International Journal of Scientific Engineering and Science
Volume 8, Issue 4, pp. 67-71, 2024. ISSN (Online): 2456-7361

The Applications for Diabetes Prediction by Machine Learning Algorithms

Yanting CAO[1], Kazumitsu NAWATA[2]
[1]Department of Engineering, University of Tokyo, Japan
[2]Department of Engineering, University of Tokyo, Japan

Abstract—*Since diabetes is a very widespread disease causing serious symptoms, accurate diagnosis becomes more and more important as the very first step of effective treatments dealing with diabetes. To relief doctors from their burden of diagnosis workloads, which require them to make medical estimates based on experience, computers and algorithms contribute a lot in the field of medical diagnosis with the development of new technology nowadays. However, due to the diversity and complexity of real-world data, usual statistical methods are often unable to handle or produce precise results. In this paper, we remove most of the outliers of Pima Indians Diabetes dataset for the subsequent classification through complex data pre-processing and data cleaning. The final experimental results show that SVM and Random Forest algorithm perform best, but XGboost algorithm also performs well.*

Keywords— *Diabetes, Machine learning, Feature selection, Classification, SVM, Random Forest.*

I. Introduction

Diabetes mellitus, clinically referred to as diabetes, is a grievous, chronic disease that impacts how an individual converts food into energy. Since food intake is broken down into glucose mostly then discharged into circulation, when a diabetic could not respond to insulin, the blood sugar would rise above an acceptable level and may get out of control. With long-standing unchecked or improper management of diabetes, many complications such as visual loss, nephropathy, retinopathy, cardiovascular diseases and neuropathy that may result in death. [1]

The diagnosis of diabetes is a very typical classification problem. So to make an accurate discrimination, a careful cleaning for raw data is indispensable. How to deal with the missing data and remove the interfering attributes is the focus of this research. In this study, based on the precise analysis of the original dataset's features, we used quite diverse machine learning technique to obtained a good experimental results: SVM and Random Forest algorithms perform well.

II. Proposed Framework and Methodology

The proposed framework is based on data cleaning, feature selections and XGboost/SVM/Ramdomforset classifiers. The whole process is shown in the figure. [Fig. 1].

A. GBDT

Gradient Boosting Decision Tree (GBDT), also called Multiple Additive Regression Tree (MART), is an iterative decision-tree algorithm, which is composed of multiple decision trees. The final solution is the accumulation of all the trees' results. GBDT is considered as a strong generalization algorithm together with SVM when it is initially proposed. The trees in GBDT are regression trees, not classification trees, and it is used for regression prediction as well as can also be used for classification after adjustment. The idea of GBDT algorithm gives it the natural advantage of being able to find a variety of distinctive features and feature combinations.

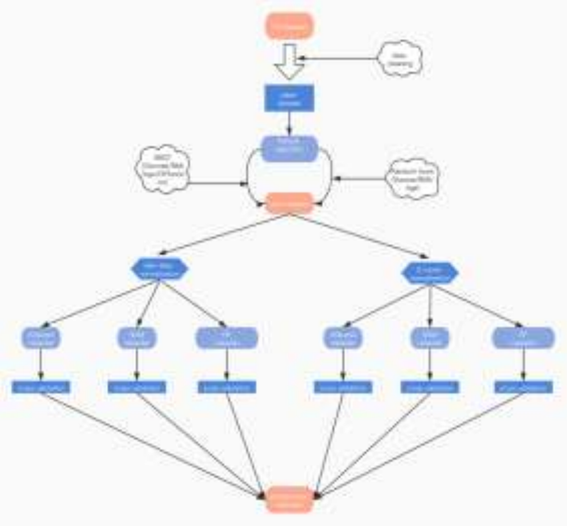

Fig. 1. The task flow chart

According to [2], the mathematical expression of GBDT algorithm shows as follow：

1. Initialize $f_0(x) = argmin_r \sum_{i=1}^{N} L(y_i,\ \gamma)$.
2. For m=1 to M:
   (a) For i=1, 2,3, …, N compute
   $$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}.$$
   (b) Fit a regression tree to the targets r_{im} giving terminal regions R_{jm}, j=1, 2, 3…, J_m.
   (c) For j=1, 2, 3…, J_m compute
   $$\gamma_{jm} = arg \min_{\gamma} \sum_{x \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma).$$

(d) Update

$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$

3. Output $\hat{f}(x) = f_M(x)$.

B. Random forest

Random forest (RF) is a supervised learning algorithm, which is an ensemble learning algorithm based on decision tree. RF is very easy to implement, and with small computational cost. But it performs very well on classification and regression problems, so RF is hailed as " representing the level of ensemble learning technology". Here is the process of RF algorithm [3]:

```
Input: Data set
    $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$;
    Feature subset size K.
Process:
1. $\mathcal{N} \leftarrow$ create a tree node based on $\mathcal{D}$;
2. if all instances in the same class, then return $\mathcal{N}$;
3. $\mathcal{F} \leftarrow$ the set of features that can be split further.
4. if $\mathcal{F}$ is empty then return $\mathcal{N}$
5. $\hat{\mathcal{F}} \leftarrow$ select $\mathcal{K}$ features from $\mathcal{F}$ randomly.
6. $\mathcal{N}.f \leftarrow$ the feature which has the best split point in $\hat{\mathcal{F}}$;
7. $\mathcal{N}.p \leftarrow$ the best split point on $\mathcal{N}.f$;
8. $\mathcal{D}_l \leftarrow$ subset of with $\mathcal{D}$ values on $\mathcal{N}.f$ smaller than $\mathcal{N}.p$;
9. $\mathcal{D}_r \leftarrow$ subset of with $\mathcal{D}$ values on $\mathcal{N}.f$ no smaller than $\mathcal{N}.p$;
10. $\mathcal{N}_l \leftarrow$ call the process with parameters $(\mathcal{D}_l, \mathcal{K})$;
11. $\mathcal{N}_r \leftarrow$ call the process with parameters $(\mathcal{D}_r, \mathcal{K})$;
12. return $\mathcal{N}$
Output: A random decision tree
```

C. Normalization

Data normalization is a very basic work for data mining. Different evaluation index tends to be with different dimensions and dimensional units, so it will affect the result of data analysis. In order to eliminate the dimension influence between indicators, we always need data standardization to solve the comparability among the data indices. After the standardization of original data, each index stays the same order of magnitude, which is better for comprehensive comparative evaluation.

There are some common methods of data normalization, such as : rescaling, min-max normalization, log function conversion, mean normalization, atan function conversion, Z-score normalization (the most common one), and fuzzy quantization. In this paper, we choose two of them to normalize the cleaned PID dataset.

Min-max normalization is also known as deviation normalization, which is a linear transformation of the raw data so that the resulting value can be mapped to interval [0,1]. The conversion function is:

$$x_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

where x is the test value in a set X; x_{max} is the maximum value in X; and x_{min} is the minimum value in X.

Z-score normalization is also known as standard deviation normalization because the processed data conform to a standard normal distribution, that is, with a mean of 0 and a standard deviation of 1. The conversion function is:

$$x_{normalized} = \frac{(x - \mu)}{\sigma}$$

where x is the test value in a set X; μ is mean, and σ is standard deviation (SD).

D. XGboost classifier

As an efficient implementation of GBDT, XGBoost is with a particularly high ceiling so that it is a favorite in the algorithm race. The basic idea of XGBoost is the same as GBDT, but some optimizations of algorithm itself, efficiency and robustness had been made. For example , second derivatives make the loss function more precise; regularization term avoids tree overfitting; block storage allows parallel computations, etc.

The loss function of XGboost is expressed as :

$$\mathcal{L}^{(t)} = \sum_{j=1}^{T} \left[G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2 \right] + \gamma T$$

So the loss function of each leaf node is:

$$f(w_j) = G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2$$

which is also a quadratic function of one variable w_j.

Because $(H_j + \lambda) > 0$, $f(w_j)$ minimizes at $w_j = -\frac{G_j}{H_j + \lambda}$, the minimum value is $-\frac{1}{2}\frac{{G_j}^2}{H_j + \lambda}$. [4]

E. SVM classifier

Support vector machine (SVM) is a popular algorithm for regression and classification which fits right in with this study. This learning method contains several models construction, from simple to complex: linear support vector machine in a linearly separable case, linear support vector machine and non-linear support vector machine. [5]

The original optimization problem of SVM is:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i$$

$$\text{s.t.}\quad y_i(w * x_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0,\ \ \mathrm{i} = 1{,}2, \dots, \mathrm{N}$$

is equal to optimize

$$\min_{w,b}\quad \textstyle\sum_{i=1}^{N}[1 - y_i(w * x_i + b)]_+ + \lambda\|w\|^2$$

F. Performance evaluation

There are many ways to evaluate the performance of classifiers. In this research, we use confusion matrix (TABLE I), ROC, AUC to evaluate how well the XGboost, SVM and RF work. [3]

TABLE I. Confusion matrix

	Positive	Negative
Positive	Ture Positive	False Negative
Negative	False Negative	True Negative

According to confusion matrix, it deduces as follow:

$$\text{Accuracy} = \frac{TP + NP}{TP + TN + FP + FN}$$

$$\text{TPR} = \frac{TP}{TP + FN} \qquad FPR = \frac{FP}{FP + TN}$$

Then we can get the receiver operating characteristic curve (ROC) and area under the curve (AUC).

III. Experimental results

In our experiment, we use the Pima Indian Diabetes (PID dataset) [6]. This dataset is initially offered by the National Institute of Diabetes and Digestive and Kidney Diseases.

All the subjects come from Pima Indian female group who are aged 21 and above. It has 768 instances of patient data with 8 attributes and one output giving the outcome of diabetic status, 0 or 1, of the patients, 268 of whom had been diagnosed with diabetes.

The eight attributes are: Pregnancies, BMI, Skin Thickness, Diabetes Pedigree Function, Blood Pressure, Insulin, Glucose and Age.

The label column, which means to rank participants numerically as 1,2,3……, are insignificance. The Outcome represents the participant with diabetes or not. (TABLE II). All the datasets are implemented by software Python.

TABLE II. Attributes of PID dataset

Data type	Attributes Description	Range
Label	Participant serial number: 1, 2, 3……768	1-768
Boolean	Outcome: 0 represents not diabetic and 1 represents diabetic	1 or 0
Integer	Pregnancies: number of times pregnant	0-17
	Glucose: plasma glucose concentration a 2 hours in an oral glucose tolerance test	0-199
	BloodPressure: diastolic blood pressure (mm Hg)	0-122
	SkinThickness: triceps skin fold thickness (mm)	0-99
	Insulin: 2-Hour serum insulin (mu U/ml)	0-846
	Age: years	21-81
Float	BMI: body mass index=weight (kg) / [height (m)]^2	0-67.1
	DiabetesPedigree Function: scores likelihood of diabetes based on family history	0.078-2.42

A. Preprocessing

To avoid the potential risk that the training process would possibly be negatively influenced by the flaws raw dataset such as missing values and very different range of features influencing predictions, here we apply several preprocessing methods on the origin dataset.

Missing values are common occurrences in data. However, this issue must be addressed prior to modeling because most predictive modeling techniques cannot handle any missing values. [7]

Missing values appears very frequently in PID dataset, some of which even influenced hundreds of data instances. We found 7 instances with too many missing attributes, so we firstly deleted these 7 instances having 4 missing attributes. In addition, we found that 4 instances have abnormal value of BMI that are zeros, and 5 instances with 0 Glucose respectively. We filled them out with average value of BMI is 32.5 and Glucose is 122. Next, another 28 instances were found missing Blood pressure attribute, which would lead to significant errors in model training and final predictions, so we manually removed these instances. Overall, 733 of 761 instances of data were used in training and testing.

What needs to be pointed out is that for the rest 2 attributes, 194 instances have 0 records on Skin Thickness and 339 zeros for Insulin records. The number of missing values has reached a level that any values to be filled in the missing ones will significantly influence the total distribution of these two attributes and severely migrate the validity of our trained model. Here we abandoned the handling of these 2 attributes and left to the feature selection algorithms, as a result of which, these 2 attributes are also not selected by the algorithms.

B. Normalization

After missing values handling, a new dataset comes out and we rename it Newdata in Python software.

In total 8 attributes, BMI and Diabetes Pedigress Function are floats while Pregnancy, Blood pressure, Skin Thickness, Insulin, Age and Glucose are integers. To avoid possible difficulties in learning process of the model which different scales of attributes with same learning rate will cause difficulties in optimization, here we apply the standardization to all data by min-max method and z-score respectively.

C. Feature selection

A few predictive models, especially tree-based techniques, can specifically account for missing data [8], To analyze the correlations between diagnose attributes and diabetic status of patients and also enhance the performance of our trained machine learning model, we apply feature selections methods on the dataset to analyze the importance of the attributes and also as a prepossessing procedure before we start training. Here we use the random forest algorithm and take the inner coefficient of trained model as the importance weight of each attribute. The result is shown in the table:

TABLE III. Output of random forest algorithm

Pregnancy	0.03704324	False
BMI	0.19482633	Ture
Diabetes Pedigrees Function	0.05895204	False
Blood pressure	0.01076994	False
Skin Thickness	0.00523024	False
Insulin	0.02407381	False
Age	0.143935	Ture
Glucose	0.52516939	Ture

Another algorithm we use for feature selection is GBDT, the result is shown in the next table:

TABLE IV. Output of GBDT algorithm

Pregnancy	0.05372133	False
BMI	0.18598156	Ture
Diabetes Pedigrees Function	0.12896012	Ture
Blood pressure	0.03426388	False
Skin Thickness	0.01332675	False
Insulin	0.04208914	False
Age	0.14112957	Ture
Glucose	0.40052764	Ture

Finally, we took the intersection of two feature selection algorithm and selected: Diabetes Pedigree Function, Age, BMI and Glucose.

D. Classification and evaluation

With the standardization of min-max and z-score, we use the XGboost, SVM and Random Forest as the final machine learning models, trained by the data and to be used for predictions of the diabetic status of the patients with given diagnostic attributes.

The result of average accuracy of 10-fold cross validation is shown in the table:

TABLE V. Result of average accuracy of 10-fold cross validation

	mix-max	**z-score**
XGboost	0.75405405	0.74324324
SVM	0.76621622	0.77702703
RF	0.78783784	0.76756757

Their ROC and AUC of each 10-fold cross validation are drawn by the Python in following Fig 2 to 7.

How is the Roc curve drawn? Through the previous knowledge about classification algorithms, we learned that dropping a sample into a classifier can produce a prediction of probability between 0 and 1. Then given a threshold for the classification, a value less than this prediction is classified as a positive class, otherwise it is an inverse class. So we rank the predictions of the classifier from largest to smallest, and then set these predictions as the threshold in order to classify these samples positively and negatively. Each classification can get a set of TPR and FPR values until the threshold takes all the predicted values of the samples, and then we mark all the points on the coordinates, finally we will get the roc curve by connecting these points into a line.

As Fawcett [9] pointed out, the AUC value is equivalent to the probability that a randomly chosen positive example is ranked higher than a randomly chosen negative example. Consequently, AUC is actually a probability value. The AUC value is chosen as the evaluation criterion since oftentimes the ROC curve cannot clearly indicate which classifier is more effective. Given that AUC is defined as the area under the ROC curve, it is obvious that the value of this area is not greater than 1. Since the ROC curve is generally above the line y=x, the value of AUC ranges between 0.5 and 1. As a value, the larger the AUC value is, the more likely the present classification algorithm will rank the positive samples in front of the negative ones, i.e., it will be able to classify them better.

AUC = 1 means that this is a perfect classifier, i.e., when this prediction model is adopted, a perfect prediction is produced regardless of the threshold set. However, in realistic forecasting situations, perfect classifiers hardly exist. AUC = 0.5 means that it is a random guess, and the model has no predictive value. Whereas when 0.5<AUV<1, it means that the model is better than a random guess. This classifier will provide predictive value if the threshold is properly set.

The SVM and Random Forest algorithms work better, their AUC are 0.83, without significant difference.

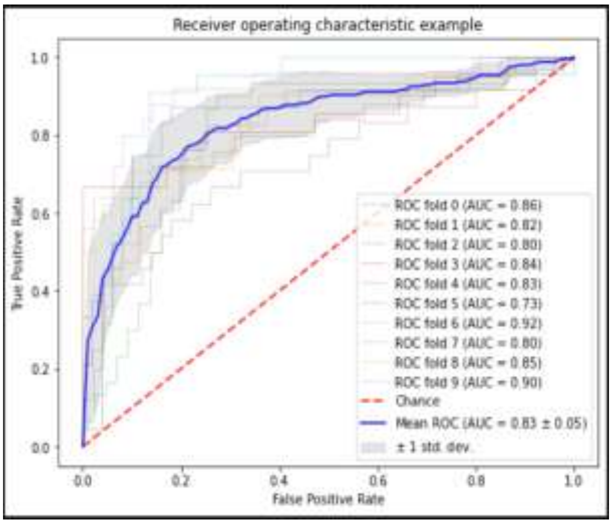

Fig. 2. ROC Curve of SVM Classification (Min-max Newdata)

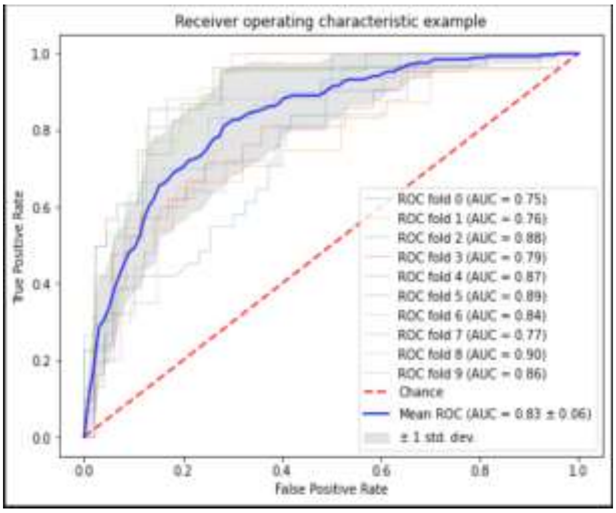

Fig. 3. ROC Curve of RF Classification (Min-max Newdata)

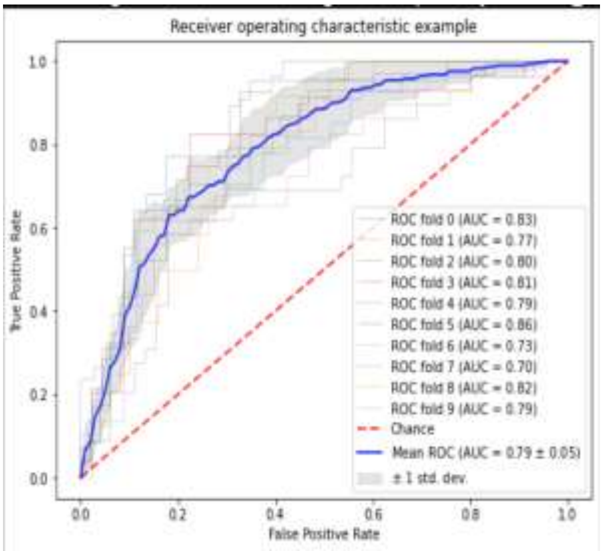

Fig. 4. ROC Curve of XGboost Classification (Min-max Newdata)

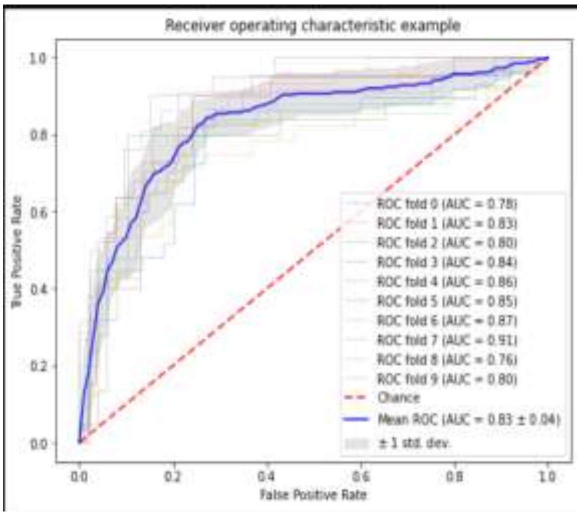

Fig. 5. ROC Curve of SVM Classification (Z-Score Newdata)

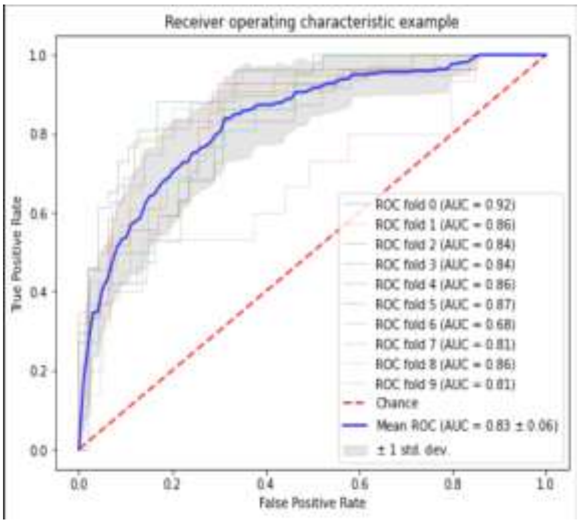

Fig. 6. ROC Curve of RF Classification (Z-Score Newdata)

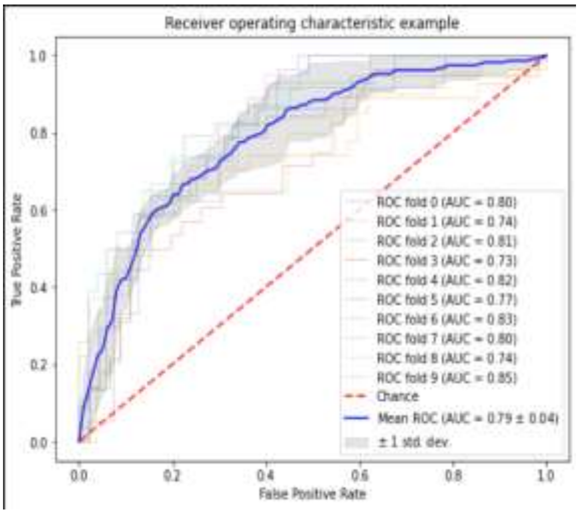

Fig. 7. ROC Curve of XGboost Classification (Z-Score Newdata)

IV. CONCLUSION

The highlight of this study is the use of machine learning algorithms for data mining to achieve accurate prediction of diabetes. Before training for classification, we cleaned the original dataset very carefully and processed the cleaned dataset contrastingly with two normalization methods: Mix-Max and Z-Score. Through a series of experiments implemented by Python, we found that SVM and Random Forest algorithms performed best with 0.83 of AUC value, more accurately, and more strongly.

For the further research, more diverse machine learning algorithms can be tried, and even deep learning methods are also very good choices.

REFERENCES

[1] International Diabetes Federation. IDF Diabetes Atlas, 9th edn. Brussels, Belgium: International Diabetes Federation, 2019.

[2] Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. (New York: Springer series in statistics, 2001). p361

[3] Zhou Z H. Ensemble methods: foundations and algorithms. (Chapman and Hall/CRC, 2019).

[4] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.

[5] Li H. Statistical learning methods [M]. (Tsinghua University Press, 2012). p97-135

[6] UCI repository of machine learning Databases, Pima Indian Diabetes Dataset. [online] Available at: https://data.world/data-society/pima-indians-diabetes-database, created in 2016.

[7] Kuhn M, Johnson K. Feature engineering and selection: A practical approach for predictive models. (CRC Press, 2019). p203

[8] Kuhn M, Johnson K. Applied predictive modeling. (New York: Springer, 2013). p42

[9] Wang, L., Gao, P., Zhang, M., Huang, Z., Zhang, D., Deng, Q., ... & Wang, L. (2017). Prevalence and ethnic pattern of diabetes and prediabetes in China in 2013. Jama, 317(24), 2515-2523.