Risk of Diabetes- A Cross-Sectional Analysis of Leading Physiological and Demographic Factors

Samriddhi Soni (ss6557)

MS in Applied Statistics, Teachers College, Columbia University

Abstract

This cross-sectional study examines the risk of diagnosis of diabetes by analyzing physiological health metrics and demographic factors including Glycohemoglobin levels, Blood Urea levels, BMI, Age and Ethnicity of individuals with methods including logistic regression, HCA Heatmap, Classification tree and accuracy analysis. Our findings and present literature indicates that these factors significantly contribute in increasing the risk of diagnosis of diabetes, highlighting the complex and substantial role of presence of chronic diseases and demographic conditions.

Keywords: *Diabetes, Classification, Clustering, Chronic Disease, Physical Health, Clustering*

Introduction

Diabetes is a chronic medical condition that is caused due to increased blood glucose levels. Diabetes prevalence has increased significantly in the past few years among adults in the United States. There are several physiological factors and demographic conditions that can be a contributing factor in increasing the risk of diabetes among individuals, therefore this study aims to identify those risk factors and their relationship with prevalence of diabetes. Physical health factors such as levels of physical activity, BMI and body weight leads to obesity, present literature indicates a strong association between obesity and diabetes. Presence of higher levels of body fat can increase the risk of type of diabetes among individuals Klein, S. et al (2022). Apart from diabetes, other chronic diseases such as kidney failure have a strong relationship with the prevalence of diabetes. Thomas, M. C. et al (2015) has elaborated on the fact that 1 in 3 adults diagnosed with diabetes have presence of chronic kidney disease. This study recognizes the interplay of these chronic diseases and aims at studying these risk factors contributing to prevalence of diabetes.

Apart from analyzing the physiological measures, demographic factors such as age and ethnicity play a role in the diagnosis of diabetes. Advanced age is a major risk factor for diabetes and prediabetes; therefore, the elderly has a higher prevalence of diabetes than the young and middle-aged as discussed in Yan, Z. et al (2023). The race and ethnicity of an individual can be detriment in determining the risk of diagnosis of diabetes or pre-diabetes. Diabetes is more common among African American and Asian-American individuals that non-Hispanic white individuals, DeCoster, V. A., & Cummings, S. (2005).

Using the analysis of the above-mentioned risk factors, this study aims at contributing to the research for disease prevention and maintaining a healthy lifestyle. The research questions that would aid the analysis for this study are as follows:

1. *What is the relationship between physiological factors and health metrics to the risk of diabetes?*
2. *Does demographic factors and ethnic background have a relationship with the diagnosis of diabetes?*

Data Source and Description

The data for this study has been sourced from the Department of Biostatistics, Vanderbilt University, and the additional information on the methodology of the survey collection came from the National Center for Health Statistics (NCHS) of Centers for Disease Control and Prevention (CDC).

The dataset contains 11 variables and 6706 responses from 6706 respondents who were surveyed during the National Health and Nutrition Examination Survey 2009-2010 to assess the

health and nutritional status of adults and children in the United States. The variables included in the study are:

1. Gender - Gender of the subjects
2. Age - Age of the subjects in years
3. Ethnicity - Race and ethnicity of the subjects
4. Family_income - Household income of the subjects
5. Diabetes- Diagnosis of subjects, where Diagnosed with Diabetes = TRUE or Pre-Diabetes = FALSE
6. BMI - Body Mass Index of individuals
7. Weights - Body weight of the subject in kilograms
8. Glycohemoglobin - Percentage of glycohemoglobin levels in subjects
9. Albumin - Albumin levels in subjects, given in grams per deciliter
10. Blood_Urea_Nitrogen - Blood urea nitrogen levels in subjects, given in milligrams per deciliter
11. Serum_Creatinine - Serum creatinine levels in subjects, given in milligrams per deciliter

Methodology

1. Exploratory Data Visualization

Exploratory data visualizations allow users to explore data interactively. It can be used to manipulate variables, zoom in on specific aspects, and gain a deeper understanding of the data through exploration and analysis of variables. It is very important to accurately represent data in a visualization, avoid distracting elements and highlight the important information as mentioned in Schwabish (2021). Gestalt principles are very useful set of rules that help in achieving effective

and useful visualizations. The laws of connection, similarity and proximity are followed in the visualizations to achieve the purpose of communicating the results from the data.

Apart from using the visualizations to understand the distribution of data, correlation heatmaps can provide a very effective graphic to understand the strength of relationship between the diabetes diagnosis and demographic and physiological health measures.

2. HCA Heatmap

In this study HCA heatmap are used as a data visualization technique to display the results of hierarchical clustering. For understanding the relationship between the diagnosis of diabetes, demographic factors and physiological health conditions, this author chose cluster analysis heatmaps after following recommendations for HCA heatmaps from Bowers (2010) using uncentered correlation as the distance metric and average linkage as the unsupervised hierarchical clustering algorithm. This heatmap provide a graphical representation of the relationships and similarities between covariates by arranging theme in similar clusters.

The clustering distance metric used is uncentered correlation (Equation 1) as the distance metric, $r(x_i, y_i)$, which is preferred over the similar Pearson correlation as it assumes the mean is zero for each vector x and y after recommendation from Bowers (2010). Thus, the uncentered correlation for any two vectors x_i and y_i of sample size *n* is:

$$r(x_i, y_i) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i}{\sigma_x^{(0)}}\right)\left(\frac{y_i}{\sigma_y^{(0)}}\right) \qquad \textbf{Equation 1}$$

where:

$$\sigma_x^{(0)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i)^2}$$

$$\sigma_y^{(0)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i)^2}$$

3. Logistic Regression

Logistic Regression is an unsupervised machine learning technique which is used for classification and predictive analytics. Logistic regression is essentially used for classification of binary response variables based on the predictor variables. For this technique log odds, or the natural logarithm of odds, and this logistic function is used for classification, Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013).

After the model fitting and classification, it's best practice to evaluate the model prediction using confusion matrix. Confusion matrix is a popular technique used for the computation of accuracy, sensitivity, and specificity of the model.

4. Classification Tree

A classification tree is a decision tree Classification and regression tree) that predicts a qualitative (i.e., categorical) variable. In general, the CART approach relies on stratifying or segmenting the prediction space into several simple regions. To make regression-based or classification-based predictions, we use the mean or the mode of the training observations in the region to which they belong.

Gini Index: This is essentially a measure of total variance across the K classes. A small Gini index indicates that a node contains predominantly observations from a single class. When building a classification tree, Gini Index is typically used to evaluate the quality of a particular split, as they are more sensitive to the changes in the splits than the classification error rate.

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

After the classification, it's best practice to evaluate the accuracy of model using confusion matrix. Confusion matrix is a popular technique used for the computation of accuracy, sensitivity, and specificity of the model. The variable importance score provides the score for variables that significantly contributed to the classification.

Result and Conclusion

1. Exploratory Data Visualization

Figure 1

Line Chart of Glycohemoglobin vs Diabetes

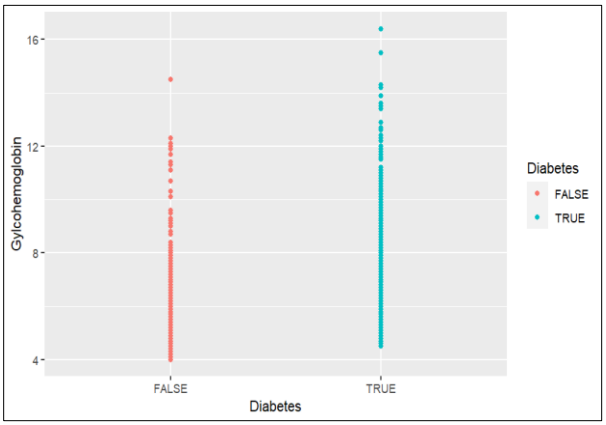
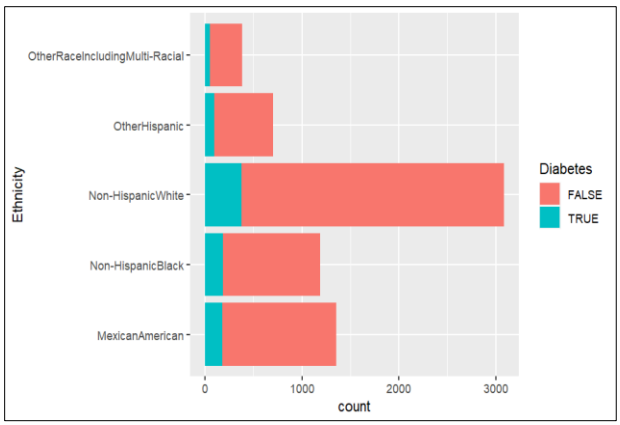


Figure 2

Bar Chart of Ethnicity vs Diabetes



The glycohemoglobin levels also know as hbA1c is a common indicator for diagnosing diabetes and pre-diabetes. Higher hbA1c levels indicates the presence of diabetes, hbA1c levels are lower for patients suffering pre-diabetes as compared to diabetes. Figure 1 indicates higher level of hnA1c for subjects diagnosed with diabetes, which is similar to the findings in present literature, Sherwani, S. I., Khan et al (2016).

Figure 2 illustrates the total count of subjects diagnosed with diabetes and pre-diabetes in various ethnic groups. The count of Non-Hispanic White individuals diagnosed with diabetes is much higher than any other ethnic group. This finding is quite interesting as this result can be a

mixture of various factors such as dietary choices among racial groups and their genetic and family history of diseases.

Figure 3

Density Plot of Age vs Diabetes

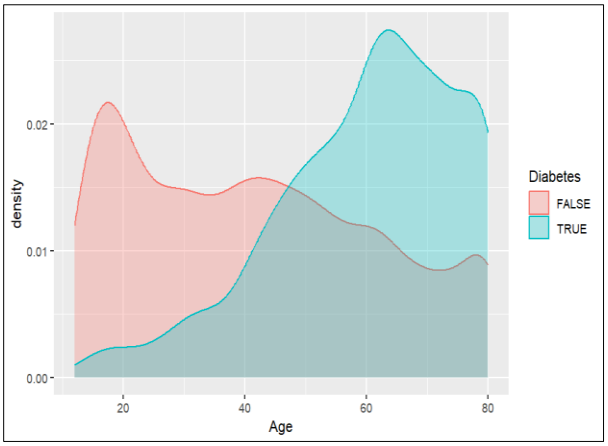


Figure 4

Correlation Heatmap

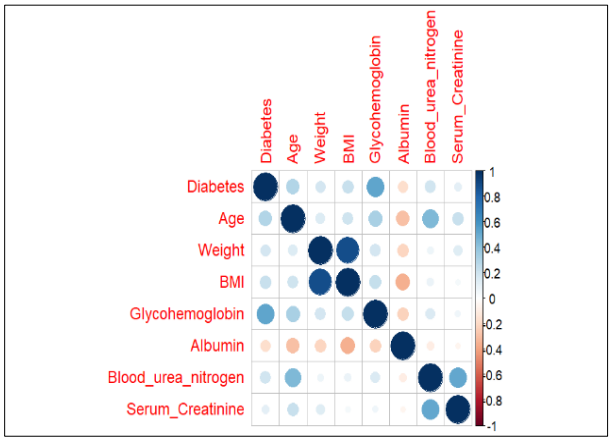


Figure 3 illustrates the individuals diagnosed with diabetes have higher density for age much higher than the individuals diagnosed with pre-diabetes. The maximum individuals diagnosed with pre-diabetes are young and maximum individuals diagnosed with diabetes are older individuals. This indicates that elderly individuals have more risk and higher prevalence of diabetes, Yan, Z. et al (2023).

The positive correlation indicated in Figure 4 between Glycohemoglobin levels and diabetes, has been discussed earlier. The factors such as Weight and BMI have high correlation among themselves and a positive relationship with diabetes. This relationship since presence of higher levels of body fat can increase the risk of diabetes among individuals, Klein, S. et al (2022).

Figure 4 displays negative correlations between albumin levels and diabetes. Low albumin level indicates a problem with kidney and liver conditions. Kidney failures have a strong relationship with the prevalence of diabetes. Thomas, M. C. et al (2015). Therefore, the negative correlation is an accurate representation of the relationship between these factors. Higher blood

urea nitrogen is associated with increased risk of incident diabetes mellitus, Xie. Y. (2018). Therefore, Figure 4 indicates a positive correlation between diabetes and blood urea levels.

2. HCA Heatmap

Figure 5

HCA Heatmap with Diagnosis of Diabetes as Criteria

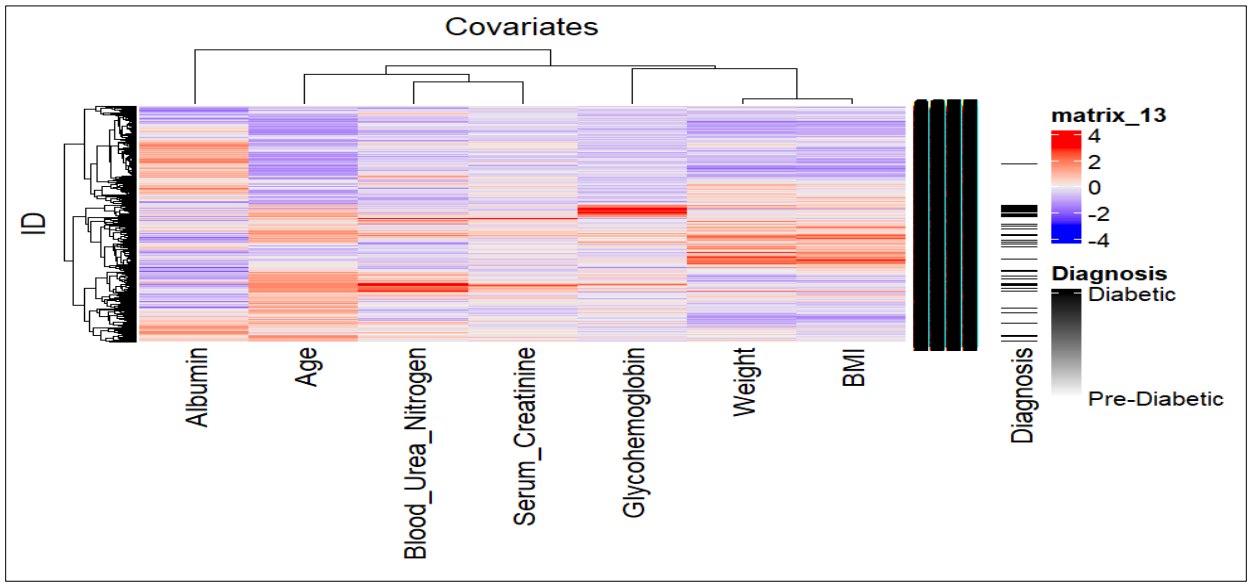


Figure 5 is an HCA heatmap that has seven numerical variables which are essentially the physiological and demographic factors that are related to the diagnosis of diabetes. The physical health information represented by BMI and weight have similar pattern with the diagnosis of diabetes. Glycohemoglobin levels have a very similar pattern to the diagnosis of diabetes, as it is used for the diagnosis of diabetes.

The Serum creatinine and blood urea levels have very similar pattern with each other, higher level of both the measures indicate kidney failure. Due to the association of chronic kidney disease and diabetes, both the measures and diabetes have similar pattern. Albumin have opposite pattern

to the blood urea and nitrogen and serum creatinine levels, since lower levels of albumin indicates kidney failure.

3. Logistic Regression

Table 1

Logistic Regression table with diabetes as criteria

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-13.474087	1.049139	-12.843	< 2e-16 ***
Age	0.027105	0.003614	7.499	6.43e-14 ***
Weight	-0.003376	0.005173	-0.653	0.514
BMI	0.066655	0.016835	3.959	7.51e-05 ***
Glycohemoglobin	1.436196	0.078965	18.188	< 2e-16 ***
Albumin	-0.097049	0.189226	-0.513	0.608
Blood_urea_nitrogen	0.019547	0.009752	2.004	0.045 *
Serum_Creatinine	0.104091	0.146009	0.713	0.476

Table 2

Confusion Matrix for Prediction through Logistic Regression

	Predicted FALSE	Predicted TRUE
Actual FALSE	1672	111
Actual TRUE	74	154

The coefficients obtained for age, BMI, glycohemoglobin levels and blood urea levels are significant predictors for classifying the diagnosis of diabetes. The increase in the log odds of these predictors indicates an increase in the log odds of diagnosis of diabetes. The model accuracy is 90.08%, sensitivity is 95.76% and specificity is 58.11%.

4. Classification Tree

Figure 6

Classification Tree with Diabetes as Criteria (cp-0043)

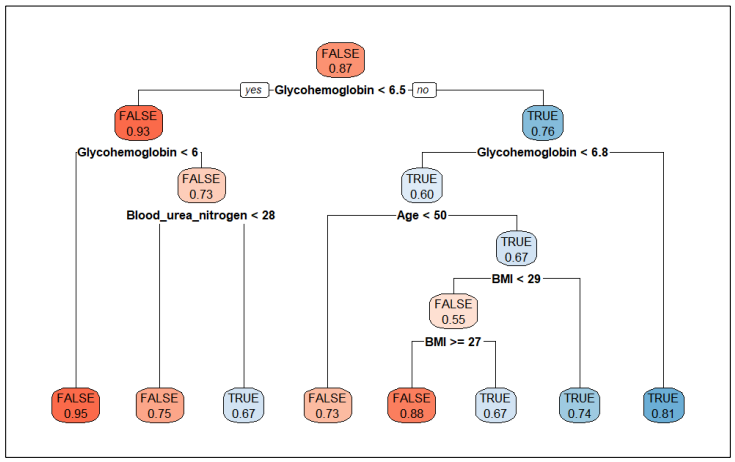


Figure 7

Optimal Value for Complexity Parameter

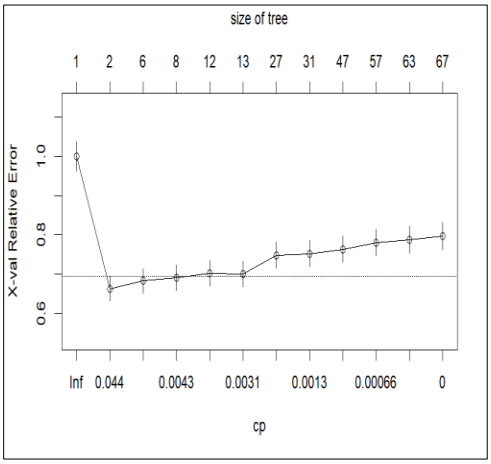


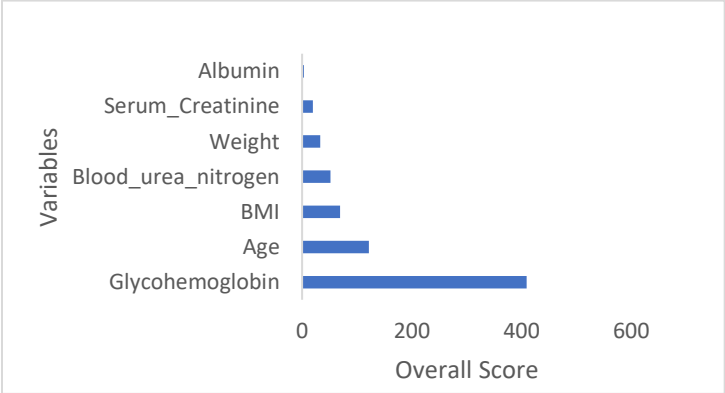
Table 3

Variable Importance table

	Overall
Glycohemoglobin	409.271225
Age	122.010891
BMI	69.288645
Blood_urea_nitrogen	51.753646
Weight	32.995068
Serum_Creatinine	19.469247
Albumin	3.370485

Table 4

Bar Chart depicting Variable Importance Score



The classification tree displayed in Figure 6, has four predictors that have highest variables importance scores. The glycohemoglobin levels, age, BMI, and levels of blood urea nitrogen has been used for the classification, these variables have the have highest variables importance scores in the order. The predictors are similar to the significant predictors obtained from Logistic Regression. The optimal complexity parameter 0.0043 is chosen from Figure 7, which neither overfit the tree nor over-simplified the prediction model.

The confusion matrix in Table 4 illustrates that the overall accuracy is around 91.6%, sensitivity is 97.37 % and specificity is 53.58% for this classification model.

Table 4

Confusion Matrix for Classification Tree

	Predicted FALSE	Predicted TRUE
Actual FALSE	1700	123
Actual TRUE	46	142

Discussion and Limitations

1. *What is the relationship between physiological factors and health metrics to the risk of diabetes?*

High BMI levels, Glycohemoglobin (hbA1c > 6.5%) and high levels of Blood urea nitrogen that turned out to be the leading physiological factors that contributed to an increase in the risk of diagnosis of diabetes.

2. *Does demographic factors and ethnic background have a relationship with the diagnosis of diabetes?*

Demographic factors such as age and ethnic background does affect the diagnosis of diabetes. The risk of diagnosis of diabetes is more in elderly individuals than in young adults and children.

For future scope of study, including socio-economic factors and detailed health metric can provide more diverse risk factors for prevalence of diabetes. This study was a cross-sectional study therefore for determining the impact of leading factors a longitudinal study can be conducted in future.

References

Klein, S., Gastaldelli, A., Yki-Järvinen, H., & Scherer, P. E. (2022). Why does obesity cause diabetes?. *Cell metabolism*, *34*(1), 11-20.

Thomas, M. C., Brownlee, M., Susztak, K., Sharma, K., Jandeleit-Dahm, K. A., Zoungas, S., ... & Cooper, M. E. (2015). Diabetic kidney disease. *Nature reviews Disease primers*, *1*(1), 1-20.

Yan, Z., Cai, M., Han, X., Chen, Q., & Lu, H. (2023). The Interaction Between Age and Risk Factors for Diabetes and Prediabetes: A Community-Based Cross-Sectional Study. *Diabetes, Metabolic Syndrome and Obesity*, 85-93.

DeCoster, V. A., & Cummings, S. (2005). Coping with type 2 diabetes: do race and gender matter?. *Social Work in Health Care*, *40*(2), 37-53.

Schwabish, J. (2021). The practice of visual data communication: what works. *Psychological Science in the Public Interest*, *22*(3), 97-109.

Bowers, A. J. (2010). Analyzing the longitudinal K-12 grading histories of entire cohorts of students: Grades, data driven decision making, dropping out and hierarchical cluster analysis. *Practical Assessment, Research, and Evaluation*, *15*(1), 7.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). "Applied Logistic Regression." Wiley.

Sherwani, S. I., Khan, H. A., Ekhzaimy, A., Masood, A., & Sakharkar, M. K. (2016). Significance of HbA1c test in diagnosis and prognosis of diabetic patients. *Biomarker insights*, *11*, BMI-S38440.

Xie, Y., Bowe, B., Li, T., Xian, H., Yan, Y., & Al-Aly, Z. (2018). Higher blood urea nitrogen is associated with increased risk of incident diabetes mellitus. *Kidney international*, *93*(3), 741-752.