رواد مصر الرقمية

Healthcare Predictive Analytics

Diabetes & Heart Disease Prediction

Team Members

Group: ALX2_AIS4_S2

Name	Role
احمد محمد محمود	Exploratory Data Analysis
عاصم أحمد مرسي	Data Preprocessing
احمد السيد احمد	PowerBI Dashboard
زياد محمد السيد	Diabetes Prediction Model
ياسين ممدوح محمد	Heart Disease Prediction Model
عبدالرحمن احمد محمد	Website + API Deployment (Mlflow and Docker)

Table of Content

1. Project Overview……..……..……..4
2. Data Collection…….…….5
3. Data Preprocessing……..6
4. Exploratory Data Analysis…….….…11
5. PowerBI Analysis Dashboard……20
6. Project Pipeline……27
7. Models Building & Evaluation……..……..…..28
   7.1. Diabetes Model……..……...……..…..28
   7.2. Heart Disease Model……..…….……..…..35
8. Deployment……..43
9. Appendix & References…….53

Project Overview

Objective

This project aims to enhance the early detection of chronic diseases—specifically heart disease and diabetes—through predictive machine learning models. It also incorporates data visualizations and interactive dashboards to uncover patterns and provide insights into key factors influencing disease risk. By translating health data into actionable insights, the project empowers individuals to take proactive steps toward preventing chronic illnesses.

Solution

The project delivers an interactive, web-based platform with:

- **Real-time Risk Assessment:** Users input health metrics (e.g., blood pressure, BMI) to receive instant predictions.
- **Visual Insights:** Dashboards made with *PowerBI* and visualizations made using *matplotlib* and *seaborn* explore trends and risk factors.
- **Technical Implementation:** A Flask-powered API serves machine learning models built with *Scikit-learn*, ensuring fast, accessible predictions.

Impact

Deployed as a user-friendly tool, the platform bridges data science and preventive healthcare. Tests achieved **86.7%** and **87.3%** accuracy in predicting diabetes and heart disease risks respectively, demonstrating its potential for real-world adoption.

Data Collection

The primary dataset used in this project is the **Diabetes Health Indicators Dataset**, sourced from Kaggle. This dataset originates from the **Behavioral Risk Factor Surveillance System (BRFSS) 2015** survey conducted by the CDC (Centers for Disease Control and Prevention). It contains health-related indicators for over **253,000 survey respondents**, making it a robust source for predictive modeling in diabetes risk assessment.

Dataset Overview

- **File Used:** `diabetes_binary_health_indicators_BRFSS2015.csv`
- **Features:** 21 health-related variables, including:
  - **Target Variables:**
    - `Diabetes_binary` (0 = no diabetes, 1 = pre diabetes/diabetes)
    - `HeartDiseaseorAttack` coronary heart disease (CHD) or myocardial infarction (MI) (0 = no 1 = yes)
  - **Predictors:** High blood pressure, BMI, age, income, physical activity, smoking habits, and other lifestyle factors.
- **Sample Size:** 253,680 records
- **Data Type:** Structured (tabular) data

Data Preprocessing

Initial Dataset Overview

- Purpose: Understand the structure and composition of the dataset.
- Used `df.info()` to inspect the dataset.
- Revealed 253,680 entries with 22 float-type columns.
- No missing values or different data types at this stage.
- Memory usage was ~42.6 MB, which is relevant for optimization decisions.

Dataset statistics

Number of variables	22
Number of observations	253680
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	11369
Duplicate rows (%)	4.5%
Total size in memory	42.6 MiB
Average record size in memory	176.0 B

Variable types

Categorical	16
Numeric	6

Statistical Summary

- Purpose: Identify key statistics (e.g., mean, median, outliers) for each feature.
- Used `df.describe()` to analyze range, central tendency, and dispersion.
- Found BMI ranges from 12 to 98, and outliers in MentHlth, PhysHlth.
- Helps in understanding skewness and data scaling needs.

	Diabetes_binary	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity
count	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000
mean	0.139333	0.429001	0.424121	0.962670	28.382364	0.443169	0.040571	0.094186	0.756544
std	0.346294	0.494934	0.494210	0.189571	6.608694	0.496761	0.197294	0.292087	0.429169
min	0.000000	0.000000	0.000000	0.000000	12.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	1.000000	24.000000	0.000000	0.000000	0.000000	1.000000
50%	0.000000	0.000000	0.000000	1.000000	27.000000	0.000000	0.000000	0.000000	1.000000
75%	0.000000	1.000000	1.000000	1.000000	31.000000	1.000000	0.000000	0.000000	1.000000
max	1.000000	1.000000	1.000000	1.000000	98.000000	1.000000	1.000000	1.000000	1.000000

Data Preprocessing

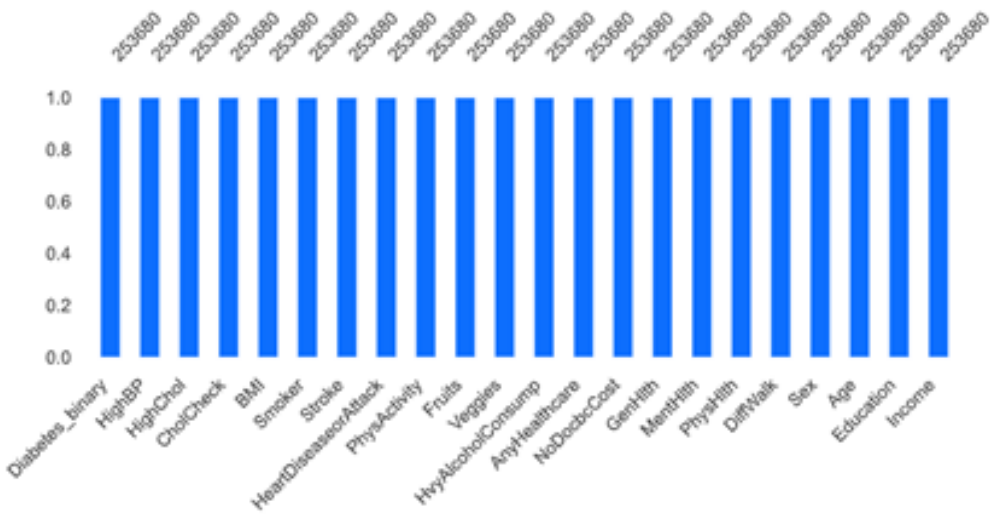
Duplicate Detection and Removal

- Purpose: Ensure dataset quality by removing redundant rows.
- Detected 24,206 duplicate rows using `df.duplicated().sum()`.
- Removed duplicates with `df.drop_duplicates(inplace=True)`.
- Confirmed no remaining duplicates afterward.

```
Click to add a breakpoint  or Duplicates"
  count the number of duplicated rows in the dataset.",

df.duplicated().sum()

24206

"### 🧹 Removing Duplicates"
"Drop duplicated rows to ensure each record is unique, then re-check to confirm."

df.drop_duplicates(inplace=True)
df.duplicated().sum()

0
```

Missing Values Check

- Purpose: Confirm dataset completeness before further processing.
- Used `df.isna().sum()` to scan each column.
- Result: No missing values across all 22 columns.
- Ensures no imputation or row-dropping was necessary.



A simple visualization of nullity by column.

Data Preprocessing

Memory Optimization

- Purpose: Reduce resource usage for faster processing.
- All columns were originally **float64**.
- Converted the entire dataset to **int32** using `df.astype('int32')`.
- Reduced memory usage from **42.6 MB** to **21.0 MB**.

```
"### 🧠 Memory Optimization"
"Convert data types to more memory-efficient formats, in this case from float32 to int32."
df = df.astype('int32')
```

```
"### 🔍 Post-Optimization Check"
"Review the updated data types and memory usage after optimization."

df.info()
```

Unique Value Analysis

- Purpose: Measure variability and potential feature encoding strategies.
- Used value count logic to compute unique values per feature.
- Categorical features like Sex, Smoker have 2 unique values.
- MentHlth, PhysHlth have higher variability (31 unique values).

	unique value count
Diabetes_binary	2
HighBP	2
HighChol	2
CholCheck	2
BMI	84
Smoker	2
Stroke	2
HeartDiseaseorAttack	2
PhysActivity	2
Fruits	2
Veggies	2
HvyAlcoholConsump	2
AnyHealthcare	2
NoDocbcCost	2
GenHlth	5
MentHlth	31
PhysHlth	31
DiffWalk	2
Sex	2
Age	13
Education	6
Income	8

Data Preprocessing

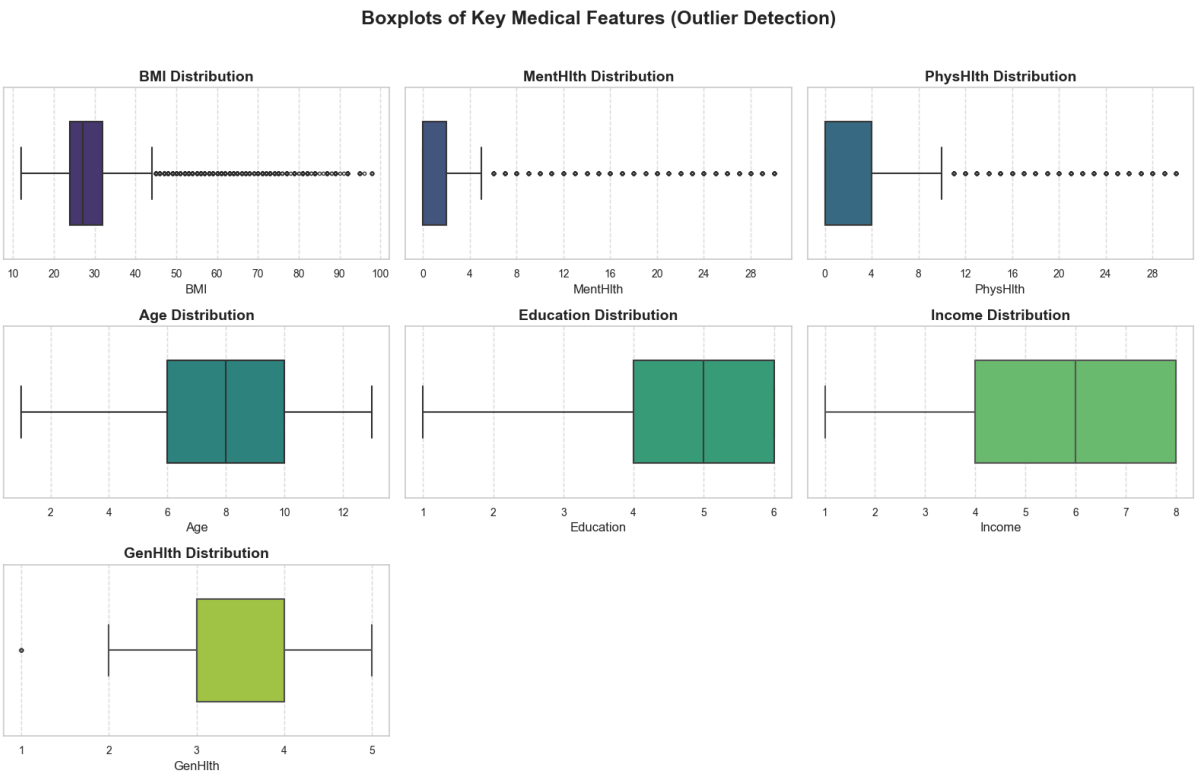
Value Distribution Overview

- Purpose: Assess class imbalance and feature frequency patterns.
- Created a reusable v_counts() function to print value distributions.
- Identified class imbalances in Diabetes_binary, Stroke, HeartDiseaseorAttack.
- Helpful for planning resampling or balancing strategies.

`CholCheck` is highly imbalanced (77.0%)	Imbalance
`Stroke` is highly imbalanced (75.5%)	Imbalance
`HeartDiseaseorAttack` is highly imbalanced (55.0%)	Imbalance
`HvyAlcoholConsump` is highly imbalanced (68.8%)	Imbalance
`AnyHealthcare` is highly imbalanced (71.8%)	Imbalance
`NoDocbcCost` is highly imbalanced (58.3%)	Imbalance

Feature Distribution Visualization

- Purpose: Visually detect outliers and data spread.
- Used Seaborn boxplots to visualize BMI, MentHlth, PhysHlth, etc.
- Outliers are visible especially in mental/physical health days (0-30 range).
- Useful for deciding normalization or outlier handling.



Data Preprocessing

Label Distribution (Binary/Ordinal Columns)

- Purpose: Evaluate balance and interpret feature scales.
- Features like HighBP, Smoker, PhysActivity are binary (0/1).
- Ordinal features like GenHlth, Education, and Income have ranked values.
- Important for choosing appropriate encodings (e.g., label vs. one-hot).

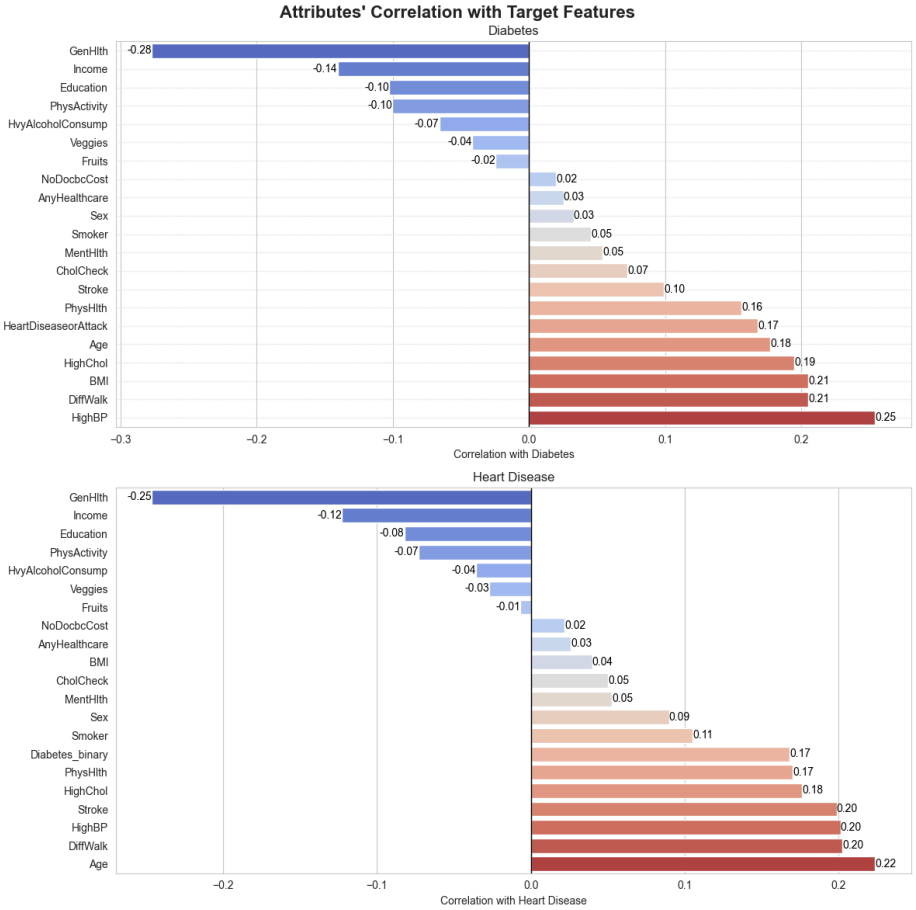
Dataset Final Check

- Purpose: Confirm dataset readiness for modeling.
- Ran `df.info()` again post-processing: all types now **int32**.
- Final count: **229,474** unique rows, no nulls, optimized memory.
- Dataset is clean, lean, and ready for training or feature engineering.

```
<class 'pandas.core.frame.DataFrame'>
Index: 229474 entries, 0 to 253679
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Diabetes_binary       229474 non-null  int32
 1   HighBP                229474 non-null  int32
 2   HighChol              229474 non-null  int32
 3   CholCheck             229474 non-null  int32
 4   BMI                   229474 non-null  int32
 5   Smoker                229474 non-null  int32
 6   Stroke                229474 non-null  int32
 7   HeartDiseaseorAttack  229474 non-null  int32
 8   PhysActivity          229474 non-null  int32
 9   Fruits                229474 non-null  int32
 10  Veggies               229474 non-null  int32
 11  HvyAlcoholConsump     229474 non-null  int32
 12  AnyHealthcare         229474 non-null  int32
 13  NoDocbcCost           229474 non-null  int32
 14  GenHlth               229474 non-null  int32
 15  MentHlth              229474 non-null  int32
 16  PhysHlth              229474 non-null  int32
 17  DiffWalk              229474 non-null  int32
 18  Sex                   229474 non-null  int32
 19  Age                   229474 non-null  int32
 20  Education             229474 non-null  int32
 21  Income                229474 non-null  int32
dtypes: int32(22)
memory usage: 21.0 MB
```

Exploratory Data Analysis

Understanding Risk Factors: How Different Health Factors Relate to Diabetes and Heart Disease



This visualization shows which health factors have the strongest relationships with diabetes and heart disease. The correlation values range from **-0.3** to **0.3**, where:

Positive values (red/orange bars to the right) indicate factors that increase risk.
Negative values (blue bars to the left) indicate factors that decrease risk.

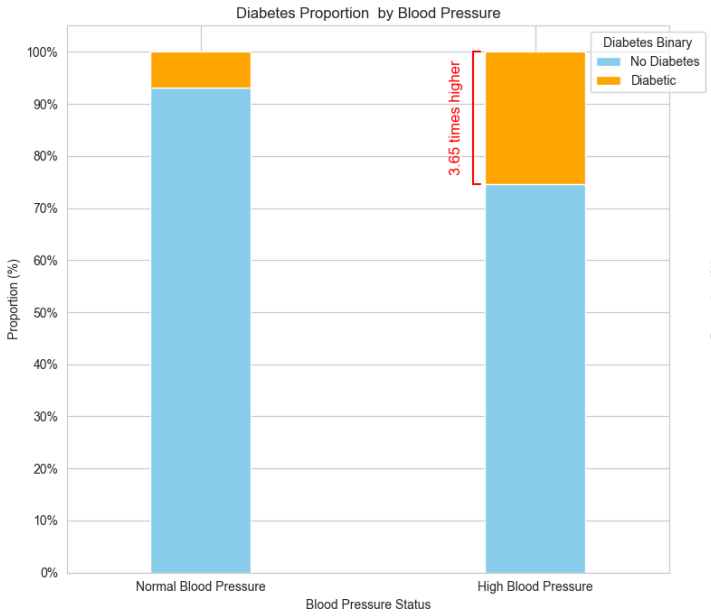
Key Insights:

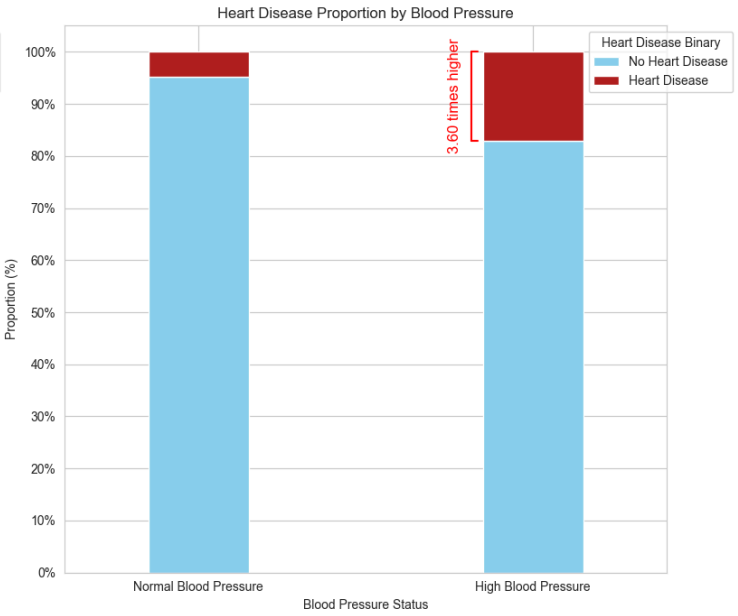
- For diabetes, high blood pressure **(0.25)**, high cholesterol **(0.19)**, and age **(0.18)** show the strongest positive associations
- For heart disease, age **(0.22)**, high blood pressure **(0.20)**, and stroke history **(0.20)** have the strongest relationships
- Good general health (**-0.28** for diabetes, **-0.25** for heart disease) and higher income (**-0.14** for diabetes, **-0.12** for heart disease) appear to be protective factors
- Physical activity shows a negative correlation with both conditions, suggesting that being active may help reduce risk

These patterns highlight the importance of managing blood pressure and cholesterol levels, especially as we age, to reduce the risk of both diabetes and heart disease.

Exploratory Data Analysis

Blood Pressure: A Critical Factor in Diabetes and Heart Disease Risk





These charts reveal a striking connection between high blood pressure and increased risk of both diabetes and heart disease.

For Diabetes:

- Among people with high blood pressure, about **25%** have diabetes (orange portion)
- In contrast, only about **7%** of those with normal blood pressure have diabetes
- This means you're about **3.65 times** more likely to have diabetes if you have high blood pressure

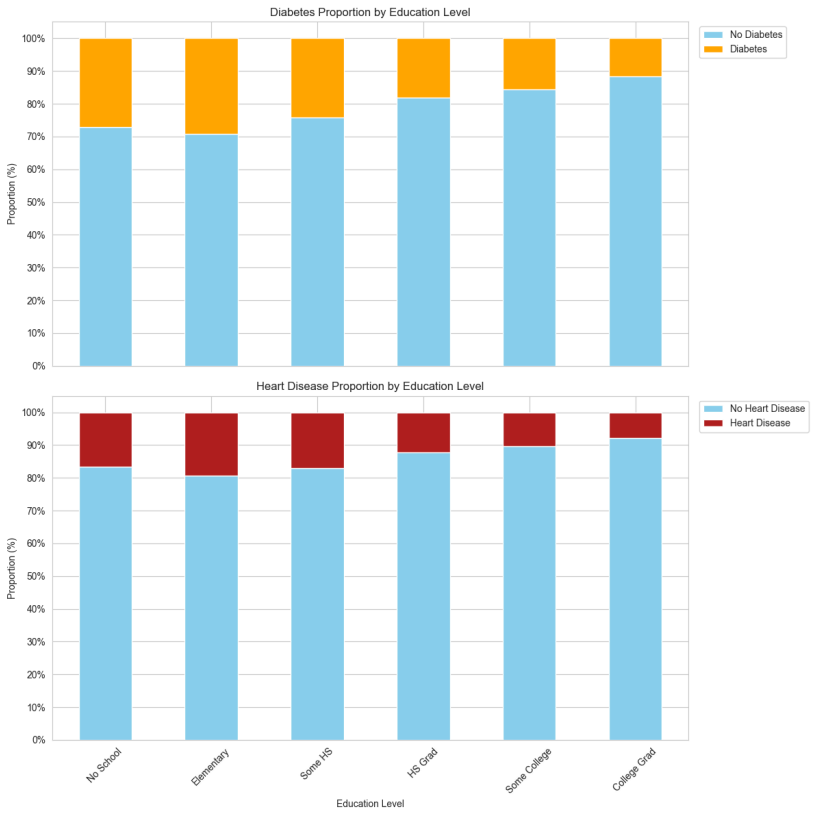
For Heart Disease:

- The risk is even more pronounced: around **17%** of people with high blood pressure have heart disease (red portion)
- Only about **5%** of those with normal blood pressure have heart disease
- This translates to being **3.6 times** more likely to have heart disease if you have high blood pressure

These findings emphasize the critical importance of maintaining healthy blood pressure levels. Regular blood pressure monitoring and management could significantly reduce your risk of developing these serious conditions.

Exploratory Data Analysis

Education's Impact on Diabetes and Heart Disease Risk



These charts reveal an interesting pattern: higher education levels appear to be associated with lower rates of both diabetes and heart disease.

For Diabetes:

- Those with no formal education have the highest diabetes rate at about 27% (orange portion)
- The rate gradually decreases as education increases
- College graduates have the lowest rate at about 12%
- This represents more than a 50% reduction in diabetes prevalence from the lowest to highest education level

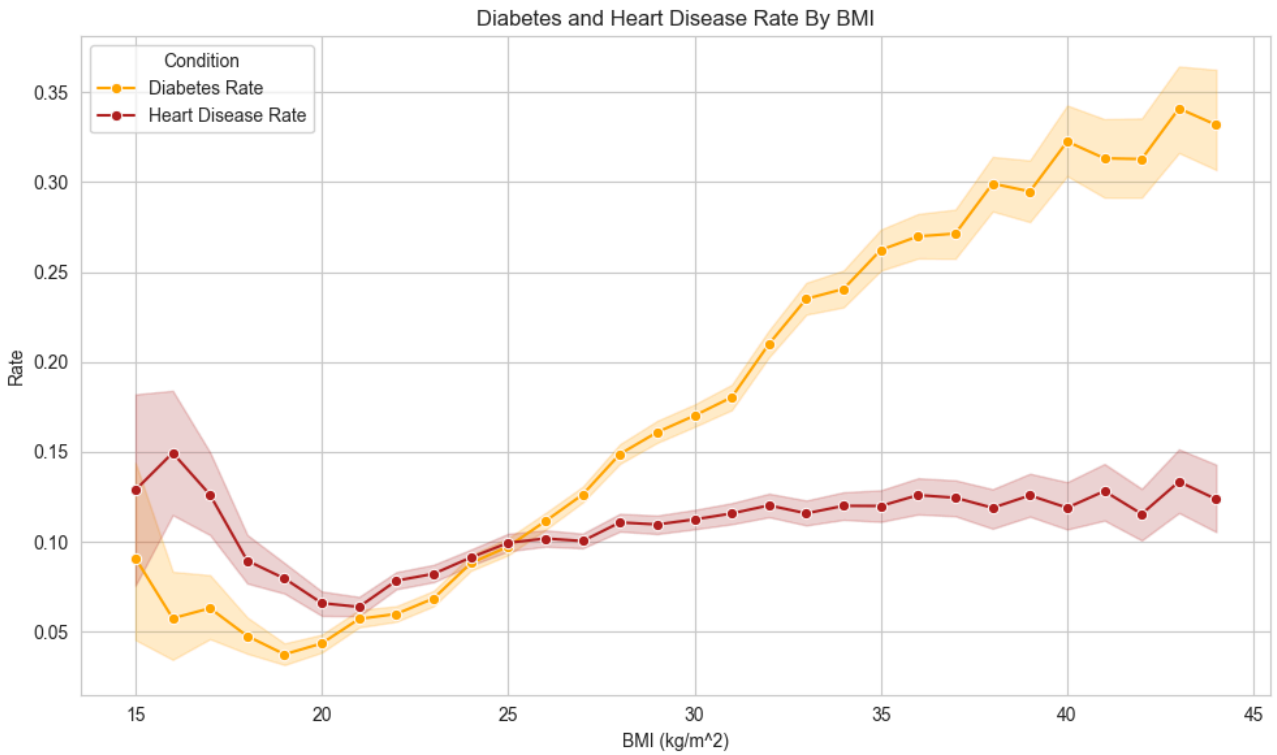
For Heart Disease:

- A similar pattern emerges, with about 17% of those with no formal education having heart disease (red portion)
- The rate decreases to about 7% among college graduates
- This trend is consistent across all education levels, showing a clear inverse relationship

These findings suggest that education may play a protective role against both conditions. This could be due to better health literacy, access to healthcare resources, and healthier lifestyle choices that often come with higher education levels.

Exploratory Data Analysis

BMI and Its Clear Connection to Diabetes and Heart Disease



This line graph tells a compelling story about how Body Mass Index (BMI) relates to the risk of both diabetes and heart disease.

Key Findings:

- For diabetes (orange line), there's a dramatic increase as BMI rises:
  - Low risk **(under 5%)** for BMI below **20**
  - Moderate risk **(5-15%)** for BMI between **20-30**
  - High risk **(15-35%)** for BMI above **30**
- For heart disease (red line), the relationship is more stable:
  - Risk generally stays between **10-15%** across all BMI levels
  - Slight increase with higher BMI, but not as steep as diabetes

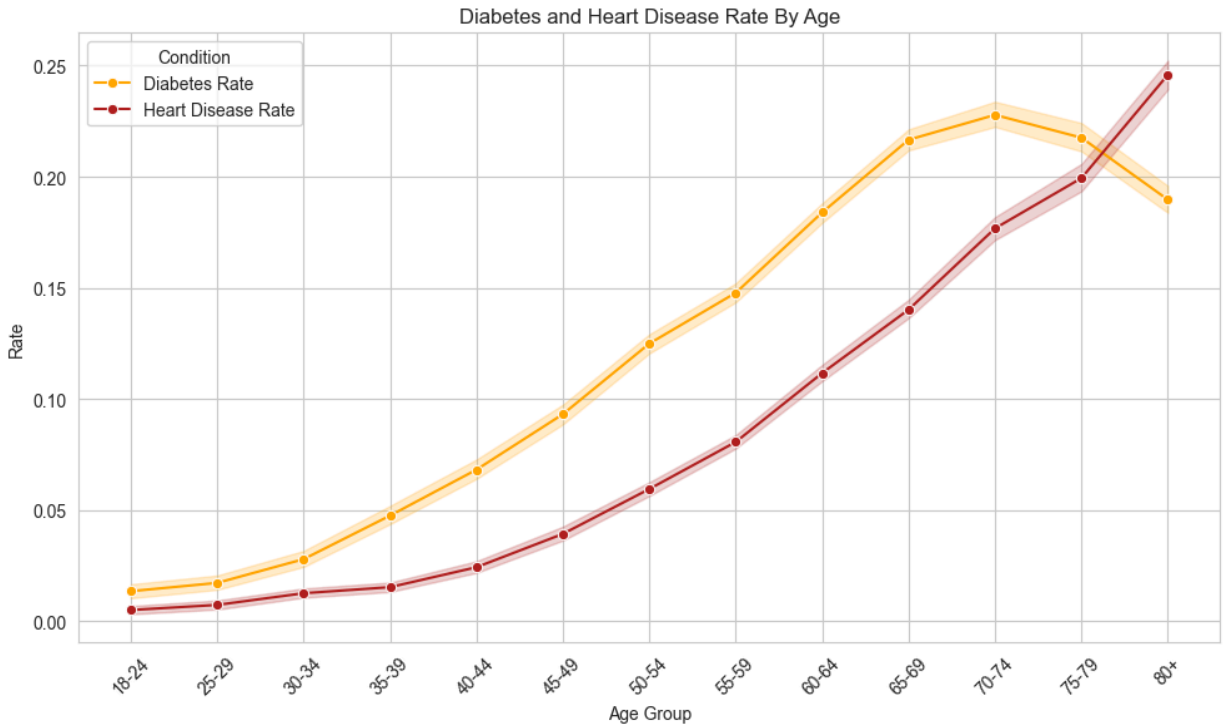
Important Insights:

- At a BMI of **20**, diabetes risk is at its lowest **(around 4%)**
- After BMI **25**, diabetes risk begins to climb sharply
- By BMI **35**, diabetes risk triples to over **25%**
- Heart disease maintains a relatively consistent **10-13%** risk regardless of BMI

This data strongly suggests that maintaining a healthy BMI, particularly below **25**, is crucial for diabetes prevention. While BMI also affects heart disease risk, the impact is less pronounced.

Exploratory Data Analysis

How Age Influences Your Risk of Diabetes and Heart Disease



Diabetes Risk (Orange Line):

- Young adults (18-24) have minimal risk at just **1.4%**
- Risk gradually increases through middle age
- Peaks at age 70-74 with **23%** risk
- Slightly lower rates in the 74+ age group may be due to higher mortality among individuals with diabetes, resulting in fewer surviving older adults with the condition.

Heart Disease Risk (Red Line):

- Starts very low at **0.7%** for young adults
- Rises more slowly than diabetes through middle age
- Continues climbing steadily to **25%** for those 80+
- Unlike diabetes, doesn't decline in the oldest age group

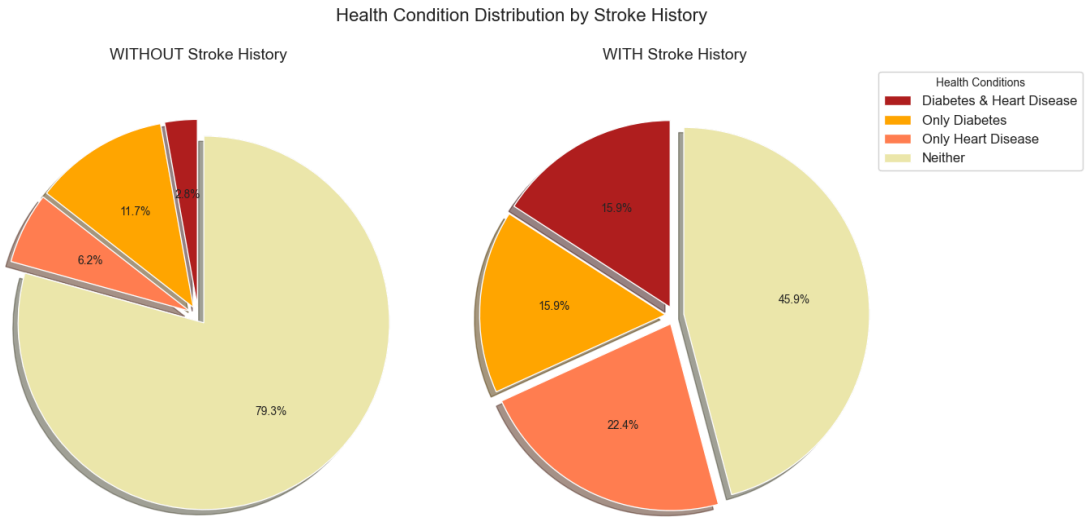
Critical Observations:

- Both conditions become significantly more common after age 45
- Diabetes risk accelerates faster in middle age (45-64)
- Heart disease risk continues to rise through all age groups
- By age 80+, both conditions affect nearly 1 in 4 individuals

This pattern emphasizes the importance of preventive health measures as we age, particularly for those entering their 40s and beyond. Regular health screenings become increasingly vital with each passing decade.

Exploratory Data Analysis

The Connection Between Stroke History and Other Health Conditions



These pie charts reveal a striking relationship between stroke history and the prevalence of diabetes and heart disease.

For People Without Stroke History:

- **79.3%** have neither diabetes nor heart disease
- **11.7%** have only diabetes
- **6.2%** have only heart disease
- **2.8%** have both conditions

For People With Stroke History:

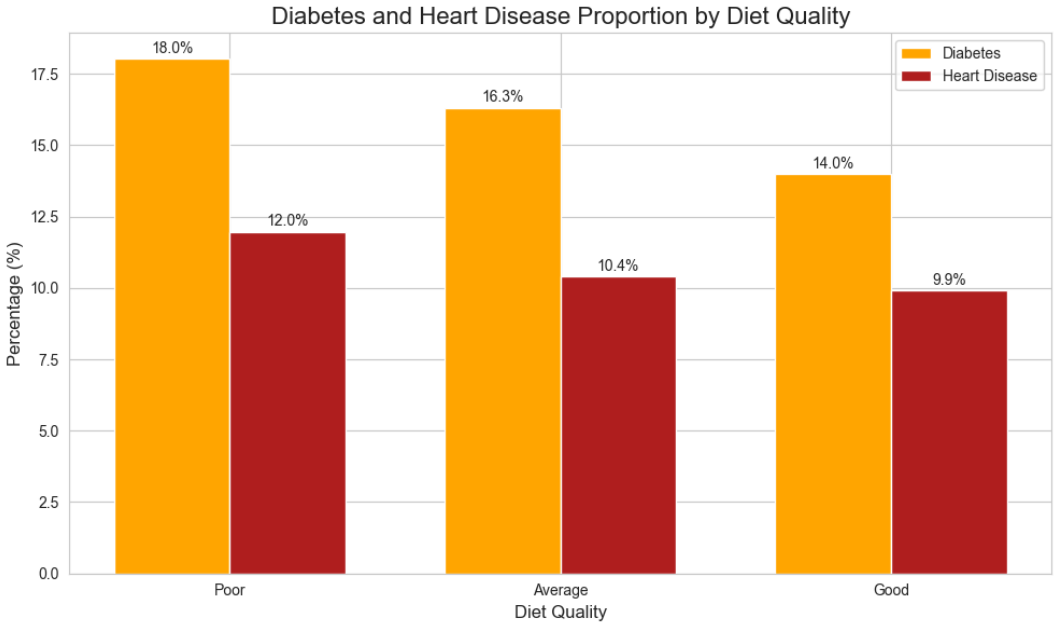
- Only **45.9%** have neither diabetes nor heart disease
- **15.9%** have only diabetes
- **22.4%** have only heart disease
- **15.9%** have both conditions

Key Insights:

- Having a stroke history dramatically increases the likelihood of having other conditions
- Heart disease is **3.6 times** more common in stroke survivors **(22.4% vs 6.2%)**
- People with stroke history are **5.7 times** more likely to have both diabetes and heart disease **(15.9% vs 2.8%)**
- Over half **(54.1%)** of stroke survivors have at least one other condition, compared to just **20.7%** of those without stroke history

This data highlights that stroke survivors face an elevated risk of developing diabetes and heart disease, indicating a strong interconnectedness between these conditions.

Diet Quality Matters: The Impact of Healthy Eating on Diabetes and Heart Disease



This bar chart demonstrates a clear relationship between diet quality and the prevalence of both diabetes and heart disease.

For Diabetes (Orange Bars):

- Poor diet: **18.0%** prevalence
- Average diet: **16.3%** prevalence
- Good diet: **14.0%** prevalence

For Heart Disease (Red Bars):

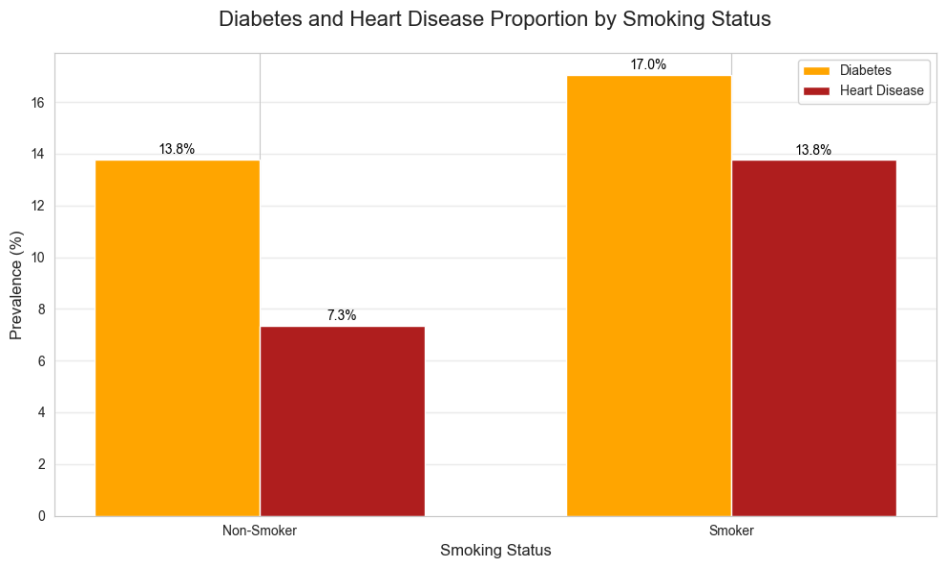
- Poor diet: **12.0%** prevalence
- Average diet: **10.4%** prevalence
- Good diet: **9.9%** prevalence

Important Insights:

- Both conditions show a consistent decrease as diet quality improves
- People with poor diets are **1.3 times** more likely to have diabetes compared to those with good diets
- Heart disease risk is about **1.2 times** higher in those with poor diets versus good diets
- Improving diet from poor to good quality could reduce diabetes risk by **22%** and heart disease risk by **17.5%**

This data clearly shows that making better dietary choices – such as eating more fruits and vegetables – can significantly reduce your risk of developing these serious health conditions.

Smoking's Significant Impact on Diabetes and Heart Disease Risk



This bar chart reveals the concerning relationship between smoking status and the prevalence of both diabetes and heart disease.

Key Statistics:

For Diabetes (Orange Bars):

- Non-smokers: **13.8%** prevalence
- Smokers: **17.0%** prevalence
- Increase: **3.2** percentage points **(23% higher risk)**

For Heart Disease (Red Bars):

- Non-smokers: **7.3%** prevalence
- Smokers: **13.8%** prevalence
- Increase: **6.5** percentage points **(89% higher risk)**

Critical Insights:

- Smoking nearly doubles your risk of heart disease **(from 7.3% to 13.8%)**
- Diabetes risk increases by nearly a quarter for smokers
- Smokers have the same rate of heart disease **(13.8%)** as non-smokers have diabetes
- The impact on heart disease is more severe than on diabetes

This data provides compelling evidence that quitting smoking or never starting can significantly reduce your risk of developing these serious health conditions, particularly heart disease.

Closing Thoughts

Health risks like high blood pressure or diabetes often develop silently over time. What makes this data powerful is how it reveals patterns we can actually do something about—whether it's getting checked more often or making small, sustainable changes. Your future self will thank you.

PowerBI Analysis Dashboard

This comprehensive analysis dashboard provides valuable insights into diabetes prevalence, risk factors, and health correlations across a sample of **229,000 patients**. The dashboard consists of multiple pages that explore key metrics, demographic distributions, behavioral factors, and comorbidities associated with diabetes and heart disease. The analysis reveals significant relationships between diabetes prevalence and factors such as BMI, physical activity, diet quality, blood pressure, and heart health.

Dashboard Overview:

Homepage

The dashboard begins with a navigation pane allowing access to:

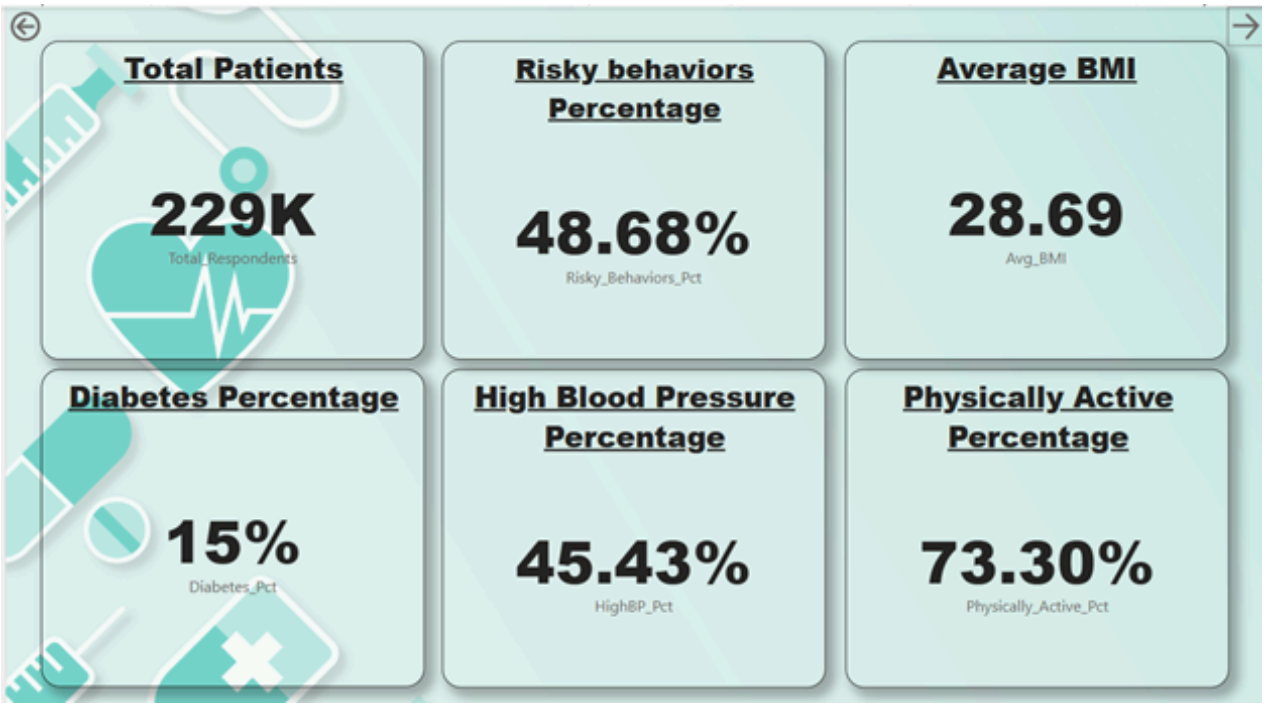
- HomePage
- Page 1 (Key Metrics)
- Page 2 (Diabetes Demographics and Correlations)
- Page 3 (Additional Analysis)
- Page 4 (Heart Disease Analysis)
- Page 5 (Lifestyle Factors)

PowerBI Analysis Dashboard

Page 1: Key Metrics

This page presents high-level statistics about the patient population:

- **Total Patients**: **229,000** respondents
- **Diabetes Prevalence**: **15%** of respondents have diabetes
- **High Blood Pressure**: Affects **45.43%** of patients
- **Physical Activity**: **73.30%** of patients report being physically active
- **Average BMI**: **28.69** (classified as overweight)
- **Risky Behaviors**: **48.68%** of patients engage in risky health behaviors

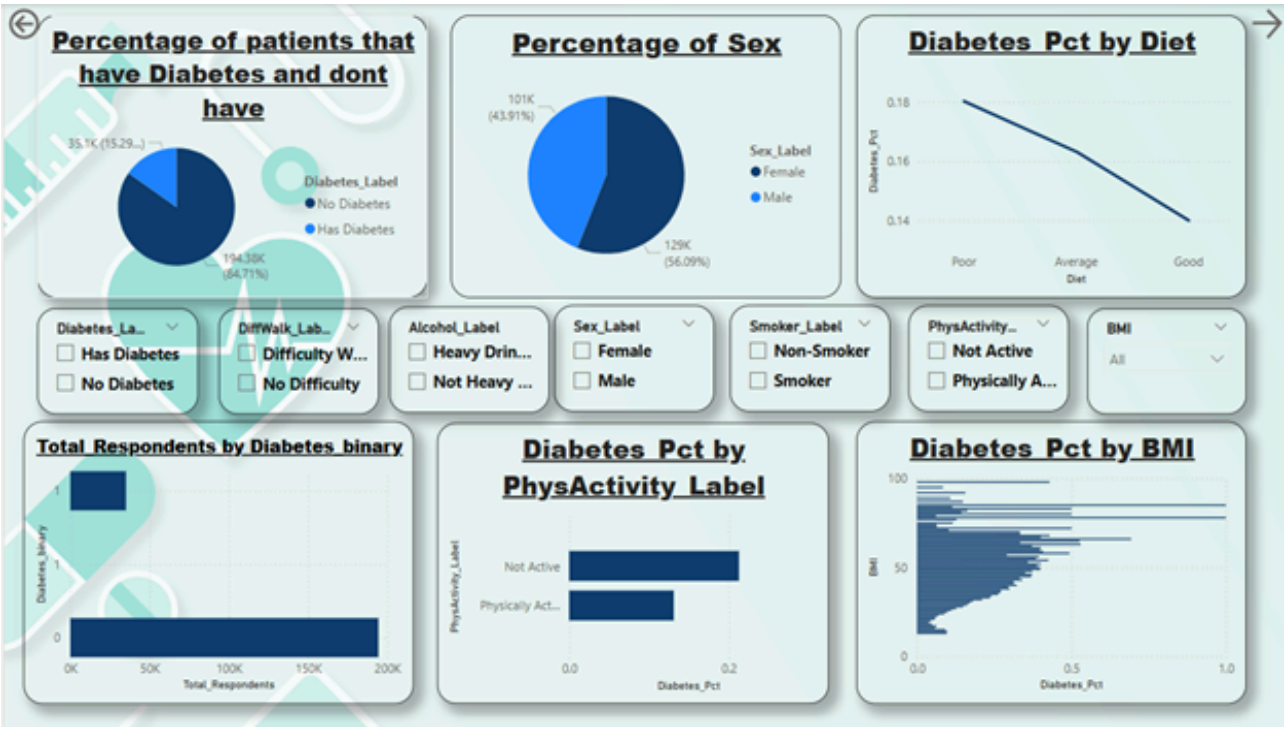


PowerBI Analysis Dashboard

Page 2: Diabetes Demographics and Correlations

This page provides detailed breakdowns of diabetes prevalence across various factors:

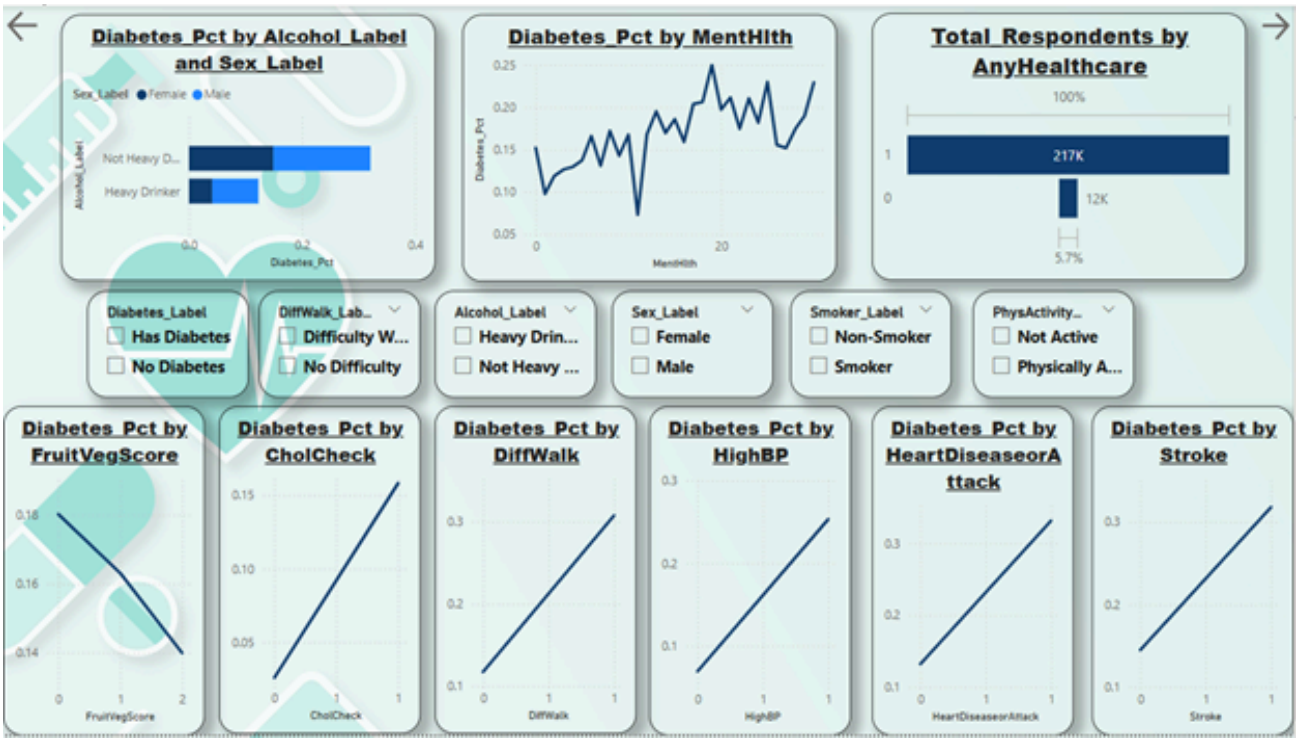
- **Diabetes Distribution**: **15% (35.1K)** have diabetes vs. **85% (194.58K)** without
- **Gender Distribution**: **56.09% male** vs. **43.91% female**
- **Diet Impact**: Diabetes prevalence decreases with better diet quality (**18% poor diet vs. 14% average diet**)
- **Physical Activity**: Shows clear correlation with lower diabetes rates
- **Smoking Status**: Reveals differences in diabetes prevalence between smokers and non-smokers
- **Alcohol Consumption**: Heavy drinkers show distinct diabetes patterns
- **Mobility Issues**: Patients reporting difficulty walking have higher diabetes prevalence



PowerBI Analysis Dashboard

Page 3: Additional Visualizations

- **Mental Health Impact**: Shows relationship between mental health and diabetes prevalence
- **Healthcare Access**: Reveals disparities in diabetes prevalence based on healthcare access
- **Comorbidity Analysis**: Demonstrates how high blood pressure and stroke correlate with diabetes

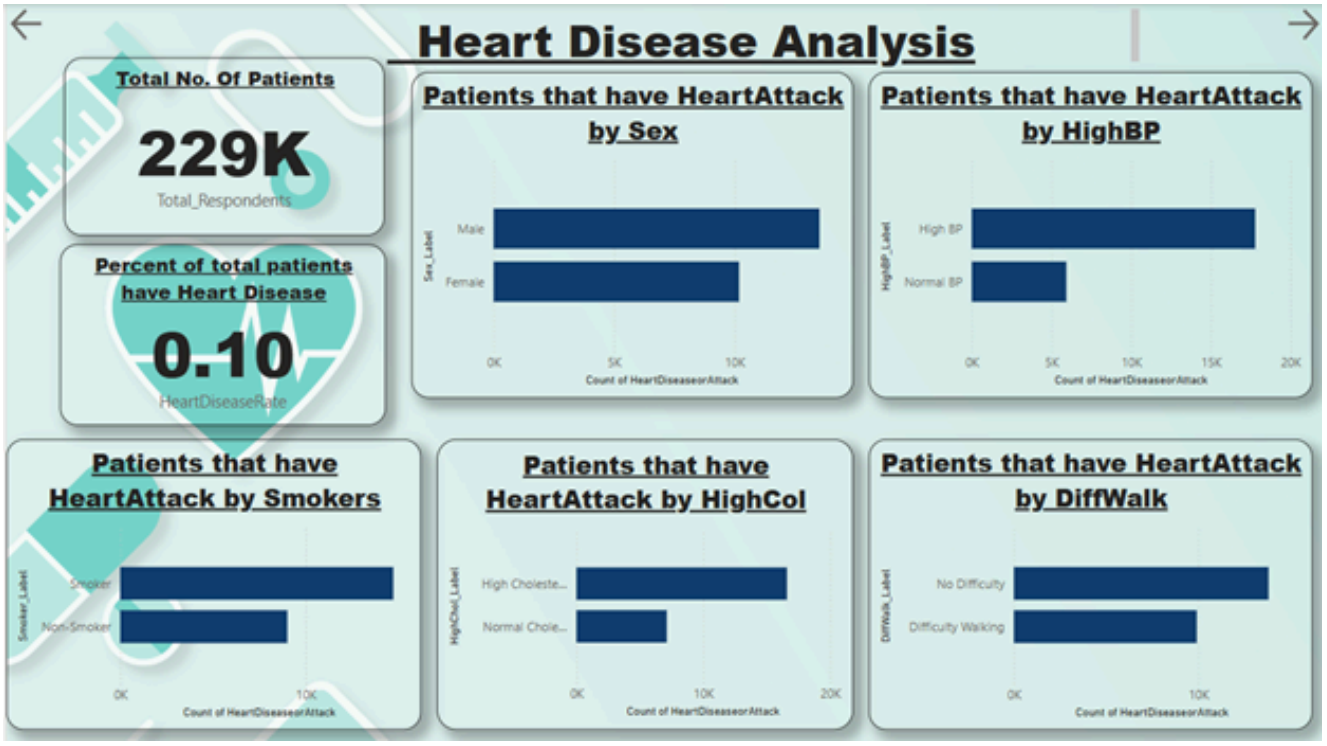


PowerBI Analysis Dashboard

Page 4: Heart Disease Analysis

This section explores the relationship between cardiovascular health and different factors

- **Heart Disease Prevalence**: **10%** of patients have heart disease
- **Gender Differences**: Males show **higher rates** of heart disease than females
- **Blood Pressure Impact**: Patients with **high BP** have significantly **higher heart disease rates**
- **Cholesterol Levels**: High cholesterol correlates strongly with heart disease
- **Physical Activity**: Active patients show lower heart disease incidence

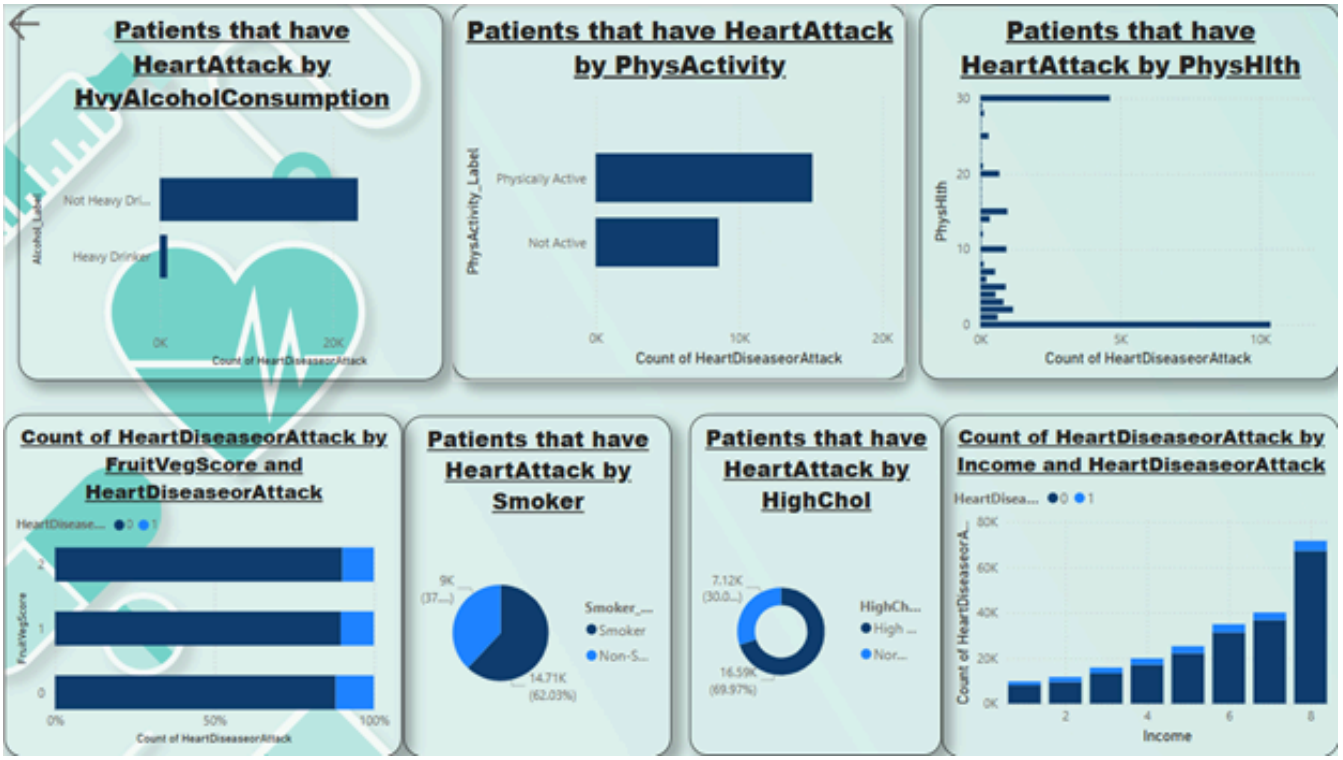


PowerBI Analysis Dashboard

Page 5: Lifestyle Factors

This page examines how lifestyle choices affect health outcomes:

- **Fruit/Vegetable Consumption:** Shows inverse relationship with both diabetes and heart disease
- **Income Levels:** Reveals socioeconomic disparities in health outcomes
- **Physical Health:** Demonstrates clear benefits of physical activity
- **Alcohol Consumption:** Heavy drinking correlates with negative health outcomes



PowerBI Analysis Dashboard

Key Insights:

1. **Diabetes Prevalence**: The **15%** diabetes rate in this population is significantly higher than the global average of **9.3%**, suggesting this may be a high-risk or clinical population.

2. **BMI Correlation**: The average BMI of **28.69 (overweight)** strongly correlates with the elevated diabetes rates, supporting known links between obesity and diabetes.

3. **Physical Activity Benefit**: The **73.30%** physical activity rate is relatively high, yet the active population shows markedly lower diabetes prevalence (visualized on Page 2).

4. **Gender Disparities**: Males **(56.09%)** are overrepresented in this dataset and show different health patterns than females across multiple metrics.

5. **Diet Quality Impact**: The clear gradient from poor diet **(18% diabetes)** to average diet (14% diabetes) demonstrates the protective effect of better nutrition.

6. **Comorbidity Patterns**: The strong correlation between high blood pressure **(45.43% prevalence)** and both diabetes and heart disease suggests metabolic syndrome is prevalent in this population.

7. **Behavioral Factors**: Nearly half **(48.68%)** of patients engage in risky behaviors, which aligns with the elevated disease prevalence rates.

8. **Socioeconomic Factors**: Visualizations showing income and healthcare access disparities reveal important social determinants of health in diabetes prevalence.

9. **Lifestyle Interventions**: The consistent patterns across physical activity, diet, and substance use suggest significant opportunities for preventive health interventions.

Conclusion

This Power BI dashboard successfully integrates multiple data dimensions to provide a comprehensive view of diabetes prevalence and its associated factors. The visualizations clearly demonstrate the multifactorial nature of diabetes risk and the importance of addressing both medical and lifestyle factors. The insights generated can directly inform clinical decision-making, population health strategies, and preventive care programs. Future enhancements could include longitudinal tracking of intervention outcomes and more detailed socioeconomic analysis.

Project Pipeline

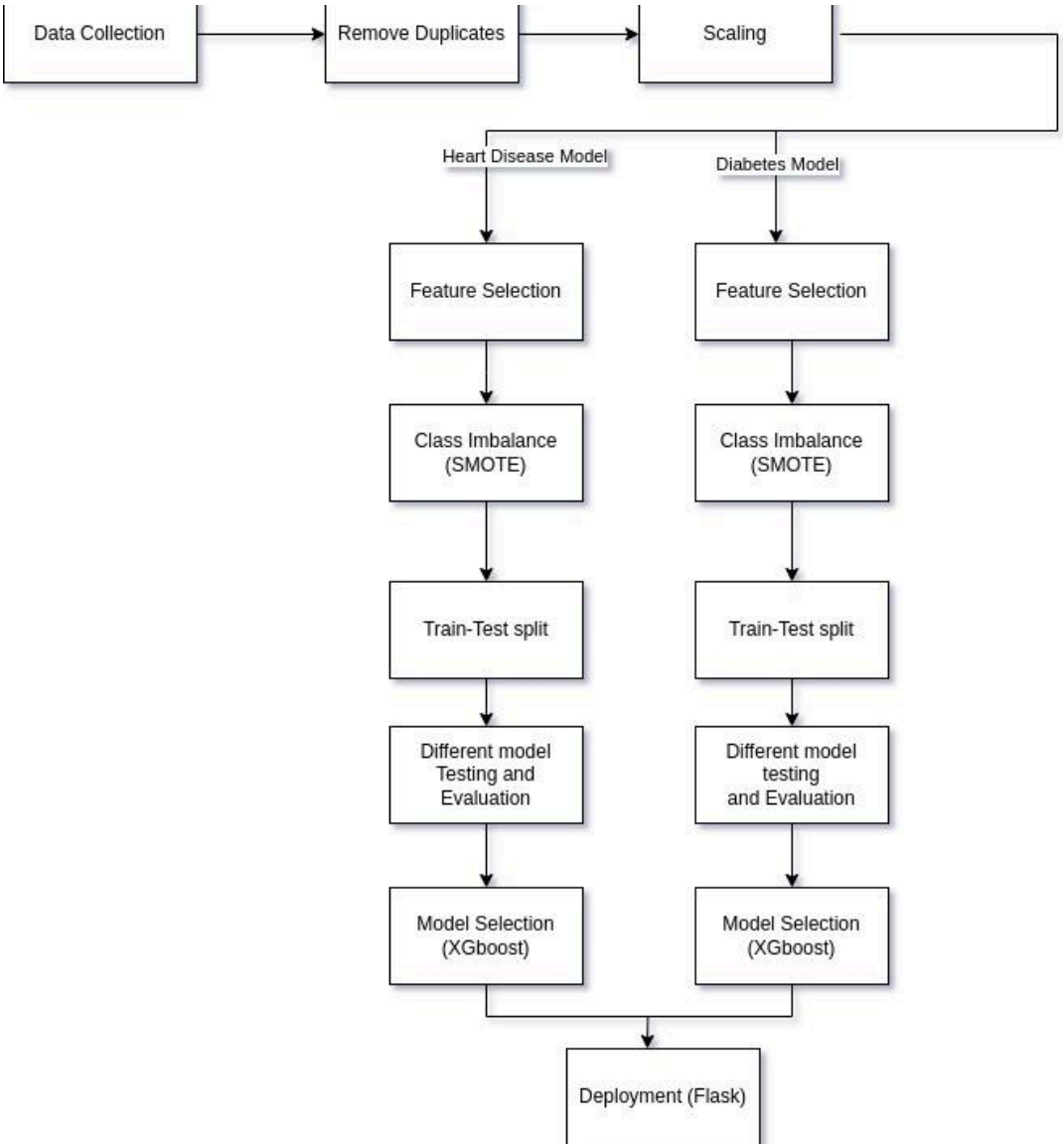
The project pipeline is designed to build effective prediction models for both heart disease and diabetes. It starts with **data collection**, followed by cleaning steps like **removing duplicates** and **scaling** the data to ensure consistency.

After preprocessing, the pipeline splits into two parallel paths—one for the **heart disease model** and the other for the **diabetes model**. Both paths follow the same key steps:

- **Feature selection** to identify the most important variables.
- **SMOTE (Synthetic Minority Over-sampling Technique)** to handle class imbalance.
- **Train-test splitting** to evaluate model performance fairly.
- **Testing and evaluating different machine learning models** to find the best fit.

Ultimately, **XGBoost** was selected for both models due to its strong performance.

Finally, the trained models were deployed using **Flask**, making them accessible through a simple web interface for real-time predictions.

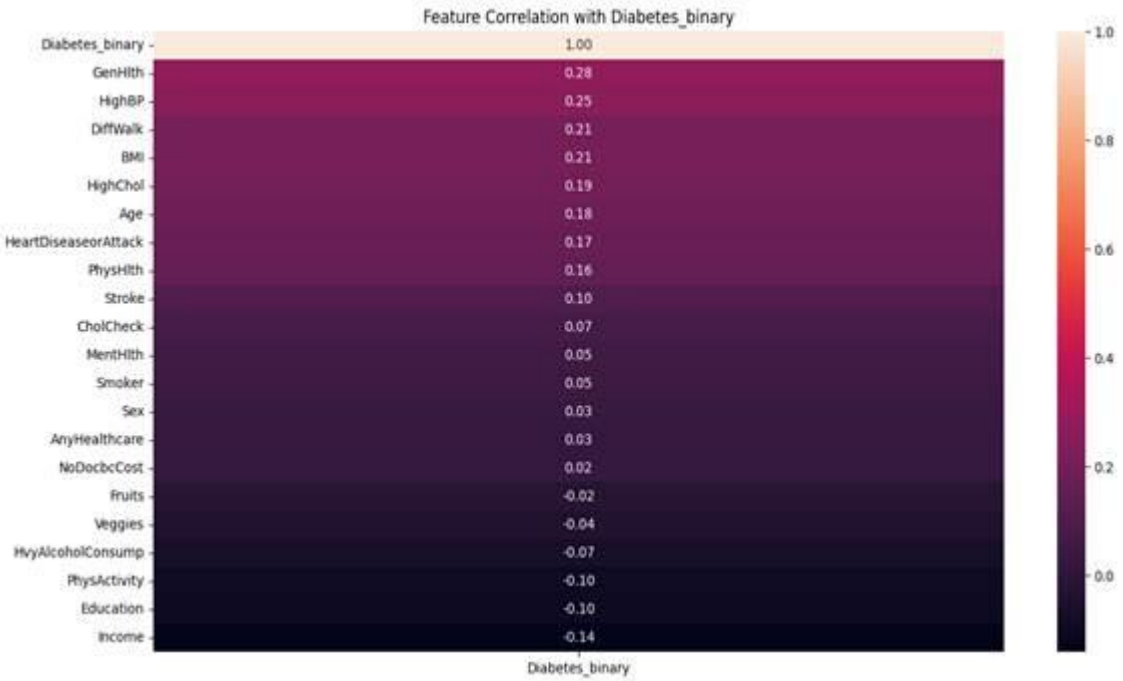


Diabetes Model

Feature Selection

In this phase, I used a (correlation matrix – Chi-Squared – ANOVA) to choose the features that had an influence on the target column.

- ➢ Correlation matrix



Chosen Features:

- Positive correlated: `HighBP`, `HighChol`, `BMI`, `DiffWalk`, `GenHlth`, `Age`, `HeartDiseaseorAttack`, `PhysHlth` and `Stroke`

- Negative correlated: `PhysActivity`, `Education` and `income`

Diabetes Model

➢ Chi-Squared Test

Best Features By Chi-Squared Test:

1. HighBP
2. DiffWalk
3. HeartDiseaseorAttack
4. HighChol
5. PhysHlth
6. GenHlth
7. Stroke

Feature	Value
HighBP	8098.548237
DiffWalk	7875.496177
HeartDiseaseorAttack	5822.145697
HighChol	4869.312739
PhysHlth	3266.292056
GenHlth	3115.279621
Stroke	2156.678382
HvyAlcoholConsump	937.401148
Age	812.103743
PhysActivity	617.563886
Income	581.129755
MentHlth	380.652825
BMI	309.992581
Smoker	253.826098
Sex	137.837135
Education	119.900184
NoDocbcCost	83.662830
Veggies	82.098846
Fruits	54.688897
CholCheck	48.904140
AnyHealthcare	7.949731

Diabetes Model

➢ ANOVA Test

Best Features By ANOVA Test

1. GenHlth
2. HighBP
3. DiffWalk
4. BMI
5. HighChol
6. Age
7. HeartDiseaseorAttack

Feature	Value
GenHlth	19061.436601
HighBP	15868.017696
DiffWalk	10097.616798
BMI	10075.425186
HighChol	9065.177413
Age	7444.458687
HeartDiseaseorAttack	6682.140008
PhysHlth	5739.591424
Income	4631.714144
Education	2445.440895
PhysActivity	2336.872150
Stroke	2280.282576
CholCheck	1213.321927
HvyAlcoholConsump	1002.426607
MentHlth	674.908702
Smoker	476.133601
Veggies	400.370333
Sex	245.998296
AnyHealthcare	147.341358
Fruits	141.282132
NoDocbcCost	92.269537

Diabetes Model

Final Features For The ML Models

- ❖ HighBP
- ❖ GenHlth
- ❖ DiffWalk
- ❖ BMI
- ❖ HighChol
- ❖ HeartDiseaseorAttack
- ❖ PhysHlth
- ❖ Age
- ❖ Stroke
- ❖ Income

Model Testing

Four models were tested with pipeline

- Logistic Regression
- Decision Tree
- Random Forest
- XGBoost

Each model was tested on:
- Balanced data (after applying SMOTE)

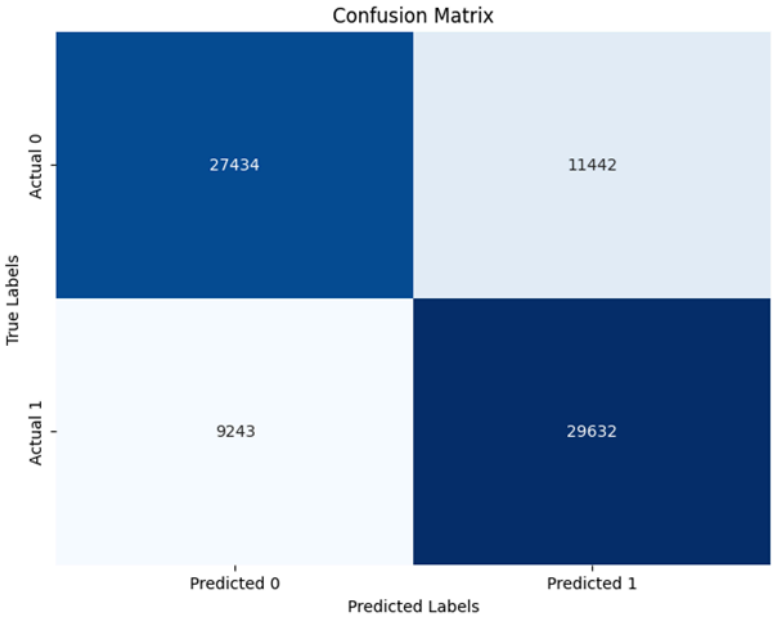
Metrics used: Accuracy, F1-Score, Recall, Precision

Diabetes Model

1. Logistic Regression

Label	Precision	Recall	F1-Score
0.0	0.75	0.71	0.73
1.0	0.72	0.76	0.74
Macro Avg	0.73	0.73	0.73
Weighted Avg	0.73	0.73	0.73

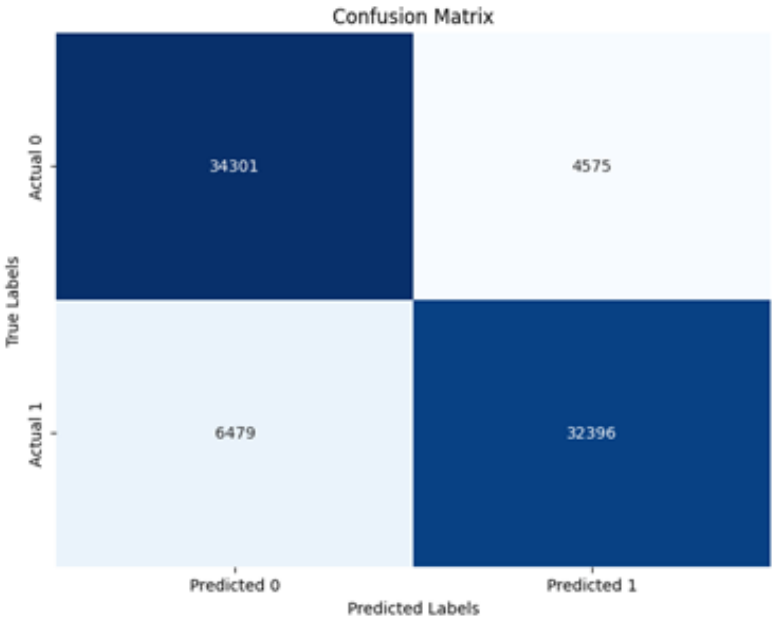
Accuracy: 0.73



2. Decision Tree

Label	Precision	Recall	F1-Score
0.0	0.84	0.88	0.86
1.0	0.88	0.83	0.85
Macro Avg	0.86	0.86	0.86
Weighted Avg	0.86	0.86	0.86

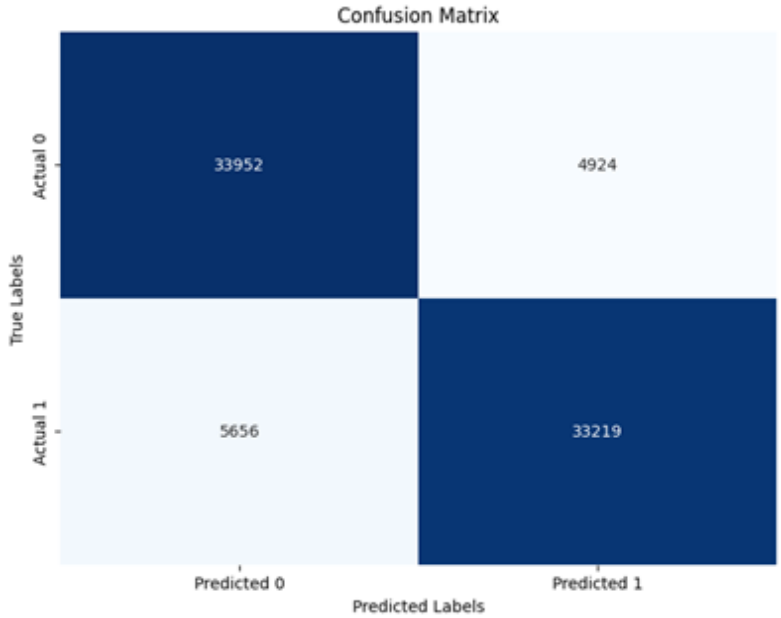
Accuracy: 0.86



3. Random Forest

Label	Precision	Recall	F1-Score
0.0	0.86	0.87	0.87
1.0	0.87	0.85	0.86
Macro Avg	0.86	0.86	0.86
Weighted Avg	0.86	0.86	0.86

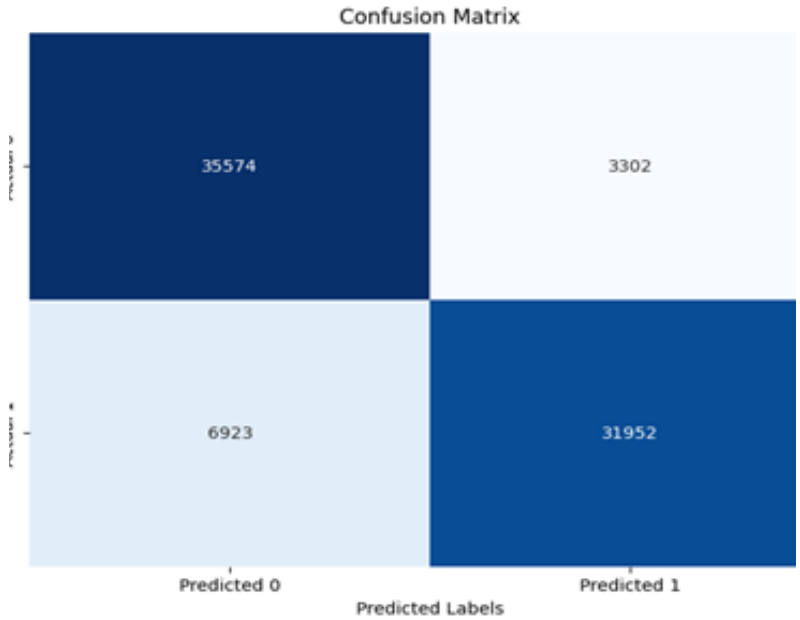
Accuracy: 0.86



4. XGBoost

Label	Precision	Recall	F1-Score
0.0	0.84	0.92	0.87
1.0	0.91	0.82	0.86
Macro Avg	0.87	0.87	0.87
Weighted Avg	0.87	0.87	0.87

Accuracy: 0.87



Diabetes Model

Summary

After testing all models, the results show that balanced data leads to significantly better overall performance.

The best model was:
- XGBoost

These were selected for hyperparameter tuning using GridSearchCV.

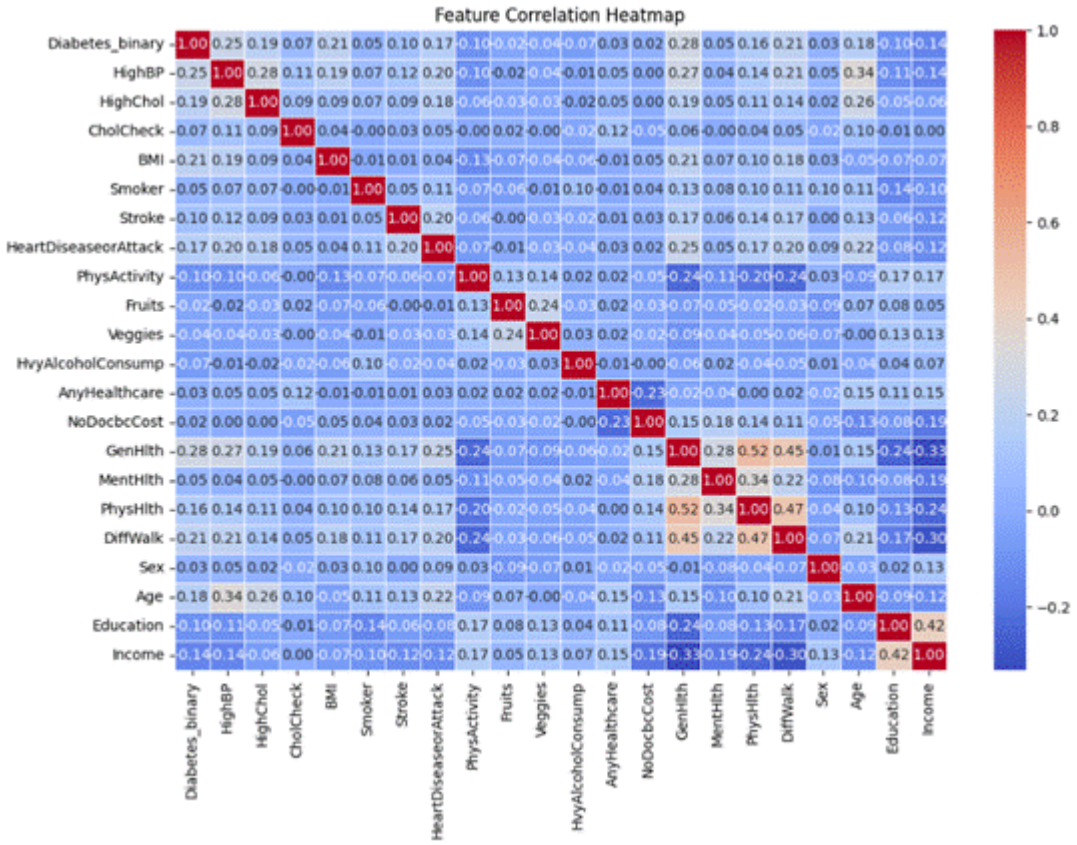
XGBoost (Best Parameters)

- colsample_bytree: 1
- learning_rate: 0.1
- max_depth: 10
- n_estimators: 200
- subsample: 0.8

Heart Disease Model

Feature Selection

After data was preprocessed, I used a correlation matrix to choose the features that had a high correlation with my target column.



Chosen Features:

Feature	Correlation
highbp	0.20
diabetes	0.17
highcol	0.18
smoker	0.11
stroke	0.20
gen health	0.25
age	0.22
diffwalk	0.20
phyhealth	0.17
income	-0.12

Heart Disease Model

Model Testing

- Four models were tested:
- Random Forest
- Logistic Regression
- XGBoost
- Pipeline (Polynomial Features + Logistic Regression)

Each model was tested on:

- Imbalanced data
- Balanced data (after applying SMOTE)

Metrics used: Accuracy, F1-Score, Recall, Precision

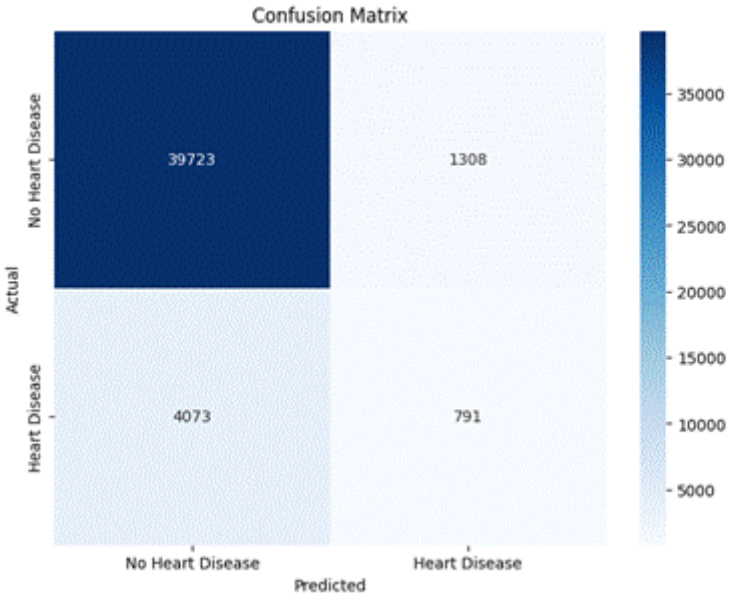
Heart Disease Model

1. Random Forest

Imbalanced

Label	Precision	Recall	F1-Score
0.0	0.91	0.97	0.94
1.0	0.38	0.16	0.23
Macro Avg	0.64	0.57	0.58
Weighted Avg:	0.85	0.88	0.86

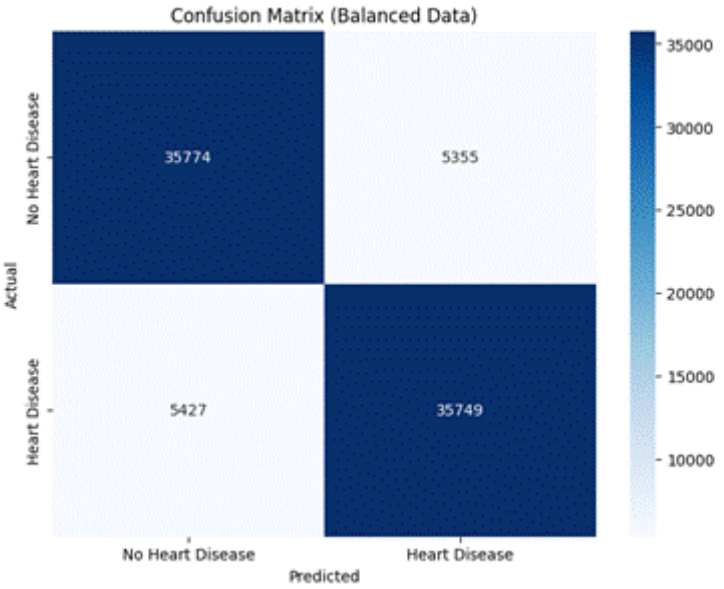
Accuracy: 0.88



Balanced

Label	Precision	Recall	F1-Score
0.0	0.87	0.87	0.87
1.0	0.87	0.87	0.87
Macro Avg	0.87	0.87	0.87
Weighted Avg:	0.87	0.87	0.87

Accuracy: 0.87

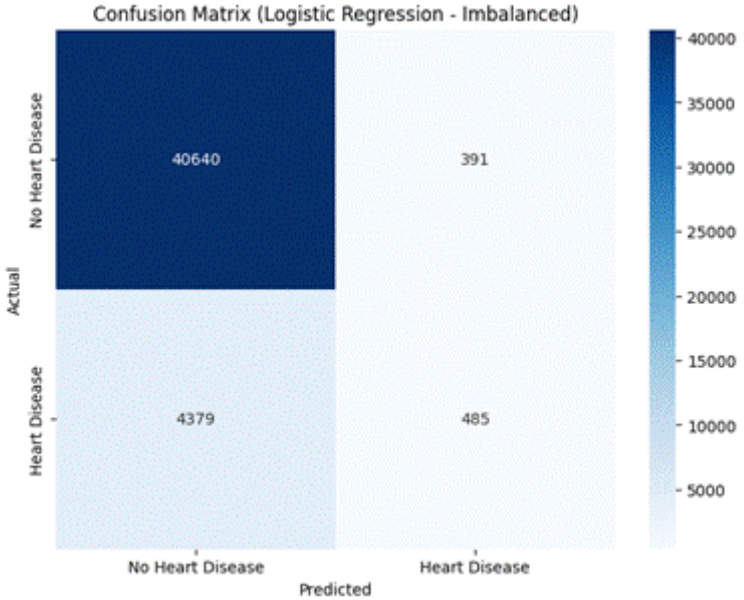


2. Logistic Regression

Imbalanced

Label	Precision	Recall	F1-Score
0.0	0.90	0.99	0.94
1.0	0.55	0.10	0.17
Macro Avg	0.72	0.55	0.56
Weighted Avg:	0.87	0.90	0.86

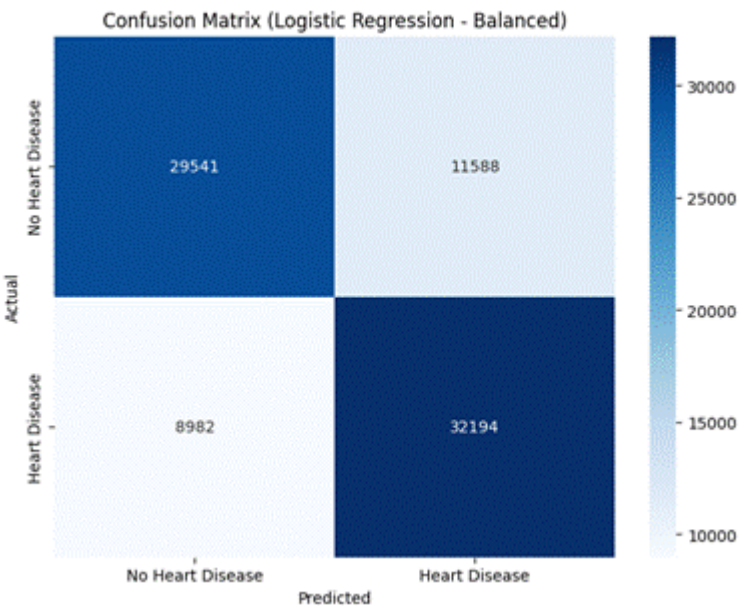
Accuracy: 0.90



Balanced

Label	Precision	Recall	F1-Score
0.0	0.77	0.72	0.74
1.0	0.74	0.78	0.76
Macro Avg	0.75	0.75	0.75
Weighted Avg:	0.75	0.75	0.75

Accuracy: 0.75

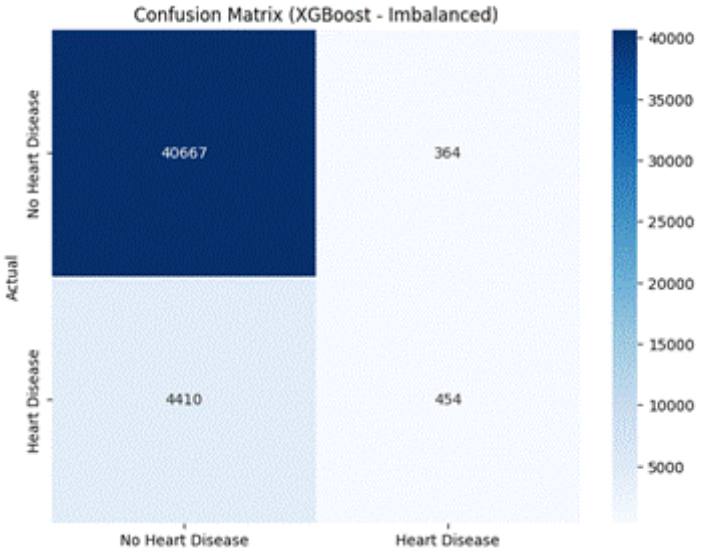


3. XGBoost

Imbalanced

Label	Precision	Recall	F1-Score
0.0	0.90	0.99	0.94
1.0	0.56	0.09	0.16
Macro Avg	0.72	0.54	0.55
Weighted Avg:	0.87	0.90	0.86

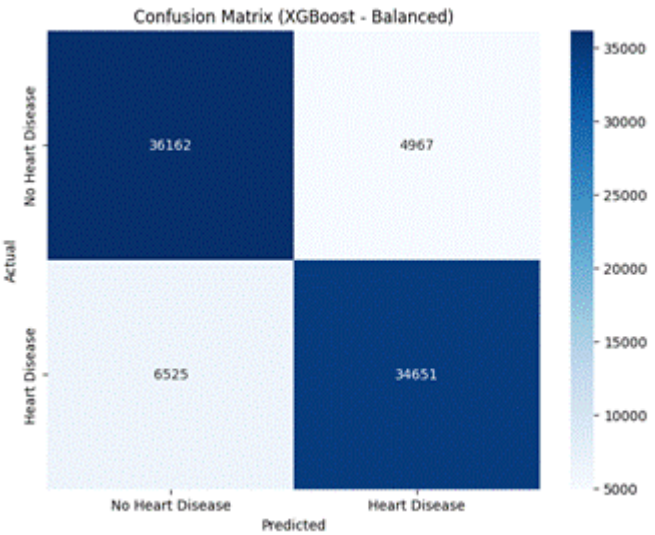
Accuracy: 0.90



Balanced

Label	Precision	Recall	F1-Score
0.0	0.85	0.88	0.86
1.0	0.87	0.84	0.86
Macro Avg	0.86	0.86	0.86
Weighted Avg:	0.86	0.86	0.86

Accuracy: 0.86

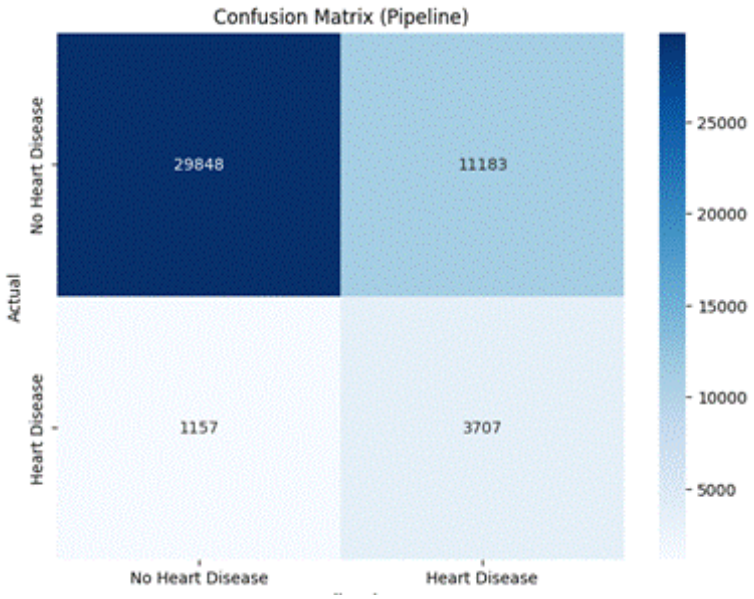


4. Pipeline (Polynomial Features + Logistic Regression)

Imbalanced

Label	Precision	Recall	F1-Score
0.0	0.96	0.73	0.83
1.0	0.25	0.76	0.38
Macro Avg	0.61	0.74	0.60
Weighted Avg:	0.89	0.73	0.78

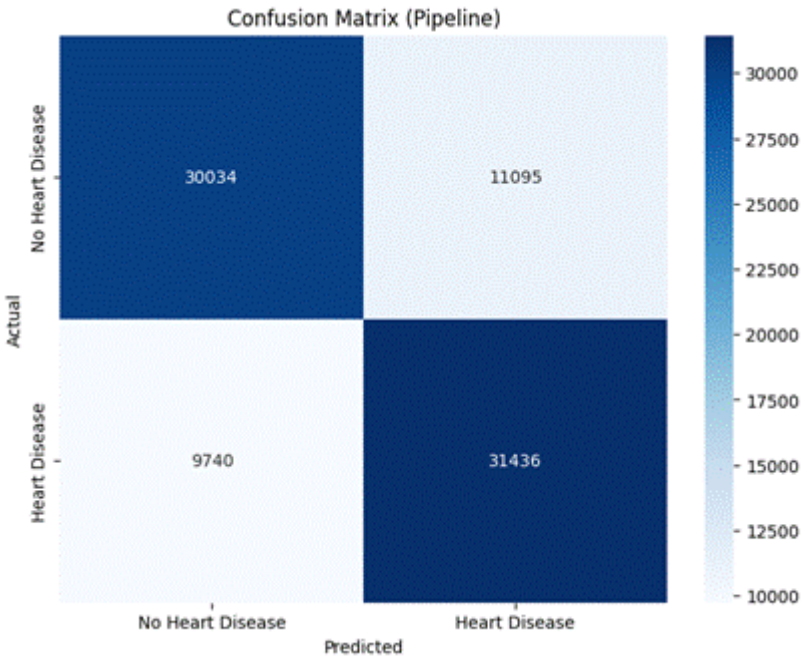
Accuracy: 0.73



Balanced

Label	Precision	Recall	F1-Score
0.0	0.76	0.73	0.74
1.0	0.74	0.76	0.75
Macro Avg	0.75	0.75	0.75
Weighted Avg:	0.75	0.75	0.75

Accuracy: 0.75



Heart Disease Model

Summary

After testing all models, the results show that balanced data leads to significantly better overall performance. The best two models were:
- XGBoost
- Random Forest

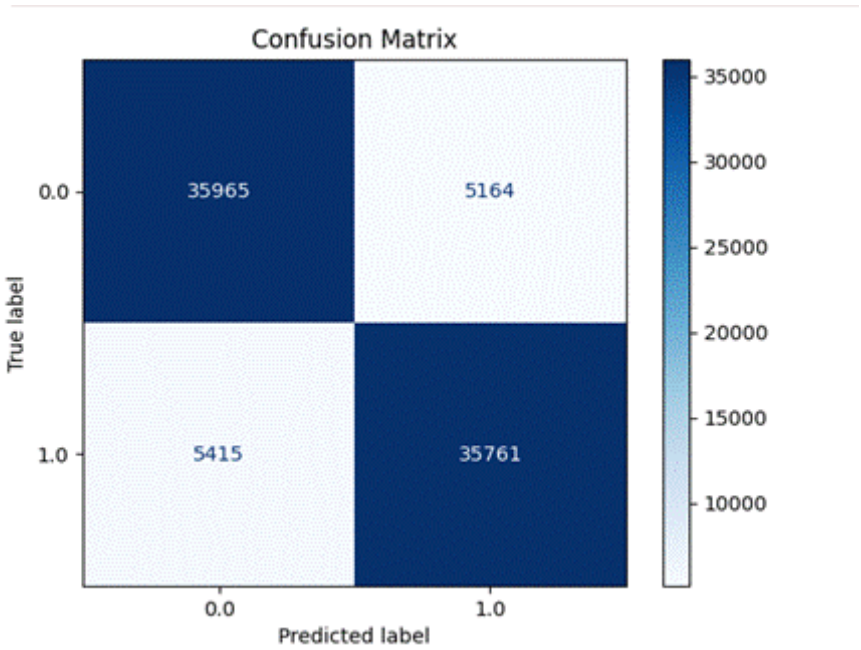
These were selected for hyperparameter tuning using GridSearchCV.

Random Forest (Best Parameters)

- max_depth: None
- max_features: sqrt
- min_samples_split: 5
- n_estimators: 300

Label	Precision	Recall	F1-Score
0.0	0.87	0.87	0.87
1.0	0.87	0.87	0.87
Macro Avg	0.87	0.87	0.87
Weighted Avg:	0.87	0.87	0.87

Accuracy: 0.87

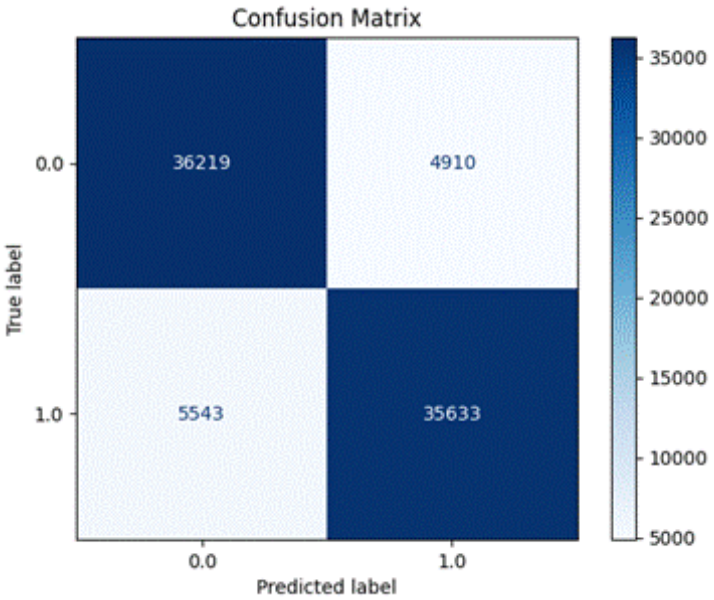


XGBoost (Best Parameters)

- colsample_bytree: 0.7
- learning_rate: 0.1
- max_depth:20
- min_child_weight: 1
- n_estimators: 100

Label	Precision	Recall	F1-Score
0.0	0.87	0.88	0.87
1.0	0.88	0.87	0.87
Macro Avg	0.87	0.87	0.87
Weighted Avg:	0.87	0.87	0.87

Accuracy: 0.87
Macro Avg: 0.87, 0.87, 0.87
Weighted Avg: 0.87, 0.87, 0.87

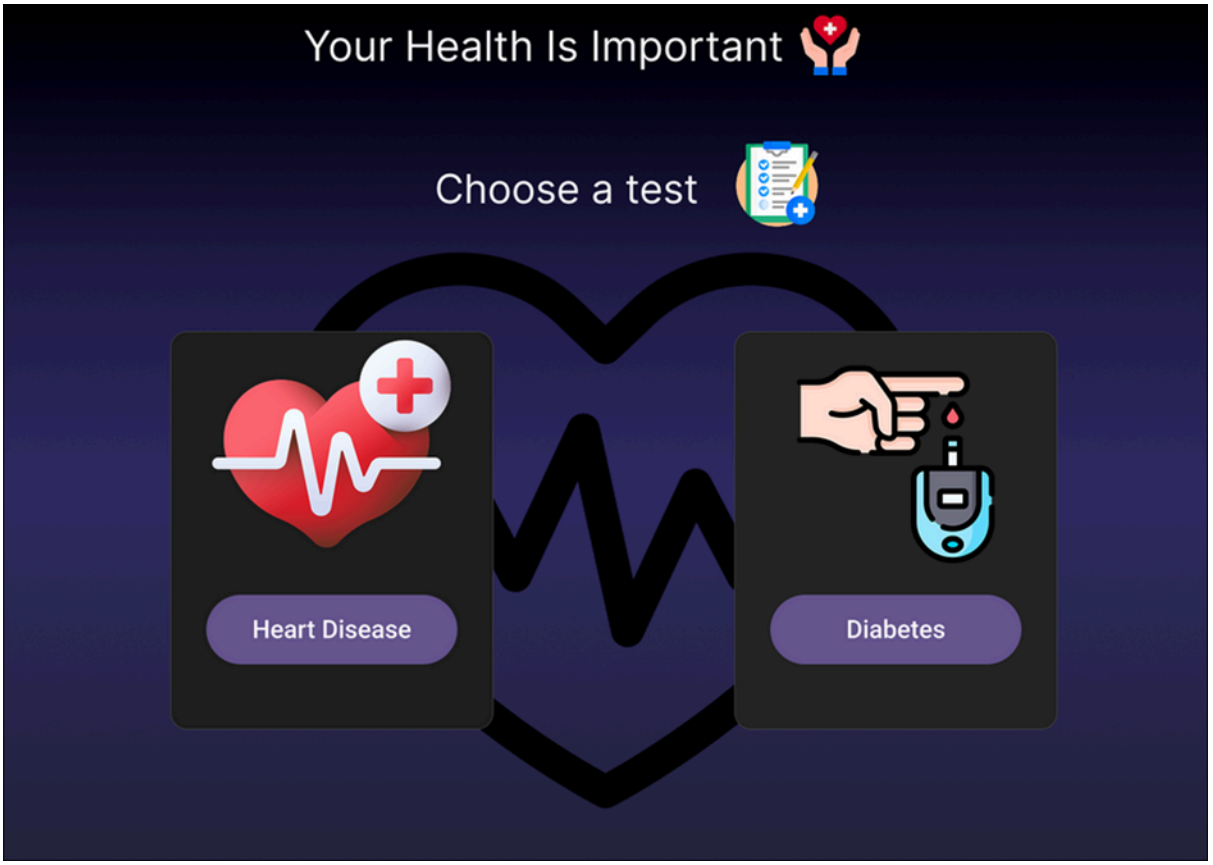


Deployment

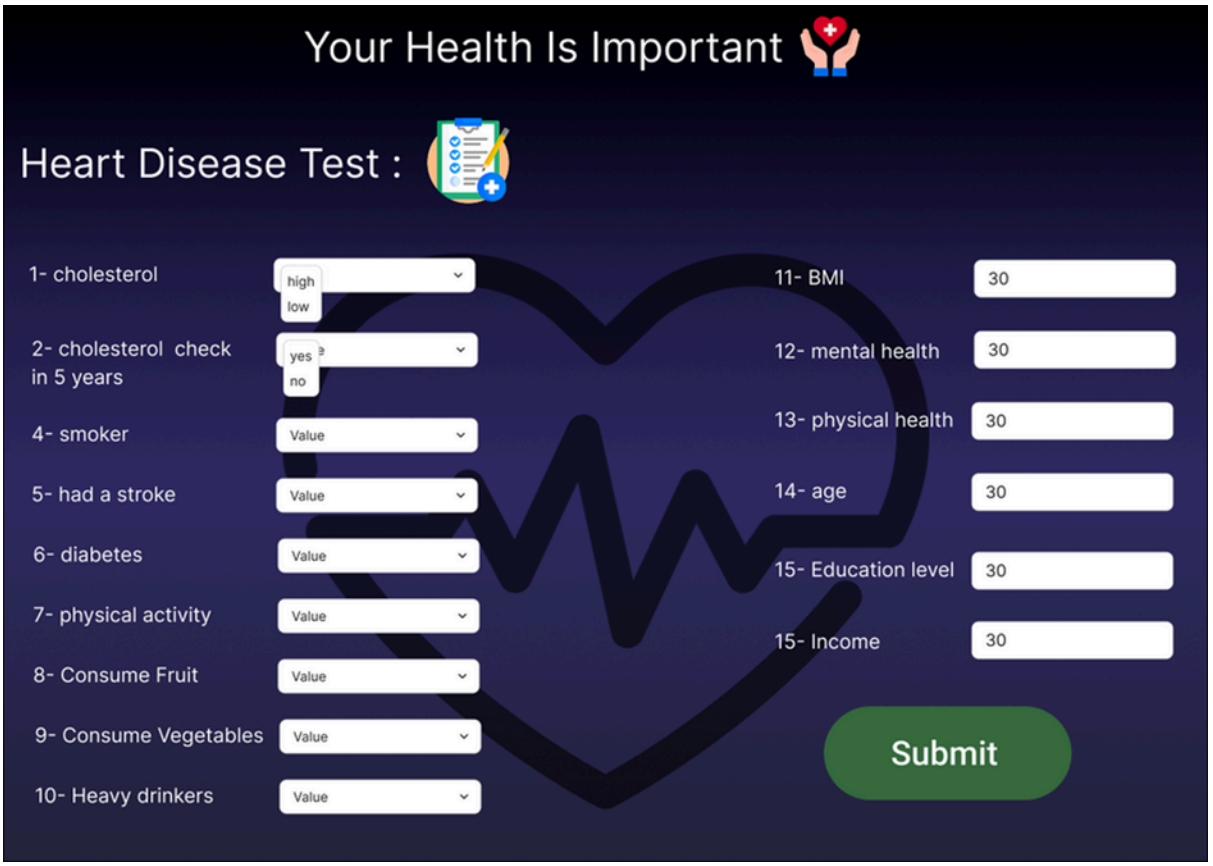
1. **Introduction:**

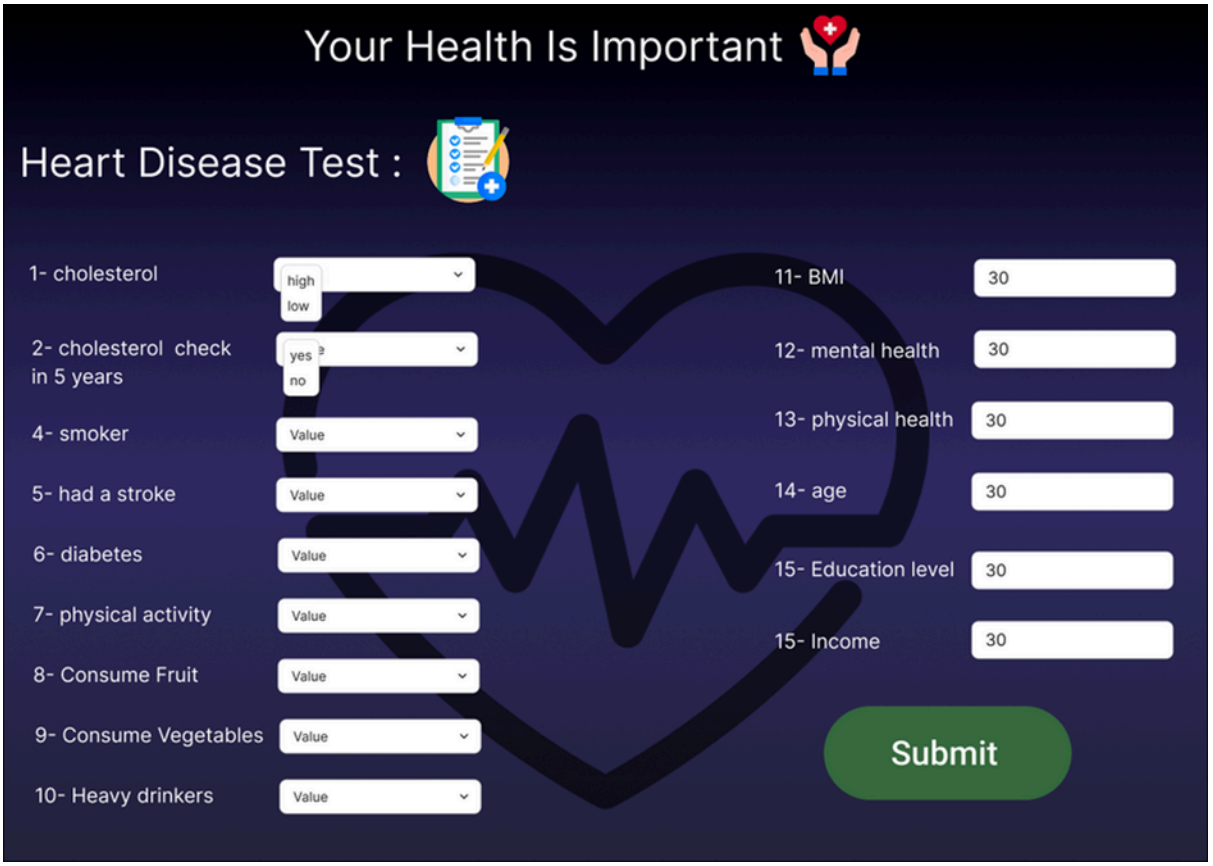
   This project aims to build an interactive web-based platform that predicts whether a user is likely to have diabetes or be at risk for heart disease based on health-related input. It combines a simple and user-friendly UI/UX, machine learning models, and a Flask-powered API to deliver fast, accessible predictions.

2. **UI/UX Design**
   - **Tools Used: Figma**
     - I started the project by designing the user interface using Figma. The focus was on simplicity, clarity, and responsiveness, ensuring users could easily input their health data without confusion.
     - **First version:**



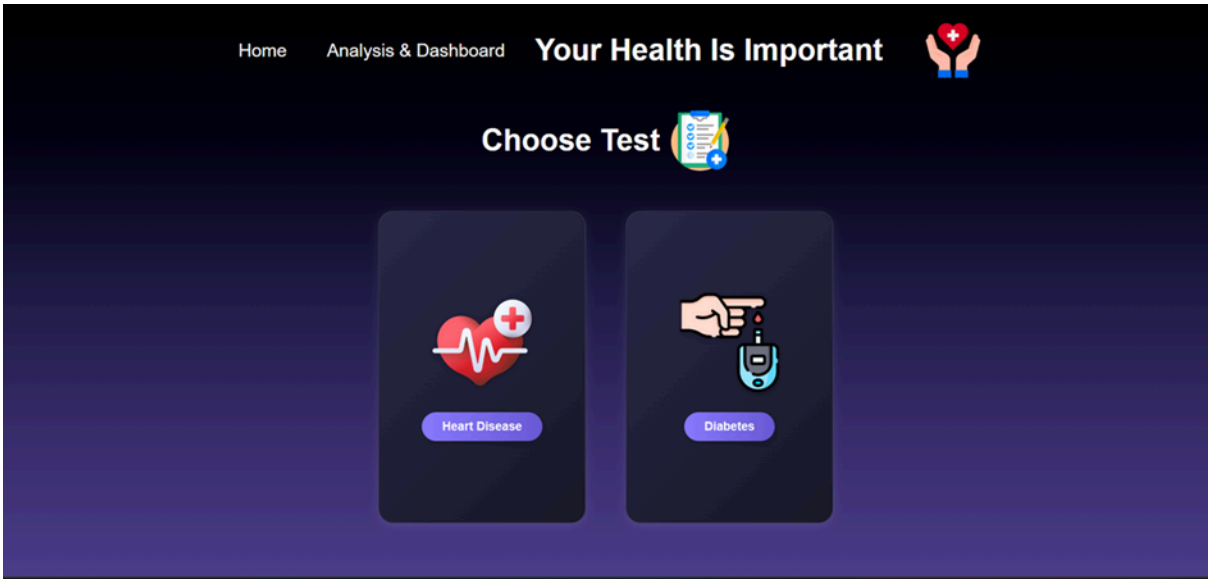
Deployment

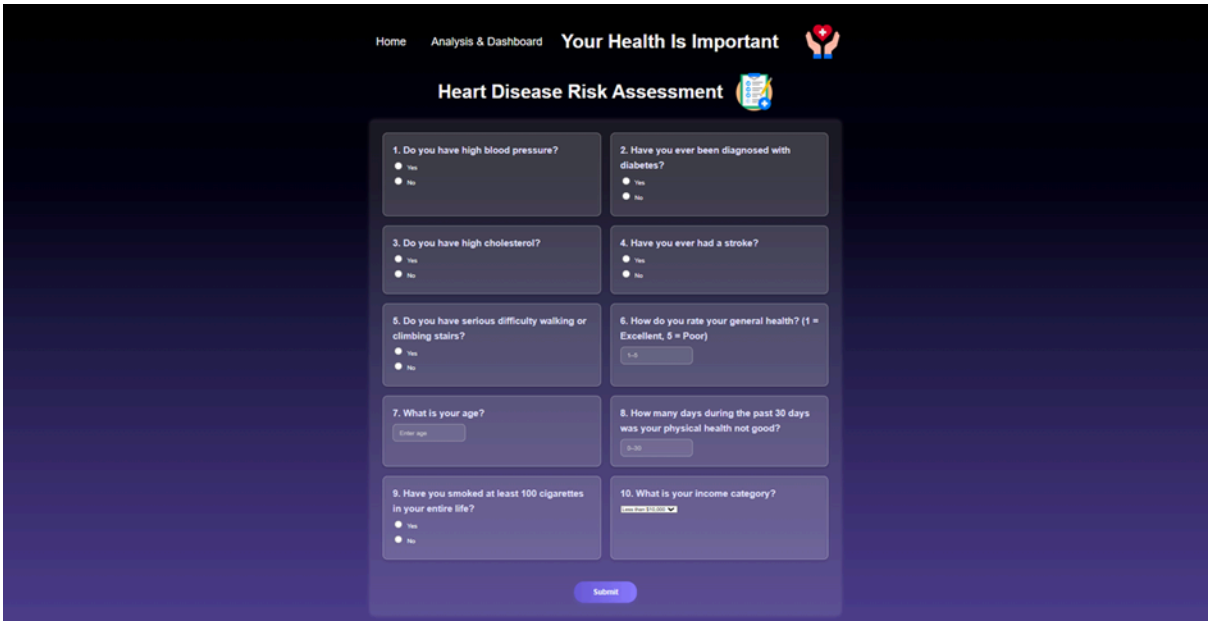

Your Health Is Important
Heart Disease Test :
1- cholesterol
high
low
2- cholesterol check in 5 years
yes
no
4- smoker
Value
5- had a stroke
Value
6- diabetes
Value
7- physical activity
Value
8- Consume Fruit
Value
9- Consume Vegetables
Value
10- Heavy drinkers
Value
11- BMI
30
12- mental health
30
13- physical health
30
14- age
30
15- Education level
30
15- Income
30
Submit

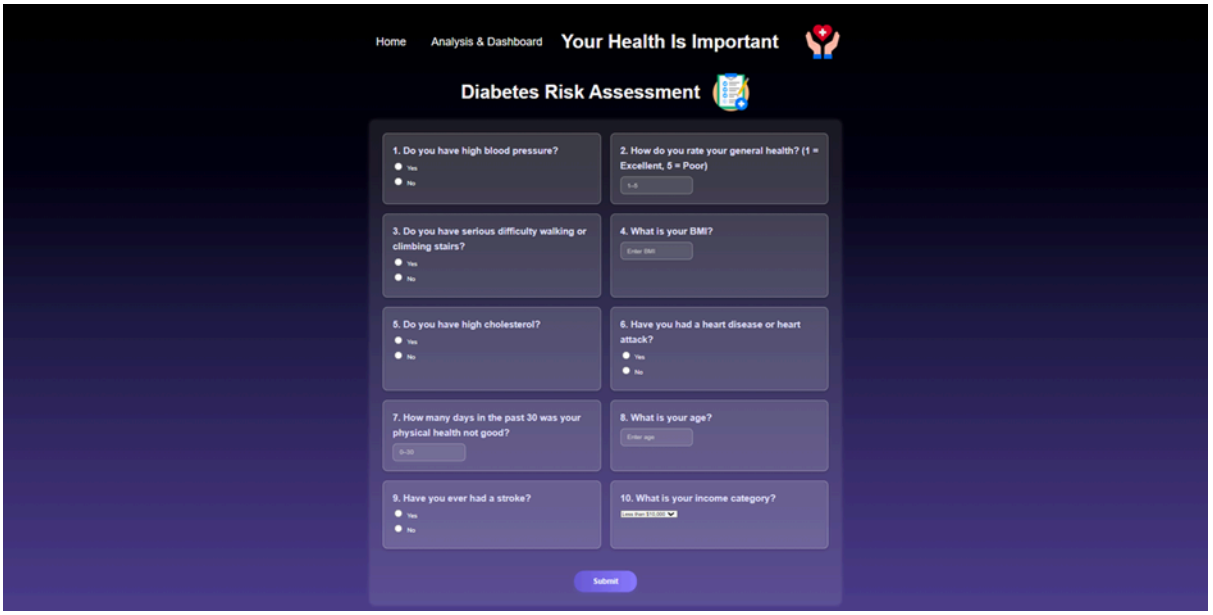

Your Health Is Important
Heart Disease Test :
1- cholesterol
high
low
2- cholesterol check in 5 years
yes
no
4- smoker
Value
5- had a stroke
Value
6- diabetes
Value
7- physical activity
Value
8- Consume Fruit
Value
9- Consume Vegetables
Value
10- Heavy drinkers
Value
11- BMI
30
12- mental health
30
13- physical health
30
14- age
30
15- Education level
30
15- Income
30
Submit

Deployment

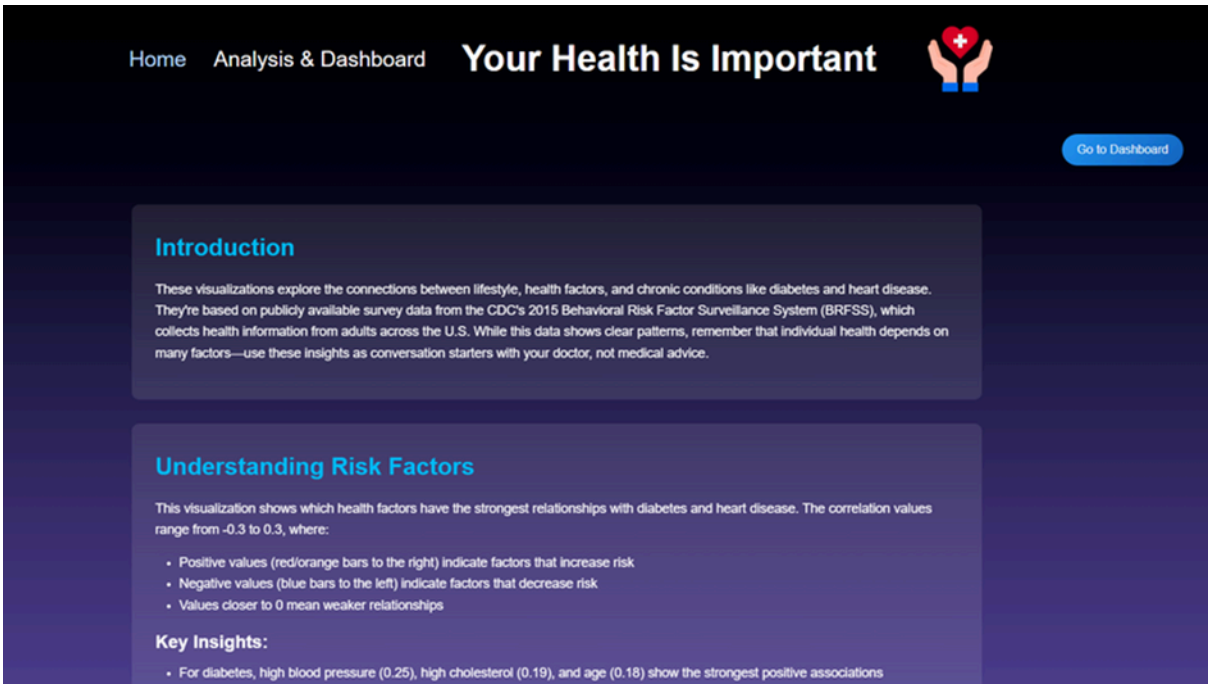
Final version:

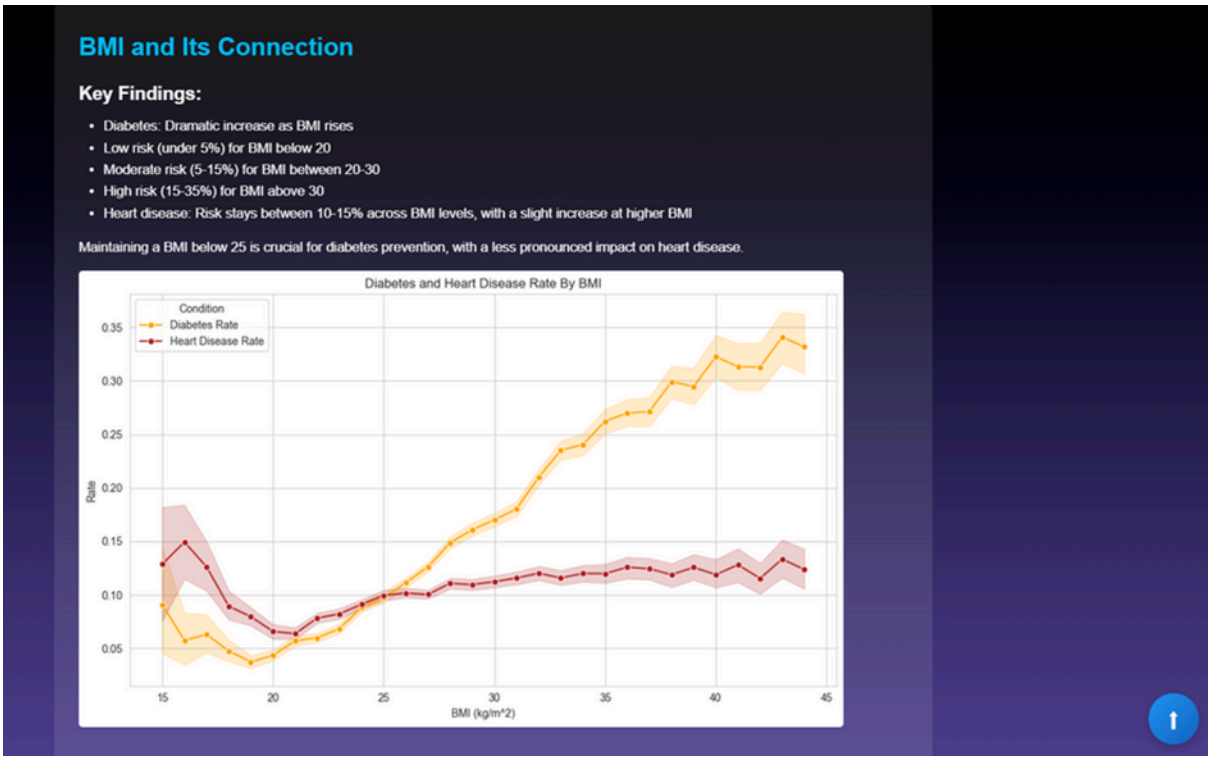




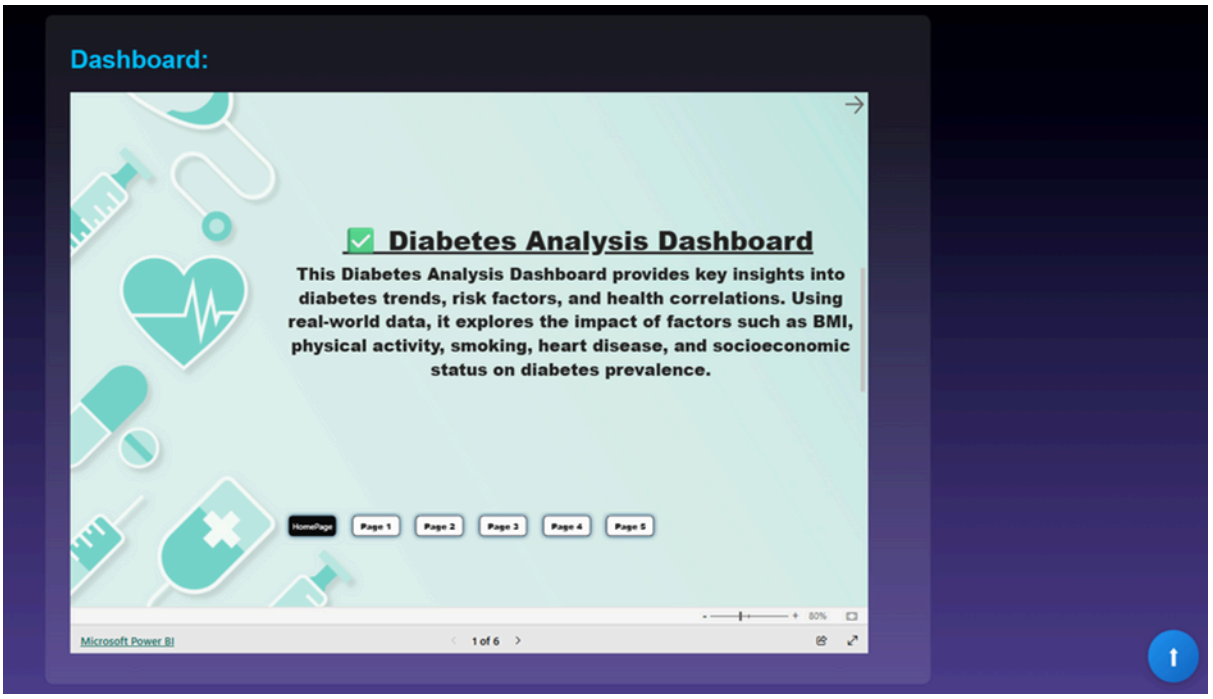


Deployment


Home
Analysis & Dashboard
Your Health Is Important
Go to Dashboard
Introduction
These visualizations explore the connections between lifestyle, health factors, and chronic conditions like diabetes and heart disease. They're based on publicly available survey data from the CDC's 2015 Behavioral Risk Factor Surveillance System (BRFSS), which collects health information from adults across the U.S. While this data shows clear patterns, remember that individual health depends on many factors—use these insights as conversation starters with your doctor, not medical advice.
Understanding Risk Factors
This visualization shows which health factors have the strongest relationships with diabetes and heart disease. The correlation values range from -0.3 to 0.3, where:
• Positive values (red/orange bars to the right) indicate factors that increase risk
• Negative values (blue bars to the left) indicate factors that decrease risk
• Values closer to 0 mean weaker relationships
Key Insights:
• For diabetes, high blood pressure (0.25), high cholesterol (0.19), and age (0.18) show the strongest positive associations


BMI and Its Connection
Key Findings:
• Diabetes: Dramatic increase as BMI rises
• Low risk (under 5%) for BMI below 20
• Moderate risk (5-15%) for BMI between 20-30
• High risk (15-35%) for BMI above 30
• Heart disease: Risk stays between 10-15% across BMI levels, with a slight increase at higher BMI
Maintaining a BMI below 25 is crucial for diabetes prevention, with a less pronounced impact on heart disease.
Diabetes and Heart Disease Rate By BMI
Condition
Diabetes Rate
Heart Disease Rate
Rate
0.35
0.30
0.25
0.20
0.15
0.10
0.05
15
20
25
30
35
40
45
BMI (kg/m^2)

Deployment


Dashboard:
Diabetes Analysis Dashboard
This Diabetes Analysis Dashboard provides key insights into diabetes trends, risk factors, and health correlations. Using real-world data, it explores the impact of factors such as BMI, physical activity, smoking, heart disease, and socioeconomic status on diabetes prevalence.
HomePage
Page 1
Page 2
Page 3
Page 4
Page 5
Microsoft Power BI
1 of 6

Deployment

3. Backend API Development

This project involves building a RESTful API using Flask to predict the likelihood of a user having diabetes or heart disease, based on user-submitted health data. The prediction is powered by pre-trained machine learning models and made accessible via simple HTTP POST requests.

Two versions were developed:

- **Version 1:** Loads models from the MLflow Model Registry.
- **Version 2:** Loads models from local .pkl files and is containerized using Docker for local deployment and portability.

Understanding the Model Pipelines

Before deployment, I thoroughly reviewed the development process of both models to ensure smooth integration and reliable predictions.

Key Areas Reviewed:

- **Feature Selection:**
  Analyzed which features were retained for training, using domain knowledge and techniques like correlation matrices and feature importance scores.
- **Model Training:**
  Reviewed training algorithms, data splits, and preprocessing steps. Ensured there was no data leakage and that input scaling or encoding was consistent between training and inference.
- **Optimizations:**
  Applied hyperparameter tuning via grid search or random search, and used cross-validation to ensure generalizability and reduce overfitting.
- **Challenges Encountered:**
  - Imbalanced classes
  - Handling missing values
  - Preventing overfitting in early models
  - Ensuring consistency between training and production environments
  - Exporting models with correct input formats

Deployment

Version 1: MLflow-Based Model Loading

Purpose

Designed for scalable and production-like environments with centralized model tracking.

Technical Highlights

- **Model Registry**: Models are stored and versioned using MLflow.
- **API Framework**: Flask
- **CORS**: Enabled using flask_cors
- **Logging**: Combines file and console logging with rotation
- **Type Hints**: Used throughout the codebase

Model Loading

```python
# --- Load Models from MLflow ---
def load_model_from_registry(model_name: str, stage: str = "1"):
    try:
        logger.info(f"Loading model: {model_name}")
        model = mlflow.pyfunc.load_model(f"models:/{model_name}/{stage}")
        logger.info(f"{model_name} model loaded successfully.")
        return model
    except Exception as e:
        logger.error(f"Error loading {model_name} model: {e}")
        logger.error(traceback.format_exc())
        raise

diabetes_model = load_model_from_registry("diabetes_production")
heart_model = load_model_from_registry("heart_disease_production")
```

Endpoints

- GET / – Returns a welcome message
- POST /diabetes_test – Returns diabetes prediction
- POST /heart_test – Returns heart disease prediction

Deployment

Version 2: Local Pickle Model Loading (Dockerized)

Purpose

A lightweight, offline-ready version suitable for demos and rapid prototyping.

Technical Highlights

- **Model Files**: .pkl models loaded directly from the file system
- **API Framework**: Flask
- **Docker**: Full Dockerfile and docker-compose setup for local deployment
- **CORS**: Enabled
- **Logging**: Same as Version 1
- **Type Hints**: Used throughout

Model Loading

```
# Load the models
diabetes_model = pickle.load(open("models/diabetes_model.pkl", "rb"))
heart_model = pickle.load(open("models/heart_disease_model.pkl", "rb"))
```

Key Challenges & Solutions

1. Model Serialization Issues

- **Problem**: Models were initially not saved with the correct input shape or format.
- **Solution**: Ensured consistent preprocessing and used pickle to serialize models with input formats matching what the API expects.

2. Data Format Conversion

- **Problem**: The frontend sent form-encoded data, but the API required JSON.
- **Solution**: Used JSON.stringify() on the frontend and request.get_json() on the backend to ensure compatibility.

3. CORS Error Handling

- **Problem**: CORS issues arose because the frontend and backend were on different domains.
- **Solution**: Integrated flask_cors to allow cross-origin requests with appropriate headers.

4. Error Handling in API

- **Problem**: Invalid or malformed data could cause crashes.
- **Solution**: Wrapped endpoint logic in try-except blocks, catching KeyError, ValueError, and other exceptions to return user-friendly error messages and HTTP status codes.

Logging and Error Management

- **Logging**: Logs are recorded to both a rotating file and console using logging.handlers.RotatingFileHandler.
- **Decorator**: A @safe_route decorator wraps route handlers to centralize error handling.
- **Output**: Errors are clearly logged, and users receive structured error responses.

Type Annotations

Python type hints were used consistently to:

- Improve editor autocomplete
- Aid static analysis tools
- Increase code clarity for future maintenance

4. Frontend Development:

- **Tools Used: HTML, CSS, JavaScript**
  - **The web app did not require frameworks like React or Angular because:**
    - The structure was simple and lightweight.
    - There was no complex state management.
    - Loading speed and simplicity were prioritized.
    - Vanilla JavaScript was sufficient to handle API interactions and user input.
    - Thus, sticking to HTML, CSS, and JS made the development faster and avoided unnecessary overhead.
- **Website Structure**
  - Home Page – Project description and links
  - Diabetes Test Page – User inputs health data to get a prediction
  - Heart Disease Test Page – Similar interface for heart risk evaluation
  - Analysis Page – Displays model analysis and visual dashboard insights

Deployment

Conclusion

This project focused on the backend and **API** development required to support a practical machine learning web application. The core of the system was built using Flask to serve well-structured **RESTful APIs** that connect a trained predictive model with a responsive web interface. Key backend engineering tasks included managing model serialization, implementing robust error handling, and resolving cross-origin (CORS) issues to ensure seamless frontend-backend communication.

A flexible deployment strategy was established, supporting both local execution through **Docker** with serialized `.pkl` models and production-grade deployment using MLflow for model tracking and loading. The API system was designed with scalability and maintainability in mind, allowing for smooth integration of additional features such as user authentication or prediction storage in future iterations.

This API-driven approach demonstrates strong backend architecture principles and provides a reliable foundation for expanding the application into a full-featured, real-world health prediction platform.

Appendix and References

Our website: https://ahmedmohamady1.github.io/Healthcare-DEPI-Project/

GitHub Repository:
https://github.com/AhmedMohamady1/Healthcare-DEPI-Project/

Data Source:
https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data?select=diabetes_binary_health_indicators_BRFSS2015.csv

Additional Acknowledgement: Elements of our code and methodology were adapted with reference to materials provided by the **DEPI Initiative**, which helped guide parts of our data analysis and model development approach.