Project Phase # 2 & Datathon 3 – High Fidelity Report
Team 16
CHL5230 Dataset - Diabetes Study File 10K Dec 14 2017

Yacine Marouf – 1010718298
Hunter Pozzebon – 1010648444
Priyonto Saha – 1004881531

October 31st, 2023

Introduction

As of 2019, 8.8% of Canadians are living with diabetes and there are approximately 549 new diagnoses daily (LeBlanc et al., 2019). Although there is no conclusive cure for type 2 diabetes (Joslin Education Team, n.d.), insulin resistance can take years to develop, meaning that the prevention and delay of the disease is the best defence against this epidemic (DPPRG, 2002). The predominant methods of preventing diabetes include lifestyle and diet adjustments (ADA, 2021). Multiple studies report significant correlations between metabolic biomarkers, exercise rates, and diabetes incidence (Biavashi, 2023; Ahmed, 2021).

Previous studies have been conducted using the same database of electronic medical records (EMRs) that attempted to predict diabetes while accounting for temporal inconsistencies by utilizing novel methods in hidden Markov models (Perveen et al, 2019; Perveen et al 2020). However, they attempted to predict future prognostic results, not time to event for diabetes. Some previous publications have attempted to predict diabetes with data from EMRs (Naveed et al., 2023) but the models used in these studies did not properly adjust for the right-censoring that is prevalent with time-to-event data.

As such, our main research questions consist of the following: 1) Can we predict the time to diabetes onset based on metabolic biomarker levels and the dates that they were measured compared to the date of diabetes onset? 2) Which variables affect the time to diabetes onset, and 3) How do they affect the time to diabetes onset?

This study will incorporate methods used in survival analysis. We plan to design two machine learning models for comparison: 1) a baseline random forest classifier that assumes independence between data points and ignores the temporal correlation that will be used, and 2) a random survival forest utilizing survival trees which account for time to event data (Ishwaran, 2008). With a focus on survival and prevention, our model aims to predict the risk of diabetes in conjunction with the amount of time before diabetes onset based on current biomarker test results. This will allow clinicians to advise a patient on how to prolong time before diabetes and determine the best times for a patient to check in for future monitoring and testing.

Methods

Dataset and Pre-Processing

For this project, we used the CHL5230 dataset, which is a dataset from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) consisting of 10,000 records from 8602 patients with 43 features. To generate the dataset, systolic blood pressure (sBP) measurements from all patients over the age of 17 were joined with other clinical measurements that were closest in time within a specific time period, alongside the specific dates for the measurements. Then, all records with missing data were removed from consideration, along with patients on insulin who may have type 1 diabetes and patients using corticosteroids which can affect blood sugar levels. This dataset was then randomly sampled to generate the 10,000 records that make up the CHL5230 dataset.

Due to our goal to conduct random forest variants on the biomarkers in relation to diabetes, we chose to drop all variables that were not biomarkers including age at exam, biomarker test dates, depression state, hypertension medications, sex, patient ID, other diseases, other comorbidities, and diabetes onset date. The dataset we analyzed includes continuous variable clinical measures such as body mass index (BMI), low density lipoprotein (LDL), high density lipoprotein (HDL), triglyceride (TG), fasting blood sugar (FBS), HbA1c (A1c), and total cholesterol (TC).

Survival time will be calculated for each record by taking the number of days between the sBP date and the diabetes onset date, noting that the date of FBS measurement must be within 1 month and the A1c date must be within 3 months. If a patient does not have diabetes, the diabetes onset date will be considered the last date before the end of records. This data will be used for the random survival forest model.

Exploratory Data Analysis

Exploratory data analysis consisted of missing data analysis, summary statistics, and a correlation matrix. In addition, histograms were plotted to visualise the distribution of the predictors, along with boxplots to determine outliers in the data.

To identify the type of missing data we conducted logistic regression models with binary missingness indicators as the outcome, where a value of 1 implied a missing value. If the logistic regression showed no significance the missingness was deemed missing completely at random (MCAR). If the logistic regression showed any significance, for the purpose of analysis the missingness was deemed missing at random (MAR). However, we cannot fully ignore the possibility of these missing observations being missing not at random (MNAR). We then dropped any records where the only missing values were MCAR and used multiple imputation by chained equations through Scikit Learn's IterativeImputer command to impute MAR missing data.

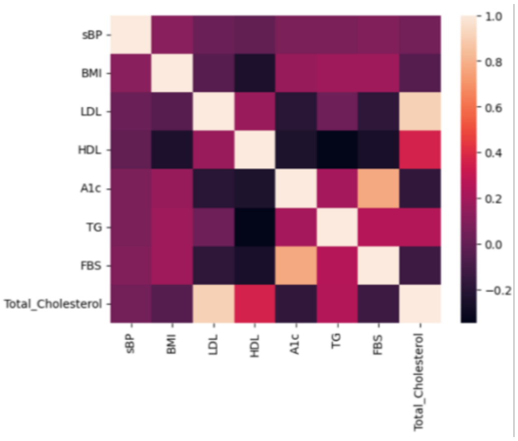

Figure 1: Correlation Matrix

There is no unexplained correlation found in the correlation matrix (figure 1). The higher correlation between total cholesterol, HDL, and LDL is explained by the fact that total cholesterol in blood predominantly consists of HDL and LDL. Patients with higher A1c and FBS show higher rate diabetes outcome (figure 2). This is not surprising as they are used as a diagnostic tool for diabetes.

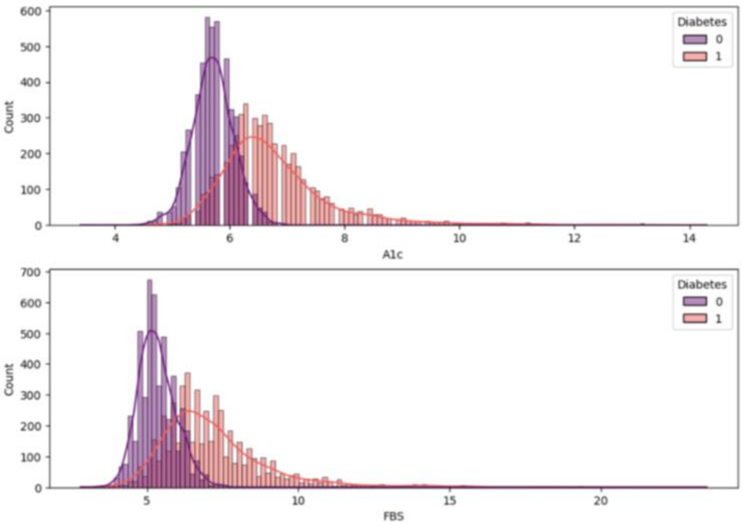

Figure 2: A1c and FBS vs Diabetes Outcome

Models

The initial model to be tested as the baseline is a random forest that utilize the biomarkers: sBP, BMI, LDL, HDL, A1c, TG, FBS and total cholesterol as the features and diabetes onset, with categories "yes" and "no" as the outcome variable to classify. We first split the data 80:20 as a train-test split, then conducted gridsearch with 5-fold cross validation to determine the optimal hyperparameters for the random forest.

The random survival forest model is an extension of the random forest model that was developed to account for the presence of censoring in survival data. To do so survival trees use the log-rank splitting rule to determine the best split within the tree (Ishwaran, 2008). The random survival forest will include all previously mentioned biomarkers, the date those biomarkers were measured, and the date of diabetes onset. Our survival model will be used in conjunction with our baseline random forest model in the future.

Results

Of the 10,000 records, 4 records had a missing sBP measurement, 61 had missing LDL measurements, 72 had missing HDL measurements, 53 had missing TG measurements, and 207 had missing total cholesterol measurements. When further analyzing the missing data, we determined that HDL and sBP were the only two variables that were MCAR, which allowed us to drop their missing records. LDL, TG, and total cholesterol were MAR and imputed. Thus our data sample size was 9941 records.

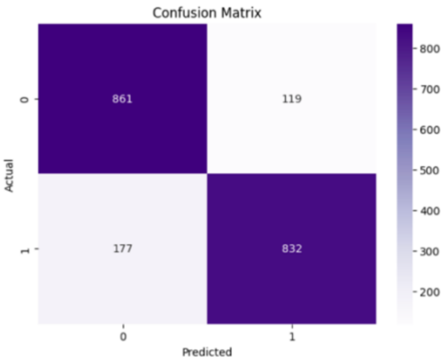

Figure 3: Confusion matrix of random forest predictor

The tuned random forest grew 100 trees with a max depth of 10 and a minimum leaf sample of 15. The classification report showed that the model correctly predicted the patient's diabetes a majority of the time with an accuracy score of 0.85. The model was able to accurately predict patients who have diabetes with a precision score of 0.82 and recall score of 0.87. The model was also able to predict patients without diabetes accurately with a precision score 0.88 and a recall score of 0.83. Out of 1009 patients with diabetes the model accurately predicts 832 of them. Out of the 980 patients without diabetes the model accurately predicted 862 of them (figure 3).

Discussion

In the initial exploratory data analysis, we noted that all the datasets appeared to be normally distributed, with only A1c and FBS having visually significant distributional differences depending on diabetes diagnosis. Moreover, when looking at the boxplots, we noted that there were some data points outside of the outer fences, but they were all within accepted biomarker values.

The initial baseline random forest was highly accurate in predicting diabetes based on the biomarker dataset with an accuracy of 85% and a high recall score of 87% for the diabetes diagnosis. This implies that there is a low false negative rate when diagnosing someone with diabetes. This allows clinicians to capture almost all patients who have diabetes from their biomarkers alone. Another important score to note is the high precision score of 88% for healthy individuals, which implies that it can accurately detect when someone is healthy based on biomarkers.

When analyzing feature importance using mean decrease in impurity (MDI) we were able to determine that the most important biomarkers needed to predict diabetes were the A1c and FBS (figure 4). This is expected since of A1c levels and FBS is understandable since diabetes is primarily diagnosed through these blood tests.

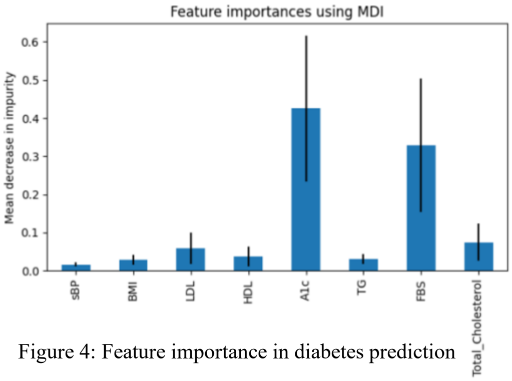

Figure 4: Feature importance in diabetes prediction

A major limitation encountered was having multiple records from the same individuals in the dataset. These records are different time points in which sBP was measured for the same individual but will have the same dates for diagnosis onsets and possible duplicates in measurement dates. This violates the independence assumption for random forests, and the accuracy of our models may be overinflated due to predicting for patients that are in both the train and test sets. One potential solution is to preprocess our data such that there is only one record from each individual, which can be chosen randomly, determined by some decision rule (i.e. earliest record in time), or as an average of the records.

The next stage of our research is to train a random survival forest. We note that the random forest model is a classification model, whereas the random survival forest is a regression model predicting survival time, and thus the two cannot be directly compared. Instead, the intention is for clinicians to use both models together as a predictive tool for diabetes prevention.

Contributions

Hunter Pozzebon: data preprocessing, writing, literature review, editing
Priyonto Saha: exploratory data analysis, model building, writing, literature review, editing
Yacine Marouf: data preprocessing, exploratory data analysis, writing, literature review, editing

All member of the group split work evenly in terms of coding and report writing.

References

3. Prevention or delay of type 2 Diabetes:Standards of medical care in diabetes—2021. (2020). Diabetes Care, 44(Supplement_1), S34-S39. https://doi.org/10.2337/dc21-s003

Ahmed, F., AL-Habori, M., Al-Zabedi, E., & Saif-Ali, R. (2021). Impact of triglycerides and waist circumference on insulin resistance and β-cell function in non-diabetic first-degree relatives of type 2 diabetes. BMC Endocrine Disorders, 21(1). https://doi.org/10.1186/s12902-021-00788-5

Biavaschi, M., Melchiors Morsch, V. M., Jacobi, L. F., Hoppen, A., Bianchin, N., & Chitolina Schetinger, M. R. (2023). Predisposition to type 2 diabetes in aspects of the glycemic curve and glycated hemoglobin in healthy, young adults: A cross-sectional study. Canadian Journal of Diabetes, 47(7), 587-593. https://doi.org/10.1016/j.jcjd.2023.05.009

Diabetes Prevention Program Research Group. (2002). Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. New England Journal of Medicine, 346(6), 393-403. https://doi.org/10.1056/nejmoa012512

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. The Annals of Applied Statistics, 2(3), 841–860.

Joslin Education Team. (n.d.). Can Type 2 Diabetes Be Reversed? Beth Israel Lahey Health, Joslin Diabetes Center. https://www.joslin.org/patient-care/diabetes-education/diabetes-learning-center/can-type-2-diabetes-be-reversed#:~:text=According%20to%20recent%20research%2C%20type,by%20losing%20significant%20amounts%20of

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

Naveed, I., Kaleem, M. F., Keshavjee, K., & Guergachi, A. (2023). Artificial intelligence with temporal features outperforms machine learning in predicting diabetes. PLOS Digital Health, 2(10), e0000354. https://doi.org/10.1371/journal.pdig.0000354

Perveen, S., Shahbaz, M., Saba, T., Keshavjee, K., Rehman, A., & Guergachi, A. (2020). Handling irregularly sampled longitudinal data and prognostic modeling of diabetes using machine learning technique. IEEE Access, 8, 21875-21885. https://doi.org/10.1109/access.2020.2968608

Yang, M., Chang, W., Kuo, T., Shen, M., Yang, C., Tien, Y., Lai, B., Chen, Y., Chang, Y., & Yang, W. (2021). Identification of novel biomarkers for pre-diabetic diagnosis using a Combinational approach. Frontiers in Endocrinology, 12. https://doi.org/10.3389/fendo.2021.641336

Code

Colab Link: https://colab.research.google.com/drive/1GAhB-6ImGI7mXwgZbsN7EV5lrkdWWEpd?usp=sharing
Github Link: https://github.com/P-Saha/16_CHL5230-F23_Phase-2_Datathon-3
This GitHub repository is private but Jasper has been as a contributor and should have access