Disease Prediction System

ITS307 DATA ANALYTICS

BACHELOR OF SCIENCE IN INFORMATION TECHNOLOGY

(YEAR III, SEMESTER I)

RESEARCHER (S)

PEMA DENDUP (12200070)

SONAM PELKI (12200081)

UGYEN KEZANG (1222094)

UGYEN LHAMO (12200095)

GUIDED BY

NIMA DEMA

Gyalpozhing College of Information Technology

Gyalpozhing, Mongar

1. Proposed Methods

1.1. System Overview

For the disease prediction, there are going to be two models; one for Heart Disease Prediction and another for Diabetes Disease Prediction.

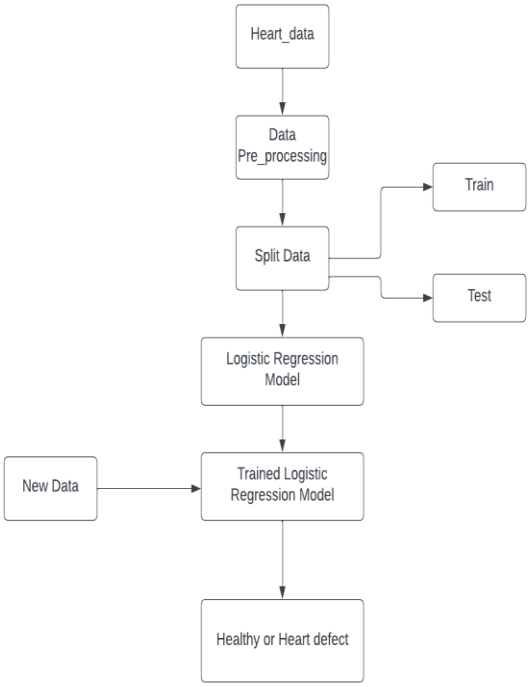

Figure 1: Heart Disease

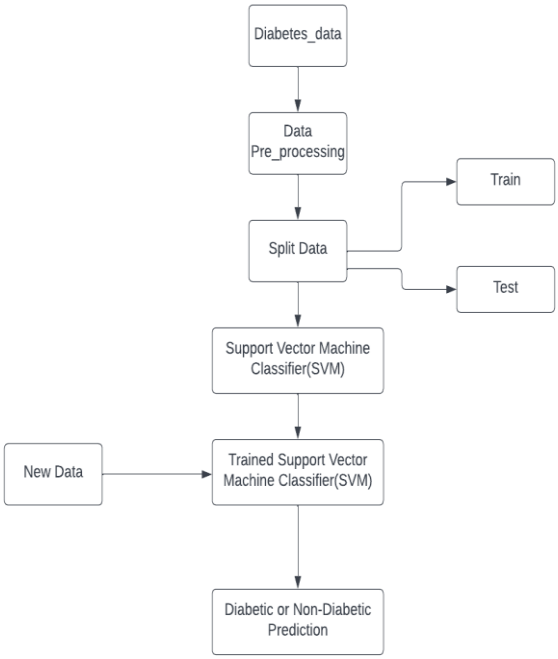

Figure 2: Diabetes Disease

1.2. Algorithm

Some of the algorithm used in our projects:

I. Logistic Regression

For the heart disease prediction system we used Logistic Regression. Because our data is binary classification, we are going to classify whether the person has heart disease or not. Since Logistic Regression is best algorithm when it comes to binary classification

II. Support Vector Machine Classifier (SVM)

For the diabetes prediction system, the SVM algorithm is used. The SVM algorithm primarily creates a line to separate the dataset into classes, enabling it to decide the test data into which classes it belongs. The line or decision boundary is called a hyperplane. The algorithm works on two types: linear and nonlinear.

Linear SVM is used when the dataset comprises two classes and is separable. In case of inseparable dataset, a nonlinear SVM is applied, where the algorithm converts the original coordinate area into a separable space. There can be multiple hyperplanes, and the best hyperplane is chosen with the maximum margin between data points. The dataset closest to the hyperplane is called a support vector.

1.3. Dataset

The dataset is from the kaggle, for both the models (i.e., heart disease prediction and diabetes prediction system).

Both the dataset are structured as it is in csv file format.

- Heart Disease Prediction
  The dataset comprises 10 numerical features including the target.
  The features include:
  i. age
  ii. Sex
  iii. Chest pain type (cp) with four values
  iv. Maximum Heart Rate achieved (thalach)
  v. Exercise induced angina (exang)
  vi. Oldpeak(oldpeak) = ST depression induced by exercise relative to rest.
  vii. Slope of the peak exercise ST segment.
  viii. Number of Major Vessels (0-3) colored by fluoroscopy (ca)
  ix. Thal: 0=normal; 1=fixed defect; 2=reversal defect.
  x. Target: 0=normal; 1=heart disease.

- Diabetes Disease Prediction
  The dataset comprises 9 features including the feature.
  Features include Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, and Outcome (Target).

1.4. Evaluation Metrics

To evaluate the performance of the model we have used an accuracy score. It is calculated by dividing the number of correct predictions by the total prediction number.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

For binary classification, accuracy can also be calculated in terms of positive and negative as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TP=True Positives, TN=True Negatives, FP=False Positives, FN=False Negatives.

1.5. Experimental Setup

- Programming language: Python
- Platform for training a model: Jupyter Notebook
- Machine Learning Libraries: Pandas, NumPy, Sci-Kit Learn.
- Python Libraries for Visualization: Matplotlib and Seaborn.