Disease Prediction using Machine Learning

Aligned with SDG 3 - Good Health and Well-Being

1. Introduction

This project supports Sustainable Development Goal 3 by leveraging machine learning to predict the likelihood of diabetes in individuals based on health-related features. Early prediction can help improve health outcomes and reduce healthcare costs.

2. Problem Statement

The goal is to develop a classification model that can predict whether a person is diabetic using their medical attributes such as glucose level, BMI, age, and insulin levels.

3. Dataset

- Source: Kaggle - Pima Indians Diabetes Database
- Records: 768 samples, 9 columns
- Target Variable: Outcome (0 = Non-diabetic, 1 = Diabetic)
- Features: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age

4. Methodology

- Tools: Python, Pandas, Scikit-learn, Matplotlib, Seaborn
- Steps:
  1. Data loading and preprocessing
  2. Feature scaling with StandardScaler
  3. Train/test split (80/20)
  4. Model training using Random Forest Classifier
  5. Evaluation using accuracy, confusion matrix, classification report

5. Results

- Accuracy: ~78%

- Confusion Matrix: Shows high precision and recall for both classes

- Top Features: Glucose, BMI, Age

6. Conclusion

The Random Forest model achieved good predictive performance and can be a useful tool in health screening processes. Early disease prediction aligns with SDG 3 by enabling preventive action and improving healthcare access.

7. Future Work

- Use larger and more diverse datasets

- Try deep learning models

- Deploy as a simple web application using Streamlit or Flask