Inteligencia artificial avanzada para la ciencia de datos 1

Identificación del problema y sus datos

Grupo: 101

Profesor: Gildardo Sánchez

Cristóbal Camarena Hernández - A01642653

agosto 2025

- ¿De que tamaño es el (o los) dataset(s)?

```
[8]  df.shape
     (319464, 32)

[9]  v1.shape
     (329356, 32)

     v2.shape
     (278704, 34)

[11] v3.shape
     (282867, 34)

     v4.shape
     (339062, 32)
```

- ¿Qué contiene cada uno?

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 476836 entries, 0 to 476835
Data columns (total 32 columns):
 #   Column                             Non-Null Count   Dtype
---  ------                             --------------   -----
 0   fraud_bool                         476836 non-null  int64
 1   income                             476836 non-null  float64
 2   name_email_similarity              476836 non-null  float64
 3   prev_address_months_count          476836 non-null  int64
 4   current_address_months_count       476836 non-null  int64
 5   customer_age                       476836 non-null  int64
 6   days_since_request                 476836 non-null  float64
 7   intended_balcon_amount             476836 non-null  float64
 8   payment_type                       476836 non-null  object
 9   zip_count_4w                       476836 non-null  int64
 10  velocity_6h                        476836 non-null  float64
 11  velocity_24h                       476836 non-null  float64
 12  velocity_4w                        476836 non-null  float64
 13  bank_branch_count_8w               476836 non-null  int64
 14  date_of_birth_distinct_emails_4w   476836 non-null  int64
 15  employment_status                  476836 non-null  object
 16  credit_risk_score                  476836 non-null  int64
 17  email_is_free                      476836 non-null  int64
 18  housing_status                     476836 non-null  object
 19  phone_home_valid                   476836 non-null  int64
 20  phone_mobile_valid                 476836 non-null  int64
 21  bank_months_count                  476836 non-null  int64
 22  has_other_cards                    476836 non-null  int64
 23  proposed_credit_limit              476836 non-null  float64
 24  foreign_request                    476835 non-null  float64
 25  source                             476835 non-null  object
 26  session_length_in_minutes          476835 non-null  float64
 27  device_os                          476835 non-null  object
 28  keep_alive_session                 476835 non-null  float64
 29  device_distinct_emails_8w          476835 non-null  float64
 30  device_fraud_count                 476835 non-null  float64
 31  month                              476835 non-null  float64
dtypes: float64(14), int64(13), object(5)
memory usage: 116.4+ MB
```

```
v1.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 329356 entries, 0 to 329355
Data columns (total 32 columns):
 #   Column                             Non-Null Count   Dtype
---  ------                             --------------   -----
 0   fraud_bool                         329356 non-null  int64
 1   income                             329356 non-null  float64
 2   name_email_similarity              329356 non-null  float64
 3   prev_address_months_count          329356 non-null  int64
 4   current_address_months_count       329356 non-null  int64
 5   customer_age                       329356 non-null  int64
 6   days_since_request                 329356 non-null  float64
 7   intended_balcon_amount             329356 non-null  float64
 8   payment_type                       329356 non-null  object
 9   zip_count_4w                       329356 non-null  int64
 10  velocity_6h                        329356 non-null  float64
 11  velocity_24h                       329356 non-null  float64
 12  velocity_4w                        329356 non-null  float64
 13  bank_branch_count_8w               329356 non-null  int64
 14  date_of_birth_distinct_emails_4w   329356 non-null  int64
 15  employment_status                  329356 non-null  object
 16  credit_risk_score                  329356 non-null  int64
 17  email_is_free                      329355 non-null  float64
 18  housing_status                     329355 non-null  object
 19  phone_home_valid                   329355 non-null  float64
 20  phone_mobile_valid                 329355 non-null  float64
 21  bank_months_count                  329355 non-null  float64
 22  has_other_cards                    329355 non-null  float64
 23  proposed_credit_limit              329355 non-null  float64
 24  foreign_request                    329355 non-null  float64
 25  source                             329355 non-null  object
 26  session_length_in_minutes          329355 non-null  float64
 27  device_os                          329355 non-null  object
 28  keep_alive_session                 329355 non-null  float64
 29  device_distinct_emails_8w          329355 non-null  float64
 30  device_fraud_count                 329355 non-null  float64
 31  month                              329355 non-null  float64
dtypes: float64(19), int64(8), object(5)
```

```
v2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 278704 entries, 0 to 278703
Data columns (total 34 columns):
 #   Column                              Non-Null Count   Dtype
---  ------                              --------------   -----
 0   fraud_bool                          278704 non-null  int64
 1   income                              278704 non-null  float64
 2   name_email_similarity               278704 non-null  float64
 3   prev_address_months_count           278704 non-null  int64
 4   current_address_months_count        278704 non-null  int64
 5   customer_age                        278704 non-null  int64
 6   days_since_request                  278704 non-null  float64
 7   intended_balcon_amount              278704 non-null  float64
 8   payment_type                        278704 non-null  object
 9   zip_count_4w                        278704 non-null  int64
 10  velocity_6h                         278704 non-null  float64
 11  velocity_24h                        278704 non-null  float64
 12  velocity_4w                         278703 non-null  float64
 13  bank_branch_count_8w                278703 non-null  float64
 14  date_of_birth_distinct_emails_4w    278703 non-null  float64
 15  employment_status                   278703 non-null  object
 16  credit_risk_score                   278703 non-null  float64
 17  email_is_free                       278703 non-null  float64
 18  housing_status                      278703 non-null  object
 19  phone_home_valid                    278703 non-null  float64
 20  phone_mobile_valid                  278703 non-null  float64
 21  bank_months_count                   278703 non-null  float64
 22  has_other_cards                     278703 non-null  float64
 23  proposed_credit_limit               278703 non-null  float64
 24  foreign_request                     278703 non-null  float64
 25  source                              278703 non-null  object
 26  session_length_in_minutes           278703 non-null  float64
 27  device_os                           278703 non-null  object
 28  keep_alive_session                  278703 non-null  float64
 29  device_distinct_emails_8w           278703 non-null  float64
 30  device_fraud_count                  278703 non-null  float64
 31  month                               278703 non-null  float64
 32  x1                                  278703 non-null  float64
 33  x2                                  278703 non-null  float64
dtypes: float64(24), int64(5), object(5)
```

```
v3.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 282867 entries, 0 to 282866
Data columns (total 34 columns):
 #   Column                              Non-Null Count   Dtype
---  ------                              --------------   -----
 0   fraud_bool                          282867 non-null  int64
 1   income                              282867 non-null  float64
 2   name_email_similarity               282867 non-null  float64
 3   prev_address_months_count           282867 non-null  int64
 4   current_address_months_count        282867 non-null  int64
 5   customer_age                        282867 non-null  int64
 6   days_since_request                  282867 non-null  float64
 7   intended_balcon_amount              282867 non-null  float64
 8   payment_type                        282867 non-null  object
 9   zip_count_4w                        282867 non-null  int64
 10  velocity_6h                         282867 non-null  float64
 11  velocity_24h                        282867 non-null  float64
 12  velocity_4w                         282867 non-null  float64
 13  bank_branch_count_8w                282867 non-null  int64
 14  date_of_birth_distinct_emails_4w    282867 non-null  int64
 15  employment_status                   282867 non-null  object
 16  credit_risk_score                   282867 non-null  int64
 17  email_is_free                       282867 non-null  int64
 18  housing_status                      282867 non-null  object
 19  phone_home_valid                    282867 non-null  int64
 20  phone_mobile_valid                  282867 non-null  int64
 21  bank_months_count                   282867 non-null  int64
 22  has_other_cards                     282867 non-null  int64
 23  proposed_credit_limit               282867 non-null  float64
 24  foreign_request                     282867 non-null  int64
 25  source                              282867 non-null  object
 26  session_length_in_minutes           282867 non-null  float64
 27  device_os                           282867 non-null  object
 28  keep_alive_session                  282867 non-null  int64
 29  device_distinct_emails_8w           282867 non-null  int64
 30  device_fraud_count                  282867 non-null  int64
 31  month                               282867 non-null  int64
 32  x1                                  282867 non-null  float64
 33  x2                                  282867 non-null  float64
dtypes: float64(11), int64(18), object(5)
```

```
v4.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 339062 entries, 0 to 339061
Data columns (total 32 columns):
 #   Column                              Non-Null Count   Dtype
---  ------                              --------------   -----
 0   fraud_bool                          339062 non-null  int64
 1   income                              339062 non-null  float64
 2   name_email_similarity               339062 non-null  float64
 3   prev_address_months_count           339062 non-null  int64
 4   current_address_months_count        339062 non-null  int64
 5   customer_age                        339062 non-null  int64
 6   days_since_request                  339062 non-null  float64
 7   intended_balcon_amount              339062 non-null  float64
 8   payment_type                        339062 non-null  object
 9   zip_count_4w                        339062 non-null  int64
 10  velocity_6h                         339062 non-null  float64
 11  velocity_24h                        339062 non-null  float64
 12  velocity_4w                         339062 non-null  float64
 13  bank_branch_count_8w                339062 non-null  int64
 14  date_of_birth_distinct_emails_4w    339062 non-null  int64
 15  employment_status                   339062 non-null  object
 16  credit_risk_score                   339062 non-null  int64
 17  email_is_free                       339062 non-null  int64
 18  housing_status                      339062 non-null  object
 19  phone_home_valid                    339062 non-null  int64
 20  phone_mobile_valid                  339062 non-null  int64
 21  bank_months_count                   339062 non-null  int64
 22  has_other_cards                     339062 non-null  int64
 23  proposed_credit_limit               339062 non-null  float64
 24  foreign_request                     339062 non-null  int64
 25  source                              339061 non-null  object
 26  session_length_in_minutes           339061 non-null  float64
 27  device_os                           339061 non-null  object
 28  keep_alive_session                  339061 non-null  float64
 29  device_distinct_emails_8w           339061 non-null  float64
 30  device_fraud_count                  339061 non-null  float64
 31  month                               339061 non-null  float64
dtypes: float64(13), int64(14), object(5)
```

- ¿Cuáles son las diferencias entre ellos?

Las principales diferencias son que todos los datasets tienen diferentes filas. También las datasets 2 y 3 tienen más variantes (x1 y x2). Los tipos de datos también varian según el dataset.

- ¿Qué significa el concepto de Fair ML o Fair AI y porque en este contexto resulta muy relevante?

Fair ML / Fair AI es hacer modelos que no traten peor a unas personas que a otras por pertenecer a cierto grupo (edad, ingreso, empleo, etc.). La idea es que la decisión del modelo no dependa injustamente de esos atributos. Por eso hoy evaluar la fairness ya es práctica común en ML, sobre todo en problemas con impacto real en la vida de la gente.

Es importante porque un "positivo" del modelo significa negar abrir una cuenta bancaria. Un falso positivo no es solo un error técnico: le puede cerrar la puerta del sistema financiero a alguien. El propio estudio lo plantea como un dominio "punitivo", donde la decisión afecta directamente el acceso a un servicio básico.

- ¿Implicaciones de usar solo uno de los dataset?

Usar un solo dataset puede generar representatividad limitada, pérdida de patrones, sesgo incrementado, menor robustez, conjunto de variables incompleto, y generalización pobre.