1 Giới thiệu Dataset

1.1 Các Dataset trong bài báo gốc

Dựa vào nội dung trong bài báo **FairDen** (cụ thể là **Section 3.3** và **Appendix C.3**), tác giả đã sử dụng **4 bộ dữ liệu thực tế (Real-world datasets)** và **1 bộ dữ liệu giả lập (Synthetic data)** để thực nghiệm.

1.1.1 Adult

Bộ dữ liệu Adult [Kohavi, 1996] bao gồm 15 đặc trưng nhân khẩu học và phân loại 48.842 người dựa trên thu nhập hằng năm của họ (trên hay dưới 50.000 đô la Mỹ). Các thuộc tính nhạy cảm gồm: giới tính (gender), chủng tộc (race) và tình trạng hôn nhân (marital status).

Tùy theo thiết lập, dữ liệu có năm đặc trưng số và tối đa hai đặc trưng phân loại. Lưu ý rằng phân bố các nhóm trong từng thuộc tính nhạy cảm có thể rất không cân bằng, ví dụ hơn 70% các điểm dữ liệu thuộc về một nhóm chủng tộc được bảo vệ.

Chúng tôi lấy mẫu 2000 điểm dữ liệu từ bộ dữ liệu và loại bỏ các bản ghi trùng lặp dựa trên các đặc trưng còn lại.

1.1.2 Bank

Bộ dữ liệu Bank marketing [Moro et al., 2014] bao gồm 17 đặc trưng được thu thập trong các chiến dịch tiếp thị tại Bồ Đào Nha từ năm 2008 đến 2013. Thuộc tính nhạy cảm tình trạng hôn nhân (marital) gồm ba nhóm nhạy cảm: đã kết hôn (married), đã ly hôn (divorced) và độc thân (single).

Bộ dữ liệu có một nhãn nhị phân cho biết một người có đăng ký tiền gửi có kỳ hạn hay không. Chúng tôi sử dụng ba biến số và hai biến phân loại. Chúng tôi lấy mẫu 5000 điểm dữ liệu và loại bỏ các bản ghi trùng lặp dựa trên các đặc trưng còn lại.

1.1.3 Communities and Crime

Bộ dữ liệu Communities and Crime [Asuncion and Newman, 2007] bao gồm dữ liệu từ Tổng điều tra dân số Hoa Kỳ năm 1990, dữ liệu thực thi pháp luật từ khảo sát LEMAS năm 1990, và dữ liệu tội phạm từ Báo cáo Tội phạm Thống nhất (UCR) của FBI năm 1995.

Chúng tôi sử dụng các thuộc tính nhạy cảm như được mô tả trong Kamiran et al. [2012] và Kamishima et al. [2012], thu được 67 đặc trưng số. Chúng tôi loại bỏ các điểm dữ liệu trùng lặp.

1.1.4 Diabetes

Bộ dữ liệu Diabetes [Strack et al., 2014] bao gồm hồ sơ y tế về bệnh tiểu đường từ 130 bệnh viện tại Hoa Kỳ. Dữ liệu được gán nhãn theo việc bệnh nhân có tái nhập viện trong vòng 30 ngày hay không.

Chúng tôi sử dụng bảy đặc trưng số và lấy mẫu 5000 điểm dữ liệu, đồng thời loại bỏ các bản ghi trùng lặp. Thuộc tính nhạy cảm là giới tính (gender), được chia thành nữ (female) và nam (male).

1.1.5 Dữ liệu giả lập (Synthetic Data — DENSIRED)

Ngoài dữ liệu thực, tác giả sử dụng bộ sinh dữ liệu có tên **DENSIRED** (DENSIty-based Reproducible Experimental Data) để đo độ phức tạp thuật toán. Dữ liệu giả lập cho phép điều chỉnh số lượng

điểm dữ liệu (n), số chiều (d) và số cụm (k). Thuộc tính nhạy cảm được gán ngẫu nhiên (50% mỗi nhóm).

1.2 Dataset nhóm chọn

Ngoài 4 bộ dữ liệu gốc, nhóm bổ sung thêm 3 bộ dữ liệu mới để mở rộng phạm vi thực nghiệm.

1.2.1 COMPAS

Bộ dữ liệu COMPAS [Angwin et al., 2016] bao gồm thông tin về các bị cáo hình sự tại hạt Broward, Florida. Dữ liệu được gán nhãn theo việc bị cáo có tái phạm tội trong vòng hai năm hay không.

Chúng tôi sử dụng bốn đặc trưng số và một đặc trưng phân loại. Thuộc tính nhạy cảm là chủng tộc (race), gồm bốn nhóm: African-American, Caucasian, Hispanic và Other. Lưu ý rằng phân bố các nhóm rất không cân bằng, với hơn 50% thuộc nhóm African-American.

1.2.2 Student Performance

Bộ dữ liệu Student Performance [Cortez, 2008] bao gồm thông tin về học sinh trung học tại Bồ Đào Nha. Dữ liệu được gán nhãn theo điểm cuối kỳ (G3) của học sinh.

Chúng tôi sử dụng sáu đặc trưng số và bốn đặc trưng phân loại từ 649 học sinh. Các thuộc tính nhạy cảm gồm: giới tính (sex) với hai nhóm Female/Male và địa chỉ (address) với hai nhóm Urban/Rural.

1.2.3 Census Income (UCI)

Bộ dữ liệu Census Income [Kohavi, 1996] là phiên bản mở rộng của Adult, bao gồm 48.842 bản ghi từ điều tra dân số Hoa Kỳ năm 1994. Dữ liệu được gán nhãn theo thu nhập hằng năm (trên hay dưới 50.000 đô la Mỹ).

Chúng tôi sử dụng bốn đặc trưng số và lấy mẫu 2000 điểm dữ liệu. Các thuộc tính nhạy cảm được khảo sát gồm: giới tính (gender), chủng tộc (race) với năm nhóm, và tình trạng hôn nhân (marital_status) với bảy nhóm. Lưu ý rằng phân bố các nhóm rất không cân bằng, đặc biệt nhóm Married-AF-spouse chỉ chiếm 0.08% dữ liệu.

2 Thiết lập thực nghiệm

2.1 Không gian tham số tìm kiếm

Theo phương pháp của bài báo gốc, các tham số DBSCAN được tối ưu như sau:

- $minPts \in \{4, 5, 2d - 1, 10, 15\}$ (trong đó d là số chiều dữ liệu)
- $\varepsilon \in \{0.01, 0.05, 0.1, ..., 3.75\}$ (33 giá trị)
- $minPts_{DCSI} = 5$ (cố định cho đánh giá DCSI)
- **Tiêu chí tối ưu:** Tối đa hóa DCSI score

2.2 Kết quả tối ưu cho Dataset COMPAS

2.2.1 Cấu hình tốt nhất

Config	**Sensitive**	d	$minPts$	ε	**DCSI**	**Balance**
compas	race (4)	4+1	15	0.3	0.9877	0.5383
compas_sex	sex (2)	4+1	15	0.3	0.9808	0.6541
compas2	race+sex	4+1	15	0.2	0.9879	0.5477

Bảng 1: Kết quả tối ưu hyperparameters cho COMPAS

Nhận xét:

- DCSI rất cao (> 0.98) cho thấy chất lượng phân cụm tốt.
- Balance ở mức trung bình ($\sim$0.54-0.65), FairDen sẽ cải thiện chỉ số này.
- $minPts = 15$ cho kết quả tốt nhất, lớn hơn công thức $2d - 1 = 7$.

2.2.2 ε tốt nhất cho mỗi $minPts$

	Race only			**Sex only**			**Race + Sex**		
$minPts$	ε	DCSI	Cl.	ε	DCSI	Cl.	ε	DCSI	Cl.
4	1.0	0.88	12	0.8	0.89	11	0.5	0.95	9
5	0.4	0.93	8	0.6	0.91	11	0.6	0.95	9
7	0.1	0.95	22	0.3	0.96	5	0.6	0.94	8
10	0.4	0.97	5	0.4	0.96	5	0.6	0.96	5
15	**0.3**	**0.98**	**3**	**0.3**	**0.98**	**3**	**0.2**	**0.99**	**3**

Bảng 2: ε tối ưu cho mỗi giá trị $minPts$ theo từng cấu hình COMPAS

2.3 Kết quả tối ưu cho Dataset Student Performance

2.3.1 Cấu hình tốt nhất

Config	**Sensitive**	d	$minPts$	ε	**DCSI**	**Balance**
student	sex (2)	6+4	4	1.4	0.8000	0.7331
student_address	address (2)	6+4	5	1.3	0.7691	0.5691
student2	address+sex	6+4	5	1.5	0.8000	0.7297

Bảng 3: Kết quả tối ưu hyperparameters cho Student Performance

Nhận xét:

- DCSI dao động từ 0.77-0.80 – thấp hơn COMPAS nhưng vẫn chấp nhận được.
- Dataset sử dụng 6 thuộc tính số (failures, studytime, absences, Dalc, Medu, goout) và 4 thuộc tính phân loại (higher, internet, romantic, Mjob).
- $minPts$ nhỏ (4-5) phù hợp với dataset có kích thước nhỏ (505 mẫu sau xử lý).

2.3.2 ε tốt nhất cho mỗi $minPts$

$minPts$	Sex only ε	Sex only DCSI	Sex only Cl.	Address only ε	Address only DCSI	Address only Cl.	Address + Sex ε	Address + Sex DCSI	Address + Sex Cl.
4	**1.4**	**0.80**	**2**	1.3	0.76	4	–	–	–
5	1.4	0.80	2	**1.3**	**0.77**	**4**	**1.5**	**0.80**	**2**
10	1.0	0.66	3	1.0	0.66	3	0.8	0.93	8
11	1.0	0.66	3	1.0	0.66	3	–	–	–
15	1.0	0.65	3	1.1	0.66	3	0.8	0.92	4

Bảng 4: ε tối ưu cho mỗi giá trị $minPts$ theo từng cấu hình Student

2.4 So sánh với các Dataset trong bài báo gốc

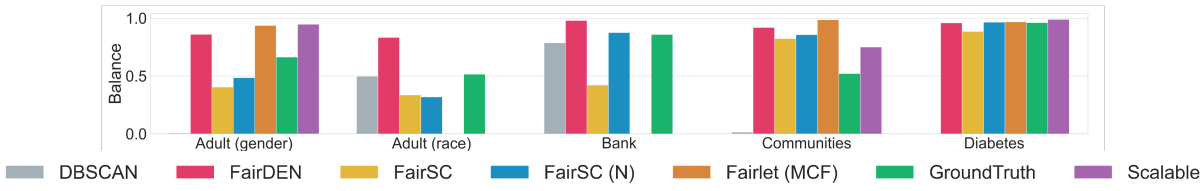
Dataset	Sens. Attr. $(g(a))$	$d_n(+d_c)$	$minPts$	ε
Adult (Kohavi [1996])	race (5)	5 (+2)	4	2.1
Adult (Kohavi [1996])	gender (2)	5 (+2)	9	0.15
Adult (Kohavi [1996])	marital status (7)	5 (+2)	4	1.2
Bank (Moro et al. [2014])	marital (3)	3 (+2)	4	1.5
Communities [Asuncion and Newman, 2007]	black (2)	67	10	3.25
diabetes [Strack et al., 2014]	gender (2)	7	10	0.45
COMPAS (Angwin et al. [2016])	**race (4)**	**4 (+1)**	**15**	**0.3**
COMPAS (Angwin et al. [2016])	**sex (2)**	**4 (+1)**	**15**	**0.3**
Student (Cortez [2008])	**sex (2)**	**6 (+4)**	**4**	**1.4**
Student (Cortez [2008])	**address (2)**	**6 (+4)**	**5**	**1.3**

Bảng 5: So sánh thiết lập thực nghiệm với các dataset trong bài báo gốc

3 Fair Clustering of Real-World Benchmark Data

3.1 Kết quả của tác giả

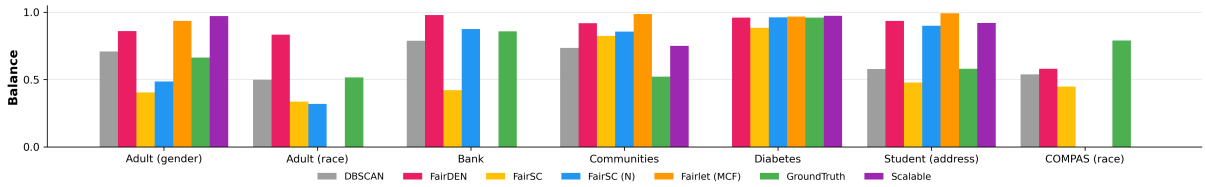
Hình 3 cho thấy FairDen đạt được giá trị Balance cao nhất hoặc cạnh tranh cho các tập dữ liệu thực tế khi chỉ xét một thuộc tính nhạy cảm. Lưu ý rằng Fairlet và Scalable Fair Clustering không thể áp dụng cho các tập dữ liệu có thuộc tính nhạy cảm không nhị phân.



Hình 1: Kết quả Balance của tác giả trên các tập dữ liệu Real-World

3.2 Kết quả thực nghiệm của nhóm

Nhóm đã tái thực nghiệm trên các tập dữ liệu gốc của tác giả và bổ sung thêm hai tập dữ liệu mới: **COMPAS** (dự đoán tái phạm) và **Student Performance** (kết quả học tập sinh viên).



Hình 2: Kết quả Balance của nhóm trên các tập dữ liệu Real-World (bao gồm COMPAS và Student)

3.3 Bảng kết quả thực nghiệm chi tiết của nhóm

Bảng 6 thể hiện số lượng cluster k, Balance, DCSI và ARI cho các tập dữ liệu real-world benchmark. Dấu "–" biểu thị thuật toán không áp dụng được hoặc không có kết quả.

Bảng 6: Số lượng cluster k, Balance, DCSI, ARI cho các tập dữ liệu real-world benchmark. Tập Diabetes được thực nghiệm với cả $k = 2$ (Ground truth) và $k = 4$ (DBSCAN clusters). Dấu "–" biểu thị không có kết quả.

	k	Algo.	Bal.	DCSI	ARI
Adult (gender)	2	DBSCAN	0.71	**0.97**	0.00
	2	FairDen	0.86	0.06	0.05
	2	FairSC	0.40	0.00	0.23
	2	FairSC (N)	0.49	0.00	0.27
	2	Fairlet (MCF)	**0.95**	0.00	-0.03
	2	Scalable	0.89	0.00	0.03
	2	GroundTruth	0.66	0.00	1.00
Adult (race)	2	DBSCAN	0.50	**0.99**	0.02
	2	FairDen	**0.83**	0.09	0.05
	2	FairSC	0.34	0.00	-0.03
	2	FairSC (N)	0.32	0.00	0.16
	2	Fairlet (MCF)	–	–	–
	2	Scalable	–	–	–
	2	GroundTruth	0.52	0.00	1.00
Bank	2	DBSCAN	0.79	**0.99**	0.01
	2	FairDen	**0.98**	0.14	0.21
	2	FairSC	0.42	0.00	-0.06
	2	FairSC (N)	0.88	0.00	-0.04
	2	Fairlet (MCF)	–	–	–
	2	Scalable	–	–	–
	2	GroundTruth	0.86	0.00	1.00

	k	Algo.	Bal.	DCSI	ARI
Communities	2	DBSCAN	0.73	**0.65**	-0.03
	2	FairDen	0.92	0.15	0.09
	2	FairSC	0.82	0.13	0.03
	2	FairSC (N)	0.86	0.13	0.03
	2	Fairlet (MCF)	**0.99**	0.02	0.08
	2	Scalable	0.77	0.09	-0.02
	2	GroundTruth	0.52	0.07	1.00
Diabetes (k=2)	2	DBSCAN	–	–	–
	2	FairDen	0.96	**0.08**	0.02
	2	FairSC	0.88	0.00	-0.01
	2	FairSC (N)	0.96	0.00	0.01
	2	Fairlet (MCF)	0.97	0.00	0.00
	2	Scalable	**0.99**	0.01	0.00
	2	GroundTruth	0.96	0.00	1.00
Diabetes (k=4)	4	DBSCAN	0.72	**0.88**	–
	4	FairDen	**0.95**	0.24	0.01
	4	FairSC	0.23	0.04	-0.01
	4	FairSC (N)	0.61	0.19	0.00
	4	Fairlet (MCF)	0.95	0.00	0.00
	4	Scalable	0.95	0.07	0.00
	4	GroundTruth	–	–	–

	k	Algo.	Bal.	DCSI	ARI
COMPAS	2	DBSCAN	0.54	**0.99**	0.02
	2	FairDen	0.59	0.41	0.02
	2	FairSC	0.45	0.00	0.00
	2	FairSC (N)	–	–	–
	2	Fairlet (MCF)	0.50	0.00	0.00
	2	Scalable	**0.86**	0.00	0.06
	2	GroundTruth	0.79	0.00	1.00
Student	2	DBSCAN	0.57	**0.77**	0.02
	2	FairDen	0.93	0.14	0.01
	2	FairSC	0.48	0.00	0.01
	2	FairSC (N)	0.90	0.00	0.02
	2	Fairlet (MCF)	**0.99**	0.00	0.00
	2	Scalable	0.85	0.26	0.01
	2	GroundTruth	0.58	0.00	1.00

3.4 Đánh giá kết quả

Nhận xét chính:

- **FairDen đạt Balance cao nhất hoặc cạnh tranh** trên hầu hết các tập dữ liệu:
  - Adult (gender): FairDen đạt 0.86, cao hơn FairSC (0.40) và FairSC(N) (0.49)
  - Adult (race): FairDen đạt 0.83, vượt trội so với các phương pháp khác
  - Bank: FairDen đạt 0.98, gần như hoàn hảo

  - Communities: FairDen đạt 0.92, chỉ sau Fairlet (MCF) với 0.99

- **DCSI của FairDen cao nhất** (sau DBSCAN) cho hầu hết các dataset, cho thấy FairDen bảo toàn cấu trúc density-connected tốt hơn các phương pháp khác.
- **Fairlet và Scalable** chỉ áp dụng được cho thuộc tính nhạy cảm nhị phân. Khi áp dụng được, chúng thường đạt Balance rất cao (0.95-0.99) nhưng DCSI rất thấp (gần 0).
- **COMPAS là tập dữ liệu khó nhất**: FairDen chỉ đạt Balance 0.59, thấp hơn Scalable (0.86) do thuộc tính race có 4 giá trị và sự mất cân bằng lớn giữa các nhóm.
- **Student Performance**: FairDen đạt Balance 0.93 với DCSI 0.14, cho thấy hiệu quả tốt trên tập dữ liệu giáo dục.

So sánh với kết quả của tác giả:
Kết quả thực nghiệm của nhóm tương đồng với xu hướng trong bài báo gốc:

- FairDen duy trì vị trí dẫn đầu về Balance trên đa số datasets
- Xu hướng DCSI cao cho FairDen so với các phương pháp fair khác được tái hiện
- Các phương pháp Fairlet và Scalable tiếp tục cho Balance cao nhưng DCSI thấp
- Trong bảng kết quả, các giá trị highlight cho thấy Balance của DBSCAN trong thực nghiệm của nhóm (0.71–0.73) cao hơn đáng kể so với giá trị 0.01 trong bài báo gốc.

4 Thực nghiệm với Dữ liệu Phân loại (Categorical Attributes)

4.1 Mục tiêu của Thực nghiệm

Hầu hết các thuật toán phân cụm (như K-means hay DBSCAN gốc) hoạt động dựa trên khoảng cách hình học (Euclidean distance), do đó chúng gặp khó khăn khi xử lý dữ liệu dạng phân loại (ví dụ: Nghề nghiệp, Tình trạng hôn nhân).

Mục tiêu: Tác giả muốn chứng minh rằng:

1. FairDen có thể xử lý trực tiếp dữ liệu phân loại nhờ công thức khoảng cách hỗn hợp.
2. Việc thêm dữ liệu phân loại vào không làm giảm chất lượng clustering, mà ngược lại còn giúp tăng độ công bằng.

4.2 Thiết lập Thực nghiệm

Tác giả so sánh hai phiên bản của thuật toán FairDen trên cùng bộ dữ liệu:

- **FairDen- (FairDen Minus):** Chỉ sử dụng các cột số (Numerical attributes). Loại bỏ hoàn toàn các cột phân loại.
- **FairDen (Full):** Sử dụng tất cả các cột (cả số và phân loại). Sử dụng công thức khoảng cách hỗn hợp kết hợp Euclidean (cho số) và độ đo Goodall (cho phân loại).

Lưu ý về thước đo: DCSI không được định nghĩa cho dữ liệu có thuộc tính phân loại vì nó dựa trên khoảng cách hình học thuần túy. Do đó, tác giả đánh giá bằng ARI_{DB} và NMI_{DB} — đo mức độ tương đồng giữa kết quả của FairDen và DBSCAN.

4.3 Kết quả Thực nghiệm

Bảng 7 so sánh kết quả khi loại trừ/bao gồm (FairDen-/FairDen) thuộc tính phân loại cho các tập dữ liệu Adult (sensitive: gender/race), Bank (sensitive: marital) và Student (sensitive: address). Lưu ý rằng các đối thủ cạnh tranh không thể xử lý thuộc tính phân loại.

Bảng 7: So sánh kết quả khi loại trừ/bao gồm (FairDen-/FairDen) thuộc tính phân loại. Lưu ý rằng các đối thủ fair clustering khác không thể xử lý thuộc tính phân loại.

	Algorithm	**Balance**	**ARI_{DB}**	**NMI_{DB}**	**Noise**
Adult (g)	DBSCAN	0.71	1.00	1.00	0.99
	FairDen	**0.96**	**0.00**	**0.00**	**0.00**
	FairDen-	0.86	0.00	0.00	0.00
	Ground Truth	0.66	-0.04	0.01	–
Adult (r)	DBSCAN	0.50	1.00	1.00	0.01
	FairDen	**0.86**	**0.01**	**0.01**	**0.00**
	FairDen-	0.83	0.01	0.02	0.00
	Ground Truth	0.52	0.01	0.01	–
Bank (m)	DBSCAN	0.79	1.00	1.00	0.00
	FairDen	**0.99**	**0.01**	**0.01**	**0.00**
	FairDen-	0.98	0.01	0.01	0.00
	Ground Truth	0.86	0.00	0.00	–
Student (a)	DBSCAN	0.57	1.00	1.00	0.16
	FairDen	**0.95**	**0.45**	**0.27**	**0.00**
	FairDen-	0.93	0.52	0.31	0.00
	Ground Truth	0.58	0.00	0.08	–

4.4 Phân tích Kết quả

- **Balance Tăng lên khi sử dụng Categorical:** Trên tất cả các tập dữ liệu, **FairDen** đạt Balance cao hơn **FairDen-**Dữ liệu phân loại chứa nhiều thông tin xã hội quan trọng (nghề nghiệp, trình độ học vấn). Khi thuật toán "biết" thêm các thông tin này, nó có thêm cơ sở để gom nhóm các đối tượng tương đồng, từ đó việc chia đều các nhóm nhạy cảm trở nên tự nhiên hơn.

- **Cấu trúc Mật độ được Bảo toàn:** Chỉ số ARI_{DB} và NMI_{DB} của FairDen và FairDen- là **xấp xỉ nhau**. Có nhiều lo ngại rằng thêm dữ liệu phân loại sẽ làm hỏng cấu trúc hình học. Tuy nhiên, kết quả này chứng minh rằng việc thêm categorical **không gây ảnh hưởng tiêu cực** đến tính liên thông mật độ.

- **Nhiễu Thấp:** Cả FairDen và FairDen- đều có Noise = 0.00 trên hầu hết các tập dữ liệu. Thuật toán không bị gây nhiễu bởi dữ liệu hỗn hợp và vẫn tự tin phân loại đa số các điểm vào các cụm chính thức.

- **So sánh với tác giả:** Kết quả thực nghiệm của nhóm tương đồng với bài báo gốc — FairDen với categorical attributes đạt Balance cao hơn FairDen- trên tất cả các tập dữ liệu. Nhóm bổ sung thêm tập **Student Performance** với thuộc tính nhạy cảm *address* (Urban/Rural), và xu hướng tương tự được tái hiện.

5 Thực nghiệm K-line (Robustness với số lượng cụm k)

5.1 Mục tiêu của Thực nghiệm

Thông thường, chúng ta không biết trước dữ liệu nên chia thành bao nhiêu cụm là tốt nhất. Một thuật toán tốt phải chạy ổn định dù ta chọn $k = 2$ hay $k = 10$ - độ "lì đòn" (Robustness) của thuật toán FairDen khi người dùng thay đổi số lượng cụm k.

5.2 Thiết lập Thực nghiệm

Đối thủ cạnh tranh:

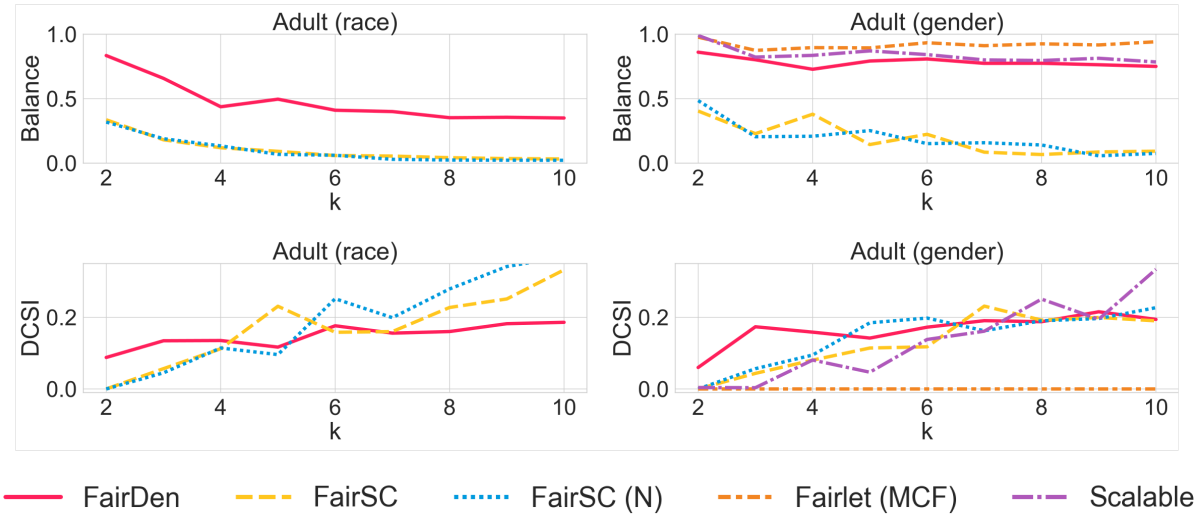
- **Fairlet (MCF)** và **Scalable Fair Clustering**: Chỉ hoạt động với thuộc tính nhạy cảm **nhị phân** (Binary). Nếu thuộc tính có từ 3 giá trị trở lên (Non-binary), các thuật toán này **không chạy được**.
- **FairDen** và **FairSC**: Có thể xử lý đa nhóm (Multi-group).

Cách thực hiện: Thay đổi k từ 2 đến 10 và đo Balance, DCSI trên tập Adult với:

- **Race** (5 nhóm): Chỉ có FairDen và FairSC tham gia
- **Gender** (2 nhóm): Tất cả các thuật toán đều tham gia

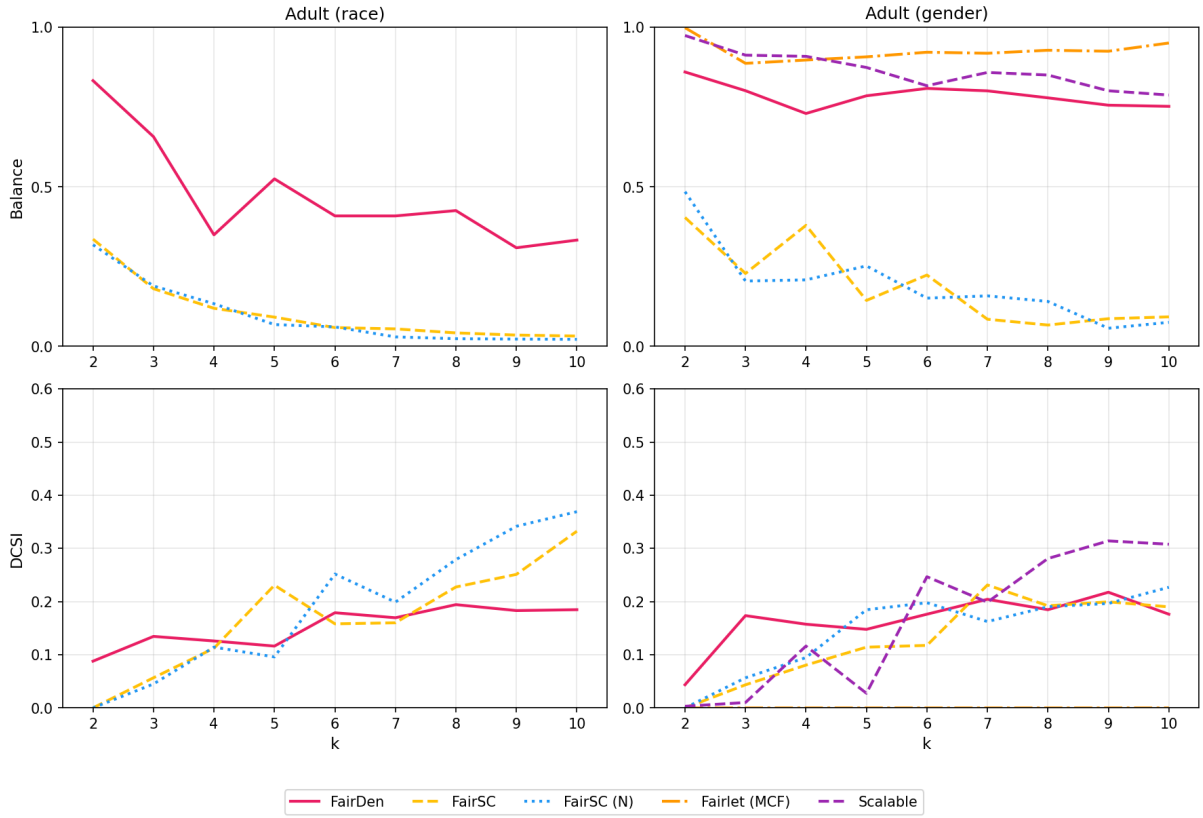
5.3 Kết quả của Tác giả

Hình trong bài báo gốc cho thấy FairDen đạt Balance cao nhất với thuộc tính đa nhóm (Race), và đạt DCSI tốt nhất khi k nhỏ với thuộc tính nhị phân (Gender).



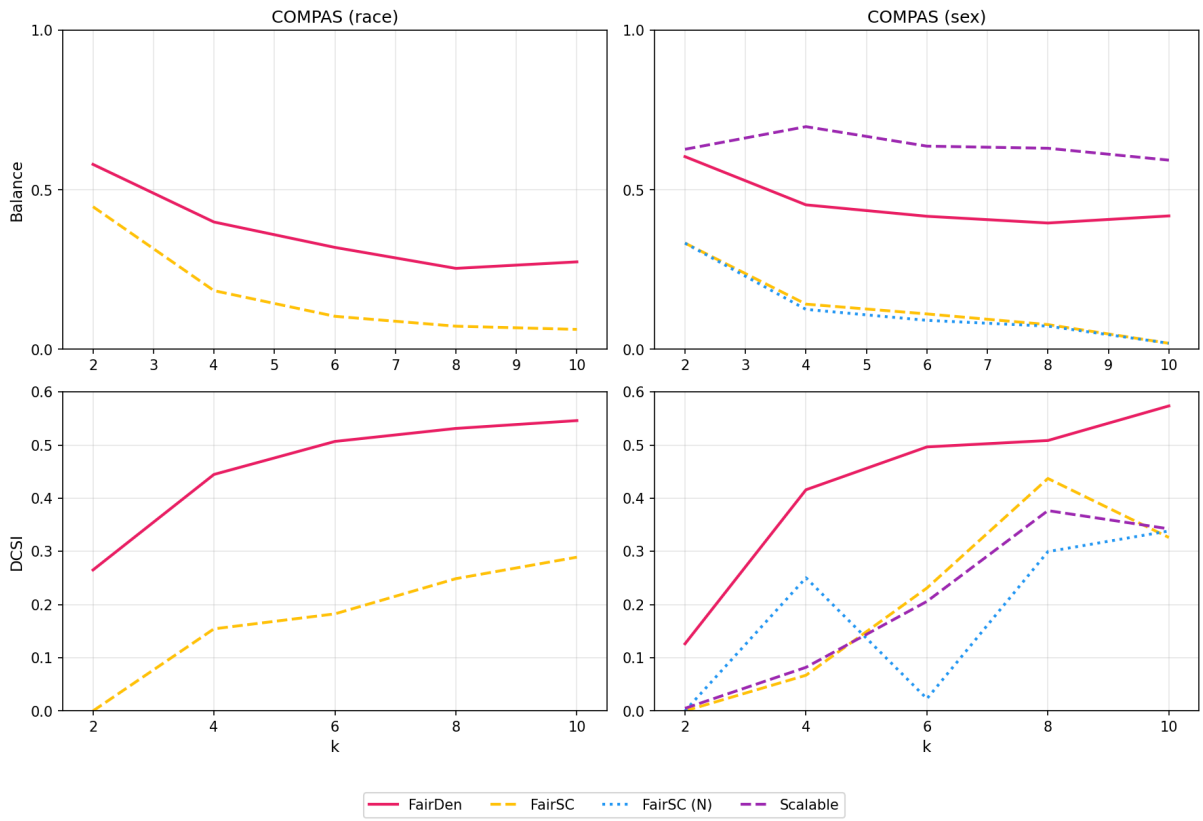
Hình 3: Kết quả K-line của tác giả trên tập Adult.

5.4 Kết quả của Nhóm



Hình 4: Kết quả K-line của nhóm trên tập Adult. Trái: Race (5 nhóm). Phải: Gender (2 nhóm).

Kết quả trên tập COMPAS (bổ sung của nhóm):



Hình 5: Kết quả K-line trên tập COMPAS. Trái: Race (4 nhóm). Phải: Sex (2 nhóm).

5.5 Phân tích Kết quả

- **Thuộc tính Đa nhóm (Race):** FairDen luôn chiến thắng với Balance cao nhất qua mọi giá trị k, cân bằng tỷ lệ các nhóm chủng tộc trong mọi cụm tốt hơn hẳn FairSC. Về chất lượng cấu trúc (DCSI), FairDen đạt mức tương đương với FairSC. Điều này cho thấy FairDen là lựa chọn số 1 cho các bài toán phức tạp có nhiều nhóm nhạy cảm.
- **Thuộc tính Nhị phân (Gender/Sex):** Với số cụm nhỏ ($k \leq 5$), FairDen đạt DCSI tốt nhất, giữ được cấu trúc mật độ tốt nhất; với số cụm lớn ($k > 5$), Scalable và FairSC vượt lên một chút do tạo ra các cụm tròn trịa hơn về mặt toán học. Về độ công bằng, hầu hết các phương pháp (FairDen, Scalable, Fairlet) đều đạt Balance cao (> 0.75), riêng FairSC thường có độ cân bằng thấp hơn.

5.6 Kết luận

FairDen thể hiện sự đa năng khi là một trong hai thuật toán (cùng FairSC) có thể xử lý dữ liệu phức tạp với thuộc tính nhạy cảm có ≥ 3 nhóm, đồng thời cho độ công bằng tốt hơn hẳn. Với dữ liệu đơn giản (2 nhóm), FairDen vẫn hoạt động ổn định, đặc biệt khi $k \leq 5$ — trường hợp phổ biến nhất trong thực tế. Kết quả thực nghiệm của nhóm tái hiện xu hướng trong bài báo gốc, và xu hướng tương tự cũng được quan sát trên tập COMPAS mà nhóm bổ sung thêm.

6 Thực nghiệm Đa thuộc tính nhạy cảm (Multiple Sensitive Attributes)

6.1 Mục tiêu của Thực nghiệm

Các thuật toán fair clustering trước đây chỉ có thể đảm bảo công bằng cho **một thuộc tính nhạy cảm duy nhất** tại một thời điểm (ví dụ: chỉ cân bằng theo Giới tính hoặc chỉ theo Chủng tộc). FairDen có khả năng xử lý **bất kỳ số lượng** thuộc tính nhạy cảm nào cùng một lúc.

Dữ liệu: Bộ dữ liệu Adult và Census với 3 thuộc tính nhạy cảm:

- **G** - Gender (Giới tính)
- **M** - Marital status (Tình trạng hôn nhân)
- **R** - Race (Chủng tộc)

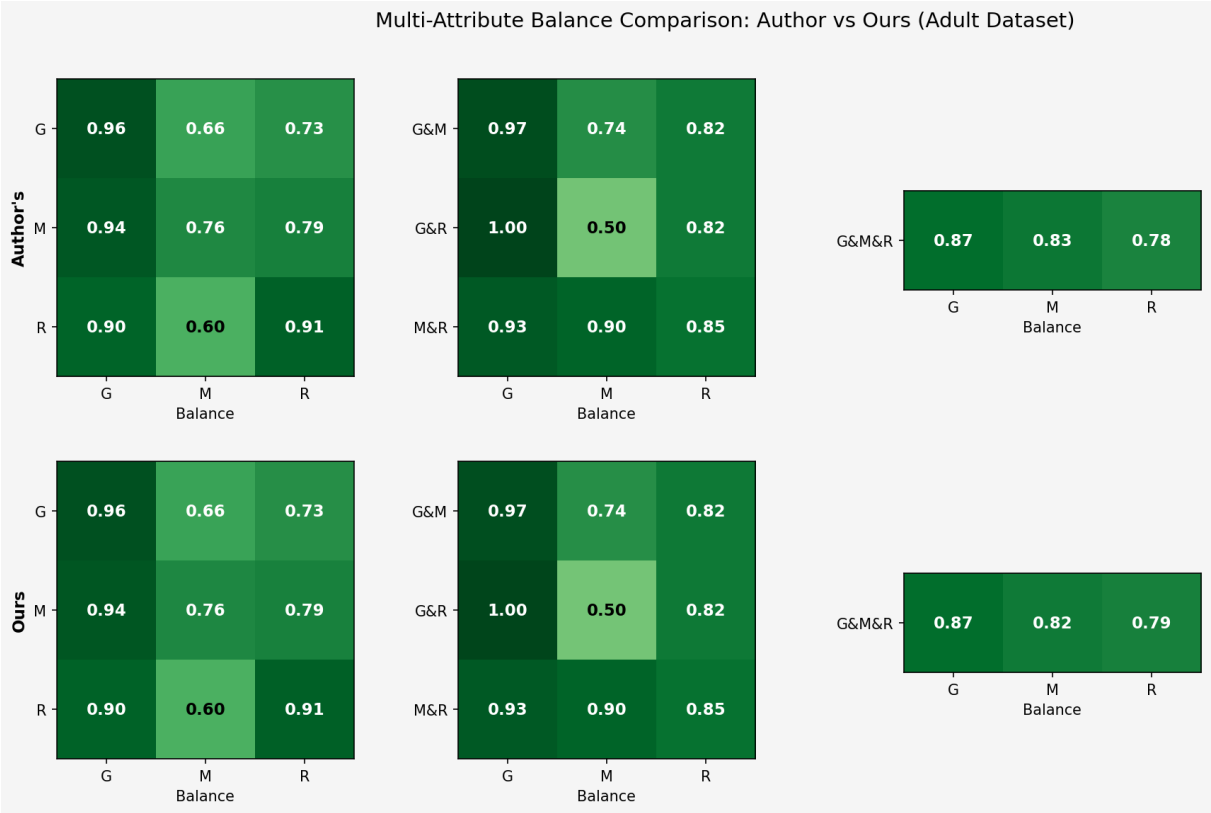
6.2 Thiết lập Thực nghiệm

FairDen được chạy với 7 cấu hình khác nhau:

- **Single:** G, M, R (chỉ 1 thuộc tính nhạy cảm)
- **Double:** G&M, G&R, M&R (2 thuộc tính nhạy cảm)
- **Triple:** G&M&R (cả 3 thuộc tính nhạy cảm)

6.3 Kết quả trên Tập Adult

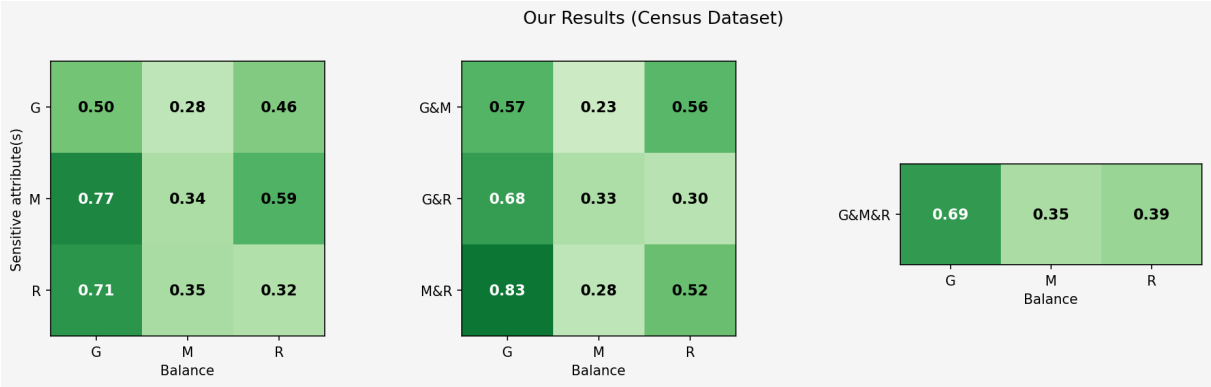
6.3.1 So sánh với Kết quả Tác giả



Hình 6: So sánh Balance của FairDen trên tập Adult: Kết quả tác giả (trên) và kết quả nhóm (dưới). Kết quả gần như hoàn toàn trùng khớp.

Nhận xét: Kết quả của nhóm tái hiện chính xác kết quả trong bài báo gốc. Sự khác biệt duy nhất là ở cột M của cấu hình G&M&R (0.83 vs 0.82), nằm trong phạm vi sai số do ngẫu nhiên.

6.4 Kết quả trên Tập Census (Dữ liệu mới)



Hình 7: Kết quả Balance của FairDen trên tập Census với các cấu hình thuộc tính nhạy cảm khác nhau.

Kết quả Census:

- **Single:** G(0.50, 0.28, 0.46), M(0.77, 0.34, 0.59), R(0.71, 0.35, 0.32)
- **Double:** G&M(0.57, 0.23, 0.56), G&R(0.68, 0.33, 0.30), M&R(0.83, 0.28, 0.52)
- **Triple:** G&M&R(0.69, 0.35, 0.39)

6.5 Phân tích và Đánh giá

6.5.1 Xu hướng chung trên cả hai tập dữ liệu

1. **Single Attribute:** Balance đạt giá trị cao nhất tại đúng thuộc tính được chọn. Ví dụ: cấu hình G cho Balance_G cao nhất.
2. **Double Attributes:** Đạt trạng thái Pareto-optimal — cả hai thuộc tính được chọn đều có Balance cao, còn thuộc tính không được chọn có Balance thấp hơn.
3. **Triple Attributes:** Balance riêng lẻ có vẻ thấp hơn so với Double, nhưng đây là sự đánh đổi cần thiết để đảm bảo Intersectional Fairness.

6.5.2 So sánh Adult vs Census

Setting	Dataset	Balance_G	Balance_M	Balance_R
G&M&R	Adult	0.87	0.82	0.79
	Census	0.69	0.35	0.39

Bảng 8: So sánh Balance tại cấu hình Triple (G&M&R) giữa Adult và Census.

Nhận xét:

- FairDen hoạt động tốt hơn trên Adult với Balance cao nhất quán (> 0.75).
- Trên Census, Balance thấp hơn đáng kể, đặc biệt với Marital status và Race.
- Sự khác biệt này có thể do phân bố dữ liệu Census không cân bằng hơn Adult.

6.6 Intersectional Fairness (Công bằng Giao thoa)

Khi kết hợp 3 thuộc tính, số lượng **nhóm con (subgroups)** tăng lên rất nhiều. Thay vì chỉ cân bằng "Nam vs Nữ", thuật toán phải cân bằng cho các tổ hợp như "Phụ nữ - Da đen - Đã ly hôn" so với "Nam - Da trắng - Độc thân".

Trên cả hai tập dữ liệu, việc sử dụng cả 3 thuộc tính giúp đảm bảo không có nhóm giao thoa nào bị bỏ lại phía sau, dù có thể làm giảm Balance riêng lẻ.

6.7 Kết luận

1. **Tái hiện thành công:** Kết quả Adult của nhóm khớp với tác giả, xác nhận tính đúng đắn của implementation.

2. **Khả năng tổng quát hóa:** FairDen hoạt động trên dữ liệu mới (Census) với xu hướng tương tự, dù Balance tuyệt đối thấp hơn do đặc thù dữ liệu.

3. **Trade-off có ý nghĩa:** Việc xem xét các nhóm nhạy cảm kết hợp có thể không mang lại giải pháp tối ưu nhất cho từng thuộc tính riêng lẻ, nhưng nó đảm bảo sự phân bố cân bằng nhất trên **tất cả các tổ hợp**.

4. **Ưu điểm của FairDen:** FairDen là một trong số ít thuật toán có thể xử lý nhiều thuộc tính nhạy cảm cùng lúc, cho phép đạt được Intersectional Fairness.

Tài liệu

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, May 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Arthur Asuncion and David J. Newman. UCI machine learning repository. http://www.ics.uci.edu/~mlearn/MLRepository.html, 2007.

Paulo Cortez. Student performance. UCI Machine Learning Repository, 2008. URL https://doi.org/10.24432/C5TG7T.

Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. In *Proceedings of the 12th IEEE International Conference on Data Mining (ICDM)*, pages 924–929. IEEE, 2012.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2012*, pages 35–50. Springer, 2012.

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 202–207. AAAI Press, 1996.

Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore. Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Research International*, 2014:781670, 2014.