CS361: Diabetes Prediction

Bharat Narah (210101030) [1]

Abstract

In the realm of healthcare, early detection and management of chronic diseases like diabetes are paramount. This project focuses on developing a predictive model for assessing diabetes risk using machine learning techniques. By analyzing relevant medical data such as patient demographics, physiological indicators, and lifestyle factors, the model employs algorithms such as Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), Support Vector Machines (SVM) and Gradient Boosting Classifier to predict the likelihood of an individual developing diabetes. Through comprehensive data preprocessing, feature selection, and model evaluation, the proposed system aims to provide accurate risk assessments, enabling proactive interventions and personalized healthcare strategies. The model's effectiveness in early detection and risk stratification can significantly impact public health initiatives, fostering timely interventions and improved patient outcomes in the management of diabetes.

1. Introduction

Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood sugar levels, leading to various complications affecting multiple organ systems. The prevalence of diabetes has been steadily increasing worldwide, posing significant challenges to public health systems and individual well-being. Early detection and effective management are crucial for preventing complications and improving patient outcomes. However, traditional diagnostic methods often rely on periodic screenings or clinical symptoms, which may delay diagnosis and intervention.

Motivated by the need for proactive healthcare solutions, this project aims to develop a predictive model for diabetes risk assessment using machine learning techniques. By leveraging advancements in data science and healthcare informatics, the objective is to create a tool that can accurately predict an individual's likelihood of developing diabetes based on relevant demographic, physiological, and lifestyle factors. This proactive approach enables early identification of high-risk individuals, facilitating timely interventions such as lifestyle modifications, medical treatments, and patient education programs.

The overarching goal of this project is to contribute to the advancement of personalized medicine and preventive healthcare strategies. By harnessing the power of machine learning algorithms, the proposed model seeks to empower healthcare providers with a predictive tool that enhances risk stratification and facilitates targeted interventions for individuals at risk of developing diabetes. Ultimately, the project aims to improve patient outcomes, reduce healthcare costs associated with diabetes-related complications, and promote a proactive approach to chronic disease management in clinical practice.

2. Methods

2.1. Title

Predict the onset of diabetes based on diagnostic measures.

2.2. Algorithms

2.2.1. K-Nearest Neighbors (KNN):

We will utilize KNN to classify individuals' diabetes risk based on their nearest neighbors in the dataset.

2.2.2. Random Forest:

We will employ Random Forest to capture complex relationships between the attributes in the dataset and predict diabetes risk accurately.

2.2.3. Logistic Regression:

We will apply Logistic Regression to estimate the probability of an individual having diabetes based on their attributes such as pregnancies, glucose levels, blood pressure, etc

2.2.4. Support Vector Machines (SVM):

Support Vector Machines (SVMs) are chosen for diabetes prediction due to their ability to handle non-linearity, robustness to overfitting, flexibility in kernel functions, effectiveness with high-dimensional data, and well-established theory and implementation.

Algorithm 1 K-Nearest Neighbors

Input: Training dataset D, number of neighbors k
Calculate the distance between the new data point and all training points
Identify the k nearest neighbors
Output: Predicted class based on majority class among neighbors

Algorithm 2 Logistic Regression

Input: Training data X, learning rate, logistic regression parameters
Initialize weights w and bias b
for each iteration **do**
  Compute logistic function to predict probabilities
  Update weights using gradient descent
end for
Output: Predicted Class based on learn model

Algorithm 3 Random Forest Classifier

Input: Training data D, number of trees N, number of features to consider at each split m
Initialize a list $forest$ to store the N trees
for $i = 1$ **to** N **do**
  Randomly sample m features from the total M features
  Generate a bootstrapped dataset D_{sample} by sampling with replacement from D
  Train a decision tree T_i using D_{sample} considering only the sampled features at each split
  Add T_i to the $forest$
end for
Output: Random forest $forest$

Algorithm 4 SVM

Input: Training dataset D, SVM parameters
Choose a kernel function(eg., linear, polynomial, radial basis function)
Formulate and solve the optimization problem to find the hyperplane
Classify new data points based on the sign of the decision function
Output: Sign of the decision function

3. Tables

Dataset

key attributes for diabetes risk assessment, including pregnancies, glucose, blood pressure, insulin, skin thickness, BMI, age, and diabetes pedigree function. Each attribute contributes crucial information for predictive modeling. This table aids in understanding the dataset's features essential for developing accurate risk assessment models. Attributes such as glucose and BMI are particularly significant predictors of diabetes risk. Utilizing these attributes effectively can enhance proactive interventions and personalized healthcare strategies.

4. Intended Experiments

4.1. Data collection and pre-processing / Data Cleaning:

Gather medical data including demographics, physiological indicators (glucose levels, blood pressure), and lifestyle factors (BMI, age) from diabetes patients. Gather medical data including demographics, physiological indicators (glucose levels, blood pressure), and lifestyle factors (BMI, age) from diabetes patients.

Refining Data:

1.1 Replacing the 0 values from ['Glucose','BloodPressure','SkinThickness','Insulin','BMI'] by NaN.
2.1 To fill these Nan values the data distribution needs to be understood
2.2 Plotting histogram of dataset before replacing NaN values
3.1 Replacing NaN value by mean, median depending upon distribution

4.2. Model Selection:

Choose Algorithms: Select appropriate machine learning algorithms for classification tasks such as Logistic Regression, Random Forest, SVM, or K-Nearest Neighbors.

To calculate the best accuracy for the four models, you would typically loop through each model, perform cross-validation with grid search to find the best hyperparameters, and then select the model with the highest accuracy. Here's a step-by-step approach to calculating the best accuracy:

1. Define the models along with their hyperparameters.
2. Use cross-validation with grid search (GridSearchCV) to find the best hyperparameters for each model.
3. Keep track of the best accuracy and the corresponding best model.

According to Dataset SVM is giving highest accuracy, then logistic Regression and then random forest model. But for further process i will choose random forest.

4.3. Model Evaluation:

Split Data: Divide the dataset into training and testing sets to evaluate model performance. Train Models: Train selected algorithms on the training data using hyperparameter tuning if necessary.

5. References

Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar," Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop", International Conference On I-SMAC, 978-1-5090-3243-3, 2017. .