



Diabetes Prediction

Data Mining B-565

Manav Mandal

Atharva Gurav



Agenda

- Diabetes
- Dataset Overview
- Progress so Far

Diabetes



1. Definition of Diabetes:

1. Body cannot properly regulate blood sugar (glucose) levels.
2. The pancreas' failure towards insulin.

2. Types of Diabetes:

1. **Type 1 Diabetes:** This is an autoimmune condition where the immune system attacks and destroys insulin-producing beta cells in the pancreas.
2. **Type 2 Diabetes:** The body doesn't use insulin properly, and over time, the pancreas may not produce enough insulin.

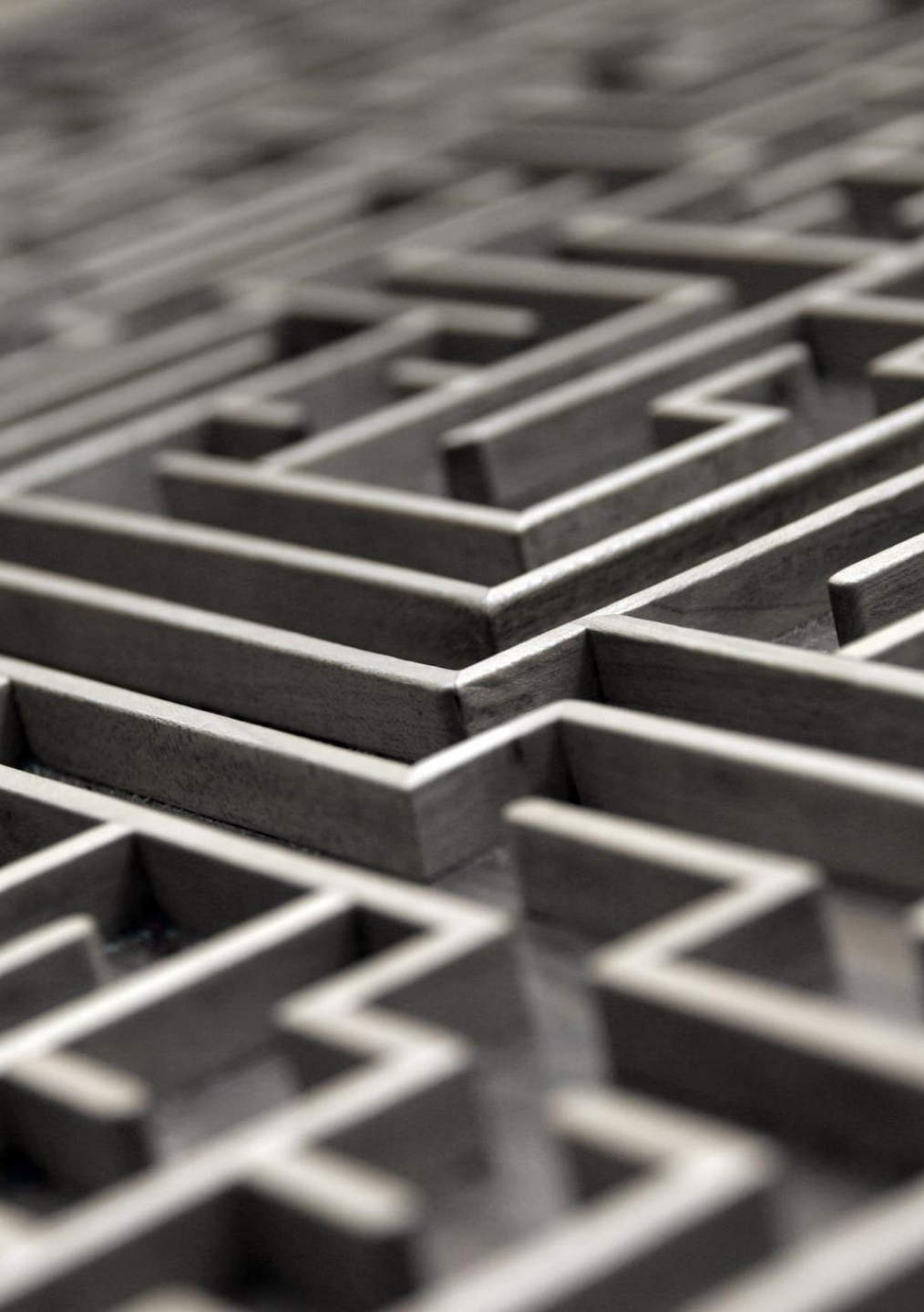
3. Global Prevalence of Diabetes:

1. Over 536 million adults (20-79 years) were living with diabetes globally.

Dataset Overview

1. **Dataset Description:** The CDC Diabetes Health Indicators Dataset contains healthcare statistics and lifestyle survey information about people, along with their diagnosis of diabetes. It includes 35 features such as demographics, lab test results, and answers to survey questions for each patient.
2. **Target Variable:** The target variable for classification is whether a patient has diabetes, is pre-diabetic, or healthy.
3. **Characteristics:** The dataset is tabular and consists of 253,680 instances and 21 features.
4. **Data processing performed** – Bucketing of Age. (AGEG5YR: THIRTEEN-LEVEL AGE CATEGORY)
5. **Purpose:** The dataset was created to better understand the relationship between lifestyle and diabetes in the US and was funded by the CDC.
6. **Features:** features include diabetes diagnosis, demographics, personal information, and health history.
7. **Potential Uses:** The dataset can be used for tasks such as cross-validation or a fixed train-test split.





Progress so far

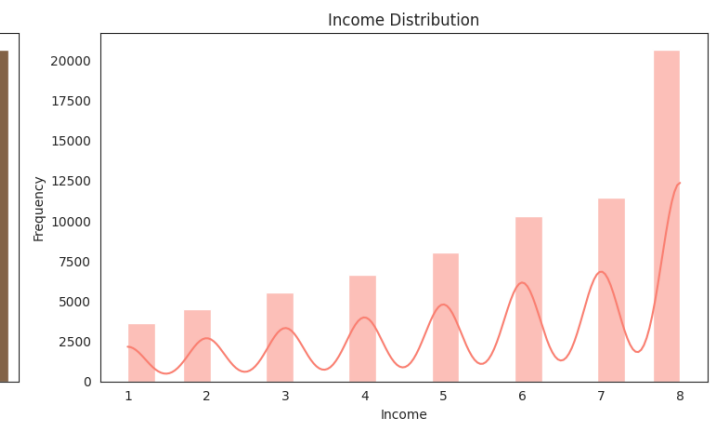
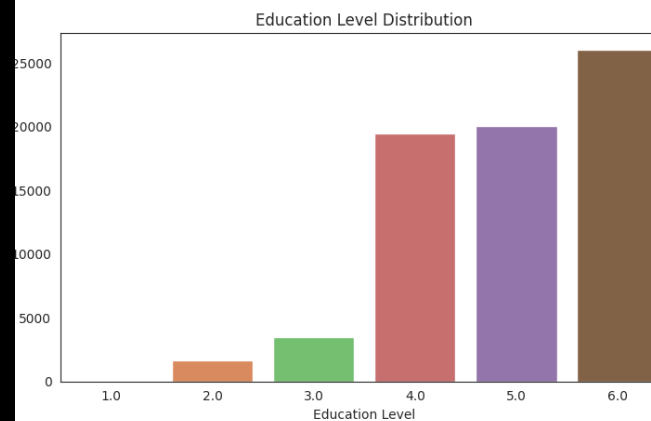
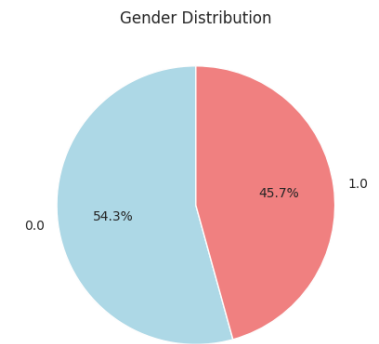
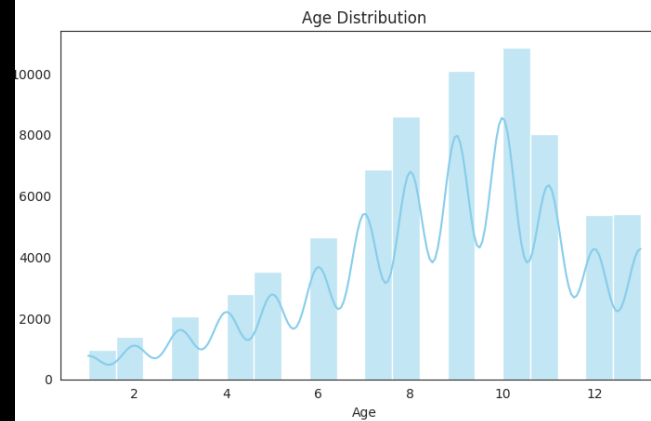
1. EDA
2. Model Fitting
3. Feature Importance using SHAP



Demographic Analysis

12/11/2023

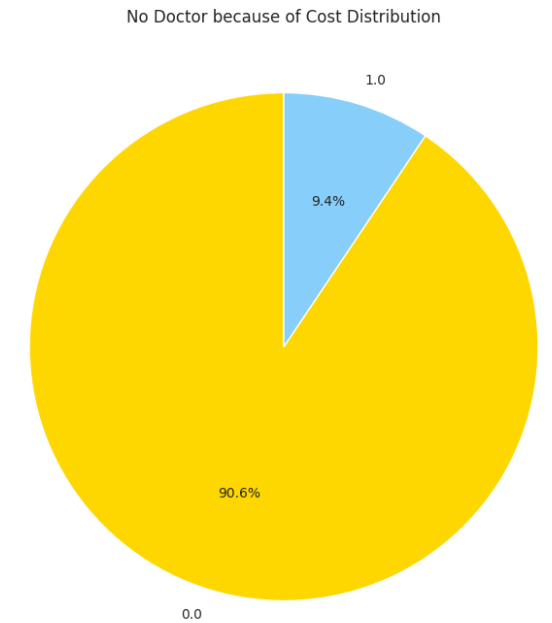
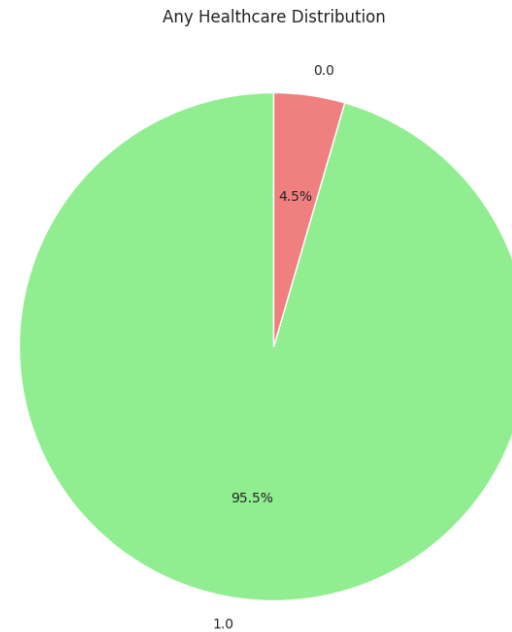
Demographics Analysis



Healthcare Access Analysis

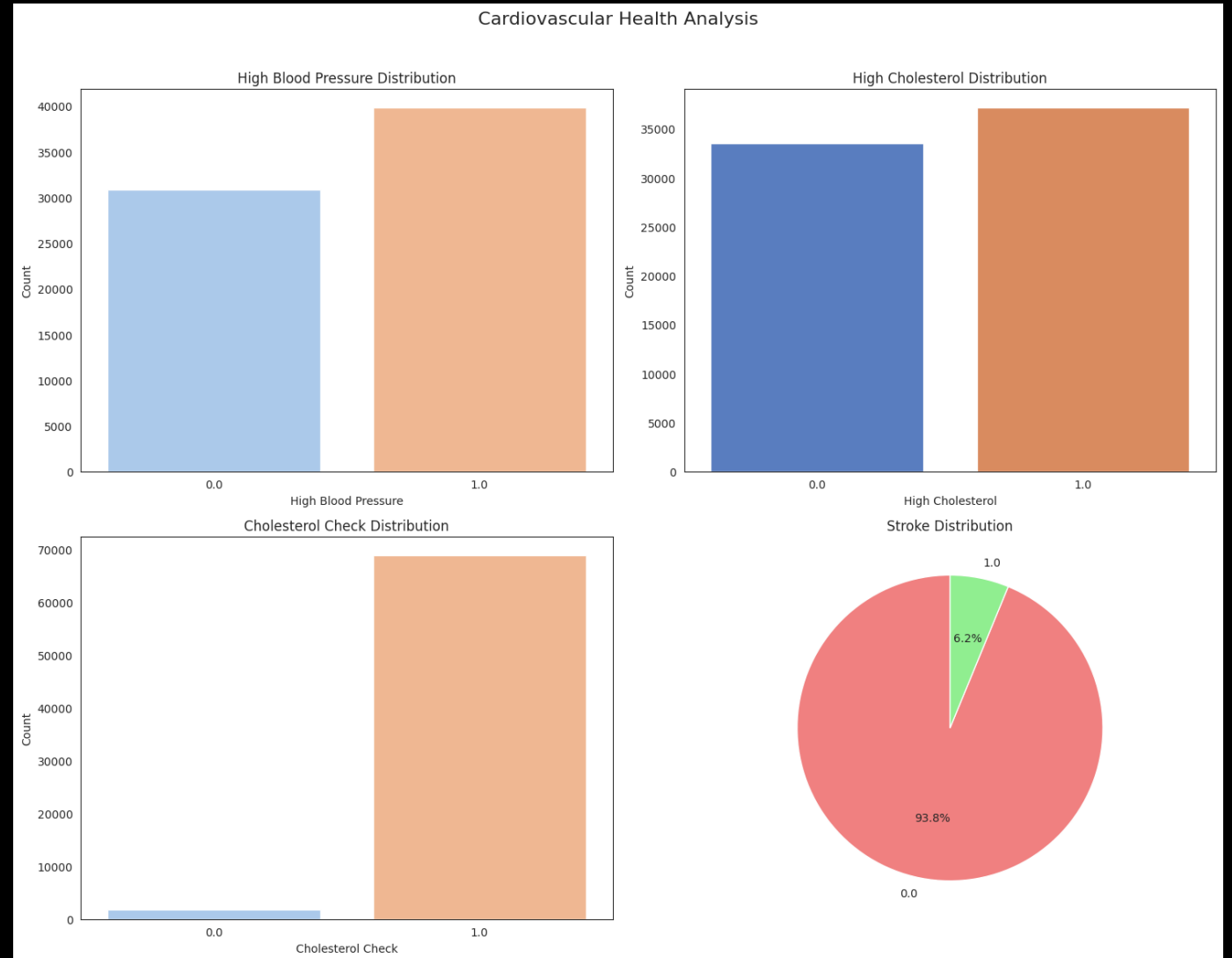
12/11/2023

Healthcare Access Analysis



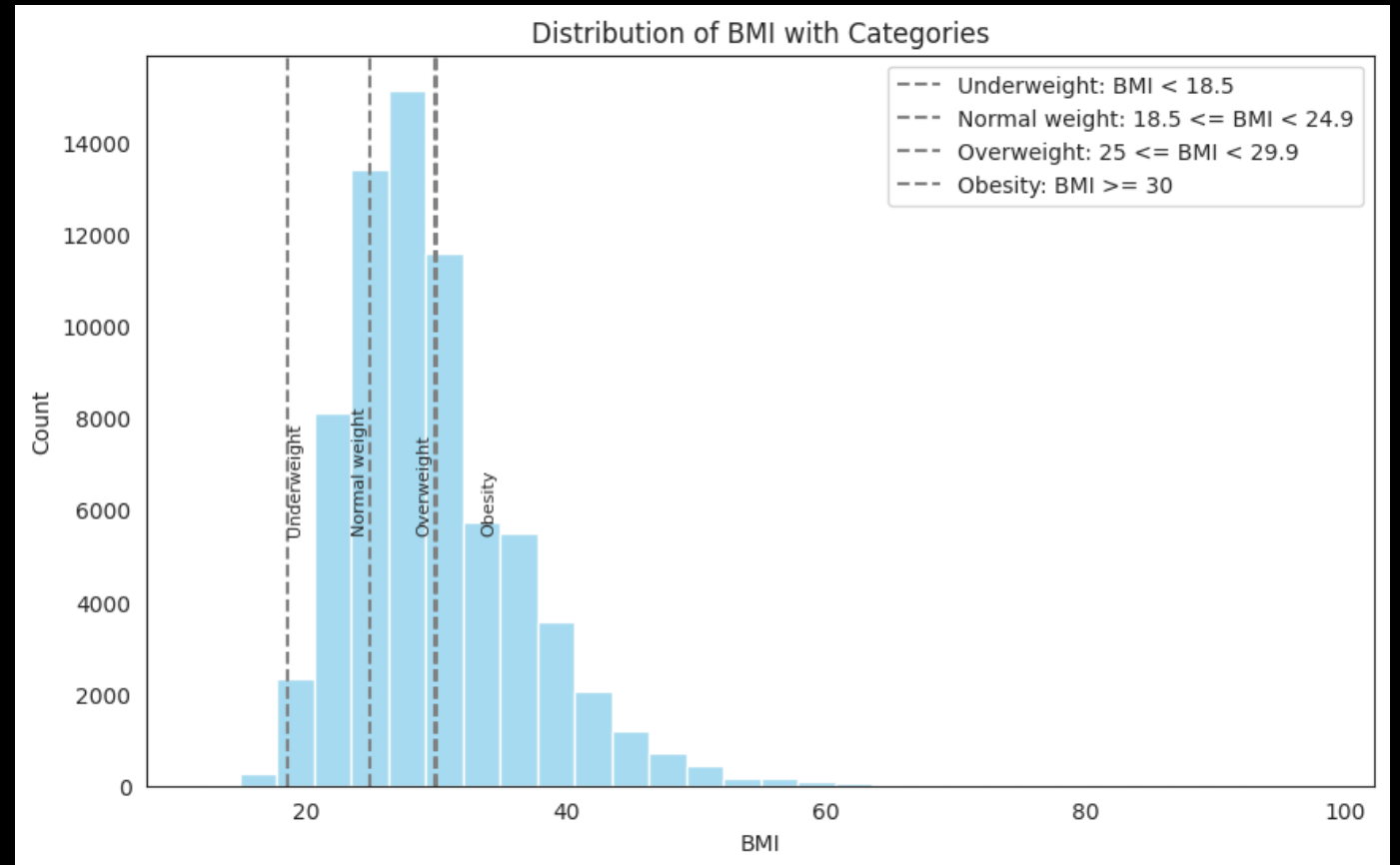
Cardiovascular Health Analysis

12/11/2023



BMI Distribution

12/11/2023



	Correlation between Variables																				
	Diabetes_binary	Age	Sex	Education	Income	AnyHealthcare	NoDocbcCost	HighBP	HighChol	CholCheck	Stroke	HeartDiseaseorAttack	BMI	DiffWalk	GenHlth	MentHlth	PhysHlth	PhysActivity	Smoker	Fruits	Veggies
Diabetes_binary	1	0.28	0.044	-0.17	-0.22	0.023	0.041	0.38	0.29	0.12	0.13	0.21	0.29	0.27	0.41	0.087	0.21	-0.16	0.086	-0.054	-0.079
Age	0.28	1	-0.0023	-0.11	-0.13	0.14	-0.13	0.34	0.24	0.1	0.12	0.22	-0.039	0.2	0.16	-0.1	0.085	-0.1	0.11	0.061	-0.019
Sex	0.044	-0.0023	1	0.044	0.16	-0.0066	-0.048	0.041	0.017	-0.008	0.0038	0.098	0.00083	-0.082	-0.015	-0.089	-0.046	0.052	0.11	-0.089	-0.053
Education	-0.17	-0.11	0.044	1	0.46	0.11	-0.097	-0.14	-0.084	-0.0087	-0.074	-0.097	-0.1	-0.2	-0.29	-0.11	-0.16	0.19	-0.14	0.099	0.079
Income	-0.22	-0.13	0.16	0.46	1	0.13	-0.2	-0.19	-0.11	0.0076	-0.14	-0.15	-0.12	-0.34	-0.38	-0.22	-0.28	0.2	-0.1	0.079	0.15
AnyHealthcare	0.023	0.14	-0.0066	0.11	0.13	1	-0.22	0.036	0.032	0.11	0.0065	0.016	-0.013	0.0081	-0.033	-0.05	-0.0033	0.027	-0.013	0.015	0.027
NoDocbcCost	0.041	-0.13	-0.048	-0.097	-0.2	-0.22	1	0.027	0.033	-0.063	0.036	0.036	0.066	0.13	0.17	0.19	0.16	-0.063	0.036	-0.046	-0.037
HighBP	0.38	0.34	0.041	-0.14	-0.19	0.036	0.027	1	0.32	0.1	0.13	0.21	0.24	0.23	0.32	0.064	0.17	-0.14	0.087	-0.041	-0.067
HighChol	0.29	0.24	0.017	-0.084	-0.11	0.032	0.033	0.32	1	0.086	0.1	0.18	0.13	0.16	0.24	0.084	0.14	-0.09	0.093	-0.047	-0.043
CholCheck	0.12	0.1	-0.008	-0.0087	0.0076	0.11	-0.063	0.1	0.086	1	0.023	0.043	0.046	0.044	0.059	-0.011	0.035	-0.0082	0.0043	0.017	0.015
Stroke	0.13	0.12	0.0038	-0.074	-0.14	0.0065	0.036	0.13	0.1	0.023	1	0.22	0.023	0.19	0.19	0.087	0.16	-0.08	0.065	-0.019	-0.019
HeartDiseaseorAttack	0.21	0.22	0.098	-0.097	-0.15	0.016	0.036	0.21	0.18	0.043	0.22	1	0.06	0.23	0.28	0.075	0.2	-0.098	0.12	-0.019	-0.019
BMI	0.29	-0.039	0.00083	-0.1	-0.12	-0.013	0.066	0.24	0.13	0.046	0.023	0.06	1	0.25	0.27	0.1	0.16	-0.17	0.012	-0.085	-0.019
DiffWalk	0.27	0.2	-0.082	-0.2	-0.34	0.0081	0.13	0.23	0.16	0.044	0.19	0.23	0.25	1	0.48	0.25	0.49	-0.28	0.12	-0.051	-0.019
GenHlth	0.41	0.16	-0.015	-0.29	-0.38	-0.033	0.17	0.32	0.24	0.059	0.19	0.28	0.27	0.48	1	0.32	0.55	-0.27	0.15	-0.099	-0.019
MentHlth	0.087	-0.1	-0.089	-0.11	-0.22	-0.05	0.19	0.064	0.084	-0.011	0.087	0.075	0.1	0.25	0.32	1	0.38	-0.13	0.091	-0.062	-0.019
PhysHlth	0.21	0.085	-0.046	-0.16	-0.28	-0.0033	0.16	0.17	0.14	0.035	0.16	0.2	0.16	0.49	0.55	0.38	1	-0.23	0.12	-0.049	-0.019
PhysActivity	-0.16	-0.1	0.052	0.19	0.2	0.027	-0.063	-0.14	-0.09	-0.0082	-0.08	-0.098	-0.17	-0.28	-0.27	-0.13	-0.23	1	-0.08	0.13	0.1
Smoker	0.086	0.11	0.11	-0.14	-0.1	-0.013	0.036	0.087	0.093	-0.0043	0.065	0.12	0.012	0.12	0.15	0.091	0.12	-0.08	1	-0.075	-0.019
Fruits	-0.054	0.061	-0.089	0.099	0.079	0.029	-0.046	-0.041	-0.047	0.017	-0.009	-0.019	-0.085	-0.051	-0.099	-0.062	-0.049	0.13	-0.075	1	0.2
Veggies	-0.079	-0.019	-0.053	0.15	0.15	0.029	-0.037	-0.067	-0.043	0.00035	-0.048	-0.036	-0.057	-0.084	-0.12	-0.052	-0.067	0.15	-0.03	0.24	1
HvyAlcoholConsump	-0.095	-0.058	0.014	0.036	0.064	-0.013	0.0097	-0.027	-0.025	-0.027	-0.023	-0.037	-0.058	-0.049	-0.059	0.016	-0.036	0.019	0.078	-0.033	0.019

1. Exploratory Data Analysis

Important Attributes according to Correlation Matrix

- GenHlth 0.407612
- HighBP 0.381516
- BMI 0.293373
- HighChol 0.289213
- Age 0.278738
- DiffWalk 0.272646
- Income 0.224449
- PhysHlth 0.213081
- HeartDiseaseorAttack 0.211523
- Education 0.170481
- PhysActivity 0.158666
- Stroke 0.125427
- CholCheck 0.115382
- HvyAlcoholConsump 0.0948

Classification Report				
	precision	recall	f1-score	suppo
0	0.88	0.97	0.92	436
1	0.50	0.16	0.25	70
accuracy			0.86	507
macro avg	0.69	0.57	0.59	507
weighted avg	0.82	0.86	0.83	507
	precision	recall	f1-score	suppo
0	0.89	0.87	0.88	436
1	0.29	0.33	0.31	70
accuracy			0.79	507
macro avg	0.59	0.60	0.59	507
weighted avg	0.80	0.79	0.80	507
	precision	recall	f1-score	suppo
0	0.88	0.97	0.92	436
1	0.49	0.18	0.26	70
accuracy			0.86	507
macro avg	0.68	0.57	0.59	507
weighted avg	0.82	0.86	0.83	507

2.Models used

- Validation Accuracies

- SVM - 0.8613
- Logistic Regression - 0.8601
- Random Forest - 0.8592
- k-NN - 0.8490
- Decision Tree - 0.7936
- Naive Bayes - 0.7713



Progress so far

1. EDA
2. Model Fitting
3. Feature Importance using SHAP



Feature Importance using SHAP (XAI)

- Feature importance is a concept in machine learning that aims to understand and quantify the impact of different features on the predictions made by a model.
- SHAP (SHapley Additive exPlanations) is a popular approach for explaining the output of machine learning models by assigning a value, called Shapley value, to each feature indicating its contribution to the model's prediction.
- Positive SHAP values indicate features pushing the model's prediction higher, while negative values indicate features pulling the prediction lower.

A bright, modern dining room with two windows, a table set for a meal, a large indoor plant, and pendant lights.

THANK YOU

ANY QUESTIONS?