APLIKASI MACHINE LEARNING DENGAN PYTHON UNTUK DETEKSI KANKER PAYUDARA MENGGUNAKAN MODEL REGRESI LOGISTIK

Ai Ilah Warnilah[1 *], Herlan Sutisna[2] , Ratningsih[3] , Vincent Christian[4] , Ranti Maharani[5]

[1,2,3,4,5] Sistem Informasi, Universitas Bina Sarana Informatika Indonesia

** Corresponding Author.E-mail:ai.aiw@bsi.ac.id*

Abstract

One of the deadliest diseases and everyone in the world must be afraid of Cancer. Breast cancer is one of the most common types of cancer in the world and is a leading cause of death among women. Early detection of breast cancer is very important to increase the chances of recovery. In recent years, the implementation of machine learning in the medical field has grown rapidly, especially in the prediction and diagnosis of diseases. This study aims to discuss the implementation of the logistic regression algorithm in breast cancer prediction. Through the use of relevant datasets, this algorithm is expected to provide accurate predictions and assist in clinical decision making. The data used comes from Kaggle.com with the file name Breast Cancer Predictor. The process carried out is Business Understanding, Data Understanding, Data Preparation and Data Visualization for modeling using Logistic Regression, after modeling the data the next step is to use the application using the Streamlit Framework. In the data training process using the Logistic Regression method, an accuracy result of 0.97% was obtained.

Keywords:
Breast Cancer, Framework, Machine Learning, Logistic Regression.

Abstrak

Salah satu penyakit paling mematikan dan semua orang di dunia pasti takut dengan penyakit Kanker. Kanker payudara merupakan salah satu jenis kanker yang paling umum di dunia dan menjadi penyebab utama kematian di kalangan wanita. Deteksi dini kanker payudara sangat penting untuk meningkatkan peluang kesembuhan. Dalam beberapa tahun terakhir, implementasi machine learning dalam bidang medis telah berkembang pesat, khususnya dalam prediksi dan diagnosis penyakit. Penelitian ini bertujuan untuk membahas implementasi algoritma regresi logistik dalam prediksi kanker payudara. Melalui penggunaan dataset yang relevan, algoritma ini diharapkan dapat memberikan prediksi yang akurat dan membantu dalam pengambilan keputusan klinis. Data yang digunakan bersumber dari Kaggle.com dengan nama file *Breast Cancer Predictor*. Proses yang dilakukan adalah Pemahaman Bisnis, Pemahaman Data, Persiapan Data dan Visualisasi Data untuk pemodelan menggunakan Regresi Logistik, setelah memodelkan data langkah selanjutnya adalah menggunakan penerapan menggunakan Framework Streamlit. Pada proses pelatihan data menggunakan metode Regresi Logistik diperoleh hasil akurasi sebesar 0,97%.

Kata Kunci:
Kanker Payudara, Framework, Machine Learning, Regresi Logistik

1. Pendahuluan

Salah satu masalah kesehatan publik yang paling signifikan di seluruh dunia adalah kanker payudara, yang merupakan salah satu penyebab kematian utama wanita. Deteksi dini sangat penting untuk meningkatkan kesembuhan dan perawatan yang lebih efektif. Penyakit Kanker Payudara merupakan kanker yang terbentuk di jaringan Payudara(Taye, 2023) .Kanker payudara ini terjadi ketika sel-sel pada jaringan di payudara yang tumbuh secara tidak terkendali dan kanker ini akan mengambil alih jaringan payudara di sekitarnya. Kanker Payudara dapat terbentuk di kelenjar yang menghasilkan susu *(lobulus)* atau di saluran *(ductus)* yang membawa air susu dari kelenjar ke puting payudara, Kanker ini juga dapat membentuk jaringan lemak didalam payudara. Ada beberapa jenis kanker payudara seperti *Ductal Carcinoma In Situ, Lobular Carnicoma In Situ, Invasive Ductal Carcinoma dan Invasie Lobukar Ca2rcinoma*. Faktor yang menjadi pemicu Kanker Payudara dapat dipicu sebagai berikut kebiasaan merokok atau minum minuman beralkohol (Starek-Świechowicz et al., 2023), kelebihan berat badan atau obsesitas, baru mulai mestruasi sebelum usia 12 tahun. Sekitar 2,3 juta perempuan terdiagnosis penyakit kanker payudara (Arnold et al., 2022). Upaya screening rutin dengan pemeriksaan fisik dengan metode konvensional masih sangat terbatas sehingga mengalami kesulitan untuk menganalisis kanker payudara tersebut. Tujuan dari penelitian ini untuk menggabungkan keunggulan *machine learning* dengan data klinis untuk mendiagnosis kanker payudara untuk mengembangkan model prediktif yang dapat lebih akurat menemukan kasus dengan risiko tinggi yang memungkinkan intervensi dan perawatan yang lebih baik.

Pada penelitian sebelumnya algoritma Naive bayes sangat efektif dalam klasifikasi kanker payudaratingkat efektifitasnya dengan rata-rata nilai precision dan recall sekitar 0.96. Nilai precision dan recall paling tinggi adalah sekitar 0.96 (Oktavianto & Handri, 2019). Selanjutnya penelitian Pengambilan keputusan dalam klasifikasi jenis kanker payudara dapat dipercepat dengan menggunakan data mining yang menggunakan metode seperti logistic regresion, decision tree, naïve bayes, dan k-nearest neighbor 95,00 persen(Ranti et al., 2022). Pada penelitian Pendekatan Machine Learning yang Efisien untuk Prediksi Kanker Payudara C4.5—ZS—GA untuk dataset BCD memiliki akurasi 77,27%, lebih baik 3,49% dari C4.5 sebagai yang terbaik dari ML normal, dan lebih baik 5,63% dari Weighted Vote (Decision Tree, Memory Based Learner, NB, SVM)—Fisher Score sebagai yang terbaik dari penelitian yang terkait (Dirjen et al., 2017). Selanjutnya

dalam dalam penelitian ini, algoritma pengklasifikasi tertentu digunakan untuk menilai keakuratan strategi klasifikasi yang berbeda. Mengembangkan pengklasifikasi yang tepat dan efektif untuk tujuan medis adalah salah satu tantangan terbesar bagi peneliti data mining dan *machine learning*. Kinerja SVM lebih baik daripada pengklasifikasi lain(Solikin et al., n.d.). Berdasarkan penelitian sebelumnya bahwa dalam menentukan akurasi klasifikasi jenis kanker payudara termasuk sudah baik dalam penelitian ini penulis akan menggunakan model yang dapat memprediksi hasil kanker apakah kanker tersebut ganas atau tidak dengan menggunakan bahasa pemrograman Python. Python terpakai karena memiliki banyak modul atau library yang dapat dipakai seperti Plotly (W. Zhang, n.d.)untuk visualisasi data yang menarik dengan plotly juga user atau pengguna dapat melihat akurasi data dengan mengarahkan kursor ke data yang dibuat. Lalu di penelitian ini akan menggunakan machine learning yaitu Sckit-Learn dengan menggunakan Algoritma Regresi Logistic. Sesudah tahap model menggunakan machine learning lanjut ke tahap deployment, di dalam penelitian ini akan digunakan framework streamlit (Panda et al., 2022)untuk deploy data dan logika algoritma didalam nya sehingga hasil akurasi dapat terlihat di website.

2. Bahan Dan Metode

2.1. Deksripsi Sumber Data

Pada penelitian ini, pengambilan *sample* dataset yaitu menggunakan dataset *public* yang dapat dicari di situs Kaggle. Dataset yang digunakan dalam penelitian ini merupakan Kumpulan data pasien kanker payudara yang diperoleh dari pembaruan November 2017 dari Program SEER NCI, yang menyediakan informasi tentang statistic kanker berbasis populasi. Kumpulan data tersebut melibatkan pasien Perempuan dengan kanker payudara karsinoma ductus infiltrasi dan lobular, yang didiagnosis pada tahun 2006-2010.

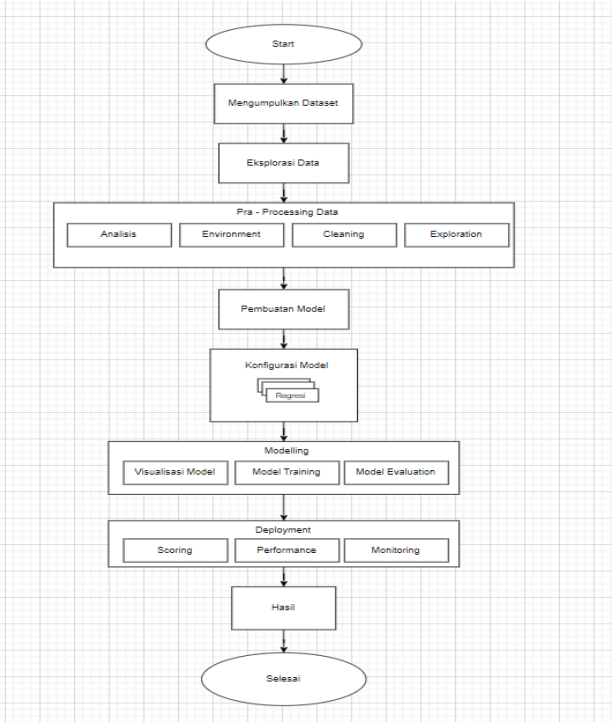
2.2.Teknik Pengumpulan Data

Penelitian ini termasuk penelitian kuantitatif. Penelitian kuantitatif adalah metode penelitian ilmiah yang menggunakan angka dan statistic untuk mengumpulkan dan menganalisis data. Dalam penelitian ini mempunyai beberapa cakupan masalah numerik oleh karena itu metode kuantitatif adalah solusi yang tepat untuk memecahkan masalah ini

2.3.Diagram Alir Penelitian

Dalam metode penelitian ini terdiri dari beberapa tahapan yaitu tahap pertama melakukan research atau mengumpulkan data, kemudian data tersebut d*ieksplorasi* selanjutnya akan dilakukan *preprocessing* seperti *analisis, environment, cleaning, exploration*. Tahap kedua, melakukan

pembuatan model dengan menggunakan algoritma Regresi. Tahap ketiga, melakukan deployment dengan data atau model yang sudah diolah disertai visualisasi grafik dan akurasi penyakit kanker payudara



Gambar 1. Diagram Alir Penelitian

Berikut penjelasan metodologi penelitian "Aplikasi Machine Learning dengan Python untuk Deteksi Kanker Payudara Menggunakan Metode Regresi" :

1. Tahap pertama melupakan proses mengumpulkan dataset yang akan digunakan untuk membuat model (Gong et al., 2023)dikumpulkan pada tahap pertama. Sumber data bisa berupa data internal perusahaan, data yang dikumpulkan dari aplikasi, survei/wawancara atau melalui dataset public seperti di dalam situs Kaggle.com,openML dan sumber data lainnya
2. Tahap kedua *eksplora*si Data
   Pada tahap ini, data yang telah dikumpulkan dieksplorasi untuk memahami distribusi,pola dan anomali dalam data. Biasanya data ini di analisis dan di visualisasikan yang digunakan untuk memahami karakteristik dari dataset(Y. Zhang et al., 2022)
3. Tahap ketiga merupakan *Pra-Processing* Data, Data akan disiapkan untuk dimodelkan, tahapan ini terdiri dari beberapa sub-langkah seperti analisis yang (Cai et al., 2021) digunakan untuk menemukan keterkaitan antar variabel, kolerasi dan faktor-faktor lain yang dapat mempengaruhi model, environment (Canali & Leonelli, 2022) digunakan untuk pengaturan lingkungan kerja sekitar, termasuk alat-alat dan teknologi yang akan digunakan untuk permodelan, cleaning digunakan untuk membersihkan data dari missing values, outliers dan data inkonsistensi lainnya, data yang kotor dapat mempengaruhi hasil dari permodelan menjadi tidak akurat, tahap terakhir exploration (Cai et al., 2021)yang digunakan untuk menemukan pola pola yang tidak terlihat pada analisi awal exploration dapat menggunakan beberapa teknik seperti PCA (Principal Component Analysis), clustering dan lainnya.

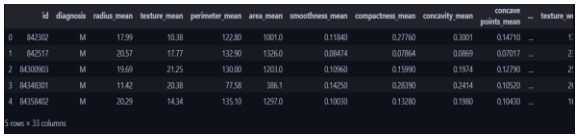
4. Tahap selanjutnya adalah Pembuatan Model. Pada tahap ini, model yang sesuai ddengan masalah atau studi kasus yang ingin dianalisis dan dibuatkan modelnya. Pilihan model tergantung pada jenis masalah yang ada di dalam dataset
5. Pada tahap Konfigurasi Model akan dilakukan Model yang telah dibuat kemudian dikonfigurasi dengan parameter parameter yang sesuai. Misalnya, jika model menggunakan regresi, maka parameter seperti learning rate, jumlah iterasi dan jenis regularisasi akan ditentukan. Pemilihan parameter dapat meningkatkan performa model
6. Tahap Modelling akan dilakukan Proses permodelan terdapat tiga sub-langkah utama yaitu visualisasi model yang digunakan untuk memberikan hasil dari model yang divisualisasikan untuk memberikan gambaran tentang bagaimana model bekerja seperti plotting kurva ROC, confusion matrix atau distribusi prediksi, model training yang digunakan untuk melatih model menggunakan dataset yang telah diproses, data training digunakan untuk mengajarkan model agar mampu membuat prediksi yang akurat, Langkah terakhir yaitu model evaluation yang digunakan untuk di evaluasi kinerja model yang telah dilatih dengan menggunakan data testing atau cross-validation.
7. Tahap Deployment akan digunakan Setelah model dievaluasi, model akan siap untuk di *deploy* atau digunakan dalam lingkungan produksi. Terdapat tiga aspek Utama dalam tahap deployment pertama scoring (Jammal et al., 2021),model digunakan untuk membuat prediksi pada data baru yang dating secara real-time atau secara langsung, kedua performance , kinerja model dipantau secara terus menerus untuk memastikan bahwa model tetap bekerja dengan baik. kadang kadang model perlu disesuaikan atau retrain jika ada perubahan signifikan pada data, terakhir ialah monitoring digunakan untuk proses untuk mendeteksi masalah seperti drift pada data atau penurunan performa model dari Waktu ke Waktu
8. Tahap terakhir ialah hasil yang betujuan untuk pengguna ketika mencoba aplikasi pendeteksi kanker payudara, akan ada beberapa opsi yang dapat dipilih contohnya gejala payudara nya (Shah et al., 2021)apakah sangat sakit atau tidak jika pengguna memilih sangat sakit maka akan terdeteksi langsung bahwa penderita terkena kanker ganas dengan akurasi 2.66%

3. Hasil Dan Pembahasan

3.1 Dataset

Dalam penelitian ini menggunakan public dataset, pengambilan sample dataset terdapat di website Kaggle yang berisi pasien kanker payudara yang diperoleh dari pembaruan November 2017 dari Program SEER NCI, yang menyediakan informasi tentang statistic kanker berbasis populasi. Kumpulan data tersebut melibatkan pasien Perempuan dengan kanker payudara karsinoma ductus infiltrasi dan lobular, yang didiagnosis pada

tahun 2006-2010. Dataset yang diperoleh terdapat 33 atribut yaitu : id, diagnosis, radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concave points_mean, symmetry_mean, fractal_dimension_mean, radius_se, texture_se, perimeter_se, area_se, smoothness_se, compactness_se, concavity_se, concave points_se, symmetry_se, fracntal_dimension_se, radius_worst, texture_worst, perimeter_worst, area_worst, smoothness_worst, compactness_worst, concavity_worst, concave points_worst, symmetry_worst, fractal_dimension_worst. Dalam pengaplikasikan dataset menggunakan python, dataset yang akan digunakan dilakukan proses import terlebih dahulu. Gambar 2. Merupakan contoh dataset yang digunakan



	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	texture_w
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...	1
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...	2
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...	2
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...	2
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...	1

5 rows × 33 columns

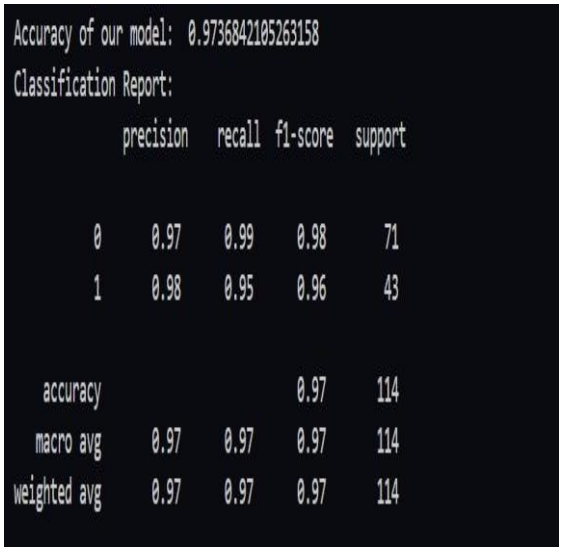
Gambar 2. Contoh Dataset

3.2. Implementasi Model

Regresi adalah metode statistik yang digunakan untuk memodelkan hubungan antara variabel depende (yang ingin diprediksi) dan satu atau lebih variabel independent (predictor). Dalam konteks prediksi kanker payudara, variabel dependen biasanya berupa diagnosis kanker (misalnya, kanker atau bukan kanker), sementara variabel independent bisa berupa berbagai fitur medis seperti ukuran tumor, tekstur, kepadatan sel, dan lain-lain.

Di dalam penelitian ini akan menggunakan regresi logistic karena regresi ini dapat digunakan untuk prediksi kanker payudara. Metode ini digunakan untuk memodelkan probabilitas terjadinya suatu kejadian dengan variabel dependen biner (misalnya, 0 untuk bukan kanker dan 1 untuk kanker baik, 2 untuk kanker yang ganas). Dengan menggunakan fungsi sigmoid, regresi logistic dapat menghasilkan output dalam bentuk probabilitas yang kemudia diklasifikasikan sebagai kanker atau bukan kanker berdasarkan ambang batas tertentu (biasanya 0,5).

Pada penelitian ini, implementasi model Regresi logistic terdiri dari presicion, recall, f1-score, support. Implementasi model ditunjukkan pada gambar 3



Accuracy of our model: 0.9736842105263158
Classification Report:

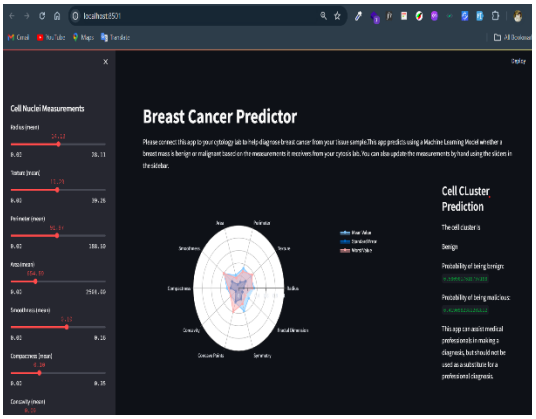
	precision	recall	f1-score	support
0	0.97	0.99	0.98	71
1	0.98	0.95	0.96	43
accuracy			0.97	114
macro avg	0.97	0.97	0.97	114
weighted avg	0.97	0.97	0.97	114

Gambar 3. Implementasi Model

3.3.Deployment

Setelah melakukan Implementasi Model maka di dalam penelitian ini akan dilakukan deployment untuk mengembangkan model yang

sudah diolah, model ini dapat digunakan untuk memprediksi pasien kanker payudara dengan berbagai macam opsi sebagai contoh dapat dilihat pada gambar 4.



Gambar 4. Deployment

4.4. Hasil Eksperimen

Setelah Implementasi model dan deployment, pada bagian ini akan menjelaskan hasil dari model dan deploy yang sudah dibuat dan diolah, akurasi model dengan mempeloreh angka 0.9736 atau sekitar 97.37%. ini berarti model berhasil memprediksi dengan benar sekitar 97 dari 100 kasus. Akurasi tinggi menunjukkan bahwa model sangat baik dalam mengklasifikasikan apakah seseorang memiliki kanker payudara atau tidak berdasarkan data uji. Presicion untuk kelas 0 sekitar 0.97 atau 97% dan Presicion untuk kelas 1 sekitar 0.98 atau 97%. Presicion mengukur berapa banyak dari prediksi positif yang benar benar positif. Precision yang tinggi (hampir mendekati 1) menunjukkan bahwa model sangat jarang memberikan prediksi positif palsu. Recall untuk kelas 0 berkisar 0.99 atau 99& dan Recall untuk kelas 1 berkisar 0.95 atau 95%. Recall mengukur berapa banyak dari kasus yang sebenarnya positif yang berhasil diprediksi dengan benar, Recall yang tinggi menunjukkan model sangat efektif dalam menangkap semua kasus positif. F1-Score untuk kelas 0 berkisar 0.98 atau 98% dan F1-Score untuk kelas 1 berkisar 0.96 atau 96%. F1-Score adalah rata-rata harmonic dari precision dan recall. Skor ini memberikan Gambaran keseimbangan antara presicion dan recall. F1-Score yang tinggi di kedua kelas menunjukkan bahwa model ini seimbang dalam memprediksi kedua kelas tanpa bias. Terakhir support untuk kelas 0 berkisar 71 yang berarti jumlah sampel dengan label 0 dan support untuk kelas 1 berkisar 43 yang berarti jumlah sampel dengan label 1. Support menunjukkan jumlah kasus actual dari masing masing kelas yang digunakan untuk evaluasi. Di data ini menunjukkan bahwa data pengujian terdiri dari 114 sampel, dengan 71 di antaranya berada dalam kelas 0 dan 43 dalam kelas 1

4. Kesimpulan

Penelitian ini menunjukkan bahwa algoritma regresi logistik merupakan alat yang efektif dalam prediksi kanker payudara. Meskipun hasil yang diperoleh sangat bergantung pada kualitas dan jumlah data, regresi logistik menawarkan pendekatan yang sederhana namun kuat untuk aplikasi medis. Dengan peningkatan lebih lanjut dalam preprocessing data dan penambahan fitur yang relevan, akurasi model dapat terus ditingkatkan.. Untuk penelitian selanjutnya, disarankan melakukan lebih banyak pengolahan data medis yang baru

sehingga mudah untuk di implementasikan untuk bidang medis .

Ucapan Terima Kasih

Kami mengucapkan terima kasih yang sebesar-besarnya kepada Yayasan BSI atas kepercayaan dan dukungan yang telah diberikan kepada tim dosen dan mahasiswa kami dalam pelaksanaan Penelitian Dana Yayasan ini. Tanpa kepercayaan dari Yayasan BSI, penelitian ini tidak akan dapat berjalan dengan baik dan mencapai hasil yang diharapkan

References

[1] Arnold, M., Morgan, E., Rumgay, H., Mafra, A., Singh, D., Laversanne, M., Vignat, J., Gralow, J. R., Cardoso, F., Siesling, S., & Soerjomataram, I. (2022). Current and future burden of breast cancer: Global statistics for 2020 and 2040. *Breast*, *66*, 15–23. https://doi.org/10.1016/j.breast.2022.08.010

[2] Cai, Q., Cui, C., Xiong, Y., Wang, W., Xie, Z., & Zhang, M. (2021). *A Survey on Deep Reinforcement Learning for Data Processing and Analytics*. http://arxiv.org/abs/2108.04526

[3] Canali, S., & Leonelli, S. (2022). Reframing the environment in data-intensive health sciences. *Studies in History and Philosophy of Science*, *93*, 203–214. https://doi.org/10.1016/j.shpsa.2022.04.006

[4] Dirjen, S. K., Riset, P., Pengembangan, D., Dikti, R., Azis, A. I. S., Surya, I., Idris, K., Santoso, B., Mustofa, Y. A., & Informatika, J. T. (2017). Terakreditasi SINTA Peringkat 2 Pendekatan Machine Learning yang Efisien untuk Prediksi Kanker Payudara. *Masa Berlaku Mulai*, *1*(3), 458–469.

[5] Gong, Y., Liu, G., Xue, Y., Li, R., & Meng, L. (2023). A survey on dataset quality in machine learning. *Information and Software Technology*, *162*. https://doi.org/10.1016/j.infsof.2023.107268

[6] Jammal, M., Kanso, A., Heidari, P., & Shami, A. (2021). Evaluating High Availability-Aware Deployments Using Stochastic Petri Net Model and Cloud Scoring Selection Tool. *IEEE Transactions on Services Computing*, *14*(1), 141–154. https://doi.org/10.1109/TSC.2017.2781730

[7] Oktavianto, H., & Handri, R. P. (2019). Analisis Klasifikasi Kanker Payudara Menggunakan Algoritma Naive Bayes. In *Informatics Journal* (Vol. 4, Issue 3). https://archive.ics.uci.edu/ml/.

[8] Panda, N. R., Pati, J. K., Mohanty, J. N., & Bhuyan, R. (2022). A Review on Logistic Regression in Medical Research. In *National Journal of Community Medicine* (Vol. 13, Issue 4, pp. 265–270). MedSci Publications. https://doi.org/10.55489/njcm.13420222

[9] Ranti, N., 1∗, M., & Hanif, K. H. (2022). *Klasifikasi Penyakit Kanker Payudara Menggunakan Perbandingan*

Algoritma Machine Learning. *3*(1), 1–6. http://creativecommons.org/licences/by/4.0/

[10] Shah, S. M., Khan, R. A., Arif, S., & Sajid, U. (2021). *Artificial Intelligence For Breast Cancer Detection: Trends & Directions*. https://doi.org/10.1016/j.compbiomed.2022.105221

[11] Solikin, I., Bhumi, R. P., & Power, J. (n.d.). *Teknik Data Mining untuk Prediksi Kanker Payudara yang Efisien*.

[12]Starek-Świechowicz, B., Budziszewska, B., & Starek, A. (2023). Alcohol and breast cancer. In *Pharmacological Reports* (Vol. 75, Issue 1, pp. 69–84). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s43440-022-00426-4

[13] Taye, M. M. (2023). Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions. In *Computers* (Vol. 12, Issue 5). MDPI. https://doi.org/10.3390/computers12050091

[14] Zhang, W. (n.d.). *Interactive Data Visualization with Python Plotly*.

[15] Zhang, Y., Sheng, M., Liu, X., Wang, R., Lin, W., Ren, P., Wang, X., Zhao, E., & Song, W. (2022). A heterogeneous multi-modal medical data fusion framework supporting hybrid data exploration. *Health Information Science and Systems*, *10*(1). https://doi.org/10.1007/s13755-022-00183-x