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.

Dưới đây là mô tả chi tiết từng bộ dữ liệu:

1.1.1 Adult (Census Income)

Đây là bộ dữ liệu kinh điển nhất trong các bài toán về công bằng (Fairness).

- **Nguồn gốc:** Dữ liệu điều tra dân số Mỹ năm 1994 (UCI Machine Learning Repository).
- **Mục tiêu gốc:** Dự đoán xem thu nhập của một người có vượt quá 50.000$/năm hay không.
- **Đặc điểm dữ liệu:** Chứa các thông tin nhân khẩu học như tuổi, trình độ giáo dục, nghề nghiệp, giờ làm việc...
- **Thuộc tính nhạy cảm (Sensitive Attributes) được tác giả dùng:**
  - **Giới tính (Gender):** Nam / Nữ.
  - **Chủng tộc (Race):** 5 nhóm (White, Black, Asian-Pac-Islander, Amer-Indian-Eskimo, Other).
  - **Tình trạng hôn nhân (Marital Status):** Đã kết hôn, Độc thân, Ly hôn...
- **Xử lý của tác giả:**
  - Lấy mẫu ngẫu nhiên **2.000 điểm dữ liệu**.
  - Loại bỏ các bản ghi trùng lặp.
- **Mục đích chính:** Dùng để kiểm thử tính năng **Đa thuộc tính nhạy cảm (Multiple sensitive attributes)** và **Thuộc tính nhạy cảm phi nhị phân** (Non-binary, ví dụ như Chủng tộc có 5 nhóm).

1.1.2 Bank Marketing (Bank)

Dữ liệu liên quan đến chiến dịch marketing trực tiếp của một ngân hàng Bồ Đào Nha.

- **Nguồn gốc:** Moro et al., 2014.
- **Mục tiêu gốc:** Dự đoán khách hàng có đăng ký gửi tiết kiệm kỳ hạn (term deposit) hay không.
- **Đặc điểm dữ liệu:** Gồm 17 thuộc tính hỗn hợp (số và phân loại) như số dư, công việc, nhà ở, khoản vay...
- **Thuộc tính nhạy cảm:**
  - **Tình trạng hôn nhân (Marital):** 3 nhóm (Married, Divorced, Single).
  - **Tuổi (Age):** Được chia nhóm (ví dụ: trẻ, trung niên, già).

- **Xử lý của tác giả:**
  - Lấy mẫu ngẫu nhiên **5.000 điểm dữ liệu**.
  - Sử dụng 3 biến số và 2 biến phân loại.
- **Mục đích chính:** Kiểm chứng khả năng xử lý **Dữ liệu hỗn hợp (Mixed-type data)** bao gồm cả dữ liệu số và dữ liệu phân loại (categorical).

1.1.3 Communities and Crime

Dữ liệu kết hợp giữa điều tra dân số, thực thi pháp luật và dữ liệu tội phạm FBI.

- **Nguồn gốc:** UCI (1990 US Census + 1995 FBI UCR).
- **Mục tiêu gốc:** Dự đoán tỷ lệ tội phạm bạo lực trong cộng đồng.
- **Đặc điểm dữ liệu:** Rất nhiều thuộc tính số (67 numerical features).
- **Thuộc tính nhạy cảm:**
  - **Chủng tộc (Race):** Tác giả tạo ra thuộc tính nhị phân dựa trên tỷ lệ người da đen trong cộng đồng (chia thành nhóm cao/thấp).
- **Xử lý của tác giả:**
  - Loại bỏ các điểm trùng lặp.
- **Mục đích chính:** Kiểm thử trên dữ liệu có số chiều cao (High dimensionality) và thuần túy là dữ liệu số.

1.1.4 Diabetes (Tiểu đường)

Dữ liệu y tế từ 130 bệnh viện tại Mỹ.

- **Nguồn gốc:** Strack et al., 2014.
- **Mục tiêu gốc:** Dự đoán xem bệnh nhân có tái nhập viện trong vòng 30 ngày hay không.
- **Đặc điểm dữ liệu:** Hồ sơ bệnh án.
- **Thuộc tính nhạy cảm:**
  - **Giới tính (Gender):** Nam / Nữ.
  - **Tuổi (Age):** Chia thành 2 nhóm ($<$50 tuổi và $\geq$50 tuổi).
- **Xử lý của tác giả:**
  - Lấy mẫu **5.000 điểm dữ liệu**.
  - Chọn 7 thuộc tính số.
- **Mục đích chính:** Kiểm thử tính công bằng trong lĩnh vực y tế (Healthcare).

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).

- **Mục đích:** Chỉ dùng cho **Thực nghiệm thời gian chạy (Runtime Experiments)** (Phụ lục A.3).
- **Lý do:** Để đo độ phức tạp thuật toán, họ cần tự do điều chỉnh:
  - Số lượng điểm dữ liệu (n) tăng dần (từ 1.000 lên 10.000...).
  - Số chiều dữ liệu (d).
  - Số cụm (k).
- **Thuộc tính nhạy cảm:** Được gán ngẫu nhiên (50% nhóm 0, 50% nhóm 1) vì mục tiêu ở đây chỉ là đo tốc độ chứ không phải chất lượng.

Tóm lại: Tác giả chọn các bộ dữ liệu này để đảm bảo tính đa dạng:

1. **Adult:** Đa dạng về nhóm nhạy cảm (Chủng tộc, Giới tính).
2. **Bank:** Đa dạng về loại dữ liệu (Số + Chữ).
3. **Communities:** Dữ liệu số chiều cao.
4. **Diabetes:** Ứng dụng y tế.

1.2 Dataset nhóm chọn

Ngoài 4 bộ dữ liệu gốc, nhóm bổ sung thêm 2 bộ dữ liệu mới để mở rộng phạm vi thực nghiệm: **COMPAS** (lĩnh vực tư pháp) và **Student Performance** (lĩnh vực giáo dục).

1.2.1 COMPAS (Correctional Offender Management Profiling)

Tổng quan Đây là bộ dữ liệu COMPAS đã được tiền xử lý từ dữ liệu gốc của ProPublica [Angwin et al., 2016], chứa thông tin về các bị cáo hình sự tại hạt Broward, Florida. Dữ liệu được lọc để chỉ giữ lại những người có đủ thời gian theo dõi 2 năm nhằm xác định chính xác hành vi tái phạm tội.

Thuộc tính	Giá trị
Số lượng mẫu ban đầu (n)	7.214 bị cáo
Số thuộc tính số (d_n)	4
Số thuộc tính phân loại (d_c)	1

Bảng 1: Tổng quan dataset COMPAS

Thuộc tính Nhạy cảm (Sensitive Attributes)

- **Race (Chủng tộc):** 4 nhóm (African-American 51.2%, Caucasian 34.0%, Hispanic 8.8%, Other 5.2%)
- **Sex (Giới tính):** 2 nhóm (Male 80.7%, Female 19.3%)

Thuộc tính Đầu vào cho Phân cụm

- **Thuộc tính Số ($d_n = 4$):** `age`, `priors_count`, `juv_fel_count`, `juv_misd_count`
- **Thuộc tính Phân loại ($d_c = 1$):** `c_charge_degree` (F: Felony 64.7%, M: Misdemeanor 35.3%)

1.2.2 Student Performance (Kết quả Học tập)

Tổng quan Bộ dữ liệu này [Cortez, 2008] dự đoán kết quả học tập của học sinh trung học dựa trên các đặc điểm xã hội, nhân khẩu học và hành vi.

Thuộc tính	Giá trị
Số lượng mẫu (n)	649 học sinh
Số thuộc tính số (d_n)	5
Số thuộc tính phân loại (d_c)	2

Bảng 2: Tổng quan dataset Student Performance

Thuộc tính Nhạy cảm

- **Sex (Giới tính):** Female 59%, Male 41%
- **Address (Địa chỉ):** Urban 70%, Rural 30%

Thuộc tính Đầu vào cho Phân cụm Thuộc tính được lựa chọn dựa trên phân tích tương quan và phương sai:

- **Thuộc tính Số ($d_n = 5$):**
  - `failures`: Số lần rớt môn (tương quan -0.39 với điểm)
  - `studytime`: Thời gian học tập (+0.25)
  - `absences`: Số buổi nghỉ học
  - `Dalc`: Mức độ uống rượu bia
  - `Medu`: Trình độ học vấn mẹ (+0.24)
- **Thuộc tính Phân loại ($d_c = 2$):**
  - `higher`: Mong muốn học đại học (Yes/No)
  - `internet`: Có kết nối Internet (Yes/No)

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 3: 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 4: 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)	5+2	4	1.4	0.8000	0.7331
student_address	address (2)	5+2	5	1.4	0.8000	0.9916
student2	address+sex	5+2	5	1.5	0.8000	0.7297

Bảng 5: Kết quả tối ưu hyperparameters cho Student Performance

Nhận xét:

- DCSI ổn định ở 0.80 – thấp hơn COMPAS nhưng vẫn chấp nhận được.
- Config `student_address` có Balance rất cao (0.99) – gần đạt công bằng hoàn hảo giữa nhóm Urban và Rural.
- $minPts$ nhỏ (4-5) phù hợp với dataset có kích thước nhỏ (649 mẫu).

2.3.2 ε tốt nhất cho mỗi $minPts$

	Sex only			**Address only**			**Address + Sex**		
$minPts$	ε	DCSI	Cl.	ε	DCSI	Cl.	ε	DCSI	Cl.
4	**1.4**	**0.80**	**2**	1.4	0.73	2	–	–	–
5	1.4	0.80	2	**1.4**	**0.80**	**2**	**1.5**	**0.80**	**2**
9	0.6	0.96	2	0.6	0.96	2	0.4	0.93	8
10	0.8	0.93	4	0.8	0.93	4	0.8	0.93	8
15	1.0	0.67	3	1.0	0.67	3	0.8	0.92	4

Bảng 6: 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)**	**5 (+2)**	**4**	**1.4**
Student (Cortez [2008])	**address (2)**	**5 (+2)**	**5**	**1.4**

Bảng 7: So sánh thiết lập thực nghiệm với các dataset trong bài báo gốc

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`.

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.