Prediction of Diabetes Complications using Machine Learning

Supervisor: Dr. Md. Ashraful Alam, Asst. Professor, Dept. of CSE

Aniqa Zaida Khanom (15101106) Sheikh Mastura Farzana (15101077) Tahsinur Rahman (15101128)

Abstract

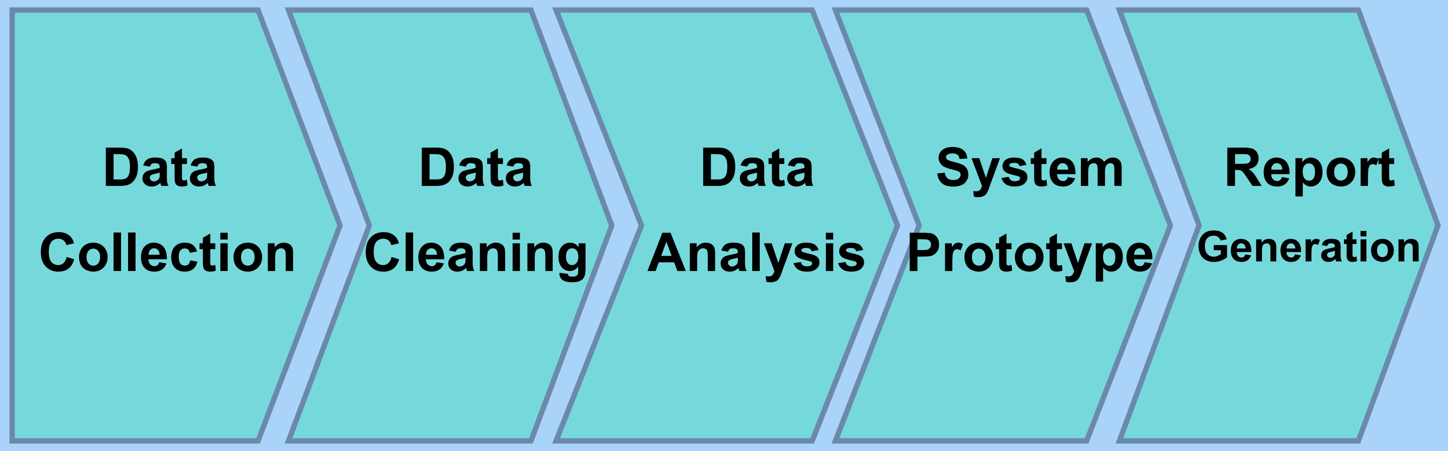
Diabetes Mellitus type 2 (T2DM) is the most common form of diabetes [WHO (2008)]. Bangladesh has a disproportionately high diabetes population with more than 8.4 million people according to research published in WHO bulletin in 2013. Unfortunately, most part of this huge population are not aware of the possible physical complications related to this disease. These complications can be either microvascular such as nephropathy, neuropathy and retinopathy or macrovascular such as coronary complications. Our proposed system aims to collect data from patients already diagnosed with T2DM and then apply data mining methods to establish approximately accurate prediction model to predict possible future complications caused by T2DM. The variables we will consider are age, gender, time since diagnosis, BMI, Hba1c, ACR, hypertension, cholesterol, etc. Our prediction model will be set up by applying Machine Learning algorithms to predict the onset of micro and macrovascular complications. The model will use supervised learning to classify the data using Logistic Regression, Naïve Bayes and Support Vector Machine (SVM) algorithms. We believe such a model can be of great use to the diabetic patients in Bangladesh to help them manage the disease better. Moreover, it can also assist doctors with diagnosis, prognosis and treatment planning of the patients.

Literature Review

The papers we went through concerning diabetes talk about how diabetes can trigger other microvascular and macrovascular diseases and their relationship with the variables. In the papers regarding machine learning, missForest algorithm was used broadly to determine and solve incomplete instances of datasets. Suitable strategies were used to handle class imbalance as well. Furthermore, decision trees and logistic regression (LR) are classification algorithms widely used in most predictive models. One similar model which used LR approach could achieve accuracy up to 0.838.

Dataset

The data for our project will be provided by BIRDEM following a brief data collection protocol. Initially we will be using Glycated Haemoglobin (HbA1c), Systolic/Diastolic Blood Pressure, Age, Time since onset of Diabetes, Body Mass Index(BMI), Hypertension, Lipid Profile, Family History, Gender, Smoking, ACR, etc. of around 4000 patients diagnosed with T2DM.

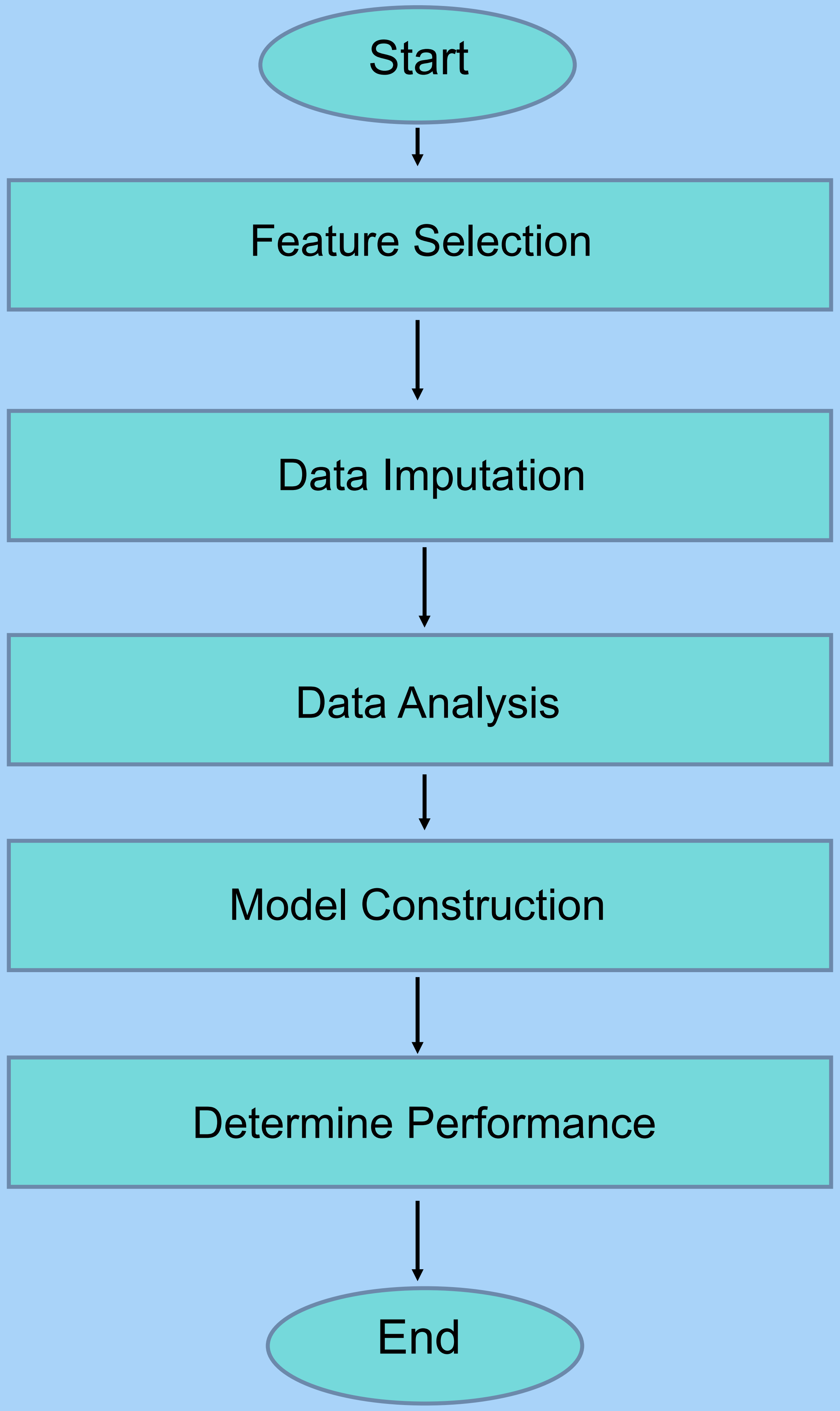

Data Collection

Data Cleaning

Data Analysis

System Prototype

Report Generation


Start

Feature Selection

Data Imputation

Data Analysis

Model Construction

Determine Performance

End

Data Cleaning and Analyzing

We propose to use Random Forest (RF) approach for the imputation of missing values in each variable. Additionally, in this phase, we will also determine and remove anomalies, to get a more accurate dataset. From the cleaned data, we will perform some basic statistical regression analysis to determine the nature of the dataset. We intend to not consider whether the data is balanced or imbalanced.

Implementation

The thresholds within which these complications may develop are 2, 3 and 5 years. Using 10-fold cross validation, we will randomly partition the dataset into 10 equal parts where one subsample will be retained as validation data for testing and the rest will be used to train the model. Logistic Regression (LR), Support Vector Machine (SVM) and Naïve Bayes (NB) are the classification algorithms we will be using to train the model.

Performance Assessment

In order to determine the performance of our model we will use the classification functions- sensitivity, specificity and accuracy.

References

[1] Dagliati, A., Marini, S., Sacchi, L., Cogni, G., Teliti, M., Tibollo, V., De Cata, P., Chiovato, L. and Bellazzi, R. (2017). Machine Learning Methods to Predict Diabetes Complications. *Journal of Diabetes Science and Technology*, p.193229681770637.

[2] Cichosz SL, Johansen MD, Hejlesen O. Toward big data analytics: review of predictive models in management of diabetes and its complications. *J Diabetes Sci Technol*. 2015;10(1):27-34.

[3] Stratton IM, Kohner EM, Aldington SJ, et al. UKPDS 50: risk factors for incidence and progression of retinopathy in type II diabetes over 6 years from diagnosis. *Diabetologia*. 2001;44(2):156-163.

[4] Yau JWY, Rogers SL, Kawasaki R, et al. Meta-Analysis for Eye Disease (META-EYE) Study Group. Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care*. 2012;35(3):556-564.

[5] Scirica BM, Bhatt DL, Braunwald E, et al. Prognostic implications of biomarker assessments in patients with type 2 diabetes at high cardiovascular risk: a secondary analysis of a randomized clinical trial. *JAMA Cardiol*. 2016;1(9):989-998.