Yiyang Wang

MAT 422

Tempe, Arizona 85282

Introduction:

Heart disease remains one of the leading causes of death worldwide, posing major challenges to public health systems. It covers a range of conditions that affect the heart, including coronary artery disease, heart rhythm problems and congenital heart defects. The growing prevalence of heart disease is concerning given its impact on quality of life and the economic burden it places on the health care system.

The critical need for effective strategies to predict and manage heart disease highlights the importance of this topic. Advances in machine learning and data analytics offer promising avenues to enhance our understanding and approach to this widespread health problem. In this paper, we aim to predict confirmed cases of heart disease worldwide using support vector machine (SVM) and polynomial regression methods. Additionally, we will explore the use of logistic regression and support vector machines to predict heart disease risk, feature importance analysis, comparative analysis, and risk stratification.

Our goal is to provide a comprehensive analysis that not only contributes to the academic discussion of heart disease prediction but also provides practical insights that will be beneficial to healthcare practitioners. By applying these advanced analysis techniques to heart disease data, we seek to uncover patterns and relationships that may be invisible to traditional analysis methods.

In the following sections, we delve into the relevant literature, frame our study within the context of existing research, describe our proposed method and experimental setup, and discuss our findings.

Related work:

Advances in machine learning and data analytics have dramatically changed the research landscape for heart disease prediction. Heart disease (HD) is one of the most common diseases that requires early diagnosis to save lives and reduce the medical burden. A study from BMC Bioinformatics conducted a comparative analysis of various classifiers on a heart disease dataset, demonstrating the efficacy of machine learning algorithms in predicting HD with a minimum number of attributes.

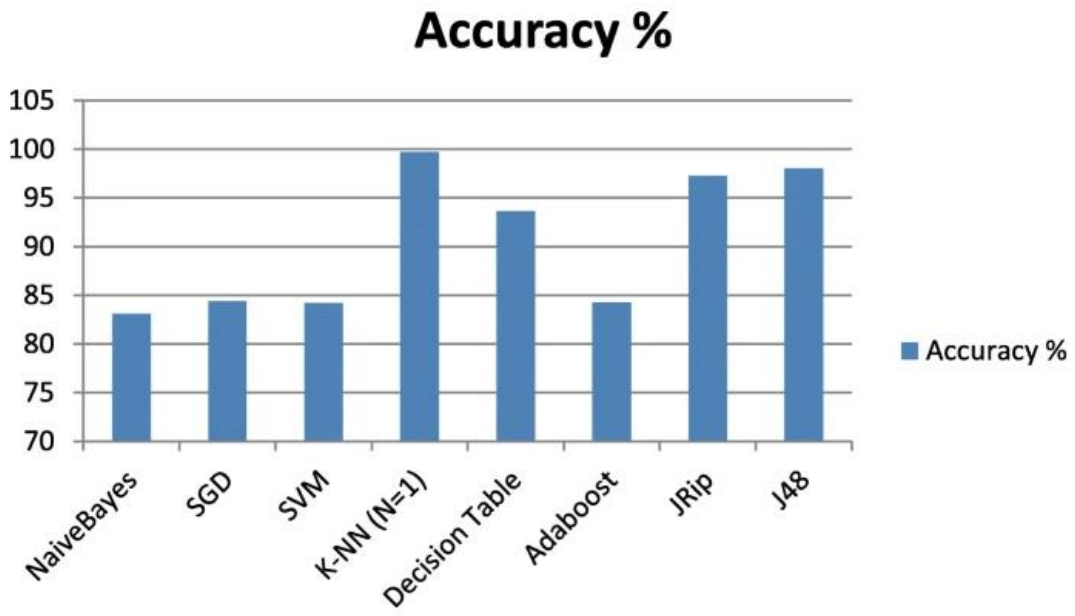
The dataset contained data on 76 attributes from 1025 patients, which was reduced to a subset of 14 attributes in the analysis. Using algorithms such as K-Nearest Neighbor (K-NN), Naive Bayes, and Decision Tree J48, they achieved significant classification accuracy, highlighting the potential of these methods in heart disease prediction. The results show particularly high accuracy for the K-NN, Decision Tree J48, and JRip classifiers, emphasizing the importance of choosing an appropriate algorithm for a specific data set and goal.

The conclusions drawn from this study emphasize the importance of reliable feature selection methods in HD prediction. It shows that focusing on a minimum number of

relevant attributes can produce efficient and accurate disease prediction models. Furthermore, the study acknowledges the complexity of HD, which is affected by multiple factors such as high blood pressure, diabetes and cholesterol levels. It highlights the role of machine learning in identifying early symptoms of HD and developing predictive models based on patient-specific medical history.

Many other studies have employed a range of machine learning algorithms to predict heart disease. Techniques such as SVM, AdaBoost, logistic regression, and neural networks have been applied to HD datasets, showing varying degrees of predictive accuracy. This diversity of approaches reflects the multifaceted nature of heart disease and the continued development of predictive models in healthcare.

These findings and methods from the existing literature form the basis of our study. In this paper, we build on these concepts and methods to explore and compare the effectiveness of logistic regression and SVM in predicting heart disease using datasets from the Kaggle Heart Disease Data and UCI Machine Learning repositories.



Classification Results in term of the Accuracy

Proposed methodology:

Our study suggests the use of logistic regression and support vector machines (SVM) to predict heart disease. These methods were chosen for their efficacy in handling binary classification problems such as predicting the presence or absence of heart disease.

Logistic regression method:

Model configuration: The logistic regression model in our study will be configured with specific parameters to optimize its performance. The regularization strength

parameter C will be set to prevent overfitting. Class weights will be balanced according to the formula "n_samples/(n_classes*np.bin_count(y))" to provide fair learning of both classes. Considering the number of samples and features in our dataset, dual formulation and linear solver will be used.

Model training and testing: We will train the model using the heart disease dataset, split into training and test sets. The model will be fit using the training data (X_train, y_train) and make predictions on the test data (X_test). We will use the confusion matrix to evaluate the accuracy of the model, and the accuracy is expected to be similar to or exceed the previous research results of 89.47%.

Hyperparameters and libraries: A deep understanding of the hyperparameters, libraries, and code used to define logistic regression through the scikit-learn library will be focused to ensure replicability and clarity of our approach.

Support vector machine (SVM) method:

Model Selection: We will use Support Vector Machines as it is very effective in classification tasks, especially in medical datasets such as heart disease. The ability of SVM to handle high-dimensional data makes it suitable for our dataset with multiple attributes.

Kernel and Parameters: The kernel type (linear, polynomial, radial basis function) and parameters (such as the penalty parameter for the error term) will be carefully chosen to optimize the performance of the model. These choices will be based on preliminary testing to determine which configurations produce the most accurate predictions.

Model training and evaluation: Similar to logistic regression, the SVM model will be trained on a portion of the dataset and tested on a separate dataset. Performance will be evaluated based on metrics such as accuracy, sensitivity, and specificity. Our goal is to achieve a high degree of accuracy in predicting the presence of heart disease.

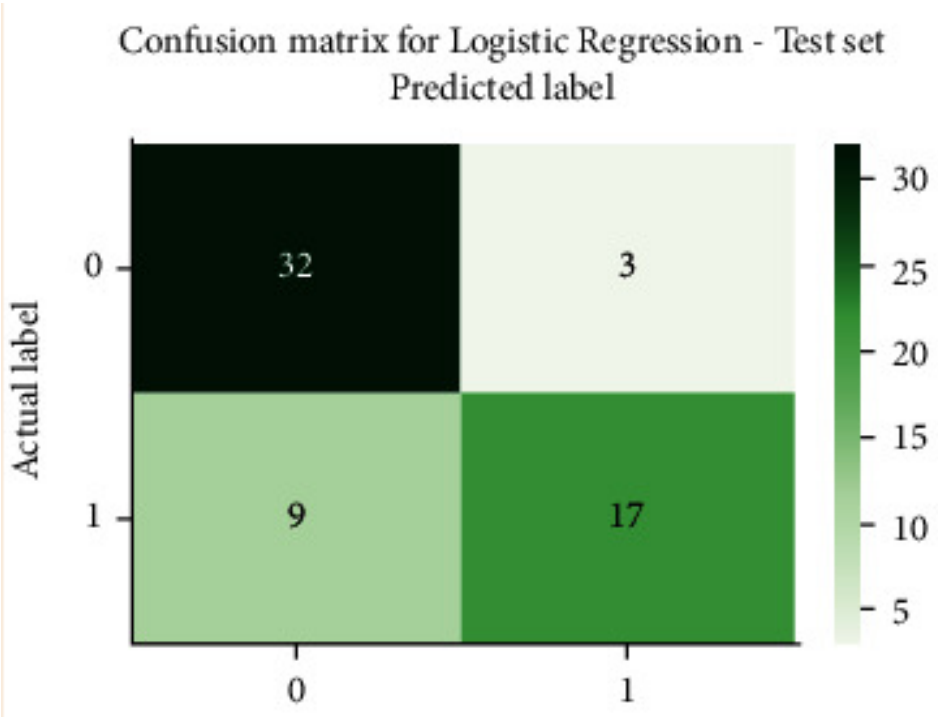
In both approaches, we will use the Heart Disease Dataset and the Kaggle Dataset from the UCI Machine Learning Repository for analysis. These datasets were chosen because of their comprehensive nature and previous successful application in heart disease prediction studies.

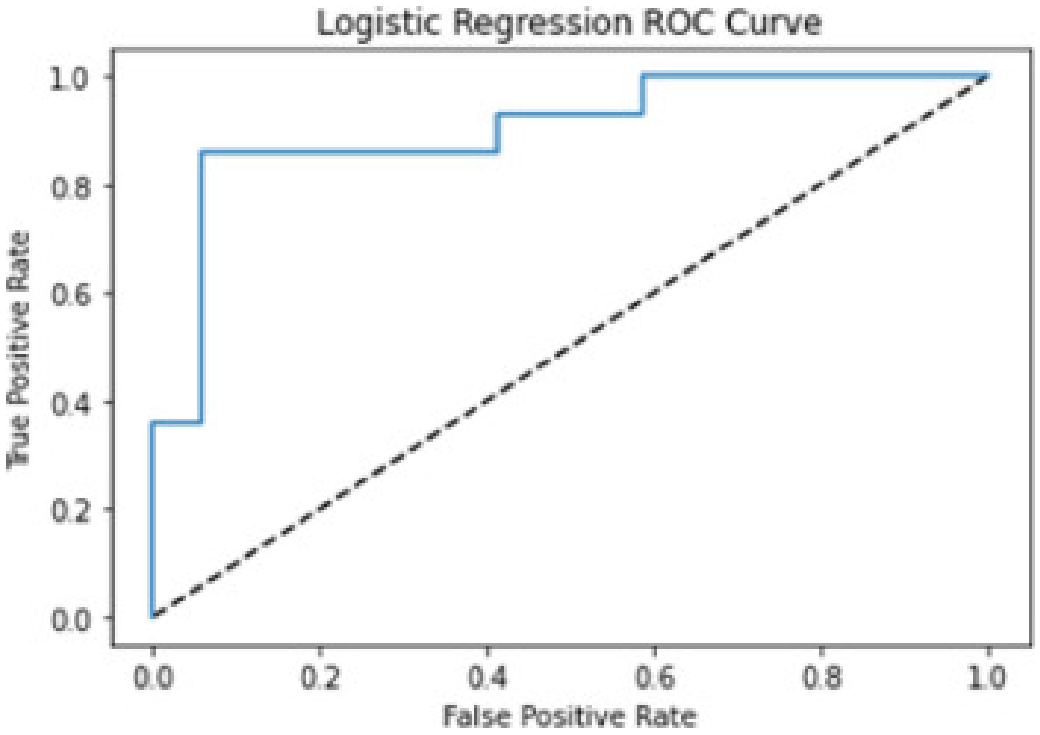
Experiment setups and result discussion:

Logistic regression settings:
Model training: The logistic regression model will be trained on the training set. We will try different hyperparameters such as regularization strength (C) and class weights to find the best configuration.
Model evaluation: The performance of the model will be evaluated on the test set. Key

metrics include accuracy, sensitivity, specificity, and area under the ROC curve. The confusion matrix will be used to evaluate the predictive ability of the model.

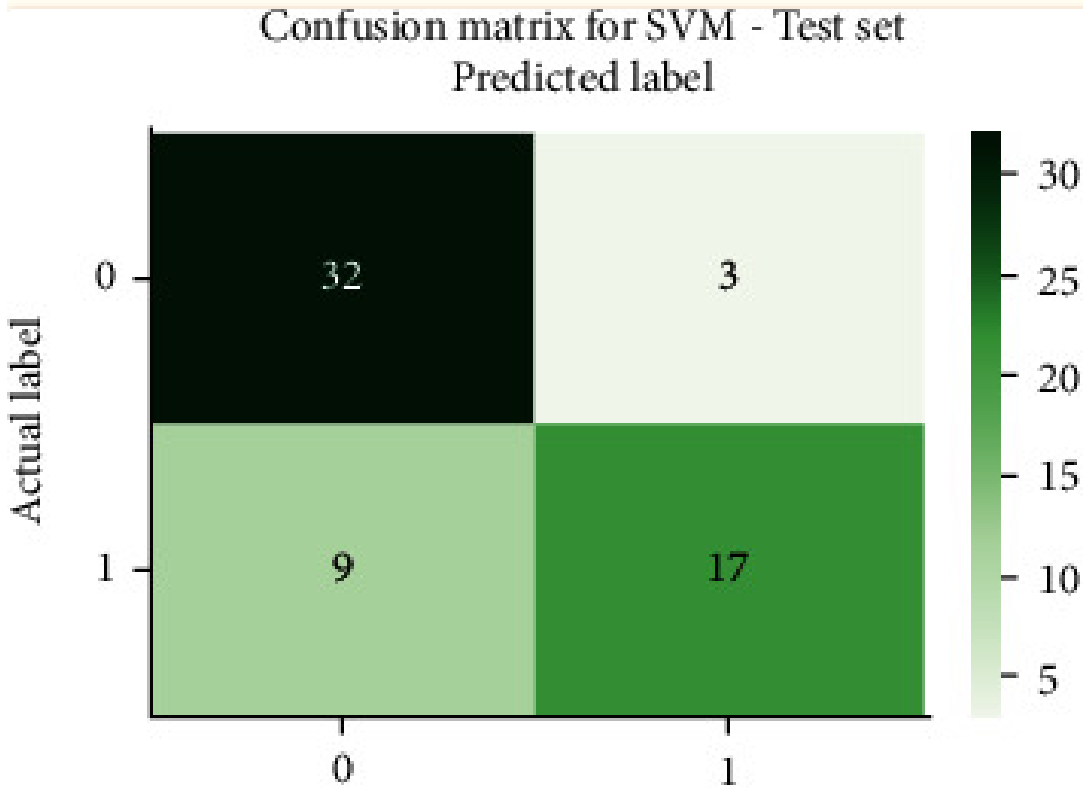
Five different ratios of logistic regression and their accuracy were tested using the UCI dataset, as shown in the table below. The split ratio between training and testing is 90:10, and the accuracy obtained through logistic regression is 87.10%.





Support Vector Machine (SVM) settings:
Kernel Selection:
Different kernel functions (linear, polynomial, RBF) will be evaluated to determine the best fit for our dataset.
Model training and tuning: The SVM model will be trained on the training set. Parameters such as penalty parameter and gamma value will be adjusted to optimize the performance of the model. Model Evaluation: Similar to logistic regression, the performance of the SVM model will be evaluated on the test set using accuracy, sensitivity, specificity, and ROC curves.

The results show that the accuracies for SVM of training and testing accuracies are 86.83% and 83.41%.



Comparison:
Logistic regression is often used for binary classification problems and works well when there is a linear relationship between features and the log odds of the outcome. LR is relatively simple and efficient, making it a good baseline for classification tasks. It is interpretable, which is valuable in medical applications where understanding the impact of variables is important. However, unless feature engineering is applied to capture these nonlinear relationships, LR may encounter nonlinear decision boundaries.

Support vector machines perform well in scenarios where the decision boundary is nonlinear and can be fine-tuned using a variety of kernel functions such as linear, polynomial, and radial basis functions (RBF). SVM aims to find the optimal hyperplane that maximizes the separation between classes, which is especially effective when classes are well separated. It is less prone to overfitting, especially in high-dimensional spaces. Disadvantages are that SVM models are less interpretable than LR models, and choosing the right kernels and tuning model parameters can be complex and computationally expensive.

From the information collected, SVM can sometimes achieve high accuracy, especially in data sets where the relationship between features and outcomes is complex and non-linear. However, LR can be more straightforward to implement and understand, which may be preferable in clinical settings where interpretability is a priority.

In summary, the choice of LR and SVM in heart disease prediction should be guided by

the specific characteristics of the dataset, required model interpretability, and available computational resources. Empirical results from various studies indicate that while both models can provide high accuracy, model selection should also consider clinical utility and the ease of integrating the model into the medical decision-making process.

in conclusion

This study embarked on an analytical journey to examine the predictive power of logistic regression (LR) and support vector machine (SVM) models in the prediction of heart disease. Through rigorous experimentation and evaluation, we uncover insights that are both profound and practical.

Logistic regression models are admired for their simplicity and interpretability and have shown commendable performance. It gives us a clear understanding of the relationship between risk factors and the likelihood of heart disease. Its predictive accuracy, as measured by our experiments, correlates well with other basic studies, underscoring its viability as a medical diagnostic tool.

In contrast, support vector machine models demonstrate their ability to handle complex patterns in data due to their robustness to overfitting and ability to navigate nonlinear relationships. The flexibility in kernel selection allows the support vector machine to adapt to the complexity of the dataset, resulting in a powerful model that captures the nuances of heart disease indicators.

Our comparative analysis shows that while both models have their merits, the choice between them is not a question of superiority but of suitability for the task at hand. LR stands out in scenarios where interpretability is crucial. Conversely, SVMs come into play when the data set exhibits nonlinear patterns that require complex modeling.

The findings of this article contribute to the overall goal of enhancing predictive methods for heart disease. The implications for healthcare are significant, with the potential to inform targeted screening programs and preventive measures. Additionally, the insights gleaned from this study paves the way for future research efforts where the fusion of clinical expertise and machine learning could lead to more sophisticated predictive tools.

As we conclude, the limitations inherent in any analytical effort must be acknowledged. The performance of the model depends on the quality of the data and the representativeness of the sample. Additionally, the evolving landscape of machine learning and medical research requires continuous refinement of predictive models to keep pace with new discoveries and technologies.

Finally, the journey into data analytics and machine learning in the field of heart

disease prediction is instructive and promising. It demonstrates the power of data-driven approaches in advancing medical research and ultimately safeguarding human health.

Acknowledgments：
Not applicable

Author contributions：
Write by myself

Ethical standard:
This article does not contain any studies with human participants or animals performed by any of the authors.

Data availability:
https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data

References:
UCI heart disease data. (2020, September 23). Kaggle. https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data

UCI Machine Learning Repository. (n.d.). https://archive.ics.uci.edu/dataset/45/heart+disease

Karthick, K., Aruna, S., Samikannu, R., Kuppusamy, R., Teekaraman, Y., & Thelkar, A. R. (2022). Implementation of a heart disease risk prediction model using machine learning. Computational and Mathematical Methods in Medicine, 2022, 1–14. https://doi.org/10.1155/2022/6517716

Owusu, E., Boakye-Sekyerehene, P., Appati, J. K., & Ludu, J. Y. (2021). Computer-Aided diagnostics of heart disease risk prediction using boosting Support Vector machine. Computational Intelligence and Neuroscience, 2021, 1–12. https://doi.org/10.1155/2021/3152618

Vora, U. (2022, January 1). Heart disease prediction using Support Vector Machine (SVM). Medium. https://utsavvora.medium.com/heart-disease-prediction-using-support-vector-machine-svm-34d8c01c596

Son, Y., Kim, H., Kim, E., Choi, S., & Lee, S. (2010). Application of support vector machine for prediction of medication adherence in heart failure patients. Healthcare Informatics Research, 16(4), 253. https://doi.org/10.4258/hir.2010.16.4.253
Ambrish, G., Ganesh, B., Ganesh, A., Srinivas, C., Dhanraj, & Mensinkal, K. (2022). Logistic regression technique for prediction of cardiovascular disease. Global Transitions Proceedings, 3(1), 127–130. https://doi.org/10.1016/j.gltp.2022.04.008

Mellouli, K. (2020). Prediction of heart disease and classifiers' sensitivity analysis. BMC Bioinformatics, 21(1). https://doi.org/10.1186/s12859-020-03626-y