Prediction Models for Type 2 Diabetes Progression

Angeline Mathew
Department of Biomedical Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India

Abraham Paul
Department of Biomedical Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India

K. Sivaprasath
Department of Biomedical Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India

Abstract:

Type 2 Diabetes Mellitus (T2DM) is a progressive metabolic disorder characterized by declining insulin sensitivity and gradual β-cell exhaustion, which together drive chronic hyperglycemia and substantially elevate the risk of cardiovascular, renal, and neurological complications over time. Despite growing clinical awareness, predicting how the disease will evolve in individual patients remains difficult using conventional assessment methods alone. This study addresses that gap by systematically evaluating four supervised classification algorithms — Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM) — for their ability to forecast T2DM progression from structured clinical data. The dataset, sourced from a publicly available medical repository, was prepared through a rigorous preprocessing pipeline that included median imputation for physiologically implausible zero values, IQR-based outlier removal, and Min-Max feature normalization to ensure scale consistency across variables. Each model was trained on 80% of the data and assessed on the remaining 20% using five complementary metrics: Accuracy, Precision, Recall, F1-Score, and AUC-ROC, chosen to capture both predictive performance and clinical reliability across imbalanced class distributions. Among the four classifiers evaluated, Random Forest achieved the highest performance, reflecting the strength of ensemble learning in capturing nonlinear interactions among clinical variables such as glucose concentration, BMI, age, and insulin levels. These findings indicate that well-tuned ensemble classifiers hold meaningful potential as supplementary decision-support tools in early diabetes risk stratification, with implications for proactive patient management and resource allocation in clinical practice.

Keywords:

Type 2 Diabetes Mellitus, Disease Progression, Machine Learning, Support Vector Machine (SVM), Random Forest, Logistic Regression, Medical Data Analytics, Predictive Modeling, Healthcare Informatics, Data Preprocessing.