

Compliance dossier — training data (AI Act, Article 10)

Automatically generated skeleton from the dataset analysis. The "to be completed by the expert" blocks require human review (provenance, compliance judgment, mitigation measures).

1. Identity & purpose

- **System / model:** _____
- **Dossier version:** _____
- **Analysis date:** _____
- **Dataset:** src
- **Intended purpose:** _____

To be completed by the expert — intended purpose and what the data is meant to represent

2. Provenance & legal basis

Origin of each source, licences/contracts, and — for personal data — initial purpose and GDPR legal basis.

To be completed by the expert — provenance per source (separate flow §16), licences, GDPR legal basis

3. Composition (filled automatically)

- **Volume:** 32561 rows, 15 columns.
- **Variables:**

Column	Type
age	BIGINT
workclass	VARCHAR
fnlwgt	BIGINT
education	VARCHAR
education_num	BIGINT
marital_status	VARCHAR
occupation	VARCHAR
relationship	VARCHAR
race	VARCHAR
sex	VARCHAR
capital_gain	BIGINT
capital_loss	BIGINT
hours_per_week	BIGINT
native_country	VARCHAR
income	VARCHAR

- **Declared sensitive attributes:** age, workclass, education, marital_status, occupation,

relationship, race, sex.

To be completed by the expert — geographic / contextual / behavioural scope

4. Preparation (automatic observations)

- **Duplicates:** 24 duplicate row(s) -> deduplication step to be traced.

To be completed by the expert — transformation log: collection, cleaning, labelling, enrichment, aggregation

5. Quality (filled automatically)

- **Accuracy / outliers:** fnlwgt (152).
- **Uniqueness:** 24 duplicate(s).

To be completed by the expert — representativeness and relevance to the purpose

6. Bias

Automatic analysis

Characteristic	Measure	Target	Verdict
Completeness	100.0 %	"e 95.0 %	compliant
Uniqueness	99.9 %	"e 99.0 %	compliant
Validity	97.7 %	"e 98.0 %	to address

workclass — « income »: **disparate impact** (ratio) 5.36 · **statistical parity** (gap) 55.7 % (!)

Group	Count	« income » rate
Self-emp-inc	1116	55.7 %
Federal-gov	960	38.6 %
Local-gov	2093	29.5 %
Self-emp-not-inc	2541	28.5 %
State-gov	1298	27.2 %
Private	22696	21.9 %
?	1836	10.4 %
Without-pay	14	0.0 %
Never-worked	7	0.0 %

education — « income »: **disparate impact** (ratio) 20.75 · **statistical parity** (gap) 74.1 % (!)

Group	Count	« income » rate
Doctorate	413	74.1 %
Prof-school	576	73.4 %
Masters	1723	55.7 %

Bachelors	5355	41.5 %
Assoc-voc	1382	26.1 %
Assoc-acdm	1067	24.8 %
Some-college	7291	19.0 %
HS-grad	10501	16.0 %
12th	433	7.6 %
10th	933	6.6 %
7th-8th	646	6.2 %
9th	514	5.3 %
11th	1175	5.1 %
5th-6th	333	4.8 %
1st-4th	168	3.6 %
Preschool	51	0.0 %

marital_status — « income »: **disparate impact** (ratio) 9.72 · **statistical parity** (gap) 40.1 % (!)

Group	Count	« income » rate
Married-civ-spouse	14976	44.7 %
Married-AF-spouse	23	43.5 %
Divorced	4443	10.4 %
Widowed	993	8.6 %
Married-spouse-absent	418	8.1 %
Separated	1025	6.4 %
Never-married	10683	4.6 %

occupation — « income »: **disparate impact** (ratio) 72.12 · **statistical parity** (gap) 47.7 % (!)

Group	Count	« income » rate
Exec-managerial	4066	48.4 %
Prof-specialty	4140	44.9 %
Protective-serv	649	32.5 %
Tech-support	928	30.5 %
Sales	3650	26.9 %
Craft-repair	4099	22.7 %
Transport-moving	1597	20.0 %
Adm-clerical	3770	13.4 %

Machine-op-inspct	2002	12.5 %
Farming-fishing	994	11.6 %
Armed-Forces	9	11.1 %
?	1843	10.4 %
Handlers-cleaners	1370	6.3 %
Other-service	3295	4.2 %
Priv-house-serv	149	0.7 %

relationship — « income »: **disparate impact** (ratio) 35.94 · **statistical parity** (gap) 46.2 % (!)

Group	Count	« income » rate
Wife	1568	47.5 %
Husband	13193	44.9 %
Not-in-family	8305	10.3 %
Unmarried	3446	6.3 %
Other-relative	981	3.8 %
Own-child	5068	1.3 %

race — « income »: **disparate impact** (ratio) 2.88 · **statistical parity** (gap) 17.3 % (!)

Group	Count	« income » rate
Asian-Pac-Islander	1039	26.6 %
White	27816	25.6 %
Black	3124	12.4 %
Amer-Indian-Eskimo	311	11.6 %
Other	271	9.2 %

sex — « income »: **disparate impact** (ratio) 2.79 · **statistical parity** (gap) 19.6 % (!)

Group	Count	« income » rate
Male	21790	30.6 %
Female	10771	10.9 %

Judgment & measures

To be completed by the expert — impact on health/safety/fundamental rights and detection/prevention/mitigation measures

7. Gaps & limitations

Detected points to examine as potential gaps:

- [medium] duplicates
- [low] outliers — fnlwgt
- [low] outliers — hours_per_week
- [high] bias — workclass
- [high] bias — education
- [high] bias — marital_status
- [high] bias — occupation
- [high] bias — relationship
- [high] bias — race
- [high] bias — sex

To be completed by the expert — identified and addressed gaps; out-of-scope uses

8. Governance & traceability

- **Analysis tool version:** 0.4.1
- **Dataset fingerprint (schema + volume):** c09f6f88fcbcff58

To be completed by the expert — responsibilities, audit log, versioning, maintenance/updates