Predictive Modeling for Detection of Type-2 Diabetes: Evaluation of Three Machine Learning Algorithms

CIND 820: Big Data Analytics Project

Abstract-revised

By:

Yitayal Mengistu [501211839]

Submitted to:

Dr. Ceni Babaoglu (supervisor)

June10, 2024

Abstract

Diabetes is a widespread chronic disease that imposes significant health and economic burdens globally (Khan et al., 2021). Uncontrolled diabetes leads to severe complications, including cardiovascular diseases, kidney dysfunction, and vision impairment (LeBlanc et al., 2019). In Canada alone, diabetes and prediabetes affect approximately 11.7 million people, with projections indicating a surge to 13.6 million cases by 2030 (Diabetes Canada, 2022). Early prediction and understanding of the predictive power of associated risk factors are crucial for effective prevention and management of diabetes-related complications (Alssema et al., 2011; Khan et al., 2021).

Leveraging various machine learning (ML) algorithms and a large dataset, this study aims to develop and evaluate the accuracy of three predictive models to identify the risk of diabetes. Specifically, it seeks to answer the following research questions:

1. How do different ML algorithms compare in terms of accuracy and interpretability for predicting the risk of diabetes?

2. Which features are most predictive of diabetes risk, and what are the implications for prevention and public health policy?

3. How does the choice of feature selection methods impact the performance of ML algorithms?

4. How does the choice of hyperparameters impact the performance of selected ML algorithms?

This study uses a publicly available dataset from the 2015 USA Behavioral Risk Factor Surveillance System (BRFSS) phone call survey, originally collected by the Centers for Disease Control and Prevention (CDC). The dataset, accessed via Kaggle, contains 70,692 observations

and 21 variables, with the target variable being binary: diabetes or non-diabetes (Burrows et al., 2017). The main variables include age, gender, education, income, alcohol consumption, body mass index, heart disease, smoking history, cholesterol, and blood glucose level.

Initial data analysis involved data cleaning and preprocessing steps. These steps included detailed descriptions of variable characteristics, identifying data types, handling missing values and outliers, and assessing the distribution of each variable. Pairwise correlations between variables were examined, and bivariate analysis were employed to compare groups based on features, providing insights into dataset dynamics.

Exploratory data analysis will further unveil dataset insights, potentially involving normalization, scaling, subsetting, and clustering to identify underlying trends or patterns (de Jonge & van der Loo, 2013). Dimensionality reduction techniques, such as recursive feature elimination and feature extraction methods, will be used to ascertain the significance of each feature in subsequent ML model analyses.

Prior to implementing various the ML classification algorithms, the dataset will be partitioned into training and testing sets, with 5-fold cross-validation employed. Three ML classification algorithms, selected based on interpretability and performance criteria, will be trained, tested, and compared. Six top candidates include Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, Support Vector Machine, and K-Nearest Neighbors (Gong, 2022a, 2022b).

The dataset will be analyzed utilizing Python libraries, Subsequently, the predictive performance of the three selected algorithms will be evaluated and compared based on different metrics. Important diabetes risk predictors will be identified, the influence of feature selection

techniques will be assessed, and the impact of hyperparameters on ML algorithm performance will be evaluated.

This research project builds on prior studies while addressing their limitations. Previous research (Kibria et al., 2022; Rani, 2020; Wang et al., 2024; Xie et al., 2019; Zou et al., 2018) demonstrated varying degrees of success in utilizing ML for diabetes prediction, often constrained by small, homogeneous datasets and limited focus on model interpretability. This study utilizes a large, diverse dataset to enhance the generalizability of findings. It systematically investigates the impact of feature selection and hyperparameter tuning on model performance, comparing three ML algorithms selected for their balance of interpretability and accuracy.

Key limitations include potential self-report biases in the dataset and the limited scope of features analyzed. However, this comprehensive approach aims to provide valuable insights for diabetes prevention and public health policy, contributing to more effective management and intervention strategies.

References

Alssema, M., Vistisen, D., Heymans, M. W., Nijpels, G., Glümer, C., Zimmet, P. Z., Shaw, J. E., Eliasson, M., Stehouwer, C. D. A., Tabák, A. G., Colagiuri, S., Borch-Johnsen, K., & Dekker, J. M. (2011). The Evaluation of Screening and Early Detection Strategies for Type 2 Diabetes and Impaired Glucose Tolerance (DETECT-2) update of the Finnish diabetes risk score for prediction of incident type 2 diabetes. *Diabetologia*, *54*(5), 1004–1012. https://doi.org/10.1007/s00125-010-1990-7

Burrows, N. R., Hora, I., Geiss, L. S., Gregg, E. W., & Albright, A. (2017). Incidence of End-Stage Renal Disease Attributed to Diabetes Among Persons with Diagnosed Diabetes — United States and Puerto Rico, 2000–2014. *MMWR. Morbidity and Mortality Weekly Report*, *66*(43), 1165–1170. https://doi.org/10.15585/mmwr.mm6643a2

de Jonge, E., & van der Loo, M. (2013). *An introduction to data cleaning with R*.

Diabetes Canada. (2022, March 3). *Diabetes rates continue to climb in Canada. [Press release]. Retrieved from [https://www.diabetes.ca/media-room/press-releases/diabetes-rates-continue-to-climb-in-canada]*.

Gong, D. (2022a). op 6 Machine Learning Algorithms for Classification. *Towards Data Science*.

Gong, D. (2022b). Top 6 Machine Learning Algorithms for Classification. *Towards Data Science*.

Khan, F. A., Zeb, K., Al-Rakhami, M., Derhab, A., & Bukhari, S. A. C. (2021). Detection and Prediction of Diabetes Using Data Mining: A Comprehensive Review. *IEEE Access*, *9*, 43711–43735. https://doi.org/10.1109/ACCESS.2021.3059343

Kibria, H. B., Nahiduzzaman, M., Goni, M. O. F., Ahsan, M., & Haider, J. (2022). An Ensemble Approach for the Prediction of Diabetes Mellitus Using a Soft Voting Classifier with an Explainable AI. *Sensors (Basel, Switzerland)*, *22*(19). https://doi.org/10.3390/S22197268

LeBlanc, A. G., Gao, Y. J., McRae, L., & Pelletier, C. (2019). At-a-glance - Twenty years of diabetes surveillance using the Canadian Chronic Disease Surveillance System. *Health Promotion and Chronic Disease Prevention in Canada*, *39*(11), 306–309. https://doi.org/10.24095/hpcdp.39.11.03

Rani, K. J. (2020). Diabetes Prediction Using Machine Learning. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 294–305. https://doi.org/10.32628/CSEIT206463

Wang, X., Wang, W., Ren, H., Li, X., & Wen, Y. (2024). Prediction and analysis of risk factors for diabetic retinopathy based on machine learning and interpretable models. *Heliyon*, *10*(9). https://doi.org/10.1016/j.heliyon.2024.e29497

Xie, Z., Nikolayeva, O., Luo, J., & Li, D. (2019). Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. *Preventing Chronic Disease*, *16*, 190109. https://doi.org/10.5888/pcd16.190109

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting Diabetes Mellitus With Machine Learning Techniques. *Frontiers in Genetics*, *9*. https://doi.org/10.3389/FGENE.2018.00515