Datathon 3

Nitya Kuruvila & Sashini Kosgoda (Team 8)

Dalla Lana School of Public Health, University of Toronto

CHL 5230: Applied Machine Learning for Health Data

Dr. Zahra Shakeri

October 31, 2023

Introduction

The primary objective of this paper is to investigate the relationship between diabetes and various medical factors. Diabetes is a chronic disease that affects approximately 422 million people worldwide, primarily individuals living in low-or middle-income countries (WHO, n.d.). Annually, 1.5 million global deaths are attributed directly to diabetes and this number continues to rise (WHO, n.d.). Diabetes can lead to severe health consequences over time including, damage to the heart, kidneys and blood vessels (WHO, n.d.). The diagnosis rates of diabetes vary significantly across the world with 87.5% of all undiagnosed cases being in low-or middle-income countries (Magliano et al., 2021). This disparity can be attributed to numerous factors including socioeconomic conditions, the local health system, and public awareness among others (Magliano et al., 2021). This paper focuses specifically on type 2 diabetes (T2DM), a chronic condition characterized by insulin resistance, leading to abnormal insulin response in cells (CDC, 2023, p. 2).

Firstly, this paper examines the relationship between diabetes and other medical conditions. The occurrence, the treatment and the management of diabetes can be significantly impacted by the presence of other medical conditions (Nowakowska et al., 2019). Secondly, this paper also aims to use biomarkers to effectively categorize individuals as diabetic or non-diabetic. The biomarkers explored in this paper are those whose association with diabetes are lesser known. Given the global prevalence and severe health impacts that can result from T2DM, it is important to better understand how other conditions and biomarkers may influence a diabetes diagnosis. The insights gained through this analysis can inform and improve new methods for diagnosis and treatment of T2DM, subsequently improving the accessibility and availability of diabetes focused health care globally. These findings can be especially beneficial to low- and middle- income countries or areas in which universal healthcare is not available.

Methods

The analyzed dataset comes from the *Canadian Primary Care Sentinel Surveillance Network* (Keshavjee, 2023) and consists of 10,000 records, including 5139 diabetic cases and 4861 non-diabetic cases. It provides information on diabetes status, other health conditions, biomarkers, the dates of condition onset and testing and medications. The dataset includes information on 5336 females and 4664 males, ranging in age from 18 to 90 years old.

Our approach, conducted using Python, involved data cleaning, identifying missing values and obtaining summary statistics for all continuous and binary variables. Variables with extensive missing data (> 30%) and low binary variability (< 5%) were excluded due to their limited information and non-relevance for our study. Patients were grouped by Patient ID and data was aggregated to reflect the mean value of all variables. Categorical values were transformed into numerical equivalents for ease of analysis. Data exploration involved using boxplots, correlation heat maps and pair plots to identify outliers and assess patterns and distributions. Class imbalance for the target variable and the count of missing data in the different target variable classes were observed to inform further analysis.

To address the first question, variables indicative of a medical condition were first selected, namely: diabetes, BMI, total cholesterol (TC), depression, hypertension, osteoarthritis, and chronic obstructive pulmonary disease. Diabetes was defined as the target while all other variables were considered as features. The data was divided into training and testing sets, using 20% for testing. A 5-fold cross-validation was then conducted in which the training data was split for training and validation. In each iteration, missing data was imputed using the median, and the data was normalized to reduce the influence of any outliers and improve the accuracy of the analysis. A logistic regression model was fitted

and evaluated to determine the average training and validation accuracy. A grid search was then completed to fine-tune hyperparameters including regularization strength, penalty and solver. A new logistic model was fitted based on the identified hyperparameters and tested on the testing data. Model evaluation was then conducted using the following metrics: the area under the curve (AUC), the accuracy, the precision score, the recall score, a confusion matrix and an ROC curve. Additional analysis included plotting logistic regression curves and obtaining a summary of the model.

To explore the second question, variables which were biomarkers were selected, including diabetes, systolic blood pressure (sBP),BMI, low-and high- density lipoproteins (LDL, HDL), triglycerides (TG), TC and A1c. All selected variables were used as predictors except diabetes which was the target variable. Fasting blood sugar was omitted from the model as it is frequently used to diagnose diabetes and its association with diabetes is well known. Variables were organized into numerical and target variables, and missing values were inspected using a bar chart to determine if the distribution of target variable classes differed. The data was split into testing and training sets, 20% was used for testing. A 5-fold cross-validation was performed, each iteration involving imputing missing values using the median, similar to the first analysis, and data normalization. A Gaussian Naive Bayes model was applied and evaluated; evaluation metrics included the accuracy, a confusion matrix, a classification report, AUC score and the ROC curve. Once validated, the model was tested on the testing data and the model performance was evaluated using the aforementioned metrics.

Results

The initial data exploration identified several variables with over 30% missing data, primarily dates and medication information. For certain tests, particularly those related to TC, the testing date was indicated but the measurement was not recorded. Additionally, missing data was identified in five variables used for further analysis (sBP, HDL, LDL, TG and TC). When comparing the class distribution of the missing values for each of these variables, there were no large differences in the distribution, indicating that the missingness is likely due to random error. The boxplots, supported by the summary statistics, identified outliers in several variables; however, this data was not removed due to their medical significance. The correlation heat map determined strong negative correlations between diabetes and LDL (-0.31), HDL (-0.26) and TC (-0.31) and strong positive correlations between diabetes and A1C (0.57) and fasting blood sugar (0.55). The data also presented a balanced distribution of diabetes classes (1:51.4%, 0: 48.6%).

The 5-fold cross validation for logistic regression, conducted to explore the first question, determined an average training accuracy score of 0.663 and an average validation accuracy score of 0.659. These scores do not exhibit signs of overfitting in the logistic model. The grid search determined the best hyperparameters for the model to be a regularization strength of 0.01, penalty of L1 and a liblinear solver. After adjusting these hyperparameters, model evaluation determined AUC and accuracy scores of 0.63 and 0.62 for the training set, and 0.63 and 0.62 for the test set, respectively. The precision scores for individuals with and without diabetes were 0.58 and 0.71 in the training set, and 0.57 and 0.71 in the test set. The recall values were 0.80 and 0.45 in the training set, and 0.79 and 0.46 in the test set for individuals with and without diabetes respectively. Consistent with our findings from the initial data exploration, the logistic regression curves demonstrated that all analyzed features, except for TC and depression, had a positive correlation to diabetes except TC level and depression. The features whose odds ratio values reached statistical significance at the 5% level for an association with diabetes were

BMI (1.05, 95% CI: 1.05, 1.06, p=0.00), TC (0.57, 95% CI: 0.54, 0.59, p=0.00), depression (0.87, 95% CI: 0.77, 0.97, p=0.01), and hypertension (1.85, 95% CI: 1.68, 2.03, p=0.00).

The analysis to answer the second question indicated that individuals with diabetes have lower LDL, HDL and TC levels and higher A1c and FBS levels compared to non-diabetic individuals. Based on the model performance of the Gaussian Naives Bayes model, the precision of predicting diabetic patients is 0.84, recall is 0.78, and F1-score is 0.81. The average AUC and accuracy of the validation set was 0.805 and 0.807 respectively. The AUC and accuracy of the testing set was 0.826 and 0.828 respectively.

Discussion

Given the health consequences that can occur from diabetes, it is important that those with diabetes are correctly diagnosed. Model evaluation determined that the logistic model had notable sensitivity, accuracy and AUC scores, indicating its effectiveness in correctly identifying individuals with diabetes. The similar evaluation scores for the training and test sets provide no indication of overfitting suggesting good generalization, this notion is supported by the results of the cross validation. Results from the logistic model would also indicate that increased BMI and hypertension are risk factors for diabetes, consistent with existing literature (Klein et al., 2022; Petrie et al., 2018). In contrast depression and increased TC levels were found to be protective factors against diabetes. The vast body of literature has found that depression and TC are also significant risk factors for T2DM, contrasting our findings (Rhee et al., 2017; Rotella & Mannucci, 2013). The results of this analysis suggest that the presence of other medical conditions may be an effective diagnostic method for diabetes, warranting further investigation.

The Gaussian Naive Bayes model's performance, as indicated by AUC and accuracy, is fairly consistent between the validation and testing sets, suggesting that the model generalizes well to new, unseen data. Furthermore, the performance metrics of the testing set lies within the range of the model performance measures of the validation set, further supporting the models generalizability. The AUC values highlight the model's ability to distinguish between diabetic and non-diabetic cases reasonably well. Further refinement and analysis, such as feature engineering and hyperparameter tuning, could potentially enhance the model's performance. It should be noted that while our second question focused on biomarkers with lesser known association with diabetes, A1c was included although it has a high accuracy for diagnosing diabetes. The sensitivity and specificity of tests using A1c varies, making its test performance uncertain (Selvin et al., 2011).

Our findings also determined negative correlations between A1c, LDL and TC which contradicts the majority of existing literature (Ghari Arab et al., 2018), although, one study found there is no significant association between HbA1c, TC and LDL (Hassan et al., 2015). Similarly, our data exploration and subsequent analysis found that diabetic individuals had lower levels of LDL and TC which does not align with current literature. Typically, individuals with T2DM present with high levels of LDL, which increases after diagnosis (Mohamed et al., 2004). This disparity in our initial data may be considered a limitation of our study, although it may be difficult to address as it is embedded in the data; it is however, important to note this inconsistency in future analysis.

Given the potential of using biomarkers to differentiate between diabetic and non-diabetic cases, next steps for this work include identifying the specific thresholds of the biomarkers that can help distinguish diabetic and non-diabetic patients.

Individual Contributions

Sashini Kosgoda contributed to working on Question #2 (creating code, running analysis and writing of the report). Nitya Kuruvila contributed to working on Question #1 (creating code, running analysis and writing the report). Both contributed to data exploration, writing the introduction and discussion and editing.

Github link: https://github.com/sashini472/Project-Part-1.git

References

CDC. (2023, April 18). *Type 2 Diabetes*. Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/basics/type2.html

Ghari Arab, A., Zahedi, M., Kazemi Nejad, V., Sanagoo, A., & Azimi, M. (2018). Correlation between Hemoglobin A1c and Serum Lipid Profile in Type 2 Diabetic Patients Referred to the Diabetes Clinic in Gorgan, Iran. *Journal of Clinical and Basic Research*, *2*(1), 26–31. https://doi.org/10.29252/jcbr.2.1.26

Hassan, D., Elhussein, A., Fadlelseed, O., Babikr, W., & Idris, O. (2015). Lipid Profile and Glycated Hemoglobin (HbA1c) in Diabetic Sudanese Patients. *International Journal of Science and Research (IJSR)*, *438*, 1813.

Keshavjee, K. (2023, June 20). *PREVENT - Institute of Health Policy, Management and Evaluation*. Institute of Health Policy, Management and Evaluation. https://ihpme.utoronto.ca/initiative/prevent/

Klein, S., Gastaldelli, A., Yki-Järvinen, H., & Scherer, P. E. (2022). Why does obesity cause diabetes? *Cell Metabolism*, *34*(1), 11–20. https://doi.org/10.1016/j.cmet.2021.12.012

Magliano, D. J., Boyko, E. J., & Committee, I. D. A. 10th edition scientific. (2021). Global picture. In *IDF DIABETES ATLAS [Internet]. 10th edition*. International Diabetes Federation. https://www.ncbi.nlm.nih.gov/books/NBK581940/

Mohamed, E., Mohamed, M., & Rashid, F. A. (2004). Dyslipidaemic Pattern of Patients with Type 2 Diabetes Mellitus. *The Malaysian Journal of Medical Sciences : MJMS*, *11*(1), 44–51.

Nowakowska, M., Zghebi, S. S., Ashcroft, D. M., Buchan, I., Chew-Graham, C., Holt, T., Mallen, C., Van Marwijk, H., Peek, N., Perera-Salazar, R., Reeves, D., Rutter, M. K., Weng, S. F., Qureshi, N., Mamas, M. A., & Kontopantelis, E. (2019). The comorbidity burden of type 2 diabetes mellitus: Patterns, clusters and predictions from a large English primary care cohort. *BMC Medicine*, *17*, 145. https://doi.org/10.1186/s12916-019-1373-y

Petrie, J. R., Guzik, T. J., & Touyz, R. M. (2018). Diabetes, Hypertension, and Cardiovascular Disease: Clinical Insights and Vascular Mechanisms. *The Canadian Journal of Cardiology*, *34*(5), 575–584. https://doi.org/10.1016/j.cjca.2017.12.005

Rhee, E.-J., Han, K., Ko, S.-H., Ko, K.-S., & Lee, W.-Y. (2017). Increased risk for diabetes development in subjects with large variation in total cholesterol levels in 2,827,950 Koreans: A nationwide population-based study. *PLOS ONE*, *12*(5), e0176615. https://doi.org/10.1371/journal.pone.0176615

Rotella, F., & Mannucci, E. (2013). Depression as a risk factor for diabetes: A meta-analysis of longitudinal studies. *The Journal of Clinical Psychiatry*, *74*(1), 31–37. https://doi.org/10.4088/JCP.12r07922

Selvin, E., Steffes, M. W., Gregg, E., Brancati, F. L., & Coresh, J. (2011). Performance of A1C for the Classification and Prediction of Diabetes. *Diabetes Care*, *34*(1), 84–89. https://doi.org/10.2337/dc10-1235

WHO. (n.d.). *Diabetes*. World Health Organization. Retrieved October 27, 2023, from https://www.who.int/health-topics/diabetes