GC-258

Exploring Feature selection and classification methods for predicting heart disease

Abstract

Research conducted by many health organizations and hospitals such as the European Health Journal has shown that the early detection of Heart Disease is critical to treating and understanding the causes. On top of this urgency, Harvard Health states that" cardiovascular disease is the most expensive chronic disease for America's health care system". Through the use of advanced machine learning models and comprehensive data sets collected on patients of varying backgrounds and health statuses, this research shows the listed correlations between attributes of patients and positive identification of heart disease. This research paper analyzes 1026 unique records and 13 attributes plus a classifier to examine machine learning techniques and look for correlation between them to assist with identifying positive cases.

Introduction

The current test for Heart Disease involves a number of invasive and expensive tests that can prove prohibitive to an individual. What if there was a way to reliably predict Heart Disease based on other known factors that are less invasive and cheaper as well as faster to collect. For example, a simple blood test combined with other factors such as information about chest pain as well as resting blood pressure. Using a smaller set of data points, can a given algorithm predict a positive diagnosis of Heart Disease? Also, which type of Machine Learning model is best for providing an accurate diagnosis?

Research Question(s)

1. How can Data Science improve medical diagnosis or provide educated information?
2. What algorithms provide a valuable classification of data?
3. How can ensemble methods improve results?
4. Can reliable results be obtained with less data?

Materials and Methods

All test were performed on the WEKA Machine Learning Platform. WEKA is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization [6]. For the purposes of preparation the 13 testing attributes were converted from any format into numeric format to ensure they are assessed as integers, the target value or classifier was converted into a nominal value to avoid either binary value having a weight greater or worse than the other.

Classifier/Test Option	5 Folds	10 Folds	75%	80%
Logistic	84.88%	85.37%	84.38%	88.29%
Stochastic Gradient Descent	84.68%	84.88%	82.42%	85.85%
Simple Logistic	85.56%	85.17%	84.77%	88.29%
Sequential Minimal Optimization	84.78%	84.88%	85.55%	88.29%
Naive Bayes	83.32%	83.41%	83.20%	85.85%

Results

Classifier/Test Option	5 Folds	10 Folds	75%	80%
Logistic	84.98%	84.68%	85.94%	88.29%
Stochastic Gradient Descent	85.66%	84.49%	86.72%	88.78%
Simple Logistic	85.17%	84.98%	85.16%	88.78%
Sequential Minimal Optimization	84.98%	84.59%	87.11%	89.76%
Naive Bayes	83.22%	83.61%	83.20%	85.85%

Classifier/Test Option	5 Folds	10 Folds	75%	80%
Logistic	84.88%	85.37%	84.38%	88.29%
Stochastic Gradient Descent	85.27%	83.51%	85.16%	84.88%
Simple Logistic	85.56%	85.07%	84.77%	88.29%
Sequential Minimal Optimization	84.68%	84.98%	85.55%	88.29%
Naive Bayes	83.80%	85.17%	86.33%	84.39%

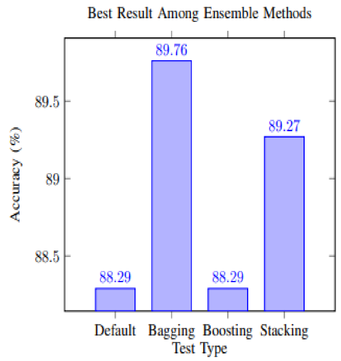
Classifier/Meta Classifier/Test Option	80%
Logistic/SGD	86.83%
Logistic/Simple Logistic	88.29%
Logistic/SMO	88.29%
Logistic/Naive Bayes	88.29%
Logistic/J48	88.29%
SMO/Logistic	88.29%
SMO/SGD Text	56.10%
SMO/Simple Logistic/J48	89.27%
SMO/Simple Logistic/Naive Bayes	88.29%
Naive Bayes/J48	86.83%

Classifier/Test Option	80% split	Before Removal
Logistic	88.29%	88.29
SGD	85.85%	85.85
Simple Logistic	88.29%	88.29
SMO	88.29%	88.29
NaiveBayes	85.85%	85.85

Classifier/Test Option	80% split	Before Removal
Logistic	87.32%	88.29%
SGD	86.83%	88.85%
Simple Logistic	88.78%	88.29%
SMO	88.29%	88.29%
Naive Bayes	84.39%	85.85%

Conclusions

The primary purpose of this study is to provide a tool to detect cardiac problems at an early stage and machine learning techniques have shown promising results in heart disease detection. To conclude the clear takeaway from these experiences was that the ensemble methods of classifying the data proved to be great improvements on the overall test structure. Bagging and stacking in particular offer a lot of opportunity to push the accuracy of these predictive models well into the 90% range.



Acknowledgments

Special thanks to Dr. Jack Zheng and Dr. Seyedamin Pouriyeh for their expert advice and support through this research.

Contact Information

Author: Devin Jackson, djack271@students.Kennesaw.edu
Author: Richard Stupka, rstupka@students.Kennesaw.edu
Author: Trinadh Chigurupati, tchiguru@students.Kennesaw.edu
Author: Demontae Moore, dmoor195@students.Kennesaw.edu
Professor: Jack Zheng, gzheng@kennesaw.edu
Project Sponsor: Seyedamin Pouriyeh, spouriye@kennesaw.edu

References

J. C. Mason and P. Libby, "Cardiovascular disease in patients with chronic inflammation: mechanisms underlying premature cardiovascular events in rheumatologic conditions," European Heart Journal, vol. 36, no. 8, pp. 482–489, 11 2014. [Online]. Available: https://doi.org/10.1093/eurheartj/ehu403
[2] J. Corliss, "Decoding the price of heart tests and procedures," Nov 2022. [Online]. Available: https://www.health.harvard.edu/hearthealth/decoding-the-price-of-heart-tests-and-procedures
[3] S. Pouriyeh, s. vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease," 07 2017.
[4] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in 2008 IEEE/ACS International Conference on Computer Systems and Applications, 2008, pp. 108–115.
[5] D. Lapp, "Heart disease dataset," Jun 2019. [Online]. Available: https://www.kaggle.com/datasets/johnsmith88/heart-diseasedataset?resource=download
[6] E. Frank, M. A. Hall, and I. H. Witten, "Weka 3: Machine learning software in java." [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/
[7] J. Starkweather and A. K. Moske, "University of north texas." [Online]. Available: https://it.unt.edu/sites/default/files/mlrjdsaug2011.pdf
[8] M. Stojiljkovic, "Stochastic gradient descent algorithm with python and numpy," Sep 2022. [Online]. Available: https://realpython.com/gradientdescent-algorithm-python
[9] R. Sidhu, "Layman's introduction to linear regression — by rishi sidhu — towards ..." Jun 2019. [Online]. Available: https://towardsdatascience.com/laymans-introduction-tolinear-regression-8b334a3dab09
[10] A. Pentrakan, C.-C. Yang, and W.-K. Wong, "How well does a sequential minimal optimization model perform in predicting medicine prices for procurement system?" May 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8196718/

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING

Author: Devin Jackson, Richard Stupka, Trinadh Chigurupati, Demontae Moore
Advisors(s): Dr. Jack Zheng, Dr. Seyedamin Pouriyeh