Predictive Modeling of Diabetes Risk Using Health Behavior Indicators: A Clinical Decision Support Approach

Frank Ofosu

Abstract

Diabetes is a major chronic condition in the United States, affecting over 34 million adults and contributing to significant morbidity, mortality, and healthcare costs. Early identification of individuals at risk for diabetes and prediabetes can enable targeted preventive interventions, improving outcomes and reducing system burden. The CDC Behavioral Risk Factor Surveillance System (BRFSS 2015) provides large-scale survey data with health, lifestyle, and demographic indicators relevant to diabetes risk. Leveraging this dataset, predictive analytics can be used to build interpretable models that inform clinical decision support (CDS) tools for preventive care. The dataset contains 3 files and will begin with the balanced binary dataset (Dataset #2), which contains equal proportions of respondents with and without diabetes. This dataset will be used to train baseline classifiers such as logistic regression, random forest, and gradient boosting, providing a fair foundation for model comparison. If time allows, the project will extend to the imbalanced binary dataset (Dataset #3), which more closely reflects real-world distributions and will require handling class imbalance through techniques such as class weighting, resampling, or threshold adjustment. Finally, as an advanced step, the multi-class dataset (Dataset #1) may be explored to distinguish between no diabetes, prediabetes, and diabetes, supporting progression risk stratification. Model evaluation will include AUROC, precision-recall metrics, calibration, and subgroup performance to ensure fairness across demographics. Interpretability will be emphasized using SHAP values to identify the most influential features, such as BMI, hypertension, and physical activity. The ultimate aim is to demonstrate how predictive modeling can inform CDS by providing risk scores, explanations, and actionable recommendations to clinicians, thereby enhancing early diagnosis and preventive care.

References

Xie, Z., Nikolayeva, O., Luo, J., & Li, D. (2019). Building risk prediction models for type 2 diabetes using 2014 BRFSS data. *Preventing Chronic Disease, 16*, E74. https://doi.org/10.5888/pcd16.190109

Rahman, S. S., Khatun, F., & Rahman, M. M. (2023). Diabetes prediction using machine learning and explainable AI. *Journal of Healthcare Engineering, 2023,* 1–12. https://pmc.ncbi.nlm.nih.gov/articles/PMC10107388/

Ullah, I., & Lee, Y. (2024). Robust diabetic prediction using ensemble machine learning. *Scientific Reports, 14,* 78519. https://doi.org/10.1038/s41598-024-78519-8

Kumar, V., Singh, R., & Gupta, A. (2023). Diabetes prediction using machine learning and Flask. *Biomedical and Pharmacology Journal, 17*(2), 859–867. https://biomedpharmajournal.org/vol17no2/diabetes-prediction-using-machine-learning-and-flask/

Wang, X., Li, H., & Zhang, Y. (2024). A feature optimization study based on a diabetes risk questionnaire. *Frontiers in Public Health, 12,* 1328353. https://doi.org/10.3389/fpubh.2024.1328353

Centers for Disease Control and Prevention. (2015). *Behavioral Risk Factor Surveillance System (BRFSS) annual data*. U.S. Department of Health & Human Services. https://www.cdc.gov/brfss/annual_data/annual_data.htm

Al-Fuhaidi, B., Farae, Z., Al-Fahaidy, F., Nagi, G., Ghallab, A., & Alameri, A. (2024). *Anomaly-based intrusion detection system in wireless sensor networks using machine learning algorithms*. Applied Computational Intelligence and Soft Computing, 2024, Article ID 2625922. https://doi.org/10.1155/2024/2625922