



**ROBERT GORDON
UNIVERSITY ABERDEEN**

SCHOOL OF COMPUTING SCIENCE AND DIGITAL MEDIA

MSc Business Analytics

Student Name:Amani Prathiba Subawickrama

Matriculation Number:

Supervisor: Dr. Alroy Mascrenge

Second Marker:

Project Title: Predicting Heart Attack with Imbalanced Dataset Using Machine Learning Techniques

Start Date:2023.01.15

Submission Date:2023.08.20

CONSENT

I agree

I do not agree

That the University shall be entitled to use any results, materials or other outcomes arising from my project work for the purposes of non-commercial teaching and research, including collaboration.

DECLARATION

I confirm:

- **That the work contained in this document has been composed solely by myself and that I have not made use of any unauthorised assistance.**
- **That the work has not been accepted in any previous application for a degree.**
- **All sources of information have been specifically acknowledged and all verbatim extracts are distinguished by quotation marks.**

Student Signature:

Date Signed:17.08.2023

ABSTRACT

Predicting Heart Attack with Imbalanced Dataset Using Machine Learning Techniques

Heart attacks, also known as myocardial infarctions, pose a significant health risk and require accurate prediction methods to enable timely intervention and prevention. This paper presents a comprehensive study on the application of machine learning techniques for heart attack prediction. This research focuses on the development, evaluation, and validation of a predictive model capable of accurately classifying individuals at risk of experiencing a heart attack based on their medical profiles.

Objective: The primary aim of this study is to identify the most suitable machine learning technique for predicting heart attacks, integrating various feature selection methods, and mitigating the impact of dataset imbalance, thereby enhancing the accuracy of predictive models.

Methods: In this study, analyzing heart attack classification done by collecting a substantial dataset pertaining to heart attack prediction, while selecting 10 independent variables crucial to the classification process. Comprehensively evaluate the predictive performance of eight prominent machine learning classifiers, namely J48, Random Tree, Random Forest, Naïve Bayes, REP Tree, k-NN, SVM, and Multilayer Perceptron. To enhance the classification accuracy, incorporate a range of feature selection methods, including Information Gain, Gain Ratio, Chi-Square, OneR, Wrapper, and CfsSubsetEval with Particle Swarm Optimization. Moreover, to effectively handle the class imbalance by employing a hybrid sampling approach, integrating Synthetic Minority Over-sampling Technique (SMOTE) and SpreadSubsample.

Results: In this study, the performance of various ML classification techniques was evaluated on an original dataset with class imbalance. Surprisingly, six classifiers, J48, Random Forest, REP Tree, Naïve Bayes, Neural Network, and SVM achieved 100% accuracy, precision, recall, F-measure, and AUC. Employing CfsSubsetEval feature selection with particle swarm optimization further refined the models, reducing the training time of neural network from 2022.75s to 330.41s while maintaining perfect accuracy. Additionally, after addressing class imbalance, the accuracy of k-nearest neighbor increased to 100%, and the overall training and processing times improved. Furthermore, utilizing CfsSubsetEval feature selection with particle swarm optimization on a balanced dataset enhanced accuracy for most techniques to 100%, with the exception of Naïve Bayes achieving 98.8877%.

Conclusion: The study highlights the strong performance of different machine learning techniques for heart attack classification. Hence, picking the best method depends on practical factors such as available resources and ease of use, as each technique has its own strengths. A combined or further testing approach could help choose the right method for real-life situations.

Keywords: heart attack prediction, machine learning, feature selection, class imbalance