Indonesian Journal of Humanities and Social Sciences
https://ejournal.uit-lirboyo.ac.id/index.php/IJHSS
E-ISSN: 2746-2188
Volume 6 Issue 4 Desember 2025

Model Ensemble untuk Prediksi Risiko Diabetes dengan Pertimbangan Efisiensi Biaya

Rofiatul Qosimah,
Universitas Negeri Surabaya, Indonesia
rofiatulqosimah.22038@mhs.unesa.ac.id

Riska Dhenabayu,
Universitas Negeri Surabaya, Indonesia
riskadhenabayu@unesa.ac.id

Achmad Kautsar,
Universitas Negeri Surabaya, Indonesia
achmadkautsar@unesa.ac.id

Anita Safitri,
Universitas Negeri Surabaya, Indonesia
anitasafitri@unesa.ac.id

Abstract

This study addresses the growing global burden of diabetes by evaluating whether ensemble-based machine learning models can support reliable and cost-efficient early risk prediction. Moving beyond accuracy-centered evaluation, the study integrates cost-sensitive threshold optimization and probability calibration to enhance clinical relevance. Random Forest and XGBoost are evaluated using two datasets with contrasting population characteristics. Model performance is examined in terms of discriminative ability, calibration quality, and total misclassification cost. The results indicate that while XGBoost remains competitive on small-scale datasets, Random Forest provides more stable calibration and more consistent cost efficiency. These findings suggest that cost-sensitive and calibrated ensemble approaches have the potential to support more rational and economically efficient diabetes screening policies.

Keywords: *Cost-Sensitive Learning, Diabetes Prediction, Probability Calibration, Random Forest, XGBoost.*

Abstrak

Penelitian ini menanggapi meningkatnya beban global diabetes dengan mengevaluasi apakah model *machine learning* berbasis *ensemble* mendukung prediksi risiko dini yang andal dan efisien secara biaya. Berbeda dari evaluasi berbasis akurasi semata, studi ini mengintegrasikan optimasi ambang *cost-sensitive* dan kalibrasi probabilitas untuk meningkatkan relevansi klinis. Random Forest dan XGBoost dievaluasi menggunakan dua dataset dengan karakteristik populasi yang kontras. Kinerja model dianalisis berdasarkan kemampuan diskriminasi, kualitas kalibrasi, dan total biaya kesalahan klasifikasi. Hasil menunjukkan bahwa meskipun XGBoost kompetitif pada data berskala kecil, Random Forest memberikan kalibrasi yang lebih stabil dan efisiensi biaya yang lebih konsisten. Temuan ini menunjukkan bahwa pendekatan ensemble yang sensitif terhadap biaya, dan terkalibrasi, berpotensi mendukung kebijakan skrining diabetes yang lebih rasional dan efisien secara biaya.

Kata kunci: *Cost-Sensitive Learning, Kalibrasi Probabilitas, Prediksi Diabetes, Random Forest, XGBoost.*

Pendahuluan

Diabetes melitus merupakan penyakit metabolik kronis yang ditandai hiperglikemia akibat gangguan sekresi atau kerja insulin. Prevalensinya terus meningkat secara global dan menimbulkan tantangan kesehatan yang serius. International Diabetes Federation (IDF) melaporkan bahwa pada tahun 2025 sekitar 11,1% populasi dewasa dunia hidup dengan diabetes, dengan lebih dari 40% penderitanya belum menyadari kondisi tersebut [1]. Jumlah ini diproyeksikan mencapai 853 juta orang pada tahun 2050 [2]. Peningkatan kasus yang cepat ini menempatkan diabetes sebagai salah satu penyebab utama morbiditas dan mortalitas global, sehingga deteksi dini menjadi komponen penting dalam upaya pencegahan komplikasi dan penurunan beban penyakit.

Selain dampak klinis, diabetes juga menimbulkan beban ekonomi yang sangat besar. Secara global, jumlah penyandang diabetes meningkat dari sekitar 200 juta pada tahun 1990 menjadi lebih dari 830 juta pada 2022 [3]. Kadar glukosa darah tinggi berkontribusi sekitar 11% terhadap kematian kardiovaskular [4]. Dari sisi ekonomi, total beban diabetes di Amerika Serikat pada tahun 2022 mencapai USD 412,9 miliar per tahun [5], sementara pengeluaran kesehatan terkait diabetes secara global telah melampaui USD 1 triliun [6]. Fakta ini menunjukkan bahwa tanpa pendekatan deteksi dan pengelolaan risiko yang lebih efisien, diabetes berpotensi mengancam keberlanjutan sistem kesehatan.

Dalam konteks tersebut, pendekatan *machine learning* (ML) semakin banyak digunakan untuk mendukung deteksi dini diabetes. Metode *ensemble learning* berbasis *decision tree* seperti *Random Forest* (RF) dan *gradient boosting*, menunjukkan kemampuan yang kuat dalam menangani kompleksitas data kesehatan. Kajian sistematis melaporkan bahwa model *gradient boosting* dan *tree-based ensemble models* termasuk yang paling efektif prediksi diabetes tipe 2 [7]. Studi empiris juga menunjukkan performa tinggi, misalnya, akurasi hingga 92,85% pada prediksi diabetes menggunakan data Pima Indians Diabetes [8]. Temuan-temuan ini mengindikasikan bahwa model ensemble memiliki potensi besar sebagai alat prediksi risiko diabetes.

Namun demikian, terdapat celah penelitian yang signifikan. Sebagian besar studi sebelumnya masih berfokus pada metrik performa umum dan belum secara memadai mempertimbangkan perbedaan dampak klinis dan ekonomi dari kesalahan klasifikasi. Dalam praktik medis, kesalahan *False Negative* (FN) memiliki konsekuensi yang jauh lebih serius dibandingkan *False Positive* (FP) karena dapat menyebabkan keterlambatan diagnosis dan

[1] IDF, "Diabetes Facts and Figures | International Diabetes Federation," 2025, https://idf.org/about-diabetes/diabetes-facts-figures/.
[2] IDF.
[3] WHO, "Diabetes," WHO, 2024, https://www.who.int/news-room/fact-sheets/detail/diabetes.
[4] WHO.
[5] Emily D. Parker et al., "Economic Costs of Diabetes in the U.S. in 2022," *Diabetes Care* 47, no. 1 (January 2, 2024): 26–43, https://doi.org/10.2337/DCI23-0085.
[6] IDF, "The Diabetes Atlas | Global Diabetes Data & Statistics," International Diabetes Federation (IDF), 2025, https://diabetesatlas.org/.
[7] Panagiotis D. Petridis et al., "A Review on Trending Machine Learning Techniques for Type 2 Diabetes Mellitus Management," *Informatics 2024, Vol. 11, Page 70* 11, no. 4 (September 27, 2024): 70, https://doi.org/10.3390/INFORMATICS11040070.
[8] Shahid Mohammad Ganie et al., "An Ensemble Learning Approach for Diabetes Prediction Using Boosting Techniques," *Frontiers in Genetics* 14 (2023): 1252159, https://doi.org/10.3389/FGENE.2023.1252159.

meningkatkan risiko komplikasi, sementara FP umumnya hanya memicu pemeriksaan lanjutan dengan risiko lebih rendah. Pendekatan *cost-sensitive learning*, yang memberikan bobot biaya lebih besar pada kesalahan FN, telah terbukti mampu menurunkan *misclassification cost* secara signifikan [9] dan terbukti dapat menurunkan *misclassification cost* secara signifikan [10]. Selain itu, aspek kalibrasi probabilitas sering diabaikan, padahal model dengan probabilitas yang tidak terkalibrasi dapat menghasilkan estimasi risiko yang bias dan mengganggu pengambilan keputusan klinis meskipun memiliki akurasi tinggi [11] Integrasi *cost-sensitive learning* dan kalibrasi probabilitas dalam evaluasi model prediksi diabetes masih relatif terbatas, terutama dalam pengujian lintas populasi dengan karakteristik data yang berbeda.

Berdasarkan celah tersebut, penelitian ini mengevaluasi kinerja model ensemble RF dan XGBoost menggunakan dua dataset dengan karakteristik kontras, yaitu Pima yang berskala kecil dan relatif homogen 768 data, serta NHANES yang berskala besar dan heterogen sekitar 19.000 data. Pendekatan *cost-sensitive* dan kalibrasi probabilitas diterapkan untuk meningkatkan sensitivitas terhadap kesalahan FN, menjaga reliabilitas probabilitas prediksi, serta menekan *misclassification cost*.

Metode

Penelitian ini mengikuti alur metodologis terstruktur yang dimulai dari pra-pemrosesan data, pemodelan, optimasi berbasis fungsi biaya, kalibrasi probabilitas, hingga evaluasi kinerja berbasis klinis dan ekonomi. Tahap pra-pemrosesan mencakup penanganan nilai hilang, transformasi variabel kategorikal, normalisasi fitur numerik, serta penyeimbangan kelas untuk mengatasi ketidakseimbangan data. Selanjutnya, dua model *ensemble* berbasis pohon keputusan, yaitu RF dan XGBoost, dilatih dan dievaluasi. Untuk meningkatkan relevansi klinis, dilakukan optimasi ambang keputusan berbasis biaya serta kalibrasi probabilitas. Kinerja model kemudian dinilai menggunakan pendekatan *multi-metrik* dan analisis sensitivitas biaya guna memastikan keseimbangan antara performa prediktif dan efisiensi ekonomi.

Pra-Pemrosesan Data

Penanganan Nilai Hilang

Nilai hilang (*missing values*) dapat menyebabkan distorsi distribusi data, data, menurunkan akurasi model, serta menimbulkan bias estimasi parameter, terutama pada model prediktif medis yang sensitif terhadap kualitas input [12]. Pada dataset Pima, yang

[9] Imane Araf, Ali Idri, and Ikram Chairi, "Cost-Sensitive Learning for Imbalanced Medical Data: A Review," *Artificial Intelligence Review 2024 57:4* 57, no. 4 (March 1, 2024): 1–72, https://doi.org/10.1007/S10462-023-10652-8; Nguyen Thai-Nghe, Zeno Gantner, and Lars Schmidt-Thieme, "Cost-Sensitive Learning Methods for Imbalanced Data," *Proceedings of the International Joint Conference on Neural Networks*, 2010, https://doi.org/10.1109/IJCNN.2010.5596486.

[10] Thai-Nghe, Gantner, and Schmidt-Thieme, "Cost-Sensitive Learning Methods for Imbalanced Data."

[11] Ben Van Calster et al., "Calibration: The Achilles Heel of Predictive Analytics," *BMC Medicine* 17, no. 1 (December 16, 2019), https://doi.org/10.1186/S12916-019-1466-7; Lathan Liou et al., "Assessing Calibration and Bias of a Deployed Machine Learning Malnutrition Prediction Model within a Large Healthcare System," *Npj Digital Medicine 2024 7:1* 7, no. 1 (June 6, 2024): 149-, https://doi.org/10.1038/s41746-024-01141-5.

[12] Dajung Ryu and Sohyune Sok, "Prediction Model of Quality of Life Using the Decision Tree Model in Older Adult Single-Person Households: A Secondary Data Analysis," *Frontiers in Public Health* 11 (August 31, 2023): 1224018, https://doi.org/10.3389/FPUBH.2023.1224018/BIBTEX; Luke Oluwaseye Joel, Wesley

memiliki distribusi fitur cenderung *skewed* dan mengandung *outlier*, imputasi dilakukan menggunakan nilai median [13]. Pada dataset NHANES, variabel numerik diimputasi menggunakan median, sedangkan variabel kategorikal menggunakan modus. Selain itu, fitur dengan proporsi nilai hilang lebih dari 50% dieliminasi untuk mengurangi potensi bias dan meningkatkan reliabilitas pemodelan [14].

Transformasi Variabel Kategorikal

Variabel kategorikal ditransformasikan menggunakan *One-Hot Encoding* (OHE) agar dapat diproses oleh algoritma pembelajaran mesin [15]. Pendekatan ini mencegah munculnya hubungan ordinal semu dan membantu mengurangi bias interaksi antar fitur pada model medis yang kompleks [16]. Selain itu, OHE merupakan tahap penting dalam feature engineering untuk menjaga kelengkapan informasi kategorikal selama proses pembelajaran model [17].

Normalisasi Fitur Numerik

Normalisasi fitur numerik dilakukan menggunakan *StandardScaler* agar seluruh fitur memiliki skala yang sebanding. Feature scaling berperan penting dalam meningkatkan stabilitas optimasi dan mempercepat konvergensi algoritma [18]. Selain itu, normalisasi membantu menjaga stabilitas proses kalibrasi probabilitas, karena metode seperti *Platt Scaling* dan *Isotonic Regression* sensitif terhadap distribusi nilai input [19]. Dalam sistem prediksi medis yang kompleks, normalisasi juga berkontribusi dalam mengurangi bias struktural akibat perbedaan skala antar fitur [20].

Doorsamy, and Babu Sena Paul, "A Comparative Study of Imputation Techniques for Missing Values in Healthcare Diagnostic Datasets," *International Journal of Data Science and Analytics 2025 20:7* 20, no. 7 (June 11, 2025): 6357–73, https://doi.org/10.1007/S41060-025-00825-9; Riska Dhenabayu et al., "Harnessing the Power of Transformer Networks in AI-Driven Decision Support Systems for Badminton Action Recognition," *2024 12th International Conference on Cyber and IT Service Management, CITSM 2024*, 2024, https://doi.org/10.1109/CITSM64103.2024.10775332; Manahel Altalhan, Abdulmohsen Algarni, and Monia Turki-Hadj Alouane, "Imbalanced Data Problem in Machine Learning: A Review," *IEEE Access* 13 (2025): 13686–99, https://doi.org/10.1109/ACCESS.2025.3531662.

[13] Joel, Doorsamy, and Paul, "A Comparative Study of Imputation Techniques for Missing Values in Healthcare Diagnostic Datasets."

[14] Ryu and Sok, "Prediction Model of Quality of Life Using the Decision Tree Model in Older Adult Single-Person Households: A Secondary Data Analysis"; Riska Dhenabayu, Hujjatullah Fazlurrahman, and Purwohandoko, "Potential Researches of GAN in Fashion Areas," *Proceedings - 2023 6th International Conference on Computer and Informatics Engineering: AI Trust, Risk and Security Management (AI Trism), IC2IE 2023*, 2023, 276–81, https://doi.org/10.1109/IC2IE60547.2023.10331411.

[15] Fabian Pedregosa et al., "Scikit-Learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot," *Journal of Machine Learning Research* 12 (2011): 2825–30.

[16] Ryu and Sok, "Prediction Model of Quality of Life Using the Decision Tree Model in Older Adult Single-Person Households: A Secondary Data Analysis"; Altalhan, Algarni, and Turki-Hadj Alouane, "Imbalanced Data Problem in Machine Learning: A Review."

[17] Alice Zheng and Amanda Casari, *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*, *ACM Computing Surveys*, First Edit, vol. 53 (Sebastopol, CA: O'Reilly Media, Inc., 2018).

[18] Md Manjurul Ahsan et al., "Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance," *Technologies 2021, Vol. 9, Page 52* 9, no. 3 (July 24, 2021): 52, https://doi.org/10.3390/TECHNOLOGIES9030052.

[19] Van Calster et al., "Calibration: The Achilles Heel of Predictive Analytics."

[20] Zheng and Casari, *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*.

Penyeimbangan Kelas

Data prediksi diabetes umumnya bersifat tidak seimbang, dengan jumlah sampel positif lebih sedikit dibandingkan negatif. Ketidakseimbangan ini dapat menurunkan recall dan meningkatkan risiko FN, yang berbahaya dalam konteks klinis. Untuk mengatasi hal tersebut, digunakan *Synthetic Minority Oversampling Technique* (SMOTE). Metode ini menghasilkan sampel sintetis baru yang mengikuti distribusi kelas minoritas dan terbukti meningkatkan stabilitas prediksi pada dataset kesehatan [21]. Penerapan SMOTE dilakukan hanya pada data latih di setiap lipatan cross-validation untuk mencegah data leakage, sesuai praktik terbaik pemodelan medis [22].

Meskipun SMOTE efektif dalam mengurangi bias kelas mayoritas dan menekan risiko false negative, perlu dicermati bahwa sampel sintetis yang dihasilkan oleh SMOTE merupakan interpolasi matematis, bukan variasi klinis yang terjadi di dunia nyata. Oleh karena itu, peningkatan performa tetap memerlukan evaluasi lanjutan sebelum diadopsi dalam skenario implementasi klinis nyata.

Optimasi Ambang Keputusan

Dalam konteks klinis, biaya kesalahan klasifikasi tidak bersifat simetris. Kesalahan FN memiliki konsekuensi yang jauh lebih serius dibandingkan FP. Oleh karena itu, ambang prediksi default tidak digunakan secara langsung, melainkan dioptimalkan berdasarkan fungsi biaya $Cost = (FN \times 12.022) + (FP \times 15)$. Nilai biaya FN merepresentasikan estimasi biaya tahunan penanganan pasien diabetes USD 12.022 , sedangkan biaya FP mencerminkan biaya pemeriksaan lanjutan berisiko rendah USD 15 [23]**.** Optimasi ambang dilakukan melalui *threshold sweep* untuk menemukan nilai yang meminimalkan total *misclassification cost*, sejalan dengan prinsip *cost-sensitive learning* pada data medis tidak seimbang [24].

Pemodelan

Pemodelan dilakukan menggunakan dua algoritma ensemble berbasis pohon keputusan, yaitu RF dan XGBoost, yang dipilih karena kemampuannya dalam menangani hubungan non-linear dan interaksi multivariabel yang umum dijumpai pada data klinis. Pemilihan kedua model ini juga bertujuan untuk membandingkan karakteristik performa algoritma *ensemble* pada populasi dengan skala dan tingkat heterogenitas yang berbeda.

RF merupakan metode *ensemble* berbasis *bagging* yang membangun banyak pohon keputusan secara paralel pada subset data dan fitur yang berbeda. Pendekatan ini memungkinkan RF mengurangi varians model dan meningkatkan stabilitas prediksi,

---

[21] Ilias Tougui, Abdelilah Jilbab, and Jamal El Mhamdi, "Impact of the Choice of Cross-Validation Techniques on the Results of Machine Learning-Based Diagnostic Applications," *Healthcare Informatics Research* 27, no. 3 (July 1, 2021): 189, https://doi.org/10.4258/HIR.2021.27.3.189; Araf, Idri, and Chairi, "Cost-Sensitive Learning for Imbalanced Medical Data: A Review."

[22] Amey Vrudhula et al., "Machine Learning and Bias in Medical Imaging: Opportunities and Challenges," *Circulation. Cardiovascular Imaging* 17, no. 2 (February 1, 2024): e015495, https://doi.org/10.1161/CIRCIMAGING.123.015495.

[23] Parker et al., "Economic Costs of Diabetes in the U.S. in 2022."

[24] Araf, Idri, and Chairi, "Cost-Sensitive Learning for Imbalanced Medical Data: A Review"; Joffrey L. Leevy et al., "Threshold Optimization and Random Undersampling for Imbalanced Credit Card Data," *Journal of Big Data 2023 10:1* 10, no. 1 (May 6, 2023): 1–22, https://doi.org/10.1186/S40537-023-00738-Z.

khususnya pada dataset berukuran kecil hingga menengah seperti Pima. Studi sebelumnya menunjukkan bahwa RF relatif tahan terhadap overfitting dan mampu mempertahankan performa yang konsisten tanpa memerlukan penalaan parameter yang sangat kompleks, sehingga sering direkomendasikan untuk aplikasi klinis dengan keterbatasan jumlah sampel [25]. Selain itu, RF menyediakan informasi *feature importance* yang berguna untuk interpretasi klinis dan identifikasi faktor risiko utama diabetes.

Sebaliknya, XGBoost menggunakan pendekatan *boosting* bertahap, di mana setiap pohon baru dibangun untuk memperbaiki kesalahan prediksi dari pohon sebelumnya. Mekanisme ini membuat XGBoost lebih adaptif dalam menangkap pola kompleks dan interaksi halus antar variabel, terutama pada dataset berskala besar dan heterogen seperti NHANES. Penelitian sebelumnya menunjukkan bahwa XGBoost memiliki kemampuan generalisasi yang tinggi ketika diterapkan lintas populasi dan mampu mengelola variasi distribusi data yang luas secara efektif [26]. Karakteristik ini menjadikan XGBoost relevan untuk mengevaluasi kinerja model pada data populasi besar dengan kompleksitas tinggi.

Dengan menggunakan RF dan XGBoost secara paralel, penelitian ini tidak hanya membandingkan performa prediktif kedua algoritma, tetapi juga mengevaluasi perbedaan stabilitas, sensitivitas terhadap kesalahan FN, serta konsistensi kinerja pada konteks populasi yang berbeda. Pendekatan komparatif ini memungkinkan analisis yang lebih komprehensif mengenai kesesuaian model ensemble untuk penerapan skrining risiko diabetes pada skenario data klinis yang beragam.

Kalibrasi Probabilitas

Model *ensemble* berbasis pohon cenderung menghasilkan probabilitas prediksi yang *overconfident* [27]. Oleh karena itu, dilakukan kalibrasi probabilitas menggunakan dua pendekatan, yaitu *Platt Scaling dan Isotonic Regression* [28]. Metode *Platt Scaling* memetakan skor prediksi model $f(x)$ ke dalam ruang probabilitas menggunakan fungsi logistik berikut:

$$\hat{p} = \frac{1}{1+e^{-(Af(x)+B)}}$$

Sebaliknya, *Isotonic Regression* dilakukan dengan mencari fungsi monotonik $g(x)$ yang meminimalkan selisih kuadrat antara probabilitas awal dan probabilitas hasil kalibrasi:

$$\hat{p} = \underset{\widehat{p_1} \le \widehat{p_2} \le \cdots \le \widehat{p_n}}{\arg} \; min \sum_{i=1}^{n} (p_i - \widehat{p_i})\, 2$$

Platt Scaling digunakan untuk data yang relatif homogen, sedangkan *Isotonic Regression* lebih sesuai untuk data berskala besar dan heterogen. Kalibrasi ini bertujuan memastikan bahwa probabilitas prediksi mencerminkan risiko klinis yang sebenarnya [29].

---

[25] Philipp Probst, Marvin Wright, and Anne-Laure Boulesteix, "Hyperparameters and Tuning Strategies for Random Forest," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9, no. 3 (February 26, 2019), https://doi.org/10.1002/widm.1301.

[26] Abdoul Malik, "Comparative Evaluation of Ensemble Learning Algorithms for Early Detection of Diabetes," *EDRAAK* 2025 (September 6, 2025): 103–10, https://doi.org/10.70470/EDRAAK/2025/013.

[27] Van Calster et al., "Calibration: The Achilles Heel of Predictive Analytics."

[28] Jing Li, "Area under the ROC Curve Has the Most Consistent Evaluation for Binary Classification," *PLOS ONE* 19, no. 12 (December 1, 2024): e0316019, https://doi.org/10.1371/JOURNAL.PONE.0316019.

[29] Bianca Zadrozny and Charles Elkan, "Transforming Classifier Scores into Accurate Multiclass Probability Estimates," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002, 694–99, https://doi.org/10.1145/775047.775151.

Evaluasi Kinerja dan Analisis Sensitivitas Biaya

Kinerja model dievaluasi menggunakan pendekatan *multi-metrik* yang mencakup AUC, *Precision-Recall* AUC (PR-AUC), *accuracy, precision, recall (sensitivitas), specificity, F1-score, Matthews Correlation Coefficient* (MCC), serta *Brier Score*. Pendekatan ini dipilih karena tidak ada satu metrik tunggal yang dapat merepresentasikan performa model secara komprehensif pada data kesehatan yang tidak seimbang [30]. Selain itu, dilakukan analisis sensitivitas biaya untuk menilai sejauh mana optimasi ambang keputusan mampu menurunkan total *misclassification cost*. Evaluasi berbasis biaya ini memastikan bahwa model tidak hanya unggul secara teknis, tetapi juga relevan secara klinis dan ekonomis [31].

Analisis dan Diskusi

Hasil Evaluasi Performa RF dan XGBoost

Hasil evaluasi performa model pada kedua dataset ditunjukkan pada Tabel 1 dan Tabel 2.

Table 1. Performa RF dan XGBoost Tanpa Mempertimbangkan Sensitivitas Biaya

Indikator	RF (Pima)	XGBoost (Pima)	RF (NHANES)	XGBoost (NHANES)
AUC	0.814	0.816	0.996	0.990
PR-AUC	0.693	0.689	0.993	0.974
Accuracy	0.734	0.740	0.985	0.971
Precision	0.603	0.625	0.956	0.914
Recall	0.704	0.648	0.979	0.957
Specificity	0.750	0.790	0.987	0.975
F1-score	0.650	0.636	0.967	0.935
MCC	0.440	0.435	0.958	0.917
Brier score	0.170	0.203	0.012	0.025

Tabel 1 menunjukkan performa dasar RF dan XGBoost tanpa mempertimbangkan sensitivitas biaya. Pada dataset Pima, kedua model menunjukkan kemampuan diskriminasi yang relatif seimbang, dengan AUC XGBoost sedikit lebih tinggi dibandingkan RF. Namun, RF menunjukkan nilai Recall dan PR-AUC yang lebih tinggi, mengindikasikan keseimbangan yang lebih baik antara pendeteksian kasus positif dan ketepatan prediksi pada kondisi ketidakseimbangan kelas.

Pada dataset NHANES, perbedaan performa antara kedua model menjadi lebih jelas. RF menunjukkan performa yang lebih stabil dengan nilai AUC, PR-AUC, MCC, dan Brier Score yang lebih baik dibandingkan XGBoost. Hasil ini mengindikasikan bahwa RF tidak

[30] Ryu and Sok, "Prediction Model of Quality of Life Using the Decision Tree Model in Older Adult Single-Person Households: A Secondary Data Analysis"; Victor Chang et al., "An Assessment of Machine Learning Models and Algorithms for Early Prediction and Diagnosis of Diabetes Using Health Indicators," *Healthcare Analytics* 2 (November 1, 2022): 100118, https://doi.org/10.1016/J.HEALTH.2022.100118.

[31] Parker et al., "Economic Costs of Diabetes in the U.S. in 2022"; Araf, Idri, and Chairi, "Cost-Sensitive Learning for Imbalanced Medical Data: A Review."

hanya unggul dalam diskriminasi kelas, tetapi juga menghasilkan probabilitas prediksi yang lebih konsisten pada populasi besar dan heterogen.

Setelah dilakukan penyesuaian ambang prediksi berbasis sensitivitas biaya, performa model berubah sebagaimana terlihat pada Tabel 2.

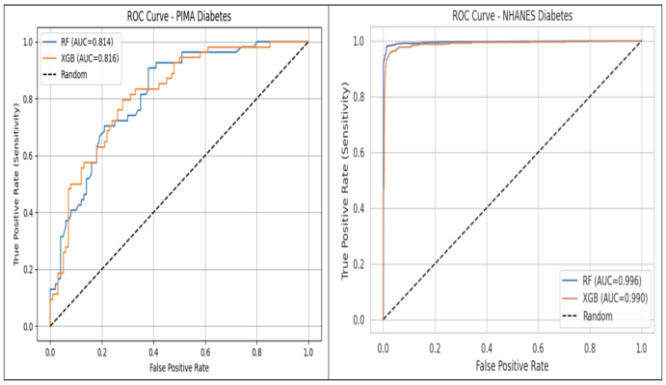
Table 2. Performa RF dan XGBoost Setelah Kalibrasi Probabilitas dengan Mempertimbangkan Sensitivitas Biaya

Indikator	RF (Pima)	XGBoost (Pima)	RF (NHANES)	XGBoost (NHANES)
Best Threshold	0.05	0.01	0.01	0.02
Accuracy	0.481	0.649	0.764	0.729
Precision	0.403	0.500	0.483	0.448
Recall	1.000	0.944	0.996	0.995
Specificity	0.200	0.490	0.996	0.995
F1-score	0.574	0.654	0.698	0.653
MCC	0.284	0.438	0.650	0.618

Tabel 2 memperlihatkan bahwa optimasi ambang berbasis biaya secara signifikan meningkatkan Recall pada kedua model, khususnya pada dataset Pima. Namun, peningkatan Recall tersebut diikuti oleh penurunan Specificity dan peningkatan False Positive, terutama pada RF. Temuan ini menegaskan adanya trade-off klinis yang tidak terhindarkan antara sensitivitas dan ketepatan klasifikasi, yang harus dipertimbangkan dalam konteks implementasi skrining.

Kurva ROC

Gambar 1 menunjukkan hubungan antara *True Positive Rate (Sensitivity)* dan *False Positive Rate* pada berbagai nilai ambang prediksi.

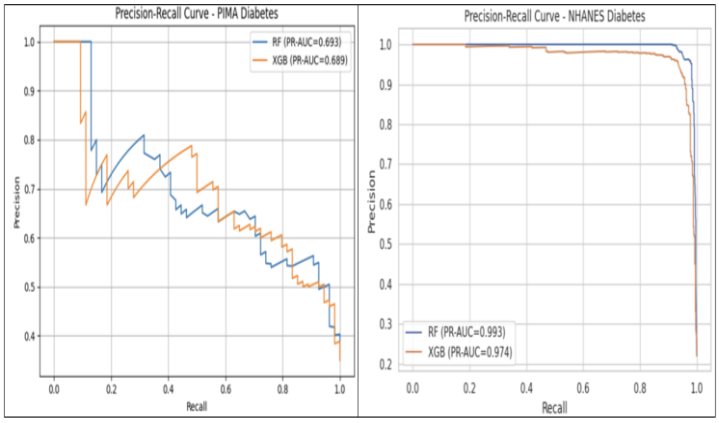


Gambar 1. Kurva ROC Dataset Pima dan NHANES

Kurva ROC memperlihatkan bahwa pada dataset Pima, XGBoost memiliki keunggulan marginal dalam AUC dibandingkan RF. Namun, pada dataset NHANES, RF menunjukkan performa yang lebih konsisten dengan kurva yang mendekati sudut kiri atas, mencerminkan kemampuan diskriminasi yang sangat baik pada populasi besar.

Kurva Precision-Recall

Gambar 2 menunjukkan hubungan antara nilai *Precision* dan *Recall* pada nilai ambang prediksi.

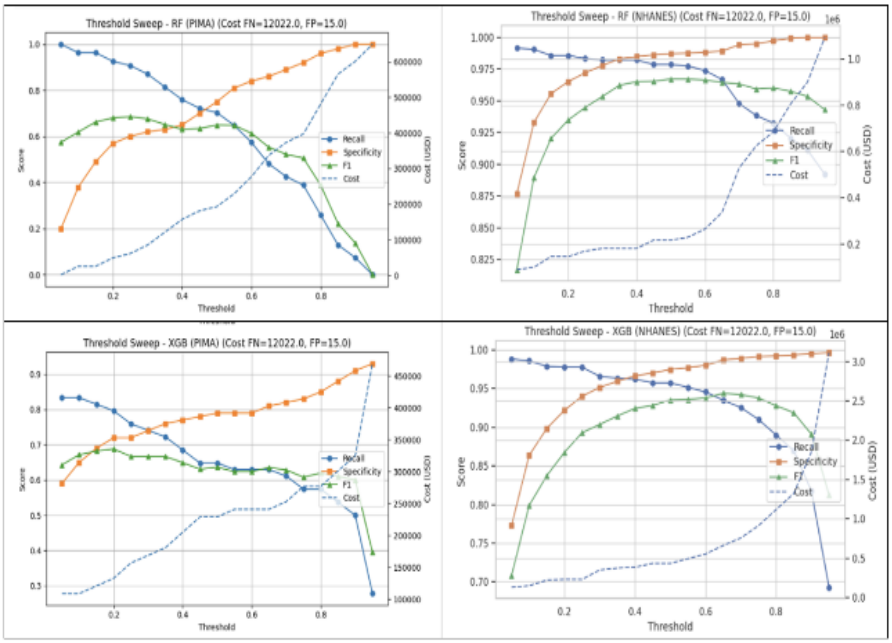


Gambar 2. Kurva PR Dataset Pima dan NHANES

Pada dataset Pima, RF menunjukkan PR-AUC yang sedikit lebih tinggi dibandingkan XGBoost, menandakan kemampuan yang lebih baik dalam mempertahankan *Precision* ketika *Recall* meningkat. Pada dataset NHANES, keunggulan RF menjadi lebih jelas dengan PR-AUC yang lebih tinggi dan kurva yang relatif stabil di hampir seluruh rentang *Recall*. Hal ini penting dalam konteks klinis, karena kestabilan *Precision* pada *Recall* tinggi berkontribusi pada penurunan risiko FN.

Analisis Threshold Sweep

Gambar 3 menampilkan hasil analisis perubahan nilai *threshold sweep* untuk kedua model pada dataset Pima dan NHANES.



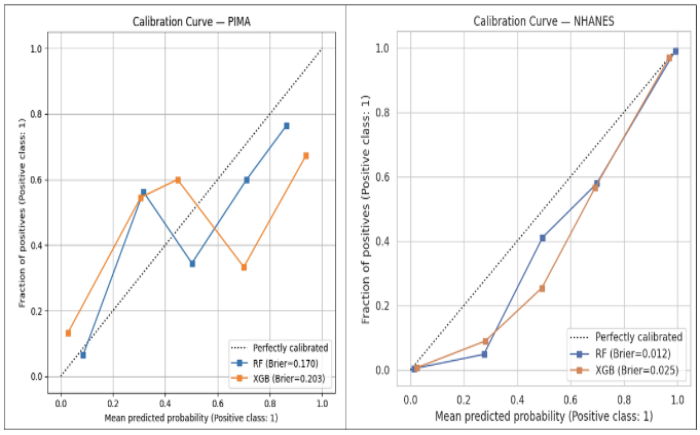
Gambar 3. Kurva Threshold Sweep Pima dan NHANES

Pada dataset Pima, RF mencapai titik biaya minimum pada threshold sekitar 0,05, di mana Recall meningkat secara signifikan sebelum biaya kembali meningkat pada threshold ekstrem. XGB menunjukkan pola serupa, tetapi dengan fluktuasi F1-score yang lebih besar, mengindikasikan stabilitas yang lebih rendah pada data kecil dan homogen.

Pada dataset NHANES, kedua model mempertahankan Recall yang sangat tinggi pada threshold rendah, namun RF menunjukkan penurunan biaya yang lebih konsisten dibandingkan XGB. Temuan ini menguatkan argumen bahwa RF lebih stabil dalam mengelola trade-off Recall–Specificity pada populasi besar. Dengan mempertimbangkan bahwa kesalahan FN memiliki konsekuensi klinis dan ekonomi yang lebih serius, RF lebih layak diprioritaskan untuk skenario skrining berbasis populasi.

Kurva Kalibrasi

Gambar 4 menunjukkan hubungan antara probabilitas prediksi rata-rata dan proporsi aktual kasus positif. sedangkan sumbu vertikal menunjukkan proporsi aktual sampel yang termasuk dalam kelas positif pada setiap interval probabilitas.



Gambar 4. Kurva Kalibrasi

Pada dataset Pima, kurva kalibrasi kedua model menunjukkan fluktuasi yang cukup besar, yang kemungkinan dipengaruhi oleh ukuran sampel yang relatif kecil. RF memiliki Brier Score yang lebih rendah dibandingkan XGB, menunjukkan estimasi probabilitas yang lebih mendekati risiko aktual.

Pada dataset NHANES, kualitas kalibrasi meningkat secara signifikan pada kedua model, dengan RF kembali menunjukkan konsistensi probabilitas yang lebih baik. Kurva RF lebih dekat dengan garis kalibrasi sempurna dan memiliki Brier Score yang lebih rendah, menandakan reliabilitas probabilistik yang lebih tinggi pada populasi besar dan heterogen.

Interpretasi ROC dan Precision-Recall

Gambar 1 ROC dan Gambar 2 *Precision-Recall* pada kedua dataset menunjukkan perbedaan karakteristik performa antara RF dan XGBoost dalam tugas klasifikasi diabetes. Pada dataset Pima, nilai AUC pada ROC Curve menunjukkan bahwa XGBoost memperoleh AUC = 0,816, sedikit lebih tinggi dibanding RF dengan AUC = 0,814, yang mengindikasikan kemampuan diskriminatif XGBoost sedikit lebih unggul dalam membedakan pasien dengan dan tanpa diabetes. Namun, pada *Precision-Recall Curve*, RF menunjukkan performa yang lebih baik dengan PR-AUC = 0,693, dibanding XGBoost yang memperoleh PR-AUC = 0,689, menandakan keseimbangan yang lebih baik antara *Recall* dan *Precision* pada kondisi ketidakseimbangan kelas.

Perlu dicermati bahwa pada dataset berukuran kecil dan relatif homogen seperti Pima, peningkatan performa pada metrik berbasis sensitivitas juga berpotensi dipengaruhi oleh penerapan SMOTE pada data latih. Pada dataset NHANES, RF menunjukkan performa

yang lebih unggul secara konsisten. Gambar 1 memperlihatkan bahwa RF mencapai AUC = 0,996, lebih tinggi dibanding XGBoost yang memperoleh AUC = 0,990, menandakan kemampuan diskriminasi hampir sempurna pada populasi besar. Gambar 2 juga menunjukkan bahwa RF memperoleh PR-AUC = 0,993, lebih tinggi daripada XGBoost yang mencapai PR-AUC = 0,974. Kestabilan nilai *Precision* pada hampir seluruh rentang *Recall* pada dataset NHANES menunjukkan kemampuan generalisasi yang lebih kuat, yang lebih mencerminkan kondisi populasi heterogen dan mendekati skenario implementasi dunia nyata.

Efektivitas Penyesuaian Ambang Threshold Sweep

Gambar 3 menunjukkan bahwa pada dataset Pima, penurunan *threshold* pada RF hingga sekitar 0,05 meningkatkan *Recall* secara signifikan dan menghasilkan total *cost* terendah sebelum biaya meningkat kembali pada *threshold* ekstrem. Pola serupa terlihat pada XGBoost, namun titik stabilnya berada pada sekitar 0,10, dengan fluktuasi *F1-Score* yang lebih besar. Hal ini menunjukkan bahwa RF lebih stabil dalam menjaga keseimbangan sensitivitas dan ketepatan prediksi pada dataset yang kecil dan homogen.

Penyesuaian *cost-sensitive threshold optimization* terbukti meningkatkan sensitivitas model pada kedua dataset. Namun demikian, nilai *threshold* optimal yang diperoleh bersifat sangat kontekstual terhadap asumsi biaya kesalahan klasifikasi serta karakteristik data yang digunakan. Dalam praktik klinis nyata, ambang keputusan tidak dapat diterapkan secara universal tanpa mempertimbangkan prevalensi lokal diabetes, kapasitas layanan kesehatan, serta kebijakan *skrining* yang berlaku. Oleh karena itu, hasil *threshold sweep* dalam penelitian ini lebih tepat dipahami sebagai kerangka pendukung pengambilan keputusan berbasis risiko, bukan sebagai nilai ambang baku yang siap diterapkan tanpa adaptasi kontekstual.

Dengan mempertimbangkan bahwa kesalahan FN memiliki konsekuensi klinis dan ekonomi yang jauh lebih besar dibandingkan FP, serta mempertimbangkan stabilitas model dalam berbagai rentang *threshold*, RF lebih direkomendasikan sebagai model utama untuk skrining risiko diabetes dan penerapan sistem pendukung keputusan klinis.

Interpretasi Kalibrasi Probabilitas

Gambar 4 menunjukkan kualitas kalibrasi probabilitas pada kedua model. Pada dataset Pima, pola kalibrasi kedua model tampak berfluktuasi akibat ukuran sampel yang relatif kecil, yang membatasi stabilitas estimasi probabilitas. Model RF memiliki nilai *Brier Score* sebesar 0,170, lebih rendah dibanding XGB dengan nilai 0,203, menunjukkan bahwa RF menghasilkan estimasi probabilitas yang lebih reliabel dan lebih mendekati proporsi aktual kasus positif.

Pada dataset NHANES, performa kalibrasi meningkat signifikan seiring ukuran data yang lebih besar dan distribusi populasi yang lebih representatif. RF menunjukkan *Brier Score* sebesar 0,012, sedangkan XGB menunjukkan nilai 0,025, menandakan kedua model terkalibrasi sangat baik, namun RF tetap menunjukkan konsistensi probabilitas yang lebih tinggi. Meskipun demikian, penerapan probabilitas terkalibrasi dalam sistem pendukung keputusan klinis di dunia nyata tetap memerlukan validasi prospektif dan integrasi dengan alur kerja klinis, agar estimasi risiko yang dihasilkan benar-benar selaras dengan pengambilan keputusan medis.

Kesimpulan

Penelitian ini membandingkan RF dan XGBoost dalam prediksi risiko diabetes pada dua karakteristik populasi, yaitu Pima dan NHANES. Hasil menunjukkan bahwa RF memberikan performa *precision–recall* yang lebih konsisten serta kualitas kalibrasi probabilitas yang lebih baik, khususnya pada populasi heterogen. Optimasi ambang berbasis biaya terbukti meningkatkan sensitivitas model dan menurunkan total biaya kesalahan klasifikasi hingga sekitar 20–25%, meskipun dengan konsekuensi peningkatan FP yang perlu dikendalikan. XGBoost menunjukkan keunggulan marginal pada dataset kecil untuk kemampuan diskriminasi murni, namun RF lebih stabil dan efisien secara klinis dan ekonomis pada skenario skrining populasi besar. Integrasi pembelajaran sensitif-biaya dan kalibrasi probabilitas terbukti krusial dalam meningkatkan keterterapan model prediksi diabetes sebagai sistem pendukung keputusan berbasis risiko.

Meskipun hasil penelitian menunjukkan peningkatan performa dan efisiensi biaya, studi ini memiliki beberapa keterbatasan. Pertama, penggunaan teknik *oversampling* seperti SMOTE berpotensi menghasilkan sampel sintetis yang tidak sepenuhnya merepresentasikan variasi klinis nyata, sehingga peningkatan performa pada data latih perlu ditafsirkan secara hati-hati dalam konteks validitas klinis, khususnya ketika model diarahkan untuk aplikasi skrining di dunia nyata. Kedua, evaluasi dilakukan secara retrospektif pada dataset sekunder, sehingga validitas prediktif model dalam konteks klinis nyata belum dapat dipastikan. Penelitian selanjutnya perlu melakukan validasi prospektif serta menguji integrasi model ke dalam alur kerja klinis aktual, termasuk penyesuaian ambang keputusan berdasarkan prevalensi lokal dan kapasitas layanan kesehatan. Pendekatan ini penting untuk memastikan bahwa model prediksi tidak hanya unggul secara statistik, tetapi juga berdaya guna dalam praktik klinis dan kebijakan kesehatan masyarakat.

Daftar Pustaka

Ahsan, Md Manjurul, M. A.Parvez Mahmud, Pritom Kumar Saha, Kishor Datta Gupta, and Zahed Siddique. "Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance." *Technologies 2021, Vol. 9, Page 52* 9, no. 3 (July 24, 2021): 52. https://doi.org/10.3390/TECHNOLOGIES9030052.

Altalhan, Manahel, Abdulmohsen Algarni, and Monia Turki-Hadj Alouane. "Imbalanced Data Problem in Machine Learning: A Review." *IEEE Access* 13 (2025): 13686–99. https://doi.org/10.1109/ACCESS.2025.3531662.

Araf, Imane, Ali Idri, and Ikram Chairi. "Cost-Sensitive Learning for Imbalanced Medical Data: A Review." *Artificial Intelligence Review 2024 57:4* 57, no. 4 (March 1, 2024): 1–72. https://doi.org/10.1007/S10462-023-10652-8.

Calster, Ben Van, David J. McLernon, Maarten Van Smeden, Laure Wynants, Ewout W. Steyerberg, Patrick Bossuyt, Gary S. Collins, Petra MacAskill, Karel G.M. Moons, and Andrew J. Vickers. "Calibration: The Achilles Heel of Predictive Analytics." *BMC Medicine* 17, no. 1 (December 16, 2019). https://doi.org/10.1186/S12916-019-1466-7.

Chang, Victor, Meghana Ashok Ganatra, Karl Hall, Lewis Golightly, and Qianwen Ariel Xu. "An Assessment of Machine Learning Models and Algorithms for Early Prediction and

Diagnosis of Diabetes Using Health Indicators." *Healthcare Analytics* 2 (November 1, 2022): 100118. https://doi.org/10.1016/J.HEALTH.2022.100118.

Dhenabayu, Riska, Hujjatullah Fazlurrahman, and Purwohandoko. "Potential Researches of GAN in Fashion Areas." *Proceedings - 2023 6th International Conference on Computer and Informatics Engineering: AI Trust, Risk and Security Management (AI Trism), IC2IE 2023*, 2023, 276–81. https://doi.org/10.1109/IC2IE60547.2023.10331411.

Dhenabayu, Riska, Nanang Hoesen Hidroes Abbrori, Hujjatullah Fazlurrahman, and Achmad Fitro. "Harnessing the Power of Transformer Networks in AI-Driven Decision Support Systems for Badminton Action Recognition." *2024 12th International Conference on Cyber and IT Service Management, CITSM 2024*, 2024. https://doi.org/10.1109/CITSM64103.2024.10775332.

Ganie, Shahid Mohammad, Pijush Kanti Dutta Pramanik, Majid Bashir Malik, Saurav Mallik, and Hong Qin. "An Ensemble Learning Approach for Diabetes Prediction Using Boosting Techniques." *Frontiers in Genetics* 14 (2023): 1252159. https://doi.org/10.3389/FGENE.2023.1252159.

IDF. "Diabetes Facts and Figures | International Diabetes Federation," 2025. https://idf.org/about-diabetes/diabetes-facts-figures/.

———. "The Diabetes Atlas | Global Diabetes Data & Statistics." International Diabetes Federation (IDF), 2025. https://diabetesatlas.org/.

Joel, Luke Oluwaseye, Wesley Doorsamy, and Babu Sena Paul. "A Comparative Study of Imputation Techniques for Missing Values in Healthcare Diagnostic Datasets." *International Journal of Data Science and Analytics 2025 20:7* 20, no. 7 (June 11, 2025): 6357–73. https://doi.org/10.1007/S41060-025-00825-9.

Leevy, Joffrey L., Justin M. Johnson, John Hancock, and Taghi M. Khoshgoftaar. "Threshold Optimization and Random Undersampling for Imbalanced Credit Card Data." *Journal of Big Data 2023 10:1* 10, no. 1 (May 6, 2023): 1–22. https://doi.org/10.1186/S40537-023-00738-Z.

Li, Jing. "Area under the ROC Curve Has the Most Consistent Evaluation for Binary Classification." *PLOS ONE* 19, no. 12 (December 1, 2024): e0316019. https://doi.org/10.1371/JOURNAL.PONE.0316019.

Liou, Lathan, Erick Scott, Prathamesh Parchure, Yuxia Ouyang, Natalia Egorova, Robert Freeman, Ira S. Hofer, et al. "Assessing Calibration and Bias of a Deployed Machine Learning Malnutrition Prediction Model within a Large Healthcare System." *Npj Digital Medicine 2024 7:1* 7, no. 1 (June 6, 2024): 149-. https://doi.org/10.1038/s41746-024-01141-5.

Malik, Abdoul. "Comparative Evaluation of Ensemble Learning Algorithms for Early Detection of Diabetes." *EDRAAK* 2025 (September 6, 2025): 103–10. https://doi.org/10.70470/EDRAAK/2025/013.

Parker, Emily D., Janice Lin, Troy Mahoney, Nwanneamaka Ume, Grace Yang, Robert A. Gabbay, Nuha A. Elsayed, and Raveendhara R. Bannuru. "Economic Costs of Diabetes in the U.S. in 2022." *Diabetes Care* 47, no. 1 (January 2, 2024): 26–43. https://doi.org/10.2337/DCI23-0085.

Pedregosa, Fabian, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion,

Olivier Grisel, Mathieu Blondel, et al. "Scikit-Learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot." *Journal of Machine Learning Research* 12 (2011): 2825–30.

Petridis, Panagiotis D., Aleksandra S. Kristo, Angelos K. Sikalidis, and Ilias K. Kitsas. "A Review on Trending Machine Learning Techniques for Type 2 Diabetes Mellitus Management." *Informatics 2024, Vol. 11, Page 70* 11, no. 4 (September 27, 2024): 70. https://doi.org/10.3390/INFORMATICS11040070.

Probst, Philipp, Marvin Wright, and Anne-Laure Boulesteix. "Hyperparameters and Tuning Strategies for Random Forest." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9, no. 3 (February 26, 2019). https://doi.org/10.1002/widm.1301.

Ryu, Dajung, and Sohyune Sok. "Prediction Model of Quality of Life Using the Decision Tree Model in Older Adult Single-Person Households: A Secondary Data Analysis." *Frontiers in Public Health* 11 (August 31, 2023): 1224018. https://doi.org/10.3389/FPUBH.2023.1224018/BIBTEX.

Thai-Nghe, Nguyen, Zeno Gantner, and Lars Schmidt-Thieme. "Cost-Sensitive Learning Methods for Imbalanced Data." *Proceedings of the International Joint Conference on Neural Networks*, 2010. https://doi.org/10.1109/IJCNN.2010.5596486.

Tougui, Ilias, Abdelilah Jilbab, and Jamal El Mhamdi. "Impact of the Choice of Cross-Validation Techniques on the Results of Machine Learning-Based Diagnostic Applications." *Healthcare Informatics Research* 27, no. 3 (July 1, 2021): 189. https://doi.org/10.4258/HIR.2021.27.3.189.

Vrudhula, Amey, Alan C. Kwan, David Ouyang, and Susan Cheng. "Machine Learning and Bias in Medical Imaging: Opportunities and Challenges." *Circulation. Cardiovascular Imaging* 17, no. 2 (February 1, 2024): e015495. https://doi.org/10.1161/CIRCIMAGING.123.015495.

WHO. "Diabetes." WHO, 2024. https://www.who.int/news-room/fact-sheets/detail/diabetes.

Zadrozny, Bianca, and Charles Elkan. "Transforming Classifier Scores into Accurate Multiclass Probability Estimates." *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002, 694–99. https://doi.org/10.1145/775047.775151.

Zheng, Alice, and Amanda Casari. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. *ACM Computing Surveys*. First Edit. Vol. 53. Sebastopol, CA: O'Reilly Media, Inc., 2018.