Transactions on Materials, Biotechnology and Life Sciences
ISSN: 2960-1932, eISSN: 2960-2319 | Volume 3
BLSME 2024

Multi-factors correlation to diabetes using Machine Learning: findings from BRFSS

Shiying Wu *

Mathematics Department, Boston University, 665 Commonwealth Ave, Boston, MA, 02215, USA

* Corresponding Author Email: wangel@bu.edu

Abstract. The escalating prevalence of diabetes has emerged as a burgeoning national concern, necessitating comprehensive investigation and intervention. Leveraging data extracted from the 2021 Behavioral Risk Factor Surveillance System (BRFSS), this study harnesses the power of data mining techniques to discern salient variables within expansive datasets. Subsequently, a series of machine learning models, comprising Random Forest, Decision Tree, Lasso Regression, and Neural Network, were constructed, and evaluated. The principal objective of this research is to facilitate the identification of the optimal predictive model for diabetes. By scrutinizing and contrasting the performance of these diverse models, this study aspires to contribute valuable insights into the field of diabetes prediction, potentially aiding in the development of more accurate and effective diagnostic tools. Consequently, this research and to make a meaningful stride towards mitigating the adverse impacts of the burgeoning diabetes epidemic and underscores the pivotal role of data-driven analytics and machine learning in the realm of public health research and policy formulation.

Keywords: Machine Learning; Data Mining; Diabetes.

1. Introduction

Diabetes, a complex metabolic disorder, manifests when the intricate system of insulin regulation in the human body falters. This condition arises either due to insufficient insulin production by the pancreas or the inability of the endocrine system to effectively utilize this crucial hormone [1]. The ramifications of diabetes extend beyond mere disruption of glucose homeostasis, encompassing the insidious erosion of vital blood vessels. This vascular deterioration impedes the delivery of blood throughout the body, potentially triggering nonfunctional nerves and a host of severe complications [2].

The economic burden of diabetes on US is staggering. According to the American Diabetes Association, this chronic ailment exerts an annual financial impact of approximately $327 billion on the nation's economy. This colossal cost encompasses healthcare expenditures and the losses incurred due to diabetes-related healthcare utilization, absenteeism, and diminished workplace productivity [3].

However, the implications of diabetes reach far beyond the realm of economics. A disconcerting trend emerges when one delves into the epidemiological landscape of this condition. Recent data derived from the National Health and Nutrition Examination Surveys, administered by Centers for Disease Control and Prevention (CDC), underscores a concerning surge in diabetes rates among adults aged 18 and older in the United States, spanning the two decades from 2001 to 2020. Presently, diabetes affects a staggering 13% of the U.S. population, with an alarming annual rise in new diagnoses. Even more concerning is the revelation that one in four individuals afflicted with diabetes remains undiagnosed [4].

Diabetes, being a multifaceted medical condition influenced by a multitude of factors encompassing biological, physical behaviors, and social aspects, necessitates a comprehensive and nuanced approach to data collection. To address this complexity, health survey domain experts have chosen the Behavioral Risk Factor Surveillance System (BRFSS) as a prominent and reliable source of data.

BRFSS stands out among comparable behavior datasets in the United States due to its extensive coverage and thoroughness in capturing critical health-related information.

Insights from Schauer et al.'s analysis of BRFSS data revealed that obesity and smoking were statistically significant factors positively correlated with diabetes [5]. Tran et al. argued that rural residents exhibited a larger diabetes prevalence probability compared to urban residents [6]. Additionally, study of Tung L et al highlighted that Asian Americans, characterized by lower BMI than other racial population, had a lower probability of diabetes [7]. According to previous study focused on diabetes and using histogram visualizing each variable, multiple significant factors toward diabetes has been analyzed in this study.

Historically, researchers have achieved notable success in predicting diabetes using various machine learning models. For instance, Yoon et al. [8] achieved an accuracy rate exceeding 80% by utilizing classification trees based on the 2013 BRFSS dataset. Lu et al. [9] identified the Random Forest model as the most effective, attaining an impressive accuracy rate of 91% for predicting type 2 diabetes. This model outperformed other machine learning algorithms, including decision trees, based on data from the Australia health fund. Similarly, Xie et al. [10] concluded that a neural network exhibited the highest predictive accuracy at 82% using the 2014 BRFSS data.

Despite these achievements, a pressing question looms: Can these past performance benchmarks withstand the test of time and remain reliable models in ever-evolving world? Recent years have witnessed seismic shifts in societal paradigms, driven by events such as the COVID-19 pandemic. These shifts have had far-reaching effects on people's dietary habits, lifestyles, and fundamental health perspectives. These changes in society have the potential to influence the risk factors associated with diabetes in unpredictable ways, underscoring the importance of reevaluating the applicability of past predictive models in the contemporary context.

In today's dynamic environment, characterized by the ongoing impact of the pandemic and evolving social, dietary, and behavioral norms, it is essential to review whether past best-performing models remain effective. The study at hand recognizes the need for a timely reevaluation of these models and the significance of employing modern data analytics techniques to gain fresh insights into diabetes prediction. By incorporating the most recent data, specifically the 2021 BRFSS dataset, this research aims to evolving understanding of diabetes in context of a rapidly changing world.

To construct the optimal diabetes prediction model, the study selected the best-performing models from previous research. This comprehensive approach aimed to harness the strengths of various machine and deep learning models, each with varying degrees of accuracy in predicting diabetes, while considering the impact of key risk factors elucidated by previous studies.

In essence, the overarching goal of this study is two-fold: to assess the continued relevance of historical predictive models and to harness contemporary data to enhance comprehension of diabetes prediction. This endeavor holds implications for an early identification of high-risk individuals and the optimization of diabetes management strategies in a society that continues to evolve.

2. Method

2.1. Data collection

The process of gathering data for this research was conducted utilizing the 2021 BRFSS, which is an extensive dataset administered by the CDC, the federal public health organization of US. BRFSS as a nationally recognized and authoritative source of health-related information, obtained through telephone surveys. This dataset is designed to capture a wide array of health-related risk behaviors and chronic health conditions among United States citizens aged 18 and older. Comprises responses from over 400,000 participants, who were selected using a uniform state-specific random sampling methodology and using the 2021 BRFSS data, this study aimed to ensure that analysis is based on the most updated and representative information available, allowing model to explore the intricate

relationships and evolving dynamics of diabetes in the context of the contemporary sociocultural and health landscape.

2.2. Data cleaning and Data Mining

Following process all based on R-studio, version 4.3.1. Given the sheer volume of variables available, encompassing a total of 305, and a dataset comprising over 400,000 records, selecting the most effective variables for analysis posed a considerable challenge. In this study, the data cleaning process was conducted following a systematic approach based on the six steps of the data mining process as outlined by Fayyad [11].

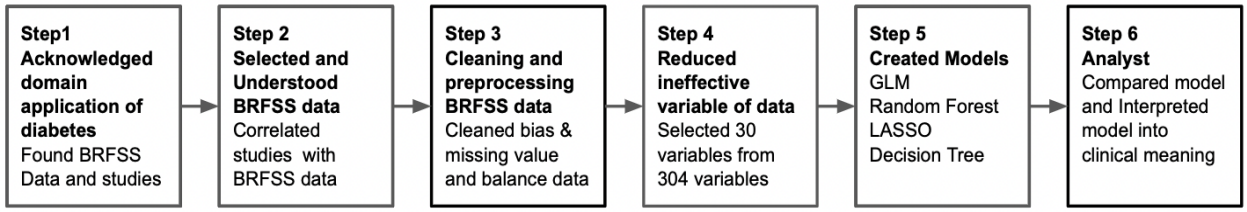

Figure 1. Fayyad's 6 steps data mining approach to constructing machine learning model.

After assessing each variable's relationship with diabetes, this study addressed missing values. These predictive models were utilized to fill missing value gaps in the dataset. However, for data containing variables that could not be reliably predicted using correlated factors has been removed. Additionally, given that diabetes patients constituted only 13% of the dataset, a significant imbalance arose in diabetes prediction. As Mariwan mentioned that imbalanced data would harm model's accuracy [12]. This study employed up-sampling to artificially increase the sampling rate, achieving a balanced representation of both diabetes and non-diabetes cases in model [13].

The dimensionality of the data was streamlined by considering three key principles: the identification of irrelevant variables, the management of highly correlated variables, and the elimination of redundant variables. Additionally, this study incorporated significant factors gleaned from prior research to further refine analysis.

2.3. Model

Once the pre-training data preparation was completed, 80% of the dataset was allocated to construct five distinct models for diabetes prediction: General Linear Model (GLM), Lasso Regression, Neural Network, Decision Tree, and Random Forest.

The General Linear Model is a versatile and widely used statistical framework for modeling the relationship between a diabetes dependent variable and multiple covariate variables. In the context of classification, GLM can be adapted for logistic regression, which is valuable for binary diabetes classification tasks of this study.

Lasso Regression, as a regularization technique used in linear regression, helps prevent overfitting by adding a penalty term to the linear regression equation, since this model contains large number of covariate variables. Lasso encourages sparsity in the model, meaning it can automatically select a subset of the most relevant features and set others to zero, effectively simplifying the model as this model.

Decision Tree used as a supervised learning algorithm used for classification and regression in this study's diabetes prediction tasks. Decision Trees divide the data into subsets according to the values of their features, forming a tree-like visual structure.

Random Forest merges numerous decision trees to enhance predictive accuracy while mitigating overfitting concerns. Random Forest is known for its robustness, versatility, and ability to handle high-dimensional data.

Neural Networks, like the structure of human brain, consist of layers of interconnected nodes or neurons, mimicking the brain's functionality.

2.4. Model performance

Table 1. Confusion Matrix for measuring classification model.

	Actual Diabetes	Actual Non-Diabetes
Predicted Diabetes	True Positive	False Positive
Predicted Non-Diabetes	False Negative	True Negative

Accuracy is to measure that the percentage of predicted output same as actual value.

$$Accuracy = \frac{Ture\ Positive + True\ Negative}{Total\ Number\ of\ Population}$$

Sensitivity, also called True Positive Rate is to identify positive instances from the total actual positive instances in the dataset.

$$Sensitivity = \frac{Ture\ Positive}{Ture\ Positive + False\ Positive}$$

Specificity, also called True Negative Rate

$$Specificity = \frac{Ture\ Negative}{Ture\ Negative + False\ Negative}$$

ROC, or Receiver Operating Characteristic, is a graphical representation and statistical tool used to visualize performance of machine learning models, particularly in scenarios where the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity) is essential.

Area Under Curve (AUC) is measures the area under the ROC curve. The AUC quantifies the classifier's classification ability, and a higher AUC suggests that the model is more effective at making this discrimination.

The remaining 20% of the data was reserved for evaluating and comparing the models in terms of AUC and accuracy, with the objective of identifying the best-performing model for diabetes prediction.

3. Result

3.1. Covariates

After Fayyad's 6 steps data mining approach [11], through an analysis of the CDC-provided codebook and a correlation assessment between variables, the dataset was manually reduced from an initial 304 variables to 31 numerical and categorical variable**s**. Model will be analyzed in 5 big categories that summarized in table 2.

Table 2. Category of Final Model Variables

Category	Variable
Demographic	Gender, Age, Education, Income, Rural, Race, Marital, Own house, Rural, Veteran, Weight KG, Height CM
Diet	Average vegetable consumption per day, Average fruit consumption per day
Habit	Smoking habit, Drinking habit, Exercise habit
Illness	High Blood Pressure, High Cholesterol, Stroke, Asthma, Angina, Heart Attack, Kidney Illness, Depression, C.O.P.D., Depression, Kidney diseases, Physical disability
Medical Care	Number of care provider, Routine checkup, Affordable of medical cost, Routine checkup, Flu shot

3.2. Model Performance

In conclusion, as shown in table 3, the results of final model analysis indicated that the Random Forest model outperforms other models, demonstrating superior accuracy and AUC. Specifically, Random Forest model achieved an impressive 96% accuracy when predicting diabetes using the reserved 20% dataset for evaluation.

Table 3. Performance of Classification Models for Diabetes

Name	Accuracy	AUC
Random Forest	0.9672	0.9672
Decision Tree	0.9054	0.9054
LASSO Regression	0.7184	0.7184
General Linear Regression	0.7148	0.7148
Neutral Network	0.6451	0.6451

This study reinforced the robustness and effectiveness of Random Forest in the context of large and complex datasets, replete with numerous factors and variables. Importantly, these results underscored the model's applicability to contemporary data, highlighting its continued relevance in the ever-evolving landscape of diabetes prediction.

According to these findings, it could be asserted with confidence that Random Forest stood out as a robust and dependable model for diabetes prediction, particularly when confronted with extensive and diverse datasets.

4. Discussion

According to previous results, it was evident that the Random Forest model, which demonstrated the best accuracy and AUC, emerged as the most promising and valuable model for future diabetes detection. This finding aligned with previous research, notably Yoon et al [8] and Lu et al [9], which also identified Random Forest as the best model for diabetes prediction. However, unlike the study based on Xie et al [10], the neural network did not show good performance in this case, due to different dataset resources and other reasons. The past models still applied in today's environment.

This study made significant strides in leveraging the construction of classification models for diabetes using advanced machine learning techniques. These models demonstrated commendable sensitivity and specificity, offering the potential for early realization and intervention in diabetes. Importantly, this study's approach utilized readily accessible variables, such as age and smoking habits, making it a promising avenue for facilitating early detection and intervention efforts.

By harnessing a comprehensive set of 31 variables, the model presented a novel approach to diabetes prediction—one that relied on easily accessible information rather than invasive or resource-intensive physical testing. This pioneering approach held the promise of equipping individuals and healthcare professionals with the necessary tools to proactively manage diabetes, leading to enhancements in public health outcomes and a reduction in the burden imposed by this chronic condition.

However, it was also important to acknowledge its inherent limitations. The foremost limitation stemmed from the cross-sectional nature of BRFSS data, which inherently precluded the establishment of causality. The analysis in this study, therefore, could only reveal associations between variables, but causative relationships could not be inferred.

Another critical limitation pertained to the self-reported nature of the BRFSS data. Self-reported data were susceptible to recall bias, a phenomenon where participants might have inaccurately recalled or reported information, potentially affecting the performance of predictive models. While measures were taken to address this bias through advanced machine learning techniques, it remained a challenge inherent to the dataset.

For future research, there was ample room to delve deeper into the causal factors of diabetes, which could have been instrumental in proactively preventing diabetes from its inception. Furthermore, the model held the potential for more in-depth analyses, facilitating the discovery of correlations between variables and the underlying reasons behind them.

At its core, the primary goal of this research was to contribute to the early detection and intervention of diabetes. By shedding light on the factors that contributed to the development of diabetes and by exploring the intricate web of relationships between variables, this aimed to empower individuals and healthcare professionals with the knowledge and tools needed to tackle this chronic condition effectively from the outset.

5. Conclusion

In conclusion, this study determined that Random Forest outperformed other machine learning models in predicting diabetes based on the 2021 dataset, exhibiting the highest accuracy. Therefore, this research advocates the utilization of the Random Forest model to identify individuals at potential risk of diabetes, facilitating proactive healthcare interventions.

References

[1] World Health Organization 2023 Diabetes. https://www.who.int/news-room/fact- sheets/detail/diabetes.

[2] Diabetes UK 2023 Complications of diabetes https://www.diabetes.org.uk/guide-to-diabetes/complications.

[3] American Diabetes Association 2018 Economic costs of diabetes in the US in 2017 *Diabetes Care* pp 917 - 928.

[4] Centers for Disease Control and Prevention Prevalence of Both Diagnosed and Undiagnosed Diabetes https://www.cdc.gov/diabetes/data/statistics-report/diagnosed-undiagnosed-diabetes.html.

[5] Schauer, Gillian L, Abigail H, Lloyd M and Mark D 2013 Health professional advice for smoking and weight in adults with and without diabetes: findings from BRFSS *Journal of Behavioral Medicine*, vol. 36, no. 1 pp. 10+ *http: //dx.doi.org/10.1007/s10865 - 011 - 9386 - 9.*

[6] Tran P, Tran L and Tran L 2019 Impact of rurality on diabetes screening in the US. *BMC Public Health 19*(1) pp 1190 – 1190. https: //doi.org/10.1186/s12889 - 019 - 7491 - 9.

[7] Tung L, Baig A, Huang S, Laiteerapong N, Chua K 2017 Racial and Ethnic Disparities in Diabetes Screening Between Asian Americans and Other Adults: BRFSS 2012–2014 *J GEN INTERN MED* 32, pp 423 – 429 https://doi-org.ezproxy.bu.edu/10.1007/s11606-016-3913-x.

[8] Yoon S, Taha B and Bakken S 2014 Using a data mining approach to discover behavior correlates of chronic disease: a case study of depression Stud Health Technol Inform 201: 71 - 8.

[9] Lu H, Uddin S, Hajati F, Moni MA and Khushi M 2022 A patient network-based machine learning model for disease prediction: the case of type 2 diabetes mellitus Appl Intell https: //doi.org/10.1007/s10489 - 021 - 02533 - w.

[10] Xie Z, Nikolayeva O, Luo J and Li D 2019 Building risk prediction models for type 2 diabetes using Machine Learning Techniques *Preventing Chronic Disease 16* https: //doi.org/10.5888/pcd16.190109.

[11] Fayyad U, Piatetski-Shapiro G, Smyth P and Uthurusamy R 1998 Advances in Knowledge Discovery and Data Mining *Technometrics*, 40 (1), p. 83. https: //doi.org/10.2307/1271414.

[12] Mariwan S 2023 Diabetes type 2 classification using machine learning algorithms with up-sampling technique *Journal of Electrical Systems and Information Technology*, vol. 10 no. 1 http://dx.doi.org/10.1186/s43067-023-00074 - 5.

[13] Brownlee J 2020 Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost- Sensitive Learning. Machine Learning Mastery.