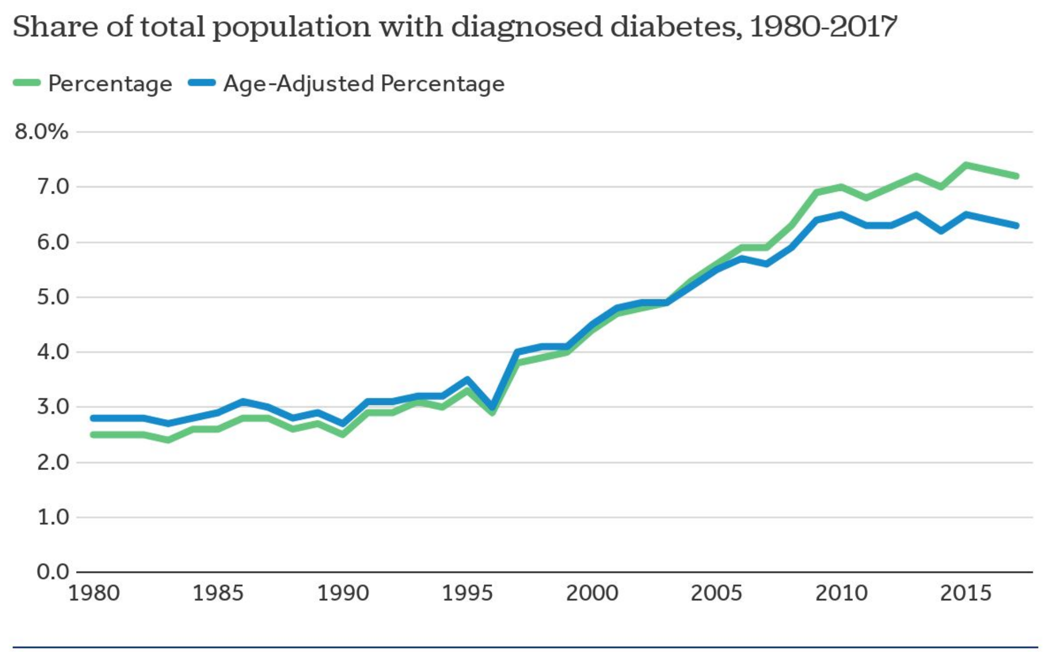
Diabetes Prediction using Medical History

Group 11 - Yunrui Jiang, Jonah Soong, Justin Sin, Anyin Huang, Zeyu Chang

Motivation

- Diabetes
  a. Chronic condition leading to life threatening issues
  b. Caused by increased resistance to insulin leading to high glucose levels
- 38.4 million people (11.6% of the US)
- 97.6 million adults have prediabetes (38.0% of US adults)



Share of total population with diagnosed diabetes, 1980-2017

Source: US Diabetes Surveillance System

Peterson-KFF
Health System Tracker

Objective

1. What are some key indicators for diabetes?

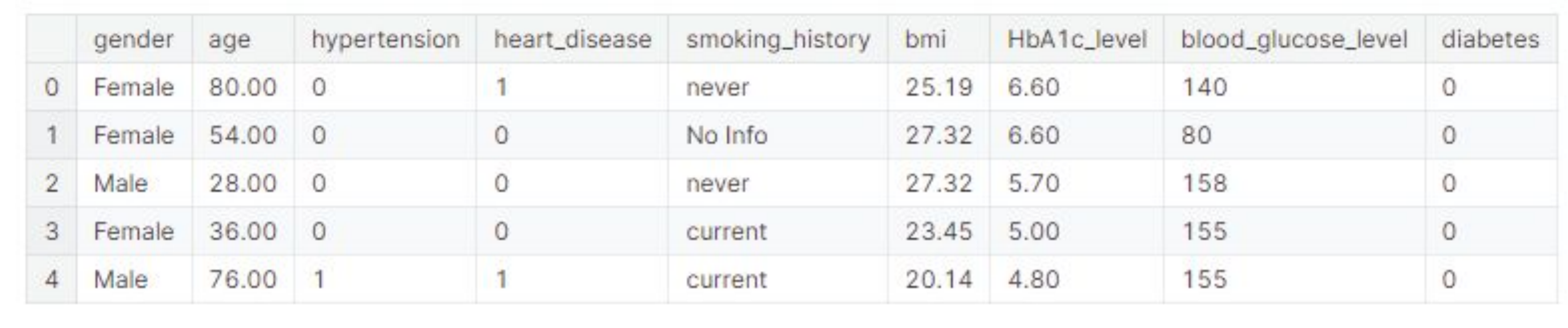
2. How well do existing prediction techniques forecast diabetes?

Methodology

- Pearson Correlation Coefficients were computed on different subsets of data as main heuristic for strong indicators
  - Categorical variables were binarized
  - Data was stratified to see if coefficients changed
- 5 different prediction ML models were tested for accuracy and precision

Datasets overview

(https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data)



	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.00	0	1	never	25.19	6.60	140	0
1	Female	54.00	0	0	No Info	27.32	6.60	80	0
2	Male	28.00	0	0	never	27.32	5.70	158	0
3	Female	36.00	0	0	current	23.45	5.00	155	0
4	Male	76.00	1	1	current	20.14	4.80	155	0

Dataset attributes

Age

Gender

Diabetes

Heart Disease

BMI (Body mass index)

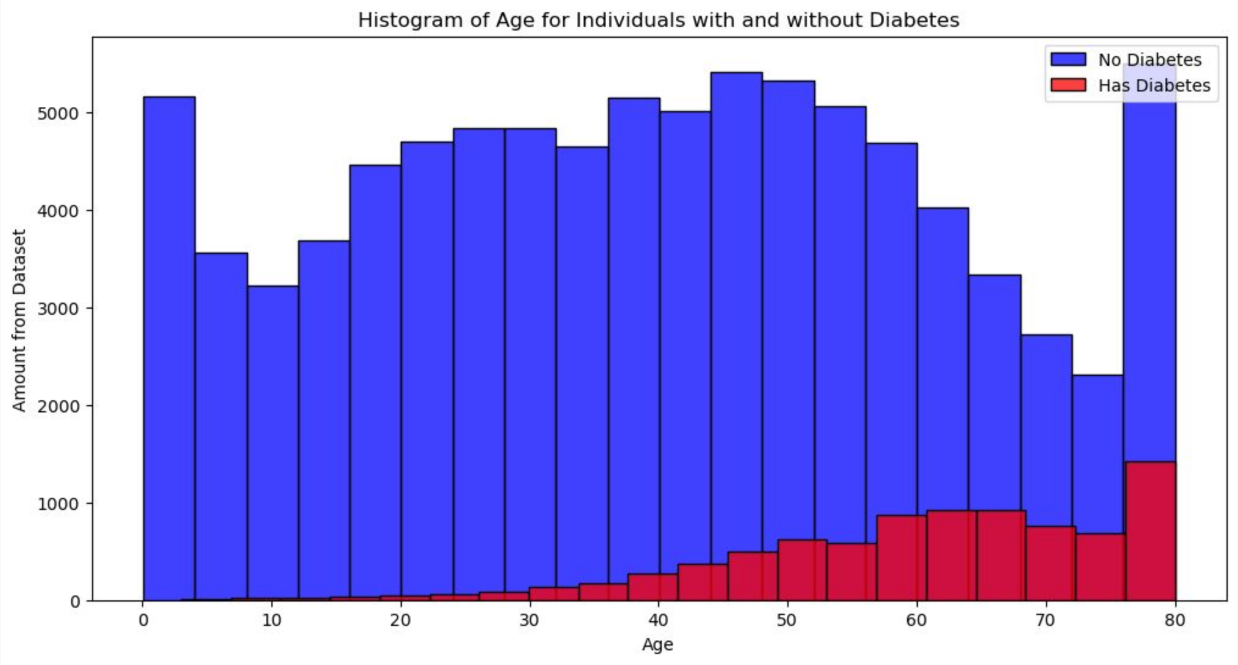
Hypertension (high blood pressure)

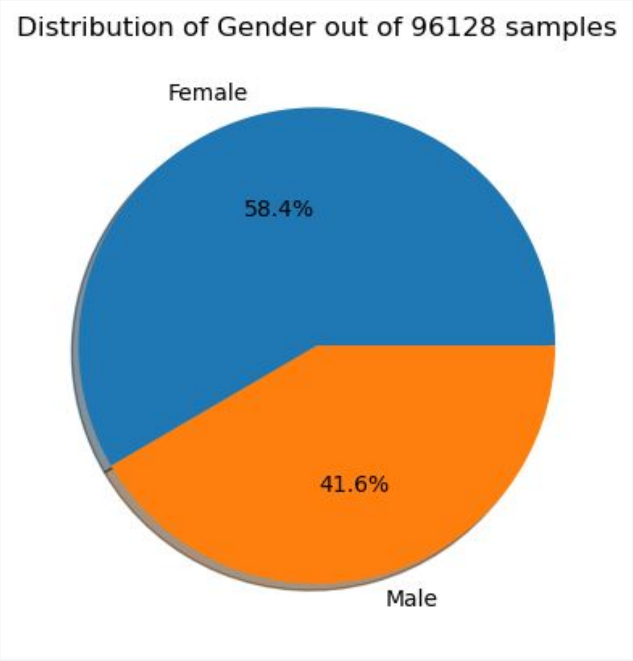
Smoking history

Blood glucose level

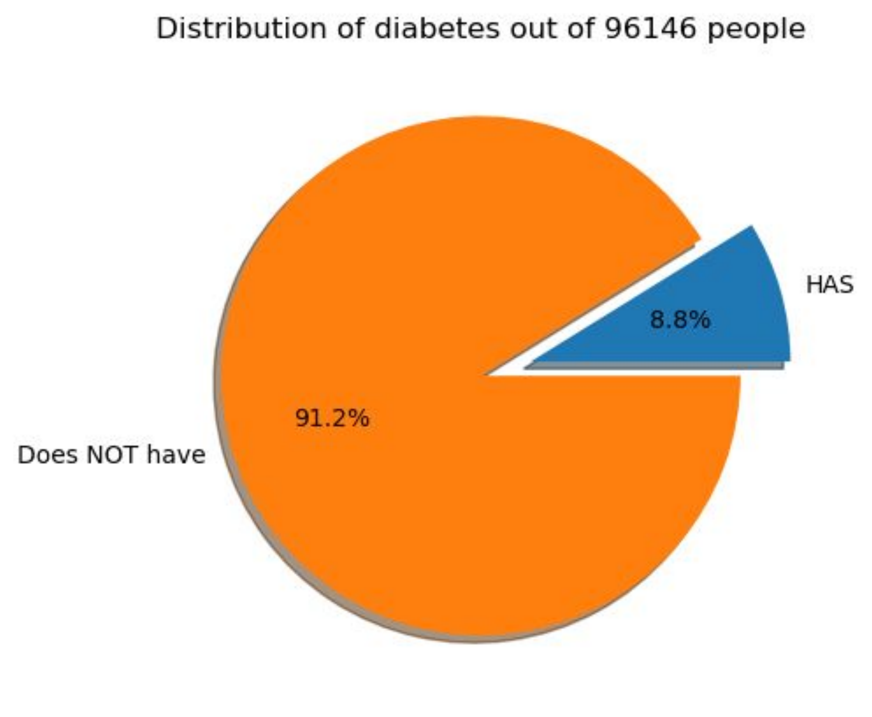
HbA1c level (average blood glucose level over the past 2-3 months)

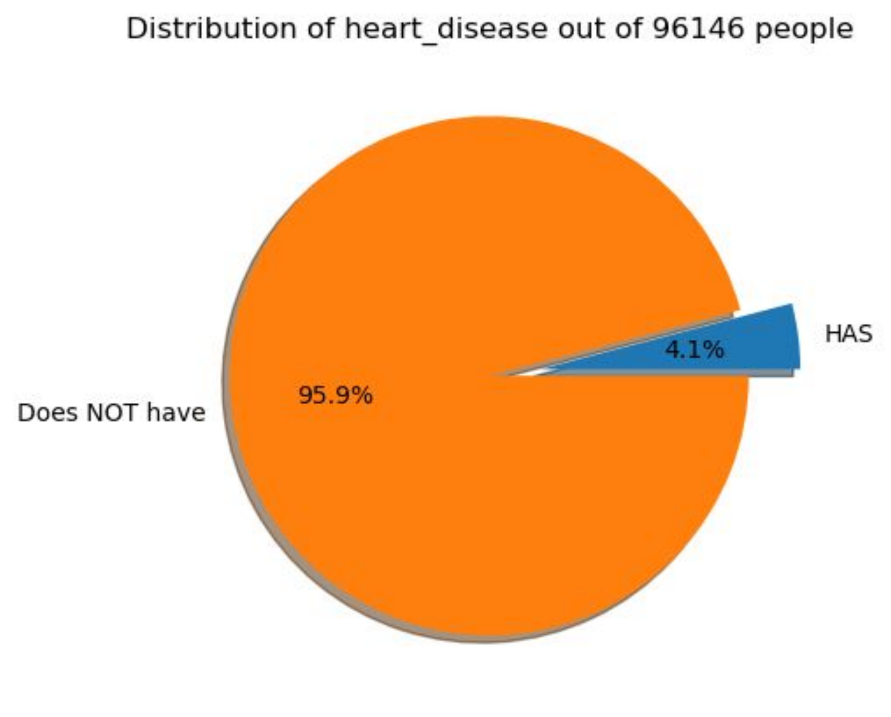
Age and Gender





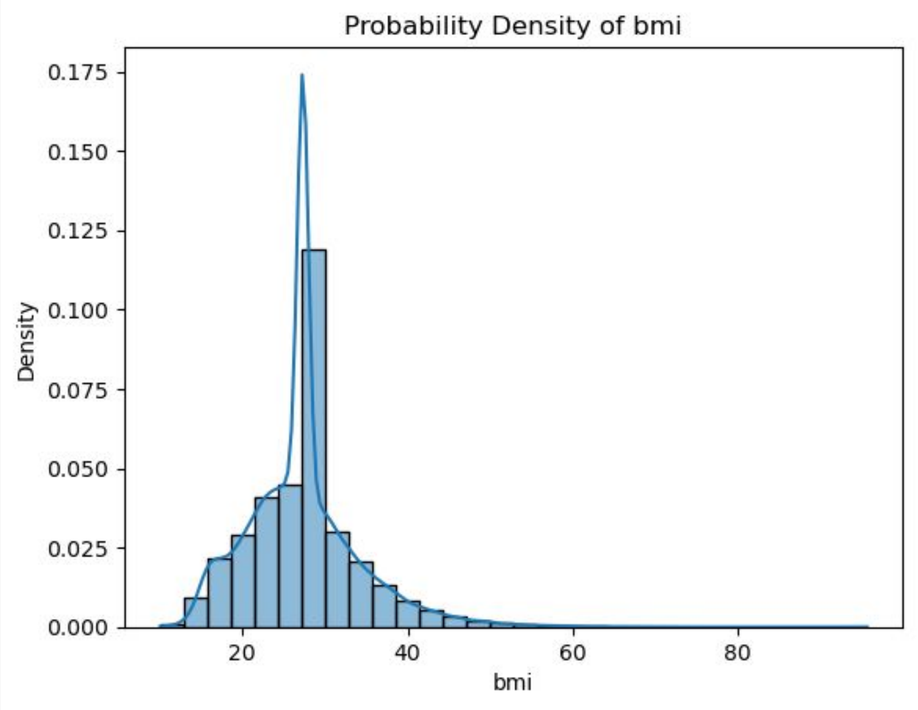
Diabetes and Heart Disease

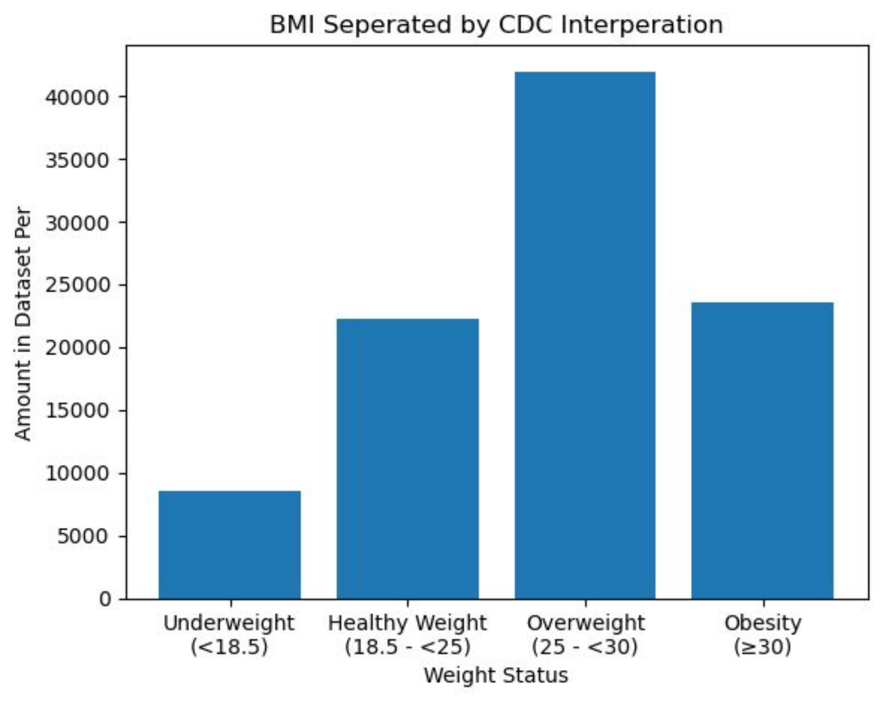




BMI

Body Mass Index - kg/m^2

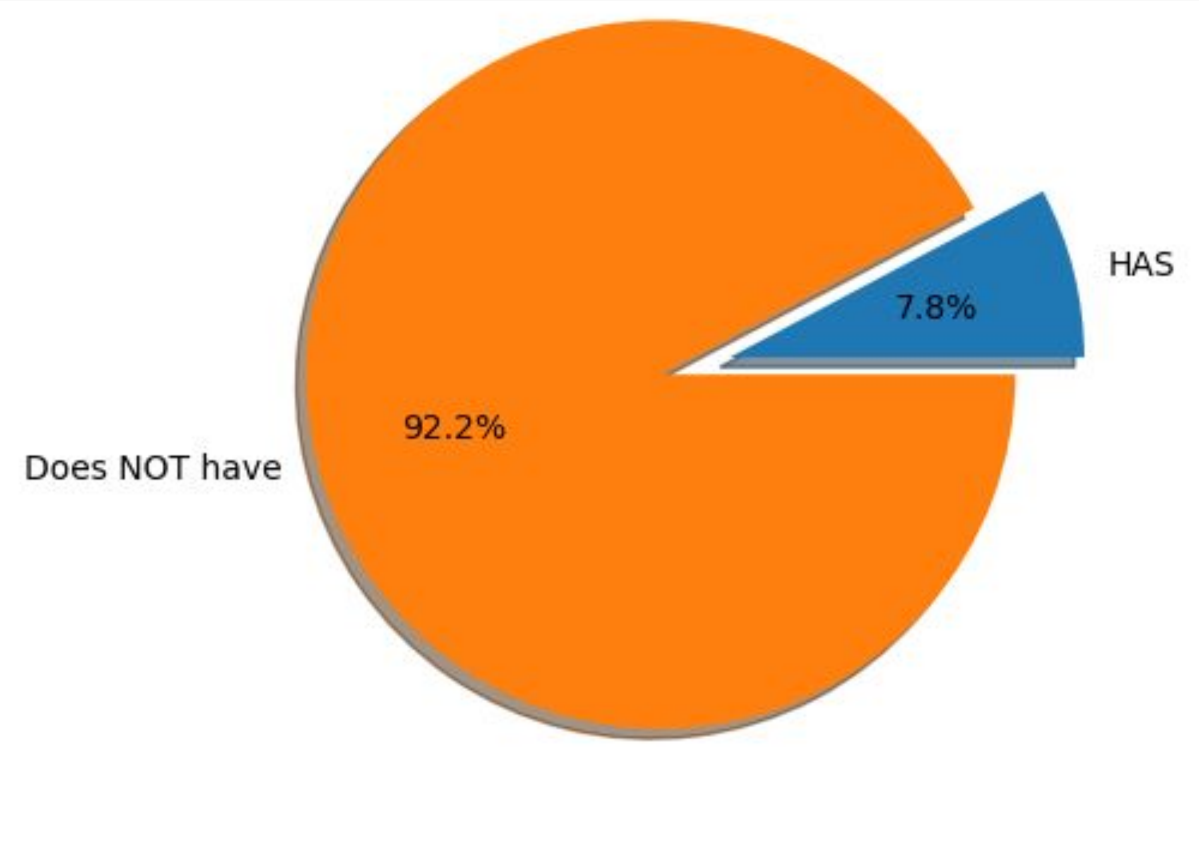




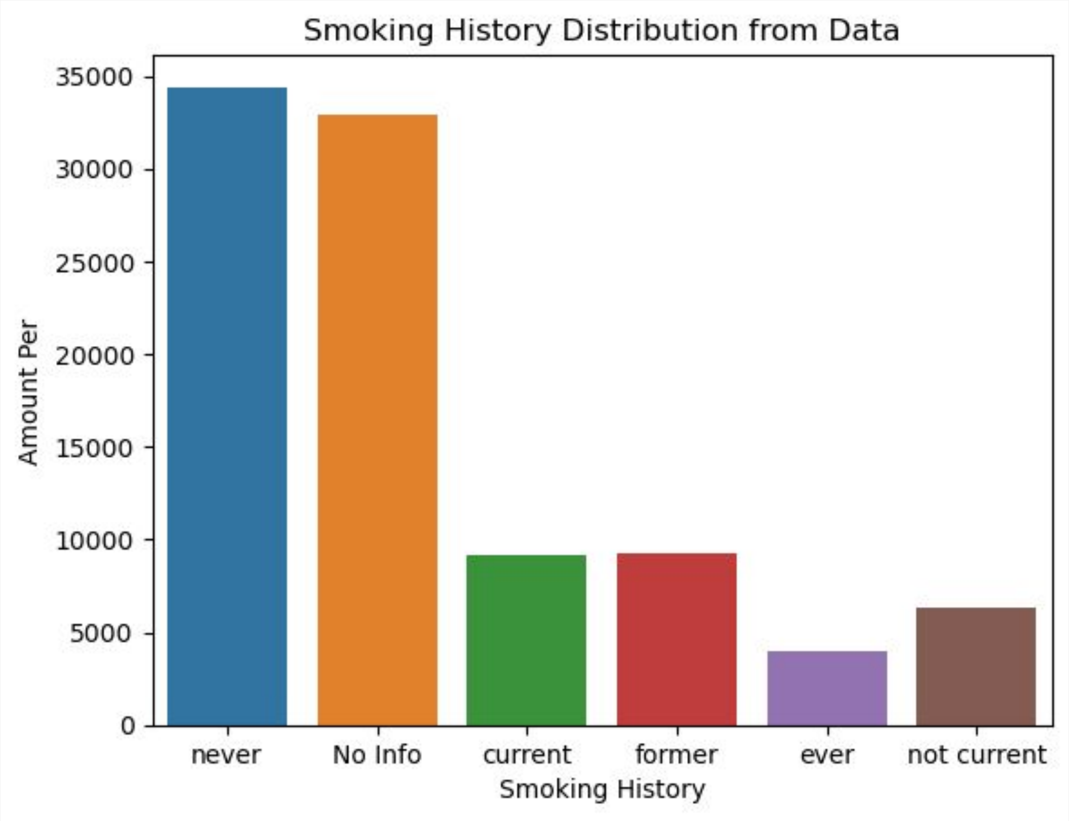
https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html

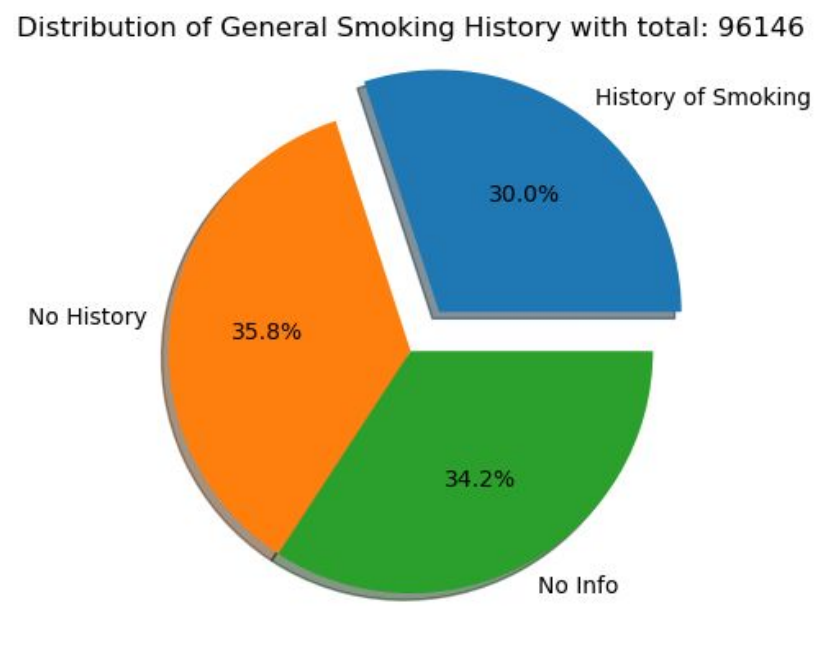
Hypertension

High Blood Pressure



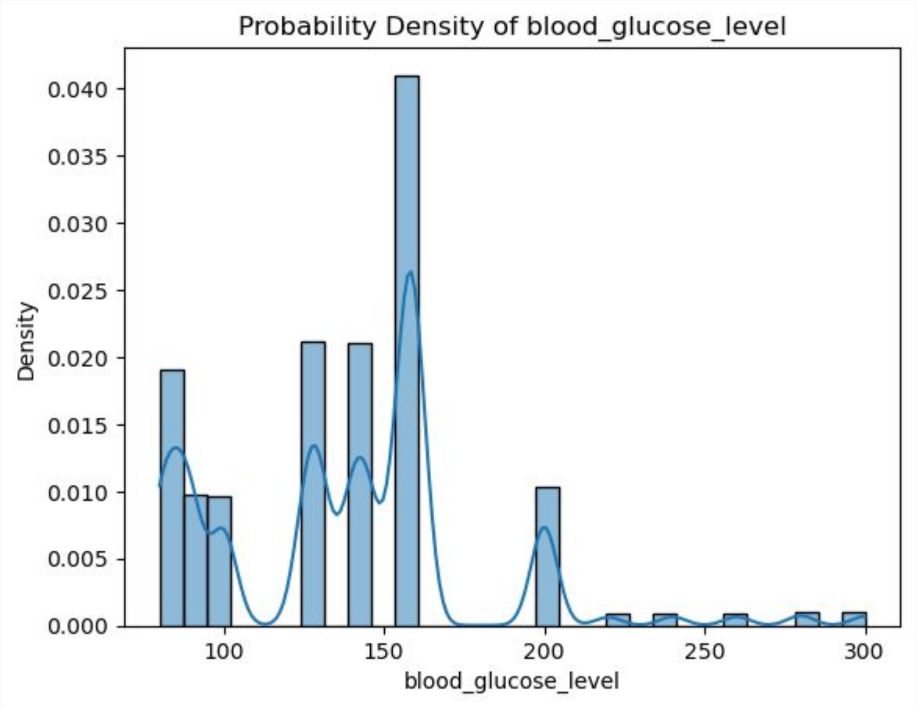
Smoking History

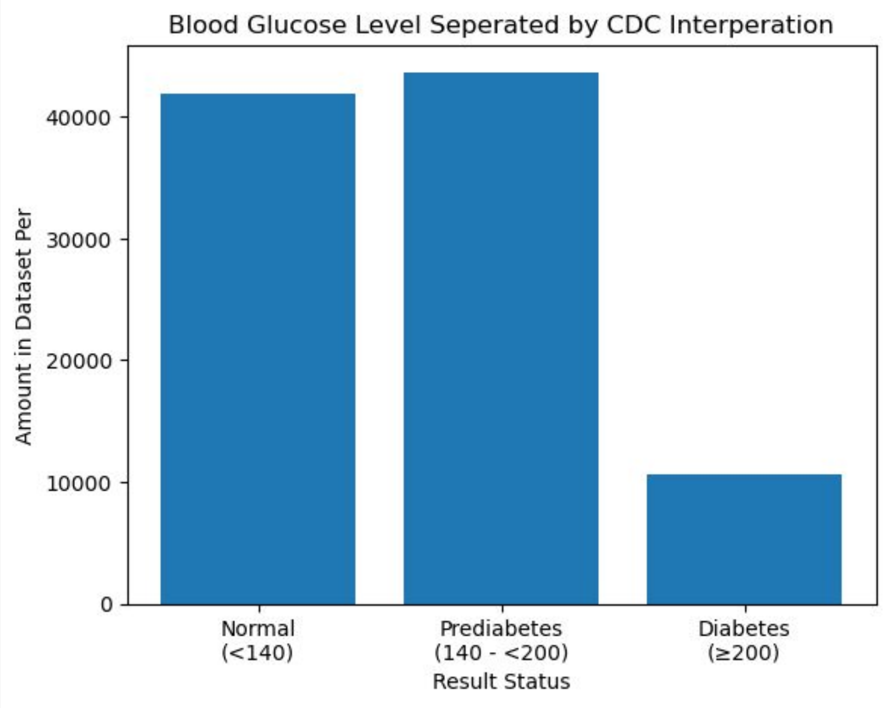




Blood Glucose

glucose in bloodstream - mg/dl

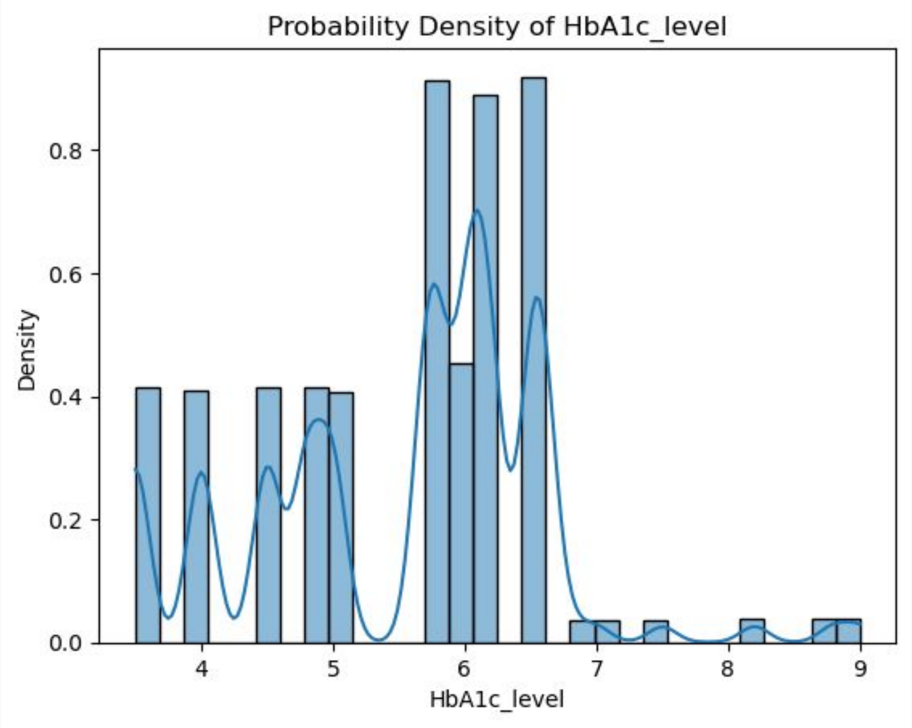


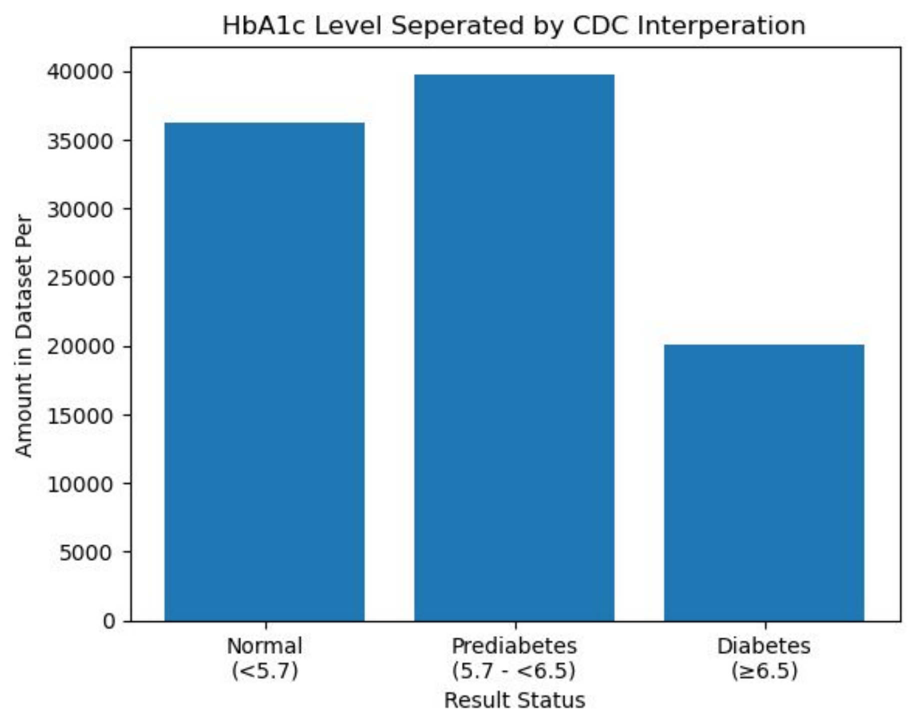


https://www.cdc.gov/diabetes/basics/getting-tested.html

HbA1c

fraction of hemoglobin(blood protein) with glucose attached - %



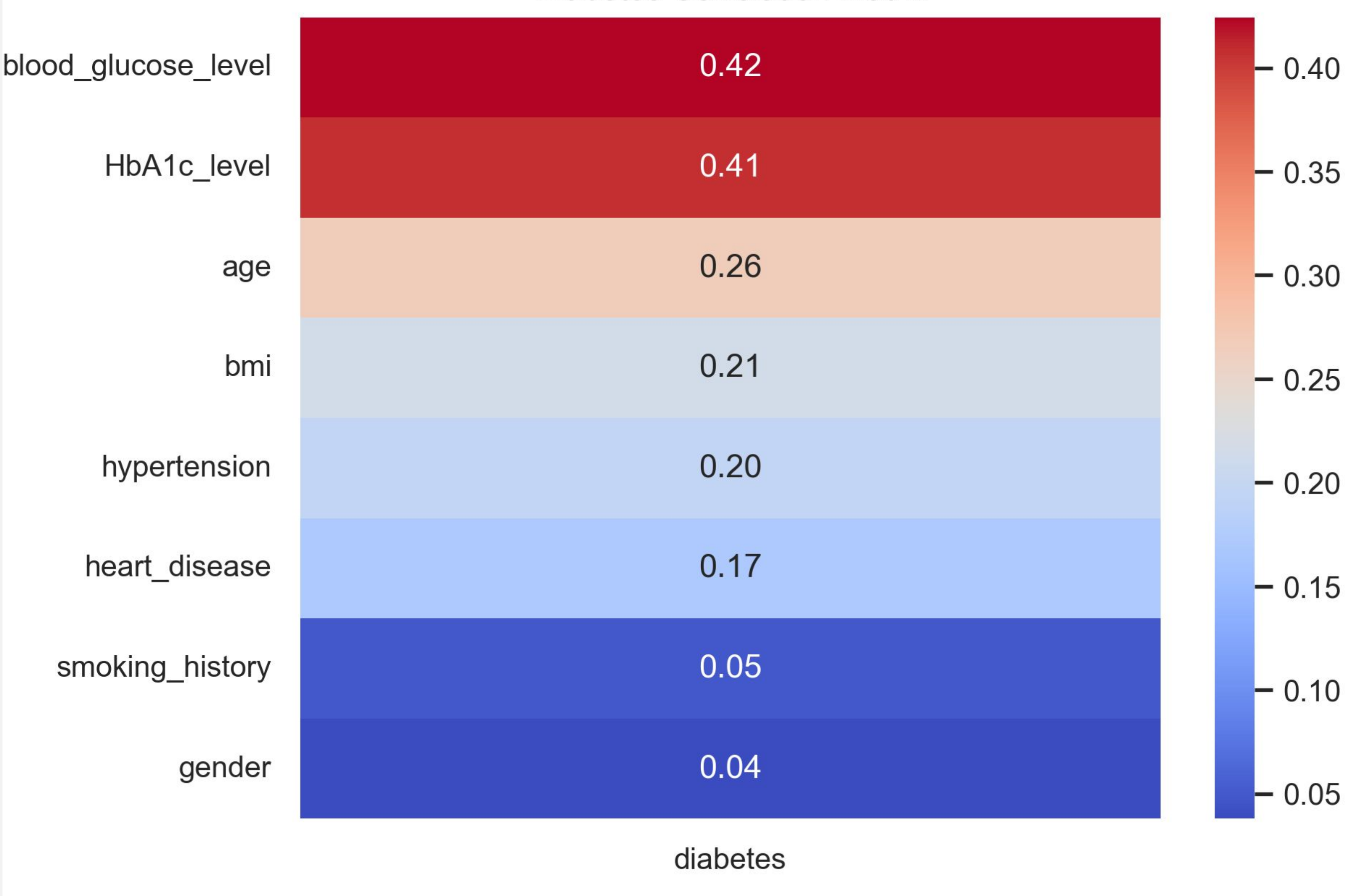


https://www.cdc.gov/diabetes/basics/getting-tested.html

What are some key indicators for diabetes?

Diabetes Correlation Matrix

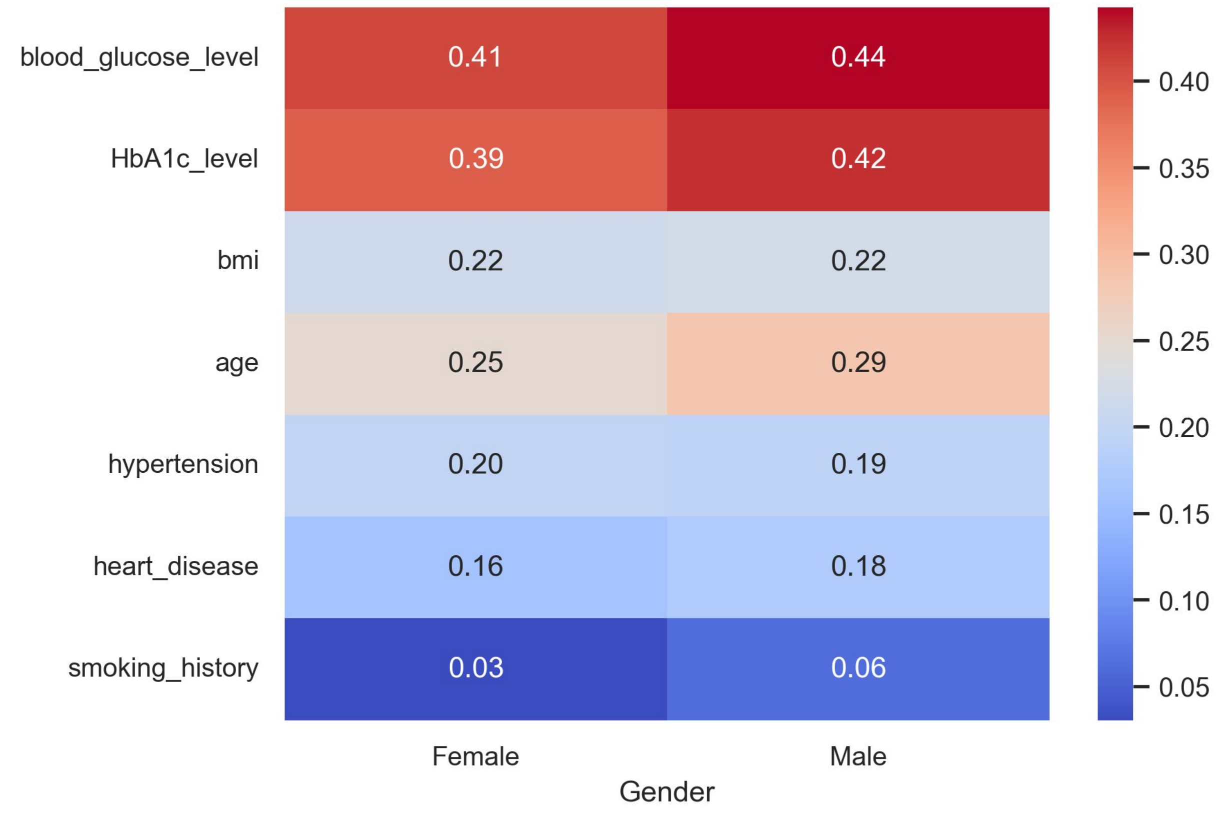
- Blood Glucose and HbA1c have positive linear relationship with diabetes
- Not significant correlation with other metrics



How does stratification affect indicators of diabetes?

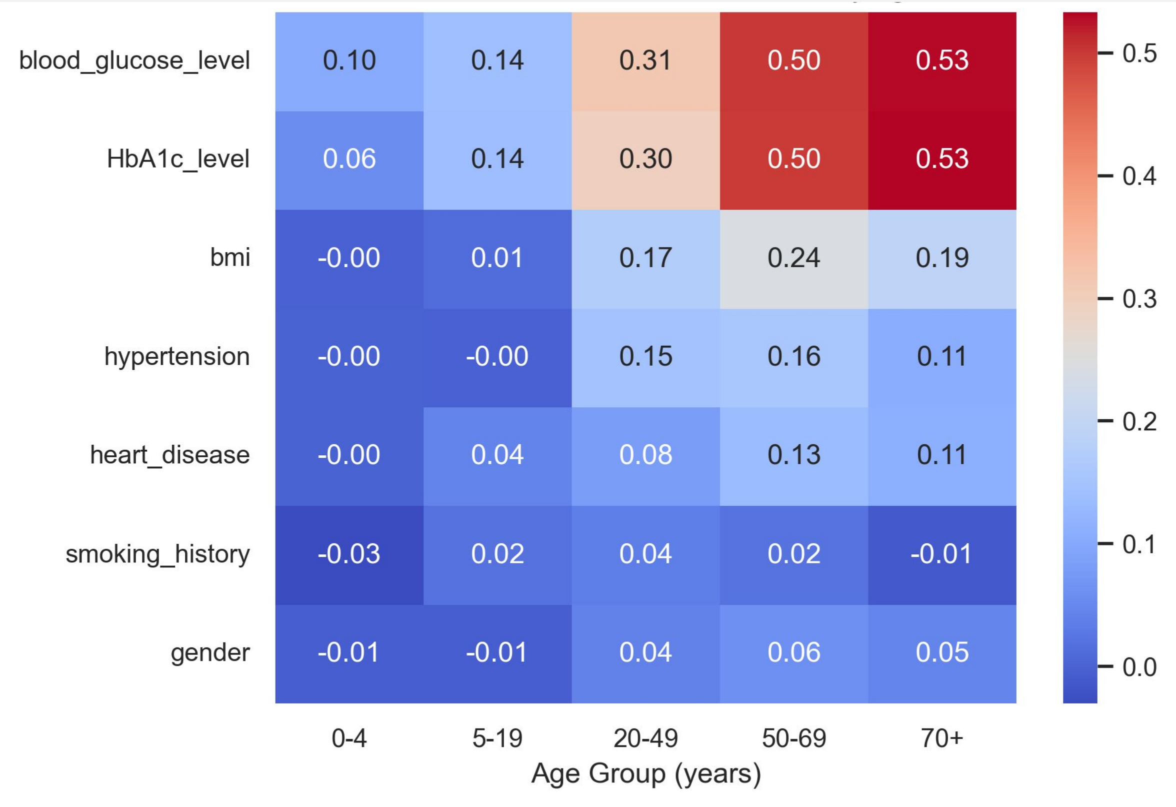
Diabetes Correlation Matrix by Gender

- Very similar correlation matrix with non stratified data

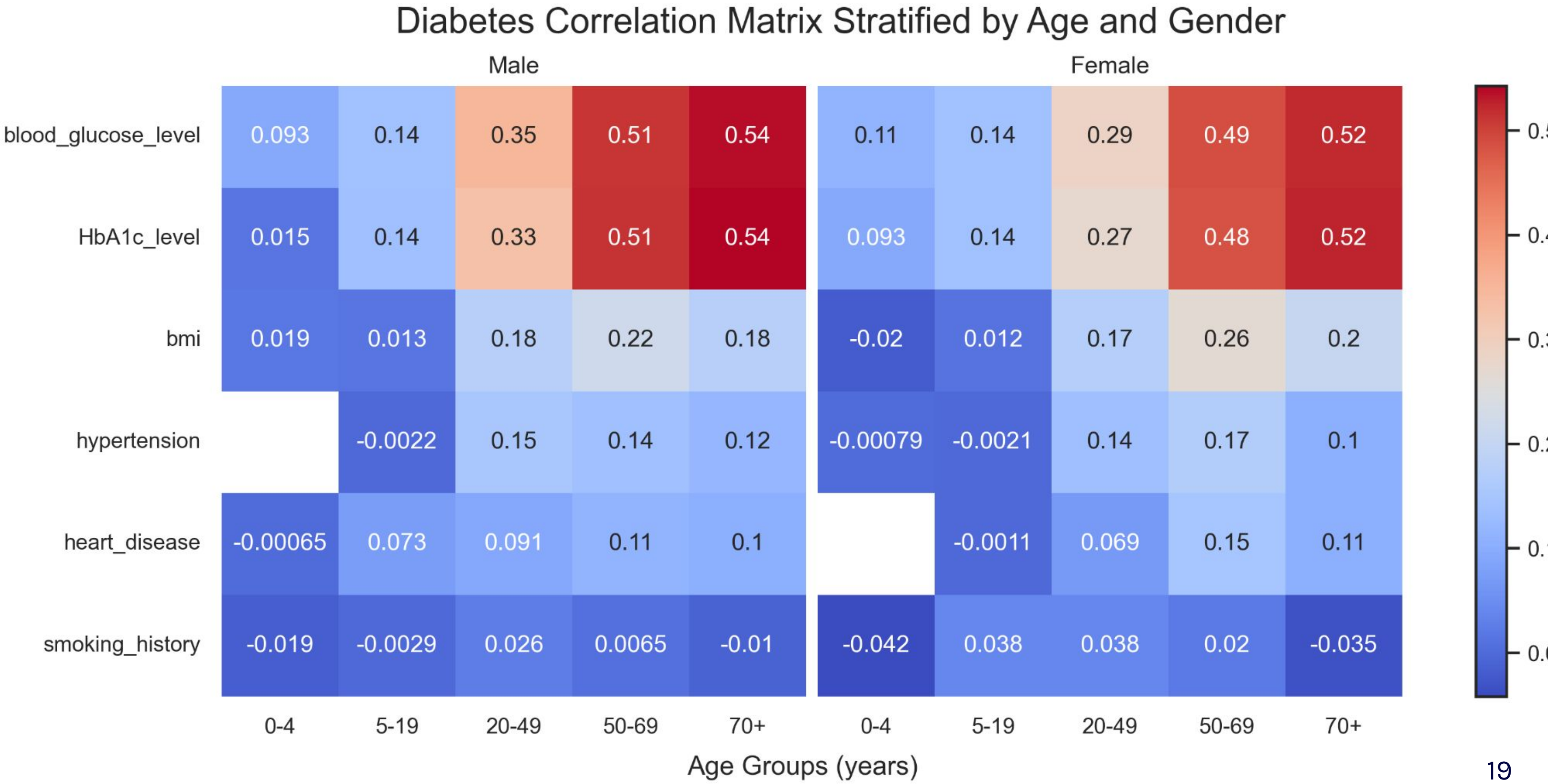


Diabetes Correlation Matrix by Age

- Stronger correlation with blood glucose and HbA1c in older people
- Younger ages don’t have strong indicators



	0-4	5-19	20-49	50-69	70+
blood_glucose_level	0.10	0.14	0.31	0.50	0.53
HbA1c_level	0.06	0.14	0.30	0.50	0.53
bmi	-0.00	0.01	0.17	0.24	0.19
hypertension	-0.00	-0.00	0.15	0.16	0.11
heart_disease	-0.00	0.04	0.08	0.13	0.11
smoking_history	-0.03	0.02	0.04	0.02	-0.01
gender	-0.01	-0.01	0.04	0.06	0.05



Diabetes Correlation Matrix Stratified by Age and Gender

	Male					Female				
	0-4	5-19	20-49	50-69	70+	0-4	5-19	20-49	50-69	70+
blood_glucose_level	0.093	0.14	0.35	0.51	0.54	0.11	0.14	0.29	0.49	0.52
HbA1c_level	0.015	0.14	0.33	0.51	0.54	0.093	0.14	0.27	0.48	0.52
bmi	0.019	0.013	0.18	0.22	0.18	-0.02	0.012	0.17	0.26	0.2
hypertension		-0.0022	0.15	0.14	0.12	-0.00079	-0.0021	0.14	0.17	0.1
heart_disease	-0.00065	0.073	0.091	0.11	0.1		-0.0011	0.069	0.15	0.11
smoking_history	-0.019	-0.0029	0.026	0.0065	-0.01	-0.042	0.038	0.038	0.02	-0.035

How well do standard ML models predict diabetes?

Predictive Modeling

5 models are used

- Logistic Regression - Predicts the probability of an event's occurrence using a logistic function.
- Random Forest - Builds multiple decision trees and combines their predictions to improve accuracy and generalizability.
- Support Vector Machine (SVM) - Finds the optimal boundary between classes to maximize the margin between categories.
- K-Nearest Neighbors (KNN) - Predicts by finding the K nearest training samples to the test data point.
- Gradient Boosting - Incrementally adds weak predictive models (usually decision trees) to minimize the loss function.

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Negative
	Positive	False Positive	True Positive

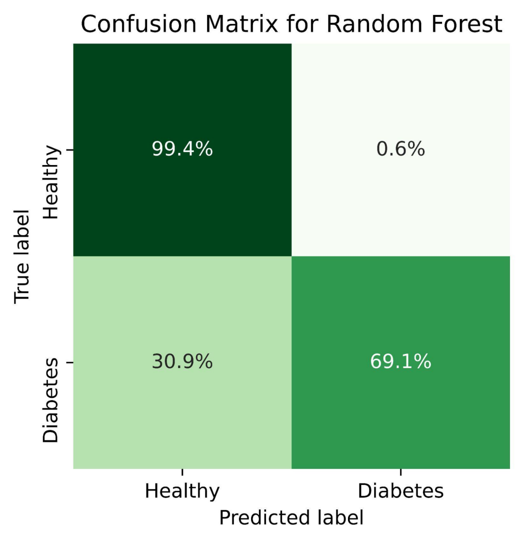
$$Accuracy = \frac{Number\ of\ Correct\ Prediction}{Number\ of\ the\ Results}$$

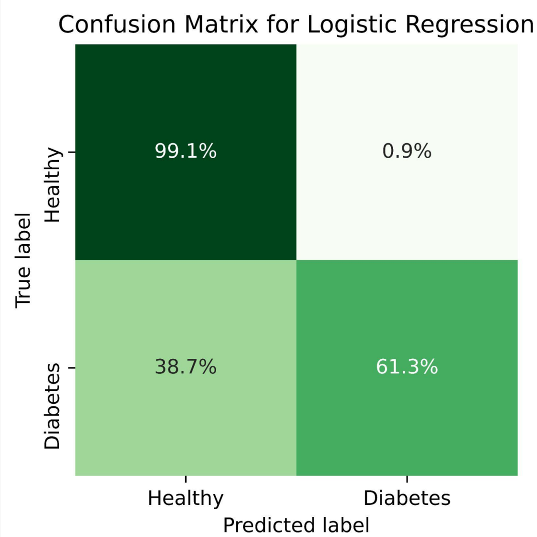
$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

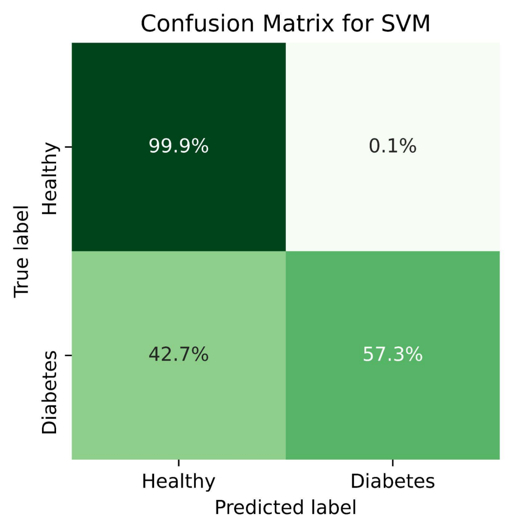
$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

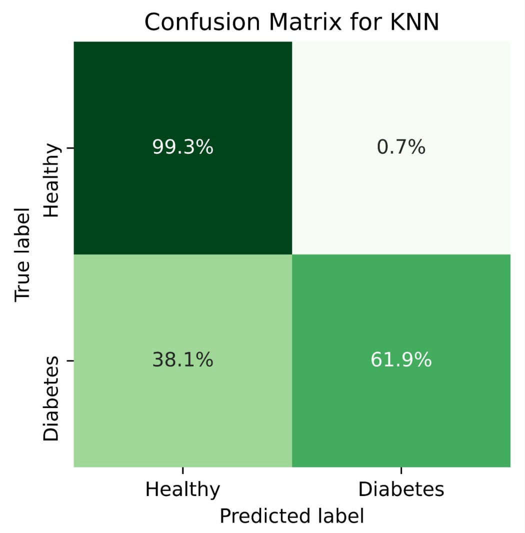
$$F1 = \frac{Precision * Recall}{Precision + Recall}$$

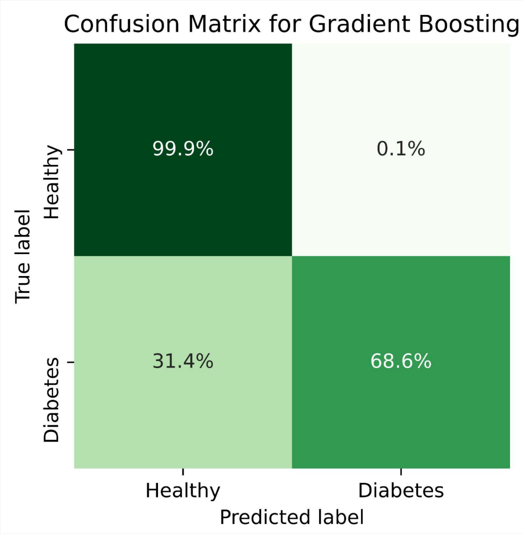
	Negative	Positive
Negative	True Negative	False Negative
Positive	False Positive	True Positive


Confusion Matrix for Random Forest

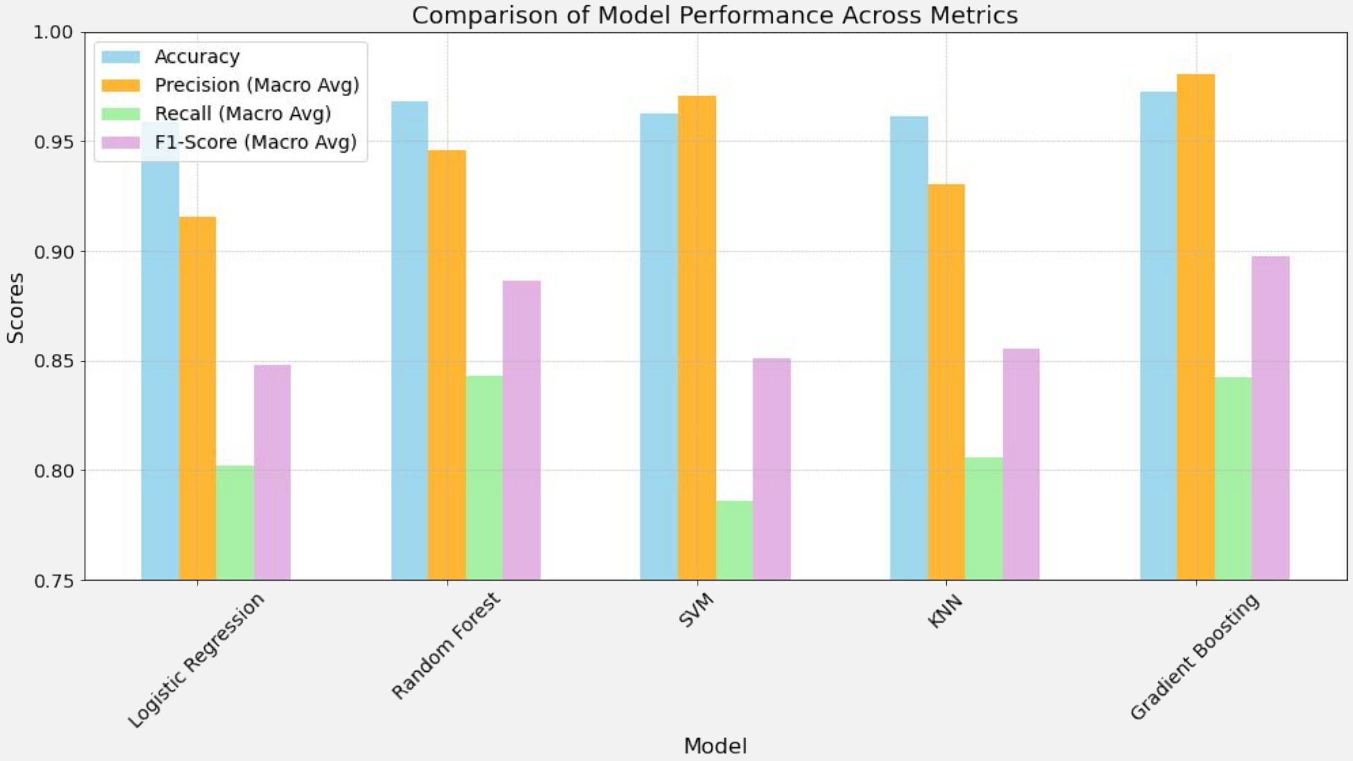

Confusion Matrix for Logistic Regression


Confusion Matrix for SVM

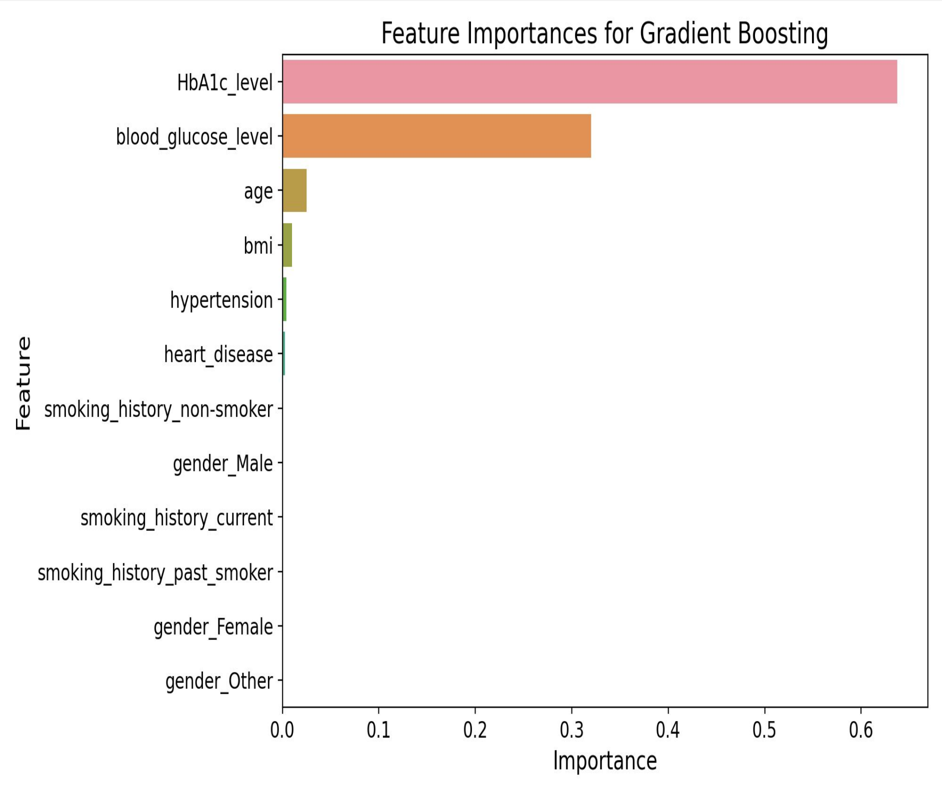

Confusion Matrix for KNN

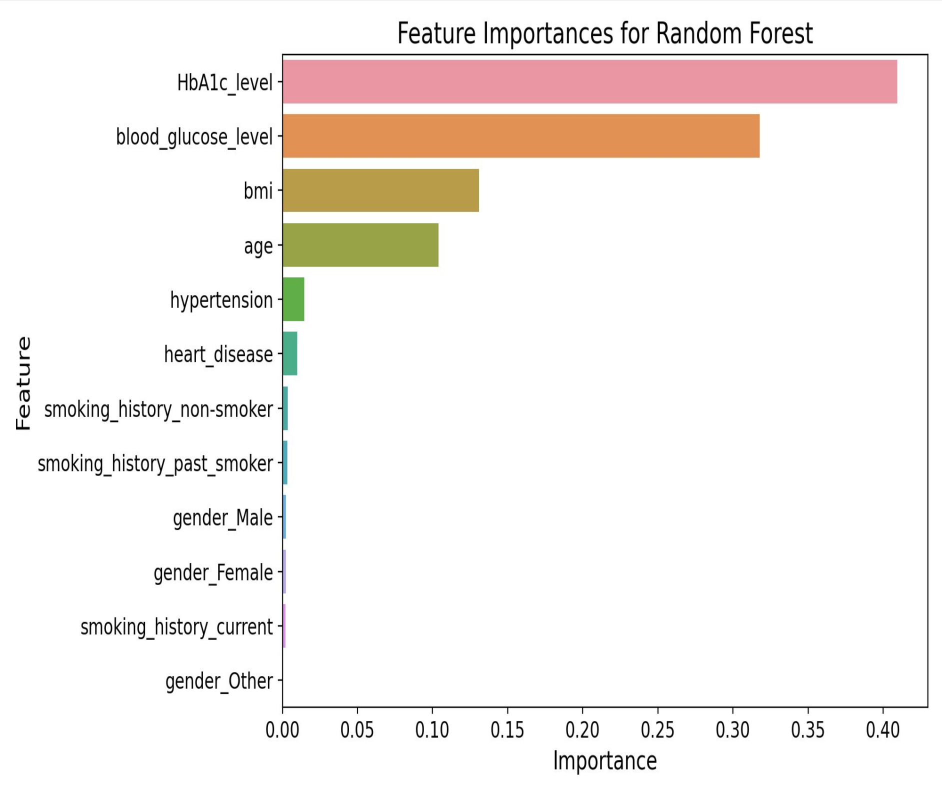

Confusion Matrix for Gradient Boosting

Models Norms	Logistic Regression	Random Forest	SVM	KNN	Gradient Boosting
Accuracy	0.802	0.843	0.786	0.806	0.842
F1 Score	0.756	0.815	0.728	0.761	0.813



Feature importance





Questions?