**JIK: Jurnal Ilmu Komputer**
Vol. 09, No. 2, Bulan Tahun, pp. 50~56 p-ISSN: 2527-9653
DOI: 10.47007/komp.v7i01.xxxxx e-ISSN: 2797-8389

# ANALISIS PERBANDINGAN KINERJA MODEL MACHINE LEARNING DALAM PREDIKSI DIABETES STUDI KASUS DATASET KAGGLE 2022

**Nicholas Abigail Kosasih[1], Nicholas Syahputra[2], Justyn Anthonyo[3], Hengki Wijaya[4], Harianto Kornelius[5], Dewi Nasien[6]**

[1,2,3,4,5,6]Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Institut Bisnis dan Teknologi Pelita Indonesia
Email: [1]nicholas.abigail@student.pelitaindonesia.ac.id, [2]nicholas@student.pelitaindonesia.ac.id, [3]justyn@student.pelitaindonesia.ac.id, [4]hengki@student.pelitaindonesia.ac.id, [5]harianto@student.pelitaindonesia.ac.id, [6]dewinasien@lecturer.pelitaindonesia.ac.id

**Abstrak**

Diabetes merupakan salah satu penyakit kronis yang ditandai oleh kadar glukosa darah yang tinggi di atas batas normal akibat gangguan produksi atau fungsi insulin. Faktor penyebab munculnya diabetes dapat melalui gaya hidup tidak sehat, obesitas, kurangnya aktivitas fisik, serta faktor genetik. Pada penelitian ini akan dilakukan pengujian terhadap dataset medis pasien diabetes yang berasal dari *Kaggle*, untuk menguji metode machine learning terhadap dataset dengan fitur Analisis Komponen Utama *(Principal Component Analysis – PCA).* Dengan dataset dan fitur ini akan digunakan empat metode, yaitu *Support Vector Machine (SVM)*, *Logistic Regression*, *Naïve Bayes*, dan *K-Nearest Neighbor*. Pengujian ini akan dilakukan dengan pembagian data 70:30. Hasil penelitian ini ditujukan untuk melakukan perbandingan penggunaan metode klasifikasi data terhadap dataset awal dengan dataset yang menggunakan fitur *PCA*.

**Kata kunci**: *diabetes, PCA, SVM, regresi logistik, naïve bayes, KNN*

# *COMPARATIVE ANALYSIS OF MACHINE LEARNING MODEL PERFORMANCE IN DIABETES PREDICTION CASE STUDY OF KAGGLE DATASET 2022*

***Abstract***

*Diabetes is a chronic disease characterized by elevated blood glucose levels above normal limits due to impaired insulin production or function. The causes of diabetes can include unhealthy lifestyles, obesity, lack of physical activity, and genetic factors. This study aims to evaluate a medical dataset of diabetes patients obtained from Kaggle to test machine learning methods on the dataset using Principal Component Analysis (PCA) as a feature. Four methods will be used: Support Vector Machine (SVM), Logistic Regression, Naïve Bayes, and K-Nearest Neighbor. The testing will be conducted with a 70:30 data split. The results of this study are intended to compare the performance of classification methods on the original dataset with the dataset utilizing PCA features.*

**Keywords**: *diabetes, PCA, SVM, logistic regression, naïve bayes, KNN*

## 1. *INTRODUCTION*

Diabetes yaitu penyakit gangguan metabolisme kronis yang ditandai peningkatan glukosa darah (Hiperglikemi), disebabkan ketidakseimbangan suplai kebutuhan untuk memfasilitasi masuknya glukosa dalam sel agar dapat digunakan untuk metabolisme dan pertumbuhan sel, sementara sel menjadi kekurangan glukosa yang sangat dibutuhkan dalam kelangsungan dan fungsi sel. [1]. Diperkirakan 578,4 juta penduduk dengan diabetes pada tahun 2030 dibandingkan 463 juta di tahun 2019 dan tahun 2045 jumlahnya akan meningkat menjadi 700,2 juta [2].

Trend *machine learning* di dunia kesehatan, khususnya pada perangkat medis dan sensor untuk memanfaatkan data dalam evaluasi pasien semakin meningkat [3]. *Machine learning* adalah cabang dari kecerdasan buatan *(Artificial Intelligence)* yang berfokus pada pengembangan sistem dan algoritma yang memungkinkan komputer untuk belajar dari data yang telah disediakan, mengidentifikasi pola, dan mengambil keputusan atau melakukan tugas tanpa perlu diprogram secara eksplisit [4]. Tujuan utama dari *machine learning* adalah membuat model prediktif atau deskriptif berdasarkan data yang diberikan, sehingga komputer dapat mengerti atau belajar dari data tersebut dan melakukan tugas yang

diberikan dengan lebih baik seiring berjalannya waktu [5].

Dengan menggunakan fitur Analisa Komponen Utama *(Principal Component Analysis – PCA)*, maka dimensi pada penginputan dataset dapat dikurangi dan mengeluarkan data terpenting untuk memprediksi diabetes [6]. Bantuan dari *PCA* akan dapat menarik faktor tersembunyi dan mengidentifikasi serta mengekstraknya dari berbagai variabel kasar [7].

Dengan menerapkan klasifikasi pada model *machine learning* dan memanfaatkan data medis yang telah terkumpul dalam dataset maka proses pengelompokkan data penyakit diabetes akan semakin efisien [8]. Tujuan dari melakukan klasifikasi dari *machine learning* adalah untuk mengelompokkan setiap variabel menjadi satu kelas atau kategori cocok untuk dapat melakukan prediksi dan diagnosa [9].

Penelitian ini akan melakukan perbandingan kinerja *machine learning* dari empat metode klasifikasi, yaitu *Support Vector Machine (SVM)*, *Logistic Regression*, *Naïve Bayes*, dan *K-Nearest Neighbor (KNN)*. Pengujian akan berfokus pada perbandingan dari dataset yang menggunakan fitur *PCA* dan yang tidak, serta data akan dibagi untuk data latih dan data uji sebanyak 70-30.

Evaluasi dari hasil kinerja metode akan ditampilkan dalam bentuk *confusion matrix* untuk memberikan gambaran analisa perbandingan hasil kinerja terhadap masing-masing dataset yang digunakan. Dengan analisa ini akan didapatkan metode metode dengan kinerja terbaik untuk melakukan prediksi dari klasifikasi dataset tersebut.

Dengan menggunakan fitur reduksi *PCA* dan melakukan perbandingan dengan data tanpa reduksi *PCA*, lalu menggunakan beberapa metode klasifikasi machine learning, maka diharapkan penelitian ini dapat mengetahui dampak dan manfaat penggunaan *PCA* untuk mereduksi data serta untuk mengetahui metode yang paling cocok digunakan pada dataset diabetes.

## 2. *RESEARCH METHOD*

### Machine Learning

*Machine learning* merupakan salah satu fitur terpenting dalam kecerdasan buatan *(Artificial Intelligence)* yang mendukung pengembangan sistem computer yang memiliki kemampuan untuk memperoleh pengetahuan dari pengalaman data terdahulu tanpa perlu pemrograman untuk setiap kasus [10].

### Principal Component Analysis

*Principal component analysis* merupakan salah satu metode untuk reduksi data yang mengubah variabel-variabel menjadi variabel tidak berkorelasi yang disebut komponen utama dan mengurangi jumlah variabel prediktif. Komponen-komponen utama bersifat independen dan diperingkatkan menurut besarnya varian. Ini berarti bahwa komponen utama 1 bertanggung jawab atas jumlah variasi yang lebih besar daripada komponen utama 2. Oleh karena itu, *PCA* berguna untuk pengurangan dimensionalitas. Komponen utama yang mempunyai nilai eigen mendekati nol dapat dihilangkan. Dengan cara ini, data dapat direduksi tanpa menyebabkan variabel yang mungkin berguna hilang [11][12].

### Support Vector Machine

*Support Vector Machine (SVM)* merupakan metode machine learning yang banyak digunakan dalam rekognisi pola dan klasifikasi. Algoritma *SVM* melakukan klasifikasi dengan membedakan antara dua kelas dengan memaksimalkan margin antara dua kluster data [13][14].

Prinsip kerja dari metode *SVM* dalam melakukan klasifikasi maupun prediksi adalah dengan mencari ruang pemisah paling optimal dari suatu dataset dalam kelas yang berbeda [15].

### Logistic Regression

*Logistic Regression* (Regresi logistik) merupakan algoritma yang dapat digunakan dalam machine learning untuk melakukan tugas klasifikasi. Analisis ini tidak memerlukan asumsi distribusi multivariat normal atau kesamaan matrik varian kovarian, serta dapat juga diterapkan dalam berbagai skala data [16].

Logistic Regression dapat digunakan untuk melakukan analisis untuk menjelaskan hubungan dari variabel yang dikotomis (skala nominal atau ordinal dengan dua kategori) atau polikotomis (skala nominal atau ordinal dengan lebih dari dua kategori) dengan satu atau lebih variabel prediktor pada data kontinu atau kategori [17].

### Naïve Bayes

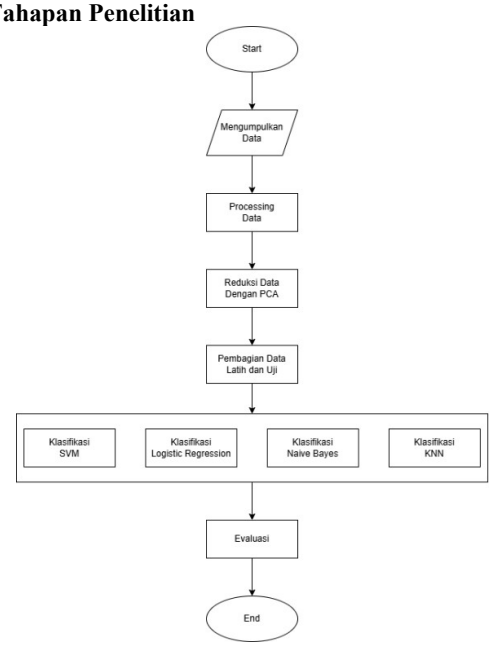
*Naïve bayes* merupakan algoritma yang menggunakan penggolongan statistic sederhana berdasarkan teorema bayes dengan asumsi antar variabel saling bebas atau conditional independence (ketidaktergantungan), kelebihan algoritma naive bayes adalah algoritma ini sangat sederhana dan mudah diimplementasikan, kecepatan proses dan akurasi yang cukup baik untuk digunakan pada tipe data dengan volume yang besar, beragam, dan tidak terstruktur seperti teks [18].

### K-Nearest Neighbors

*K-Nearest Neighbor (KNN)* merupakan algoritma dari machine learning yang dapat diimplementasikan untuk melakukan memprediksi, klasifikasi, dan regresi. *KNN* merupakan metode sederhana yang dilakukan dengan cara mencari nilai neighbor (*k*) dari tetangga terdekatnya dan mencari hasil prediksi berdasarkan mayoritas dan jumlah yang ganjil [19]. *KNN* merupakan salah satu metode

yang dapat dipakai dalam pengelompokan data yang menggunakan algoritma supervised [20].

**Tahapan Penelitian**




Gambar 1. Tahapan penelitian

Pada gambar 1 dijelaskan tahapan penelitian yang terdiri dari mengumpulkan data, processing data, reduksi data dengan PCA, pembagian data latih dan data uji, implementasi klasifikasi dengan 4 metode yaitu *Support Vector Machine*, *Logistic Regression*, *Naïve Bayes*, dan *K-Nearest Neighbors*, lalu evaluasi.

1. Mengumpulkan data

   Data yang digunakan dalam penelitian jurnal ini berasal dari *Kaggle*, dengan fokus pada dataset penyakit diabetes. Data tersebut mencakup variabel-variabel yang relevan untuk klasifikasi dan prediksi penyakit diabetes.

2. Processing data

   Data akan mengalami tahap processing, termasuk penanganan nilai-nilai yang hilang, normalisasi, dan pemrosesan lainnya agar saat proses klasifikasi berlangsung dapat digunakan secara optimal tanpa kendala.

3. Reduksi data dengan PCA

   Metode ekstraksi fitur yang digunakan adalah *Principal Component Analysis (PCA)* yang bertujuan untuk mengurangi dimensi fitur dan mempertahankan informasi yang signifikan untuk digunakan sebagai input pada metode klasifikasi.

4. Pembagian data latih dan uji

   Data akan dibagi menjadi data latih dan data uji sebesar 70-30 yang diperlukan untuk mengevaluasi kinerja metode klasifikasi.

5. Klasifikasi SVM

   Klasifikasi dengan metode *SVM* akan dilakukan dengan bahasa pemrograman *Python*. Klasifikasi akan dilakukan terhadap dataset yang telah di reduksi dengan *PCA* dan yang tidak.

6. Klasifikasi Logistic Regression

   Klasifikasi dengan metode *Logistic Regression* akan dilakukan dengan bahasa pemrograman *Python*. Klasifikasi akan dilakukan terhadap dataset yang telah di reduksi dengan *PCA* dan yang tidak.

7. Klasifikasi Naïve Bayes

   Klasifikasi dengan metode *Naïve Bayes* akan dilakukan dengan bahasa pemrograman Python. Klasifikasi akan dilakukan terhadap dataset yang telah di reduksi dengan PCA dan yang tidak.

8. Klasifikasi KNN

   Klasifikasi dengan metode *KNN* akan dilakukan dengan bahasa pemrograman Python. Klasifikasi akan dilakukan terhadap dataset yang telah di reduksi dengan *PCA* dan yang tidak.

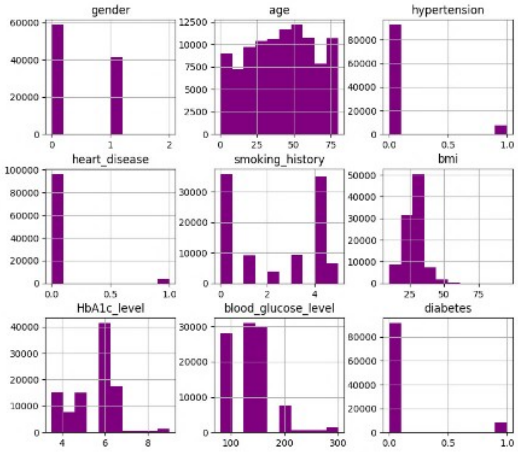
9. Evaluasi

   Setiap metode klasifikasi yang telah dilakukan akan ditampilkan dan dievaluasi dalam bentuk confusion matrix yang mencakup parameterparameter seperti akurasi, presisi, recall, dan F1-score. Analisis akan dilakukan terhadap hasil evaluasi untuk memahami kekuatan dan kelemahan masing-masing metode terhadap dataset yang direduksi dengan *PCA* dan yang tidak.

## 3. *RESULT AND ANALYSIS*

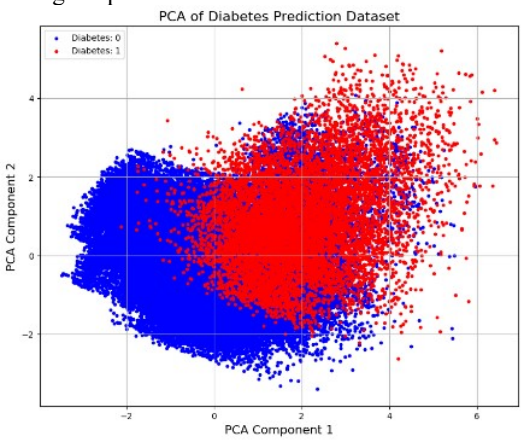
**Processing Dataset**

Dataset penyakit diabetes yang didapatkan melalui *Kaggle*, memiliki variabel-variabel yang dapat dilihat dalam ringkasan keluaran berikut.




Gambar 2. Visualisasi data dataset

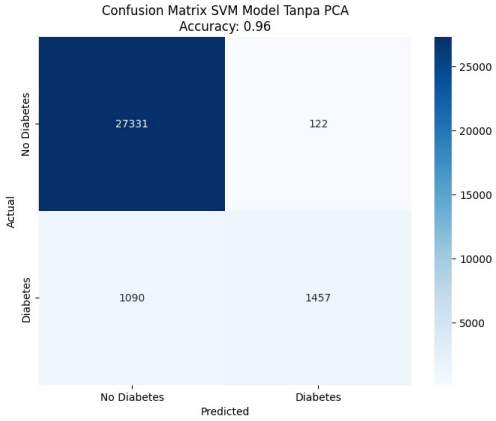
Dataset yang telah dikumpulkan akan di proses dengan fitur reduksi data *PCA* terhadap variabel-variabel yang ada menjadi dua komponen *PCA*. Lalu dengan variabel yang telah di ubah menjadi komponen *PCA* akan diklasifikan variabel diabetes, dimana plot berwarna biru berarti "tidak diabetes" sedangkan plot berwarna merah berarti "diabetes".



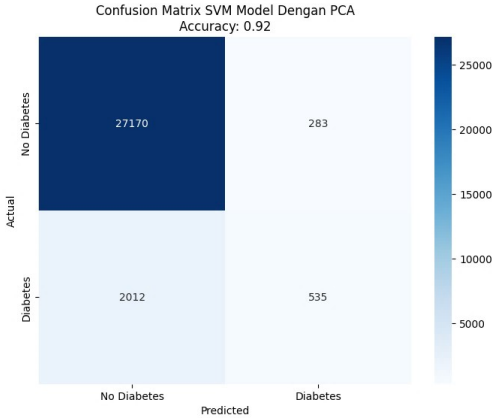

Gambar 3. Visualisasi data dataset dengan fitur PCA

## Klasifikasi Support Vector Machine

Pengujian yang dilakukan dengan metode *Support Vector Machine* menggunakan bahasa pemrograman *Python*. Hasil dari pengujian metode ini akan ditampilkan dalam bentuk *confusion matrix*, yang ditampilkan pada gambar berikut.



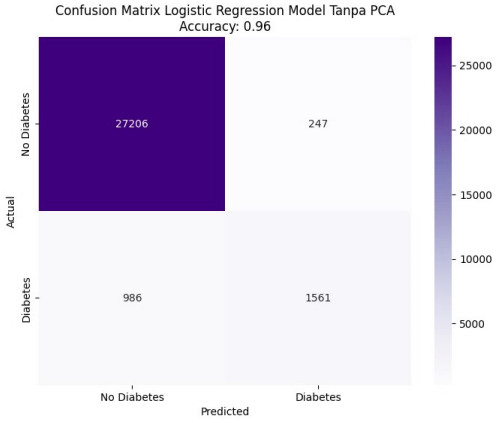

Gambar 4. Confusion Matrix SVM tanpa PCA



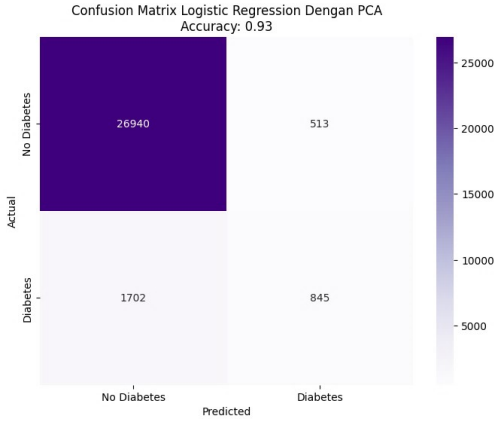

Gambar 5. Confusion Matrix SVM dengan PCA

## Klasifikasi Logistic Regression

Pengujian yang dilakukan dengan metode *Logistic Regression* menggunakan bahasa pemrograman *Python*. Hasil dari pengujian metode ini akan ditampilkan dalam bentuk *confusion matrix*, yang ditampilkan pada gambar berikut.



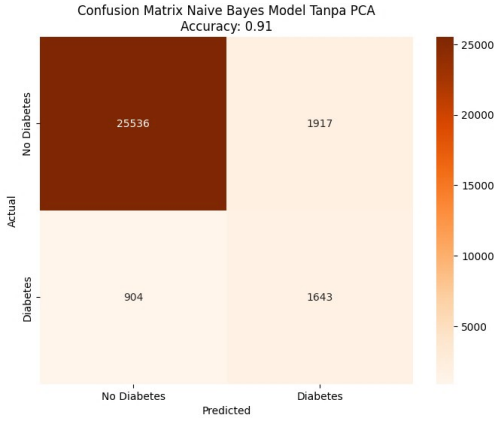

Gambar 6. Confusion Matrix Logistic Regression tanpa PCA



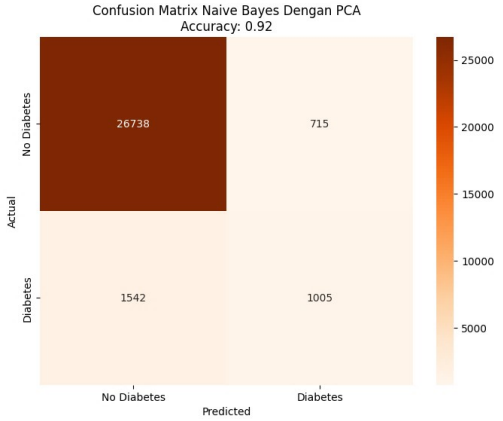

Gambar 7. Confusion Matrix Logistic Regression dengan PCA

## Klasifikasi Naive Bayes

Pengujian yang dilakukan dengan metode *Naïve Bayes* menggunakan bahasa pemrograman *Python*. Hasil dari pengujian metode ini akan ditampilkan dalam bentuk *confusion matrix*, yang ditampilkan pada gambar berikut.



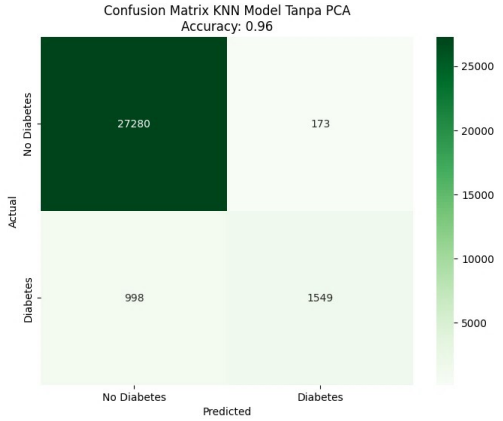

Gambar 8. Confusion Matrix Naïve Bayes tanpa PCA



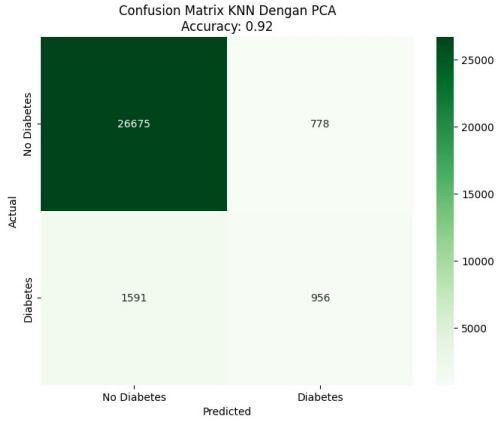

Gambar 9. Confusion Matrix Naïve Bayes dengan PCA

## Klasifikasi K-Nearest Neighbors

Pengujian yang dilakukan dengan metode *K-Nearest Neighbors* menggunakan bahasa pemrograman *Python*. Hasil dari pengujian metode ini akan ditampilkan dalam bentuk *confusion matrix*, yang ditampilkan pada gambar berikut.




Gambar 10. Confusion Matrix KNN tanpa PCA




Gambar 11. Confusion Matrix KNN dengan PCA

## 4. *CONCLUSION*

Dari hasil penelitian dalam jurnal ini, setelah melakukan evaluasi hasil dari visualisasi data dan akurasi dari beberapa metode klasifikasi yang digunakan, yaitu *Support Vector Machine (SVM)*, *Logistic Regression*, *Naïve Bayes*, dan *K-Nearest Neighbors (KNN)*, dapat ditemukan beberapa temuan terhadap perbandingan metode-metode yang digunakan pada dataset yang digunakan *PCA* dan tidak menggunakan *PCA*.

Untuk metode *SVM* pada data tanpa *PCA* mendapatkan akurasi sebesar 96%, sedangkan pada data dengan *PCA* mendapatkan akurasi sebesar 92%.Dengan ini dapat disimpulkan metode *SVM* lebih cocok digunakan pada data tanpa *PCA*.

Untuk metode *Logistic Regression* pada data tanpa *PCA* mendapatkan akurasi sebesar 96%,

sedangkan pada data dengan *PCA* mendapatkan akurasi sebesar 93%. Dengan ini dapat disimpulkan metode *Logistic Regression* lebih cocok digunakan pada data tanpa *PCA*.

Untuk metode *Naive Bayes* pada data tanpa *PCA* mendapatkan akurasi sebesar 91%, sedangkan pada data dengan *PCA* mendapatkan akurasi sebesar 92%. Dengan ini dapat disimpulkan metode *Naive Bayes* lebih cocok digunakan pada data dengan *PCA*.

Untuk metode *KNN* pada data tanpa *PCA* mendapatkan akurasi sebesar 96%, sedangkan pada data dengan *PCA* mendapatkan akurasi sebesar 92%. Dengan ini dapat disimpulkan metode *KNN* lebih cocok digunakan pada data tanpa *PCA*.

## 5. *ACKNOWLEDGEMENTS*

Puji syukur kami panjatkan kepada tuhan yang maha esa kami dapat menyelesaikan penulisan jurnal yang merupakan hasil dari kerja keras dan usaha kelompok penelitian kami hingga tuntas.

Kami ingin mengucapkan rasa terima kasih kepada semua pihak yang telah berpartisipasi dan berkontribusi dalam penelitian ini. Kami menyadari bahwa penulisan jurnal ini jauh dari kesempurnaan, oleh karena itu kami dengan tulus menerima kritik dan saran yang membangun guna perbaikan di masa yang akan datang.

## 6. *REFERENCES*

[1] P. R. Putri, "Jurnal Pengabdian Komunitas," vol. 03, no. 01, pp. 1–6, 2024.

[2] F. B. S. Rizki Aqsyari D, Siti Fatimah Aminah Nikita Putri Adhila, Putri Inrian Tari and B. Murti, "Edukasi Pencegahan Diabetes Pada Lansia di RW 13 Jebres," *Pengabdi. Komunitas*, vol. 02, no. 01, pp. 64–70, 2023.

[3] K. Shah, R. Punjabi, P. Shah, and M. Rao, "Real Time Diabetes Prediction using Naïve Bayes Classifier on Big Data of Healthcare," *Int. Res. J. Eng. Technol.*, vol. 07, no. 05, pp. 102–107, 2020.

[4] A. Pratama, A. C. Nurcahyo, and L. Firgia, "Penerapan Machine Learning dengan Algoritma Logistik Regresi untuk Memprediksi Diabetes," *Pros. CORISINDO 2023*, pp. 116–121, 2023, [Online]. Available: https://stmikpontianak.org/ojs/index.php/corisindo/article/view/30%0Ahttps://stmikpontianak.org/ojs/index.php/corisindo/article/download/30/22

[5] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 20–28, 2021, doi: 10.38094/jastt20165.

[6] M. R. Belgaum *et al.*, "Enhancing the Efficiency of Diabetes Prediction through Training and Classification using PCA and LR Model," *Ann. Emerg. Technol. Comput.*, vol. 7, no. 3, pp. 78–91, 2023, doi: 10.33166/AETiC.2023.03.004.

[7] S. Xie, H. Lin, T. Ma, K. Peng, and Z. Sun, "Journal of Rock Mechanics and Geotechnical Engineering Prediction of joint roughness coef fi cient via hybrid machine learning model combined with principal components analysis," *J. Rock Mech. Geotech. Eng.*, no. xxxx, 2024, doi: 10.1016/j.jrmge.2024.05.059.

[8] M. R. Hunafa and A. Hermawan, "KLIK: Kajian Ilmiah Informatika dan Komputer Perbandingan Algoritma Naïve Bayes dan K-Nearest Neighbor Pada Imbalace Class Dataset Penyakit Diabetes," *Media Online*, vol. 4, no. 3, pp. 1551–1561, 2023, doi: 10.30865/klik.v4i3.1486.

[9] J. Flamino, R. DeVito, B. K. Szymanski, and O. Lizardo, "A Machine Learning Approach to Predicting Continuous Tie Strengths," 2021, [Online]. Available: http://arxiv.org/abs/2101.09417

[10] A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," *Procedia Comput. Sci.*, vol. 165, pp. 292–299, 2019, doi: 10.1016/j.procs.2020.01.047.

[11] J. Y. Le Chan *et al.*, "Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review," *Mathematics*, vol. 10, no. 8, 2022, doi: 10.3390/math10081283.

[12] A. Gorgoglione, A. Castro, V. Iacobellis, and A. Gioia, "A comparison of linear and non-linear machine learning techniques (PCA and SOM) for characterizing urban nutrient runoff," *Sustain.*, vol. 13, no. 4, pp. 1–19, 2021, doi: 10.3390/su13042054.

[13] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes," *BMC Med. Inform. Decis. Mak.*, vol. 10, no. 1, 2010, doi: 10.1186/1472-6947-10-16.

[14] T. H. Tanjung and M. Furqan, "Classification of Heart Disease Using Support Vector Machine," *Sinkron*, vol. 8, no. 3, pp. 1803–1812, 2024, doi: 10.33395/sinkron.v8i3.13904.

[15] H. Apriyani and K. Kurniati, "Perbandingan Metode Naïve Bayes Dan Support Vector Machine Dalam Klasifikasi Penyakit

Diabetes Melitus," *J. Inf. Technol. Ampera*, vol. 1, no. 3, pp. 133–143, 2020, doi: 10.51519/journalita.volume1.isssue3.year2020.page133-143.

[16] Q. R. Cahyani *et al.*, "Prediksi Risiko Penyakit Diabetes menggunakan Algoritma Regresi Logistik Diabetes Risk Prediction using Logistic Regression Algorithm Article Info ABSTRAK," *JOMLAI J. Mach. Learn. Artif. Intell.*, vol. 1, no. 2, pp. 2828–9099, 2022, doi: 10.55123/jomlai.v1i2.598.

[17] D. Ariyanto, A. Sofro, A. N. Hanifah, J. B. Prihanto, D. A. Maulana, and R. W. Romadhonia, "Logistic and Probit Regression Modeling To Predict the Opportunities of Diabetes in Prospective Athletes," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 18, no. 3, pp. 1391–1402, 2024, doi: 10.30598/barekengvol18iss3pp1391-1402.

[18] R. Aristawidya, I. Indahwati, E. Erfiani, A. Fitrianto, and M. A. A., "Perbandingan Analisis Regresi Logistik Biner Dan Naïve Bayes Classifier Untuk Memprediksi Faktor Resiko Diabetes," *J. Lebesgue J. Ilm. Pendidik. Mat. Mat. dan Stat.*, vol. 5, no. 2, pp. 782–794, 2024, doi: 10.46306/lb.v5i2.617.

[19] A. Oktaviana, D. P. Wijaya, A. Pramuntadi, and D. Heksaputra, "Prediksi Penyakit Diabetes Melitus Tipe 2 Menggunakan Algoritma K-Nearest Neighbor (K-NN)," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 3, pp. 812–818, 2024, doi: 10.57152/malcom.v4i3.1268.

[20] P. D. Rinanda, B. Delvika, S. Nurhidayarnis, N. Abror, and A. Hidayat, "Perbandingan Klasifikasi Antara Naive Bayes dan K-Nearest Neighbor Terhadap Resiko Diabetes pada Ibu Hamil," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 2, no. 2, pp. 68–75, 2022, doi: 10.57152/malcom.v2i2.432.