## Motivation

1. For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
   Cardiovascular diseases (CVDs) are the number 1 cause of death globally. 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.

2. Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
   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.
   Donor:
   David W. Aha (aha '@' ics.uci.edu) (714) 856-8779

## Composition

1. What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
   The instances are people's personal information about 13 features. Some of them are general features, but most of them are related to heart.
2. What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)or features? In either case, please provide a description.
   Age: age of the patient [years]
   Sex: sex of the patient [M: Male, F: Female]
   Race: race of the patient [Asian, Other, White, Hispanic, Black]
   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]
3. Is there a label or target associated with each instance? If so, please provide a description.
   The label is specified according to different features, as described above.

| | |
|---|---|
| | 4. How many instances are there in total (of each type, if appropriate)?<br>There are 1190 observations in total, with 272 observations duplicated. The final dataset includes 918 observations. |
| **Collection process**<br>1. How was the data associated with each instance acquired?<br>This dataset was created by combining different datasets already available independently but not combined before.<br>2. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?<br>[INSERT ANSWER HERE]<br>3. Did the individuals in question consent to the collection and use of their data?<br>[INSERT ANSWER HERE] | **Preprocessing/cleaning/labeling**<br>1. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?<br>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. There are duplicated 272 observations removed.<br>2. Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.<br>[INSERT ANSWER HERE] |
| **Uses**<br>1. Has the dataset been used for any tasks already? If so, please provide a description.<br>[INSERT ANSWER HERE]<br>2. Are there tasks for which the dataset should not be used? If so, please provide a description.<br>This data is collected solely in the CVDs domain, so systems trained on it may or may not generalize to other sentiment prediction tasks. Consequently, such systems should not—without additional verification—be used to make consequential decisions about people. | **Distribution/maintenance**<br>1. Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?<br>Yes, the dataset is publicly available on the internet.<br>2. How can the owner/curator/manager of the dataset be contacted (e.g., email address)?<br>[INSERT ANSWER HERE] |

**Considering contextual integrity framing of privacy, identify two forms of information flow in this scenario: one that is acceptable and another that is unacceptable?**

According to the interview transcripts in the full case study, for a patient, telling his/her symptoms to the doctor and letting doctors measure his/her vital signs, such as heart rate, blook pressure, respiratory rate, and temperature are acceptable. However, leaving those personal information to the machine to diagnose without informing patients how those information is going to be used is unacceptable.

**Part 2:**

**Identify any identifiers, quasi-identifiers and sensitive attributes in the given data. Calculate k-anonymity and l-diversity for row 8-27 of this dataset. Provide a modified 3-anonymous dataset from these 20 rows.**

There’s no identifier. However, other attributes except for FastingBS could be quasi-identifiers. Sensitive attributes are those other than age, sex, race.
Without preprocessing, k as well l equals 1.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **>=45** | **F** | ***** | **ATA** | **130** | **237** | **0** | **Normal** | **170** | **N** | **0** | **Up** | **0** |
| **>=45** | **M** | ***** | **ATA** | **110** | **208** | **0** | **Normal** | **142** | **N** | **0** | **Up** | **0** |
| **<45** | **M** | ***** | **ASY** | **140** | **207** | **0** | **Normal** | **130** | **Y** | **1.5** | **Flat** | **1** |
| **>=45** | **F** | ***** | **ATA** | **120** | **284** | **0** | **Normal** | **120** | **N** | **0** | **Up** | **0** |
| **<45** | **F** | ***** | **NAP** | **130** | **211** | **0** | **Normal** | **142** | **N** | **0** | **Up** | **0** |
| **>=45** | **M** | ***** | **ATA** | **136** | **164** | **0** | **ST** | **99** | **Y** | **2** | **Flat** | **1** |
| **<45** | **M** | ***** | **ATA** | **120** | **204** | **0** | **Normal** | **145** | **N** | **0** | **Up** | **0** |
| **<45** | **M** | ***** | **ASY** | **140** | **234** | **0** | **Normal** | **140** | **Y** | **1** | **Flat** | **1** |
| **<45** | **F** | ***** | **NAP** | **115** | **211** | **0** | **ST** | **137** | **N** | **0** | **Up** | **0** |
| **>=45** | **F** | ***** | **ATA** | **120** | **273** | **0** | **Normal** | **150** | **N** | **1.5** | **Flat** | **0** |
| **<45** | **M** | ***** | **ASY** | **110** | **196** | **0** | **Normal** | **166** | **N** | **0** | **Flat** | **1** |
| **<45** | **F** | ***** | **ATA** | **120** | **201** | **0** | **Normal** | **165** | **N** | **0** | **Up** | **0** |
| **>=45** | **M** | ***** | **ASY** | **100** | **248** | **0** | **Normal** | **125** | **N** | **1** | **Flat** | **1** |
| **<45** | **M** | ***** | **ATA** | **120** | **267** | **0** | **Normal** | **160** | **N** | **3** | **Flat** | **1** |
| **<45** | **F** | ***** | **TA** | **100** | **223** | **0** | **Normal** | **142** | **N** | **0** | **Up** | **0** |
| **<45** | **M** | ***** | **ATA** | **120** | **184** | **0** | **Normal** | **142** | **N** | **1** | **Flat** | **0** |
| **>=45** | **F** | ***** | **ATA** | **124** | **201** | **0** | **Normal** | **164** | **N** | **0** | **Up** | **0** |
| **<45** | **M** | ***** | **ATA** | **150** | **288** | **0** | **Normal** | **150** | **Y** | **3** | **Flat** | **1** |
| **<45** | **M** | ***** | **NAP** | **130** | **215** | **0** | **Normal** | **138** | **N** | **0** | **Up** | **0** |
| **<45** | **M** | ***** | **NAP** | **130** | **209** | **0** | **Normal** | **178** | **N** | **0** | **Up** | **0** |

**Part 3:**
Please check the readme file.