Diabetes Prediction & Risk Factor Analysis

Understanding Differences in Prediction of Prediabetes and Diabetes

Taylor Erickson
Applied Computer Science
University of Colorado Boulder
Boulder Colorado US
taer7274@colorado.edu

PROBLEM STATEMENT/MOTIVATION

Type 2 diabetes is a common chronic illness, affecting roughly 1 in 10 American adults. It is caused by a prolonged inability to maintain healthy blood sugar levels. Diabetes is associated with many adverse outcomes, including heart disease, blindness, and kidney disease [1]. While there is no known cure for Type 2 diabetes, it is possible for patients to experience reduction of symptom severity or even reversal of the disease through treatment [2]. Thus, early detection and regular screening of high-risk patients is imperative.

One of the most widely studied risk factors for type 2 diabetes is the presence of prediabetes. Considered an intermediate stage of hyperglycemia, prediabetes is diagnosed when blood sugar levels are higher than normal but are not high enough to be considered diabetes. 25% of patients with prediabetes will develop diabetes within 3-5 years, and up to 70% will develop diabetes over the span of their lifetime. In addition to an increased risk of diabetes, prediabetes is also associated with myriad adverse health outcomes, including increased risk of cardiovascular disease, peripheral neuropathy, and stroke. While a very common condition, affecting roughly 1 in 3 American adults, prediabetes often goes undiagnosed, with 90% unaware they have the disease [3, 4]. Widespread screening of high-risk individuals has been recommended for detection of diabetes [5, 6]. However, universal screening for prediabetes is currently not recommended [7]. This is in part due to inconsistent findings on potential risk factors, making it difficult to identify who to screen and how to justify the cost [8, 9]. However, emerging evidence suggests that prediabetes is a distinct disease with its own set of risk factors, diagnostic criteria and disease progression [10, 11, 12, 13]. Unfortunately, there is still a gap in understanding of risk factors and appropriate screening criteria for prediabetes as a distinct disease.

The goal of this project is to leverage logistic regression techniques to identify the strongest predictors of prediabetes and diabetes as separate outcomes and compare them to current screening standards. This will help identify whether screening approaches could be expanded to improve the likelihood of diagnosis and early detection in individuals who are at high risk for prediabetes.

1 Literature Survey

There are a set of widely accepted risk factors for diabetes. They are commonly grouped as behavioral, cardiometabolic, and non-modifiable attributes. The most significant behavioral risk factors include physical activity, diet, smoking behavior, and alcohol consumption. The most significant cardiometabolic risk factors include obesity, hypertension, and dyslipidemia. Finally, the most significant attribute risk factors include gender, race/ethnicity, age, and genetic factors. [16, 17, 18, 19] When reviewing the available literature, these risk factors are well researched and thoroughly proven to be associated with diabetes. However, there is much less clarity on prediabetes risk factors. The first issue is that many studies combine prediabetes and diabetes into a single outcome, making it difficult to identify the differences between prediabetes and diabetes risk factors [24]. The second issue is that there are differing definitions of prediabetes, which causes inconsistency in modeled risk factor strengths [20]. The third issue is that studies investigating prediabetes often consider prediabetes as a risk factor to progression of diabetes, rather than an outcome of interest [21, 22, 23].

However, if we consider studies that have investigated prediabetes risk factors specifically, we find there are some similar risk factors between diabetes and prediabetes. Examples of overlapping risk factors include gender, age, hypertension, dyslipidemia, obesity, and diet [10, 25, 29, 31, 32, 36]. However, there is not consistency between studies on which attributes are strongly associated with prediabetes. Some studies find that factors like gender, age, or diet are not strongly associated with prediabetes, and others find that factors like marital status, income, and education level are significantly associated [25, 27, 29, 31, 32, 34, 37]. This inconsistency is likely due to unbalanced class sizes, low incidence of prediabetes patients, inconsistency in the definition of prediabetes, and overconfidence in the similarity of the pathogenesis of prediabetes and diabetes [8, 11, 20, 28].

Generally, the variance in risk factors for prediabetes is related to a greater difficulty identifying statistically significant risk factors for prediabetes and a reduced strength of prediction for risk factors that overlap with diabetes [27, 30, 32].

For example, a study by Okwechime et al. (2015) found that being overweight, obese, hypertensive, hypercholesterolemic, and arthritic were significantly associated with prediabetes. However, they also found that these predictors were more strongly associated with diabetes. Additionally, many predictors that were not strongly associated with prediabetes were strongly associated with diabetes, including age, income level, and level physical activity [27]. This inconsistency demonstrates the need for more thorough analysis of prediabetes risk factors and how they are distinct from or related to diabetes risk factors.

2 Proposed Work

The goal of this project is to leverage logistic regression modeling techniques to measure the strength of potential risk factors in predicting prediabetes and diabetes as separate classes. We will create two separate logistic regression models, one for diabetes as an outcome and one for prediabetes as an outcome. Additionally, we will fit a multinomial logistic regression model on all three of our classes of interest, diabetes, no diabetes, and prediabetes. We leverage logistic regression modeling because it is a good fit for the data mining task and has been previously used as a successful method in this research domain [26, 27, 31, 35]. Then, we will compare the risk factors for diabetes and prediabetes to see whether they are the same or different. Finally, we will compare against existing research and documented risk factors to confirm whether we were able to replicate results reported by previous studies.

We will first complete basic data cleaning and exploration. The datasets have already undergone extensive data cleaning, with the removal of null or missing values and numeric transformation of categorical attributes. The primary focus of this step will be to examine the variance and distribution of each predictor compared to our outcomes of interest, to identify whether it is a potential predictor. We will leverage techniques like the t-test and levene statistic to measure equality of variance and influence of the feature on our outcome of interest. We will compare the variance across all three outcomes of interest, prediabetes, diabetes and no diabetes. Variables with low or no variance will be dropped.

Additionally, we will leverage techniques like Variance Inflation Factor (VIF) measures, Principal Component Analysis (PCA), and correlation heatmaps to identify collinear features and drop them from our dataset. In addition, there are only a small number of responses for the prediabetes class, so we will leverage oversampling techniques like SMOTE to create a more balanced dataset.

Once we have identified the best potential predictors, we will use logistic regression modeling techniques to build our models. We will first create a best fit logistic regression model on our diabetes vs. no diabetes dataset. This mirrors other previously conducted work on modeling diabetes risk factors and should replicate similar results to the logistic regression models created by the CDC and others on similar datasets [27, 33]. We will attempt to replicate the findings with the balanced dataset that combines prediabetes and diabetes into one outcome of interest to see whether unbalanced class sizes might be a contributing factor to the inconsistency of previous findings. This mirrors previous work that predicts diabetes and pre-diabetes as a single outcome of interest [24]. Then, we will create a logistic regression model to compare between our outcome classes of prediabetes vs. no diabetes. This will allow us to compare diabetes and prediabetes as separate outcomes. Finally, we will create a multinomial logistic regression model to predict between our three classes, no diabetes, diabetes and pre-diabetes. This model will leverage a similar approach to the work done by Okwechime et al [2015], but our study will be conducted on a more recent, larger and more representative sample. Additionally, we will compare the multinomial logistic regression model to the single class logistic regression model to introduce a novel comparison of multiple logistic regression models comparing diabetes to prediabetes predictors within one study.

Once we have created and fit our models, we will evaluate their performance. We will use statistical measures like area under the receiver operating characteristic (ROC) curve (AUC) to evaluate the strength of our model. We will evaluate p-values for our coefficients to understand the strength of our feature coefficients, with $p <= 0.05$ as significant. Once we have identified the strength of our relative models, we will revise any poorly performing models, leveraging techniques like forward stepwise refinement to improve the performance of our models.

3 Data Sets

This project will leverage four separate data sets, comprised of results from the 2015 CDC Behavioral Risk Factor Surveillance System (BRFSS) survey. The annual survey has been conducted since 1984 and collects telephone survey responses from over 400,000 Americans on a range of health conditions, health behaviors, and health services [14]. The data sets used in this project are a subset of the publicly available records, comprised of features that are specifically related to diabetes, prediabetes, and related risk factors. The data is available via Kaggle and can be downloaded as separate files [15].

The first dataset is a cleaned dataset of 253,680 survey responses with 3 classes in the target variable, no diabetes (or diabetes only during pregnancy), prediabetes, and diabetes. There is a class imbalance in this dataset, with an uneven distribution amongst the three classes. There are 21 attributes of interest.

The second dataset is a cleaned subset of the first, comprised of 70,692 responses with an even split between respondents with no diabetes or respondents with prediabetes or diabetes. There are two target classes in this dataset, no diabetes vs. prediabetes or diabetes.

The third dataset is comprised of the same responses as the first dataset, except that the target variable is transformed to a binary target variable, no diabetes vs. prediabetes or diabetes.

4 Evaluation Methods

The primary evaluation method for our project will be to leverage statistical measures like ROC/AUC curve analysis and p-values of our coefficients to determine the strength of our models and of our predictors. Additionally, we plan to compare our identified risk factors against the existing literature.

We expect to see similarities between our selected features and those identified by similar methods. Additionally, we will compare our results against identified practitioner best practices and screening standards to identify any similarities or novel differences.

5 Tools

The project will leverage industry best practices for statistical modeling. The analysis will be conducted in Python, and requires use of these key statistical, modeling, and visualization libraries: pandas, seaborn, matplotlib, scipy, statsmodels, sklearn, numpy, yellowbrick.

The project will be completed in an interactive Jupyter notebook file, to improve readability and understanding of the modeling process. Visualizations and written explanation will be displayed next to the model generation and statistical analysis code blocks.

Given the size of the dataset, the data cleaning and processing does not require any other specialized data mining tools outside of what will be completed in the Juypter notebook.

6 Milestones

6.1 Milestone 1. July 17th – Ingestion and initial visual exploration of features.

6.2 Milestone 2. July 24th – Initial logistic regression models.

6.3 Milestone 3. July 31st – Analysis and re-work to strengthen model performance.

6.4 Milestone 4. August 7th – Comparison between results and existing research.

6.5 Milestone 5. August 14th – Finished write-up and project delivered.

REFERENCES

[1] CDC. 2023. Type 2 Diabetes. (April 2023). Retrieved July 10, 2023 from https://www.cdc.gov/diabetes/basics/type2.html

[2] Sarah J Hallberg, Victoria M Gershuni, Tamara L Hazbun, and Shaminie J Athinarayanan. 2019. Reversing Type 2 Diabetes: A Narrative Review of the Evidence. *Nutrients* 11, 4, (April 2019), 766. DOI: https://doi.org/10.3390/nu11040766

[3] Adam G. Tabák, Christian Herder, Wolfgang Rathmann, Eric J. Brunner, and Mika Kivimäki. Prediabetes: a high-risk state for diabetes development. *Lancet* 379, 9833, (June 2012), 2279-2290. DOI: https://doi.org/10.1016/S0140-6736(12)60283-9

[4] Ulrike Hostalek. Global epidemiology of prediabetes – present and future perspectives. *Clinical Diabetes and Endocrinology* 5, 5, (May 2019). DOI: https://doi.org/10.1186/s40842-019-0080-0

[5] US Preventive Services Task Force. Screening for Prediabetes and Type 2 Diabetes: US Preventive Services Task Force Recommendation Statement. *JAMA*. 326, 8, (August 2021), 736–743. DOI: https://10.1001/jama.2021.12531

[6] Kenneth Lam and Sei J. Lee. Prediabetes – A Risk Factor Twice Removed. *JAMA Internal Medicine* 181, 4, (August 2021), 520-521. DOI: https://doi.org/10.1001/jamainternmed.2020.8773

[7] Aditya K. Khetan and Sanjay Rajagopalan. Prediabetes. *Canadian Journal of Cardiology* 34, 5, (May 2018), 615-623. DOI: https://doi.org/10.1016/j.cjca.2017.12.030

[8] Ashkann Zand, Karim Ibrahim and Bhargavi Patham. Prediabetes: Why Should We Care? Methodist Debakey Cardiovascular Journal,14, 4, (October 2018), 289-297. DOI: https://doi.org/10.14797/mdcj-14-4-289

[9] Justin B. Echouffo-Tcheugui and Elizabeth Selvin. Prediabetes and What It Means: The Epidemiological Evidence. *Annual Review of Public Health* 42, 59, (April 2021), 59-77. DOI: https://doi.org/10.1146/annurev-publhealth-090419-102644

[10] Alicia Diaz-Redondo, Carolina Giráldez-García, Lourdes Carrillo et al. Modifiable risk factors associated with prediabetes in men and women: a cross-sectional analysis of the cohort study in primary health care on the evolution of patients with prediabetes (PREDAPS-Study). *BMC Family Practice*, 16, Article 5, (January 15). DOI: https://doi.org/10.1186/s12875-014-0216-3

[11] J.F. Elgart, R. Torrieri, M. Ré et al. Prediabetes is more than a pre-disease: additional evidences supporting the importance of its early diagnosis and appropriate treatment. *Endocrine* 79, (January 2023), 80-85. DOI: https://doi.org/10.1007/s12020-022-03249-8

[12] Jan Brož, Jana Malinovska, Marisa A. Nunes et al. Prevalence of diabetes and prediabetes and its risk factors in adults aged 25-64 in the Czech Republic: A cross-sectional study. *Diabetes Research and Clinical Practice*, 170, Article 108470, (September 2020). DOI: https://doi.org/10.1016/j.diabres.2020.108470

[13] Khaled K. Aldossari, Abdulrahman Alidiab, Jamaan M. Al-Zahrani et al. Prevalence of Prediabetes, Diabetes, and Its Associated Risk Factors among Males in Saudi Arabia: A Population-Based Survey. *Journal of Diabetes Research* 2018, Article 2194604 (April 2018), 12 pages. DOI: https://doi.org/10.1155/2018/2194604

[14] CDC. Behavioral Risk Factor Surveillance System. (May 2023). Retrieved July 10, 2023 from https://www.cdc.gov/brfss/index.html

[15] Alex Teboul. Diabetes Health Indicators Dataset. (2021). Retrieved July 10, 2023 from https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

[16] Sania Siddiqui, Hadzliana Zainal, Sabariah Noor Harun, Siti Maisharah Sheikh Ghadzi, and Saadia Ghafoor. Gender differences in the modifiable risk factors associated with the presence of prediabetes: A systematic review. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 14, 5, (July 2020), 1243-1252. DOI: https://doi.org/10.1016/j.dsx.2020.06.069

[17] CDC. Diabetes Risk Factors. (April 2022). Retrieved July 10, 2023 from https://www.cdc.gov/diabetes/basics/risk-factors.html

[18] Barbara Fletcher, Meg Gulanick, and Cindy Lamendola. Risk Factors for Type 2 Diabetes Mellitus. *The Journal of Cardiovascular Nursing*, 16, 2, (January 2002), 17-23.

[19] Phillipa J Talmud, Aroon D. Hingorani, Jackie A Cooper et al. Utility of genetic and non-genetic risk factors in prediction of type 2 diabetes: Whitehall II prospective cohort study. *BMJ* 340, Article 4838, (January 2010). DOI: https://doi.org/10.1136/bmj.b4838

[20] Crystal Man Ying Lee, Stephen Colagiuri, Mark Woodward et al. Comparing different definitions of prediabetes with subsequent risk of diabetes: an individual participant data meta-analysis involving 76,513 individuals and 8,208 cases of incident diabetes. *BMJ Open Diabetes Research and Care,* 7, 1, (November 2019). DOI: https://doi.org/10.1136/ bmjdrc-2019-000794

[21] Bennasar-Veny M, Fresneda S, López-González A, Busquets-Cortés C, Aguiló A, Yañez AM. Lifestyle and Progression to Type 2 Diabetes in a Cohort of Workers with Prediabetes. *Nutrients,* 12, 5, Article 1538 (May 2020,. DOI: https://doi.org/10.3390/nu12051538

[22] Rooney MR, Rawlings AM, Pankow JS, et al. Risk of Progression to Diabetes Among Older Adults With Prediabetes. *JAMA Intern Med.*, 181, 4, (February 2021), 511-519. DOI: https://doi.org/10.1001/jamainternmed.2020.8774

[23] N. Anthony, V. Lenclume, A. Fianu, N.Le Moullec, X. Debussche, P. Gérardin, C. Marimoutou, E. Nobécourt, Association between prediabetes definition and progression to diabetes: The REDIA follow-up study. *Diabetes Epidemiology and Management,* 3, Article 100024 (November 2021). DOI: https://doi.org/10.1001/jamainternmed.2020.8774

[24] Hai Wang, Xin Zheng, Zheng-Hai Bai, Jun-Hua Lv, Jiang-Li Sun, Yu Shi, and Hong-Hong Pei. A retrospective population study to develop a predictive model of prediabetes and incident type 2 diabetes mellitus from a hospital database in Japan between 2004 and 2015. *Medical Science Monitor*, 26, (April 2020). DOI: https://doi.org/10.12659/MSM.920880

[25] Jiahua Wu, Jiaqiang Zhou, Xueyao Yin, Yixin Chen, Xihua Lin, Zhiye Xu, and Hong Li. A Prediction Model for Prediabetes Risk in Middle-Aged and Elderly Populations: A Prospective Cohort Study in China. *International Journal of Endocrinology*, 2021, Article 2520806 (November 2021). DOI: https://doi.org/10.1155/2021/2520806.

[26] Xue-Hui Meng, Yi-Xiang Huang, Dong-Ping Rao, Qiu Zhang, and Qing Liu. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors, *The Kaohsiung Journal of Medical Sciences*, 29, 2 (October 2012), 93-99. DOI: https://doi.org/10.1016/j.kjms.2012.08.016

[27] Ifechukwude Obiamaka Okwechime , Shamarial Roberson, and Agricola Odoi. Prevalence and Predictors of Pre-Diabetes and Diabetes among Adults 18 Years or Older in Florida: A Multinomial Logistic Modeling Approach. *PLOS ONE,* 10, 12, Article e0145781 (2015). DOI: https://doi.org/10.1371/journal.pone.0145781

[28] Gary S. Collins, Susan Mallett, Omar Omar, and Ly-Mee Yu. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med*icine 9, 1, (December 2011), 1-14.

[29] Mayo Clinic. 2022. Prediabetes. (November 2022). Retrieved July 10, 2023 from https://www.mayoclinic.org/diseases-conditions/prediabetes/symptoms-causes/syc-20355278

[30] Vijay Viswanathan, Satyavani Kumpatla, Vigneswari Aravindalochanan et al. Prevalence of Diabetes and Pre-Diabetes and Associated Risk Factors among Tuberculosis Patients in India. *PLOS ONE,* 7, 7, (2012), Article ID e41367. DOI: https://doi.org/10.1371/journal.pone.0041367

[31] Zhao, Ming, Hongbo Lin, Yanyan Yuan, Fuyan Wang, Yang Xi, Li Ming Wen, Peng Shen, and Shizhong Bu. Prevalence of Pre-Diabetes and Its Associated Risk Factors in Rural Areas of Ningbo, China*.* *International Journal of Environmental Research and Public Health* 13, 8, Article 808 (August 2016). DOI: https://doi.org/10.3390/ijerph13080808

[32] Hemavathi Dasappa, Farah Naaz Fathima, Rugmani Prabhakar, Sanjay Sarin. Prevalence of diabetes and pre-diabetes and assessments of their risk factors in urban slums of Bangalore. *Journal of Family Medicine and Primary Care*, 4, 3 (July 2015), 399-404. DOI: https://doi.org/10.4103/2249-4863.161336

[33] Xie Z, Nikolayeva O, Luo J, Li D. Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. *Prev Chronic Dis* 16, Article 130, (September 2019). DOI: https://doi.org/10.5888/pcd16.190109

[34] F Hadaegh, A Derakhshan, N Zafari, D Khalili, M Mirbolouk,N Saadat, and F Azizi. Pre-diabetes tsunami: incidence rates and risk factors of pre-diabetes and its different phenotypes over 9 years of follow-up. *Diabet Med*. 34, 1, (January 2017), 69-78. DOI: https://doi.org/10.1111/dme.13034

[35] Rui Wang, Peng Zhang, Zhijun Li et al. The prevalence of pre-diabetes and diabetes and their associated factors in Northeast China: a cross-sectional study. *Scientific Reports*, 9, Article 2513, (February 2019). https://doi.org/10.1038/s41598-019-39221-2

[36] Parisa Amiri, Sara Jalali-Farahani, Mehrdad Karimi et al. Factors associated with pre-diabetes in Tehranian men and women: A structural equations modeling. *PLOS ONE*, 12, 12, Article e0188898, (December 2017). https://doi.org/10.1371/journal.pone.0188898

[37] Anna Zamora-Kapoor, Amber Fyfe-Johnson, Adam Omidpanah, Dedra Buchwald, and Ka'imi Sinclair. Risk factors for pre-diabetes and diabetes in adolescence and their variability by race and ethnicity. *Preventive Medicine*, 115, (August 2018), 47-52. DOI: https://doi.org/10.1016/j.ypmed.2018.08.015