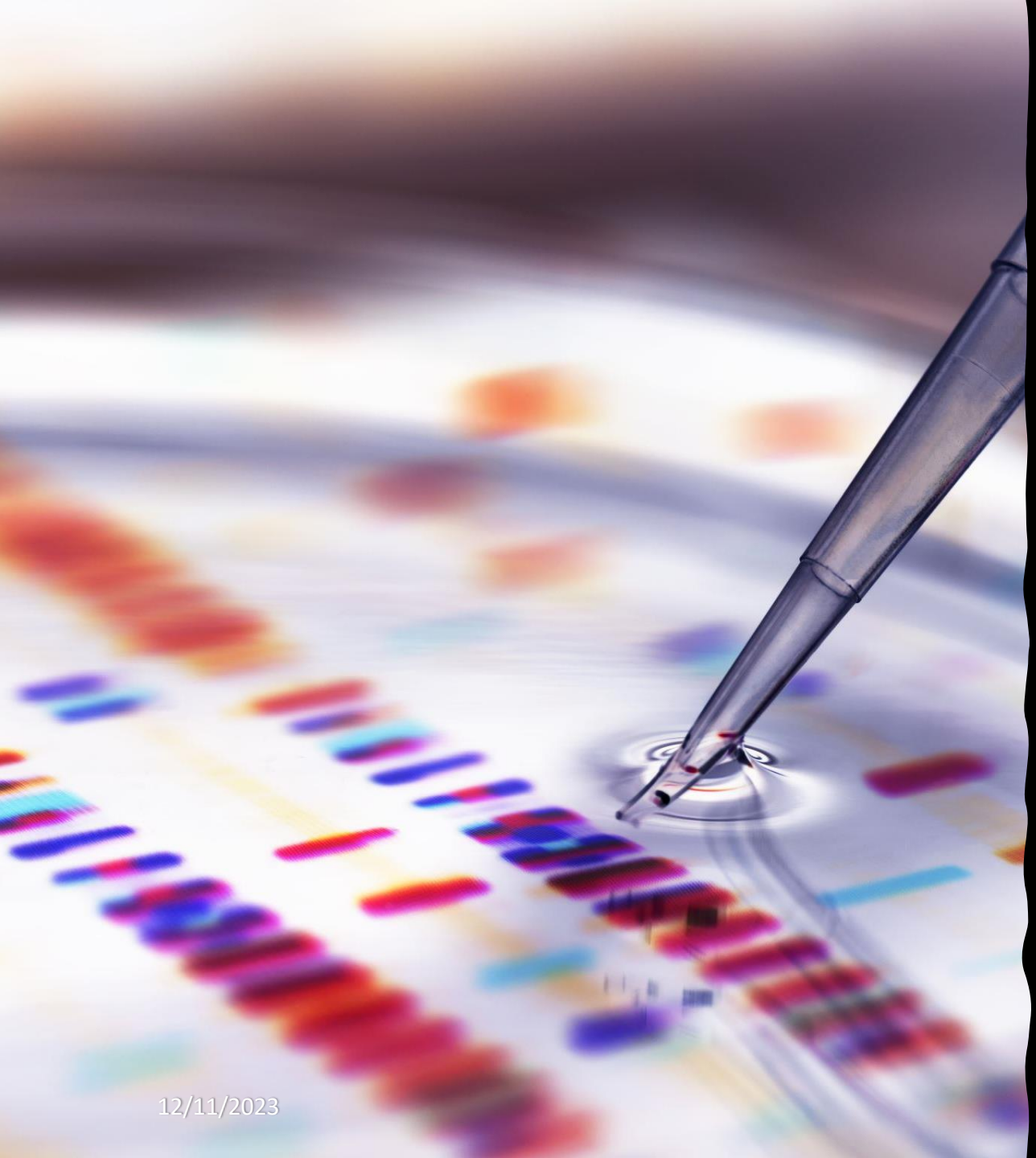


Diabetes Prediction

Data Mining B-565

Manav Mandal

Atharva Gurav

Agenda

- Diabetes
- Dataset Overview
- Progress so Far

Diabetes

Global

Diabetes data report 2000 – 2045

At a glance	2000	2011	2021	2030	2045
Diabetes estimates (20-79 y)					
People with diabetes, in 1,000s	151,000.0	366,000.0	536,600.0	642,800.0	783,700.0

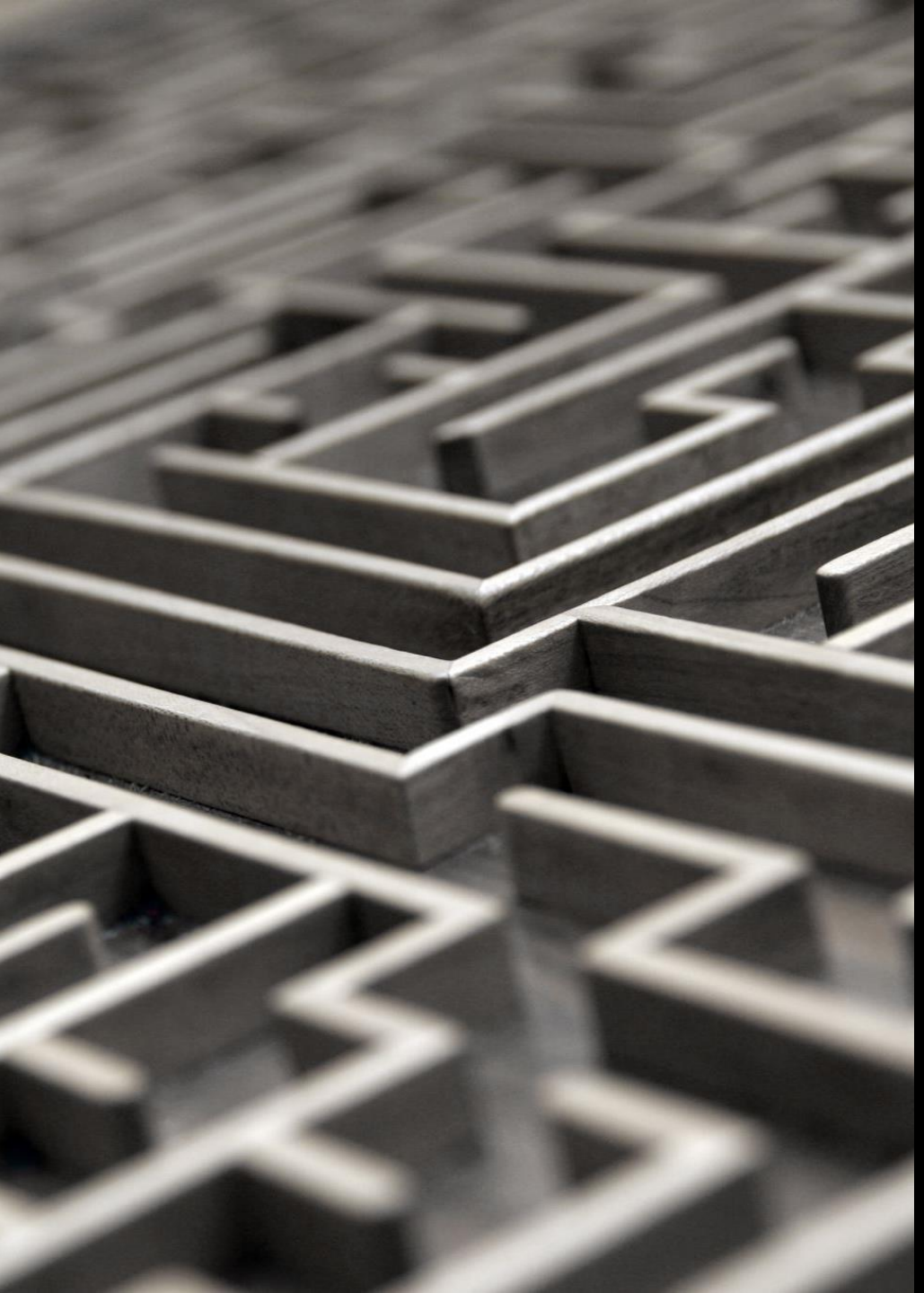
1. **Definition of Diabetes:**
    1. Body cannot properly regulate blood sugar (glucose) levels.
    2. The pancreas' failure towards insulin.

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.

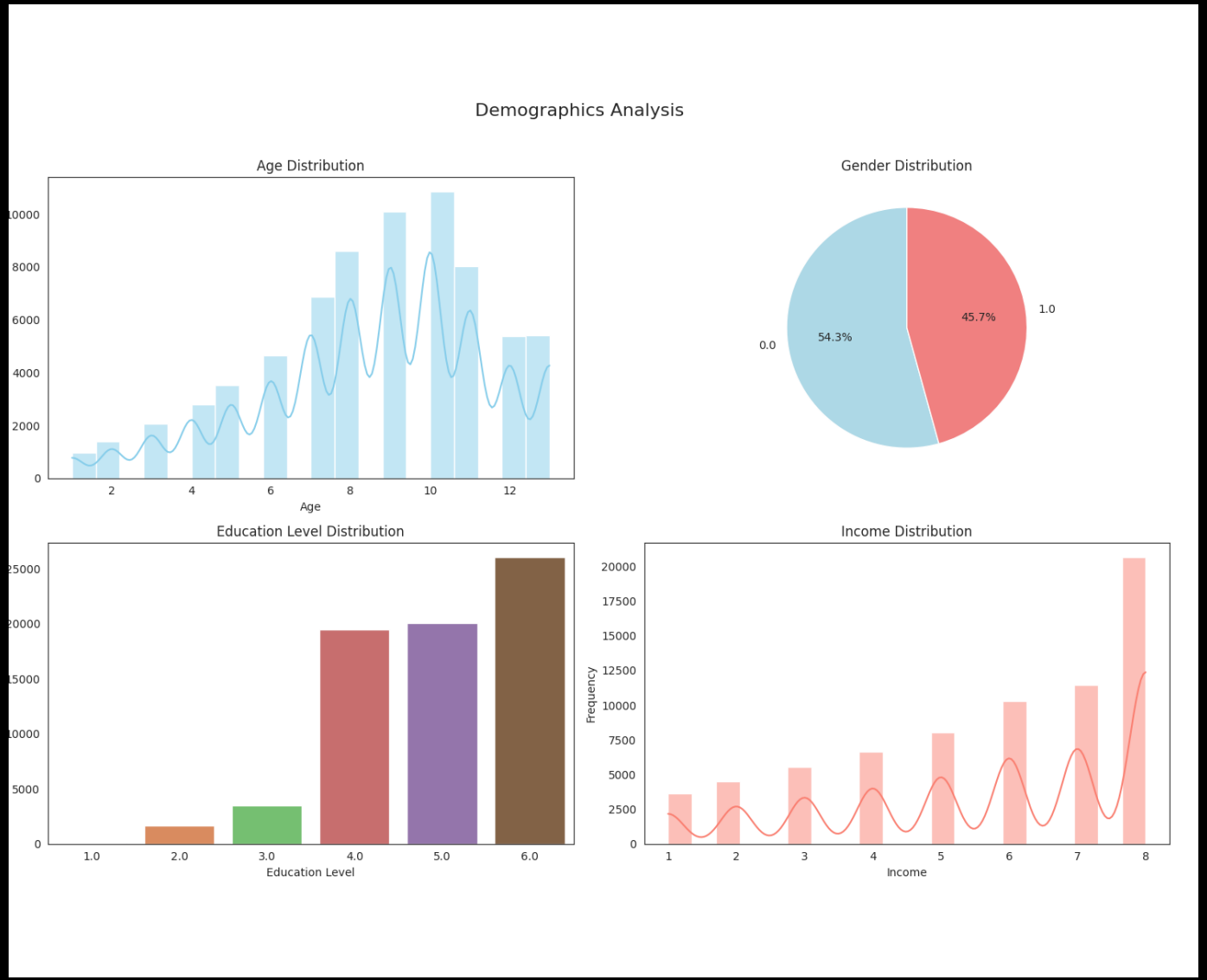


Progress so far

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

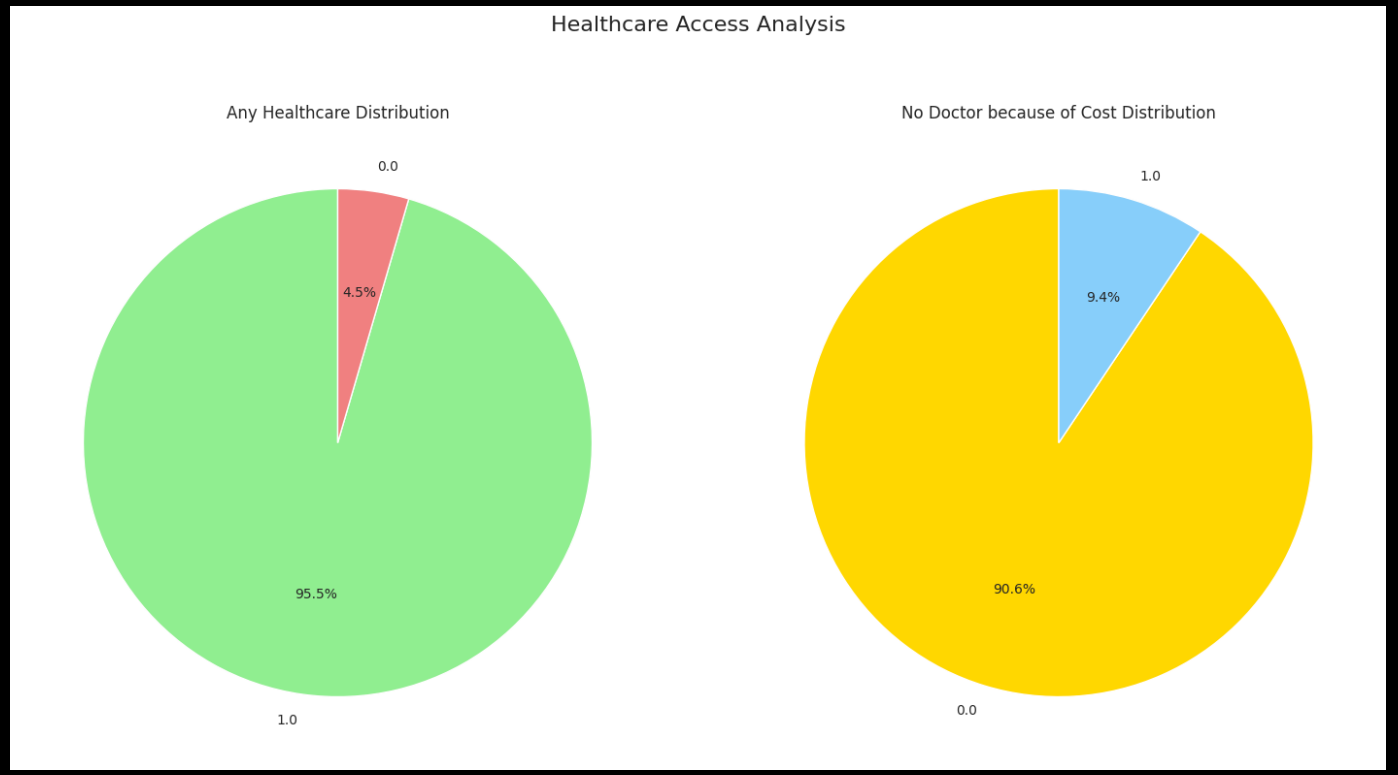
Demographic Analysis

12/11/2023



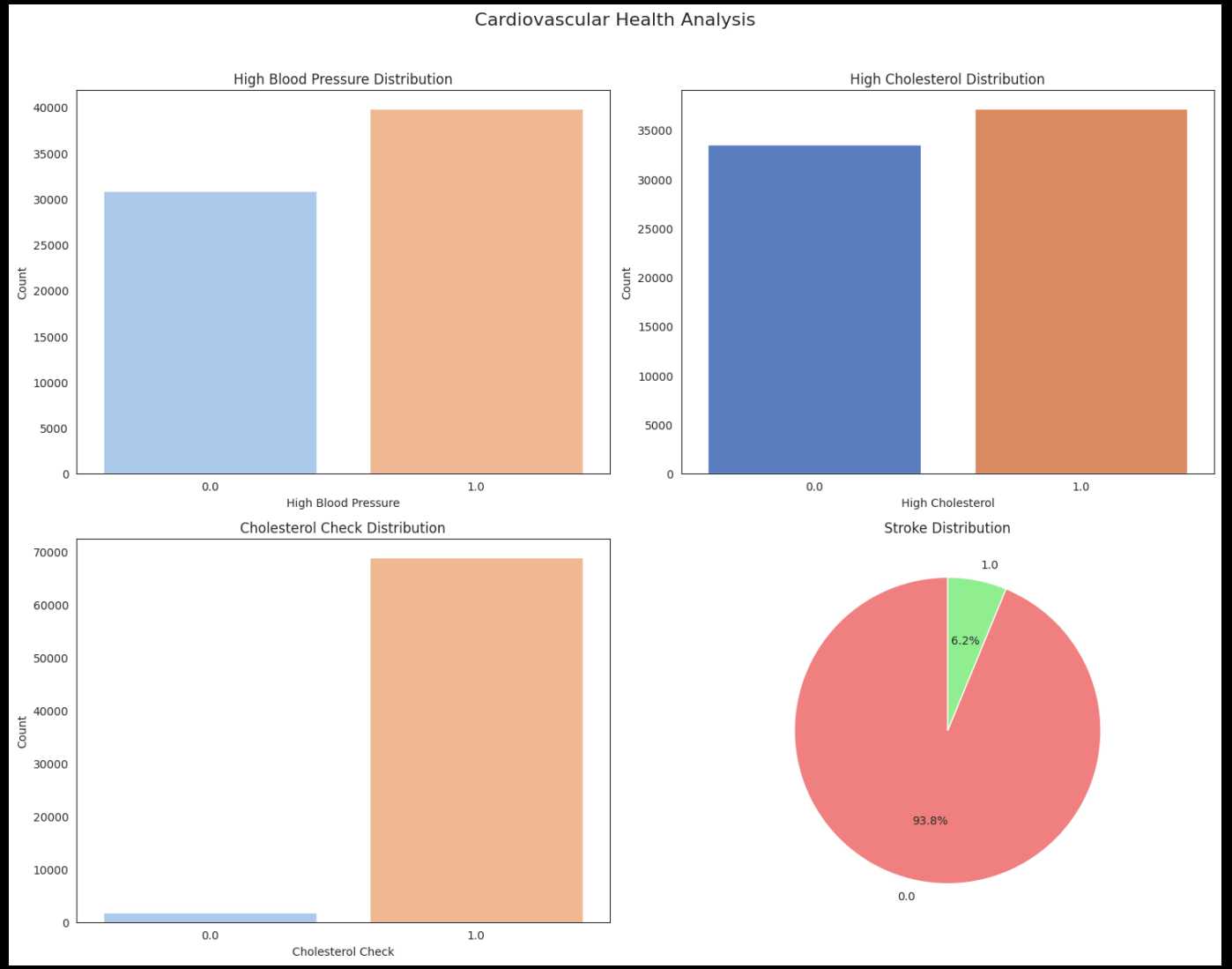
Healthcare Access Analysis

12/11/2023



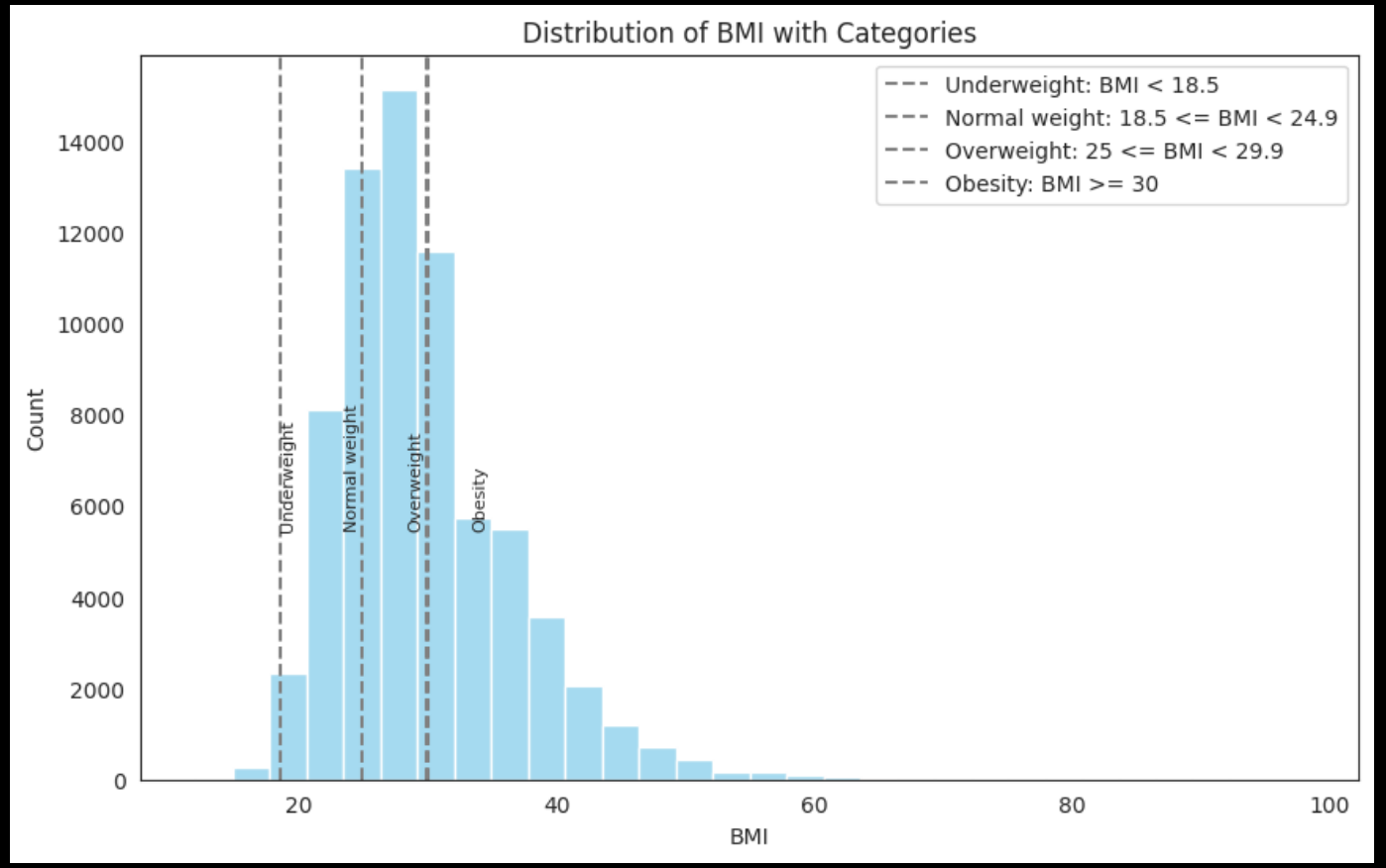
Cardiovascular Health Analysis

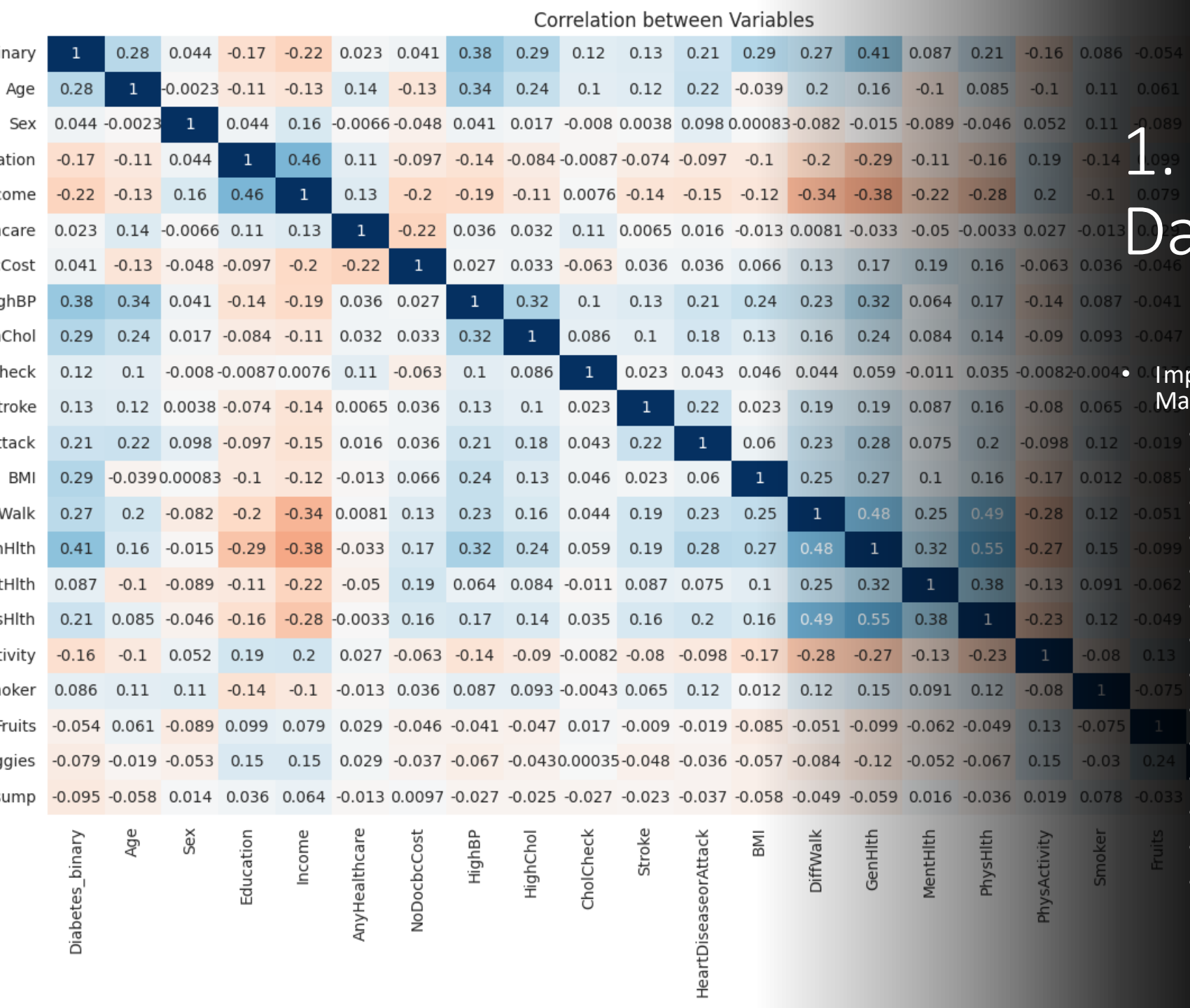
12/11/2023



BMI Distribution

12/11/2023





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.

THANK YOU

ANY QUESTIONS?