Udacity Nanodegree Capstone Project Proposal

2019

Predict the onset of diabetes based on diagnostic measures

Debu Sinha
debusinha2009@gmail.com

Table of Content

Domain Background

Joshi and Parikh have described India as the “capital of the world with 41 million Indians having diabetes, every fifth diabetic in the world is an Indian”.

In 2000, India (31.7 million) topped the world with the highest number of people with diabetes mellitus followed by China (20.8 million) with the United States (17.7 million) in second and third places, respectively. According to Wild et al. .3, the prevalence of diabetes is predicted to double globally from 171 million in 2000 to 366 million in 2030, with a maximum increase in India. It is predicted that by 2030 diabetes mellitus may afflict up to 79.4 million individuals in India, while China (42.3 million) and the United States (30.3 million) will also see significant increases in those affected by the disease.3,4 India currently faces an uncertain future about the potential burden that diabetes may impose upon the country. Many influences affect the prevalence of disease throughout the country, and the identification of those factors is necessary to facilitate change when facing health challenges.

Although the Indian urban population has access to reliable screening methods and anti-diabetic-medications, such health benefits are not often available to rural patients. There is a disproportionate allocation of health resources between urban and rural areas, and also, poverty in rural areas may be multifaceted. Aged care facilities in rural areas report disparity in the diabetes management compared with their urban counterparts,11 with these populations more likely to suffer from diabetic complications compared to their urban counterparts. More needs to be done to address the rural-urban inequality in diabetes intervention entry.

Problem Statement

The goal of this project is to build a reliable machine learning model to diagnose if a person has diabetes or not based on clinical data input provided with high precision.

This is a classification problem to determine if a patient has diabetes or not.

Datasets and Inputs

The clinical data that will be used in this project has been provided by Kaggle https://www.kaggle.com/uciml/pima-indians-diabetes-database.

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

The datasets consists of several medical predictor variables and one target variable, `Outcome`. Predictor variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

The dataset has 768 entries. The outcome class has two categories 0 and 1. The data is unbalanced as the number of 1's(tested positive) are 268 and the 0's(tested negative) are 500. Since the ratio of **number of outcome with value 0 / number of outcome with value 1** is 1.82 we can move forward with our model training without performing upsampling or downsampling.

The models we are proposing to work with are logistic regression and Xgboost both of which handle unbalanced data well.

Upsampling can suffer from overfitting and downsampling can suffer from loss of important information in the data points that we don't include.

Solution Statement and Benchmark model

For this project we can take a simple logistic regression model as baseline for predicting whether the patient has diabetes or not. The next stage will be to utilize Xgboost algorithm to get a better precision than logistic regression.

Evaluation Metrics

As the use case we are working on is dealing with a classification problem, we will evaluate the performance of the model based on **accuracy_score** that we can get from the sklearn.metrics Apart from the accuracy score it will also be important to have a **high recall and high precision** in order to correctly diagnose the diabetes condition. We can use **f1 score** as a metric to evaluate our model. F1 score will raise a flag if either of the precision and recall value is low.

Project Design

After importing the dataset for the first time into pandas data frame, we can check the shape of the data to see the number of rows and columns in the dataset. We can also do head to inspect the content of each of the columns.

Next, we move on to the feature engineering part. We can find if there are any null values in the dataset or not. After dealing with the null values, we can draw a correlation matrix using a seaborn heatmap and visualize.

If there are any categorical values, we need to convert them to numeric and run one hot encoding if more than two categories exist. In case any of the independent feature values in a data record is 0, we will make use of imputer to fill these values by the mean value for that feature.

Next, we can check if the dataset is balanced or not. Since we already figured out that our dataset is slightly unbalanced, we will not perform any upward or downward sampling of data.

Once all our features have been processed, we will take all the independent features from the dataset and put into X and take the dependent feature(Outcome) and put it in y.

To train the machine learning model and avoid overfitting, we will make use of a 70/30 train test split.

Once our data is ready, we will train the logistic regression model on the training dataset and later use the trained model to predict test data whether a patient has diabetes or not and use f1 score and accuracy as metric.

Next, we will use a grid search by RandomSearchCV for Xgboost to search for the best estimator and hyperparameters using the training dataset and initiate a classifier to predict the label of target feature in the test set and validate new accuracy and f1 score.

References

1. Joshi SR, Parikh RM. India - diabetes capital of the world: now heading towards hypertension. J Assoc Physicians India. 2007;55:323–4. [PubMed] [Google Scholar]

2. Kumar A, Goel MK, Jain RB, Khanna P, Chaudhary V. India towards diabetes control: Key issues. Australas Med J. 2013;6(10):524–31. [PMC free article] [PubMed] [Google Scholar]

3. Wild S, Roglic G, Green A, Sicree R, King H. Global prevalence of diabetes-estimates for the year 2000 and projections for 2030. Diabetes Care. 2004;27(3):1047–53.[PubMed] [Google Scholar]

4. Whiting Dr, Guariguata L, Weil C, Shawj. IDF Diabetes atlas: Global estimates of the prevalence of diabetes for 2011 and 2030. Diabetes Res Clin Pract. 2011;94:311–21.[PubMed] [Google Scholar]

5. Anjana RM, Ali MK, Pradeepa R, Deepa M, Datta M, Unnikrishnan R, Rema M, Mohan V. The need for obtaining accurate nationwide estimates of diabetes prevalence in India - rationale for a national study on diabetes. Indian J Med Res. 2011;133:369–80.[PMC free article] [PubMed] [Google Scholar]

6. Zargar AH, Khan AK, Masoodi SR, Laway BA, Wani AI, Bashir MI, Dar FA. Prevalence of type 2 diabetes mellitus and impaired glucose tolerance in the Kashmir Valley of the Indian subcontinent. Diabetes Res Clin Pract. 2000;47(2):135–46. [PubMed] [Google Scholar]

7. Ramachandran A, Snehalatha C, Kapur A, Vijay V, Mohan V, Das AK, Rao PV, Yajnik CS, Prasanna Kumar KM, Nair JD. Diabetes Epidemiology Study Group in India (DESI). High prevalence of diabetes and impaired glucose tolerance in India: National Urban Diabetes Survey. Diabetologia. 2001;44(9):1094–101. [PubMed] [Google Scholar]

8. Arora V, Malik JS, Khanna P, Goyal N, Kumar N, Singh M. Prevalence of diabetes in urban Haryana. Australas Med J. 2010;3(8):488–94. [Google Scholar]

9. Bramley D, Hebert P, Jackson R, Chassin M. Indigenous disparities in disease-specific mortality, a cross-country comparison: New Zealand, Australia, Canada, and theUnited States. N Z Med J. 2004;117(1207):U1215. [PubMed] [Google Scholar]

10. Sukala WR, Page RA, Rowlands DS, Lys I, Krebs JD, Leikis MJ, Cheema BS. Exercise intervention in New Zealand Polynesian peoples with type 2 diabetes: Cultural considerations and clinical trial recommendations. Australas Med J. 2012;5(8):429–35. [PMC free article][PubMed] [Google Scholar]

11. Khalil H, George J. Diabetes management in Australian rural aged care facilities: A cross-sectional audit. Australas Med J. 2012;5(11):575–80. [PMC free article] [PubMed] [Google Scholar]