ASSIGNMENT 2

PART 1

Gauri Sharma (261026894)

Question 1

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.
   The dataset was created to enable research in the healthcare domain. Given a set of person's health related info like cholesterol level, resting blood pressure, we can predict if someone has a heart disease or not. The dataset was created with the intention of predicting heart disease.

2. Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
   The creators of the dataset are Andras Janosi, M.D. (Hungarian Institute of Cardiology. Budapest), William Steinbrunn, M.D. (University Hospital, Zurich, Switzerland), Matthias Pfisterer, M.D. (University Hospital, Basel, Switzerland), Robert Detrano, M.D., Ph.D. (V.A. Medical Center, Long Beach and Cleveland Clinic Foundation). This dataset was donated to University of California Irvine Data Repository by David W. Aha.

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 health data of different people extracted during medical checkup which point towards whether the person had a heart disease or not. It consists of features like age, sex, gender, heart disease, chest pain type, resting blood pressure. The heart disease prediction is binary.
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.
   All the instances have text or numeric values based on the instance it is describing. The age and cholesterol have numeric data. Whereas, sex, race, chest pain type has test data. Each instance contributes to the binary value of heart disease.
3. Is there a label or target associated with each instance? If so, please provide a description.
   The label is the binary value for heart disease derived from the medical data.

ASSIGNMENT 2

PART 1

Gauri Sharma (261026894)

Question 1

Collection process

1. How was the data associated with each instance acquired?
   The data was acquired during a medical checkup.
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)?
   No information
3. Did the individuals in question consent to the collection and use of their data?
   No information

Preprocessing/cleaning/labeling

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)?
   No information
2. Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.
   No information

Uses

1. Has the dataset been used for any tasks already? If so, please provide a description.
   Yes the data has been used for different tasks and projects. A list of papers that used this dataset can be found at *https://archive.ics.uci.edu/ml/datasets/heart+disease*
2. Are there tasks for which the dataset should not be used? If so, please provide a description.
   This dataset contains instances solely for heart disease prediction. Therefore, the medical data in the dataset shouldn't be used for the prediction of other diseases.

Distribution/maintenance

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?
   The data was donated to UCI repository by David W. Aha.
2. How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
   The donor can be contacted by email and phone at (aha@ics.uci.edu) (714) 856-8779). The UCI repository can be contacted at ml-repository@ics.uci.edu

Question 2

According to the contextual integrity framework, two forms of information flow in this scenario are as follows:

- **Acceptable** - Medical information about the patient, patient's age, patient's sex
- **Unacceptable** - Demographic information (Example: address), financial details (incorporated during payment of bills at hospital), contact information, spouse details

Question 3

- **Identifiers** - No identifiers are present that can help identify the patient directly.
- **Quasi-identifiers** - Age, Sex, Race, RestingBP, Cholestrol, MaxHR.
- **Sensitive Attributes** - HeartDisease

K-anonymity and L-diversity was calculated for various combination of quasi identifiers and identifiers. One of them is given below.

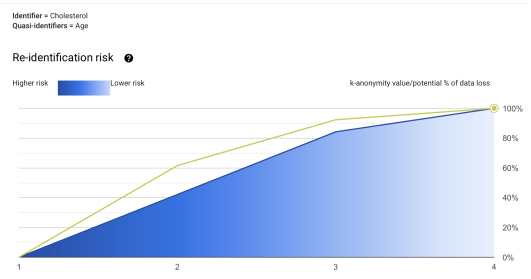

Figure 1: K-Anonymity Plot

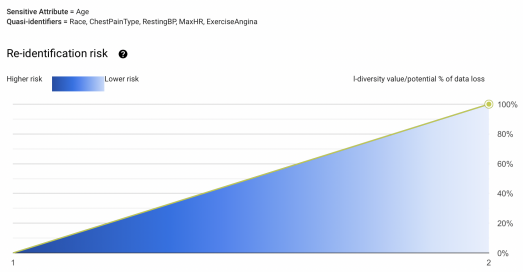

Figure 2: L-Diversity Plot

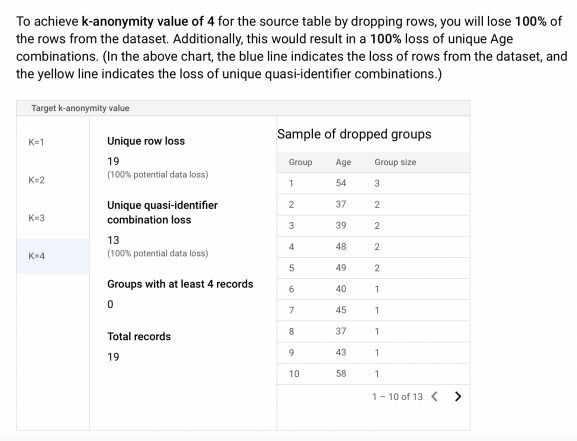

Figure 3: K-anonymity values for 4

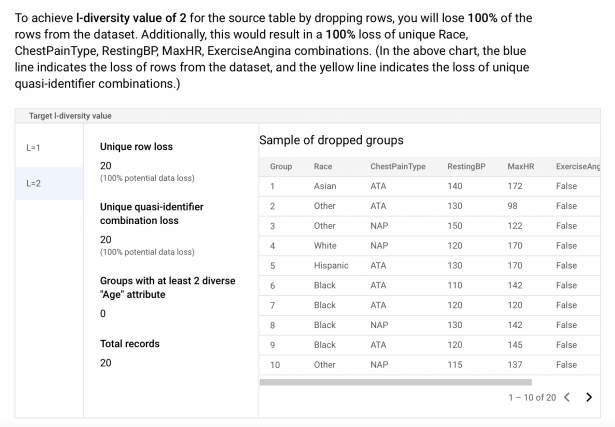

Figure 4: L-diversity values for 2

heart_disease_anonymized

Age	Cholesterol	HeartDisease	count
[28.0]	[132.0]	1	1
[34.0]	[182.0]	0	6
[40.0]	[0.0]	0	1
[40.0]	[0.0]	1	5
[43.0]	[0.0]	0	2
[43.0]	[0.0]	1	4
[39.0]	[147.0]	0	4
[39.0]	[147.0]	1	2
[39.0]	[182.0]	0	3
[39.0]	[182.0]	1	4
[37.0]	[194.0]	0	3
[37.0]	[194.0]	1	3
[39.0]	[199.0]	0	7
[39.0]	[199.0]	1	1
[43.0]	[186.0]	0	7
[43.0]	[186.0]	1	1
[43.0]	[211.0]	0	7
[43.0]	[211.0]	1	1
[48.0]	[0.0]	1	7
[48.0]	[159.0]	0	5
[48.0]	[159.0]	1	3

Figure 5: 5-Anonymous Data

Question 4

```
#Train logistic regression classifier without DP
clf = LogisticRegression(solver="lbfgs")
clf.fit(X_train, y_train)

baseline = clf.score(X_test, y_test)
logreg_y_pred=clf.predict(X_test)
print("Non-private test accuracy: %.2f%%" % (baseline * 100))
print(classification_report(y_test, logreg_y_pred))

cm=confusion_matrix(y_test, logreg_y_pred)
conf_matrix=pd.DataFrame(data=cm, columns=['Predicted:0', 'Predicted:1'] , index=['Actual:0', 'Actual:1'])
plt.figure(figsize = (8,5))
sns.heatmap(conf_matrix, annot=True, fmt='d')
```

Figure 6: Code snippet (from *Assignment2_Part1.ipynb*)

Question 5

```
#Train logistic regression classifier with DP
clf_dp = dp.LogisticRegression()
clf_dp.fit(X_train, y_train)

baseline_dp = clf_dp.score(X_test, y_test)
logreg_y_pred=clf_dp.predict(X_test)
print("Non-private test accuracy: %.2f%%" % (baseline_dp * 100))
print(classification_report(y_test, logreg_y_pred))

cm=confusion_matrix(y_test, logreg_y_pred)
conf_matrix=pd.DataFrame(data=cm, columns=['Predicted:0', 'Predicted:1'] , index=['Actual:0', 'Actual:1'])
plt.figure(figsize = (8,5))
sns.heatmap(conf_matrix, annot=True, fmt='d')
```

Figure 7: Code snippet (from *Assignment2_Part1.ipynb*)

Question 6 and Question 7

```
#Plot accuracy with DP vs non-private accuracy
epsilons, baseline, accuracy = pickle.load(open("lr_accuracy.p", "rb"))

plt.semilogx(epsilons, accuracy, label="Differentially private")
plt.plot(epsilons, np.ones_like(epsilons) * baseline, dashes=[2,2], label="Non-private")
plt.title("Differentially private logistic regression accuracy")
plt.xlabel("epsilon")
plt.ylabel("Accuracy")
plt.ylim(0, 1)
plt.xlim(epsilons[0], epsilons[-1])
plt.legend(loc=3)
plt.show()
```

Figure 8: Code snippet (from *Assignment2_Part1.ipynb*)

Question 8

The value of epsilon that is appropriate for this scenario is 10. As we can see in the figure **??**, the accuracy shifts for both non-differential private (NDP) and differential private (DP) with different values of epsilon. When the value of epsilon reaches 10, the accuracy for both NDP and DP classifiers is the same, which supports our idea of differential privacy.
Although the chosen epsilon value gives us a good accuracy, it is very high, which will reduce the overall security of the model and might lead to data loss.

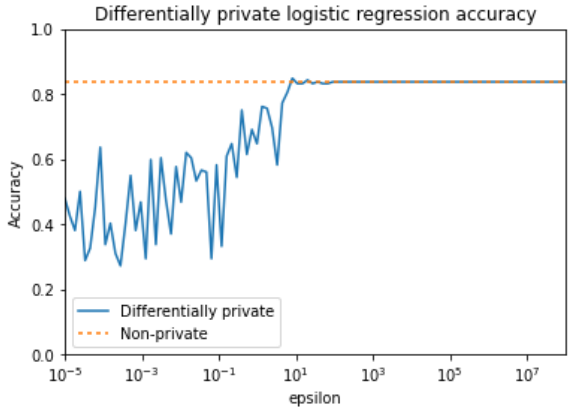

Figure 9: Accuracy vs epsilon plot for NDP and DP Classifiers