Kieren Singh Gill, ksg7699
Responsible Data Science
Professor George Wood
Homework #2

Problem 1 (10 points): Racial disparities in predictive policing

1a) (5 points) Give three distinct reasons why racial disparities might arise in the predictions of such a system.

Reason 1: Pre-existing bias

- Fig 1(b) shows that the estimated number of drug users were spread out throughout Oakland, whereas Fig 1(a) shows that arrests were made in a small concentrated area in black neighbourhoods. If this was because police offers were prejudiced against black people, this pre-existing bias will be reflected in the dataset of drug-related arrests. If the algorithm is trained with past criminal data from the police department, this pre-existing bias will influence the machine learning system, and it might lead the system to predict more drug-related crimes in black neighborhoods as well.

Reason 2: Technical bias

- As stated in the research paper "Data distribution debugging in machine learning pipelines", data distribution bugs are often introduced during preprocessing, which can introduce a skew in the data. In particular, it can exacerbate the under-representation of historically disadvantaged groups. In this case, this would mean that preprocessing could skew the data so that the algorithm will target black people, and this would be a form of technical bias.

Reason 3: Emergent bias

- If the predictive policing algorithm targets black neighbourhoods, this means that the police will allocate their resources towards policing these areas more, which would lead to an increase in arrests within this specific concentration. This will create a feedback loop, as new disproportionate data will be fed into this algorithm, and the police will continue to use this as a justification for excessive policing of black neighborhoods. This is an example of emergent bias, and it will continue to perpetuate the racial divide.

1b) (5 points) Propose two mitigation strategies to counteract racial disparities in the predictions of such a system. Note: It is insufficient to state that we could use a specific pre, in, or post-processing technique that we covered in class when we discussed fairness in classification. Additional details are needed to demonstrate your understanding of how the ideas from fairness in classification would translate to this scenario.

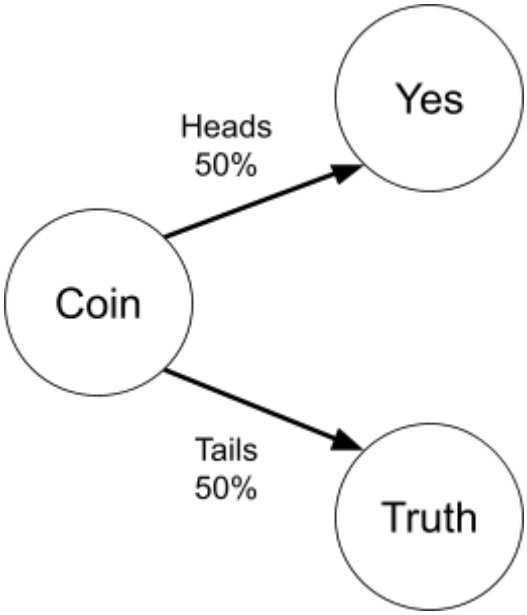
Strategy 1:

- There should be a limit on the number of stationary police vehicles there can be in black neighborhoods specifically. This mitigation strategy will work specifically for the emergent bias problem described in question 1a), as the police will not be allowed to allocate most of their resources towards policing black neighborhoods. This should reduce the amount of overpolicing in those areas, and will hopefully result in arrest data that is more representative of the estimated number of drug users as shown in Fig 1(b).

Strategy 2:

- We can use AIF360 and set the protected attribute to race. Then, we can use AIF360 to create a disparate impact baseline for our machine learning model, and as we continue to train our model with incoming data, we can use AIF360's DisparateImpactRemover if the disparate impact value exceeds the baseline. This will ensure that our algorithm is mitigating the biases within the incoming data, and it will prevent the algorithm from creating skewed predictions.

Problem 2 (15 points): The simplest version of randomized response involves flipping a single fair coin (50% probability of heads and 50% probability of tails). Suppose an individual is asked a potentially incriminating question and flips a coin before answering. If the coin comes up tails, he answers truthfully, otherwise, he answers "yes". Is this mechanism differentially private? If so, what epsilon value does it achieve? Carefully justify your answer.



In the mechanism described in the question, individuals can only answer truthfully if they flip tails - otherwise, their response would be a "yes". There is no way for an individual to answer "no" outside of telling the truth because there is no randomization mechanism by which they could answer "no". This means that these individuals are not given plausible deniability, hence, individuals who answer "no" can be removed from the possibility of being incriminated, and they can be easily distinguished from the rest of the responses in the dataset because we know for certain that they are telling the truth. For an individual to respond "yes", they are either telling the truth, or they flipped a heads. While there is still some degree of plausible deniability in this scenario for the people who answered truthfully, this degree is reduced because we can still definitively distinguish them from the people who answered "no". Hence, their privacy is still not fully protected.

An algorithm is differentially private if an outside observer seeing the algorithm's output cannot distinguish which individual's information was used in the computation. This is done by creating a neighboring dataset with random variables, which gives plausible deniability to individuals by acting as

noise that obscures the identity of the individuals in the database. It is only differential privacy if it pertains to the computation of the data and its outputs - as discussed in lecture, aggregation without randomization and de-identification is insufficient due to the presence of auxiliary information. Because the mechanism designed in the question doesn't allow plausible deniability for all participating individuals, it is not differentially private.

(a) (15 points): Execute the following queries on synthetic datasets and compare the results to those on the corresponding real datasets:

Q1 (**hw_compas** only): Execute basic statistical queries over synthetic datasets.

Data	Column	Median	Mean	Min	Max
Ground Truth	Age	32	35.1433	18	96
	Score	4	4.37127	-1	10
A	Age	51	50.1731	0	100
	Score	5	4.9392	-1	10
B	Age	33	35.7354	18	76
	Score	4	4.3657	1	10
C	Age	36	41.5788	18	96
	Score	5	4.9487	-1	10
D	Age	39	44.1532	18	96
	Score	4	4.466	-1	10

Fig. 1

Figure 1 shows the median, mean, minimum and maximum values in all four datasets - the Ground Truth dataset which represents the real world values, dataset A (random), dataset B (independent attribute), dataset C (correlated attribute with a Bayes net degree of k=1) and dataset D (correlated attribute with a Bayes net degree of k=1).

When comparing the ground truth values with the values generated by the synthesizer, the random mode sythesizer is the worst with a mean of 50 for age when compared to the ground truth mean of 35 for age. This makes sense because the random mode synthesizer generates completely random values, whereas the other synthesizers are based on attributes from the ground truth dataset. The independent attribute mode performed the best, with the closest values to the ground truth values for both age and score. This makes sense because the independent attribute mode adds noise to a histogram of values that are derived from each attribute. This keeps the values as close to the ground truth values as possible, which is why

it is the most accurate when it comes to the general representation of the ground truth dataset. The correlated attribute mode performed the best at representing the extremes (maximum and minimum) of the ground truth dataset, but was not as accurate when it came to representing the median and mean. It is interesting to note that between both the correlated attribute modes, the one with a Bayes net degree of k=1 represented the mean and median for age better, and the one with a Bayes net degree of k=2 represented the mean and median for score better.

Q2 (hw_compas only): Compare how well random mode (A) and independent attribute mode (B) replicate the original distribution.

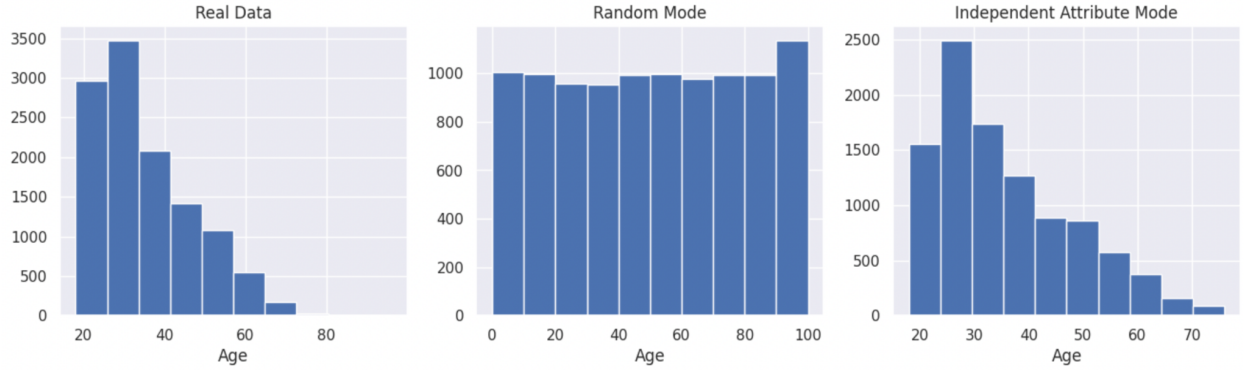


Fig. 2

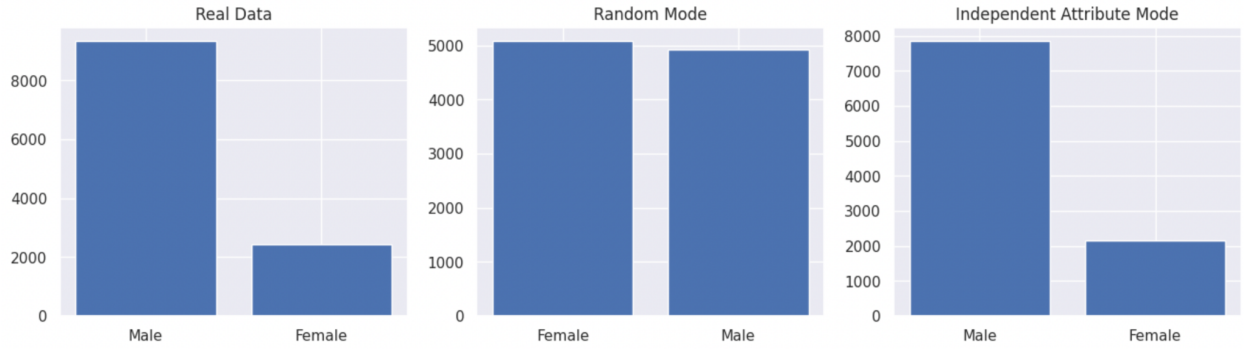


Fig. 3

From observing Figures 2 and 3, it is evident that the ground truth values for both age and sex are much better represented by the dataset created with independent attribute mode as compared to the datsaet created with random mode on the DataSynthesizer. It makes sense that the random mode creates a dataset with a nearly uniform distribution, because the values in the dataset are randomly generated. This is good when handling sensitive data that needs to be protected, but there is a large tradeoff between protection and accurate representation of ground truth values. In this case, although the data would be well protected, we would not be able to learn much from it because it is not representative of the actual population. On the other hand, independent attribute mode accurately represents ground truth, which is useful for us because we can use this to learn about the population. However, if we were handling sensitive data, the data would not be well protected whatsoever.

```
klTest1 = kl_test(df_real, df_real_A, "sex")
print(klTest1)

klTest2 = kl_test(df_real, df_real_B, "sex")
print(klTest2)
```

```
0.22319792405369002
0.0002494300869420041
```

```
ksTest1 = ks_test(df_real, df_real_A, "age")
print(ksTest1)

ksTest2 = ks_test(df_real, df_real_B, "age")
print(ksTest2)
```

```
0.3735091775112699
0.026252445351705345
```

Fig. 4

Figure 4 shows the results of the KS test, which can tell us the similarities between the distributions of the quantitive real and synthetic data. Meanwhile, KL test tells us the similarities between the distributions of the categorical real and synthetic data. The higher the test score, the less similar these datasets are to the ground truth values. In both cases, for sex and age, the independent attribute mode has a significantly lower score when compared to the random mode score. Once again, this shows the tradeoff between accuracy and protection of data - the random mode offers more protection but the independent attribute offers more protection.

Q3 (**hw_fake** only): Compare the accuracy of correlated attribute mode with k=1 (C) and with k=2 (D).

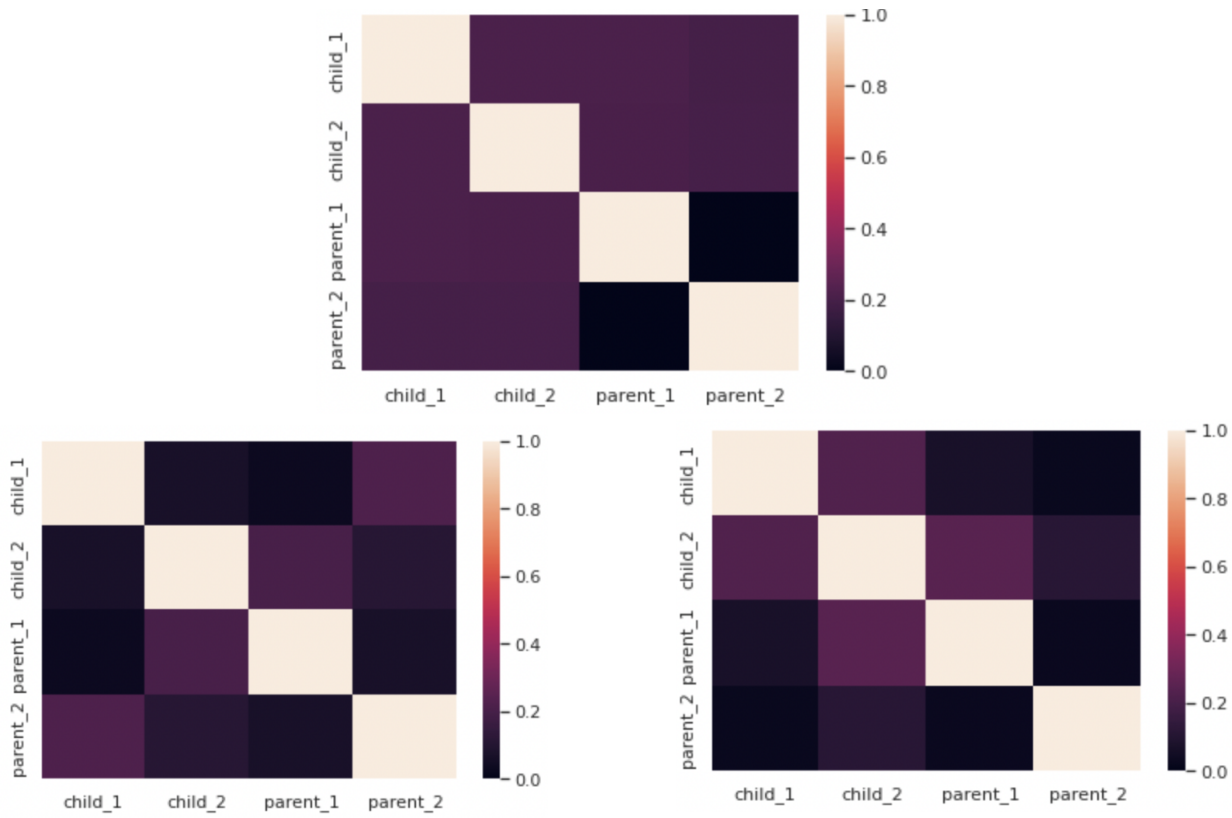


Fig. 5 (pairwise_fake, pairwise_fake_C, pairwise_fake_D)

The correlated attribute mode constructs Bayesian networks to model the attribute values. The k value represents the maximum number of BN node parents. The synthetic data generated under C has a higher correlation with hw_fake than the synthetic data generated under D. We can tell that this is the case because the colors in the heatmap are most similar between C and hw_fake than D and hw_fake. This means the mutual information is better preserved in C than it is in D. However, generally, we can say that a fair amount of mutual information is lost in both C and D - we can observe that around the edges of both heatmaps, certain areas that had correlations of roughly 0.3 have dropped to near 0.

(b) (10 points, hw_compas only**)**: Study the variability in the **mean** and **median** of **age** for synthetic datasets generated under settings A, B, and C.

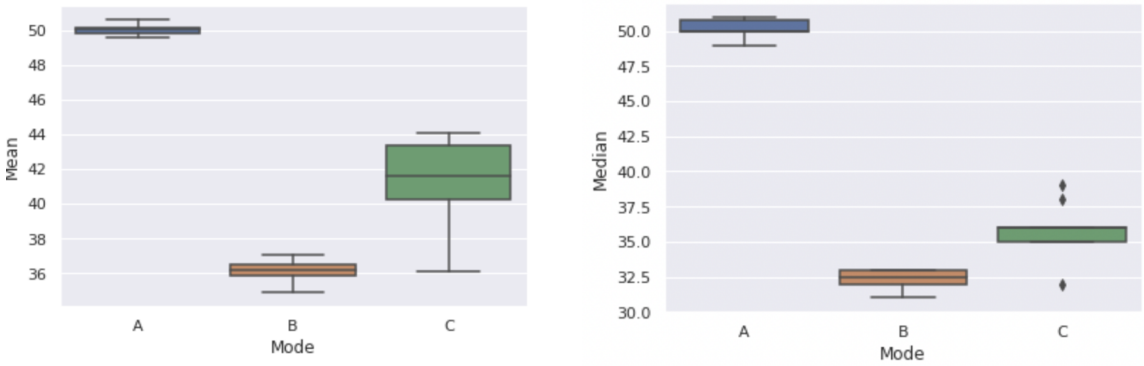


Fig 6

The box and whiskers plots in figure 6 show the distributions of statistical measures across synthetic datasets in all three modes. The random mode stays consistently around 50 since the DataSynthesizer chooses random values in the statistical range. The independent attribute mode is the most accurate in comparison to the ground truth values. This is good when it comes to representing the original dataset. However, there is a tradeoff between accuracy and privacy. The more accurate the synthetic dataset, the less protection it gives to the individuals in the dataset. In the correlated attribute mode, which constructs a Bayesian network at degree k=1, the mean is distributed over a larger range and is less accurate than the independent attribute mode. In addition, the median is also not as accurate as independent attribute mode. This means correlated attribute mode provides more privacy than independent attribute mode and more accuracy than random mode.

(c) (10 points, hw_compas only**)**: Study how well statistical properties of the data are preserved as a function of the privacy budget, epsilon. To see robust results, execute your experiment with 10 different synthetic datasets (with different seeds) for each value of epsilon, for each data generation setting (B, C, and D).

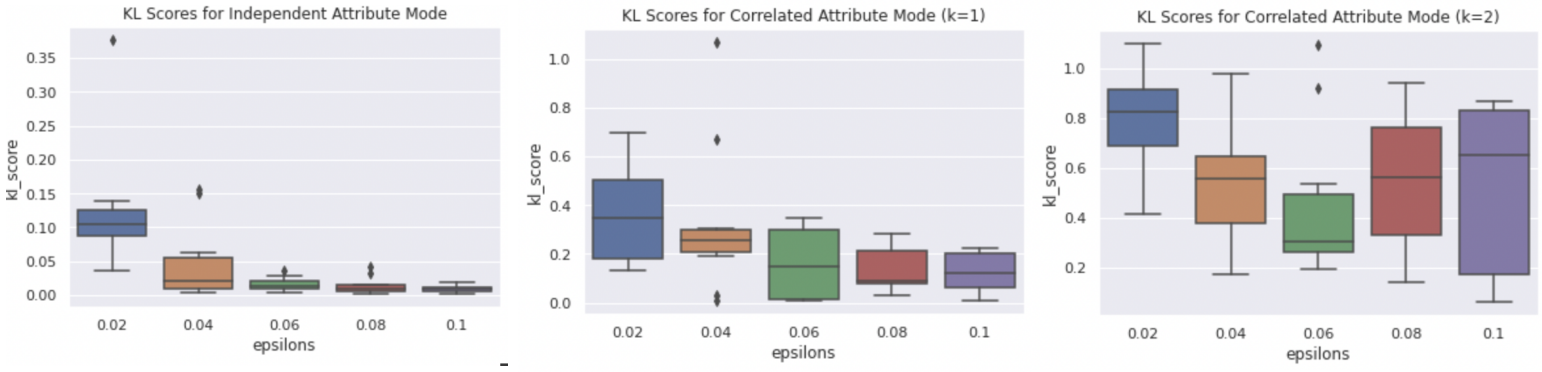


Fig 7

The box and whisker plots above show the distribution of KL scores at different levels of epsilon in 10 synthetic datasets. The higher the epsilon value, larger the privacy budget to spend on the data synthesizer. As a result, the synthetic datasets will more accurately represent the ground truth values. In independent attribute mode, the KL score is diverging less and less as epsilon is increased, as well as approaching 0. This means that by the time the dataset is synthesized with a differential privacy of epsilon=0.1, the KL test score tells us that there is essentially no differences between the race attribute for the ground truth and synthetic datasets. Mode C follows this pattern as well, with the KL score's range decreasing as epsilon is increased with the exception of epsilon=0.04. Meanwhile, the KL scores for correlated attribute at k=2 does not and only diverges more as epsilon is increased.