Research Article

Comprehensive Exploration of Influential Factors to Diabetes Status

Ava Ng, Junior

Diamond Bar High School, 21400 Pathfinder Rd, Diamond Bar, CA 91765

ABSTRACT

The present study addresses the escalating public health concern of diabetes, which affects approximately 34.2 million Americans. Recognizing the multifaceted etiology of diabetes, encompassing genetic, environmental, and lifestyle factors, the research aims to identify influential determinants of this chronic condition. A comprehensive dataset of 100,000 entries from Kaggle is utilized, thereby embracing the advantages of observational data analysis. The study employs a hybrid methodological framework, integrating machine learning and traditional statistical techniques. Specifically, the Random Forests for feature importance analysis and binary logistic regression are leveraged to understand the relationships between variables and diabetes risk. This dual approach allows us to harness the predictive power and variable ranking capability of machine learning while maintaining the interpretability and statistical rigidity of logistic regression. The analysis encompasses a broad range of factors, including biological aspects like BMI and blood sugar levels, and other socio-economic determinants. By combining diverse methodologies, this study intends to provide a more detailed understanding of diabetes risk factors, facilitating the development of targeted prevention strategies and informing policy decisions. The findings hold the potential to significantly impact health education, innovative healthcare solutions, and policy development, addressing a critical need in diabetes management and prevention.

Keywords: Diabetes; Random Forest; Binary Logit Regression; Socio-economic Factors; Medical Conditions.

INTRODUCTION

Diabetes is a chronic medical condition characterized by an inability of the body to properly process and use glucose, a type of sugar that is a primary source of energy [1]. This results in elevated levels of glucose in the blood, which can lead to a variety of health complications over time. Diabetes is caused by a combination of genetic, environmental, and lifestyle factors, generally dependent on the different types of Diabetes, such as Type-1, Type-2, and others [2]. Diabetes was a major public health concern in the United States, affecting millions of people. For instance, approximately 34.2 million Americans have diabetes, accounting for about 10.5% of the U.S. population. Of these, around 7.3 million were undiagnosed [3]. The consequences of diabetes in the U.S. are profound, impacting individuals' health and quality of life, imposing substantial economic costs, and putting a significant strain on the healthcare system [4].

Reducing diabetes requires a comprehensive and integrated approach, involving individuals, communities, healthcare providers, and policymakers. The wide

Corresponding author: Ava Ng, Junior, E-mail: avang880793@gmail.com.

Received January 10, 2023**; Accepted** February 2, 2023
https://doi.org/10.70251/HYJR2348.21110

variety of strategies to prevent diabetes include, but no limited to, promoting healthy eating, increasing physical activity, increasing awareness and education, promoting innovative healthcare solutions, and advocating for policy change [5]. Among them, exploration of influential factors of diabetes plays an important role in the strategies related with education, innovation solutions, and policy development [6]. General methods of the identification of the impactful factors for diabetes include clinical trials, experimental research in the lab, and data analysis based on survey or other collected data. While experimental and clinical research is invaluable for establishing causality and understanding mechanisms, data analysis of existing datasets offers a practical, cost-effective, and comprehensive approach to identifying factors related to diabetes. These methods leverage the richness of available data to provide insights that are grounded in real-world conditions, cover diverse populations, and span extensive periods, facilitating a nuanced understanding of diabetes risk factors.

Due to the significant advantages of research based on observational data, a plethora of studies have embraced this methodology. Many of these investigations leverage statistical analysis and modeling techniques such as regression and correlation analysis, survival analysis, and propensity score matching [7, 8, 9]. Additionally, a substantial portion of research has utilized machine learning and predictive modeling approaches, including decision trees, support vector machines, neural networks, deep learning, and feature importance analysis [10-14] Moreover, various studies have focused on a diverse array of factors, encompassing lifestyle elements like smoking, diet, and physical activity, biological aspects such as age, gender, BMI, and blood sugar levels, as well as environmental and social determinants including socio-economic status and access to healthcare, among others [15-17].

While the aforementioned multitude of methods has provided invaluable insights to mitigate the risk of diabetes, there is still room for enhancement in these approaches due to challenges like limited data volume and a narrow spectrum of methodologies. To bridge this research gap, the current study endeavors to investigate the causal factors of diabetes, employing an extensive dataset comprising 100,000 entries sourced from Kaggle. Furthermore, I leverage a hybrid approach, utilizing both machine learning techniques, specifically random forests, and statistical methods, including binary logistic regression and more, to capitalize on their distinct advantages and mitigate their inherent limitations.

DATA DESCRIPTION

Data Source

The Comprehensive Dataset for Predicting Diabetes with Medical & Demographic Data consists of 100,000 observations. The dataset is a collection of medical and demographic data from patients, along with their diabetes status (positive or negative), serving as the target. The other features contained include age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. The excessive sample size perfectly satisfies the main objective of the paper to explore the relationships between various medical and demographic factors and the likelihood of developing diabetes. It is also anticipated that the results could be useful for healthcare professionals in identifying patients who may be at risk of developing diabetes and in developing personalized treatment plans. The data is publicly available online from Kaggle for the facilitation of pertinent research [18]. The data credibility can be indicated by the extensive views (206k) and downloads (41k) at the time of paper-writing. It is worth mentioning that while Kaggle datasets offer convenience and a wealth of resources, they may lack the specificity and control that come with collecting my own data. However, the large sample size associated with the dataset used perfectly satisfies one of the main objectives of the paper. Detailed information of the data used is exhibited in Table 1.

METHODOLOGY

It is almost infeasible to clearly explore the various perspectives of the data with a large sample size using a single technique or tool. Therefore, the present research uses a set of tools including boxplot, chi-square test, correlation analysis, random forest, and logit regression model. The details of each method are described in turn as follows.

Comparative Boxplots

Comparative boxplots, also known as side-by-side boxplots, are a statistical tool used to visually compare distributions of data across different groups or categories [19]. They are particularly useful for highlighting differences in medians, the spread of data, and the presence of outliers among different groups. In the paper, such boxplots are used to compare the distribution of the numerical features (age, bmi, HbA1c_level, blood_glucose_level) in the groups with the presence and absence of diabetes.

Table 1. The List of Variables and The Associated Definition and Descriptive Statistics

Variables	Type	Definition	Descriptive Statistics
diabetes	Categorical	It is the target variable being predicted, with values of 1 indicating the presence of diabetes and 0 indicating the absence of diabetes.	0: 91,500; 1: 8,500.

gender	Categorical	It refers to the biological sex of the individual, which can have an impact on their susceptibility to diabetes. There are three categories in it male, female and other.	Female: 58,552;
			Male: 41,431;
			Other: 18.

heart_disease	Categorical	It is another medical condition that is associated with an increased risk of developing diabetes. It has values a 0 or 1, where 0 indicates they don't have heart disease and for 1 it means they have heart disease.	0: 96,058; 1: 3,942.

smoking_history	Categorical	It is also considered a risk factor for diabetes and can exacerbate the complications associated with diabetes. There are 6 categories: not current, former, No Info, current, never, and ever.	No Info: 35816;
			Never: 35095;
			Ever: 4004;
			Former: 9352;
			Not current: 6447.
			Current: 9286.

hypertension	Categorical	It is a medical condition in which the blood pressure in the arteries is persistently elevated. It has values a 0 or 1 where 0 indicates they don't have hypertension and for 1 it means they have hypertension.	0:92,515; 1:7,485.

age	Numerical	It indicates the age of the subjects.	Mean: 41.9;
			SD: 22.5;
			Min: 0.08;
			Max: 80.

blood_glucose_level	Numerical	It refers to the amount of glucose in the bloodstream at a given time.	Mean: 27.3;
			SD: 6.64;
			Min: 10;
			Max: 95.7.

bmi	Numerical	It is a measure of body fat based on weight and height. Higher BMI values are linked to a higher risk of diabetes. BMI less than 18.5 is underweight, 18.5-24.9 is normal, 25-29.9 is overweight, and 30 or more is obese.	Mean: 138;
			SD: 40.7;
			Min: 80;
			Max: 300.

HbA1c_level	Numerical	The Hemoglobin A1c level is a measure of a person's average blood sugar level over the past 2-3 months.	Mean: 5.53;
			SD: 1.07;
			Min: 3.5;
			Max: 9.

Chi-Square Test

The Chi-square test is a statistical test used to determine whether there is a significant association between two categorical variables [20]. It is usually performed based on the contingency table, where the frequency count for each combination of categories is stored. For each cell in the contingency table, the expected frequencies (E) can be calculated using the following expression:

$$E = \frac{Column\ total \times Row\ total}{Grand\ total} \tag{1}$$

Based on E, the Chi-square statistics (χ^2) is then computed as follows:

$$X^2 = \sum \frac{(E - O)^2}{O} \tag{2}$$

Where O is the observed frequency for each cell. With the specific degrees of freedom and usually selected significance level of 0.05, the obtained Chi-square statistics can then be compared with the critical value to determine whether each of the categorical features (gender, heart_disease, smoking_history, and hypertension) is statistically significantly related with the diabetes. It is noteworthy that the Chi-square test generally requires a sufficiently large sample size. As a rule of thumb, each cell in the contingency table should have an expected frequency of 5 or more. The large sample size used in the paper ensures the successful satisfaction of the important assumption.

Correlation Analysis

In addition to the Chi-square test used for testing the independence between the diabetes status and the categorical features, the correlation analysis is also implemented to evaluate the strength and direction of the linear relationship among pairs of quantitative variables including age, blood_glucose_level, bmi, and HbA1c_level. The most popular correlation measures Pearson correlation coefficient, often denoted as r, is used herein. The Pearson correlation coefficient ranges from -1 to +1. A value of +1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship. The formula for calculation of r is denoted as:

$$r = \frac{\sum (x_i - x)(y_i - y)}{\sqrt{\sum (x_i - \bar{x})(x_i - \bar{x}) \sum (y_i - \bar{y})(y_i - \bar{y})}} \tag{3}$$

Where are individual value and mean of the values for the x-variable, respectively, while are the corresponding values for y-variable [21].

Variance Importance Ranking by Random Forest

The above-mentioned correlation and chi-square tests are done based on the linear relationship between two features. They show enormous limitations in complicated nonlinear interactions among multiple variables. To address this shortcoming, the random forest is also utilized to examine the importance of all features to the diabetes target when considering all predictors simultaneously.

Random Forest is a versatile and widely-used machine learning algorithm that operates by constructing multiple decision trees during training. It provides an excellent method for ranking the importance of variables (also known as features) in a dataset for predictive modeling. The general procedure to determine variable importance contains multiple steps: 1. Building numerous decision trees using a random subset of the data and a random subset of variables at each split. 2. Keeping track of which features are used to split data at each node and how much the split improves the purity of the node (e.g., by reducing the Gini impurity in classification tasks or the mean squared error in regression). 3. Calculating the importance of each feature upon the construction of the forest of trees. In between the two measurements, the Gini Importance, is selected over the Mean Decrease in Accuracy as the former offers a more efficient, model-intrinsic way to assess feature importance, especially when a direct interpretation of the model's decision-making process is needed. In general, the calculation of Gini is done following the expression below:

$$Gini\ Impurity = 1 - \sum_{i=}^{M} p_i^2 \tag{4}$$

Where M is the number of classes, and p_i is the proportion of the ith class [22]. This method measures how each feature contributes to the homogeneity of nodes and leaves in the decision trees. A higher value means the feature is more important for making splits.

Binary Logit Regression

While correlation analysis and Random Forest provide valuable insights into variable relationships and importance, binary logistic regression adds a layer of interpretability, statistical testing, and a different perspective on the data, especially useful for understanding the effect size and direction of predictors on the binary diabetes outcome. The regression models the log-odds (logit) of the probability of the event occurring, rather than modeling the probability directly. The log-odds are the natural logarithm of the odds, where odds are defined as the probability of the event occurring divided by the probability of the event not occurring. The formula of the binary logit regression is expressed as follows:

$$\log(\frac{p}{1-p} = \beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_mX_m \qquad (5)$$

Where p is the probability of the dependent variable equaling the presence of diabetes; p/(1-p) is the odds of the diabetes occurring; $X_i^{'}s$ are the set of independent variables, and $\beta_i^{'}s$ are the corresponding coefficients [23].

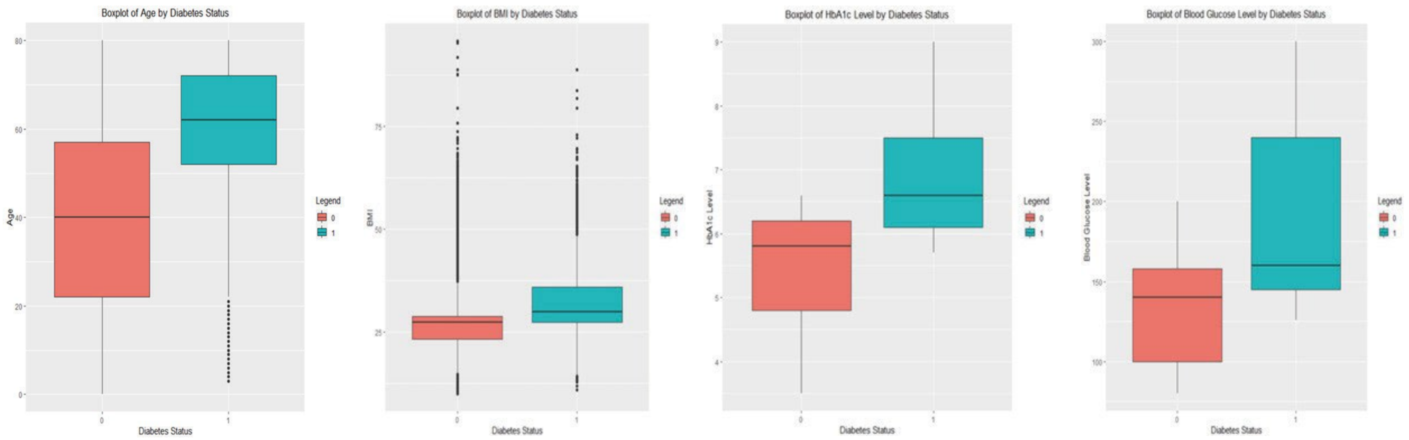
Albeit with multiple benefits, the binary logit models also suffer some limitations such as assumption of linearity and predictor independence. The previous correlation analysis among the independent variables indicates the multicollinearity issue could occur that may lead to inflated standard errors and hence can make some variables appear statistically insignificant when they should be significant, or vice versa. However, removing some of the highly correlated variables, also known as omitting a confounder, could yield biased coefficients, incorrect conclusions, and reduced predictive accuracy. After careful consideration, I chose to retain all collected independent variables to maintain the model and coefficient accuracy at the cost of potential failure to capture some of the truly statistically significant variables.

RESULTS

Each of the previously stated methods is applied to the comprehensive diabetes dataset to reveal the relationship between different variables and response variables from different perspectives. The detailed results are presented in the following subsections.

Exploration of Relationship between Diabetes and Individual Predictor Variables

The four panels in Figure 1 indicate the visual display of relationship of diabetes status and the individual quantitative factors using the comparison of boxplots (Figure 1). The group with diabetes status "0" representing those without diabetes is indicated by the red boxplot, while the group with diabetes status "1" representing with diabetes is displayed by the cyan boxplot. In panel (a), the median age of the group without diabetes is around 27 years old as indicated by the line in the middle of the red box. The interquartile range (IQR), which is the range of the middle 50% of the data, spans approximately from 22 to 32 years old, indicated by the top and bottom of the red box. There are no visible outliers or extreme values in this group. The median age of the group with diabetes is higher, around 55 years old, indicated by the line in the middle of the cyan box. The IQR for this group is wider, spanning from about 45 to 65 years old, indicating greater variability in age within this group. There are a few outliers or extreme age values indicated by the individual points below the bottom of the cyan box, suggesting that there are some younger individuals with diabetes. Overall, the boxplot shows that individuals with diabetes tend to be older than those without diabetes, with a wider range of ages and some younger outliers in the diabetes group. Likewise, it can be inferred from Panel (b) that individuals with diabetes tend to have a higher BMI, as indicated by both the median and the range of the IQR. The presence of more extreme outliers in the diabetic group suggests greater variability in BMI among those with diabetes. Panels (c) and (d) indicate that individuals with a diabetes



(a) Diabetes vs. Age (b) Diabetes vs. BMI (c) Diabetes vs. HbA1c (d) Diabetes vs. Blood Glucose

Figure 1. Comparative Boxplots Indicating the Relationship Between Diabetes Status and Other Individual Numerical Influential Factors.

status of ‘1’ have significantly higher HbA1c and blood glucose levels compared to those with a status of ‘0’.

Table 2 presents the results of chi-square tests comparing diabetes status with three different categorical variables: smoking, hypertension, and gender. The degrees of freedom for each test differ, with the test involving hypertension having only 1 degree of freedom, while the test involving smoking has 5, and the test involving gender has 2. This suggests that different numbers of categories were compared in each test. The p-values are all less than 2.2e-16, which is far below the 0.05 threshold, indicating a very strong statistical significance in the association between diabetes status and each of the three variables.

Correlation analysis among Predictor Variables

In addition to the exploration of the relationship between diabetes status and individual independent variables, it is important to identify the correlation among the set of predictors as well. Pearson’s correlation analysis results and their associated p-values are illustrated in Table 3.

The upper portion of the table presents the correlation coefficient values, which are a statistical measure that describes the size and direction of a relationship between two or more variables. The relationship of the pairs of variables are shown below:

- **age** and **bmi** have a correlation of 0.337, suggesting a moderate positive relationship.
- **age** and **HbA1c_level** have a correlation of 0.101, indicating a weak positive relationship.
- **age** and `blood_glucose_level` have a correlation of 0.111, which is also a weak positive relationship.
- **bmi** and **HbA1c_level** have a correlation of 0.083, suggesting a very weak positive relationship.
- **bmi** and **blood_glucose_level** have a correlation of 0.091, another weak positive relationship.
- **HbA1c_level** and **blood_glucose_level** have a correlation of 0.167, which is a weak positive

Table 2. Chi-Square Test Results for Diabetes Status and Other Individual Categorical Variables

Categories	Chi-squared Test		
	X-Squared	Degree of Freedom	p-value
Diabetes vs. Smoking	1956.1	5	<2.2e-16
Diabetes vs. Hypertension	3910.7	1	<2.2e-16
Diabetes vs. Gender	143.2	2	<2.2e-16

Table 3. Correlation Analysis Results among Predictor Pairs

Variables	age	bmi	HbA1c_level	blood_glucose_level
	Correlation Coefficient Values			
age	1	0.337	0.101	0.111
bmi	0.337	1	0.083	0.091
HbA1c_level	0.101	0.083	1	0.167
blood_glucose_level	0.111	0.091	0.167	1
	Correlation Coefficient p-Values			
age	1.00	0.000000e+00	1.520e-226	5.64e-270
bmi	1.00	1.00	2.430e-152	6.81e-184
HbA1c_level	1.52e-226	2.43e-152	1.00	1.00
blood_glucose_level	5.64e-270	6.81e-184	1.00	1.00

relationship but stronger than the other correlations involving **HbA1c_level** and **blood_glucose_level**.

- The diagonal of the matrix, which compares each variable to itself, has a correlation coefficient of 1, as any variable is perfectly correlated with itself.

The p-values as shown in the lower portion of the table are extremely small, indicating highly significant results from the statistical tests comparing these variables. The results are consistent with the visual representation of a correlation matrix among the various numerical values. As shown in Figure 2, each cell in the matrix shows the correlation coefficient between the variables, while the size and color of the circles in each cell represent the strength of the correlation, with larger and darker circles indicating a stronger relationship. Again, it is obvious that the strongest positive correlation shown in the matrix is between age and BMI, while the correlations between the other variables are relatively weak. The relationship between HbA1c level and blood glucose level, while still weak, is the strongest among the three non-identity correlations.

Variable Importance Ranking Results Uisng Random Forest

Aside from the above tools for demonstrating the relationship among the various pairs of the predictors and between the response and individual predictor variables, it is also worthwhile to assess the importance of various influential variables on the response variable (or, diabetes status) with the presence of all other variables. The decision tree oriented random forest is a popular method to satisfy such purpose.

Figure 3 shows a decision tree for diabetes prediction. It's a simple tree with two levels of decision nodes based on threshold values of certain medical test results. The first node (root node) is based on the HbA1c level. The decision criterion here is whether the HbA1c level is less than 6.7. If the HbA1c level is less than 6.7, we follow the branch to the left. If the HbA1c level is 6.7 or higher, we follow the branch to the right, which leads directly to a leaf node with diabetes status "1". On the left side, the next decision node is based on the blood glucose level, with a threshold of 201. If the blood glucose level is less than 201, we follow the branch to the left, which leads to preparation of status "0". If the blood glucose level is 201 or higher, we follow the branch to the right, leading to prediction of status "1". Overall, the tree structure implies that HbA1c level is the primary deciding factor, followed by the blood_glucose_level as the second most important variable. Even though the single decision tree is easy to understand and implement, it suffers multiple disadvantages including overfitting, lack of stability, and biased results, especially if some classes dominate (or, status "0" in the present study). Hence, the authors also proceed with variable importance analysis using the more popular random forest method, which provides a reliable method for ranking variable importance due to their ensemble nature, ability to handle different types of data, and robustness to overfitting and multicollinearity.

Specifically, the mean decrease in Gini provides a way to rank features based on their contribution to the homogeneity of nodes and leaves in a random forest, thereby indicating their significance in the model's

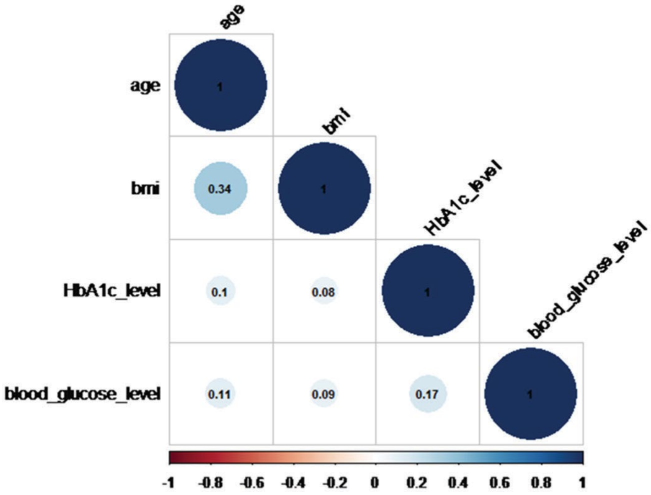


Figure 2. Correlation Matrix Plot Among Numerical Variables.

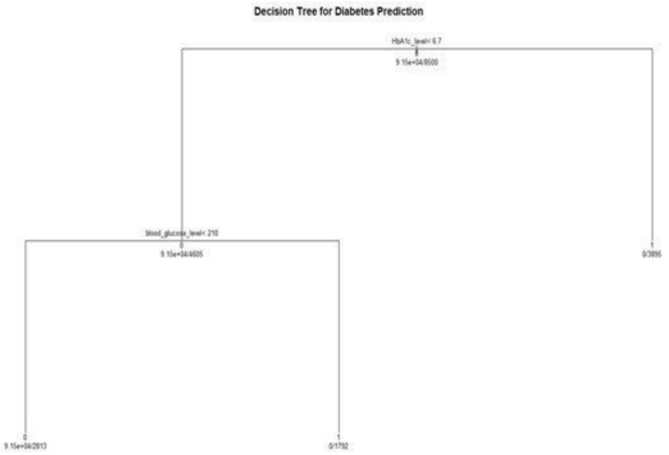


Figure 3. Single Decision Tree of Diabetes Prediction.

decision-making process. The higher this value, the more important the feature is deemed to be for the prediction task performed by the random forest. As shown in Figure 4, each feature's importance is represented as a horizontal bar, with the length of the bar indicating the magnitude of the mean decrease in Gini. The results suggest that according to this random forest model, HbA1c level and blood glucose level are the most important features for predicting the target variable, while gender is the least important. Such findings are generally in consistent with the previous results from other statistical analysis.

Binary Logit Regression Results

While the previous random forests are useful for capturing non-linear relationships and interactions without a need for specifying a functional form, and correlation analysis provides a measure of the strength and direction of linear relationships between variables, logistic regression analysis brings in a structured approach to quantify relationships, control for confounding, and provides a framework for statistical inference. Hence, the binary logit regression was also performed for the assessment of the diabetes status , and the detailed modeling results are shown in Table 4.

From this model output, it is known that almost all predictors are statistically significant at the level of 0.05, meaning that there is a strong association between these factors and the likelihood of having diabetes, as modeled by this logistic regression. Specifically, the intercept value is quite negative, suggesting that when all other variables are at zero, the log-odds of having diabetes are low. However, this is a theoretical value because the variables cannot actually be zero (e.g., age cannot be zero). Among the categorical variables, the positive coefficient for gender Male suggests that being male is associated with higher log-odds of having diabetes compared to the baseline gender (presumably female). This implies that males in this study have a higher likelihood of diabetes when controlling for other factors. The positive statistical relationship also appears in the other two categorical variables, hypertension and heart_disease, suggesting individuals with hypertension and heart diseases have higher log-odds of having diabetes. Interestingly, for the smoking_hostory, all other types of history are associated with a negative coefficient (statistically significant or insignificant), indicating the current smoking status leads to the highest chance of having diabetes, compared with all other smoking statuses.

For all quantitative variables, age, bmi, blood_glucose_level, and hba1c_level, the estimated coefficients are shown to be statistically significantly positive. Such findings indicate that the log-odds of having diabetes increase with the patients' bmi, blood_glucose_level, and hba1c_level. Overall, this is consistent with the general medical knowledge and previous statistical analyses.

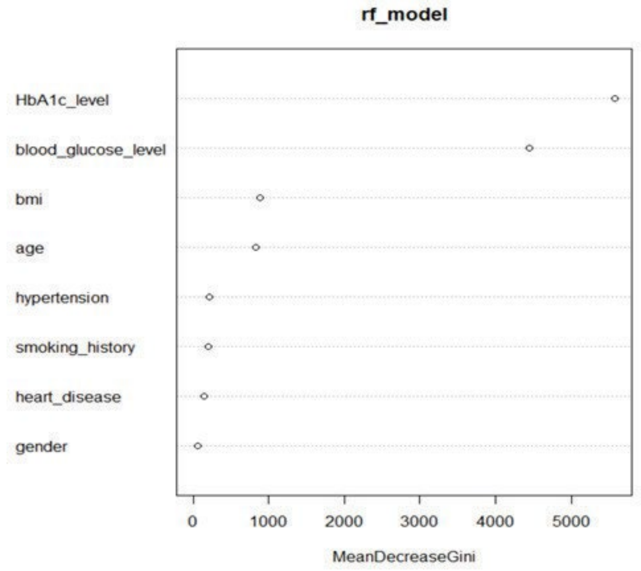


Figure 4. Feature Importance Ranking Results Using Mean Decrease of Gini in Random Forest.

CONCLUSIONS AND RECOMMENDATIONS

This comprehensive study, integrating a diverse array of machine learning and statistical methodologies, marks a significant advancement in diabetes research. By analyzing an extensive dataset of 100,000 entries, the authors have explored the multifactorial nature of diabetes, a condition that continues to pose a substantial public health challenge in the United States. The research findings underscore the complexity of diabetes, influenced by a set of factors including genetic predispositions, medical indicators, and socio-economic conditions. The utilization of advanced machine learning techniques, such as Random Forests, alongside traditional statistical methods like binary logistic regression, has enabled a more robust and nuanced understanding of the causal factors of diabetes. This hybrid approach not only enhances predictive accuracy but also provides deeper insights into the relative importance and interplay of various determinants. The study's implications extend beyond the academic realm, offering valuable insights for policymakers, healthcare providers, and individuals, highlighting the critical need for targeted intervention

Table 4. Binary Logit Regression Results for Parameter Estimation

Variable	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	-27.080	0.293	-92.455	< 2e-16
genderMale	27.240	0.036	7.540	4.69e-14
genderOther	-9.475	102.900	-0.092	0.927
age	0.046	0.001	41.040	< 2e-16
hypertension	0.741	0.047	15.737	< 2e-16
heart_disease	0.735	0.061	12.099	< 2e-16
smoking_historyever	-0.051	0.029	-0.551	0.582
smoking_historyformer	-0.108	0.070	-1.546	0.122
smoking_historynever	-0.157	0.061	-2.586	0.010
smoking_historyno_info	-0.730	0.067	-10.981	< 2e-16
smoking_historynot_current	-0.211	0.033	-2.538	0.011
bmi	849.500	0.003	331.819	< 2e-16
hba1c_level	2.390	0.036	66.413	< 2e-16
blood_glucose_level	0.033	0.000	69.207	< 2e-16
Null deviance: 58163 on 99999 degrees of freedom				
Residual deviance: 22627 on 99986 degrees of freedom				
AIC: 22655				

Notes: 1. The bold font indicates the statistical significance at the level of 0.05. 2. Refer to Table 1 for the definition of the variables.

strategies and policy initiatives. The integration of diverse methodologies sets a precedent for future research, emphasizing the importance of multifaceted approaches in tackling complex health issues like diabetes.

Even though the study leads to the elevated understanding regarding diabetes status using the rarely used observational data with excessively large sample size, it suffers some limitations that could be enhanced in the future. First, the typical binary logit regression model is used, which lacks the capability to capture the underlying heterogeneity existing in various patients. The more advanced models like those with random effects or random parameters could generate more accurate estimates. Second, due to the under-sampling nature associated with diabetes status, it's worth exploring the impact of the sophisticated sampling strategies to the distinct data analytical methods used in the paper.

ACKNOWLEDGEMENT

I extend my deepest gratitude to the contributors of the Kaggle dataset used in this study. The richness and diversity of the data provided a robust foundation for our analysis, contributing significantly to the understanding of the complex factors influencing diabetes. I would also like to express my sincere appreciation to the anonymous reviewers and editors for their invaluable contributions to my work.

REFERENCES

1. Mukhtar, Y., A. Galalain, and UJEJoB Yunusa. A modern overview on diabetes mellitus: a chronic endocrine disorder. European Journal of Biology 5.2 (2020): 1-14.
2. Reichard, P., Pihl, M., Rosenqvist, U., & Sule, J. (1996). Complications in IDDM are caused by elevated blood glucose level: the Stockholm Diabetes Intervention Study (SDIS) at 10-year follow up. Diabetologia, 39, 1483-1488.
3. Aldosari, M., Aldosari, M., Aldosari, M. A., & Agrawal, P. (2022). Diabetes mellitus and its association with dental caries, missing teeth and dental services utilization in the US adult population: Results from the 2015–2018 National Health and Nutrition Examination Survey. Diabetic Medicine, 39(6), e14826.
4. Bloomgarden, Z. T. (2004). Consequences of diabetes: car-

diovascular disease. Diabetes Care, 27(7), 1825-1831.

5. ElSayed, N. A., Aleppo, G., Aroda, V. R., Bannuru, R. R., Brown, F. M., Bruemmer, D., Gabbay, R. A. (2023). 6. Glycemic targets: standards of care in diabetes—2023. Diabetes Care, 46(Supplement_1), S97-S110.
6. Yuan, S., Merino, J., & Larsson, S. C. (2023). Causal factors underlying diabetes risk informed by Mendelian randomisation analysis: evidence, opportunities and challenges. Diabetologia, 66(5), 800-812.
7. Tabaei, B. P., & Herman, W. H. (2002). A multivariate logistic regression equation to screen for diabetes: development and validation. Diabetes Care, 25(11), 1999-2003.
8. Tachkov, K., Mitov, K., Koleva, Y., Mitkova, Z., Kamusheva, M., Dimitrova, M., Petrova, G. (2020). Life expectancy and survival analysis of patients with diabetes compared to the non-diabetic population in Bulgaria. PloS one, 15(5), e0232815.
9. Lohia, P., Kapur, S., Benjaram, S., Cantor, Z., Mahabadi, N., Mir, T., & Badr, M. S. (2021). Statins and clinical outcomes in hospitalized COVID-19 patients with and without Diabetes Mellitus: a retrospective cohort study with propensity score matching. Cardiovascular Diabetology, 20(1), 1-15.
10. Seto, H., Oyama, A., Kitora, S., Toki, H., Yamamoto, R., Kotoku, J. I., Moriyama, T. (2022). Gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data. Scientific reports, 12(1), 15889.
11. Yu, W., Liu, T., Valdez, R., Gwinn, M., & Khoury, M. J. (2010). Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC medical informatics and decision making, 10(1), 1-7.
12. Lee, J. E., Jeon, H. J., Lee, O. J., & Lim, H. G. (2024). Diagnosis of diabetes mellitus using high frequency ultrasound and convolutional neural network. Ultrasonics, 136, 107167.
13. Saini, M., & Susan, S. (2022). Diabetic retinopathy screening using deep learning for multi-class imbalanced datasets. Computers in Biology and Medicine, 149, 105989.
14. Akyol, K. (2017). Assessing the importance of attributes for diagnosis of diabetes disease. International Journal of Information Engineering and Electronic Business, 9(5), 1.
15. Laws, R. A., George, A. B. S., Rychetnik, L., & Bauman, A. E. (2012). Diabetes prevention research: a systematic review of external validity in lifestyle interventions. American journal of preventive medicine, 43(2), 205-214.
16. Robben, J. H., Knoers, N. V., & Deen, P. M. (2006). Cell biological aspects of the vasopressin type-2 receptor and aquaporin 2 water channel in nephrogenic diabetes insipidus. American Journal of Physiology-Renal Physiology, 291(2), F257-F270.
17. Hill-Briggs, F., Adler, N. E., Berkowitz, S. A., Chin, M. H., Gary-Webb, T. L., Navas-Acien, A., Haire-Joshu, D. (2021). Social determinants of health and diabetes: a scientific review. Diabetes care, 44(1), 258.
18. Mustaf, M. (2023). Comprehensive Dataset for Predicting Diabetes with Medical & Demographic Data. Published by Kaggle. https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data (accessed in 11/2023)
19. Pfannkuch, M. (2006). Comparing box plot distributions: A teacher’s reasoning. Statistics Education Research Journal, 5(2), 27-45.
20. Sharpe, D. (2015). Chi-square test is statistically significant: Now what? Practical Assessment, Research, and Evaluation, 20(1), 8.
21. Cohen, I., Huang, Y., Chen, J., Benesty, J., Benesty, J., Chen, J., Huang, Y. and Cohen, I., 2009. Pearson correlation coefficient. Noise reduction in speech processing, pp.1-4.
22. Nembrini, S., König, I. R., & Wright, M. N. (2018). The revival of the Gini importance? Bioinformatics, 34(21), 3711-3718.
23. Harrell, Jr, F. E., & Harrell, F. E. (2015). Binary logistic regression. Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis, 219-274.