# Performance Analysis of Supervised Classification Models on Heart Disease Prediction

**Ezekiel Adebayo Ogundepo1; Waheed Babatunde Yahya2; Olubayo Adekanmbi1**

1Data Science Nigeria (DSN), Lagos, Nigeria

2Department of Statistics, University of Ilorin, Ilorin, Nigeria.

e-mail: ezekiel@datasciencenigeria.ai, ogundepoezekiel@gmail.com1; wbyahya@unilorin.edu.ng2; olubayo@datasciencenigeria.ai1

**Abstract**

This paper presents a detailed predictive analysis of heart disease conditions of some patients to determine the possible risk factors that are suggestive of whether a patient in the sample has heart disease or not. The predictive strengths of the identified risk factors were further determined using suitable classification methods. Two independent (but similar) published heart disease data, the Cleveland (training) and the Statlog (validation) datasets were considered in this paper. Detail exploratory analysis using the Chi-square test of independence was performed to get more insight about the features that are useful at determining whether a patient has heart disease or not before employing ten standard machine learning techniques for the class prediction. The results showed that some of the bio-clinical categorical variables are strongly associated with the heart disease conditions of the patients in the two data sets (p < 0.001). To contruct the classification models, the Cleveland data were randomly partitioned into 70% (208) training samples which were used to build the models and 30% (89) test samples that were used for evaluating their predictive performances over 200 replications. These results were validated on the Statlog data in 10-fold cross-validation. The classification results from the test samples indicated that the Support Vector Machine (SVM) yielded the best predictive performances with 85% Accuracy, 82% Sensitivity, 87% Specificity, 87% Precision, 91% Area under the ROC curve (AUC), and 38.5% LogLoss value. These results are all consistent with those yielded by the SVM classifier on an independent validation (Statlog) dataset which is a clear indication that the results of this study can be easily reproduced on similar independent heart disease datasets.

**Keywords:** Classifiers, Model Selection, Feature Selection, Exploratory Data Analysis, Evaluation metrics.