Comparative Analysis of Diabetes Prediction Models Using the Pima Indian Diabetes Database

Yize Zhao

Department of Statistical Science, University College London, WC1E 6BT London, UK

Abstract. This study provides an in-depth review and comparison of diabetes prediction models using the Pima Indian Diabetes database. The main aim is to contrast and evaluate the performance of two distinct predictive models: K-means clustering and Random Forest. The research begins by introducing the significance of accurate diabetes prediction and the methodologies used in the analysis. The K-means model operates by grouping data points into separate clusters according to their characteristics, achieving an accuracy of 90.04% in diabetes prediction. In comparison, the random forest model, which builds multiple decision trees (DT) to do their predictions, demonstrates superior performance over several widely used algorithms such as K-Nearest Neighbours (KNN), Logistic Regression (LR), DT, Support Vector Machines (SVM), and Gradient Boosting (GB). The study reveals that while both models are effective, the Random Forest model provides enhanced predictive accuracy. These findings underscore these models' potential for use in real-world medical diagnosis, where they can assist in identifying people at risk of diagnosing diabetes and starting early prevention. Future research directions include further refinement of these models and their application to larger and more diverse datasets to improve prediction accuracy and generalizability.

1 Introduction

In modern society, due to its extensive spread in children and adults, diabetes has emerged as one of the most significant public health issues [1]. The number of individuals diagnosed with diabetes has experienced a rapid growth from 108 million to 422 million from 1980 to 2014 [2]. In 2019, diabetes caused about 1.4 million people to lose their lives [2]. Diabetes mainly has 2 types: Type 1 diabetes and Type 2 diabetes. Type 1 diabetes happens because there is not enough insulin produced in the body [3]. Type 2 diabetes is triggered by cells that cannot use insulin [3]. Under the condition that diabetes has a negative influence on people's lives, it is of great importance to use the existing data to predict diabetes, which can show people whether there are possibilities for them to have risks of having diabetes.

In this area, with the rapid development of artificial intelligence, scientists have created various algorithms and methods. There are numerous related work and progress. Lai et al. [4] conducted predictive models such as Gradient Boosting Machine (GBM) techniques as well as Logistic Regression by applying the area under the receiver operating characteristic curve

Corresponding author: zcakzhc@ucl.ac.uk

(AROC) to compare different models' accuracy. The results turn out that the GBM model's AROC is 84.7%, whose sensitivity is 71.6% and the AROC for the logistic regression is 84.0%, its sensitivity is 73.4% [4]. In addition, another study illustrates that the random forest has a high accuracy using more than 60000 people as the dataset, with the number of 0.8084 [5]. In addition, using logistic regression, artificial neural networks (ANN), regression trees, and classification to forecast diabetes is another example [6]. These studies have a huge contribution to the enhancement of predicting diabetes. Based on these studies, it can be shown that the main method used today to have prediction is machine learning, which is a branch of artificial intelligence. Machine learning is a method that trains machines to more effectively deal with the data [7]. Although there is some progress, some limitations still exist. For instance, it is crucial to find the right classifier, data mining technique, and attributes [5]. Also, for some methods, the accuracy and data validity are not high enough to be applied to real life [8].

This study aims to explore and evaluate diabetes prediction models. It begins with an introduction to diabetes, including relevant definitions and background information. The core of the study involves analysing various machine learning prediction models, with a focus on K-means clustering and Random Forest models. The paper provides a detailed comparison of these models, discussing their respective advantages and limitations. Additionally, it considers future developments in predictive technologies and their implications for diabetes prevention. The article is structured as follows: the first section covers background information and definitions; the second section examines the principles of the predictive methods; the third section mainly illustrates the experimental results and discusses them; the final section shows the conclusion and future perspectives. This research offers practical insights into the application of predictive models for identifying high-risk individuals and devising prevention strategies for diabetes.

2 Methodology

2.1 Dataset description and preprocessing

Currently, the dataset mainly applied to the prediction of diabetes is the Pima Indians Diabetes Database [9]. In this dataset, several medical predictor variables are introduced to predict diabetes. The outcome of prediction depends on these variables, including the times of pregnancies people have, blood sugar, blood pressure, skin thickness, insulin, body mass index (BMI), diabetes pedigree function, and age of people. These variables have a relationship with diabetes. By analysing these diagnostic data, this dataset can be used by models to forecast whether a person is diagnosed with diabetes.

2.2 Proposed approach

This article mainly analyses and discusses different models predicting diabetes. In this way, this part primarily describes the technologies applied in the prediction area, including main ideas and main modules. There are different kinds of models in this area for prediction. In this way, this part introduces different models in detail and illustrates the process of their prediction work, including their specific characteristics, meanings, and structures. By introducing the principles and comparing the performances of various models used in diabetes prediction, specifically the K-mean model and random forest model in this article, their benefits and drawbacks can become clearer. Fig. 1 shows the structure of this article. In addition, the figures of the process of different models will be illustrated for visualization in Fig. 2 and Fig. 3.

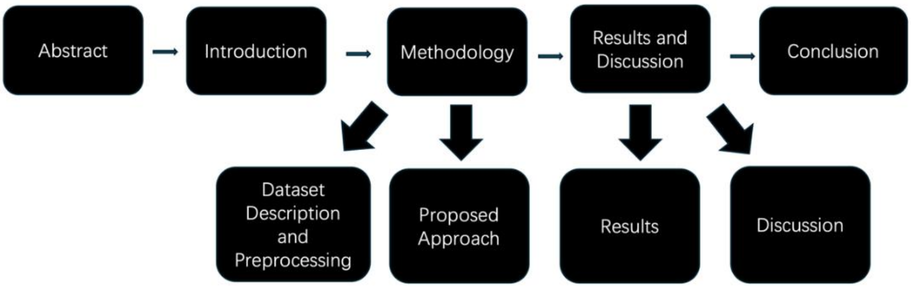

Fig. 1. The structure of this article (Picture credit: Original).

2.2.1 K-means model

K-means is one of the most often used clustering algorithms established independently by various scientists about 50 years ago [10]. For a given dataset, K-means use a certain number of clusters to classify it. K-means randomly select K objects as the K initial cluster centre. Then, it will use Euclidean distance to connect each point of a dataset with the nearest cluster centre. After that, it will repeat the process [10]. The decision tree algorithm is used in data mining. It is simple to construct and change [10]. Fig. 2 shows the pipeline of this model:

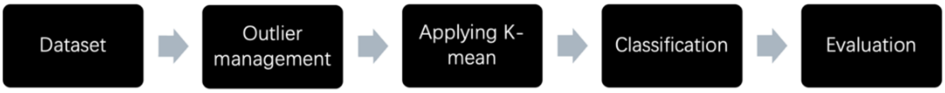

Fig. 2. The pipeline of K-means model (Picture credit: Original).

First, this model deals with outliers. Mean values are substituted for all missing and impossible values. [10]. After that, data reduction is conducted. The algorithm is used to delete the wrong samples to create a basis for applying classification. After that, the decision trees are applied and do classification. In this Pima dataset, about 236 values have the wrong classification [10]. In this way, the decision tree uses 532 instances, where the 236 values are subtracted from the total number [10]. Eventually, the evaluation will be tested to determine how the model performs and decide whether this model is a good fit for predicting diabetes.

2.2.2 Random forest model

From Liberian [11], there is a definition of a random forest. Random forest is a classifier made up of a number of tree-structured classifiers that are random and independent identically distributed, and each tree votes individually for the most popular class at input x. Random Forest is a method that can enhance the performance of a decision tree [12]. More specifically, it selects a certain amount of data from the whole dataset, then it repeatedly splits the node to create a certain number of decision trees for further prediction. The random forest mainly builds numerous decision trees during the training time. Then it outputs the classification or mode of classes or prediction of individual trees [12]. Fig. 3 below describes the process of how the random forest model works.

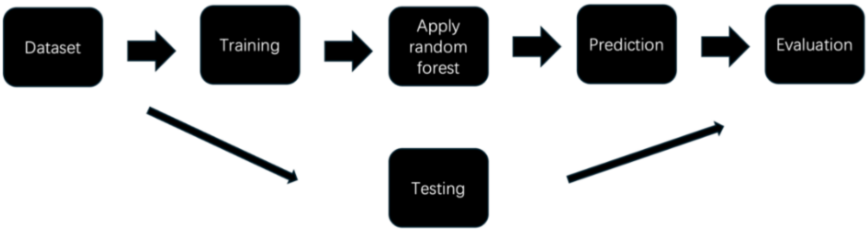

Fig. 3. The pipeline of random forest model (Picture credit: Original).

In the dataset, the initial stage is to analyse the data in the dataset and classify them. Then, several decision trees are created in different random forests. After that, the model conducts the prediction using the data and decision tree created before. Finally, evaluation is processed to find out whether the performance of this model is suitable for predicting diabetes.

3 Result and Discussion

3.1 Results

Before applying the K-means model, the Pima dataset needs to be dealt with. The variables in the Pima dataset are assigned different numbers and the means are calculated. Chen et al. [10] collected these data, which will be shown in Table 1. After dealing with the Pima dataset, valid 532 instances are selected for the experiment [10]. The results of diabetes prediction are illustrated in Table 2. Based on Table 2, the accuracy of this model is 90.04% [10], which shows this model has a decent accuracy. In the random forest model, comparing the result with other algorithms, the random forest algorithm has better accuracy [12]. For the feature importance, it is shown that glucose contributes the highest feature importance [12].

Table 1. Conduct on the Pima dataset [10].

Number	Variables	Mean
1	Number of pregnancies	3.80
2	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	120.90
3	Diastolic blood pressure (mm Hg)	69.10
4	Triceps skin fold thickness (mm)	20.50
5	2-Hour serum insulin (mu U/ml)	79.80
6	BMI	32.00
7	Diabetes pedigree function	0.50
8	Age (years)	33.20

Table 2. The result of diabetes prediction using K-means model [10].

	Predicted Class	Yes	No
Actual Class	/	/	/
Yes	/	144	21
No	/	32	335

3.2 Discussion

Both the K-means clustering model and the random forest model demonstrate strong accuracy in diabetes prediction when applied to the Pima dataset. This indicates that both models are effective for this specific dataset. However, it is of significance that the reliability of these results may be limited by the fact that only one dataset was used in the experiments. The generalizability of the models could be enhanced by testing them on a variety of datasets, which would provide a more comprehensive assessment of their accuracy and robustness across different populations and conditions.

In the experiments conducted, the K-means model performed a binary classification, distinguishing between two categories. While this approach is effective for simple classification tasks, there is potential for improvement by exploring multi-class classification scenarios. Incorporating additional class labels and examining the model's functionality in more complicated classification tasks could provide deeper insights into its capabilities and limitations. Future research could focus on expanding the classification framework of the K-means model to handle multiple classes and evaluate its effectiveness in a wider range of applications. This approach would both improves the model's versatility and contributes valuable knowledge to the field of diabetes prediction. Furthermore, integrating these models with other predictive techniques or combining their strengths in ensemble methods could also be an area of exploration. This could potentially lead to improved predictive accuracy and robustness. Overall, while the K-means and Random Forest models show promise, further experimentation with diverse datasets and expanded classification strategies is essential for advancing their applicability and reliability in diabetes prediction.

4 Conclusion

This study provides a comprehensive exploration and discussion of diabetes prediction models, focusing on two primary approaches: the K-means clustering model and the random forest model. The K-means model operates by randomly selecting K initial cluster centres and then each data point is supposed to be assigned to its closest cluster centre. This process is repeated until the cluster centres are stable, ultimately providing predictions based on the clustering results. In contrast, the random forest model constructs multiple decision trees from randomly sampled subsets of the data and aggregates their predictions as reference to finalize the decision. By combining the strengths of multiple decision trees, this ensemble technique improves prediction accuracy. The Pima dataset was applied to assess both models, and the findings show that they both demonstrate strong accuracy in diabetes prediction. Notably, the Random Forest model outperforms several widely-used algorithms, such as K-Nearest Neighbours (KNN), Logistic Regression, Decision Trees (DT), Support Vector Machines (SVM), and GB, in terms of accuracy. Looking forward, future research will focus on advancing both models. For the K-means model, this involves exploring multi-class classification to enhance its versatility. For the random forest model, to increase its forecast accuracy, it will be used for bigger and more complicated datasets. These enhancements aim to refine the models' performance and broaden their applicability in diabetes prediction.

References

1. F.A. Khan, et al. Detection and prediction of diabetes using data mining: a comprehensive review. IEEE Access 9, 43711-43735 (2021)
2. World Health Organization, “Diabetes”, 2023, Retrieved on 2024, Retrieved from: https://www.who.int/news-room/fact-sheets/detail/diabetes

3. R. Kumar, et al. A review on diabetes mellitus: type1 & Type2. World Journal of Pharmacy and Pharmaceutical Sciences 9(10), 838-850 (2020)
4. H. Lai, et al. Predictive models for diabetes mellitus using machine learning techniques. BMC endocrine disorders 19, 1-9 (2019)
5. Q. Zou, et al. Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics, 9, 515 (2018)
6. L.Y. Zhang, et al. Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: The Henan Rural Cohort Study. Scientific reports, 10(1), 4406 (2020).
7. B. Mahesh, Machine learning algorithms-a review. International Journal of Science and Research, 9(1), 381-386 (2020).
8. H. Wu, et al. Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked, 10, 100-107 (2018).
9. J.W. Smith, “Pima Indians Diabetes Database”, 2016, Retrieved on 2024, Retrieved from: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
10. W.Q. Chen, et al. A hybrid prediction model for type 2 diabetes using K-means and decision tree. IEEE international conference on software engineering and service science (2017)
11. L. Breiman, Random forests. Machine learning, 45, 5-32 (2001).
12. M. Soni and V. Sunita, Diabetes prediction using machine learning techniques. International Journal of Engineering Research & Technology, 9(9), 2278-0181 (2020).