

# Predicting diabetes risk based on health examination data using machine learning

Nguyen Xuan An

**Abstract:** This study applies machine learning to predict the early risk of diabetes from clinical indicators. The author processes raw data, replaces missing values with the median, and normalizes the scale of attributes to ensure reliability. The project compares the performance of Logistic Regression and Random Forest models, with results showing that Random Forest achieves higher accuracy due to its ability to handle non-linear relationships well. The system provides an effective medical screening support tool based on the Python programming platform and the Scikit-learn library.

**Keywords:** Diabetes prediction, Machine Learning, Random Forest

## References

- [1] American Diabetes Association, "Standards of medical care in diabetes—2024," *Diabetes Care*, vol. 47, no. S1, pp. S1–S322, 2024.
- [2] International Diabetes Federation, *IDF Diabetes Atlas*, 11th ed. Brussels, 2025.
- [3] Pandas Documentation, "Pandas: Python data analysis library," 2025. [Online]. Available:
- [4] World Health Organization, *Global report on diabetes*, 2024. [Online]. Available:
- [5] American Diabetes Association, "Early prediction of diabetes using KNN, SVM, Random Forest and XGBoost," 2024.