Health Care Data Warehouse User Manual

# Table of Contents

1. Introduction
2. System Requirements
3. Installation Guide
4. Getting Started
5. Features Overview
6. User Interface Guide
7. How to Perform Tasks

* Task 1 : Data Extraction with SSIS
* Task 2 : EDA and Cleaning using Python
* Task 3 : Data Analysis with Power BI
* Task 4 : Disease Prediction using Machine Learning
* Task 5 : Web Design using Streamli t

1. Troubleshooting
2. FAQ
3. Contact Support

# 1. Introduction

The **Health Care Data Warehouse** is designed to help users extract, manage, and analyze health-related data, enabling efficient data analysis and disease prediction using machine learning. The primary features include data extraction from multiple health datasets, analysis with Power BI, and disease prediction with machine learning models deployed via Streamlit. It’s suitable for healthcare professionals, data engineers, and data analysts.

# 2. System Requirements

This section lists the hardware and software requirements to run the Health Care Data Warehouse system.

* **Operating System**: Windows 10 or later / macOS X 10.14 or later
* **Processor**: Intel i5 or higher
* **Memory**: 8 GB RAM minimum, 16 GB recommended
* **Storage**: 2 GB of available space for data storage
* **Software**: SQL Server, SSIS, Power BI, Python (3.8+), Azure Synapse, Streamlit
* **Additional Requirements**: Internet connection for cloud-based processing and Power BI online integration

# 3. Installation Guide

Follow these steps to set up the Health Care Data Warehouse project:

1. Download and install the necessary software:
   * SQL Server and SSIS
   * Power BI Desktop
   * Python (install necessary libraries like pandas, scikit-learn, streamlit)
   * Azure Synapse for cloud processing
2. Set up your SQL database and Azure Synapse environment.
3. Download the CSV files (diabetes, brain stroke, heart disease) from the data source.
4. Follow the SSIS guide to load data into the data warehouse.
5. Set up Power BI for analysis and visualization.
6. Launch the Streamlit application to access the disease prediction model.

# 4. Getting Started

This section guides users through the initial steps after installation.

1. **SQL Setup**: Create tables for each disease (diabetes, brain stroke, heart disease) in the data warehouse.
2. **SSIS ETL**: Use SSIS to extract, transform, and load (ETL) data from CSV files into the SQL data warehouse.
3. **Power BI**: Connect Power BI to your data warehouse for visualization and analysis.
4. **Streamlit App**: Run the Streamlit web app for real-time disease prediction using the machine learning model.

# 5. Features Overview

Core features of the Health Care Data Warehouse project:

* **ETL Process**: Automated data extraction, transformation, and loading using SSIS.
* **Data Warehouse**: Centralized storage of health-related datasets (diabetes, brain stroke, heart disease).
* **Data Analysis**: Interactive reports and dashboards using Power BI for data analysis.
* **Machine Learning Models**: Predict the likelihood of diseases using trained models (logistic regression, decision trees, etc.).
* **Web Interface**: User-friendly interface for disease prediction using Streamlit.

# 6. User Interface Guide

Visual guide to navigating the project’s key components:

* **SSIS Interface**: Used for creating and managing ETL packages.
* **Power BI Dashboards**: Provides visualization of patient data, including disease statistics.
* **Streamlit Web Page**: A simple interface where users can input health data and receive predictions on disease likelihood.
* **Azure Synapse**: Manage cloud-based data processing.

# 7. How to Perform Tasks

**Task 1: Extracting Data from CSV Files**

1. Open SSIS and create a new ETL package.
2. Set up data sources pointing to the CSV files (diabetes, brain stroke, heart diseases).
3. Define transformation rules and map the data to the SQL database tables.
4. Run the ETL process to load data into the data warehouse.

**Task 2: Analyzing Data in Power BI**

1. Open Power BI Desktop and connect it to your SQL database.
2. Import tables containing health data.
3. Create visualizations, such as bar charts and line graphs, to analyze trends.
4. Publish the dashboard to Power BI Online for collaborative access.

**Task 3: Predicting Disease Using Machine Learning Models**

1. Open the Streamlit application using Python (streamlit run app.py).
2. Enter the patient data (e.g., age, blood pressure, cholesterol).
3. Select the disease type (diabetes, brain stroke, heart disease) for prediction.
4. View the prediction result and confidence score based on the machine learning model.

# 8. Troubleshooting

Common issues and solutions:

* **Issue**: SSIS package fails during data extraction.
  + Solution: Check file paths and ensure that the CSV files are accessible.
* **Issue**: Power BI dashboard not refreshing data.
  + Solution: Verify the connection to the SQL database and refresh settings.
* **Issue**: Streamlit web page is not loading.
  + Solution: Ensure Python and necessary libraries (Streamlit, pandas, scikit-learn) are installed correctly.

# 9. FAQ

 **Q**: How do I update the data in the warehouse?

* **A**: Re-run the SSIS ETL package to extract the latest data from the CSV files.

 **Q**: Can I customize the machine learning models?

* **A**: Yes, the ML models are built using Python libraries and can be retrained with new datasets or updated using different algorithms.

 **Q**: How can I add more data sources?

* **A**: Update the SSIS package to include additional data files, ensuring proper mapping to the data warehouse.

# 10. Contact Support

For any assistance, please contact our support team:

* **Email**: support@healthdatawarehouse.com
* **Phone**: +20 112 900 6242
* **Website**: www.healthdatawarehouse.com/support