Machine Learning Engineer Nanodegree 2019

Capstone Proposal

Early identification of Prediabetic using machine learning

Asis Pattnaik
July 2019

Domain Background

India has 61.3 million diabetes patients. The numbers are estimated to rise to 101.2 million by 2030.

Out of all the cases of diabetes, 95% cases are of Type 2 diabetes mellitus. The type 2 diabetes can go unnoticed for a long time, till it manifests in the form of metabolic syndromes like Retinopathy, Neuropathy, Nephropathy and Cardio vascular diseases. The prediabetic stages also carry high risk for cardiovascular diseases (CVDs).

There have been many studies that shows this large-scale epidemic is faced not only in India, but across the world at large.

The type 2 Diabetes is a curable lifestyle disease. Due to ignorance, socio economic reasons and unhealthy food habits it has become such a big problem. If it can be diagnosed at a prediabetes level, many young and old folks alike can get the treatment in time and will be benefited from it.

Reference articles:

1. http://www.apiindia.org/medicine_update_2013/chap40.pdf
2. https://globalizationandhealth.biomedcentral.com/articles/10.1186/s12992-014-0080-x
3. https://apps.who.int/iris/bitstream/handle/10665/204871/9789241565257_eng.pdf;jsessionid=770028777591CC613FC338061D0255F1?sequence=1
4. https://www.who.int/diabetes/country-profiles/ind_en.pdf
5. https://www.researchgate.net/publication/51607936_The_Indian_Council_of_Medical_Research-India_Diabetes_ICMR-INDIAB_study_methodological_details
6. https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(18)30387-5/fulltext
7. https://www.sciencedirect.com/science/article/pii/S2214999616000035
8. https://www.who.int/en/news-room/fact-sheets/detail/diabetes

Problem Statement

Due to the asymptomatic nature of disease and the low disease awareness among the population, diagnosis of this disease is delayed by several years. As a result, many subjects already have vascular complications at the time of diagnosis of diabetes.

The current challenges

1. High prevalence of prediabetes and very high conversion rate from prediabetes to diabetes (to the tune of 18% per year), is a great challenge. India has nearly 3.0% of adults with prediabetes that constitutes impaired fasting glucose (IFG) and impaired glucose tolerance (IGT).
2. Many studies show that Indians have more body fat for any given body mass index (BMI) compared with other countries in Europe, America and Africa. Indians also have higher levels of central obesity (measured as waist circumference [WC], WHR, visceral fat, and posterior subcutaneous abdominal fat). The prevalence of overweight in adults in India increased from 9·0% in 1990 to 20·4% in 2016; this prevalence increased in every state of the country. For every 100 overweight adults aged 20 years or older in India, there were 38 adults with diabetes, compared with the global average of 19 adults in 2016.
3. The food habits, along with the climate in India differs significantly from region to region. This could be also a reason why the diabetes in some of the states is more prevalent than the others.

Datasets and Inputs

The dataset that will be used is available from the "data.gov.in" on the Annual health survey 2014: Clinical, Anthropometric & Bio-chemical (CAB) Survey.

https://data.gov.in/resources/clinical-anthropometric-bio-chemical-cab-2014-survey-data-district-sitamarhi-bihar

The survey was carried out in the following nine states in India:

RAJASTHAN
UTTAR PRADESH
ODISHA
ASSAM
MADHYA PRADESH
CHHATTISGARH
UTTARAKHAND
BIHAR
JHARKHAND

The survey was carried out in 288 districts in these states. Out of the indices captured in this survey the following indices will be analyzed for prediction of prediabetes.

Age
BP_Diastolic
BP_Diastolic_2reading
BP_systolic
BP_systolic_2_reading
Haemoglobin_level
Length_height_cm
Marital_status
Pulse_rate
Pulse_rate_2_reading
Sex
Weight_in_kg
date_survey
district_code
fasting_blood_glucose_mg_dl
state_code
year_of_birth
BMI

Solution Statement

The objective of this project is to come up with a simple and noninvasive diabetes risk score that can predict the risk percentage specific to a region and sex. It can help early identification of the prediabetic stage and also help in management of related abnormalities such as dyslipidemia and hypertension and to preserve the beta cell function.

Along with the high blood glucose level, in diabetes there is also prevalence of hypertension. The BMI also a factor that influences the occurrence of diabetes. These parameters can be easily be accessed. If there is a correlation identified between these abnormalities among a populace, it can be a good identifier for possible diabetes cases.

We will develop a deep neural network model that can predict the risk of prediabetes based on the available parameters like BMI index, Age, Sex, BP, hemoglobin level, Geo Social factors.

This diabetes risk score will help significantly in alarming the high-risk individual to take appropriate action much before the diabetes surfaces. This will help in reducing the unprecedented health care challenges in India that has arose due to diabetes.

Benchmark Model

At present the following are some of the criteria considered for the detection of early diabetes / prediabetes cases.

Fasting blood sugar test: A blood sample will be taken after an overnight fast.

i. Normal: A fasting blood sugar level less than 100 mg/dL (5.6 mmol/L)
ii. Prediabetes: A fasting blood sugar level from 100 to 125 mg/dL (5.6 to 6.9 mmol/L)
iii. Diabetes: A fasting blood sugar level > 126 mg/dL (7 mmol/L) or higher on two separate tests

Body mass index higher than 25 (23 for Asian-Americans)

Blood pressure : Systolic (Top Number) / Diastolic (Bottom number)

- **(90/60) or less:** May have low blood pressure.
- **(90/60 - 120/80):** The blood pressure reading is ideal and healthy
- **(120/80 -140/90):** A normal blood pressure reading but it is a little higher than it should be
- (>=**140/90 over several weeks):** May have high blood pressure (hypertension).
- **Systolic** >= **140** - High blood pressure, regardless of your bottom number.
- **Diastolic** >= **90** - High blood pressure, regardless your top number.
- **Systolic** <= **90** - Low blood pressure, regardless of your bottom number.
- **Systolic** <= **60** - Low blood pressure, regardless of your top number.

Age > **45** is advised to receive an initial blood sugar screening, and then, if the results are normal, to be screened every three years thereafter.

From the analysis of the survey, we will predict more accurate parameters for the persons for the specific geography zone based on age, sex, marital status, BP, Hemoglobin level. India has a diverse food habit based on different geographies. Hence one of the objectives of this model is to identify any specificity with the geographical region based on the socio-cultural reasons.

Evaluation Metrics

The evaluation metric for this problem is simply the Accuracy Score.

Project Design

The dataset will be used for the following analysis

1. Identify the correlation between the prediabetes and the following factors
    a. Location
    b. Age
    c. BMI index
    d. Sex
    e. Marital status
    f. Blood Pressure: Systolic
    g. Blood Pressure: Diastolic
    h. Hemoglobin level

The Data will be Split into a training set and validation set with an 80-20 split.

Model training and evaluation

I will start with the simple model architecture first before training and evaluating it. Then iterate
this process trying different architectures and hyper-parameters to reach an accuracy score.