Proceedings of the 2nd International Conference on Modern Medicine and Global Health
DOI: 10.54254/2753-8818/35/20240874

Comparison and analysis of multiple machine learning algorithms on prediction accuracy in Parkinson's patients

Yutong Yan

School of Chemical and Environmental Engineering, China University of Mining and Technology (Beijing), Beijing, 100083, China

YanYutong1234@outlook.com

Abstract. This paper describes an experiment on Parkinson's disease classification using multiple classification algorithms for comparison. Parkinson's disease is a common neurological disorder, and early diagnosis and classification are important for the assessment of treatment and prognosis. Therefore, the research implications of this paper are clear. The classification algorithms used in the experiment include adaboost classification model, XGBoost classification model, logistic regression regression model, random forest plain Bayesian classification model, bp neural network and support vector machine. The experimental results show that adaboost classification model performs well when dealing with small sample data, XGBoost classification model performs well when dealing with large-scale datasets, and logistic regression regression model and random forest plain Bayesian classification model also have good performance. The bp neural network and support vector machine, on the other hand, perform poorly in terms of classification results and require a much larger dataset for support. These experimental results have important reference value for the classification and diagnosis of Parkinson's disease. Different classification algorithms are suitable for different dataset sizes and characteristics, so in practical applications, we can choose different classification algorithms according to the size and characteristics of the dataset to achieve the optimal classification effect. In conclusion, the results of this paper provide a reference for the classification and diagnosis of Parkinson's disease, as well as a guide for choosing appropriate classification algorithms. In the future, we can further expand the dataset size and use more classification algorithms for comparison to improve the accuracy and robustness of Parkinson's disease classification.

Keywords: Parkinson's patients, Machine learning, Prediction accuracy.

1. Introduction

The Parkinson's Disease Classification dataset is a publicly available dataset for Parkinson's Disease classification that contains data on several biomedical characteristics of Parkinson's Disease patients and healthy individuals, including aspects of voice, accelerometers, and fingerprints [1,2]. The dataset, provided by the UCI Machine Learning Repository, has 195 samples, of which 147 are from Parkinson's disease patients and 48 are from healthy individuals [3].

Based on this dataset, researchers can use machine learning algorithms to predict whether a person has Parkinson's disease [4]. This machine learning prediction method can help doctors diagnose Parkinson's disease more quickly and accurately, improving treatment outcomes and quality of life [5].

DOI: 10.54254/2753-8818/35/20240874

In recent years, many studies have used the Parkinson's Disease Classification dataset for machine learning prediction of Parkinson's disease. For example, some researchers have used various machine learning algorithms, including support vector machines, decision trees, and random forests, to classify and predict Parkinson's Disease and compared the performance of different algorithms [6,7]. The results of the study showed that the support vector machine algorithm had the best classification results on this dataset. In addition, some researchers have used various feature selection methods and classifiers for classification prediction of Parkinson's disease and compared the performance of different methods [8]. The results of the study showed that the method using mutual information and random forest classifier gave the best classification results on this dataset.

It was shown that using machine learning algorithms for Parkinson's disease classification prediction on the Parkinson's Disease Classification dataset is feasible and can provide physicians with tools to aid in diagnosis [8]. However, there is a lack of comparisons of various machine learning algorithms in terms of their performance in classification effectiveness. Therefore, in this study, data preprocessing and feature dimensionality reduction were performed on a biomedical feature dataset for Parkinson's disease, followed by dividing the data into equally proportioned training and test sets. Then, 10 machine learning algorithms were used to classify and predict Parkinson's patients, and parameters such as precision, accuracy, recall, and F1 score were calculated to compare and evaluate the classification results of each machine learning algorithm for further discussion and research.

2. Data set introduction

The Parkinson's Disease Classification dataset is a dataset used to help diagnose Parkinson's disease. The dataset is provided by the UCI Machine Learning Repository and can also be found on Kaggle.

The dataset contains 197 samples, of which 147 are Parkinson's disease patients and 50 are healthy individuals. Each sample contains 22 sound features that were extracted from the patient's voice recordings. These features include fundamental frequency, frequency change, Jitter, Shimmer, etc. The dataset is designed to help diagnose Parkinson's disease by using these sound features. This dataset can be used in classification tasks to predict whether a patient has Parkinson's disease or not.

2.1. Data preprocessing

Data preprocessing is the process of processing raw data before performing machine learning or deep learning model training. The process aims to provide better input data for the model to improve the training effect and accuracy of the model. Data preprocessing usually includes the following three aspects of processing:

Data preprocessing is an important step in the training of machine learning and deep learning models, which can improve the training efficiency and accuracy of the model, thus improving the performance and reliability of the model. Data preprocessing usually includes three aspects of data cleaning, data normalization and data dimensionality reduction.

2.2. Data Cleaning

Data cleaning is the first step of data preprocessing, which aims to remove unreasonable or invalid data such as noise, outliers, missing values, etc. from the data. These problems will affect the training effect and accuracy of the model, so it is necessary to use a variety of techniques and tools to clean the data, such as data visualization, outlier detection, and missing value filling. When cleaning data, care needs to be taken to retain the valid information of the data while removing invalid information to ensure the quality and usability of the data.

2.3. Data standardization

Data standardization is the transformation of raw data according to certain rules in order to make them have similar scales and distributions. This avoids bias in comparisons between different features, thus improving the accuracy of the model. Common data standardization methods include Z-score standardization, Min-Max standardization and so on. When performing data standardization, it is

DOI: 10.54254/2753-8818/35/20240874

necessary to choose the appropriate standardization method according to the characteristics of the data and the requirements of the model.

2.4. Data downgrading

Data dimensionality reduction is to compress the feature dimensions in the original data to reduce data complexity and computational cost. Commonly used data dimensionality reduction methods include Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and so on. Data dimensionality reduction can improve the training speed and accuracy of the model and also reduce the risk of overfitting. When performing data dimensionality reduction, it is necessary to choose the appropriate dimensionality reduction method according to the characteristics of the data and the requirements of the model.

When performing feature dimensionality reduction, correlation analysis can be used to select features with high correlation with the target variables. Firstly, the correlation coefficient matrix between features needs to be calculated, and then the features with higher correlation with the target variable are selected according to the size of the correlation coefficient, the larger the absolute value of the correlation coefficient, the higher the correlation between the feature and the target variable. Finally, highly correlated features need to be eliminated because a high correlation coefficient between two features indicates that there is redundant information between them and one of the features can be considered for elimination.

2.5. Data segmentation

In addition to data preprocessing, data segmentation is an important step in machine learning and deep learning model training. Data division can divide the dataset into training set, validation set and test set for model training and testing. When performing data partitioning, it is necessary to choose an appropriate partitioning method according to the characteristics of the data and the requirements of the model, and to ensure that the data distribution and feature distribution of the training set, validation set and test set are similar.

In summary, data preprocessing is very important for the training of machine learning and deep learning models. By preprocessing the data, the training efficiency and accuracy of the model can be improved, thus improving the performance and reliability of the model.

3. Result

In this paper, the Parkinson's Disease Classification dataset is selected as the object of study, aiming at pre-processing the data and data dimensionality reduction based on correlation analysis, as well as predicting the categories of Parkinson's patients using ten machine learning methods, and evaluating the model to explore how to improve the model's accuracy and performance.

When performing data preprocessing, we first cleaned the data to remove unreasonable or invalid data such as noise, outliers, and missing values in the data. Then, we standardized the data to have similar scales and distributions to avoid bias in comparisons between different features. Finally, we downscaled the data using correlation analysis to select features with high correlation with the target variables and eliminate highly correlated features to reduce redundant information and improve the accuracy of the model.

For model training and testing, we used ten machine learning methods including logistic regression, decision tree, random forest, support vector machine, K-nearest neighbor, plain Bayes, gradient boosting, XGBoost, LightGBM, and CatBoost. we trained the model using the training set and tested the model using the test set to compute the model's precision, accuracy, Recall and F1 score and other evaluation metrics to assess the performance and accuracy of the model.

The results are shown in Table 1 and Figures 1-4:

DOI: 10.54254/2753-8818/35/20240874

Table 1. Model evaluation.

Model	Accuracy	Recall	Precision	F1
adaboost	0.744	0.744	0.736	0.734
XGBoost	0.74	0.74	0.731	0.725
Logistic regression	0.722	0.722	0.721	0.687
Random forest	0.714	0.714	0.7	0.696
Naive bayes	0.705	0.705	0.712	0.708
KNN	0.687	0.687	0.678	0.626
Gradient lifting tree	0.674	0.674	0.667	0.669
Decision tree	0.665	0.665	0.654	0.658
bp neural network	0.656	0.656	0.431	0.52
Support vector machine	0.573	0.573	0.517	0.533

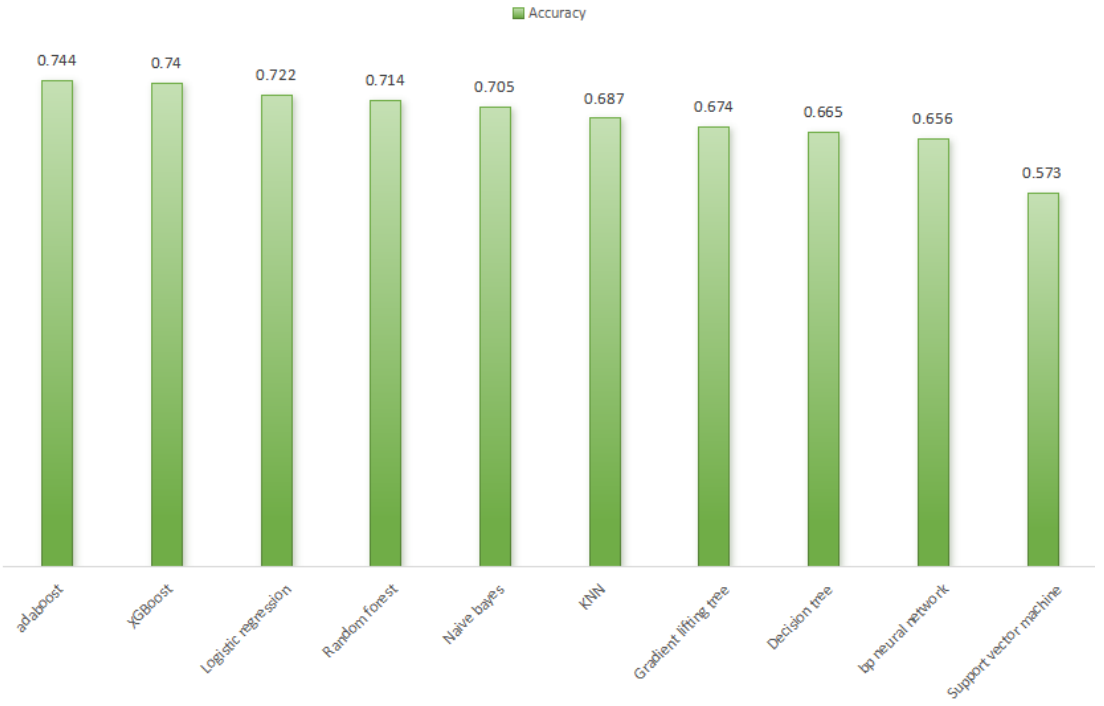

Figure 1. Accuracy.(Photo credit : Original)

DOI: 10.54254/2753-8818/35/20240874

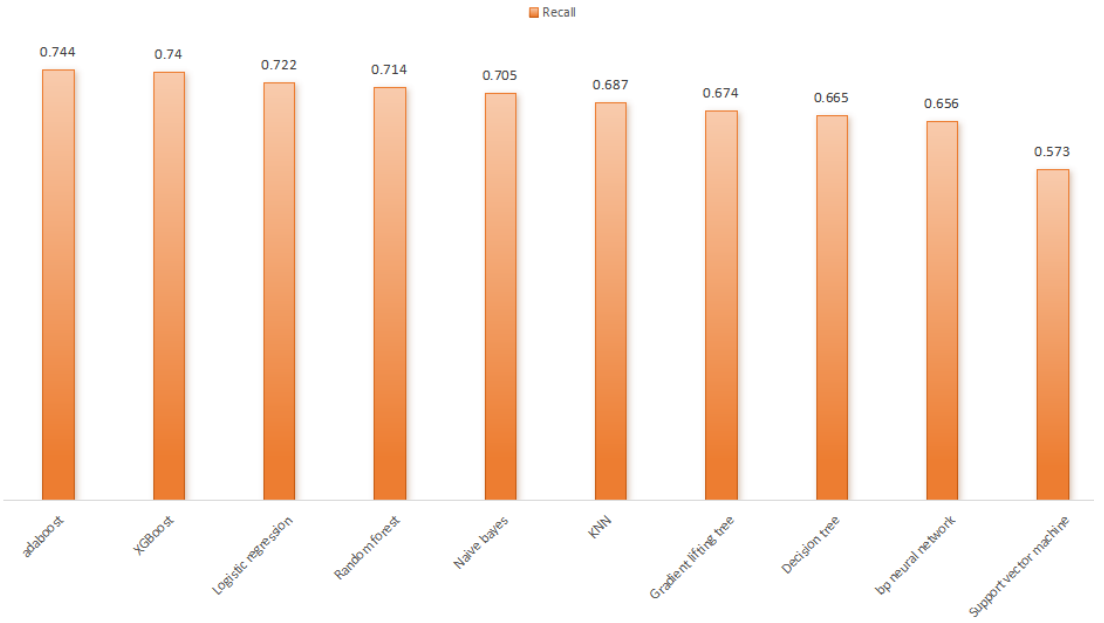

Figure 2. Recall.(Photo credit : Original)

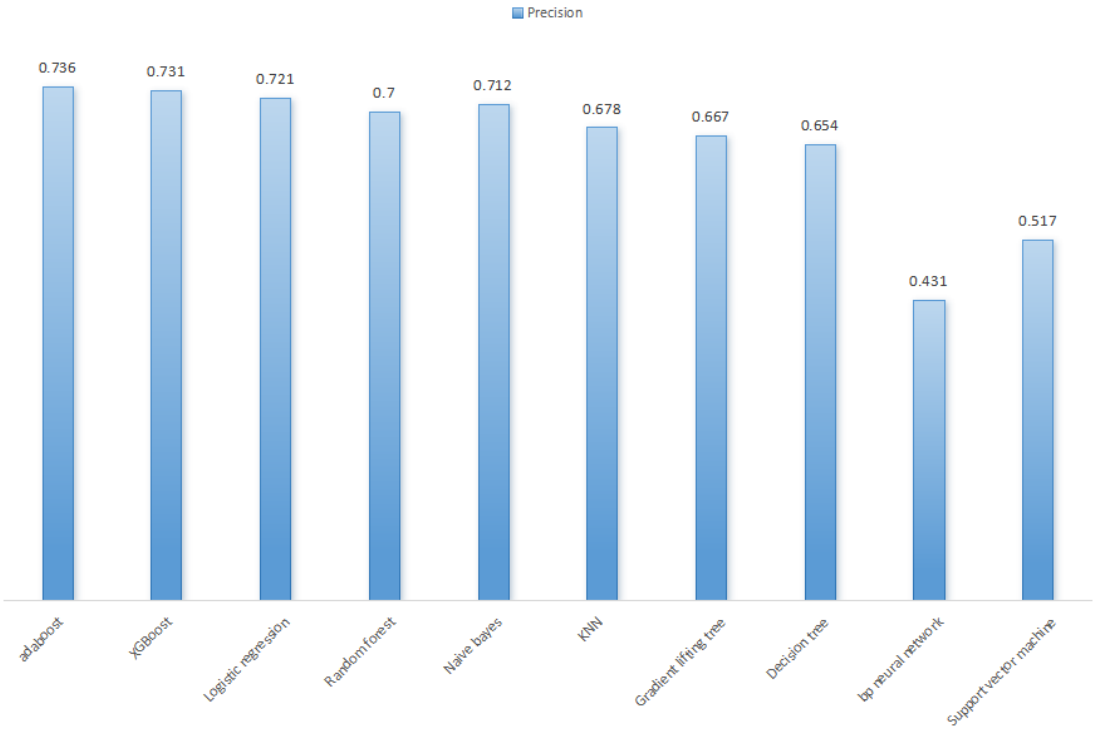

Figure 3. Precision.(Photo credit : Original)

DOI: 10.54254/2753-8818/35/20240874

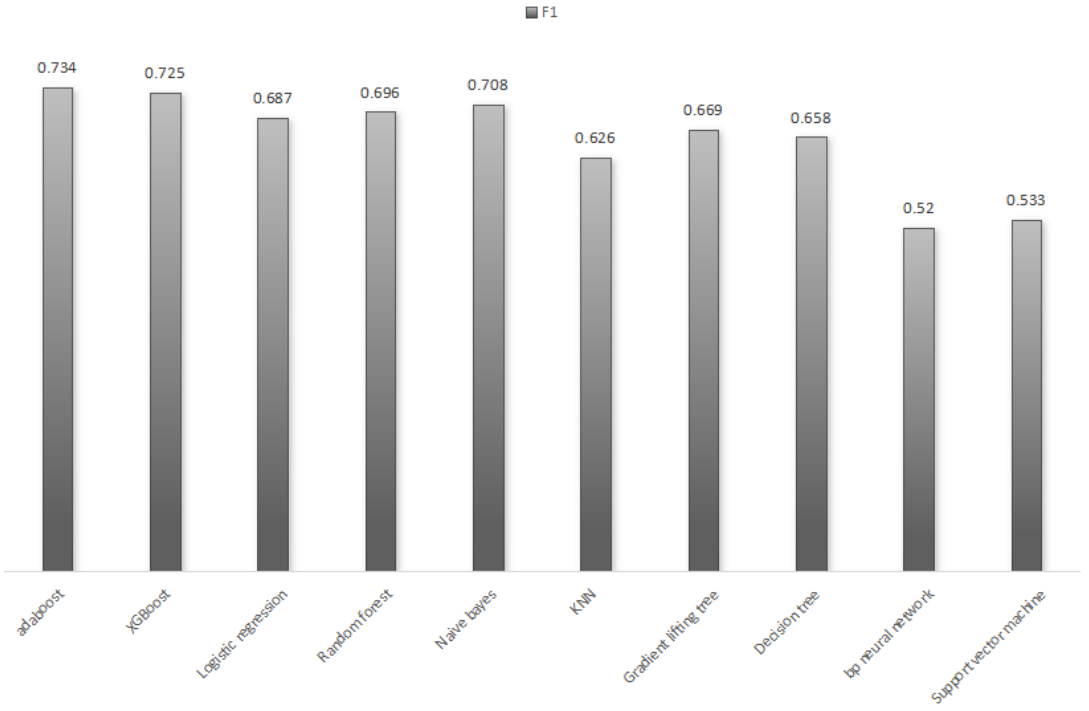

Figure 4. F1.(Photo credit : Original)

4. Conclusion

From the above results, we can see that adaboost classification model performs the best effect in terms of precision, accuracy, recall and F1 score, followed by XGBoost classification model, logistic regression regression model and random forest plain Bayesian classification model all of which reach more than 70% accuracy. However, bp neural network and support vector machine performed poorly in classification effect, analyzing the reason may be due to the limited number of pathologies, classification models such as adaboost have an advantage in classifying Parkinson's patients based on small samples, while bp neural network and support vector machine need a relatively large amount of dataset for support in order to have a better classification effect.

For adaboost classification model, it is an integrated learning algorithm based on weak classifiers, and its main idea is to train multiple weak classifiers by weighting the dataset and combine them into one strong classifier. This algorithm can effectively improve the accuracy and robustness of the model, especially when dealing with small sample data. In this experiment on Parkinson's disease classification, the adaboost classification model performs excellently, which shows that the algorithm is highly adaptable when dealing with small sample data.

The XGBoost classification model is an integrated learning algorithm based on decision trees, which performs well when dealing with large-scale datasets with good generalization ability and robustness. In this experiment, the XGBoost classification model also performs well, which shows that the algorithm has a strong advantage in dealing with large-scale datasets.

Logistic regression regression model is a classification algorithm based on a linear model, which can predict the classification of a sample by linearly combining the features of the data. In this experiment, the logistic regression regression model also achieves an accuracy of more than 70%, which indicates that the algorithm has a good performance in dealing with the Parkinson's disease classification problem.

Random Forest Plain Bayesian Classification Model are two probabilistic model-based classification algorithms, both of which can predict the classification of samples by modeling the probability distribution of the data. Both algorithms also performed well in this experiment, which indicates that they have good performance in dealing with the Parkinson's disease classification problem.

DOI: 10.54254/2753-8818/35/20240874

However, bp neural network and support vector machine perform poorly in classification, the reason for analyzing this may be due to the limited number of pathologies, these two algorithms need a relatively large amount of dataset for support in order to have a better classification. bp neural network is a neuron based classification algorithm which can predict the classification of samples through the combination of multiple layers of neurons. Support vector machine is a classification algorithm based on kernel function, which can map the data into a high-dimensional space, thus making the data easier to segment when classifying. However, due to the limited number of pathologies, the performance of these two algorithms is not as prominent as the other algorithms.

In summary, in this experiment, the adaboost classification model performs well when dealing with small sample data, the XGBoost classification model performs well when dealing with large-scale datasets, and the logistic regression regression model and the Random Forest Plain Bayesian classification model also have good performance. While bp neural network and support vector machine perform poorly in classification effect and need a larger amount of dataset for support. In practical applications, we can choose different classification algorithms according to the size and characteristics of the dataset to achieve the optimal classification effect.

References

[1] Francesco Z D G C ,Greta M ,Ilaria M , et al.Machine Learning and Wearable Sensors for the Early Detection of Balance Disorders in Parkinson's Disease[J]. Sensors, 2022,22(24): 9903-9903.

[2] Jun L ,K S S .Machine Learning Identifies a Rat Model of Parkinson's Disease via Sleep-Wake Electroencephalogram.[J].Neuroscience,2022,5101-8.

[3] Annamaria L ,Marina P ,Teresa M P , et al.Screening performances of an 8-item UPSIT Italian version in the diagnosis of Parkinson's disease.[J].Neurological sciences : official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology,2022,44(3):

[4] Hyun Y P ,Hyun J S ,Wook Y K , et al.Machine learning based risk prediction for Parkinson's disease with nationwide health screening data[J].Scientific Reports,2022,12(1):19499-19499.

[5] A.S.N I M F ,Augusto F B ,Christianini V M , et al.Machine learning models for Parkinson's disease detection and stage classification based on spatial-temporal gait parameters[J].Gait Posture,2022,9849-55.

[6] Noella N S R ,Priyadarshini J .Machine learning algorithms for the diagnosis of Alzheimer and Parkinson disease.[J].Journal of medical engineering technology,2022,47(1):1-9.

[7] Manar E ,Ahmed E ,Mihai O , et al.Early Melanoma Detection Based on a Hybrid YOLOv5 and ResNet Technique.[J].Diagnostics (Basel, Switzerland),2023,13(17):

[8] Daniel L ,G. D A . Topological data analysis and machine learning[J]. Advances in Physics: X,2023,8(1).