PRESIDENCY UNIVERSITY

Private University Estd. in Karnataka State by Act No. 41 of 2013

BANGALORE

A Project Report

On

# "Early Prediction of Lifestyle Diseases"

## Batch Details

| Sl. No. | Roll Number | Student Name |
|---|---|---|
| 1 | 20201CSD0153 | DEEKSHITH D L |
| 2 | 20201CSD0182 | MITHUN R |
| 3 | 20201CSD0187 | MADHUSHREE |

## School of Computer Science,

## Presidency University, Bengaluru.

Under the guidance of,

Mr. Lakshmisha S K

School of Computer Science,

Presidency University, Bengaluru

# CONTENTS

# 1. INTRODUCTION

**General Introduction**:

The increasing prevalence of lifestyle diseases has become a significant global health concern, leading to a surge in healthcare costs. Early prediction and preventive measures can play a crucial role in reducing the burden of these diseases. This report focuses on utilizing detailed demographic and vital statistics, obtained during physical checkups, to predict the likelihood of lifestyle diseases. Technology companies, such as Google, can employ machine learning models to enhance early detection, enabling more effective preventive healthcare strategies.

**Introduction to the Domain**

Lifestyle diseases encompass conditions like heart disease, diabetes, hypertension, and obesity, often influenced by factors such as age, genetics, diet, physical activity, and smoking. The objective is to leverage machine learning to analyze these factors and predict the likelihood of individuals developing such diseases, facilitating timely interventions.

**Problem Statement:**

How can we predict the likelihood of lifestyle diseases early to enable preventive healthcare. This can reduce the cost of treatment significantly. Potential solution: using detailed demographic and vital stats about people who have a particular disease and those who dont, technology companies like Google can create machine learning models to predict specific diseases in an individual during physical checkups

# 2. LITERATURE REVIEW

## Existing Methods

### Advantages:

- **Early Detection:** Existing methods leverage medical data for early disease detection.
- **Cost Reduction:** Early identification enables cost-effective preventive measures.
- **Improved Patient Outcomes:** Timely interventions lead to better patient outcomes.
- **Data-Driven Insights:** Data analytics provides valuable insights into disease trends.

### Limitations:

- Data Privacy Concerns: Use of personal health data raises privacy issues.
- **Limited Predictive Accuracy:** Current methods may have limitations in accurately predicting disease risk.
- **Dependency on Data Quality:** The reliability of predictions depends on the quality of input data.
- **Ethical Considerations:** Ethical concerns arise from the use of personal health information.

# 3.OBJECTIVES

- Develop a machine learning model for predicting lifestyle diseases.
- Enhance predictive accuracy by incorporating advanced features.
- Address privacy concerns through secure data handling practices.
- Evaluate the model's effectiveness in a real-world setting.

# 4. METHODOLOGY

## Experimental Details/Methodology

**Hardwares and Softwares Used:**

### Hardware:

- High-performance computing system for model training and validation.

### Software:

- Python for programming.
- Scikit-learn, TensorFlow, or PyTorch for machine learning model development.
- Jupyter Notebooks for experimentation and analysis.

## Design Procedure

- **Data Collection:** Gather detailed demographic and vital stats from individuals during physical checkups.
- **Data Preprocessing:** Clean and preprocess the data, handling missing values and outliers.
- **Feature Engineering:** Extract relevant features, considering factors like age, gender, BMI, smoking status, etc.
- **Model Development:** Implement machine learning models such as logistic regression, decision trees, or neural networks.
- **Model Training:** Train the models using historical data, fine-tuning parameters for optimal performance.
- **Validation:** Evaluate the model's performance using a separate dataset.
- **Privacy Measures:** Implement encryption and secure data handling practices to address privacy concerns.
- **Deployment:** Deploy the model for real-world prediction during physical checkups.

# 5. OUTCOMES

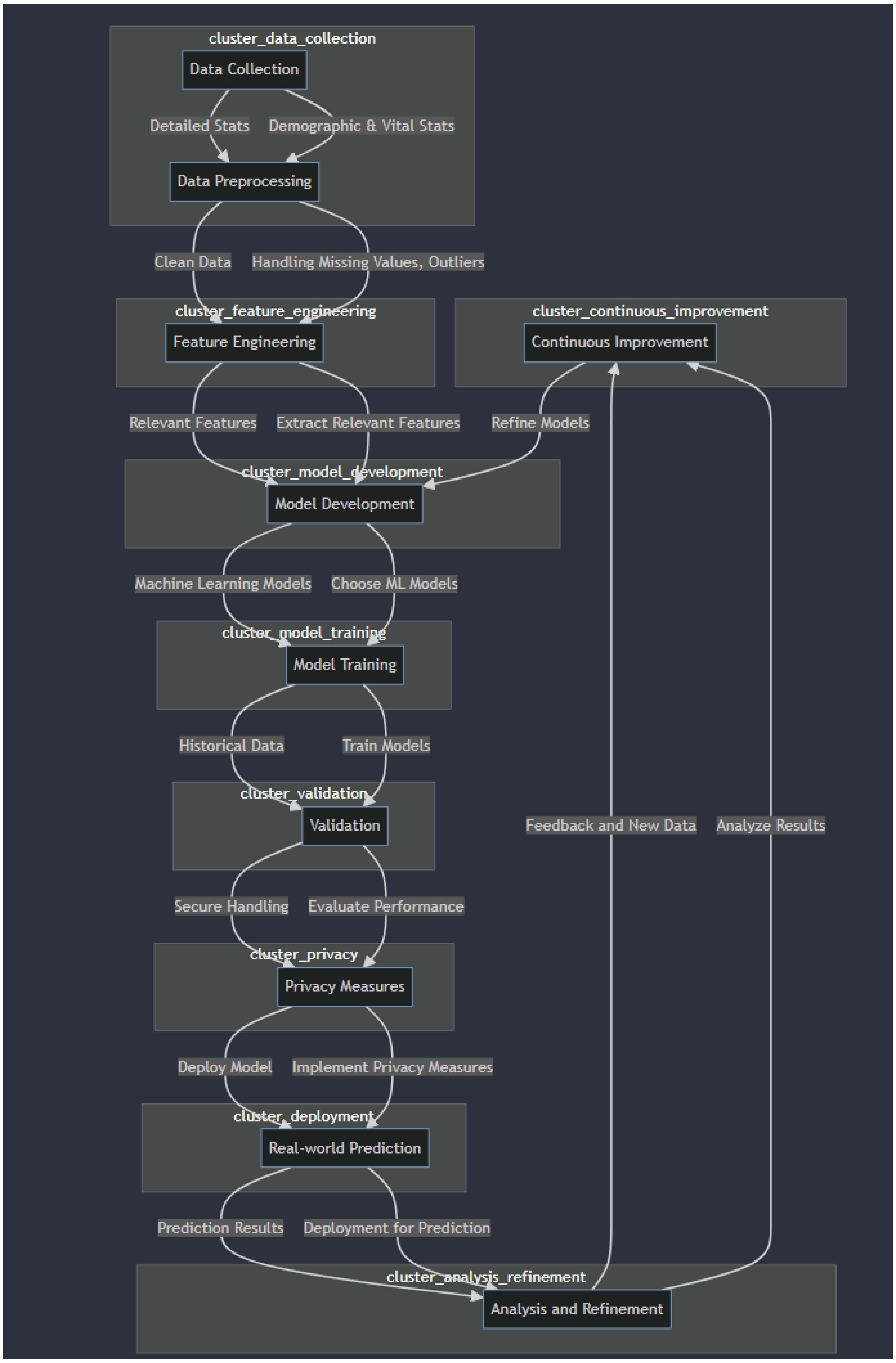
**The expected outcomes include:**

- A machine learning model capable of predicting the likelihood of lifestyle diseases.
- Improved accuracy and reliability compared to existing methods.
- Enhanced understanding of the impact of various features on disease prediction.

# 6. PROJECT EXECUTION PLAN

- Data collection and preprocessing.
- Feature engineering and model development.
- Model training and validation.
- Privacy measures implementation and testing.
- Deployment and real-world testing.
- Analysis of results, refinement of models, and final report preparation.

## Work Flow:



cluster_data_collection
Data Collection
Detailed Stats
Demographic & Vital Stats
Data Preprocessing
Clean Data
Handling Missing Values, Outliers
cluster_feature_engineering
Feature Engineering
cluster_continuous_improvement
Continuous Improvement
Relevant Features
Extract Relevant Features
Refine Models
cluster_model_development
Model Development
Machine Learning Models
Choose ML Models
cluster_model_training
Model Training
Historical Data
Train Models
cluster_validation
Validation
Feedback and New Data
Analyze Results
Secure Handling
Evaluate Performance
cluster_privacy
Privacy Measures
Deploy Model
Implement Privacy Measures
cluster_deployment
Real-world Prediction
Prediction Results
Deployment for Prediction
cluster_analysis_refinement
Analysis and Refinement

## Data Collection:

| A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | Age | Gender | BMI | Smoking_St | BloodPressur | Physical_Act | Diet | Cholesterol | Alcohol_Cc | SleepDurat | Disease | |
| 1 | 55 | Male | 28 | Non-Smoke | 130 | Active | Healthy | Normal | Low | 7 | Healthy | |
| 2 | 65 | Female | 32 | Former-Sm | 140 | Inactive | Unhealthy | High | Moderate | 6 | Heart_Disease | |
| 3 | 40 | Male | 24 | Non-Smoke | 120 | Active | Balanced | Normal | Low | 8 | Healthy | |
| 4 | 70 | Female | 35 | Smoker | 150 | Inactive | Unhealthy | High | High | 5 | Heart_Disease | |
| 5 | 60 | Male | 30 | Non-Smoke | 135 | Active | Balanced | Normal | Moderate | 7 | Hypertension | |
| 6 | 50 | Female | 28 | Non-Smoke | 128 | Active | Balanced | Normal | Low | 7 | Healthy | |
| 7 | 75 | Male | 33 | Non-Smoke | 145 | Inactive | Unhealthy | High | Low | 6 | Diabetes | |
| 8 | 58 | Male | 29 | Non-Smoke | 132 | Active | Balanced | Normal | Moderate | 7 | Hypertension | |
| 9 | 68 | Female | 31 | Non-Smoke | 138 | Inactive | Unhealthy | High | Low | 6 | Hypertension | |
| 10 | 43 | Female | 25 | Non-Smoke | 123 | Active | Balanced | Normal | Low | 8 | Healthy | |
| 11 | 57 | Male | 27 | Former-Sm | 128 | Active | Balanced | Normal | Moderate | 7 | Heart_Disease | |
| 12 | 48 | Female | 26 | Non-Smoke | 125 | Active | Balanced | Normal | Moderate | 7 | Healthy | |
| 13 | 53 | Female | 28 | Non-Smoke | 130 | Active | Unhealthy | High | Moderate | 6 | Obesity | |
| 14 | 63 | Male | 31 | Non-Smoke | 137 | Inactive | Healthy | Normal | Low | 7 | Hypertension | |
| 15 | 62 | Female | 29 | Non-Smoke | 134 | Active | Balanced | Normal | Low | 8 | Hypertension | |
| 16 | 42 | Male | 25 | Smoker | 125 | Active | Unhealthy | High | High | 5 | Obesity | |

## Exploratory Data Analysis:

In [2]:
```python
import pandas as pd
import numpy as np
```

In [3]:
```python
data = pd.read_csv("LifeStyle_Data_Demo.csv")
```

In [4]:
```python
data.head(3)
```

Out[4]:

| | ID | Age | Gender | BMI | Smoking_Status | BloodPressure_Systolic | Physical_Activity | Diet | Cholesterol | Alcohol_Consumption | SleepDuration | Disease |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 55 | Male | 28 | Non-Smoker | 130 | Active | Healthy | Normal | Low | 7 | Healthy |
| 1 | 2 | 65 | Female | 32 | Former-Smoker | 140 | Inactive | Unhealthy | High | Moderate | 6 | Heart_Disease |
| 2 | 3 | 40 | Male | 24 | Non-Smoker | 120 | Active | Balanced | Normal | Low | 8 | Healthy |

In [6]:
```python
data.isnull().sum()
```

Out[6]:
```
ID                        0
Age                       0
Gender                    0
BMI                       0
Smoking_Status            0
BloodPressure_Systolic    0
Physical_Activity         0
Diet                      0
Cholesterol               0
Alcohol_Consumption       0
SleepDuration             0
Disease                   0
dtype: int64
```

In [7]:
```python
data.describe()
```

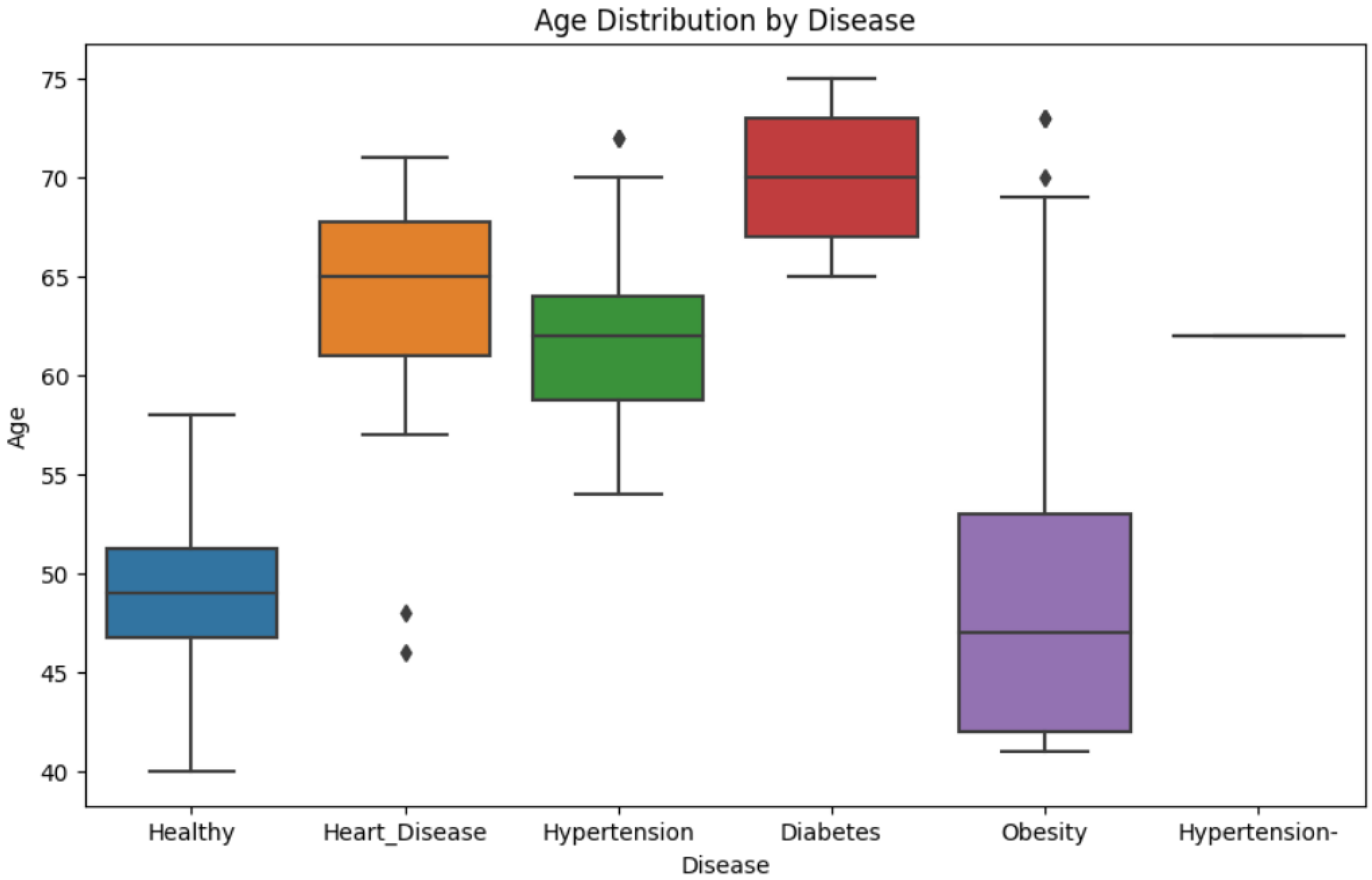
Out[7]:

| | ID | Age | BMI | BloodPressure_Systolic | SleepDuration |
|---|---|---|---|---|---|
| count | 200.000000 | 200.000000 | 200.000000 | 200.000000 | 200.000000 |
| mean | 100.500000 | 57.035000 | 29.440000 | 134.230000 | 6.615000 |
| std | 57.879185 | 9.398694 | 2.811749 | 7.943083 | 0.889229 |
| min | 1.000000 | 40.000000 | 24.000000 | 120.000000 | 5.000000 |
| 25% | 50.750000 | 49.000000 | 27.000000 | 128.000000 | 6.000000 |
| 50% | 100.500000 | 58.000000 | 29.000000 | 133.000000 | 7.000000 |
| 75% | 150.250000 | 65.000000 | 31.000000 | 139.000000 | 7.000000 |
| max | 200.000000 | 75.000000 | 36.000000 | 153.000000 | 8.000000 |

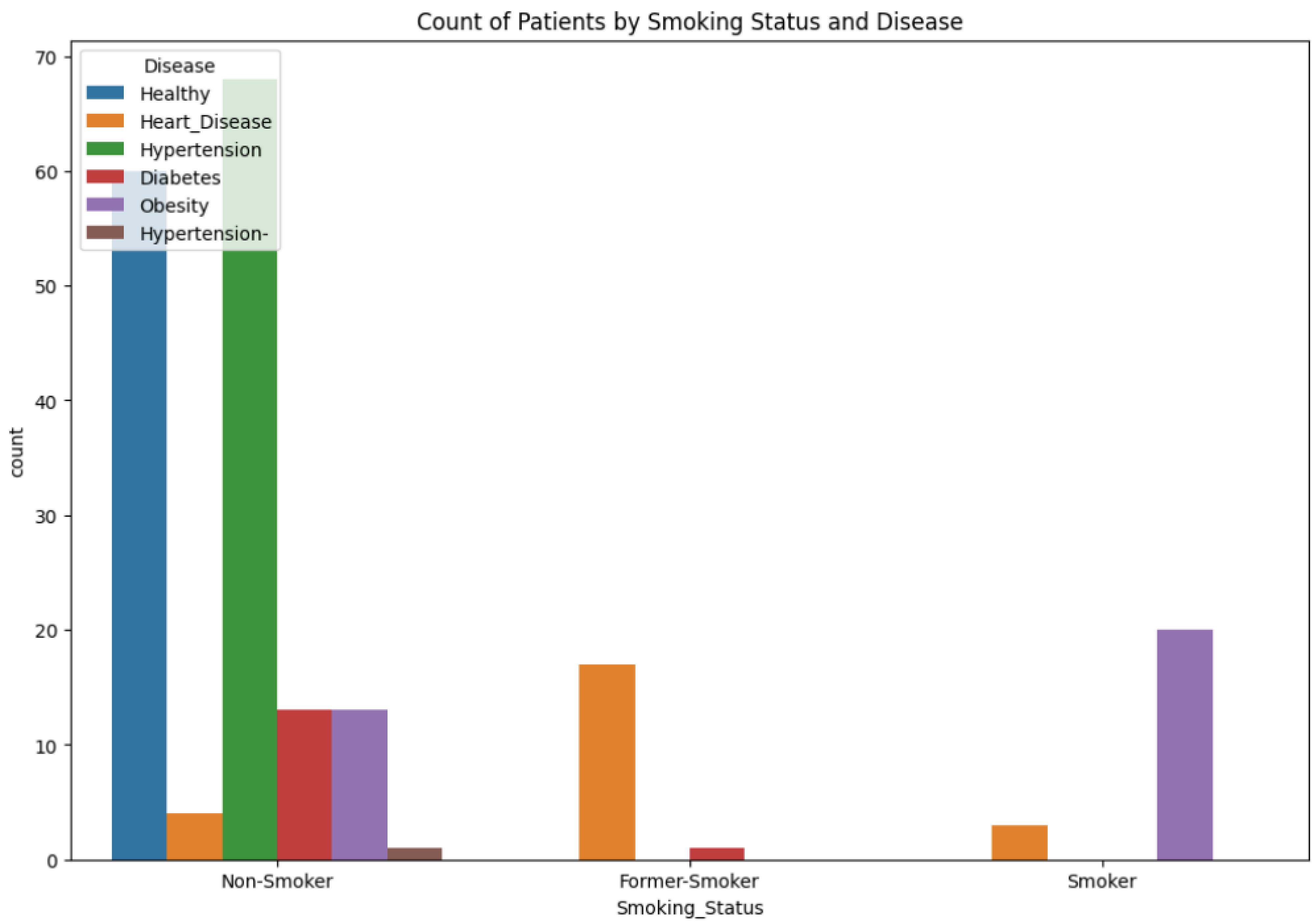
In [8]:
```python
data.shape
```

Out[8]: (200, 12)

In [14]:
```
# Boxplot for age distribution by disease
plt.figure(figsize=(10, 6))
sns.boxplot(x='Disease', y='Age', data=data)
plt.title('Age Distribution by Disease')
plt.show()
```

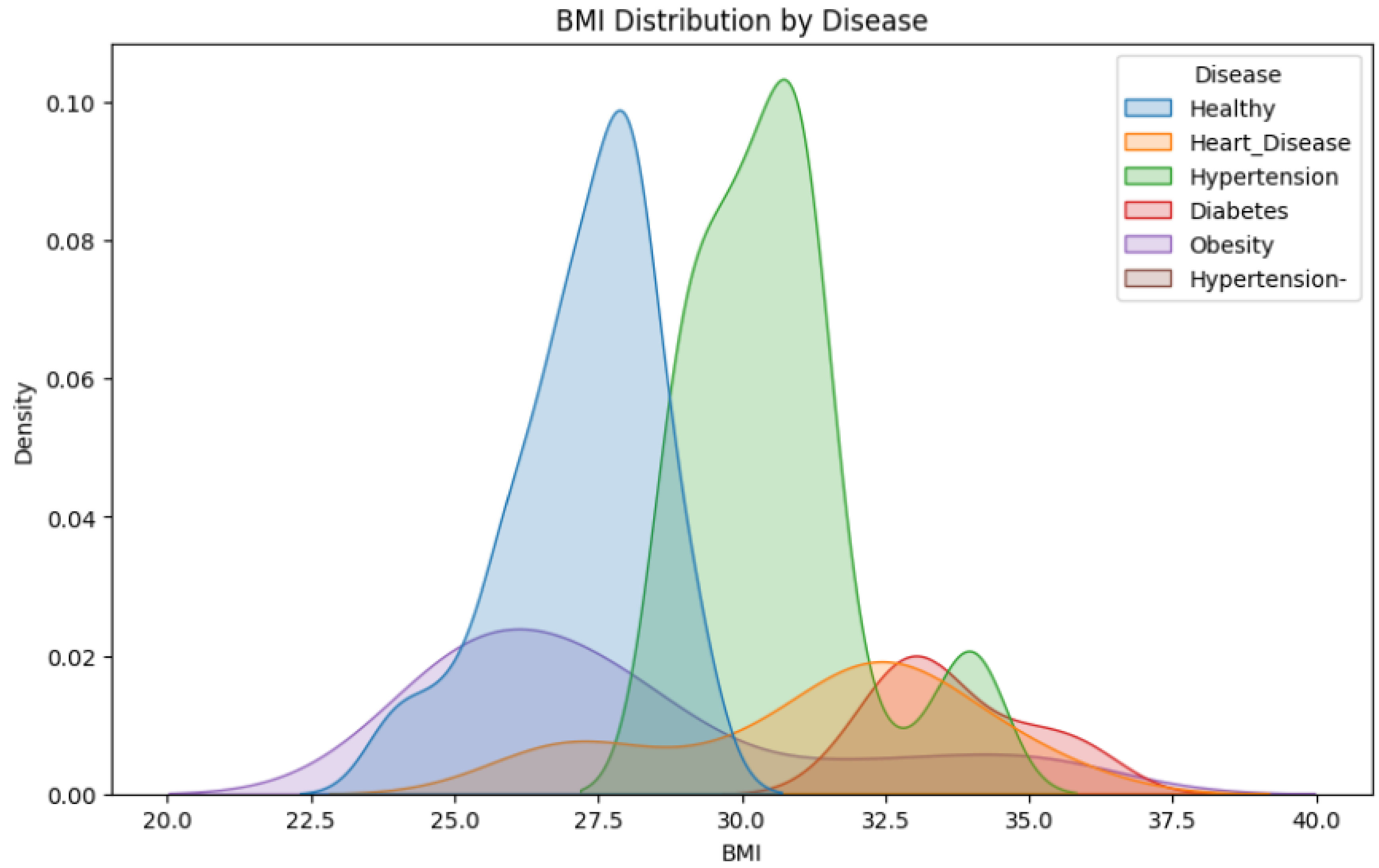



In [15]:
```
# Countplot for categorical variables
plt.figure(figsize=(12, 8))
sns.countplot(x='Smoking_Status', hue='Disease', data=data)
plt.title('Count of Patients by Smoking Status and Disease')
plt.show()
```

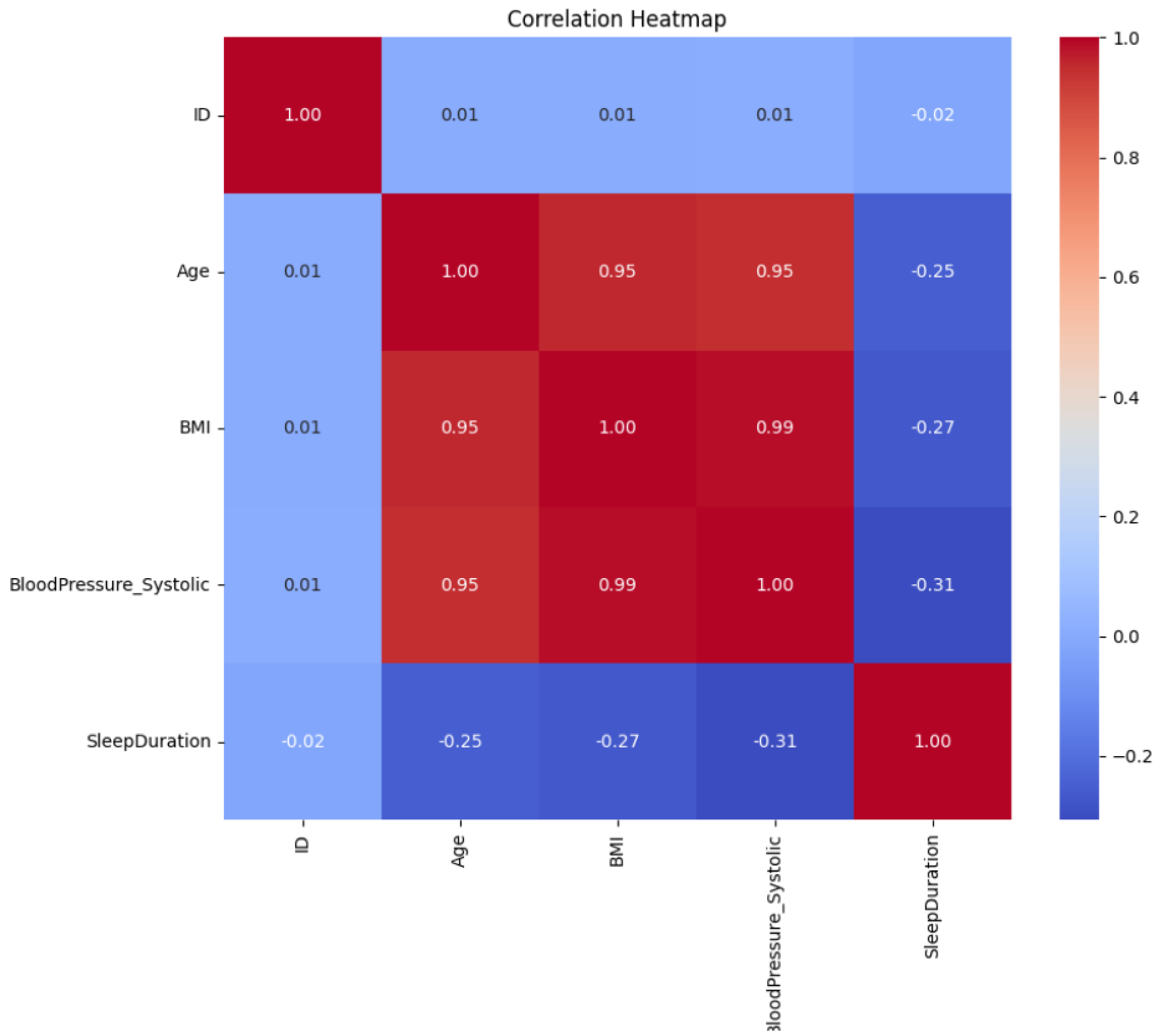


In [16]:
```
# Distribution of BMI by Disease
plt.figure(figsize=(10, 6))
sns.kdeplot(x='BMI', hue='Disease', data=data, fill=True)
plt.title('BMI Distribution by Disease')
plt.show()
```

```
d:\Users\e3t\anaconda3\lib\site-packages\seaborn\distributions.py:316: UserWarning: Dataset has 0 variance; skipping density es
timate. Pass `warn_singular=False` to disable this warning.
  warnings.warn(msg, UserWarning)
```




In [17]:
```
# Correlation heatmap for numerical variables
plt.figure(figsize=(10, 8))
correlation_matrix = data.corr()
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Heatmap')
plt.show()
```



## Data Preprocessing:

```
In [19]: # Handle missing values (replace with mean for simplicity)
         data.fillna(data.mean(), inplace=True)

C:\Users\e3t\AppData\Local\Temp\ipykernel_10660\3797966520.py:2: FutureWarning: Dropping of nuisance columns in DataFrame reduc
tions (with 'numeric_only=None') is deprecated; in a future version this will raise TypeError.  Select only valid columns befor
e calling the reduction.
  data.fillna(data.mean(), inplace=True)

In [ ]: #Data Preprocessing

In [20]: from sklearn.model_selection import train_test_split
         from sklearn.preprocessing import StandardScaler, LabelEncoder

In [21]: # Encode categorical variables (Label Encoding for simplicity)
         label_encoder = LabelEncoder()
         data['Gender'] = label_encoder.fit_transform(data['Gender'])
         data['Smoking_Status'] = label_encoder.fit_transform(data['Smoking_Status'])
         data['Diet'] = label_encoder.fit_transform(data['Diet'])
         data['Alcohol_Consumption'] = label_encoder.fit_transform(data['Alcohol_Consumption'])

In [22]: # Split the dataset into features (X) and target variable (y)
         X = data.drop('Disease', axis=1)
         y = data['Disease']

         # Split the dataset into training and testing sets
         X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

In [23]: # Standardize numerical features using StandardScaler
         scaler = StandardScaler()
         X_train[['Age', 'BMI', 'BloodPressure_Systolic']] = scaler.fit_transform(X_train[['Age', 'BMI', 'BloodPressure_Systolic']])
         X_test[['Age', 'BMI', 'BloodPressure_Systolic']] = scaler.transform(X_test[['Age', 'BMI', 'BloodPressure_Systolic']])

In [24]: # Display the preprocessed dataset
         print(X_train.head())
         print(X_test.head())

       ID       Age   Gender       BMI   Smoking_Status  BloodPressure_Systolic  \
```



## Model Building:

## Continue……

## 7. CONCLUSION

This project aims to significantly contribute to preventive healthcare by leveraging machine learning for early prediction of lifestyle diseases. Through comprehensive data analysis, the model is expected to provide accurate predictions, allowing for timely interventions and cost-effective healthcare.

## 8.REFERENCES


**Books:**

- "Introduction to Machine Learning with Python: A Guide for Data Scientists" by Andreas C. Müller & Sarah Guido.
- "Python Machine Learning" by Sebastian Raschka and Vahid Mirjalili.
- "Healthcare Data Analytics" by Chandan K. Reddy, Charu C. Aggarwal, and Hillol Kargupta

**Journals and Papers:**

- "Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology" by Alaa M. Ahmed et al.
- "Machine Learning in Medicine: A Practical Introduction" by Thomas H. McCoy Jr. et al.
- "Machine Learning for Predictive Modeling of Health Outcomes" by Rajkomar et al.

**Online Resources:**

- PubMed Central (PMC) - Machine Learning and Healthcare Section.
- arXiv.org - Machine Learning (cs.LG)
- Google Scholar
- Wikipedia