Reconhecimento de Padrões
Inteligência Geoespacial
Aprendizagem Computacional em Biologia

2021/2022

Project Assignment

Heart Disease prediction

1 Background

According to the Centers for Disease Control (CDC) and Prevention, heart disease is one of the leading causes of death. About half of all Americans (47%) have at least 1 of 3 key risk factors for heart disease: high blood pressure, high cholesterol, and smoking. Other key indicator include diabetic status, obesity (high BMI), not getting enough physical activity or drinking too much alcohol. Detecting and preventing the factors that have the greatest impact on heart disease is very important in healthcare.

The goal of this project assignment is to detect patterns from the data that can predict a patient's condition.

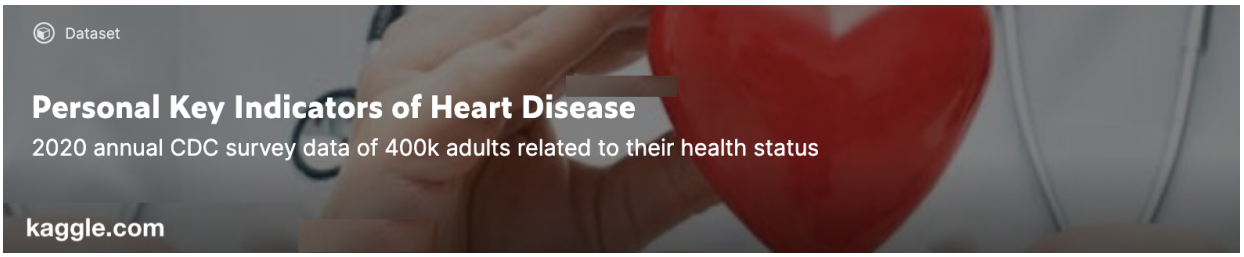

Figure 1: Indicators of Heart Diseases, adapted from kaggle.com.

2 Dataset Description

The original dataset comes from the CDC and is a major part of the Behavioral Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to gather data on the health status of U.S. residents. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world.". The most recent dataset (as of February 15, 2022) includes data from 2020. The vast majority of columns are questions asked to respondents about their health status, such as "Do you have serious difficulty walking or climbing stairs?" or "Have you smoked at least 100 cigarettes in your entire life?". https://www.kaggle.com/kamilpytlak/personal-key-indicators-of-heart-disease

The provided dataset has already been filtered containing a total of 318958 entries with 4 dependent variables and 15 independent variable. The first 4 columns are binary and correspond to the following dependent variables:

1. **CoronaryHeartDisease**, 18085 positive cases
2. **MyocardialInfarction**, 16861 positive cases
3. **KidneyDisease**, 11672 positive cases
4. **SkinCancer**, 29676 positive cases

The rest of the information includes:

Dependent variables:

- **CoronaryHeartDisease** - Respondents that have ever reported having coronary heart disease (CHD);
- **MyocardialInfarction** - Respondents that have ever reported having myocardial infarction (MI);
- **KidneyDisease** - Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?;
- **SkinCancer** - (Ever told) (you had) skin cancer?

Independent variables:

- **BMICategory** - Body Mass Index (BMI);
- **Smoking** - Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes];
- **AlcoholDrinking** - Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week);
- **Stroke** - (Ever told) (you had) a stroke;
- **PhysicalHealth** - Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?;
- **MentalHealth** - Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?;
- **DiffWalking** - Do you have serious difficulty walking or climbing stairs?;
- **Sex** - Are you male or female?;
- **AgeCategory** - Fourteen-level age category;
- **Race** - Imputed race/ethnicity value (This value is the reported race/ethnicity or an imputed race/ethnicity, if the respondent refused to give a race/ethnicity. The value of the imputed race/ethnicity will be the most common race/ethnicity response for that region of the state) 7;
- **Diabetic** - (Ever told) (you had) diabetes? (If ´Yes´ and respondent is female, ask ´Was this only when you were pregnant?´. If Respondent says pre-diabetes or borderline diabetes, use response code 4.);
- **PhysicalActivity** - Adults who reported doing physical activity or exercise during the past 30 days other than their regular job;
- **GenHealth** - Would you say that in general your health is;
- **SleepTime** - On average, how many hours of sleep do you get in a 24-hour period?;
- **Asthma** - (Ever told) (you had) asthma?;

Figure 2: Information of the dataset.

In this project we propose the use of two derived variables that should be calculated accordingly:

1. **HeartDisease**, 26536 positive cases. Positive when reported `CoronaryHeartDisease OR MyocardialInfarction`;
2. **HeartDiseaseComorbities**, 19111 with Coronary Heart Disease but no Comorbities, 7425 with Coronary Heart Disease and Comorbities and 292422 with no Heart Disease. otherwise. Comorbidities are considered if reported `KidneyDisease OR SkinCancer`.

3 Objective

Your task is to develop classifiers for Heart Disease. Consider three scenarios:

- **Scenario A (Coronary Heart Disease)**: where a single classifier should be used to predict if a patient has **Coronary Heart Disease**;
- **Scenario B (Heart Disease)**: where a classifier, is designed to distinguish if a patient with Heart Disease has either **CoronaryHeartDisease** or **MyocardialInfarction**;
- **Scenario C (Heart Disease with comorbidities)**: where classifiers should be designed to distinguish three classes: **HeartDiseaseNoComorbidities**, **HeartDiseaseComorbidities**, and **NoHeartDisease**.

4 Practical Assignment

4.1 Data import

Implement scripts for feature data import. Organize data into sub-sets, relating to each source type you intend to test, e.g..: features from patient data. Use 200k samples for running the Experiments and the remaining data for Validation.

4.2 Feature Selection and Reduction

Some of the supplied features may be useless, redundant or highly correlated with others. In this phase, you should consider to use feature selection and dimensionality reduction techniques, and see how they affect the performance of the pattern recognition algorithms. Analyse the distribution of the values of your features and compute the correlation between them. Make sure you know your features! Do not forget to present your findings in the final report.

4.3 Experimental Analysis

You should be able to design experiments in order to run the pattern recognition algorithms in the given data and evaluate their results. Define the appropriate performance metrics and justify your choices!

Run the experiments multiple times! To be able to present average results and standard deviations (of the metrics used) you should split the training set in parts and use cross-validation. At the end you should be able to choose the best classifier and evaluate them in a testing set.
Do not forget that manually inspecting the predictions of your algorithms can give you precious insights of where they might be failing (and why), and what you can do to improve them (e.g. what makes the algorithm fail in this particular case? what special characteristic does it have that makes it so hard? how can I make the algorithm better deal with those cases?). Go back and forward to the Pre-processing, Feature reduction and Feature Selection phases until you are satisfied with the results. It is a good idea to keep track of evolution of the performance of your algorithm during this process. Try to show these trends in your final report, to be able to justify all the issues involved (choosing parameters, model fit, etc.)

4.4 Pattern Recognition Methods

You can write your own code in your language of choice or use the functions and methods available in MATLAB and in the Statistical Pattern Recognition STPRTool used in the classes (since you are already familiarized with it). The methods used in your work should be described as well as discussion of the parameters used. Try out different pattern recognition algorithms. You should try to understand how they perform differently in your data.

4.5 Results and Discussion

Present and discuss final results obtained in your Project assignment. This problem was already studied by other authors. Compare your results with the results from other sources. In this problem one important aspect is to evaluate among the data available the one more appropriate for the different scenarios.

4.6 Code & Graphical User Interface (GUI)

You should deliver your software code in MATLAB, or in any other programming language you used during the project.

For your project you should write code for a graphical user interface (GUI). The GUI should improve the interaction of the user with the code by providing options for data-loading, feature selection/dimensionality reduction, classification, post-processing, validation and visualization.
Remember to comment your code. Write also a help section to your code that tells the purpose of the function, usage, and explanation of parameters.

5 Documentation

Write documentation (in Portuguese or in English) about your project. The documentation should include a cover page where course name, project title, date, names and student numbers of the authors are mentioned.

Describe the methods used for classification in such detail that the reader would be able to implement the same kind of functions for feature extraction and classification just based on your documentation and some basic background in pattern recognition. Always justify your choices, even when their are based on intuition. Do not forget to verify your assumptions! Include classification results with the given data to your documentation. At the end of your documentation you should have a list of all references used.

5.1 Requirements

Practical assignment is meant to be done in groups of two persons, but up to three are allowed. If someone wants to work alone, this is also possible. Larger groups are in principle not allowed.

5.2 Project Submission & Deadlines

1. **Project First Milestone (Deadline: 24th April 2022!)**
   Deliverables:
   - Data Preprocessing (Scaling, Feature Reduction (PCA & LDA), Feature Selection, etc.);
   - Minimum Distance classifier, Fisher LDA for Scenario A.
   - Code + short report.
2. **Project Final Goal (Deadline: 20th May 2022!)**
   Deliverables:
   - Data Preprocessing (Scaling, Feature Reduction (PCA & LDA), Feature Selection, etc.);
   - Several classifiers (Scenario A, Scenario B, Scenario C);
   - Final Report
   - Matlab code + GUI.
3. **Presentation and Discussion (from 30th May to 2nd June 2022!)**