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.

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

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 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.

Classification of the populace based on the risk that they will develop diabetes will help in providing the right Medicare in right time.

Since the data collected are categorical i.e the Age, BMI Index, Haemoglobin level, Blood pressure, iodine level, Marital status, these information need to be One hot encoded before applying the classification techniques. We also have to identify the specificity and sensitivity factors to measure the accuracy.

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.

Parameter	Number of entries
Age	1896385
Age_Code	1896470
BP_Diastolic	946667
BP_Diastolic_2reading	946555
BP_systolic	946680
BP_systolic_2_reading	946588
Bad	0
Client IP	0
Description	106
Diabetes_test	1032692
Good	0
Haemoglobin	1576027
Haemoglobin_level	1168346
Haemoglobin_test	1641477

Length_height_cm	1508807
Length_height_measured	1665385
Marital_status	415368
PSU_ID	1896470
Pulse_rate	946660
Pulse_rate_2_reading	946541
Sex	1896470
Sl_no	106
Status	0
Timestamp	0
Variable_name	106
Web Service	0
Weight_in_kg	1510418
Weight_measured	1666439
ahs_house_unit	1896470
ani_milk_month	70639
...	
day_or_mn_for_breast_feeding_cd	72718
day_or_month_for_breast_feeding	72311
district_code	1896470
duration_pregnanacy	51754
fasting_blood_glucose	1007905
fasting_blood_glucose_mg_dl	880918
first_breast_feeding	73366
gauna_perfor_not_perfor	279670
house_hold_no	1896470
identification_code	1896038
illness_duration	42971
illness_type	123139
is_cur_breast_feeding	73217
length_height_code	1517217
month_of_birth	1896470
record_code_iodine	1896470
record_code_iodine_reason	6424

rural_urban	1896470
semisolid_month_or_day	71718
sl_no	1896470
solid_month	71287
state_code	1896470
stratum	1896470
test_salt_iodine	1896470
treatment_type	51570
usual_residance	1896470
usual_residance_Reason	83514
vegetables_month_or_day	71051
water_month	71113
year_of_birth	1896470

After performing the data cleansing of the data that are not available and are important for this analysis the final dataset size is = 871582

The number of participants that are found to be normal (not diabetic) = 603545
The number of participants who were found to be pre diabetic or diabetic = 268037
The participants will be classified into three classes i.e. who are diabetic , Prediabetic and non-diabetic.

The present dataset is not balanced as the number of non-diabetic instances are three time that of the other class. Random selection to match the occurrence for both the classes.

Since after the balancing the entries for the non-diabetic and diabetic, the dataset size will be 536072 counts

The training set will be 85% of the dataset : 455660 counts

Validation set will be 5% of the dataset : 26804 counts

Testing set will be 10% of the dataset : 53607 counts

stratified k-fold cross-validation will be used for creation of the train, test and validation set.

Dataset after the one-Hot-Encoding implementation , data enrichment and cleansing the following dataset is obtained which will be used for analysis.

Parameter	Number of entries
BP_HIGH	871582
BP_LOW	871582
Diabetes	871582
Prediabetes	871582
Nodiabetes	871582
Iodine_GE_15	871582
Sex_Male	871582
Age_LT_20	871582
Age_20_25	871582
Age_25_30	871582
Age_30_35	871582
Age_35_40	871582
Age_40_45	871582
Age_45_50	871582
Age_50_55	871582
Age_55_60	871582
Age_60_65	871582
Age_65_70	871582
Age_GT_70	871582
BMI_LT_18	871582
BMI_18_20	871582
BMI_20_22	871582
BMI_22_24	871582
BMI_24_26	871582
BMI_26_28	871582
BMI_28_30	871582
BMI_30_32	871582
BMI_GT_32	871582
Hb_Normal	871582

Hb_Low	871582
Hb_High	871582
Married_And_Gauna	871582
Married_No_Gauna	871582
Never_married	871582
Widow	871582
Separated	871582
Divorced	871582
Remarried	871582
UTTAR_PRADESH	871582
MADHYA_PRADESH	871582
BIHAR	871582
ODISHA	871582
RAJASTHAN	871582
ASSAM	871582
CHHATTISGARH	871582
JHARKHAND	871582
UTTARAKHAND	871582
BP_Normal	871582

Solution Statement

The objective of this project is to model 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 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.
- **Diastolic <= 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.

Hamoglobin level

- Men : >= 13.5 g/dL
- Wemen: >= 12 g/dL

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.

Bench mark: Considering the health industries recommendation that a person having BMI > 25 and above and Age > 45 are prone to prediabetic or diabatic. In the current dataset the number of instances of people who are more than 45 years old and have BMI > 25 are 34746. There are 16176 people who are nondiabetic and 18570 are diabetic / prediabetic. The prediction accuracy of the criteria that people with BMI>25 and age more than equals to 45: 53.44 %

Evaluation Metrics

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

The accuracy of the prediction should be more than this bench mark value of accuracy 53.44 %. While predicting we must take care to minimise the type=2 error. The aim is to minimise the total error is calculated based on both precision and recall (precision i.e. What proportion of positive identifications was actually correct and the recall i.e. What proportion of actual positives was identified correctly ?)

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.

The Classification techniques:

K-Nearest Neighbors (Option 1)

We will use the K-Nearest Neighbors model. In order to select the best value for K, we'll use 10-fold Cross-Validation combined with Grid Search where K=(1, 2, … 30). In pseudo code:

1. Partition the training data into 10 stratified folds. Call these test folds.
2. For K = 1, 2, … 10
    1. For each test fold
        1. Combine the other four folds to be used as a training fold
        2. Fit a K-Nearest Neighbors model on the training fold (using the current value of K)
        3. Make predictions on the test fold and measure the resulting accuracy rate of the predictions
    2. Calculate the average accuracy rate from the five test fold predictions
3. Keep the K value with the best average CV accuracy rate

XGBoost (Option 2):

We will use The XGBoost (Extreme Gradient Boosting) which is a Gradient boosting technique for the prediction. XGBoost is a type of decision tree ensembles. The tree ensemble model consists of a set of classification and regression trees (CART).

We will use the training data (with multiple features) xi to predict a target variable yi.

In order to train the model, we will to define the objective function to measure how well the model fit the training data.

The objective function : obj(θ)=$L(\theta)$+$\Omega(\theta)$

where L is the training loss function, and Ω is the regularization term. The training loss measures how predictive our model is with respect to the training data. A common choice of L is the mean squared error, which is given by

$$L(\theta) = \sum_i (y_i - \hat{y}_i)^2$$

The regularization term controls the complexity of the model, which helps us to avoid overfitting or known as bias-variance trade-off .

The model :

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F}$$

where K is the number of trees, f is a function in the functional space F, and F is the set of all possible CARTs. The objective function to be optimized is given by

$$\text{obj}(\theta) = \sum_i^n l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

More on the implementation: https://xgboost.readthedocs.io/en/latest/tutorials/model.html