2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE)

2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE) | 978-1-6654-2204-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/CSAIEE54046.2021.9543343

Prediction of Diabetes with its Symptoms Based on Machine Learning

Xingchen Xu[1*]
School of Mechanic Engineering and Automation,
Beihang University,
Beijing, 100083, China
18375426@buaa.edu.cn

Jinhui Ma[2]
School of Engineering, UC Santa Cruz,
Santa Cruz, CA, 95064, United States

Xiao Huang[3]
The Stony Brook School,
Stony Brook, NY, 11790, United States

Xuejianwei Luo[4]
Guiyang Happy Valley International Experimental School,
Guiyang, Guizhou, 550024, China

Abstract: **As the destruction of diabetes is significant to the whole world, we want to focus on it and extract useful information from the correlation between symptoms and disease. The dataset obtained from UCI is the fundamental resource for the research. In order to ensure the accuracy of the project conclusions, three different approaches were used to verify each other: literature analysis, data analysis and machine learning. Literature part mainly contains previous work and large quantities of medical research done on diabetes. Data analysis included data preprocessing and visualization so as to unfold the concealed information of the dataset. Machine learning is to use the inspiration from the previous two parts to attain a suitable model for diabetes prediction. The project finally provides knowledge of different symptoms of diabetes and their relation with diabetes. It also elaborates how symptoms can be used to predict disease. Finally, we put forward suggestions for the prevention of diabetes and monitoring of potential disease.**

Keywords: Diabetes, Prediction, Machine Learning

I. Introduction

Diabetes is a syndrome of insufficient absolute or relative secretion of insulin to cause a series of metabolic disorders such as protein, fat, water and electrolytes, which can lead to chronic complications of blood vessels, heart, nerves, eyes and other tissues. In typical clinical cases, polyuria, polydipsia, polyphagia, weight loss and other manifestations may occur. Diabetes has become the fifth killer of human of all disease, so it is urgent to promote the understanding of it so that more people can be free from its threat or at least receive earlier treatment.

In general, people of all ages could have diabetes, just the different kinds of diabetes. A lack of insulin production and requires daily insulin administration are features of Type 1 diabetes (previously called insulin-dependent, childhood-onset). Symptoms like polyuria, polydipsia, weight loss, etc., may occur suddenly [1].

Type 2 diabetes arises from the body's ineffective insulin use, is primarily the result of excess body weight and physical inactivity. In the past, type 2 diabetes has only occurred in adults. However, in recent years, more children also get type 2 diabetes [1]. The reason for this phenomenon may be the improvement of living conditions, exercise is less while unhealthy food is preferred, which makes children become obesity.

Therefore, there is a relationship between diabetes and age. However, it is not about whether or not to get sick. It is about what kind of disease to get. Younger people are more likely to get type 1 diabetes, and older adults are more likely to get type 2 diabetes [2].

In obesity, it plays an essential role in developing type 2 diabetes, and it is also the leading aetiological cause of type 2 diabetes [3]. According to the analysis result, this attribute is weakly related to diabetes. Because the sample patients' average age is 48 years old, they are more likely to get type 2 diabetes. Type 2 diabetes is essentially the result of excess body weight and physical inactivity. However, most of the samples are not obese, so it was deduced that they belong to Non-obese type 2 diabetes. Non-obese type 2 diabetes may have post-receptor dysfunction of target cells initially, like insulin antibody. The islet B cells with hereditary diabetes have defects themselves and cannot adapt to islet resistance, so non-obese type 2 diabetes occurs.

According to the 2019 IDF report, in 2019, the prevalence of diabetes in patients between 20 and 79 is around 9.3%. It is expected to increase to 10.2 and 10.9 by 2030 and 2045 [4].

Although there are researches about machine learning used in diabetes, none of them concern many attributes and symptoms. Therefore, in the work, we combined data analysis and machine learning to give more comprehensive conclusions.

In order to get fully understanding of those symptoms and diabetes, four main questions were brought out:

①Is it true that people who are older are more likely to develop such disease?

②Among men and women, who have higher possibility to develop diabetes?

③Which attribute or attributes can have more impact on determining whether a person is positive?

④In order to prevent and judge in advance, what symptoms should people be vigilant to?

II. Exploring Data Analysis

Medical papers about diabetes were analyzed first to give clues and guide the research through the project. Then, EDA

provided detailed information about each attribute and their correlation with diabetes. Based on the information gained from EDA, machine learning was used to predict diabetes, which was the ultimate goal of the project. To form conclusions, each part's conclusions were not just combined, instead, in order to maintain a unified and correct condition, conclusions from all the three parts were cross-compared and adjusted.

The dataset used in this project was acquired using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh and approved by a doctor. It contains 16 attributes and 1 target class (table 1, table 2). The total object number is 520 with no missing values. The ratio of positive class to negative class is 2:1, which means the dataset is balanced. As the data is nominal, in order to process it, it was changed into binary code: 0 and 1.

Table 1: original dataset

Age	Gender	Polyuria	Polydipsia	Sudden weight loss	weakness	Polyphagia	Genital thrush	Visual blurring	Itching
40	Male	No	Yes	No	Yes	No	No	No	Yes
58	Male	No	No	No	Yes	No	No	Yes	No
41	Male	Yes	No	No	Yes	Yes	No	No	Yes

Table 2: original dataset

Irritability	Delayed healing	Partial paresis	Muscle stiffness	Alopecia	Obesity	class
No	Yes	No	Yes	Yes	Yes	Positive
No	No	Yes	No	Yes	No	Positive
No	Yes	No	Yes	Yes	No	Positive

The 16 attributes related with diabetes are binary code while 'Age' is not. As in the figure 1, the standard deviation of age is much higher than any other attribute. Therefore, to avoid the possible effect of age, normalization of age was also performed.

```
file.std()

Age                 11.816607
Gender               0.489470
Polyuria             0.499773
Polydipsia           0.499773
sudden weight loss   0.494841
weakness             0.492670
Polyphagia           0.497821
Genital thrush       0.426037
visual blurring      0.497821
Itching              0.500762
Irritability         0.428003
delayed healing      0.498782
partial paresis      0.499319
muscle stiffness     0.484060
Alopecia             0.468645
Obesity              0.348607
```

Figure 1: standard deviation

To get a rough impression of the data, qualitative analysis was implemented by radar map (see figure 2).

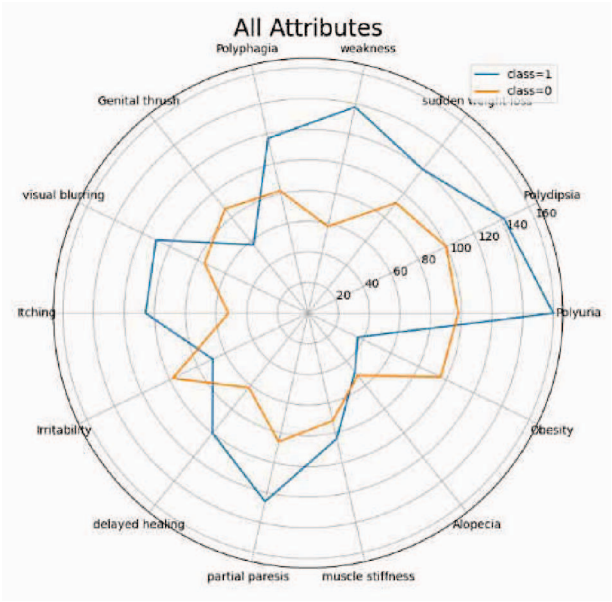


Figure 2: radar map

The blue line represents people with diabetes and the orange line represents people without diabetes. The graph suggests that Polyuria, Polydipsia, Partial Paresis, Delayed Healing, Itching, Visual Blurring, Polyphagia, Weakness and Sudden Weight Loss have a strong correlation with diabetes since the ratio of number of people having diabetes to people who do not is high. Other attributes like Obesity, Alopecia, Muscle Stiffness, Irritability and Genital Thrush are in weak relativity.

To quantize the extent of correlation, heat map was used. Crammer's V coefficient was used instead of Pearson method since it is a 0-1 distribution. As illustrated in figure 3, the rightest column of the map is the impact of attributes to diabetes solely. In the center of the graph, those figures represent the correlation

between the two attributes. And a bigger number indicates a stronger correlation. Since medical research and related features are always obscure, compared to each other, those number greater than 0.4 were regarded as strong correlation and those greater than 0.6 as very strong correlation.

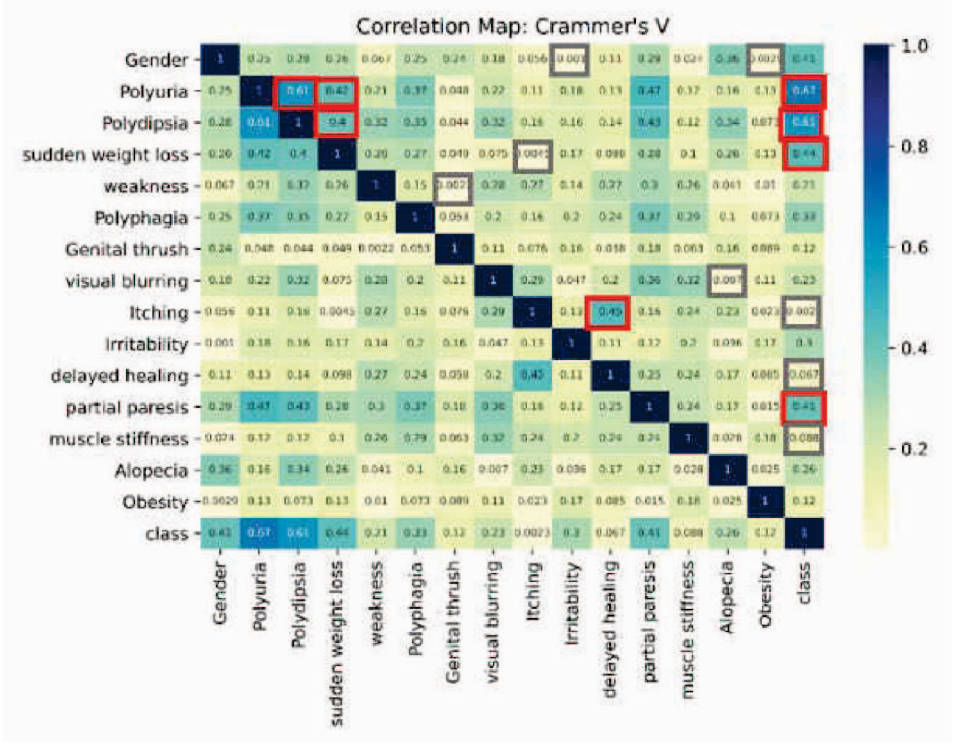


Figure 3: heat map

It can be gathered from the map that strong factors are Polyuria, Polydipsia, Sudden Weight Loss and Partial Paresis. Weak factors are Itching, Delayed Healing and Muscle Stiffness [5]. Factors with strong relativity to each other are Polyuria and Polydipsia, Polyuria and Sudden Weight Loss, Polyuria and Partial Paresis, Polydipsia and Sudden Weight Loss, Polydipsia and Partial Paresis, Itching and Delayed Healing [6]. Factors existing weak relation with each other are Gender and Irritability [7], Gender and Obesity, Sudden Weight Loss and Itching, Weakness and Genital Thrush, Visual Blurring and Alopecia [8].

In order to check the influence of gender to diabetes, histogram was used (figure 4). The green bars represent the samples who test positive for diabetes and the pink bars represent the samples who test negative. According to the data, we could calculate the percent of positive samples in each gender over each gender samples. The incidence rate of female who have diabetes dividing all female is 91.1% and for the male samples is 50.8%. Therefore, conclusion can be drawn that the incidence rate of female is greater that the male's.

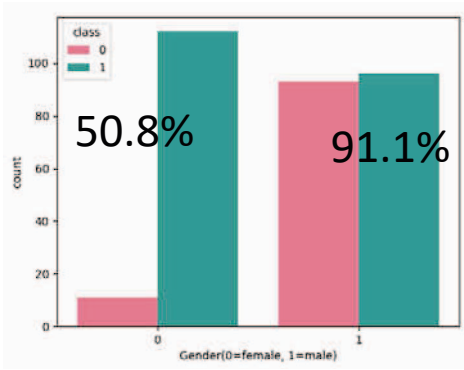


Figure 4: histogram of gender

To get a more specific understanding of Gender and Age, we separated them in two graphs (figure 5). From 30 to 60, the percentage of women who have diabetes in the whole gender is much higher than it is for men, proving that women are more likely to get such disease. Also, as the figure 6 shows, adults mainly develop type 2 diabetes and the possibility of getting diabetes for female is far higher than it is for male, which confirms that the EDA conclusion in Gender and Age is right.

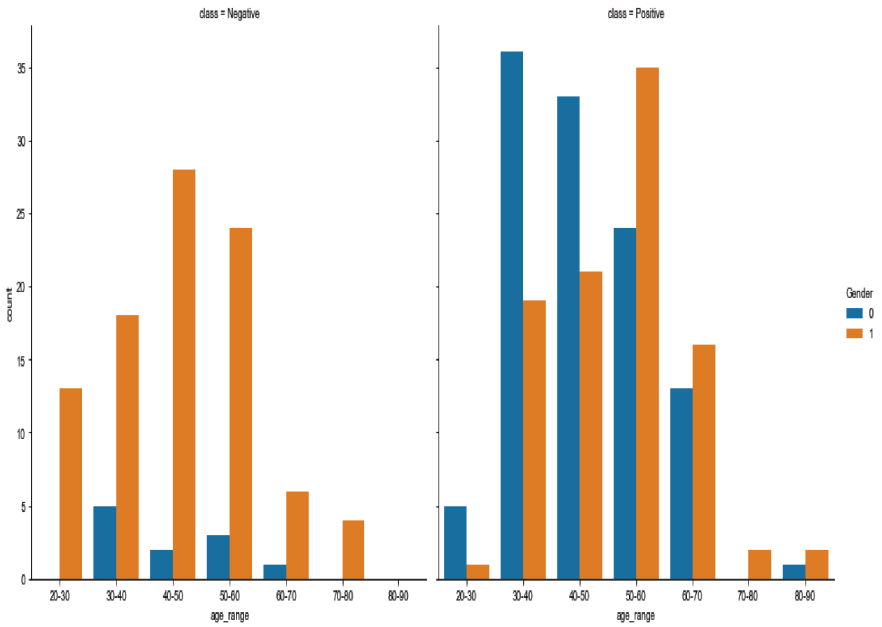


Figure 5: histogram of gender and age

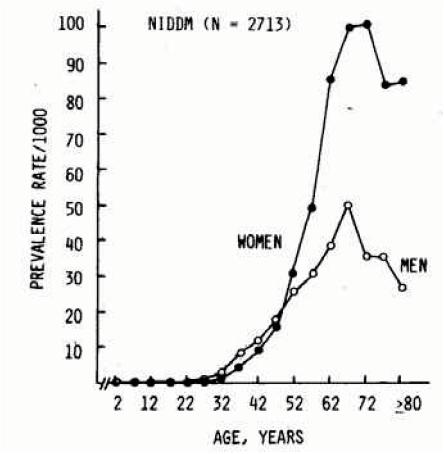


Figure 6: influence of gender and age on diabetes' type

Another attribute of high correlation is Polyuria. As shown in figure 7, the percent of people who are positive with Polyuria is 70.49% while the percent of people who are positive without Polyuria is 29.51%. Therefore, it is apparent that Polyuria is highly correlated with the diabetes which the correlation is more than 60 percent.

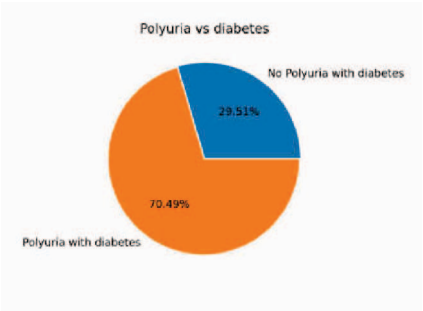


Figure 7: polyuria and diabetes

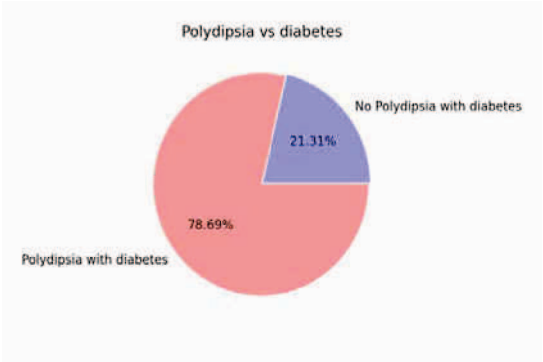


Figure 8: polydipsia and diabetes

Polydipsia is strong related with diabetes (figure 8). The percent of people who are positive with Polydipsia is 78.69% compared with 29.51% of those positive but without Polydipsia, proving Polydipsia's importance in diabetes, as the same as it is for Partial Paresis (figure 9).

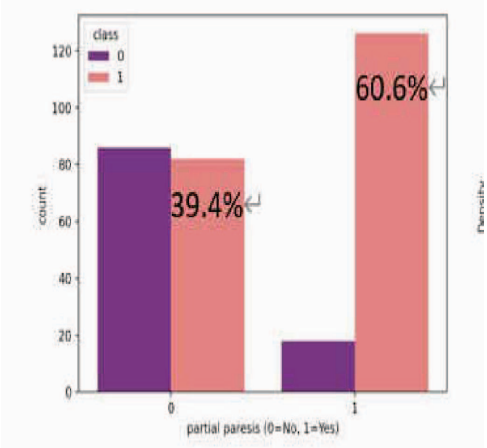


Figure 9: histogram of partial paresis

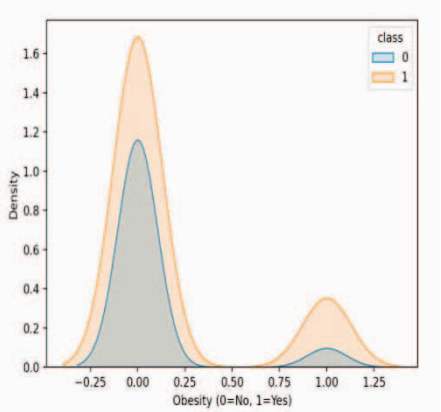


Figure 10: kernel density of obesity

The diagrams above depicted a strong correlation between the attributes assigned and diabetes. To give a sense of comparison, Obesity and Muscle Stiffness were chosen. Kernel Density Estimation curve was used to analyze Obesity (figure 10). The orange shade represents the positive diabetes and the blue one represents the negative ones. The samples who do not have Obesity have a greater amount of density than those who have, indicating that Obesity is not correlated with diabetes. The same thing also happens to Muscle Stiffness (figure 11) [5].

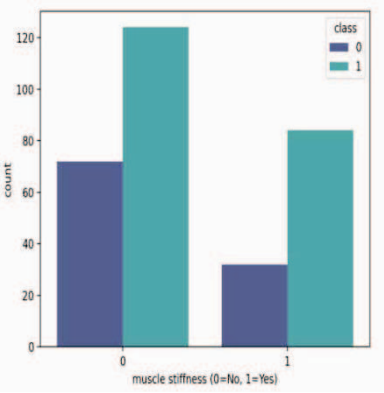


Figure 11: histogram of muscle stiffness

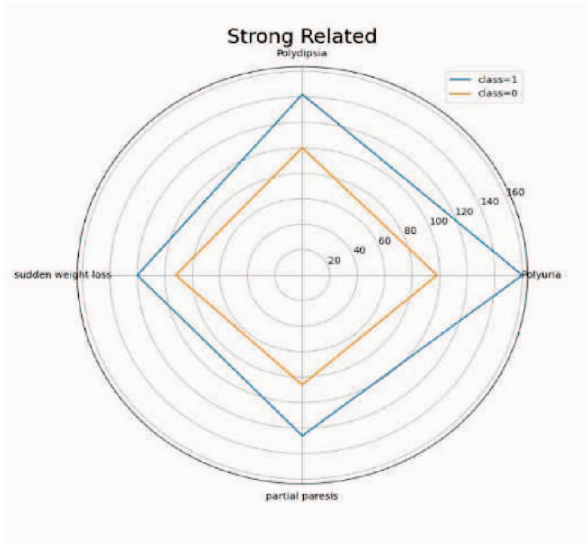


Figure 12: radar map of strong correlated factors

To gain a directly perception, the four strong correlated factors were put in one radar map (figure 12). It is obvious that people who already have diabetes are more likely to develop such symptoms than people who do not.

III. Further Research Based on Machine Learning

To fully accomplish the goal of predicting diabetes, machine learning is applied to the model.

A. Simple Models

1)SGD

The first basic ML method is Stochastic Gradient Descent (SGD). The default setting on non-normalized training set shows 71% accuracy with 13.8% standard deviation. The main factor of the SGD is max iteration. As shown in the figure 13, it is based on the independent variable Max Iteration. The highest score is 87.5% when adjusting max iteration.

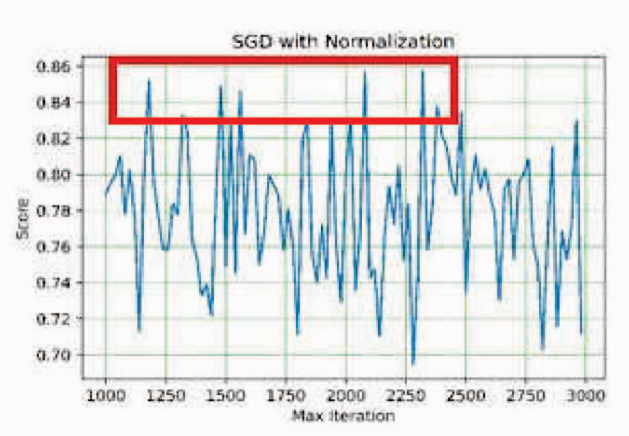


Figure 13: impact of max iteration without normalization

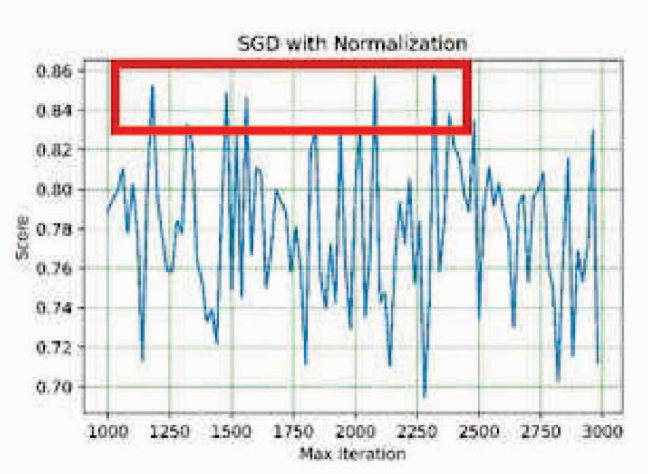


Figure 14: impact of max iteration with normalization

The performance on normalized dataset is worse, which is 85.7% (see figure 14). Therefore, the SGD is not a wise choice for this case.

2)KNN

The second basic ML method is KNN. The accuracy on default setting is 84.1% and on testing set is 87% (table 3). The

main arguments affecting accuracy are algorithm, leaf size and n neighbors. By using Exhaustive Grid Search (EGS), best parameters were settled. The highest score by cross validation is 88.2% with 88% on testing set. Using the same optimization on the normalized dataset can significantly improve the accuracy. Based on EGS, the highest score by cross validation is 92.3% and the precision on testing set is 96% (table 4).

Table 3: default setting on KNN

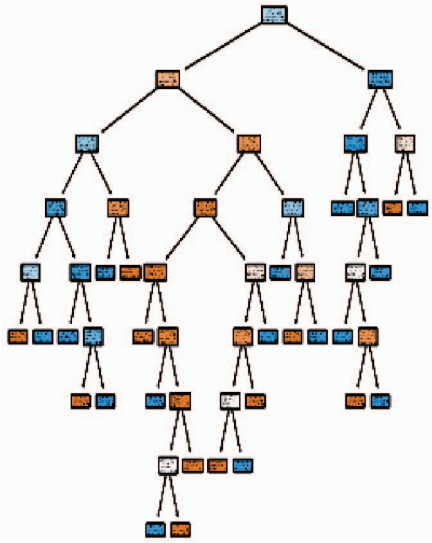
Default	Precision	Recall	F1
Negative	0.76	0.92	0.83
Positive	0.94	0.81	0.87
Accuracy	0.87	0.85	0.85

Table 4: EGS on KNN

Optimization	Precision	Recall	F1
Negative	0.91	1.00	0.95
Positive	1.00	0.94	0.97
Accuracy	0.96	0.96	0.96

3)Decision Tree

The last basic method is Decision Tree based on CART. The initial accuracy using default is 96.4% even the GINI value can be up to 0.496 (figure 15), which is rather high compared with the two methods before.



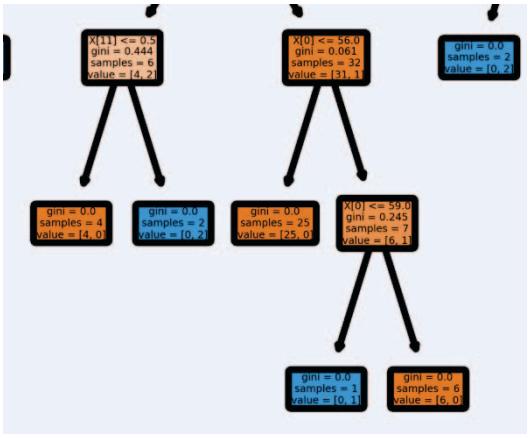


Figure 15: tree map and partial enlarged view

There are three main parameters in Decision Tree to adjust: max depth, min samples split and min samples leaf. However, Exhaustive Grid Search did not improve the performance, which remained 96.9% on training set and 97% on testing set (table 5). Since normalization has no contribution to tree like methods, it was not discussed.

Table 5: EGS on decision tree

Default	Precision	Recall	F1
Negative	0.95	0.97	0.96
Positive	0.98	0.97	0.97
Accuracy	0.97	0.97	0.97

4)ROC on Simple Methods

According to the analysis above, the two competitive methods are KNN and Decision Tree. To have a more qualitative impression, ROC curve was used (figure 16). The areas of the two curves are the same and their turning points are close to each other, indicating that the two methods have similar performance.

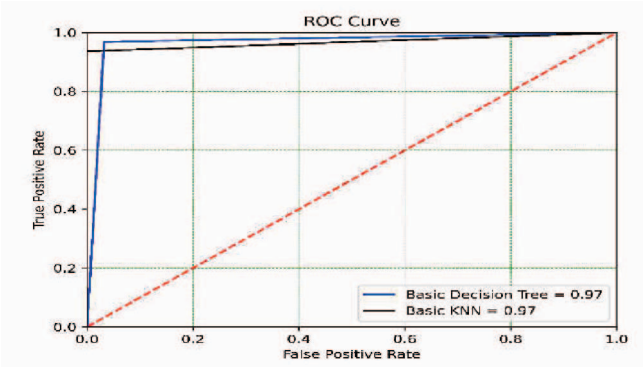


Figure 16: ROC curve

B. Complex Models

1)Bagging Based on Random Forest

The first approach is bagging based on Random Forest. The initial accuracy using default setting is 97.8%, higher than any optimized simple methods. In order to adjust those parameters easier and quicker, we separately examined the influence of

variables on the predicted scores (figure 17). The three graphs show the impact of n estimators, min samples split and min samples leaf on accuracy of training set.

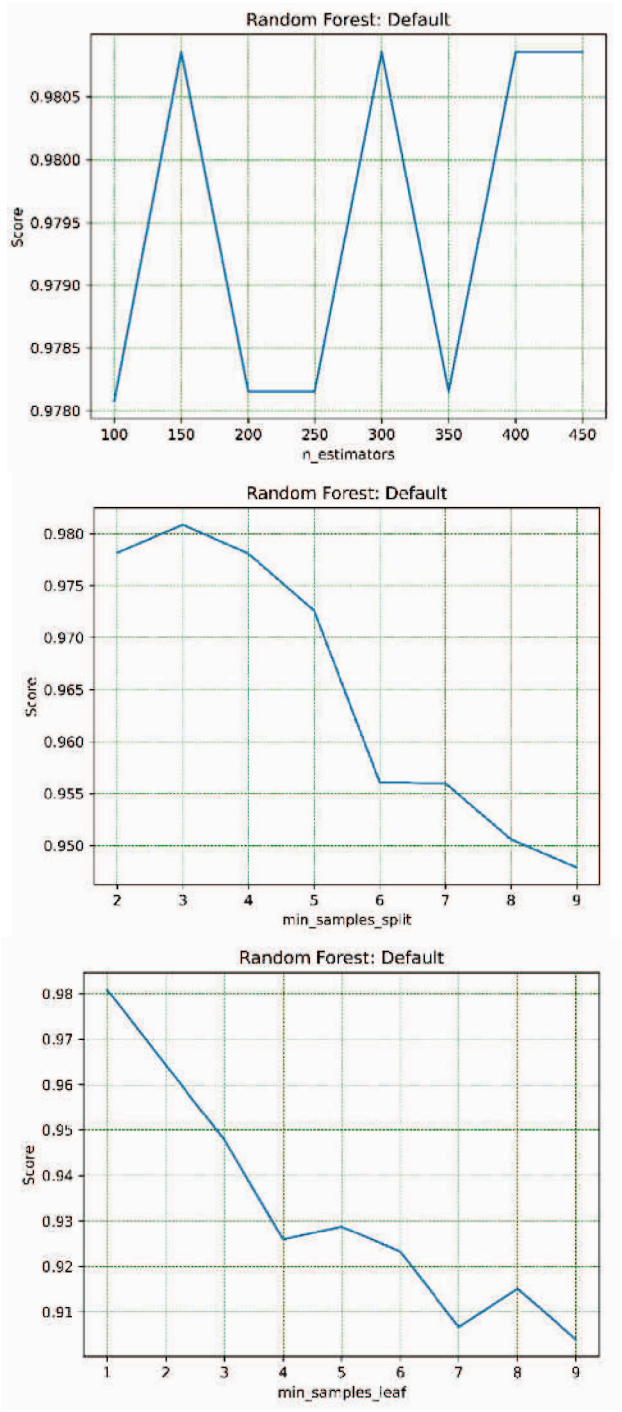


Figure 17: influence of variables on scores

Based on information from graphs and EGS, the score on training set by cross validation is 98.1% with standard deviation of 1.8%. The score on testing set is 98% (table 6). This result and parameters correspond with what shows in those diagrams, proving that examining parameters separately is feasible.

2)Neural Networks

Considering that the values of most of the attributes are binary code, a fully connected neural network (FCNN) was used. Due to the scale of the dataset, we believe that an FCNN with a structure of layers with reducing neuron numbers would have good performance based on intuition and former experience[9]. FCNN with 5 layers with the structure tested was shown on the chart: 16 attributes go into the first layer, and outputs 0 or 1 for the prediction of the final class. Testing with 5 validation sets results in an average cross-validation score of 0.96. When applied to a separate testing set, final testing score was 0.96 (table 7).

Table 6: EGS on random forest

Default	**Precision**	**Recall**	**F1**
Negative	0.97	0.98	0.98
Positive	0.99	0.98	0.98
Accuracy	0.98	0.98	0.98

Table 7: scores on neural networks

Default	**Precision**	**Recall**	**F1**
Negative	0.94	0.95	0.94
Positive	0.97	0.96	0.96
Accuracy	0.96	0.96	0.96

3)AdaBoost Based on Decision Tree

Two weak learners were used to analyze. One is Decision Tree and another is MLP. On Decision Tree, the initial accuracy using default setting is 92.3%. As there were more arguments to adjust, it is wise to deal with the frame parameters first then base estimator. Validation curve can provide an approximate optimal range of parameters. So frame parameters were adjusted by validation curve (figure 18, figure 19).

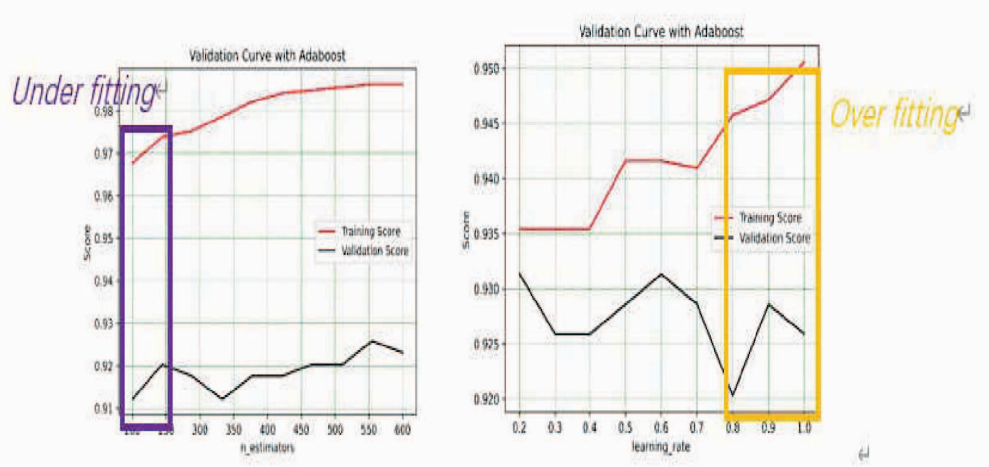
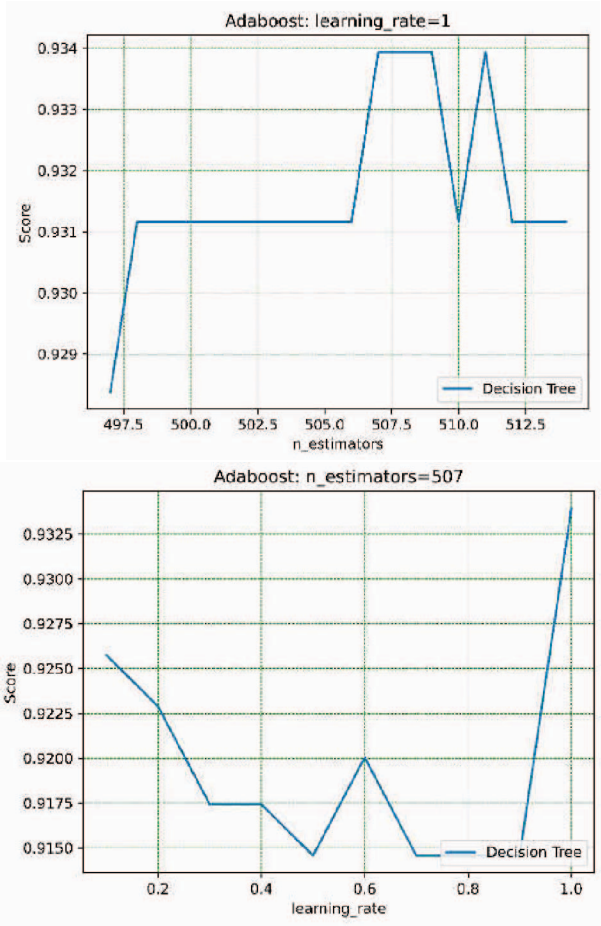


Figure 18: adjusting frame parameters with validation curves

If the training score is high and the validation score is low, the estimator is overfitting. If the training score and the validation score are both low, the estimator will be under fitting. Therefore, the middle parts may be the feasible range.



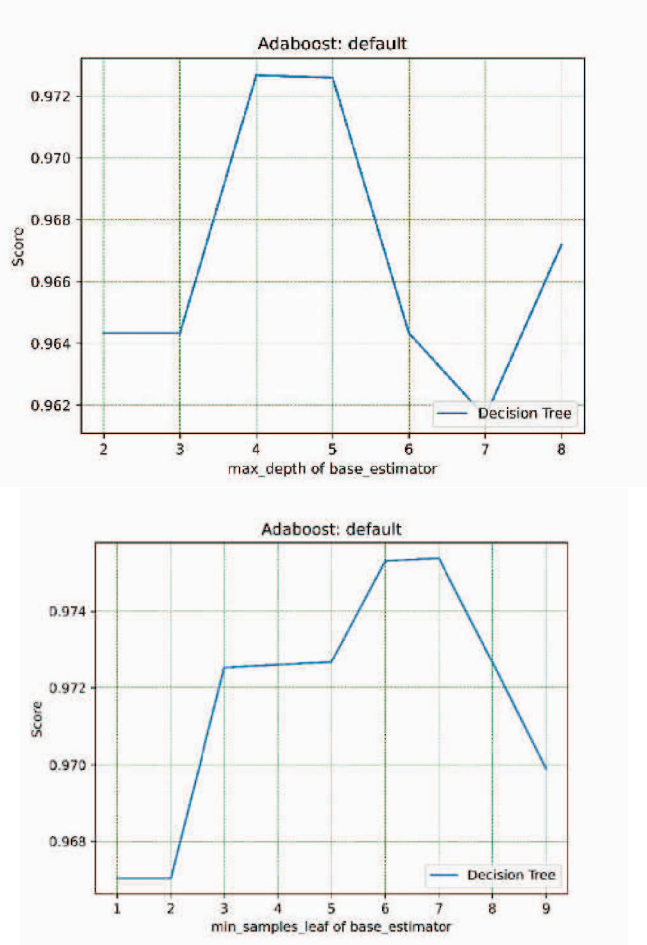


Figure 19: optimal frame parameters

After the frame is settled, base estimators were checked separately such as max depth, min samples split and min samples leaf (figure 20).

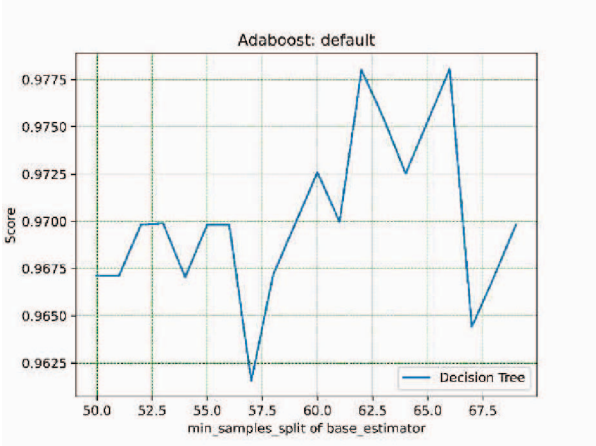


Figure 20: adjusting base estimators with validation curve

Using the possible range from the three diagrams, Grid Search yields the best score on training set: 97.8% and best accuracy on testing set: 98% (table 8).

Table 8: EGS on adaboost on decision tree

Default	**Precision**	**Recall**	**F1**
Negative	0.97	0.98	0.98
Positive	0.99	0.98	0.98
Accuracy	0.98	0.98	0.98

4)AdaBoost Based on MLP

With MLP, much more parameters should be adjusted. The sequence was the same with Decision Tree, which was to adjust frame first and then base estimator. As there were 6 arguments to adjust, Grid Search was not suitable for its huge time expanse. Instead, Random Search was used. The best score on training set is 97.3%.

5)ROC Based on Complex Models

ROC curve (figure 21) shows that AdaBoost with Decision Tree has the best performance and Random Forest has the second place. But all three methods have similar score and show better results than those simple models. It suggests that boosting and bagging both have great performance than simple model. However, the less parameters a model has, the easier the adjusting is and the better the performance may output.

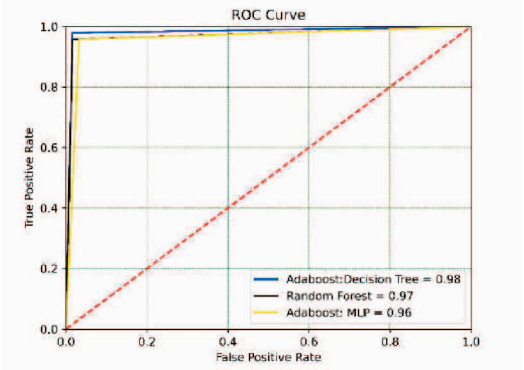


Figure 21: ROC curve

IV. analyses on EDA and machine learning

All three methods prove that Polyuria, Polydipsia, Sudden Weight Loss and Partial Paresis are strongly related to diabetes. It is also clear that some attributes are highly correlated with other attributes and may have a gigantic impact on the disease such as Polyuria and Polydipsia, Polyuria and Partial Paresis. Based on this, it seems that some attributes embody high relation but in fact. Due to the existence of complications of diabetes, group with diabetes may have those attributes and may not have them, which means those attributes are not the core factors of diabetes and should not be allocated with a high correlation coefficient.

In other hands, some attributes may incur bias on diabetes but it is not true. For instance, Obesity seems to be lowly related with diabetes but it may be the result of other possible causes, such as failure of key protein. As a matter of fact, type 2 diabetes has been proved to be a disease and obesity is the main cause of type 2 diabetes. Therefore, the correlation coefficient of Obesity should be raised rationally. The conclusions in EDA and medical research are contrary involving Itching. However, in clinic research, doctors also found that the real phenomenon contradicting to it in the observing group. The two findings indicate that even some attributes show on many people, it doesn't mean that those attributes herald high possibility of developing diabetes. Therefore, even many people who have diabetes also have Itching, Itching may be just one of the small complications of diabetes since lots of people have Itching but not diabetes.

The attributes Age and Gender should be discussed together as they are basic properties of a person. Female have higher possibility of developing diabetes than male and the whole trend is that elder group has higher possibility of getting this disease. It is been proved that elder ones have diabetes easier by medical research. The little bias in the EDA is just because there are not enough samples in elder group. Nevertheless, normally it is assumed that women are more likely to develop diabetes than men according to statistics. However, recent medical research suggests that the possibility of getting diabetes is the same in male and female. The data outputted in EDA shows that Gender has some correlation with diabetes indeed. The fact is that female can have depression and other mental illness easier than male, which cause them to develop diabetes easier.

Synthesizing all those points above, the relative margin of each attribute can be calculated separately (table 9).

Table 9: relative possibility of contribution to diabetes of each attribute

Feature	Margin Ratio	Feature	Margin Ratio
Gender	0.41	Itching	0.0023
Polyuria	0.67	Irritability	0.30
Polydipsia	0.61	Delayed Healing	0.067
Sudden Weight Loss	0.44	Partial Paresis	0.41
Weakness	0.21	Muscle Stiffness	0.088
Polyphagia	0.33	Alopecia	0.26
Genital Thrush	0.12	Obesity	0.12
Visual Blurring	0.23		

Getting rid of one attribute with high correlation coefficient shown in heat map may not change anything. This is because another feature with similar correlation coefficient may perform the same role in ML. Therefore, getting rid of them simultaneously can cause acute change or little change to the prediction accuracy, proving them to be imperative (acute change) or irrelevant (little change).

V. Conclusions

The whole research synthesizes medical research, data analysis and machine learning. It is normal if some attributes show different relativity in the three fields. Nevertheless, some attributes are of strong connection with diabetes and should be paid more attention (such as polyuria, polydipsia). Some attributes are of little connection with it since they can be found in patients with other diseases (such as itching, delayed healing). Attributes whose relativity sit between strong and weak connection can be the results of complications of diabetes and other diseases, which should be concerned as well (such as partial paresis, polyphagia). Gender and age can affect the possibility of developing diabetes with the influence of environment.

Admittedly, there are more to do to improve the work, such as obtaining more samples, including people from different countries and continents and examining the influence of living environments. To give basic guidance, we strongly suggest people exercise frequently and have more low-fat and rich-protein food.

Reference

[1] World Health Organization, (n.d.). Diabetes. https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] Laakso, M., & Pyorala, K. (1985). Age of onset and type of diabetes. Diabetes Care, 8(2): 114-117.

[3] Astrup, A., & Finer, N. (2000). Redefining Type 2 Diabetes: 'Diabesity' Or 'Obesity Dependent Diabetes Mellitus'? Obesity Reviews, 1: 57-59.

[4] International Diabetes Federation, (2019). IDF Atlas 9th edition. https://www.diabetesatlas.org

[5] Bursać, Snježana Novaković, et al. “Complications of Diabetes Mellitus on Muscles and Joints of Lower Extremities.” Scripta Medica, vol. 49, no. 2, Oct. 2018, pp. 105–111. EBSCOhost, doi:10.7251/SCMED1802105B.

[6] Brown, Matthew L., et al. “Delayed Fracture Healing and Increased Callus Adiposity in a C57BL/6J Murine Model of Obesity-Associated Type 2 Diabetes Mellitus.” PLoS ONE, vol. 9, no. 6, June 2014, pp. 1–11. EBSCOhost, doi:10.1371/journal.pone.0099656.

[7] D H Surridge, D L Erdahl, J S Lawson, M W Donald, T N Monga, C E Bird and F J Letemendia BJP 1984, 145:269-276. “Psychiatric aspects of diabetes mellitus” Access the most recent version at DOI: 10.1192/bjp.145.3.269

[8] Wang, Sue-Jane, et al. “Increased Risk for Type I (Insulin-Dependent) Diabetes in Relatives of Patients with Alopecia Areata (AA).” American Journal of Medical Genetics, vol. 51, no. 3, 1994, pp. 234–239., doi:10.1002/ajmg.1320510313.

[9] Goodfellow, I. J 2016) Deep Learning. MIT Press