Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/46/20241071

Machine learning-based readmission risk prediction for diabetic patients

Chenyang Li

Maynooth University International Engineering College, Fuzhou University, Fuzhou, 350108, China

LI.LI.2022@MUMAIL.IE

Abstract. Among the numerous hospitalized patients with chronic diseases, diabetes patients are under a higher readmission rate, which poses challenges and pressures to both patients and the healthcare system. To predict the likelihood of diabetic patients being readmitted within a short amount of time, this paper utilizes various machine learning-based models for performance analysis and comparison. By selecting appropriate datasets, cleaning and preprocessing data, the models were trained to forecast the probability of patient readmission. The paper compares the performance metrics of six classifiers: XGBoost, logistic regression, GBDT, decision tree, random forest, and deep neural network. The metrics include accuracy, f1 score, precision, recall, and ROC curve. Experimental results demonstrate that XGBoost exhibits better adaptability to complex data and achieves higher mean values of Accuracy (64.43%), f1 score (59.16%), recall (55.9%), and ROC (70.14%) in readmission prediction.

Keywords: Machine learning, Diabetic Readmissions, Deep Neural Networks, XGboost.

1. Introduction

Readmission is a very common risk among inpatients with chronic diseases. The financial burden of patients and the waste of medical resources is greatly increased by the readmission caused by inadequate treatment. Among the many hospitalized patients with chronic diseases, the readmission risk of diabetic patients deserves close attention. According to relevant literature data [1], the probability of readmission within 30 days of inpatients with diabetes and its complications is greater than of inpatients with other diseases. Chronic diabetes is difficult to control and cannot be cured, causing most patients to be repeatedly admitted to the hospital, which prolongs the medical service process for patients, greatly reduces the utilization rate of medical treatment, and exacerbates the mismatch between supply and demand of medical resources. Therefore, it is quite necessary to reduce the risk of readmission in diabetic patients.

Machine learning algorithms like decision trees, naive Bayesian, and support vector machines were used by Sisodia et al to process and predict diabetes data [2]. Ali et al. used the K-nearest neighbor algorithm to detect and classify diabetes, and the results showed that the K-nearest neighbor algorithm has higher accuracy [3]. The early detection results of diabetes were used by Sneha et al to study feature selection, and the naive Bayesian algorithm achieved the best results [4]. Khanam et al. have employed various machine learning algorithms to investigate the diagnosis and classification of diabetes, and discovered that logistic regression and support vector machine models were more effective in predicting

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/46/20241071

diabetes [5]. Krishnamoorthi et al. utilized different machine learning algorithms to classify and predict diabetes data, and developed an intelligent diabetes prediction framework [6]. Du et al. proposed an interpretable clinical decision support system based on machine learning, using algorithms like logistic regression, adaptive amplification and amplification of extreme gradients for data modelling, and improved the multi-planar model [7].

In summary, machine learning algorithms have broad application prospects in diabetes care, and are called upon to further improve the accuracy of diabetes prediction and diagnosis. However, there are numerous factors that affect the prediction of whether a patient will be readmitted. Numerous researchers choose a small data set, and the accuracy and stability of the algorithm cannot be guaranteed. Different researchers have used different machine learning algorithms and achieved different results in terms of accuracy and precision.

Therefore, in order to seek the best predictive model for readmission of diabetic patients, analyze the influencing factors of readmission, and effectively reduce the readmission rate. This paper takes a variety of methods, including XGBoost, logistic regression, GBDT, decision tree, random forest, and deep neural network. This variety of method choices provides a more comprehensive perspective for research. By adopting these methods, it is possible to analyze and model the readmission of diabetic patients from diverse perspectives. In addition, the performance indicators used throughout this paper are also more abundant, including Accuracy, F1, Precision, Recall, and ROC curves. These metrics give a more comprehensive assessment of model performance and provide information on prediction accuracy, precision, and recall. This paper conducts experiments on a larger dataset, which contains 101,766 pieces of data and 50 feature attributes. Furthermore, the analysis of feature attributes in this paper is more comprehensive. By analyzing comprehensive feature attributes, the model's performance can be optimized and the accuracy of diabetic patient readmission diagnosis can be improved.

2. Data preprocessing

2.1. Data Sources and Analysis

This dataset records the clinical care and readmission records of diabetic patients admitted to 130 hospitals in the United States over the past 10 years and contains 101,766 data and 50 feature attributes. Table 1 presents the primary features attributes.

Table 1. Feature analysis.

Feature Name	Description	Type
time_in_hospital	Integer number of days between admission and discharge	Continuous variable
num_lab_procedures	Number of lab tests performed during the encounter	Continuous variable
num_procedures	Number of procedures (other than lab tests) performed during the encounter	Continuous variable
num_medications	Number of distinct generic names administered during the encounter	Continuous variable
number_outpatient	Number of outpatient visits of the patient in the year preceding the encounter	Continuous variable
number_emergency	Number of emergency visits of the patient in the year preceding the encounter	Continuous variable
number_inpatient	Number of inpatient visits of the patient in the year preceding the encounter	Continuous variable
number_diagnoses	Number of diagnoses entered to the system	Continuous variable
change	Indicates if there was a change in diabetic medications	Categorical variable
diabetesMed	Indicates if there was any diabetic medication prescribed	Categorical variable
readmitted	Days to inpatient readmission	Categorical variable

DOI: 10.54254/2755-2721/46/20241071

2.2. Data processing

When conducting medical data research, the problem of missing data is often encountered. Figure 1 shows the lack of all attributes. It can be clearly seen that there are a large number of missing feature attributes such as 'payer_code', 'medical_specialty' and weight'. Statistics show that the rate missing 'weight' is 96.86%, the rate missing 'payer_code' is 39.56%, and the rate missing 'medical_specialty' is 49.08%. If the missing rate of these attributes is too high, it will have an impact on classification performance, so they need to be deleted. 'ecountet_id' and 'patient_nbr' belong to identifiers and have no specific meaning. It's best to delete them since they have a small impact on our data training. 'citoglipton' and 'examamide' represent the two drugs for treating diabetes respectively. All values of them are NO. It is not helpful for classification prediction, so the attribute is removed. For the missing feature attributes such as 'race', 'diag_1', 'diag_2', and 'diag_3', the interpolation method is utilized to fill them. The trend of existing data is utilized to predict missing values, thereby preserving the overall characteristics and statistical properties of the data. There are not any duplicate values in the dataset.

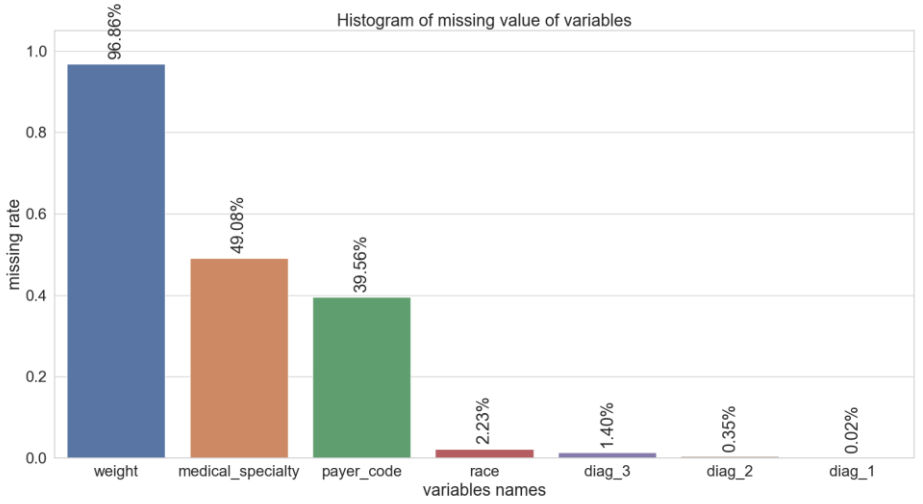

Figure 1. Data missing value display.

In machine learning, it is generally necessary to normalize the training data, mainly because normalization can speed up the speed of finding the optimal solution and can effectively improve the accuracy. This paper uses the LabelEncoder function in sklearn for encoding to obtain the mapping relationship between the original label and the current encoding. Then use fit_transform() to standardize the data.

For predictor label variables. Label variables are: <30, >30, NO. Readmission within 30 days of discharge, readmission after 30 days, and non-aggression are all represented by these three variables. In the experiments in this paper, the label variable of patients readmitted within 30 days is placed at 1 (positive cases) and the rest is set to 0 (negative cases). As illustrated in figure 2, the data is relatively balanced. This chapter adopts the SMOTE sampling method, which can effectively increase the amount of samples in the minority category, thereby improving the prediction ability of the model for the minority category. Through the generation of synthetic samples, the SMOTE method can increase the diversity of data and alleviate the problem of overfitting.

By screening the coefficients, the coefficients whose absolute value is larger than 0.045 are selected. As shown in figure 2, the following results are obtained: 'num_medications' (0.046772), 'time_in_hospital' (0.051289), 'diabetesMed' (0.061508), 'number_outpatient' (0.082142), 'number_emergency' (0.103011), 'number_diagnoses' (0.112564), 'number_inpatient' (0.217194), change (-0.046011).

DOI: 10.54254/2755-2721/46/20241071

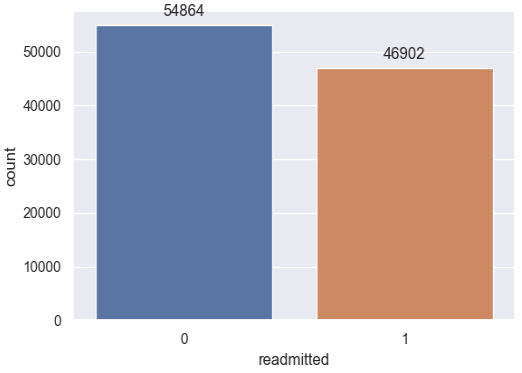

Figure 2. Data missing value display.

2.3. Feature processing

2.3.1. Feature screening

Feature filter. After data processing, this chapter adopts embedding method. Obtain the weight coefficients for each feature using the logistic regression model, and then select features based on the coefficients from large too small. The coefficients for features that are more irrelevant to the output variable will decrease, while those that are more important will have larger coefficients in the model. This paper uses the feature_selection library in sklearn for feature selection.

By screening the coefficients, the coefficients whose absolute value is larger than 0.045 are selected. As shown in figure 3, the following results are obtained: 'num_medications' (0.046772), 'time_in_hospital' (0.051289), 'diabetesMed' (0.061508), 'number_outpatient' (0.082142), 'number_emergency' (0.103011), 'number_diagnoses' (0.112564), 'number_inpatient' (0.217194), change (-0.046011).

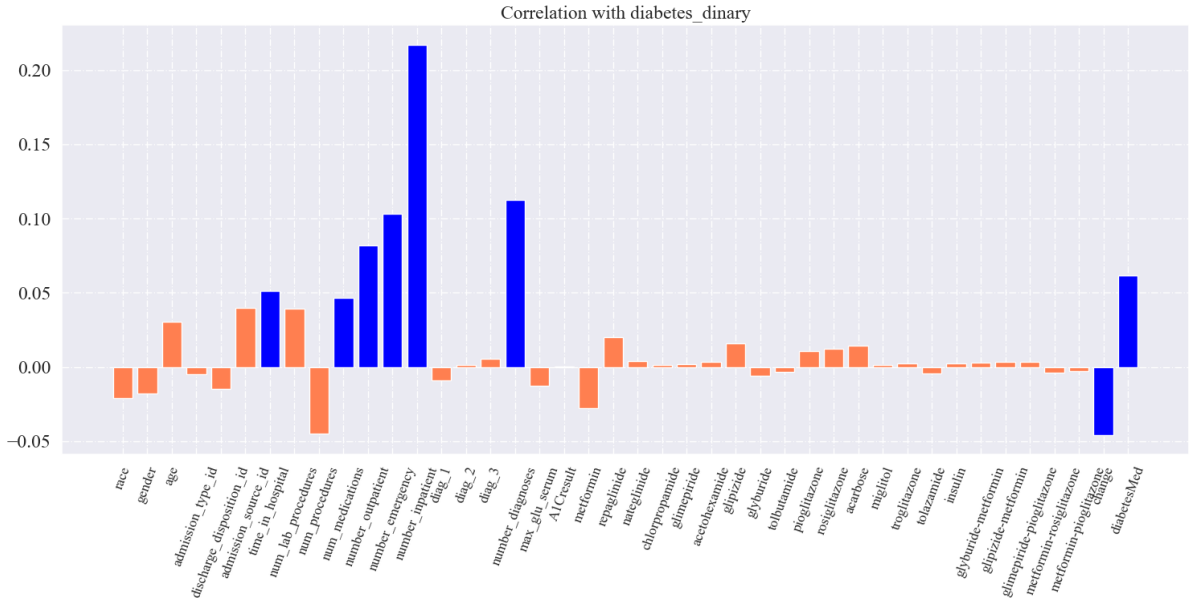

Figure 3. Feature Coefficient Screening.

DOI: 10.54254/2755-2721/46/20241071

2.3.2. Analysis of Feature Selection Results

(1) num_medications and time_in_hospital

As shown in figure 4 below, for the characteristic attribute of 'num_medications', as the number of patients taking drugs increases, the number of readmissions and non-readmissions shows a certain trend. Specifically, the number of readmissions and non-readmissions increases with the number of medications taken between 0 and 13. However, the number of readmissions and non-readmissions begins to decline after the number of medications taken exceeds 13. Through the observation of the number of drugs taken by patients, it is found that the changes in the number of drugs taken by both readmitted patients and non-rehospitalized patients show a similar pattern. Therefore, the discrimination of this feature is low.

For the characteristic attribute of 'time_in_hospital', with the increase of normal hospitalization time, the normal hospitalization time of non-readmitted patients and readmitted patients gradually tends to be the same. In most cases, patients' hospital stays are concentrated between 1 and 5 days. This suggests that the normal length of stay provides no additional information or discriminatory power for distinguishing between non-readmission and readmission samples.

Therefore, when building the model, the simplicity and predictive ability of the model can be improved by not adding the number of drugs taken by the patient and the normal length of hospitalization as features to the model.

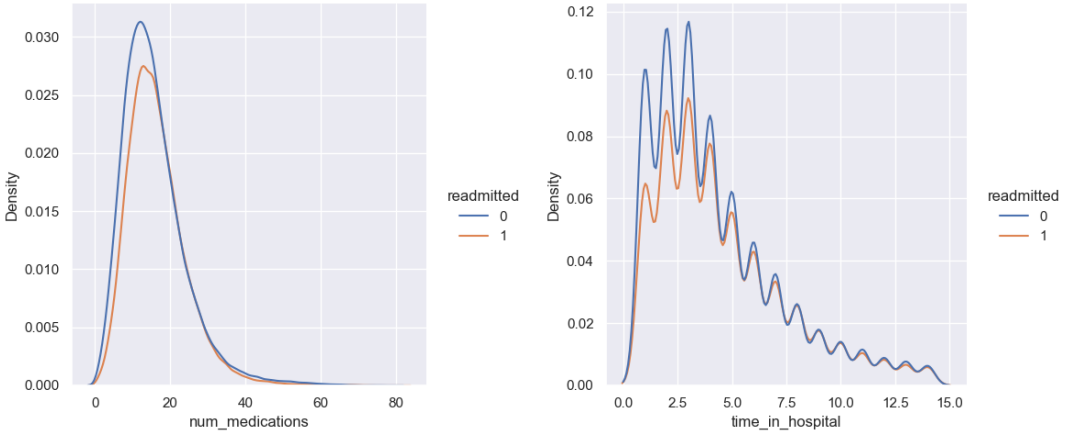

Figure 4. num_medications and time_in_hospital (kernel density estimation map).

(2) diabetesMed and change

As can be observed in figure 5 below, according to the characteristic attribute observation of 'diabetesMed'. Most patients who had been prescribed diabetes medications during their hospitalization were more likely to be readmitted. This suggests an association between diabetes medication use and readmission. This feature has a strong ability to predict readmission and can provide additional information to make a distinction between non-readmission and readmission samples.

Observations of changes in characteristic attributes inform the following: First, diabetes medication use has a limited effect on readmission. The number of readmissions remains relatively stable regardless of changes in diabetes medication use. This suggests that medication use has a smaller impact on readmission. Second, non-readmissions outnumber readmissions regardless of changes in diabetes medication use. This indicates that most patients are not readmitted regardless of changes in diabetes medication use. This may be because their diseases are effectively controlled and managed. Furthermore, the number of patients who are not readmitted is higher when diabetes medication use is changed compared to when it is unchanged. This may imply that patients are adopting more effective treatment strategies by adjusting their diabetes medication use, thereby maintaining disease stability and reducing the likelihood of readmission. In conclusion, the characteristics of diabetes medication use provide

DOI: 10.54254/2755-2721/46/20241071

valuable insights into patients' risk of readmission. This indicates that changes in medication use play a positive role in preventing readmissions.

Therefore, when building the model, ‘diabetesMed’ and ‘change’ are added to the model as features to enhance the predictive capability of the model.

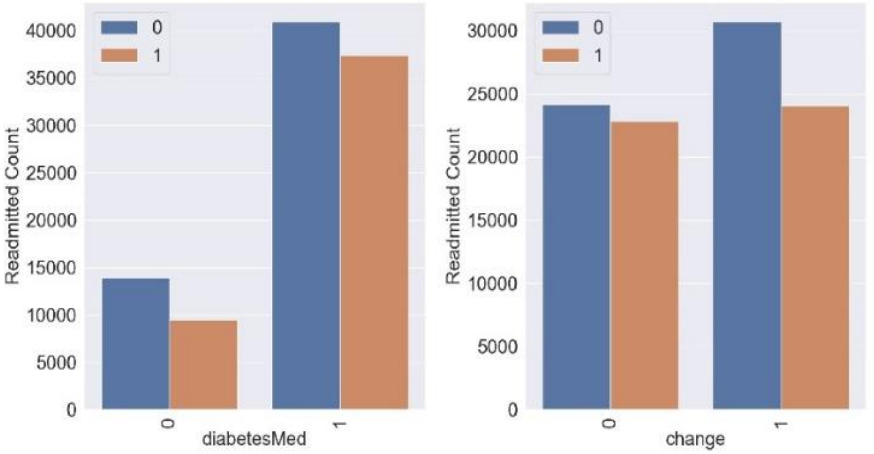

Figure 5. diabetesMed and change (kernel density estimation map).

(3) number_outpatient and number_emergency and number_inpatient

As can be seen in figure 6 below, the number of outpatient visits, emergency visits, and hospitalizations for most patients is 0. This shows that the samples of these three feature attributes are very unbalanced, and the trend cannot be seen intuitively. Therefore, in order to mine the data in more detail, this paper will further analyze the relationship between the number of outpatient visits and readmissions.

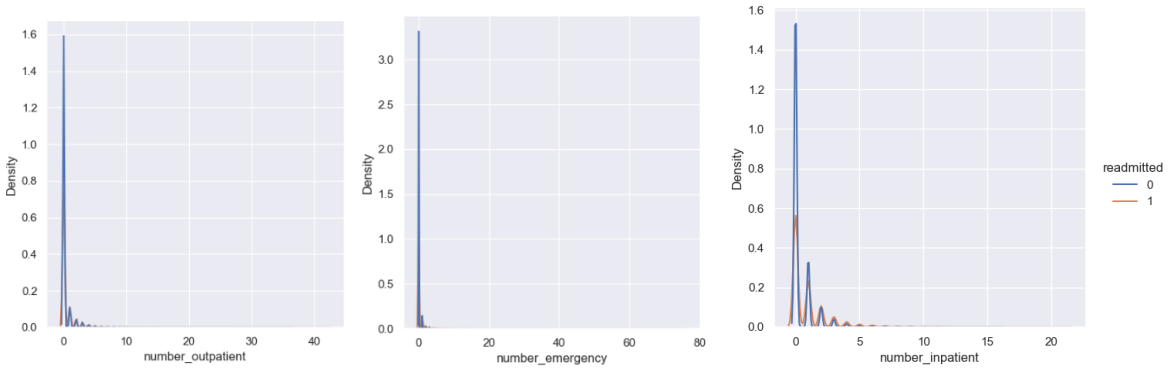

Figure 6. number_outpatient and number_emergency and number_inpatient (kernel density estimation map).

As observed in figure 7 below, this paper will first focus on data with less than 5 outpatient visits. When the number of outpatient visits is 0, there is a significantly larger number of patients who were not readmitted compared to those who were readmitted. This indicates that patients with 0 outpatient visits have a relatively low probability of readmission. This can be attributed to the fact that patients with 0 outpatient visits were either less severely ill or in the recovery stage.

On the other hand, when the number of hospitalizations is equal to or greater than 1, the number of patients who were readmitted begins to exceed the number of patients who were not readmitted. However, it is important to note that many patients were more prone to readmission when the number

DOI: 10.54254/2755-2721/46/20241071

of outpatient visits reached ≥5. This suggests that these patients were more severely ill and required more frequent outpatient treatment or follow-up.

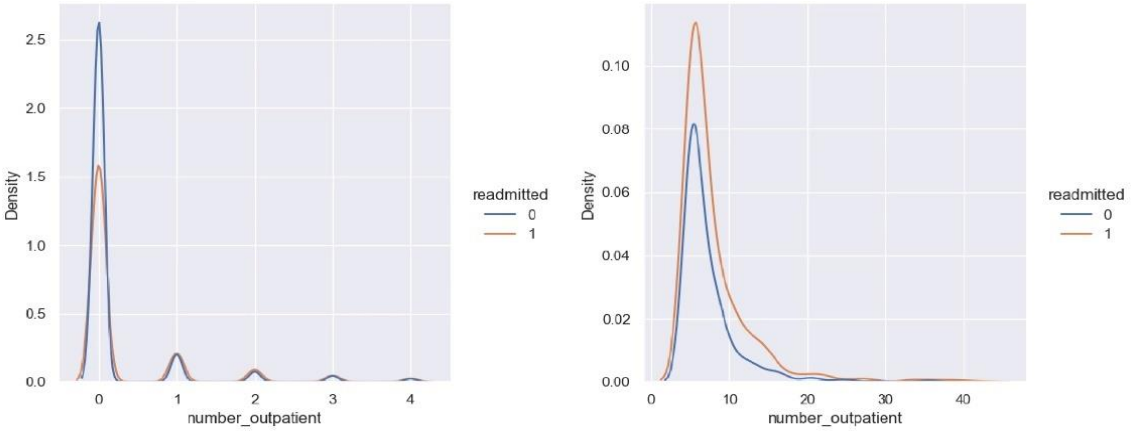

Figure 7. number_outpatient (Left picture: <5, Right picture: ≥5).

As illustrated in figure 8 below, the primary focus of this paper is to analyze the dataset comprising patients with less than five emergency visits. Within this specified range, it was found that when the number of emergency visits reached zero, the number of patients who were not readmitted exceeded those who were readmitted by a significant margin. This observation implies that patients with zero emergency visits have a reduced probability of readmission. Plausible explanations for this trend include the absence of emergencies, the presence of mild medical conditions, or the administration of timely and effective outpatient care.

However, in cases where the number of emergency visits exceeded five, it was observed that a higher proportion of patients were prone to readmission as their emergency visits approached ten. Consequently, there exists a positive correlation between the frequency of emergency visits and the probability of readmission among patients. This association could potentially be attributed to the increased severity of illness or reduced responsiveness to treatment, leading to a heightened necessity for frequent emergency visits.

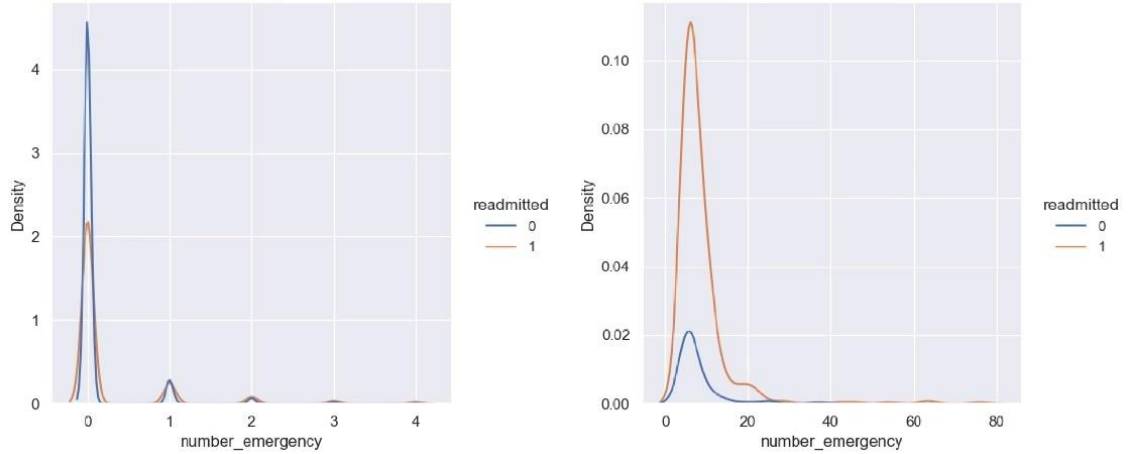

Figure 8. number_emergency (Left picture: <5, Right picture: ≥5).

As illustrated in figure 9, this paper focuses on examining cases with less than five hospitalizations. When the number of hospitalizations is zero, the number of patients not admitted is higher than the

DOI: 10.54254/2755-2721/46/20241071

number of patients readmitted, suggesting a low probability of readmission. This could be attributed to effective outpatient treatment, eliminating the need for re-hospitalization. However, when the number of hospitalizations reaches one or more, the number of patients readmitted surpasses those not readmitted.

Subsequently, attention is directed towards data involving five or more hospitalizations. Within this range, a considerable number of patients demonstrate a propensity for readmission as the number of hospitalizations approaches five. This phenomenon may arise from patients having complex medical conditions requiring long-term treatment and monitoring. As the number of hospitalizations nears or exceeds five, readmission becomes necessary to facilitate comprehensive and continuous care.

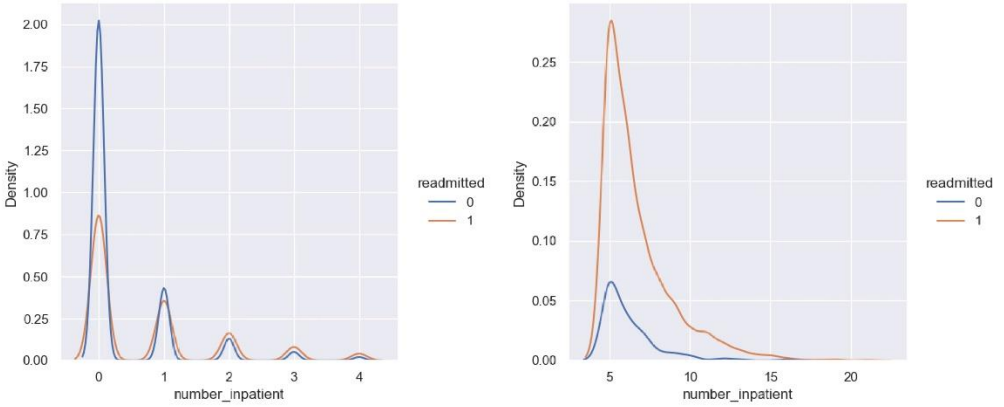

Figure 9. number_inpatient (Left picture: <5, Right picture: ≥5).

Based on the aforementioned observations, it is evident that the number of outpatient visits, emergency visits, and hospitalizations exhibit valuable information and patterns within a specific range. Hence, it is deemed worthwhile to incorporate them into the model. In constructing the model, this paper categorizes the number of outpatient visits into two groups: zero outpatient visits and one or more outpatient visits. Similarly, the number of emergency visits is categorized as zero emergency visits and one or more emergency visits. Additionally, the number of hospitalizations is divided into two categories: zero hospitalizations and one or more hospitalizations.

(4) number_diagnoses

Illustrated in figure 10, there is no substantial disparity between patients who were readmitted and those who were not in terms of the number of diagnoses. However, slight variations can be observed, such as a relatively higher likelihood of readmission when the number of diagnoses falls within the range of 8 to 10. Nevertheless, due to the uneven distribution of diagnoses, visual observation alone may not provide a comprehensive understanding of the trend. To gain a more detailed insight into the data, this paper will delve deeper into the analysis to explore the relationship between the number of diagnoses and readmissions.

DOI: 10.54254/2755-2721/46/20241071

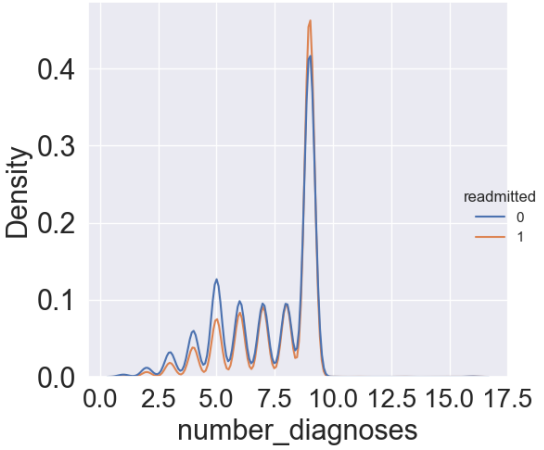

Figure 10. number_diagnoses (kernel density estimation map).

As indicated in figure 11, there is no significant difference between the number of patients who were readmitted and those who were not. However, a higher proportion of readmissions can be observed when the number of diagnoses falls within the range of 8 to 14. This suggests a potential correlation between the number of diagnoses and readmissions within this specific range. Conversely, when the number of diagnoses exceeds 14, a greater number of patients were not readmitted compared to those who were readmitted. This implies that a higher number of diagnoses (greater than 14) may be associated with a lower probability of readmission.

Based on these findings, it is evident that the number of diagnoses presents valuable information and patterns within a certain range. Therefore, it is deemed worthwhile to incorporate the number of diagnoses as a feature in the model. Accordingly, when constructing the model, this paper includes the number of diagnoses as a categorical variable divided into three categories: less than 8 diagnoses, 8 to 14 diagnoses, and greater than 14 diagnoses.

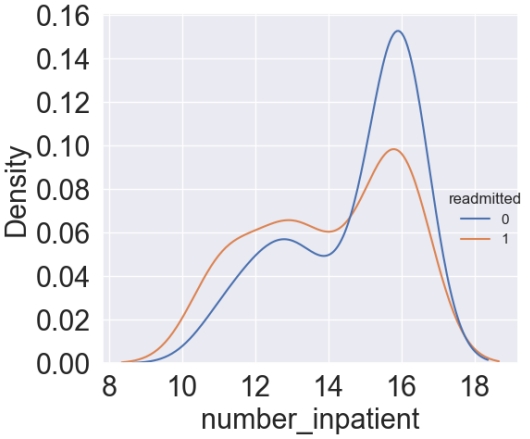

Figure 11. number_diagnoses (The number of diagnoses is greater than 8).

2.3.3. Feature Correlation Analysis. As depicted in figure 12 below, a strong correlation of 0.51 is observed between "Change" and "diabetesMed". This indicates that a significant change in the treatment plan of patients corresponds to a substantial alteration in the usage of diabetes drugs. This finding highlights the connection between treatment requirements and medication usage, warranting further investigation.

DOI: 10.54254/2755-2721/46/20241071

Likewise, a strong correlation of 0.47 is observed between "time_in_hospital" and "num_medications". Additionally, there is a certain correlation of 0.39 between "num_medications" and "num_lab_procedures". Furthermore, a moderate correlation of 0.32 exists between "time_in_hospital" and "num_lab_procedures".

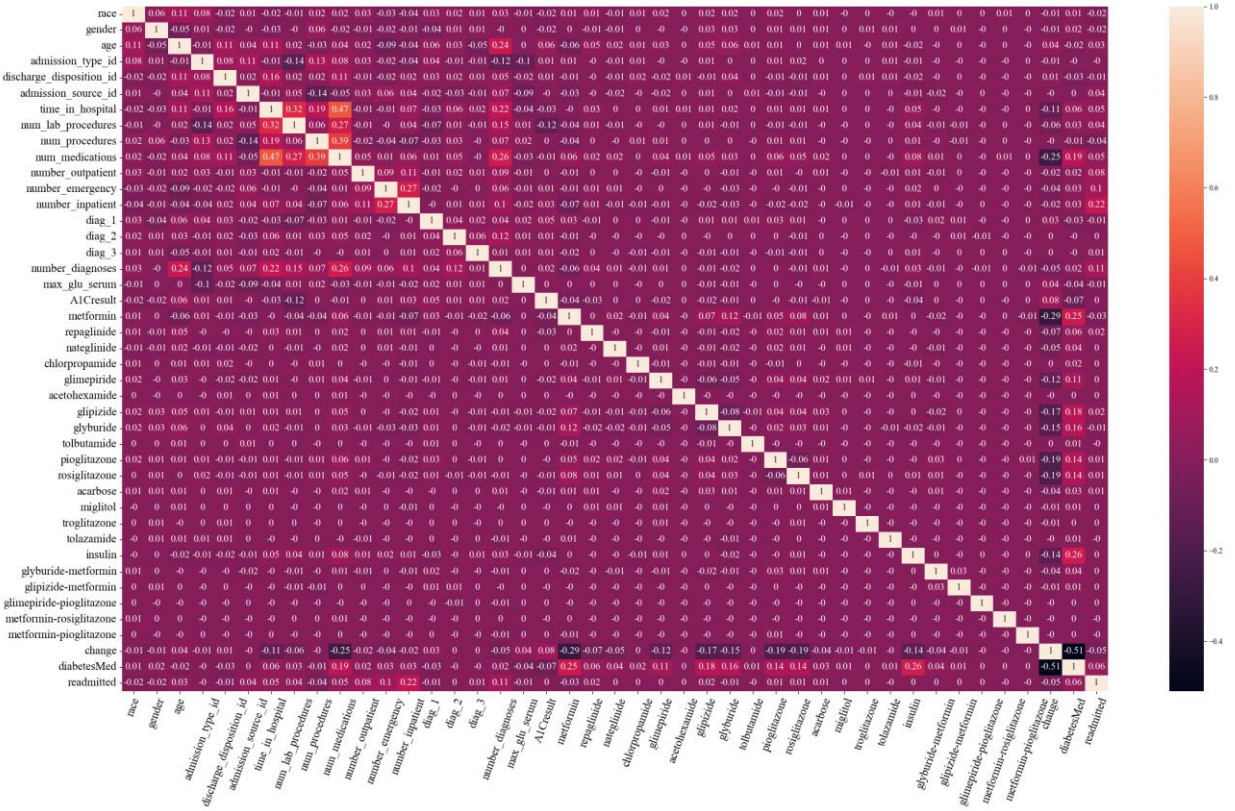

Figure 12. Correlation heat map.

3. Model construction

In terms of models, this paper chooses six types: XGBoost, logistic regression, GBDT, decision tree, random forest, and deep neural network.

(1) XGBoost:

XGBoost was selected because it is a gradient boosting tree algorithm [8]. It gradually optimizes the performance of the model by iteratively training multiple decision trees. XGBoost uses regularization techniques during training to prevent overfitting and supports parallel computing. It performs well when dealing with structured data and high-dimensional sparse features, achieving high accuracy, F1-score, precision, AUC, and recall. XGBoost also handles missing and outlier values and helps explain model predictions through feature importance analysis.

(2) GBDT:

GBDT is selected because it is a decision tree ensemble model based on the gradient boosting algorithm [9]. It trains multiple decision trees iteratively and continuously optimizes the model through the gradient of the loss function. GBDT can handle classification and regression problems and has strong nonlinear fitting ability. It has high accuracy and interpretability and can handle large-scale datasets and high-dimensional features. GBDT is also capable of handling non-linear relationships and aids in feature selection and interpretation of model predictions through feature importance analysis.

(3) Random Forest:

Random forest is chosen because it is a decision tree-based ensemble learning model [10]. It builds multiple decision trees by randomly selecting features and sample sets and makes predictions by voting

DOI: 10.54254/2755-2721/46/20241071

or averaging. Random forest has better robustness and accuracy and can handle high-dimensional data and feature selection problems. It is capable of dealing with outliers and missing values and assists in feature selection and interpretation of model predictions through feature importance analysis.

(4) Decision Tree:

Decision tree is chosen because it is a classification and regression model based on a tree structure [11]. It makes predictions by stepwise partitioning the feature space. Decision trees have the advantages of being easy to interpret, visualize, and handle mixed data types. They are also robust to outliers and missing values. Decision trees can help in understanding and interpreting the model's predictions through feature selection and visualization.

(5) Logistic Regression:

Logistic regression was selected because it is a generalized linear model for binary classification problems [11]. It works by mapping the output of a linear regression model to a probability value, followed by classification. Logistic regression is simple, fast, easy to interpret, and interpretable. It has low requirements for feature engineering and can handle large-scale data and high-dimensional features. Logistic regression can help understand and explain the prediction results of the model through feature weight analysis.

(6) Deep Neural Network (DNN):

The deep neural network was chosen because it is a machine learning model based on a multi-layer neuron network, which makes predictions through multi-layer nonlinear transformation and weight learning. DNN performs well when dealing with complex nonlinear problems and large-scale datasets. It can automatically learn feature representations and has strong pattern recognition and abstraction capabilities.

DNN's neural network layers can be categorized into three categories: the input layer, hidden layers, and output layer. The first layer is the input layer, the last layer is the output layer, and the layers in between are all hidden layers, each containing multiple neurons. This paper uses two hidden layers containing 64 and 20 neurons, respectively, as indicated in the figure 13.

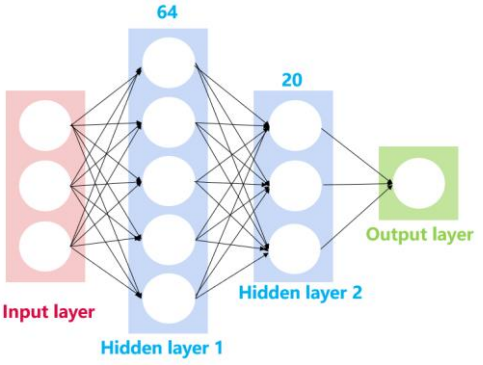

Figure 13. Deep Neural Network Architecture.

4. Experimental results and analysis

4.1. Experiment details

The experiments in this paper were primarily conducted on a GPU server using the Python programming language. PyCharm was utilized for development and experimentation purposes. The preprocessed dataset was split into a training set and a set of tests with a ratio of 7:3. The training set was used for model training and parameter adjustment, while the test set was used to evaluate model performance and generalizability.

The trained model was evaluated using the test set to calculate various performance indicators, including accuracy, f1 score, precision, recall. Additionally, the ROC curve was plotted to evaluate the model's classification ability and threshold selection.

DOI: 10.54254/2755-2721/46/20241071

4.2. Parameter settings

1.Training times: The training times were selected using the cross-validation method. The range of training times ranged from 100 to 1000, with intervals of 100. The training set was used to train the model, and its performance was evaluated on the validation set for each epoch. The optimal number of training times was determined based on the evaluation metrics on the validation set.

2.XGBoost and GBDT: The maximum tree depth was set to 3, and the learning rate was set to 0.1. Both the sample sampling method and the sampling feature ratio were set to 0.5. The number of trees was set to 100.

3.Random Forest: The number of decision trees used was set to 100. sqrt(n_features) was used as the feature selection method, where n_features is the total number of features. True was the sampling mode for the sample.

4.Decision tree: The Gini coefficient was selected as the splitting criterion. The splitting criterion was determined by the Gini coefficient. Set the minimum number of samples required for a partition to 2 and the minimum number of samples required for a leaf node to 1. The tree's maximum depth was set to 3.

5.Logistic regression: L2 regularization was selected, and the regularization parameter was set to 0.1. The maximum number of iterations was set to 100.

6.Deep neural network: The model was composed of two hidden layers, one with 64 neurons and the other with 20 neurons. ReLU was the activation function used in the first hidden layer, and sigmoid was used in the last layer. The model was compiled using the compile function. During training, several callback functions were employed to monitor and optimize the model.

Adam optimizer at adaptive learning rate has been chosen. The model's initial learning rate was set to 0.0001, which enabled the loss function to quickly converge to a minimum value during training.

4.3. Model comparative analysis

Compare the performance indicators of six classifiers including XGBoost, logistic regression, GBDT, decision tree, random forest, and deep neural network. As shown in the table 2 below, the classifier with the best performance is XGBoost. It achieved the highest values in accuracy (0.644), f1-score (0.591), AUC (0.701), and recall (0.559). The highest precision value among the classifiers is 0.642, which is achieved by GBDT. The ROC curve can be seen from Figure 14.

Table 2. Performance comparison.

Model Name	Accuracy	F1-score	Precision	Recall	AUC
Xgboost	**0.6443**	**0.5916**	0.6283	**0.5590**	**0.7014**
GBDT	0.6399	0.5585	**0.6420**	0.4943	0.6956
RF	0.6357	0.568	0.6264	0.5195	0.6881
DT	0.5653	0.5300	0.5282	0.5317	0.5628
LR	0.6170	0.4885	0.6352	0.3969	0.6537
DNN	0.6376	0.5738	0.6264	0.5293	0.6883

DOI: 10.54254/2755-2721/46/20241071

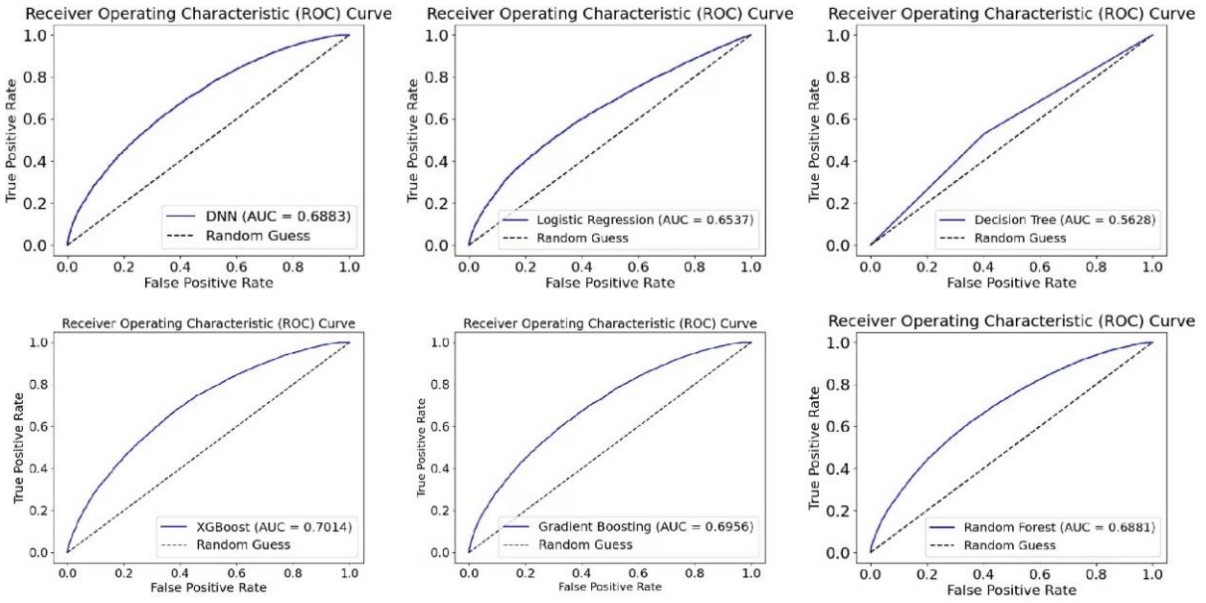

Figure 14. ROC curve.

The reason why XGBoost performs best overall is that it automatically combines features. By splitting and combining features, it can better capture the nonlinear relationships between them and improve the model's generalization ability. On the other hand, the overall performance of decision trees is the worst. Decision trees are prone to overfitting problems when dealing with large-scale datasets, especially when the tree depth is too large. This overfitting leads to poor performance on the test set and a decline in the model's generalization ability. Moreover, decision trees may struggle to find the global optimal solution when dealing with large-scale datasets, affecting the model's performance on various evaluation indicators.

There are several reasons for the underperformance of DNNs compared to XGBoost. Firstly, DNNs typically require a large amount of training data to fully learn complex features and patterns. If the dataset is small, DNNs may be unable to leverage their advantages, resulting in inferior performance compared to XGBoost. Secondly, XGBoost relies less on feature engineering and can automatically capture nonlinear relationships between features, while DNNs often require more feature engineering for processing large-scale data. Additionally, XGBoost is more suitable for handling datasets with a large number of sparse features, as it can store data in a sparse format and perform efficient calculations. In contrast, DNNs usually require converting sparse features into dense features, which increases storage and computational complexity. Lastly, DNNs have high requirements for data preprocessing, while XGBoost has relatively fewer requirements and can directly process raw data, making it more convenient and efficient for handling large-scale data. Overall, these factors contribute to the inferior performance of DNNs compared to XGBoost.

The reason why logistic regression performed worst in terms of F1-score (0.4885) and Recall (0.3969) is because it is a linear classifier with limited ability to fit nonlinear relationships in the data. Additionally, when there is just an imbalance in the data, logistic regression may be biased towards predicting categories with a large number of samples, resulting in poor predictive performance for minority categories. In contrast, other models like XGBoost, decision tree, GBDT, random forest, and deep neural network can address data imbalance through sample weight adjustment or sampling strategies, thereby improving the prediction performance of minority categories.

Compared with XGBoost, GBDT achieves a higher accuracy rate of 0.642. This advantage of GBDT can be attributed to the use of regression trees as base classifiers, which have a tendency to produce highly accurate results for classification problems. Additionally, in GBDT, the training of each base

DOI: 10.54254/2755-2721/46/20241071

classifier adjusts the samples based on the predictions of previous rounds, leading to improved classification accuracy.

4.4. Cause Analysis of Low Performance Index

1.The low performance index may be caused by issues such as noise, missing values, or outliers in the dataset itself. These problems can interfere with the model's training and prediction process, resulting in degraded performance.

2.Feature engineering plays a crucial role in improving model performance. Problems in feature selection, feature transformation, and feature combination can contribute to low performance indicators. If feature extraction is inadequate or if the relationships between features are not accurately represented, the model may struggle to learn effective patterns and rules from the data.

3.Adjusting parameters has a significant impact on performance indicators for models such as XGBoost, logistic regression, GBDT, decision tree, random forest, and deep neural network. Insufficient parameter tuning can lead to suboptimal model performance.

4.The presence of class imbalance in the dataset can result in poor prediction performance for minority classes. This imbalance can affect the overall performance of the model across various evaluation metrics.

5.Models with higher complexity generally require more data and longer training time to achieve higher performance indicators.

5. Conclusion

Diabetes is a chronic disease with a high incidence rate and is difficult to cure, which has a significant effect on global health. With the advancements in computer technology, it can provide doctors with tools to assist in diabetes diagnosis and utilize big data technology to improve the accuracy of diabetes diagnosis, thereby alleviating the burden on doctors.

This paper employs various methods, including XGBoost, logistic regression, GBDT, decision tree, random forest, and deep neural network, among others. The diverse range of methods chosen provides multiple perspectives for research. Whether applied to prediction, diagnosis, or treatment, these methods have demonstrated remarkable power and accuracy, contributing significant advancements to the field of diabetes research. Additionally, this paper utilizes a diverse set of evaluation metrics, including Accuracy, F1, Precision, Recall, and ROC curves. The model's performance is thoroughly evaluated by these metrics, resulting in more comprehensive evaluation results. XGBoost's overall performance is the best, with an accuracy of 0.6443, an F1-score of 0.5916, a recall rate of 0.559, and an AUC of 0.7014. GBDT performs exceptionally well in terms of precision, reaching 0.642. Utilizing these metrics can enable us to gain a comprehensive understanding of the model's accuracy, precision, recall rate, and predictive ability under different thresholds. Furthermore, the feature attribute analysis conducted in this paper is comprehensive. By thoroughly analyzing the feature attributes, we can gain insights into the relationships between features, the degree of influence on the prediction results, and the presence of redundant or invalid features. This comprehensive analysis of feature attributes serves as a critical foundation for model optimization.

In summary, this paper offers a comprehensive and accurate solution for diabetes research and diagnosis through the use of diverse methods, a wide range of evaluation metrics, and a comprehensive analysis of feature attributes.

References

[1] Friedman B, Jiang HJ, Elixhauser A. Costly hospital readmissions and complex chronic illness. Inquiry: The journal of health care organization, provision, and financing. 2008;45(4):408-421.

[2] Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. Procedia Computer Science. 2018;132(1):1578-1585.

[3] Ali A, Alrubei M, Hassan LFM, et al. Diabetes classification based on KNN. IIUM Engineering

DOI: 10.54254/2755-2721/46/20241071

Journal. 2020;21(1):175-181.
[4] Sneha N, Gangil T. Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big Data. 2019;6(1):467-480.
[5] Khanam JJ, Foo SY. A comparison of machine learning algorithms for diabetes prediction. ICT Express. 2021;7(4):432-439.
[6] Krishnamoorthi R, Joshi S, Almarzouki HZ, et al. A novel diabetes healthcare disease prediction framework using machine learning techniques. Journal of Healthcare Engineering. 2022;3(1):126-138.
[7] Du Y, Rafferty AR, McAuliffe FM, et al. An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus. Scientific Reports. 2022;12(1):1-14.
[8] Li J, Han K, Shen J, Sun W, Gao L, Gao Y. Application value of XGBoost machine learning model in the diagnosis of hepatitis B cirrhosis. World Chinese Journal of Digestology. 2023;31(13):544-554.
[9] Guan J, Yao L, Chung CR, et al. StackTHPred: Identifying tumor-homing peptides through GBDT-based feature selection with stacking ensemble architecture. International Journal of Molecular Sciences. 2023;24(12):10348.
[10] Yu J, Zhou B, Wang C, et al. Comparison of the efficacy of random forest and logistic regression models in predicting prolonged hospital stay for hip fracture patients. Chinese Journal of Tissue Engineering Research. 2023;27(34):5413.
[11] Chen J, Xu L, Wu X, et al. Analysis of factors influencing unplanned hospital readmission in postoperative colorectal cancer patients based on logistic regression and decision tree models. Journal of Nursing. 2022;29(2):1-6.