Diabetes Risk Assessment: A Comparative Study of Decision Trees and Ensemble Learning Models

Tianxing Lei

School of Information Science and Engineering, Yunnan University, 650091 Yunnan, China

Abstract. Diabetes poses a significant threat to global health, making accurate prediction and effective treatment of the disease critical. This study explores the application of machine learning algorithms in assessing diabetes risk, with a particular focus on Decision Trees (DT) and Ensemble Learning techniques. DT methodically evaluate various indicators that impact classification outcomes, using sequential decisions to classify each indicator based on the results of previous classifications. This process ensures that all possible combinations of indicators are mapped to a single classification result. Ensemble Learning, on the other hand, leverages multiple classifiers with assigned weights to form a robust ensemble. Each classifier provides its prediction, and the final classification result is derived from a weighted voting mechanism based on the performance of each learner. The study's experimental results demonstrate that applying Principal Component Analysis (PCA) to preprocess the data, followed by training a Random Forest (RF) model with 80% of the dataset, achieves an impressive accuracy of 89.86%. This high accuracy highlights the effectiveness of machine learning algorithms in predicting diabetes risk. The findings underscore the potential of these methods in enhancing diabetes management and offer a valuable contribution to the field of medical predictive analytics.

1 Introduction

Diabetes is an illness that raises blood glucose levels because of abnormalities in either the production of insulin, its function, or both. It is a prevalent illness across all age groups. A diabetic's body is unable to generate or use insulin, the hormone that "unlocks" cells, effectively. of the body, enabling the arrival of glucose to power them. A diabetic's chance of developing other illnesses such renal disease, heart disease, nerve damage, blood vessel damage, and blindness illness. Type 1 diabetes, which is insulin-dependent, and type 2 diabetes, which is not diabetes treated with insulin [1].

With the advancement of medical detection technologies in recent years, a substantial volume of health data has been produced. Processing the data produced by medical tests can yield important insights into a number of disorders. In addition to being utilized for other

Corresponding author: leitianhang@stu.ynu.edu.cn

© The Authors, published by EDP Sciences. 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/).

reasons, these data points can help medical facilities identify diseases more accurately [2]. Disease prediction has made extensive use of a variety of Artificial Intelligence (AI) methods during the last 20 years, including machine learning and deep learning. One such example is the use of logistic regression to predict the risk of heart disease and hence achieve early identification of the condition. The study of diabetes prediction will also help medical institutions to provide more accurate prevention and treatment suggestions for patients, and more effectively protect people's health. Compared to machine learning, deep learning algorithms represented by deep neural networks (DNNs) have more powerful big data analysis capabilities. Even if the amount of data processed is very large, the efficiency of deep neural networks will not be greatly affected. Recursive neural networks (RvNNs), recurrent neural networks (RNNs), and convolutional neural networks (CNN) are the three most well-known varieties of deep learning networks [3].

According to Socher team's explanation, RvNNS can train computers to understand the recursive relationships inherent in natural scenes, human language, and other things [4]. Afterward, based on the learned recursive relationships, objects such as buildings and vehicles in photos, as well as different grammatical components in human language, can be freely split and combined. RNNs typically have three layers: an input, an output, and a hidden layer. They also incorporate cyclic connections between the hidden levels, making them resemble short-term memory units [3]. RNNs can recognize long-term dependencies and dynamic actions in sequential data because to this architecture. RNNs are widely used in speech recognition, time series prediction, and natural language processing (NLP) for precisely this reason. CNN is the most widely utilized algorithm in the deep learning space. The primary benefit of CNN over other algorithms is its ability to autonomously identify pertinent characteristics without human oversight [5]. Three main advantages of CNN were noted by Goodfellow et al. [6]: parameter sharing, sparse interactions, and comparable representations. CNN uses local connections and weight sharing to its greatest potential, making full use of two-dimensional input data structures like picture signals. The CNN training procedure is quite straightforward since this operation utilizes few parameters, but the network speed is also very quick. Multiple convolutional layers, pooling layers, and termination layers make up a typical CNN.

This paper first discusses the related concepts of diabetes prediction and introduces the related technologies. Next, this paper discusses some machine learning-based diabetes prediction technologies in depth and discusses the principles, advantages, and disadvantages of these technologies. This chapter introduces the concept of diabetes prediction and analyzes some algorithms that can be used for this prediction. Chapter 2 analyzes the core concepts and principles of the methods used in this article, followed by an analysis and discussion of experimental results in Chapter 3, and finally a summary in Chapter 4.

2 Methodology

2.1 Dataset description

Pima Indians Diabetes Datasets [7] are selected as the research object in this study. This data set is from the National Institute of Diabetes, Digestive and Kidney Diseases, an authoritative medical research institution. This dataset is highly reliable and credible and is widely used in many related studies. There are 768 data items in the diabetes dataset, including 8 medical predictive variables and 1 result variable. The specific medical characteristics are shown in Table 1.

Table 1. The specific meanings of various indicators in the dataset.

Pregnancies	The information about the number of pregnancies females had to date.
Glucose	The glucose level of the patient is generally higher, and glucose levels show the chances of sugar.
Blood Pressure	Blood pressure data of the patient.
Skin Thickness	Skin thickness of patient.
Insulin	Insulin level of the patient.
BMI(Body Mass Index)	A technique for measurement that divides people into four groups: underweight, normal weight, obese, and overweight.
DPF	Diabetes Pedigree Function.
Age	Age of patient
Outcome	This input uses 0 to indicate that the respondents do not have diabetes, and 1 to indicate that the respondents have diabetes.

2.2 Proposed approach

Diabetes prediction has become more and more dependent on sophisticated computer technologies, namely machine learning and deep learning. Both strategies provide the capacity to draw conclusions and predictions from data, however depending on the size of the dataset, they may or may not be applicable. Machine learning algorithms can perform effectively with smaller datasets, whereas deep learning algorithms typically require larger datasets and more substantial computing resources. As computer performance continues to improve and diabetes-related datasets expand, deep learning algorithms are gaining wider adoption in the field. This paper focuses on the core concepts and distinguishing features of several classic machine learning algorithms, specifically the Decision Tree Algorithm and the Random Forest Algorithm. By analysing these algorithms, the paper aims to shed light on their effectiveness in diabetes prediction, particularly in scenarios where data availability and computational power vary. The research methodology and process are illustrated in Fig. 1, providing a clear overview of the study's approach.

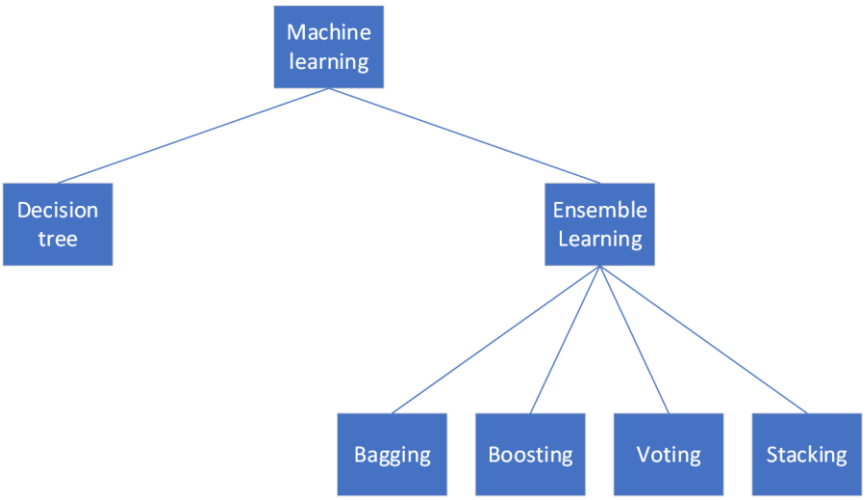

Fig. 1. Research process (Picture credit: Original).

2.2.1 Machine learning

Supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning are the four subcategories of machine learning. These four learning methods are applied in different scenarios, among which both supervised and semi-supervised learning can be used for prediction. In supervised learning, each training data has a clear label indicating whether this attribute is a basis for evaluation or a judgment result. This paper will take supervised learning as an example to illustrate the principle of using machine learning to predict diabetes. Firstly, all machine learning models must extract features from various aspects of the training data and determine the labels that are used to be evaluated for the research object. For example, the purpose of this paper is to assess the risk of diabetes in different people according to their physical conditions. Therefore, the characteristics of the training data extracted in this experiment are the physical conditions of all aspects of the subjects mentioned in 2.1 and whether they are ill or not. The label to be assessed is to predict the risk of diabetes in the subjects. For supervised learning, it is necessary to distinguish which attributes in the training data are used as evaluation criteria and which attributes are given as judgment results. For the 9 items mentioned in section 2.1, it is clear that the first eight items are used as evaluation criteria, while the outcome is the given judgment result.

After identifying the tags to be evaluated and giving the training data, the machine learning algorithm calculates a correlation model between the physical indicators of the respondents and whether they have diabetes. This process is similar to asking the machine learning model to vote for various physical indicators. The more votes a physical indicator gets in this process, the greater the impact of the health of this physical indicator on the risk of diabetes of the subject. In addition, once this model is created, all its predictions and results can also be used as new training data for it to learn on its own, and the accuracy will be further improved in this process. The process of machine learning can be represented by (Fig. 2):

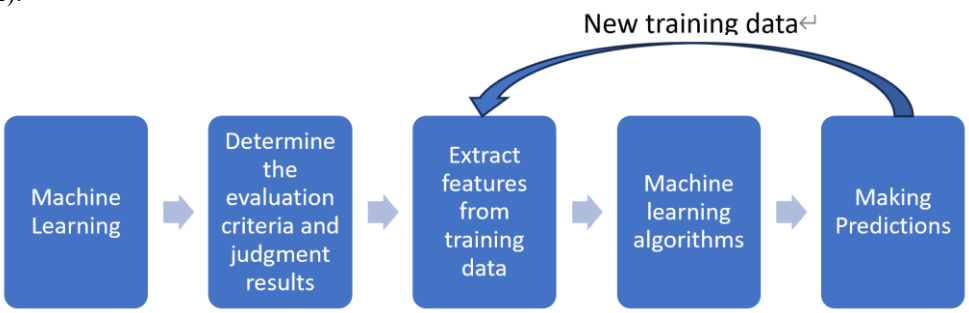

Fig. 2. The process of machine learning (Picture credit: Original).

2.2.2 Decision tree (DT)

An algorithm for supervised learning is the decision tree algorithm. To accomplish the tasks of data screening and decision-making, it makes use of the concept of categorization to build mathematical models based on the properties of the data. Hunt et al. presented this method in 1966. C4.5, ID3, and Classification and Regression Tree (CART) are among the decision tree algorithms that are often utilized.

A predictive analytic model that may be represented as a tree structure, including binary and multi-branch trees, is the decision tree algorithm. Every leaf node maintains a category, every branch represents the output of this feature attribute within a specific range, and every non-leaf node represents a test on a feature attribute. Taking the risk assessment of diabetes in this paper as an example, the first consideration may be the value of Body Mass Index

(BMI), so BMI is the root node of the decision tree. Next, if the BMI is higher than 23.9, the next test may be age: if the age is higher than 60, the risk of diabetes is higher; If the age is less than 20 years old, the risk of diabetes is low; If the age is between 20-60 years old, proceed to the next test, such as the number of pregnancies. For BMI in other ranges, decision trees also have corresponding classification methods. The number of pregnancies will also appear in the decision tree as an internal node, and the risk of diabetes will be saved in the leaf node as the output result, as shown in Fig. 3.

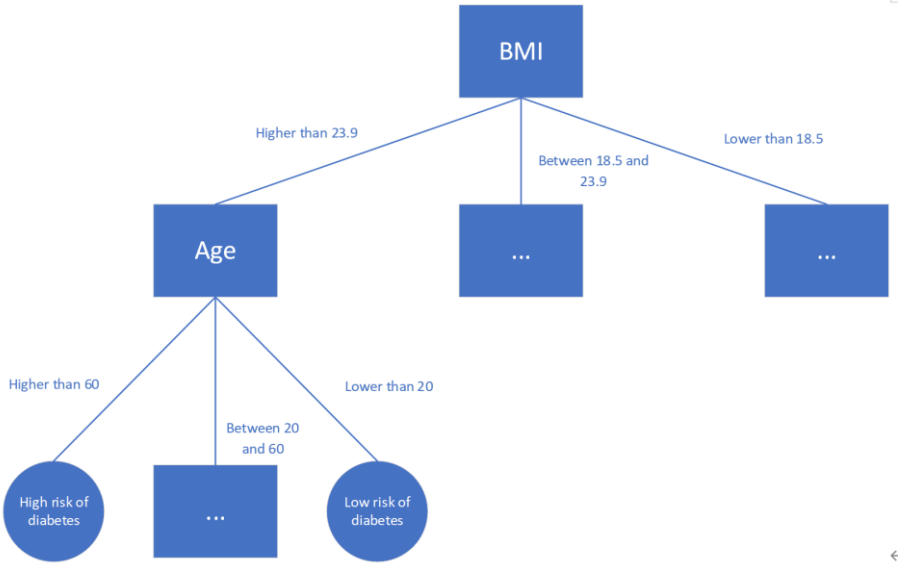

Fig. 3. Decision tree example (Photo/Picture credit: Original).

Finally, the decision tree will complete the classification of all input data combinations and give the corresponding combination's risk of diabetes. A decision tree typically has a root node, a number of internal nodes, and a number of leaf nodes. Information entropy is used by the decision tree method as an indicator to choose features; the higher the information entropy, the more selectivity that attribute has. Put another way, the decision tree method will give preference to classifying a collection of items with classification into the class with the largest information entropy when there are numerous classes available. The formula for calculating information entropy is as follows:

$$I(x = x_i) = -\log_2 p(x_i) \tag{1}$$

The information entropy of a random variable is denoted by $I(x)$, and the probability that x_i will occur is denoted by $p(x_i)$

The decision tree method has the benefit of having a low computing complexity and being simple to translate into classification rules. The splitting circumstances at each stage, from the tree roots to the leaves, can specifically identify a predicate for categorization. Additionally, the categorization rules it mines are clear and have a high accuracy. Which fields are more essential may be seen in the decision tree. Furthermore, decision trees do not require parameter settings because they are a non-parametric learning method. Decision tree algorithms, however, are extremely sensitive to samples and prone to overfitting. Occasionally, even a small alteration to the sample might have a significant impact on the overall tree structure.

2.2.3 Ensemble learning

According to Pelin Yildirim Taser's explanation [8], ensemble learning no longer uses a single learner to predict the attributes of the target, but instead consists of multiple learners forming a strong classifier. Combine the outputs of each learner using a voting mechanism to make the final class label prediction. Compared to using a single learner prediction algorithm, the prediction results of random forests are often more accurate. The four primary types of ensembles learning approaches are voting, stacking, boosting, and bagging.

Bagging is a popular ensemble learning approach that is used for bootstrap aggregation. It generates several training sets using bootstrap techniques. Accurate sample from the original dataset are chosen at random to create numerous training sets in the Bootstrap framework. Following the creation of training subsets, each learner within the ensemble structure is trained using these subsets to build numerous learning models. Ultimately, the ultimate determination is derived from a summary of each model's forecast outcomes. Leo Breiman and Adele Cutler's standard bagging approach is the random forest algorithm [9]. The core idea of Boosting is to train weak learners one by one and add them to the prediction model. After each training session, adjust the distribution of the data based on the idea of valuing data that is inconsistent with the predicted results and ignoring data that is consistent with the predicted results. After training, assign different voting weights to each learner based on their accuracy, and use the weighted voting results of all learners as the final prediction result.

Voting is an ensemble learning model that integrates numerous models to increase model resilience and minimizes variance by adhering to the concept of the minority following the majority. Voting should, in theory, outperform all base models in terms of prediction accuracy. The impacts between the basic models must not differ significantly in order to be voted on. When a certain base model performs poorly compared to other base models, it is likely to become noise [10]. The limitation of voting is that all models contribute equally to the prediction. Stacking first divides the dataset into multiple subsets, and then uses different base models for training and prediction for each subset. Afterward, the predicted results of each basic model will be used as new features and combined into a new training dataset. These predicted results serve as supplements to the original features, providing more information to train the meta-model. After combining the new dataset, use the new training dataset to train the final metamodel.

3 Result and Discussion

True positive (TP), false positive (FP), true negative (TN), and false negative (FN) are the four categories into which the prediction results are often separated in the diabetes prediction process based on the comparison between the prediction findings and the actual results. The four prediction outcomes mentioned above were subjected to equation (2–5), yielding the following four metrics for assessing algorithm performance: Accuracy denotes the model's prediction accuracy; Recall is the model's capacity to accurately identify every person who is impacted; The precision of a model's positive forecast indicates its trustworthiness. Recall and accuracy are balanced using the F1 Score.

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{2}$$

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{3}$$

$$\Pr ecision = \frac{TP}{TP + FP} \tag{4}$$

$$F1 - Score = \frac{2TP}{2TP + FN + FP} \tag{5}$$

This article is evaluated both before to and during the Principal Component Analysis submission (PCA). accompanying the creation of the Pima diabetes dataset, the accompanying tables (Tables 2 and 3) display the outcomes of four classifiers: support vector machine (SVM), naïve Bayes (NB), random forest (RF), and decision tree (DT). Table 2 displays the 70% training and 30% testing outcomes based on the acquired categorization findings. The outcomes of 80% training and 20% testing are displayed in Table 3.

Table 2. All model performance for the 70% and 30% of training and testing ratio.

Algorithm	Accuracy	Precision	Recall	F1-Score	AUC
Without Using PCA					
SVM	76.32	72.53	77.65	70.91	85.49
NB	75.42	71.76	78.16	75.55	87.63
RF	82.42	85.33	74.62	77.38	79.32
DT	73.57	68.55	71.43	70.87	74.65
Using PCA					
SVM	85.54	87.33	86.47	89.12	91.13
NB	84.54	88.55	86.42	84.76	88.55
RF	88.76	87.33	89.13	90.22	92.43
DT	80.38	74.32	74.73	74.30	76.77

Table 3. All model performance for the 80% and 20% of training and testing ratio.

Algorithm	Accuracy	Precision	Recall	F1-Score	AUC
Without Using PCA					
SVM	79.02	74.43	78.03	71.33	87.22
NB	78.18	73.43	78.92	76.53	87.87
RF	83.65	86.98	75.65	80.02	77.94
DT	72.55	71.12	73.04	72.01	80.82
Using PCA					
SVM	86.08	88.88	86.90	88.65	92.91
NB	89.54	88.23	85.93	85.83	92.33
RF	89.86	89.18	89.77	89.91	93.72
DT	82.02	84.65	73.91	73.92	87.55

It is clear from comparing the data produced by the two tables that Random Forest performs the best out of these four algorithms, irrespective of the training and testing ratios that are employed. Improving classification ability may be achieved by reducing dimensionality. While naive Bayes is usually quite dependable, decision trees perform somewhat when employing PCA, yet accuracy and area under the curve (AUC) are sometimes low. It is clear from the data in Tables 2 and 3 that the best machine learning method should be chosen in accordance with the particular goals and characteristics of the dataset.

4 Conclusion

This paper's main goal is to create a classification model that works well for quickly and accurately assessing diabetes risk. The Pima Indian diabetes dataset greatly improves the performance of the classification model by feature selection and outlier reduction, which are made possible by the application of PCA. After preprocessing the data, several classifiers were applied to the training and testing datasets, such as SVM, Random Forests, Naive Bayes, and Decision Trees. According to the findings, using 80% of the dataset produced an astounding 89.86% accuracy rate. This great degree of accuracy highlights the suggested method's dependability in determining diabetes risk, which is essential for enhancing patient outcomes and illness treatment. Accurate diabetes risk assessment is vital for early intervention and personalized treatment plans, which can greatly enhance patient prognosis. Future research will aim to refine the classification model further by focusing on more granular distinctions among patients who test positive for diabetes. This will involve examining anomalies across various indicators to provide more detailed insights, thereby offering valuable references for medical institutions in treatment and management strategies. The continued development and validation of these models hold the potential to advance the field of diabetes diagnosis and significantly benefit patient care.

References

1. K. Gaganjot, Improved J48 Classification Algorithm for the Prediction of Diabetes, International Journal of Computer Applications, 98(22), 13-17 (2014)
2. C.Y. Zeng, W. Ke, B.W. Zhi, X.X. Shu, H.L. Zhi, Cluster Computing 26, 1231–1251 (2023)
3. A. Laith, L.Z. Jing, J. Amjad, A.D. Ayad, D. Ye, A.S. Omran, J. Santamaría, A. Mohammed, A.A. Muthana, and F. Laith, Review of deep learning: concepts, CNN architectures, challenges, applications, future directions, Journal of Big Data (2021)
4. S.R. Richard, C.Y. Cliff, C.Y. Lin, Y. Andrew, D. Christopher, Parsing natural scenes and natural language with recursive neural networks (2011)
5. J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, et al. Recent advances in convolutional neural networks. Pattern Recogn.77, 354–77 (2018)
6. I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, Deep learning, nature, 521(7553), 436-444 (2015).
7. M. Tahir, “Diabetes Dataset For Beginners”, 2023, Retrieved on 2024, Retrieved from: https://www.kaggle.com/code/mehmettahirmemili/ml-knn-k-nearest-neighbors
8. Y.T. Pelin, Mutations in the nebulin gene associated with autosomal recessive nemaline myopathy, Proceedings of the National Academy of Sciences, 74(1), 6 (2021)
9. B. Leo, Random Forests. Machine Learning, 45. 5–32 (2011)
10. S.S. Merdin, K.I. Rowaida, R.M. Subhi, A.Z. Dilovan, M. Lozan, N.M.A. Abdulrahman, Diabetic Prediction Based on Machine Learning Using PIMA Indian Dataset. Communications on Applied Nonlinear Analysis, 31(5), 138-156 (2024)