Exploring the Relationship Between Diabetes and Risk Factors

Abstract

Aims/ objectives: This study investigates the relationship between diabetes and its risk factors, aiming to identify key predictors and assess their impact on disease progression.The influence of variables such as age, smoking status, cholesterol levels, blood pressure, BMI, and glucose levels on diabetes outcomes was analyzed.The study utilizes a dataset of 4,240 observations from the Kaggle data repository. A logistic regression modeling technique was employed in this study as the methodology for analyses. The findings suggest that glucose levels and age are key predictors of diabetes, implying that individuals with higher glucose levels or advancing age are at a greater risk. Other factors may also play a role but show varying levels of correlation. Our findings emphasize the importance of regular monitoring and early intervention, especially for older populations.

Keywords: Logistic Regression; Diabetes; Models; Risk Factors; Predictors; Confusion Matrix

2010 Mathematics Subject Classification: 53C25; 83C05; 57N16

1 Introduction

Over 500 million cases of diabetes, primarily among adults, are reported globally each year [Hossain et. al., (2004)]. Diabetes is a chronic illness characterized by the body's inability to regulate glucose levels effectively. The condition is associated with several complications, including cardiovascular diseases, kidney failure, neuropathy, and vision loss [Deshpande et. al., (2008)], which significantly impact individuals' productivity and quality of life. Beyond the health challenges, diabetes imposes a substantial financial burden on individuals and families due to medical expenses. Numerous studies in the literature have explored the relationship between diabetes and its risk factors using methods such as logistic and multiple regression [Wu et. al., (2021); Agarwal et. al., (2020); Jeerasuwannakul et. al., (2021); McGurnaghan et. al., (2021); Wakasugi et. al., (2021); Ghosal et. al., (2021)], retrospective analysis [He et. al., (2021); Ata et. al., (2023); Koska et. al., (2022); Tremblay et. al., (2021);

Nianogo and Arah (2022)], and supervised machine learning models [Ooka et. al., (2021); Rastogi and Bansal (2023); Wang et. al., (2021)]. However, a key gap remains. There is a lack of integrated studies comparing the performance of these methods in terms of predictive accuracy, interpretability, and generalizability—factors crucial for selecting the most effective approach in diverse healthcare settings. Furthermore, much of the existing research overlooks the integration of diverse and potentially significant risk factors, such as genetic predisposition and socio-economic status, which could provide a more comprehensive understanding of diabetes risk. Additionally, most studies rely on cross-sectional data that do not capture temporal changes, highlighting the need for longitudinal studies to examine the progression of diabetes risk over time. In this study, we aim to address these gaps by providing valuable insights into diabetes prediction based on various risk factors.

The remainder of this paper is organized as follows: Section 2 presents related studies on diabetes, while Section 3 provides a detailed description of the methodologies employed. The results and discussion of the findings are presented in Section 4, followed by the conclusion and potential areas for future research in Section 5.

2 Related Works

Several studies have been conducted to investigate diabetes and its risk factors, aiming to enhance prevention, control, and intervention strategies. This study reviews multiple articles on diabetes and its risk factors to evaluate the current state of research and contribute to the field. To begin, [Wu et. al., (2021)] explored risk factors for neuropathy in type 2 diabetes mellitus (T2DM) using Lasso and Logistic regressions to construct predictive models. Four predictive models—A, B, C, and D—were developed, with findings indicating that models C and D were most effective in identifying influential factors for early screening of diabetic peripheral neuropathy (DPN) in T2DM patients. In another study, [Ooka et. al., (2021)] employed a Random Forest modeling approach to predict risk factors for type 2 diabetes using a large-scale health checkup dataset from Japan. The analysis compared the Random Forest model with a Multivariate Logistic Regression model as a benchmark. Results demonstrated that the Random Forest model exhibited superior predictive power for changes in HbA1c levels compared to Multivariate Logistic Regression. The study concluded that the appropriate application of the Random Forest method can enable highly accurate risk prediction for HbA1c changes and potentially identify novel risk predictors for diabetes. The prediction of type 2 diabetes (T2D) risk was studied by [He et. al., (2021)], who compared polygenic risk scores (PGS), poly-exposure scores (PXS), and clinical risk scores (CRS). The study included a population of 7,513 individuals with incident T2D. The C-statistics for the predictive models were 0.709 for PGS, 0.762 for PXS, and 0.839 for CRS. The findings concluded that while PXS provides a modest incremental predictive value over established clinical risk factors in T2D risk prediction, the concept of PXS merits further exploration. It is likely to have potential utility in other chronic disease risk stratification models as well. In another study by [Rastogi and Bansal (2023)], four data mining techniques—Random Forest, Support Vector Machine (SVM), Logistic Regression, and Naive Bayes—were employed to predict diabetes. Performance evaluations based on the confusion matrix, sensitivity, and accuracy metrics revealed that Logistic Regression was the best-performing model, achieving an accuracy of 82.3%, outperforming the other three models. The results suggest that incorporating a wider range of classification algorithms could enhance the accuracy of diabetes prediction. Similarly, supervised learning classifiers were explored by [Wang et. al., (2021)] to study diabetes mellitus and its complex related factors. The results demonstrated that the Random Forest classifier, combined with SVM-SMOTE resampling technology and the LASSO feature screening method, achieved the best performance in identifying individuals at high risk of diabetes mellitus (DM). The model's metrics were as follows: Accuracy = 0.890, Precision = 0.869, Recall = 0.919, F1-Score = 0.893, and AUC = 0.948. A retrospective study on clinical and biochemical determinants of extended hospital stays was conducted by [Ata et. al., (2023)] to investigate diabetic ketoacidosis (DKA) admissions across four hospitals in Qatar from 2015 to 2021. The study included 922 patients with a median age of 35 years (IQR: 25–45), of whom 62% were male. Among the patients, 52% had type 1 diabetes mellitus (T1DM), and 48% had type 2 diabetes mellitus (T2DM). The findings from this study, derived from a diverse population, offer valuable insights for physicians and healthcare systems aiming to reduce the diabetes-related healthcare burden in DKA patients. Pre-admission diabetes and COVID-19-specific risk factors for mortality were examined by [Agarwal et. al., (2020)] using multivariate modeling to assess the independent association of HbA1c levels in a hospital setting. The study analyzed 1,126 hospitalized patients with diabetes and COVID-19, with a mean age of 68 years. Of these patients, 50% were male, 75% were Black, 98% had type 2 diabetes, and the mean BMI was 30 kg/m^2. The mean HbA1c level was 7.5%, and 33.1% of the patients died. Multivariate logistic regression was applied to analyze eight factors associated with proteinuria in patients with type 2 diabetes mellitus (T2DM) [Jeerasuwannakul et. al., (2021)]. The results revealed that, among these factors, only fasting plasma glucose (FPG) was significantly correlated with proteinuria, with an adjusted odds ratio of 1.009 (95% CI: 1.004–1.0156). An FPG level of 136 mg/dL demonstrated a sensitivity of 80.43%, suggesting a potential link between FPG levels and the presence of proteinuria. Thus, an FPG cutoff of 136 mg/dL showed good sensitivity as a predictor of proteinuria in patients with T2DM. A study by [Koska et. al., (2022)] utilized a composite AGE (Advanced Glycation End Products) score to investigate its potential in predicting loss of renal function and high-risk chronic kidney disease (hrCKD) in patients with type 2 diabetes. The results, adjusted for baseline and follow-up HbA1c levels and other risk factors in the ACCORD study, demonstrated significant associations between the AGE score and various renal outcomes. These included a reduction in eGFR (b-estimate: 20.66 mL/min/1.73 m^2 per year; P = 0.001), 30% RFL (hazard ratio: 1.42 [95% CI: 1.13–1.78]; P = 0.003), 40% RFL (HR: 1.40 [95% CI: 1.13–1.74]; P = 0.003), macroalbuminuria (HR: 1.53 [95% CI: 1.13–2.06]; P = 0.006), and hrCKD (HR: 1.88 [95% CI: 1.37–2.57]; $P < 0.0001$). These findings provide robust evidence supporting a causal role of AGEs in diabetic nephropathy, independent of glycemic control. Moreover, they suggest that the composite AGE panel could be a valuable tool for predicting long-term renal function decline in type-2 diabetes patients. A cardiovascular disease (CVD) risk predictive model for type-1 diabetes, based on a Poisson regression approach, was studied by [McGurnaghan et. al., (2021)] using data from diabetes patient registers in Sweden. The results showed that the age-standardized rate of CVD per 100,000 person-years was 4,070 for men and 3,429 for women with type 1 diabetes in Scotland, and 4,014 for men and 3,956 for women in Sweden. The prediction tool developed through this study has the potential to provide individualized risk predictions, offering valuable insights for managing CVD risk in type-1 diabetes patients. The development of a multi-polygenic risk score (multi-PRS) for diabetes complications and response to intensive blood pressure and glucose control was studied by [Tremblay et. al., (2021)]. The study combined ten weighted polygenic risk scores (10 wPRS), consisting of 598 single nucleotide polymorphisms (SNPs) associated with key risk factors and outcomes of type 2 diabetes. These scores were derived from genome-wide association study (GWAS) summary statistics and analyzed using a logistic regression model. The results demonstrated comparable predictive performance for cardiovascular and renal complications across different cohorts. Furthermore, the multi-PRS model effectively stratified individuals with type 2 diabetes based on their risk of complications, offering a valuable tool for personalized risk assessment and management. An agent-based model, calibrated on the Los Angeles ViLA cohort population, was used to predict the incidence and burden of obesity and type 2 diabetes mellitus (T2DM) in Los Angeles County [Nianogo and Arah (2022)]. The results revealed that the age-specific incidence of obesity generally increased from 10% to 30% across the lifespan, with two notable peaks occurring at ages 6–12 and 30–39 years. Similarly, the incidence of T2DM increased from less than 2% at ages 18–24 to a peak of 25% at ages 40–49. The ViLA Obesity model offers valuable insights into the future burden of obesity and T2DM in Los Angeles County, one of the most diverse regions in the United States, highlighting the need for targeted interventions to address these health challenges. A cross-sectional study by [Wu et. al., (2021)] utilized chi-square and covariance analysis to examine the association between physical activity levels and cardiovascular risk factors in adolescents living with type 1 diabetes mellitus (T1DM). The results revealed that regular physical activity is associated with a beneficial cardiovascular profile in T1DM, including notable improvements in lipid profiles. A multivariate regression model was employed to examine the association between continuous glucose monitoring (CGM)-derived metrics and arterial stiffness among patients with type 2 diabetes in Japan [Wakasugi et. al., (2021)]. The results indicated that all CGM-derived metrics were significantly associated with brachial-ankle pulse wave velocity (baPWV), a marker of arterial stiffness, whereas HbA1c showed no significant association. These findings suggest that CGM-derived metrics could be valuable in identifying patients at high risk of developing cardiovascular disease. Similarly, a simulation study conducted by [Ghosal et. al., (2021)] utilized a multivariate regression methodology to investigate the effects of nationwide lockdowns during the COVID-19 pandemic on the worsening of glycosylated hemoglobin (HbA1c) levels and an increase in diabetes-related complications. The results predicted an increase in HbA1c from baseline, with projections of 2.26% at the end of a 30-day lockdown and 3.68% at the end of a 45-day lockdown. The materials and methods used in this study are clearly presented and discussed in the following section.

3 Materials and Methods

This section outlines the materials and methods used in this study, including the data source, analysis techniques, and computational tools employed.

3.1 Kaggle Repository

A publicly available dataset from the Kaggle repository [Kaggle (2025)] will be used in this study. Kaggle is an online platform that offers a wide range of datasets for research, particularly for data science and machine learning projects. The selected dataset focuses on diabetes and its associated risk factors, including age, BMI, smoking status, systolic blood pressure, and physical activity levels.

3.2 Logistic Regression Model

A statistical technique called logistic regression [Hosmer et. al., (2013)] will be used to examine the simultaneous effects of multiple independent variables—such as age, BMI, smoking status, systolic blood pressure, and physical activity levels—on the dependent variable (diabetes). Logistic regression is typically employed when the dependent variable is dichotomous or binary in nature, as in the case of diabetes status (e.g., diabetic vs. non-diabetic). The general form of the multiple logistic regression model is given by:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{3.1}$$

where P(Y=1) is the probability of the event of interest (e.g., diabetes), β_0 is the intercept, β_1, $\beta_2 \cdots \beta_k$ are the coefficients of the predictor variables (X_1, X_2, $\cdots$, X_k), which represent the change in the log-odds of the dependent variable for a one-unit change in the corresponding independent variable.

3.2.1 Multicollinearity test

Multicollinearity is a phenomenon that occurs when two or more independent or predictor variables in a regression model are highly correlated [Morrow-Howell (1994)]. The presence of multicollinearity can lead to unreliable estimates of the regression coefficients, making it difficult to isolate the individual effect of each predictor on the dependent variable. This issue can inflate the standard errors of the coefficients, which may reduce the statistical significance of the predictors. Several methods exist for detecting multicollinearity, and in this study, we will use the correlation matrix [Mansfield and Helms (1982)] and the variance inflation factor (VIF) [Oke (2019)] to assess its presence. The VIF calculated as:

$$\text{VIF} = \frac{1}{1 - \text{R}^2} \tag{3.2}$$

3.3 Accuracy Measure

The accuracy of the logistic regression model in this study is evaluated using the confusion matrix in Table 2 and computed by Equation 3.3.

$$\text{Accuracy} = \frac{\text{TP - TN}}{\text{TP + TN + FP + FN}} \tag{3.3}$$

Here, the symbols TP, TN, FP, and FN in Equation 3.3 denote true positive, true negative, false positive, and false negative, respectively.

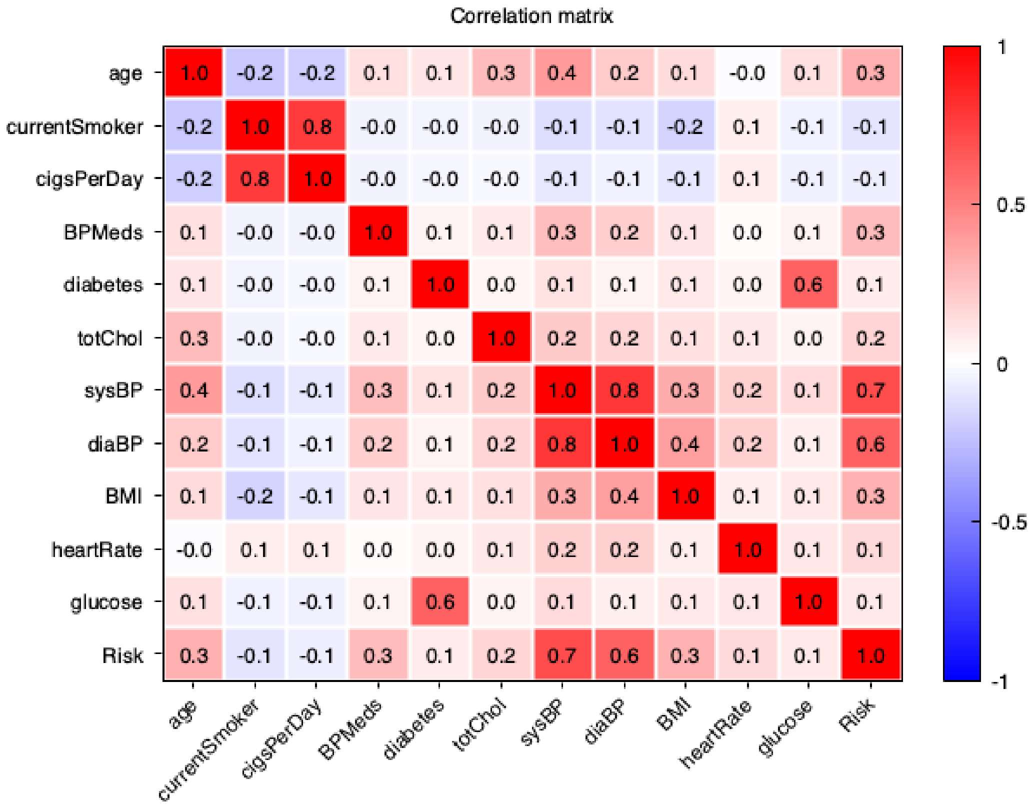
3.4 Gretl Software

Gretl (GNU Regression, Econometrics, and Time-series Library) is the tool proposed for data analysis in this study. Gretl is an open-source software package [Lemenkova (2019)] designed for econometric analysis, providing a user-friendly interface and powerful capabilities for statistical modeling, particularly in regression and time-series analysis. The following section presents the study results and discusses the findings.

4 Results Presentation and Discussion

The dataset was preprocessed, and summary statistics revealed that 3,751 valid observations were used for the analysis after filtering out missing values. The average age of participants was 49.58 years, with an average glucose level of 81.96 mg/dL. Additionally, approximately 49% of participants were current smokers, and 3% were on blood pressure medication. The prevalence of diabetes in the dataset was 2.6%.

A correlation plot, presented in Figure 1, illustrates the degree of association among the independent variables in the dataset. The correlation index reveals strong positive relationships between certain variables, such as *currentSmoker* and *cigsPerDay*, *sysBP* and *diaBP*, and *sysBP* and Risk. However, correlations among other independent variables were relatively weak. The association between diabetes and glucose levels (r = 0.6) was moderate, as was the correlation between diabetes and age (r = 0.3).



	age	currentSmoker	cigsPerDay	BPMeds	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	Risk
age	1.0	-0.2	-0.2	0.1	0.1	0.3	0.4	0.2	0.1	-0.0	0.1	0.3
currentSmoker	-0.2	1.0	0.8	-0.0	-0.0	-0.0	-0.1	-0.1	-0.2	0.1	-0.1	-0.1
cigsPerDay	-0.2	0.8	1.0	-0.0	-0.0	-0.0	-0.1	-0.1	-0.1	0.1	-0.1	-0.1
BPMeds	0.1	-0.0	-0.0	1.0	0.1	0.1	0.3	0.2	0.1	0.0	0.1	0.3
diabetes	0.1	-0.0	-0.0	0.1	1.0	0.0	0.1	0.1	0.1	0.0	0.6	0.1
totChol	0.3	-0.0	-0.0	0.1	0.0	1.0	0.2	0.2	0.1	0.1	0.0	0.2
sysBP	0.4	-0.1	-0.1	0.3	0.1	0.2	1.0	0.8	0.3	0.2	0.1	0.7
diaBP	0.2	-0.1	-0.1	0.2	0.1	0.2	0.8	1.0	0.4	0.2	0.1	0.6
BMI	0.1	-0.2	-0.1	0.1	0.1	0.1	0.3	0.4	1.0	0.1	0.1	0.3
heartRate	-0.0	0.1	0.1	0.0	0.0	0.1	0.2	0.2	0.1	1.0	0.1	0.1
glucose	0.1	-0.1	-0.1	0.1	0.6	0.0	0.1	0.1	0.1	0.1	1.0	0.1
Risk	0.3	-0.1	-0.1	0.3	0.1	0.2	0.7	0.6	0.3	0.1	0.1	1.0

Figure 1: Correlation plot

Furthermore, a multicollinearity test was performed using the Variance Inflation Factor (VIF) and diagnostic checks, with the results presented in Tables 3 and 1, respectively. The VIF values indicate that all variables are moderately correlated, as all values are less than 5. Results in Table 1 show that a BKW condition number ≥ 30 indicates ”strong” near-linear dependence, while a condition number between 10 and 30 suggests ”moderately strong” dependence. Parameter estimates whose variance is largely associated with problematic condition values may themselves be considered problematic.

Table 1: Belsley-Kuh-Welsch Collinearity Diagnostics Check

λ	**Condition**	**Constant**	**Age**	**currentSmoker**	**BPMeds**	**totChol**	**sysBP**	**diaBP**	**heartRate**	**glucose**
8.523	1.000	0.000	0.000	0.001	0.001	0.000	0.000	0.000	0.000	0.000
1.213	2.651	0.000	0.000	0.067	0.036	0.000	0.000	0.000	0.000	0.000
0.932	3.024	0.000	0.000	0.005	0.771	0.000	0.000	0.000	0.000	0.000
0.168	7.121	0.000	0.001	0.761	0.002	0.000	0.000	0.001	0.001	0.001
0.046	13.572	0.002	0.003	0.008	0.029	0.466	0.018	0.003	0.114	0.003
0.039	14.800	0.000	0.091	0.084	0.043	0.023	0.004	0.005	0.021	0.441
0.030	16.865	0.004	0.557	0.000	0.000	0.071	0.010	0.040	0.004	0.045
0.020	20.520	0.034	0.042	0.024	0.069	0.045	0.105	0.062	0.041	0.004
0.017	22.331	0.000	0.005	0.000	0.000	0.202	0.011	0.021	0.675	0.362
0.006	36.352	0.536	0.052	0.030	0.016	0.191	0.139	0.303	0.029	0.063
0.004	43.745	0.424	0.247	0.020	0.033	0.000	0.713	0.565	0.114	0.080

The performance of the logistic regression model in predicting diabetes status was assessed using the confusion matrix presented in Table 2. The model achieved an accuracy of 98.6%, calculated using Equation 3.3, indicating its effectiveness in classifying individuals with and without diabetes. Furthermore, the model's confusion matrix demonstrated a prediction accuracy of 98.6%, correctly classifying 3,699 out of 3,751 cases (see Table 2). These results confirm that glucose levels and age are the strongest predictors of diabetes among the variables analyzed. Moreover, these findings align with existing literature, which highlights the significant role of glucose levels and age-related physiological changes in the development of diabetes.

Table 2: Confusion Matrix

	Predcited Positive [== 1]	Predicted Negative [== 0]
Predicted Positive [== 1]	3685	2
Predicted Negative [== 0]	56	47

Table 3: Results of the analysis

Variable	**Coefficient**	**Std. Error**	**Z-Statistic**	**p-value**	**VIF**	95% **CI**
Constant	-7.774	0.701	-11.090	0.000	2.012	-9.284, -6.516
Age	-0.015	0.007	-2.262	0.024	1.323	-0.051, -0.042
currentSmoker	-0.358	0.116	-3.085	0.002	2.573	-0.432, -0.267
BPMeds	0.698	0.270	2.587	0.010	1.083	0.562, 0.867
totChol	-0.003	0.001	-2.061	0.039	1.103	-0.015, -0.012
sysBP	0.010	0.004	2.321	0.020	3.192	0.002, 0.032
diaBP	-0.041	0.005	-3.997	0.000	2.859	-0.063, -0.028
heartRate	-0.018	0.005	-3.997	0.000	1.065	-0.033, -0.004
glucose	0.032	0.002	13.200	0.000	1.040	0.021, 0.053

The R^2 and adjusted R^2 values, 54.49% and 52.15%, respectively, indicate a relatively strong goodness of fit. The logistic regression model demonstrates reasonable accuracy in classifying diabetes cases, addressing misclassification rates and improving model calibration could enhance its predictive reliability in clinical settings. The results suggest that risk factors such as age, BMI, blood pressure, and glucose levels play a significant role in predicting diabetes. The model's performance may be further improved by incorporating additional clinical and lifestyle variables, optimizing feature selection, or employing advanced machine learning techniques.

5 Conclusion

This study employed a logistic regression model to analyze the relationship between diabetes and various risk factors, including age, BMI, smoking status, systolic blood pressure, and physical activity levels. The key findings of this study are summarized as follows:

a. Among the predictors of diabetes, glucose levels and age were the most significant, showing strong associations with the likelihood of developing the condition.

b. The logistic regression model provided a good fit for the data, as indicated by the performance accuracy derived from the confusion matrix.

c. These findings highlight the importance of addressing modifiable risk factors, such as glucose levels and physical activity, in public health interventions aimed at reducing the prevalence of diabetes.

d. Future studies could focus on comparative analyses using advanced machine learning techniques to enhance understanding and develop more effective predictive models, ultimately improving accuracy and overall model performance.

Declaration

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

Hossain MJ, Al-Mamun M, Islam MR. Diabetes mellitus, the fastest growing global public health concern: Early detection should be focused. *Health Science Reports*. Vol. 7, Issue 3(Mar. 2024).

Deshpande AD, Harris-Hayes M, Schootman M. Epidemiology of diabetes and diabetes-related complications. *Physical therapy*. Vol. 88, Issue 11(Nov. 2008); pp. 1254-64.

Wu B, Niu Z, Hu F. Study on risk factors of peripheral neuropathy in type 2 diabetes mellitus and establishment of prediction model. it Diabetes & metabolism journal. Vol. 45, Issue 4(Jul. 2021); pp. 526-38.

Ooka T, Johno H, Nakamoto K, Yoda Y, Yokomichi H, Yamagata Z. Random forest approach for determining risk prediction and predictive factors of type 2 diabetes: large-scale health check-up data in Japan. *BMJ Nutrition, Prevention & Health*. Vol. 4, Issue 1(2021); pp. 140.

He Y, Lakhani CM, Rasooly D, Manrai AK, Tzoulaki I, Patel CJ. Comparisons of polyexposure, polygenic, and clinical risk scores in risk prediction of type 2 diabetes. *Diabetes Care*. Vol. 44, Issue 4(Apr. 2021); pp. 935-43.

Rastogi R, Bansal M. Diabetes prediction model using data mining techniques. Measurement: *Sensors*. Vol. 25, (Feb. 2023):100605.

Wang X, Zhai M, Ren Z, Ren H, Li M, Quan D, Chen L, Qiu L. Exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier. *BMC medical informatics and decision making*. Vol. 21, (Dec. 2021); pp. 1-4.

Ata F, Khan AA, Khamees I, Iqbal P, Yousaf Z, Mohammed BZ, Aboshdid R, Marzouk SK, Barjas H, Khalid M, El Madhoun I. Clinical and biochemical determinants of length of stay, readmission and recurrence in patients admitted with diabetic ketoacidosis. *Annals of medicine*. Vol. 55, Issue 1(Dec. 2023); pp. 533-42.

Agarwal S, Schechter C, Southern W, Crandall JP, Tomer Y. Preadmission diabetes-specific risk factors for mortality in hospitalized patients with diabetes and coronavirus disease 2019. *Diabetes Care*. Vol. 43, Issue 10(Oct. 2020); pp. 2339-44.

Jeerasuwannakul B, Sawunyavisuth B, Khamsai S, Sawanyawisuth K. Prevalence and risk factors of proteinuria in patients with type 2 diabetes mellitus. *Asia Pac J Sci Technol*. Vol. 26, Issue 4(Jan. 2021); pp. 1-5.

Koska J, Gerstein HC, Beisswenger PJ, Reaven PD. Advanced glycation end products predict loss of renal function and high-risk chronic kidney disease in type 2 diabetes. *Diabetes Care*. Vol. 45, Issue 3(Mar. 2022); pp. 684-91.

McGurnaghan SJ, McKeigue PM, Read SH, Franzen S, Svensson AM, Colombo M, Livingstone S, Farran B, Caparrotta TM, Blackbourn LA, Mellor J. Development and validation of a cardiovascular risk prediction model in type-1 diabetes. *Diabetologia*. Vol. 64, Issue 9,(Sep. 2021); pp. 2001-11.

Tremblay J, Haloui M, Attaoua R, Tahir R, Hishmih C, Harvey F, Marois-Blanchet FC, Long C, Simon P, Santucci L, Hizel C. Polygenic risk scores predict diabetes complications and their response to intensive blood pressure and glucose control. *Diabetologia*. Vol. 64, (Sep. 2021); pp. 2012-25.

Nianogo RA, Arah OA. Forecasting obesity and type 2 diabetes incidence and burden: the ViLA-obesity simulation model. *Frontiers in Public Health*. Vol. 10, (Apr. 2022): 818816.

Wu N, Bredin SS, Jamnik VK, Koehle MS, Guan Y, Shellington EM, Li Y, Li J, Warburton DE. Association between physical activity level and cardiovascular risk factors in adolescents living with type 1 diabetes mellitus: a cross-sectional study. *Cardiovascular Diabetology*. Vol. 20(Dec. 2021); pp. 1-1.

Wakasugi S, Mita T, Katakami N, Okada Y, Yoshii H, Osonoi T, Kuribayashi N, Taneda Y, Kojima Y, Gosho M, Shimomura I. Associations between continuous glucose monitoring-derived metrics and arterial stiffness in Japanese patients with type 2 diabetes. *Cardiovascular diabetology*. Vol. 20 (Dec. 2021); pp. 1-2.

Ghosal S, Sinha B, Majumder M, Misra A. Estimation of effects of nationwide lockdown for containing coronavirus infection on worsening of glycosylated haemoglobin and increase in diabetes-related complications: a simulation model using multivariate regression analysis. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*. Vol. 14, Issue 4(Jul. 2020); pp. 319-23.

Kaggle Repository. Open Access via the link: *https://www.kaggle.com/datasets/mathchi/diabetes-data-set*. Acessed online (Jan. 2025)

Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. *John Wiley & Sons*; (Feb. 2013).

Morrow-Howell N. The M word: Multicollinearity in multiple regression. *Social work research*. Vol. 18, Issue 4(Dec. 1994).

Mansfield ER, Helms BP. Detecting multicollinearity. *The American Statistician*. Vol. 36, Issue 3a(Aug. 1982); pp.158-60.

Oke J, Akinkunmi WB, Etebefia SO. Use of correlation, tolerance and variance inflation factor for multicollinearity test. *GSJ*. Vol. 7, Issue 5(May 2019); pp. 652-9.

Lemenkova P. Regression models by Gretl and R statistical packages for data analysis in marine geology. *International Journal of Environmental Trends (IJENT)*. Vol. 3, Issue 1(Jun. 2019); pp. 39-59.