Institut Riset dan Publikasi Indonesia (IRPI)
SENTIMAS: Seminar Nasional Penelitian dan Pengabdian Masyarakat
Homepage: https://journal.irpi.or.id/index.php/sentimas
Prosiding Seminar Nasional/ ISSN(E): 2830-3083
03 Agustus 2023/ Pages: 86-92

Application of Decision Tree Algorithm and Linear Regression for Breast Cancer Classification

Penerapan Algoritma Decision Tree aan Regresi Linear untuk Klasifikasi Kanker Payudara

Putri Kurnia Illahi[1], Ayu Rina Viana[2], Nur Fitria Mita Permata[3], Muhammad Yudha Pratama[4]

[1,2,3,4] Program Studi Sistem Informasi, Fakultas Sains dan Teknologi,
UIN Sultan Syarif Kasim Riau,
Jl. HR Soebrantas, KM. 18.5, No. 155, Simpang Baru, Pekanbaru, Indonesia, 28293

E-Mail: [1]12050322876@students.uin-suska.ac.id, [2]12050323128@students.uin-suska.ac.id, [3]12050320446@students.uin-suska.ac.id, [4]12050312952@students.uin-suska.ac.id

Corresponding Author: Putri Kurnis Illahi

Abstract

Breast cancer is the most common cancer for women and the leading cause of cancer death worldwide. Patients with breast cancer can be predicted by means of data mining. In data mining there are many concepts and models, one of which is the decision tree concept. Linear regression is a statistical calculation to determine the influence between variable 1 and other variables. One of the methodologies in this study is divided into five paths, namely data collection, data preprocessing, decision tree processing and linear regression, validity testing, and analysis results. The results and discussion in this study used a dataset of breast cancer. From the experimental results, the accuracy rate between the Decision Tree algorithm is 93.51% and Linear Regression is 95.61%.

Keyword: Breast Cancer, Classification, Decision Tree, Linear Regression,

Abstrak

Kanker payudara merupakan kanker yang paling umum untuk wanita dan penyebab utama kematian kanker diseluruh dunia. Penderita penyakit kanker payudara dapat diprediksi dengan cara data mining. Dalam data mining ada banyak konsep dan model, salah satunya ialah konsep Pohon Keputusan (Decision Tree). Regresi linear merupakan hitungan statistic untuk menetapkan pengaruh antara vaiabel 1 dan variabel lainnya. Salah satu metodologi dalam penelitian ini dibagi dalam lima alur yaitu pengumpulan data, preprocessing data, proses decision tree dan regresi linear, uji validitas, dan hasil analisis. Hasil dan pembahasan dalam penelitian ini menggunakan dataset penyakit kanker payudara. Dari hasil percobaan, didapat tingkat akurasi antara algoritma Decision Tree sebesar 93,51% dan Regresi Linear sebesar 95,61%.

Kata Kunci: Decision Tree, Kanker Payudara, Klasifikasi, Regresi Linear

1. PENDAHULUAN

Kanker payudara adalah kanker yang paling umum untuk wanita dan penyebab utama kematian kanker diseluruh dunia. *World Health Organization* memperkirakan bahwa 84 juta orang meninggal yang disebabkan karena kanker dari 2005 hingga 2015 [1]. Pengamatan yang dilakukan WHO menyatakan bahwa 8-9 persen wanita menderita penyakit kanker payudara. Karena itu membuat kanker payudara sebagai jenis kanker yang sering banyak ditemui pada wanita setelah terkena kanker leher Rahim [2].

Kanker payudara ini secara umum dibagi menjadi 2, yang pertama *benign* atau disebut jinak, yang kedua *malignant* disebut juga ganas. Kanker payudara jinak ditandai dengan bentuk benjolan kecil bulat lembut [3]. Pada kanker payudara tingkat ganas berbentuk yang tidak simetris, kasar, rasa nyeri dan lainnya, dan biasanya kanker payudara ini menyebar dan merusak jaringan dan organ lain yang ada didekatnya [4]. Dan untuk saat

ini ada satu cara pengobatan kanker payudara adalah dengan pembedahan dan jika perlu dilanjutkan dengan kemoterapi maupun radiasi [5].

Penderita penyakit kanker payudara dapat diprediksi dengan cara data mining. Menurut *Daryl Pregibon* bahwa data mining merupakan gabungan dari stastistik, kecerdasan buatan, dan riset basis data yang masih berkembang (Gonunescu, 2011) [6]. Data mining mempunyai makna sama dengan *Knowledge-discovery in database* (KDD) yang memiliki tujuan untuk memanfaatkan data dalam *database* dengan mengelolanya dan mendapatkan hasil informasi baru yang berguna [6].

Salah satu peran data mining yaitu memprediksi. Prediksi serupa dengan klasifikasi dan estimasi, hanya saja untuk hasil prediksinya itu ada dimasa depan (Larose,2005).

Dalam data mining ada banyak konsep dan model, salah satunya ialah konsep Pohon Keputusan (*Decision Tree*), Pohon keputusan merupakan konsep yang dimanfaatkan sebagaimana untuk prosedur kecerdasan untuk mendapatkan jawaban dari suatu masalah, pohon yang dihasilkan tidak hanya dalam bentuk biner [7]. Dalam kasus penentuan keputusan *decision tree* lebih banyak digunakan [8]. Metode ini menarik karena lebih fleksibel dalam memberikan keuntungan berupa visualisasi saran sehingga prediksinya dapat diamati. Tetapi ketika kelas dan kriteria digunakan sangat banyak menyebabkan keputusan yang lebih lama dan memori yang dibutuhkan sangat banyak [9]. Selain itu hasil kualitas hasil keputusan yang didapat dengan menggunakan metode ini sangat tergantung pada rancangannya. Ada dua jenis model prediksi yaitu klasifikasi dan regresi. Klasifikasi dipakai pada variabel target diskret, kalau regresi dipakai untuk pada variabel target continue [10].

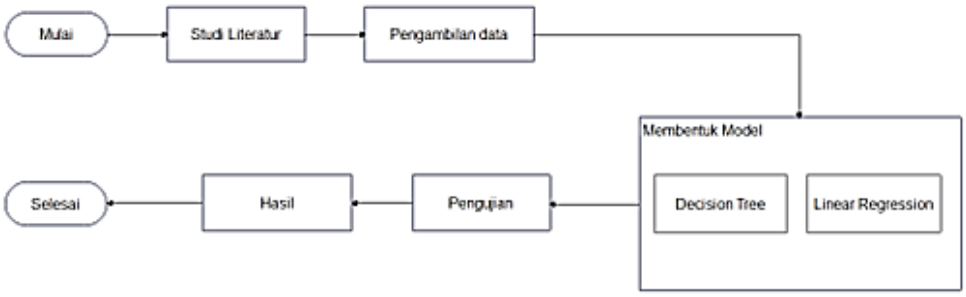
Regresi linear merupakan hitungan statistik untuk menetapkan pengaruh antara vaiabel 1 dan variabel lainnya. Dengan adanya analisis regresi linear untuk dapat memprediksi nilai antara variabel dengan lebih akurat. Regresi linear merupakan model yang terhubung antara variabel terikat (*dependen*) dengan variabel bebas (*independen*) [11].

Berdasarkan uraian diatas, maka penelitian ini melakukan penerapan berupa algoritma *decision tree* dan regresi linear untuk mengetahui tingkat akurasi dan hasil klasifikasi pada penyakit kanker payudara tersebut [12]. Pada penelitian sebelumnya, algoritma *decision tree* dapat diimplementasikan pada rekomendasi penerimaan mitra penjualan di PT. Atria Artha Persada, dilihat dari tingkat akurasi mencapai 96.26% yang mampu memprediksi untuk penjualan. [13].

Dan penelitian sebelumnya yang lain, beberapa algoritma klasifikasi diujicoba terhadap data dengan hasil terbaik pada model klasifikasi dengan algoritma decision tree. Hasilnya adalah penerapan algoritma decision tree akan membanu pihak koperasi dalam menentukan jumlah kredit yang akan dicairkan [14]. Penelitian terdahulu mengenai prediksi menggunakan regresi linear, dalam penelitiannya untuk memprediksi bahan bakar dengan hasil metode ini mampu menghasilkan akurasi sebesar 94,82% [15]. Dan penelitian terdahulu yang serupa dengan hasil yang menunjukan penggunaan metode regresi linear dengan mencapai akurasi sebesar 80% [16].

2. BAHAN DAN METODE

Dalam penelitian ini metodologi penelitian dibagi dalam lima alur yaitu pengumpulan data, preprocessing data, proses decision tree dan regresi linear, uji validitas, dan hasil analisis. Metodologi dapat dilihat dibawah ini.



Gambar 1. Metode Penelitian

Pada metode penelitian kali ini terdapat 5 tahapan dimana tahapan pertama yaitu studi literatur, lalu tahapan kedua pengambilan datan tahapan ketiga membentuk model, dimana model disini ialah Decision Tree dan Regresi Linear, tahapan keempat yaitu pengujian dan yang terakhir ialah hasil penelitian

2.1 Pengumpulan Data

Sumber data untuk penelitian ini di dapatkan dari Kaggle. Data yang digunakan dalam penelitian ini sebanyak 569 data dengan 32 atribut. Pengumpulan data bertujuan untuk mengklasifikasikan data-data yang

diperlukan dalam penelitian. Daftar kanker payudara didapat melalui website Kaggle. Pengumpulan data yang digunakan dengan tahapan studi literature. Studi literatur adalah langkah awal dalam tektik pengumpulan data pada suatu penelitian. Metode ini dilakukan dengan cara mengambil sumber referensi di berbagai buku, karya ilmiah maupun jurnal-jurnal.

2.2 Decision Tree

Algoritma Decision Tree adalah salah satu algoritma Data Mining yang sering digunakan sebagai penyelesian untuk klasifikasi suatu masalah. Decision Tree juga merupakan metode klasifikasi dan prediksi yang sangat akurat dan banyak digunakan. Decision Tree juga merupakan struktur seperti flowchart. Decision Tree menggunakan metode supervised machine learning dengan proses pembelajaran dimana data baru dikelompokkan menurut training samples yang sudah ada.

2.3 Regresi Linear

Regresi Linear merupakan sebuah metode data yang melakukan prediksi menggunakan peluasan hubungna sistematis antara variabel, yaitu variabel dependen (y) dengan variabel independen (x) [17]. Variabel dependen merupakan variabel akibat akibat atau variabel yang dipengaruhi, sedangkan variabel independen merupakan variabel sebab atau yang mempengaruhi. Regresi linear menjadi salah satu cara yang digunakan dalam produksi untuk melakukan prediksi kualitas ataupun kuantitas [18].

2.4 Uji Validitas (DBI)

Untuk menguji hasil dari pengelompokan untuk pemeriksaan hasil dari sebuah cluster itu sendiri.

2.4.1 Davies-Bouldin Index

Menurut Permatadevi, et al. (2013) jika proses pengclusteran untuk masing-masing k, maka untuk dapat menentukan jumlah cluster yang terbaik dapat dikemukakan penilaian dengan menggunakan davies-bouldin index. Rencana pengukuran ini bertujuan untuk mengembangkan jarak antara cluster yang satu dengan yang lainnya dan pada masa yang telah dicoba untuk mengecilkan jarak antara objek pada sebuah cluster (Hilmi,et al., 205) [19]

3. HASIL DAN PEMBAHASAN

3.1 Pengambilan Data

Penelitian ini menggunakan dataset penyakit kanker payudara dengan 32 atribut dan 569 record yang bersumber dari situs Kaggle.

Tabel 1. Dataset Kanker Payudara

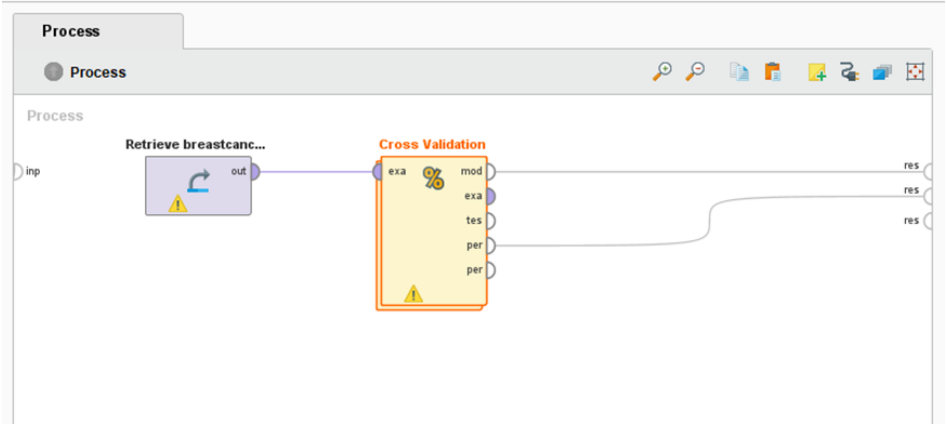
Id	diagnosis	radius_mean	…	concave points_worst	symmetry_worst	fractal_dimension_worst
842302	M	0,777083		1,843056	3,195139	0,825694
…			…			
8912280	M	16.24	…	1,202778	0,192361	0,738194
8912284	B	0,561806	…	0,70625	1,388194	0,07127
92751	B	0,344444	…	0	1,99375	0,07039

3.2 Algoritma Decision Tree

Setelah mendapatkan data, kemudian dilakukan pemodelan algoritma Decision Tree menggunakan tools RapidMiner untuk mengetahui bagaimana cara terbentuknya Decision Tree tersebut. Proses Pengujian:

Percobaan pertama

Percobaan pertama dilakukan menggunakan selection attributes pada RapidMiner pemodelan dibuat menggunakan algoritma klasifikasi yaitu algoritma decision tree.



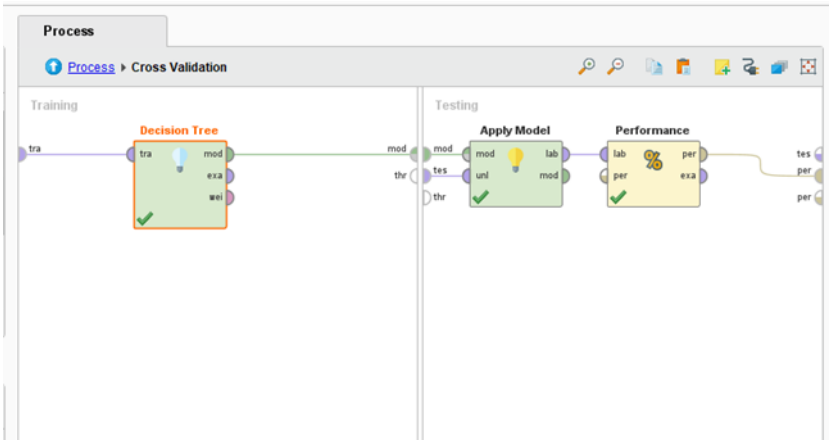
Gambar 2. Model Proses 1

Berikut hasil import menggunakan RapidMiner

Tabel 2. Import Dataset

Row No.	Id	doagnosis	Radius_ mean	Texture_ mean	Primer _ mean	area_mean	smoothness_mean	compactnes_mean	condisional_ mean
1	842302	M	18	10	123	1001	0.118	0.278	0.30
2	842517	M	21	18	133	1326	0.085	0.079	0.08
3	84300903	M	20	21	130	1203	0.110	0.150	0.19
4	84348301	M	11	20	78	386	0.142	0.284	0.24
5	84358402	M	20	14	135	1297	0.100	0.133	0.19
6	847386	M	12	16	83	477	0.128	0.170	0.15
7	844359	M	18	20	120	1040	0.095	0.109	0.11
8	84458202	M	14	21	90	578	0.119	0.165	0.09
9	844981	M	13	22	88	520	0.127	0.193	0.18
10	84501001	M	12	24	84	476	0.119	0.240	0.22
11	845636	M	16	23	103	798	0.082	0.067	0.03

Pada penelitian tersebut dilakukannya sebuah proses validasi dengan menggunakan cross validation. Berikut gambar 3 pemodelan yang didapat pada cross validation.



Gambar 3. Model Cross Validation Dengan Decision Tree

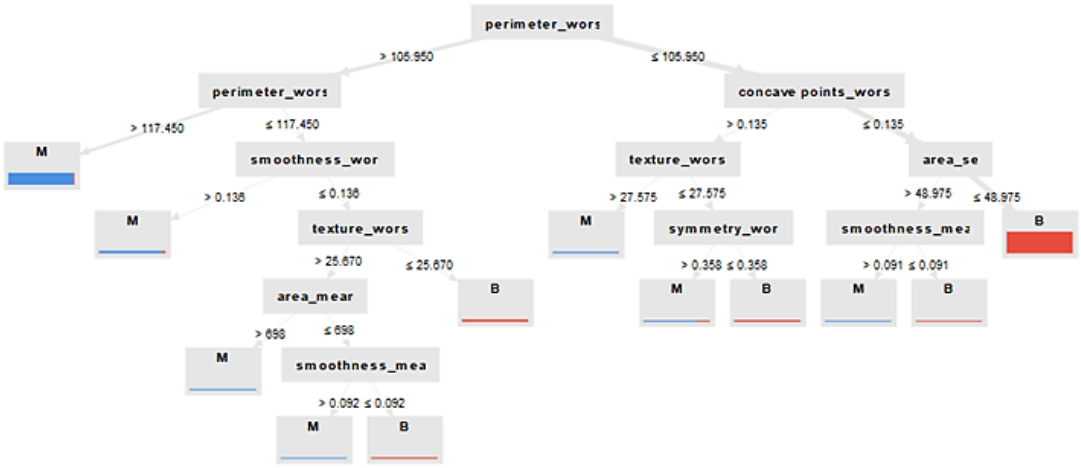
Pada pemodelan cross validation didalamnya terdapat dua bagian, yaitu pertama bagian training yang kedua testing. Bagian training menggunakan algoritma klasifikasi *decision tree* dan bagian testing menggunakan *Apply Model* untuk menampilkan *description tree*, yang digunakan untuk menampilkan hasil akurasi. Berikut Gambar 4 yang menunjukkan *description decision tree*.

PerformanceVector (Performance) | Tree (Decision Tree)

Tree

```
perimeter_worst > 105.950
|   perimeter_worst > 117.450: M {M=165, B=2}
|   perimeter_worst ≤ 117.450
|   |   smoothness_worst > 0.136: M {M=22, B=1}
|   |   smoothness_worst ≤ 0.136
|   |   |   texture_worst > 25.670
|   |   |   |   area_mean > 698: M {M=6, B=0}
|   |   |   |   area_mean ≤ 698
|   |   |   |   |   smoothness_mean > 0.092: M {M=2, B=0}
|   |   |   |   |   smoothness_mean ≤ 0.092: B {M=0, B=6}
|   |   |   texture_worst ≤ 25.670: B {M=0, B=20}
perimeter_worst ≤ 105.950
|   concave points_worst > 0.135
|   |   texture_worst > 27.575: M {M=9, B=0}
|   |   texture_worst ≤ 27.575
|   |   |   symmetry_worst > 0.358: M {M=4, B=1}
|   |   |   symmetry_worst ≤ 0.358: B {M=0, B=11}
|   concave points_worst ≤ 0.135
|   |   area_se > 48.975
|   |   |   smoothness_mean > 0.091: M {M=2, B=0}
|   |   |   smoothness_mean ≤ 0.091: B {M=0, B=2}
|   |   area_se ≤ 48.975: B {M=2, B=314}
```

Gambar 4. Description Decision Tree



Gambar 5. Hasil Decision Tree

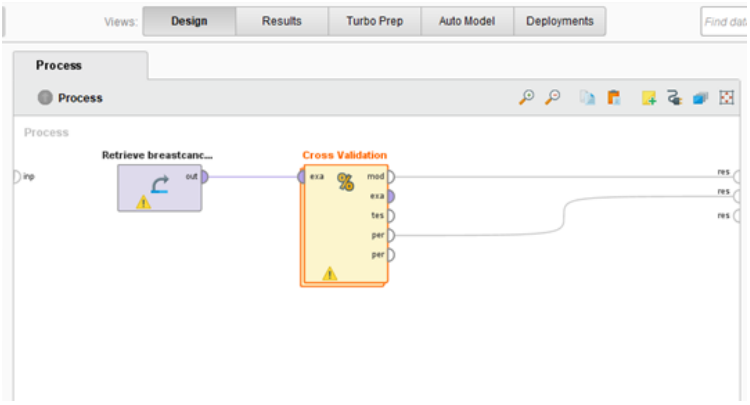
Hasil accuracy menggunakan Algoritma *decision tree* dalam pengujian pertama dengan *Cross Validation* yaitu sebesar 93,51%. Dapat dilihat dari Gambar 5 dengan nilai micro average: 93.50%.

Tabel 3. Hasil Accuracy Decision Tree

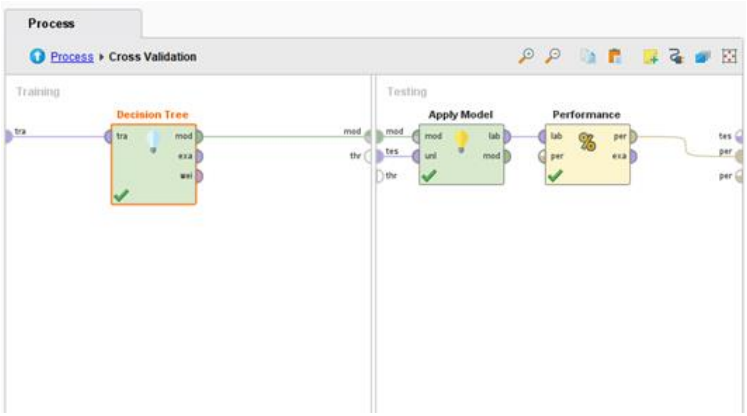
	True M	True B	Class Precision
Pred. M	194	19	91.08%
Pred. B	18	338	94.94%
Class Recall	91.51%	94.68%	

Percobaan Kedua

Percobaan kedua dilakukan menggunakan *select attributes* pada Rapid Miner. Pemodelan menggunakan algoritma klasifikasi yaitu algoritma Regresi Linear ditunujukkan pada Gambar 6 dan 7.



Gambar 6. Proses Cross Validation



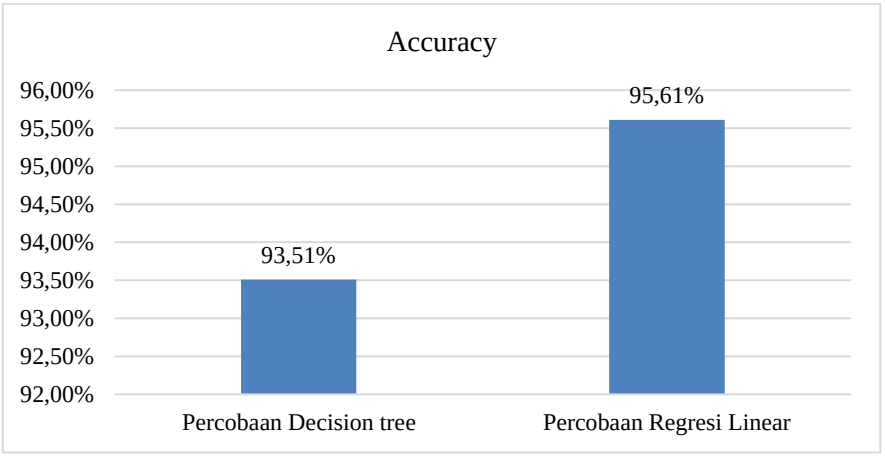
Gambar 7. Hasil Poses Model

Hasil accuracy menggunakan algoritma regresi linear menggunakan RapidMiner yaitu sebesar 95.61%, dengan micro avarage:95.61%.

Tabel 4. Hasil Accuracy Regresi Linear

	True M	True B	Class Precision
Pred. M	189	2	98.95%
Pred. B	23	355	93.92%
Class Recall	89.15%	99.44%	

Berikut gambar 9 merupakan perbedaan akurasi antara algoritma Decision Tree dan Regresi Linear. Disini terlihat bahwa tingkat akurasi Regresi Linear lebih tinggi dibanding tingkat akurasi Decision Tree



Gambar 8. Hasil Percobaan Keseluruhan

4. KESIMPULAN

Dari hasil percobaan menggunakan tools rapidminer yang peneliti lakukan menggunakan algoritma Decision tree dan regresi linear didapatkan bahwa hasil akurasi regresi linear lebih tinggi daripada algoritma decision tree dengan perbandingan 95,61% : 93,51%. Dari hasil akurasi kedua algoritma tersebut menunjukkan bahwa hasil klasifikasinya baik, sehingga dapat diprediksi pasien yang didiagnosis kanker payudara ganas dan pasien yang didiagnosis kanker payudara jinak.

REFERENSI

[1] Jody Alwin irawadi and S. Sunendiari, "Penerapan dan Perbandingan Tiga Metode Analisis Pohon Keputusan pada Klasifikasi Penderita Kanker Payudara," *Jurnal Riset Statistika*, vol. 1, no. 1, pp. 19–27, Jul. 2021, doi: 10.29313/jrs.v1i1.22.

[2] F. S. Nugraha, M. J. Shidiq, and S. Rahayu, "ANALISIS ALGORITMA KLASIFIKASI NEURAL NETWORK UNTUK DIAGNOSIS PENYAKIT KANKER PAYUDARA," *Jurnal Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 149–156, Aug. 2019, doi: 10.33480/pilar.v15i2.601.

[3] D. Cahyanti, A. Rahmayani, and S. Ainy Husniar, "Indonesian Journal of Data and Science Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara," vol. 1, no. 2, pp. 39–43, 2020.

[4] A. Maulana, A. Nugroho, and I. Romli, "Optimalisasi Support Vector Machine Menggunakan Particle Swarm Optimization Untuk Mendiagnosa Penyakit Kanker Payudara," 2021.

[5] H. Susanto SMK Negeri, "DATA MINING UNTUK MEMPREDIKSI PRESTASI SISWA BERDASARKAN SOSIAL EKONOMI, MOTIVASI, KEDISIPLINAN DAN PRESTASI MASA LALU DATA MINING TO PREDICT STUDENT'S ACHIEVEMENT BASED ON SOCIO-ECONOMIC, MOTIVATION, DISCIPLINE AND ACHIEVEMENT OF THE PAST," 2014.

[6] C. C. Aggarwal, *Data Mining*. Cham: Springer International Publishing, 2015. doi: 10.1007/978-3-319-14142-8.

[7] Z. Azmi, "DECISION TREE BERBASIS ALGORITMA UNTUK PENGAMBILAN KEPUTUSAN".

[8] I. Sutoyo, "IMPLEMENTASI ALGORITMA DECISION TREE UNTUK KLASIFIKASI DATA PESERTA DIDIK," vol. 14, no. 2, 2018, [Online]. Available: www.bsi.ac.id

[9] Y. Arie Wijaya *et al.*, "Analisa Klasifikasi menggunakan Algoritma Decision Tree pada Data Log Firewall", [Online]. Available: https://ejournal.stmikgici.ac.id/

[10] P. Meilina, "PENERAPAN DATA MINING DENGAN METODE KALSIFIKASI MENGGUNAKAN DECISION TREE DAN REGRESI," 2015.

[11] J. P. Matematika, D. Matematika, T. N. Padilah, and R. I. Adam, "ANALISIS REGRESI LINIER BERGANDA DALAM ESTIMASI PRODUKTIVITAS TANAMAN PADI DI KABUPATEN KARAWANG".

[12] R. Dwi Shaputra and S. Hidayat, "Implementasi regresi linier untuk prediksi penjualan dan cash flow pada aplikasi point of sales restoran."

[13] Y. Sulistyo Nugroho, "PENERAPAN ALGORITMA C4.5 UNTUK KLASIFIKASI PREDIKAT KELULUSAN MAHASISWA FAKULTAS KOMUNIKASI DAN INFORMATIKA UNIVERSITAS MUHAMMADIYAH SURAKARTA," 2014.

[14] S. Wahyuningsih and D. Retno Utari, "Konferensi Nasional Sistem Informasi 2018 STMIK Atma Luhur Pangkalpinang," 2018.

[15] M. Amin, F. N. Ridho, H. Hasanah, and I. Oktaviani, "Prediksi Area Kebakaran Hutan dengan Temperatur Menggunakan Regresi Linear."

[16] N. Nafi'iyah, "PENERAPAN REGRESI LINEAR DALAM MEMPREDIKSI HARGA JUAL MOBIL BEKAS."

[17] G. Najla, A. #1, and D. Fitrianah, "Penerapan Metode Regresi Linear Untuk Prediksi Penjualan Properti pada PT XYZ," *Jurnal Telematika*, vol. 14, no. 2.