AI Course

Team Project Final Report

For students (instructor review required)

ⓒ2019 SAMSUNG. All rights reserved.

Samsung Electronics Corporate Citizenship Office holds the copyright of this document.

This document is a literary property protected by copyright law so reprint and reproduction without permission are prohibited.

To use this document other than the curriculum of Samsung innovation Campus, you must receive written consent from copyright holder.

| Understanding the relationship between numerous biological factors to the occurrence of heart failure |
| --- |

April 23rd, 2022

Heartful

Eric Yang

Mohamed Hasan

Amarachi Mbakwe

Azin Semsar

Yubin Lee

Luna Kim

Content

1. Introduction

1.1. Background Information

1.2. Motivation and Objective

1.3. Members and Role Assignments

1.4. Schedule and Milestones

2. Project Execution

2.1. Data Acquisition

2.2. Training Methodology

2.3. Workflow

2.4. System Diagram

3. Results

3.1. Data Preprocessing

3.2. Exploratory Data Analysis (EDA)

3.3. Modeling

3.4. User Interface (Interface).

3.5. Testing and Improvements

4. Projected Impact

4.1. Accomplishments and Benefits

4.2. Future Improvements

5. Team Member Review and Comment

6. Instructor Review and Comment

1. Introduction

**1.1. Background Information**

Heart failure occurs when the heart muscle doesn't pump blood as well as it should. Blood normally backs up and causes fluid to build up in the lungs and in the legs. The fluid buildup can cause shortness of breath and swelling of the legs and feet. Poor blood flow may cause the skin to appear blue. The number one cause of death globally is cardiovascular diseases (CVD’s), it accounts for 31 percent of deaths in the world. It is in the best interest of any one group of people to have access to data that shows significant or even sufficient evidence towards the potential occurrences of heart failures.

**1.2 Motivation and Objective**

All of the members are gathered from their deep interest in researching disease predictions. We believe that using a large amount of data with predictions about this disease would certainly be useful and valuable information for many patients and people. Thus, the capstone project goal is predicting whether one has a heart disease based on measured biological data. Moreover, we need to mainly focus on understanding the relationship between numerous biological factors to the occurrence of heart failure.

**1.3 Members and Role Assignments**

Eric Yang - Researcher and writer for plan and report, and presentation

Mohamed Hasan - Preliminary data analysis/EDA

Amarachi Mbakwe - Model creating and training

Azin Semsar - Model creating and training

Yubin Lee - model evaluation & visualization

Luna Kim - Preprocessing with write some parts of presentation and report

**1.4 Schedule and Milestones**

4/16 - 4/17: Preliminary data analysis/EDA

4/17 - 4/19: Training and creating the model

4/19 - 4/21: Testing and evaluation

4/16 - 4/29: Finalize draft

5/1 - 5/7: Additional checking up and prepare pitch presentation

5-8 - 5/13: Last wrapping up process with final report

2. Project Execution

**2.1 Data Acquisition**

Our main dataset is from the kaggle; it allows users to find and publish data sets and enter competitions to solve data science challenges. The website address is <https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction>. The data file is formed as ".cvs"(comma-separated values file), so it is easy to edit manually, simple to implement and parse. In detail, the dataset contains 918 observations with 12 attributes (columns).

**2.2 Training Methodology**

The task involved is to develop the project model that is able to effectively classify the target(goal). The target is categorized into 1 having a heart failure or 0 not having a heart failure. The model will be a classification model.

**2.3 Workflow**



2.4 System Diagram



3. Results

**3.1. Data Preprocessing**

First, we split the columns into input and output for the model. Following this, we normalized continuous numeric biometric values to be between 0 (minimum) and 1(maximum). There was time to change string value categorization for Sex, Chest pain type, Resting ECG, Exercise Angina, ST Slope to be numerical integer representations since there were a lot of variables. Finally, we splitted 918 entries of data into train and test data with k-fold.

**3.2 Exploratory Data Analysis (EDA)**

In our EDA, we evaluated numerous aspects of our dataset in order to gain an insight on how to analyze and effectively discover insights. The shape of the dataset is a 2D matrix in the shape of 918 rows by 12 columns, one of which being whether someone has a heart failure. We then inspect the descriptions of each column as well as categories that exist within them. In a box plot analysis, all of the different variables in consideration have significant overlaps with the presence of heart disease though there are only two possible values in the heart disease column. The correlations between each individual variable and the output are not strong. We use pair plots to understand the best set of features for predicting the output as well as the clusters among the variables.

**3.3 Modeling**

Linear Regression is an easily implemented model and is most often used for datasets that are linearly separable. KNN is used to solve classification model problems. This model creates a boundary to classify data, therefore larger k values mean smoother curves of separation and vice versa. XGBoost is a decision tree based ensemble machine learning algorithm that uses a gradient boosting framework. Random Forest is commonly used because of its ability to handle both classification and regression problems. It combines the outputs of multiple decision trees to reach a result that avoids overfitting and bias. It provides flexibility and makes it easy to determine feature importance. SVM is a supervised machine learning algorithm used for classification and regression models. SVM serves to find a hyperplane in an N-dimensional space that classifies data points. It is effective in high dimension cases, is memory efficient and uses different kernel functions.

**3.4 User Interface (Interface).**

The state of the best model we have trained is deserialized and exported in the format of a pkl file using the joblib library and the trained model could be serialized and loaded as a trained model. In an interactive terminal Python script, the user inputs numerous information about a patient such as their age, sex, resting heartbeat rate, and other biological metrics and receives a prediction along with a numeric confidence for whether someone has received a heart attack.

**3.5. Testing and Improvements.**

The testing and improvements process were integrated within the model development and were a crucial part to optimizing each of the model’s configurations. Through testing metrics, we have found that the support vector machine as well as random forest models were the most accurate at a rate of 89.1%.



4. Projected Impact

**4.1. Accomplishments and Benefits**

We have created a model that predicts whether someone with a number of different biological metrics would have a heart failure or not. Using this model, medical professionals could simply use one’s data to assess a risk before continuing doing more thorough examination. In addition, one could find which metric has a heavier impact on the final predicted answer by examining the weights the model training iteratively adjusted.

**4.2 Future Improvements**

Our team had to change our project dataset at the last minute because our previous dataset may very likely be computer generated as the correlation was completely perfect it appeared to not be real measured data. Even despite the deadline extension, our team has been swamped with our responsibilities from classes and final season to really improve upon it much more. If we had more time, we would expand our horizons to explore categorical models and implement a user interface.

5. Team Member Review and Comment 

|  |
| --- |

| NAME | REVIEW and COMMENT |
| --- | --- |
| Eric Yang | I am happy with the project and I’m proud of the results my team has worked hard together to build. |
| Mohamed Hasan | Wonderful atmosphere of learning, great team and a very supportive environment! |
| Amarachi Mbakwe | Working on this project was an exciting teamwork experience. Teamwork is a highly professional skill, and I am amazed to work with people I have never met before. Indeed, it was a great learning experience. Thanks to the instructor, my team members, and the SIC organizers. |
| Azin Semsar | I truly enjoyed working on this project with our amazing team. My experience was a great team work in applying Machine learning techniques and improving my knowledge in the field. |
| Yubin Lee |  |
| Luna Kim | Since it was my first outside-project, it was a little bit tough. But, it was a great opportunity to apply the variable modeling (algorithm process) on the real-world project. And I want to thank all of the members for hard working and supporting each other! |

6. Instructor Review and Comment

| CATEGORY | SCORE | REVIEW and COMMENT |
| --- | --- | --- |
| IDEA | \_\_/20 |  |
| CODING | \_\_/20 |  |
| PROJECT MANAGEMENT | \_\_/30 |  |
| PRESENTATION & REPORT | \_\_/30 |  |
| TOTAL | \_\_/100 |  |