Dean&Francis
ISSN 2959-6157

Performance Investigation of Feature Selection Based on Random Forest in Heart Disease Prediction Using KNN Model

Shun Kong

School of Data Science and Big Data Technology, North China University of Technology, Beijing, China

2625102634@brunel.ac.uk

Abstract:

Heart disease is one of the leading causes of death worldwide, claiming millions of lives each year. To address this serious public health challenge, early prediction of heart disease using machine learning techniques has become a hot topic of research. This study explores the impact of different numbers of features on the performance of the K-Nearest Neighbors (KNN) model in predicting heart disease. Initially, a random forest algorithm was employed to rank the importance of a large set of features and identify the key factors most influential in predicting heart disease. Subsequently, starting with the most important features, the study incrementally increased the number of features applied to the KNN model, comparing the model's accuracy and recall across different feature combinations. The results show that as the number of features increases, the model's predictive performance does not consistently improve. When the number of features is initially increased, accuracy experiences a sharp decline; although it slightly recovers later, the overall performance does not return to the high level observed with fewer features. Meanwhile, recall significantly improves when the number of features first increases but then starts to fluctuate and noticeably decreases when a certain number of features is reached. This study demonstrates that simply increasing the number of features does not guarantee improved model performance; instead, it may introduce redundant information or noise, weakening the model's effectiveness.

Keywords: Heart disease; K-Nearest Neighbors (KNN); random forest; feature selection.

ISSN 2959-6157

1. Introduction

Heart disease is one of the leading causes of death worldwide, with an extremely high mortality rate, claiming millions of lives each year [1]. It is not limited to the elderly, as cases of sudden death among younger individuals are also common. With the rise of urbanization, unhealthy diets, and lack of physical activity, the incidence of heart disease is steadily increasing. Related risk factors such as hypertension, high cholesterol, and diabetes are becoming more prevalent, contributing to the frequent occurrence of heart disease. Beyond the heavy emotional and financial burden, it places on patients and their families, heart disease also imposes significant pressure on society and healthcare systems. The high cost of treatment and the resulting loss of productivity make heart disease a major challenge in global public health.

To achieve more efficient, accurate, and cost-effective predictions of heart disease, numerous researchers have made significant progress by employing various machine learning methods and data mining techniques. For example, Jindal and Agrawal utilized logistic regression and K-Nearest Neighbors (KNN) algorithms to classify and predict heart disease in patients [2]. Srinivas et al. collected extensive healthcare data and applied techniques such as decision trees, Naïve Bayes, and neural networks for heart disease prediction [3]. Masethe et al. used data mining algorithms like J48, Naïve Bayes, REPTREE, CART, and Bayes Net, achieving a prediction accuracy of 99% in forecasting heart attacks [4]. Additionally, Du proposed a non-parametric improved logistic regression model based on Nadaraya-Watson (NW) estimation for heart disease prediction [5]. Meanwhile, scholars like Wang have identified heart disease prediction as an imbalanced classification problem with small sample recognition, leading them to develop an improved AdaBoost algorithm for multi-class imbalance classification using active learning, thereby enhancing prediction accuracy [6].

In heart disease prediction research, identifying the key factors influencing heart attacks is crucial for improving the accuracy of predictive models. With the continuous accumulation of medical data, selecting the most impactful factors from a large set of features has become one of the key challenges in this field. Existing studies have shown that the combination of feature selection methods and machine learning models can effectively enhance predictive performance. However, the impact of different numbers of features on model outcomes requires further investigation. The aim of this study is to explore how to accurately select the most important influencing factors from large-scale feature data and optimize the model training process to improve the accuracy and stability of heart disease prediction. Specifically, the study first employs random forest to filter a large set of features, extracting the key factors most influential to heart attacks. Then, a Support Vector Machine (SVM) is used to train the selected features, with the predictive results under different feature quantities compared to assess the impact of feature selection on model performance.

This study provides an efficient and reliable method for heart disease prediction, aiding the medical field in making more accurate decisions in diagnosing and preventing heart disease. Additionally, the findings of this research offer valuable insights for the application of feature selection in predicting other complex diseases.

2. Method

2.1 Dataset Preparation

This study conducted an exploratory analysis based on the dataset collected from the Kaggle [7], implemented by the Centers for Disease Control and Prevention (CDC). This dataset is a large-scale annual survey that collects information on the health behaviors of the U.S. adult population through telephone interviews with randomly selected household members. In 2015, the survey covered over 440,000 participants, including a wide range of critical lifestyle and health indicators. A total of 330 feature variables from this dataset were included in the analytical framework of this study.

To ensure that the feature set used for building the predictive model was both highly relevant and reliable, this paper carried out extensive data preprocessing on this dataset. Initially, the number of missing values for each feature variable was assessed to evaluate their quality, and features with a high proportion of missing values were excluded to reduce uncertainty in the dataset and ensure that the model was trained on high-quality data. Following this, the remaining features were manually selected based on the knowledge and experience of domain experts to ensure that the selected features were highly relevant to the prediction target and had strong predictive potential. Ultimately, approximately twenty features were selected, which not only reduced the risk of overfitting but also improved the interpretability and predictive accuracy of the model, ensuring robust generalization to new datasets.

2.2 Random Forest-based Feature Selection

The Random Forest algorithm [8, 9], a widely used machine learning method, was employed to achieve effective predictions by integrating the results of multiple decision trees. When constructing each decision tree, the algorithm

uses a bootstrapping method to randomly sample from the original dataset and selects the best split feature from a randomly chosen subset of features at each node. This process is designed to increase the diversity among trees, reduce correlations between them, and ultimately enhance the overall performance of the model.

In this study, a Random Forest classifier was applied for an in-depth analysis of the dataset. The dataset was first loaded and preprocessed, then divided into a feature set and target variable, followed by splitting into training and testing sets. Subsequently, a Random Forest model comprising 100 decision trees was constructed and trained on the training set, enabling the model to effectively learn the relationships between features and the target variable. The importance of each feature was then analyzed, and its contribution to classification decisions was visually presented. The Random Forest model was selected for this study due to its superior performance in handling high-dimensional data, strong resistance to overfitting, and its ability to provide feature importance scores, which help identify the most influential features. Additionally, the Random Forest algorithm demonstrated excellent performance in handling missing data, allowing the processing of incomplete datasets without significant information loss.

2.3 KNN-based Machine Learning Model Prediction

The K-Nearest Neighbors (KNN) algorithm is a simple yet widely applied supervised learning method, commonly used for classification and regression tasks [10, 11]. The KNN algorithm predicts the class or value of a new sample by calculating its distance to all samples in the training set, selecting the K nearest neighbors, and using their labels. Since KNN does not require a complex training process and only computes distances during prediction, it is particularly suitable for small to medium-sized datasets.

In this study, the KNN algorithm was utilized to evaluate the impact of varying the number of features on the performance of a heart disease prediction model. This study incrementally increased the number of features, starting with the most important variables, and used the KNN algorithm (with K=3) to model the training set and classify samples in the test set. In each iteration, KNN classified samples based on the three nearest neighbors, allowing for an effective assessment of how different feature combinations influenced model performance. The simplicity and intuitive nature of the KNN algorithm provide certain advantages when exploring the relationship between feature selection and model performance.

To more accurately evaluate the model's ability to identify heart disease patients, recall was selected as the primary evaluation metric. Recall measures the sensitivity of the model in detecting heart disease patients (positive class samples), which is particularly critical in medical diagnostics. Given that missed diagnoses can have severe consequences, recall was chosen to ensure that the model identifies as many actual heart disease cases as possible, thereby enhancing its utility and reliability in high-risk scenarios.

3. Results and Discussion

The feature importance analysis using the Random Forest model reveals that Body Mass Index (BMI) is the most significant predictor in the dataset shown in Fig. 1, with an importance score approaching 0.30, substantially higher than that of other variables. Age and Income follow as the next most important features, with importance scores close to 0.10. The other features, such as Education, show moderate importance, with scores ranging between 0.05 and 0.10. Features like Smoke (Smoking), Fruit (i.e. Fruit consumption), Diabetes, and Blood Pressure (BP) have relatively lower importance scores.

ISSN 2959-6157

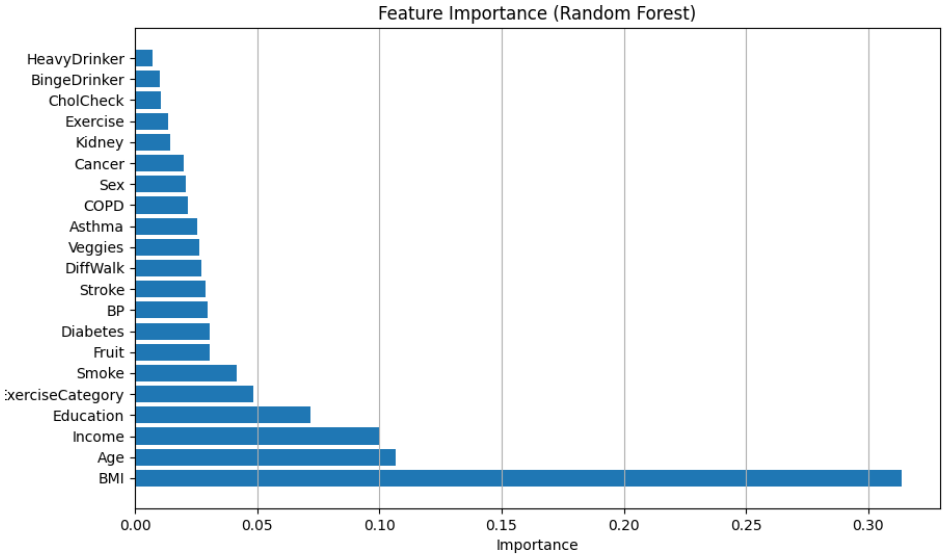

Fig. 1 The feature importance of the random forest (Photo/Picture credit: Original).

These results indicate that BMI, age, and income are critical factors in predicting cardiovascular health outcomes, underscoring the significant impact of obesity, aging, and socioeconomic status on heart disease risk. Lifestyle factors such as education level and exercise habits also contribute to the model's predictive power, albeit to a lesser extent. Features with lower importance scores, such as heavy drinking, may have a smaller impact on the dataset or may reflect higher noise levels or data imbalance.

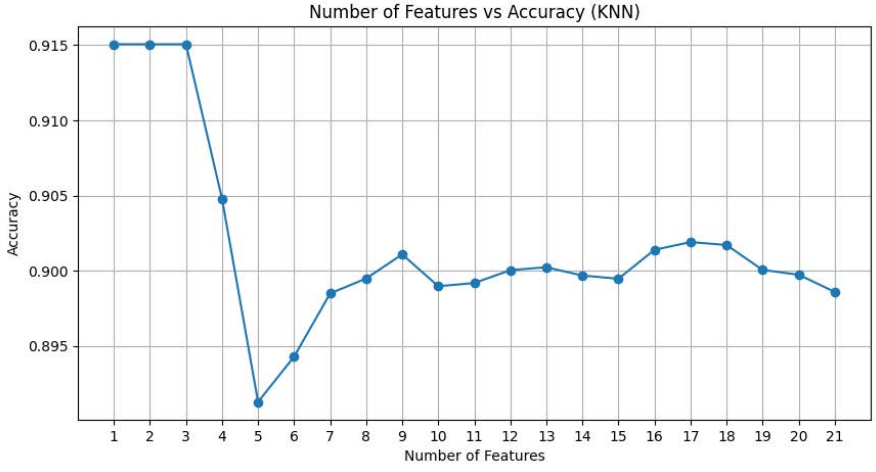

Fig. 2 Accuracy performance of the KNN model (Photo/Picture credit: Original).

Starting with BMI, the feature ranked highest in importance, this paper incrementally increased the number of features and applied them to the KNN model to assess the impact on model accuracy shown in Fig. 2. The results indicated that when the number of features ranged from 1 to 3, the model maintained a high accuracy level of approximately 0.915. However, when the number of features increased to 4, accuracy significantly dropped to around 0.895, suggesting the inclusion of potentially irrelevant or noisy features. As the number of features continued to increase, the model's accuracy fluctuated between 0.895 and 0.905, without recovering to the higher levels observed with fewer features.

However, in the task of heart disease prediction, using accuracy as an evaluation metric has significant limitations, primarily because accuracy only measures the overall correctness of the model's predictions but fails to adequately reflect the class imbalance between positive and negative

samples in the dataset. Given that the dataset contains a disproportionately high number of negative samples (i.e., healthy individuals), the model could achieve a high accuracy even if it simply predicts all samples as negative. However, in such cases, the model's performance in identifying positive samples (i.e., heart disease patients) may be poor. Therefore, to more comprehensively evaluate the effectiveness of the model in heart disease prediction, it is necessary to employ recall as an additional evaluation metric. Recall is particularly advantageous in this context, especially for detecting positive samples. Recall measures the model's ability to correctly identify actual positive samples, i.e., the proportion of true heart disease patients that the model correctly detects. A higher recall indicates a lower rate of missed detections in identifying heart disease patients, which is crucial for clinical applications, as missed detections can have severe consequences.

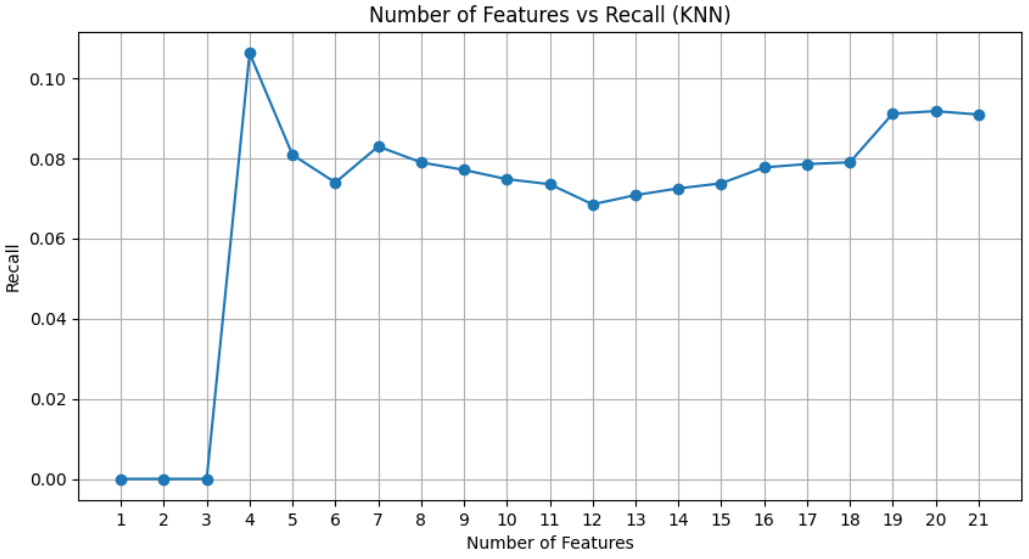

Fig. 3 Recall performance of the KNN model (Photo/Picture credit: Original).

From the recall results shown in Fig. 3, it can be observed that with 1 to 3 features, the recall is nearly zero. When the number of features increases to 4, recall rapidly rises to its highest point (slightly above 0.10). However, as the number of features continues to increase, recall begins to decline and fluctuates between 5 and 12 features. Afterward, recall gradually rises and stabilizes between 12 and 21 features, but it never exceeds the recall value observed when using 4 features.

4. Conclusion

This study demonstrates that increasing the number of features does not consistently enhance model prediction performance. Simply adding more features does not necessarily lead to better performance. In terms of recall, the recall rate significantly improves when the number of features is increased to four. However, as more features are added, the recall rate fluctuates and notably decreases when the number of features reaches twelve. This indicates that while a moderate increase in features can help improve the model's ability to identify positive instances, too many features may introduce redundant information or noise, which can weaken model performance. Similarly, the trend in accuracy reveals a comparable situation: as the number of features increases to four, accuracy sharply declines. Although it slightly recovers afterward, the overall performance does not match the high level observed with fewer features. This suggests that merely increasing the number of features may not lead to higher accuracy; instead, it may reduce performance due to the introduction of redundancy and noise. In the future, other feature selection methods such as LASSO regression or Principal Component Analysis (PCA) could be employed in addition to Random Forest, potentially revealing different effective features. Additionally, techniques such as oversampling, undersampling, or synthetic data generation (e.g., SMOTE) could be utilized to address class imbalance issues in the dataset, creating a more balanced dataset. Moreover, applying the research methods to datasets with different characteristics could help verify the generalizability of the conclusions and determine whether the results are applicable beyond the current dataset.

ISSN 2959-6157

References

[1] WHO, Cardiovascular diseases (CVDs), https://www.who.int/zh/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds), 2021.

[2] Jindal H, Agrawal S, Khera R, et al. Heart disease prediction using machine learning algorithms. IOP conference series: materials science and engineering. IOP Publishing, 2021, 1022(1): 012072.

[3] Srinivas K, Rani B K, Govrdhan A. Applications of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering (IJCSE), 2010, 2(02): 250-255.

[4] Masethe H D, Masethe M A. Prediction of heart disease using classification algorithms. Proceedings of the world Congress on Engineering and computer Science. 2014, 2(1): 25-29.

[5] Du YY. Analysis of heart disease prediction performance using different classification models (in Chinese). Modeling and Simulation, 2023, 12: 5600.

[6] Wang LL, Fu ZL, Tao P, et al. Heart disease classification based on an imbalanced multi-class AdaBoost algorithm using active learning (in Chinese). Computer Applications, 2017, 37(7): 1994-1998.

[7] CDC, Behavioral Risk Factor Surveillance System. https://www.cdc.gov/brfss/annual_data/annual_data.htm, 2024

[8] Rigatti SJ. Random forest. Journal of Insurance Medicine. 2017 Jan 1;47(1):31-9.

[9] Biau G, Scornet E. A random forest guided tour. Test. 2016 Jun;25:197-227.

[10] Guo G, Wang H, Bell D, Bi Y, Greer K. KNN model-based approach in classification. InOn The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings 2003 (pp. 986-996). Springer Berlin Heidelberg.

[11] Zhang S, Li X, Zong M, Zhu X, Cheng D. Learning k for knn classification. ACM Transactions on Intelligent Systems and Technology (TIST). 2017 Jan 12;8(3):1-9.