Dielektrika – Department of Electrical Engineering University of Mataram
Vol. 11, No. 1, February 2024, pp. 82 - 88
P-ISSN: 2086-9487, E-ISSN: 2579-650x

Prediksi Diabetes Melitus Tipe-2 Menggunakan Sequential Forward Selection (SFS) Dengan Algoritma Support Vector Machine (SVM)

Saputra [1], L. Ahmad S. Irfan Akbar [1], Cipta Ramadhani[1]
[1] Jurusan Teknik Elektro, Fakultas Teknik Universitas Mataram Jl. Majapahit 62, Mataram, INDONESIA 83115

ARTICLE INFO

Article history :
Received February 19, 2024
Revised February 27, 2024
Accepted February 27, 2024

Keywords :
Diabetes;
Support Vector Machine;
Sequential Forward Selection;
Data Cleaning;

ABSTRACT

Diabetes, as a chronic condition, can lead to severe complications if not managed effectively. Consequently, the development of accurate predictive models is crucial for its early detection and prevention of long-term consequences. This study investigates the impact of feature dimensionality reduction and data cleaning on the performance of such predictive models. Leveraging the Pima Indian Dataset, two distinct models were constructed, each undergoing different preprocessing stages. The initial model was built without data cleansing, whereas the second model incorporated data cleansing techniques. Subsequently, Sequential Forward Selection was employed to determine the optimal feature subset size, followed by model training utilizing the Support Vector Machine algorithm. Evaluation of model performance was conducted post-training, and a comparative analysis was performed between the two models. The findings indicate that feature dimensionality reduction significantly enhances model performance, with the data-cleansed model exhibiting superior predictive capabilities.

Diabetes, sebagai penyakit kronis, dapat menyebabkan komplikasi parah jika tidak ditangani secara efektif. Oleh karena itu, pengembangan model prediksi yang akurat sangat penting untuk deteksi dini dan pencegahan dampak jangka panjang. Studi ini menyelidiki dampak pengurangan dimensi fitur dan pembersihan data terhadap kinerja model prediktif tersebut. Dengan memanfaatkan Kumpulan Data Pima Indian, dua model berbeda dibuat, masing-masing menjalani tahap prapemrosesan yang berbeda. Model awal dibangun tanpa pembersihan data, sedangkan model kedua menggunakan teknik pembersihan data. Selanjutnya, Sequential Forward Selection digunakan untuk menentukan ukuran subset fitur yang optimal, dilanjutkan dengan pelatihan model menggunakan algoritma Support Vector Machine. Evaluasi kinerja model dilakukan pasca pelatihan, dan analisis komparatif dilakukan antara kedua model. Temuan menunjukkan bahwa pengurangan dimensi fitur secara signifikan meningkatkan kinerja model, dengan model yang dibersihkan datanya menunjukkan kemampuan prediktif yang unggul.

Corresponding Author
L. Ahmad S.Irfan Akbar, Universitas Mataram. Mataram. Indonesia
Email: irfan@unram.ac.id

1. PENDAHULUAN

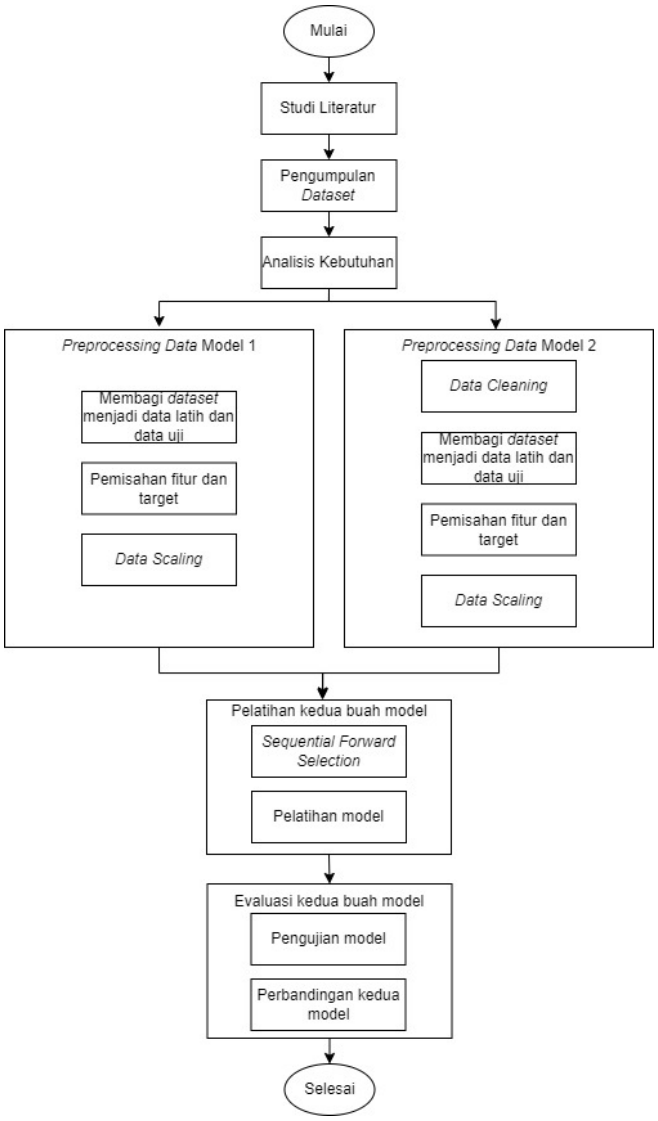
Diabetes [1] merupakan salah satu penyakit kronis yang memberikan dampak signifikan terhadap kesehatan dan kualitas hidup individu. Komplikasi jangka panjang biasanya berkembang secara bertahap saat diabetes tidak ditangani dengan baik [2], beberapa diantaranya adalah gangguan pada mata, kerusakan ginjal dan saraf, hingga penyakit kardiovaskular. Untuk mencegah hal tersebut diperlukan sebuah model *machine learning* [3] yang dapat memberikan prediksi diabetes dengan akurasi yang tinggi.

Dataset yang digunakan dalam penelitian ini, telah juga digunakan pada penelitian oleh Smith [4], untuk memprediksi seseorang akan menderita diabetes atau tidak dalam 5 tahun kedepan dengan menggunakan *ADAP Learning* dan dihasilkan sebuah model yang memiliki performa yang cukup baik dengan nilai *sensitivitas* dan *specificity* 76%. Penelitian tersebut menggunakan *Pima Indian Dataset* [5], penelitian tersebut tidak melakukan *data cleaning* dan tidak mengimplementasikan teknik reduksi dimensi. *Dataset* dengan dimensi yang tinggi dapat mengakibatkan dampak yang buruk bagi model yang dibuat, seperti kompleksitas dari model menjadi tinggi, proses training yang lama, dan juga dapat menyebabkan *overfitting*.

Sequential Forward Selection (SFS) [6] adalah salah satu algoritma pencarian, algoritma ini merupakan algoritma yang sederhana dalam konteks pengurangan dimensi. Penelitian tentang kombinasi SFS-SVM juga pernah dilakukan oleh Kabir et al. [7] untuk memprediksi diabetes dan didapatkan hasil bahwa model kombinasi SFS dengan SVM lebih bagus daripada model yang mengkombinasikan SFS dengan algoritma *machine learning* lainnya, pada penelitian ini juga dihasilkan bahwa model SFS-SVM dengan data yang fiturnya direduksi menghasilkan performa lebih baik daripada model dengan data yang fiturnya tidak direduksi (hanya menggunakan SVM). Penelitian ini melakukan klasifikasin bagi objek dataset yang diprediksi mengalami penyakit diabetes di 5 tahun ke depan. SFS digunakan sebagai metode untuk menyeleksi fitur-fitur yang paling berpengaruh terhadap performa model dan menggunakan SVM [8] sebagai algoritma klasifikasi.

2. METODE

Penelitian ini dilakukan sesuai dengan alur penelitian yang digambarkan dalam Gambar 1. Alur penelitian digunakan oleh penulis dalam pelaksanaan penelitian ini agar hasil yang dicapai tidak menyimpang dari tujuan yang telah ditetapkan sebelumnya. Alur penelitian dapat dilihat dalam Gambar 1 berikut.



Gambar 1. Alur Penelitian

Tahap pertama adalah Studi literatur tentang "Prediksi Diabetes Melitus Tipe-2 Menggunakan Sequential Forward Selection (SFS) Dengan Algoritma Support Vector Machine (SVM)" akan melibatkan pencarian, seleksi, dan penelaahan tentang literatur terkait dengan metode tersebut serta aplikasinya dalam

memprediksi diabetes. Studi literatur yang dilakukan didapatkan melalui E-Book, Jurnal dan lain-lain. Tahap kedua adalah pengumpulan dataset, dataset yang digunakan pada penelitian kali ini adalah dataset Pima Indian yang diperoleh dari National Institute of Diabetes and Digestive and Kidney Disease. Dataset ini diperoleh melalui situs Kaggle. Dataset yang digunakan memiliki 2 label, yaitu 0 (tidak diabetes) dan 1 (positif diabetes). Pada penelitian ini dataset yang digunakan memiliki 8 fitur atau variabel bebas dan 1 variabel terikat (label/target) yang memiliki 768 baris, dengan 268 sampel didiagnosis penyakit diabetes dan 500 sampel didiagnosis sehat atau tidak diabetes. Dalam dataset ini, sampel diambil dari populasi perempuan Pima Indian yang berada di dekat Phoenix, Arizona. Populasi tersebut telah diteliti secara terus menerus sejak tahun 1965 oleh National Institute of Diabetes and Digestive and Kidney Diseases karena tingginya angka kejadian diabetes [9].

Tahap ketiga adalah analisis kebutuhan, adapun berbagai kebutuhan yang mendukung untuk dilakukannya penelitian ini mencakup hardware, software, dan library dari Python. Tahap selanjutnya adalah preprocessing data, pada penelitian kali ini akan dibuat beberapa model dengan teknik preprocessing yang beberda. Model 1 akan dibuat tanpa menggunakan teknik data cleaning, sedangkan model 2 akan dibuat dengan menggunakan data cleaning. Data cleaning yang dilakukan adalaha menghapus data yang memiliki nilai null. Tahap kelima adalah pelatihan kedua buah model. Sebelum model dilatih akan melalui tahap seleksi fitur menggunakan Sequential Forward Selection (SFS) dan akan dicari model dengan jumlah fitur optimal yang menghasilkan performa terbaik. Setelah melalui tahap seleksi fitur, model akan dilatih dengan data latih menggunakan algoritma Support Vector Machine (SVM). Tahap terakhir adalah evaluasi kedua buah model. Akan dilakukan pengujian terhadap model-model yang sudah dibuat untuk melihat kinerja dari model. Tahap pengujian akan menggunakan data uji. Setelah masing-masing model diuji lalu akan dibandingkan performanya dan akan dilihat model manakah yang menghasilkan performa lebih baik.

3. HASIL DAN PEMBAHASAN (10 PT)

3.1. *Data Cleaning*

Data cleaning diperlukan untuk mencegah pengaruh buruk kepada model yang dikarenakan oleh nilai fitur yang hilang tersebut. Untuk menghapus sampel tersebut, pertama-tama dilakukan perhitungan jumlah nilai yang hilang pada fitur fitur tersebut.

```
Jumlah nilai 0 pada kolom Pregnancies: 111
Jumlah nilai 0 pada kolom Glucose: 5
Jumlah nilai 0 pada kolom BloodPressure: 35
Jumlah nilai 0 pada kolom SkinThickness: 227
Jumlah nilai 0 pada kolom Insulin: 374
Jumlah nilai 0 pada kolom BMI: 11
Jumlah nilai 0 pada kolom DiabetesPedigreeFunction: 0
Jumlah nilai 0 pada kolom Age: 0
Jumlah nilai 0 pada kolom Outcome: 500
```

Gambar 2. Jumlah Nilai 0 Pada Setiap Fitur

Setelah melakukan perhitungan jumlah nilai 0, akan dipilih sampel yang tidak memiliki nilai 0 pada fitur Glucose, BloodPressure, SkinThickness, Insulin, dan BMI. Dikarenakan nilai 0 pada fitur-fitur tersebut adalah kemungkinan nilai yang hilang atau nilai yang tidak diukur. Setelah melalui tahap ini, sampel yang tersisa adalah 392 sampel.

3.2. *Pembagian Dataset*

Dataset dibagi menjadi dengan perbandingan 80% untuk data latih, dan 20% untuk data uji dari total sampel. Setelah dibagi dataset akan disimpan menjadi file csv dan akan di-import kembali ke jupyter notebook. Dari kedua buah model yaitu model dengan data cleaning dan model tanpa data cleaning, didapatkan jumlah data latih dan data uji yaitu sebagai berik*ut.*

Tabel 1. Splitting data dari model dengan Data Cleaning

Data Latih	Data Uji
576 Sampel	192 Sampel

Tabel 1 merupakan hasil dari pembagian dataset dari model tanpa data cleaning, dapat dilihat sampel total pada data latih adalah 576 sampel, sedangkan pada data uji adalah 192 sampel. **Tabel 2**

merupakan hasil dari pembagian dataset dari model dengan data cleaning, dapat dilihat sampel total pada data latih adalah 313 sampel, sedangkan pada data uji adalah 79 sampel

***Tabel 2**.Splitting data dari model dengan Data Cleaning*

Data Latih	Data Uji
313 Sampel	79 Sampel

3.3. *Evaluasi Model*

Setelah kedua model SFS-SVM dengan teknik preprocessing yang berbeda dibuat, akan dilihat hasil performa model dengan masing-masing mempertahankan jumlah fitur yang berbeda. Dilakukan evaluasi model dengan parameter akurasi, presisi, recall, specificity dan F1-Score. Berikut adalah performa dari model SFS-SVM yang tanpa melakukan data cleaning.

***Tabel 3**.Performa Model SFS-SVM Tanpa Data Cleaning*

Jumlah fitur	Akurasi (%)	Presisi (%)	*Recall* (%)	*Specificity* (%)	F1-*Score* (%)
1	71.88	85.37	74.47	64.71	79.55
2	74.48	84.55	77.61	67.24	80.93
3	74.48	82.93	78.46	66.13	80.63
4	75.00	83.74	78.63	67.21	81.10
5	76.04	84.55	79.39	68.85	81.89
6	75.00	82.93	79.07	66.67	80.95
7	73.96	79.67	79.67	63.77	79.67
8	73.96	79.67	79.67	63.77	79.67

Tabel 3 merupakan hasil performa model SFS-SVM tanpa melakukan data cleaning dengan jumlah fitur hasil seleksi yang berbeda-beda. Dapat dilihat performa model terbaik pada tabel tersebut terletak pada jumlah fitur hasil seleksi 5. Sehingga dapat diartikan bahwa pengurangan jumlah dimensi atau fitur meningkatkan performa model.

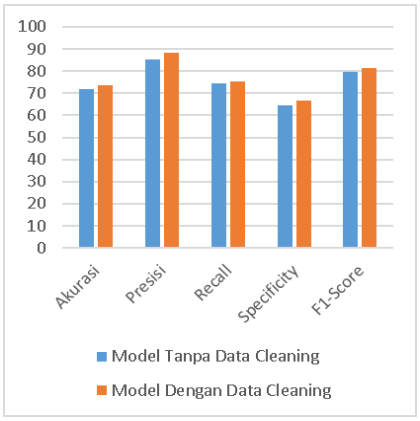
***Tabel 4**.Performa Model SFS-SVM Dengan Data Cleaning*

Jumlah fitur akhir	Akurasi (%)	Presisi (%)	*Recall* (%)	*Specificity* (%)	F1-*Score*(%)
1	73.42	88.46	75.41	66.67	81.42
2	75.95	90.38	77.05	72.22	83.19
3	75.95	90.38	77.05	72.22	83.19
4	75.95	88.46	77.97	70.00	82.88
5	73.42	86.54	76.27	65.00	81.08
6	77.22	86.54	80.36	69.57	83.33
7	73.42	84.62	77.19	63.64	80.73
8	74.68	84.62	78.57	65.22	81.48

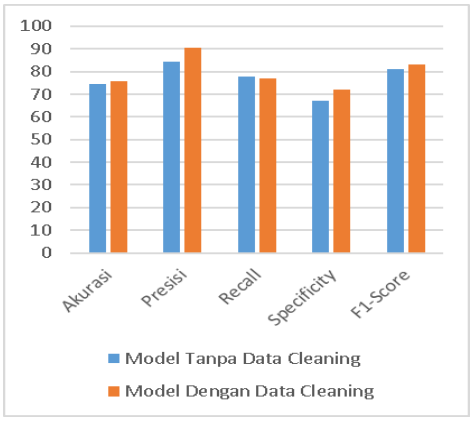
Tabel 4 diatas merupakan hasil performa model SFS-SVM yang melakukan data cleaning dengan jumlah fitur hasil seleksi yang berbeda-beda. Dapat dilihat performa model terbaik pada tabel tersebut terletak pada jumlah fitur hasil seleksi 6. Sehingga dapat diartikan bahwa pengurangan jumlah dimensi atau fitur meningkatkan performa model.

3.4. *Perbandingan Model*

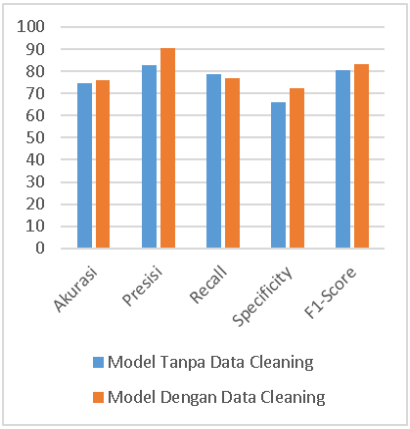
Kedua model yaitu model SFS-SVM yang tanpa melakukan data cleaning , dan model SFS-SVM yang melakukan data cleaning akan dilihat perbandingan performa yang dihasilkan dari kedua buah model yang sudah dibuat tersebut. Perbandingan akan ditinjau berdasarkan jumlah fitur atau dimensi dari masing-masing model, adapun perbandingannya dapat dilihat dari grafik-grafik berikut berikut.



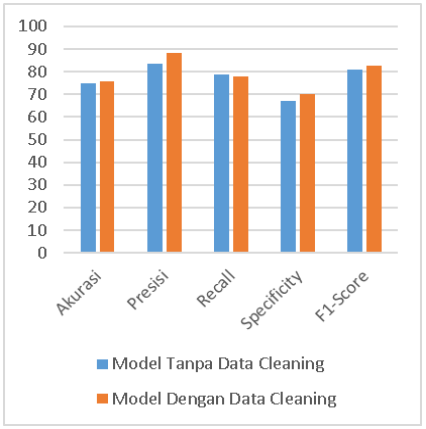
Gambar 3. *Model Dengan 1 Fitur*



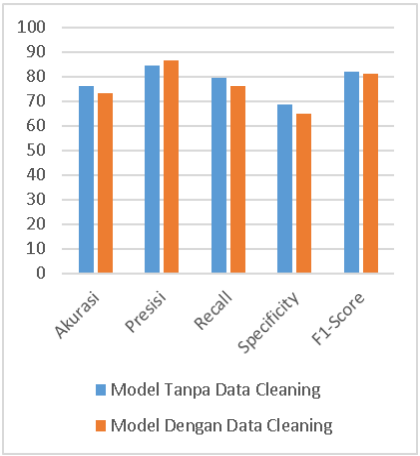
Gambar 4.*Model Dengan 2 Fitur*



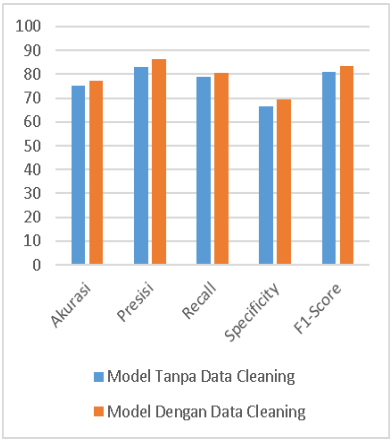
Gambar 5.Model Dengan 3 Fitur



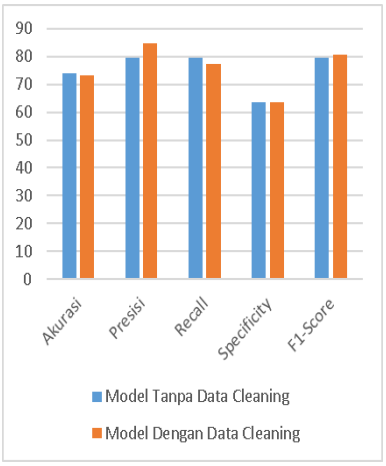
Gambar 6. Model Dengan 4 Fitur



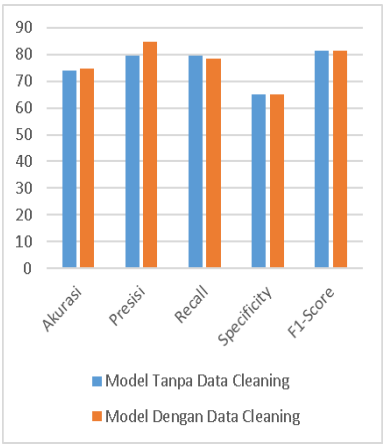
Gambar 7.Model Dengan 5 Fitur



Gambar 8. Model Dengan 6 Fitur



Gambar 9.Model Dengan 7 Fitur



Gambar 10.Model Dengan 8 Fitur

Berdasarkan grafik-grafik diatas dapat dilihat perbandingan performa model PCA-SVM dan SFS-SVM berdasarkan jumlah fitur. Didapatkan bahwa performa model meningkat jika fiturnya dikurangi dengan menggunakan teknik reduksi fitur baik PCA ataupun SFS. Performa metode PCA-SVM terbaik didapatkan pada *principal component* berjumlah 4, sedangkan performa metode SFS-SVM terbaik didapatkan pada hasil seleksi fitur berjumlah 6

4. KESIMPULAN

Dibuat dua buah model dengan tahap preprocessing yang berbeda yaitu model tanpa tahap data cleaning dan model dengan tahap data cleaning. Digunakan metode seleksi fitur yaitu Sequential Forward Selection (SFS) untuk mengurangi jumlah dimensi atau fitur pada data yang digunakan dan algoritma Support Vector Machine (SVM) sebagai algoritma klasifikasi. Pengurangan jumlah fitur pada model menunjukan bahwa dengan dikurangi jumlah fitur dapat meningkatkan performa model. Pada penelelitian ini didapatkan performa terbaik pada model yang tanpa melakukan data cleaning yaitu pada jumlah fitur hasil seleksi 5 dengan akurasi 76.04%, presisi 84.55%, recall 79.39%, specificity 68.85%, dan F1-Score 81.89%. Sedangkan pada model yang melakukan data cleaning yaitu pada jumlah fitur hasil seleksi 6 dengan akurasi 77.22%, presisi 86.54%, recall 80.36%, specificity 69.57%, dan F1-Score 83.33%.Dua buah model yang dibangun dengan tahap preprocessing yang berbeda, ditunjukkan bahwa model yang melakukan tahap data cleaning memiliki performa yang lebih baik daripada model yang tidak melakukan tahap data cleaning.

DAFTAR PUSTAKA

[1] S. A. Lestari and Zulkarnain and Sijid, "Diabetes Melitus: Review Etiologi, Patofisiologi, Gejala, Penyebab, Cara Pemeriksaan, Cara Pengobatan dan Cara Pencegahan," in *Prosiding Biologi Achieving the Sustainable Development Goals with Biodiversity in Confronting Climate Change*, Gowa, Indonesia, 2021.

[2] W. A. Kadek, "Diabetes Melitus Tipe 2: Faktor Risiko, Diagnosis, dan Tatalaksana," *Ganesha Medicina Journal*, vol. 1, pp. 114-120, 2021.

[3] S. A. Ahuja, "Machine learning and its applications: A review," in *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*, 2017.

[4] S. A. Johannes, "Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus," *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pp. 261-265, 7-11 November 1988.

[5] "kaggle," Pima Indians Diabetes Database, [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. [Accessed 2022].

[6] Y. A. Saifudin, "Sequential Feature Selection in Customer Churn Prediction Based on Naive Bayes," in *3rd International Conference on Informatics, Engineering, Science, and Technology (INCITEST 2020)*, Bandung, Indonesia, 2020.

[7] E. K. Hashi, "Developing Diabetes Disease Classification Model using Sequential Forward Selection

Algorithm," *International Journal of Computer Applications,* vol. 180, pp. 1-6, 2017.

[8] N. A. Vinod, "Performance Analysis of Support Vector Machine in Diabetes Prediction," in *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2020.

[9] M. A. Furqon, "Klasifikasi Penyakit Diabetes menggunakan Metode Support Vector Machine," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 5, pp. 622-633, 2021.

Saputra was born on March 09, 2002 in Mataram and lives in Selaparang, Mataram City, West Nusa Tenggara Province. He has completed his undergraduate education in Electrical Engineering at the University of Mataram. Currently doing Internship IT Bootcamp at PT. Bank Rakyat Indonesia (Persero) Tbk. as PHP - Laravel Developer.

Ahmad Syamsul Irfan Akbar, completed the Bachelor's program in the Department of Engineering Physics, Gadjah Mada University in 2006. Completed the Master's Program in Electrical Engineering in 2009. Since 2009 he has served as a lecturer in the Department of Electrical Engineering, University of Mataram. He has a strong interest in reading and specializes in teaching various subjects, including Computer Networks, Machine Learning, Internet of Things.
irfan@unram.ac.id

Cipta Ramadhani, completed Bachelor degree program in Electrical Engineering Department at The University of Mataram, 2009. Completing a Master's Degree Program in Electrical Engineering at The University of Gadjah Mada, 2011. Since 2012, he has been serving as a lecturer at the Electrical and Computer Engineering Department of the University of Mataram. He has a strong interest in reading and specializes in teaching various courses, including Data Science and Deep Learning. cipta.ramadhani@unram.ac.id