ISSN: 2664-9640

Research on diabetes risk prediction based on ensemble learning

Huazhong Yang, Hualu Chen*, Hualian Yang, Huajian Yang and Deyuan Ren

Yangtze University, Jingzhou 434000, China.

* Corresponding Author

Abstract

To improve the prediction accuracy of early diabetes risk, a diabetes risk prediction model was established based on an ensemble learning algorithm. A diabetes prediction model was established based on the ensemble learning algorithms Random Forest, LightGBM, AdaBoost, CatBoost, and XGBoost respectively, and the classification and prediction performance of the above five algorithms were compared. Validation experiments were carried out on the Pima Indians diabetes dataset in UCI. By smoothing and standardizing the data, a model was constructed to predict the risk of diabetes, and five indicators of accuracy, precision, recall, F1 score, and AUC were used to evaluate Model performance and rank variable importance based on the best performing model. Among all models, the LightGBM algorithm based on ensemble learning has the highest accuracy, precision, recall, F1 value, and AUC value, reaching 96.62%, 94.73%, 98.63%, 96.64%, and 99.53%, respectively, which is higher than the prediction accuracy of AdaBoost. 0.68%, which is 1.35% higher than the prediction accuracy of CatBoost, XGBoost, and Random Forest algorithms. In the ranking of variable importance given by the LightGBM algorithm, the ranking of influence weights is Blouse, Insulin, BMI, SkinThickness, Age, BloodPressure, DiabetesPedigreeFunction, and Pregnancies. Based on LightGBM and AdaBoost, these two ensemble learning algorithms show good performance in predicting the risk of diabetes. Compared with CatBoost, XGBoost, and random forest classifiers, it can more accurately identify early high-risk patients, which is helpful for clinicians to conduct more accurate diagnoses and treatments. medical decisions.

Keywords

Diabetes;Risk prediction; Integrated learning; LightGBM; AdaBoost.

1. Introduction

Diabetes mellitus (DM) is a metabolic disease characterized by hyperglycemia, and its prevalence is increasing[1]. Diabetes causes many problems because it causes other diseases such as coronary heart disease, cardiovascular and cerebrovascular diseases, etc., and eventually becomes the leading cause of death (directly or indirectly). Therefore, early and regular screening for diabetes risk factors, including Glucose levels, BloodPressure status, BMI, and insulin levels, is important, emphasizing the need for early diagnosis and treatment. In addition, diabetes prevention and management are equally important. Once the risk of developing diabetes is recognized, early intervention can help slow the progression of diabetes. In this case, informing patients of their diabetes risk and providing them with appropriate lifestyle changes would be more effective than other treatments[2].

With the advancement of algorithms and the substantial reduction of data storage costs, many machine learning and data mining techniques have been widely used in the medical field[3]. Data mining technology has become an essential tool in medical fields such as disease diagnosis, cancer prediction, auxiliary diagnosis and treatment, drug mining[4], hospital information

ISSN: 2664-9640

systems, and biomedicine. Data mining technology conducts data analysis from a large amount of unstructured medical data, extracts the hidden knowledge of diseases, and finally draws conclusions from the analysis. Therefore, predicting the risk of DM through data mining technology can not only save money, but also a new research direction in the future[5].

Like many people undergo general routine health screenings provided by the government or based on their needs, this study aimed to develop applicable DM prediction models by conducting validation experiments on UCI's Pima Indian diabetes dataset[6]. The data was fully utilized to try to implement various DM prediction models. We compare the advantages and disadvantages of each model, analyze the effectiveness of the model to predict diabetes, extract the important factors affecting diabetes, and inform people to stay away from the mediators that trigger diabetes and improve the quality of life[7]. In this paper, integrated learning is used to predict the risk of developing diabetes. The rest of the paper is presented as follows: Section 2 discusses existing work relevant to the prediction of diabetes and its diagnosis. The experimental approach to the study is described in Section 3. Section 4 describes the experimental model and its. Section 5 discusses the experimental findings, and Section 6 summarizes and research strengths and concludes the study with guidelines for future work.

2. Related work

Machine learning techniques are the study of how computers can simulate or implement human learning behaviors to acquire new knowledge or skills and continuously improve their performance, and are currently used by many scholars to predict the risk of developing diabetes[8],[9]. henock M. Deberneh used LR, RF, XGBoost, SVM, CIM, Stacked Classifier (ST), and Soft Voting (SV) algorithms to predict The results showed that the integrated classifier approach (CIM, ST and SV) was the best predictor of the risk of diabetes[10]; An Dinh et al. used the XGBoost algorithm to predict the risk of developing diabetes and showed that the AU-ROC (receiver operating characteristic) score for XGBoost prediction was 86.2%[11]; Raja Krishnamoorthi et al. proposed the use of machine learning to develop a unique intelligent diabetes prediction framework (IDMPF) to predict the risk of developing diabetes, and the results showed that the proposed model was much better than the single model decision tree (DT) with random forest (RF) and support vector machine (SVM)[12]; Jun Li et al. used Stacking model and ResNet50 model for diabetes risk prediction. demonstrated that the differential changes in tongue signs reflect abnormalities in glucose metabolism, and therefore the combination of TCM tongue diagnosis and machine learning techniques to form a diabetes risk prediction model is feasible[13]; Leon Kopitar used machine learning prediction models (i.e., Glmnet, RF, XGBoost, LightGBM) to compare with commonly used regression models for predicting undiagnosed T2DM. The results showed the lowest mean RMSE of 0.838, followed by RF (0.842), LightGBM (0.846), Glmnet (0.859), and XGBoost (0.881)[14]; Sanjay Basu et al. used machine learning techniques to identify patterns in large datasets to predict outcomes or classify patient characteristics, and the results showed that machine learning methods based on machine learning methods with tree learners (generating decision trees to help guide clinical interventions) generally have higher sensitivity and specificity than traditional risk prediction regression models[15]; Satish Kumar Kalagotla et al. used novel stacking techniques to predict diabetes, and experimental results showed that the novel stacking techniques (MLP, SVM, and LR models, respectively) achieved an accuracy of 78.2%, which is better than other models[16].

As mentioned above, many machine learning algorithms can not only be applied to disease diagnosis but also get good results by building and combining multiple base learners to accomplish the learning task[17]. In this paper, we use the latest integrated learning algorithms

ISSN: 2664-9640

in machine learning to predict the risk of developing diabetes and analyze the experimental results to provide a new way of thinking and approach for diabetes risk prediction.

3. Methodology

The flowchart of the experiment is shown in Figure 1. python Jupyter NoteBook was used to implement the entire experiment. different packages such as NumPy, pandas, scikit and Matplotlib were used to analyze the data. The tasks performed at each stage and the associated functionality explored from the Python toolkit are described below.

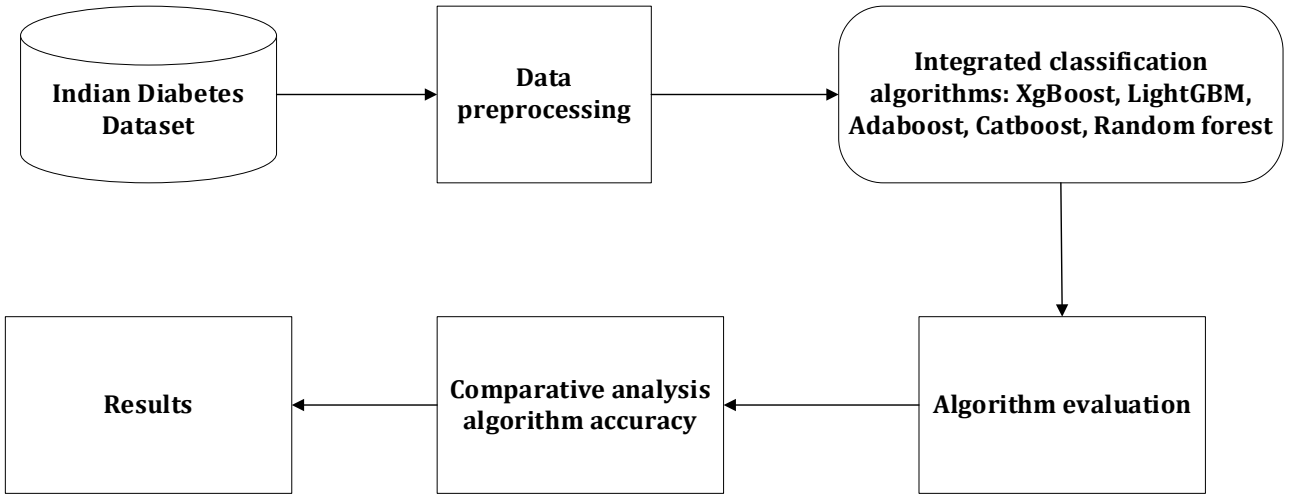

Figure 1: Project flow chart

3.1. Datasets

The dataset for this paper is from the Pima Indian Diabetes Database, a dataset commonly used for the prediction of diabetes. The dataset consists of 768 rows and 9 columns with 500 samples of healthy and 268 samples of unhealthy, as shown in Table 1 , with OutCome denoted by Y and explanatory variables denoted by $X_i(1,2\ldots,8)$ in order, and columns containing attributes such as glucose, pregnancy, skin thickness, blood pressure, BMI, insulin, age and outcome that predict whether the outcome is healthy or diseased[18]. And functions such as Numpy are used to process the dataset.

Table 1: Variable assignment

Variables	Feature Description
OutCome(Y,Class)	Healthy = 0, Sick = 1
Age(X_1, Age)	$21 \leq Age \leq 81$
Pregnancies(X_2, Pregnancies)	Number of pregnancies
Glucose(X_3, Glucose)	Glucose content
BloodPressure(X_4, BloodPressure)	Blood pressure value
SkinThickness(X_5, SkinThickness)	Skin Thickness Value
Insulin(X_6, Insulin)	Insulin Levels
BMI(X_7, BMI)	Weight Index
DiabetesPedigreeFunction(X_8, DiabetesPedigreeFunction)	Diabetes genetic function coefficient

3.2. Data Visualization

Data visualization helps to understand data more intuitively by placing it in a visual form. At this stage, the data is represented in the form of a bar graph[19]. The analysis revealed missing

ISSN: 2664-9640

values and feature correlations in the data. It also shows information about datasets such as Glouse, insulin, BMI and blood pressure. Among other things, it predicts how many people are affected by diabetes from 768 data items. To display the output, graphical representations such as plot axes, pyplot, etc. are used. In this paper, the visualization of data missing values is shown in Figure 2; the eigenvalue correlation is shown in Figure 3.

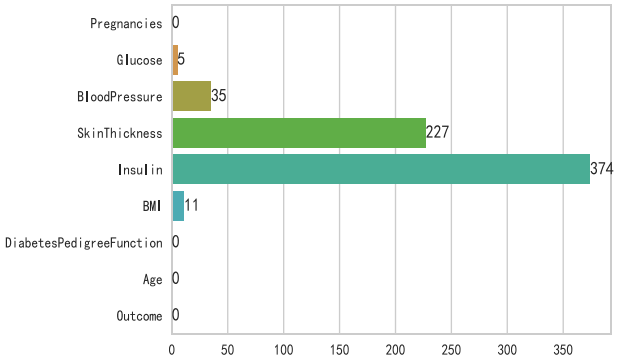

Figure 2: Missing Data Chart

From Figure 2, it can be observed that in 768 data, 5 out of 9 eigenvalues are missing, namely Glucose, BloodPressure, SkinThickness, Insuln, BMI. missing data are 5, 35, 227, 374, and 11 feature values respectively, and we use the mean value to fill the missing values.

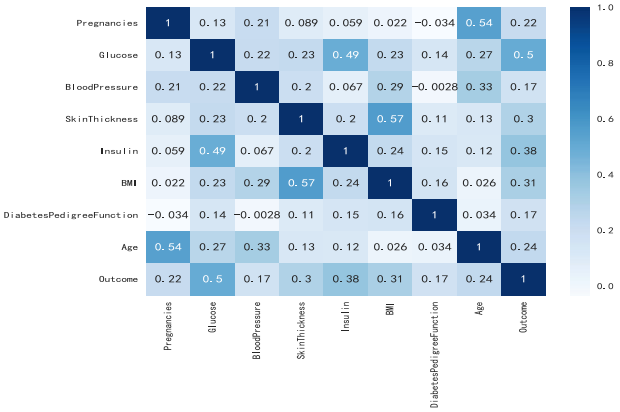

Figure 3: Correlation of eigenvalues

Analyzing the correlation coefficient of each eigenvalue, the darker the color, the stronger the correlation, from Figure 3 we can see the correlation between each of the eigenvalues.

3.3. Preprocessing

Pre-processing includes removing outliers, outliers, and normalized data[20]. The processed data is used to create the model. Before applying a classifier to the data, the data should be properly pre-processed. In this paper, we use the mean fill method for missing data, the 3σ principle[21] for outlier data, Min-Max for normalized[22] data, and the SMOTEENN technique for smoothing the data process to obtain more accurate results. The dataset contains missing values and the algorithm requires that the feature values should not have null values. Then, we normalize all the values by scaling the dataset.

3.4. Min-Max Standardization

For the sequence $\{x_1, x_2, \ldots\ldots, x_n\}$ the transformations are carried out.

$$y_i = \frac{x_i - \min\limits_{1\le j\le n}\{x_j\}}{\max\limits_{1\le j\le n}\{x_j\} - \min\limits_{1\le j\le n}\{x_j\}}$$

Then new sequence $\{y_1, y_{2,\ldots,} y_n \in [0,1]\}$ and is dimensionless.

ISSN: 2664-9640

3.5. SMOTEENN Technology

SMOTEENN Developed by Batista et al (2004), this method combines the SMOTE ability to generate synthetic examples for minority classes and ENN ability to delete some observations from both classes that are identified as having different classes between the observation's class and its K-nearest neighbor majority class. The process of SMOTE-ENN can be explained as follows.

3.6. Ensemble Learning Classification Algorithms

After preprocessing the data, the integrated learning classifier in the scikit-learn Python toolkit is used. scikit is a simple toolkit for processing and analyzing numbers. The data set is first split into a training dataset and a test dataset using model selection training test splitting. Due to the limited source of the dataset, about 90% of the dataset is used for training purposes and the remaining 10% is used for testing with randomly selected data. Then, the training was performed in XGBoost, LightGBM, Adaboost, Catboost, and Random Forest integrated learning classification algorithms, respectively, and a test set was used to test the classification algorithm classification effects and compare the advantages and disadvantages between different algorithms.

3.7. Performance Evaluation

The evaluation metrics of classification models used in this paper are Accuracy, Precision, Recall, F1-Score, and AUC: Accuracy is a classification evaluation model metric that refers to the proportion of results (both positive and negative cases) that are correctly predicted by the model. Namely：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

The checking rate is the proportion of samples predicted to be in the positive class that belongs to the positive class. Namely：

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

The check-completion rate is the proportion of all positive category samples that are correctly identified as positive categories. Namely：

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

From the above equation (2)(3), we can see that Precision, Recall will exist in a certain contradiction, there will not be Precision, Recall at the same time with high accuracy, so F1-Score takes into account these two indicators, Namely:

$$\mathrm{F1-Score} = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

The AUC metric represents the area under each ROC curve and takes a value between 0.5 and 1. The horizontal axis of the ROC curve represents the probability of the wrong classification of negative cases and the vertical axis represents the probability of the right classification of positive cases.

4. Machine Learning Classification Models

4.1. XGBoost

The XGBoost algorithm[23] is an optimized distributed gradient boosting library that uses decision trees as the base classifier, and the new function formed by the new trees is used to fit

ISSN: 2664-9640

the residuals of the previous predictions, and the accumulated results of all trees are added to obtain the final prediction. the objective function of XGBoost is as follows：

$$minL = \sum_{i=1}^{n} l(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k) \tag{1}$$

$$\Omega(f_k) = \Upsilon \mathrm{T} + \frac{1}{2}\Upsilon \sum_{j=1}^{T} W_j^2 \tag{2}$$

Where n is the number of training samples and k is the number of decision trees f_k is the base learner. The loss function l is used to measure the difference between the true score and the predicted score. The regularization term Ω contains two components, where T denotes the number of leaf nodes and W denotes the leaf node fraction; ϒ and λ denote the penalty strength, which controls the number of leaf nodes and limits the node fraction to prevent the model from overfitting and losing the prediction effectiveness.

4.2. LightGBM

LightGBM[24] is an improved decision tree algorithm based on decision trees developed by Microsoft in 2017. More powerful than XGBoost, it is faster, takes up less memory, has better accuracy, and supports parallelized computation. The main features of the LightGBM algorithm are a decision tree algorithm with histogram, one-sided gradient sampling, and mutual exclusion feature bundle, and a leaf-wise leaf growth strategy with depth limitation, LightGBM also directly supports category features, efficient parallelization, optimized cache hit rate, and differential acceleration with histograms. The loss function of LightGBM is as follows:

$$Obj^{(t)} = \sum_{i=1}^{n} L\left(y_i, \hat{y}^{(t-1)} + f_t(x_i)\right) + \sum_{i=1}^{t} \Omega(f_i) + constant \tag{1}$$

$$G = \frac{1}{2}\left[\frac{{G_L}^2}{H_l + \lambda} + \frac{{G_R}^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] \tag{2}$$

y_i is the target value, i is the predicted value,t denotes the number of the leaf nodes,q denotes the structure-function of the tree, w is the leaf weight, and n is the number of samples.

4.3. Adaboost

AdaBoost is an algorithm that iteratively adds a new weak classifier in each round until a predetermined small enough error rate is reached[25][26]. It improves the classification ability of the data through continuous training, has a high detection rate, and is less prone to overfitting. In general, AdaBoost, which uses decision trees as weak learners, is often called the best classifier.To prevent Adaboost overfitting, a regularization term is usually added, which we call the learning rate, defined as v. The iterative mathematical formula for the previous weak learner is as follows:

$$f_m(x) = f_{m-1}(x) + \alpha_m G_m(x) \tag{1}$$

when we add a regularization term, then :

$$f_m(x) = f_{m-1}(x) + v\alpha_m G_m(x) \tag{2}$$

The range of v is $0 \leq v \leq 1$. For the same learning effect on the training set, a lower v means more iterations for the slow learner are required. The number of steps and the maximum number of iterations is usually used together to determine the fit of the algorithm.

ISSN: 2664-9640

4.4. Random Forest

The RF Algorithm is based on building Bagging Integration with decision trees as the base learner, and further incorporating the selection of random attributes in the training process of decision trees[27]. As a supervised learning algorithm, RF can avoid some drawbacks of single classification prediction and obtain higher classification prediction accuracy. It is a machine learning algorithm that integrates multiple decision trees through integration learning, which has better performance than individual decision trees. rf has randomness in sample and feature selection, and the introduction of these two randomness makes it less likely to fall into overfitting and has good noise immunity.

4.5. CatBoost

CatBoost is a GBDT framework with fewer parameters, support for category-based variables, and high accuracy implemented based on a symmetric decision tree-based learner, which mainly addresses the efficient and reasonable processing of category-based features, CatBoost is composed of Categorical and Boosting[28]. In addition, CatBoost also solves the problems of Gradient Bias and Prediction shift to reduce the occurrence of overfitting and thus improve the accuracy and generalization ability of the algorithm. Compared with XGBoost and LightGBM, CatBoost automatically processes categorical features into numerical features, CatBoost automatically processes categorical features into numerical features and also uses combined categorical features, which greatly enriches the feature dimensions.Compared with XGBoost and LightGBM, CatBoost automatically transforms categorical features into numerical features and also uses combined category features, which greatly enriches the feature aspects.

5. Results

5.1. Prediction accuracy results

PIDD data set consists of 768 patients of which 268 patients were affected with diabetes and 500 patients are nondiabetic. Figure 4 represents a graph comparing the results of diabetes risk prediction using the integrated learning algorithms in 5.

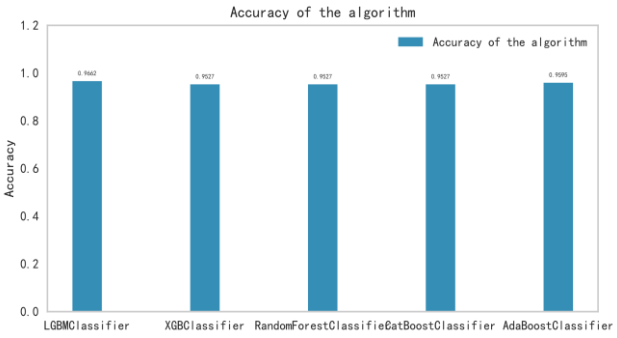

Figure 4: Algorithm prediction result chart

As can be seen in Figure 4, among the five used algorithms, the accuracy of LightGBM reached 96.62%, which is the most accurate among the five algorithms, and the accuracy of the remaining four differed not much, and the contrast experiment illustrated that the use of LightGBM was the best for diabetes risk prediction and provided an optimal prediction algorithm for diabetes risk prediction.

5.2. Algorithm evaluation

By comparing the prediction results of the 5 integrated learning algorithms for diabetes risk, it can be seen that using the LightGBM algorithm is the best for diabetes risk prediction, although the AUC value of LightGBM is not the largest, the remaining 4 evaluation index values are the

ISSN: 2664-9640

largest among the 5 algorithms, so LightGBM is the best for diabetes risk prediction. The results of the integrated learning algorithm for predicting diabetes are shown in Table 2.

Table 2: Algorithm evaluation

Algorithm name	Accuracy	Precision	Recall	F1-Score	AUC
LightGBM	0.96621	0.94737	0.98630	0.96644	0.99525
AdaBoost	0.95946	0.93506	0.98630	0.96000	0.99178
XGBoost	0.95270	0.92307	0.98630	0.95364	0.98886
RandomForest	0.95270	0.93421	0.97260	0.95302	0.99159
CatBoost	0.95270	0.92308	0.98630	0.95364	0.99616

5.3. The AUC curves of 5 algorithms

The AUC curve shows the effect of algorithm prediction, and the larger its value, the better the prediction effect, among the 5 algorithms used in this paper, the AUC value of the CatBoost algorithm is the largest, which is 0.99616, followed by the LightGBM algorithm, whose value is 0.99525, indicating that these 2 algorithms have the best effect in predicting diabetes risk, considering that the other 4 values of the LightGBM values are the highest, so LightGBM is the best algorithm for predicting diabetes risk, and the ROC curves of the 5 algorithms for predicting diabetes mellitus risk are shown in Figure 5.

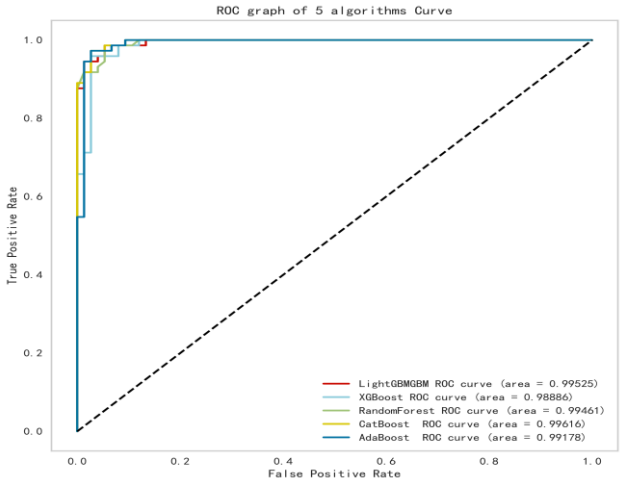

Figure 5: ROC curves of the 5 algorithms

5.4. Significant Eigenvalues

By using LightGBM to predict the risk of developing diabetes, we can derive the eigenvalues that affect the prediction results, the eigenvalues that affect the risk of developing diabetes are ranked as Glucose, Insulin, BMI, SkinThickness, Age, BloodPressure, DiabetesPedigreeFunction, Pregnancies. Pregnancies. observation can be concluded that Glucose, Insulin, and BMI ranked in the top three, indicating that the level of blood glucose has a greater impact on the risk of diabetes, if suffering from diabetes, in life, we should reduce the intake of sugary foods, likewise, Insulin and BMI have a greater impact on the risk of developing diabetes, so we should in daily life Strengthen exercise and improve physical fitness. The graph of important characteristics affecting diabetes is shown in Figure 6.

ISSN: 2664-9640

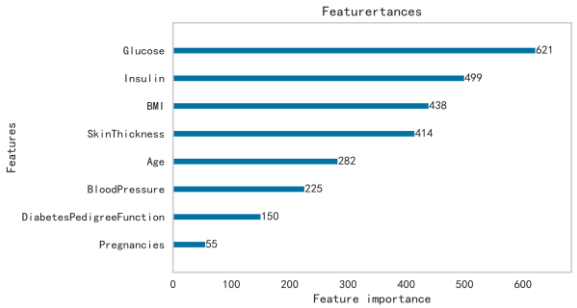

Figure 6: Important Features Chart

6. Summary and Conclusion

Machine learning techniques are valuable in diagnosing diseases. Early diagnosis is beneficial for patients to seek medical treatment early, improve the diagnosis rate, and help patients' follow-up treatment. In this paper, we used the UCI database to build diabetes disease risk prediction models based on integrated learning algorithms RF, LightGBM, CatBoost, AdaBoost, and XGBoost, and the results showed that the LightGBM algorithm had the highest accuracy, and precision The results showed that the LightGBM algorithm had the highest accuracy, recall, and F1 values, reaching 96.621%, 94.737%, 98.630%, and 96.644%, respectively. Patients who consume large amounts of water lead to a feeling of dry mouth in the brain. Patients who drink a lot of water lead to an increase in systemic blood flow and an increase in renal perfusion pressure, which leads to an increase in urine output. In addition, diabetic patients have insufficient insulin secretion or insulin resistance, and the cells of the body cannot use blood glucose normally for energy supply, which makes the patients feel hungry and prompt them to eat continuously to ensure energy supply, and the body will break down fat and protein for energy supply, so the patients lose weight. In conclusion, the early diabetes risk prediction model constructed by the integrated learning algorithm in this paper can classify potential patients with diabetes more accurately. In the future, a larger data set and more accurate feature classification algorithm can be used for diabetes risk prediction to help clinicians identify early diabetic patients, reduce the occurrence of diabetes-related complications, improve the quality of life of patients, and reduce the burden on society.

Data Availability

The data that support the findings of this study are available on request from the author.

Conflicts of Interest

The authors of this manuscript declare that they do not have any conflicts of interest.

References

[1] Hathaway, Quincy A et al. “Machine-learning to stratify diabetic patients using novel cardiac biomarkers and integrative genomics.” Cardiovascular diabetology vol. 18,1 78. 11 Jun. 2019.

[2] Bilandzic, Anja, and Laura Rosella. “The cost of diabetes in Canada over 10 years: applying attributable health care costs to a diabetes incidence prediction model.” “Les coûts du diabète sur 10 ans au Canada : intégration des coûts en soins de santé imputables au diabète à un modèle de prédiction de son incidence.” Health promotion and chronic disease prevention in Canada : research, policy and practice vol. 37,2 (2017): 49-53.

[3] Choi, Rene Y et al. “Introduction to Machine Learning, Neural Networks, and Deep Learning.” Translational vision science & technology vol. 9,2 14. 27 Feb. 2020.

[4] Patel, Lauv et al. “Machine Learning Methods in Drug Discovery.” Molecules (Basel, Switzerland) vol. 25,22 5277. 12 Nov. 2020.

[5] Dagliati, Arianna et al. “Machine Learning Methods to Predict Diabetes Complications.” Journal of diabetes science and technology vol. 12,2 (2018): 295-302.

[6] Zou, Quan et al. “Predicting Diabetes Mellitus With Machine Learning Techniques.” Frontiers in genetics vol. 9 515. 6 Nov. 2018.

[7] Xie, Zidian et al. “Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques.” Preventing chronic disease vol. 16 E130. 19 Sep. 2019.

[8] Information on: https://www.mathworks.com/discovery/machine-learning

[9] Lai, Hang et al. “Predictive models for diabetes mellitus using machine learning techniques.” BMC endocrine disorders vol. 19,1 101. 15 Oct. 2019.

[10] Deberneh, Henock M, and Intaek Kim. “Prediction of Type 2 Diabetes Based on Machine Learning Algorithm.” International journal of environmental research and public health vol. 18,6 3317. 23 Mar. 2021.

[11] Dinh, An et al. “A data-driven approach to predicting diabetes and cardiovascular disease with machine learning.” BMC medical informatics and decision making vol. 19,1 211. 6 Nov. 2019.

[12] Krishnamoorthi, Raja et al. “A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques.” Journal of healthcare engineering vol. 2022 1684017. 11 Jan. 2022.

[13] Li, Jun et al. “Establishment of noninvasive diabetes risk prediction model based on tongue features and machine learning techniques.” International journal of medical informatics vol. 149 (2021): 104429.

[14] Kopitar, Leon et al. “Early detection of type 2 diabetes mellitus using machine learning-based prediction models.” Scientific reports vol. 10,1 11981. 20 Jul. 2020.

[15] Basu, Sanjay et al. “Use of Machine Learning Approaches in Clinical Epidemiological Research of Diabetes.” Current diabetes reports vol. 20,12 80. 3 Dec. 2020.

[16] Kalagotla, Satish Kumar et al. “A novel stacking technique for prediction of diabetes.” Computers in biology and medicine vol. 135 (2021): 104554.

[17] Rhee, Sang Youl et al. “Development and Validation of a Deep Learning Based Diabetes Prediction System Using a Nationwide Population-Based Cohort.” Diabetes & metabolism journal vol. 45,4 (2021): 515-525.

[18] Howlader, Koushik Chandra et al. “Machine learning models for classification and identification of significant attributes to detect type 2 diabetes.” Health information science and systems vol. 10,1 2. 9 Feb. 2022.

[19] Birnbaum, David. “Regarding data visualization.” Infection control and hospital epidemiology vol. 42,9 (2021): 1154-1155.

[20] Malley, Brian, et al. “Data Pre-processing.” Secondary Analysis of Electronic Health Records, edited by MIT Critical Data, Springer, 10 September 2016.pp. 115–141.

[21] Pukelsheim, Friedrich. "The three sigma rule." The American Statistician 48.2 (1994): 88-91.

[22] Patro, S., and Kishore Kumar Sahu. "Normalization: A preprocessing stage." arXiv preprint arXiv:1503.06462 (2015).

[23] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.

[24] Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." Advances in neural information processing systems 30 (2017).

[25] Schapire, Robert E. "Explaining adaboost." Empirical inference. Springer, Berlin, Heidelberg, 2013. 37-52.

[26] Ying, Cao, et al. "Advance and prospects of AdaBoost algorithm." Acta Automatica Sinica 39.6 (2013): 745-758.

[27] A. Saffari, C. Leistner, J. Santner, M. Godec and H. Bischof, "On-line Random Forests," 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, 2009, pp. 1393-1400.

[28] Prokhorenkova, L. , et al. "CatBoost: unbiased boosting with categorical features." 2017.