Heart Analysis and Heart Failure Prediction

Advances in Data Science/ Architecture [INFO 7390]
Professor Ram Kumar Hariharan

Final Project Team 23:
Akanksha Telagam Setty [002131614]
Monisha Gali[002193887]
Pramod Gopal[001586157]

Background

- Heart diseases are the leading cause of death globally. Each year, over a million deaths occur worldwide. One third of these deaths occur below the age of 70. A lot of effort is provided by researchers all over the world to provide prevention, help, relieve, and hopefully one day cure heart diseases.
- People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.
- Adding on, ECG is widely used by cardiologists and medical practitioners for monitoring the cardiac health. The main problem with manual analysis of ECG signals, similar to many other time-series data, lies in difficulty of detecting and categorizing different waveforms and morphologies in the signal. For a human, this task is both extensively time-consuming and prone to errors.

Introduction

This project is an analytical study to predict the risk of having a heart failure. Various data analysis techniques have been used to observe trends between various risk factors for heart diseases. Based on the features, different machine learning models were then implemented to predict whether a person has heart disease.

We also study the ECG of patients and classify them into 5 different categories. To address the problems raised with the manual analysis of ECG signals, many studies in the literature explored using machine learning techniques to accurately detect the anomalies in the signal. We have implemented 3 different implementations for classifying the ECG signals.

Modules:

- Heart Failure Prediction
- ECG HeartBeat Classification

Heart Failure Prediction

Dataset Features

- **Age**: age of the patient [years]
- **Sex**: sex of the patient [M: Male, F: Female]
- **ChestPainType**: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
- **RestingBP**: resting blood pressure [mm Hg]
- **Cholesterol**: serum cholesterol [mm/dl]
- **FastingBS**: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
- **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]
- **MaxHR**: maximum heart rate achieved [Numeric value between 60 and 202]
- **ExerciseAngina**: exercise-induced angina [Y: Yes, N: No]
- **Oldpeak**: oldpeak = ST [Numeric value measured in depression]
- **ST_Slope**: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
- **HeartDisease**: output class [1: heart disease, 0: Normal]

Algorithms Used

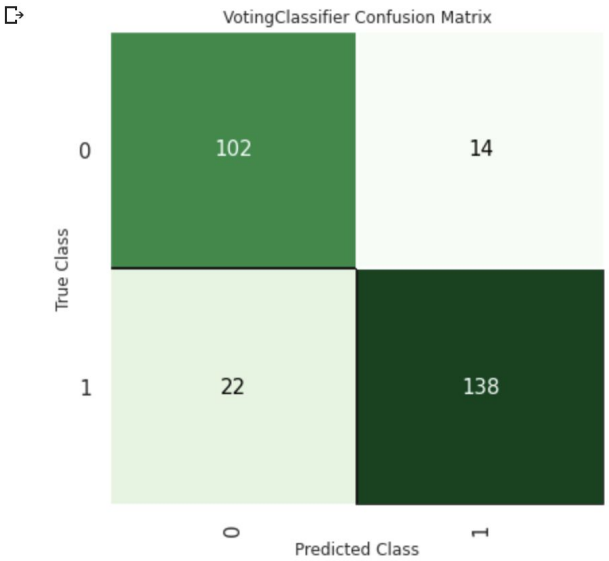
- Models implemented:
  - Extra trees classifier
  - Gradient Boosting Classifier
  - Random Forest Classifier
  - CatBoost Classifier
  - Light Gradient Boosting Machine
  - Decision Tree Classifier
- Best performing model : Ensemble of the models implemented (except Decision tree) with Soft Voting
- Final model : Calibrated version of the best model using pycaret

Comparative Performance

	Model	Accuracy	AUC	Recall	Prec.	F1
catboost	CatBoost Classifier	0.8444	0.8974	0.8949	0.8533	0.8712
et	Extra Trees Classifier	0.8113	0.8912	0.8705	0.8239	0.8442
nb	Naive Bayes	0.8305	0.9037	0.8635	0.8621	0.8576
rf	Random Forest Classifier	0.8022	0.9014	0.8462	0.8271	0.8352
lightgbm	Light Gradient Boosting Machine	0.8024	0.8661	0.8308	0.8394	0.8318
gbc	Gradient Boosting Classifier	0.7965	0.8667	0.8282	0.8334	0.8271
dt	Decision Tree Classifier	0.7459	0.7372	0.7814	0.7934	0.7833

Confusion Matrix

Ensemble model



Ensemble model with Soft Voting

ECG Classification

Algorithms:

- Logistic Regression
- Random Forest
- XgBoost

Performed Minority Sampling using SMOTE technique as data had imbalance

Best performing model - Random Forest

HeartBeat Classification Results

Algorithm	Without SMOTE (Accuracy)	With SMOTE(Accuracy)
Logistic Regression	91.29	66.2
Random Forest Classifier	97.60	98.2
XgBoost Classifier	84.44	75.65

Thank You