Analysis And Research on Overfitting and Underfitting Issues in Heart Disease Prediction Models

Yuying Hu

Nanjing Foreign Language School, Nanjing, China

joyce.hu25@nflsicc.com

Abstract. Research Background: Cardiovascular diseases remain the leading cause of mortality globally, necessitating the advancement of predictive models that can accurately assess heart disease risk. Factors such as cholesterol levels, smoking habits, and demographic variables add complexity and variability to modeling efforts, often resulting in overfitting or underfitting. These issues compromise the models' applicability to new, unseen datasets, limiting their utility in clinical settings. The challenge lies not only in integrating diverse health indicators into a cohesive analytical framework but also in managing the intrinsic trade-offs between model complexity and generalizability. **Study Focus and Methodology:** This study addresses the pivotal challenge of overfitting and underfitting in predictive models for heart disease using a comprehensive dataset sourced from Kaggle. By employing a variety of modeling techniques, including logistic regression, random forest, K-nearest neighbors, and decision tree classifiers, this research evaluates how different models assimilate and predict based on the multifaceted data related to heart disease. Through the application of principal component analysis (PCA), this study effectively reduces dimensionality, thereby simplifying the models without sacrificing the integrity of the information. Furthermore, rigorous cross-validation methods are utilized to ensure the models maintain their accuracy and generalizability when applied to new data.

Keywords: Predictive Modeling; Heart Disease; Overfitting; Dimensionality Reduction.

1. Introduction

Heart disease remains one of the leading causes of mortality worldwide, with an estimated 17.9 million lives claimed annually according to the World Health Organization [1]. The complexity of cardiovascular disease is compounded by various biologically proven risk factors, including high cholesterol, excessive alcohol consumption, and obesity. As populations strive for healthier living, the development of predictive models becomes imperative to assess individual risks associated with these diseases [2]. These models demand sophisticated computational approaches to accurately parse through the myriad of contributing factors, necessitating the use of comprehensive datasets to uncover the principal causes of heart disease through advanced data analytics and dimensionality reduction techniques [3][4].

Research Problem: The task of predicting heart disease is challenged by the diversity and complexity of influencing factors, which complicates model accuracy. Effective prediction requires models that can accommodate the extensive data involved, ranging from individual health metrics to lifestyle choices. This process, often referred to as Knowledge Discovery in Databases, involves extracting valuable insights from large-scale data through data mining, highlighting the necessity of multifaceted analysis including factor analysis to pinpoint primary risk factors [5]. Addressing these challenges requires an exploration of various modeling techniques to ensure the reliability and applicability of predictive outcomes.

Contributions of This Paper: This paper addresses the critical issues of overfitting and underfitting in machine learning models designed to predict heart disease. Overfitting occurs when models, often too complex, fail to generalize beyond their training data, while underfitting results from overly simplistic models that fail to capture underlying data patterns [6][7][8]. By employing a range of modeling techniques such as logistic regression, random forest, K-nearest neighbors, and decision tree classifiers, this study compares their effectiveness in managing the balance between model complexity and performance [9]. Utilizing principal component analysis and other

dimensionality reduction methods, the research aims to optimize the accuracy and generalizability of these models, ensuring they are neither overfitted nor underfitted. This exploration not only enhances the predictive accuracy of heart disease risks but also provides actionable insights for individuals to undertake preventive measures tailored to their specific risk profiles [10]. This paper will delve into the methodologies employed to mitigate the issues of fitting and discuss the implications of these strategies on the performance of heart disease predictive models.

2. Data Processing and Prediction Model Design

The data set from Kaggle has 25 rows which are different features related to heart health of patients, for example, their age, sex and previous heart problem. The number of samples is 8763 and is undoubted a massive data sample because the participants in it are diverse in age, gender and geographic location.

2.1. Data Preprocessing and Dimensionality Reduction

In real datasets, data often presents challenges such as messiness, incompleteness, missing values, noise, outliers, and inconsistent formats, which, if unaddressed, can degrade model performance. Data preprocessing is a crucial step to clean and standardize numerical features, such as normalizing or standardizing these features, thereby enhancing the model's learning efficacy and predictive accuracy. Initial data cleaning involves removing duplicate records, addressing missing values, and splitting the 'blood pressure' attribute into 'systolic' and 'diastolic' components to create a dataset with appropriately formatted categorical variables. Additionally, histograms are generated using Rmarkdown packages to identify potential outliers within the dataset.

Datasets with many attributes are referred to as high-dimensional data, which can increase both the training time and computational complexity of models. Moreover, such data are prone to overfitting if dimensionality reduction is not implemented. Principal Component Analysis (PCA) is an effective technique employed to identify the main components of the data through linear transformation. These main components are combinations of features that account for the greatest variance in the new dimension, thereby retaining as much crucial information as possible. After applying PCA, the dimensionality was effectively reduced from 21 to 17.

The diversity and similarity within the dataset are illustrated through the spread and concentration of data points across the PCA plot. Many individuals share similar characteristics along these principal components, while those further from the center represent outliers or individuals with unique characteristics.

The visualization of variable contributions to the PCA reveals that factors such as smoking and sex significantly influence the first two principal components. A color scale enhances this visualization, with red indicating higher contributions and green showing lower contributions. Other variables like age, cholesterol, and stress level also make notable contributions to the risk of heart disease, emphasizing their importance in explaining the variance of the principal components.

2.2. Cluster Analysis

Clustering is an unsupervised learning technique for grouping samples from a data set into different subsets such that samples in one subset are similar to each other, vice versa. It can uncover hidden patterns or structures in the data, providing the basis for further analysis.

2.2.1 K means clustering

Means clustering iterates the sample into K clusters so that the distance between the sample of each cluster and its center of mass is minimal. The K value is predefined as 3, which is also the number of clusters.

The cluster plot in Figure 1 displays the results of the clustering analysis. Three distinct clusters are identified and represented by different colors and shapes. The clusters show varying densities and overlaps, particularly between clusters 2 and 3. This indicates that the observations within these

clusters share more similarities and that the boundaries between them are less distinct. Cluster 1 is more distinct, suggesting a different grouping characteristic in the data.

As show in the fig.1. By doing the K-means clustering, the patients are divided into different groups that can reflect different health states or risk levels of patients. The model can find which characteristics differ significantly across groups, the use this information to help identify people at high risk.

Fig. 1 Clustering (Photo credit: Original).

2.2.2 Elbow method

As show in the fig.2. When doing the K-means clustering, it is important to choose the suitable number of clusters. If not, the model is easy to fall into local optimality. By analyzing the clustering performance indexes under different K values, a reasonable K value is selected. Figure 2 illustrates the relationship between Number of Clusters and Sum of Squared Errors (SSE). SSE is the sum of the squares of the distance between each sample and the center of mass of its corresponding cluster. The point (3, 136000) in the graph where the SSE decline has slowed significantly is the elbow position indicates that 3 is likely the optimal number of clusters for the data set, as adding more clusters beyond this point results in only a marginal decrease in WSS.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} ||x - \mu_i||^2 \quad (1)$$

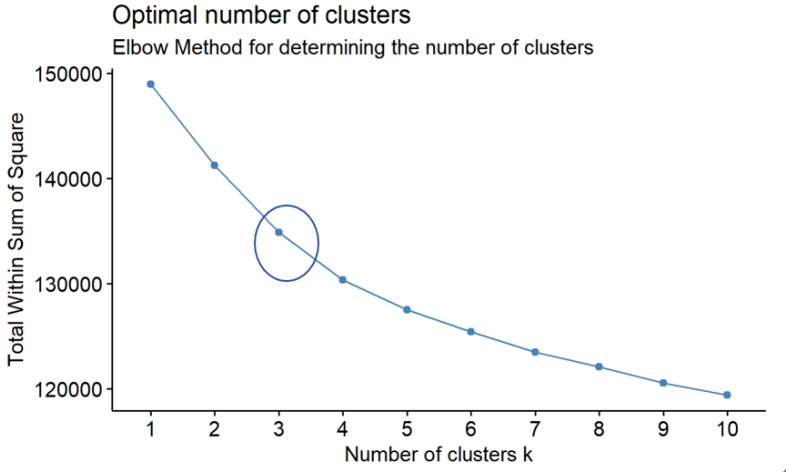

Fig. 2 Elbow Method for determining the number of clusters (Photo credit: Original).

2.2.3 Silhouette method

Another useful method to determine the number of clusters is Silhouette Method. It further assesses the degree to which each data point is assigned to the appropriate cluster and it can verify the K values chosen by the Elbow Method and compare the quality of clusters with different K values. By calculating and averaging the silhouette coefficients of each sample, an overall clustering quality index can be obtained. The Silhouette coefficient is used to measure how each sample is distributed within the cluster, and the value of it is between -1 and 1, with higher value indicating that the sample is correctly assigned to the appropriate cluster. The best K value is the K value with the highest average silhouette coefficient. The highest average silhouette width is observed at k=3. suggesting that 3 clusters provide the best separation and cohesion for the data set. After k=3, the average silhouette width decreases, indicating that additional clusters do not significantly improve the clustering quality.

$$\text{Average Silhouette Coefficient} = \frac{1}{n}\sum_{i=1}^{n} s(i) \tag{2}$$

By using Elbow Method to quickly determine a reasonable range of K values, and using the Silhouette Method to further evaluate the clustering quality of these K values, the optimal cluster number can be selected more scientifically, improving the reliability and interpretability of clustering results.

2.2.4 Factor analysis

Since clustering helps the data set separate the patients into three clusters, factor analysis is used to explore the underlying factors hidden in the data, which means find the main factors that cause patients get heart disease in each cluster. Another reason of choosing factor analysis is because there are some correlations among these factors. For instance: MR1: Strongly influenced by "Smoking" and "Sex," with correlation coefficients of 0.9 and 0.6 respectively. MR2: Includes variables like "Low Blood Pressure" and "Cholesterol," with coefficients of 0.7. MR3: Encompasses variables such as "Exercise Hours Per Week," "Family History," and "BMI."

2.3. Model Construction and Validation

Cross validation is a technique used to evaluate the performance of machine learning models, especially when data is limited. In normal test, the model is only evaluated in a fixed set of tests. This may result in the model to be overly optimistic or pessimistic on that particular partition. Cross-validation forces the model to learn on multiple different training sets, which means that the model is not optimized for just one specific training set, but to adapt to different combinations of data in each training. This diverse training process helps to improve the generalization ability of the model so that it can better adapt to previously unseen data and reduce the risk of overfitting on new data of patients. It can also be used to compare the performance of different models and identify which model is most likely to perform best on unseen data. The data set is separated into training_ data (80%) and test_ data (20%). Then the training_ data is further more separated into training data (80%) and validation data (20%).

Four types of models are built: Logical Regression, Random Forest, K Nearest Neighbour and Decision Tree Classifier. The final accuracy of these four models is 0.5103, 0.6345, 0.5888, 0.6476 respectively. The decision tree model had the best fit.

The analysis of the performance metrics for various predictive models, as reported in the data, shows varied results across different machine learning approaches, including Logistic Regression, Random Forest, K Nearest Neighbour, and Decision Tree Classifier. These results demonstrate each model's capabilities and limitations in predicting heart disease based on a comprehensive dataset. The models were evaluated on their Area Under the Curve (AUC) and Accuracy metrics across training, testing, and validation datasets.

Logistic Regression

- **Training with Validation Set**: The Logistic Regression model showed an AUC of 0.5019 and an Accuracy of 0.6391.
- **Training without Validation Set**: The model maintained an AUC of 0.5056 and identical Accuracy of 0.6391, indicating stability in its performance irrespective of the validation set.
- **Testing with Validation Set**: AUC was slightly higher at 0.5103, but the accuracy data is missing, which limits complete evaluation.
- **Testing without Validation Set**: The AUC remained steady at 0.5103.
- **Validation Set**: This model scored an AUC of 0.5139, suggesting a slight improvement in model performance under validation conditions.

Random Forest

- **Training with and without Validation Set**: The Random Forest model consistently showed an Accuracy of 0.6345. The stability in accuracy regardless of the validation set suggests robustness.
- **Testing with Validation Set**: Accuracy data is missing, which restricts a thorough comparison.
- **Testing without Validation Set**: It scored an Accuracy of 0.6345.
- **Validation Set**: There's missing data, making it difficult to fully assess its performance under validation conditions.

K Nearest Neighbour

- **Training with Validation Set**: Reported an AUC of 0.4949 and Accuracy of 0.5711.
- **Training without Validation Set**: The model mirrored its training performance with validation, showing the same AUC and Accuracy.
- **Testing with Validation Set**: Accuracy improved slightly to 0.5888.
- **Testing without Validation Set**: The accuracy was consistent at 0.5888.
- **Validation Set**: The model achieved an Accuracy of 0.6052, indicating a slight improvement when validated.

Decision Tree Classifier

- **Training with Validation Set**: Showed an AUC of 0.4890 and Accuracy of 0.6202.
- **Training without Validation Set**: Performance was consistent with the training that included validation, at the same AUC and Accuracy.
- **Testing with Validation Set**: The Accuracy was higher at 0.6476, indicating better performance in testing conditions.
- **Testing without Validation Set**: The Accuracy remained high at 0.6476.
- **Validation Set**: The Accuracy was slightly lower at 0.6452 but still above the training results.

Overall, the Decision Tree Classifier appears to perform relatively better in testing conditions compared to other models, especially noted in its higher Accuracy rates during testing. Logistic Regression shows marginal improvement in AUC when tested or validated. The Random Forest model indicates stability, but due to missing data in testing and validation, a full assessment is constrained. K Nearest Neighbour shows the least performance in training but improves during validation, suggesting sensitivity to the dataset's specifics. These insights underline the importance of selecting appropriate models based on specific criteria such as AUC and Accuracy and the necessity of comprehensive data to evaluate model performance thoroughly. Further, the missing data points in Random Forest and Logistic Regression testing and validation phases highlight the need for complete datasets to ensure accurate performance assessments and optimal model tuning.

2.4. Theoretical Foundations of the Models

2.4.1 Logistic regression

Logistic Regression uses a logical function to map the output of linear regression to a probability value between 0 and 1. It is used as the model of binary classification problem which is predicting whether a person is at risk for heart disease. The formula is given as:

$$P(y = 1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\ldots+\beta_n X_n)}} \quad (3)$$

Random Forest.

Random forest is an ensemble learning method that combines the prediction results of multiple decision trees to improve the accuracy of the model, which is especially suitable for dealing with complex and nonlinear classification problems. In heart disease risk prediction, Random Forest can train decision tree models with large amounts of data (smoking, blood pressure, family medical history, etc.). Each tree makes a prediction based on a different combination of features, and the final prediction result is the average or majority vote of all tree predictions. It can generalize data and effectively reduce the overfitting risk of a single model

2.4.2 K-Nearest neighbour (KNN)

KNN is an example learning method that directly uses training data to make predictions. It finds the nearest K neighbors by calculating the Euclidean distance between the new sample and the training sample of patients. The Euclidean distance formula is:

$$d(i,j) = \sqrt{\sum_{n=1}^{N}(x_{i,n} - x_{j,n})^2} \quad (4)$$

The risk of new patients is then predicted based on data about their medical history (whether or not they develop heart disease).

2.4.3 Decision tree classifier

The decision tree is gradually split according to different characteristics of the patient (such as age, gender, lifestyle), and finally provides a clear classification result (heart disease risk) for each patient. The visual nature of the tree structure allows it to interpret predictions.

3. Experimental Results and Analysis

3.1. Quantitative Analysis of Heart Disease Risk Factors

Overall, by using statistical and computational methods to identify the impact of various risk factors on heart disease is useful. The data set is very comprehensive in the variables of heart disease, although the number of samples is too big. From the experimental results, it is clear the variables that are most strongly associated with heart disease are smoking, sex and age. And it is proved by the models' accuracy. It is shown that heart disease happens more in males than females. From some medical researches, female smokers show a 25% higher risk of developing CHD than men and older females are at a greater risk for CVS. So, it is clear that the relationships among factors are quite strong.

3.2. Model Performance Evaluation

K-fold cross-validation and the validation set method were employed to assess the decision tree model's performance. Despite these validation techniques, the best model achieved only 64% accuracy, falling short of the anticipated 70%. This suggests that the model was underfitting. The omission of the validation set method in subsequent tests did not significantly improve the accuracy, which remained below the predicted target.

Performance metrics from both the training set and test set were analyzed, showing high reliability. Notably, the accuracy of the test set was slightly higher than that of the training set, even though the test data was entirely independent. The F1 score from the training set indicated suboptimal balanced performance, falling below 80%. However, the sensitivity exceeded 90%, demonstrating the model's strong ability to identify true positives. The overall lack of expected accuracy underscores the model's limitations, likely due to underfitting.

4. Discussion

4.1. Diagnosis and Solutions for Overfitting Issues

A data set that contains a huge amount of data always has a problem called high dimensionality. It means the distance between data points increases, the input space that the model needs to deal with expands rapidly, making the combination of possibilities that the model considers grow exponentially. As a result, the model may not have enough samples when trained to effectively cover the entire high-dimensional space, resulting in the model failing to learn the underlying structure of the data correctly. Additionally, in high-dimensional data, too many parameters may lead to too strong fitting ability of the model, and even features that are not important in the training data will be captured by the model as meaningful. This results in models that perform well on training data but poorly on new data. In order to deal this problem, PCA is used to project a higher-dimensional Euclidean space into a lower-dimensional Euclidean space to ensure that each feature is compared on the same scale.

Cross validation is also a good method to prevent overfitting through the performance of the model on unseen data is ensured by balancing training and validation on multiple subsets.

4.2. Underfitting and Data Set Challenges

During the modeling process, the maximum depth of the tree was adjusted to try to make the model more complex and achieve the expected accuracy. A smaller depth can prevent overfitting but may cause the model to be too simple to capture complex relationships in the data. A greater depth may lead to a model that is overly complex and prone to overfitting. Currently, the model is in a state of underfitting, and increasing the depth did not significantly change the accuracy. It is inferred that this is a data limitation issue, where increasing the tree's depth does not bring additional predictive power. The data set used in this study comprises 8,763 samples. A larger data set with more balanced data could potentially enhance the model's performance by allowing for better-fitting models. In this instance, insufficient attention was paid to parameter optimization in the decision tree model, leading to suboptimal parameter values and a consequent reduction in accuracy. To address the underfitting issue and improve model accuracy, exploring more advanced methods such as neural networks, coupled with interdisciplinary collaboration, may offer promising solutions.

5. Conclusion

This study conclusively identifies age, sex, smoking, stress level, and cholesterol as significant predictors of heart attack risk, leveraging a sophisticated predictive model to facilitate personalized risk assessments and tailored health management plans. By integrating these individual characteristics, the developed model enhances the precision and efficacy of heart disease treatment and prevention strategies. Despite the potential of the model, challenges such as overfitting and underfitting were addressed using a range of methodologies, including data standardization and the meticulous handling of outliers. Principal Component Analysis (PCA) and cross-validation played critical roles in refining the model's construction, although it exhibited slight underperformance due to underfitting. This issue was thoroughly examined, with discussions highlighting the necessity of selecting a diverse yet relevant feature set to ensure the model's complexity does not detract from its practical applicability.

Looking ahead, there is substantial scope to augment the capabilities of predictive models by incorporating real-world patient data and a broader array of health-related variables, which could enhance the model's applicability to more complex and variable cases of heart disease. Additionally, adopting more advanced predictive algorithms may offer deeper insights into the nuanced patterns of heart attack risks. Ongoing refinement and enhancement of these predictive models are imperative for advancing proactive healthcare strategies. Such developments will not only improve the accuracy of heart disease predictions but also contribute significantly to the reduction of heart attack incidences, paving the way for more effective preventive healthcare measures globally.

References

[1] Abdul Salam M., Azar A. T., Elgendy M. S., et al. The effect of different dimensionality reduction techniques on machine learning overfitting problem. International Journal of Advanced Computer Science and Applications, 2021, 12(4): 1-8.

[2] Bharti R., Khamparia A., Shabaz M., et al. Prediction of heart disease using a combination of machine learning and deep learning. Computational Intelligence and Neuroscience, 2021, Article ID 8387680, 11 pages.

[3] Zhu X., Huang Y., Wang X., et al. Emotion recognition based on brain-like multimodal hierarchical perception. Multimed. Tools Appl., 2024, 83(18): 56039-56057.

[4] 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, 2010, 2(2): 250-255.

[5] Wang R., Zhu J., Wang S., Wang T., Huang J., Zhu X. Multi-modal emotion recognition using tensor decomposition fusion and self-supervised multi-tasking. International Journal of Multimedia Information Retrieval, 2024, 13(4): 39.

[6] Acharya A. Comparative study of machine learning algorithms for heart disease prediction. 2017.

[7] Yewale D., Vijayaragavan S. P., Bairagi V. K. An effective heart disease prediction framework based on ensemble techniques in machine learning. International Journal of Advanced Computer Science and Applications, 2023, 14(2).

[8] Ramprakash P., Sarumathi R., Mowriya R., et al. Heart disease prediction using deep neural network[C]//2020 international conference on inventive computation technologies (ICICT). IEEE, 2020: 666-670.

[9] Zhu, X., Guo, C., Feng, H., Huang, Y., Feng, Y., Wang, X., & Wang, R. (2024). A Review of Key Technologies for Emotion Analysis Using Multimodal Information. Cognitive Computation, 1-27.

[10] Ufumaka I. Comparative analysis of machine learning algorithms for heart disease prediction. Int. J. Sci. Res., 2021, 11: 339-346.