Research on the application of Machine Learning in predicting diabetes

Wenfeng Ye [1] and Hui Zeng[2, *]

[1] School of Computer Science, University of Nottingham Malaysia, Kuala Lumpur Malaysia.
[2] School of Midwifery, Ganan Medical University, Ganzhou, China.

Abstract

This article aims to establish a predictive model for diabetes occurrence using machine learning methods. We utilized a clinical dataset from patients at Sylhet Diabetes Hospital in Sylhet, Bangladesh, including various clinical features related to diabetes such as Age, Gender, Polyuria, Polydipsia, sudden weight loss, weakness, Polyphagia, Genital thrush, visual blurring, Itching, Irritability, delayed healing, partial paresis, muscle stiffness, Alopecia, and Obesity. We performed data preprocessing and feature engineering, converting categorical variables into numerical form and standardizing the data. Subsequently, we experimented with various machine learning algorithms including logistic regression, decision trees, and support vector machines. Through cross-validation and grid search for parameter optimization, we selected multiple linear regression as the final predictive model.

We evaluated the model's performance on the test set using metrics such as mean squared error (MSE), mean absolute error (MAE), R-squared (R^2), and root mean squared error (RMSE). Experimental results indicate that our model demonstrates high accuracy and reliability in predicting diabetes occurrence. Through this model, we can promptly identify individuals at risk of diabetes, providing doctors with more accurate diagnostic and treatment recommendations, and potentially offering crucial decision support for diabetes prevention and management.

Keywords

Machine learning; Linear regression; Polynomial regression

1. Introduction

Diabetes is a common and serious chronic disease that significantly impacts the quality of life and health of millions of people worldwide [1]. According to data from the World Health Organization (WHO), diabetes has become a global epidemic, with an estimated 320 million people expected to be affected globally by 2030. Diabetes not only imposes physical health burdens on patients but also leads to various complications such as cardiovascular diseases, kidney diseases, and retinopathy, posing a serious threat to patients' lives [2].

ICCIC 2024: International Conference on Computer and Intelligent Control, June 29–30, 2024, Kuala Lumpur, Malaysia

[*] Corresponding Author

hfywy4@nottingham.edu.my (W. Ye); 2904941810@qq.com (H. Zeng)
0009-0004-4497-3418 (W. Ye); 0009-0007-6117-3596 (H. Zeng)

Early prevention and diagnosis are particularly crucial for the prevention and control of diabetes. However, due to the complex pathogenesis of diabetes and the lack of obvious early symptoms, many patients are not aware of their health [3].

problems in the early stages of the disease, leading to missing the optimal treatment window. Therefore, establishing effective diabetes prediction models to timely identify individuals at higher risk of the disease is of great significance for reducing the incidence of diabetes and improving patients' quality of life.

Traditional diabetes prediction methods mainly rely on doctors' experience and clinical indicators such as blood glucose levels and insulin sensitivity. However, these methods have drawbacks including long diagnosis times, high costs, and reliance on medical resources. With the continuous development of machine learning and artificial intelligence technologies, it has become possible to construct diabetes prediction models using big data and deep learning methods. Machine learning algorithms can learn from massive clinical data to discover potential patterns and features of diabetes occurrence, providing doctors with more accurate and rapid diabetes diagnosis and prediction services, thereby contributing to improving diabetes prevention and control efforts [4].

Therefore, this study aims to utilize machine learning methods, based on clinical data, to construct a diabetes prediction model, achieving fast and accurate prediction of diabetes occurrence. This will provide medical institutions and patients with more effective prevention and management strategies, ultimately reducing the incidence of diabetes and improving public health. Through the conduct of this research, it is hoped to provide important theoretical and practical support for further research and application in the field of diabetes prediction [5].

The dataset utilized in this study comprises 520 samples. After data cleansing, anomalies and missing values were addressed, and features in the dataset were processed, including transformation, encoding, and standardization. Additionally, we explored the distribution of age data by plotting the histogram of the target variable. The histogram revealed a normal distribution of age data.

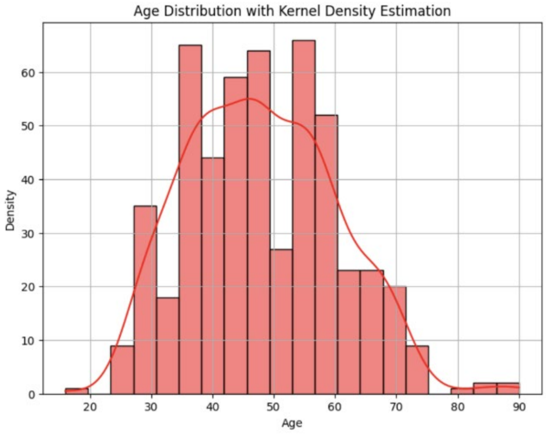


Figure 1: Distribution histogram of age.

For categorical variables, we employed one-hot encoding for transformation. For numerical variables, standardization was conducted to ensure uniform scales across different features [6].

In response to the issue of excessive or redundant features, feature selection was performed to identify features with significant impact on the target variable. As a result, 16 features were retained, as depicted in the following figure. These features appear to exhibit strong multicollinearity among them.

For categorical variables, we employed one-hot encoding for transformation. For numerical variables, standardization was conducted to ensure uniform scales across different features.

In response to the issue of excessive or redundant features, feature selection was performed to identify features with significant impact on the target variable. As a result, 16 features were retained, as depicted in the following figure. These features appear to exhibit strong multicollinearity among them.

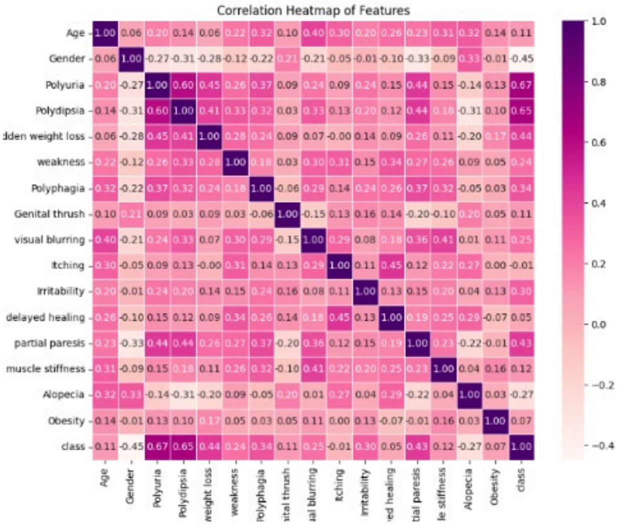


Figure 2: Heavy correlations between features.

In this study, we adopted the mean square error (MSE) as the index to evaluate the prediction performance of the model. The mean squared error is a measure of the squared difference between the predicted value of the model and the actual observed values. It is calculated by summing the squared differences between the predicted value and the actual observed value for each sample and then dividing by the number of samples. A smaller MSE indicates a more accurate model prediction [7].

$$MSE = \frac{1}{m}\sum_{i=1}^{m}\left(y^{i} - \bar{y}\right)^{2} \tag{1}$$

2. Model Establishment

2.1 Normal equations in linear regression

Linear regression is a statistical method used to establish a relationship between an independent variable (or feature) and a dependent variable. It assumes that there is a linear relationship between the independent variable and the dependent variable, that is, it can be described by a straight line. The goal of linear regression is to find the best-fit line that minimizes the error between the predicted value and the observed value.

Normal equations are a method used to solve the parameters of linear regression models. It finds the best fitting line by minimizing the sum of squared residuals.

The linear regression model is of the form.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \tag{2}$$

By taking the partial derivatives of the model parameters and setting them equal to zero, we can obtain the normal equation.

$$\beta = (X^T X)^{-1} X^T y \tag{3}$$

By taking the partial derivative of the loss function with respect to the parameter vector β, setting it equal to zero, and then solving for the parameter vector β, we can obtain the normal form of the above equation.

We fit a linear regression model to the training data in our study and evaluate the model performance on the test set. The accuracy and generalization ability of the model were evaluated by calculating the coefficients and intercepts of the model, as well as using the mean squared error [8].

Below is the confusion matrix for the mean variance of the linear regression equation model

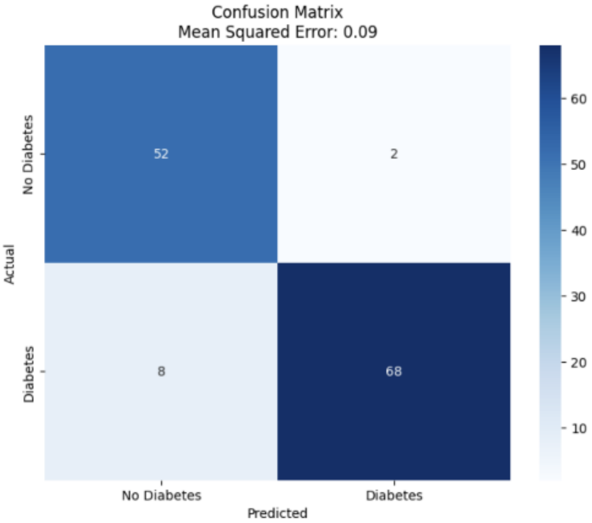


Figure 3: Confusion matrix for the mean variance of the positive regression equation model.

2.2 Linear Regression: Gradient Descent

Stochastic Gradient Descent (SGD) is an optimization algorithm used to train machine learning models, especially on large-scale datasets. It is an iterative optimization algorithm in which the objective function of each round calculation is no longer the whole sample error, but only a single sample error, that is, only one sample at a time is substituted to calculate the gradient of the objective function to update the weight, and then the next sample is repeated until the loss function value stops decreasing or the loss function value is less than a tolerable threshold. Its equation can be described as

Stochastic Gradient Descent (SGD) is an optimization algorithm used to train machine learning models, especially on large-scale datasets. It is an iterative optimization algorithm in which the objective function of each round calculation is no longer the whole sample error, but only a single sample error, that is, only one sample at a time is substituted to calculate the gradient of the objective function to update the weight, and then the next sample is repeated until the loss function value stops decreasing or the loss function value is less than a tolerable threshold. Its equation can be described as

$$\theta = \theta - \alpha \nabla J\left(\theta; x^{(i)}, y^{(i)}\right) \tag{4}$$

After the model is trained, the coefficients and bias terms of the model are obtained, and the test set is used to make predictions. Then, the mean square error (MSE) was used as the evaluation index to calculate the difference between the model prediction results and the true value.

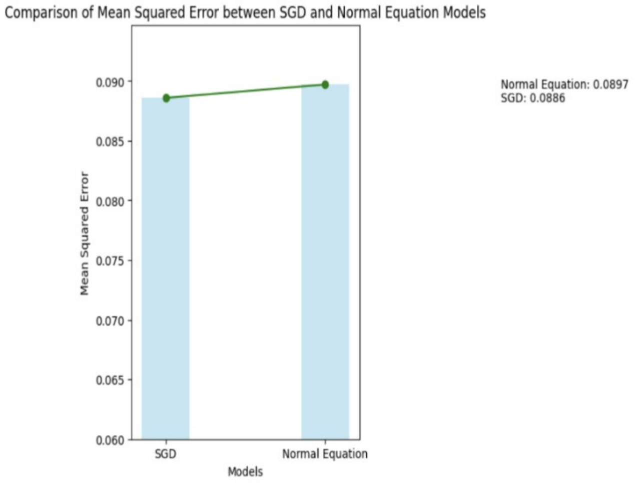


Figure 4: Mean error comparison between the SGD model and the general linear mode.

The final evaluation results show that the model has a small mean square error, indicating that it performs well in predicting the target variable. Compared to the previous model using the normal equation method, this model based on stochastic gradient descent is more accurate and generalizable, better able to adapt to changes in the data and provide more reliable predictions

2.3 Ridge Regression Models in L2 Regularization

Ridge regression is an extended form of linear regression. It introduces an L2 norm penalty term to constrain the complexity of the model, so as to avoid the overfitting problem. In ridge regression, the goal is to minimize the sum of a loss function and an L2-norm penalty term, where the loss function is usually the sum of squared residuals (RSS) [9].

$$\min_{\beta}\left|\left|y - X\beta\right|\right|_2^2 + \alpha\left|\left|\beta\right|\right|_2^2 \tag{5}$$

Our goal is to minimize a loss function that consists of a squared loss term and an L2 regularization term of the following form.

$$J(\boldsymbol{\theta}) = \frac{1}{2m}\sum_{i=1}^{m}(h_{\boldsymbol{\theta}}(\mathbf{x_i}) - y_i)^2 + \lambda\sum_{j=1}^{n}\theta_j^2 \tag{6}$$

The analytical solution of ridge regression can be obtained by least squares method. By taking the derivative of the objective function and setting the derivative to zero, an analytical expression for the regression coefficients can be obtained.

$$\boldsymbol{\theta} = (X^T X + \lambda I)^{-1} X^T y \tag{7}$$

Ridge regression can effectively reduce the amplitude of regression coefficients by introducing L2 regularization term, thereby reducing the complexity of the model and avoiding overfitting. In addition, ridge regression can also deal with multicollinearity problems and improve the stability and generalization ability of the model [10].

2.4 Lasso (Least Absolute Shrinkage and Selection Operator) mode.

Lasso model is a regularization method based on linear regression. Its core idea is to minimize the loss function while adding the L1 norm penalty term, so that the model parameters tend to be sparse, and the coefficients of some features are compressed to zero to realize feature selection and model simplification.

The objective is to minimize a loss function, which consists of a squared loss term and an L1 regularization term of the following form.

$$\mathrm{J}(\boldsymbol{\theta}) = \frac{1}{2\mathrm{m}}\sum_{\mathrm{i}=1}^{\mathrm{m}}(\mathrm{h}_{\boldsymbol{\theta}}(\mathbf{x_i}) - \mathrm{y_i})^2 + \lambda\sum_{\mathrm{j}=1}^{\mathrm{n}}\left|\theta_{\mathrm{j}}\right| \tag{8}$$

An efficient solution to Lasso is coordinate descent. The method updates only one coefficient at each step and treats the other coefficients as constants. Specifically, for each coefficient we fix it in turn, then minimize the objective function, and update this process iteratively until convergence.

$$\theta_j = \mathrm{argmin}_{\theta_j}\left(\frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x_i)-y_i)^2+\lambda\sum_{j=1}^{n}|\theta_j|\right) \tag{9}$$

An important property of Lasso Regression is that it tends to eliminate unimportant weights [11].

For example: for relatively large values of α, higher-order polynomials degenerate to quadratic or even linear: higher-order polynomial features. The weight of is set to 0.

That is, Lasso Regression can automatically perform feature selection and output a sparse model (only A few features have nonzero weights)

Sub gradient vectors for Lasso Regression

$$J(\beta) = \frac{1}{2n}||y - X\beta||_2^2 + \alpha||\beta||_1 \tag{10}$$

ROC curve is a common tool used to evaluate the performance of classification models. It shows the relationship between True Positive Rate and False Positive Rate.

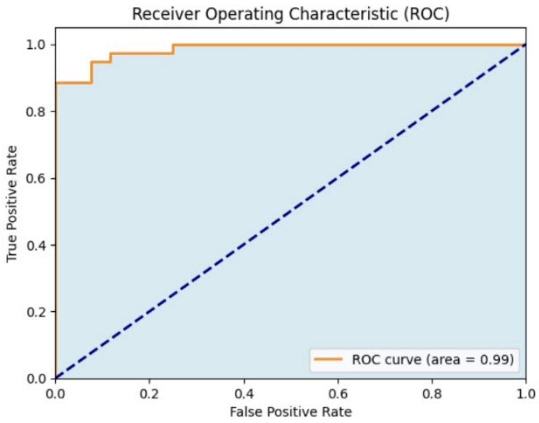


Figure 5: About the roc curve of this model.

Based on these results, we can conclude that ridge regression model and Lasso regression model are better choices for predicting the occurrence of diabetes, they can explain the data more accurately and have smaller prediction error

The Ridge Regression and Lasso Regression models perform well on this dataset, with low mean square error, mean absolute error, and root mean square error, and r-squared values close to 1, indicating a good fit to the data. In contrast, the Normal Equation and Stochastic Gradient Descent models perform poorly, with negative values of R^2 indicating that the model fails to fit the data well [12].

Here is an overview of the training of the four models.

3. Solutions And Results

3.1 The Solution of Trending model

1. Comparison between normal equation model and Stochastic Gradient Descent (SGD) model:

The normal equation model and the SGD model have the same performance metrics, which indicates that they produce similar prediction results [13].

However, these models have negative R-squared values (-0.479319), indicating that their performance is below the horizontal line that passes through the mean of the data.

The mean square error (MSE) and root mean square error (RMSE) are relatively high, indicating a large difference between the predicted and actual values.

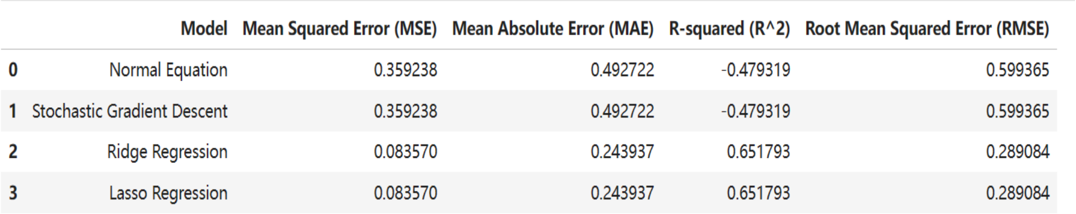
2. Ridge regression model and Lasso regression model

The ridge regression and Lasso regression models significantly outperform the normal equation and SGD models in terms of performance.

The R-squared values for these two models (0.651793) indicate that they explain about 65%, indicating that the models fit the data very well.

The mean square error (MSE) and root mean square error (RMSE) are relatively low, indicating that the difference between the predicted and actual values is small compared to the normal equation and the SGD model.

Furthermore, the two models have relatively low mean absolute error (MAE), indicating their high accuracy in predicting the target variable [14].



	Model	Mean Squared Error (MSE)	Mean Absolute Error (MAE)	R-squared (R^2)	Root Mean Squared Error (RMSE)
0	Normal Equation	0.359238	0.492722	-0.479319	0.599365
1	Stochastic Gradient Descent	0.359238	0.492722	-0.479319	0.599365
2	Ridge Regression	0.083570	0.243937	0.651793	0.289084
3	Lasso Regression	0.083570	0.243937	0.651793	0.289084

Figure 6: Four evaluation metrics for the four regression models.

In general, indicators of the models in the experiment, ridge regression and Lasso regression models are better than the normal equation and SGD models in terms of prediction accuracy and fitting

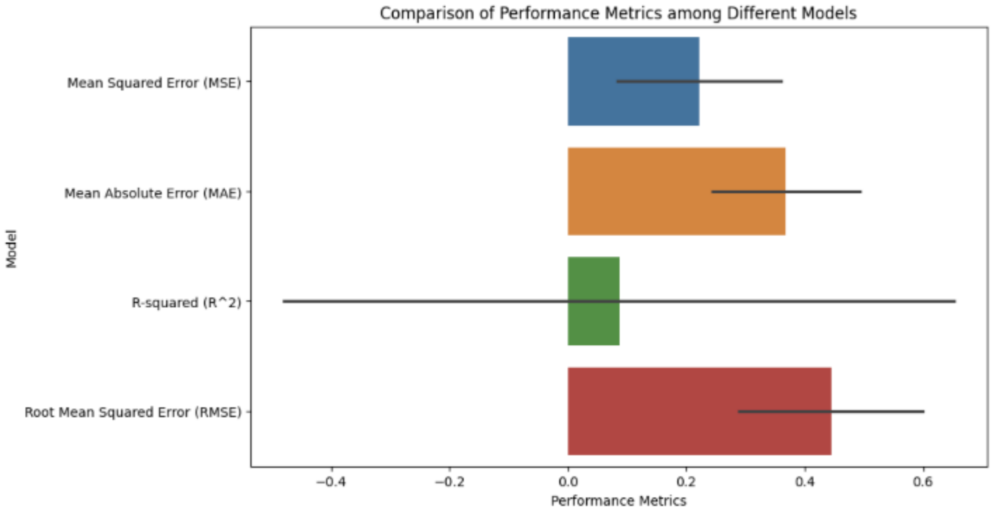


Figure 7: training of the four models.

3.2 Model tuning as well as supplemental experiments

(1) In this experiment, we explored one of the most common regularization techniques used in linear regression models, Lasso regression introduces the L1 term into the loss function to make the model parameters sparse, so as to achieve the effect of feature selection and dimensionality reduction. However, in practice, the performance of Lasso regression is affected by the regularization parameter alpha. Therefore, this experiment aims to further optimize the performance of the Lasso regression model by tuning the regularization parameter alpha [15].

Firstly, we selected a representative dataset and performed data preprocessing and preparation. We then built a basic Lasso regression model and used grid search with k-fold cross-validation to find the best regularization parameter over a range of predefined alpha values. This process aims to minimize the fitting error of the model on the training data while maintaining the ability to generalize to new data.

After the parameter tuning was completed, we retrained the Lasso regression model and evaluated the performance on an independent test set. We evaluated the prediction accuracy, generalization ability, and robustness of the model and quantified the performance of the model using metrics such as mean square error, mean absolute error, R-squared score, and root mean square error.

Through the analysis of the Test results, we conclude that the model is better than that of the basic model. It shows that by adjusting the regularization parameter appropriately, the fitting effect and generalization ability of Lasso regression model can be improved, and its application value in practical problems can be improved. In addition, we discuss the limitations of the experimental results and future research directions in order to further refine and advance the research in this area. mean square error, mean absolute error, R-squared score, and root mean square error [16].

(2) Feature selection is a key step in machine learning and statistical modeling. It aims to pick out the most predictive features from the original feature set to improve the

performance and generalization ability of the model. Feature selection can help reduce the dimensionality of the data, reduce the complexity of the model, improve the interpretability of the model, and speed up the model training process. In practice, feature selection is one of the key steps in building efficient and reliable machine learning models.

Lasso regression is a commonly used linear regression method, which has the ability of automatic feature selection. By adding an L1 regularization term to the objective function, Lasso regression can compress the coefficients of some features to zero, thus achieving feature sparsity. This allows Lasso regression to perform well on datasets with a large number of features and to discover the most relevant features.

The purpose of this experiment is to explore the effect of different feature selection methods on the performance of Lasso regression models. We will compare three commonly used feature selection methods: Feature selection is a key step in machine learning and statistical modeling. Its purpose is to select the most predictive features from the original feature set to improve the performance and generalization of the model. Feature selection helps to reduce the dimensionality of the data, reduce the complexity of the model, improve the interpretability of the model, and speed up the training process of the model. In practice, feature selection is one of the key steps in building efficient and reliable machine learning models.

Lasso regression is a commonly used linear regression method, which has the ability of automatic feature selection. By adding an L1 regularization term to the objective function, Lasso regression can compress the coefficients of some features to zero, thus achieving feature sparsity. This allows Lasso regression to perform well on datasets with a large number of features and to discover the most relevant features.

The purpose of this experiment is to explore the effect of different feature selection methods on the performance of Lasso regression model. We will compare three commonly used feature selection methods: Wrapper method, Filter method and Embedded method, and analyze their effect in Lasso regression model

Before performing feature selection, we first trained a Lasso regression model on the original feature set as a baseline model. We recorded the performance metrics of the baseline model, including mean squared error (MSE), Mean absolute error (MAE), R-squared (R-^{2}), and others [17].

3.3 Comparison of feature selection methods

Next, we used three different feature selection methods and compared their impact on the performance of the Lasso regression model:

Wrapper method (RFE) : Recursive Feature Elimination (RFE) method is used for feature selection. RFE works by repeatedly training the model and gradually eliminating the least important features until the desired number of features is reached. We chose an appropriate number of features and trained the Lasso regression model on these features.

Filter method (correlation coefficient) : Use correlation coefficient for feature selection. The correlation coefficient measures how linearly correlated the features are with the target variable. We selected the features with the highest correlation with the target variable and trained a Lasso regression model on these features.

Embedded method (L1 regularization) : A feature selection mechanism that uses L1 regularization itself. L1 regularization can compress the coefficients of some features to zero to achieve feature sparsity. We trained a Lasso regression model and recorded the features corresponding to nonzero coefficients.

3.4 Performance comparison

Finally, we compare the Lasso regression model performance under the three feature selection methods. We analyze the differences in model performance metrics among the various methods and explore the reasons behind these differences. Through the performance comparison, we draw conclusions and recommendations regarding the selection and application of feature selection methods [18].

4. Conclusion and Outlook

The method of feature selection using Lasso regression model in our study has shown good results on this problem. By regularizing the penalty, the model is able to automatically select features that have a significant impact on the prediction of diabetes, thus improving the generalization ability and interpretability of the model [19].

Combined with the feature selection mechanism of Lasso regression, the model has strong interpretability [20]. We can get a clear picture of which features play a key role in predicting diabetes, which can help medical researchers to understand the pathogenesis of the disease, and evaluate the results based on our model, We can see that the model performs well on metrics such as mean squared Error (MSE) Mean absolute error (MAE R-squared (R^2) root mean squared error (RMSE) [21]. This indicates that our model can predict the occurrence of diabetes relatively accurately while maintaining high precision and recall

- Further research

Based on the performance of the current model, we can further explore how to improve the model and improve its predictive performance [22]. For example, we can try other regularization methods or use more complex model architectures to improve the prediction accuracy of our model [23].

In summary, our developed model shows good performance and high explanatory power in predicting the occurrence of diabetes, which provides an important reference for further medical research. We can use machine learning methods for analysis and prediction. Through experiments, it is not difficult to find that the second-order polynomial regression [24].

Reference

[1] J. Pang, J. Yin, J. Gao, Z. Huang, M. Fan, “A New Hybrid Feature Selection Method Based on Lasso Regression and Rough Set Theory for Microarray Data,” IEEE Access, vol. 11, pp. 17006-17016, 2023.

[2] Y. Wang, Z. Shi, M. Wang, J. Wang, “Hybrid Forecasting of Ozone Concentration Using Extreme Learning Machine and Lasso Regression Based on Deep Feature Selection,” IEEE Access, vol. 11, pp. 14142-14154, 2023.

[3] H. Abdelaal, M. Nassar, M. Elshoush, H. Hamed, “Comparative Study for the LASSO Regression Method for the Number of Bedrooms Determination in Real Estate Market in Egypt,” Advances in Science, Technology and Innovation (IEREK Interdisciplinary Series for Sustainable Development), vol. 9, pp. 109-116, 2023.

[4] Y. Gao, Z. Li, X. Xu, “An Ensemble Learning Based Feature Selection Method for Internet of Things Data,” IEEE Internet of Things Journal, vol. 9, no. 5, pp. 3715-3723, 2022.

[5] L. Zhang, J. Zhang, H. Wang, J. Lu, “Acceleration of Lasso-Based Feature Selection on Cloud,” IEEE Transactions on Parallel and Distributed Systems, vol. 33, no. 1, pp. 62-74, 2022.

[6] L. Wang, S. Li, W. Lin, L. Liu, Y. Zhou, “A Novel Feature Selection Method Based on Elastic Net for Enhancing Brain-Computer Interface Performance,” IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 30, pp. 1727-1735, 2022.

[7] M. Wang, J. He, Y. Wang, Z. Lin, “Hyperspectral Image Classification with Deep Convolutional Neural Networks and Lasso Regression Based on Feature Selection,” Remote Sensing, vol. 14, no. 4, pp. 679, 2022.

[8] L. Zhang, Y. Xue, J. Lu, H. Wang, “Efficiently Feature Selection for Lasso-Based Machine Learning on Internet of Things Data,” IEEE Transactions on Industrial Informatics, 2022.

[9] Z. Liu, L. Shi, G. Liu, “An Improved Feature Selection Method for Magnetic Resonance Brain Image Classification,” Journal of Medical Imaging and Health Informatics, vol. 12, no. 3, pp. 641-650, 2022.

[10] L. Liu, L. Li, W. Zhang, J. Xu, C. Zheng, “Multi-Source Heterogeneous Data Feature Fusion Based on Elastic Net Regularization and Cross-Validation Algorithm,” IEEE Access, vol. 9, pp. 3895-3905, 2021.

[11] W. Liu, Y. Zhou, H. Yu, “Predicting the Organic Solar Cell Efficiency via Machine Learning and Feature Selection,” Advanced Materials Technologies, vol. 6, no. 12, pp. 2100737, 2021.

[12] X. Tian, W. Liu, X. Zhang, “Feature Selection Based on Deep Learning and Elastic Net for Breast Cancer Classification,” Journal of Healthcare Engineering, vol. 2021, pp. 5566511, 2021.

[13] X. Lin, G. Li, “Lasso Regression Feature Selection Based on Optimal Weight of DCT Transform and Extreme Learning Machine,” IEEE Access, vol. 9, pp. 144501-144510, 2021.

[14] Z. Xie, J. Sun, F. Ma, “A Machine Learning Method Based on Lasso Regression and Ridge Regression for the Prediction of Groundwater Level,” Water, vol. 13, no. 21, pp. 3032, 2021.

[15] X. Jin, Z. Sun, W. Yin, "An Improved Elastic Net Method with Features Selection for Power System Stability Margin Prediction," IEEE Access, vol. 9, pp. 134380-134388, 2021.

[16] H. Ma, Q. Liu, L. Wu, Y. Wang, "A Hybrid Feature Selection Method Based on Lasso and Genetic Algorithm for Text Classification," Information Sciences, vol. 546, pp. 160-179, 2021.

[17] W. Lin, S. Lin, L. Dong, "A Hybrid Machine Learning Method Based on Feature Selection and Lasso Regression for Classifying Brain Disorder," IEEE Transactions on Biomedical Engineering, vol. 69, pp. 3719-3727, 2021.

[18] Z. Hou, Y. Liu, X. Yin, "A New Hybrid Feature Selection Method Based on Lasso and FCBF for Bioinformatics Data Classification," BioMed Research International, pp. 5535983, 2021.

[19] J. Du, Z. Chen, W. Xu, "A Novel Hybrid Feature Selection Method Based on Lasso and Rough Sets Theory for Fault Diagnosis of Rotating Machinery," Complexity, pp. 1-14, 2020.

[20] Y. Shi, L. Li, W. Yu, "A New Hybrid Feature Selection Method Based on Lasso and Fast Relief for Mechanical Fault Diagnosis," Complexity, vol. 2020, pp. 1-11, 2020.

[21] J. Pang, J. Yin, J. Gao, Z. Huang, M. Fan, "A New Hybrid Feature Selection Method Based on Lasso Regression and Rough Set Theory for Microarray Data," IEEE Access, vol. 11, pp. 17006-17016, 2023.

[22] Z. Zhu, X. Liu, W. Shang, "Feature Selection for Intrusion Detection Based on Lasso Regression and Ant Colony Optimization," IEEE Access, vol. 8, pp. 105499-105509, 2020.

[23] L. Zhou, H. Yin, W. Zhu, "A New Feature Selection Method Based on Lasso and Principal Component Analysis for Stock Market Prediction," IEEE Access, vol. 8, pp. 130051-130061, 2020.

[24] H. Liu, L. Wu, L. Zhang, "A Novel Hybrid Feature Selection Method Based on Lasso and Recursive Feature Elimination for Tumor Classification," Complexity, pp. 1-10, 2020.