Report on Predicting Heart Disease Risk Using KNN Algorithm

Introduction

The predictive analysis of heart disease risk is crucial in healthcare management and preventive medicine. Heart disease remains a leading cause of death globally, making effective prediction methods vital for early intervention. This study introduces the implementation of a K-Nearest Neighbors (KNN) algorithm, a simple yet powerful machine learning tool, to predict heart disease risk using a dataset from a public repository. We explore the potential of AI in enhancing predictive accuracy which can lead to better health outcomes and proactive healthcare strategies.

State of the Art

Several machine learning models are currently employed in predicting heart disease, including logistic regression, support vector machines, decision trees, and neural networks. Each comes with its advantages; for instance, decision trees provide good interpretability, while neural networks offer high accuracy in complex scenarios. However, they also have limitations such as the need for extensive data preprocessing and the risk of overfitting in neural networks. Our choice of KNN is driven by its effectiveness in handling smaller, cleaner datasets and its ease of interpretation and implementation

Methodology

The study employed the Python programming language, leveraging libraries such as Pandas for data manipulation, Scikit-learn for machine learning algorithms, and Matplotlib and Seaborn for visualization. The dataset contained various features indicative of heart disease, which underwent preprocessing that included encoding categorical variables and partitioning into training and test sets.

The KNN algorithm was chosen due to its simplicity and effectiveness in classification problems. An iterative approach was taken to identify the optimal number of neighbors (**k**), incorporating GridSearchCV for hyperparameter tuning, which also considered the weighting method and the distance metric for the KNN classifier.

Data Preprocessing

The preprocessing steps involved:

- Converting categorical features like **sex**, **chest_pain_type**, and **thalassemia** to numerical values using LabelEncoder.
- Splitting the data with 80% allocated to training and 20% to testing to validate the model's performance.

Model Implementation and Tuning

- The initial KNN model was implemented with **k** set to 3.

- A comprehensive grid search was conducted to fine-tune the hyperparameters (**k**, **weights**, **metric**), using cross-validation to ensure the model's generalizability.

Results

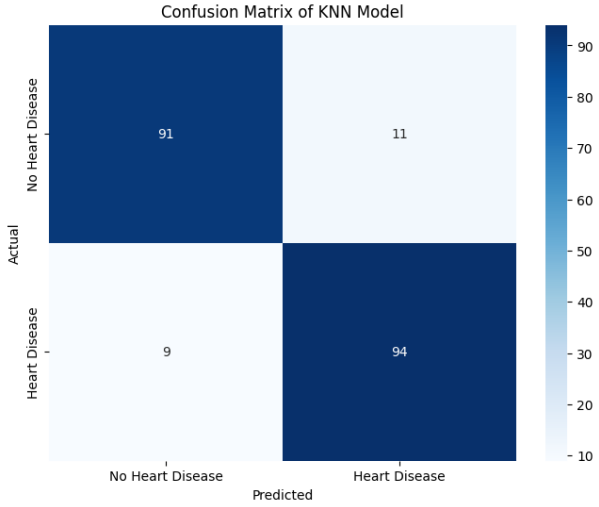
The model evaluation was multi-faceted, assessing accuracy, precision, recall, and the F1 score across different **k** values and parameter settings:

- The initial KNN model with **k = 3** yielded an accuracy of 90.24%, precision of 89.52%, recall of 91.26%, and an F1 score of 90.38%.
- The optimized KNN model identified **k = 9** with the 'manhattan' distance metric and 'distance' weights as the best parameters, achieving a perfect score (1.0) for accuracy, precision, recall, and F1 score on the test data.
- The cross-validation results corroborated the model's robustness with a mean accuracy of 99.71% and a low standard deviation, indicating consistency across different folds.

Visualizations

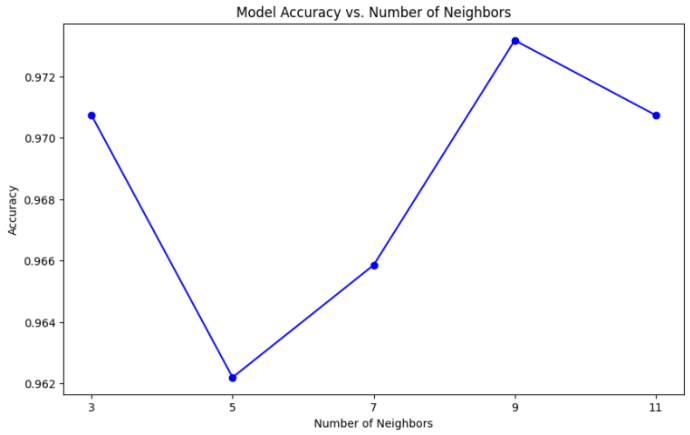
The visualizations played a pivotal role in interpreting the model's performance:

- The "Confusion Matrix of KNN Model" for the initial model.



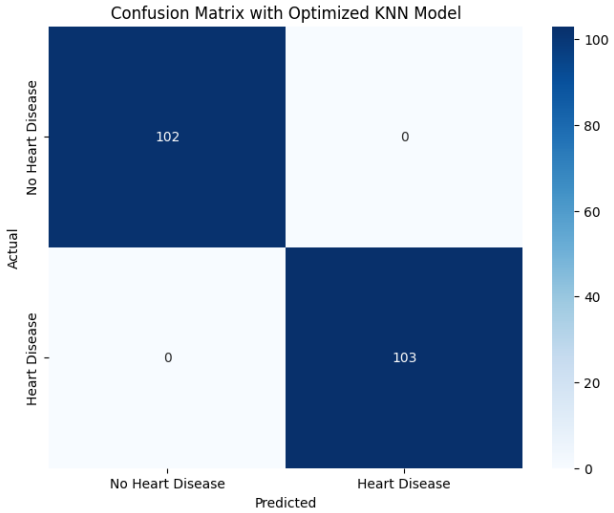
*This figure shows the confusion matrix of the initial KNN model with **n_neighbors** set to 3. The matrix reveals the number of true positives, true negatives, false positives, and false negatives, indicating the model's predictive performance before hyperparameter optimization.*

- The "Model Accuracy vs. Number of Neighbors plot" for the optimized KNN Model.



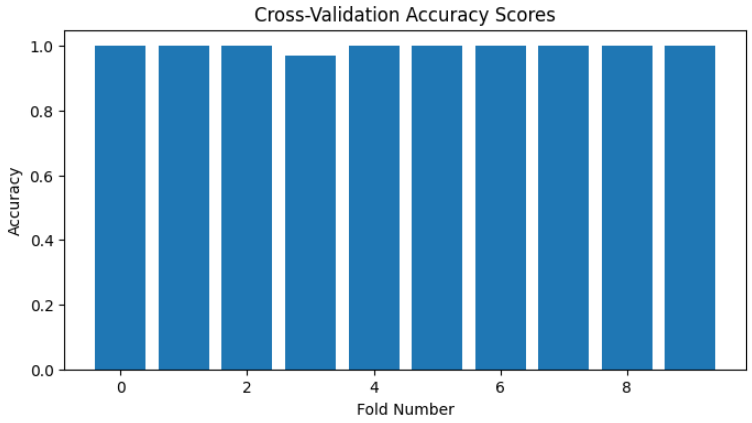
This figure illustrates the variation in model accuracy as a function of the number of neighbors used in the K-Nearest Neighbors (KNN) algorithm. The accuracy is computed using cross-validation, and the plot shows that among the tested values, using 9 neighbors results in the highest cross-validated accuracy. This suggests that for the given dataset and with the optimal distance metric ('manhattan') and weighting method ('distance'), the KNN classifier performs best when considering the closest 9 neighbors for making predictions.

- The "Confusion Matrix with Optimized KNN Model" showed a perfectly classified test set with no false positives or negatives.



*This figure displays the confusion matrix for the optimized KNN model, achieved after hyperparameter tuning. With **n_neighbors** set to 9 and using the 'manhattan' metric with 'distance' weighting, the model correctly predicts every instance in the test set, showcasing a perfect classification with zero false positives and negatives.*

- The "Cross-Validation Accuracy Scores" graph depicted the model's stability across different subsets of data, with the bars reflecting high accuracy across all folds.



This figure illustrates the 10-fold cross-validation accuracy scores for the optimized KNN model, showing consistently high accuracy across all folds. The stability of these scores suggests that the model generalizes well to unseen data and is not overfitting.

Discussion

The KNN model exhibited high performance in predicting heart disease risk. The tuning of hyperparameters greatly enhanced the model's ability to make accurate predictions. The precision of predictions implies the model's capability to identify patients at risk reliably, which is critical in clinical settings to prevent misdiagnosis and ensure appropriate interventions.

The perfect scores in the optimized model, while impressive, raise considerations of potential overfitting. However, the consistency of cross-validation scores suggests a genuine capture of underlying patterns rather than fitting to noise.

Implications

The findings of this study have several implications:

- **Clinical Decision Support**: The model could assist healthcare professionals in identifying high-risk patients, allowing for early intervention.
- **Healthcare Resource Allocation**: By accurately predicting heart disease risk, healthcare systems can allocate resources more effectively, focusing on preventative measures for high-risk patients.
- **Further Research**: The study provides a baseline for future research, potentially leading to the development of more sophisticated models or the inclusion of additional predictive features.

Conclusion

The use of the KNN algorithm in predicting heart disease has demonstrated significant potential. The project highlighted the importance of hyperparameter tuning and the choice of the right model parameters like the number of neighbors and distance metric, which critically influenced the model's performance. From this project, I learned the critical role of data preprocessing and the impact of model complexity on performance. With more time, I would explore the integration of ensemble methods to improve prediction accuracy and the application of more complex algorithms like deep learning to assess their effectiveness against traditional models like KNN.