Milestone Report

Group 1: Jianxing Wu, Jieming Yu

Topic Selection

Our project topic is "Prediction of Heart Disease with Machine Learning". We choose option 2: empirical evaluation.

Introduction

Heart disease is a broad term encompassing a range of conditions that affect the heart. Cardiovascular diseases are currently the leading cause of mortality worldwide, with the World Health Organization reporting 17.9 million deaths attributed to them annually [1]. The American Heart Association notes several potential warning signs of heart disease, including sleep disturbances, irregular heartbeat, and swollen legs [2]. However, these symptoms resemble various diseases, particularly in the senior, making accurate diagnosis of heart disease challenging. Nowadays, many researchers have trained machine learning models to diagnose heart disease on datasets created by a substantial amount of hospital records, which yielded promising results in prediction accuracy. In our project, we plan to implement 4 machine learning methods to predict the presence of heart disease including Logistic Regression, Support Vector Machine, Decision Tree and Random Forest. The dataset we will be working on is the classic Cleveland Heart Disease dataset.

Literature Review

We reviewed previous studies implementing the machine learning methods for heart disease prediction to provide some insights and guidance. The findings are as follows.

Logistic Regression (open source)

Logistic Regression was used as a benchmark model to compare with neural networks in the study by Allam et al. [3]. The model was imported from scikit learn using default settings. For data preprocessing, ANOVA was implemented to find out the most informative features. And to tune hyperparameters, the authors tried both L1 and L2 regularization with a range of C values in order to find an optimal pair of regularization strength and penalty type.

SVM (open source)

In a study by Rodriguez and Nafea[4], the author trained a SVM model with linear kernel, and a scaled gamma. A linear kernel assumes the data is linear separable. Gamma controls the complexity of decision boundary. For example, a large gamma will lead to a winding decision boundary sensitive each datapoint, which may cause the model to overfit. The author used the scikit learn built in function to scale the gamma according to the dataset in order to avoid overfitting and underfitting. The SVM model achieved a testing accuracy of 83.3%, outperforming Logistic Regression by 2.8%.

Decision Tree (open source)

In clinical practice, decision trees offer advantages over other classification methods. Their transparency aids in clinical decision-making, and their ability to handle unfamiliar data makes them valuable for diagnosing

new cases [5]. Training on a relatively small dataset, decision trees can easily overfit. We will control the maximum depth of the tree or limit leaf nodes in minimize the odds of overfitting. In a study by Mishra [6], the author used grid search to optimize the parameters of the Decision Tree model from scikit learn, reaching a accuracy of 78% on a similar heart disease dataset.

Random Forest (open source)

Kibria and Matin explored Random Forest as one of the models in the prediction of heart disease [7]. Different from decision tree, Random Forest randomly chooses some rows and unique features from the dataset to create several trees and then combines the result. Hyperparameter tuning is also important for random forest, grid search or random search techniques will be utilized to systematically explore hyperparameters such as number of estimators, maximum depth, maximum features. In this study, the number of estimators for Random Forest was tuned to 100. The accuracy reached 65.57%.

Conclusion

After reading previous papers, we found that SVM has the best performance on the heart disease prediction task. In our project, we will implement the for methods and make our own comparison. Additionally, we will use techniques learned in class such as such as PCA and cross validation to see if we can achieve a better result.

References

[1] “Cardiovascular diseases.” Accessed: Oct. 25, 2024. [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases

[2] T. K. Bekmurodovna and Z. A. Chorievich, “Study of frequency indicators of comorbid states at different functional classes of heart failure,” *Acad. Int. Multidiscip. Res. J.*, vol. 11, no. 3, pp. 2556–2560, 2021.

[3] A. Allam, M. Nagy, G. Thoma, and M. Krauthammer, “Neural networks versus Logistic regression for 30 days all-cause readmission prediction,” *Sci. Rep.*, vol. 9, no. 1, p. 9277, 2019.

[4] M. Padilla Rodriguez and M. Nafea, “Centralized and Federated Heart Disease Classification Models Using UCI Dataset and their Shapley-value Based Interpretability,” *ArXiv E-Prints*, p. arXiv-2408, 2024.

[5] A. Alanazi, “Using machine learning for healthcare challenges and opportunities,” *Inform. Med. Unlocked*, vol. 30, p. 100924, 2022.

[6] S. Mishra, “A comparative study for time-to-event analysis and survival prediction for heart failure condition using machine learning techniques,” *J. Electron. Electromed. Eng. Med. Inform.*, vol. 4, no. 3, pp. 115–134, 2022.

[7] H. B. Kibria and A. Matin, “The severity prediction of the binary and multi-class cardiovascular disease- A machine learning-based fusion approach,” *Comput. Biol. Chem.*, vol. 98, p. 107672, 2022.