Enhancing Heart Disease Prediction: A Comparative Analysis of Data Mining Techniques

Mateusz Słowik
mateusz.slowik@ru.nl

ABSTRACT

This study addresses the critical challenge of predicting heart disease using data mining techniques. In the analysis two prominent machine learning models: the Random Forest Classifier and K-Nearest Neighbors (KNN) were used. Through a careful process of feature selection, model training, and optimization, the project aims to maximize the predictive accuracy. The models are evaluated based on accuracy, precision, recall, and F1 score. Our findings reveal that the optimized Random Forest model outperforms its initial configuration and the KNN model, achieving superior accuracy and precision. This research demonstrates the potential of machine learning in enhancing predictive analytics in healthcare, specifically for heart disease prediction.

1 INTRODUCTION

Heart disease remains one of the most significant health challenges globally, being a leading cause of mortality. The complexity of cardiovascular diseases necessitates advanced diagnostic methods for early detection and effective management. In this context, the role of data mining in healthcare has grown exponentially, offering new avenues for analyzing extensive clinical data to predict and prevent heart diseases.

Advancements in machine learning have opened up promising prospects for improving diagnostic accuracy. These technologies are capable of processing large datasets, identifying hidden patterns, and learning from clinical records, thereby enhancing predictive analytics in cardiology. This project aims to leverage these advancements to address the urgent need for reliable heart disease prediction tools.

1.1 Research Problem

The primary challenge in heart disease prediction is the accurate analysis of clinical data, which often contains complex, multi-dimensional variables. Traditional diagnostic approaches, while effective, may not fully utilize the latent patterns and correlations present in the data.

This project employ two sophisticated machine learning models: the Random Forest Classifier and the K-Nearest Neighbors (KNN). Each model has its unique strengths in handling classification problems, and their comparative analysis in this study aims to identify the most effective approach for heart disease prediction.

The research specifically focuses on:

- Analyzing a comprehensive dataset of heart disease patients, encompassing a wide range of clinical attributes.
- Implementing feature selection techniques to identify the most relevant predictors of heart disease.
- Training and optimizing the Random Forest and KNN models to achieve the highest possible accuracy.
- Comparing the performance of these models based on accuracy, precision, recall, and F1 score to determine their effectiveness in practical diagnostic applications.
- Investigating the models' behavior through receiver operating characteristic (ROC) analysis, providing insights into their diagnostic capabilities.

By addressing these objectives, the research contributes to the broader field of medical data analytics, offering new perspectives and tools for heart disease prediction. The findings of this study are expected to enhance the diagnostic processes and aid healthcare professionals in making more informed decisions, ultimately leading to improved patient outcomes in the battle against heart disease.

2 RELATED PREVIOUS WORK

In the domain of cardiovascular disease prediction, several previous studies have leveraged machine learning (ML) techniques to enhance diagnostic accuracy.

2.1 Overview of Previous Studies

Krittanawong et al. (2020) conducted a comprehensive meta-analysis, involving 103 cohorts with over 3 million individuals, evaluating various ML algorithms for cardiovascular disease prediction. This study highlighted the promising predictive ability of ML algorithms, particularly Support Vector Machine (SVM) and boosting algorithms, in coronary artery disease and stroke prediction ("Machine learning prediction in cardiovascular diseases: A meta-analysis").

2.2 Specific Machine Learning Approaches

Srinivasan et al. (2023) focused on cardiovascular heart disease prediction using eight different ML classifiers, including neural network models like Naïve Bayes and Radial Basis Functions, achieving accuracies up to 94.78% and 90.78% respectively ("An active learning machine technique based prediction of cardiovascular heart disease from UCI-repository database"). Louridi et al. (2021) employed a diverse range of ML models, including AdaBoost, Gradient Boost, Random Forest, KNN, and SVM. These models, optimized using grid search and Bayesian optimization, achieved accuracies

Final Data Mining Report, November & December 2023, Nijmegen, Netherlands

and AUC scores up to 88.9% and 0.925 ("Machine learning-based identification of patients with a cardiovascular defect").

2.3 Challenges and Opportunities

Despite these advances, challenges remain in the early diagnosis and prediction of heart disorders using ML. Louridi et al. (2021) demonstrate varying degrees of accuracy using different ML approaches, underscoring the need for ongoing optimization and validation of these models ("Machine learning-based identification of patients with a cardiovascular defect").

3 DATASET DESCRIPTION

The dataset used in this study is sourced from Kaggle, specifically the "Heart Attack Analysis & Prediction Dataset" (available at https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset). It comprises various medical attributes that are potentially indicative of heart attack risk. The dataset includes several clinical parameters such as age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, and others. The dataset is structured with 303 entries, each representing an individual's medical profile.

3.1 Data Collection

The dataset was collected and made available on Kaggle, a platform for data science and machine learning. This dataset is commonly used in research and studies related to heart disease prediction and provides a comprehensive set of features for analysis.

4 METHODOLOGY

- **Feature Selection:** The study began with a detailed exploratory data analysis to identify the most influential features for heart disease prediction. This step is critical in understanding the underlying patterns and relationships within the data.
- **Model Training and Testing:** Both models were trained using the selected features from the dataset. The training process involved fitting the models to the training data and subsequently evaluating them on the test set. This phase is crucial for assessing the models' performance and their ability to generalize to new, unseen data.
- **Hyperparameter Tuning:** To enhance the models' performance, hyperparameter tuning was performed using GridSearchCV. This process involves searching through a specified parameter grid to find the combination of parameters that yields the best performance. This step is essential to optimize the models and improve their predictive accuracy.
- **Evaluation Metrics:** The models were evaluated based on accuracy, precision, recall, and F1 score. These metrics provide a comprehensive understanding of the models' performance, especially in a medical context where the cost of false negatives or false positives can be significant.

The chosen approach and methods aim to provide a balanced and thorough analysis of heart disease prediction, ensuring both accuracy and reliability in the models' predictions.

5 DATA PREPROCESSING

Data preprocessing involved several key steps to prepare the dataset for machine learning model training and testing:

- **Feature Selection:** Initial exploratory data analysis was performed to understand the correlations and relevance of different features. Key features such as 'oldpeak', 'thalachh', 'caa', 'cp', 'thall', 'age', 'exng', and 'slp' were identified as crucial for the study.
- **Data Splitting:** The dataset was split into training and testing sets, with 80% of the data used for training and the remaining 20% for testing. This split ensured a balanced approach to training and validation.
- **Feature Scaling:** A StandardScaler was applied to the dataset to normalize the feature values, ensuring that the model is not biased towards variables with higher magnitudes.
- **Feature Extraction:** Only the selected features were used for model training and testing, ensuring a focused approach towards the most relevant predictors of heart disease.

These preprocessing steps were crucial in preparing the dataset for effective and accurate modeling, thus ensuring reliable predictions from the machine learning algorithms used in this study.

6 APPROACH

The primary objective of this study is to predict heart disease using machine learning models. The approach involved selecting and comparing two distinct models: Random Forest Classifier and K-Nearest Neighbors (KNN). These models were chosen due to their proven efficacy in classification tasks, especially in medical diagnostics where interpreting complex patterns in data is crucial.

6.1 Rationale Behind Model Selection

- **Random Forest Classifier:** This model was selected for its robustness in handling overfitting. Random Forest, being an ensemble learning method, combines multiple decision trees to improve the model's generalizability and accuracy. The model's ability to handle large datasets with a higher dimensionality without significant performance degradation makes it suitable for complex datasets like those used in heart disease prediction.
- **K-Nearest Neighbors (KNN):** KNN was chosen for its simplicity and effectiveness in classification tasks. It's a non-parametric method that works well for smaller datasets and can be very effective when the dataset has a meaningful distance metric. This model's ability to make predictions based on the proximity of data points makes it a valuable tool for medical data analysis.

7 MAIN RESULTS AND INSIGHTS

The study's primary objective was to leverage machine learning techniques for effective heart disease prediction. The key findings are as follows:

- **Optimized Model Performance:** The Random Forest model, post-optimization, emerged as the most effective, with an accuracy of 86.9%, precision of 87.5%, and an F1 score of

87.5%. Its recall rate was equally impressive, indicating a high true positive rate.
- **Comparison with KNN Model:** The KNN model, though slightly less effective than the Random Forest, still showed commendable performance, underscoring the utility of diverse modeling approaches in predictive analytics.
- **Effective Hyperparameter Tuning:** The application of Grid Search and Random Search for hyperparameter tuning played a pivotal role in enhancing model performance, demonstrating the value of these techniques in machine learning workflows.

7.1 Insights Drawn

The project yielded several critical insights, essential for the field of medical data analysis:

- **Importance of Model Choice:** The variance in performance between the Random Forest and KNN models highlights the importance of selecting appropriate algorithms based on the nature of the data and the specific requirements of the predictive task.
- **Value of Optimization:** The improvement in model metrics post-optimization emphasizes the necessity of model tuning for achieving optimal performance, especially in complex tasks like disease prediction.
- **Utility of ROC Curve Analysis:** The use of ROC curves provided a clear visual representation of model performance, especially in distinguishing between true positive and false positive rates, a crucial factor in medical diagnoses.

7.2 Practical Implications

These results have practical implications in healthcare, particularly in enhancing diagnostic accuracy and supporting clinical decision-making. The high performance of the optimized models suggests a potential for these tools to be integrated into medical practice, aiding in early and accurate diagnosis of heart disease.

8 ANALYSIS OF RESULTS AND EVALUATION

The project undertook a comprehensive approach in evaluating the performance of various machine learning models for predicting heart disease. Key metrics such as accuracy, precision, recall, and the F1 score were employed to assess model efficacy. Notably, the optimized Random Forest model exhibited superior performance with an accuracy of 86.9%, precision of 87.5%, and an identical recall and F1 score of 87.5%. In comparison, the optimized K-Nearest Neighbors (KNN) model demonstrated slightly lower metrics, indicating a marginally reduced efficiency in prediction accuracy.

8.1 Hyperparameter Optimization

Crucial to the project's success was the rigorous process of hyperparameter tuning. Techniques such as Grid Search and Random Search were instrumental in refining the models. This optimization led to significant improvements in model performance. The careful tuning process underscores the importance of parameter optimization in achieving optimal model functionality.

8.2 Comparative Analysis of Models

The project presented a comparative analysis between the optimized Random Forest and KNN models. Through various evaluation metrics, it was evident that while both models performed admirably, the Random Forest model had a slight edge in predictive accuracy and reliability. This comparison is vital in understanding the strengths and limitations of each model in the context of heart disease prediction.

8.3 Receiver Operating Characteristic (ROC) Curve Analysis

An integral part of the evaluation was the ROC curve analysis. The ROC curves for both models provided a visual representation of their performance, particularly in terms of the trade-off between the true positive rate and the false positive rate. The area under the curve (AUC) for the Random Forest model was indicative of its superior performance compared to the KNN model, affirming its efficacy in the classification task.

9 POSSIBLE FUTURE DIRECTIONS

The findings from the current study on heart disease prediction using machine learning models offer several avenues for future research and development. These directions not only aim to enhance the existing models but also to explore new methodologies and applications in the realm of medical diagnostics.

9.1 Integration of Advanced Machine Learning Techniques

Future studies could explore the integration of more advanced machine learning techniques such as deep learning. These techniques have the potential to uncover complex patterns in data that traditional models may overlook, potentially leading to even more accurate predictions.

9.2 Exploration of Additional Data Sources

Incorporating a broader range of data sources, including patient medical history, lifestyle factors, and genetic information, could provide a more holistic view of the risk factors associated with heart disease. This approach could lead to more personalized and precise diagnostic tools.

9.3 Cross-Disease Application

The methodologies and insights gained from this study could be applied to other diseases as well. Future research could focus on adapting the models for the prediction of other chronic conditions, thereby broadening the scope of machine learning in healthcare.

9.4 Enhancement of Patient Engagement and Telemedicine

Another direction could be the development of patient-facing applications and telemedicine tools that use these models. Such applications could aid in early detection and continuous monitoring of heart disease, making healthcare more accessible and efficient.

9.5 Collaboration Across Disciplines

Future work could benefit from a multidisciplinary approach, involving collaboration between data scientists, healthcare professionals, and bioinformaticians. Such collaborations can lead to a more comprehensive understanding of the challenges and opportunities in applying machine learning to healthcare.

10 CONCLUSION

This study has successfully demonstrated the potential of machine learning, particularly the Random Forest and K-Nearest Neighbors (KNN) models, in predicting heart disease.

Through rigorous feature selection, model training, and hyperparameter tuning, this research has highlighted the crucial role these factors play in enhancing the accuracy and reliability of predictive models. The comparative analysis between the Random Forest and KNN models has provided valuable insights, emphasizing the importance of choosing the appropriate model for specific data characteristics.

In summary, this project not only underscores the capabilities of machine learning in healthcare analytics but also opens up exciting possibilities for future research in this evolving field.

BIBLIOGRAPHY

(1) Krittanawong, C., Virk, H. U. H., Bangalore, S., Wang, Z., Johnson, K. W., Pinotti, R., ... & Tang, W. H. W. (2020). Machine learning prediction in cardiovascular diseases: A meta-analysis. *Scientific Reports*, 10(1), 16057.

(2) Srinivasan, S., Gunasekaran, S., Mathivanan, S. K., Malar M. B., B. A., Jayagopal, P., ... & Dalu, G. T. (2023). An active learning machine technique based prediction of cardiovascular heart disease from UCI-repository database. *Scientific Reports*, 13(1), 13588.

(3) Louridi, N., Douzi, S., & El Ouahidi, B. (2021). Machine learning-based identification of patients with a cardiovascular defect. *Journal of Big Data*, 8(1), 133.

A APPENDIX

This appendix provides a detailed overview of the primary code segments used in this study for heart disease prediction using machine learning models. The code is developed in Python, using libraries such as Pandas, Scikit-learn, and Matplotlib for data processing, model training, and result visualization.

A.1 Data Preprocessing

```
heart_data = pd.read_csv('/path/to/heart.csv')
X = heart_data.drop('output', axis=1)
y = heart_data['output']
X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.2, random_state=42)
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

This segment covers the initial data loading, feature selection, and splitting the dataset into training and testing sets. StandardScaler is also used for feature scaling.

A.2 Feature Selection and Correlation Analysis

```
corr_matrix = heart_data.corr()
plt.figure(figsize=(15, 15))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation matrix')
plt.show()
```

Here, a correlation analysis is preformed visualized through a heatmap, helping in the selection of features most relevant to heart disease prediction.

A.3 Model Training and Evaluation

```
# Model training
rf_model = RandomForestClassifier(random_state=42)
knn_model = KNeighborsClassifier()
rf_model.fit(X_train_selected, y_train)
knn_model.fit(X_train_selected, y_train)

# Model evaluation
rf_predictions = rf_model.predict(X_test_selected)
knn_predictions = knn_model.predict(X_test_selected)
```

This code demonstrates the training and initial evaluation of the Random Forest and KNN models using accuracy, precision, recall, and F1 score metrics.

A.4 Hyperparameter Tuning

Hyperparameter tuning is a crucial step in optimizing machine learning models for enhanced performance. In this project, 'GridSearchCV' from Scikit-learn is employed for systematic tuning of hyperparameters for both the Random Forest and KNN models. The following code snippets and explanations detail this process.

A.5 Random Forest Hyperparameter Tuning

```
rf_param_grid = {
    'n_estimators': [50, 100, 200, 300, 400, 500],
    'max_depth': [None, 10, 20, 30, 40, 50],
    'min_samples_split': [2, 5, 10, 15, 20],
    'min_samples_leaf': [1, 2, 4, 6, 8]
}

rf_grid_search = GridSearchCV(estimator=rf_model,
param_grid=rf_param_grid, cv=3, n_jobs=-1, verbose=2)
rf_grid_search.fit(X_train_selected, y_train)
```

This segment performs exhaustive search over specified parameter values for the Random Forest model. The parameters include the number of trees in the forest (‘n_estimators‘), the maximum depth of the tree (‘max_depth‘), the minimum number of samples required to split an internal node (‘min_samples_split‘), and the minimum number of samples required to be at a leaf node (‘min_samples_leaf‘).

A.6 K-Nearest Neighbors Hyperparameter Tuning

```
knn_param_grid = {
    'n_neighbors': [3, 5, 7],
    'weights': ['uniform', 'distance'],
    'algorithm': ['ball_tree', 'kd_tree', 'brute']
}

knn_grid_search = GridSearchCV(estimator=knn_model,
param_grid=knn_param_grid, cv=3, n_jobs=-1, verbose=2)
knn_grid_search.fit(X_train_selected, y_train)
```

For the KNN model, the grid search explores various combinations of hyperparameters, including the number of neighbors (‘n_neighbors‘), the weight function used in prediction (‘weights‘), and the algorithm used to compute the nearest neighbors (‘algorithm‘). The aim is to identify the optimal settings that yield the best prediction results.

A.7 Evaluating the Best Hyperparameters

After the completion of the grid search, the best parameters are identified and used to retrain the models. This process ensures that the models are fine-tuned to the specific characteristics of the dataset, leading to improved predictive performance.

```
best_rf_params = rf_grid_search.best_params_
best_knn_params = knn_grid_search.best_params_

print("Best Random Forest Parameters:", best_rf_params)
print("Best KNN Parameters:", best_knn_params)
```

This step concludes the hyperparameter tuning process, where the best parameters are reported and subsequently used for model optimization.

A.8 Model Optimization and ROC Curve Analysis

```
fpr_rf, tpr_rf, _ = roc_curve(y_test,
rf_model_optimized.predict_proba(X_test_selected)[:, 1])
fpr_knn, tpr_knn, _ = roc_curve(y_test,
knn_model_optimized.predict_proba(X_test_selected)[:, 1])
plt.figure(figsize=(10, 8))
plt.plot(fpr_rf, tpr_rf, label='Random Forest')
plt.plot(fpr_knn, tpr_knn, label='KNN')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curves')
plt.legend(loc='lower right')
plt.show()
```

The final section includes the optimization of models using GridSearchCV and the ROC curve analysis for the optimized models, crucial for assessing the models’ diagnostic capabilities.

This appendix encapsulates the key components of the code utilized in this project, providing clarity on how the data is processed, models are built, evaluated, and optimized.