GR-130 Prediction of Heart Disease with Machine Learning Techniques

Abstract

Heart disease consistently ranks as one of the leading causes of deaths globally. This project utilizes a newly merged heart disease dataset with 918 unique instances and 12 attributes, comprised of five globally recognized datasets from UCI Machine Learning Repository. Various machine learning techniques such as OneR, Decision Tree J48, K-Nearest Neighbors, Support Vector Machines, Logistic Regression, and ensemble methods are applied on Waikato Environment for Knowledge Analysis (Weka) open-source ML tool and python sickit-learn with 10-fold cross validation to the dataset to evaluate the likelihood of a patient having heart disease. Overall, python sickit-learn and Weka results were similar, and ensemble methods have better accuracy than simple classifiers. The Weka results were slightly higher in all methods.

Introduction

Early detection of heart disease is important so treatment can begin. Frequently, there are no observable symptoms, and a heart attack or stroke is the first indication of a problem. Therefore, the ability of machine learning techniques to leverage key factors to accurately predict heart disease can be critical to saving lives.

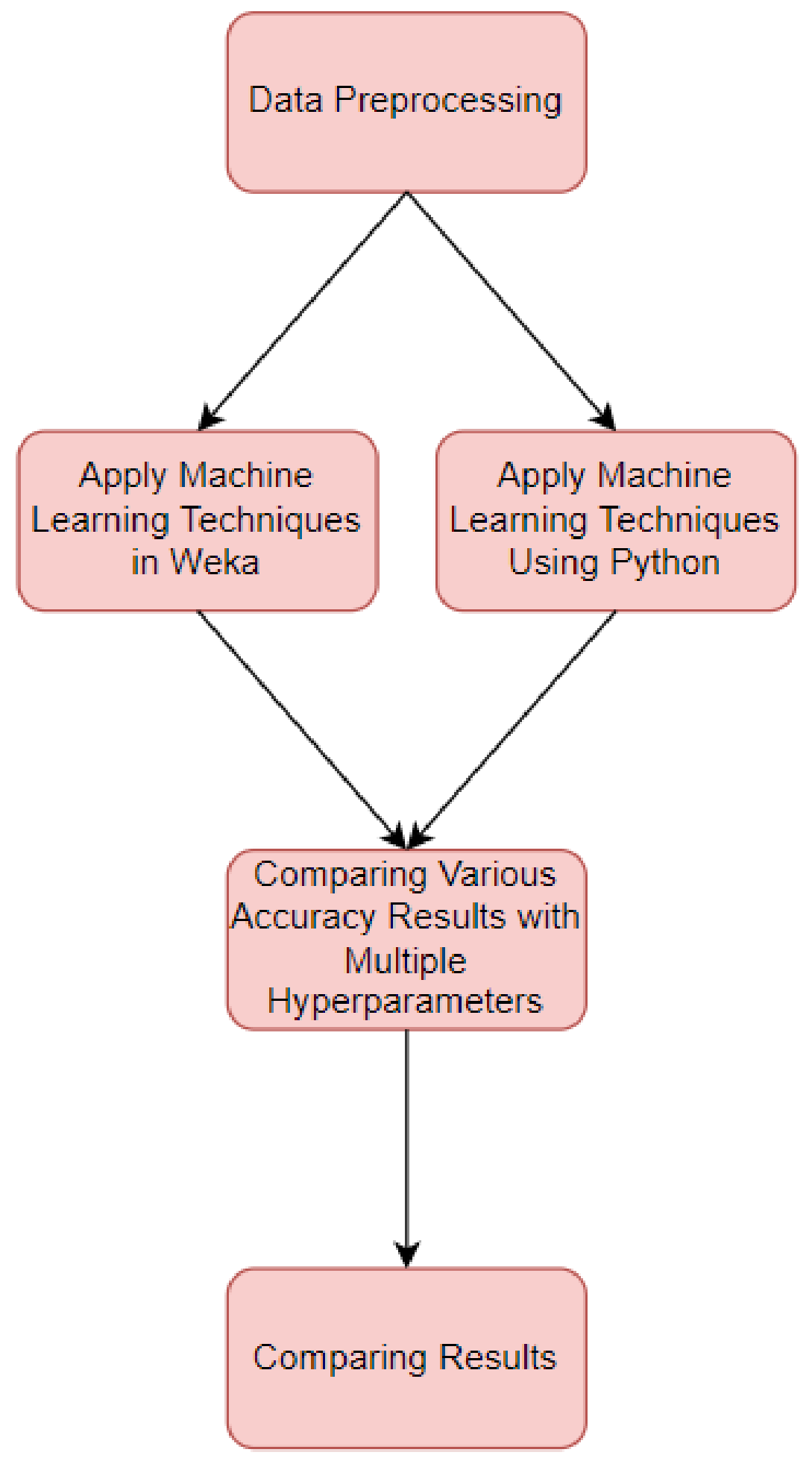
In this project we will use Weka open-source machine learning tool and python scikit-learn package to analyze a heart disease prediction dataset and evaluate the results obtained after preparing the dataset and applying different machine learning techniques with 10-fold cross validation.

These advanced methods are applied in accessible open-source Weka environment that can be used worldwide by doctors as an assisting tool to predict heart disease and save lives.

Research Questions

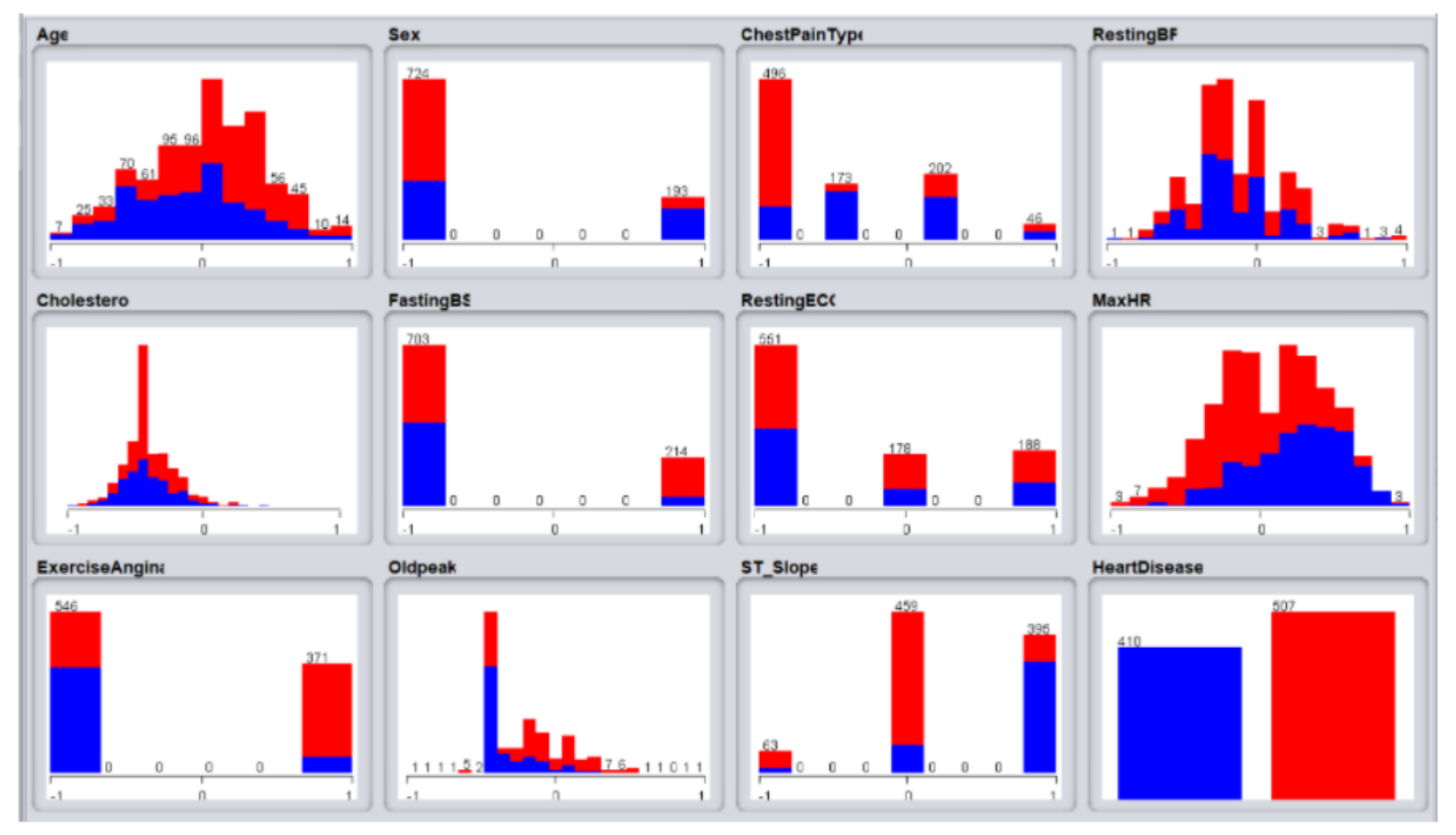
1. What is the role of machine learning in predicting heart disease?
2. Is there a difference in machine learning models when completed through Weka versus completed using python sickit-learn?
3. What machine learning model predicts heart disease with the highest accuracy?

Materials and Methods


Data Preprocessing
Apply Machine Learning Techniques in Weka
Apply Machine Learning Techniques Using Python
Comparing Various Accuracy Results with Multiple Hyperparameters
Comparing Results

Heart Failure Prediction Dataset: Cleveland, Hungarian, Switzerland, Long Beach VA, and Stalog (Heart)

Fig 1. Data after preprocessing



Results

Weka Results

Results achieved upon applying ML techniques with seed 1 suggest that except for OneR, all methods have shown similar accuracy, with stacking taking the lead. The stacking method performed a KNN using 19 nearest neighbors and Euclidean distance, then applied the Logit Boost which performs additive logistic regression, and then the vote classifier. These results then feed into the IBK (KNN) model using 19 and Euclidean distance to get the final results.

Accuracy of Machine Learning Techniques in Weka

Machine Learning Method	Accuracy %
OneR	81.35
Logistic Regression	84.27
Support Vector Machines	85.01
Decision Tree - J48 (pruned, n=2)	85.06
Random Forest	85.23
K-Nearest Neighbors (Manhattan, k=11)	85.28
Adaptive Boosting	85.28
Bagging (i=10, RF)	86.37
Stacking (KNN, LR, KNN)	87.24

Python sickit-learn Results

Results achieved upon applying ML in python sickit-learn, show that the bagging method with 100 iterations and Random Forest classifier with a base learner demonstrated the best accuracy. During the analysis, various parameters have been tested aiming get the best possible results.

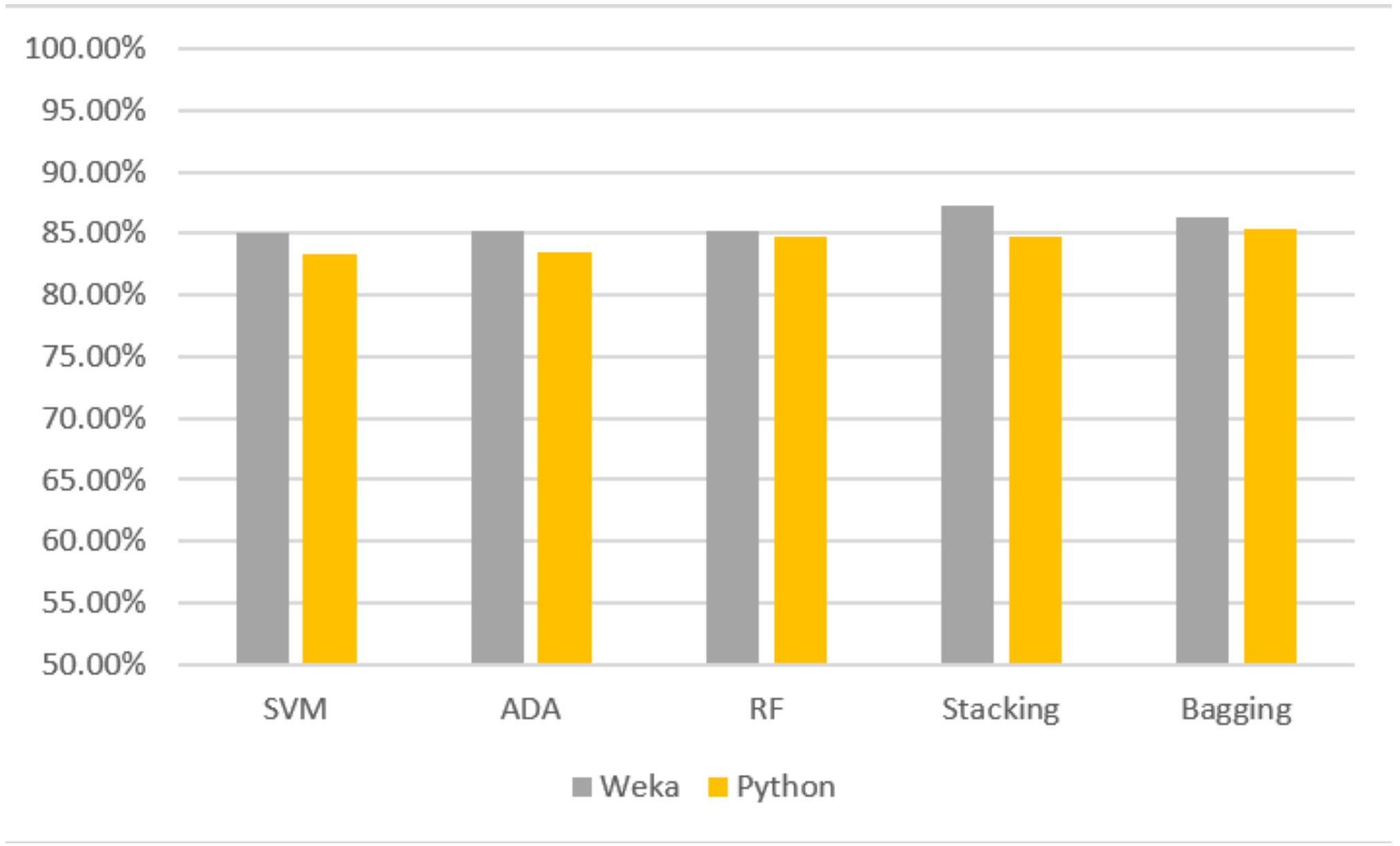
Accuracy of Top 5 Machine Learning Techniques in Python scikit-learn

Machine Learning Method	Accuracy %
Support Vector Machine	83.40
Adaptive Boosting	83.52
Random Forest	84.71
Stacking (RF, DT, LR)	84.82
Bagging (i=100, RF)	85.36

Comparison

Overall, python sickit-learn and Weka results were similar. The Weka results were slightly higher in all methods.

Fig 2. Comparison between Weka and python sickit-learn accuracy results



Conclusions

The ability to better predict cardiovascular disease for a patient can save the cost of unnecessary expensive medical tests and save lives by starting treatment early before more serious events such as a heart attack or stroke occur.

The results of this study show that python sickit-learn and Weka results were somewhat similar. Ensemble methods stacking and bagging have better accuracy than weak classifiers alone.

The stacking (KNN, LR, KNN) machine learning technique performed the best in Weka with 87.24% accuracy, followed by bagging (i=10, RF) at 86.37%. In python sickit-learn, the best method was bagging (i=100, RF) method with 85.36% accuracy, followed by stacking (RF, DT, LR) at 84.82%.

The intellectual merit of this project expands upon a previously published work. The previously published work examines weighted attributes while this project seeks to look at all the attributes presented within this dataset. By examining the dataset this way, we can ensure that no given attribute is given too much weight. For this reason, this paper can advance the understanding of how all these common risk factors can determine if someone has heart disease or not.

The most significant action we can undertake to further the potential of our research and have the most beneficial impact on society is to promote our research using Weka to the global medical community. Open-source machine learning tool with graphical user interfaces provide easy accessibility and usability to users. Doctors can add new instances to existing datasets to predict a cardiovascular disease for a patient.

Contact Information

Marcella Araujo - maraujo9@students.kennesaw.edu
Lauren Pope - lpope19@students.kennesaw.edu
Stephen Still - sstill10@students.kennesaw.edu
Cynthia Yannone - cyannone@students.kennesaw.edu

References

Fedesoriano. (2021) Heart failure prediction dataset. [Online]. Available: https://www.kaggle.com/fedesoriano/heart-failure-prediction

M. D. Ritchey, H. K. Wall, M. G. George, and J. S. Wright, “US trends in premature heart disease mortality over the past 50 years: Where do we go from here? ”Trends in Cardiovascular Medicine, vol. 30, no. 6, pp.364–374, 2020.

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,” in 22nd IEEE Symposium on Computers and Communication (ISCC), 2017.

P. Ghosh, S. Azam, M. Jonkman, A. Karim, F. M. J. M. Shamrat, E. Ignatious, S. Shultana, A. R. Beeravolu, and F. De Boer, “Efficient prediction of cardiovascular disease using machine learning algorithms with Relief and LASSO feature selection techniques,” IEEE Access, vol. 9,pp. 19304–19326, 2021.

Scan for our website

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING

Authors: Marcella Araujo, Lauren Pope, Stephen Still, Cynthia Yannone
Advisor: Dr. Seyedamin Pouriyeh