DEPI Diabetes Detection - Technical Documentation

Date: May 15, 2025

Table of Contents

---

1. Introduction

1.1 Project Overview

The **DEPI Diabetes Detection** system is a Streamlit-based web application developed to help users assess their risk of developing diabetes. It offers two primary functionalities:

1. An interactive self-assessment form powered by a machine learning model to estimate diabetes risk.

2. A dynamic data analysis dashboard that visualizes trends and insights from a comprehensive diabetes-related dataset.

The system combines an easy-to-use interface with data-driven analytics, making it accessible for both the general public and healthcare enthusiasts.

1.2 Objectives

- Deliver an intuitive and informative risk prediction tool for diabetes.
- Enable users to explore diabetes-related trends via an interactive dashboard.
- Apply robust machine learning techniques for reliable and interpretable predictions.
- Build a modular, scalable codebase that supports further development and deployment.

1.3 Target Audience

- **General Users:** Individuals seeking a simple tool to understand their diabetes risk based on health and lifestyle inputs.
- **Students and Enthusiasts:** Learners or analysts interested in exploring patterns in diabetes-related health data.

1.4 Technologies Used

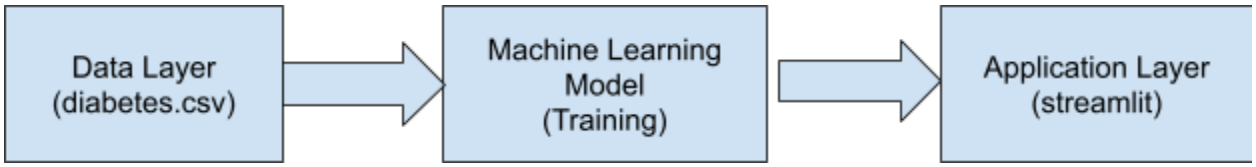
- **Programming Language:** Python
- **Web Framework:** Streamlit
- **Data Manipulation:** Pandas, NumPy
- **Machine Learning:** Scikit-learn, XGBoost
- **Visualization:** Matplotlib, Seaborn
- **Model Persistence:** Joblib, Pickle

---

2. System Architecture

2.1 High-Level Overview

The system follows a multi-layered architecture:



2.2 Components

- **Dataset:** `diabetes.csv`
- **Notebook:** `diabetes2.ipynb`
- **Serialized Models:** `depi_xgb.pkl`, `diabetes_model.pkl`
- **Streamlit Scripts:** `home.py`, `input_form.py`, `analysis.py`, `prediction.py`

---

3. Data Layer

3.1 Dataset: `diabetes.csv`

- **Format:** CSV
- **Records:** 235,495
- **Features:** 22

3.2 Data Dictionary (Key Features)

- `Diabetes_binary`, `HighBP`, `HighChol`, `CholCheck`, `BMI`, `Smoker`, `Stroke`, `HeartDiseaseorAttack`, `PhysActivity`, `Fruits`, `Veggies`, `HvyAlcoholConsump`, `AnyHealthcare`, `NoDocbcCost`, `GenHlth`, `MentHlth`, `PhysHlth`, `DiffWalk`, `Sex`, `Age`, `Education`, `Income`

3.3 Data Source

- Kaggle

---

4. Machine Learning Model

4.1 Model Selection

- **Primary:** XGBoost Classifier (Accuracy: 84%)
- **Secondary:** Random Forest Classifier (Accuracy: 70%)

4.2 Training Pipeline (`diabetes2.ipynb`)

4.2.1 Data Loading

- Load dataset, check stats and missing values.

4.2.2 Preprocessing

- StandardScaler
- SMOTE for class balancing
- `train_test_split`

4.2.3 Feature Selection

- All available features used after cleaning.

4.2.4 Training

- `XGBClassifier` with `GridSearchCV`

4.2.5 Evaluation

- Classification Report
- Confusion Matrix
- **Accuracy: 84% (XGBoost Classifier)**

4.3 Model Persistence

- `depi_xgb.pkl` (XGBoost)

4.4 Prediction Features

For `depi_xgb.pkl` (`input_form.py`)

- Uses all major features (21 fields)

For `diabetes_model.pkl` (`prediction.py`)

- Uses subset of 12 features

---

5. Application Layer (Streamlit Web Application)

5.1 Overview

Streamlit-based web interface divided into several functional pages.

5.2 Core Python Scripts

5.2.1 `home.py`

- Welcome screen with navigation to `input_form` and `analysis`

5.2.2 `input_form.py`

- Risk assessment form
- Predict using `depi_xgb.pkl`
- Displays risk, confidence, and recommendations

5.2.3 `analysis.py`

- Loads `diabetes.csv`
- Offers filters by Age and Health
- Visualizes data distributions, correlation heatmaps

5.2.4 `prediction.py`

- Uses `diabetes_model.pkl` for predictions

5.3 User Interface Flow

1. Home page → Navigation
2. Input form → User input → Prediction
3. Analysis page → Filter → Visual insights

---

6. Deployment

Running the Application

- Deployed Version: https://depiproject-a7txbipsjwvaawrmsftjef.streamlit.app/
- Codebase: streamlit run home.py

---

7. Dependencies

- `streamlit`
- `xgboost`
- `scikit-learn`
- `matplotlib`, `seaborn`
- `pandas`, `numpy`

- `joblib`, `pickle`
- `imbalanced-learn`

---

8. Directory Structure

```
project_root/
├── model/
│   └── depi_xgb.pkl
├── data/
│   └── diabetes.csv
├── scripts/
│   ├── home.py
│   ├── input_form.py
│   ├── analysis.py
│   └── prediction.py
├── model/
│   └── diabetes2.ipynb
└── README.md
```

---

9. Limitations

- Predictions are probabilistic, not diagnostic.
- Model trained on a specific dataset – may not generalize to all populations.
- Imbalanced dataset