Project proposal

Domain background

The World Health Organization (WHO) has reported that Cardiovascular diseases (CVDs) stand as the foremost global cause of mortality. In 2019, an estimated 17.9 million individuals succumbed to CVDs, accounting for 32% of all worldwide fatalities. Of these fatalities, 85% were attributed to heart attacks and strokes. Among the 17 million untimely deaths (occurring before the age of 70) resulting from noncommunicable diseases in 2019, 38% were directly linked to CVDs. [1]

Heart failure is a common occurrence stemming from CVDs. The utilization of Machine Learning models to predict heart diseases has the potential to significantly mitigate Cardiovascular risk. [2]

Problem statement

The data is structured in a tabular format, and the problem is a classification task, specifically a binary classification problem with two classes: 0 (“Normal”) and 1 (“Heart disease”). With a binary classification, there are many approaches: Logistic Regression, Random Forest, LightGBM, ...

Solution statement

In addressing this problem, I will employ a variety of algorithms to determine the optimal solution. The key stages I will encompass include data preparation, exploratory data analysis (EDA), feature engineering, model training, evaluation, and ultimately, model deployment.

- Platform:

I intend to utilize AWS SageMaker for various stages of this project, including data preparation, EDA, feature engineering, and model-related tasks, such as training, hyperparameter tuning and evaluation. Subsequently, I plan to deploy the model using two distinct methods. The first entails leveraging a AWS SageMaker endpoint and AWS Lambda function, while the second involves using the best model's weights and/or

hyperparameters for deployment on the FastAPI framework locally.

- Algorithm:

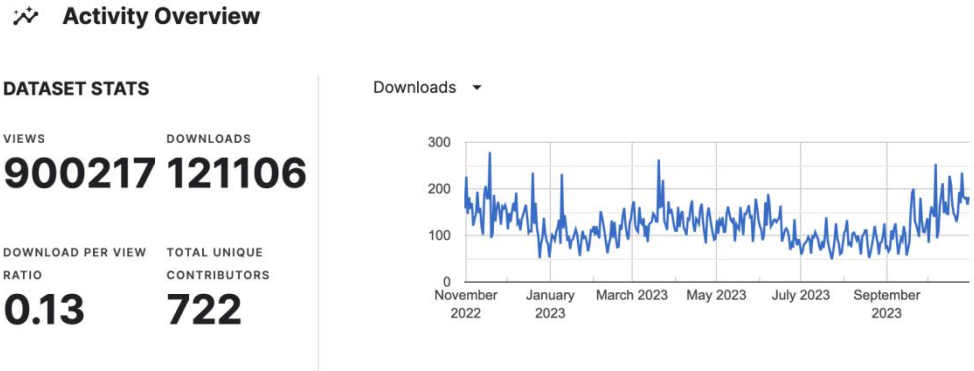
I plan to explore various algorithms, including Logistic Regression, Decision Tree Classifier, Extra Trees Classifier, LightGBM, ... and Ensemble methods. I will experiment with hyperparameter tuning using Optuna and leverage automation libraries such as Pycaret and AutoGluon. The primary metric for optimization is the F1 score.

Datasets and inputs

The dataset is sourced from Kaggle and it was created by combining different datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over 11 common features which makes it the largest heart disease dataset available so far for research purposes. The five datasets used for its curation are:

- Cleveland: 303 observations
- Hungarian: 294 observations
- Switzerland: 123 observations
- Long Beach VA: 200 observations
- Stalog (Heart) Data Set: 270 observations

Total: 1190 observations.
Duplicated: 272 observations.
Final dataset: **918 observations**.



First five rows of the data:

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0

Feature information:

- **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]

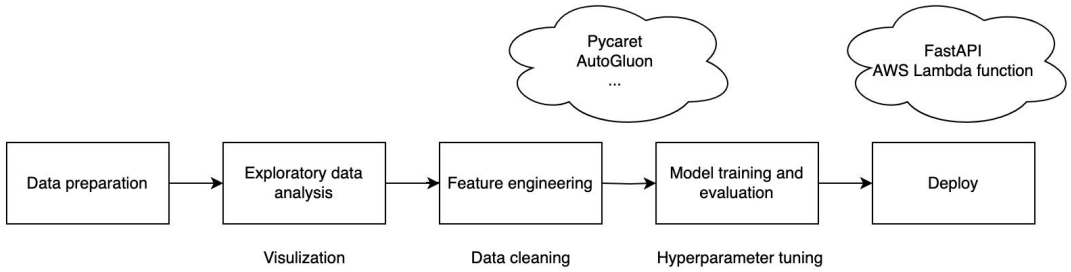
Benchmark model

A benchmark model is this logistic regression model, exhibiting an Accuracy of 55.6%, which surpasses the baseline of random guessing at 50%.

Evaluation metrics

Given the nature of this classification problem, I will use both Accuracy and F1 score as metrics to assess the performance of models.

Project design



Reference

[1] : World Health Organization. (n.d.). Cardiovascular diseases (cvds). World Health Organization. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] : Fedesoriano. (2021, September 10). Heart failure prediction dataset. Kaggle. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data