Jurnal Informatika dan Komputer (INFOKOM)
Volume 12 Nomor 2 Tahun 2024
E ISSN : 2722-4147 P ISSN : 2339-188X
OJS : http://journal.piksi.ac.id/index.php/INFOKOM

EVALUASI PERBANDINGAN KINERJA MODEL MACHINE LEARNING UNTUK PREDIKSI DIABETES: STUDI KASUS XGBOOST, RANDOM FOREST, DAN SVM

[1,]Adil Setiawan, [2]Adelina, [3]Damri Mulia Hutabalian,[4]Ricky Irnanda,[5]Heru Fredi, [6]Iswanto
[1,2,3,4,5,6] Program Studi Ilmu Komputer
[1,2,3,4,5,6] Universitas Potensi Utama
Email:[1]Adio165@gmail.com;[2] linaada289@gmail.com;[3] damrinainggolan23@gmail.com; [4]rickyirnanda1@gmail.com;[5]herufredi85@gmail.com; [6]iswantoo1982@gmail.com;

ABSTRACT

This study evaluates and compares the performance of three major machine learning (ML) models—XGBoost, Random Forest, and Support Vector Machine (SVM)—for diabetes risk prediction using the Pima Indians Diabetes Dataset. The core problem addressed is the need for accurate and effective early detection to mitigate serious complications such as cardiovascular disease and kidney failure. The proposed solution involves training and evaluating these models on a pre-processed dataset, using metrics like accuracy, precision, recall, F1-score, and Area Under the Curve (AUC) on the ROC Curve. Random Forest achieved the best performance, showing the highest accuracy (0.76) and AUC (0.82). Furthermore, Random Forest was superior in detecting positive cases (diabetes), as evidenced by the confusion matrix analysis, which is critical in a medical context. Glucose and BMI were identified as the most crucial features for prediction across the models. The key finding is that Random Forest is the most effective and stable model, providing better discriminative abilities for clinical decision support in early diabetes risk prediction.

Keywords: *Diabetes, Machine learning, Random Forest, XGBoost, SVM.*

ABSTRAK

Penelitian ini mengevaluasi dan membandingkan kinerja tiga model *machine learning* (ML) utama—XGBoost, Random Forest, dan Support Vector Machine (SVM)—untuk prediksi risiko diabetes menggunakan *Pima Indians Diabetes Dataset*. Pokok permasalahan yang diangkat adalah kebutuhan akan deteksi dini yang akurat dan efektif untuk mengurangi komplikasi serius seperti penyakit kardiovaskular dan gagal ginjal. Pendekatan yang diusulkan melibatkan pelatihan dan evaluasi model-model ini pada *dataset* yang telah melalui *preprocessing*, dengan menggunakan metrik seperti akurasi, *precision*, *recall*, *F1-score*, dan *Area Under the Curve* (AUC) pada *ROC Curve*. Random Forest mencapai kinerja terbaik, menunjukkan akurasi tertinggi (0.76) dan AUC tertinggi (0.82). Lebih lanjut, Random Forest unggul dalam mendeteksi kasus positif (diabetes)

sebagaimana dibuktikan oleh analisis *confusion matrix*, yang sangat penting dalam konteks medis. *Glucose* dan *BMI* diidentifikasi sebagai fitur paling krusial untuk prediksi pada kedua model yang dianalisis. Temuan utama adalah Random Forest merupakan model yang paling efektif dan stabil, memberikan kemampuan diskriminasi yang lebih baik untuk mendukung keputusan klinis dalam prediksi dini risiko diabetes.

Katakunci : *Diabetes*, *Machine learning*, Random Forest, XGBoost, SVM.

PENDAHULUAN

Diabetes mellitus adalah salah satu penyakit kronis yang paling umum di dunia, dengan prevalensi yang terus meningkat. Penyakit ini dapat menyebabkan berbagai komplikasi serius seperti penyakit kardiovaskular, gagal ginjal, dan neuropati jika tidak didiagnosis dan dikelola dengan baik (Bontha et al., 2025; Ismail & Materwala, 2025; Kadam et al., 2025). Oleh karena itu, deteksi dini dan prediksi yang akurat sangat penting untuk mengurangi dampak penyakit ini pada individu dan sistem kesehatan (Bontha et al., 2025; Kadam et al., 2025; Shrivastava et al., 2023).

Dengan meningkatnya jumlah dataset besar dalam sektor kesehatan, metode machine learning (ML) telah mendapatkan perhatian signifikan dalam menganalisis, memprediksi, dan menemukan pola dalam dataset diabetes (Güler et al., 2025; Kadam et al., 2025; Shrivastava et al., 2023). Model ML dapat meningkatkan akurasi diagnosis dan prediksi, yang pada gilirannya dapat membantu dalam pengambilan keputusan klinis yang lebih baik (Bontha et al., 2025; Güler et al., 2025; Ismail & Materwala, 2025). Namun, ada tantangan yang signifikan dalam memilih model ML yang paling efektif karena variasi dalam teknik yang digunakan dan kurangnya transparansi dalam fitur yang digunakan oleh model (Fregoso-Aparicio et al., 2021; Sami et al., 2024).

Studi ini berfokus pada evaluasi komparatif terhadap tiga algoritma *machine learning* utama—XGBoost, Random Forest, dan Support Vector Machine (SVM)—untuk memprediksi risiko diabetes. Tujuannya adalah untuk menentukan model mana yang menunjukkan kinerja paling efektif. Evaluasi ini penting untuk menentukan model yang paling efektif dan efisien dalam berbagai kondisi dataset dan parameter model yang seragam (Bontha et al., 2025; Güler et al., 2025; Singh & Baweja, 2025).

KAJIAN PUSTAKA

XGBoost: Model ini terkenal karena kemampuan untuk menangani dataset yang sangat besar dan kompleks dengan tingkat akurasi yang tinggi. Beberapa studi menunjukkan bahwa XGBoost sering kali mengungguli model lain dalam hal akurasi dan kemampuan generalisasi (Güler et al., 2025; Singh & Baweja, 2025; Wang et al., 2020).

Jurnal Informatika dan Komputer (INFOKOM)
Volume 12 Nomor 2 Tahun 2024
E ISSN : 2722-4147 P ISSN : 2339-188X
OJS : http://journal.piksi.ac.id/index.php/INFOKOM

Random Forest: Model ini menawarkan keseimbangan antara akurasi dan interpretabilitas. Random Forest telah terbukti konsisten dalam berbagai ukuran dataset dan mampu menangani data yang tidak seimbang dengan baik 1 10 11(Bontha et al., 2025; Devi et al., 2025; Harika et al., 2025).

Support Vector Machine (SVM): SVM efektif dalam menangani data berdimensi tinggi dan sering digunakan dalam aplikasi medis. Namun, kinerjanya dapat bervariasi tergantung pada ukuran dan kualitas dataset (Dalve et al., 2023; Devi et al., 2025; Güler et al., 2025).

Pengembangan Hipotesis

Hipotesis penelitian ini adalah:

- **H1**: XGBoost akan menunjukkan kinerja prediksi diabetes yang lebih baik dibandingkan dengan Random Forest dan SVM dalam hal akurasi dan F1-score.
- **H2**: Random Forest akan menunjukkan kinerja yang lebih konsisten dibandingkan dengan XGBoost dan SVM dalam berbagai ukuran dataset.
- **H3**: SVM akan menunjukkan kinerja yang lebih baik dalam dataset berdimensi tinggi tetapi mungkin kurang efektif dibandingkan dengan XGBoost dan Random Forest dalam dataset yang lebih kecil dan tidak seimbang.

Dengan mengevaluasi dan membandingkan ketiga model ini, penelitian ini diharapkan dapat memberikan wawasan yang lebih baik tentang model ML mana yang paling efektif untuk prediksi diabetes, serta memberikan panduan bagi praktisi kesehatan dalam memilih model yang tepat untuk aplikasi klinis.

METODE / ANALISIS PERANCANGAN

Dataset

Dataset yang digunakan dalam penelitian ini adalah **Pima Indians Diabetes Database**, yang berisi 768 data pasien dengan 8 fitur medis dan 1 variabel target yang menunjukkan apakah pasien menderita diabetes (1) atau tidak (0). Fitur-fitur yang digunakan dalam penelitian ini mencakup:

- **Pregnancies**: Jumlah kehamilan.
- **Glucose**
- **BloodPressure**
- **SkinThickness**.
- **Insulin**
- **BMI**
- **DiabetesPedigreeFunction**: Keturunan diabetes.
- **Age**.

Variabel targetnya adalah **Outcome**, yang menunjukkan apakah seseorang memiliki diabetes atau tidak (1 atau 0).

Jurnal Informatika dan Komputer (INFOKOM)
Volume x Nomor x Tahun xxxx
E ISSN : 2722-4147 P ISSN : 2339-188X
OJS : http://journal.piksi.ac.id/index.php/INFOKOM

Tahapan Penelitian

1. Pengumpulan Data
   Data yang digunakan pada penelitian ini diperoleh dari **Kaggle** dan diunduh dalam format **CSV**. Proses pengumpulan data dimulai dengan **uploading** dataset ke **Google Colab** untuk pemrosesan lebih lanjut.
2. Eksplorasi Data
   Sebelum melakukan pelatihan model, data dieksplorasi dan dianalisis untuk memeriksa **keberadaan missing values** dan **distribusi data**. Deskripsi statistik dari data digunakan untuk memahami karakteristik fitur dan target variabel. Visualisasi distribusi target (Outcome) dilakukan dengan menggunakan **countplot** untuk memeriksa keseimbangan antara pasien yang memiliki diabetes dan yang tidak.

3. Preprocessing Data
   Dilakukan beberapa proses pembersihan data, yaitu:
   a. Penanganan Nilai Nol: Nilai 0 pada fitur medis seperti Glucose, BloodPressure, SkinThickness, Insulin, dan BMI dianggap sebagai nilai yang tidak valid atau hilang, karena nilai ini tidak mencerminkan kondisi medis normal. Oleh karena itu, nilai-nilai ini diganti dengan median masing-masing kolom untuk menghindari distorsi pada model.
   b. Scaling Data.
4. Pembagian Data
   Data kemudian dibagi menjadi training set dan test set menggunakan fungsi train_test_split dari scikit-learn dengan pembagian 80% untuk pelatihan dan 20% untuk pengujian. Proses ini memastikan bahwa model diuji pada data yang belum pernah dilihat sebelumnya untuk mengukur kinerja generalisasi model.
5. Pemilihan Model
   Tiga model machine learning yang dipilih untuk penelitian ini adalah:
   a. Random Forest
   b. Support Vector Machine (SVM)
   c. XGBoost.
6. Training Model
   Setiap model dilatih menggunakan training set, yang terdiri dari fitur-fitur yang telah diproses dan label Outcome. Proses pelatihan ini dilakukan dengan menyesuaikan parameter model untuk memaksimalkan kinerja dalam memprediksi hasil pada test set.
7. Evaluasi Model
   Setelah model dilatih, model diuji pada test set untuk mengukur kinerja model menggunakan beberapa metrik evaluasi, yaitu:
   - Akurasi

INFOKOM
Jurnal Informatika dan Komputer
Jurnal Informatika dan Komputer (INFOKOM)
Volume 12 Nomor 2 Tahun 2024
E ISSN : 2722-4147 P ISSN : 2339-188X
OJS : http://journal.piksi.ac.id/index.php/INFOKOM

- Precision
- Recall
- F1-score
- Confusion Matrix

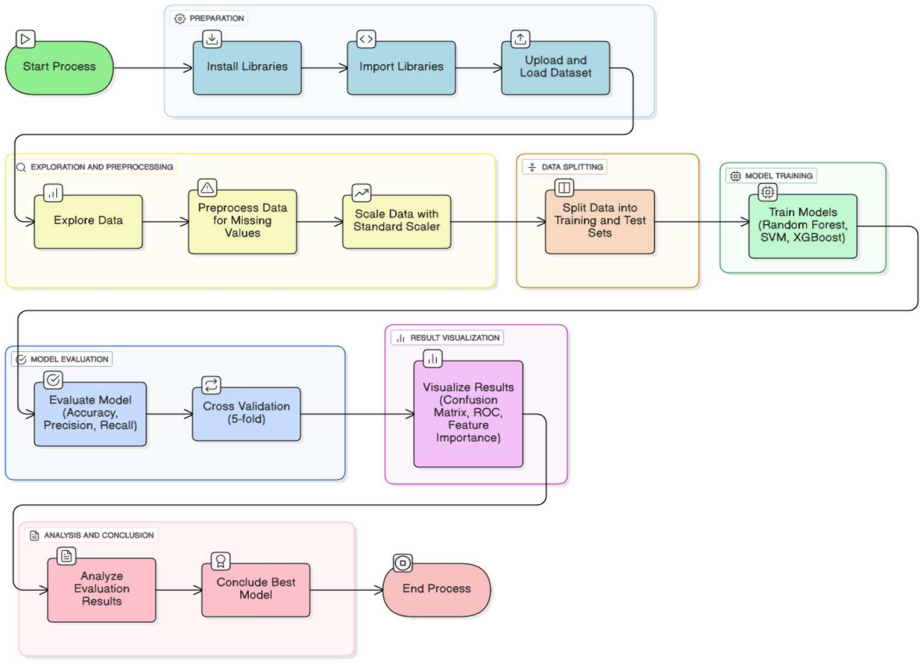
Selain itu, dilakukan cross-validation menggunakan 5-fold cross-validation untuk memperoleh evaluasi yang lebih stabil dan menghindari overfitting. Proses ini melibatkan pembagian data menjadi 5 bagian, dengan setiap bagian digunakan sekali sebagai data uji dan sisanya digunakan untuk pelatihan model.

8. Visualisasi Hasil Evaluasi

   Beberapa visualisasi digunakan untuk membandingkan hasil evaluasi dari masing-masing model, seperti:

   a. Confusion Matrix: Visualisasi dari confusion matrix untuk setiap model untuk mengevaluasi kinerja masing-masing model dalam mengklasifikasikan data.
   b. ROC Curve
   c. Perbandingan Akurasi: Visualisasi bar chart yang menunjukkan akurasi model untuk perbandingan yang lebih jelas antara ketiga model.
   d. Feature Importance: Menyediakan wawasan mengenai fitur-fitur mana yang paling berkontribusi pada prediksi model, yang dapat digunakan untuk memahami model lebih dalam.

Workflow Model



Gambar 1. ***Workflow Model***

INFOKOM
Jurnal Informatika dan Komputer

Jurnal Informatika dan Komputer (INFOKOM)
Volume x Nomor x Tahun xxxx
E ISSN : 2722-4147 P ISSN : 2339-188X
OJS : http://journal.piksi.ac.id/index.php/INFOKOM

HASIL DAN PEMBAHASAN

Dataset berisi delapan fitur klinis sebagai variabel prediktor dan satu variabel target (Outcome) yang menunjukkan kondisi diabetes pada pasien (0 = tidak diabetes, 1 = diabetes). Distribusi kelas target menunjukkan adanya ketidakseimbangan kelas, yaitu 65% tidak diabetes dan 35% diabetes, sehingga model memerlukan kemampuan klasifikasi yang baik agar tidak bias terhadap kelas mayoritas.
Setelah dilakukan pra-pemrosesan, termasuk penanganan nilai 0 pada fitur biologis dan standarisasi data, ketiga model kemudian diuji menggunakan data uji sebesar 20% dari dataset dan divalidasi dengan metode **5-fold cross validation** untuk memperoleh performa yang lebih stabil.

Hasil Evaluasi Model

Ringkasan akurasi ketiga model adalah sebagai berikut:

Tabel 1. Ringkasan Akurasi ketiga model

Model	Akurasi
Random Forest	0.76
SVM	0.73
XGBoost	0.73

Dari tabel tersebut, **Random Forest memperoleh akurasi tertinggi sebesar 0.7597**, sementara **SVM memperoleh akurasi 0.7272** dan **XGBoost sebesar 0.7337**. Perbedaan ini mengindikasikan bahwa model Random Forest lebih mampu menangkap pola hubungan antara fitur klinis dengan risiko diabetes.

Selain itu, berdasarkan hasil cross-validation, diperoleh skor berikut:

Tabel 2. Skor Cross-Validation

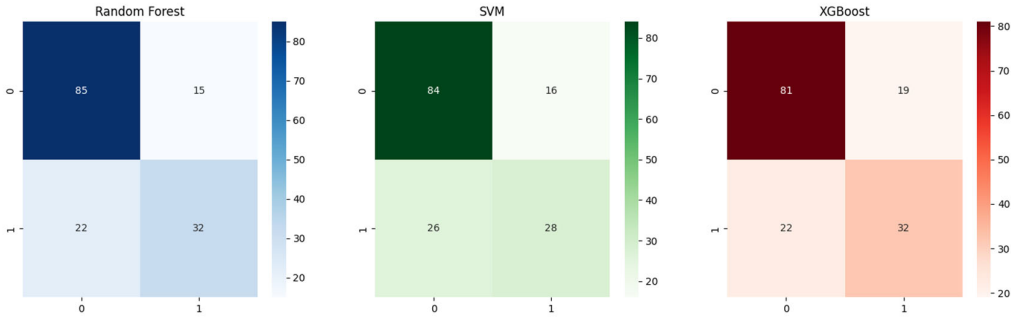
Model	Skor Cross-Validation (cv=5)
Random Forest	0.7604
SVM	0.7656
XGBoost	0.7266

Walaupun skor cross-validation SVM sedikit lebih tinggi dibandingkan Random Forest, performa pada data uji tetap menunjukkan Random Forest lebih baik secara konsisten terutama dalam mendeteksi pasien diabetes dengan nilai recall yang lebih besar dibandingkan dua model lainnya. Hal ini penting karena pada kasus medis, kemampuan mengenali kasus positif (penderita diabetes) lebih prioritaskan dibanding hanya mencapai akurasi keseluruhan yang tinggi.

INFOKOM
Jurnal Informatika dan Komputer
Jurnal Informatika dan Komputer (INFOKOM)
Volume 12 Nomor 2 Tahun 2024
E ISSN : 2722-4147 P ISSN : 2339-188X
OJS : http://journal.piksi.ac.id/index.php/INFOKOM

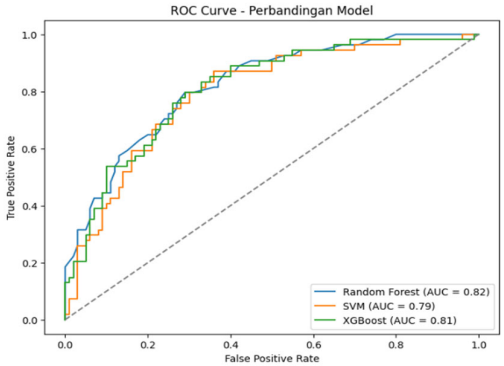
Kualitas Prediksi Berdasarkan Confusion Matrix

Berdasarkan hasil confusion matrix ketiga model yang diuji, terlihat bahwa Random Forest memberikan kinerja terbaik dalam mendeteksi pasien dengan diabetes (kelas 1). Model Random Forest berhasil mengidentifikasi 32 kasus positif dengan benar, dengan jumlah kesalahan False Negative sebanyak 22 kasus. Sementara itu, XGBoost menghasilkan jumlah True Positive yang sama yaitu 32, namun memiliki False Positive lebih tinggi dibandingkan Random Forest, sehingga lebih banyak pasien non-diabetes yang salah diprediksi sebagai diabetes. SVM merupakan model dengan performa terendah dalam mendeteksi kelas positif, dengan True Positive sebesar 28 dan False Negative 26. Kondisi ini menunjukkan bahwa SVM kurang sensitif terhadap kelas minoritas akibat ketidakseimbangan data. Dari hasil ini dapat diinterpretasikan bahwa Random Forest memiliki kemampuan klasifikasi yang lebih baik dan lebih stabil dibandingkan model lainnya, terutama dalam konteks medis di mana kemampuan mendeteksi pasien diabetes secara benar sangat penting untuk meminimalkan risiko kegagalan diagnosis dini.



Gambar 2. *Confusion Matrix* ketiga model

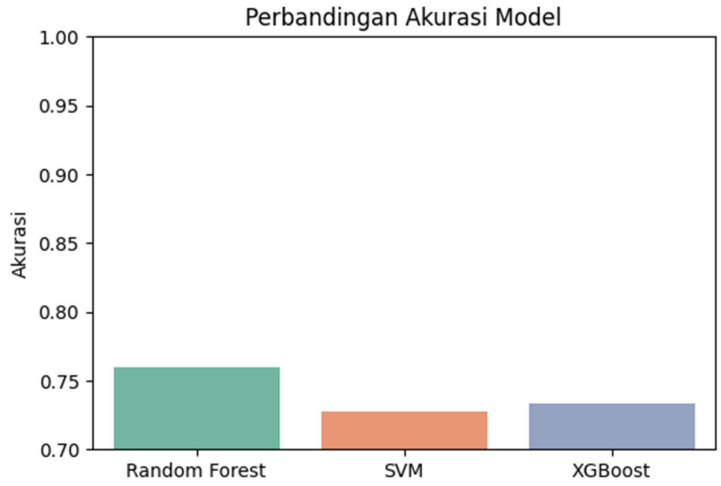
Analisis Visual



Gambar 3. *ROC Curve* ketiga model

Jurnal Informatika dan Komputer (INFOKOM)
Volume x Nomor x Tahun xxxx
E ISSN : 2722-4147 P ISSN : 2339-188X
OJS : http://journal.piksi.ac.id/index.php/INFOKOM

Perbandingan kurva *Receiver Operating Characteristic* (ROC) menunjukkan bahwa model **Random Forest** mencatatkan *Area Under the Curve* (AUC) tertinggi, yaitu **0.82**. Capaian ini menegaskan superioritas Random Forest dalam memisahkan pasien diabetes dan non-diabetes (kemampuan diskriminasi) dibandingkan model lain. **XGBoost** mengikuti dengan nilai AUC yang juga baik, yaitu **0.81**, meskipun kinerjanya sedikit di bawah Random Forest. Sementara itu, **SVM** memiliki nilai AUC terendah pada **0.79**, menunjukkan bahwa kemampuannya untuk mengidentifikasi pola antar kelas kurang optimal. Meskipun ketiga model ini menunjukkan kemampuan klasifikasi yang memadai (AUC di atas 0.75) , disparitas nilai AUC memperkuat kesimpulan bahwa Random Forest adalah model klasifikasi yang paling unggul dalam studi ini, menjadikannya alat pendukung keputusan yang lebih efektif untuk prediksi risiko diabetes.



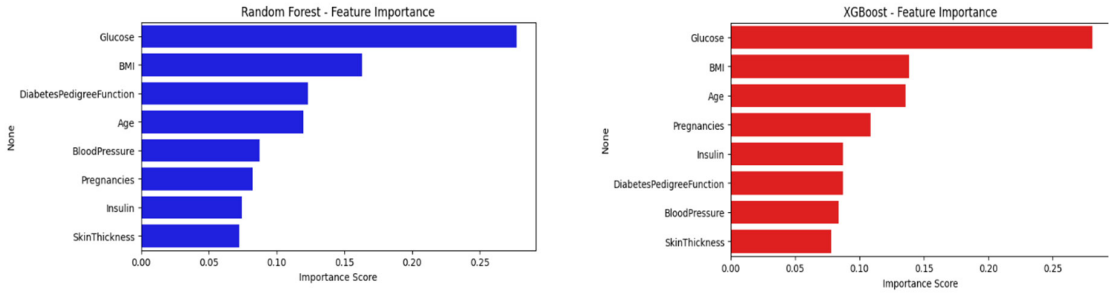
Gambar 4. Perbandingan Akurasi Model

Visualisasi perbandingan akurasi model, yang didasarkan pada pengujian menggunakan 20% data uji, mengkonfirmasi bahwa **Random Forest** mencapai performa tertinggi. Random Forest mencatatkan akurasi sebesar **0.76** , menjadikannya lebih unggul dalam mempelajari dan memprediksi pola data medis yang kompleks. Keunggulan ini disebabkan oleh penggunaan teknik *ensemble* yang menggabungkan banyak pohon keputusan, yang berkontribusi pada peningkatan stabilitas dan mitigasi risiko *overfitting*.

Sebaliknya, **XGBoost** dan **SVM** menunjukkan akurasi yang hampir sama, yaitu sekitar **0.73**, dengan SVM sedikit di bawah XGBoost. Model SVM dianggap kurang optimal, terutama pada *dataset* dengan ketidakseimbangan kelas dan fitur numerik yang saling terkait, yang mengakibatkan kinerja yang lebih rendah.

.

Jurnal Informatika dan Komputer (INFOKOM)
Volume 12 Nomor 2 Tahun 2024
E ISSN : 2722-4147 P ISSN : 2339-188X
OJS : http://journal.piksi.ac.id/index.php/INFOKOM



Gambar 5. *Feature Importance* ketiga model

Kedua grafik menunjukkan bahwa baik Random Forest maupun XGBoost menilai Glucose dan BMI sebagai fitur paling penting dalam memprediksi diabetes. Fitur seperti Age, Pregnancies, dan DiabetesPedigreeFunction memiliki tingkat pengaruh sedang, meskipun urutannya sedikit berbeda antar-model. Sementara itu, BloodPressure, SkinThickness, dan Insulin konsisten menjadi fitur dengan pengaruh paling rendah.

Secara keseluruhan, kedua model sepakat bahwa kadar glukosa dan indeks massa tubuh adalah faktor utama yang menentukan hasil prediksi. SVM tidak ditampilkan pada grafik *feature importance* karena SVM tidak memiliki perhitungan *feature importance* bawaan, terutama jika menggunakan kernel non-linear yang umum dipakai untuk akurasi lebih tinggi.

KESIMPULAN

Penelitian ini berhasil mengevaluasi dan membandingkan kinerja tiga model *machine learning* (ML)—Random Forest, XGBoost, dan Support Vector Machine (SVM)—dalam memprediksi risiko diabetes menggunakan *dataset* Pima Indians Diabetes. Berdasarkan pengujian pada data uji dan validasi silang (*5-fold cross-validation*), Random Forest menunjukkan kinerja prediksi yang paling unggul dibandingkan dengan dua model lainnya, dengan mencapai akurasi tertinggi sebesar 0.76 dan nilai AUC tertinggi sebesar 0.82. Selain itu, analisis *confusion matrix* menunjukkan bahwa Random Forest paling efektif dalam mendeteksi kasus positif (penderita diabetes) dengan jumlah *True Positive* yang baik dan *recall* yang lebih besar, suatu hal yang krusial dalam konteks diagnosis medis untuk meminimalkan *false negatives* (kegagalan diagnosis dini). Analisis *feature importance* dari Random Forest dan XGBoost juga konsisten menunjukkan bahwa kadar Glukosa dan BMI adalah fitur paling penting dalam memprediksi diabetes. Dengan demikian, Random Forest direkomendasikan sebagai model ML yang paling efektif dan stabil untuk prediksi awal risiko diabetes, memberikan panduan berharga bagi praktisi kesehatan dan mendukung pengambilan keputusan klinis yang lebih baik.

DAFTAR PUSTAKA

Bontha, S. S., Jammalamadaka, S. K. R., Vudatha, C. P., Jammalamadaka, S. B., Duvvuri, B. K., & Vudatha, B. C. (2025). Predicting Risk and Complications of Diabetes Through Built-In Artificial Intelligence. *Computers*, *14*(7). Scopus. https://doi.org/10.3390/computers14070277

Dalve, P., Bobby, D., Marathe, A., Dusane, A., & Daga, S. (2023). *Comparison of Performance of Machine Learning Algorithms for Diabetes Detection*. 2023 3rd International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies, ICAECT 2023. Scopus. https://doi.org/10.1109/ICAECT57570.2023.10118315

Devi, N. M., Asha, V., Govindaraj, M., Manwani, P., Singh, N., & Anoopa, K. M. (2025). *Optimization of Diabetics Diagnosis Using SVM, Random Forest and XGBoost*. 1261–1266. Scopus. https://doi.org/10.1109/ICDT63985.2025.10986609

Fregoso-Aparicio, L., Noguez, J., Montesinos, L., & García-García, J. A. (2021). Machine learning and deep learning predictive models for type 2 diabetes: A systematic review. *Diabetology and Metabolic Syndrome*, *13*(1). Scopus. https://doi.org/10.1186/s13098-021-00767-9

Güler, H., Avcı, D., Ulaş, M., & Omma, T. (2025). In-depth analysis of machine learning models and explainable artificial intelligence methods in diabetes diagnosis. *Journal of the Faculty of Engineering and Architecture of Gazi University*, *40*(3), 1995–2011. Scopus. https://doi.org/10.17341/gazimmfd.1552790

Harika, N. V., Fathimabi, S. K., & Hannah, P. J. (2025). *Symptom-Based Diabetes Likelihood Prediction Using Machine Learning and Big Data*. *1431 LNEE*, 193–202. Scopus. https://doi.org/10.1007/978-981-96-7253-0_16

Ismail, L., & Materwala, H. (2025). IDMPF: intelligent diabetes mellitus prediction framework using machine learning. *Applied Computing and Informatics*, *21*(1–2), 78–89. Scopus. https://doi.org/10.1108/ACI-10-2020-0094

Kadam, P., Godse, S., & Mahalle, P. (2025). *Machine learning in the early detection and prediction of diabetes: A systematic review*. *3325*(1). Scopus. https://doi.org/10.1063/5.0291792

Sami, A., Naseer, A., Hussain, M. Z., Hasan, M. Z., Mustafa, M., Khalid, A., Awan, R., Ashraf, F., Khan, Z. A., & Javaid, A. (2024). *Enhancing Diabetes Detection: A Weighted Averaging Approach for Combined Model Accuracy*. 2024 IEEE 9th International Conference for Convergence in Technology, I2CT 2024. Scopus. https://doi.org/10.1109/I2CT61223.2024.10543684

Shrivastava, P., Kumari, A., Kumari, S., & Bajaj, P. (2023). *A Comprehensive Review on the Prediction of Diabetes Disease Using Machine Learning*. ISED 2023 -

International Conference on Intelligent Systems and Embedded Design. Scopus. https://doi.org/10.1109/ISED59382.2023.10444546

Singh, A. K., & Baweja, D. (2025). *Predicting Diabetic Health Through Ensemble Methods of Machine Learning for Enhanced Well-Being in the Digital Age*. 132–138. Scopus. https://doi.org/10.1109/ICDICI66477.2025.11135166

Wang, L., Wang, X., Chen, A., Jin, X., & Che, H. (2020). Prediction of type 2 diabetes risk and its effect evaluation based on the xgboost model. *Healthcare (Switzerland)*, *8*(3). Scopus. https://doi.org/10.3390/healthcare8030247