Motivation

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) is the worldwide leading cause of death. As such, this dataset, which is a composition of other smaller datasets, was created in an attempt to better predict heart failures given medical information.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

FEDESORIANO (Kaggle username) is the creator of this dataset. No affiliated organization is mentioned on Kaggle.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Any other comments?

Composition

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 data consists of people, their corresponding general and medical information, and whether the suffered from heart failure.

How many instances are there in total (of each type, if appropriate)?

There are a total of 918 unique rows in this dataset, each having 11 features and one associated label (12 features if you include the synthetic feature for race).

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.

In all cases, the data for each instance consists in one of two formats: a value describing a medical reading or general information, or a string describing a medical state or general information.

Is there a label or target associated with each instance? If so, please provide a description.

All the instances of data are labelled by the “HeartDisease” output class, which tells whether the given patient did or did not suffer from a heart condition.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

All rows included in the dataset are completely filled.

Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)? If so, please describe how these relationships are made explicit.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

There are no recommended data splits.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

All data is either qualitative (and cannot have any source of error) or is a medical measurement or assignment. Though sources of error are possible in these cases, the likelihood of the medical equipment outputting false values or readings is relatively low.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

This dataset is comprised of five other, smaller datasets and is now self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

This dataset does contain age and sex as features. An additional synthetic feature for race was added for the purposes of this assignment.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

Though no names or obvious identifiers are present in the dataset, this dataset is 1-anonymous, meaning that there exist unique rows in the data.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

Though not originally in the dataset, the dataset used for the assignment does have a race column. Other sensitive data may include gender and age, both of which are present in the dataset.

Any other comments?

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data for this dataset was collected by compiling information from 5 other datasets, for a combined 11 features and a total of 918 people. The data on race was created synthetically for the purposes of this assignment.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

Who was involved in the data collection process (e.g., students, crowd workers, contractors) and how were they compensated (e.g., how much were crowd workers paid)?

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the time-frame in which the data associated with the instances was created.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

Every individual dataset used to create the larger, all-encompassing dataset 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/

Were the individuals in question notified about the data collection?

If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

For this particular dataset, seemingly no consent was given, however it cannot be confirmed whether consent was acquired for the 5 datasets from which this dataset was created.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Any other comments?

Preprocessing/cleaning/labeling

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)? If so, please provide a description. If not, you may skip the remaining questions in this section.

There is no mention or obvious place where any preprocessing/cleaning of the data was performed. All of the feature data is taken directly from patient information files and medical readings. Labelling has been performed to tell whether the given patient suffered from a heart disease.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data

Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

This dataset has been downloaded nearly 90000 times on Kaggle. Though people can proceed how they see fit with the data, it's primary purpose is to take the given data and develop a classifier which can predict whether a certain patient does or does not have heart disease.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

What (other) tasks could the dataset be used for?

This dataset could be used to accomplish any sort of medical diagnosis classifier. Assuming the condition in question is associated with the gathered feature data, this same dataset could be used to diagnosis diseases or conditions other than heart disease.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

Are there tasks for which the dataset should not be used? If so, please provide a description.

As is the case with any medical data, this data should not be used as a means to discriminate against people who fall under the umbrella of any of the features of the dataset.

Any other comments?

Distribution

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? If so, please provide a description.

Though Kaggle does not specify any intentions to specifically distribute this dataset to a third party, anyone, including any third party, can access, download and use this dataset through Kaggle. This can be seen as both an upside and downside of open-source datasets, depending on the intentions of the dataset's user.

How will the dataset will be distributed (e.g., tar ball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

When will the dataset be distributed?

Will the dataset be distributed under a copyright or other intellectual property (IP)

license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? Though it is not necessarily evident, presumably the owner can be contacted since their Kaggle username is public information.

Is there an erratum? If so, please provide a link or other access point.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)? According to Kaggle, there is no intention for the author to update the dataset.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

Will older versions of the dataset continue to be sup-ported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

Any other comments? If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description. Others may do so and should contact the original authors about incorporating fixes/extensions.

Any other comments?