Diabetes Prediction Using Random Forest in Healthcare

Shengyu Wang *

Department of Natural Sciences, The University of Manchester, Chengdu, China

* Corresponding author: shengyu.wang-3@student.manchester.ac.uk

Abstract. Accurate diabetes prediction has emerged as a crucial problem in the field of healthcare. It is important to detect individuals at risk of having diabetes, which can allow for prompt intervention and tailored treatment strategies. Nowadays, machine learning models are usually employed for diabetes prediction. Lots of work has been developed various machine learning models for diabetes prediction. The random forest, as a popular ensemble learning algorithm, has illustrated its superiority for diabetes prediction. To this end, this paper demonstrates a study that employs the random forest algorithm for diabetes prediction. A random forest model is trained on a publicly available diabetes dataset and compared to its performance with logistic regression. The missing data imputation techniques are further leveraged to improve data integrity. Regarding model performance, it can be found that the random forest model significantly outperforms the logistic regression model. This highlights the superiority of tree-based models, such as random forest, for predicting diabetes compared to logistic regression.

Keywords: Diabetes prediction, random forest, logistic regression.

1. Introduction

In the field of healthcare, accurate diabetes prediction is a pressing concern. It is important to detect individuals at risk of having diabetes, which can allow for prompt intervention and tailored treatment strategies.

The prediction of diabetes usually employs machine learning models. The research on machine learning model-based diabetes prediction can date back to [1]. Subsequently, lots of work has been developed on various of machine learning models for diabetes prediction [2, 3]. The random forest, as a popular ensemble learning algorithm, has illustrated its superiority for diabetes prediction.

This paper explores the random forest algorithm in the task of diabetes prediction. Our goal is to develop a reliable and interpretable model to help healthcare practitioners identify individuals at risk. Specifically, this paper employs the publicly available Pima Indians Diabetes dataset for model training and evaluation. The missing data imputation techniques are leveraged to improve data integrity, train a random forest algorithm, and further conduct a comparative analysis between random forest and logistic regression. Experimental results show that the Random Forest model exhibits superior performance compared to the logistic regression model.

The rest of this paper first reviews related work in Section 2, and then presents the methodologies in Section 3. The experimental results are illustrated in Section 4. Finally, this paper is concluded with a discussion and analysis of the experimental results in Section 5.

2. Literature References

The study of diabetes prediction can be traced back to [1], which focused on high-risk groups affected by diabetes. The authors identified some critical features for diabetes prediction and employed the ADAP algorithm to train a neural network adaptively. The results showed the potential of machine learning based methods in the field of healthcare.

In a subsequent study, Kayaer et al. [4] employ another neural network model, the general regression neural network (GRNN), for diabetes prediction. Unlike the ADAP algorithm, GRNN can approximate any function without the need for an iterative training process. Thus, the GRNN may have a better performance than the ADAP algorithm.

Kahramanli in proposed a hybrid neural network model for medical data analysis such as diabetes prediction [2]. Kahramanli found that the fuzzy set theory can be helpful for effectively handling uncertain and imprecise medical data [2]. To this end, the authors introduced a hybrid method that combines artificial neural networks (ANN) and fuzzy neural networks (FNN). The hybrid model can leverage the advantages of these two technologies to improve accuracy and reliability.

Following the idea of fuzzy logic, Lee et al. [5] proposed a fuzzy expert system to support decision-making of diabetes-related applications. The author claimed that, when encountering imprecise and fuzzy knowledge in practical scenarios, the traditional ontology has some disadvantages. To overcome these challenges, they proposed to use a fuzzy ontology and introduced a five-layer fuzzy ontology tailored specifically for the field of diabetes. Specifically, a fuzzy diabetes ontology (FDO) was proposed to represent knowledge related to diabetes. The authors introduced a Semantic Decision Support Agent (SDSA) method that can integrate mechanisms for knowledge construction, fuzzy ontology generation, and semantic fuzzy decision-making.

In a recent study, Hasan et al. [3] proposed to utilize of various machine-learning methods for diabetes prediction. The authors proposed a comprehensive data pre-processing workflow, including outlier detection, missing value imputation, and feature selection. Various machine learning classifiers such as decision trees, random forests, k-nearest neighbors, naive Bayes, AdaBoost, XGBoost, and multi-layer perceptron are utilized to train the diabetes prediction models. Furthermore, the authors proposed an ensemble classifier that can combine the outputs of multiple machine learning models through weighted voting.

Overall, the advancements in machine learning have significantly contributed to the field of diabetes prediction. These developments hold great potential for improving early detection, decision support, and personalized treatment strategies for individuals at risk of or living with diabetes.

3. Methods

This section introduces the employed dataset and presents the methodologies of diabetes prediction.

3.1. Data Description

Table 1. The statistics of the Pima Indians Diabetes Database

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	DPF	Age
Mean	3.84	120.89	69.11	20.54	79.80	31.99	0.47	33.24
Standard Deviation	3.37	31.97	19.35	15.95	115.24	7.88	0.33	11.76
Minimum	0	0	0	0	0	0	0.08	21
Median	3	117	72	23	30.5	32	0.37	29
Maximum	17	199	122	99	846	67.10	2.42	81

The data used in this paper is sourced from the publicly available Pima Indians Diabetes database. This database consists of 768 patient records of Pima Indian Women. Each record contains 8 features: Age, Blood Pressure, Body Mass Index (BMI), Diabetes Pedigree Function (DPF), Glucose, Insulin, Pregnancies and Skin Thickness. It is labeled by whether the patient has diabetes or not (Outcome). Table 1 shows the statistics of the dataset.

Among the 768 records, 500 have an outcome of 0 (indicating no diabetes), while only 268 have an outcome of 1 (indicating diabetes). This indicates an imbalanced dataset, with most records indicating the absence of diabetes. The detailed distributions among all 8 features grouped by the outcome are shown in Figure 1.

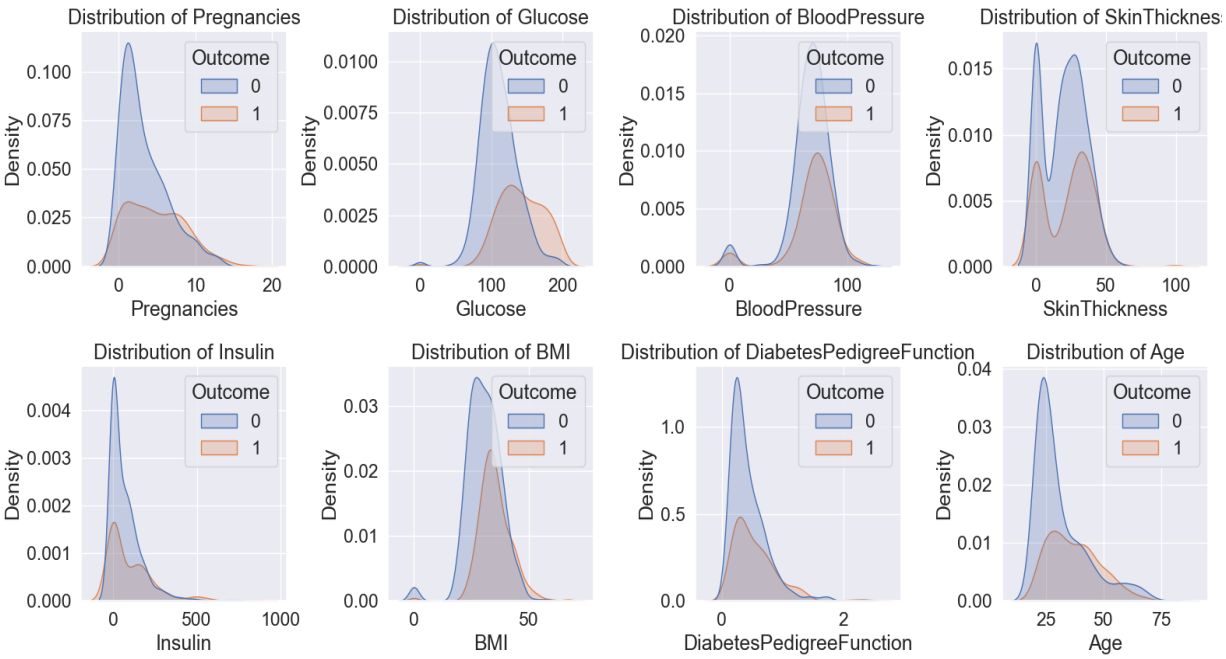


Figure 1. The data distribution of each feature grouping by the outcome.

3.2. Missing Value Imputation

From Table 1, it can be noted that all features except DPF and Age have minimum values of zero. While the presence of zero values for features like Pregnancies (indicating the number of times pregnant) can be reasonably explained, it is implausible for other features such as Glucose (indicating the plasma glucose concentration after 2 hours), Blood Pressure, and BMI to be zero for a normal human. For example, the average Glucose in blood level for humans is around 80-180 mg/dl. And the blood pressure is usually above 60 mm Hg. Such zero values are missing data. This subsection discusses how to handle such missing values for accurate model training.

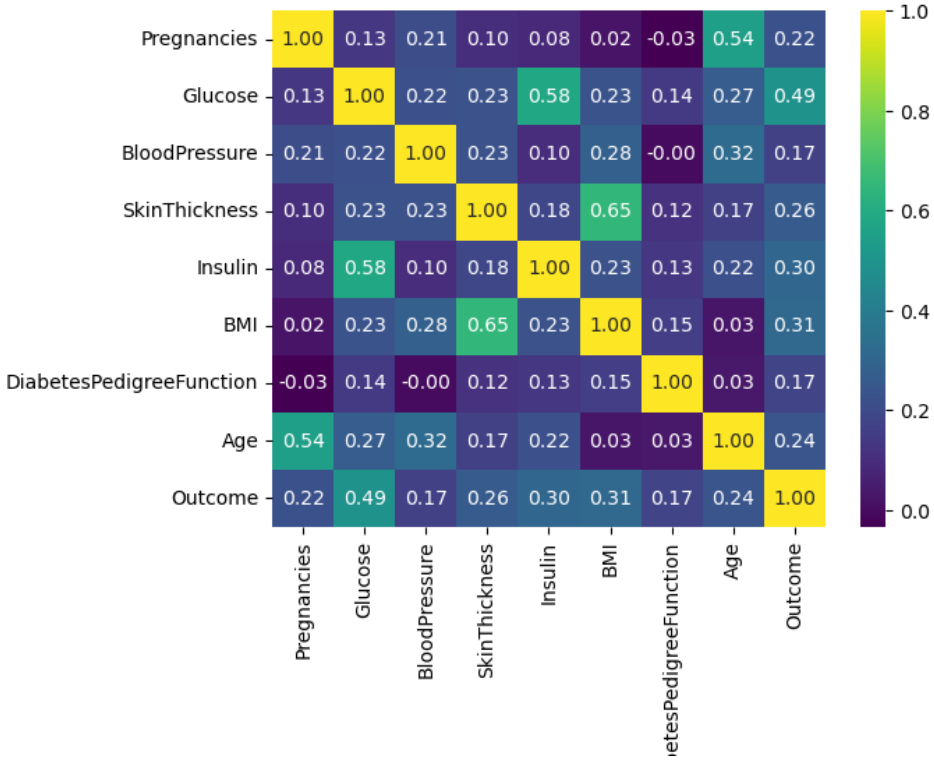


Figure 2. The Pearson correlation coefficients of features.

Imputation for low missing rate data. From Figure 1, it can be found that the proportion of missing (zero) values of feature Glucose, Blood Pressure, and BMI is relatively small. Specifically, there are 5 missing values for Glucose, 35 missing values for Blood Pressure, and 11 missing values for BMI. For such features, the missing values can be simply filled with their means or medians, which is a frequently used missing value imputation technique [6].

Imputation for high missing rate data. For the other two features with missing values, i.e., Skin Thickness and Insulin, the proportion of missing values is larger. Specifically, there are 227 missing values for Skin Thickness (30% missing rate), and 374 missing values for Insulin (49% missing rate). For such features with high missing rates, it cannot be simply replaced by the mean or median values, because it may heavily skew the original data distribution. Thus, this paper tries to use the other correlated features to impute them.

Specifically, this paper first draws the joint probability distribution of each pair of features, as shown in Figure 2. It can be found that the correlation between Skin Thickness and BMI is strong. And the feature of Insulin may be correlated with Glucose. To verify these correlation assumptions, the Pearson correlation coefficients of these features are computed, as shown in Figure 3. The correlation coefficients of Insulin and Glucose, and that of BMI and Skin Thickness are significantly higher than other features. Thus, the features of BMI and Glucose can be used to fill the missing values of Skin Thickness and Insulin, respectively.

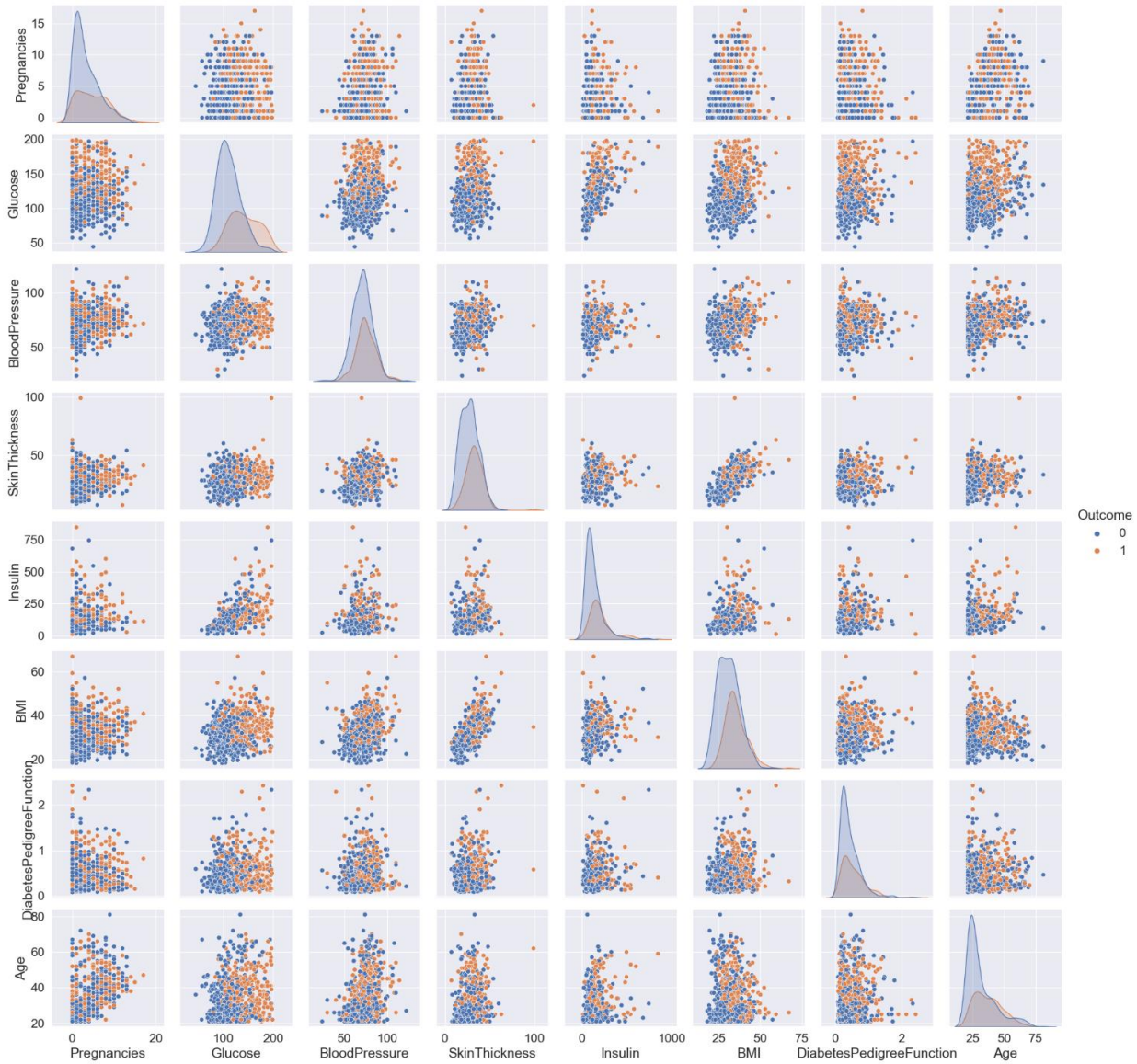


Figure 3. The joint probability distribution of each pair of features.

Specifically, it utilizes a linear regression model to learn the correlation between these values. Take the model of filling the missing values of Skin Thickness as an example, the linear regression model can be formulated as

$$\hat{y}_{skin} = a \cdot x_{BMI} + b \tag{1}$$

And it can use the Least Square method to find an optimal a and b to minimize the square error between the $\hat{y}_{skin}$ and the ground truth yskin. The model for Insulin is similar. And it just replaces the feature x_{BMI} as $x_{Glucose}$.

3.3. Model Selection

After the data preprocessing, two machine learning models are utilized to make the diabetes prediction, *Logistic Regression* and *Random Forest*.

Logistic Regression. Logistic regression is a linear model that maps the input features to the probability of a binary outcome [7]. This is achieved by employing the sigmoid function, a.k.a., the logistic function. The sigmoid function acts as a converter, transforming the linear combination of the input features into a value that falls within the range of 0 to 1. It is defined as

$$P\ (y = 1|x) = \frac{1}{1+exp(-z)} \tag{2}$$

Where $P\ (y = 1|x)$ denotes the probability of a particular event occurring (e.g., diabetes or not). The input features are denoted by the variable x, and z represents a linear combination of these input features. Mathematically, z is defined as:

$$z = a_1 \cdot x_1 + a_2 \cdot x_2 + \cdots + a_n \cdot x_n + b \tag{3}$$

Here, $a_1, \cdots, a_n$, and b are the weights assigned to each input feature $(x_1, x_2, \cdots, x_n)$. These weights can be estimated through the maximum likelihood estimation (MLE). Specifically, the likelihood function is defined as the product of the probabilities of the observed outcomes given the input features and the current coefficients. The MLE seeks to find the coefficients that maximize this likelihood function.

To make predictions using a logistic regression model, a decision boundary is established. This boundary separates the two classes based on the estimated probabilities. Typically, a threshold of 0.5 is employed. If the estimated probability is equal to or greater than 0.5, the predicted class is assigned as the positive class (e.g., diabetes). Conversely, if the estimated probability is below 0.5, the predicted class is assigned as the negative class (e.g., doesn't have diabetes).

Random Forest. The random forest algorithm is a widely machine learning technique, which has also been applied to predict diabetes [8]. The key idea is to combine multiple individual models to make more accurate predictions.

Specifically, the core of the random forest model is the decision tree. A decision tree is structured hierarchically, where each node represents a specific feature or attribute. The branches emanating from each node symbolize decision rules based on the feature values. Finally, the leaf nodes of the decision tree depict the outcomes associated with the given set of features and decision rules. The process of constructing decision trees involves recursively partitioning the dataset according to various feature values, to maximize information gain at each stage.

In the case of random forests, a collection of decision trees is combined to create the model. The key steps are as follows:

• *Random Sampling.* From the original training dataset, random sampling with replacement (known as bootstrapping) is performed to create multiple subsets called "bootstrap samples." Each bootstrap sample is used to train an individual decision tree within the random forest.

• *Feature Subset Selection.* During the construction of each decision tree in the random forest, each node will select a random subset of features. This step introduces diversity and randomness into the ensemble. Typically, the square root of the total number of features is used as a guideline to determine the number of features considered.

• *Decision Tree Training*. Subsequently, decision trees are trained independently on each subset of data. Specifically, the tree model partition training data recursively until meeting a stopping criterion. Stopping criteria include reaching the maximum depth of the tree or having a minimum number of samples per leaf node.

• *Ensemble Prediction.* After training the decision trees, predictions are generated by combining the individual predictions from each tree. In classification tasks such as predicting diabetes, the most commonly employed aggregation method is voting. Each tree "casts a vote" for a specific class, and the final prediction is determined by selecting the class that receives the majority of votes.

3.4. Evaluation Methodology

To evaluate model performance, the dataset is divided into training and testing subsets using a split ratio of 7:3. As a result, 576 records were assigned to the training dataset, while 192 records were assigned to the validation dataset. It is worth mentioning that the label distribution of this dataset is imbalanced, with a skewed proportion.

To mitigate the potential bias caused by this imbalanced label distribution, a method known as up sampling is employed [9]. In particular, the number of positive samples is augmented in the training dataset, aiming to enhance the model's capacity to learn from the minority class.

4. Results

This section presents the experimental results of our diabetes prediction task.

4.1. Results of Missing Value Imputation

The model performance for imputing Insulin and Skin Thickness can be compared using visualization techniques and R-squared correlation values. Regarding Insulin interpolation, Figure 4 presents the QQ plot, scatter plot, and KDE plot to depict the residuals of the interpolated model. The R-squared correlation for insulin is found to be 0.317.

In contrast, Figure 5 shows the estimated model residuals for Skin Thickness, which exhibits a relatively higher R-squared correlation of 0.426. Consequently, the model used for Skin Thickness interpolation demonstrates better performance.

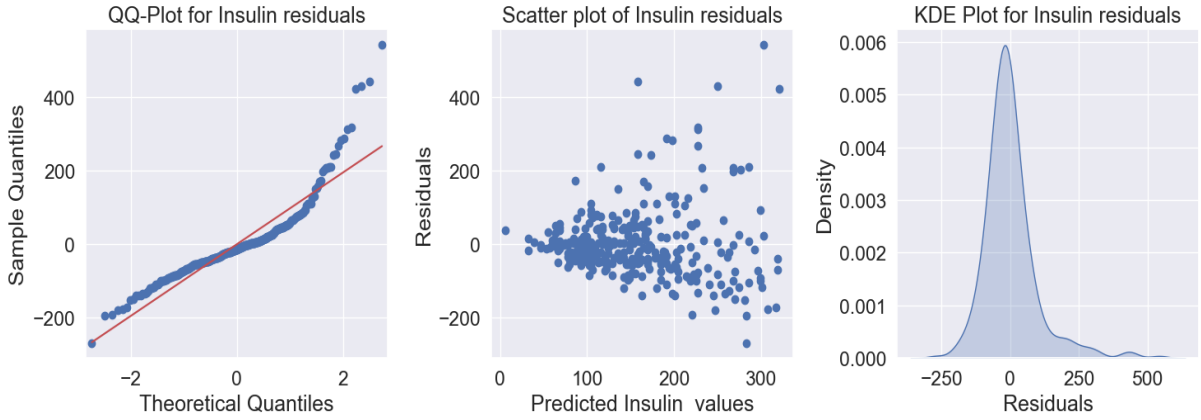


Figure 4. The imputed values for Insulin.

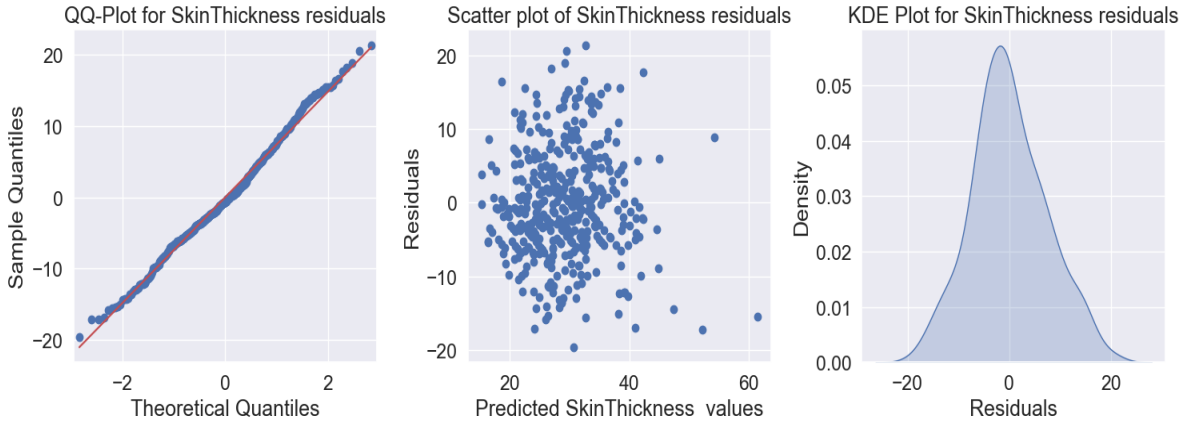
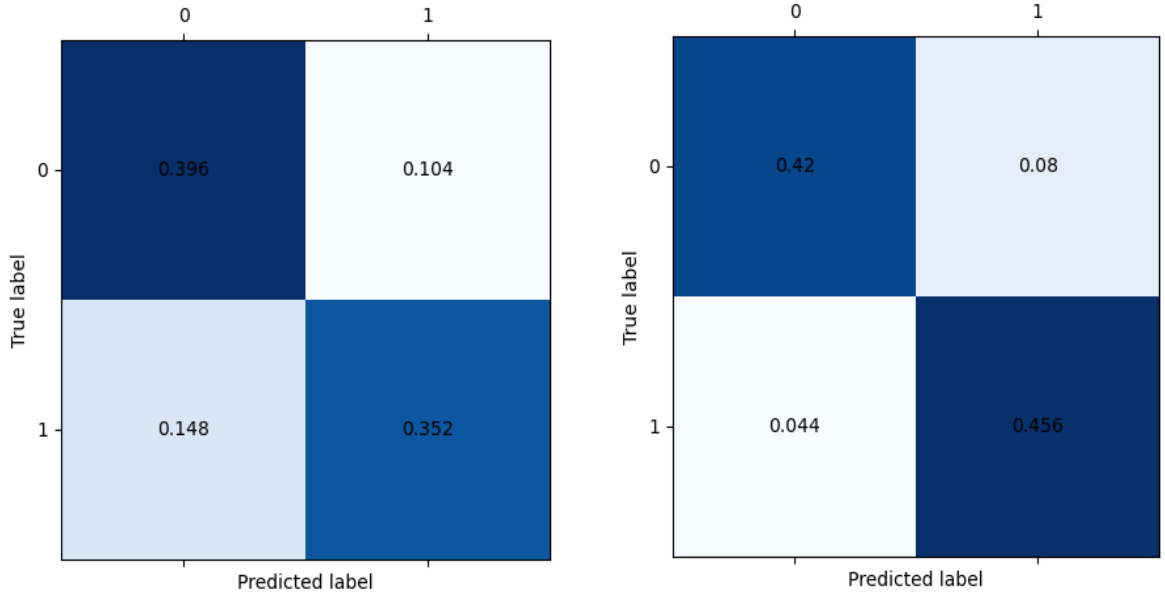


Figure 5. The imputed values for Skin Thickness.

4.2. Results of Diabetes Prediction

Both logistic regression and random forest models were trained on the same training dataset. The logistic regression model utilized the L-BFGS algorithm as the optimization method and employed L2 regularization. The model was trained for a total of 100 iterations.

On the other hand, the random forest model was constructed using 100 decision tree estimators, each considering a maximum of 3 features.



(a) Confusion matrix of logistic regression (b) Confusion matrix of random forest

Figure 6. The model performance.

The performance of the logistic regression and random forest models is presented in Figure 6. The logistic regression model achieves an accuracy of 74.8%. On the other hand, the random forest model exhibits superior performance compared to the logistic regression model, achieving an accuracy of 83.6%. These results indicate that the random forest model was more effective in making accurate predictions in the given context.

5. Conclusion

This paper focuses on predicting diabetes using the random forest algorithm. To ensure the completeness of the data, it begins by performing missing value imputation. From the visualization in Figure 5, it can be observed that the imputed values for Skin Thickness exhibit a distribution close to normal. This suggests a positive correlation between Skin Thickness and BMI (Body Mass Index).

In contrast, the distribution of Insulin values, as depicted in Figure 4, shows a long-tail pattern. Additionally, there is a high correlation between Insulin and Glucose levels. Regarding model performance, it can be found that the random forest model significantly outperforms the logistic regression model. This highlights the superiority of tree-based models, such as random forest, for predicting diabetes compared to logistic regression.

References

[1] Smith JW, Everhart JE, Dickson WC, et al. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the annual symposium on computer application in medical care, pages 261 – 265, 1988.

[2] Kahramanli H and Allahverdi N. Design of a hybrid system for the diabetes and heart diseases. Expert System with Applications, 35 (1-2): 82 – 89, 2008.

[3] Hasan MK, Alam MA, Das D, et al. Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8: 76516 – 76531, 2020.

[4] Kayaer K, Yildirim T, et al. medical diagnosis on pima indian diabetes using general regression neural networks. In Proceedings of the international conference on artificial neural networks and neural information processing, volume 181, page 184, 2003.

[5] Lee CS and Wang MH. A fuzzy expert system for diabetes decision support application. IEEE Transactions on Systems, Man, and Cybernetics. Part B, 41 (1): 139 – 153, 2011.

[6] Lin WC and Tsai CF. Missing value imputation: a review and analysis of the literature (2006-2017). Artificial Intelligence Review, 53 (2): 1487 – 1509, 2020.

[7] Hosmer DW and Lemeshow S. Applied Logistic Regression, Second Edition. Wiley, 2000.

[8] Breiman L. Random forests. Machine Learning, 45 (1): 5 – 32, 2001.

[9] He H and Garcia EA. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21 (9): 1263 – 1284, 2009.