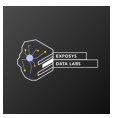
Diebetes detection with machine learning algorithms

A project report submitted by
Sachin Negi

in partial fulfillment of the requirements for the internship certificate
Data Science



Exposys Data Labs

Abstract

Diebetes is a type of chronic disease which is more common among the people of all age groups.Predicting the disease at an early stage can help a person to take necessary precautions and change his/her lifestyle according to either prevent the occurance of this disease or control the disease(for people who already have the disease).

Contents

Chapter 1

Introduction

Diabetes mellitus (DM), commonly known as just diabetes, is a group of metabolic disorders characterized by a high blood sugar level over a prolonged period of time. Symptoms often include frequent urination, increased thirst and increased appetite. If left untreated, diabetes can cause many health complications. Acute complications can include diabetic ketoacidosis, hyperosmolar hyperglycemic state, or death.Serious long-term complications include cardiovascular disease, stroke, chronic kidney disease, foot ulcers, damage to the nerves, damage to the eyes and cognitive impairment.

The project include prediction of Diebettes based on some features. More than one dataset is taken and some features are selected from them. After that both dataset are merged together. Various machine learning algorithm are used. The algorithm include Linear Regression, Adaboost, Bagging decision tree based, Bagging K-neighbours and MLP(Multi Layer Perceptron) Classifier. The accuracy obtained from different algorithm is compared.

Chapter 2

Existing Method

Glycosylated Hemoglobin Test (A1c)

The A1c test is a common test that measures how much sugar has been in your blood over the previous few months. It's used to diagnose diabetes or determine how well a person with the disease has been managing their blood sugar. Some people have different types of hemoglobin (red blood cell protein) that may affect results.

What's a glycosylated hemoglobin test?

A glycosylated hemoglobin test measures the amount of glucose (sugar) in your blood. The test is often called A1c, or sometimes HbA1c. It's a simple blood test used to:

- Detect prediabetes — high sugar levels that can lead to diabetes, heart disease and stroke.
- Diagnose diabetes.
- Tell how well a person with diabetes has been managing the disease.

How does the A1c test work?

A1c testing relies on hemoglobin. Hemoglobin is the part of the red blood cell that carries oxygen throughout the body. When you have glucose in your blood, it glycates (sticks) to hemoglobin. The more glucose in your blood, the more it sticks. And it can stay there for around three months, or about how long the average red blood cell lives.

The A1c test measures the average amount of glucose that's been attached to hemoglobin over time. So the A1c test provides more information about blood sugar over a longer period of time than home monitoring does.

What do the A1c test results mean?

An A1c test result gets reported as a percentage. The number represents the portion of hemoglobin proteins that are glycosylated, or holding glucose. The higher the percentage, the higher your blood sugar levels have been over the last few months.

- Less than 5.7 % means you don't have diabetes.
- 5.7 % to 6.4 % signals pre-diabetes.
- 6.5 % or higher means a diabetes diagnosis.
- 7 % or lower is the goal for someone trying to manage their diabetes.

Chapter 3

Proposed method with Architecture

The main idea of the method used is:
Three features are selected which are:

- ArmCircum:
- SaggitalAbdominal
- GripStrength

Mid-Upper Arm Circumference (MUAC) is the circumference of the left upper arm, measured at the mid-point between the tip of the shoulder and the tip of the elbow (olecranon process and the acromium).

MUAC is used for the assessment of nutritional status. It is a good predictor of mortality and in many studies, MUAC predicted death in children better than any other anthropometric indicator. This advantage of MUAC was greatest when the period of follow-up was short.

The MUAC measurement requires little equipment and is easy to perform even on the most debilitated individuals. Although it is important to give workers training in how to take the measurement, the correct technique can be readily taught to minimally trained health workers and community-based volunteers. It is thus suited to screening admissions to feeding programs during emergencies.

Sagittal abdominal diameter Sagittal abdominal diameter (SAD) is a measure of visceral obesity, the amount of fat in the gut region. SAD is the distance from the small of the back to the upper abdomen. SAD may be measured when standing or supine. SAD may be measured at any point from

the narrowest point between the last rib and the iliac crests to the midpoint of the iliac crests.

SAD is a strong predictor of coronary disease, with higher values indicating increased risk independent of BMI.

GripStrength Grip strength is the force applied by the hand to pull on or suspend from objects and is a specific part of hand strength. Optimum-sized objects permit the hand to wrap around a cylindrical shape with a diameter from one to three inches. Stair rails are an example of where shape and diameter are critical for proper grip in case of a fall. Other grip strengths that have been studied are the hammer and other hand tools. In applications of grip strength, the wrist must be in a neutral position to avoid developing cumulative trauma disorders (CTDs).

Grip strength is a general term also used to refer to the physical strength of an animal and, for athletes, to the muscular power and force that can be generated with the hands. In athletics, grip strength is critical for rock climbers and is an important factor in strongman competitions and weight lifting. Grip strength training is also a major feature in martial arts and can be useful in various professions where people must work with their hands.

Different machine learning algorithm is used:

The algorithm include Linear Regression, Adaboost, Bagging decision tree based, Bagging K-neighbours and MLP(Multi Layer Perceptron) Classifier. The data is feed to the algorithm to learn the features. Based on the learning of the features. The algorithm predicts the person to be Diebetic or not Diebetic.

Chapter 4

Methodology

Data

The data is collected from two different sources. Some features are selected from each dataset and then both the dataset and combined together.

Their are 3 columns(features) and 9813 rows.

Pre-processing

The data is divided into two parts.

i) Training set

ii) Test set

The training set consist of 6000 rows. The remaining 3813 rows is Test set.

Application of Algorithm

Machine Learning Algorithm is applied which include

Linear Regression

Adaboost

Bagging decision tree based

Bagging K-neighbours

MLP(Multi Layer Perceptron) Classifier

The algorithm learn the features from training set. The learned features is used to predict the test set.

Chapter 5

Implementation

Inbuilt library is used to implement different machine learning algorithm. Sklearn library is used to implement algorithm AdaBoostClassifier, DecisionTreeClassifier, BaggingClassifier, KNeighborsClassifier.

Distribution of each features are plotted to analyse the distribution

Heat map of features are plotted

Pair plot is also plotted for the features

A comparative study is done based on the accuracy obtained from different machine learning algorithm.

Chapter 6

Conclusion

The accuracy obtained from different machine learning algorithm is as follows:

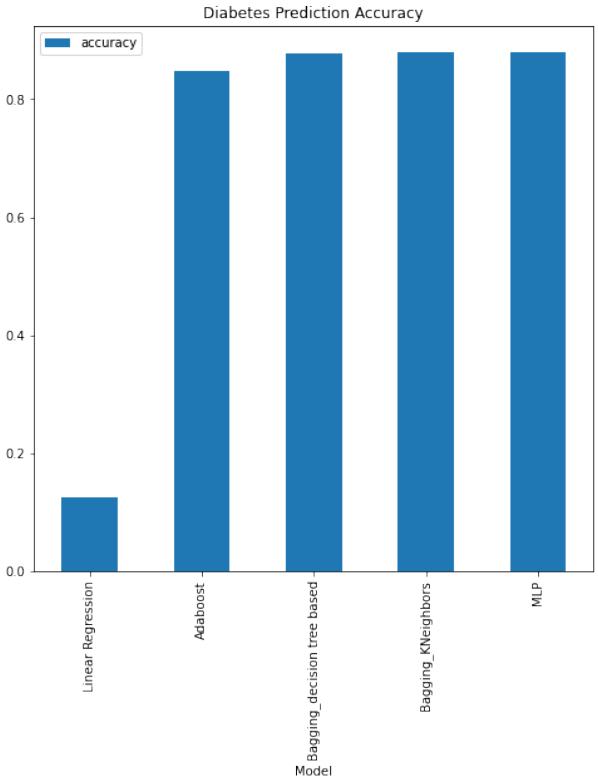


Figure 6.1: Accuracy

Accuracy obtained

- Linear Regression - 0.123966 %
- Adaboost - 0.847364 %
- Bagging Decision tree based - 0.878049 %

- Bagging K-neighbors - 0.880671 %
- Multi Layer Perceptron - 0.879098 %