Jurnal Kajian Ilmiah
e-ISSN: 2597-792X, ISSN: 1410-9794
Vol. 23 No. 1 (Januari 2023), Halaman: 87 – 94
Terakreditasi Peringkat 4 (SINTA 4) sesuai SK RISTEKDIKTI Nomor. 158/E/KPT/2021

Algoritma *Naïve Bayes* Dengan *Backward Elimination* Pada Dataset *Breast Cancer*

Recha Abriana Anggraini [1,*]

[1] Fakultas Sistem Informasi; Universitas Bina Sarana Informatika; e-mail: recha.rcb@bsi.ac.id

* Korespondensi: e-mail: recha.rcb@bsi.ac.id

Submitted: **18/12/2022**; Revised: **05/01/2023**; Accepted: **10/01/2023**; Published: **23/01/2023**

Abstract

Cancer is a type of disease that is not recognized by most people, because some people affected by this disease do not know about cancer itself and do not do early detection of cancer, as a result most cancers are found at an advanced stage and are difficult to treat, thus placing a large burden on cancer sufferers. . Early detection of cancer, especially breast cancer is very important to do to overcome the very high risk of death in women caused by breast cancer. This study aims to help classify breast cancer based on data from routine patient examinations which are summarized in the coimbra breast cancer dataset and this data was donated to the UCI machine learning repository in 2018. The method used in the classification process in this study is backward elimination modeling for optimization accuracy as well as the naive Bayes algorithm and split validation validation to validate the model. The results of this study show an accuracy of 77.14%. These results indicate that the results of this study are good enough to help classify breast cancer.

Keywords*: Backward Elimination, Breast Cancer, Classification, Naïve Bayes, Split Validation*

Abstrak

Kanker merupakan jenis penyakit yang kurang disadari oleh sebagian masyarakat, karena sebagian masyarakat yang terkena penyakit ini kurang mengetahui kanker itu sendiri dan kurangnya melakukan deteksi dini terhadap kanker, akibatnya sebagian besar kanker ditemukan pada stadium lanjut dan sulit ditanggulangi sehingga memberikan beban yang besar bagi penderita kanker. Deteksi dini terhadap kanker khususnya kanker payudara sangat penting dilakukan untuk menanggulangi resiko kematian pada wanita yang sangat tinggi disebabkan oleh kanker payudara. Penelitian ini bertujuan untuk membantu melakukan klasifikasi kanker payudara berdasarkan data hasil pemeriksaan rutin pasien yang dirangkum dalam dataset breast cancer coimbra dan data tersebut didonasikan kepada UCI *machine learning repository* pada tahun 2018. Metode yang digunakan dalam proses klasifikasi pada penelitian ini yaitu permodelan *backward elimination* untuk optimasi akurasi serta algoritma naive bayes dan validasi split validation untuk memvalidasi permodelan. Hasil dari penelitian ini menunjukkan hasil akurasi sebesar 77.14%. Hasil tersebut menunjukkan bahwa hasil penelitian ini cukup baik untuk membantu mengklasifikan kanker payudara.

Kata kunci: *Backward Elimination,* Kanker Payudara, Klasifikasi*, Naïve Bayes, Split Validation*

1. Pendahuluan

Kanker merupakan jenis penyakit yang kurang disadari oleh sebagian masyarakat, karena sebagian masyarakat yang terkena penyakit ini kurang mengetahui kanker itu sendiri dan kurangnya melakukan deteksi dini terhadap kanker, akibatnya sebagian besar kanker ditemukan pada stadium lanjut dan sulit ditanggulangi sehingga memberikan beban yang besar

bagi penderita kanker (Safutra & Prabowo, 2016). Kanker payudara (*Carcinome mamae*) dalam bahasa inggrisnya disebut *breast cancer* merupakan kanker pada jaringan payudara (Andini & Putri, 2018). Saat ini, kanker payudara merupakan penyebab kematian kedua akibat kanker pada wanita, setelah kanker leher rahim dan merupakan kanker yang paling banyak ditemui diantara Wanita (Safutra & Prabowo, 2016).

Berdasarkan data dari *American Cancer Society*, sekitar 1,3 juta wanita terdiagnosis menderita kanker payudara dan tiap tahunnya diseluruh dunia kurang lebih 465.000 wanita meninggal karena penyakit ini. Mengingat dampak yang ditimbulkan oleh penyakit tersebut maka sangat perlu dilakukan deteksi dini terhadap kanker payudara agar dapat meminimalisir resiko kematian pada wanita yang diakibatkan oleh kanker payudara. Umumnya proses deteksi dan klasifikasi suatu penyakit dilakukan dengan bantuan diagnosa seorang dokter secara langsung (Maryam & Ariono, 2022). Perkembangan teknologi medis dan teknologi informasi, di dalam dunia medis dapat digunakan peneliti di bidangnya untuk mengembangkan model deteksi dini dari data konsultasi rutin dan analisis darah (Patrício et al., 2018). Salah satu metode yang bisa dipakai untuk melakukan deteksi dini kanker tersebut adalah menggunakan teknik *deep learning* (Priatna, Purnomo, & Putra, 2021). Untuk meminimalisir kesalahan deteksi dan menghindari keterlambatan deteksi pada sel kanker dapat dilakukan penerapan dan pemanfaatan teknik *data mining. Data mining* merupakan sebuah proses untuk menemukan suatu hubungan, pola, atau tren baru yang bermakna dengan cara menyaring data yang sangat besar atau yang biasa disebut big data yang tersimpan dalam *repository* atau penyimpanan dengan menggunakan teknik pengenalan pola seperti teknik statistik atau matematika (Kamagi & Hansun, 2016).

Kegunaan teknik *data mining* telah diketahui dan banyak diterapkan di dunia medis karena teknik data mining dapat menemukan pola tersembunyi yang biasanya tidak ditemukan. Teknik *data mining* yang telah diterapkan pada data medis meliputi *association rule* untuk menemukan pola berulang, prediksi, klasifikasi, dan klastering (Safutra & Prabowo, 2016). Salah satu teknik yang sering dipakai dalam penelitian data medis adalah teknik klasifikasi.Klasifikasi memerlukan data yang baik agar diperoleh hasil yang sesuai selain juga pemilihan bahasa pemrograman yang banyak dijumpai saat ini (Herlawati & Handayanto, 2020). Teknik tersebut telah terbukti keakuratannya dalam memberikan diagnosa dalam dunia medis. Algoritma dari teknik klasifikasi dalam data mining terdiri dari beberapa algoritma, salah satunya adalah *naive bayes*.

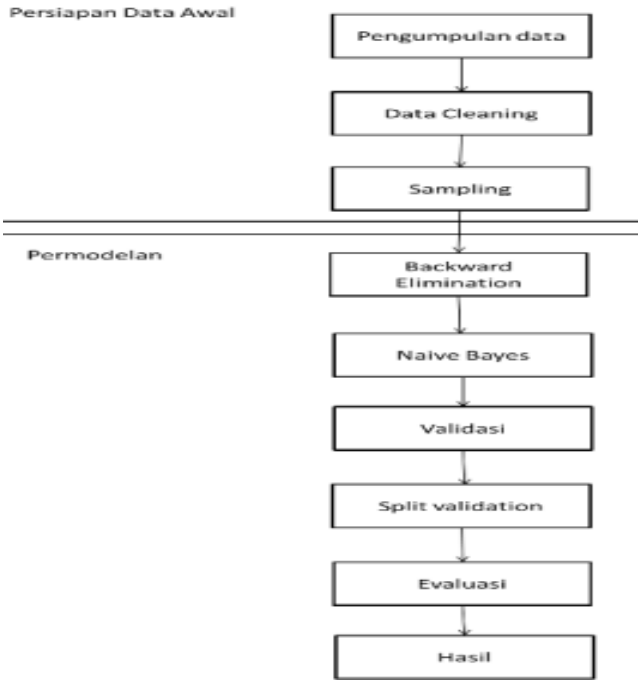
Naive bayes adalah metode yang diadopsi dari nama penemunya yaitu Thomas Bayes pada tahun 1950. Pada dasarnya, teori tersebut menyatakan bahwa kejadian dimasa depan dapat diprediksi dengan syarat kejadian sebelumnya telah terjadi. Teori *naive bayes* memiliki kemampuan klasifikasi yang serupa dengan *decision tree* dan *neural network* bahkan algoritma *naive bayes* memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam *database* dengan data yang besar (Anggraini, Widagdo, Budi, & Qomaruddin, 2019).

Backward elimination adalah salah satu metode yang memiliki fungsi untuk pengoptimalan kinerja suatu model dengan cara kerja sistem pemilihan mundur (Drajana, 2017). Metode ini biasanya digunakan untuk meningkatkan akurasi dari proses klasifikasi, dengan menambahkan metode ini dalam proses pengolahan data diharapkan dapat memperoleh hasil akurasi yang maksimal untuk mengklasifikasikan data.

Dalam penelitian ini, data yang akan diolah adalah data *breast cancer coimbra* yang diperoleh dari UCI *machine learning repository*. Data tersebut akan diolah menggunakan teknik klasifikasi data mining algoritma *naive bayes*. Tujuan dari penelitian ini adalah untuk menerapkan algoritma *Naive Bayes* untuk mengklasifikasi *breast cancer coimbra* untuk menghasilkan model *data mining* dengan akurasi terbaik. Penerapan algoritma ini terhadap *dataset breast cancer Coimbra* belum pernah dilakukan pada penelitian sebelumnya sehingga diharapkan hasil dari penelitian ini dapat memiliki nilai akurasi yang Ibih baik sehingga proses klasifikasi yang dihasilkan lebih akurat.

2. Metode Penelitian

Pada penelitian ini diperlukan kerangka penelitian untuk menggambarkan penelitian yang akan dilakukan, dibawah ini adalah gambar 1 kerangka penelitian:



Sumber: Hasil Penelitian (2022)

Gambar 1. Kerangka Penelitian

Pada gambar 1 dijelaskan tentang 4 tahap penelitian *data mining* yang dilakukan yaitu Pengumpulan Data, *Data Cleaning* dan *Sampling* yang termasuk dalam tahap Persiapan Data Awal, Permodelan yang meliputi penggunaan metode algoritma *data mining* dan validasi model, dan evaluasi. Data yang digunakan untuk penelitian ini merupakan data *public* yang bernama *breast cancer coimbra dataset* yang terdiri dari 116 data pasien dengan 9 atribut. Kelas pada

data itu sendiri dan terbagi menjadi dua kelas, yaitu 1 (*healthy control*) dan 2 (*patient*). Dari 116 pasien dalam *dataset*, terdiri dari 64 pasien terdiagnosis mengidap breast cancer dan 52 pasien didiagnosis tidak mengidap *breast cancer*.

Pada tahap kedua, *dataset* kemudian diolah menggunakan proses *data cleaning* dan *data transformation* sebagai persiapan awal sebelum proses *data mining*. Kemudian data dibagi menjadi dua, bagian pertama sebagai data training dan bagian kedua sebagai *data testing*. Kemudian data tersebut diolah menggunakan algoritma klasifikasi *data mining Naïve Bayes* dengan metode *Backward Elimination*. Teori bayes adalah kondisi probabilitas suatu kejadian hipotesis bergantung pada kejadian lain sebagai bukti (Moriesta, Selviani, & Ibrahim, 2017). Teori *naive bayes* memiliki kemampuan klasifikasi yang serupa dengan *decision tree* dan *neural network* bahkan algoritma *naive bayes* memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam *database* dengan data yang besar (Anggraini et al., 2019).

Hasil *data mining* kemudian divalidasi menggunakan metode *split validation*. Tahap akhirnya adalah mengevaluasi hasil model *data mining* yang digunakan untuk mengetahui tingkat akurasi model *data mining*. Di bawah ini adalah tabel yang menjelaskan jumlah data yang dipakai dan nama atribut yang ada pada *breast cancer coimbra dataset*.

Tabel 1. Atribut Dataset

No	Nama Atribut
1	Age
2	BMI
3	Glucose
4	Insulin
5	HOMA
6	Leptin
7	Adiponectin
8	Resistin
9	MCP-1

Sumber: (Patricio et al., 2022)

Tabel 1 menunjukkan atribut yang ada didalam *dataset breast cancer Coimbra*. Jumlah *class* pada dataset tersebut ada 2 yaitu *healthy control* dan *breast cancer*.

3. Hasil dan Pembahasan

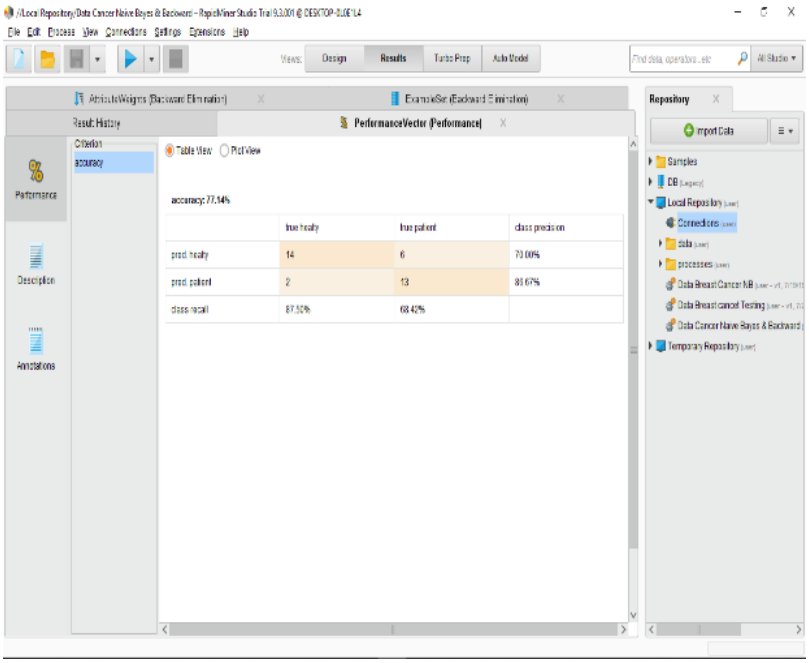
3.1. Eksperimen Rapid Miner

Penerapan metode *naive bayes* pada *breast cancer coimbra dataset* bertujuan untuk mengetahui dan mendapatkan hasil akurasi yang lebih baik dan menerapkan metode klasifikasi yang belum pernah diterapkan pada klasifikasi deteksi penyakit *breast cancer* dari penelitian-penelitian sebelumnya. Hasil akurasi yang optimal akan diraih atau tidak, semuanya akan terlihat pada hasil eksperimen.

Eksperimen pada algoritma naive bayes akan dilakukan dengan menggunakan *backward elimination* dan metode *validasi split validation*. *Backward elimination* diterapkan dengan tujuan untuk mengoptimalkan nilai akurasi dari permodelan algoritma *naive bayes* sedangkan metode validasi *split validation* diterapkan sebagai teknik evaluasi pada model *naive bayes*. Dalam penelitian ini, dataset yang telah disiapkan untuk diimplementasikan pada proses

uji model kemudian diujikan terhadap proses *backward elimination* dan algoritma *naive bayes.* Setelah dilakukan proses uji model, maka di dapatkan hasil akurasi sebesar 77.14%.

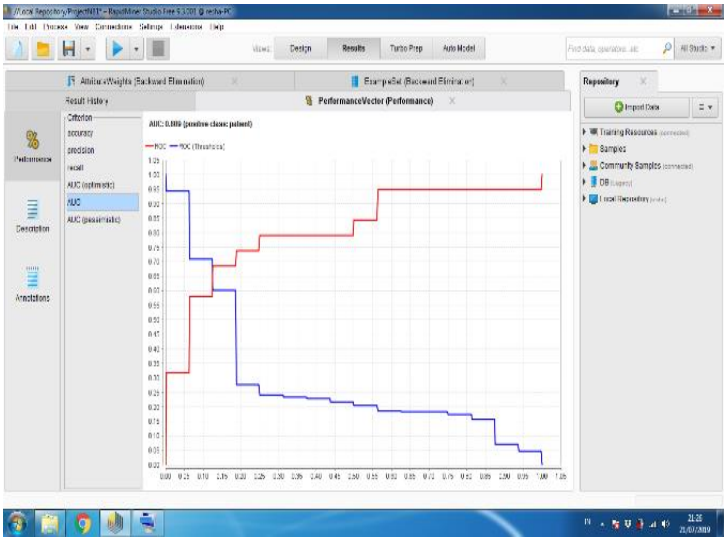
Dari nilai akurasi hasil dari permodelan dapat dilihat nilai *confussion matrix* yang menunjukkan relevansi hasil akurasi. Berikut adalah nilai *confussion matrix* dari permodelan tersebut. Dari gambar 2 dapat dilihat nilai akurasi yang didapatkan pada permodelan dengan metode validasi *split validation* sebesar 77.14%.



Sumber: Hasil Penelitian (2022)

Gambar 2. *Confussion Matrix*

Gambar 3 menunjukkan bentuk dari kurva ROC setelah dilakukan validasi terhadap data. Kurva tersebut menunjukkan nilai AUC sebesar 0.809 dan nilai tersebut menujukkan bahwa hasil klasifikasinya termasuk dalam kategori *good classification*.



Sumber: Hasil Penelitian (2022)

Gambar 3. Kurva ROC

3.2. Perhitungan Manual Algoritma Naïve Bayes

Mencari nilai prediksi pada data testing dalam permodelan *naive bayes* dilakukan dengan cara menghitung *prior probabilities* dan *conditional probabilities* untuk mencari hasil perhitungan *posterior probabilities* masing-masing *class*. Berikut contoh klasifikasi dengan menggunakan perhitungan manual untuk salah satu data *testing* dengan menggunakan seluruh *dataset* sebagai *data training*.

Tabel 2. Data Testing perhitungan manual metode naïve bayes

Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP1	Class
48	18,67	75	4,09	1,33	8,88	5,1	3,32	63,61	?

Sumber: Dataset Breast Cancer Coimbra (2022)

Berdasarkan tabel *class prior probabilities* dan tabel *conditional probabilities* pada dataset, dapat dihitung *posterior probabilities* dari *data testing* sebagai berikut:

P (*Healthy*) = 55 : 116

P (*Patient*) = 64 : 116

Perhitungan Data Testing:

P (*Healthy*) = (1/52)*(1/52)* (1/52)* (1/52)* (1/52)* (1/52)* (1/52)* (1/52)* (1/52)

= 0,0192*0,0192*0,0192*0,0192*0,0192*0,0192*0,0192*0,0192*0,0192

= 3,546

P (*Patient*) = (0/64)* (0/64)* (0/64)* (0/64)* (0/64)* (0/64)* (0/64)* (0/64)* (0/64)

= 0

Dari hasil perhitungan manual data testing tersebut dapat diketahui bahwa nilai P(*healthy*) > nilai P(*patient*) sehingga dapat diambil kesimpulan bahwa orang tersebut termasuk dalam kelompok *Healthy*.

3.3. Perbandingan Hasil Penelitian Dengan Penelitian Sebelumnya

Perbandingan hasil pada penelitian ini dengan penelitian-penelitian sebelumnya adalah untuk mengevaluasi hasil keseluruhan pada penelitian ini. Penelitian ini adalah penelitian lanjutan dari penelitian-penelitian sebelumnya dengan objek data yang sama, yaitu *Breast Cancer Coimbra Dataset*.

Pada Tabel 3 adalah tabel hasil dari beberapa penelitian sebelumnya yang telah dilakukan terhadap *breast cancer coimbra dataset*. Dilihat dari hasil penelitian-penelitian sebelumnya, percobaan klasifikasi dengan menggunakan *backward elimination* dengan algoritma *naive bayes* dan metode validasi *split validation* dapat dikatakan cukup baik hasilnya, karena nilai akurasi yang diperoleh dari permodelan ini dapat dikatakan lebih baik dari nulai akurasi beberapa permodelan dan algoritma yang dilakukan pada penelitian-penelitian sebelumnya.

Tabel 3. Perbandingan Penelitian

No	Judul	Metode	Accuracy
1	*A Novel ML Approach to Prediction of Breast Cancer: Combining of mad normalization, KMC based feature weighting and AdaBoostM1 classifier (Polat & Senturk, 2018)*	MAD & AdaBoost M1 Classifier	75%
		MAD, KMC & AdaBoostM1 Classifier	91,37%
2	*Comparison between Fuzzy Kernel C-Means and Sparse Learning Fuzzy C-Means for Breast Cancer Clustering. Objek penelitiannya adalah menbandingkan algoritma Fuzzy (Fijri & Rustam, 2018).*	SLFCM	39,72% - 89,28%
		FKCM	70,27% - 70,64%
3	*Classification of Breast Cancer Using Data Mining (Sardouk, Duru, Bayat, & others, 2019)*	AdaBoostM1	75,00%
		Classification Via Regression	76,20%
		Random forest	74,10%
		Jrip	71,50%
		RBFNN	72,90%
		J48	69,10%

Sumber: Hasil Pengolahan Data (2018-2019)

4. Kesimpulan

Pada penelitian ini dilakukan eksperimen terhadap *breast cancer coimbra dataset* dilakukan dengan menggunakan algoritma *data mining* untuk klasifikasi yaitu *naive bayes* dan metode validasi metode *split validation*. Dari hasil eksperimen tersebut dapat diperoleh kesimpulan yaitu telah diterapkan algoritma untuk optimasi *data mining backward elimination* serta algoritma klasifikasi *naive bayes* untuk klasifikasi *dataset breast cancer* dengan metode validasi *split validation* dan diketahui hasil akurasi dari eksperimen menggunakan optimasi backward elimination serta algoritma *naive bayes* dan metode validasi *split validation* terhadap *breast cancer coimbra dataset*. Pada penelitian berikutnya diharapkan dapat menerapkan *backward elimination* serta algoritma naive bayes dan metode validasi *split validation* terhadap dataset lainnya dengan jumlah atribut data yang lebih beragam agar dapat diketahui nilai akurasi dan performanya juga diharapkan dapat menggunakan algoritma *data mining* lainnya yang belum digunakan terhadap *breast cancer coimbra dataset* agar dapat diketahui performa dan nilai akurasi klasifikasinya

Daftar Pustaka

Andini, A., & Putri, D. F. A. (2018). Sistem Pakar Diagnosa Penyakit Tuberculosis Menggunakan Certainty Factor, (672014245), 1–12.

Anggraini, R. A., Widagdo, G., Budi, A. S., & Qomaruddin, M. (2019). Penerapan Data Mining Classification untuk Data Blogger Menggunakan Metode Naïve Bayes. *Jurnal Sistem Dan Teknologi Informasi (JUSTIN)*, *7*(1), 47. https://doi.org/10.26418/justin.v7i1.30211

Drajana, I. C. R. (2017). Metode Support Vector Machine Dan Forward Selection Prediksi Pembayaran Pembelian Bahan Baku Kopra. *ILKOM Jurnal Ilmiah*, *9*(2), 116–123. https://doi.org/10.33096/ilkom.v9i2.134.116-123

Fijri, A. L., & Rustam, Z. (2018). Comparison between Fuzzy Kernel C-Means and Sparse Learning Fuzzy C-Means for Breast Cancer Clustering. *Proceedings of ICAITI 2018 - 1st*

International Conference on Applied Information Technology and Innovation: Toward A New Paradigm for the Design of Assistive Technology in Smart Home Care, (4), 158–161. https://doi.org/10.1109/ICAITI.2018.8686707

Herlawati, H., & Handayanto, R. T. (2020). Penggunaan Matlab dan Python dalam Klasterisasi Data. *Jurnal Kajian Ilmiah*, *20*(1), 103–118. https://doi.org/10.31599/jki.v20i1.85

Kamagi, D. H., & Hansun, S. (2016). A robust active stabilization technique for dc microgrids with tightly controlled loads. *Proceedings - 2016 IEEE International Power Electronics and Motion Control Conference, PEMC 2016*, *VI*(1), 254–260. https://doi.org/10.1109/EPEPEMC.2016.7752007

Maryam, M., & Ariono, H. W. (2022). Sistem Pakar Pengklasifikasi Stadium Kanker Serviks Berbasis Mobile Menggunakan Metode Decision Tree. *Jurnal Kajian Ilmiah*, *22*(3), 267–278. https://doi.org/10.31599/jki.v22i3.1368

Moriesta, E., Selviani, & Ibrahim, A. (2017). Analisis Penyaringan Email Spam Menggunakan Metode Naive Bayes. *Prosiding Annual Research Seminar 2017*, *3*(1), 45–48.

Patricio, M., Pereira, J., Crisosteomo, J., Matafome, P., Seica, R., Caramelo, F., & Gomes, M. (2022). Uci Repository. Retrieved from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra

Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., & Caramelo, F. (2018). Using Resistin, glucose, age and BMI to predict the presence of breast cancer. *BMC Cancer*, *18*(1), 1–8. https://doi.org/10.1186/s12885-017-3877-1

Polat, K., & Senturk, U. (2018). A Novel ML Approach to Prediction of Breast Cancer: Combining of mad normalization, KMC based feature weighting and AdaBoostM1 classifier. *ISMSIT 2018 - 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies, Proceedings*. https://doi.org/10.1109/ISMSIT.2018.8567245

Priatna, W., Purnomo, R., & Putra, T. D. (2021). Implementasi Deep Learning Untuk Rekomendasi Aplikasi E-learning Yang Tepat Untuk Pembelajaran jarak jauh. *Jurnal Kajian Ilmiah*, *21*(3), 261–274. https://doi.org/10.31599/jki.v21i3.521

Safutra, A. R., & Prabowo, D. W. (2016). Diagnosis Penyakit Kanker Payudara Menggunakan Metode Naive Bayes Berbasis Desktop. *Jurnal Penelitian Dosen FIKOM (UNDA)*, *6*(1), 1–6.

Sardouk, F., Duru, A. D., Bayat, O., & others. (2019). Classification of breast cancer using data mining. *American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)*, *51*(1), 38–46.