Heart Disease Prediction Using Machine Learning

UNIVERSITY OF HOUSTON

Charles Tian, William Le, Dr. Nouhad Rizk

Department of Computer Science, University of Houston

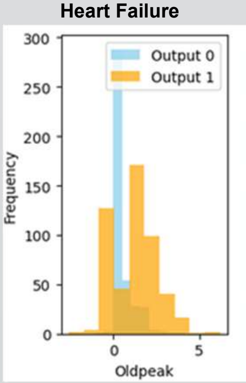
Abstract

Heart attacks and heart failure rank among the most common heart diseases in the general population. This study aims to employ machine learning to predict the likelihood of an individual developing these diseases. The results, presented as binary, can empower individuals to make informed decisions about their health. To achieve this goal, this study will combine the 'Heart Failure At first, our goal was to combine the 'Heart Failure Prediction Dataset' with the 'Heart Attack Analysis & Prediction Dataset. However, due to substantial differences in distinct features, we opted to focus exclusively on the heart failure dataset. The expectation is that our model's predictions will provide patients and users with valuable insights into their heart health, enabling them to take proactive measures. Given the high mortality rate associated with heart disease and heart failure in the absence of protective measures, this research endeavor seeks to transform unpredictable events into manageable ones.

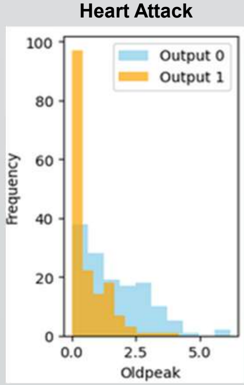
Background

- Heart attack occurs when blood flow to a part of the heart muscle is blocked, usually by a blood clot. This blockage can deprive the heart muscle of oxygen and cause damage or death to that part of the heart.
- Heart failure is a chronic condition where the heart muscle does not pump blood as well as it should

Heart Failure

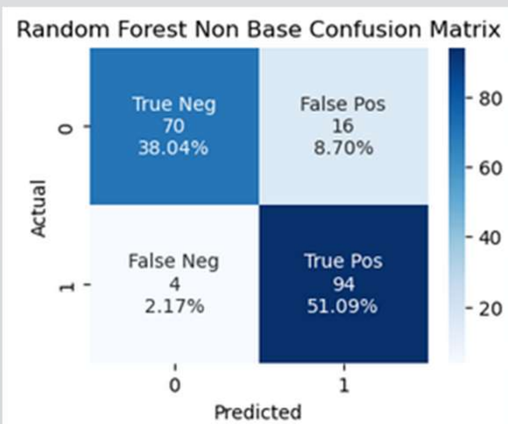


Heart Attack



Methods

- Basic Support Vector Classification model
- Basic Linear Regression model
- Basic Random Forest Classifier
- Random Forest using Grid Search for hyper parameter optimization and using trial and error to find the best model
- Basic Multi-Layer Perceptron
- Multi-Layer Perceptron using 3 layers and Logistic activation
- Multi-Layer Perceptron with Keras Sequential models: 20% Dropout rate, tuning of hyperparameter for optimization
- Multi-Layer Perceptron with Keras Functional models: 20% Dropout rate, Different architecture test and tuning of hyper parameter
- Feature Modification for Complexity Analysis: Removed specific features such as Sex, FastingBS , RestingECG, RestingBP from the original dataset. This simplification aims to evaluate the models' performance with reduced feature dimension.



Results

accuracy	recall_pos	model
0.8750	0.9388	SVM
0.8859	0.9388	LR
0.8587	0.9286	RF base
0.8913	0.9592	RF not Base
0.8478	0.8571	MLP base
0.8315	0.8980	MLP with 3 layers
0.8478	0.9184	Seq 2 Layers Relu BC
0.8370	0.9082	Fn 2 Layers Relu BC
0.8533	0.9286	Fn 3 Layers Relu MSE
0.8750	0.9388	Fn 2 Layers Sigmoid MSE

accuracy	recall_pos	model
0.8641	0.9184	SVM
0.8859	0.9388	LR
0.8696	0.9286	RF not Base
0.8696	0.9286	RF not Base
0.8696	0.9592	RF2 not Base
0.8587	0.9184	MLP base
0.8587	0.9082	MLP with 3 layers
0.8478	0.9286	Seq 2 Layers Relu BC
0.8587	0.9388	Fn 2 Layers Relu BC
0.8533	0.8980	Fn 3 Layers Relu MSE
0.8587	0.9184	Fn 4 Layers Relu MSE
0.8750	0.9184	Fn 2 Layers Sig MSE

Conclusion

- Random Forest had the best accuracy and a high recall of 96% for true positive
- MLP being a more flexible models and good at finding pattern in complex dataset was not able to match to Random Forest

Future Direction

- Expanding the dataset is essential for enhancing accuracy. Main issue encountered during the training of models such as MLP and Random Forest was the rapid overfitting.
- The mean squared error yielded a significantly lower loss compared to binary cross-entropy, suggesting that greater emphasis should be placed on the former.

Acknowledgments

Grateful for the impactful work of Umarani Nagavelli, Debabrata Samanta, and Partha Chakraborty in machine learning-based heart disease detection models.
Special thanks to Federico Soriano for the heart failure prediction dataset on Kaggle and Rashik for the heart attack analysis dataset.