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.

This dataset was generated by aggregating already existing data sets. Namely, this data set aggregated the data from the following already published datasets :

- Cleveland: 303 observations
- Hungarian: 294 observations
- Switzerland: 123 observations
- Long Beach VA: 200 observations
- Stalog (Heart) Data Set: 270 observations

Total: 1190 observations

Duplicated: 272 observations

Final dataset: 918 observations

The main aim behind the creation of this dataset is to generate the largest possible dataset so far with the most features to develop more accurate and generalizable predictive machine learning algorithms. Those algorithms' main goal is the early detection of cardiovascular disease and/or cardiovascular risk in patients.

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

The dataset was created by M.D. Andras Janosi of the Hungarian Institute of Cardiology (Budapest), M.D. William Steinbrunn of the University Hospital (Zurich, Switzerland), M.D. Matthias Pfisterer of the University Hospital (Basel, Switzerland), and M.D. Robert Detrano of the V.A. Medical Center, Long Beach and Cleveland Clinic Foundation. David W. Aha is also to be noted as a donor.

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.

As mentioned in the question above, David W. Aha of the University of California Irvine is to be noted as a donor.

Any other comments?

None.

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.

Each instance of the dataset corresponds to a specimen (human patient), and each feature of this dataset corresponds to an attribute of that patient, like sex, age, and a selection of cardiovascular features (ChestPainType, RestingBP, MaxHR, amongst others). Another feature was added synthetically which corresponds to the race of the individual. The presence of a heart disease is noted by (1), whereas the absence of a heart disease is noted by a (0).

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

As mentioned above, this dataset is the result of a merger of other datasets. After getting rid of duplicates, this dataset is composed of 918 instances, with each 12 attributes. One can also note that overall, the dataset is composed of a gaussian distribution of ages, and male samples are overrepresented in the dataset (79% male vs 21% female). Finally, about 44.66% of samples do not present a heart disease, while 55.33% do.

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).

As mentioned, this dataset is an aggregation of 5 pre-existing datasets over 11 common features. Getting rid of duplicates (272 duplicated observations), this dataset contains in total 918 observations.

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

Each instance consists of raw data about the following attributes (11 + Heart Disease + Race (added synthetically)):

1. Age: age of the patient [years]
2. Race (added synthetically for the purpose of this assignment)
3. Sex: sex of the patient [M: Male, F: Female]
4. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
5. RestingBP: resting blood pressure [mm Hg]
6. Cholesterol: serum cholesterol [mm/dl]
7. FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
8. 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]
9. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
10. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
11. Oldpeak: oldpeak = ST [Numeric value measured in depression]
12. ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
13. HeartDisease: output class [1: heart disease, 0: Normal]

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

Each sample is labeled (1) if the sample does have a heart disease, or (0) if the patient does not have a heart disease.

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.

Every sample has the same attributes populated. No sample is missing data compared to the other samples (no instance has uncomplete data). It is also worth noting that this dataset is only based on the common 11 features of the original dataset. In that sense, some data was removed from certain samples so that all attributes could be shared across all samples.

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.

None (explicitly).

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.

None (explicitly).

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

See preprocessing below.

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.

The dataset is entirely self-contained. It only relies on other datasets in the sense that it is an aggregate of other datasets, but it is its own entity.

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.

There is sensitive data in the sense that the dataset displays the presence or not of heart disease for some patients.

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

No.

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.

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.

The dataset contains some quasi-identifiers, in that no direct information about the identity of each instance is made available. However, it would be technically possible to retrace the identity of some specimen if cross-checked with other resources (original databases, or other sensitive information from hospitals for example).

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.

The dataset reveals the age, sex and race of each specimen, but that last attribute is randomly generated for each sample.

Any other comments?

None.

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.

This dataset was formed by combining data from other datasets. The original method of collecting data (respectively to each source dataset) is unknown to the authors of our dataset, but can theoretically be retrieved by retrieving this information for each source dataset.

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?

Unknown to the authors of the datasheet.

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

All instances from the source datasets were combined into one dataset removing duplicates. No sampling was done from the original datasets to this dataset.

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)?

Unknown to the authors of the datasheet.

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 timeframe in which the data associated with the instances was created.

Unknown to the authors of the datasheet.

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.

Unknown to the authors of the datasheet.

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

As described above, the data was collected from various datasets stated above.

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.

Unknown to the authors of the datasheet.

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.

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? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A.

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.

N/A.

Any other comments?

None.

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.

The only preprocessing done (to the knowledge of the author) was that only the common features across all datasets were kept for the purpose of the aggregated dataset. In addition to that, the duplicated samples were removed. Finally, a feature was added to the dataset : race. This feature was added synthetically.

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.

Yes. Every source dataset used 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/

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

No.

Any other comments?

None.

Uses

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

At the time of publication, at least 974 projects reference the use of this dataset on Kaggle.

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.

All the projects that mention the use of this dataset on Kaggle are available using the code tab of the Kaggle webpage, which can be found here : https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/code

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

The dataset could be used for anything related to modeling or understanding heart diseases. I can also be used as a mean for understanding the statistical distributions underlying heart diseases.

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?

One would have to put in context that the use of machine learning for healthcare does not constitute a medical diagnostic in itself. Rather, it should be used as a mean to better understand the statistical characteristics underlying certain heart diseases.

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

This dataset should not be used as the foundation of a predictive model that would constitute a medical diagnostic in itself. Rather, the predictions obtained from this dataset would have to be confronted to the opinion of a medical professional. Moreover, one could also state that this dataset would also benefit from being even more enlarged to more samples to provide more generalizable results.

Any other comments?

One could note that this dataset constitutes in itself an effort to generalize predictions to different samples with different characteristics. However, some populations might present closer attribute patterns (leading to heart disease) than others. Therefore, it could hardly (at this stage) constitute an accurate and reliable diagnostic tool for any specimen. One would still have to confront the predictions obtained from feeding this dataset to a model to the opinion of a medical professional.

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.

Yes, the dataset is publicly available on the internet.

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

The dataset is distributed on Kaggle here : https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data

When will the dataset be distributed?

The dataset was first released in September 2021.

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.

The dataset is published under the following license :
https://opendatacommons.org/licenses/odbl/1-0/

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.

Unknown to authors of the datasheet.

Any other comments?

None.

Maintenance

Who will be supporting/hosting/maintaining the dataset?

This dataset was last updated 2 years ago by fedesoriano, and was not updated since.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The expected update frequency of the dataset is never. Therefore, one should not expect any further update of the dataset.

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

N/A

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)?

No, the dataset will not be updated or maintained.

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.

N/A.

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

No.

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?

None.

ECSE 557

Intro to Ethics of Intelligent Systems

Assignment 2:

Data and Privacy

CHIKHAOUI Hamza 260912960

Winter 2024

Part 1 : Datasheet and contextual integrity:

1. See Datasheet
2. The stakeholders of this case study are the patients, the nurses, the doctors and the teams developing Prioritize. The patients are natural stakeholders as they are the ones that will be triaged, the nurses are also at the forefront of the triage process. Finally, the doctors are also stakeholders as the main aim of Prioritize is reducing the time between the moment where the patients gets to the hospital to the point where the patients are taken in charge by a doctor.
   A potential acceptable flow of information would be to use the data collected by Prioritize to accelerate the triage process of patients, and to offer a first framework for the doctor's diagnosis. The prediction generated would then only be used for the benefit of the patient and strictly to ensure an overall more efficient (and therefore better) experience for patients.
   A potential unacceptable flow of information would be for pharmaceutical companies to collect the data used by Prioritize for targeted marketing purposes (for example, for the advertisement of certain drugs depending on the profile of the patient). This flow of information would imply that the information is used for the benefit of the pharmaceutical company first. Furthermore, this use of data was not agreed upon by the patient which is in a vulnerable position, which constitutes a breach in the patient's privacy and consent.

Part 2 : K-anonymity and l-diversity:

3. In this dataset, there is no direct identifiers, as no directly identifiable information is available on the patients (such as their names or social security numbers for example). However, there are multiple quasi-identifiers such as age, sex and race. The sensitive attribute of this dataset is the presence (or not) of a heart disease in each patient.
   Simply by looking at the first line, it appears as if only one sample is 45 years of age, a female, and Hispanic. Therefore, one can say that the k-anonymity in this case is 1. This is to be confronted further by the fact that the dataset comprises a lot more male samples than female samples, which simplifies even further the identification of a female sample.
   Because the dataset is already 1-anonymous, then the different instances cannot be all grouped in groups of more than 1 sample. Therefore, the l-diversity is also limited to 1. Indeed, one would not be able to construct any equivalent classes such that each "class of data samples" contains at least more than 1 sample.
   We can note that the 3-anonimity table developed below was generated using the anonypy library. The results of that cell have just been uploaded on excel for illustration purposes.
   See below the 3-anonimity table :

Table 1 : 3-anonimity sample of the dataset (non-anonymized features)

Age	Sex	Race	ChestPainTyp	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngi	Oldpeak	ST_Slope	HeartDisease
45	F	Non-white	ATA	130	237	0	Normal	170	N	0	Up	0
48	F	Non-white	ATA	120	284	0	Normal	120	N	0	Up	0
54	F	Non-white	ATA	120	273	0	Normal	150	N	1.5	Flat	0
49	F	Non-white	ATA	124	201	0	Normal	164	N	0	Up	0
54	M	*	ATA	110	208	0	Normal	142	N	0	Up	0
58	M	*	ATA	136	164	0	ST	99	Y	2	Flat	1
60	M	*	ASY	100	248	0	Normal	125	N	1	Flat	1
37	M	Hispanic	ASY	140	207	0	Normal	130	Y	1.5	Flat	1
36	M	Hispanic	ATA	120	267	0	Normal	160	N	3	Flat	1
36	M	Hispanic	NAP	130	209	0	Normal	178	N	0	Up	0
37	F	Non-white	NAP	130	211	0	Normal	142	N	0	Up	0
42	F	Non-white	NAP	115	211	0	ST	137	N	0	Up	0
43	F	Non-white	ATA	120	201	0	Normal	165	N	0	Up	0
43	F	Non-white	TA	100	223	0	Normal	142	N	0	Up	0
39	M	Black	ATA	120	204	0	Normal	145	N	0	Up	0
38	M	Black	ASY	110	196	0	Normal	166	N	0	Flat	1
40	M	Black	NAP	130	215	0	Normal	138	N	0	Up	0
49	M	*	ASY	140	234	0	Normal	140	Y	1	Flat	1
44	M	*	ATA	120	184	0	Normal	142	N	1	Flat	0
44	M	*	ATA	150	288	0	Normal	150	Y	3	Flat	1

The grading rubric also mentions changing the data set into a k=5 anonymity dataset. Here would be the output of such a table :

Table 2 : 5-anonimity sample of the dataset

Age	Sex	Race	ChestPainTyp	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngi	Oldpeak	ST_Slope	HeartDisease
44<	*	*	ATA	130	237	0	Normal	170	N	0	Up	0
44<	*	*	ATA	120	284	0	Normal	120	N	0	Up	0
44<	*	*	ATA	120	273	0	Normal	150	N	1.5	Flat	0
44<	*	*	ATA	124	201	0	Normal	164	N	0	Up	0
44<	*	*	ATA	110	208	0	Normal	142	N	0	Up	0
44<	*	*	ATA	136	164	0	ST	99	Y	2	Flat	1
44<	*	*	ASY	100	248	0	Normal	125	N	1	Flat	1
44<	*	*	ASY	140	234	0	Normal	140	Y	1	Flat	1
44<	*	*	ATA	120	184	0	Normal	142	N	1	Flat	0
44<	*	*	ATA	150	288	0	Normal	150	Y	3	Flat	1
<39	*	Non-White	ASY	140	207	0	Normal	130	Y	1.5	Flat	1
<39	*	Non-White	ATA	120	267	0	Normal	160	N	3	Flat	1
<39	*	Non-White	NAP	130	209	0	Normal	178	N	0	Up	0
<39	*	Non-White	NAP	130	211	0	Normal	142	N	0	Up	0
<39	*	Non-White	ASY	110	196	0	Normal	166	N	0	Flat	1
39<x<44	*	Non-White	NAP	115	211	0	ST	137	N	0	Up	0
39<x<44	*	Non-White	ATA	120	201	0	Normal	165	N	0	Up	0
39<x<44	*	Non-White	TA	100	223	0	Normal	142	N	0	Up	0
39<x<44	*	Non-White	ATA	120	204	0	Normal	145	N	0	Up	0
39<x<44	*	Non-White	NAP	130	215	0	Normal	138	N	0	Up	0

We will note here that the larger the k, the more anonymized the quasi-identifiers have to be made. This makes sense in that to uniformize the data in groups of at least 5, we would have to homogenize more of their parameters (simply because it is more and more unlikely as k grows that all samples in a class would share many quasi-identifiable features).
An excel file submitted alongside this assignment contains all the generated k-anonymity tables (with anonymized features and non-anonymized features).

Part 3 : Differential privacy:

4. (See Jupyter Notebook)
5. (See Jupyter Notebook)
6. We can observe that the cross validation accuracy (with 10 folds) of the non-private logistic regression classifier is of about 85% with a standard deviation of 0.04 (which is similar to the results we had from the first assignment). On the other hand, we can also observe that the 10-fold cross validation accuracy of the private logistic regression classifier with epsilon = 1 (default) is of about 69% with a standard deviation of 0.10. It is also worth noting that the single fold accuracy of the private classifier is of 65.9%, which is even lower. With an epsilon of 0.5, the accuracy of the classifier (for a single fold validation) drops to 45.41%. These results are expected since the differential privacy deepens randomness into the data, which decreases the accuracy of the classifier. It is also worth noting that cross validation helps mitigate the effects of the introduction of this randomness to a certain extent.
7. See the graph below :

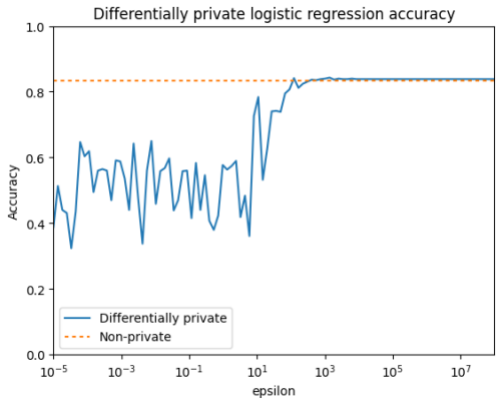

Figure 1 : Differentially private logistic regression accuracy as a function of parameter epsilon

8. From the table above, we observe that the selection of the value of epsilon has a direct impact on the accuracy of the differentially private model. In that sense, we can see that for low values of epsilon, a lot of randomness is introduced to the model which reduces the accuracy of the classifier. On the other hand, we can also see that for values of epsilon = 1 and above, the classifier becomes much more accurate. For relatively high values of epsilon, we can see that the performance of the differentially private classifier is about the same as the non-private one.

   Considering our specific application, we are facing a trade-off. On the one hand, our dataset is not very private. The presence of multiple quasi-identifiers (such as sex, age and race) in the same dataset poses an issue with regards to the privacy of the samples present in the dataset. Because the dataset has a k-anonymity of 1 and a l-diversity of 1, we can clearly state that whoever would consult the dataset would have access to a lot of information about the patients (potentially enough to identify them if completed with further information). This in turn would call for the use of a low value of epsilon to increase randomness in the dataset and limit the possibility for anyone to trace back the identity of someone in the dataset.

   On the other hand, this model's aim is to detect (or not) the presence of a heart disease. This model would be dangerous and not fit for use if it could not accurately (to a certain extent) predict the presence of a heart disease. For example, for a value of epsilon of 0.5, the accuracy of the classifier falls under 50%, which is worse than randomly guessing the presence or not of a heart disease in a patient. Therefore, the model should still achieve for the highest accuracy possible, which would imply the use of a rather large value of epsilon.

   For this specific application, it appears as if a good choice of epsilon would be a value of epsilon between 5 and 10. Selecting the low bound of this interval, I think a good tradeoff would be obtained for a value of epsilon = 5. With a value of epsilon of 5, the 10 fold cross validation accuracy of the model is of about 80% with a standard deviation of 0.05, which is satisfactory while introducing the right amount of randomness in the dataset to protect the privacy and identity of the patients.