



Machine Learning-Based Prediction of Diabetes Medication Use: A Pharmacoepidemiological Analysis of the Indonesia Family Life Survey

Husnul Khuluq¹, Imam Tri Suryadin², Tri Cahyani¹, Ayu Nissa Ainni¹

¹Department of Pharmacy, Faculty of Health Science, Universitas Muhammadiyah Gombong, Sangkalputung, Gombong, Kebumen, Jawa Tengah 54411, Indonesia

²Department Computer Science, Faculty of Science and Humaniora, Universitas Muhammadiyah Gombong, Sangkalputung, Gombong, Kebumen, Jawa Tengah 54411, Indonesia

*husnulkhuluq@unimugo.ac.id

ABSTRACT

Diabetes pharmacotherapy frequently coexists with treatment for cardiovascular and metabolic comorbidities, highlighting the potential role of data-driven methods in pharmaceutical care. This study aimed to develop and compare machine learning models for predicting diabetes medication use using the fifth wave of the Indonesia Family Life Survey (IFLS5), conducted in Indonesia in 2014–2015. A secondary analysis included 48,139 participants with valid outcome information, of whom 353 (0.73%) reported diabetes medication use. Predictors included age, sex, pregnancy status, weight, height, hemoglobin, systolic and diastolic blood pressure, pulse, and medication use for anemia, hypertension, and hypercholesterolemia. Class imbalance was addressed using Synthetic Minority Over-sampling Technique for Nominal and Continuous features (SMOTENC) exclusively within training folds. LASSO-based feature selection was also performed within stratified 10-fold cross-validation. Random Forest, linear Support Vector Machine (SVM), Logistic Regression, and Gaussian Naive Bayes were compared. Logistic Regression achieved 80.13% accuracy, 76.20% sensitivity, 80.16% specificity, and an AUC of 0.870. Linear SVM showed comparable discrimination (AUC=0.870). Random Forest achieved high accuracy (97.86%) but low sensitivity (18.13%), whereas Naive Bayes achieved high sensitivity (98.30%) but low specificity (26.77%). LASSO consistently identified pregnancy status, cholesterol medication use, hypertension medication use, and weight among the strongest predictors. Logistic Regression provided the most balanced classification performance and may support further pharmacoepidemiological investigation of diabetes medication-use patterns.

Keywords: diabetes medication; LASSO; machine learning; pharmacoepidemiology; SMOTENC

How to cite (in APA style)

Khuluq, H., Suryadin, I. T., Widiastuti, T. C., & Ainni, A. N. (2026). Machine Learning-Based Prediction of Diabetes Medication Use: A Pharmacoepidemiological Analysis of the Indonesia Family Life Survey. *Indonesian Journal of Global Health Research*, 8(6), 631–640. <https://doi.org/10.37287/ijghr.v8i6.2740>.

INTRODUCTION

Diabetes mellitus represents an important and increasing burden on healthcare systems worldwide. Indonesia is among countries with a substantial diabetes population, creating challenges not only for disease detection but also for long-term pharmacotherapy management. The increasing burden of diabetes consequently has implications for pharmacists and other healthcare professionals involved in medication management, adherence support, monitoring, and prevention of treatment-related problems (Rastogi & Bansal, 2023). From a pharmaceutical perspective, diabetes treatment rarely occurs in isolation. Patients receiving diabetes medication may also receive pharmacotherapy for cardiovascular and metabolic comorbidities, particularly hypertension and dyslipidemia. Medication-use patterns may therefore reflect the combined influence of disease burden, multimorbidity, healthcare access, prescribing practices, and continued engagement with treatment. These characteristics make diabetes medication use an important outcome for pharmacoepidemiological investigation (Alwhaibi, 2022).

Pharmacists contribute to diabetes management through medication review, patient education, adherence assessment, monitoring of treatment outcomes, and identification and resolution of medication-related problems. Pharmacist-led interventions have been associated with improvements in medication adherence and clinical outcomes among patients with diabetes, including evidence from Indonesian healthcare settings. Identification of medication-use patterns at the population level may therefore provide additional information relevant to the development and prioritization of pharmaceutical-care services(Aman & Chhillar, 2023).

Population-based surveys provide an opportunity to investigate medication use outside individual hospitals or clinics. The Indonesia Family Life Survey (IFLS) contains extensive demographic, socioeconomic, and health-related information. Its fifth wave, IFLS5, was conducted in 2014–2015 and included physical health measurements and medication-related information that can be integrated for predictive analysis(Sardu et al., n.d.). Machine learning can complement conventional statistical approaches by identifying multidimensional patterns associated with health outcomes. Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), and Naive Bayes (NB) represent classification approaches with different assumptions and decision mechanisms. However, predictive performance can be misleading when the outcome is highly imbalanced. In such circumstances, a model may achieve high overall accuracy by predominantly predicting the majority class while failing to detect observations belonging to the clinically relevant minority class(Sardu et al., n.d.).

Synthetic Minority Over-sampling Technique (SMOTE) and related approaches can reduce this problem by increasing representation of the minority class during model training. When both continuous and categorical predictors are present, SMOTENC provides an extension appropriate for mixed feature types. Importantly, oversampling should be restricted to training data within each cross-validation iteration to avoid information leakage(Elreedy et al., 2024a). Feature selection is another important component of predictive modeling. Least Absolute Shrinkage and Selection Operator (LASSO) uses L1 regularization to shrink less informative coefficients toward zero, providing a mechanism for reducing model complexity and identifying informative predictors. Nevertheless, variables retained by LASSO are predictive features rather than necessarily causal risk factors(Ou et al., 2023).

The present study aimed to develop and compare RF, linear SVM, LR, and Gaussian NB models for predicting diabetes medication use using IFLS5 data. The analytical strategy incorporated training-fold SMOTENC, LASSO-based feature selection, and stratified 10-fold cross-validation. The study also examined the stability of selected predictors, with particular attention to medication-related variables relevant to pharmacoepidemiology and pharmaceutical care(Ou et al., 2023). This study aimed to identify factors associated with self-reported diabetes medication use using LASSO-based feature selection and to compare the performance of machine learning algorithms for classifying diabetes medication use in the Indonesia Family Life Survey Wave 5 (IFLS5), while addressing severe class imbalance using SMOTE.

METHOD

Study Design and Data Source

This study was a secondary analysis of data from the fifth wave of the Indonesia Family Life Survey (IFLS5). IFLS is a longitudinal household survey containing socioeconomic, demographic, household, and health information from Indonesian individuals and households. IFLS5 field data were collected during 2014–2015. Health and social datasets were linked at the individual level using corresponding respondent identifiers. Individual and household identifiers were used exclusively for linkage and were excluded from predictive modeling. The analysis used de-identified secondary data. The appropriate documentation should be attached separately during journal submission.

Study Population and Outcome

The merged dataset initially contained 52,580 observations. Participants without valid information for the target outcome were excluded. The final analytical sample consisted of 48,139 participants, including 353 (0.73%) positive observations and 47,786 (99.27%) negative observations (Ullah et al., 2022). The outcome was self-reported diabetes medication use. Based on the coding of the corresponding IFLS questionnaire item, “Yes” was classified as the positive outcome and “No” as the negative outcome (Ullah et al., 2022). Importantly, this variable represents reported medication use rather than clinically confirmed diabetes. Therefore, the resulting models should be interpreted as predicting diabetes medication use and not as diagnostic models for diabetes (Al Assaf et al., 2022).

Predictor Variables

Predictors were selected according to their availability in the merged IFLS dataset and their potential relevance to demographic characteristics, physical health, and pharmacotherapy. Continuous predictors included weight (kg), height (cm), hemoglobin, systolic blood pressure, diastolic blood pressure, pulse rate, and age. Categorical predictors included pregnancy status, sex, anemia medication use, hypertension medication use, and cholesterol medication use (“9. Pharmacologic Approaches to Glycemic Treatment: Standards of Care in Diabetes—2024,” 2024). Pregnancy status was represented as “Yes,” “No,” or “Not applicable.” Sex was categorized as male or female. Medication-related predictors represented reported use or non-use of medication for the corresponding condition. Identifiers were not used as predictors (Ullah et al., 2022).

Data Preprocessing and Class Imbalance

All data-dependent preprocessing was performed independently within the training portion of each cross-validation iteration to minimize information leakage. Missing continuous values were imputed using the median calculated from the corresponding training fold, while missing categorical values were imputed using the training-fold mode. Because the dataset contained both numerical and categorical predictors, SMOTENC was used to address class imbalance. Oversampling was applied exclusively to training observations and was never applied to the validation fold (Coutureau et al., 2022). After SMOTENC, continuous predictors were standardized and categorical predictors were transformed using one-hot encoding. The transformations fitted on the training data were subsequently applied to the corresponding untouched validation observations (Coutureau et al., 2022). The analytical sequence for each cross-validation iteration was therefore: training fold → imputation → SMOTENC → standardization/one-hot encoding → LASSO feature selection → model fitting → untouched validation fold (Ou et al., 2023).

LASSO Feature Selection

LASSO-style feature selection was performed using L1-regularized Logistic Regression. Features with non-zero coefficients were retained for subsequent model development. Feature selection was conducted separately within each training fold rather than on the complete dataset. This prevented information from validation observations from influencing feature selection (Kwon et al., 2023). Selection stability was evaluated by determining how frequently each transformed predictor was retained across the ten folds. Mean absolute coefficients ($|\beta|$) were used to rank the contribution of predictors. These coefficients were interpreted as predictive model parameters and not as causal effects or conventional epidemiological odds ratios (Kwon et al., 2023).

Machine Learning Models

Four machine learning classifiers were evaluated: Random Forest, linear Support Vector Machine, Logistic Regression, and Gaussian Naive Bayes (Wang et al., n.d.). Random Forest represents an ensemble tree-based algorithm capable of capturing nonlinear relationships and interactions. Linear SVM identifies a separating hyperplane that maximizes the margin between classes. Logistic

Regression estimates outcome probability using a linear combination of predictors, while Gaussian Naive Bayes estimates class probabilities under conditional-independence assumptions (Wang et al., n.d.). The same cross-validation framework was applied to all four models.

Cross-Validation and Performance Evaluation

Internal validation was conducted using stratified 10-fold cross-validation. The data were divided into ten folds while approximately preserving the distribution of positive and negative outcomes. Nine folds were used for model training and the remaining fold for validation. This procedure was repeated until every participant had served as a validation observation once. Predictions obtained when participants were in held-out validation folds were combined to generate out-of-fold predictions (Vanacore et al., 2024). Performance was evaluated using accuracy, sensitivity, specificity, confusion matrices, and the area under the receiver operating characteristic curve (AUC). Because only 0.73% of participants belonged to the positive class, model interpretation emphasized sensitivity, specificity, and ROC-AUC rather than accuracy alone (Kwon et al., 2023). Data processing and machine learning analyses were conducted using Python and established machine-learning libraries. Fixed random states were used where applicable to improve computational reproducibility (Vanacore et al., 2024).

RESULT

Study Population

A total of 48,139 participants had valid diabetes medication-use information and were included in model development. Among them, 353 (0.73%) reported diabetes medication use, whereas 47,786 (99.27%) did not. This highly imbalanced outcome distribution justified the use of training-fold SMOTENC and performance measures beyond overall accuracy (Ullah et al., 2022).

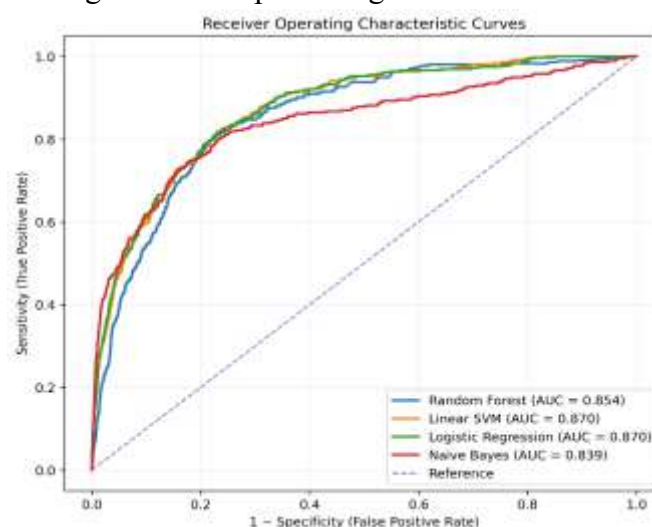
Performance of Machine Learning Models

Performance of the four machine learning algorithms is summarized in Table 1, while discrimination across classification thresholds is presented using ROC curves in Figure 1

Table 1.

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	ROC-AUC
Random Forest	97.86	18.13	98.45	0.854
Linear Support Vector Machine	80.28	75.92	80.31	0.870
Logistic Regression	80.13	76.20	80.16	0.870
Gaussian Naive Bayes	27.30	98.30	26.77	0.839

Performance of machine learning models for predicting diabetes medication use



Abbreviation: AUC, area under the receiver operating characteristic curve.

Figure 1. discrimination across classification thresholds is presented using ROC curves

Performance of Machine Learning Models

Performance of the four machine learning algorithms is summarized in Table 1, while discrimination across classification thresholds is presented using ROC curves in Figure 1. The four algorithms showed substantially different sensitivity-specificity profiles. Random Forest achieved the highest accuracy (97.86%) and specificity (98.45%) but had a sensitivity of only 18.13%. Thus, despite its apparently excellent accuracy, Random Forest failed to identify a substantial proportion of participants reporting diabetes medication use. Its out-of-fold ROC-AUC was 0.854 (Alwhaibi, 2022). Linear SVM provided a more balanced classification profile, with 80.28% accuracy, 75.92% sensitivity, and 80.31% specificity. Its ROC-AUC was 0.870, demonstrating good discrimination between participants reporting and not reporting diabetes medication use. Logistic Regression achieved 80.13% accuracy, 76.20% sensitivity, and 80.16% specificity. Its ROC-AUC was also 0.870. The close balance between sensitivity and specificity indicated that LR did not disproportionately favor either outcome class and provided the most clinically interpretable balance among the evaluated models (Al Assaf et al., 2022). Gaussian Naive Bayes demonstrated the opposite pattern to Random Forest. It achieved the highest sensitivity (98.30%) but only 26.77% specificity and 27.30% accuracy. Its ROC-AUC was 0.839. Consequently, although NB detected nearly all positive observations at the applied threshold, it incorrectly classified a large proportion of negative observations as positive (Collins et al., 2024). Overall, Table 1 and Figure 2 demonstrate that accuracy alone did not adequately characterize model performance. Random Forest favored negative-class identification, whereas Naive Bayes strongly favored positive-class detection. Linear SVM and Logistic Regression achieved considerably more balanced performance, with Logistic Regression selected as the preferred model because of its combination of discrimination, sensitivity, specificity, and interpretability (Collins et al., 2024).

LASSO Feature Selection

LASSO-based feature selection demonstrated substantial stability across the ten cross-validation folds (Table 2).

Table 2.
Stability of predictors selected using LASSO across stratified 10-fold cross-validation

Predictor	Folds selected	Stability (%)	Mean coefficient (β)
Pregnancy = Yes	10/10	100	-4.144
Cholesterol medication = Yes	10/10	100	2.228
Pregnancy = Not applicable	10/10	100	1.737
Hypertension medication = Yes	10/10	100	1.634
Weight	10/10	100	1.119
Systolic blood pressure	10/10	100	0.802
Pulse	10/10	100	0.775
Diastolic blood pressure	10/10	100	-0.478
Hemoglobin	10/10	100	-0.476
Female sex	10/10	100	0.289
Age	10/10	100	-0.123
Height	10/10	100	-0.018
Anemia medication = Yes	9/10	90	-0.472

LASSO-based feature selection demonstrated substantial stability across the ten cross-validation folds (Table 2). Pregnancy status, cholesterol medication use, hypertension medication use, weight, systolic blood pressure, pulse, diastolic blood pressure, hemoglobin, sex, age, and height were retained in all ten folds. Anemia medication use was selected in nine of ten folds (Kwon et al., 2023). The five largest predictors ranked by mean absolute LASSO coefficient are presented in Figure 1. Pregnancy “Yes” had the largest absolute coefficient ($|\beta|=4.144$), followed by cholesterol medication use ($|\beta|=2.228$), pregnancy “Not applicable” ($|\beta|=1.737$), hypertension medication use ($|\beta|=1.634$), and weight ($|\beta|=1.119$). Each of these predictors was retained in all ten cross-validation

folds(Cabot & Ross, 2023). The coefficients describe contributions to the regularized predictive model. They should not be interpreted as evidence of causal or protective effects.

DISCUSSION

This study developed and compared four machine learning algorithms for predicting diabetes medication use using a large Indonesian population dataset. Logistic Regression and linear SVM demonstrated the most balanced predictive performance, with both producing ROC-AUC values of approximately 0.87. Random Forest achieved very high accuracy and specificity but poor sensitivity, whereas Naive Bayes demonstrated very high sensitivity accompanied by poor specificity. These findings highlight the importance of considering multiple performance measures when developing prediction models for highly imbalanced pharmacoepidemiological outcomes(Welvaars et al., 2023).

The performance of Random Forest provides a clear illustration of the limitations of accuracy as a primary performance metric. Because approximately 99.3% of observations belonged to the negative class, a classifier could achieve high overall accuracy largely by correctly predicting the majority class. Random Forest achieved 97.86% accuracy but identified only 18.13% of positive observations. Such performance would be problematic if the intended purpose were to identify individuals reporting diabetes medication use(Piccialli & Sciandrone, 2022). Conversely, Naive Bayes identified 98.30% of positive observations but had a specificity of only 26.77%. Such a model would generate a large number of false-positive classifications. Linear SVM and Logistic Regression demonstrated more balanced sensitivity and specificity. The ROC curves in Figure 2 further support the discriminatory performance of these linear models(Collins et al., 2024).

Logistic Regression is particularly relevant when interpretability is important. Although increasingly complex machine learning methods are frequently proposed for health prediction, more complex algorithms do not necessarily outperform conventional statistical learning methods in structured datasets. For pharmaceutical applications, interpretability may also facilitate communication of model behavior among pharmacists, physicians, researchers, and other healthcare professionals (Elreedy et al., 2024b). The medication-related predictors identified by LASSO provide an additional pharmacoepidemiological perspective. Cholesterol medication use and hypertension medication use were among the five predictors with the largest mean absolute coefficients and were selected consistently across all ten folds. Their contribution is plausible within the broader context of cardiometabolic multimorbidity. Individuals receiving diabetes pharmacotherapy may simultaneously receive treatment for hypertension and dyslipidemia, reflecting a more complex chronic-disease and medication-use profile(Alfian et al., 2026).

Nevertheless, these findings must not be interpreted as evidence that antihypertensive or cholesterol-lowering medications cause diabetes or lead to diabetes medication use. The analysis was predictive rather than causal. Concurrent medication use may instead serve as a marker of multimorbidity, healthcare utilization, diagnosis, prescribing, or engagement with long-term pharmacotherapy (Elseddawy et al., 2022). This distinction has practical relevance to pharmaceutical care. Patients receiving medication for multiple chronic conditions may require medication review, adherence assessment, monitoring for medication-related problems, and coordinated follow-up. Previous research has demonstrated that pharmacist-led interventions can contribute to improvements in adherence and diabetes-related outcomes. Predictive analytics may eventually complement such services by assisting with population-level identification of medication-use patterns or prioritization of individuals for additional review(Elseddawy et al., 2022).

The interpretation of the pregnancy variables requires particular caution. Pregnancy “Yes” showed the largest mean absolute LASSO coefficient, and the “Not applicable” category was also among the five strongest predictors. However, pregnancy eligibility is structurally related to sex and age within the survey questionnaire. Consequently, these variables may partly encode demographic and questionnaire-routing information rather than represent biological determinants of diabetes medication use. They are therefore described as predictors, not risk factors (Kwon et al., 2023). An important consideration in the present study is the definition of the outcome. The target represented self-reported use of medication for diabetes rather than a clinical diagnosis of diabetes. Medication utilization is itself a meaningful pharmacoepidemiological outcome, but it reflects several processes, including underlying disease, diagnosis, healthcare access, prescribing, treatment initiation, and continued medication use. Consequently, the present model cannot be used as a diagnostic algorithm for diabetes (Wang et al., n.d.).

The analytical design incorporated several methodological safeguards. First, the study included more than 48,000 participants. Second, all four models were evaluated under the same stratified 10-fold cross-validation framework. Third, SMOTENC was restricted to training folds. Fourth, preprocessing and LASSO feature selection were performed using training data rather than the complete dataset. This reduced the risk of information leakage and overly optimistic validation estimates. Finally, the stability of LASSO selection across folds was evaluated rather than relying on a single feature-selection result (Elreedy et al., 2024a). Several limitations should nevertheless be acknowledged. Only 353 observations belonged to the positive class despite the large overall sample, and estimates of positive-class performance should therefore be interpreted cautiously. The outcome was self-reported medication use rather than clinically verified diabetes. Detailed information concerning individual antidiabetic drugs, doses, duration of therapy, adherence, adverse drug reactions, medication-related problems, and drug interactions was not included in the analytical variables. Consequently, the present study cannot evaluate the appropriateness, effectiveness, or safety of individual pharmacotherapy regimens (Elseddawy et al., 2022).

The study also relied on IFLS5 data collected in 2014–2015. Diabetes treatment patterns and access to pharmacotherapy may have changed since the survey was conducted. Furthermore, the models underwent internal cross-validation but were not externally validated in an independent contemporary population. Predictive performance should therefore not be assumed to generalize directly to current Indonesian clinical or community pharmacy settings (Alex et al., 2022). SMOTENC also has limitations because synthetic observations may alter training distributions and model decision boundaries. Future studies should compare oversampling with alternative approaches such as class weighting, undersampling, ensemble imbalance methods, and decision-threshold optimization. For rare outcomes, precision-recall curves and related measures would also complement ROC-AUC (Kwon et al., 2023).

Future research should incorporate richer pharmacy-specific information, including antidiabetic drug classes, total medication count, dosage, treatment duration, dispensing history, adherence, adverse drug reactions, drug interactions, and medication-related problems. Combining these variables with laboratory outcomes such as HbA1c could help distinguish disease status, treatment status, and therapeutic effectiveness (Al Assaf et al., 2022). External validation using more recent Indonesian data is essential before implementation. Future studies should additionally assess model calibration, precision-recall performance, decision-curve analysis, and clinically meaningful probability thresholds. Such work could clarify whether predictive models have practical utility for pharmacist-led medication review or prioritization of pharmaceutical-care interventions (Alex et al., 2022).

CONCLUSION

Logistic Regression and linear SVM provided the most balanced performance for predicting self-reported diabetes medication use in IFLS5. Logistic Regression achieved 80.13% accuracy, 76.20% sensitivity, 80.16% specificity, and an ROC-AUC of 0.870. Random Forest demonstrated that very high accuracy can be misleading in a severely imbalanced population, whereas Naive Bayes maximized sensitivity at the expense of specificity. LASSO consistently identified medication-related and physiological predictors, with cholesterol medication use and hypertension medication use among the five strongest predictors. These findings support further investigation of machine learning as a pharmacoepidemiological approach for characterizing medication-use patterns. However, the outcome represents medication use rather than diabetes diagnosis, and external validation with contemporary, pharmacy-specific data is required before clinical implementation.

ACKNOWLEDGMENTS

This study was funded by the Institute for Research and Community Service (LPPM), [Universitas Muhammadiyah Gombong, Indonesia

REFERENCES

9. Pharmacologic Approaches to Glycemic Treatment: Standards of Care in Diabetes—2024. (2024). *Diabetes Care*, 47, S158–S178. <https://doi.org/10.2337/dc24-S009>
- Al Assaf, S., Zelko, R., & Hanko, B. (2022). The Effect of Interventions Led by Community Pharmacists in Primary Care for Adults with Type 2 Diabetes Mellitus on Therapeutic Adherence and HbA1c Levels: A Systematic Review. In *International Journal of Environmental Research and Public Health* (Vol. 19, Number 10). MDPI. <https://doi.org/10.3390/ijerph19106188>
- Alex, S. A., Jhanjhi, N. Z., Humayun, M., Ibrahim, A. O., & Abulfaraj, A. W. (2022). Deep LSTM Model for Diabetes Prediction with Class Balancing by SMOTE. *Electronics (Switzerland)*, 11(17). <https://doi.org/10.3390/electronics11172737>
- Alfian, R., Fathin Fawwazi, M. H. A., Adikusuma, W., & Abero Phiri, Y. V. (2026). Pharmacists' Experiences and Needs in Pharmaceutical Care to Support Diabetes Medication Adherence in Indonesian Primary Care. *Journal of Public Health and Pharmacy*, 6(2), 241–251. <https://doi.org/10.56338/jphp.v6i2.8568>
- Alwhaibi, M. (2022). Potentially Inappropriate Medications Use among Older Adults with Comorbid Diabetes and Hypertension in an Ambulatory Care Setting. *Journal of Diabetes Research*, 2022. <https://doi.org/10.1155/2022/1591511>
- Aman, & Chhillar, R. S. (2023). Optimized stacking ensemble for early-stage diabetes mellitus prediction. *International Journal of Electrical and Computer Engineering*, 13(6), 7048–7055. <https://doi.org/10.11591/ijece.v13i6.pp7048-7055>
- Cabot, J. H., & Ross, E. G. (2023). Evaluating prediction model performance. *Surgery (United States)*, 174(3), 723–726. <https://doi.org/10.1016/j.surg.2023.05.023>
- Collins, G. S., Dhiman, P., Ma, J., Schluskel, M. M., Archer, L., Van Calster, B., Harrell, F. E., Martin, G. P., Moons, K. G. M., van Smeden, M., Sperrin, M., Bullock, G. S., & Riley, R. D. (2024). Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ*. <https://doi.org/10.1136/bmj-2023-074819>
- Coutureau, C., Slimano, F., Mongaret, C., & Kanagaratnam, L. (2022). Impact of Pharmacists-Led Interventions in Primary Care for Adults with Type 2 Diabetes on HbA1c Levels: A Systematic Review and Meta-Analysis. In *International Journal of Environmental Research and Public Health* (Vol. 19, Number 6). MDPI. <https://doi.org/10.3390/ijerph19063156>
- Elreedy, D., Atiya, A. F., & Kamalov, F. (2024a). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning*, 113(7), 4903–4923. <https://doi.org/10.1007/s10994-022-06296-4>

- Elreedy, D., Atiya, A. F., & Kamalov, F. (2024b). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning*, 113(7), 4903–4923. <https://doi.org/10.1007/s10994-022-06296-4>
- Elseddawy, A. I., Karim, F. K., Hussein, A. M., & Khafaga, D. S. (2022). Predictive Analysis of Diabetes-Risk with Class Imbalance. *Computational Intelligence and Neuroscience*, 2022. <https://doi.org/10.1155/2022/3078025>
- Kwon, Y., Han, K., Suh, Y. J., & Jung, I. (2023). Stability selection for LASSO with weights based on AUC. *Scientific Reports*, 13(1). <https://doi.org/10.1038/s41598-023-32517-4>
- Ou, Q., Jin, W., Lin, L., Lin, D., Chen, K., & Quan, H. (2023). LASSO-based machine learning algorithm to predict the incidence of diabetes in different stages. *Aging Male*, 26(1). <https://doi.org/10.1080/13685538.2023.2205510>
- Piccialli, V., & Sciandrone, M. (2022). Nonlinear optimization and support vector machines. *Annals of Operations Research*, 314(1), 15–47. <https://doi.org/10.1007/s10479-022-04655-x>
- Rastogi, R., & Bansal, M. (2023). Diabetes prediction model using data mining techniques. *Measurement: Sensors*, 25. <https://doi.org/10.1016/j.measen.2022.100605>
- Sardu, C., Zimbudzi, E., Moradpour, F., & Alfian, S. D. (n.d.). *Sociodemographic and behavioural risk factors associated with low awareness of diabetes mellitus medication in Indonesia: Findings from the Indonesian Family Life Survey (IFLS-5)*.
- Ullah, Z., Saleem, F., Jamjoom, M., Fakieh, B., Kateb, F., Ali, A. M., & Shah, B. (2022). Detecting High-Risk Factors and Early Diagnosis of Diabetes Using Machine Learning Methods. *Computational Intelligence and Neuroscience*, 2022. <https://doi.org/10.1155/2022/2557795>
- Vanacore, A., Pellegrino, M. S., & Ciardiello, A. (2024). Fair evaluation of classifier predictive performance based on binary confusion matrix. *Computational Statistics*, 39(1), 363–383. <https://doi.org/10.1007/s00180-022-01301-9>
- Wang, P., Adisa, R., Davar Siadat, S., Yang, Y., Tong Xingwei Wu, R., Copyright, fpubh, Li, M., Lu, X., Yang, H., Yuan, R., Tong, R., & Wu, X. (n.d.). *Development and assessment of novel machine learning models to predict medication non-adherence risks in type e diabetics*.
- Welvaars, K., Oosterhoff, J. H. F., van den Bekerom, M. P. J., Doornberg, J. N., van Haarst, E. P., van der Zee, J. A., van Andel, G. A., Lagerveld, B. W., Hovius, M. C., Kauer, P. C., Boevé, L. M. S., van der Kuit, A., Mallee, W., & Poolman, R. (2023). Implications of resampling data to address the class imbalance problem (IRCIP): an evaluation of impact on performance between classification algorithms in medical data. *JAMIA Open*, 6(2). <https://doi.org/10.1093/jamiaopen/ooad033>.

