Comparative Analysis of Machine Learning Classification Approaches for Heart Disease Prediction: A Study of Preprocessing Techniques and Algorithms

Karlavi M. M.[1] and Chatrabgoun O.[1,2*]

[1]Department of Data Science and Computational Intelligence, School of Computing, Mathematics and Data Science, Faculty of Engineering, Coventry University, United Kingdom

[2] School of Computing, Mathematics and Data Science, Faculty of Engineering, Environment and Computing Coventry University, Gulson Road, Coventry, CV1 2JH, UK

**maharoofm@coventry.ac.uk*

The growing number of cardiovascular diseases is a big concern for health worldwide, showing the need for effective tools to help detect these diseases early and provide timely treatment. This research contributes to the growing body of knowledge in healthcare analytics by investigating the effectiveness of machine learning algorithms to enhance patient diagnosis and treatment strategies in cardiovascular care. We compare four different Machin Learning Algorithms as K-Nearest Neighbors (K-NN), Logistic Regression (LR), Support Vector Machine (SVM), and Naive Bayes (NB) using a comprehensive dataset of patient attributes. Our methodology includes preprocessing techniques such as one-hot encoding and label encoding to convert categorical variables for optimal model performance. Additionally in our study, we explored the influence of these preprocessing techniques, identifying that one-hot encoding generally enhanced accuracy for most algorithms. Hyperparameter tuning was conducted for SVM, optimizing parameters as the kernel type and regularization strength, which further improved the model's accuracy. The dataset was systematically split into 80% training and 20% testing subsets, allowing us to assess each algorithm's accuracy on the testing set. The results revealed that SVM outperformed the other algorithms, achieving an accuracy of 89.69%, highlighting the critical role of methodological choices in developing effective predictive models. What sets this research apart from recent studies is its comprehensive comparison of multiple algorithms alongside detailed data preprocessing techniques, providing insights into the impact of these choices on predictive performance.

Keywords: *Cardiovascular Disease, Data Preprocessing, Hyperparameter Tuning, ML Algorithm, One-hot Encoding.*