Joint SMOTE and Random Forest for Heart Disease Prediction and Characterization

Yi Lu
College of Electronic and Information Engineering, Tongji University, Shanghai, China

Keywords: Heart disease, Synthetic Minority Over-Sampling Technique, Random Forest, Feature analysis.

Abstract: For decades, heart diseases have remained the primary global cause of mortality. Consequently, comprehending the influential elements and forecasting the onset of cardiovascular conditions is imperative, enabling individuals to proactively preserve their well-being. The primary goal of this study is to forecast the occurrence of heart disease while exploring the influencing factors associated with it. The study is conducted on the Personal Key Indicators of Heart Disease dataset from Kaggle. Following the completion of exploratory data analysis (EDA), the research tackles the problem of an uneven distribution of data by integrating the Synthetic Minority Over-sampling Technique (SMOTE) approach into the initial Random Forest (RF) model. Notably, the resultant model achieves commendable performance metrics, boasting an accuracy of 93.39%, precision of 94.25%, recall of 92.42%, and an F1 score of 93.33%. Through the RF feature analysis, it is revealed that Body Mass Index (BMI), overall health status, and age are the top three influential features significantly impacting the model's predictive performance. This finding provides valuable guidance for heart disease prevention efforts, aiding in the development of more precise intervention measures targeting individual risk factors.

1 INTRODUCTION

According to the World Heart Report 2023 from the World Heart Federation, more than half a billion people around the world suffer from cardiovascular diseases. Nevertheless, it is worth noting that up to 80% of premature heart attacks and strokes can be prevented. Therefore, it is crucial to understand the contributing factors and analyze the likelihood of the occurrence of heart diseases, which paves the way for people to take proactive measures to maintain their health.

Given that even minor errors in the diagnosis of heart disease can potentially lead to severe consequences, it is imperative to improve the accuracy of predicting heart-related conditions (Singh and Kumar 2020). In recent years, Machine Learning (ML) has been widely used in predicting and analyzing the heart disease. Due to the substantial volume of data received by medical institutions, often containing a significant amount of noise, can pose challenges for effective analysis. Utilizing ML techniques for data processing and analysis can greatly enhance the efficiency of healthcare professionals while enabling effective prediction and data management (Katarya and Meena 2021 & Ramalingam et al 2018). For instance, Jesmin Nahar et al. primarily focused on three rule mining algorithms involving Apriorist, Predictive Apriorist, and Tertius. Drawing from both gender and significant risk factors for heart disease, they identified key signals of good health using healthy regulations that boasted confidence levels exceeding 90% and accuracy levels surpassing 99% (Nahar et al 2013). Typically, there is a multitude of related features, making it crucial to extract meaningful features. Escamila et al. selected the features by using the chi-square (CHI) method and performed feature extraction through principal component analysis (PCA). They then employed several classification models to predict heart diseases and compare the final outcomes. According to the result, the best-performing method was CHI-PCA with Random Forest (RF) (Gárate-Escamila et al 2020). Amin et al. first proposed a data mining technique to select significant features. Then seven classification models were used: k-Nearest Neighbors (k-NN), Decision Tree (DT), Naive Bayes (NB), LR, Support Vector Machine (SVM), Neural Network (NN), and Vote. Results showed that Vote yield the best performance, having an 87.4% accuracy rate (Amin et al 2019). Akgül et al. introduced a hybrid approach aimed at improving the classification accuracy of the conventional Artificial Neural Network (ANN) model. Through the fusion of ANN with a Genetic Algorithm (GA), their findings indicated that

Lu, Y.
Joint SMOTE and Random Forest for Heart Disease Prediction and Characterization.
DOI: 10.5220/0012816100003885

In *Proceedings of the 1st International Conference on Data Analysis and Machine Learning (DAML 2023)*, pages 127-132
ISBN: 978-989-758-705-4

the ANN-GA model delivered the highest performance, achieving an accuracy rate of 95.82% (Akgül et al 2019).

Class imbalance is a common issue in raw data, particularly prevalent in the field of medical diagnostics, where the majority of classification data often lean towards negative class values (Thabtah et al 2020). Often, the number of healthy patients might greatly exceed the number of patients with heart disease, potentially exerting a considerable influence on model accuracy. The primary aim of the research is to anticipate the existence of cardiovascular ailments by considering a wide range of influencing factors such as Body Mass Index (BMI), Smoking, and Alcohol Drinking. In addition, this data imbalance problem is addressed to gain higher accuracy of the heart disease classification model. In the study, Exploratory Data Analysis (EDA) is first used to gain an understanding of the distribution and features of a dataset, laying the groundwork for subsequent modeling. Second, the Synthetic Minority Over-sampling Technique (SMOTE) is employed to balance the dataset. Rather than simply replicating examples from the minority class, the fundamental concept of the technique is to generate synthetic samples. SMOTE is considered the standard benchmark for learning from imbalanced data (Fernández et al 2018). Subsequently, the RF is applied to make predictions. RF has significant advantages in handling high-dimensional features and large-scale data, while also maintaining high interpretability. The model attains commendable performance metrics, highlighting its resilience in identifying individuals who are susceptible to heart disease. The resultant model has an accuracy of 93.39%, a precision of 94.25%, a recall of 92.42%, and an F1 score of 93.33%. Furthermore, by analyzing feature importance, the experimental results demonstrate that BMI emerges as the most influential factor in predicting the presence of heart disease, facilitating a better understanding of the influential variables in this critical healthcare context. Importantly, the knowledge derived from the study also furnishes a valuable framework for predicting other rare medical conditions with similar class imbalance challenges.

2 METHODOLOGY

2.1 Dataset Description and Preprocessing

Personal Key Indicators of Heart Disease dataset from Kaggle is designed to investigate key indicators associated with heart disease, a leading cause of mortality in the United States (Dataset 2023). The dataset has been refined to retain 319,795 data points with 18 relevant variables. The target variable is "heart disease," which serves as a binary indicator of the existence or non-existence of cardiovascular ailments. Besides, there are 13 categorical features: smoking status, alcohol drinking habits, stroke history, difficulty walking, gender, age category, race, diabetic status, physical activity levels, general health assessments, asthma, kidney disease, and skin cancer. The dataset also includes 4 numerical features: BMI, physical health assessments, mental health assessments, and sleep time, all of which contribute to the dataset's richness.

In the data preprocessing phase, it is observed that the dataset exhibits an imbalanced distribution of heart disease cases, with only 9% labeled as "Yes". This class imbalance has been taken into consideration during model development to prevent bias. Moreover, to ensure data accuracy and enhance model effectiveness, a total of 18,078 duplicated data points are identified and removed from the dataset. Besides, through label encoding, each distinct category is assigned a unique integer value, ensuring that ML models can effectively interpret and utilize these features. These steps play a pivotal role in ensuring the dataset's integrity for subsequent analysis and modeling tasks, enhancing the effectiveness of the model.

2.2 Proposed Approach

The primary objective of this study is to explore key indicators related to heart disease. As depicted in Fig. 1, the study initially conducts preliminary EDA to gain an in-depth understanding of the dataset's characteristics. Given the dataset's inherent class imbalance, the SMOTE method is used to balance the dataset, addressing the issue of data imbalance. Following data preprocessing, RF is utilized for heart disease prediction. The performance of the model is assessed through a range of performance metrics, including recall, precision, accuracy, F1 score, which ensures that the model exhibits robustness and generalization capability. Additionally, feature importance scores provided by the RF are leveraged to analyze the significance of various influencing factors, enhancing the understanding of which factors play a pivotal role in predicting heart disease.

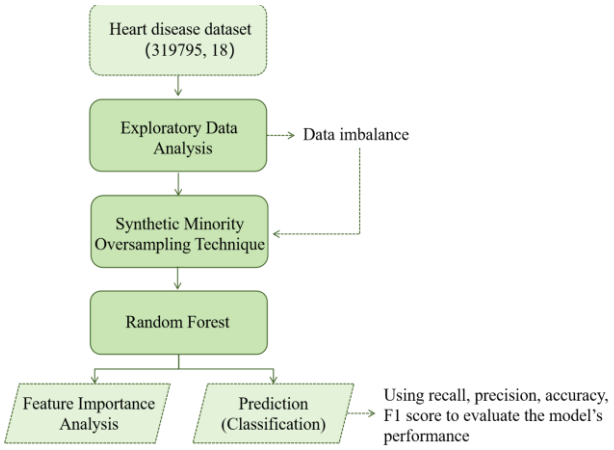

Figure 1: The pipeline of the model (Original).

2.2.1 EDA

EDA is a data analysis method that involves analyzing data and deriving patterns. The main steps are shown in Fig. 2. Data cleaning includes handling errors, duplicates, anomalies, and missing values to ensure the accuracy of data. The data exploration and analysis stage primarily involves computing statistical values, exploring data distributions, and analyzing the correlations between different variables. Based on the data analysis, feature engineering can be conducted, involving the selection or creation of new features. Finally, through data visualization, it can intuitively present data, aiding users in understanding and extracting the information embedded in the data. These steps collectively contribute to a deeper understanding of the dataset, laying a solid foundation for subsequent modeling and analysis.

Figure 2: The steps for EDA (Picture credit: Original).

2.2.2 SMOTE

SMOTE is a synthetic data technique used to address the problem of imbalanced datasets. Its core principle involves generating synthetic samples by assessing the similarity between minority class samples, thereby increasing the number of minority class instances. SMOTE selects a minority class sample as the starting point and then calculates the distances between this sample and its k-NN within the minority class, typically using Euclidean distance or other similarity metrics. Next, SMOTE randomly chooses one of these nearest neighbor samples and computes the feature vector difference between the selected sample and the starting point. It generates a random value between 0 and 1, multiplies it by the feature vector difference, and adds the result to the feature vector of the starting point, creating a new synthetic sample. The termination condition for this process loop is the dataset reaching the predetermined balance level, effectively increasing the number of minority class samples.

2.2.3 RF

RF is a powerful ensemble learning method that plays a critical role in improving model performance by combining multiple decision trees. It is suitable for both classification and regression tasks. In the context of this experiment, RF is applied to solve classification problems. In this algorithm, for each decision tree, the dataset is randomly sampled to create distinct subsets, and each tree is trained on these unique subsets. To diversify the model and mitigate overfitting, at each tree node, the splitting process is guided by Gini impurity, facilitating the selection of optimal splitting features and points to reduce node impurity.

Furthermore, to promote model diversity, RF considers only a subset of random features when selecting the splitting attribute, and then choosing the most effective one from this subset. Ultimately, RF aggregates the predictions from each tree and selects the class with the highest frequency as the final prediction, ensuring reliable and robust classification capabilities. This amalgamation of randomness and ensemble techniques positions RF as a vital tool in the domain of ML, excelling in diverse and complex tasks.

In the RF model, feature importance is typically measured based on Gini impurity. During each node splitting process, the reduction in Gini impurity is

calculated, and then the weighted average of the reduction in Gini impurity across all decision trees is computed to obtain the importance score of each feature. The higher the importance score of a feature, the greater its influence on the model performance. This measurement method helps identify the features that have the most significant influence on model predictions and can be used for feature selection while providing deeper insights into the behavior of the model.

2.2.4 Evaluation Metrics

Accuracy: Accuracy evaluates the percentage of accurately predicted instances among the total instances. The formula for accuracy is as follows:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Samples} \quad (1)$$

Recall: Out of all actual positive class samples, recall quantifies the percentage of accurately anticipated positive class samples. For medical diagnostic problems, recall is a significant metric as missing patients can have serious consequences. The formula for recall is as follows:

$$Recall = \frac{Number\ of\ True\ Positive\ Samples}{Total\ Number\ of\ Positive\ Class\ Samples} \quad (2)$$

Precision: Precision assesses the proportion of accurately predicted positive instances among all predicted positive class samples. It helps assess the correctness of the positive predictions. The formula for precision is as follows:

$$Precision = \frac{Number\ of\ True\ Positive\ Samples}{Total\ Number\ of\ Predicted\ Positive\ Samples} \quad (3)$$

F1-Score: The F1-Score, applicable to binary classifications, represents the harmonic average of precision and recall, ranging from 0 to 1.

$$F1 - Score = \frac{2\times(Precision \times Recall)}{Precision+Recall} \quad (4)$$

2.3 Implementation Details

The study is conducted using Python 3.11.3 on a Windows 11 operating system. The handling of imbalanced datasets involved applying the SMOTE technique from the imbalanced-learn library. The following parameters are utilized: The random state parameter is set to 42 to ensure the result's reproducibility and the default value of 5 is utilized for the k neighbors parameter. The RF model is configured with the following settings: the Gini impurity measure is adopted as the criterion for node splitting; At least 1 sample per leaf node is obligatory; a minimum of 2 samples is needed to conduct a split on internal nodes; the minimum impurity decrease parameter is set to 0.0, indicating the minimum impurity decrease for node splitting.

3 RESULTS AND DISCUSSION

In this experiment, a comparative analysis of the performance of the RF before and after applying SMOTE is conducted. Moreover, an analysis of the feature importance in the random forest is performed.

As indicated in Table I, the initial random forest exhibits impressive outcomes concerning accuracy, precision, recall, and F1 score on the training dataset. However, on the testing dataset, all performance metrics imply that the predictive performance of the model is relatively unsatisfactory. While the accuracy of the testing dataset remains relatively high, other metrics are notably lower. Of particular concern is the recall score on the testing dataset, which is as low as 0.1108, significantly lower than the 0.9715 recall score on the training dataset. This implies that the model struggles to effectively identify individuals with a genuine heart condition in the testing dataset, potentially leading to serious misclassifications.

Table 1: Performance of Model.

Model		Accuracy	Precision	Recall	F1
Initial Random Forest	Training	0.9970	0.9957	**0.9715**	0.9835
	Testing	0.8984	0.3204	**0.1108**	0.1646
SMOTE+Random Forest	Training	0.9983	0.9981	**0.9985**	0.9983
	Testing	0.9302	0.9302	**0.9209**	0.9295

The fundamental reason behind this outcome is the data imbalance present in the original dataset. As illustrated in Fig. 3, despite the roughly balanced number of male and female samples with and without heart conditions, the proportion of samples with heart conditions is only 9%, resulting in a significant imbalance. When dealing with imbalanced data, ML models tend to predict the majority class, a phenomenon that becomes particularly pronounced in the testing dataset. In the training dataset, the number of samples with heart conditions is much smaller than the number of healthy samples. This leads to a notable

decrease in the model's recall rate on the testing dataset, as it tends to incorrectly classify individuals with genuine heart conditions as not having heart conditions.

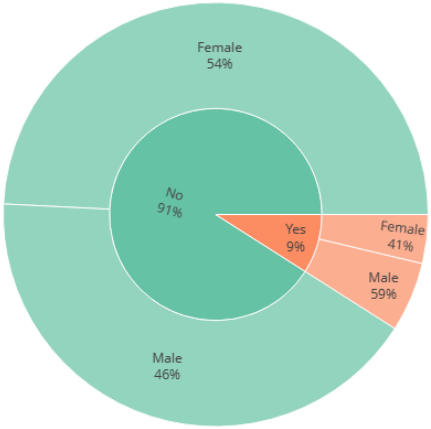

Figure 3: Heart Disease Proportion by Gender (Original).

To tackle the issue of data imbalance, SMOTE is applied to the original dataset. In the processed dataset, the ratio of positive class (individuals with heart disease) to negative class (individuals without heart disease) samples became 1:1, thereby improving the dataset's balance. Subsequently, the RF model is trained on the balanced dataset. As shown in Table 1, the optimized model exhibits outstanding results in various performance metrics on the testing dataset. All metrics on the testing dataset, encompassing accuracy, precision, recall, and F1 score, show a significant increase. Recall, in particular, improves substantially from its initial value of 0.1108 to 0.9209. This signifies that the model can more effectively identify individuals with genuine heart disease in the testing dataset, reducing the likelihood of misclassification.

The SMOTE-processed model not only enhances predictive performance but also increases the accuracy of identifying individuals with heart disease. This improvement can be attributed to data balancing, enabling the model to better adapt to the challenges posed by imbalanced data. Through this optimization, the study provides a more reliable tool for the medical field to accurately predict patients with heart disease, thus offering robust support for medical decision-making and interventions.

According to Fig. 4, through the feature importance analysis of the RF model, it is revealed that 'BMI' holds the highest importance in the model with a score of 0.3516, indicating its significant impact on the classification task of heart disease. 'GenHealth' and 'AgeCategory' also exhibit relatively high importance, with scores of 0.1076 and 0.1039 respectively. This suggests that these features play a significant role in predicting heart disease. While the importance scores of other features are relatively lower, they still contribute to the model's performance. Understanding the varying importance of these features can assist the healthcare field in gaining a better understanding of and addressing the risks associated with heart disease.

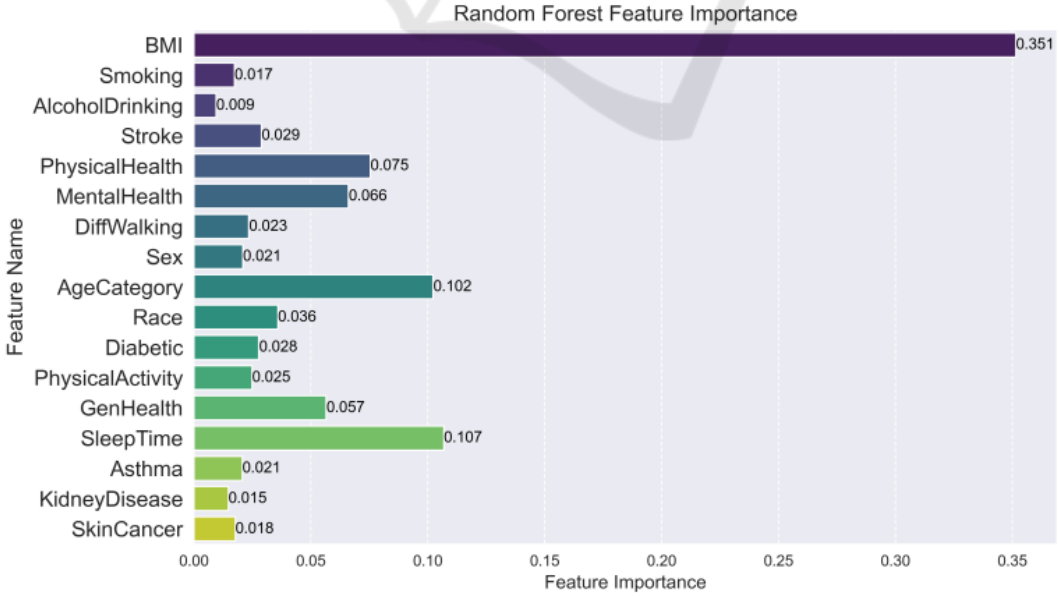

Figure 4: Random Forest Feature Importance (Original).

4 CONCLUSION

This study explores the influencing factors of heart disease and predicts its occurrence by constructing a random forest model. Initially, the data imbalance present in the dataset is identified through EDA. This skewed distribution has the potential to misclassify individuals with heart conditions, consequently impeding the predictive capabilities of the initial RF model. In this study, the combination of SMOTE and RF effectively addresses the issue of data imbalance, achieving excellent predictive performance. Eventually, the accuracy of the model is 93.39%, precision is 94.25%, recall is 92.42%, and F1 score is 93.33%. These results underscore the model's reliability in predicting heart disease occurrence. Additionally, the feature importance analysis conducted within the random forest framework highlights the substantial influence of BMI, general health status, and age on heart disease, with their combined importance exceeding 0.55. Furthermore, the study's findings offer valuable insights applicable to the prediction of other rare medical conditions confronted with similar class imbalance challenges. Moving forward, given that heart disease is a complex condition influenced by various factors, it is imperative for future research to delve deeper into the intricate relationships between heart disease and various correlated conditions. This comprehensive understanding will facilitate the development of more effective and targeted preventative strategies in combating the complexities of heart disease.

REFERENCES

A. Singh, R. Kumar. “Heart disease prediction using machine learning algorithms”, 2020 international conference on electrical and electronics engineering (ICE3), vol. 2020, pp. 452-457.

R. Katarya, S.K. Meena. “Machine learning techniques for heart disease prediction: a comparative study and analysis”, Health and Technology, vol. 11, 2021, pp. 87-97.

V. V. Ramalingam, A. Dandapath, M. K. Raja. “Heart disease prediction using machine learning techniques: a survey”, International Journal of Engineering & Technology, vol. 7(2.8), 2018, pp. 684-687.

J. Nahar, T. Imam, K. S. Tickle, et al. “Association rule mining to detect factors which contribute to heart disease in males and females”, Expert systems with applications, vol. 40(4), 2013, pp. 1086-1093.

A. K. Gárate-Escamila, A. H. El Hassani, E. Andrès. “Classification models for heart disease prediction using feature selection and PCA”, Informatics in Medicine Unlocked, vol. 19, 2020, p. 100330.

M. S. Amin, Y. K. Chiam, K. D. Varathan. “Identification of significant features and data mining techniques in predicting heart disease”, Telematics and Informatics, vol. 36, 2019, pp. 82-93.

M. Akgül, E. Sönmez Ö, T. Özcan. “Diagnosis of heart disease using an intelligent method: a hybrid ANN–GA approach”, International conference on intelligent and fuzzy systems. Cham: Springer International Publishing, vol. 2019, pp. 1250-1257.

F. Thabtah, S. Hammoud, F. Kamalov, et al. “Data imbalance in classification: Experimental evaluation”, Information Sciences, vol. 513, 2020, pp. 429-441.

A. Fernández, S. Garcia, F. Herrera, et al. “SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary”, Journal of artificial intelligence research, vol. 61, 2018, pp. 863-905.

Dataset, “personal-key-indicators-of-heart-disease”, Kaggle, 2023 https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease/data