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

CIND 820: Big Data Analytics Project

Literature Review, Data Description and Project Approach

Yitayal Mengistu [501211839]

Submitted to:

Dr. Ceni Babaoglu (supervisor)

June10, 2024

Contents

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.

1. Literature Review, Data Description and Project Approach

Historically, diabetes diagnosis has relied on clinical tests such as fasting blood glucose and Glycated Hemoglobin (HbA1c) levels(Lu et al., 2023). With the rising prevalence of type-2 diabetes, advances in computing power and generation of big data, predictive modeling for type-2 diabetes detection has garnered significant interest among data scientists and health researchers recently (Khan et al., 2021; Kibria et al., 2022; Rani, 2020; Wang et al., 2024; Xie et al., 2019; Zou et al., 2018). In the following sections, a review of selected related literatures on predictive modeling for type-2 diabetes detection during the past 10 years is undertaken focusing on specific research questions raised, techniques and tools used, conclusions arrived. Then, the structure of the dataset being used for the current project is described and the project approach is outlined.

1.1. Review of Literature on Predictive Modeling for Type 2 Diabetes

Various machine learning algorithms have been trained and tested to answer similar or closely related research questions as the current project, predicting the risk of type 2 diabetes, during the past decades (Khan et al., 2021; Kibria et al., 2022; Rani, 2020; Wang et al., 2024; Xie et al., 2019; Zou et al., 2018). Their main focus was on how to increase the prediction accuracy of the ML models to predict type 2 diabetes. However, these research projects vary with regard to model selection, size and structure of dataset, data preprocessing techniques, feature selection approach, evaluation metrices used and geographic cover. Below, some selected projects undertaken on predictive modeling for type 2 diabetes risk prediction between 2014 and 2024 are reviewed, their limitations discussed and the objective of this project towards improving and fixing these limitations explained.

Zou et al. (2018) used decision tree, random forest and neural network algorithms to train, test and predict diabetes mellitus using a hospital physical examination dataset on 14 features and 68,994 individuals in Liuzhou, China. The primary research question was “What are the comparative performances of ML algorithms in predicting diabetes mellitus?” They used imbalanced data set for training the algorithms. This study used principal component analysis (PCA) to reduce the dimensionality of the data. The results showed that prediction with random forest achieved the highest accuracy of 80.84%.

Xie et al. (2019) trained risk prediction models for type 2 diabetes using several ML classifiers including SVM, Gaussian Naive Bayes, Logistic Regression, Neural Network, Decision Tree, and Random Forest using the 2014 BRFSS survey data set in USA. This study uses a similar, US 2015 BRFSS data set. The primary research questions this study tried to answer were “what is the most significant risk factor for type-2 diabetes?” and “what is the performance comparison of different ML algorithms?”. The study used SMOTE techniques to balance the training dataset. They found out that the neural network prediction model had the highest accuracy (82.4%), specificity (90.2%), and AUC (0.7949) values. In contrast, the decision tree prediction model had the highest sensitivity (78.2%). The analysis also showed that not only under sleeping (≤6 hours per day) but also over sleeping (≥9 hours per day) increases the risk for type 2 diabetes. This study focusses only on the performance of ML algorithms and forgot to explain how to improve the interpretability of the models.

Rani (2020) employed various machine learning algorithms, including KNN, logistic regression, decision trees, random forests, and support vector machines (SVM) to answer the following research questions: “Which machine learning classification algorithms are most effective in predicting diabetes?” and “What level of accuracy can be achieved in predicting diabetes using different machine learning algorithms?” In this study, 9 features, a small data set, 2000 cases, and unbalanced classes for the training dataset is used (1400 vs 700). The decision tree model achieved the highest accuracy of 99%. This extreme level of accuracy may be caused by the data imbalance where the model trained more from the majority class.

Kibria et al. (2022) attempt to use ensemble model with 5-fold cross validation approach by combining two ML algorithms (Random Forest and, XGBoost), using a weighted soft voting classifier to predict the risk of diabetes. The primary research question of this project was: “What methods can be employed to enhance the explainability and interpretability of machine learning models for diabetes prediction?” The results of this study indicate that the model achieved an accuracy of 90% and a F1 score of 89%. The authors tried to use a technique called Shapley additive explanations (SHAP) to interpret machine learning model predictions, specifically to aid physicians in understanding these predictions.

Compared to other projects reviewed in this section, one key lesson learned from the methodological approach of Kibria et al. (2022) is the importance of balancing the accuracy and interpretability of ML models for predicting the risk of diabetes. Among other things, the ensemble method along with these techniques could be taken as an important step towards balancing accuracy and interpretability of ML models in feature studies. However, Kibria et al. (2022) used a very small dataset, 768 instances and 9 attributes and the class of the training dataset was highly imbalanced. They tried to balance the target classes but another issue with the data was that it contains information only on female patients limiting the importance or validity of the model for generalization.

Alam et al. (2023) raised the following two research questions to identify the most accurate ML predictive models and key features for predicting diabetes: “Which ML model is the most accurate and suitable for classifying the presence of diabetes?" and "What are the key features for the classification model for predicting diabetes in patients?" The main features included in this study are pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age and outcome. Alam et al. (2023) found out that the Random Forest classifier is the best classifier for the detection of diabetes in patients and selected 5 out of 8 features as important predictors (the number of pregnancies, BMI, blood pressure, glucose level and age. The limitation of this project is that it used a vey small dataset, 768 cases and 9 features, and observation were only female patient.

Wang et al. (2024) tried to build a risk prediction model for diabetic retinopathy, one of the major complications of diabetes, using 5 ML algorithms: CatBoost, SVM, RF, XGBoost and GBT and 10-fold cross validation techniques. The primary research question was "how can the accuracy and interpretability of machine learning models for diabetic retinopathy prediction be improved"? The authors used very small dataset of 1000 cases and 87 variables. Their findings indicate that Cat Boost exhibited the best overall performance, with the accuracy value of 0.8250, the precision value of 0.8211, the recall value of 0.8125 and the F1 value of 0.8168. The analysis also indicated that the authors identified the following potential risk factors for diabetic retinopathy: blood pressure level, anemia or renal illness, low FBG points to poor blood clotting capacity, and low PCV and FBG together.

The review of these research projects shows that several gaps and areas for improvement or further exploration can be identified: First, most of the studies utilized relatively small datasets, fewer than 1000 instances, with limited diversity (Kibria et al., 2022; Alam et al.,2023; and Wang et al., 2024)). Small sample sizes can lead to overfitting and reduced generalizability. Future studies should aim to use larger and more diverse datasets that include a balanced representation of different genders, ages, and ethnic groups to improve the generalizability of the models.

Some of these studies struggled with imbalanced datasets. (Rani, 2020 and Kibria et al., 2022). Imbalanced datasets can skew the results and class distributions could lead to biased models favoring the majority class. Implementing more advanced data balancing techniques such as SMOTE, or other oversampling and undersampling methods to ensure balanced datasets for training data set is important.

The studies utilized a variety of features, but there was inconsistency in the number and type of features considered. Zou et al. (2018) used PCA for dimensionality reduction, while others did not explicitly mention their feature selection processes. Alam et al. (2023) highlighted the importance of certain features but used a limited set of 8 features. Future research should adopt systematic feature selection techniques and explore more comprehensive feature sets, possibly incorporating domain expertise and newer techniques like recursive feature elimination (RFE).

Only two studies Kibria et al. (2022) and Xie et al. (2019) discussed the importance of balancing accuracy and interpretability of their models. Kibria et al. (2022) made efforts to balance accuracy with interpretability using Shapley Additive Axplanations (SHAP). Xie et al. (2019) noted the lack of focus on model interpretability in other studies. Future projects should prioritize the development of interpretable models, using methods such as SHAP or inherently interpretable models like decision trees, to ensure that predictions are understandable to clinicians.

The above reviewed projects varied in their use of evaluation metrics and validation techniques. Rani (2020) reported an extremely high accuracy (99%), likely due to data imbalance. Kibria et al. (2022) and Wang et al. (2024) used cross-validation techniques which are more robust.

Based on this it is clear to suggest that employing a standardized set of evaluation metrics (accuracy, precision, recall, F1 score, AUC) and robust validation techniques (e.g., k-fold cross-validation) to ensure reliable and comparable results across studies.

While most studies focused on general diabetes prediction, Wang et al. (2024) explored diabetic retinopathy, a specific complication of diabetes. Though this out of the scope of this project, the review highlighted the importance that future research should consider not only the prediction of diabetes but also its complications, such as neuropathy, nephropathy, and retinopathy,(Lu et al., 2023) to provide comprehensive care insights.

In summary, the review of these related literature highlighted the importance of addressing these research gaps and the need to prioritize the use of larger and more diverse datasets, employ advanced data balancing techniques, standardizing and adopt systematic feature selection methods, emphasize model interpretability and using robust evaluation metrics can significantly enhance the accuracy, generalizability, and clinical applicability of machine learning models in predicting diabetes.

This research project builds on the foundation established by previous studies in type-2 diabetes prediction using machine learning (ML) algorithms. By addressing some of the key gaps identified in the literature review above, this project aims to advance both the methodological rigor and practical applicability of these predictive models. Previous research, such as that by Zou et al. (2018) and Xie et al. (2019), has demonstrated the potential of various ML algorithms in predicting diabetes, yet these studies often utilized very small and relatively homogeneous datasets. This project will overcome these limitations by leveraging a large and diverse dataset, thereby enhancing the generalizability of the findings across different populations.

Furthermore, while earlier studies have explored the performance of different ML algorithms, there has been a lack of systematic investigation into the impact of feature selection methods and hyperparameter tuning on model performance. This research project will advance robust evaluation metrics, alongside a thorough examination of feature selection methods and hyperparameter optimization. This comprehensive approach will not only compare the accuracy of various ML algorithms in predicting diabetes risk but also identify the most predictive features, offering valuable insights for prevention and public health policy. By emphasizing model interpretability, this project will ensure that the predictive models are both accurate and actionable, ultimately contributing to more effective diabetes management and intervention strategies. Specifically:

- The performance of three selected ML algorithms in predicting diabetes risk, chosen based on their inherent interpretability, will be compared
- This study will investigate which features are most predictive of diabetes risk and the implications for prevention and public health policy and
- It will examine the impact of feature selection methods and hyperparameter choices on the performance of the selected ML algorithms and emphasize model interpretability to ensure that the predictive insights are actionable and useful for healthcare practitioners and policymakers.

Despite its comprehensive approach, this study has some limitations. First, the dataset, though large and diverse, is collected via phone call surveys alone rather than from medical records. This may introduce self-report biases and limit the accuracy of the data, as participants might misreport or inaccurately recall information related to their health status and behaviors.

Second, due to time constraints, the study is restricted to using only three ML algorithms. While these algorithms are selected for their balance of interpretability and accuracy, this limitation might overlook other potentially important algorithms that could offer better predictive performance or insights. Third, while the research aims to investigate feature selection methods, the scope of features analyzed may still be limited by the data collected and the focus of the phone survey, potentially omitting important predictors available in more comprehensive medical datasets.

By acknowledging these limitations, the study aims to provide a balanced and realistic assessment of the capabilities and implications of using ML for diabetes risk prediction, paving the way for future research to build upon these findings with more refined data and broader algorithmic exploration.

1.2. Data Description, Cleaning and Preprocessing

In this section, a brief description of the source and structure of the dataset used for this project is provided and the steps taken to clean and pre-process the data set is described in sufficient detail. The data cleaning step include checking and handling of missing values, inconsistencies, errors and outliers. The preprocessing steps include univariate, bivariate and multivariate analysis.

For this project, a publicly available 2015 USA Behavioral Risk Factor Surveillance System phone call survey dataset, originally collected by the Center for Disease Control and Prevention (CDC), is used. The dataset, designated as public domain, is accessible on the following Kagle website. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?select=diabetes_binary_5050split_health_indicators_BRFSS2015.csv.

The data set contains 70, 692 observations and 21 features. The target variable has binary classes: diabetes and non-diabetes. The data set consists of demographic variables, medical history, lifestyle factors and socioeconomic variables. Some of the main variables include age, gender, education, income, alcohol consumption, body mass index, heart disease, smoking history, and cholesterol level. For detail description of features, see Appendix-A.

The initial analysis of the dataset involved several preprocessing steps. These included renaming some of the columns for clarity, creating detailed descriptions of variable characteristics, and verifying appropriate data types. Each feature in the dataset was checked for missing values, no features exhibited missing data. However, 1635 duplicate rows were identified and dropped, and inconsistencies or anomalous values were checked. outliers and anomalies were identified based on the distance from the inter interquartile range and visual inspections using box plots.

Summary statistics were generated to gain a comprehensive understanding of the dataset. This included measures of central tendency (mean, median) and dispersion (standard deviation, interquartile range) for continuous variables, as well as frequency distributions for categorical variables. The distribution of diabetes outcome by some well known factors including smoking, cholesterol level, alcohol consumption, heart disease and blood pressure were visualized to get insights into their influence and potential impacts on diabetes outcome. The prevalence of diabetes outcome by age, education level and income level also gave important clue for the analysis and modeling of diabetes. The distribution of diabetes outcomes by good habit factors such as physical exercise, consumption of fruits and vegetables has also shown some difference among the two groups, diabetic and non diabetic groups. To understand the relationships between different features and between various features and the target variable, Pearson correlation coefficients were computed for continuous variables, while chi-square tests were used for categorical variables.

Significant correlations were visualized using heatmaps to highlight strong relationships between features expected be critical indicators of diabetes. Multivariate analysis techniques were employed to compare groups based on selected features, providing insights into the relationships and dynamics of the variables.

Major activities in the cleaning and data preprocessing steps include the following (see also the python output pdf file):

- Columns renamed to have clear, short and lowercase names;
- The data type of each column checked and/or converted to appropriate ones;
- 1635 duplicate rows Identified and dropped;
- Checked the presence of missing values. No missing values found
- Aggregation of feature categories to fewer levels for age, education and income
- Frequency distribution of the target/outcome variable, diabetes
- Outliers detected, visualized, and removed
- Histograms to visualize the distribution of numeric features.
- scatter plots to explore relationships between variables.
- Summary statistics for numeric features.
- Correlation matrix/heat map to visualize the correlation between numeric features
- Chi-square test for independence of categorical variables against 'diabetes'
- Bar graph -The distribution of diabetes outcome by ‘risk factors’

- Bar graph - The distribution of diabetes outcome by ‘good habit features’
- Bar graph - Prevalence of diabetes by age, gender, education and income level

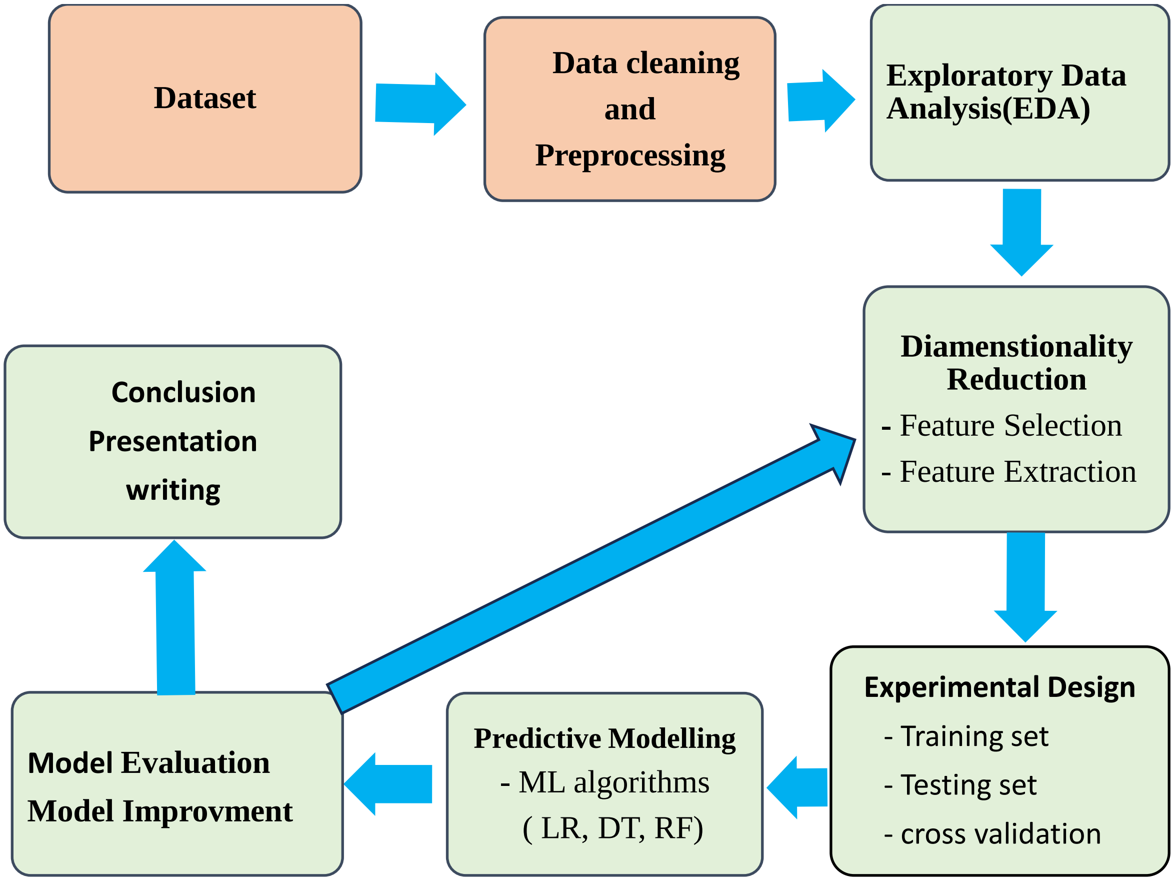
Overall, the dataset appeared robust after cleaning, with some attributes showing logical and expected distributions. By carefully cleaning and preprocessing the data, a foundation was established for further steps of exploratory data analysis, data partition and building accurate and interpretable predictive models. The python files used for generating the data cleaning and preprocessing steps can be accessed on the following GitHub link: https://github.com/ymengistu/CIND820_BigDataAnalyticsProject_DiabetesPrediction

1.3. Project Approach

The work flow chart presented in figure-1 below provides a clear and organized visual representation of the phases in this project. It outlines all the main millstones in the project, from data retrieval to conclusion, presentation and writing. The first two pink boxes indicate the progress or completed part of the project as of June 10,2024.

An important part of the flow chart is the loop created by the arrow between dimensionality reduction or feature selection and model Evaluation and model improvement steps. Since modelling in machine learning and data analysis is an iterative process, we need to repeatedly refine and improve the models by going through several cycles of selecting variables, training models, and evaluating their performances(Hohman et al., 2020). This iterative process is crucial for building effective and robust machine learning models. By continually selecting and refining variables, adjusting techniques, and evaluating performance, the accuracy of the models, generalizability, and usefulness can progressively be improved. This cyclical approach allows for continuous learning and adaptation, leading to better insights and more reliable predictions.

Figure 1: Workflow chart for predictive modeling of type-2 diabetes



Appendix 1-Data dictionary of 2015 USA BRFSS phone call survey dataset

Variable Name	Data Type	Description
ID	Integer	Patient ID
Diabetes	Binary	Target variable, 0 = no diabetes, 1 = prediabetes or diabetes
HighBP	Binary	0 = no high BP, 1 = high BP
HighChol	Binary	0 = no high cholesterol, 1 = high cholesterol
CholCheck	Binary	0 = no cholesterol checks in 5 years, 1 = yes cholesterol check in 5 years
BMI	Integer	Body Mass Index
Smoker	Binary	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no, 1 = yes
Stroke	Binary	(Ever told) you had a stroke? 0 = no, 1 = yes
HeartDiseaseorAttack	Binary	coronary heart disease (CHD) or myocardial infarction (MI)? 0 = no, 1 = yes
PhysActivity	Binary	Physical activity in the past 30 days – not including job? 0 = no, 1 = yes
Fruits	Binary	Consume Fruit 1 or more times per day 0 = no 1 = yes
Veggies	Binary	Consume Vegetables 1 or more times per day? 0 = no, 1 = yes
HvyAlcoholConsump	Binary	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) 0 = no 1 = yes
AnyHealthcare	Binary	Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no 1 = yes
NoDocbcCost	Binary	Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no 1 = yes
GenHlth	Integer	Would you say that in general your health is: scale 1-5 1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor

MentHlth	Integer	Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? scale 1-30 days
PhysHlth	Integer	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? scale 1-30 days
DiffWalk	Binary	Do you have serious difficulty walking or climbing stairs? 0 = no, 1 = yes
Sex		0 = female, 1 = male
Age	Integer	13-level age category 1 = Age 18 to 24, Notes: 18 <= AGE <= 24 2 = Age 25 to 29, Notes: 25 <= AGE <= 29 3 = Age 30 to 34, Notes: 30 <= AGE <= 34 4 = Age 35 to 39, Notes: 35 <= AGE <= 39 5 = Age 40 to 44, Notes: 40 <= AGE <= 44 6 = Age 45 to 49, Notes: 45 <= AGE <= 49 7 = Age 50 to 54, Notes: 50 <= AGE <= 54 8 = Age 55 to 59, Notes: 55 <= AGE <= 59 9 = Age 60 to 64, Notes: 60 <= AGE <= 64 10 = Age 65 to 69, Notes: 65 <= AGE <= 69 11 = Age 70 to 74, Notes: 70 <= AGE <= 74 12 = Age 75 to 79, Notes: 75 <= AGE <= 79 13 = Age 80 or older, Notes: 80 <= AGE <= 99 14 = Don’t know/Refused/Missing Notes: 7 <= AGE <= 9
Education	Integer	Education level (EDUCA see codebook) scale 1-6 1 = Never attended school or only kindergarten 2 = Grades 1 through 8 (Elementary) 3 = Grades 9 through 11 (Some high school) 4 = Grade 12 or GED (High school graduate)

		5 = College 1 year to 3 years (Some college or technical school) 6 = College 4 years or more (College graduate)
Income	Integer	annual household income from all sources Income scale 1-8 1 = Less than $10,000 2 = Less than $15,000 [$10,000, $15,000) 3 = Less than $20,000 [$15,000, $20,000) 4 = Less than $25,000 [$20,000, $25,000) 5 = Less than $35,000 [$25,000, $35,000) 6 = Less than $50,000 [$35,000, $50,000) 7 = Less than $75,000 [$50,000, $75,0000) 8 = $75,000 or more 77 = Don’t know/Not sure 99 = Refused

Source: (Burrows et al., 2017)

https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators

References

Alam, M., Khan, I. R., Alam, M. A., Siddiqui, F., & Tanweer, S. (2023). The diabacare cloud: predicting diabetes using machine learning. *Acta Scientiarum. Technology*, *46*(1), e64783. https://doi.org/10.4025/actascitechnol.v46i1.64783

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*.

Hohman, F., Wongsuphasawat, K., Kery, M. B., & Patel, K. (2020). Understanding and Visualizing Data Iteration in Machine Learning. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3313831.3376177

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

Lu, Y., Wang, W., Liu, J., Xie, M., Liu, Q., & Li, S. (2023). *Medicine ® Vascular complications of diabetes A narrative review*. https://doi.org/10.1097/MD.0000000000035285

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