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

Module 20: Transparency and Interpretability

Datasheet for Dataset Template

The questions below will help you provide context for the Pima Indian Diabetes Dataset. 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: The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.
* 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: Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). [Using the ADAP learning algorithm to forecast the onset of diabetes mellitus](http://rexa.info/paper/04587c10a7c92baa01948f71f2513d5928fe8e81). *In Proceedings of the Symposium on Computer Applications and Medical Care* (pp. 261--265). IEEE Computer Society Press.
* Any other comments?
  + Answer:

**Composition**

Answer the following questions:

* What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Please provide details.
  + Answer: The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.
* How many instances of each type are in total?
  + Answer: 768
* 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:  In particular, all patients here are females at least 21 years old of Pima Indian heritage.
* What does each instance consist of? Raw data? Unprocessed? Text, images?
  + Answer: Raw Data
* Are there any labels to the data?
  + Answer: Yes
* Is there any missing information from individual instances?
  + Answer: Yes invalid values
* 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:No
* Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?
  + Answer:No
* 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:No
* 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:No
* 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:No
* 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:
* Any other comments?
  + Answer:

**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: Inspiration was - Can you build a machine learning model to accurately predict whether or not the patients in the dataset have diabetes or not?
* If the data is a sample of a larger subset, what was the sampling strategy? Deterministic, random, etc...?
  + Answer:
* Over what time frame was the data collected?
  + Answer:
* Were there any ethical review processes conducted (e.g. by an institutional reviewing board?)
  + Answer:
* Were the individuals notified of the collection of the data?
  + Answer:
* Did the individuals consent to their data being collected?
  + Answer:
* If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?
  + Answer:
* 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:
* Any other comments?
  + Answer:

**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: Yes, missing values diabetes\_data\_copy[['Glucose','BloodPressure','SkinThickness','Insulin','BMI']] = diabetes\_data\_copy[['Glucose','BloodPressure','SkinThickness','Insulin','BMI']].replace(0,np.NaN)
* Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?
  + Answer:No
* Any other comments?
  + Answer:

**Uses**

Answer the following questions:

* What other tasks could the dataset be used for?
  + Answer: Diabetes classification
* 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:
* Are there tasks for which the dataset should not be used? If so, please provide a description.
  + Answer:
* Any other comments?
  + Answer:

**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:
* How will the dataset be distributed?
  + Answer:
* When will the dataset be distributed?
  + Answer:
* 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: [CC0: Public Domain](https://creativecommons.org/publicdomain/zero/1.0/)
* Any other comments?
  + Answer:

**Maintenance**

Answer the following questions:

* Who will be maintaining the dataset?
  + Answer: Probably,   
    UCI Machine Learning (Owner)

Kaggle Team (Admin)

* Any other comments?
  + Answer