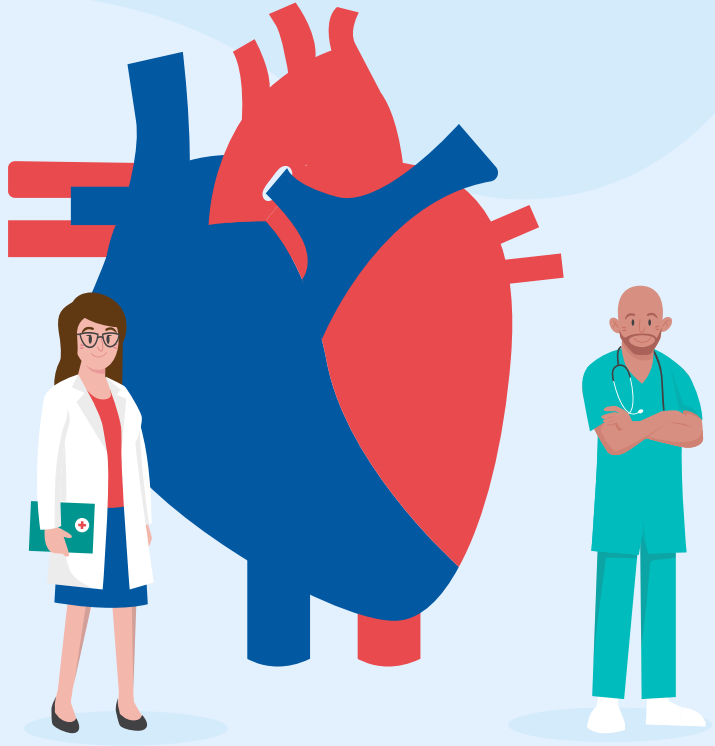
Heart Failure Prediction

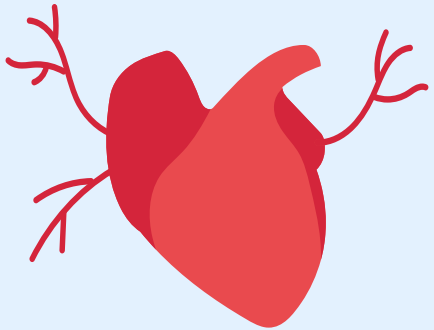


CSCI 4050U, Machine Learning
Ken Pu.
Ontario Tech University, Faculty of Science
April 16, 2023

Rija Baig, Preet Panchal, Eihab Syed

Intro to Heart Failure

- Cardiovascular diseases (CVDs) are the leading cause of death globally, accounting for 31% of all deaths worldwide and taking an estimated 17.9 million lives each year.
- Four out of 5 CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age.
- Heart failure is a common event caused by CVDs.
- The dataset mentioned in the context contains 11 features that can be used to predict a possible heart disease.
- People with cardiovascular disease or at high cardiovascular risk need early detection and management, for which a machine learning model can be helpful.



About the Dataset

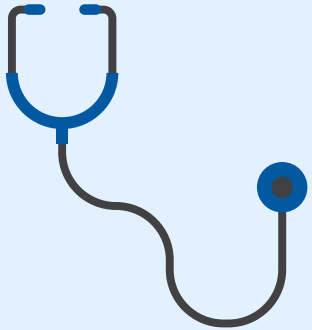
- Consists of 11 features and 918 observations.
- Helps predict presence of heart disease in patients given the features.
- **HeartDisease:** output class [1: heart disease, 0: Normal]
    - Those with heart disease (output class = 1) and those without heart disease (output class = 0).

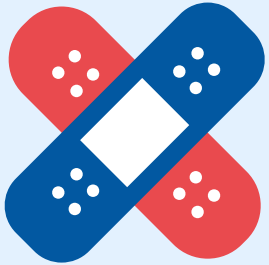
The Features

- **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]

Source of the Dataset

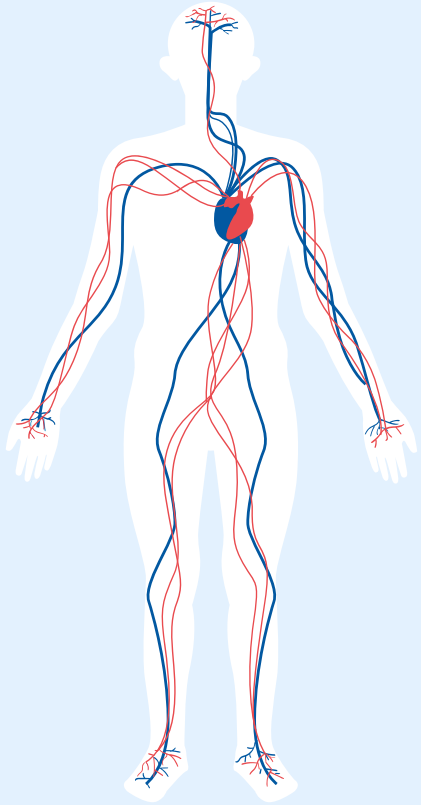
- This dataset was created by combining different datasets already available independently but not combined before. In this dataset, **5 heart datasets are combined over 11 common features** which makes it the largest heart disease dataset available so far for research purposes. The five datasets used 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**
- Every dataset used can be found under the Index of heart disease datasets from UCI Machine Learning Repository on the following link: https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/





Our Learning Problem

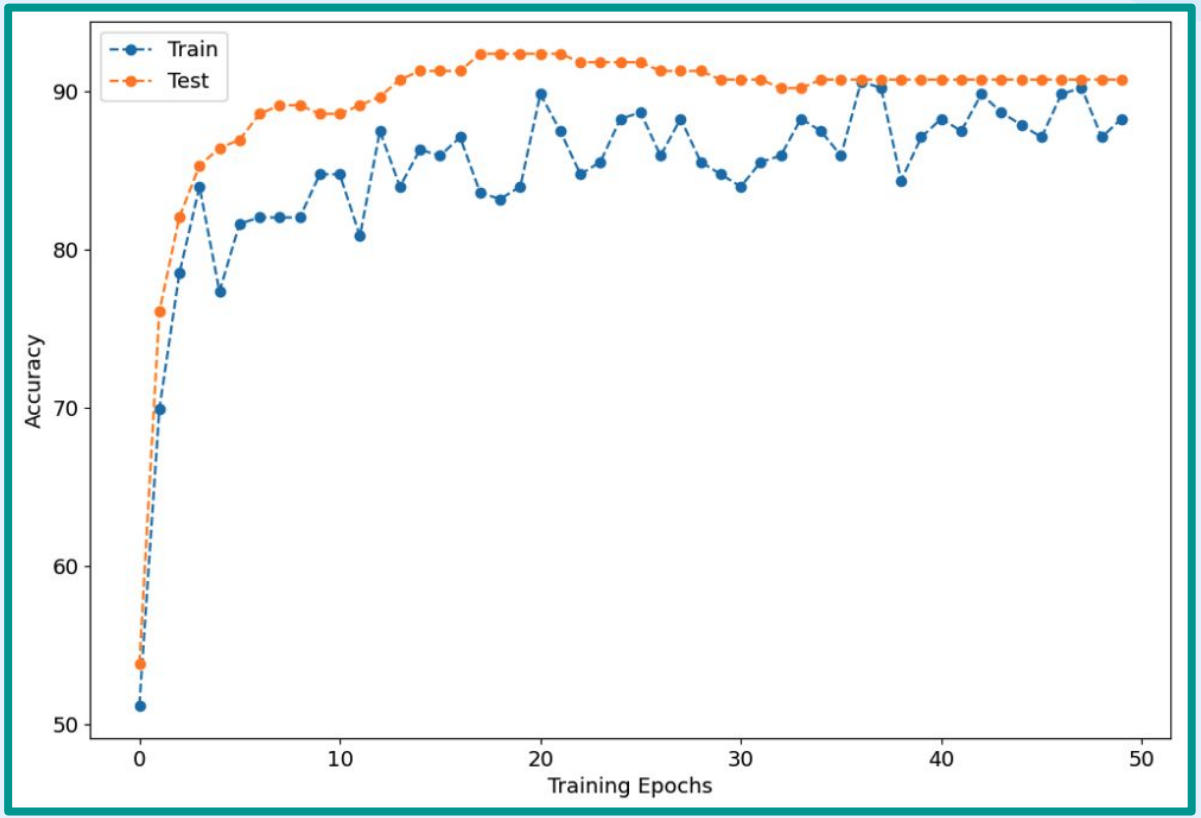
For our project, we are studying a **binary classification problem**, where the objective is to **predict whether a patient has heart disease** or not based on the 11 input features. The problem can be framed as training a machine learning model to **accurately classify patients into two categories**: those with heart disease (output class = 1) and those without heart disease (output class = 0). The aim of the project would be to **develop a machine learning model** that can generalize well on unseen data and **achieve a high accuracy rate** on the test set.



Our Team

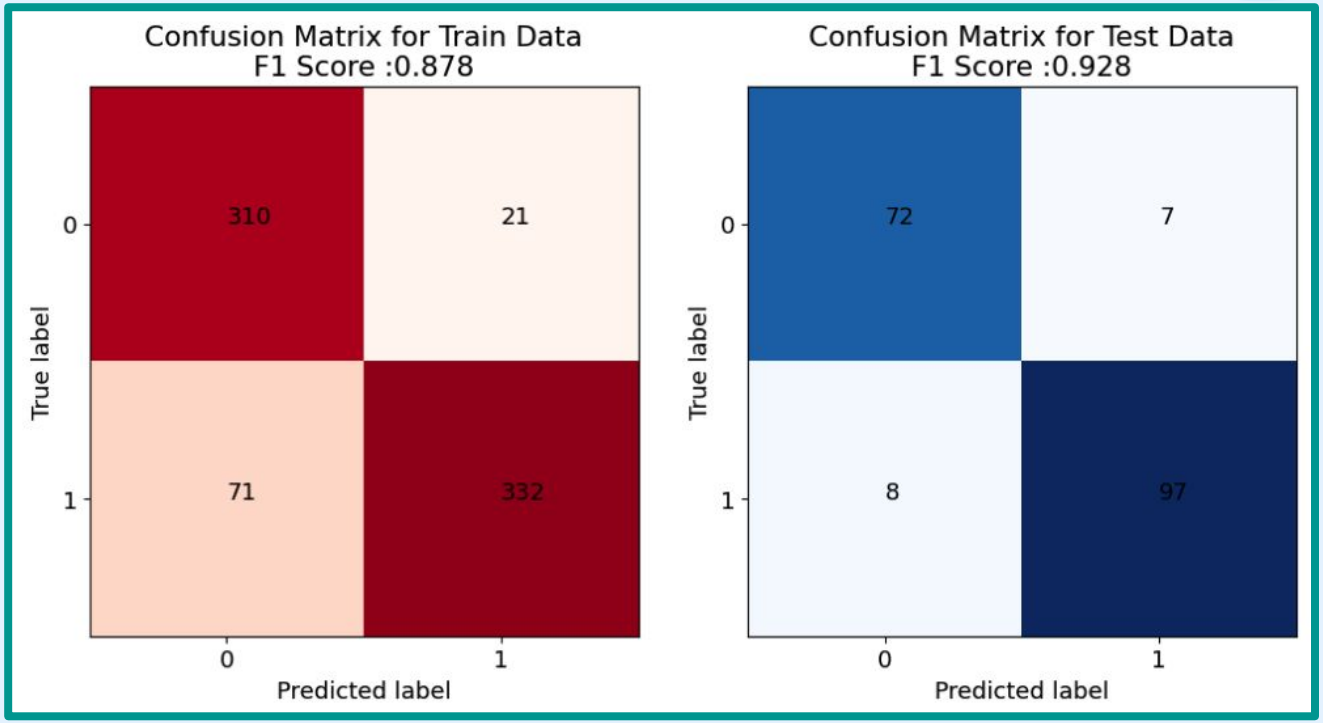
Our Findings

Visualizing the train and test accuracy for our model.



Our Findings

Visualizing the confusion matrix for training and testing datasets.



Final Remarks

For the confusion matrix with the training data, we can see that our model was able to accurately predict if a subject had heart disease. The true value being on the y axis and the predicted value being on the x axis, our model predicted 71 false negatives means our model has a **23% chance of producing false negatives**. It produced 21 false positives giving us a **6% chance at false positives**, which are very low values.

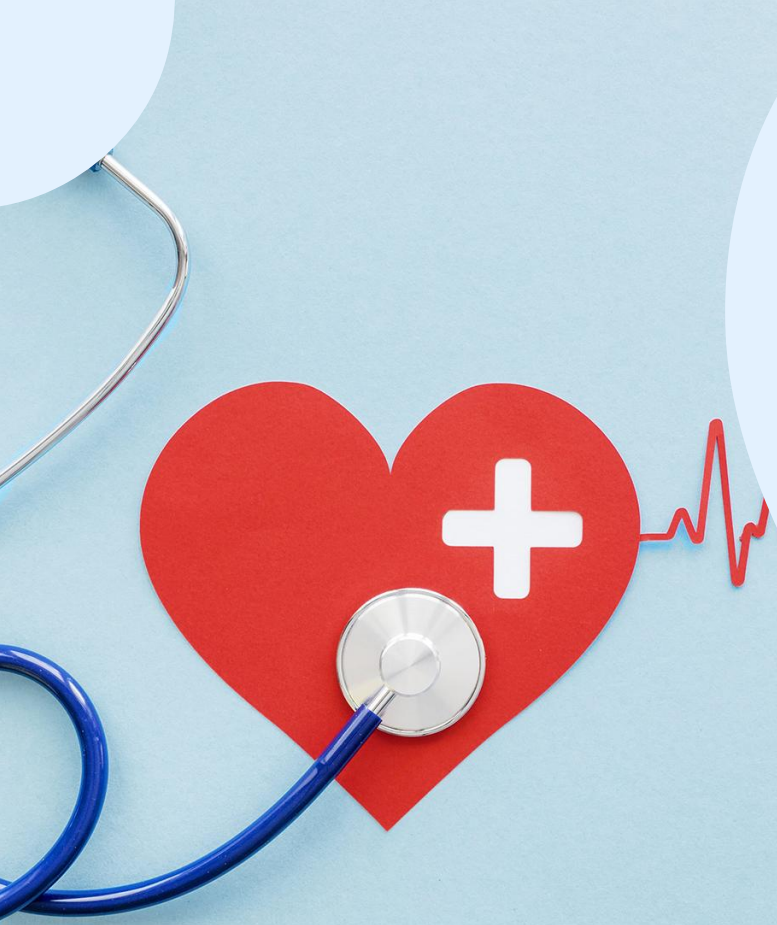
For the testing dataset, our train model provided a pretty high accuracy where we only got 8 false negative cases giving our model an **11% chance of false negatives** and 7 false positives giving us a **6% chance for false positives**.

Please Note: The numbers provided in the above statement may vary due to re-runs of the training model, however, the expectation is that the % values will be very close in correlation.

Citation & Acknowledgements

fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved [Apr 16, 2023] from https://www.kaggle.com/fedesoriano/heart-failure-prediction.

Creators:
Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.



Thank you for watching this presentation.