

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.

(!) **Potential personal data detected:** name, first, last, dob — a GDPR legal basis is required for these columns.

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

3. Composition (filled automatically)

- **Volume:** 7214 rows, 53 columns.
- **Data period** (compas_screening_date): 2013-01-01 -> 2014-12-31.
- **Variables:**

Column	Type
id	BIGINT
name	VARCHAR
first	VARCHAR
last	VARCHAR
compas_screening_date	DATE
sex	VARCHAR
dob	DATE
age	BIGINT
age_cat	VARCHAR
race	VARCHAR
juv_fel_count	BIGINT
decile_score	BIGINT
juv_misd_count	BIGINT
juv_other_count	BIGINT

priors_count	BIGINT
days_b_screening_arrest	BIGINT
c_jail_in	TIMESTAMP
c_jail_out	TIMESTAMP
c_case_number	VARCHAR
c_offense_date	DATE
c_arrest_date	DATE
c_days_from_compas	BIGINT
c_charge_degree	VARCHAR
c_charge_desc	VARCHAR
is_recid	BIGINT
r_case_number	VARCHAR
r_charge_degree	VARCHAR
r_days_from_arrest	BIGINT
r_offense_date	DATE
r_charge_desc	VARCHAR
r_jail_in	DATE
r_jail_out	DATE
violent_recid	VARCHAR
is_violent_recid	BIGINT
vr_case_number	VARCHAR
vr_charge_degree	VARCHAR
vr_offense_date	DATE
vr_charge_desc	VARCHAR
type_of_assessment	VARCHAR
decile_score_1	BIGINT
score_text	VARCHAR
screening_date	DATE
v_type_of_assessment	VARCHAR
v_decile_score	BIGINT
v_score_text	VARCHAR
v_screening_date	DATE
in_custody	DATE

out_custody	DATE
priors_count_1	BIGINT
start	BIGINT
end	BIGINT
event	BIGINT
two_year_recid	BIGINT

- **Declared sensitive attributes:** race, sex, age_cat.
- **Detected PII columns:** name, first, last, dob.

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

4. Preparation (automatic observations)

- **Missing values:** days_b_screening_arrest (307) -> handling strategy to be documented.

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

5. Quality (filled automatically)

- **Completeness:** 307 missing value(s) (days_b_screening_arrest).
- **Accuracy / outliers:** age (1).

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

6. Bias

Automatic analysis

Characteristic	Measure	Target	Verdict
Completeness	81.4 %	"e 95.0 %	to address
Uniqueness	100.0 %	"e 99.0 %	compliant
Validity	96.7 %	"e 98.0 %	to address

race — « score_text »: **disparate impact** (ratio) 4.83 · **statistical parity** (gap) 26.4 % (!)

(!) **Non-robust** disparity: smallest group n=18 — not statistically significant or driven by a small sample. Do not read as an established bias.

Group	Count	« score_text » rate
Native American	18	33.3 %
African-American	3696	27.7 %
Caucasian	2454	11.2 %
Hispanic	637	10.5 %
Asian	32	9.4 %
Other	377	6.9 %

sex — « score_text »: **disparate impact** (ratio) 1.53 · **statistical parity** (gap) 7.2 % (!)

Group	Count	« score_text » rate
Male	5819	20.8 %
Female	1395	13.6 %

age_cat — « score_text »: **disparate impact** (ratio) 3.65 · **statistical parity** (gap) 21.5 % (!)

Group	Count	« score_text » rate
Less than 25	1529	29.6 %
25 - 45	4109	20.0 %
Greater than 45	1576	8.1 %

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:

- [low] missing — days_b_screening_arrest
- [low] missing — c_jail_in
- [low] missing — c_jail_out
- [low] missing — c_case_number
- [low] missing — c_offense_date
- [high] missing — c_arrest_date
- [low] missing — c_days_from_compas
- [low] missing — c_charge_desc
- [high] missing — r_case_number
- [high] missing — r_charge_degree
- [high] missing — r_days_from_arrest
- [high] missing — r_offense_date
- [high] missing — r_charge_desc
- [high] missing — r_jail_in
- [high] missing — r_jail_out
- [high] missing — violent_recid
- [high] missing — vr_case_number
- [high] missing — vr_charge_degree
- [high] missing — vr_offense_date
- [high] missing — vr_charge_desc
- [low] missing — in_custody
- [low] missing — out_custody
- [low] outliers — age
- [low] outliers — priors_count
- [low] outliers — days_b_screening_arrest
- [low] outliers — c_days_from_compas
- [low] outliers — r_days_from_arrest
- [low] outliers — priors_count_1
- [low] outliers — start
- [high] pii — name
- [high] pii — first
- [high] pii — last
- [high] pii — dob
- [low] bias — race
- [medium] bias — sex

- [high] bias — age_cat
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):** fee3a183fbb13d6f
To be completed by the expert — responsibilities, audit log, versioning, maintenance/updates