**Professional Certificate in Machine Learning and Artificial Intelligence**

Portfolio Project – Erika Kenny: Transparency and Interpretability

Datasheet for Dataset Template

The questions below will help you provide context for the Heart Failure Prediction Dataset (https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction). Provide answers where possible.

**Motivation**

Answer the following questions:

* 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.
  + Answer: No provide early diagnosis of heart disease to start treatment as soon as possible and improve patient outcomes.
* Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? Who funded the creation of the dataset?
  + Answer:
    - 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.

**Composition**

Answer the following questions:

* What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Please provide details.
  + Answer: Numeric data (integers and floats), data objects and 1 Boolean (target)
* How many instances of each type are in total?
  + Answer: 6 objects, 1 Boolean and 5 numeric
* 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).
  + Answer: The dataset represents a good, balanced mix of patients with heart disease and patients without heart disease.
* What does each instance consist of? Raw data? Unprocessed? Text, images?
  + Answer: Clean, reliable data that was recorded by medical professionals.
* Are there any labels to the data?
  + Answer: Yes, Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, ST\_Slope and target variable HeartDisease
* Is there any missing information from individual instances?
  + Answer: No
* Are relationships between individual instances made explicit?
  + Answer: Yes
* Are there recommended data splits (e.g. train / test)? Provide a description of the splits, and the rationale behind them.
  + Answer: I used a function to split the data into train and test set (70/30).
* Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?
  + Answer: 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.
  + Answer: Yes, I think all data contained in this data set would be protected by HIPPA.
* Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
  + Answer: No
* Does the dataset identify any subpopulations (e.g., by age, gender)?
  + Answer: Yes, ages are recorded.
* 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.
  + Answer: Yes, someone could conceivably use combinations of the subject’s age and details to possibly identify the individual in question.
* 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.
  + Answer: Yes, all the data in this dataset are age or health-related and therefore, highly sensitive.
* Any other comments?
  + Answer: These data would need to be treated with the highest confidentiality.

**Collection process**

Answer the following questions:

* How was the data associated with each instance acquired? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
  + Answer: ‘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.’
* If the data is a sample of a larger subset, what was the sampling strategy? Deterministic, random, etc...?
  + Answer: Deterministic. Individuals who has reason to seek medical treatment.
* Over what time frame was the data collected?
  + Answer: ‘Unknown to the authors of the datasheet’
* Were there any ethical review processes conducted (e.g. by an institutional reviewing board?)
  + Answer: ‘Unknown to the authors of the datasheet’
* Were the individuals notified of the collection of the data?
  + Answer: ‘Unknown to the authors of the datasheet’ Being asked to participate in a regular exam is not the same as knowing how the data will be used.
* Did the individuals consent to their data being collected?
  + Answer: ‘Unknown to the authors of the datasheet’
* If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?
  + Answer: ‘Unknown to the authors of the datasheet’
* 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?
  + Answer: ‘Unknown to the authors of the datasheet’
* Any other comments?
  + Answer: None

**Preprocessing/cleaning/labelling**

Answer the following questions:

* 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.
  + Answer: ‘Unknown to the authors of the datasheet’
* Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?
  + Answer: ‘Unknown to the authors of the datasheet’
* Any other comments?
  + Answer: None

**Uses**

Answer the following questions:

* What other tasks could the dataset be used for?
  + Answer: The dataset could be used when newer, better AI/ML tools become available.
* 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?
  + Answer: These data should not be used to deny participants healthcare.
* Are there tasks for which the dataset should not be used? If so, please provide a description.
  + Answer: The data should not be used to identify any patients in this study.
* Any other comments?
  + Answer: None

**Distribution**

Answer the following questions:

* 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.
  + Answer: ‘Unknown to the authors of the datasheet’
* How will the dataset be distributed?
  + Answer: ‘Unknown to the authors of the datasheet’
* When will the dataset be distributed?
  + Answer: ‘Unknown to the authors of the datasheet’
* 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.
  + Answer: ‘Unknown to the authors of the datasheet’
* Any other comments?
  + Answer: None

**Maintenance**

Answer the following questions:

* Who will be maintaining the dataset?
  + Answer: ‘Unknown to the authors of the datasheet’
* Any other comments?
  + Answer: None