ANALISIS PERFORMA *LOGISTIC REGRESSION, NAÏVE BAYES,* DAN *RANDOM FOREST* SEBAGAI ALGORITMA PENDETEKSI KANKER PAYUDARA

Cecep Wahyu Cahyana[1], Akhsin Nurlayli[2,*]

[1,2]Departemen Pendidikan Teknik Elektronika dan Informatika, Fakultas Teknik, Universitas Negeri Yogyakarta, Jl. Colombo No.1 Karangmalang Yogyakarta 55281 INDONESIA

Abstrak

Kanker payudara merupakan jenis penyakit kronis yang sampai saat ini masih diragukan terkait upaya penyembuhan total penyakit ini, selain itu juga memerlukan waktu pengobatan yang lama dan juga biaya yang cukup tinggi. Faktor penyebab dari kanker payudara sendiri hingga kini masih belum diketahui secara spesifik, namun dapat dicermati bahwa penyebab penyakit ini bersifat multifaktorial yang saling mempengaruhi satu dengan lainnya, seperti: faktor lingkungan, genetika, virus, pola makanan, dan juga radiasi di daerah dada. Tujuan penelitian ini adalah untuk mengetahui metode mana yang memiliki akurasi tertinggi dalam prediksi kanker payudara di Coimbra, dengan metode *Logistics Regression, Naïve Bayes*, atau *Random Forest*. Penelitian ini diharapkan mampu membantu masyarakat dan tenaga medis dalam deteksi dini penyakit kanker payudara. Berdasarkan pengujian yang dilakukan menggunakan algoritma *Logistics Regression* didapatkan nilai akurasi sebesar 80%, pada algoritma *Naïve Bayes* mendapatkan nilai sebesar 75%, dan terakhir dengan algoritma *Random Forest* didapatkan nilai sebesar 75%. Dari pengujian tersebut dapat disimpulkan bahwa algoritma *Logistics Regression* terbukti memiliki tingkat akurasi yang paling baik dalam hal prediksi penyakit kanker payudara dibandingkan dengan kedua algoritma lainnya.

Kata Kunci:
breast cancer, logistics regression, naïve bayes, random forest

Abstract

Breast cancer is a chronic disease that is still in doubt regarding efforts to cure this disease completely. Besides that, it also requires a long treatment time and is quite expensive. The causative factors of breast cancer itself are still not specifically known. However, it can be observed that the causes of this disease are multifactorial and influence one another, such as environmental factors, genetics, viruses, diet, and radiation to the chest area. This study aimed to find out which method has the highest accuracy in predicting breast cancer in Coimbra, using Logistics Regression, Naïve Bayes, and Random Forest methods. This research was expected to help the community and medical personnel in the early detection of breast cancer. Based on the tests, the Logistics Regression algorithm had an accuracy value of 80%, the Naïve Bayes algorithm at 75%, and the Random Forest algorithm received a value of 75%. The Logistics Regression algorithm has the best breast cancer prediction accuracy from the testing processes compared to the other two algorithms.

Keywords:
breast cancer, logistics regression, naïve bayes, random forest

1. PENDAHULUAN

Menurut data yang dihimpun dari World Health Organization (WHO) tahun 2020, terdapat total kurang lebih 2,3 juta orang penderita kanker payudara dengan total pasien meninggal sekitar 685.000. Hingga akhir tahun 2020, ada 7,8 juta wanita hidup yang didiagnosis menderita kanker payudara dalam 5 tahun terakhir, menjadikannya kanker paling umum di dunia. Angka kematian akibat kanker payudara sedikit berubah dari tahun 1930-an hingga tahun 1970-an. Kenaikan persentasi kelangsungan hidup

* Korespondensi
E-mail: cecepwahyu.2019@student.uny.ac.id, akhsinnurlayli@uny.ac.id

penderita kanker payudara dimulai pada 1980-an di negara-negara dengan program deteksi dini yang dikombinasikan dengan berbagai cara pengobatan untuk membasmi penyakit invasif (WHO, 2021). Kanker payudara menjadi urutan pertama penyebab kematian terkait penderita kanker di Indoesia. Data Globocan pada tahun 2020 menunjukkan jumlah kasus baru pada kanker payudara di Indonesia sebanyak 68.858 dari total keseluruhan yang mencapai 396.914, serta jumlah kematian akibat kanker payudara mencapai lebih dari 22.000 jiwa. Menurut data Kementerian Kesehatan Republik Indonesia, 43% kematian dapat dicegah apabila pasien melakukan deteksi dini secara rutin dan menghindari factor penyebab kanker. Banyaknya kasus di Indonesia terjadi karena deteksi penyakit kanker payudara dilakukan tidak sejak dini (Kementerian Kesehatan Republik Indonesia, 2022).

American Cancer Society merekomendasikan beberapa metode dalam pendeteksian dini penyakit kanker, salah satunya adalah penerapan teknologi Informasi dalam pendeteksian dini penyakit kanker (The American Cancer Society medical and editorial content team, 2022). Jenis aplikasi pendeteksi dini akan memberi informasi yang memiliki sifat eksplanatoris terkait cara untuk pendeteksian kanker sejak dini. Selain itu, Teknologi Informasi dalam pendeteksian dini juga dirupakan dalam teknologi penggalian data yang bertujuan untuk mempersingkat waktu dan faktor pendeteksian dini dari penyakit kanker payudara. Data-data terkait faktor diagnosis keganasan kanker payudara sendiri dapat secara luas diakses oleh masyarakat (tidak berbayar) melalui situs resmi *UCI Machine Learning*. Denganbegitu kesempatan untuk melakukan riset akurasi dan juga deteksi dini penyakit kanker dapat dilakukan secara luas (Goel, 2018).

Selanjutnya, untuk dapat mendukung keputusan apakah seseorang tersebut dapat divonis sebagai penderita kanker payudara atau tidak, maka bisa dilakukan dengan merancang sebuah model prediksi dari beberapa parameter/variabel seperti dalam perancangan model prediksi penyakit kanker payudara tersebut dapat dilakukan dengan menggunakan metode statistik tradisional seperti *Logistic Regression*. Akan tetapi seiring dengan perkembangan teknologi saat ini, maka dapat dilakukan juga dengan menggunakan algoritma pengklasifikasian pada data mining. Data Mining sendirimerupakan sebuah proses dalam menemukan pola dan informasi menarik dalam data yang ada, dengan menggunakan teknik atau metode tertentu (Mulyo et al., 2013). Terdapat beberapa jenis metode pengklasifikasian dalam data mining yang digunakan untuk deteksi penyakit, seperti N*aïve Bayes, Decision Tree, K-Nearest Neighborhood*, *Support Vector Machine* (SVM), dan lainnya (Ajeng Wijayanti et al., 2017; Bustami, 2013; Dinh et al., 2016; Goldenberg & Punthakee, 2013; Kumar et al., 2022; Network, 2022; Permanasari et al., 2017; Sandeep & Bethel, 2021; Shah et al., 2016).

Proses diagnosa penyakit kanker payudara sendiri secara medis masih cukup sulit dilakukan saat ini. Data dari medis yang kurang relevan, dan juga adanya redundant data membuat pengaruh yang signifikan dalam proses diagnosis penyakit kanker payudara. Upaya untuk memaksimalkan proses diagnosis kanker payudara salah satunya dengan menggunakan teknik klasifikasi data mining computer menggunakan data medis yang sudah ada, untuk dapat menggali lebih dalam informasi dataset penyakit kanker payudara, dan memungkinkan untuk menghasilkan diagnosis penyakit Kanker Payudara yang akurat di masa depan.

Penelitian yang telah dilakukan sebelumnya mengenai pendeteksian dini Kanker Payudara sendiri telah banyak dilakukan dengan berbagai algoritma klasifikasi, namun belum ada penelitian yang meneliti secara ekplisit pendeteksian Kanker Payudara dengan menggunakan dataset penyakit kanker payudara di Kota Coimbra dan menggunakan algoritma *Logistics Regression*, Naïve Bayes, dan juga Random Forest. Penelitian ini memiliki tujuan berupa mempertajam hasil analisis sebelum proses diagnosa pasien dengan melakukan komparasi berbagai macam algoritma klasifikasi yang ada tersebut. Ketajaman yang dimaksud sebelumnya yaitu terkait pemilihan dengan seksama manakah variabel yang memiliki dampak yang relevan terkait pendeteksian dini dari penyakit Kanker Payudara. Dengan komparasi berbagai macam algoritma ini, diharapkan akan memiliki dampak terhadap akurasi dari masing-masing variabel, dan dapat menentukan algoritma mana yang paling sesuai untuk pendeteksian dini Kanker Payudara.

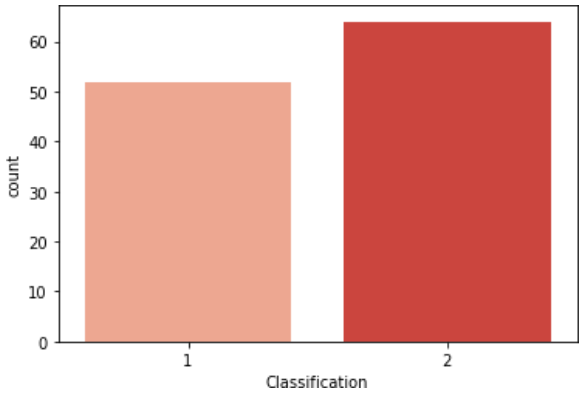
Pada penelitian ini menerapkan algoritma *Logistics Regression*, Naïve Bayes, dan juga Random Forest. Dari ketiga algoritma tersebut dilakukan perbandingan terkait performansinya untuk melihat algoritma mana yang memiliki akurasi paling tinggi dan tepat dalam hal klasifikasi terhadap diagnosa seseorang mengidap kanker payudara atau tidak. Tidak hanya itu, *output* dari penelitian ini juga berupa model prediksi apakah seseorang dengan data tertentu terklasifikasi dalam penderita Kanker Payudara atau tidak menderita kanker payudara. Data yang digunakan dalam penelitian ini adalah data para pasien terduga penderita kanker payudara di Kota Coimbra Portugal yang dihimpun oleh Yassir Hussein Shakir, dan bersumber dari Kaggle.com. Dengan demikian, dari penelitian ini dapat diketahui jenis metode yang memiliki performansi tertinggi dalam pengujian menggunakan tersebut yaitu, *Logistics Regression*, Naïve Bayes, dan Random Forest dalam hal pengklasifikasian dataset penyakit kanker payudara di Kota Coimbra.

2. METODE

Pada bagian metodologi ini akan dijelaskan terkait dataset yang digunakan dalam pembuatan model prediksi, desain model *workflow*, dan juga metode yang digunakan untuk prediksi. Tidak hanya itu, dalam bagian ini juga akan dijelaskan metode validasi dari keempat algoritma yang digunakan dalam prediksi.

A. Breast Cancer in Coimbra Dataset

Dataset yang digunakan dalam penelitian pengembanganmodel prediksi kali ini adalah Breast Cancer Coimbra Datasetyang didapatkan dari Kaggle.com dan dihimpun oleh . Datasettersebut diambil dari pasien wanita dengan rentang umur antara 24 hingga 89 tahun di Kota Coimbra, Portugal.



Gambar 1. Pasien observasi (1) dan pasien positif Kanker Payudara (2)

Dataset tersebut memiliki jumlah total 116 baris data dengan total 10 kolom, data ini terdiri dari 9 atribut data yang masing-masing mengindikasikan kondisi dari para pasien. Pada Gambar 1 di atas menunjukkan sebaran data dari para penderita kanker payudara terhadap total keseluruhan data yang ditandai dengan keterangan nilai 1 sebagai pasien yang sedang diobservasi kesehatannya, dan nilai 2 sebagai pasien yang terkonfirmasi mengidap kanker payudara. Dari data tersebut terdapat sekitar merupakan 52 pasien observasi, dan sebanyak 64 pasien sisanya adalah pasien positif mengidap kanker payudara. Tabel 1 menunjukkan deskripsi atribut pada dataset yang digunakan.

Tabel 1. Deskripsi Atribut

Atribut	Deskripsi
Age	Usia pasien (tahun)
BMI	Indeks massa tubuh dari pasien (berat kg/(tinggi m)^2)
Glucose	Kadar glukosa plasma 2 jam dalam tes toleransi glukosa oral
Insulin	Kadar insulin serum 2 jam (mu U/ml)
HOMA	Kadar derajat disfungsi sel B pankreas (20 x insulin puasa (μ IU/mL))
Leptin	Hormon dengan massa 16 kDa yang berperan dalam regulasi berat tubuh, fungsi metabolisme, dan reproduksi
Adiponectin	Protein spesifik yang disekresikan oleh sel adiposit dan bersifat anti- inflamasi
Resistin	Resistin merupakan *Adipose Tissue Specific Secretory Factor* (ADSF), sebuah hormon yang disekresi oleh jaringan adiposa yang menginduksi resistensi insulin dalam hati dan otot. Kadar Resistin ini diperkirakan dapat menggambarkan indikator resistensi insulin dan juga inflamasi
MCP.1	Kemokin utama yang mengatur migrasi dan infiltrasi makrofag/monosit

Masing-masing atribut dalam dataset kanker payudara yang telah diuraikan dalam Tabel 1 tersebut memiliki berbagaimacam tipe data dan nilai *min-max* pada masing-masingkolomnya, untuk detailnya akan diuraikan pada Tabel 2.

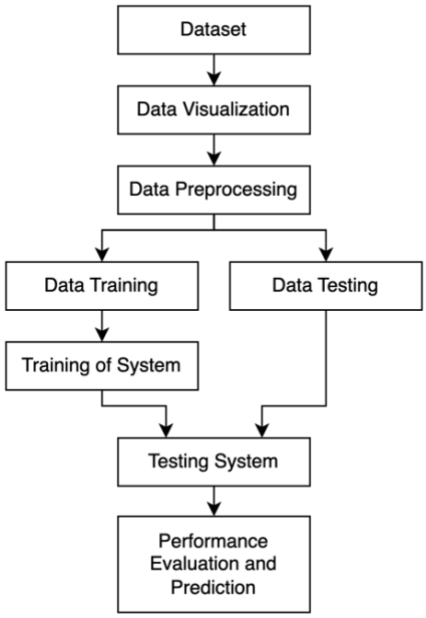
Tabel 2. Atribut Dataset

Atribut	Tipe Data	Rentang Nilai
Age	Int	24 - 89
BMI	Float	18,37 - 38,57
Glucose	Int	60 - 201

Atribut	Tipe Data	Rentang Nilai
Insulin	Float	2,43 - 58,46
HOMA	Float	0,46 - 25,05
Leptin	Float	4,31 - 90,28
Adiponectin	Float	1,65 - 38,04
Resistin	Float	3,21 - 82,1
MCP.1	Float	45,84 - 1698,44

B. Workflow

Gambar 2 adalah rancangan dari alur pengembangan sistem prediksi yang melalui berbagai macam tahap guna mencapai hasil prediksi yang paling akurat dari beberapa metode yang ada. Kemudian dilakukan tahap berupa *data preprocessing* yang berisi berupa kegiatan *Missing Value Analysis* dan juga *Outliers*, lalu langkah selanjutnya ialah data dibagi menjadi dua bagian yaitu data *training* dengan persentase 75%, dan data *testing* berjumlah 25%.



Gambar 2. *Workflow Model*

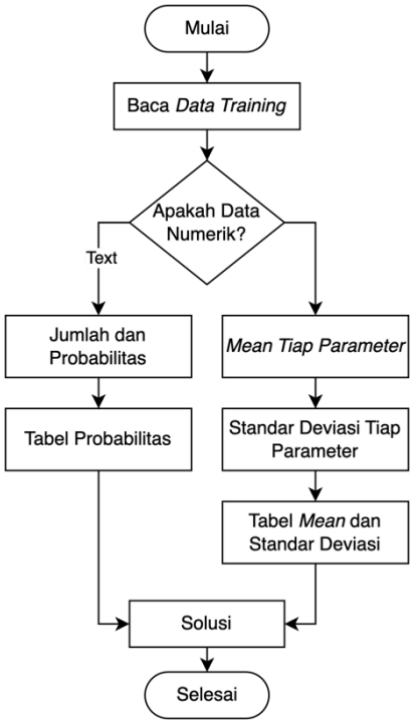
Tahap selanjutnya yaitu proses tahap percobaan/*testing* dengan menggunakan metode/algoritma yang ditentukan, Kemudian setelah proses tersebut dilakukan maka akan terbentuk sebuah model yang kedepannya dapat digunakan untuk proses prediksi dini penyakit kanker payudara. Pada penelitian ini, model *Logistic Regression*, Naïve Bayes, dan Random Forest akan dibahas pada bagian berikutnya. Setelah model sudah berhasil terbentuk, proses selanjutnya ialah tahapan *Performance Evaluation and Prediction*. Pada tahap akhir akan dilakukan uji sistem dengan menggunakan *Confusion Matrix* dan juga *Classification Matrix* untuk kemudian dilanjutkan pada tahap uji prediksi pada dataset baru yang betujuan untuk melihat angka akurasi model yang telah dibuat sebelumnya.

C. Logistic Regression

Algoritma *Logistic Regression* merupakan salah satu algoritma teknik analisis data dalam Statistika yang dirancang untuk dapat mengetahui hubungan keterkaitan antar variabel, dimana dalam variabel respon tersebut bersifat *categorical*, baik itu nominal atau ordinal, dan juga variabel penjelas yang memiliki sifat kategoris atau kontinu (Ramli et al., 2013). Metode *logistic regression* ini sendiri memiliki teknik dan prosedur yang hampir serupa dengan metode linear regression. Dimana metode *Ordinary Least Square* (OLS) sering digunakan untuk proses estimasi nilai parameter oleh linear, sementara itu untuk dapat memperkirakan estimasi nilai parameter dari *linear regression* menggunakan metode maximum *likelihood estimation* atau MLE.

D. Naïve Bayes

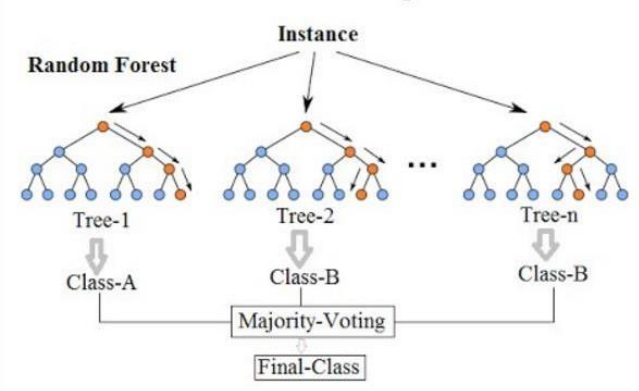
Algoritma Naïve Bayes adalah salah satu metode klasifikasi dalam data mining yang menggunakan statistik danprobabilitas dalam penentuan hasilnya. Metode ini awalnya dikenalkan oleh ilmuwan Inggris yaitu Thomas Bayes. Algoritma ini diberi nama Naïve dikarenakan dalam algoritmaini akan mengasumsikan data yang ada dengan kondisi antaratribut yang saling bebas (Alita et al., 2019)(Harimurti, 2017). Dalam Naïve Bayes, klasifikasi yang adadiasumsikan bahwa ada atau tidaknya ciri khas tertentu dalamsebuah kelas tidak berhubungan dengan ciri khas dalam kelas lain (Harimurti, 2017).



Gambar 3. Alur Metode *Naïve Bayes*

E. Random Forest

Random Forest adalah salah satu metode yang umum digunakan untuk melakukan regresi dan klasifikasi data. Metode *Random Forest* ini sendiri adalah sebuah kumpulan (*ensemble*) dari metode pembelajaran yang mengimplementasikan *decision tree* untuk *base classifier* yang dibangun dan juga dikombinasikan (Kulkarni & Sinha, 2014). Terdapat tiga aspek penting yang ada dalam metode Random Forest, seperti: (1) Melakukan *bootstrap sampling* dalam membangun *prediction tree*; (2) Masing-masing dari *prediction tree* tadi akan melakukan prediksi dengan prediktorrandom/acak; (3) Kemudian Random Forest akan membuat prediksi dengan menggunakan kombinasi hasil dari masing- masing *prediction tree* dengan cara *majority vote* untuk mengklasifikasikan atau rata-rata untuk regresi (Sadewo et al., 2016).



Gambar 4. Alur Metode Random Forest (Sadewo et al., 2016)

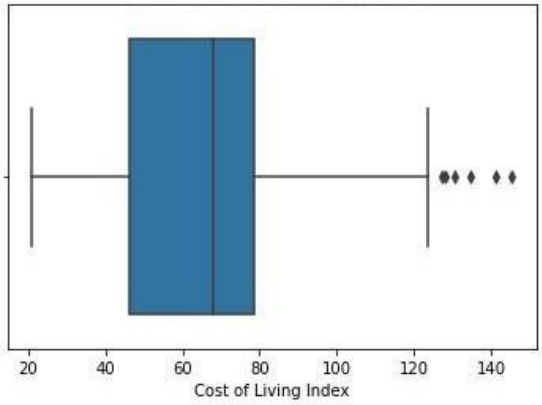
F. Preprocessing

1) *Missing Value Analysis*

Missing value merupakan nilai yang hilang di dalam sebuah dataset. Proses pengolahan data akan mengalami kesulitan apabila dalam dataset tersebutterdapat missing value (Acuna & Rodriguez, n.d.), hal tersebut bisa terjadi karena terdapat prasangka dan penurunan kualitas dari algoritma karena terdapat nilai yang tidak lengkap (Qin et al., 2007). Oleh karena itu diperlukan sebuah treatment khusus untuk dapat menemukan missing value tersebut, untuk kemudian dapat diganti atau bisa juga dengandihapus dari dataset, sehingga proses pengolahan data dalam penelitian dapat menghasilkan output classification yang lebih akurat.

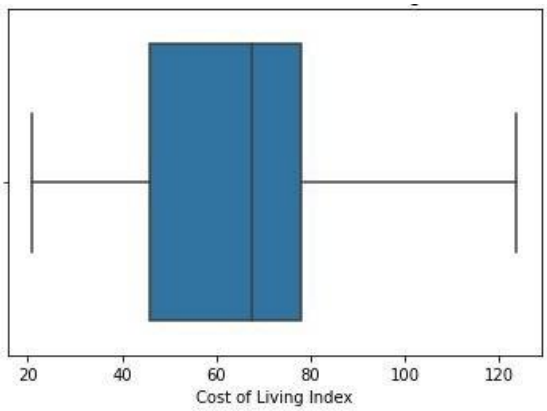
2) *Outlier*

Outlier merupakan data pengamatan dalam dataset yang memiliki value tidak konsisten atau sangat berbeda dari data lainnya. Munculnya data yang berbeda atau outlier ini tentunya akan mempengaruhi proses analisis data, dimana output classification dari pemrosesan data akan menjadi bias dan tidak akurat. Maka perlu dilakukanperbaikan dari dataset yang memiliki outlier tersebut sebelum dilakukan analisis lanjutan.



Gambar 5. Data dengan *Outlier*

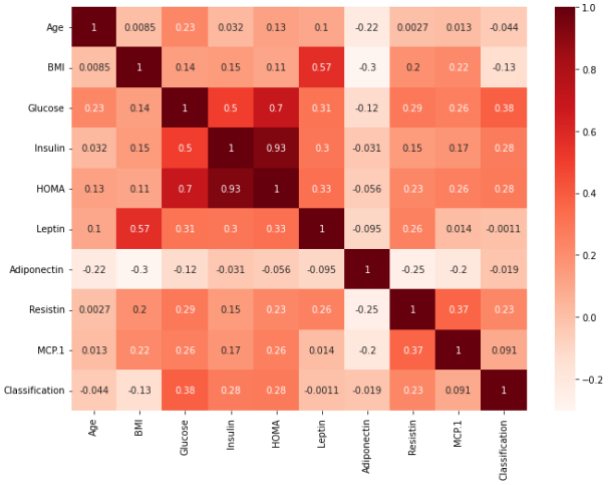
Dapat dilihat dalam Gambar 5 di atas, dari total 116 baris data yang ada dalam dataset, terlihat dalam *visual* data di atas terdapat data yang *outlier* dari data lainnya, dengan adanya *outlier* tersebut maka *output classification* yang akan dihasilkan tidak akan sempurna karena akan mengurangi akurasi prediksi, oleh karena itu diperlukan pemrosesan berupa perbaikan data hingga data *outlier* tersebut akan terlihat seperti fenomena sebenarnya, hasil perbaikan data *outlier* dapat dilihat pada Gambar 6 berikut dengan sisa sebanyak 80 baris data yang siap digunakan.



Gambar 6. Data Tanpa *Outlier* (Setelah Dibersihkan)

G. Correlation

Korelasi antar variabel dapat dilihat pada Gambar 7 di bawah ini.

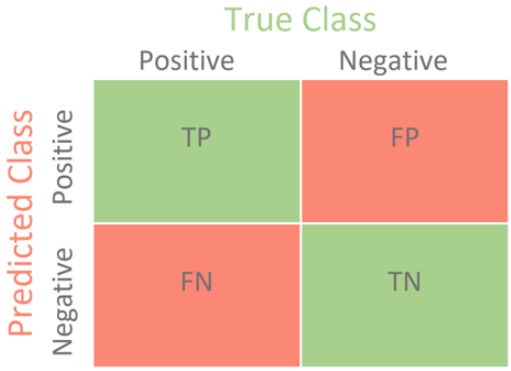


Gambar 7. *Data Correlation*

Data correlation seperti yang ditunjukkan pada Gambar 7di atas adalah data hubungan antar dua variabel yang paling berkorelasi dengan *Classification*. Data dalam *matrix* tersebut diperoleh setelah melewati tahap berupa *Missing Value Analysis* dan juga tahap perbaikan data *Outlier*. Dari *matrix* data di atas didapatkan informasi bahwa data yang paling berkorelasi dengan *Classification* adalah data Insulin dengan data HOMA dengan value 0,93. Sementara itu data dengan korelasi *Classification* terendah adalah korelasi antara *Resistin* dengan *Adiponectin* dengan *value* korelasi sebesar -0,25.

H. Confusion Matrix

Confusion Matrix merupakan sebuah Tabel yang mendeskripsikan klasifikasi jumlah data uji yang benar dan juga jumlah data uji yang salah. Berikut ini adalah contoh *confusion matrix* untuk klasifikasi biner yang ditunjukkan pada Gambar 8.



Gambar 8. *Confusion Matrix*

Keterangan:
TP (*True Positive*) : jumlah dokumen dari kelas *positive* yang benar diklasifikasikan dengan kelas *positive.*
TN (*True Negative*) : jumlah dokumen dari kelas *negative* yang benar diklasifikasikan dengan kelas *negative.*
FP (*False Positive*) : jumlah dokumen dari kelas *negative* yang salah diklasifikasikan dengan kelas *positive*.
FN (*False Negative*) : jumlah dokumen dari kelas *positive* yang salah diklasifikasikan dengan kelas *negative*.
Cara memperoleh hasil dari *confusion matrix* sendiri digunakan untuk menghitung akurasi (*accuracy*), presisi (*precision*), dan *recall.*

I. Classification Report

Dalam *library python Scikit-learn* yang digunakan, *classification report* sendiri memiliki berbagai *performance metrics*, detail dari masing-masing *performance metrics* dapat dilihat berikut:

1) Recall

Recall merupakan total nilai *positive* yang benar dandibandingkan dengan nilai total prediksi yang bernilai *positive*. Rumus *recall* sebagai berikut:

$$Recall = \frac{TP}{TP+FN} \quad (1)$$

2) Precision

Precision merupakan jumlah total dari prediksi positive yang benar dan dibandingkan dengan total keseluruhan prediksi positive yang diberikan oleh modelprediksi. Rumus precision sebagai berikut:

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

3) F1-Score

F1-Score merupakan nilai average/harmonic mean dari recall dan precision. Untuk rumus dari nilai F1- Score sebagai berikut:

$$F1 - Score = \frac{2\ x\ Precision\ x\ Recall}{Precision+Recall} \quad (3)$$

4) Accuracy

Accuracy merupakan jumlah prediksi yang tepat, yaitu nilai dengan value positive yang benar dan nilai prediksi dengan value negative yang benar dibandingkan dengan jumlah total keseluruhan dari prediksi. Rumus dari accuracy sebagai berikut:

$$Accuracy = \frac{TP + TN}{Total} \quad (4)$$

5) Support

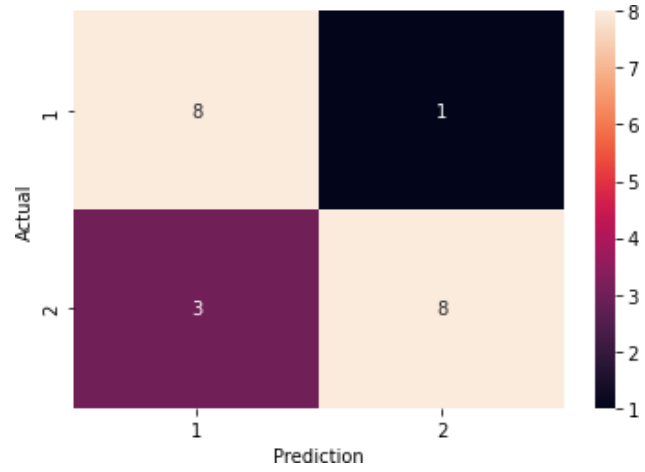
Support merupakan nilai yang dihitung secara horizontal dari Tabel confusion matrix. Support ini akan merepresentasikan jumlah nilai kelas *positive* maupun ***negative*** yang sebenarnya.

3. HASIL DAN PEMBAHASAN

A. Logistic Regression

Tahap pengujian yang dilakukan dengan menggunakan metode Logistic Regression pada Google Collaboratory. Hasil dari proses pengujian dengan menggunakan model ini untuk memprediksi penyakit kanker payudara untuk menentukan akurasi yang didapatkan mendapatkan hasil yang cukup baik yaitu dengan nilai akurasi data test sebesar 70% dan akurasi data training sebesar 76%.

Kemudian setelah dilakukan pengujian data model, selanjutnya dilaksanakan validasi pada model prediction. Gambar 9 merupakan diagram confusion matrix dari data model yang telah diuji terhadap data test. Dari confusion matrix tersebut digunakan sebanyak 20 data yang merupakan 25% dari total 80 baris data. Dapat terlihat terdapat sejumlah 10 data dengan class negative dan 10 data dengan class positive. Sistem dengan method ini dapat memprediksi masing-masing 9 data negativedan 8 data positive dengan benar.



Gambar 9. Confusion Matrix Metode Logistic Regression

Setelah dilakukan pengujian dengan menggunakan confusion matrix di atas, kemudian dilaksanakan classification report. Dari perhitungan yang telah dilakukan terhadap data test tersebut, dapat ditemukan bahwa sistem telah berjalan dengan baik ketika menentukan hasil nilainegative (1) namun masih kurang baik dalam menentukan nilai positive (2).

Tabel 3. Classification Report Metode Logistic Regression

	Precision	Recall	F1-score	Support
1	0,89	0,73	0,80	11
2	0,73	0,89	0,80	9
Accuracy			0,80	20
Macro avg	0,81	0,81	0,80	20
Weighted avg	0,82	0,80	0,80	20

Tabel di atas merupakan hasil yang telah didapatkan dari pengujian model yang telah dibuat, selanjutnya dilakukan pengujian model dengan menggunakan dataset baru dengan jumlah 5 baris. Seperti yang dapat diamati dalam Gambar 10, dataset tersebut akan digunakan untuk melakukan pengujian apakah hasil dari klasifikasi dari model telah sesuai dengan data fakta dan juga untuk melihat tingkat akurasi yang diperoleh.

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1
0	52	24.6000	75	3.549	0.805386	9.8827	22.432040	10.57635	773.920
1	66	19.6450	86	2.935	0.612725	9.5456	6.429285	12.76600	417.114
2	76	27.5620	87	3.115	0.945452	8.8438	10.702400	4.06405	664.697
3	44	24.5246	67	4.498	0.706897	17.9393	8.169560	6.25454	928.220
4	59	22.5154	96	2.707	0.467409	8.8071	7.819240	7.99585	468.786

Gambar 10. Dataset Baru Prediksi Logistic Regression

Selanjutnya dataset di atas kita lakukan pengujian dan dapat dilihat hasil pengujian pada Gambar 11. Hasil klasifikasi yang diperoleh cukup akurat dengan data fakta. Dapat disimpulkan bahwa algoritma Logistic Regression ini dinilai baik untuk melakukan prediksi dini penyakit Kanker payudara.

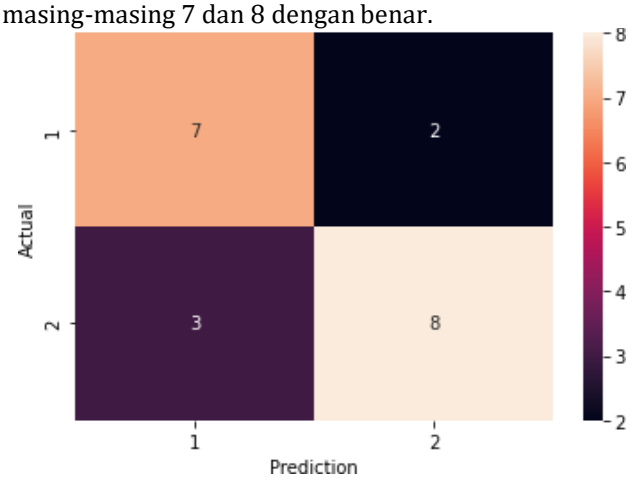
```
   Age      BMI  Glucose  Insulin      HOMA   Leptin  Adiponectin  Resistin  \
0   52  24.6000       75    3.549  0.805386   9.8827    22.432040  10.57635
1   66  19.6450       86    2.935  0.612725   9.5456     6.429285  12.76600
2   76  27.5620       87    3.115  0.945452   8.8438    10.702400   4.06405
3   44  24.5246       67    4.498  0.706897  17.9393     8.169560   6.25454
4   59  22.5154       96    2.707  0.467409   8.8071     7.819240   7.99585

     MCP.1
0  773.920
1  417.114
2  664.697
3  928.220
4  468.786
[1 2 1 1 2]
```

Gambar 11. Hasil Klasifikasi Dataset Baru Prediksi Logistic Regression

B. Naïve Bayes

Pada pengujian kedua dilakukan dengan menggunakan metode/algoritma Naïve Bayes. Pada pengujian dengan menggunakan model kedua ini mendapatkan hasil yang juga cukup baik, yaitu mendapatkan nilai akurasi data testing sebesar 75% dan akurasi data training sebesar 80%. Sama dengan pengujian metode sebelumnya, tahap selanjutnya ialah dilakukan validasi model prediksi. Validasi model prediksi ini dilakukan dengan confusion matrix, Gambar 12 merupakan confusion matrix dari model prediksi yang telah dilakukan testing terhadap 20 data. Dari pengujian tersebut, didapatkan total 9 data dengan class negative dan 11 total datadengan class negative. Dalam metode Naïve Bayes ini, sistemdapat melakukan prediksi dengan masing-masing 7 dan 8 dengan benar.



Gambar 12. Confusion Matrix Metode Naïve Bayes

Setelah pengujian dengan menggunakan confusion matrix selesai, dilanjutkan dengan classification report. Setelah dilakukan perhitungan pada data test, dapat disimpulkan bahwa sistem berjalan kurang maksimal pada pengujian keduanya, dengan prediksi negative sedikit lebih baik dibandingkan prediksi positive.

Tabel 4. Classification Report Metode Naïve Bayes

	Precision	Recall	F1-score	Support
1	0,78	0,70	0,74	10
2	0,73	0,80	0,76	10
Accuracy			0,75	20
Macro avg	0,75	0,75	0,75	20
Weighted avg	0,75	0,75	0,75	20

Tahap selanjutnya yaitu pengujian model yang telah dibuat, model Naïve Bayes ini diuji dengan dataset baru dengan jumlah 5 baris. Dataset baru seperti terlihat dalam Gambar 13, data tersebut digunakan untuk pengujian apakah hasil klasifikasi sudah sesuai dengan data fakta yang ada danjuga untuk melihat tingkat akurasi yang didapatkan.

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1
0	52	24.6000	75	3.549	0.805386	9.8827	22.432040	10.57635	773.920
1	66	19.6450	86	2.935	0.612725	9.5456	6.429285	12.76600	417.114
2	76	27.5620	87	3.115	0.945452	8.8438	10.702400	4.06405	664.697
3	44	24.5246	67	4.498	0.706897	17.9393	8.169560	6.25454	928.220
4	59	22.5154	96	2.707	0.467409	8.8071	7.819240	7.99585	468.786

Gambar 13. Dataset Baru Prediksi Naïve Bayes

Hasil dari pengujian model Naïve Bayes tersebut dapat dilihat dalam Gambar 14, hasil yang diperoleh dari pengujian tersebut cukup akurat dengan data fakta yang ada. Dapat disimpulkan bahwa penggunaan metode Naïve Bayes ini dalam prediksi penyakit kanker payudara cukup baik, hampir mirip dengan hasil klasifikasi metode Logistic Regression, namun berbeda dalam hasil klasifikasi pada data pertama dan kedua.

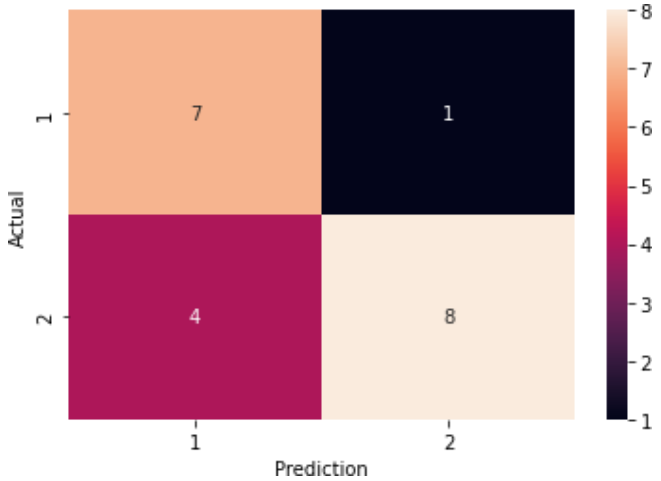
```
   Age      BMI  Glucose  Insulin      HOMA   Leptin  Adiponectin  Resistin  \
0   52  24.6000       75    3.549  0.805386   9.8827    22.432040  10.57635
1   66  19.6450       86    2.935  0.612725   9.5456     6.429285  12.76600
2   76  27.5620       87    3.115  0.945452   8.8438    10.702400   4.06405
3   44  24.5246       67    4.498  0.706897  17.9393     8.169560   6.25454
4   59  22.5154       96    2.707  0.467409   8.8071     7.819240   7.99585

     MCP.1
0  773.920
1  417.114
2  664.697
3  928.220
4  468.786
[2 1 1 1 2]
```

Gambar 14. Hasil Klasifikasi Dataset Baru Prediksi Naïve Bayes

C. Random Forest

Pengujian terakhir menggunakan algoritma Random Forest. Hasil pengujian menggunakan algoritma RandomForest dalam prediksi penyakit kanker payudara kali inimendapatkan nilai akurasi masing-masing sebesar 75% pada data test, dan akurasi data training sebesar 100%. Selanjutnya dilakukan pengujian dan validasi dari modelprediksi. Seperti yang ditunjukkan dalam Gambar 15, dalam confusion matrix tersebut ditemukan total data dengan class negative dan data dengan class positive. Sementara itu total keseluruhan sistem dapat memprediksi masing-masing dengan tepat.



Gambar 15. Confusion Matrix Metode Random Forest

Setelah dilakukan pengujian confusion matrix, dilanjutkan dengan classification report. Setelah dilakukan prosesperhitungan dari data testing, dapat ditarik kesimpulan bahwasistem dapat menentukan hasil nilai positive yang lebih baik dibandingkan ketika sistem menentukan hasil nilai negative.

Tabel 5. Classification Report Metode Random Forest

	Precision	**Recall**	**F1-score**	**Support**
1	0,88	0,64	0,74	11
2	0,67	0,89	0,76	9
Accuracy			0,75	20
Macro avg	0,77	0,76	0,75	20
Weighted avg	0,78	0,75	0,75	20

Kemudian dilakukan pengujian dengan dataset baru dengan jumlah 5 baris seperti pengujian model sebelumnya. Untuk detail dari isi dataset baru tersebut dapat dilihat dalamGambar 16. Dataset baru ini digunakan untuk pengujian akurasi model apakah sudah sesuai atau belum.

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1
0	52	24.6000	75	3.549	0.805386	9.8827	22.432040	10.57635	773.920
1	66	19.6450	86	2.935	0.612725	9.5456	6.429285	12.76600	417.114
2	76	27.5620	87	3.115	0.945452	8.8438	10.702400	4.06405	664.697
3	44	24.5246	67	4.498	0.706897	17.9393	8.169560	6.25454	928.220
4	59	22.5154	96	2.707	0.467409	8.8071	7.819240	7.99585	468.786

Gambar 16. Dataset Baru Prediksi Random Forest

Setelah dilakukan pengujian, hasil dari penggunaan modelRandom Forest ini dapat dilihat dalam Gambar 17, hasil yangdiperoleh sangat berbeda jika dibandingkan dengan kedua algoritma sebelumnya, terdapat 3 buah data yang berbeda. Jadi, dapat disimpulkan bahwa penggunaan metode RandomForest ini kurang direkomendasikan dalam prediksi dini penyakit Kanker payudara.

```
   Age      BMI  Glucose  Insulin      HOMA   Leptin  Adiponectin  Resistin  \
0   52  24.6000       75    3.549  0.805386   9.8827    22.432040  10.57635
1   66  19.6450       86    2.935  0.612725   9.5456     6.429285  12.76600
2   76  27.5620       87    3.115  0.945452   8.8438    10.702400   4.06405
3   44  24.5246       67    4.498  0.706897  17.9393     8.169560   6.25454
4   59  22.5154       96    2.707  0.467409   8.8071     7.819240   7.99585

     MCP.1
0  773.920
1  417.114
2  664.697
3  928.220
4  468.786
[2 2 2 2 2]
```

Gambar 17. Hasil Klasifikasi Dataset Baru Prediksi Random Forest

D. Comparison

Setelah dilakukan penerapan pada ketiga metode yakni Logistic Regression, Naïve Bayes, dan Random Forest, maka hasil perbandingan hasil akurasi dapat dilihat pada Tabel 6.

Tabel 6. Perbandingan Akurasi

Algoritma	**Training**	**Testing**
Logistic Regression	76%	80%
Naïve Bayes	80%	75%
Random Forest	100%	75%

Dilihat dari hasil perbandingan akurasi pada Tabel 6., dapat disimpulkan bahwa penggunaan model klasifikasi algoritma Logistic Regression memiliki tingkat akurasi yang jauh lebih tinggi dibandingkan kedua algoritma lainnya, terutama dalam testing.

4. SIMPULAN DAN SARAN

Berdasarkan pada hasil penelitian yang telah dilaksanakan dengan pengujian data *Breast Cancer Coimbra Dataset* dengan menggunakan algoritma *Logistic Regression*, Naive Bayes, dan Random Forest. Diperoleh hasil nilai akurasi pada*Logistic Regression* dengan total nilai akurasi *data testing* sebesar 80% dan akurasi *data training* sebesar 76%. Untuk algoritma Naïve Bayes diperoleh hasil nilai akurasi *data testing* sebesar 75% dan akurasi *data training* sebesar 80%. Sementara itu dengan menggunakan algoritma Random Forest memperoleh hasil akurasi *data testing* sebesar 70% dan akurasi *data training* sebesar 100%. Dengan demikian, dari hasil yang didapatkan dari evaluasi tersebut dapat disimpulkan bahwa algoritma *Logistic Regression* memiliki tingkat akurasi yang jauh lebih tinggi dibandingkan dengan algoritma Naive Bayes dan Random Forest. Hasil yang didapatkan dalam penelitian ini yaitu algoritma *Logistic Regression* dapat dijadikan sebuah komparasi atau perbandingan dengan prediksi menggunakan algoritma klasifikasi yang lain, sehingga akan diperoleh algoritma yanglebih akurat dalam prediksi penyakit Kanker payudara.

Daftar Pustaka

Acuna, E., & Rodriguez, C. (n.d.). *The Treatment of Missing Values and its Effect on Classifier Accuracy*. *1995*.

Ajeng Wijayanti, L., Bambang Hidayat, D., Suhardjo, D., & SpRKG, M. (2017). *PENGOLAHAN CITRA RADIOGRAF PERIAPIKAL PADA DETEKSI PENYAKIT GRANULOMA MENGGUNAKAN METODE DISCRETE WAVELET TRANSFORM & PRINCIPAL COMPONENT ANALYSIS BERBASIS ANDROID Image Processing of Periapical Radiograph on Granuloma Desease Detection Using Discrete Wav*. *4*(2), 547–553.

Alita, D., Priyanta, S., & Rokhman, N. (2019). *Analysis of Emoticon and Sarcasm Effect on Sentiment Analysis of Indonesian Language on Twitter*. *5*(2), 100–109.

Bustami. (2013). PENERAPAN ALGORITMA NAIVE BAYES UNTUK MENGKLASIFIKASI DATA NASABAH. *TECHSI: Jurnal Penelitian Teknik Informatika*.

Dinh, V. T., Luu, N. D., & Trinh, H. H. (2016). Vehicle classification and detection based coarse data for warning traffic jam in VietNam. *NICS 2016 - Proceedings of 2016 3rd National Foundation for Science and Technology Development Conference on Information and Computer Science*, 223–228. https://doi.org/10.1109/NICS.2016.7725654

Goel, V. (2018). *Building a Simple Machine Learning Model on Breast Cancer Data*. https://towardsdatascience.com/building-a-simple-machine-learning-model-on-breast-cancer-data-eca4b3b99fa3

Goldenberg, R., & Punthakee, Z. (2013). Definition, Classification and Diagnosis of Diabetes, Prediabetes and Metabolic Syndrome. *Canadian Journal of Diabetes*, *37*(SUPPL.1), 8–11. https://doi.org/10.1016/j.jcjd.2013.01.011

Harimurti, F. A. (2017). *METODE NAÏVE BAYES CLASSIFIER UNIVERSITAS TRUNOJOYO MADURA) MENGGUNAKAN (STUDI KASUS SCHOLARSHIP RECEPTION CLASSIFICATION USING NAÏVE BAYES CLASSIFIER METHOD (CASE STUDY UNIVERSITAS TRUNOJOYO MADURA)*.

Kementerian Kesehatan Republik Indonesia. (2022). *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

Kulkarni, V. Y., & Sinha, P. K. (2014). *Effective Learning and Classification using Random Forest Algorithm*. *3*(11), 267–273.

Kumar, M., Singhal, S., Shekhar, S., & Sharma, B. (2022). *Optimized Stacking Ensemble Learning Model for Breast Cancer Detection and Classification Using Machine Learning*.

Mulyo, A., Yanathifal, W., Anggraeni, S., Anwar, N., Ichwani, A., & Anggara, B. (2013). *Performansi K-NN , J48 , Naive Bayes dan Regresi Logistik Sebagai Algoritma Pengklasifikasi Diabetes*. *2007*.

Network, E. (2022). *Retracted : Application of Feature Selection Based on Elastic Network and Random Forest in the Evaluation of Sports Effects*. *2022*.

Permanasari, A. E., Nurlayli, A., & Description, A. D. (2017). *Decision Tree to Analyze the Cardiotocogram*

Data for Fetal Distress Determination. 459–463.

Qin, Y., Zhang, S., Zhu, X., Zhang, J., & Zhang, C. (2007). *Semi-parametric optimization for missing data imputation*. 79–88. https://doi.org/10.1007/s10489-006-0032-0

Ramli, Yuniarti, D., & Goejantoro, R. (2013). *Comparison of Classification Methods Between Logistic Regression and Artificial Neural Network (Case Study: Selection of Language and Social Studies Depertement at SMAN 2 Samarinda academic year 2011/2012)*. *4*, 17–24.

Sadewo, M. G., Windarto, A. P., & Hartama, D. (2016). *PENERAPAN DATAMINING PADA POPULASI DAGING AYAM RAS PEDAGING DI INDONESIA BERDASARKAN PROVINSI MENGGUNAKAN K-MEANS*. 60–67.

Sandeep, D., & Bethel, G. N. B. (2021). *A Survey on Accurate Breast Cancer Detection and Classification using Machine Learning Approach*.

Shah, S. A. A., Aziz, W., Arif, M., & Nadeem, M. S. A. (2016). Decision Trees Based Classification of Cardiotocograms Using Bagging Approach. *Proceedings - 2015 13th International Conference on Frontiers of Information Technology, FIT 2015*, 12–17. https://doi.org/10.1109/FIT.2015.14

The American Cancer Society medical and editorial content team. (2022). *American Cancer Society Recommendations for the Early Detection of Breast Cancer*. https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/american-cancer-society-recommendations-for-the-early-detection-of-breast-cancer.html#:~:text=at any age.-,Mammograms,years before physical symptoms develop.

WHO. (2021). *Breast cancer*. https://www.who.int/news-room/fact-sheets/detail/breast-cancer