Indonesian Journal of Data and Science (IJODAS) ISSN: 2715-9930
Vol 3, No 2, July 2022, pp. 82-88

Analisis perbandingan Reduction Technique dengan metode Dimentional Reduction dan Cross Validation pada dataset breast cancer

Yudha Islami Sulistya [a,1], Chyquitha Danuputri [b,2]

[a] *Universitas Muslim Indonesia, Jl. Urip Sumoharjo KM.05, Makassar dan 90231, Indonesia*
[b]*Universitas Bunda Mulia, Jl. Lodan Raya No.12, RT.12/RW.2, Ancol, Kec. Pademangan, Kota Jkt Utara, Daerah Khusus Ibukota Jakarta dan 14430, Indonesia*
[1]*yudhaislamisulistya@mail.ugm.ac.id;*[2]*chyquitha.vivaldy@gmail.com;*

INFORMASI ARTIKEL

Diterima : 17 – 05 – 2022
Direvisi : 19 – 06 – 2022
Diterbitkan : 31 – 07 – 2022

Kata Kunci:
Machine Learning
Dimentional Reduction
Cross Validation
Reduction Technique
Breast Cancer

ABSTRAK

Machine learning (ML) merupakan bidang ilmu yang memungkinkan komputer dalam mengembangkan sebuah sistem yang dapat belajar dari data. Dalam ML sendiri banyak teknik sangat berperan penting dalam pengembangan machine ML salah satunya adalah teknik reduksi yang dimana membuat sistem lebih baik dari data yang telah di reduksi. Penelitian ini bertujuan membandingkan performa teknik reduksi dengan metode dimentional reduction dan cross validation pada dataset breast cancer. Dimentional reduction merupakan teknik yang menyederhanakan feature atau mengurangi dimensi pada dataset sedangkan cross validation merupakan metode yang digunakan untuk memaksimalkan hasil dari prediksi pada suatu model. Setalah melakukan tahapan-tahapn dalam pengujian dengan dimentional reduction dan cross validation menggunakan algoritma K-Nearest Neighbors dengan dataset breast cancer berjumlah 500. Hasil yang diperolah untuk dimentional reduction akurasi rata-rata pada model 95.2%, sedangkan pada cross validation 96.6%.

I. Pendahuluan

Machine learning (ML) mesin adalah bidang studi dari cabang ilmu yang menggabungkan ide-ide dari beberapa cabang ilmu seperti kecerdasan buatan, statistik, teori informasi, matematika, dll [1]–[6]. Pembelajaran mesin biasanya berfokus pada teori, kinerja, dan sifat sistem dan algoritma pembelajaran. Memecahkan masalah mulai dari robotika hingga pengenalan sistem, penambangan data, dan sistem kontrol otomatis juga dapat diselesaikan melalui pembelajaran mesin. Pembelajaran mesin mandiri ini berguna dalam memperbaiki masalah yang sulit diselesaikan manusia, dengan menyederhanakan masalah yang sulit.

Salah satu bagian ML yang paling ketat adalah pembelajaran mesin yang diawasi. Ini adalah algoritma yang mempelajari fungsi yang bergantung pada data input berlabel untuk menghasilkan output yang sesuai dengan data baru yang tidak berlabel. Bagian ini memiliki dua bagian penting dalam penerapannya, yang pertama ialah klasifikasi dan regresi. klasifikasi bagian dari supervised ML yang menentukan output pada machine learning berupa *binary classification* maupun multi-class classification dan kemudian hasil dalam ML sendiri disebut dengan model. ML terkhusus pada supervised learning memiliki beberapa algoritma yang sangat terkenal dalam proses klasifikasi serta penerapannya antara lain KNN, *Naïve Bayes*, SVM, *Random Forest* dan *Decision Tree*. Suatu model dalam ML bergantung pada tingkat akurasi atau yang lebih dikenal dengan score. Sedangkan ada beberapa hal yang sangat berpengaruh pada tingkat akurasi pada suatu model, mulai dari jumlah data hingga teknik reduksi pada model.

Teknik reduksi pada model memiliki beberapa cara umum mulai dari *regularization*, dimensionality *reduction, feature selelection, cross validation, early stopping*.

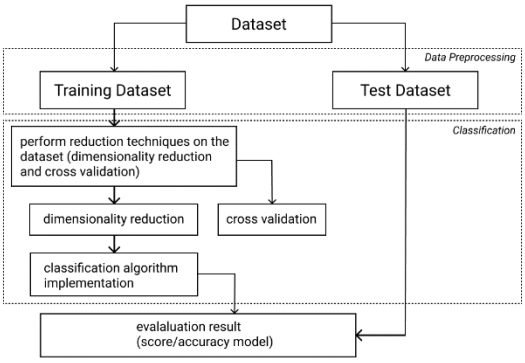
Berdasarkan latar belakang yang ada, maka pada penelitian ini akan dilakukan perbandingan teknik reduksi dengan metode *dimensionality reduction* dan *cross validation* pada dataset *breast cancer*[7] atau kanker payudara. Dataset kanker payudara merupakan salah satu dari banyak dataset yang disediakan oleh pustaka scikit-learn dimana memiliki beberapa karakteristik antara lain jumlah kelas yakni 2, yang masing-masing

kelas malignant (tumor ganas) berjumlah 212 dan benign berjumlah 357 (tumor jinak), jumlah data 569, jumlah dimensi 30 serta karakteristik pada feature yakni bilangan real dan positif.

Hasil dari penelitian ini adalah sebuah analisis perbandingan teknik reduksi dedngan metode *dimensionality reduction* dan *cross validation* pada dataset *breast cancer* atau kanker payudara dengan menerapakan beberapa algoritma klasifikasi mulai dari KNN, *Naïve Bayes*, SVM, Random Forest dan *Decision Tree*[8] dalam mengklasifikasikan penyakit kanker payudara antara lain tumor ganas maupun tumor jinak.:

II. Metode

Penelitian ini bertujuan untuk menerapkan teknik reduksi yakni *dimensionality reduction* dan cross validation pada dataset kanker payudara menggunakan bebearapa algoritma dalam klasifikasi.



Gambar 1. *Experimental framework*

A. *Experimental framework*

Seperti yang digambarkan pada Gambar 1, kerangka *experimental* penelitian menggunakan dataset yang dibagi kedalam set pelatihan dan set pengujian, proses pembagian dataset dibagi berdasarkan perbandingan 8:2 atau 80% untuk set pelatiahan dan 20% untuk set pengujian. Setiap dari dataset ini melalui proses data preprocessing yang dinamakan teknik reduksi antara lain *dimensionality reduction* dan cross validation[9].

Teknik reduksi diimplementasikan untuk mencari skor pada model dengan mereduksi menggunakan *dimensionality reduction* dan cross validation. Pada pengujiannya menerapkan algoritma klasifikasi antara lain KNN, Naïve Bayes, SVM, Random Forest dan Decision Tree. Hasil yang kemudian akan didapatkan akan di evaluasi[10]–[12].

B. *Dataset*

Dataset yang digunakan dalam penelitian ini adalah kumpulan data yang telah disediakan oleh pustaka scikit-learn. Deskripsi singkat tentang karakteristik dan kumpulan data masing-masingg dijadikan pada Tabel 1 dan Tabel 2.

Tabel 1. *Description dataset*

Description	Characteristic
Classes	2
Sample per class	212(m), 357(b)
Sample total	569
Dimensionality	30
Features	real, positive

Tabel 2. *Example dataset* (6 *out of* 569)

No	Mean radius	…	Worst fractal dimension	Label
1	17.99		0.11890	0
2	20.57		0.08902	0
3	19.69		0.08758	0
…				
4	16.60		0.07820	0
5	20.60		0.12400	0
6	7.76		0.07039	1

C. Dimensionality reduction

Secara umum, *Principal Component Analysis* (PCA) adalah metode *dimensionality reduction* yang bervariasi di beberapa variabel, kompresi data, pengenalan pola, dan analisis statistik. PCA sendiri memiliki fungsi untuk memperkecil ukuran variabel input menjadi komponen utama yang berukuran lebih kecil dengan meminimalkan kehilangan informasi tetapi menjaga variabilitas data, dimana komponen utama yang terbentuk tidak saling berkorelasi. Algoritma analisis komponen utama. PCA akan membentuk satu set dimensi baru, yang kemudian akan diberi peringkat berdasarkan varians data. PCA akan membangkitkan komponen utama yang diperoleh dari dekomposisi nilai eigen dan nilai eigen matriks kovarians. Adapun langkah-langkah algoritma PCA dalam *dimensionality reduction* adalah sebagai berikut:

1) Menghitung *mean* ($\bar{x}$) dari setiap data pada tiap dimensi menggunakan persamaan (1):

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} X_i \quad (1)$$

Keterangan:
n = jumlah data sampel

X_i = data sampel

2) Menghitung covariance matrix (C_x) menggunakan persamaan (2):

$$C_x = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T \quad (2)$$

Keterangan:
n = jumlah data sampel
X_i = data sampel

$\bar{X}$ = mean

3) Menghitung eigenvector (v_m) dan eigenvalue (λ_m) dari covariance matrix menggunakan persamaan (3):

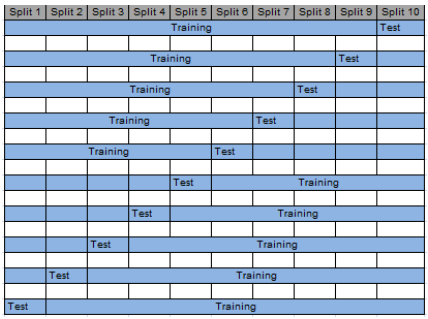
$$C_X v_m = \lambda_m v_m \quad (3)$$

4) Urutan eigenvalue secara descending. PC adalah deretan eigenvector yang sesuai dengan urutan eigenvalue pada tahap 3.

5) Menghasilkan dataset baru dengan beberapa dimensi yang lebih sederhana sesuai dengan kebutuhan.

D. Cross Validation

Cross validartion adalah teknik validasi dengan membagi data secara acak menjadi k bagian dan setiap bagian melalui proses klasifikasi. Cross validation, pengujian k akan dilakukan. Data yang digunakan dalam pengujian ini adalah data latih untuk mencari tingkat kesalahan secara keseluruhan. Secara umum pengujian nilai kfold atau k yakni 10 dan dalam penelitian ini akan menerapkan hal yang demikian. setiap pengujian akan menggunakan data uji dan bagian k1 akan menjadi data latih, kemudian data uji tersebut akan ditukar dengan data latih sehingga untuk setiap pengujian akan diperoleh data uji yang berbeda. Data latih adalah data yang akan digunakan dalam proses pembelajaran sedangkan data uji adalah data yang belum pernah digunakan untuk pembelajaran dan akan berfungsi sebagai data untuk memeriksa kebenaran atau keakuratan hasil pembelajaran . *Experimental framework* ditunjukkan pada Gambar 2.



Gambar 2. *Experimental framework*

E. Classification Algorithm

Menggunakan beberapa algoritma dalam menganalisis perbandingan teknik reduksi dengan menggunakan metode dimentionality reduction dan cross validation guna membandingkan hasil terbaik dari dua teknik reduksi dengan akurasi optimum dari algoritma machine learning khusus klasifikasi. Beberapa algoritma yang digunakan dalam penelitian ini yaitu KNN, Naïve Bayes, SVM, Random Forest Classifier, dan Decision Tree[13]–[18].

1) K-Nearest Neighbors (KNN)

KNN adalah salah satu metode argumentasi kelas tertua. Pengambilan keputusannya sangat sederhana, yaitu sampel yang akan diuji sama dengan kategori sampel terdekat. Jika set pelatihan dan metrik jarak tidak berubah, hasil dari keputusan aturan tetangga terdekat akan ditentukan secara unik untuk setiap sampel yang ditanamkan. Untuk semua sampel dalam himpunan E, jika y adalah contoh tetangga terdekat dari x, maka tipe y adalah hasil keputusannya, yaitu aturan tetangga terdekat. Dengan asumsi X adalah sampel dari kategori yang tidak diketahui, proses keputusan tertentu adalah pada persamaan (4)[19]

$$g_j = \min g_j(x)$$
$$i = 1,2, \dots C \tag{4}$$

2) Gaussian Naïve Bayes

Persamaan Gaussian naïve bayes dapat dilihat pada persamaan (5)[20]

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2} \tag{5}$$

3) Support Vector Machine (SVM)

SVM tidak terlalu mengandalkan pengalaman belajar dan memiliki struktur yang lebih fleksibel. Mengoptimalkan pemecahan masalah dengan SVM seperti adalah persamaan [21]

$$\min(w, b) = \left\{\frac{1}{2}\|w\|^2 + c \sum_i \check{c}_i\right\} \tag{6}$$

4) Random Forest Classifier

Prediksi pola tak kasat mata dapat dibuat dengan merata-ratakan prediksi dari setiap pohon klasifikasi seperti pada persamaan [6]

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x') \tag{7}$$

5) Decision Free

Pohon keputusan atau decision tree adalah salah satu teknik yang digunakan untuk mengklasifikasikan sekumpulan objek. Teknik ini dilatih pada himpunan simpul keputusan yang dihubungkan oleh cabang, turun dari simpul akar dan berakhir pada simpul terjauh (simpul daun)[22]. Adapun langkah-langkah derajat impurity secara kuantitatif adalah pada persamaan (8)(9)(10)(11)[21]

$$Gini : H(Q_m) = \sum_k P_{mk}(1 - p_{mk}) \tag{8}$$

$$Entropy : H(Q_m) = -\sum_k p_{mk} \log(p_{mk}) \quad (9)$$

$$Misclassification : H(Q_m) = 1 - \max(p_{mk}) \quad (10)$$

$$GAIN_{split} = E(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} E(i) \right) \quad (11)$$

III. Hasil dan Pembahasan

Berdasarkan teori dan metode di atas, implementasi dilakukan dalam hal menerapkan teknik reduksi dengan metode dimentionality reduction dan cross validation pada dataset kanker payudara. Penelitian ini menggunakan metode penelitian eksperimen, yakni dengan melibatkan dataset kanker payudara, beberapa algoritma dan salah dua dari teknik reduksi.

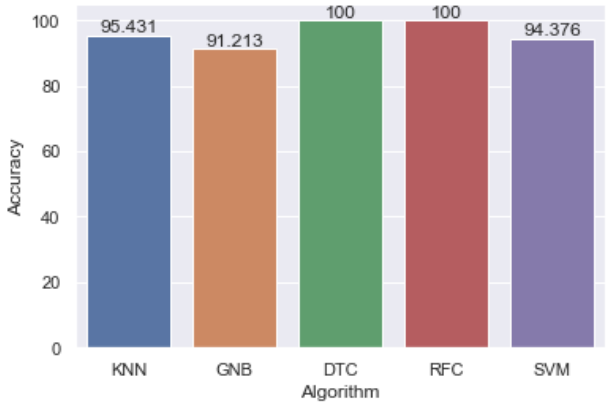
A. Pengujian Metode Dimentionaly

Berikut adalah hasil pengujian dimentionaly reduction menggunakan beberapa algoritma dalam klasifikasi antara lain KNN, Naïve Bayes, SVM, Random Forest dan Decision Tree.

Hasil dan pembahasan berisi tentang hasil akhir, output program dan analisis metode dalam penelitian ini ditunjukkan pada Tabel 3

Tabel 3. *result of dimension reduction using* PCA

PC1	PC2	Label
9.192837	1.948583	0
2.387802	-3.768172	0
5.733896	-1.075174	0
…	…	…
1.256179	-1.902297	0

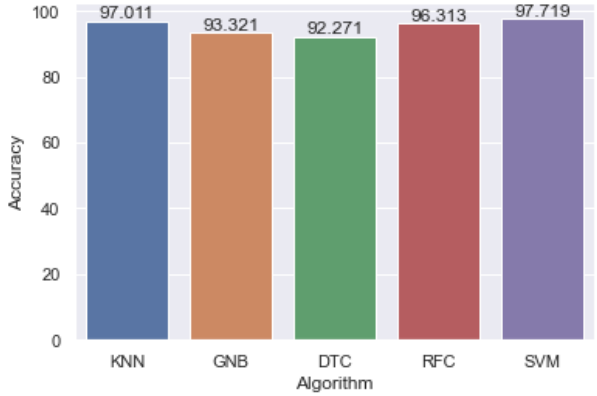


Gambar 3. *Result of score model with dimentionaly reduction*

B. Pengujian Metode Cross Validation

Berikut adalah hasil pengujian cross validation menggunakan beberapa algoritma dalam klasifikasi antara lain KNN, Naïve Bayes, SVM, Random Forest dan Decision Tree.

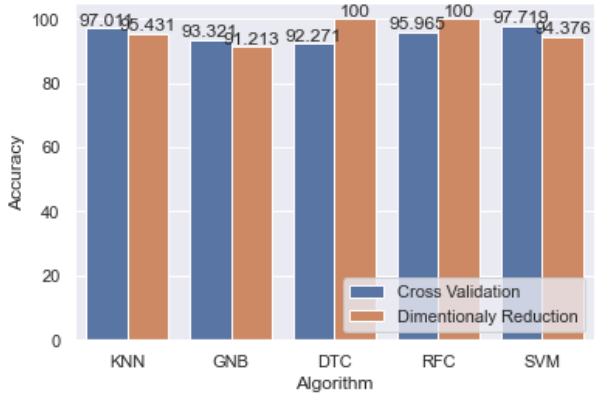
Hasil dan pembahasan berisi tentang hasil akhir, output program dan analisis metode dalam penelitian ini



Gambar 4. *Result of score model with cross validation*

C. Perbandingan Metode Teknik Reduksi

Dari hasil pengujian metode dimentionality reduction dan cross validation pada dataset kanker payudara yang diperoleh pada Gambar 3 dan Gambar 4 dengan masing-masing nilai perbandingan akurasi dapat dilihat pada Gambar 5.



Gambar 5. *Result compare score model both dimentionaly reduction and cross validation*

IV. Kesimpulan

Dalam analisiS perbandingan teknik reduksi model dengan metode dimentionaly reduction dan cross validation dengan dataset kanker payudara. Dalam penelitian in membandingkan suatu metode pada teknik reduksi untuk mengklasifikasikan tumor ganas dan tumor jinak pada dataset kanker payudara. Dari percobaan yang dilakukan dengan dimentionaly reduction menggunakan algoritma KNN sebesar 95.43%, GNB sebester 91.23%, SVM sebesar 94,37%, Random Forest sebesar 100% dan Decision Tree sebesar 100% sedangkan untuk cross validation menggunakan algoritma KNN sebesar 97.01%, GNB sebester 93.32%, SVM sebesar 97,71%, Random Forest sebesar 96.31% dan Decision Tree sebesar 92.27%.

Daftar Pustaka

[1] M. M. Baharuddin, T. Hasanuddin, and H. Azis, "Analisis Performa Metode K-Nearest Neighbor untuk Identifikasi Jenis Kaca," *Ilk. J. Ilm.*, vol. 11, no. 28, pp. 269–274, 2019.

[2] H. Zhang, "The optimality of Naive Bayes," *Proc. Seventeenth Int. Florida Artif. Intell. Res. Soc. Conf. FLAIRS 2004*, vol. 2, pp. 562–567, 2004.

[3] M. B. Bejiga, A. Zeggada, A. Nouffidj, and F. Melgani, "A convolutional neural network approach for assisting avalanche search and rescue operations with UAV imagery," *Remote Sens.*, vol. 9, no. 2, 2017, doi: 10.3390/rs9020100.

[4] J. D. Kelleher, B. Mac Namee, and A. D. Arcy, *Fundamentals of Machine Learning For Predictive Data Analytics Algorithms, Worked Examples, and Case Studies*. London: The MIT Press, 2015.

[5] Rizky Ade Putranto, Triastiti Wuryandari, and Sudarno, "Perbandingan Analisis Klasifikasi Antara

Decision Tree Dan Support Vector Machine Multiclass Untuk Penentuan Jurusan Pada Siswa Sma," *J. Gaussian*, vol. 4, no. 4, pp. 1007–1016, 2015.

[6] I. F. Nurahmadan, A. Agusta, P. A. Winarno, and B. H. Sazali, "Perbandingan Algoritma Machine Learning Untuk Klasifikasi Denyut Jantung Janin," no. April, pp. 733–740, 2021.

[7] D. Cahyanti, A. Rahmayani, and S. Ainy, "Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 39–43, 2020.

[8] A. Primajaya, B. N. Sari, and A. Khusaeri, "Prediksi Potensi Kebakaran Hutan dengan Algoritma Klasifikasi C4.5 Studi Kasus Provinsi Kalimantan Barat," *J. Edukasi dan Penelit. Inform.*, vol. 6, no. 2, p. 188, 2020, doi: 10.26418/jp.v6i2.37834.

[9] W. Purnami, A. M. Regresi, and L. Ordinal, "Perbandingan Klasifikasi Tingkat Keganasan Breast Cancer Dengan Menggunakan Regresi Logistik Ordinal Dan Support Vector Machine (SVM)," *J. Sains Dan Seni Its*, vol. 1, no. 1, 2012.

[10] H. Azis, F. Tangguh Admojo, and E. Susanti, "Analisis Perbandingan Performa Metode Klasifikasi pada Dataset Multiclass Citra Busur Panah," *Techno.Com*, vol. 19, no. 3, pp. 286–294, 2020.

[11] N. Fadhillah, H. Azis, and D. Lantara, "Validasi Pencarian Kata Kunci Menggunakan Algoritma Levenshtein Distance Berdasarkan Metode Approximate String Matching," *Pros. Semin. Nas. Ilmu Komput. dan Teknol. Inf*, vol. 1, pp. 3–7.

[12] F. Muharram, H. Azis, and A. R. Manga, "Analisis Algoritma pada Proses Enkripsi dan Dekripsi File Menggunakan Advanced Encryption Standard (AES)," *Pros. Semin. Nas. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 2, pp. 112–115, 2018.

[13] A. Prasetya Wibawa, W. Lestar, A. Bella Putra Utama, I. Tri Saputra, and Z. Nabila Izdihar, "Multilayer Perceptron untuk Prediksi Sessions pada Sebuah Website Journal Elektronik," *Indones. J. Data Sci.*, vol. 1, no. 3, pp. 57–67, 2020, doi: 10.33096/ijodas.v1i3.15.

[14] K. Ilunga, T. Lumbala, and K. Mulenda, "Application of Big data to configuration management in a PLM context," *Indones. J. Data Sci.*, vol. 3, no. 1, pp. 9–16, 2022.

[15] S. Sari, U. Khaira, P. Pradita, and T. S. Tri, "… Beauty Shaming Di Media Sosial Twitter Menggunakan Algoritma SentiStrength: Sentiment Analysis Against Beauty Shaming Comments on Twitter Social Media …," *Indones. J. …*, vol. 1, no. 1, pp. 71–78, 2021.

[16] Hasran, "Klasifikasi Penyakit Jantung Menggunakan Metode K-Nearest Neighbor," *Indones. J. Data Sci.*, vol. 1, no. 1, pp. 1–4, 2020.

[17] N. Rokhman and J. Maharanti, "Deteksi Steganografi Berbasis Least Significant Bit (LSB) Dengan Menggunakan Analisis Statistik," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 5, no. 2, pp. 57–62, 2013, doi: 10.22146/ijccs.2007.

[18] H. Nursan and Muslim, "Penerapan Metode Digital Watermarking dan Privilege pada Dokumen Skripsi," *Indones. J. Data Sci.*, vol. 1, no. 1, pp. 19–22, 2020.

[19] W. Xing and Y. Bei, "Medical Health Big Data Classification Based on KNN Classification Algorithm," *IEEE Access*, vol. 8, pp. 28808–28819, 2020, doi: 10.1109/ACCESS.2019.2955754.

[20] M. S. Vural and M. Gök, "Criminal prediction using Naive Bayes theory," *Neural Comput. Appl.*, vol. 28, no. 9, pp. 2581–2592, 2017, doi: 10.1007/s00521-016-2205-z.

[21] N. Nurajijah and D. Riana, "Algoritma Naïve Bayes, Decision Tree, dan SVM untuk Klasifikasi Persetujuan Pembiayaan Nasabah Koperasi Syariah," *J. Teknol. dan Sist. Komput.*, vol. 7, no. 2, pp. 77–82, 2019, doi: 10.14710/jtsiskom.7.2.2019.77-82.

[22] S. C. Esananda, B. Nugroho, F. T. Anggraeny, P. S. Informatika, F. I. Komputer, and U. P. Nasional, "Penerapan Algoritma Decision Tree Dalam Menentukan Prestasi Akademik Siswa," vol. 02, no. 2, 2021.