Machine Learning Approach for Diabetes Risk Prediction.

Bijay Adhikari[1], Merit Kayastha[2,*], Nischal Bhandari[3] and Saurav Dahal[4]

[1] Department of Theoretical and Applied Sciences, Ramapo College of New Jersey; badhika2@ramapo.edu
[2] Department of Theoretical and Applied Sciences, Ramapo College of New Jersey; mkayasth@ramapo.edu
[3] Department of Theoretical and Applied Sciences, Ramapo College of New Jersey; nbhanda1@ramapo.edu
[4] Department of Theoretical and Applied Sciences, Ramapo College of New Jersey; sdahal4@ramapo.edu
* Department of Theoretical and Applied Sciences, Correspondence: mkayasth@ramapo.edu

Abstract:

This project focuses on analyzing the prevalence and predictors of diabetes in the United States using a dataset from the Behavioral Risk Factor Surveillance System (BRFSS). The dataset, obtained from Kaggle, comprises 70,692 survey responses with 17 feature variables and 1 target variable, allowing for a comprehensive analysis of factors influencing diabetes. The research involved correlating feature variables with diabetes, building predictive models to identify individuals at risk, and evaluating key predictors across top-performing models. Machine learning classifiers like Logistic Regression, KNN, SGD, Random Forest, Decision Trees, Bagging, and SVM were employed. The models were evaluated based on accuracy, precision, recall, and F1 scores, with Logistic Regression, Bagging, and SVM emerging as top performers. Key predictors of diabetes were identified, including CholCheck, HvyAlcoholConsump, HighBP, GenHlth, and others. Cross-validation was used to ensure model reliability. The study also examined the impact of lifestyle and demographic factors on diabetes likelihood, revealing significant associations with smoking, physical activity, diet, age, sex, and BMI. The analysis found that age, sex, and BMI are significant predictors of diabetes, which can aid in risk assessment and personalized medicine. The study emphasizes the importance of models with high recall scores to minimize false negatives. SGD and Linear SVM were identified as the best models based on their F1 scores and ROC-AUC values. In summary, this project provides insights into the factors influencing diabetes prevalence and offers a framework for predictive modeling in healthcare, highlighting the importance of tailored strategies in diabetes management and prevention.

Keywords: diabetes; predictive modeling; machine learning classifiers; accuracy; precision; recall; f1 score; cross-validation; behavioral risk factor surveillance system (brfss); kaggle; feature variables; roc-auc curve; lifestyle factors; demographic factors; statistical significance; public health

1. Introduction

Diabetes, characterized by elevated blood glucose levels, stands as a prevalent and pressing health issue globally, particularly in the United States, where it ranks among the leading causes of mortality [1]. Despite extensive efforts to address its impact, a substantial proportion of undiagnosed cases persist among the adult population [2]. The ramifications of unmanaged diabetes span a spectrum of complications, encompassing cardiovascular ailments, neuropathy, compromised immunity, and renal impairment [3]. This project aims to comprehensively analyze factors influencing diabetes prevalence and its associated complications. Through data analysis, the objective is to pinpoint key predictors contributing to diabetes onset. The project seeks to develop predictive models that accurately identify individuals at risk, thereby enhancing preventive healthcare strategies and interventions.

2. Materials and Methods

The dataset was obtained from Kaggle at https://www.kaggle.com/datasets/prosperchuks/health-dataset/data. The underlying data is actually from a 2015 survey conducted by the Centre for Disease Control and Prevention (CDC)'s Behavioral Risk Factor Surveillance System (BRFSS). BRFSS is a collaborative project designed to measure behavioral risk factors for the adult population (aged 18 years of age and older) residing in the United States. The 2015 BRFSS data is an aggregation of the combined landline and cell phone data and includes responses for 50 states of the US, the District of Columbia Guam, and Puerto Rico.

Our dataset is a collection of 70,692 survey responses with 17 feature variables and 1 target variable. The features of our dataset are as follows:

1) Age: 13 different age categories starting from (1=18-24 yrs, 2=25-29 yrs, 3=30-34 yrs, 4=35-39 yrs, 5=40-44 yrs, 6=45-49 yrs, 7=50-54 yrs, 8=55-59 yrs, 9=60-64 yrs, 10=65-69 yrs, 11=70-74 yrs, 12=75-79 yrs, 13=80 or older).
2) Sex: 1 for male, 0 for female.
3) HighChol: 0 for no high cholesterol, 1 for high cholesterol.
4) CholCheck: 0 for no cholesterol check in 5 yrs, 1 for yes.
5) BMI
6) Smoker: 0 for not smoked 5 packs of cigarettes in lifetime, 1 for yes.
7) HeartDiseaseorAttack: 0 for no, 1 for yes.
8) PhysActivity: in the past 30 days, 0 for no and 1 for yes.
9) Fruits: 1 for fruit consumed 1 or more times a day, 0 for no.
10) Veggies: 1 for vegetables consumed 1 or more times a day, 0 for no.
11) HvyAlcoholConsump: 0 for no, 1 for yes.
12) GenHlth: 1 =excellent, 2=very good, 3=good, 4=fair, 5=poor.
13) MentHlth:1-30 (days of poor mental health in a month).
14) PhysHlth: 1-30
15) DiffWalk: serious difficulty while walking or climbing stairs (0 for no, 1 for yes).
16) Stroke: 0 for no, 1 for yes.
17) HighBP: 0 for no, 1 for yes.
18) Diabetes: 0 for no, 1 for yes.

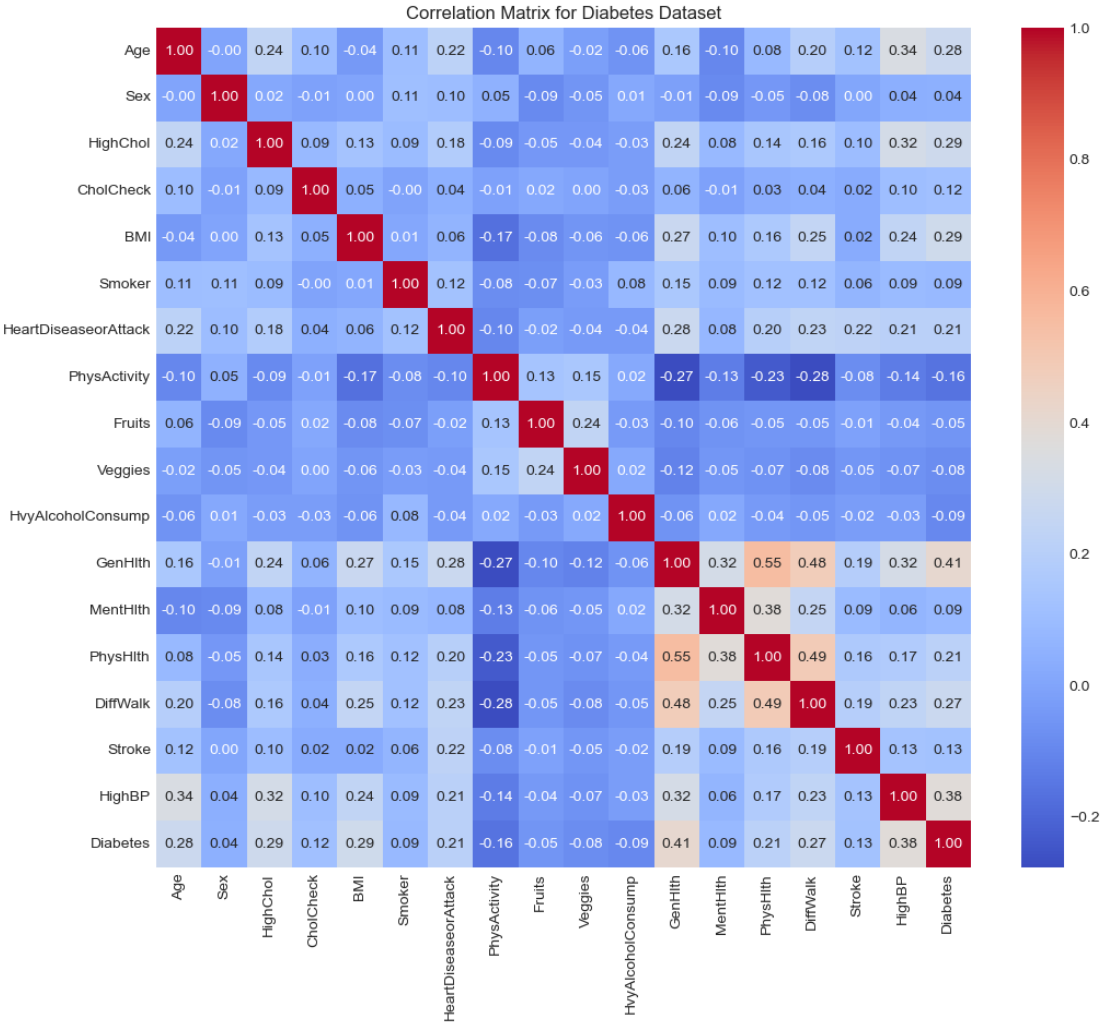
From our dataset, we used machine learning approaches to explore the following questions:

a) How do the factors interact with each other in influencing the likelihood of these health conditions? (Via correlation matrix)
b) Can we develop a predictive model that accurately identifies individuals at risk of these conditions? (Ensemble model from machine learning classifiers such as Logistic Regression, KNN, SGD, Decision Trees, Random Forest, Bagging, SVM).
c) What is the impact of lifestyle factors (like smoking, alcohol consumption, physical activity) on the likelihood of these health conditions? (Looking at feature importance coefficients from models with the best accuracies).
d) Are there significant differences in the prevalence of these health conditions across different demographic groups (like age, and gender)? (Via data visualization and chi-squared testing).

3. Results

3.1. Correlation of feature variables with response variable (Diabetes)

The strength of the correlation between "Diabetes" and certain factors like "GenHlth," "HighBP," "BMI," and "HighChol" (0.407, 0.381, 0.293, and 0.289 respectively) indicates a noticeable relationship with the presence of diabetes. Factors such as general health perception ("GenHlth"), high blood pressure ("HighBP"), body mass index ("BMI"), and high cholesterol ("HighChol") often coincide with or contribute to the development of diabetes, which explains stronger positive correlations. Conversely, features like "CholCheck," "MentHlth," "Smoker," "Sex," "Fruits," "Veggies," "HvyAlcoholConsump," and "PhysActivity" show either weak negative correlations or minimal associations, indicating a less direct relationship or influence on the occurrence of diabetes based on the correlation values. For instance, a physically active person under “PhysActivity” is less likely to be diagnosed with diabetes.



3.2. Building a Predictive Model to Identify Individuals At Risk of Diabetes

To predict if an individual is diabetic or not, we used 7 different machine learning classifiers: Logistic regression, K-Nearest Neighbor (KNN), Stochastic Gradient Descent (SGD), Random Forest, Simple Decision Trees, Bagging and Linear Support Vector Machine (SVM). Models like KNN are sensitive to the spread of data. So, the numerical variables such as Age, BMI, PhysHealth, MentHealth, and GenHealth were standardized using the Z-score standardization approach. The 70,692 rows of data were then split into training and testing set in the ratio of 80:20.

After tuning the hyperparameters to avoid cases of underfitting or overfitting, the accuracy scores, precision scores, recall scores, and F1 scores obtained for each classifier are depicted in Table 3.2:

Table 3.2. Performance scores for different machine learning models.

Classifiers	Accuracy Score	Precision Score	Recall Score	F1 Score
Logistic Regression	0.746446	0.734752	0.769045	0.751508
KNN (n = 5)	0.714761	0.700000	0.748759	0.723559
SGD	0.745668	0.709603	0.829196	0.764752
Random Forest	0.727138	0.710405	0.764222	0.736331
Bagging	0.746587	0.729963	0.780394	0.754337
Decision Tree	0.658604	0.663286	0.640233	0.651556
Linear SVM (cost parameter = 10)	0.746446	0.723773	0.794723	0.757590

In terms of accuracy, the best-performing classifiers were logistic regression, bagging, and linear SVM. Logistic Regression, Bagging, and Linear SVM were the top 3 models for precision. For the recall score, SGD, Bagging, and Linear SVM were the top-performing models. Similarly, SGD, Bagging, and Linear SVM were also the top-performing models for the F1 score.

Finally, all the different classifiers were combined to form an aggregate ensemble model with hard voting. Decision tree had the lowest score in all the categories, and thus they were not included in the ensemble model. Combining all the best models improved the accuracy of the ensemble model to 0.748143. The precision, recall, and F1 scores for the ensemble model were obtained to be 0.732347, 0.779827, and 0.755342 respectively. SGD model had the highest performance in both recall and F1 score. Since our data is fairly balanced, we plotted a ROC - AUC graph for the SGD model.

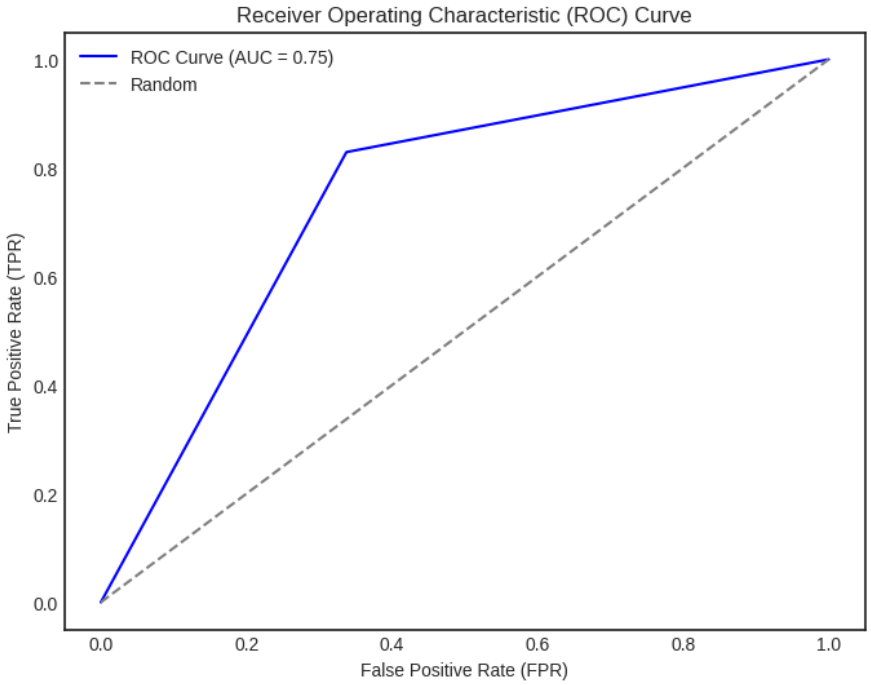


Figure 3.2. ROC - AUC Curve for SGD model with Area Under the Curve obtained to be 0.75.

3.3. Key predictors of diabetes across top-performing models

Among the 17 features, different features will have different weights across different models. To analyze the different features and their importance, it was important to analyze different individual classifiers rather than the combination of different base classifiers. That was the reason why bagging classifiers and ensemble models were not considered to find the top ten predictors of diabetes. After bagging and ensemble models, logistic regression and support vector machines were the models with the highest accuracy score, 0.746446 for both models. For the logistic regression, the top 10 features based on the absolute feature importance coefficients are listed below and displayed in *Fig 3.3(a)*:

1. Cholcheck
2. HvyAlcoholConsump
3. HighBP
4. GenHlth
5. HighChol
6. BMI
7. Age
8. HeartDiseaseorAttack
9. Sex
10. Stroke

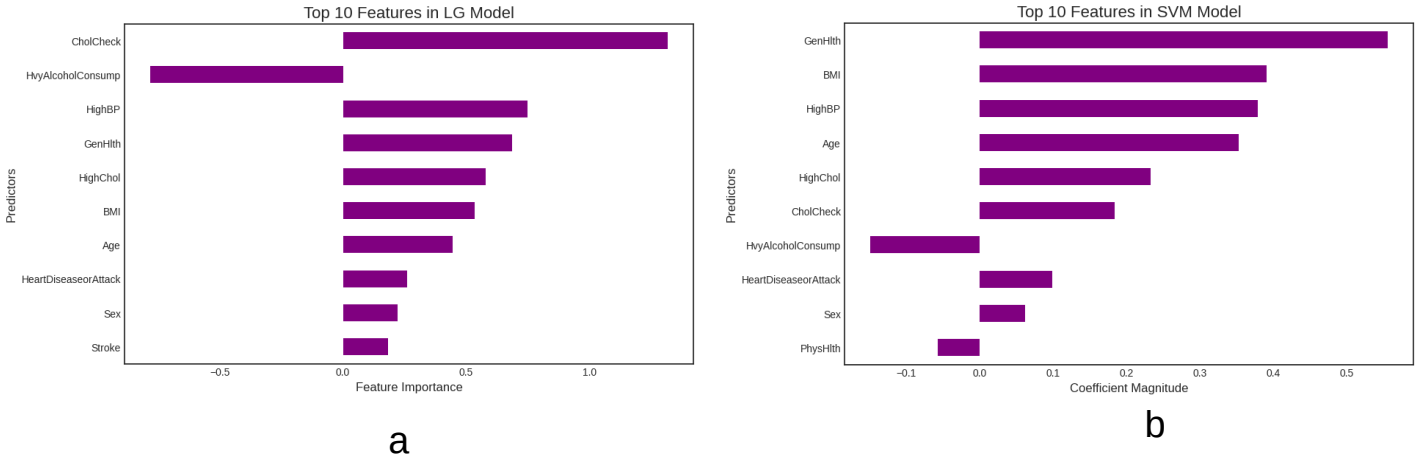


Figure 3.3. Top 10 predictors of diabetes: (a) in the logistic regression model, (b) Support Vector Machines Model

Likewise, the top 10 features based on the absolute feature importance coefficients for support vector machine with cost parameter set to 10 are listed below and displayed above in *fig 3.3(b)*:

1. GenHlth
2. BMI
3. HighBP
4. Age
5. HighChol
6. CholCheck
7. HvyAlcoholConsump
8. HeartDiseaseorAttack
9. Sex
10. PhysHlth

Some of the top 10 important features in predicting the response variable, i.e. whether a patient will have diabetes, in both of the models are common. HvyAlchoholConsump, Age, Sex, HighChol, and HighBP are some of them. Due to the presence of some lifestyle and demographic factors as important features in both of the models, these factors are studied independently in the following steps.

Before independently studying each factor, it is necessary to address why both the models have the same accuracy, 0.746446. It is important to ensure that the accuracy is not the outcome of the flaw present in the data which will jeopardize further studies. Even if the data distribution in the response variable is normal, sometimes the test set can be such that it contains 90% of negative cases, 0s. In such cases, even if a model predicts 0 all the time, it can still be 90% accurate. In such a case, the accuracy calculated will be the reflection of the data anomalies rather than that of a specific model. So cross-validation tests were performed in both of the models.

3.3.1. Cross-validation: Logistic regression

Cross-validation was run in the logistic regression model using 10 folds, and the scoring metric used was accuracy. The accuracy scores for 10 different folds are listed below with the mean and standard deviation of the scores in the validations sets.

Accuracy Scores across folds	0.74243281, 0.73564356, 0.75173292, 0.74352808, 0.74338662, 0.74918659, 0.75201584, 0.75201584, 0.74777196, 0.75343047
Mean Accuracy Score	0.7471144705562447
Standard Deviation of Accuracy Scores	0.005424513679856111

Table 3.3.1: Cross-validation results of the logistic regression model

The accuracy scores listed above in *Table 3.3.1* are consistent across all folds. It means that the logistic regression model generalizes well across different subsets of data. Overall the model performs with 74.71 % accuracy. In addition, the lower variance of accuracy scores in different folds verifies that the model is not sensitive or overfitted to a specific subset of data from the dataset.

3.3.2. Cross-validation: Tune SVM (c= 10)

Cross-validation was similarly performed with the tuned SVM as in logistic regression detailed in section 3.3.1. The results of the cross-validation sets are listed in the following table *3.3.2.*

Accuracy Scores across folds	0.7397454, 0.73705799, 0.75187438, 0.74084029, 0.7464988, 0.74678172, 0.75286462, 0.75130853 , 0.74890366, 0.75357193
Mean Accuracy Score	0.7469447332947429
Standard Deviation of Accuracy Scores	0.005596553417096679

Table 3.3.2: Cross-validation results of the tuned SVM model

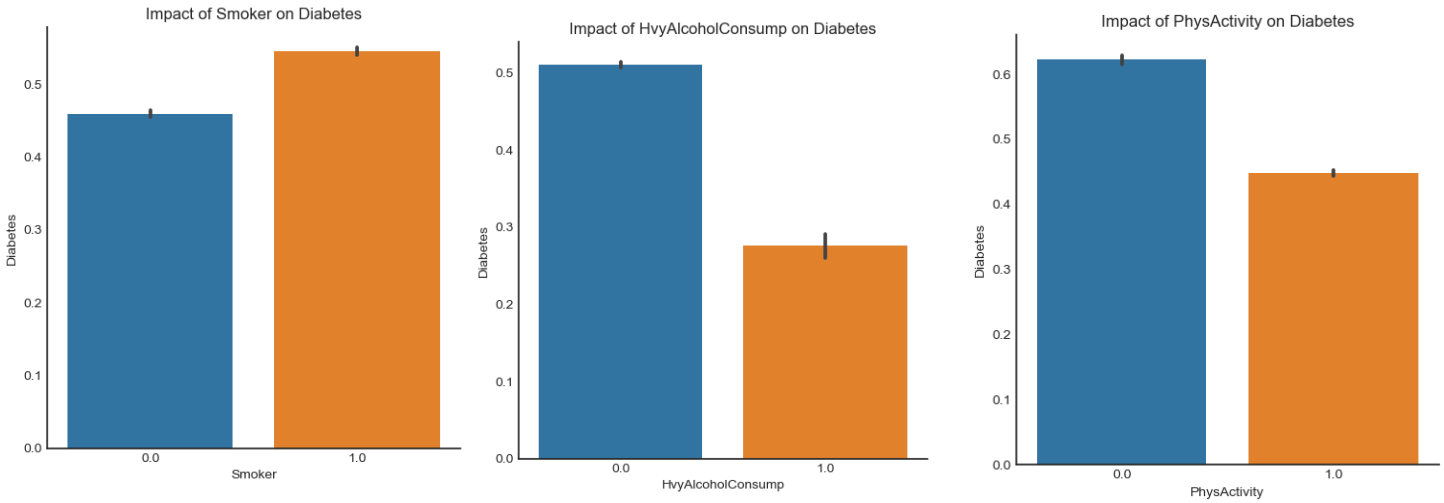
The accuracy scores listed above in *Table 3.3.2* are consistent across all folds. It means that the SVM model generalizes well across different subsets of data as did the logistic regression model. Overall the model performs with 74.69 % accuracy which is slightly less than the mean score in the logistic regression model in 3.3.1. In addition, the lower variance of accuracy scores in different folds verifies that the model is not sensitive or overfitted to a specific subset of data from the dataset.

It can also be concluded from the different accuracy scores between logistic regression and the SVM model that the accuracy scores are not just the characteristics of the

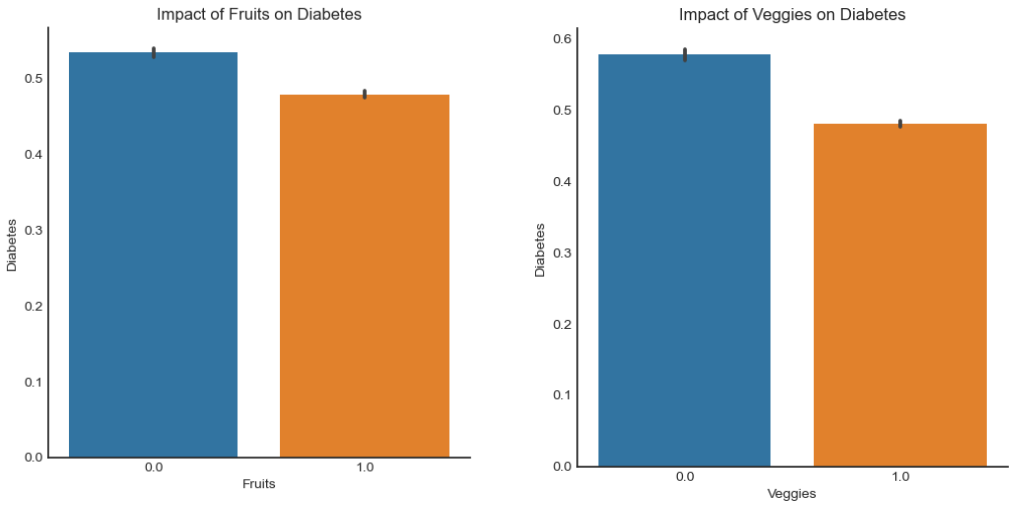
dataset but of the model itself. So the impact of lifestyle factors or demographic variables on the likelihood of diabetes can be carried out confidently.

3.4. Impact of lifestyle factors on the likelihood of Diabetes

The lifestyle factors present in the dataset are 'Smoker', 'HvyAlcoholConsump', 'PhysActivity', 'Fruits', and 'Veggies'. What these individual variables measure can be found in the introduction above. Before fitting a model to the subset of the data based on these factors, all these variables were visualized to study their presence among diabetic and nondiabetic patients. The results are shown in *Fig 3.4, a -e.*



a b c



d e

Figure 3.4 (a-e): General impact of lifestyle factors on the mean diabetic score of the population. 0 = No, 1 = Yes in all of the x-axis labels.

As shown in Figure 3.4, the population of smokers is more diabetic (a), physically passive people are more diabetic (b), and fruits and vegetables consumers are less diabetic (d, e). Interestingly, heavy alcohol consumers were found to be less diabetic in

our dataset (fig 3.4 b). Furthermore, the HvyAlcoholConsum feature had negative coefficients in both the logistic and SVM models (fig 3.3 (a,b)). The American Diabetes Association mentions that "a daily cocktail or two may improve blood glucose (blood sugar) management and insulin sensitivity"[4].

After the initial data visualization, logistic regression was performed. Then the fitted model was converted to the statsmodels format to obtain a summary using the summary2 method. The summary of the model is showcased in the *fig 3.4 f.* As presented in the model summary, all the p-values for the variables are less than 0.001, it can be concluded that they are significant in predicting the likelihood of diabetes in a person. The model was able to converge after 34 iterations.

Heavy alcohol consumption, physical activity, fruits, and veggies have negative coefficients, meaning their increase in magnitude decreases the log odds of diabetes.

Results: Logit

Model:	Logit	Pseudo R-squared:	0.032
Dependent Variable:	Diabetes	AIC:	75888.3619
Date:	2023-12-14 00:54	BIC:	75942.0195
No. Observations:	56553	Log-Likelihood:	-37938.
Df Model:	5	LL-Null:	-39200.
Df Residuals:	56547	LLR p-value:	0.0000
Converged:	1.0000	Scale:	1.0000
No. Iterations:	34.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
const	0.6079	0.0249	24.3894	0.0000	0.5590	0.6567
Smoker	0.3257	0.0173	18.7786	0.0000	0.2917	0.3597
HvyAlcoholConsump	-1.0753	0.0469	-22.9420	0.0000	-1.1672	-0.9834
PhysActivity	-0.6280	0.0193	-32.5130	0.0000	-0.6659	-0.5902
Fruits	-0.0943	0.0183	-5.1509	0.0000	-0.1302	-0.0584
Veggies	-0.2714	0.0220	-12.3624	0.0000	-0.3144	-0.2283

Figure 3.4 (f): Summary of the logistic regression model fitted on the lifestyle factors

It should be noted that the pseudo R-squared is just 0.032 which accounts for negligible variability in the dataset. In addition, the accuracy of the model fitted just in the lifestyle factors dropped to 58.4 %. When all predictors were used, the accuracy of the logistic regression was more than 70%. So we can not necessarily isolate other health conditions like stroke, mental health, physical health, cholesterol, and so on that might arise with respect to lifestyle factors and additionally accelerate diabetes. Besides, most of the measurements in the lifestyle factors columns were subjective in nature and can not necessarily be trusted in their entirety. Demographic features that will be discussed in the 3.5 section are, however, objective in nature and might be more accurate as well in accounting for the likelihood of diabetes.

3.5. Impact of demographic factors on the likelihood of Diabetes

Diabetes is closely linked to factors like age, sex, and Body Mass Index (BMI). Age-wise, the risk of developing diabetes increases with advancing years, largely due to lifestyle and physiological changes. Regarding sex, men are generally more susceptible to diabetes compared to women. BMI is a critical factor; higher BMI, particularly with central obesity, significantly escalates the risk of insulin resistance, a precursor to diabetes. These factors, often interrelated, highlight the importance of tailored prevention and management strategies in addressing the global diabetes challenge.

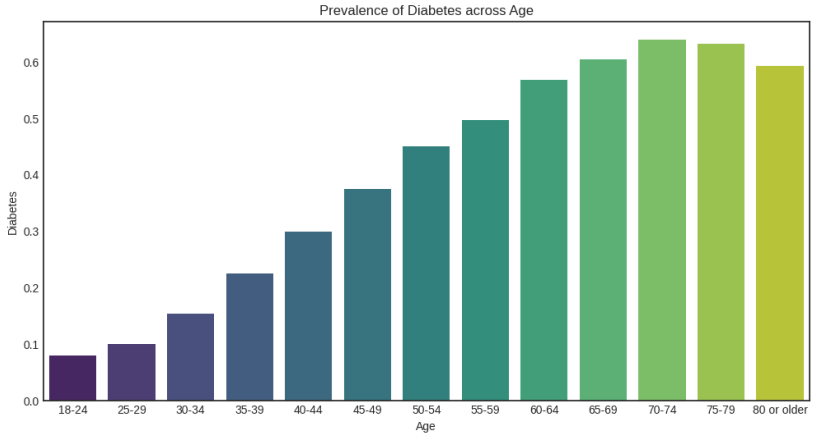


Figure 3.5(a): Increasing prevalence of diabetes with age, from ages 18 to 80 and older.

Figure 3.5(a) presents a bar chart from the dataset we used, which visualizes the prevalence of diabetes across different age categories. The data reveal a clear age-related gradient, with the lowest prevalence in the 18-24 age group and a steady increase through to the 80 and older group. This pattern underscores the impact of age as a significant factor in the incidence of diabetes, as illustrated by the dataset.

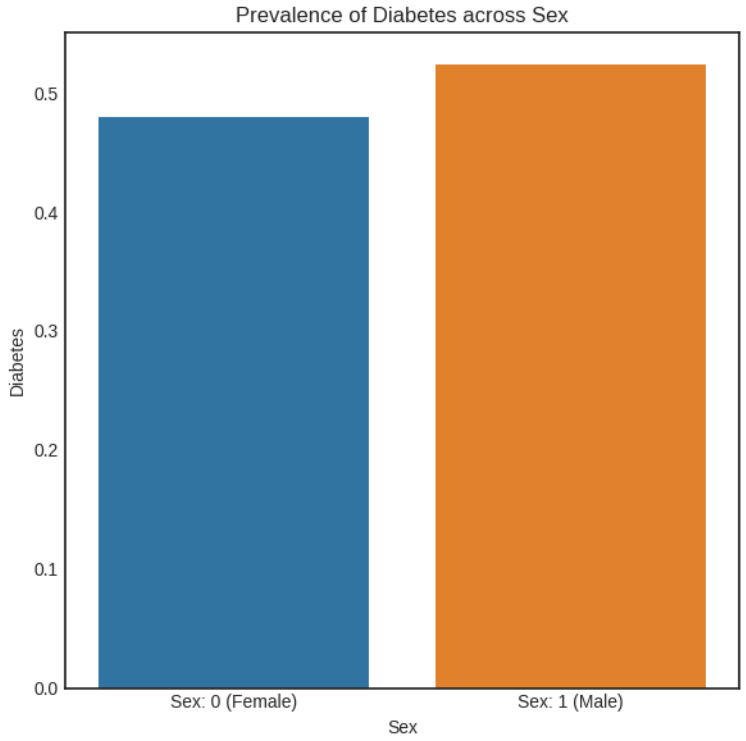


Figure 3.5(b): Bar graph showing the prevalence of diabetes by sex

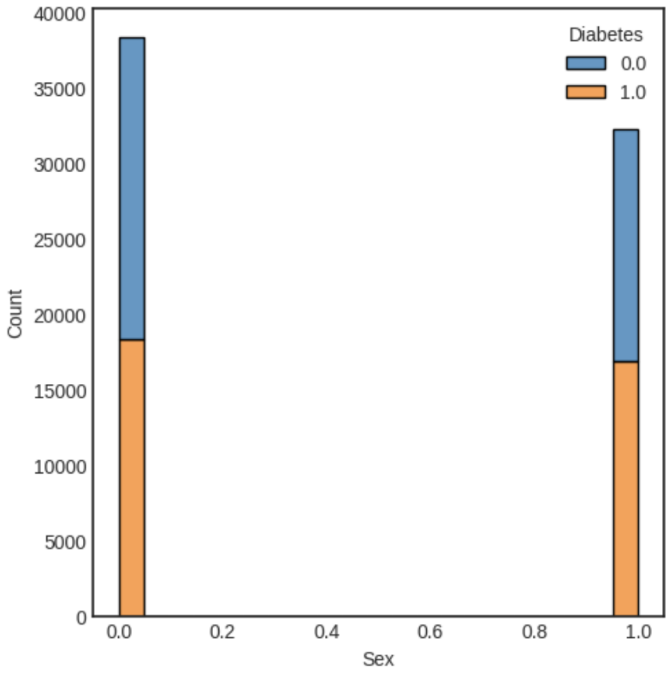


Figure 3.5(c): Stacked bar chart representing the counts of individuals with and without diabetes categorized by sex.

Figure 3.5(b) displays a bar chart comparing the prevalence of diabetes between females (Sex: 0) and males (Sex: 1). The height of each bar corresponds to the proportion of individuals within each sex who have diabetes. The chart indicates that males have a higher prevalence of diabetes than females within this specific dataset.

Figure 3.5(c) is a stacked bar chart, which breaks down the count of individuals by sex and their diabetes status. Each bar is divided into two segments: the lower segment (orange) represents the count of individuals with diabetes, and the upper segment (blue) represents those without. This visualization shows not only the total count of individuals within each sex but also the proportion of those who have been diagnosed with diabetes. From this chart, we can observe the distribution of diabetes across sexes and note that while the total counts may be similar, the proportion with diabetes is visually distinct, potentially corroborating the findings in Figure 3.5(b) regarding prevalence by Male Group.

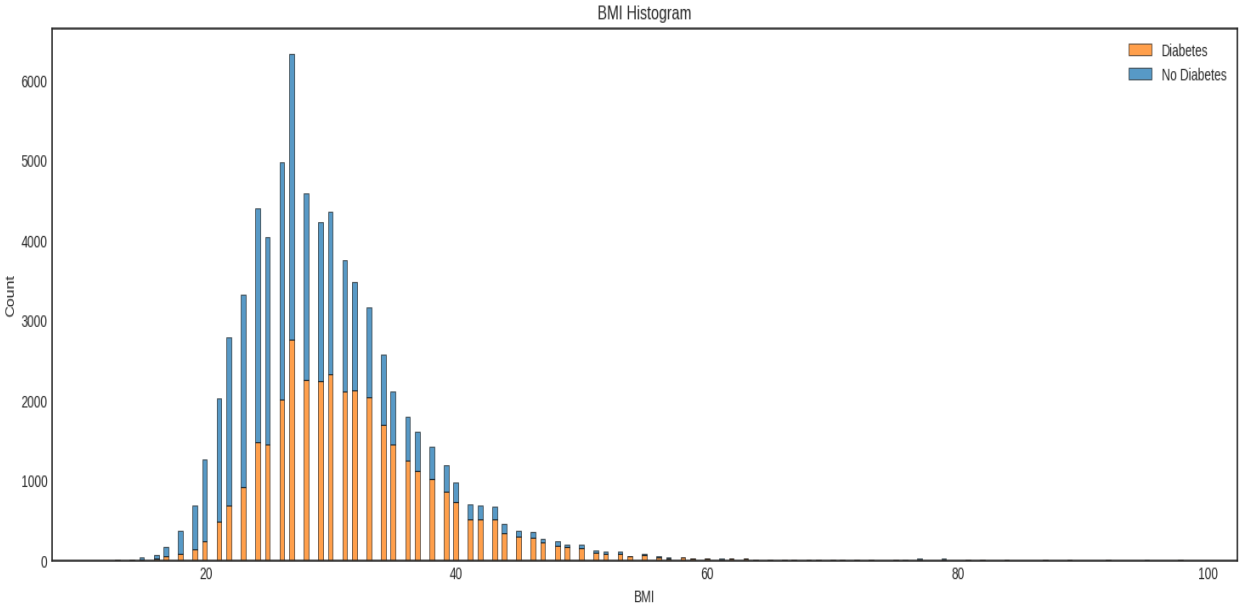


Figure 3.5(d): Histogram comparing the BMI distribution of individuals with and without diabetes.

The histogram in *Figure 3.5(d)* illustrates the distribution of Body Mass Index (BMI) among individuals, segmented into those with diabetes (orange) and those without diabetes (blue). The counts represent the number of individuals falling within various BMI ranges. A noticeable overlap occurs across the BMI spectrum, yet the distribution suggests a higher concentration of individuals with diabetes in higher BMI ranges. This visualization from the dataset supports the established link between increased BMI and the risk of diabetes.

The figures from the dataset collectively contribute to a multifaceted understanding of the epidemiology of diabetes. *Figure 3.5(a)* shows a clear trend of increasing prevalence of diabetes with age, indicating that older age groups are more affected. *Figure 3.5(b)* reveals a higher prevalence of diabetes in males compared to females, suggesting possible sex-related differences in diabetes risk or diagnosis rates. Figure 3.5(c) confirms this by showing both the total counts and the proportions of males and females with diabetes, highlighting a significant representation of diabetes within each sex. Lastly, Figure 3.5(d) which is a histogram of BMI distribution that correlates higher BMI with a greater incidence of diabetes, reinforcing the notion that BMI is a strong risk factor for the condition.

Given the insights provided by these figures, we plan to evaluate the significance of age, sex, and BMI as predictors in our logistic regression model. By selecting these variables, which have shown a quantifiable impact on the prevalence of diabetes, we will assess their importance and influence as independent variables. This analysis will

allow us to determine how critical these factors are in predicting the occurrence of diabetes, facilitating more informed decision-making for clinical assessments and public health policies.

Building upon the initial analysis of demographic factors, our logistic regression classifier yielded an accuracy of **0.6871** when predicting diabetes using age, sex, and BMI as predictors. This result, while not exhaustive, indicates a moderate level of predictive power for these factors. It confirms that age, sex, and BMI are indeed relevant considerations in the context of diabetes risk. However, the accuracy also suggests that there may be other variables that could improve the model's predictive performance.

```
Optimization terminated successfully    (Exit mode 0)
            Current function value: 0.5960195360656187
            Iterations: 17
            Function evaluations: 17
            Gradient evaluations: 17
                          Results: Logit
==================================================================
Model:              Logit            Method:           MLE
Dependent Variable: Diabetes         Pseudo R-squared: 0.140
Date:               2023-12-18 03:17 AIC:              67419.3856
No. Observations:   56553            BIC:              67446.2144
Df Model:           2                Log-Likelihood:   -33707.
Df Residuals:       56550            LL-Null:          -39200.
Converged:          1.0000           LLR p-value:      0.0000
No. Iterations:     17.0000          Scale:            1.0000
------------------------------------------------------------------
         Coef.     Std.Err.       z       P>|z|     [0.025    0.975]
------------------------------------------------------------------
Age      0.7109      0.0103    69.0243    0.0000    0.6907    0.7311
Sex      0.1083      0.0136     7.9563    0.0000    0.0817    0.1350
BMI      0.8176      0.0116    70.5546    0.0000    0.7949    0.8403
==================================================================
```

The results from the logistic regression analysis, which utilized age, sex, and BMI as predictors for diabetes, have provided quantitative insights. The model, which converged after 17 iterations, indicated that all three predictors were statistically significant, as evidenced by their p-values being less than 0.05. Specifically:

- Age: With a coefficient of 0.7109, age was shown to be a significant predictor, suggesting that the likelihood of developing diabetes increased with age.
- Sex: The coefficient for sex was 0.1083, indicating a measurable, albeit less pronounced, impact of sex on the risk of diabetes compared to age and BMI.
- BMI: The coefficient for BMI was 0.8176, indicating a strong predictive value of BMI for diabetes, aligning with the understanding that a higher BMI is associated with a higher risk of the condition.

The model's pseudo R-squared value was 0.140. Although not indicative of a high level of variance explanation, it nonetheless signifies a meaningful contribution of these predictors in the context of logistic regression. These results led to the conclusion that age and BMI, in particular, were strong indicators of diabetes, positioning the model as a solid basis for understanding the relative importance of these predictors in assessing the risk of diabetes.

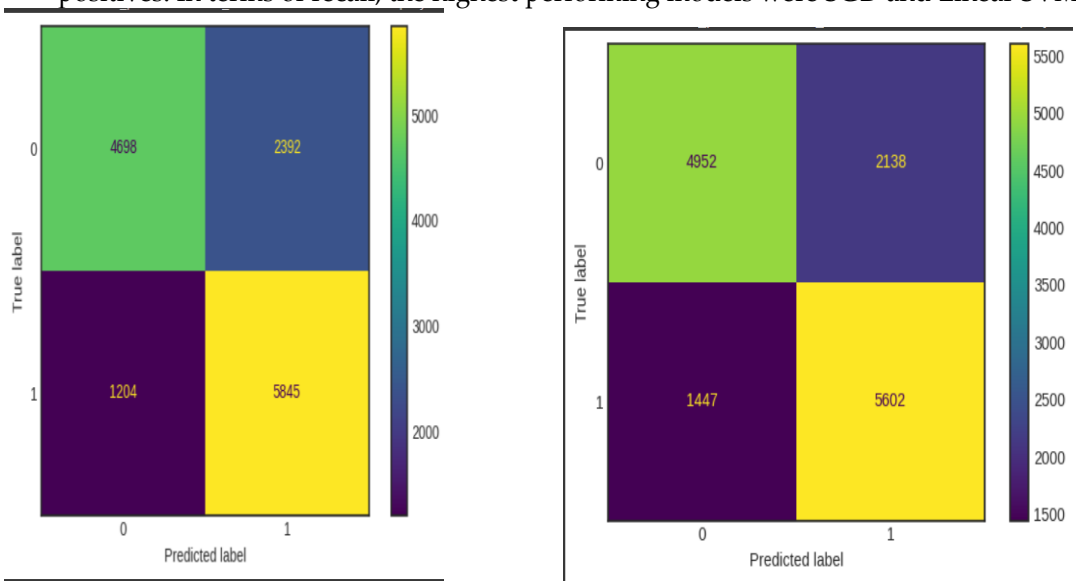
The logistic regression analysis of the dataset has confirmed the significant roles of age, sex, and BMI in predicting the likelihood of diabetes. A doctor could use these findings in several practical ways such as:

- **Risk Assessment**: The coefficients provide a quantifiable measure of how each factor increases the risk of diabetes. For instance, the strong coefficient for BMI could help a doctor prioritize weight management in patients' care plans.
- **Personalized Medicine**: Understanding the significance of these predictors enables a doctor to tailor individual patient screenings and interventions, potentially identifying at-risk patients earlier.
- **Educational Tool**: The model can be used as an educational tool for patients to understand their risk factors and the importance of lifestyle modifications.

In conclusion, this model, with its focus on age, sex, and BMI, provides a statistically grounded method for a doctor to estimate diabetes risk, thereby aiding in the early detection and proactive management of this condition.

4. Discussion

Predicting Diabetes is a sensitive medical issue, and it is imperative that the false negatives (diabetic patients being misclassified as healthy) be as minimal as possible. So, models with the highest recall scores work the best in context of our data. High recall scores would mean a high true positive rate, even if it comes at the expense of some false positives. In terms of recall, the highest performing models were SGD and Linear SVM.



a **b**

Figure 4.1. (a) Confusion matrix for SGD model; (b) Confusion matrix for Linear SVM model.

From the figure 4.1, we can see that our SDG misclassified 1204 rows as false negatives and our Linear SVM model misclassified 1447 out of 14,139 test data. While a low false negative definitely helps in the context of our data, a good model should have both low false negatives and low false positives. In the case of our data, we also do not want to prescribe a diabetes medication to a healthy individual (false positives) if our model were to be used in a real life setting. F1 score is a harmonic mean between precision and recall, and prioritizing F1 score is a decent trade-off.

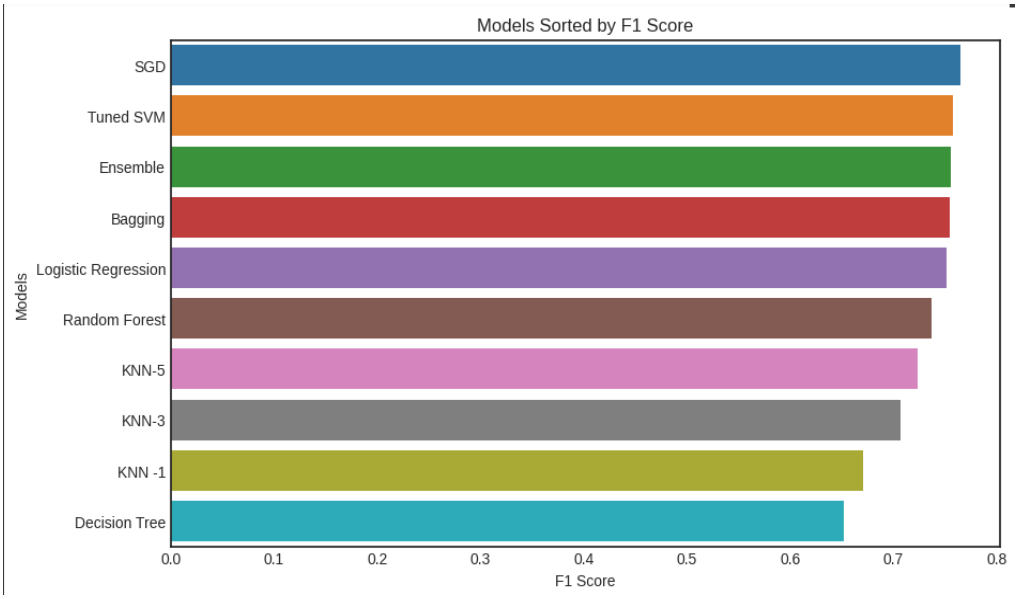


Figure 4.2. Different classifiers used to predict diabetes, sorted by their F1 score.

Considering the F1 score, SGD and Linear SVM still take the top two positions as the best performing models. The Area Under The Curve value of 0.75 for ROC Curve of the SGD model suggests that it is fairly decent at distinguishing diabetic and non-diabetic patients.

It is also important to acknowledge the limitations of the data that was used to train the models discussed above. First of all, the data is not up-to-date which might raise the question of the relevancy of the conclusions made in the contemporary situations. Overall, it can be said that the models were not sensitive to a specific subset of the data or overfitting to a subset and underperforming to new test sets as validated by the cross validation evaluation in the sections 3.3.1 and 3.3.2.

Some of the top-10 important features of the both models were lifestyle and demographic factors which lead to the study of diabetes with respect to the lifestyle factors. The accuracy of the model fitted just in the lifestyle factors dropped to 58.4 %. When all predictors were used, the accuracy of the logistic regression was more than 70%. So we can not necessarily isolate other health conditions like stroke, mental health, physical health, cholesterol, and so on that might arise with respect to lifestyle factors and additionally accelerate diabetes. However, the accuracy of the model fitted on demographic factors was 68.7 %. It can be said that demographic factors might aid in the early detection and proactive management of diabetes.

5. Conclusions

In practical healthcare settings, where time is often limited, doctors can benefit significantly from focusing on key demographic factors like age, sex, and BMI for a quick preliminary assessment of diabetes risk. Our study underscores the importance of these factors in predicting diabetes, offering a streamlined approach for healthcare providers.

Moreover, our best-performing model, the Stochastic Gradient Descent (SGD), is particularly effective in minimizing false negatives. This high recall aspect of the SGD model ensures that fewer diabetic patients are misclassified as healthy, which is crucial in medical diagnostics. By employing this model, doctors can more reliably identify patients at risk of diabetes, leading to timely interventions and better management of the disease.

Thus, combining the quick assessment using demographic factors with the high recall capabilities of our SGD model presents a robust strategy for healthcare practitioners. This approach not only saves time but also enhances the accuracy of diabetes diagnosis, potentially improving patient outcomes in the fast-paced environment of medical practice.

Our study opens the door to many new ways to improve health care, from using technology to better understand and manage diabetes, to guiding policies that keep people healthier.

Data Availability Statement: Link to the underlying survey from the CDC – 2015 BRFSS Data and Documentation: https://www.cdc.gov/brfss/annual_data/annual_2015.html

Link to the cleaned dataset from Kaggle:
https://www.kaggle.com/datasets/prosperchuks/health-dataset/data

Overview of the CDC BRFSS 2015:
https://www.cdc.gov/brfss/annual_data/2015/pdf/overview_2015.pdf

References

1. Tao, Ziqi, Aimin Shi, and Jing Zhao. "Epidemiological Perspectives of Diabetes." *Cell Biochemistry and Biophysics* 73 (2015): 181-185.
2. Beagley, Jessica, et al. "Global Estimates of Undiagnosed Diabetes in Adults." *Diabetes Research and Clinical Practice* 103.2 (2014): 150-160.
3. Bahia, Luciana R., et al. "The Costs of Type 2 Diabetes Mellitus Outpatient Care in the Brazilian Public Health System." Value in Health 14.5 (2011): S137-S140.
4. American Diabetes Association. "Alcohol and Diabetes." Alcohol and Diabetes. ADA. Accessed December 18, 2023. https://diabetes.org/health-wellness/alcohol-and-diabetes#:~:text=A%20daily%20cocktail%20or%20two,t%20mean%20you%20should%20start.