Application of Machine Learning to Predict And Diagnose Diabetes

Xu Li
Sheffield University, Science Automatic and Intelligent System
Western Bank Sheffield, S10 2TN, UK

Author’s E-mail: lixusolar@163.com

Abstract. This article uses linear regression in machine learning and LightGBM algorithm for data mining, and compares the linear regression and LightGBM algorithm’s ability of fast running speed and high accuracy under the same data conditions. The least squares method is used to make error judgments to realize the rapid and accurate prediction and judgment of the probability of diabetes in a large amount of data. At the same time, the importance and correlation of the variables are compared between the variables. Among them, it is found that Body Mass Index (BMI) has the greatest impact on diabetes, and the factors affecting BMI are weight and height. Height and age can be input into system, automatically determining the BMI value, and providing doctors with a basis for judgment.

1. Introduction
Diabetes is a serious chronic disease, divided into type 1 diabetes and type 2 diabetes. Type 1 diabetes is a congenital defect that can be treated with insulin with a little cases. Type 2 diabetes is an increase in blood sugar caused by cells in the body that cannot use insulin normally[1]. It is difficult to cure and has a large number of patients. In many countries, due to insufficient medical conditions, many patients do not realise that they have diabetes[2]. How to detect a large number of diabetic patients quickly and accurately has attracted much attention. This article uses machine learning methods, linear regression and LightGBM for data mining. When comparing the efficiency of each algorithm, the variable BMI is also selected as the most important factor affecting diabetes, and Person and Spearman are also used to compare the correlation between the variables. Finally, the system can determine the possibility of diabetes according to the variable BMI value determined between the variables, thereby greatly improving the detection efficiency and saving doctors and patients time[3].

2. Analysis
This article mainly uses machine learning to predict the probability of diabetes, which is convenient for doctors to predict patients in advance, and can treat diabetic patients at an early stage, which can solve many problems that patients do not know about the disease at an early stage. Evidently, the use of algorithms to filter the influencing factors of diabetes through data provided people with powerful
prevention and improvement information, which could help doctors and patients to quickly determine the cause of the disease. At the same time, it can reduce the treatment costs of patients in the later stage of diabetes and improve the happiness index of life[4].

Table 1 shows a small part of the data for the main data about the 442 diabetic patients. The diabetes data involved in this article are 10 baseline variables obtained for each of 442 diabetic patients, namely AGE, SEX, Body Mass Index (BMI), average Blood Pressure (BP) and 6 serum measurements (S1-S6), as well as diabetes progression of concern (y) during the year[5].

|   | AGE | SEX | BMI | BP   | S1   | S2   | S3   | S4   | S5   | S6   | y    |
|---|-----|-----|-----|------|------|------|------|------|------|------|------|
| 1 | 59  | 2   | 32.1| 101  | 157  | 93.2 | 38   | 4    | 4.8598| 87   | 151  |
| 2 | 48  | 1   | 21.6| 87   | 183  | 103.2| 70   | 3    | 3.8918| 69   | 75   |
| 3 | 72  | 2   | 30.5| 93   | 156  | 93.6 | 41   | 4    | 4.6728| 85   | 141  |
| 4 | 24  | 1   | 25.3| 84   | 198  | 131.4| 40   | 5    | 4.8903| 89   | 206  |
| 5 | 50  | 1   | 23  | 101  | 192  | 125.4| 52   | 4    | 4.2905| 80   | 135  |
| 6 | 23  | 1   | 22.6| 89   | 139  | 64.8 | 61   | 2    | 4.1897| 68   | 97   |
| 7 | 36  | 2   | 22  | 90   | 160  | 99.6 | 50   | 3    | 3.9512| 82   | 138  |
| 8 | 66  | 2   | 26.2| 114  | 255  | 185  | 56   | 4.55 | 4.2485| 92   | 63   |
| 9 | 60  | 2   | 32.1| 83   | 179  | 119.4| 42   | 4    | 4.4773| 94   | 110  |
| 10| 29  | 1   | 30  | 85   | 180  | 93.4 | 43   | 4    | 5.3845| 88   | 310  |

2.1. Regression modeling prediction

2.1.1. Linear regression. Figure 1 indicates the graphical results of linear regression. The purpose of linear regression is to minimize the residual sum of squares between the predicted value and the actual value. Residual error in mathematical statistics refers to the difference between the actual observation value and the estimated value (fitting value). The w is a vector.

\[ h(x) = w^T X + b \]  \( (1) \)
2.1.2. Ordinary least squares regression model. The least square method, also known as the least square method, is a mathematical optimization technique that finds the best function match of the data by minimizing the square sum of the error. The least square method can be used to easily obtain unknown data and minimize the sum of squares of errors between the obtained data and the actual data. This method can also be used for curve fitting[6].

\[ S(\alpha) = \|X\alpha - y\|^2 \] (2)

As shown in Table 2, 30% of the diabetes data is used to test the accuracy of the algorithm model. There are a total of 442 data, 30% of which are 133, and the total number of data is 556.1986. Figure 2 shows the result of fitting the test data (orange square) and the predicted result (blue curve).

**Table 2.** Test data representing 30% of total diabetes data

| y_predict | y_test | minus | abs(minus) | sum | numbers |
|-----------|--------|-------|------------|-----|---------|
| 1         | 138.4703227 | 219   | 80.52968   | 80.5296773 | 556.1986 | 133   |
| 2         | 181.103118  | 70    | -111.103   | 111.103118 |         |       |
| 3         | 125.346504  | 202   | 76.6535    | 76.653496  |         |       |
| 4         | 292.7540939 | 230   | -62.7541   | 62.7540939 |         |       |
| 5         | 123.8808007 | 111   | -12.8808   | 12.8808007 |         |       |
| 6         | 91.89920521 | 84    | -7.89921   | 7.89920521 |         |       |
| 7         | 257.2662357 | 242   | -15.2662   | 15.2662357 |         |       |
|   | y_test         | y_predict      |
|---|---------------|----------------|
|   | y_test         | y_predict      |
|   | y_test         | y_predict      |
|   | y_test         | y_predict      |
|   | y_test         | y_predict      |
| 8 | 177.7630936   | 272            | 94.23691      | 94.2369064    |
| 9 | 84.98399508   | 94             | 9.016005      | 9.01600492    |
| 10| 109.1573425   | 96             | -13.1573      | 13.1573425    |
| 11|               |                |               |               |
| 12|               |                |               |               |
| 13|               |                |               |               |
| 131| 122.2883565 | 63             | -59.2884      | 59.2883565    |
| 132| 80.8073436  | 93             | 12.19266      | 12.1926564    |
| 133| 233.2206723 | 232            | -1.22067      | 1.2206723     |

**Figure 2.** The fitting result of predicted data and test data

### 2.2. Evaluation index

#### 2.2.1. Mean square error

MAE is a linear indicator, and all individual differences are equally weighted on the average, so it highlights outliers more.

\[
MAE_1(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} |y_i - \hat{y}_i| \quad (3)
\]

#### 2.2.2. Median absolute error

In statistics, the absolute median MAD is a robust measure of the sample deviation of univariate numerical data. It can also represent the overall parameters estimated by the sample's MAD.

\[
MedAE(y, \hat{y}) = median(|y_1 - \hat{y}_1|, \ldots, |y_n - \hat{y}_n|). \quad (4)
\]
# Divide training set and test set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
# Create linear regression object
regr = linear_model.LinearRegression()
# Train the model using the training sets
regr.fit(X_train, y_train)

# Make predictions using the testing set
y_pred = regr.predict(X_test)

# Formula Derivation
MAE = sum(abs(y_test - y_pred)) / len(y_test)
# Median absolute variance
MAE1 = mean_absolute_error(y_test, y_pred)
# Median absolute variance
MedianAE = median_absolute_error(y_test, y_pred)
# Correlation comparison
r2 = r2_score(y_test, y_pred)

## 2.3. Algorithm comparison
LightGBM uses Leaf-wise instead of Level-wise for tree growth. This approach is mainly because
LightGBM believes that the Level-wise approach will produce some nodes with low information gain and
waste computing resources. In fact, Level-wise is still very effective in preventing overfitting, so everyone
prefers to compare it with Leaf-wise. Leaf-wise can pursue better accuracy and split nodes that produce
better accuracy. But this brings about the problem of overfitting, so use max_depth to control its
maximum height. The reason is that LightGBM is doing data merging, and various operations such as
Histogram Algorithm and GOSS actually have a natural regularization effect, so using Leaf-wise to
improve accuracy is a very good choice[1].

Table 3 compares the prediction accuracy of different algorithms of linear regression and LightGBM.
Table 3 lists the MAE values calculated by formulas in linear regression and LightGBM, respectively. The
mean-squared-error and Median absolute error are used to calculate MAE1. And MedAE, also use r2_score
at the same time to get the value of the correlation r2 between the two. The results of this article show that
the performance of LightGBM is much higher than linear regression, which is about twice that of linear regression.

**Table 3. Comparison of the prediction effects of different algorithms between linear regression and LightGBM**

|                  | MAE   | MAE1  | MedAE     | r2      |
|------------------|-------|-------|-----------|---------|
| Linear regression| 41.91925361 | 41.919253605566794 | 37.33816856677825 | 0.47729201741573335 |
| LightGBM         | 14.27389981 | 14.273899805021866 | 11.222111509158069 | 0.9340903897021662 |

2.4. Comparison of the importance of diabetic molecular weight

There are many factors causing diabetes. Diabetes can also cause many complications and a series of health problems. Therefore, it is important to study which factors will have the greatest impact on diabetes[7].

From Figure 3, we can see that among the various factors, BMI has the greatest impact on diabetes at about 265. SEX and S4 have roughly the same impact on diabetes and the smallest impact on diabetes is about 45, while BP and S5 have the smallest impact on diabetes. The effect is almost the same and the effect on diabetes is ranked second after BMI with a value of 200. The effect of S2 and AGE on diabetes is ranked third after BP and S5 with a value of 180. Finally, the importance of S1, S3 and S6 to diabetes The fourth impact ranking is 150.

Therefore, BMI has the greatest impact on the body of diabetes, while gender and serum content of S4 have the least impact on the body.

![lightgbm feature importance](image)

**Figure 3.** The comparison of the importance of each factor in diabetes after quantification

2.4.1. *Person Correlation Coefficient.* Pearson's correlation coefficient, also called Pearson's product-moment correlation coefficient, is a linear correlation coefficient used to reflect the degree of linear correlation between two variables X and Y[8].
Among them, $\sigma_x \sigma_y$ represents the standard deviation of the variables $X$ and $Y$, and $\text{cov}(X,Y)$ represents the covariance of the variables $X$ and $Y$.

$$\text{cov}(X,Y) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n} \quad (6)$$

Where $\bar{X}, \bar{Y}$ are the average of $X$ and $Y$.

$$\sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n} \quad (7)$$

Finally, we can get:

$$P_{X,Y} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2 (Y_i - \bar{Y})^2}} \quad (8)$$

**Table 4.** Pearson correlation coefficient comparison

|          | high_correlation_pearson |
|----------|--------------------------|
| S2       | S1                       | 0.896663 |
| S4       | S2                       | 0.659817 |
|          | S3                       | -0.73849 |
| S5       | S4                       | 0.617857 |

2.4.2. *Spearman Correlation Coefficient.* Spearman correlation coefficient is a kind of rank correlation coefficient. It is also usually called Spearman's rank correlation coefficient. Spearman can be understood that achievement is a kind of order or sorting, then Spearman is solved according to the sorting position of the original data.

$$P_s = \frac{\sum_{i=1}^{N} (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{N} (R_i - \bar{R})^2 \sum_{i=1}^{N} (S_i - \bar{S})^2}} = 1 - \frac{6 \sum d_i^2}{N(N^2 - 1)} \quad (9)$$

**Table 5.** Spearman correlation coefficient comparison

|          | high_correlation_spearman |
|----------|---------------------------|
| S2       | S1                        | 0.878793 |
| S4       | S2                        | 0.652283 |
|          | S3                        | -0.78969 |
| S5       | S4                        | 0.64039 |
By comparing the two correlations between person and spearman, it is found that the two comparison methods are roughly the same. Both S2 and S1 are highly correlated, S4 is highly correlated with S2 and S3, and S5 and S4 are highly correlated. The only difference is measured by person and spearman. The correlation coefficient is different, but there is no big difference.

2.5. BMI index model forecast

By using the above-mentioned machine learning algorithms, we know that BMI is the biggest factor affecting diabetes, and it is positively correlated. At the end of this article, the BMI index model is used to make a simple body mass index prediction to estimate the risk of diabetes. The following is the code for the BMI model. The value of BMI is determined by calculation. It is known that BMI<18.4, thin; BMI (18.5-23.9), normal; BMI (24-27.9), overweight; BMI>28 obese. Among them, x and y respectively correspond to weight and height. The patient's information can be input, judging the BMI value, and then predicting the probability of diabetes. The model will issue corresponding warnings[9]

```python
Weight, height = x, y
def BMI(weight, height):
    Bmi = weight/height**2
    return bmi
def Warn(bmi):
    if bmi<=18.4:
        print(you are thin)
    if 18.5<bmi<=23.9:
        print(you are normal, congratulations)
    if 24<bmi<=27.9:
        print(you are overweight, pay attention)
    if bmi>28:
        print(you are obesity, be caution)
```

3. Conclusion

This article introduces the use of machine learning and data mining to predict diabetes in the early stage, mainly using linear regression and LightGBM for algorithm comparison to compare their accuracy and computing speed. Use Mean square error and Median absolute error to compare the two algorithms of linear regression and LightGBM, and also use the correlation r2_score to reflect the correlation between the training data and the test data. The results show that the accuracy of LightGBM is very high, which is about linear regression. double.

As there are many pathogenic factors in diabetes, it is very important to conclude that the pathogenic factors have an impact on diabetes. After the importance test, it is found that BMI is the pathogenic factor with the greatest impact on diabetes. At the same time, person and spearman were used to compare the internal comparative correlation among many pathogenic factors, and the correlation results of the two were roughly the same.

Finally, according to the experimental results, BMI has the greatest impact on diabetes among many pathogenic factors, so a simple algorithm model for judging the BMI value was made to estimate whether obesity and the probability of diabetes based on weight and height.

Although this article achieves the prediction of diabetes by using linear regression and LightGBM algorithm, this is still a simple model. If a specific and comprehensive prediction of diabetes is to be achieved, the algorithm model must be improved to increase the data processing capacity of diabetes.
This article is not accurate enough to judge the impact factors of diabetes. In the future, it is hoped that the impact factors of diabetes can be quantified more accurately and directly remind patients which factors need to pay attention to which adverse effects on the body, and help patients make plans to change their physical conditions.

4. Acknowledgement
When participating in Artificial Intelligence for Driving Scientific Experiments hosted by Professor Robert F. Murphy from Carnegie Mellon University in the autumn of 2020, I completed related courses such as an Introduction to Science’s Demand for Artificial Intelligence, Predictive Models, Machine Learning and an Active Machine Learning Demonstration. After two months of hard work, I finally completed the paper. I am very grateful to Professor Robert F. Murphy for his help from Carnegie Mellon University in these two months. At the same time, I am very grateful to the reviewing teacher for their help, who provided useful suggestions for the revision and correction of my paper.

5. References
[1] Y. Wang and T. Wang, “applied sciences Application of Improved LightGBM Model in Blood Glucose Prediction,” 2020.
[2] A. Aminian et al., “Predicting 10-Year Risk of End-Organ Complications of Type 2 Diabetes With and Without Metabolic Surgery: A Machine Learning Approach,” vol. 43, no. April, pp. 852–859, 2020, doi: 10.2337/dc19-2057.
[3] K. Plis, R. Bunescu, C. Marling, J. Shubrook, and F. Schwartz, “A Machine Learning Approach to Predicting Blood Glucose Levels for Diabetes Management.”
[4] D. Preuveneers, “Mobile Phones Assisting With Health Self-Care: a Diabetes Case Study,” pp. 177–186, 2008.
[5] B. Y. B. R. E. Fron, T. R. H. Astie, I. A. I. N. J. Ohnstone, and G. R-eb-, “LEAST ANGLE REGRESSION,” vol. 32, no. 2, pp. 407–499, 2004.
[6] H. Gong et al., “A deep learning- and partial least square regression-based model observer for a low-contrast lesion detection task in CT,” pp. 14–18, 2019, doi: 10.1002/mp.13500.
[7] R. Singla, A. Singla, Y. Gupta, and S. Kalra, “Artificial Intelligence / Machine Learning in Diabetes Care,” pp. 2019–2021, 2019, doi: 10.4103/ijem.IJEM.
[8] S. S. A. B. U. Naser, “Design and Development of Diabetes Intelligent Tutoring System,” vol. IV, no. 9, pp. 8117–8128, 2016.
[9] N. Taghizadeh, H. M. Boezen, J. P. Schouten, and C. P. Schröder, “BMI and Lifetime Changes in BMI and Cancer Mortality Risk,” pp. 1–16, 2015, doi: 10.1371/journal.pone.0125261.