HEART FAILURE PREDICTION AND ANALYTICS

Team - *Data Driven* | Advanced Data Science | INFO 7390 | Summer 2023

Objective

Heart failure, a debilitating condition where the heart's ability to pump blood efficiently is compromised, has a substantial global impact. As heart failure impairs individuals' quality of life, places immense strain on healthcare systems, and impacts economies through lost productivity and medical costs, it underscores the urgency for continued research, prevention strategies, and innovative treatments to address its global burden effectively. We are employing supervised machine learning algorithms to predict the accurate results with relevance to one's heart health. We essentially aim to save lives through correct predictions of heart conditions.

Target Variable

Targeting 'HeartDisease' as the variable is justified due to its substantial healthcare impact, clinical relevance, preventive potential, quality data, ML challenge, and public health value.

Machine Learning Algorithms

Machine learning algorithms play a crucial role in heart disease prediction by leveraging patterns and insights from medical data to provide accurate assessments of an individual's risk. The algorithms that we have employed for our project is as follows:

1. Decision Tree
2. Random Forest
3. Multilayer Perceptron Classifier
4. Support Vector Machine
5. Extra Tree
6. Logistic Regression
7. K-Nearest Neighbors
8. Gradient Boosting Classifier
9. Stacking Classifier

Data Set Description

Heart failure is a common event caused by CVDs and this dataset contains 11 features that can be used to predict a possible heart disease.

1. Age: age of the patient [years]
2. Sex: sex of the patient [M: Male, F: Female]
3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. RestingBP: resting blood pressure [mm Hg]

5. Cholesterol: serum cholesterol [mm/dl]
6. FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10. Oldpeak: oldpeak = ST [Numeric value measured in depression]
11. ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. HeartDisease: output class [1: heart disease, 0: Normal]

Additional Analytics & Conclusion

From the employed list of algorithms, *Random Forest* and additionally *Stacking Classifier* showed the top performance with the following performance metrics,

Random Forest – Train Accuracy: 92.52% | Test Accuracy: 85.814% | f1 Score: 86.64%

Stacking Classifier – Train Accuracy: 93.45% | Test Accuracy: 84.42% | f1 Score: 85.71%

	Model	Train_accuracy	Test_accuracy	precision_score	recall_score	f1_score	Description
0	Decision Tree	0.897196	0.811594	0.838926	0.816993	0.827815	
1	Random Forest	0.925234	0.851449	0.863636	0.869281	0.866450	The first best
2	Neural Network	0.894081	0.858696	0.870130	0.875817	0.872964	
3	SVC	0.898754	0.855072	0.864516	0.875817	0.870130	
4	Extra Tree	0.900312	0.847826	0.858065	0.869281	0.863636	
5	GradientBoosting	0.914330	0.851449	0.873333	0.856209	0.864686	
6	Logistic Regression	0.853583	0.815217	0.844595	0.816993	0.830565	
7	KNN	0.878505	0.829710	0.848684	0.843137	0.845902	
8	Stacking	0.934579	0.844203	0.871622	0.843137	0.857143	The second best

Scope of Future Work

Although the Random Forest and Stacking Classifier exhibited the top performance yet, we believe that by integrating ECG analytics we could hike the predictions for obtaining higher accuracies, further resulting in decisiveness in saving lives on humanitarian grounds.