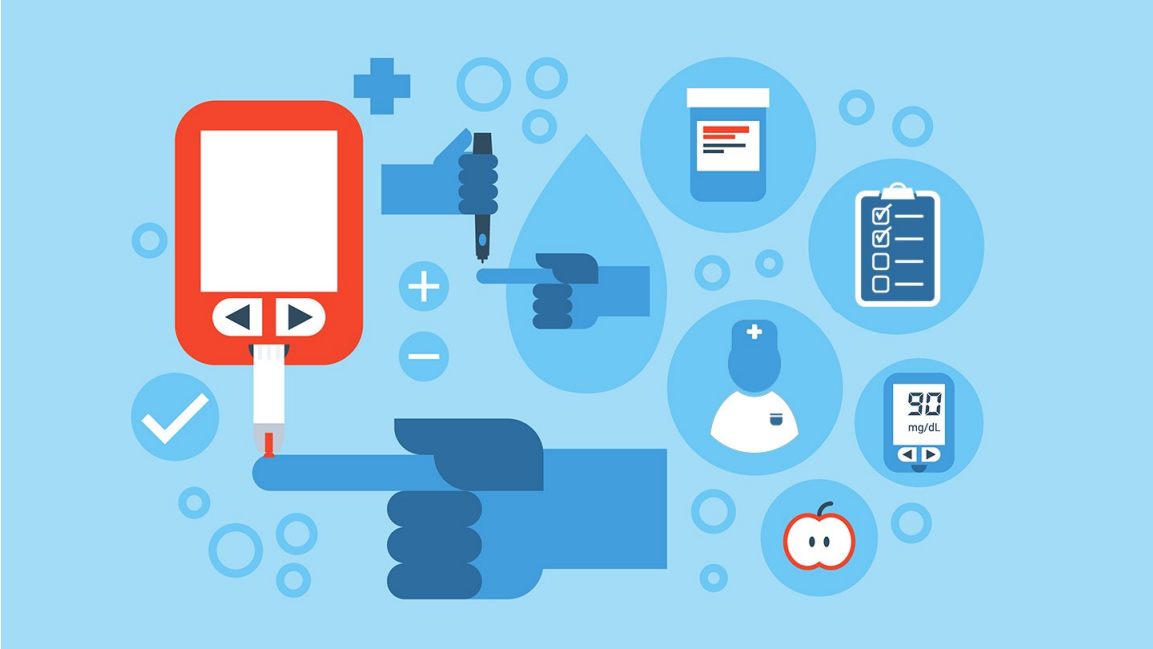
Differentiation and Classification of Diabetes, Pre-diabetes and Non-diabetes



Aditya Adhikary 2015007 | Anjali Dhall PHD17207 | Ramya Y. S. 2015115 | Surabhi S Nath 2016271 | Vaibhav Mittal MT18242

Contents

1 Introduction

Diabetes is a group of disorders characterized by high blood sugar levels. During diabetes, the body is unable to produce and use insulin efficiently. Insulin is the hormone produced by pancreas responsible for regulating blood sugar levels in our body. Diabetes can occur in different stages - Type 1 or Type 2 diabetes, with Type 1 diabetes being more severe where the pancreas produce little or no insulin. Prediabetes is the condition where blood sugar level is significantly high which may turn into a case of Type 2 diabetes if not suitably attended to.

Most developed countries and many rapidly developing countries are facing an alarming rise in the number of diabetes cases observed annually. Today, nearly 425 million adults are living with diabetes. Diabetes is known to cause a couple millions of deaths each year. India has nearly 50 million people suffering from the disease and nearly 87 million are considered prediabetic. Studies claim that by the end of 2030, India will have close to 98 million people suffering from Type 2 diabetes. India has more people with Type-2 diabetes than any other nation. Diabetes is considered India's fastest growing disease and Times of India coined India as the "Diabetes Capital in the World."

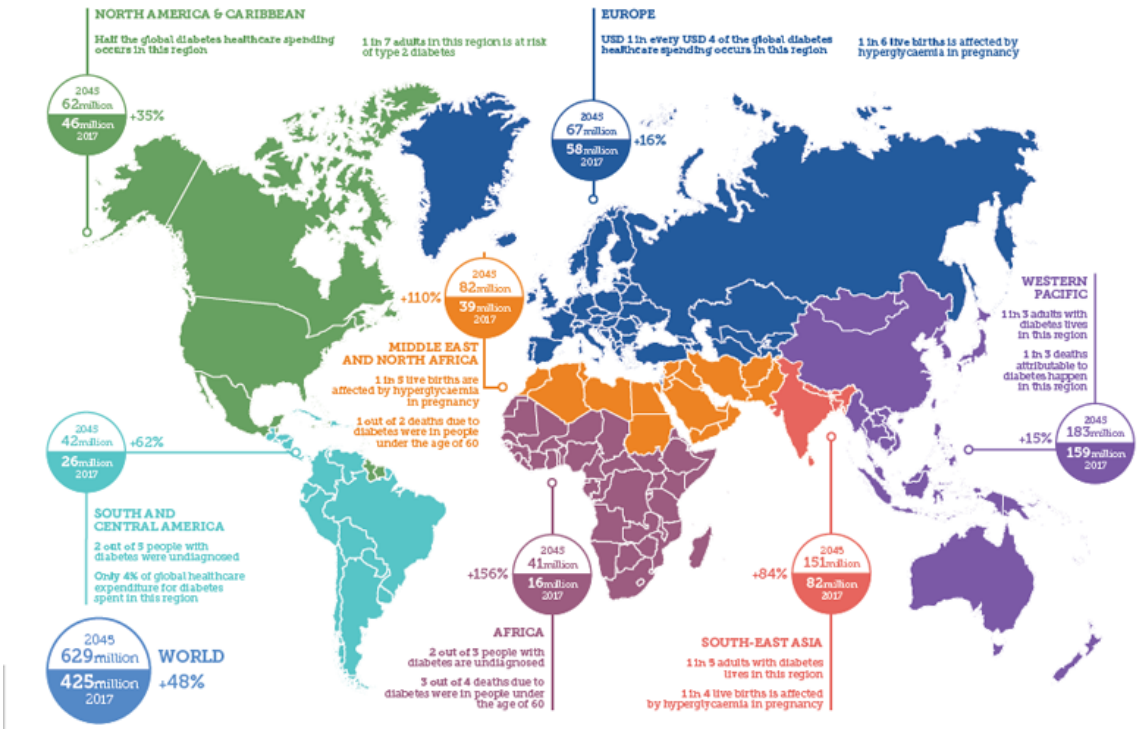


Figure 1: Diabetes in the World

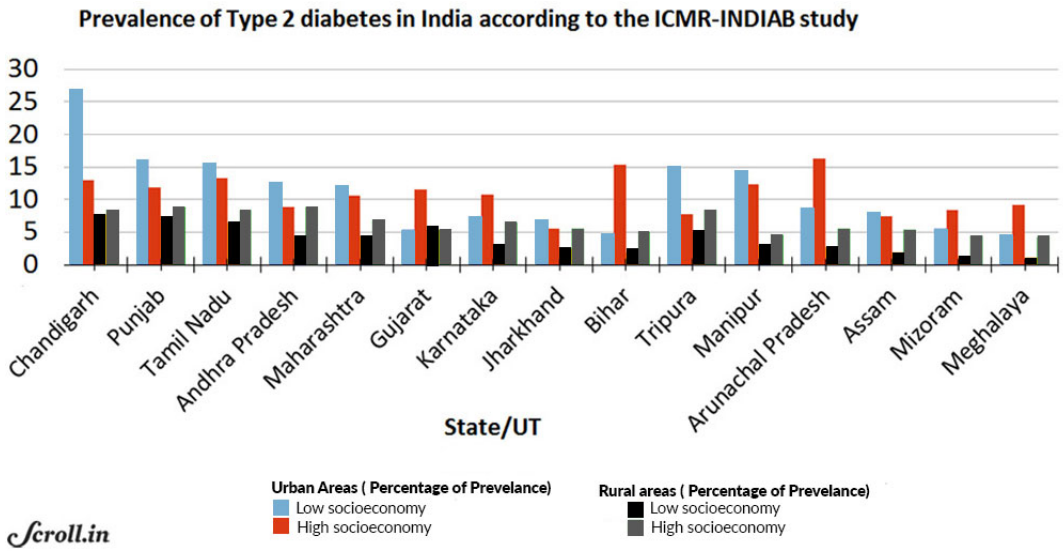


Figure 2: Diabetes in India

Diabetes is a disorder which also needs environmental factors to manifest in addition to the heritable factors. These environmental factors can be diet, exercise, and proper rest, among others. Timely detection and proper management can significantly help diabetes patients. Diabetes must be controlled and cured since it often proliferates to cause Diabetic Retinopathy, heart diseases and several other disorders. Diabetes classification algorithms can aid diagnosis of the disease.

In this project, we are utilising the powers of big data mining techniques and machine learning for classifying individuals on the basis of their diabetic disease status. We have reimplemented the paper titled "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes" by Wei Yu et. al and have extended it further by employing our own methods to overcome the shortcomings their work presented.

2 Paper Summary

Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes

Wei Yu, Tiebin Liu, Rodolfo Valdez, Marta Gwinn, Muin J Khoury

2.1 Overview

The paper explores the use of Support Vector Machine in classifying people with diabetes and pre-diabetes using data collected from National Health and Nutrition Examination Survey (NHANES), an ongoing, cross-sectional, probability sample survey of the U.S. population. The authors have formulated two Classification Schemes where they have implemented binary classifiers to classify diabetes, pre-diabetes and non diabetes cases. They have achieved an AUC of 0.83 and 0.73 for the two schemes respectively.

2.2 Background - Support Vector Machine

Support vector machine are models for supervised learning. They work on the principle of maximizing margin between samples of the various classes. The authors have employed SVM as opposed to other techniques since the SVM approach is data-driven, model-free and has discriminative power for classification, especially in cases where sample sizes are small and a large number of variables are involved. The SVM algorithm performs a classification by constructing a multidimensional hyperplane that optimally discriminates between two classes by maximizing the margin between two data clusters. This algorithm achieves high discriminative power by using special nonlinear functions called kernels to transform the input space into a multidimensional space. The basic idea is that even if data seems inseparable in a lower dimension, it may still be separable using a hyperplane in a higher dimension. SVM constructs an n-1 dimensional separating hyperplane to discriminate two classes in an n-dimensional space. SVM involves several hyperparameters like C and γ which need to tuned accordingly based on the data, the application and multiple experiments.

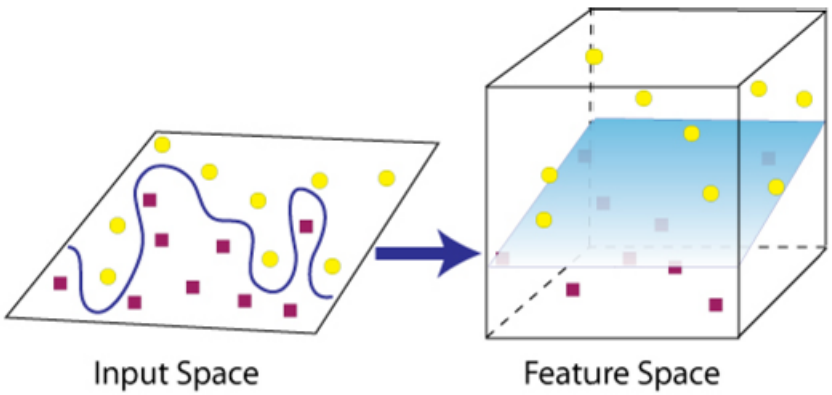


Figure 3: Feature Transformation in SVM

2.3 Methods

2.3.1 Data Collection

The authors have collected 5 year data (1999-2004) from NHANES interview which includes demographic, socioeconomic, dietary, and health-related questions. The examination component consists of medical, dental, and physiological measurements, as well as laboratory tests administered by highly trained medical personnel. The study was limited to participants of over 20 years of age and non pregnant women. Each participant was asked the question - "Have you been told by a doctor that you have diabetes?" If the participant answered "Yes", it was considered a case of Diagnosed Diabetes. If the participant answered "No", decision was made based on the glucose level - if level was higher than 126 mg/dl, it was considered a case of Undiagnosed Diabetes, if it lay between 100 mg/dl and 126 mg/dl, it was considered to be a case of pre-diabetes and a case of non diabetes if glucose level were less than 100 mg/dl.

2.3.2 Feature Selection

14 simple variables commonly associated with the risk for diabetes were chosen which included family history, age, gender, race and ethnicity, weight, height, waist circumference, BMI, hypertension, physical activity, smoking, alcohol use, education, and household income.

2.3.3 Model Generation

Two classification schemes were formulated. Classification Scheme I fuses people with pre-diabetes and non diabetes into one class and discriminates them from the people with either diagnosed or undiagnosed diabetes. In Classification Scheme II fuses the gathering of people with either undiscovered diabetes or pre-diabetes and discriminates them against the non diabetes cases. These two schemes are pictorially depicted in Figure 4.

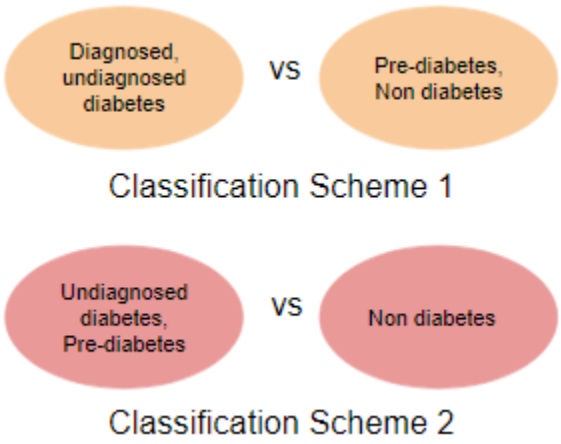


Figure 4: Classification Schemes used in the Paper

For pre-processing, the features were normalized to values from -1 to +1 and values of categorical variables such as Race are arbitrarily assigned to numbers between -1 and +1. Following, they implemented SVM using LibSVM, along with fine-tuning the kernel, gamma and C parameters. They tried various kernel like linear, polynomial, sigmoid and rbf kernel functions and executed 10 fold cross validation. For comparison purposes, they also implemented multiple logistic regression modeling (MLR) using the same selected risk variables or features and case status as the output variable. The SVM approach tends to classify entities without providing estimates of the probabilities of class membership in the dataset, which is a fundamental difference from multiple logistic regression. For evaluating the performance of their models, for 10 fold cross validation and have reported Sensitivity, Specificity, PPV, NPV and AUC scores.

2.3.4 Results

The authors have obtained the following results from their experiments -

- Classification Scheme 1 - RBF kernel, AUC - SVM 83.47%, logistic regression 83.19%
- Classification Scheme 2 - Linear kernel, AUC - SVM 73.18%, logistic regression 73.35%

Model	Data set	Sensitivity	Specificity	PPV	NPV	AUC
Classification Scheme I*	Test	0.7715	0.7503	0.4926	0.9127	0.8347
	Training	0.7938	0.7169	0.4550	0.9211	0.8383
	10-fold cross- validation	0.7765	0.7027	0.4388	0.9130	0.8242
Classification Scheme II*	Test	0.7359	0.6254	0.5061	0.8195	0.7318
	Training	0.7092	0.6590	0.6729	0.8087	0.7393
	10-fold cross- validation	0.7059	0.6589	0.5293	0.8054	0.7357

Figure 5: Reported Results

3 Our Analysis

In order to replicate the paper, we mined data from NHANES for the same years as used in the paper - 1999 to 2004 and extracted the same 14 mentioned features. We converted questionnaire data of into label form - diabetes vs no diabetes vs borderline. We generated a CSV file with 15 features for 6993 patients. As our preliminary task, we implemented SVM for tri-class classification and evaluated results using precision, recall, F1 score, confusion matrix and ROC curves. We faced the following challenges and proposed corresponding solutions to overcome them:

3.1 Challenges

1. Less data
2. Inconvenient file naming - names of csv files containing the features have been indiscriminately changed over the years
3. Complicated structure of data - demographics, dietary, examination, laboratory and questionnaire
4. Highly imbalanced data, with 89% samples of class 2 (no diabetes)
5. Missing values, incomplete information, NAs

3.2 Proposed Solutions

1. Increase data - 1999 to 2017
2. Tackle class imbalance issue
3. Fill NAs with suitable values
4. Use Decision Trees to select the best features
5. Attempt DL based approach

For our final task, we incorporated all the proposed solutions and developed several models using a variety of classifiers. To tackle the less data issue, we expanded the dataset to include data from the years 2004-2015. As a result, we were able to increase to dataset to nearly 4 times and make it around 23,000 samples. Using this data, we modelled the same two Classification Schemes described in the paper and implemented models for both schemes. To deal with the class imbalance issue, we generated synthetic data. We performed oversampling to equalize samples in each class by randomly generated data points near the actual data points in the Euclidean plane. We handled the missing NA positions using two workarounds - replace those positions with -1 or ignore those samples entirely. Ignoring the samples resulted in removal or nearly 6000 samples.

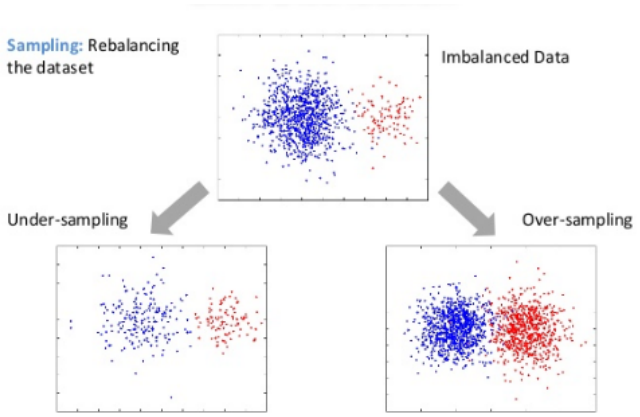


Figure 6: Synthetic Data Generation using Oversampling

For Classification Scheme 1, we employed SVM, Decision tree and Random forest classifiers and for Classification Scheme 2 we used K Nearest Neighbours, MLP, Decision Tree and Random Forest classifiers. The models were created utilizing 80% of the data under train set and 20% under test set. The results obtained are shown below. Due to inconvenient file naming used in the dataset, we were unable to extract more features because of which we were unable to apply DL based strategies since features were very few.

Classification Scheme 1						
SVM						
	Precision	Recall	f1-Score	Acc	Support	MCC
Validation	100.00	99.00	99.00	99.36	7714	98.73
DT						
Validation	100.00	100.00	100.00	100.00	7714	100.00
RF						
Validation	100.00	100.00	100.00	100.00	7714	100.00
			MLP			
Validation	93.00	98.00	0.73	62.89	7714	35.94

Figure 7: Obtained Results - Classification Scheme 1

Classification Scheme 2						
KNN						
	Sens	Spec	FPR	Acc	AUC	MCC
Training	67.34	48.61	51.39	60.05	0.62	0.16
Validation	67.46	47.73	52.27	59.78	0.61	0.15
MLP						
Training	87.76	86.51	13.49	87.27	0.95	0.74
Validation	88.76	85.76	14.24	87.59	0.95	0.74
DT						
Training	97.45	96.74	3.26	97.17	0.99	0.94
Validation	96.7	94.82	5.18	95.97	0.99	0.92
RF						
Training	100	99.17	0.83	99.68	1	0.99
Validation	100	99.27	0.73	99.72	1	0.99

Figure 8: Obtained Results - Classification Scheme 2

4 Docker Instructions

All code files, CSV files along with a README file can be found in the docker container.

To pull the project use the following command:

docker pull vmittal3/diabetes

To run the container and generate an image:

docker run -it vmittal3/diabetes

Then follow the README file to run the codes and generate results:

cd /home/project; vi README.txt