dan Lahan. *Journal of Software Engineering*, *1*(1), 7.

Liang, S. (2021). Comparative Analysis of SVM, XGBoost and Neural Network on Hate Speech Classification. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *5*(5), 896–903. https://doi.org/10.29207/resti.v5i5.3506

Lukong, K. E. (2017). Understanding breast cancer – The long and winding road. *BBA Clinical*, *7*, 64–77. https://doi.org/10.1016/j.bbacli.2017.01.001

Magboo, V. P. C., & Magboo, Ma. S. A. (2021). Machine Learning Classifiers on Breast Cancer Recurrences. *Procedia Computer Science*, *192*, 2742–2752. https://doi.org/10.1016/j.procs.2021.09.044

Ministry of Health Republic of Indonesia. (2022, February 4). *Ministry of Health Republic of Indonesia*. Kanker Payudara Paling Banyak Di Indonesia, Kemenkes Targetkan Pemerataan Layanan Kesehatan. https://www.kemkes.go.id/article/view/22020400002/kanker-payudara-paling-banyak-di-indonesia-kemenkes-targetkan-pemerataan-layanan-kesehatan.html

Momenimovahed, Z., & Salehiniya, H. (2019). Epidemiological characteristics of and risk factors for breast cancer in the world. *Breast Cancer: Targets and Therapy*, *Volume 11*, 151–164. https://doi.org/10.2147/BCTT.S176070

Nugroho, K. S. (2020, June 4). Confusion Matrix untuk Evaluasi Model pada Supervised Learning. *Medium*. https://ksnugroho.medium.com/confusion-matrix-untuk-evaluasi-model-pada-unsupervised-machine-learning-bc4b1ae9ae3f

Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In *Machine Learning* (pp. 101–121). Elsevier. https://doi.org/10.1016/B978-0-12-815739-8.00006-7

Prastyo, P. H., Paramartha, I. G. Y., Pakpahan, M. S. M., & Ardiyanto, I. (2020). Predicting Breast Cancer: A Comparative Analysis of Machine Learning Algorithms. *Proceeding International Conference on Science and Engineering*, *3*, 455–459. https://doi.org/10.14421/icse.v3.545

Rabiei, R. (2022). Prediction of Breast Cancer using Machine Learning Approaches. *Journal of Biomedical Physics and Engineering*, *12*(3). https://doi.org/10.31661/jbpe.v0i0.2109-1403

Rahmayani, O. S., Permana, R. H., & Witdiawati, W. (2020). Early Detection of Breast Cancer According to Fertile Age Women. *Media Keperawatan Indonesia*, *3*(1), 32. https://doi.org/10.26714/mki.3.1.2020.32-37

Saikin, S., Fadli, S., & Ashari, M. (2021). Optimization of Support Vector Machine Method Using Feature Selection to Improve Classification Results. *JISA(Jurnal Informatika Dan Sains)*, *4*(1), Article 1. https://doi.org/10.31326/jisa.v4i1.881

Samosir, F. V. P., Mustamu, L. P., Anggara, E. D., Wiyogo, A. I., & Widjaja, A. (2021). Exploratory Data Analysis terhadap Kepadatan Penumpang Kereta Rel Listrik. *Jurnal Teknik Informatika Dan Sistem Informasi*, *7*(2), Article 2. https://doi.org/10.28932/jutisi.v7i2.3700

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Shanbehzadeh, M., Kazemi-Arpanahi, H., Bolbolian Ghalibaf, M., & Orooji, A. (2022). Performance evaluation of machine learning for breast cancer diagnosis: A case study. *Informatics in Medicine Unlocked*, *31*, 101009. https://doi.org/10.1016/j.imu.2022.101009

Styawati, S., & Mustofa, K. (2019). A Support Vector Machine-Firefly Algorithm for Movie Opinion Data Classification. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *13*(3), Article 3. https://doi.org/10.22146/ijccs.41302

Sunata, H. (2020). Komparasi Tujuh Algoritma Identifikasi Fraud ATM Pada PT. Bank Central Asia Tbk. *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, *7*(3), 441–450. https://doi.org/10.35957/jatisi.v7i3.471

Vaka, A. R., Soni, B., & K., S. R. (2020). Breast cancer detection by leveraging Machine Learning. *ICT Express*, *6*(4), 320–324. https://doi.org/10.1016/j.icte.2020.04.009