NYU CDS, Responsible Data Science, Spring 2023

Homework 2

Gurmehr Sohi, gs3541@nyu.edu

Question 1

Part.a

1. One way racial disparities can arise if historically there was always more arrests made in certain specific community (like Black neighbourhoods). If the historical data-set is already biased against them, it might lead to pre-existing bias and unfair system for further policing.

2. Second we can have racial unfairness due to factors like poverty, race, unemployment and education level which can cause the policing to do more policing towards certain demographic groups (Blacks) and lead to more drug arrests. In this the system can be unfair due to correlations between features.

3. Also racial bias can arise due to the system being created by bias individuals. If the individuals who created the system for prediction are already biased or unfair towards specific categories in features (like being unfair to a race, sex, education category) , it can cause racial disparities.

Part.b

1. One way to achieve fairness in this scenario would be to remove race as an input feature and instead include other variables that are highly predictive of drug activity. Alternatively, the system could be modified to actively seek out and correct for any bias that arises in its predictions. This could be achieved through techniques such as adversarial debiasing.

2. We can take regular feedback of the system usage (how correct the system is predicting) , interviewing police personals about their personal experiences with the usage of system. We can also reduce racial bias by removing

unnecessary features that do not correlate to the drug usage and add new features to the data-set like more information about the neighbourhoods (hospital information regarding cases of drug overdose).

Part.c

1. Firstly the report by Lum and Isaac are very limited as it shows only the data of 2010 and 2011. We need to analyse years before 2010 to get more in-depth dynamics of the arrests made. This is a very small sample to get a proper picture of the actual representation of the drug usage in other regions. Also the findings are only specific to Oakland and therefore cannot be justifiable in other areas. Their findings cannot be used to make prediction system in other regions which have different demographics, social and economic settings.

2. Secondly in their findings we cannot state that racial bias is the only reason for more drug arrests. We need more studies and in-depth analyses of the drug arrest and this might lead to finding more reasons for bias in arrests made in 2011. The paper's analysis and methods were not entirely transparent, making it impossible to assess the correctness and validity of the results reached. The authors, for example, did not provide the particular methodologies used to generate the ratings, which might have influenced the results. Moreover, the authors did not offer a clear explanation of how they chose the variables utilized in the study or why they chose some statistical models over others. This lack of openness calls into doubt the findings validity.

Question 2

Part.a

If any category of individuals have a distribution different than the overall feature distribution, they will face a disadvantage when replacing null values with overall mean value.

In this case, the mean experience years for white females is 7.40 and non-white females is 7.91 are significantly higher than the overall mean experience years which is 6.12.

While we can see that males will get an advantage as their mean is smaller than the overall mean. This would create a bias in our final data set after imputation.

Thus, by imputing the missing values with the overall mean, Alex is assuming that the experience distribution is the same across all demographic groups, which is not the case in this data-set.

Our model would rank all females, especially non white females with the lower ranking.

Part.b

An appropriate imputation method would be to use the mean value for each demographic group separately. This would account for the differences in experience distribution across groups and provide a more accurate representation of an applicant's experience level.

Like for white males it would be better to use the overall mean of all white males and similarly for all other demographic groups.

This would also reduce any bias in the ranking model that can be created during imputation and would create a fair ranking model for all applicants.
This approach would result in a more precise and fair ranking of all applicants, as it accounts for the variation in experience distribution across demographic groups.

Part.c

Here in our case if we use the overall mean to impute values into the null values in our dataset, we can further extend a pre-existing bias that was already there in the dataset. This pre-existing bias is due to the under-representation of females in the dataset.

In our data set which has 4000 total individuals, we just have 934 females and 3066 males which is a huge imbalance and the cause of pre-existing bias. Due to this the overall mean is skewed towards the mean of males. This pre-existing bias is the representation of our world today we live in where women are always under-representated in jobs. Even though the mean experience of females is higher than males we see a biased being introduced during imputation and females being ranked lower.

Emergent bias can occur during model training on imputed data. If the imputed data is already mis-representating the experience levels of individuals, it would incorrectly train the model. The trained model would be biased and this would further create more unfair outcomes for future predictions.

Question 3

Part.a

In class we saw that for this flip coin problem we have the following steps:

1. flip a coin C1
   - if C1 is tails, then respond truthfully
2. if C1 is heads, then flip another coin C2
   - if C2 is heads then Yes
   - else C2 is tails then respond No

1. Truth=Yes by P
2. Response=Yes by A
3. C1=tails by T
4. C1=heads and C2=tails by HT
5. C1=heads and C2=heads by HH

Now we know that a Randomized algorithm M provides ϵ (epsilon is a privacy parameter) if, for all neighboring databases D1 and D2, and for any set of outputs S :

$$Pr[M(D1) \in S] <= e^{\epsilon} Pr[M(D2) \in S]$$

$$Pr[A|P] = Pr[T] + Pr[HH] = 3/4$$

$$Pr[A|\ P] = Pr[HH] = 1/4$$

$$Pr[A|P] = 3Pr[A|\ P]$$

$$\epsilon = ln(3)$$

This shows that is $ln(3)$ - differentially private.

Part.b

In this part we have 2 dices D1 and D2 with faces labelled 1,2,3,4,5,6. The mechanism here is

1. Roll a dice
   - if value is smaller than 4 (1,2,3) we will respond truthfully.

2. Otherwise we will roll dice D2
   - if D2 is smaller than 3 (1,2) we respond yes
   - otherwise we respond no

Probabilities

1. Truth=Yes by P

2. Response=Yes by A

3. D1= smaller than 4 by S

4. D1 = bigger than or equal 4 and D2 = smaller than 3 by BS

5. D1 = bigger than or equal 4 and D2 = bigger than or equal 3 by BB

Now we know that a Randomized algorithm M provides ϵ (epsilon is a privacy parameter) if, for all neighboring databases D1 and D2, and for any set of outputs S :

$$Pr[M(D1) \in S] <= e^{\epsilon} Pr[M(D2) \in S]$$

$$Pr[A|P] = Pr[S] + Pr[BB] = 3/6 + (3/6 * 2/6) = 1/2 + 1/6 = 4/6$$

$$Pr[A|\ P] = Pr[BB] = (3/6 * 2/6) = 1/6$$

$$Pr[A|P] = 4Pr[A|\ P]$$

$$\epsilon = ln(4)$$

This shows that is $ln(4)$ - differential private.

Part.c

Now we know that a Randomized algorithm M provides ϵ (epsilon is a privacy parameter) if, for all neighboring databases D1 and D2, and for any set of outputs S :

$$Pr[M(D1) \in S] <= e^{\epsilon} Pr[M(D2) \in S]$$

$$Pr[BB] = 3/6 * (1/12 + 1/12) = 3/6 * 1/6 = 3/36 = 1/12$$

$$Pr[A|P] = Pr[S] + Pr[BB] = 3/6 + 1/12 = 6 + 1/12 = 7/12$$

$$Pr[A|\ P] = Pr[BB] = 1/12$$

$$Pr[A|P] = 7Pr[A|\ P]$$

$$\epsilon = ln(7)$$

This shows that is $ln(7)$ - differentially private. This is more than the epsilon value obtained in the previous part.b and therefore it would guarantee less privacy to the respondent because lower epsilon value gives stronger privacy.

Question 4

part.a

We apply a classification association rule (CAR) to the given loan attribute in dataset, with sex (M/F) as the sensitive parameter. Sex (M/F) and/or education (HS/BS/MS) are input factors, with loan (Yes/No) as the target variable.

The mandatory input parameter must be the sensitive attribute (M/F), and the target variable must be the loan result (Yes/No). It is also expected that the selected CARs will have support >= 3 and confidence >= 0.6.

We acquire the list of CARs given below after selecting all of the probable CARs based on the constraints described above and filtering them based on the support >= 3 and confidence >= 0.6 as shown in Table A.

The frequency of the provided item set in the given data set is referred to as support. Confidence for a specific item set is computed as the ratio of support to total support for the given item set, regardless of the target variable.

1. Confidence (M, Yes) = 11/16 = 0.6875
2. Confidence (F, No) = 11/16 = 0.6875
3. Confidence (F, HS, No) = 4/4 = 1
4. Confidence (M, BS, Yes) = 4/6 = 0.66
5. Confidence (M, MS, Yes) = 4/4 = 1

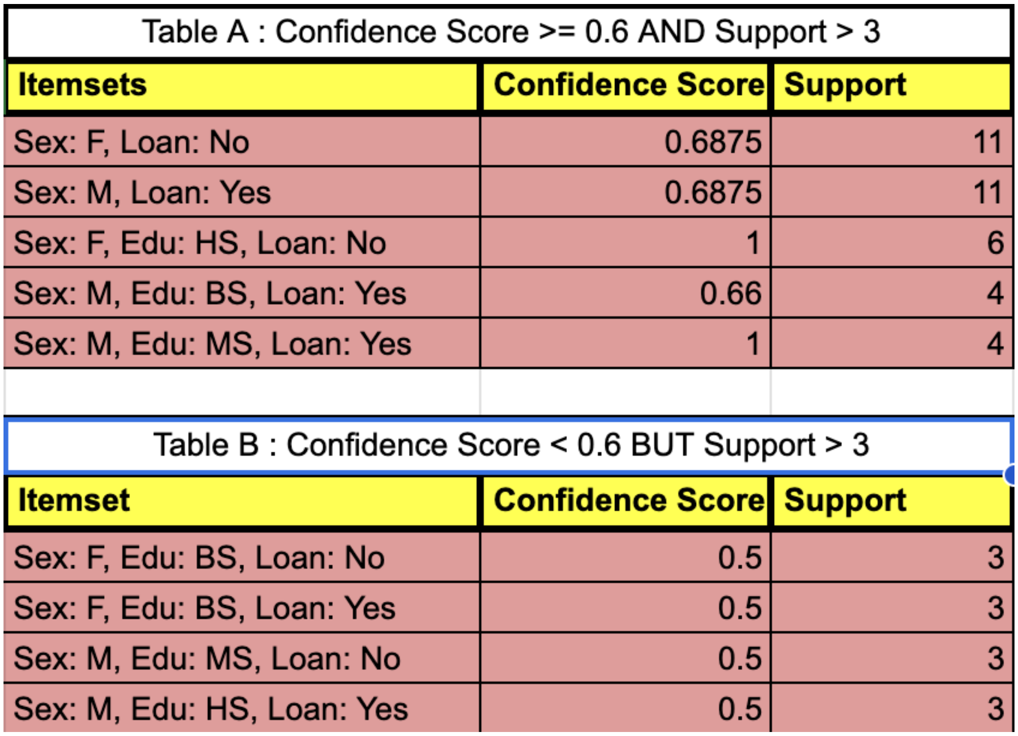


Table A : Confidence Score >= 0.6 AND Support > 3

Itemsets	Confidence Score	Support
Sex: F, Loan: No	0.6875	11
Sex: M, Loan: Yes	0.6875	11
Sex: F, Edu: HS, Loan: No	1	6
Sex: M, Edu: BS, Loan: Yes	0.66	4
Sex: M, Edu: MS, Loan: Yes	1	4

Table B : Confidence Score < 0.6 BUT Support > 3

Itemset	Confidence Score	Support
Sex: F, Edu: BS, Loan: No	0.5	3
Sex: F, Edu: BS, Loan: Yes	0.5	3
Sex: M, Edu: MS, Loan: No	0.5	3
Sex: M, Edu: HS, Loan: Yes	0.5	3

Figure 1: Table A shows the CARS Support>= 3 Confidence Score > 0.6

part.b

From Problem 4A, we have computed the following CARs, as shown in figure 1 Table A,
We can combine sequential and parallel composition to distribute portions of the privacy budget ($\epsilon = 1$) to each frequent itemset while maximizing the utility of the information released.

To perform parallel composition, group the confidential association rules (CARs) based on the sensitive attributes, such as sex in this case. Since the sensitive attribute is a mandatory field, the groups can be considered as separate, allowing for parallel composition. Distribute the privacy budget evenly between the groups based on the number of sensitive attribute values. Since there are two groups based on sex, namely M and F, we can assign each group an equal share of the privacy budget, which is ($\epsilon = 1/2 = 0.5$).
We will now distribute the privacy budget among the frequently used itemsets. To achieve this, we will combine sequential and parallel composition to allocate portions of the privacy budget ($\epsilon = 1$) : to each itemset while ensuring maximum utility of the released information. To begin, we will group the frequent itemsets based on sensitive attributes (such as sex) for parallel composition. Since the sensitive attribute is mandatory, the groups can be considered disjoint, and we can allocate the privacy budget evenly among the groups. For instance, in our case, we have two groups based on sex (male and female), so we can assign ($\epsilon/2 = 0.5$) to each group. Within each group, we will use sequential composition to allocate the privacy budget to each frequent itemset. To ensure maximum utility, we can assign the budget proportionally to the support of the itemsets within each group. Finally, we will allocate the privacy budget for each frequent itemset.
The total male support group = 19
The total female support group = 17

For the Male group with $\epsilon = 0.5$, we have the following attribute combinations:

1. sex=M, loan=yes, support=11
2. sex=M, edu=BS, loan=yes, support=4
3. sex=M, edu=MS, loan=yes, support=4

The privacy budget allocation for each frequent itemset in the Male group is as follows:

- $\epsilon = 0.5 * (11/19) = 0.290$.
- $\epsilon = 0.5 * (4/19) = 0.11$.
- $\epsilon = 0.5 * (4/19) = 0.11$.

For the Group Female ($\epsilon = 0.5$) :

1. For the Female group with attributes edu=HS, sex=F, and loan=no, the support is 6.
2. For the Female group with attributes sex=F and loan=no, the support is 11.

The privacy budget allocation for Female group :

- $\epsilon = 0.5 * (11/17) = 0.32$.
- $\epsilon = 0.5 * (6/17) = 0.18$.

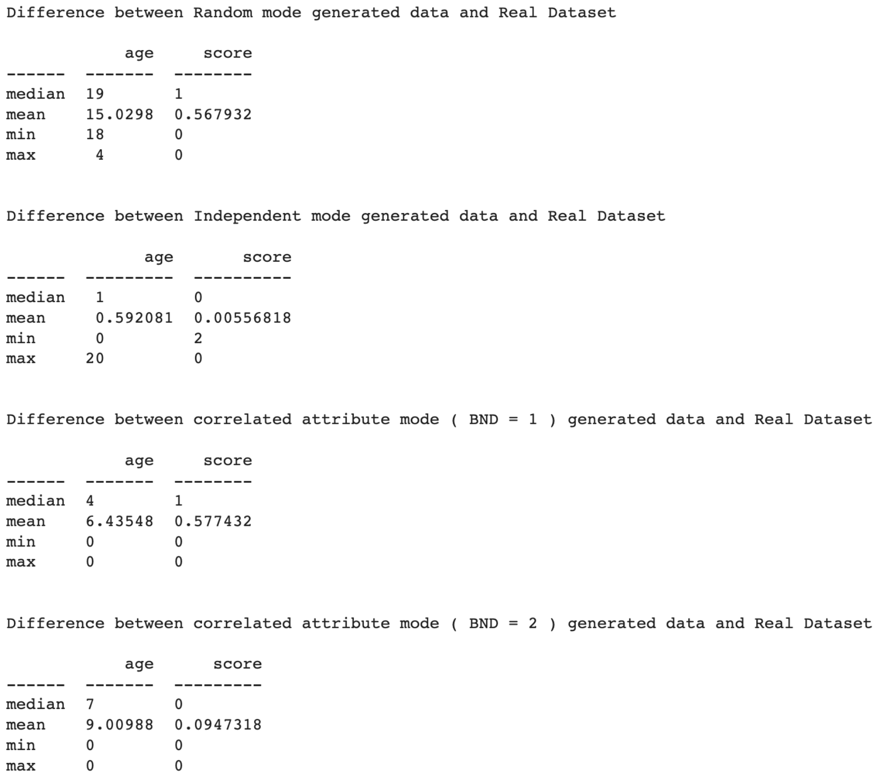
Finally we can see the results for all the itemsets

1. sex:Female , Loan:No $\epsilon = 0.32$
2. sex:Male , Loan:Yes $\epsilon = 0.29$
3. sex:Female ,Education:HS Loan:No $\epsilon = 0.18$
4. sex:Male ,Education:BS, Loan:Yes $\epsilon = 0.11$
5. sex:Male ,Education:MS, Loan:Yes $\epsilon = 0.11$

Question 5

Part.a

Q1



```
Difference between Random mode generated data and Real Dataset

            age     score
------  -------  --------
median  19       1
mean    15.0298  0.567932
min     18       0
max      4       0


Difference between Independent mode generated data and Real Dataset

              age       score
------  ---------  ----------
median   1         0
mean     0.592081  0.00556818
min      0         2
max     20         0


Difference between correlated attribute mode ( BND = 1 ) generated data and Real Dataset

            age     score
------  -------  --------
median  4        1
mean    6.43548  0.577432
min     0        0
max     0        0


Difference between correlated attribute mode ( BND = 2 ) generated data and Real Dataset

            age      score
------  -------  ---------
median  7        0
mean    9.00988  0.0947318
min     0        0
max     0        0
```

Figure 2: part.a - synthetic data (difference in mean, median, min, max)

If we compare the mean median difference between synthetic generated data and real data, we see that in independent mode there is minimum difference between synthetic data and real data. While we can say that random mode performs the worst.

Looking at the age distribution of all the synthetic datasets and comparing it to the real dataset we can say that the Random mode gives the worst results because it gives each value equal probability, while independent mode and correlated modes give quite similar results. Here it can be that the 'age' attribute can be a little or not correlated at all. That is why even after correlation we dont see a lot of improvement.

Looking at the 'score' distribution of all the synthetic datasets and comparing it to the real dataset we can say that the Random mode gives the worst results because it gives each value equal probability, while independent mode provides the best results with comparable distribution. Here the correlated datasets are not showing good distribution results. This can be because the attribute 'score' is not correlated to any other attribute and therefore treating it as an independent column and generating new data is a better method.

Q2

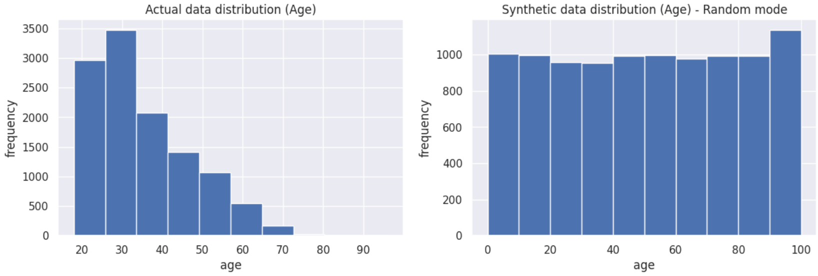


Figure 3: part.b - Age distribution in Random mode vs Actual data

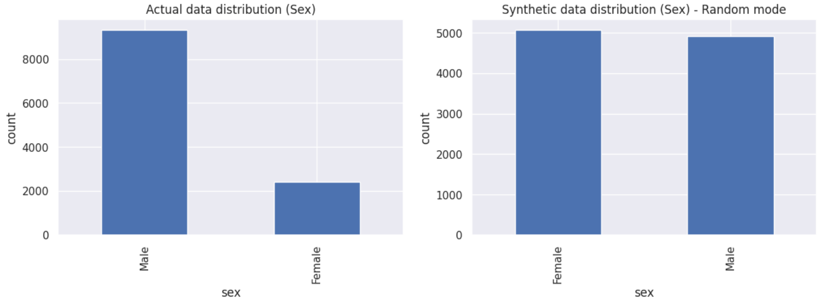


Figure 4: part.b - sex distribution in Random mode vs Actual data

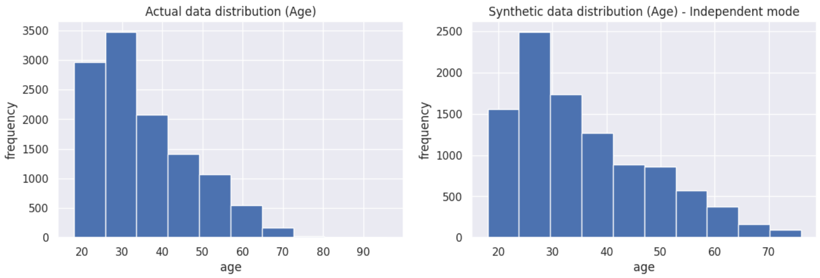


Figure 5: part.b - age distribution in Independent mode vs Actual data

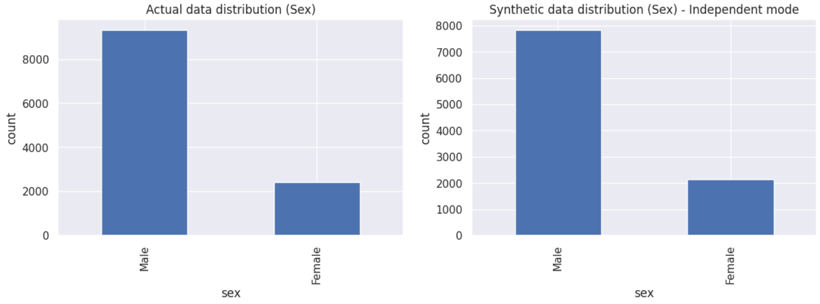
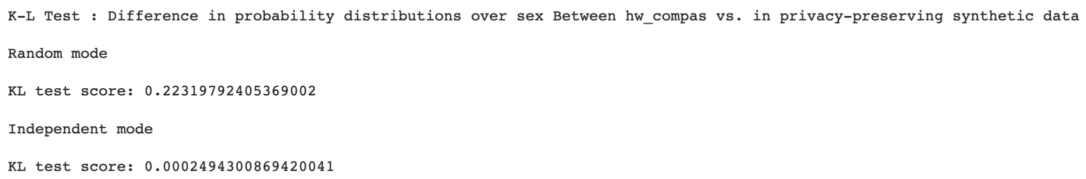


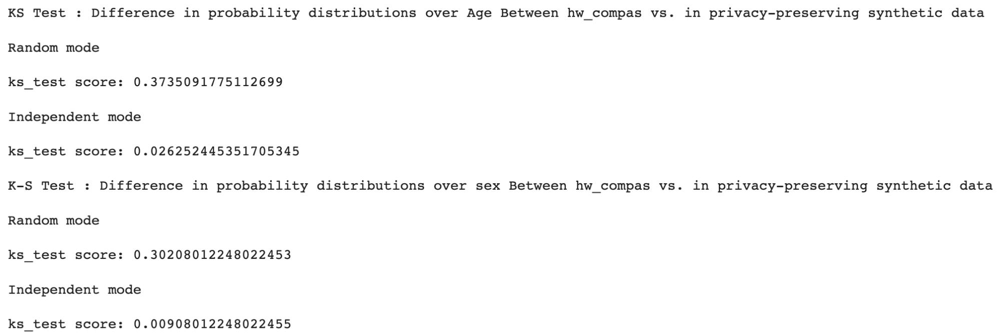
Figure 6: part.b - sex distribution in Independent mode vs Actual data

We can clearly see in the figure of age and sex distribution that in independent we see better results as compared to the random mode as it gives all categories of a feature equal probability while in independent mode we consider all the feature independently and not correlated to any other feature. Independent mode gives us better results as compared to the random mode.



```
K-L Test : Difference in probability distributions over sex Between hw_compas vs. in privacy-preserving synthetic data

Random mode

KL test score: 0.22319792405369002

Independent mode

KL test score: 0.0002494300869420041
```

Figure 7: K-L test results



```
KS Test : Difference in probability distributions over Age Between hw_compas vs. in privacy-preserving synthetic data

Random mode

ks_test score: 0.3735091775112699

Independent mode

ks_test score: 0.026252445351705345

K-S Test : Difference in probability distributions over sex Between hw_compas vs. in privacy-preserving synthetic data

Random mode

ks_test score: 0.30208012248022453

Independent mode

ks_test score: 0.00908012248022455
```

Figure 8: K-S test

For the KL Test we can see that the value of independent mode is very less as compared to the value of random mode for the attribute 'sex'. This is because the independent mode performs much better and is quite similar to actual dataset. Its probability distribution is similar to the privacy-preserving synthetic data. Also the KL test only works on continuous

For the KS Test we can see that the value of independent mode is very less as compared to the value of random mode for the attribute 'sex' and 'race' both. We know from the K-S test that if the value of the Kolmogorov-Smirnov test statistic is large then it suggests that there is higher difference between the distributions of both the datasets. So here the random mode has higher K-S test value as compared to independent mode and therefore the independent mode has better probability distributions.

Q3

We can use mutual information to further understand how the relationships between features are similar/different in the real data and the synthetic data. Mutual information is defined as follows for two discrete variables X and Y:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log(\frac{p(x,y)}{p(x)p(y)})$$

Higher values indicate greater levels of mutual information. For two independent variables, the value will be zero (look at the logged term). This metric works for categorical variables or continuous variables.

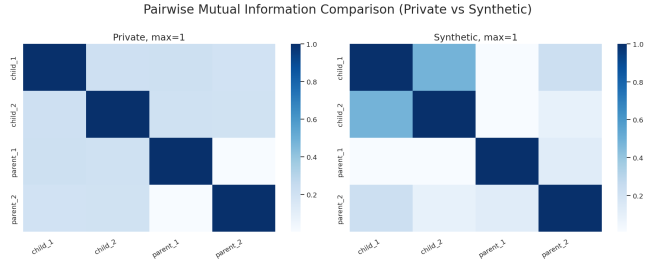


Figure 9: CASE-C : CORRELATED ATTRIBUTE MODE WITH K=1

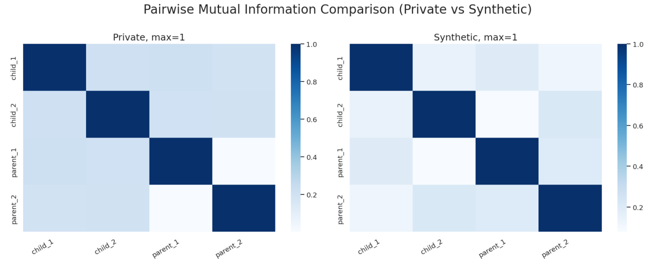


Figure 10: CASE-D : CORRELATED ATTRIBUTE MODE WITH K=2

In the above pairwise mutual information we can see that when we compare both the heatmaps we see that Case-D performs much better as compare to Case-C. The heatmap for the synthetic dataset generated using correlated attribute mode with K=1 is similar to the heatmap for the real dataset, this suggests that the synthetic dataset has similar statistical properties to the real dataset. Conversely, the heatmap generated using correlate attribute mode with K=2 and real datasets are very different, this could indicate that this synthetic dataset does not accurately capture the statistical properties of the real dataset.

part.b

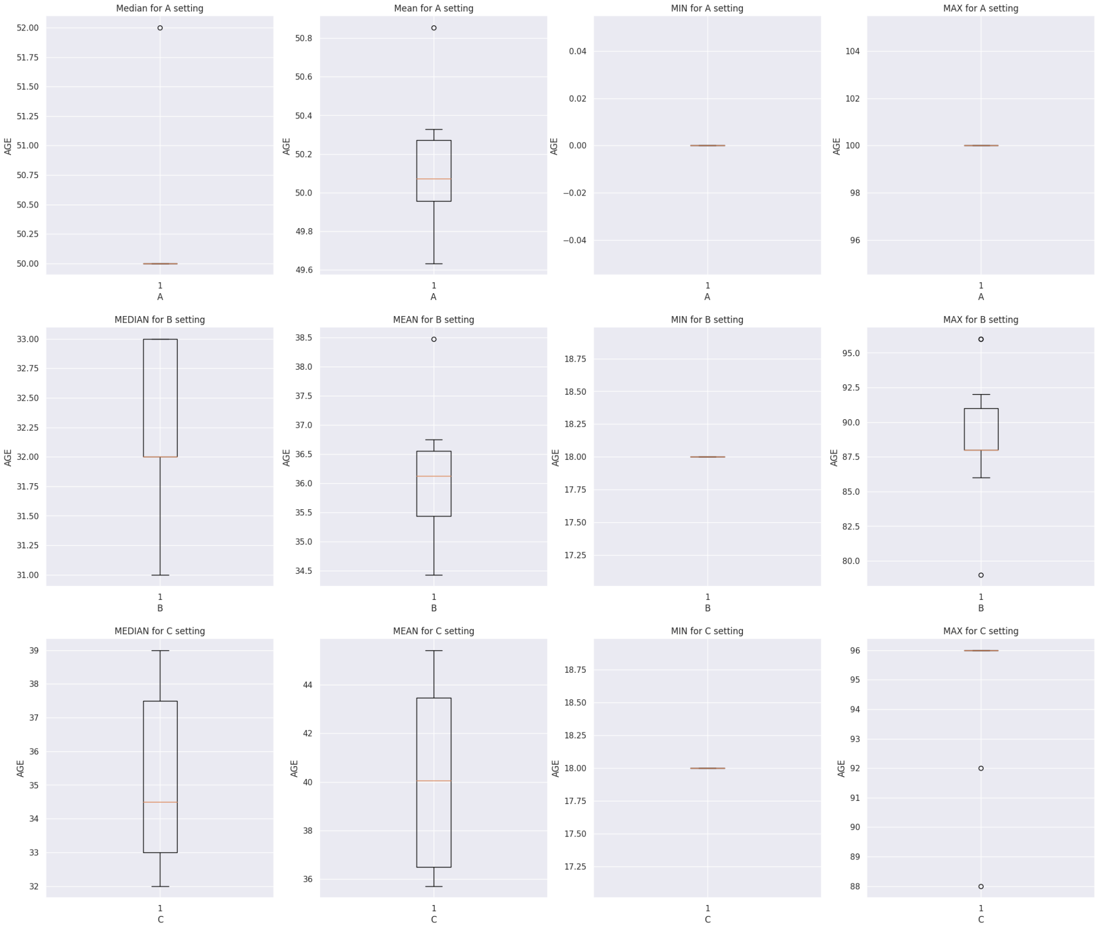


Figure 11: Mean Median Min Max Subplots with different setting (A,B,C)

1. Real data Median: 32.0
2. Real data Mean: 35.14
3. Real data Min: 18
4. Real data Max: 96

Here we observe that the setting B which is B: independent attribute mode with epsilon = 0.1. gives us the best results as the min and max value are similar to the real data and the median and mean also show same results when we compare it to the actual dataset from which these synthetic datsets are created.

So we can say that setting B shows the true results which are similar to the real dataset.

This can be because the features might have less correlation between them and taking them independently while creating the synthetic data is the best option. Also we see that the setting C which is correlated attribute mode with epsilon = 0.1 shows some good results which are better than random mode (A)

part.c

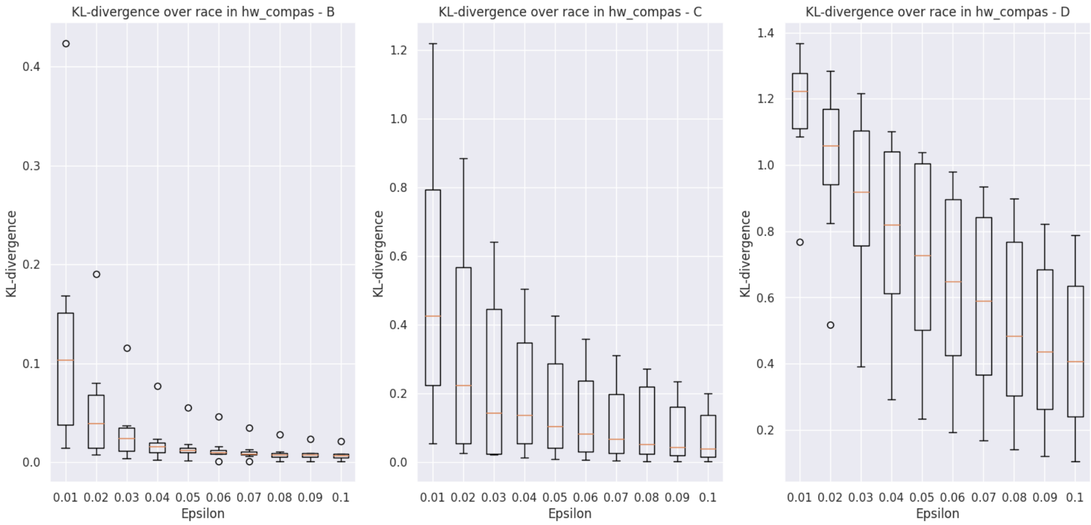


Figure 12: KL-divergence over the attribute race in $hw - compass$

When we talk about better results in the context of KL-divergence, it generally means that a smaller KL-divergence value is better. This is because a smaller KL-divergence indicates that the estimated probability distribution is closer to the true probability distribution. A KL-divergence of 0 indicates that the two distributions are identical.

In summary, our experiments show that as the privacy budget (epsilon) increases, the synthetic datasets better preserve the statistical properties of the sensitive datasets. Additionally, the correlated attribute mode generally outperforms the independent attribute mode in terms of preserving statistical properties and mutual information between attributes. However, in our case we see that the independent mode performs better and has lower kl divergence value and also improves as we increase the epsilon value further.

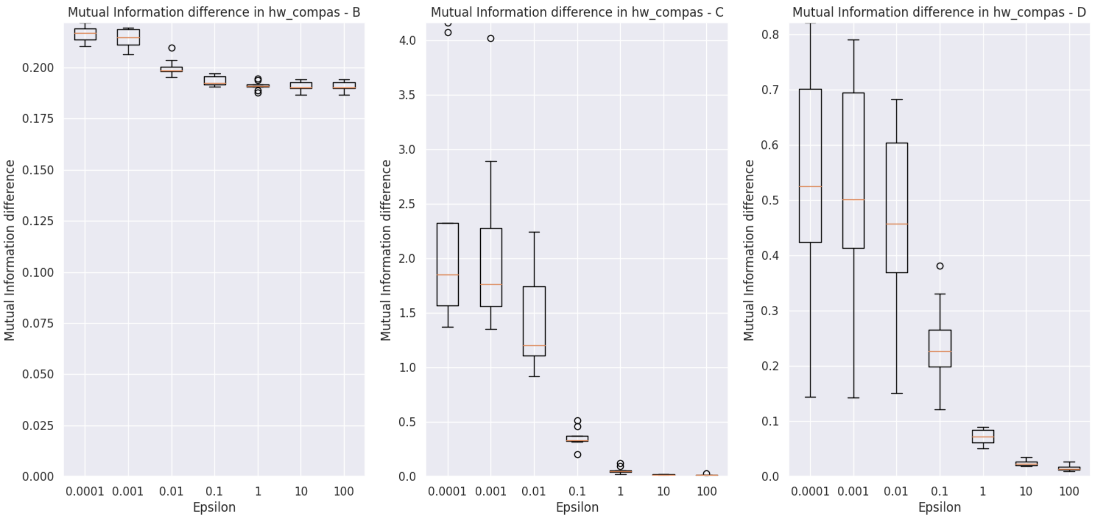


Figure 13: Aggregate difference in pairwise mutual information ($hw - compass$)

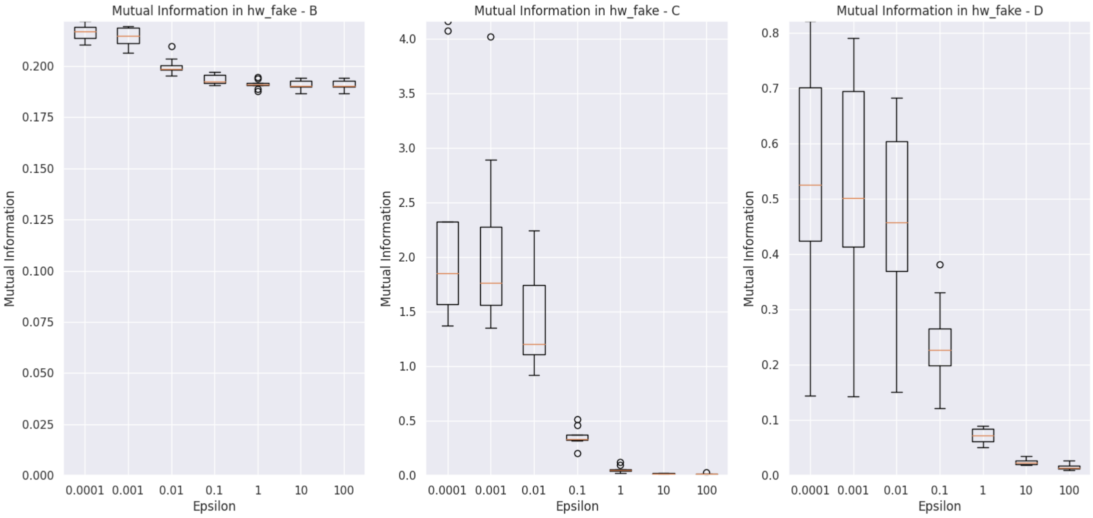


Figure 14: Aggregate difference in pairwise mutual information ($hw - fake$)

In summary for the aggregate difference in pairwise mutual information for our datasets, our experiments show that as the privacy budget (epsilon) increases, the synthetic datasets better preserve the statistical properties of the sensitive datasets. The aggregate difference in pairwise mutual information value decrease as we increase epsilon. This shows that the datasets are better preserving the properties and information of the actual private dataset. In our case we see that the Correlated mode correlated attribute mode with epsilon = 0.1, with Bayesian network degree k=1 performs better and has lower aggregate difference in pairwise mutual information and also improves as we increase the epsilon value further.

part.d

To compute the differentially private spanning tree for hw-fake with epsilon = 0.1, we utilize the Private library. The generated spanning tree depicts the links between the dataset's various properties.

We utilize the "correlated attribute mode with k=2" setting with epsilon = 0.1 for Data Synthesizer to construct a Bayesian network that models the associations between distinct qualities in the dataset.

When we compare the two models, we can observe that the MST spanning tree offers a comparable perspective of the connections between the characteristics in the dataset as the Data Synthesizer Bayesian network. Although the particular relationships reflected by each model may differ, both models capture the correlations between distinct qualities.

MST chooses the most informative 2-way marginals based on the mutual information between any two attributes, whereas DataSynthesizer calculates the conditional probability tables for each attribute given its parents in the Bayesian network. The information acquired by MST's marginals is comparable to the information captured by DataSynthesizer's conditional tables, since they both capture the correlations between distinct qualities.

Therefore, while the approaches employed to model the dataset by DataSynthesizer and MST differ, both models capture the correlations between different variables and may be used to produce synthetic datasets that retain these correlations while safeguarding the privacy of the individuals in the dataset.

part.e

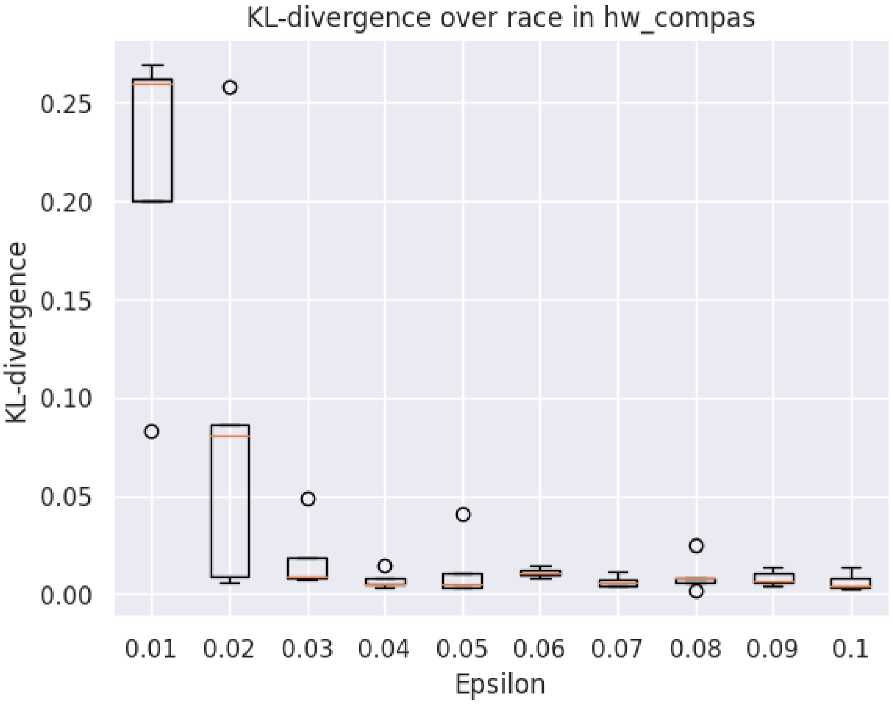


Figure 15: KL-divergence over race in hw-compass generated using MST

MST and Data Synthesizer are two separate privacy-preserving data synthesizers, each with its own set of advantages and disadvantages.

MST is a method for extracting ”useful” patterns from sensitive material while remaining anonymous. MST uses a tree-based structure to divide the data into subgroups, then makes the data anonymous inside each subgroup. MST is particularly useful in applications that want to extract patterns or rules from sensitive data, such as medical diagnosis or fraudulent actions.

Data Synthesizer, on the other hand, is a method for generating synthetic data that is statistically similar to the original data, but with added privacy guarantees. Data Synthesizer generates synthetic data by modeling the distribution of the sensitive data and then using that model to generate new data points. The resulting data can be used for many of the same purposes as the original data, but without revealing sensitive information.

We can see that the aggregate difference in pairwise mutual information for our datasets decrease with increase in epsilon except for the hw-compass where it suddenly increase for epsilon=0.1 and then decrease. For both the hw-compass and hw-fake we see that the value performs really good for epsilon values 10 and 100. These values show the best results.

While in Kl-divergence over race plot we see that it performs much better to the data synthesizer. Comparing it to the Data synthesizer generated dataset using setting D we see that here the value of kl-divergence starts decreasing significantly for 0.03 epsilon value and keeps decreasing as we increase the epsilon.

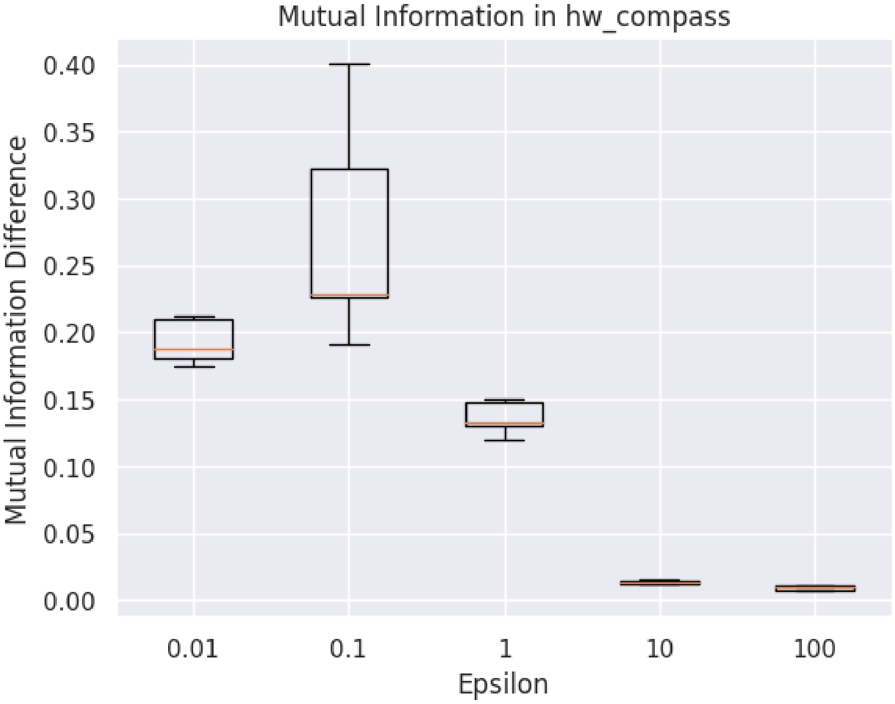


Figure 16: Aggregate difference in pairwise mutual information hw-compass

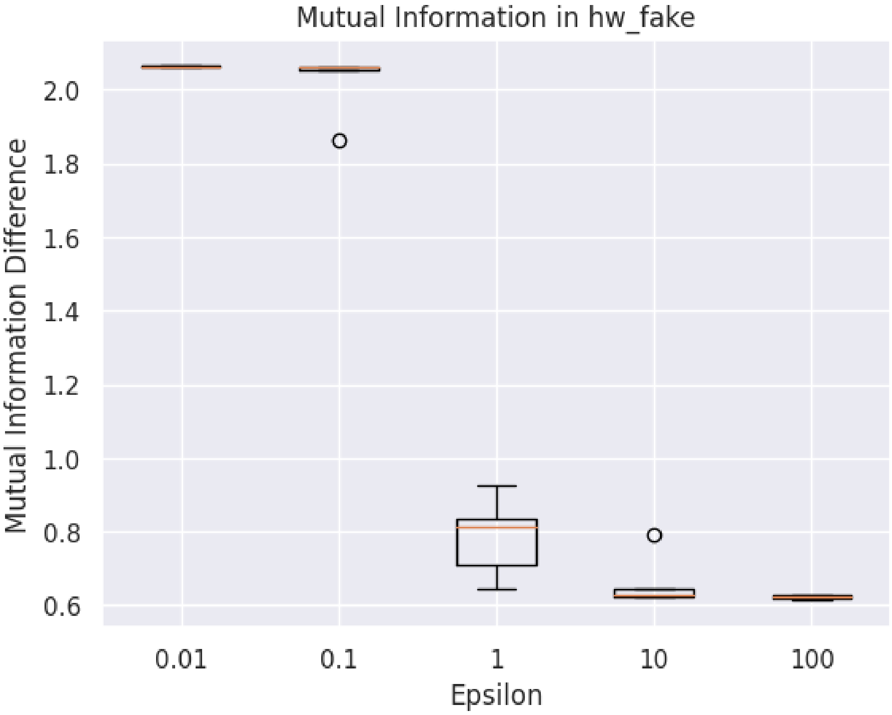


Figure 17: Aggregate difference in pairwise mutual information hw-fake