PREDICTING THIRTY-DAY HOSPITAL READMISSION IN PATIENTS WITH DIABETES USING MACHINE LEARNING AND ELECTRONIC HEALTH RECORD DATA

Nivedha Nataliya Anthony[1]*

[1]Department of Chemistry, Faculty of Science, University of Colombo, Colombo, Sri Lanka

*Corresponding author: 2023s19801@stu.cmb.ac.lk

Abstract

Hospital readmission within thirty days of discharge is a key indicator of care quality and a major source of avoidable cost, and patients with diabetes are readmitted more often than the general inpatient population owing to complex glycaemic management, comorbidities and demanding medication regimens. This study develops and compares machinelearning models to predict, at discharge, thirty-day readmission risk in patients with diabetes from electronic health record data, to identify its strongest predictors, and to explore whether such an accessible, opensource approach might be feasible in low and middleincome or datalimited settings. The publicly available Diabetes 130US Hospitals dataset, comprising 101,766 inpatient encounters from 71,518 patients across 50 variables, was used. Encounters ending in death or hospice discharge were excluded, and only each patient's first encounter was retained to avoid dependence between records. After removing variables with extensive missing data, grouping diagnosis codes into clinical categories and encoding categorical variables, 69,987 encounters and 96 features remained. Readmission was modelled as a binary outcome, and the data were split 80:20 with stratification. Logistic Regression, Random Forest and Extreme Gradient Boosting were trained with classimbalance weighting and evaluated by accuracy, precision, recall, F1score and area under the ROC curve (AUC). The ensemble models outperformed the logisticregression baseline, with Extreme Gradient Boosting achieving the highest discrimination (AUC 0.646); recall and AUC were prioritized over accuracy given the class imbalance. By permutation importance on this model, discharge disposition, prior inpatient visits, primary diagnosis, length of stay, number of medications and age were the most influential predictors. The model relied on routinely collected administrative variables rather than costly laboratory tests, which were substantially incomplete and not among the leading predictors. This suggests potential feasibility for lowerresource settings; however, as the data originate from US hospitals, this remains a hypothesis requiring local retraining and external validation.

Keywords: Diabetes, Hospital readmission, Machine learning, Electronic health records, Predictive modeling