Informe Tiresias

Auditoria de l'algorisme RisCanvi

Versió
9 de gener de 2024

Autoria
Dribia Data Research

DRIBIA

Generalitat de Catalunya
Departament de Justícia,
Drets i Memòria
Direcció General
d'Afers Penitenciaris

Continguts

Resum executiu

S'ha auditat l'algoritme RisCanvi, analitzant els resultats actuals per als següents riscos: violència autodirigida (VAD), intrainstitucional (VII), trencament condemna (TRC), reincidència general (RG) i reincidència violenta (RV). També s'ha fet una proposta de millora tant algorítmica com de transparència. Les conclusions són les següents:

1. S'han analitzat dues bases de dades diferents per avaluar els 5 algoritmes d'estimació de risc. La majoria dels resultats mostrats en aquest estudi corresponen a reincidències que queden gravades al sistema del SIPC (Sistema d'Informació Penitenciari Català) com a *incidents*: intrainstitucional, autodirigida, i trencament de condemna. També s'ha analitzat un conjunt de dades molt més petit que ha estat creat pel Centre d'Estudis Jurídics i Formació Especialitzada (CEJFE) per auditar els algoritmes de Reincidència General i Violenta.

2. Els resultats d'AUC-ROC dels algoritmes actuals per als riscos avaluats són similars o lleugerament millors als reportats l'any 2017. Els nivells d'AUC-ROC són comparables a altres algoritmes en el panorama internacional (COMPAS 0,7 [9], OASys ~0,75 [10]).

AUR-ROC	Viol. Auto.	Viol Intra.	T. Condemna	R. General	R.Violenta
Screening	0,85	0,76	0,57	0,68	0,68
Completa	0,8	0,75	0,64	0,65	0,68

3. Donats els punts de tall actual, si entenem positiu com un intern que reincideix, que serà veritable si se l'ha etiquetat com a risc alt, i fals negatiu si, per contra, se l'ha etiquetat com a risc mig o baix, les taxes de falsos positius (FPR), falsos negatius (FNR), veritables positius (TPR) i veritables negatius (TNR) són les següents:

RISC	ESCALA	FPR	FNR	TPR	TNR
Trencament condemna	SCREENING	3,63%	100,00%	0,00%	96,37%
	COMPLETA	23,70%	56,00%	44,00%	76,30%
Violència autodirigida	SCREENING	1,17%	88,87%	11,13%	98,83%
	COMPLETA	12,34%	51,36%	48,64%	87,66%
Reincidència general	SCREENING	31,12%	39,97%	60,03%	68,88%
	COMPLETA	6,39%	81,28%	18,72%	93,61%
Reincidència violenta	SCREENING	4,77%	84,25%	15,75%	95,23%
	COMPLETA	9,93%	72,29%	27,71%	90,07%
Violència intrainstitucional	SCREENING	6,95%	71,71%	28,29%	93,05%
	COMPLETA	21,48%	43,87%	56,13%	78,52%

Taxes de falsos positius i negatius i de vertaders positius i negatius [1]

En aquesta taula s'observa que els punts de tall escollits en totes les violències prioritzen tenir una taxa de falsos positius baixa (FPR entre 1 i 31 %) per reduir els casos de predicció de risc alt erronis. Això vol dir que Riscanvi té més tendència a donar un risc baix o mig. Aquest fet implica inevitablement tenir un nombre de falsos negatius alt, com es veu a la taula (FNR entre 40 i 100%). Això, sumat a l'extrema baixa reincidència (~5%), fa que el model no identifiqui com a risc alt correctament els interns que sí que acaben reincidint (TPR), sobretot en el trencament de condemna (on la prevalència és encara inferior, 0,5 %). Finalment, degut també a la baixa reincidència, els interns amb risc baix o mig no acaben reincidint (TNR > 76 % en escala completa i > 69 % en screening) en totes les violències.

4. Tot i tenir un AUC-ROC prou elevat, és impossible aconseguir uns **punts de tall** que donin uns nivells d'encert en tots els indicadors. La decisió dels punts de tall necessita un alt coneixement de l'ús de l'eina i de les seves implicacions i no recomanem optimitzar-los només des del punt de vista algorítmic. Això sí, nosaltres **recomanem la revisió per experts de l'ús de RisCanvi i de l'algoritme** per tal d'intentar ajustar els punts de tall als objectius desitjats.

5. **No s'han detectat biaixos discriminatoris greus en el RisCanvi Complet** respecte a grups protegits en termes de sexe, edat ni nacionalitat. Només hi ha dues excepcions. Una és la VII, la RG i la RV, on, sí que hi ha més error amb el grup protegit (<30 anys), tot i que dins el marge de disparitat acceptat (menys d'1,2). L'altra és amb l'**algoritme Screening** i per Violència Autodirigida, hi ha més proporció de **falsos negatius en els casos protegits** (dones, estrangers i menors de 30) i, per tant, una subestimació del risc. També, trobem una major proporció d'error en l'escala Screning per predir la Reincidència General i Violenta en estrangers.

[1] Per a aquesta anàlisi només s'han revisat els resultats de l'algoritme, sense tenir en compte el protocol d'actuació associat als resultats (p. ex. fer un RisCanvi Completa pels 5 tipus de reincidència quan només un dels Screenings dona risc alt, o altres tipus de recomanacions).

6. En la versió actual de punts de tall, l'algoritme **es comporta diferent quan s'analitza per tipus de delicte o situació de l'intern.** Pel RisCanvi Completa, l'algoritme té més taxa de falsos positius pels interns amb delictes contra la propietat. Per RisCanvi Screening, l'algoritme té més taxa de falsos negatius pels interns amb delictes contra la salut pública en el cas de Violència Autodirigida. No hem trobat patrons d'error de predicció en funció de la situació de l'intern.

7. Avui en dia, comptem amb més dades que quan es va dissenyar l'algoritme original. A més, hi ha hagut un gran nombre d'avanços tecnològics i en la matèria de la intel·ligència artificial durant l'última dècada. En aquesta auditoria, hem entrenat un **nou algoritme d'aprenentatge automàtic,** conegut com a *Catboost*.

    - Hem comprovat que la **millora** en les prediccions del model que proposem són **molt remarcables**. En termes d'**AUC-ROC** són entre un **3 i un 30 % millors que en l'algoritme actual**.
    - Mantenint uns punts de tall similars al model actual es podria, fixant l'error de falsos positius al valor actual, aconseguir aproximadament entre 10 i 20 punts percentuals més d'encert en assignar risc alt a interns que acaben reincidint. O, per contra, mantenint els positius veritables, aproximadament es podria reduir entre un 20 % i un 50 % l'assignació risc alt a interns que acaben no reincidint.
    - Dels 43 factors de risc analitzats, l'algoritme troba més importants els següents:

RISC	TOP1	TOP2	TOP3	TOP4	TO5	TOP6	TOP7	TOP8	TOP9	TOP10
Viol. autod.	F37	F2	F10	F30	F21	F36	F26	F5	F19	F41
T. condemna	F38	F34	F33	F2	F9	F18	F5	F14	F30	F40
Violència Intra.	F10	F12	F2	F30	F43	F38	F37	F5	F6	F19
Reinc. General.	F8	F30	F12	F26	F19	F14	F38	F9	F29	F2
Reinc. Violenta	F8	F7	F12	F30	F41	F10	F15	F26	F21	F9

Factors de risc (FR) més importants per la predicció de la incidència segons l'algoritme Catboost. En blau, els FR considerats en l'algoritme actual de RisCanvi Complet.

En la taula anterior, en blau, hem marcat els factors de risc que ja es consideren en cadascun dels algoritmes actuals. Veiem que mentre que pel risc de Violència Intrainstitucional coincideixen gairebé totes les variables, pel risc de Trencament de condemna **l'algoritme Catboost dona més importància a variables que ara mateix no es contemplen**. De fet, és en el risc de Trencament de condemna on experimentem una millora de resultats més notable en termes d'AUC-ROC.

També és interessant veure d'aquests factors, quins es consideren en l'algoritme RisCanvi Screening actual. En la següent taula hem marcat en taronja els factors de la Completa que tenen un pes en el model de Screening actual[2].

RISC	TOP1	TOP2	TOP3	TOP4	TOP5	TOP6	TOP7	TOP8	TOP9	TOP10
Viol. autod.	F37	F2	F10	F30	F21	F36	F26	F5	F19	F41
T. condemna	F38	F34	F33	F2	F9	F18	F5	F14	F30	F40
Violència Intra.	F10	F12	F2	F30	F43	F38	F37	F5	F6	F19
Reinc. General.	F8	F30	F12	F26	F19	F14	F38	F9	F29	F2
Reinc. Violenta	F8	F7	F12	F30	F41	F10	F15	F26	F21	F9

Factors de risc (FR) més importants per la predicció de la incidència segons l'algoritme Catboost. En taronja, els FR considerats en l'algoritme actual de RisCanvi Screening.

Veiem que dels 10 factors de risc més rellevants (d'entre els 43 totals) segons el nou model, l'algoritme RisCanvi actual de l'escala Screening en Violència Autodirigida i Intrainstitucional, només en considera dos. A més a més, pel risc de Trencament de condemna, no hi ha cap coincidència. És a dir, podem dir que hi ha informació rellevant per la predicció d'incidència que no està sent contemplada en la fase de cribratge (escala Screening).

- El nou algoritme permet, a més,
  - **Mantenir** l'**explicabilitat d**e la predicció.
  - Afegeix detalls d'explicabilitat a nivell **d'avaluació individual** (contribució específica d'un factor per a cada cas concret). Aquest punt creiem que seria un afegit diferencial, ja que habilitaria als experts a interpretar el nivell de risc que dona l'algoritme a un intern concret per a una avaluació concreta.
  - **Reentrenar el model amb noves dades de forma fàcil i ràpida.**

8. De la revisió del programari actual del càlcul de risc de Riscanvi concloem que:

- Està en un llenguatge (psql) que no permet l'ús d'algoritmes d'intel·ligència artificial moderns.

[2] Noteu que no hi ha una correspondència 1 a 1 entre els factors de l'escala Completa i Screening. Per exemple, el F5 de l'escala Screening (*Problemes amb el consum de drogues o alcohol*) correspondria al conjut de F30 (*Abús o dependència de les drogues*) i F31 (*Abús o dependència a l'alcohol*) de l'escala Completa.

- Que no té documentació i, per tant, depèn totalment de l'implementador de l'algoritme, que és un consultor concret d'un proveïdor extern. Aquest fet, per exemple, fa que qualsevol consulta o modificació depengui exclusivament d'aquesta persona dificultant-ne l'accés, la compartició, revisió i modificació.
- No té un repositori propi i, en conseqüència, està integrat amb tot el codi de la base de dades de Riscanvi, fent-lo menys independent i dificultant la seva compartició, revisió i modificació.

9. La plataforma usada no s'integra totalment amb totes les fonts de dades disponibles al SIPC. En conseqüència, no es poden validar les evidències introduïdes de forma automàtica o assistida mitjançant un algoritme de validació amb suport.

10. Feta la revisió de l'algoritme i del sistema **es recomana**

a. Sistematitzar la recollida de dades d'incidència / reincidència o no incidència / reincidència d'un intern, tant dins del sistema penitenciari català com en societat, que permet l'anàlisi periòdica de la precisió de l'algoritme per a totes les reincidències o violències.

b. Implementar un nou model predictiu basat en models d'intel·ligència artificial més moderns, com el Catboost analitzat en aquest informe, amb nous factors de risc i punts de tall. Aquest model s'ha de desenvolupar en col·laboració d'experts en reincidència amb experts en algoritmes d'intel·ligència artificial. Si això no fos possible, com a mínim s'haurien de reavaluar els punts de tall dels models actuals.

c. Desplegar el nou algoritme amb programari més modern, mitjançant un microservei connectat via API amb la base de dades i interfície gràfica actual, que permeti la implementació del nou algoritme.

d. Crear un repositori del codi de l'algoritme així com d'una documentació viva que permeti l'accés ràpid a qualsevol usuari que vulgui consultar o modificar l'algoritme actual.

e. Obrir les dades del model per a anàlisi científica i potencialment al públic en general (anonimitzades adequadament).

f. Fer públic el codi i la documentació de l'algoritme per transparència i a revisió per part de tercers i possibles propostes de millora.

g. Implementar un equip de manteniment, revisió i millora contínua de l'algoritme.

h. Habilitar la generació automàtica d'un informe periòdic, amb periodicitat mínima semestral, de precisió i control de biaixos.

Objectius i estructura del document

Els objectius d'aquesta auditoria són:

- Auditoria algorítmica de RisCanvi.
  a. Anàlisi de resultats de l'algoritme actual.
  b. Avaluació de biaixos del model actual.
  c. Tipus de reincidències considerades: Violència Autodirigida, Violència Intrainstitucional, Reincidència Violenta, Trencament condemna, Reincidència General.
- Proposta de nou model de predicció.
- Recomanacions d'accions concretes.

D'acord amb aquests objectius hem estructurat el document de la forma següent.

1. Context de l'auditoria i terminologia important
2. Dades: descripció de les dades necessàries utilitzades per a l'algoritme: d'on han sortit les dades, quines limitacions ens hem trobat, quines transformacions i una anàlisi bàsica.
3. Algoritme actual: avaluació de l'algoritme actual: com funciona, quins resultats obté, i si mostra o no biaixos.
4. Nou algoritme: Es descriuen una proposta de nou algoritme i se'n descriu el seu funcionament, els resultats que se n'obtindrien i com contribuirien a l'explicabilitat del resultat final.
5. Implementació: comentaris sobre la implementació actual de RisCanvi a nivell tècnic, de programari i d'experiència d'usuari.
6. Conclusions: Conclusions generals de l'auditoria i recomanacions específiques.

Context

RisCanvi és una eina i conjunt de protocols d'actuació professional per a la valoració i gestió del risc de comportaments violents i trencaments de condemna en l'àmbit penitenciari i comunitari de les persones internades en centres ordinaris i en llibertat condicionals al sistema penitenciari català.
L'eina es va dissenyar per millorar les prediccions individualitzades de risc de cometre reincidència general, reincidència violenta o violència intrainstitucional, tenir conductes de violència autodirigida, o trencar permisos.
RisCanvi considera dos tipus d'avaluacions:

- **Screening**. Versió simplificada d'avaluació, on només es revisen 10 factors i la sortida pot ser risc Alt o Baix. Si per a algun dels tipus de reincidència surt un factor Alt o hi ha recomanació específica, es passa a fer l'avaluació Completa. En alguns casos, recollits al protocol, com per exemple delictes de base violenta, es passa a fer directament l'avaluació Completa.
- **Completa**. Versió estesa de l'avaluació. Es consideren 43 factors i la sortida pot ser risc Alt, Mitjà o Baix o recomanació específica.

La primera versió de Riscanvi s'implementà l'any 2009. Se n'han fet dues revisions:

- 2011: modificació de punts de tall, [2]
- 2017: inclusió d'un nou risc a predir, reincidència general, i revisió de l'algoritme per Trencament de Condemna. [3]

L'any 2010 es va editar l'informe inicial de RisCanvi [1] i l'any 2017 es va fer una revisió externa de l'algoritme [4].

Per a la revisió de RisCanvis es necessita accés a

a. La descripció i documentació de l'algoritme.
b. Les dades d'entrada i de sortida de l'algoritme.
c. Les dades de validació (incidències/reincidències passades associades a cada Riscanvi).
d. La implementació en codi de l'algoritme amb la seva documentació.

Aquest document recull els resultats de l'auditoria feta durant l'últim trimestre de 2022 i el segon i tercer trimestre de 2023.

Dades

Per tal de poder dur a terme l'auditoria algorítmica, és necessari tenir accés a les dades d'entrada i sortida de l'algoritme RisCanvi. Les dades d'entrada fan referència a totes les variables que potencialment es podrien fer servir per predir el nivell de risc de reincidència. Exemples de dades d'entrada són les variables demogràfiques, els factors de risc o les variables relacionades amb la pena. Les dades de sortida de l'algoritme és la probabilitat de risc de reincidència (alt/baix per a Screening i alt/mitjà/baix per a Completa). A més, per poder avaluar i entrenar nous algoritmes, és necessari disposar de les dades reals de reincidència, i.e., per cada avaluació s'ha de poder identificar si l'intern avaluat ha comès o no alguna de les 5 reincidències en un interval de temps determinat.

En aquesta secció explicarem el corpus de dades al que s'ha tingut accés durant el projecte, el procés de neteja i transformació que s'ha aplicat i presentem l'anàlisi inicial de les dades proporcionades.

Origen de les dades

Les dades del sistema RisCanvi utilitzades en aquesta auditoria provenen de dues extraccions de dades i documentació interna de la direcció general d'Afers Penitenciaris (AP) i del Centre d'Estudis Jurídics i Formació Especialitzada (CEJFE).

Dades i documentació AP

Les dades d'Afers Penitenciaris provenen d'una extracció executada per un proveïdor extern del departament. S'han facilitat dos fitxers tabulats que contenen les dades *crues*:

- *extraccio_rc.csv*: informació de les avaluacions i Incidències registrades dins el sistema català de presons, que permet identificar 3 violències (intrainstitucional, autodirigida i trencament de condemna).
- *extraccio_permisossortides.csv*: informació dels permisos de sortides.

També s'ha elaborat amb la direcció general el document:

- *Tiresias - RisCanvi - Mapping variables.ods*: informació de la classificació dels interns segons descripció i de tipus de reincidència associada a incidents.

La descripció de les columnes dels tres fitxers (els dos tabulats i el de *mapping*) es troben al document annex: *Tiresias - RisCanvi - Dades d'entrada.ods*: per a cada un dels documents, es descriu el contingut de les columnes.

Dades i documentació CEFJE

Durant la segona iteració del projecte, el CEJFE va publicar una investigació pròpia [12] amb dades obertes per a la recerca. Els registres d'aquest fitxer tabulat contenen informació

d'excarcerats el 2015 dels quals es va fer un seguiment per detectar reincidència en societat fins al 2020. Aquest subconjunt de la població era susceptible de cometre reincidència general o reincidència violenta, rellevant per a la comprovació del funcionament de RisCanvi per a aquests dos tipus de reincidència. En col·laboració entre AP i el CEJFE es va aconseguir una versió estesa d'aquestes dades que incloïa una identificació que permetés creuar-les amb les anàlisis fetes amb les dades d'AP.

A part de les dades, l'equip del CEFJE també ha publicat el document *Plantilla_variables_catala_Taxa_2020.pdf on* es descriuen breument les variables, el format i la seva codificació.

Neteja i transformació de dades i consideracions

Per tal de poder executar l'anàlisi, s'ha realitzat una neteja de les dades *crues.* A continuació, detallem els procediments aplicats:

Dades extracció AP

- S'ha aplicat un algoritme de *hash MD5* sobre la columna *NIS*, que identifica als interns, per garantir l'anonimitzat de les dades.
- S'han retirat les columnes *FAC_44* i *FAC_45*, ja que no hi havia cap registre amb dades.
- Els valors sense dades a la columna *VC_GEN_CONSELLS,* s'ha substituït per *No,* per tal d'indicar que no hi ha recomanació d'escala específica.
- S'han corregit dos valors negatius al camp *EDAT*.
- S'han simplificat els textos lliures al camp *CLASSIFICACIÓ* d'acord amb el document de mapping (*Tiresias - RisCanvi - Mapping variables.ods*) validat per la Direcció General d'Afers Penitenciaris.
- S'ha unificat el format per les dates.
- S'ha unificat el format pels camps buits al valor *Sense Dades.*
- S'ha creat la columna *ORIGEN* a partir de la columna *NACIONALITAT.* Aquest camp dicotòmic pot prendre dos valors: *Espanya* o *Estranger.*
- S'ha creat la columna *EDAT_CAT* a partir de la columna *EDAT.* Aquest camp dicotòmic pot prendre dos valors: *<30 anys* o *≥30 anys.*
- S'ha creat la columna *RISCANVI_SORTIDA*, que indica si es tracta de l'última avaluació RisCanvi abans de la sortida de l'intern. Aquestes avaluacions tenen la característica que tots els registres als camps de factors de risc tenen el valor *Sense Dades.*

Per tal de poder avaluar l'actual algoritme de RisCanvi i entrenar nous algoritmes d'aprenentatge automàtic (o *Machine Learning*), és necessari disposar de les dades de reincidència real associades a cada RisCanvi. S'han deduït les reincidències Autodirigida, Violència intrainstitucional i Trencament de condemna a partir de la informació d'incidents. Com que en aquests tres casos la reincidència es dona en el si del centre, la reincidència queda registrada indirectament. Al document *Tiresias - RisCanvi - Mapping variables.ods* trobem les equivalències de literals de les dades crues per associar un tipus de reincidència

al literal d'un incident. D'aquesta manera, si existeix una reincidència del tipus corresponent durant els sis mesos posteriors, etiquetem l'avaluació RisCanvi com a reincident. A l'annex adjuntem una infografia per aquest procediment.

Dades CEFJE

El fitxer facilitat pel CEFJE conté 3562 files i 380 columnes. Per tal de fer aquest estudi hem aplicat els següents procediments de neteja de dades:

- S'ha anonimitzat el NIS amb l'algoritme MD5 i s'ha suprimit la columna identificativa CIC.
- S'han filtrat els interns per tipus de reingrés (variable V346_TIPUS_REINGRÉS), per ometre aquells que tornen a la presó per antecedents.
- S'ha assumit que els 7 reingressos on no s'especifica si hi ha hagut violència en la reincidència (V361_REIN_DELICTE_VIOLENCIA), es tractaven de delictes no violents.
- S'han eliminat les columnes "XXXXXXXXX_REINCIDÈNCIA_JUDICIAL" i "REINCIDÈNCIA_VIOLENTA" perquè la informació que podrien contenir ve inclosa en altres variables del fitxer.
- Igual que a les dades d'AP, s'han creat les variables ORIGEN i EDAT_CAT.

En aquestes dades hi trobem diferents variables indicadores de reincidència. Entre elles, la *reincidència penitenciària*, que ve definida al document [5] com *nova entrada al sistema penitenciari, sigui com a persona preventiva o penada, per un delicte comès amb posterioritat a la data de sortida en llibertat definitiva*. A més, hi trobem la reincidència en execució penal, que compta amb els casos de reincidència penitenciària i els que tinguin una MEPC (Mesures d'Execució Penal a la Comunitat) en les mateixes característiques [12]. Finalment, hi trobem la variable de *reincidència judicial,* indicadora d'una nova causa judicial després de la finalització de la pena base [12].

Basant-nos en la secció *4. Què avalua RisCanvi? Avaluació de les conductes criteri*, del document [7], veiem que la variable que més s'ajusta a la definició de reincidència general seria la reincidència en execució penal, ja que inclou tant reingressos a presó com mesures penals alternatives (MPA) a conseqüència de delictes comesos en societat. També identifiquem que la reincidència violenta correspon als casos de reincidència penitenciària per un delicte violent, sense incloure els interns amb una MPA. No obstant això, aquestes dades només inclouen excarcerats i, per tant, no capturen la reincidència comesa durant permisos de sortida o en qualsevol altra situació que hagin permès a l'intern sortir del centre abans d'obtenir la llibertat definitiva (llibertat condicional, tercer grau, o similars). Després de converses amb l'equip d'Afers Penitenciaris i del CEFJE, hem observat que la recollida d'aquest subconjunt de dades de reincidència és complexa a nivell tècnic i, per tant, no s'ha inclòs en la base de dades.

Finalment, s'han integrat les informacions provinents dels dos conjunts de dades (AP i CEJFE) per a fer-los més robustos per a les anàlisis de reincidència general i reincidència violenta.
En les dades del CEJFE, s'ha observat que a) 855 interns tenien tots els factors de risc absents i b) que els interns que tenien informació de l'escala Screening no tenien la

Completa i viceversa. Això causa una reducció significativa del nombre d'interns per escala. Per això, per al present estudi s'han enriquit les dades del CEFJE amb dades de l'extracció AP: per a cada intern es considera l'última avaluació RisCanvi abans de l'excarceració per cada escala (Screening i Completa) excepte en els casos que no hi ha dades de l'última avaluació a l'extracció AP, que s'agafen les variables disponibles a la base de dades del CEFJE. Les variables de reincidència general i violenta s'han agafat de les dades del CEFJE.

Anàlisi de dades

En aquesta secció, presentem una anàlisi de dades per tal de posar en context tot l'ecosistema RisCanvi.

Primer, es resumeix l'estat i quantitat de dades disponibles i, seguidament, incloem diverses anàlisis de les variables utilitzades per l'algoritme així com dels resultats de les avaluacions (dades de sortida) i de la reincidència detectada.

Descripció preliminar de les Dades

El corpus analitzat en aquest projecte de dades provinents d'AP conté informació sobre:

- **Avaluacions**
  - 220.672 avaluacions RisCanvi.
    - 104.604 Completa.
    - 70.270 Screening.
    - 45.798 de sortida.
  - 11 anys d'avaluacions (01/01/10 - 31/12/21).
  - 44.605 interns.
- **Incidents**
  - 95.798 incidents relacionats amb reincidències.
  - 18.351 incidents d'altres tipologies.
- **Permisos i sortides**
  - 463.269 registres de permisos i sortides.
  - 27 anys de permisos (01/01/94 - 31/12/21).
  - 25.285 interns.

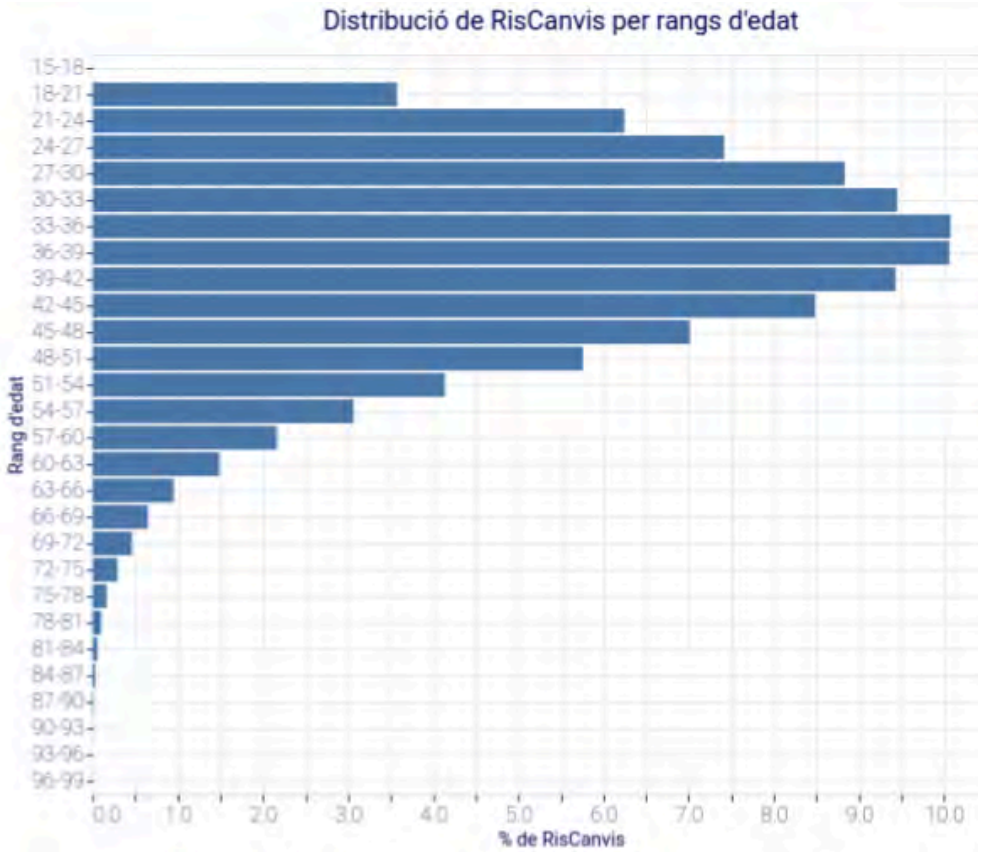
La base de dades del CEFJE una vegada enriquida amb les avaluacions de la primera extracció conté informació sobre:

- 2.159 avaluacions Screening (una per intern).
- 2.251 avaluacions Completa (una per intern).

Anàlisi de les variables demogràfiques

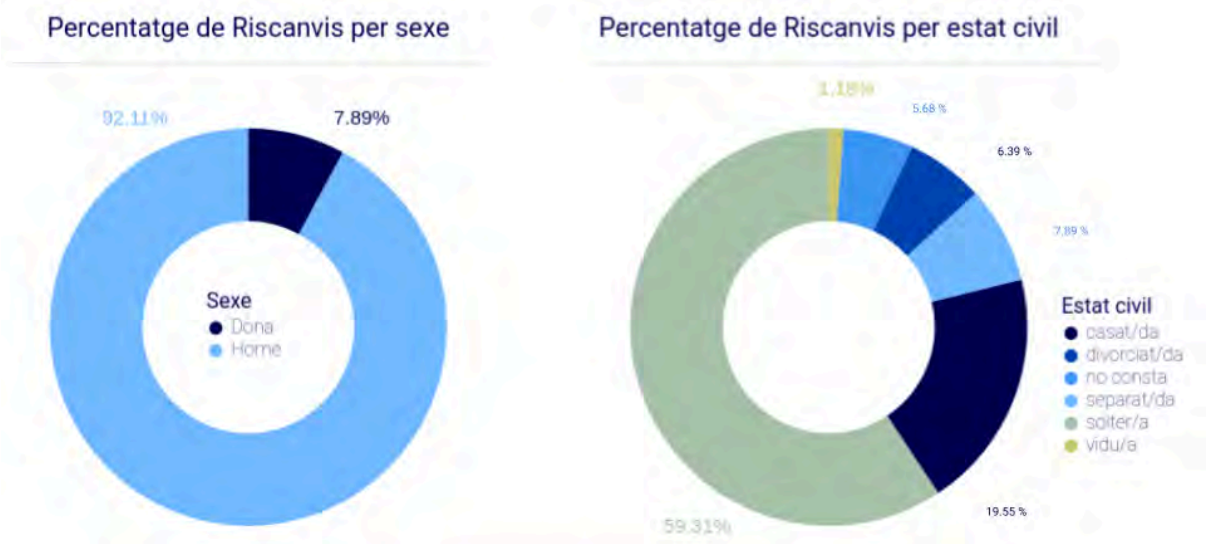
En aquesta secció analitzem les variables demogràfiques presents a les dades: edat, sexe, nacionalitat, i estat civil.

A la següent imatge mostrem la distribució de les edats a les avaluacions RisCanvi. Veiem que la mitjana es troba entre els 33 i els 39 anys i que tota la població avaluada és major d'edat.



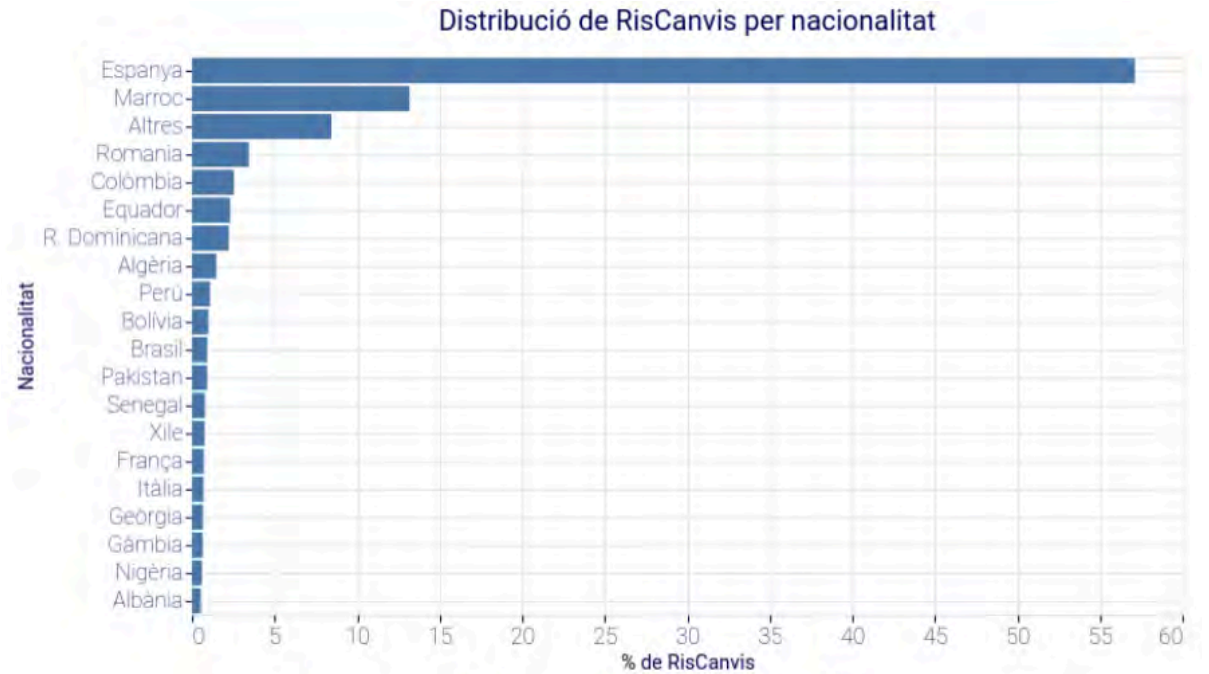
Distribució d'avaluacions RisCanvi per rangs d'edat

A les següents gràfiques observem el percentatge d'avaluacions per sexe i per estat civil. Veiem que la majoria dels RisCanvis pertanyen a població masculina i amb estat civil solter.



Percentatge d'avaluacions RisCanvi per sexe i per estat civil

Finalment, a la següent imatge mostrem la distribució del percentatge de RisCanvis per nacionalitat. Veiem que la majoria de les avaluacions són de persones amb nacionalitat espanyola o marroquina.



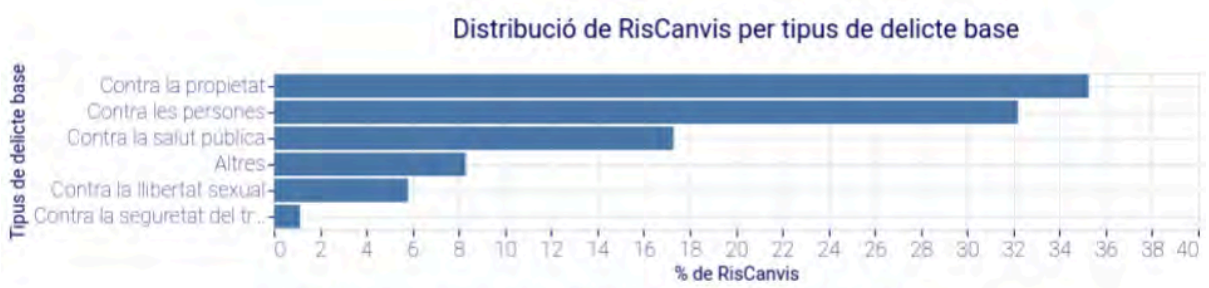
Percentatge d'avaluacions RisCanvi per nacionalitats

A l'annex, es poden trobar els gràfics d'edat, sexe i nacionalitat per les dades del CEFJE.

Anàlisi de les variables relacionades amb la pena

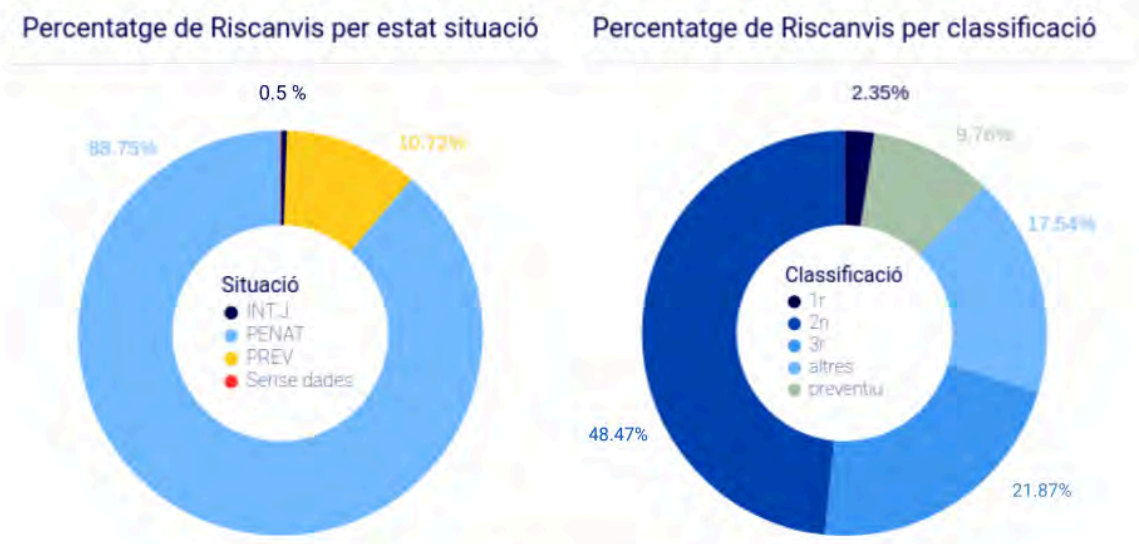
En aquesta secció analitzem les variables relacionades amb la pena: tipus de delicte base, situació, classificació, víctima i nombre d'ingressos.

Primer de tot, inspeccionem el percentatge de RisCanvis per tipus de delicte base. Veiem que hi ha una predominança de delictes contra la propietat i contra les persones.



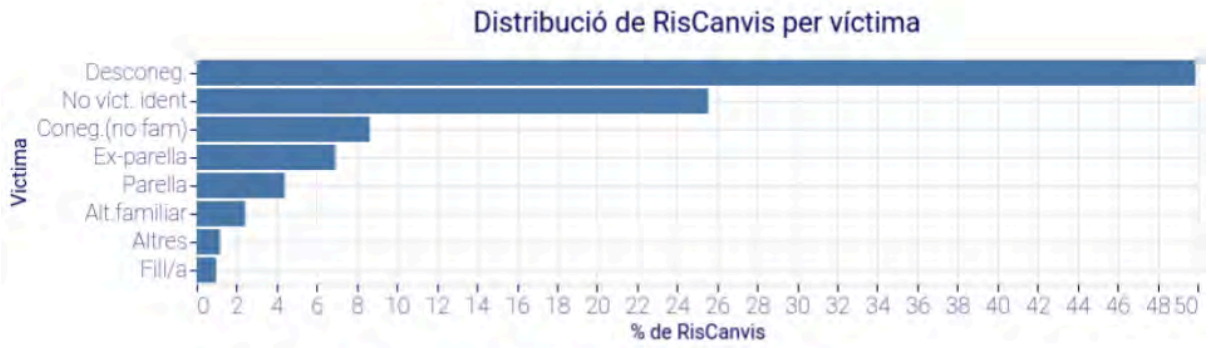
Percentatge d'avaluacions RisCanvi per tipus de delicte base

També mostrem el percentatge de RisCanvis per situació i per classificació de l'intern. La majoria d'avaluacions corresponen a persones penades i amb una classificació de segon grau. A la gràfica de classificació, hi ha un percentatge alt de persones pendents de classificar, que es troben a dins de la categoria *altres*.



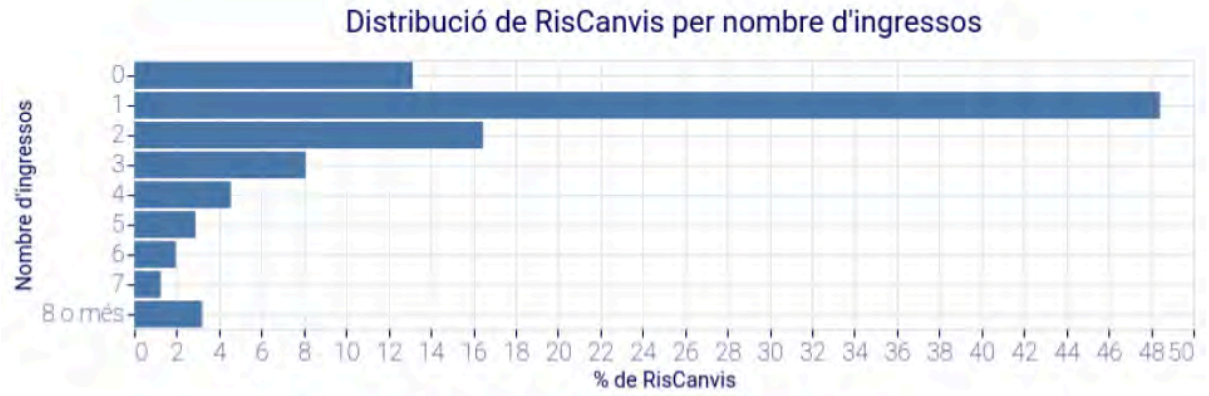
Percentatge d'avaluacions RisCanvi per situació i classificació de l'intern

A la següent gràfica visualitzem la distribució del percentatge de RisCanvis per víctima, i veiem que gairebé la meitat de les avaluacions Riscanvi estan etiquetades amb el valor *Desconeg.* per la columna *víctima*.



Percentatge d'avaluacions RisCanvi per víctima

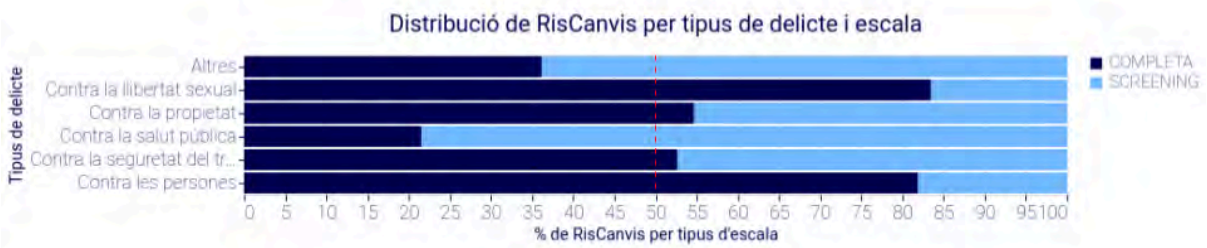
A continuació, analitzem la distribució de la variable *nombre d'ingressos*. Veiem que la majoria de RisCanvis corresponen a interns amb 1 ingrés. Crida l'atenció que aquesta variable pugui prendre el valor 0. No obstant això, després de converses amb el proveïdor informàtic i la direcció general, hem clarificat que aquesta columna mesura el nombre de reingressos al sistema penitenciari català des de llibertat. Aquest comptador no inclou els ingressos a fora de Catalunya amb trasllat al sistema penitenciari català i, per tant, pot prendre el valor 0.



Percentatge d'avaluacions RisCanvi per nombre d'ingressos

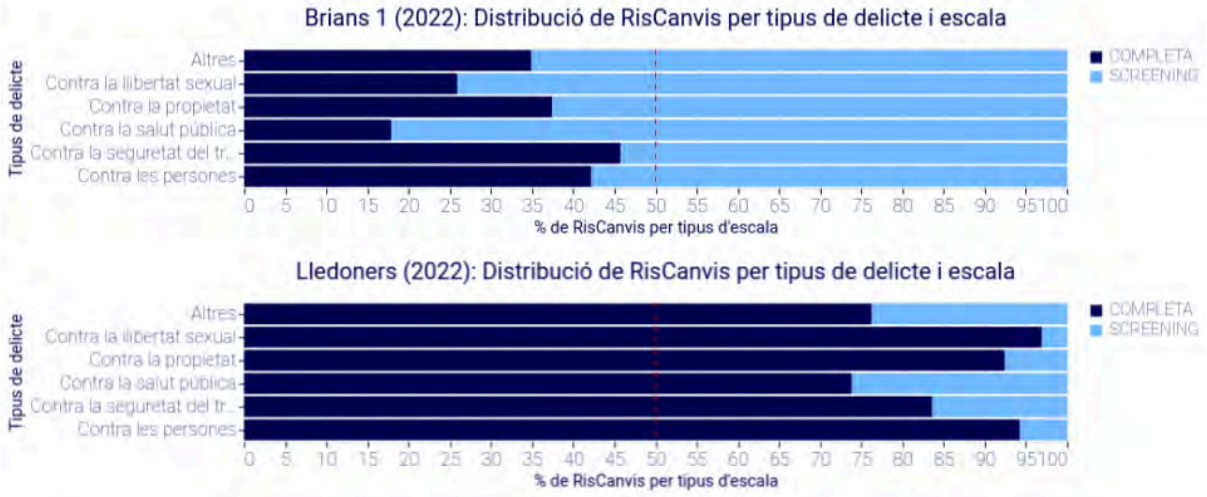
Finalment, podem fer una anàlisi més exhaustiva de qualsevol de les anteriors variables. En les següents gràfiques centrem l'atenció en la columna *tipus de delicte*, i mostrem diferents percentatges d'avaluacions per escala, centre i any.

Si revisem el percentatge de RisCanvis Completa o Screening per tipus de delicte, veiem que en els casos contra la llibertat sexual i contra les persones clarament tenim una predominança de l'escala Completa, doncs van associats a delictes amb violència.



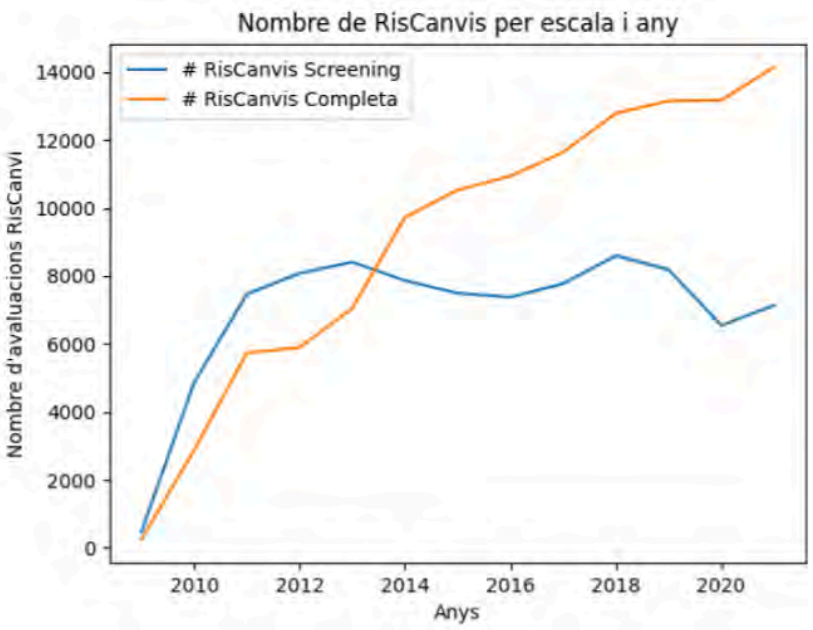
Percentatge d'avaluacions RisCanvi per tipus de delicte i escala

També, a través de la visualització de les dades podem veure que hi ha centres on es fan molt més RisCanvis Completa que Screening. En la figura a sota, veiem que al centre *Lledoners* hi ha una predominança de l'escala Completa mentre que al centre *Brians 1* trobem just el cas contrari.



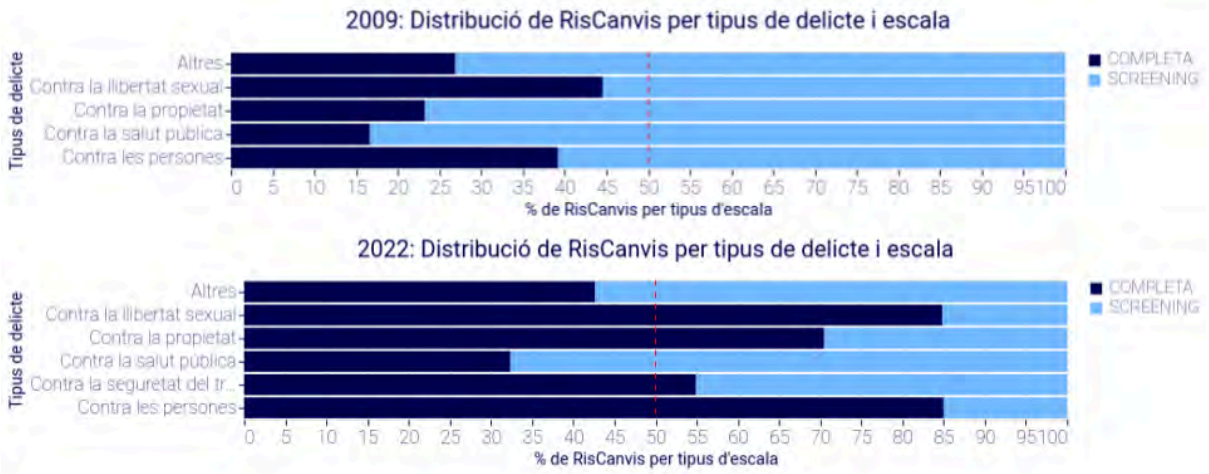
Percentatge d'avaluacions RisCanvi en 2022 per tipus de delicte i escala a dos centres del sistema penitenciari

Des de la implementació del protocol RisCanvi, hi ha hagut una evolució temporal del tipus d'avaluacions fetes. Amb el pas del temps, el percentatge d'anàlisis Completes ha augmentat.



Nombre d'avaluacions RisCanvi per escala entre 2009 i 2022

A la figura a sota observem l'evolució esperada en l'augment de Completa directa des de 2009 a 2022 per cada tipus de delicte. Veiem que tant els delictes contra la llibertat sexual com a contra les persones, són els que han experimentat un creixement més gran.



Percentatge d'avaluacions RisCanvi per tipus de delicte i escala en 2009 i en 2022

A l'annex d'aquest informe incloem la resta de gràfiques per centre i any, i també els gràfics de distribucions de tipus de delicte, victima, i nombre d'ingressos per les dades del CEFJE.

Anàlisi de les dades de reincidència

En aquesta secció, presentem els percentatges de reincidència real presents al corpus de dades d'avaluacions RisCanvi. Recordem que aquesta etiqueta de reincidència real per *Violència Autodirigida, Trencament de Condemna* i *Violència Intrainstitucional* s'ha assignat si

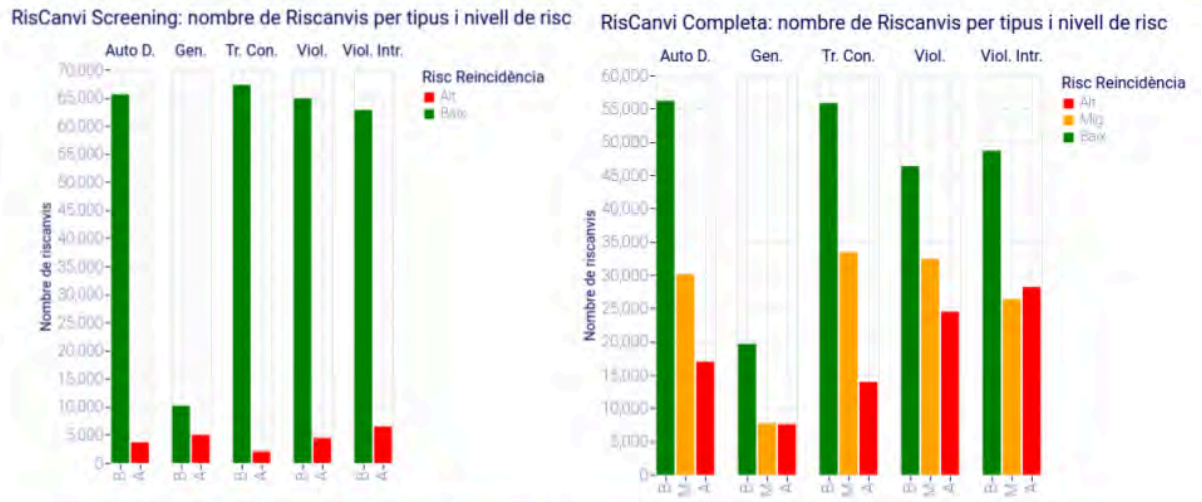
existeix un incident del tipus corresponent en els següents 6 mesos a l'avaluació. Pels casos de reincidència *General* i *Violenta,* s'han agafat les dades de reincidència en execució penal i reincidència penitenciària violenta com s'explica a l'apartat Dades CEFJE. Es pinten amb un color de fons diferent els casos de R. general i R. Violenta per remarcar que la quantitat d'avaluacions és molt inferior en aquests casos, ja que tenim menys mostres.

	Viol. Autodir.	Tr. Condemna	Viol. Intra. Ins.	Reinc. general	R. violenta
No reincidents	98,39%	99,42%	86,80%	71,14%	93,24%
Reincidents	1,61%	0,58%	13,20%	29,86%	6,76%

Observem que la majoria d'avaluacions no tenen una reincidència posterior. Aquest fet és molt evident en el cas de *Trencament de Condemna*, ja que el percentatge d'interns susceptibles a cometre aquesta reincidència és molt petit perquè ha de tenir un permís de sortida o bé estar en un grau de classificació que li permeti sortir de la presó. Això explicaria que la reincidència detectada estigui entorn del 0,58 % de les avaluacions.

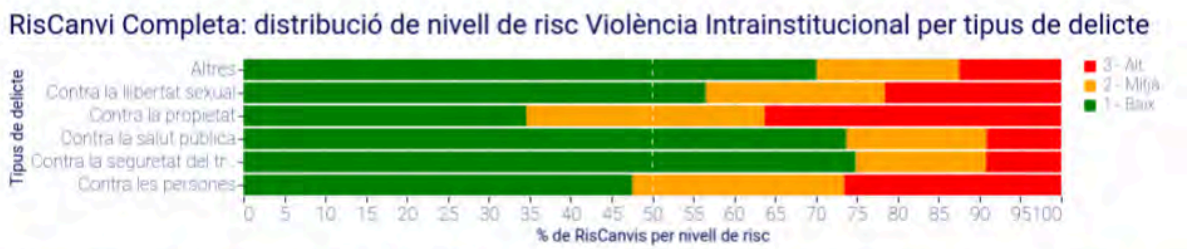
Anàlisi de les variables de sortida de RisCanvi

Per finalitzar l'anàlisi de dades, estudiem les variables de sortida de l'algoritme RisCanvi presents al corpus de dades. Per començar, a la següent imatge mostrem la distribució del nombre d'avaluacions Screening i Completa per cada nivell de risc i per cada tipus de reincidència. Veiem que la majoria de les avaluacions per totes les reincidències presenten un risc *Baix*. Aquest fet es mostra més evident en les avaluacions Screening, però hem de tenir present que els interns que passen per un RisCanvi Screening són un subconjunt de la població penitenciària (p. ex. si el delicte base és violent, l'intern passa directament a escala Completa). Aquesta anàlisi on trobem un percentatge inferior de riscos alts concorda amb les dades reals d'incidència, ja que per cada tipus de reincidència hi ha un percentatge petit d'avaluacions associades a incidència real.

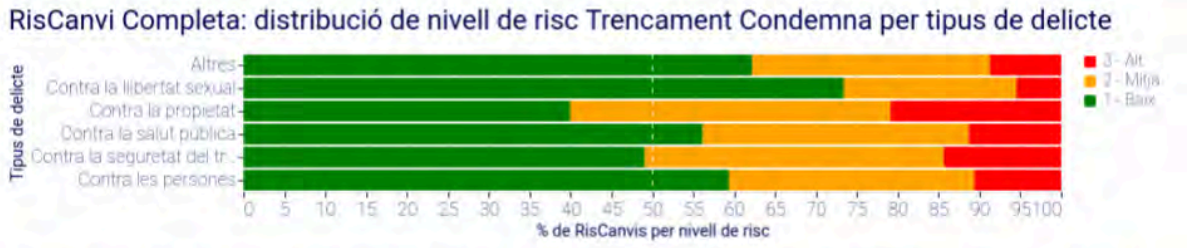


Nombre d'avaluacions RisCanvi Screening i Completa per nivell de risc i tipus de reincidència.

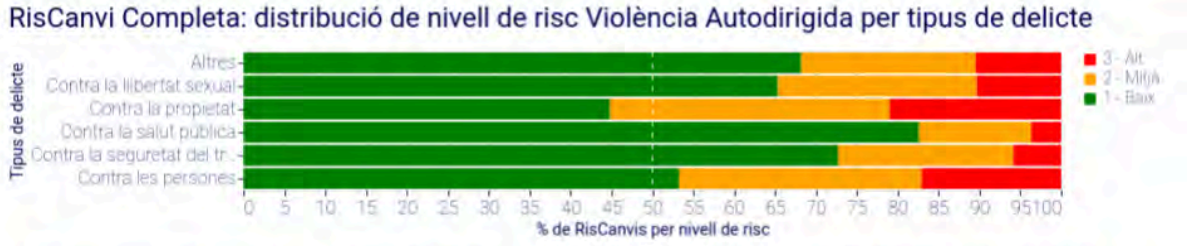
Podem analitzar els nivells de risc per tipologia del delicte base. A les següents figures, mostrem el percentatge d'avaluacions RisCanvi per nivell de risc i tipus de delicte per cada tipus de reincidència. Destaquem en aquesta anàlisi que els delictes contra la propietat sempre tenen més probabilitat d'un RisCanvi més alt (alt o mitjà). D'acord amb l'expertesa de l'equip de la direcció general, aquest fet té sentit perquè normalment els interns que cometen aquests tipus de delictes tenen una carrera delictiva consolidada i, per tant, tenen un risc més alt de reincidir.



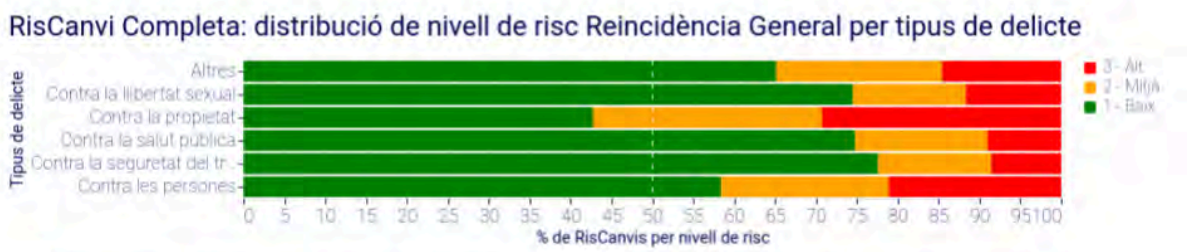
Percentatge d'avaluacions RisCanvi Completa per nivell de risc de Violència Intrainstitucional i tipus de delicte



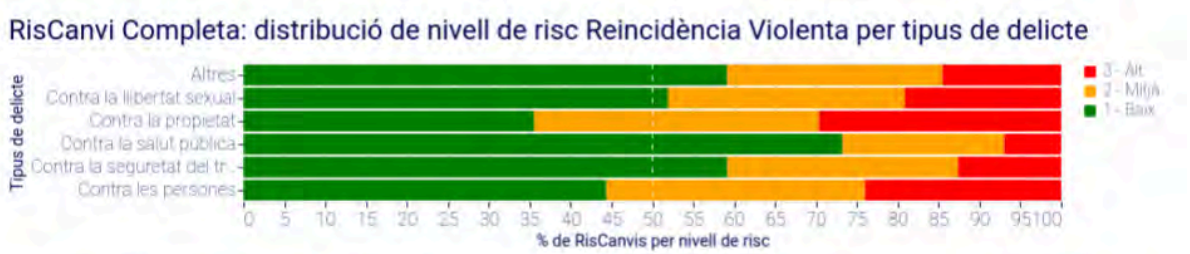
Percentatge d'avaluacions RisCanvi Completa per nivell de risc de Trencament de Condemna i tipus de delicte



Percentatge d'avaluacions RisCanvi Completa per nivell de risc de Violència Autodirigida i tipus de delicte



Percentatge d'avaluacions RisCanvi Completa per nivell de risc de Violència General i tipus de delicte



Percentatge d'avaluacions RisCanvi Completa per nivell de risc de Reincidència Violenta i tipus de delicte

A l'annex d'aquest informe incloem la resta de gràfiques per l'escala Screening.

Algoritme actual

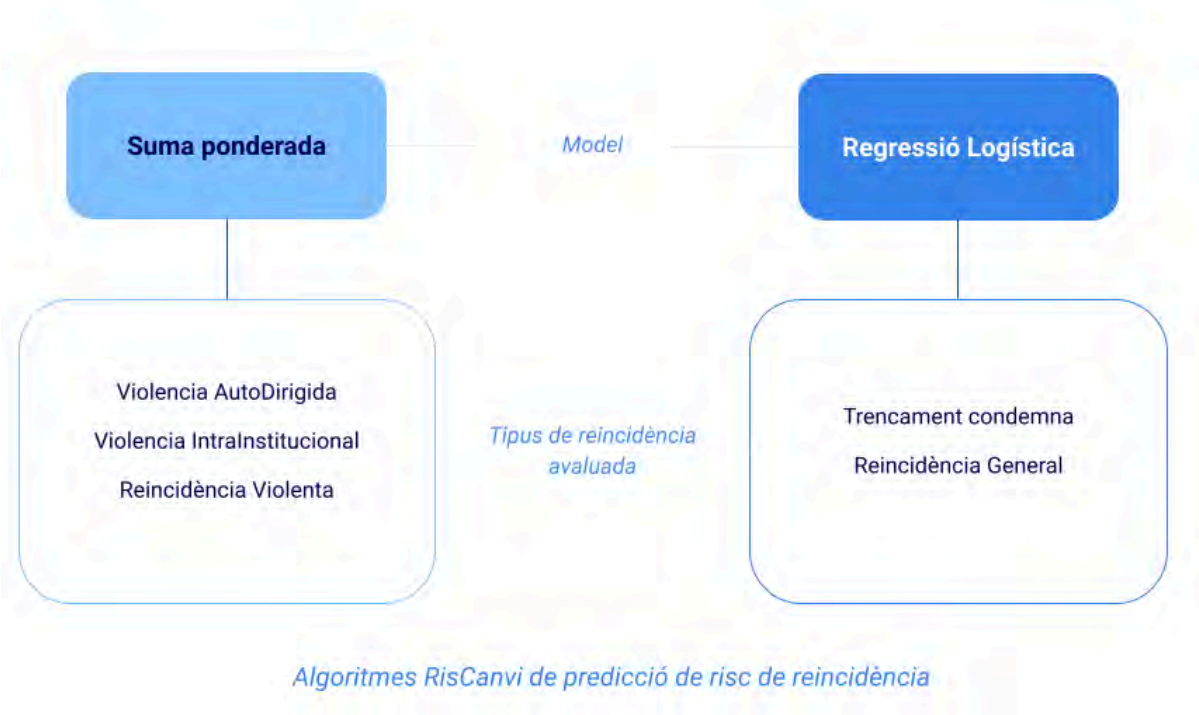
Un algoritme no és més que un conjunt d'instruccions que serveixen per donar resposta a un problema. En aquest cas, la qüestió que volem resoldre és determinar si una persona és susceptible de cometre en el futur una de les cinc reincidències descrites prèviament. L'algoritme RisCanvi processarà una sèrie de variables d'entrada per crear un variable de sortida que ha de servir per orientar la presa de decisions per part dels professionals i, d'aquesta manera, gestionar el possible risc de reincidència.

Existeixen molts tipus d'algoritmes, des d'algorismes més senzills que consten de pocs passos fins a algoritmes més complexos que s'han d'executar en supercomputadors. Un subconjunt d'algoritmes molt utilitzats avui dia pertanyen a la branca de la intel·ligència artificial i es coneixen com a algoritmes d'Aprenentatge Automàtic (Machine Learning en anglès). Gràcies a aquests algoritmes, els ordinadors poden aprendre de les dades sense que els experts hagin de codificar explícitament els paràmetres que determinen l'output de l'algoritme. A partir de les dades històriques i mitjançant diferents mètodes, l'ordinador és capaç de determinar la relació entre les variables i assignar-li els pesos més adients. Aquest procés s'anomena *entrenament* de l'algoritme. Una vegada l'algoritme ha après la configuració de paràmetres òptima, podem fer-lo servir per dur a terme la tasca corresponent, en aquest cas, predir el risc de reincidència.

En aquesta secció donem una visió general dels diferents algoritmes de predicció de risc de reincidència que han estat introduïts al sistema RisCanvi, tal com venen descrits a la documentació proporcionada ([1], [2], [3]).

Funcionament

Als inicis de la creació del sistema RisCanvi es va plantejar l'algoritme de predicció de risc com a un **model de suma ponderada**, és a dir, com a una combinació lineal de les variables d'entrada amb uns pesos assignats a cada variable a partir de coneixement expert. Per determinar els nivells de risc per cada escala (Alt-Baix o Alt-Mig-Baix), es van triar uns punts de tall sobre la puntuació de sortida de l'algoritme. El 2011, es van ajustar aquests punts de tall per tal de millorar les prediccions [2]. Finalment, el 2017 es va introduir un nou algoritme de predicció de risc conegut com a **Regressió Logística**. Aquest model forma part del conjunt d'algoritmes d'intel·ligència artificial i, de la mateixa manera que el model de Suma Ponderada, fa servir una sèrie de punts de tall per separar els diferents nivells de risc [3]. Actualment, el model de Regressió Logística s'aplica per predir el risc de *Reincidència General* i *Trencament de Condemna* mentre que el model de Suma Ponderada s'aplica per predir el risc de *Violència Intrainstitucional, Reincidència Violenta* i *Violència Autodirigida.*



Algoritmes RisCanvi de predicció de risc de reincidència

A continuació explicarem el funcionament d'aquests dos models, basant-nos en la documentació [1], [2], [3].

Suma Ponderada

L'algoritme de Suma Ponderada del sistema RisCanvi va ser dissenyat per l'equip del Dr. A. Andrés Pueyo el 2010. Aquest model calcula la variable de sortida com a una combinació lineal de les variables d'entrada. A cada variable d'entrada li correspon un pes determinat amb coneixement expert. Així doncs, l'algoritme de suma ponderada no pertany a la família de models d'intel·ligència artificial.

En el moment de la creació d'aquest algoritme, la Reincidència General no estava contemplada a dins del sistema i, per tant, es van definir 8 models diferents (un per cada escala, Screening i Completa, i tipus de reincidència).

En el model de Suma Ponderada trobem dues tipologies de variables d'entrada:

- Variables de context: sexe (home/dona), edat (major/menor de 30 anys), nacionalitat (espanyol/estranger) i situació (penat/preventiu).
- Factors de risc: els 10 factors de risc de l'escala Screening o els 43 factors de risc de l'escala Completa.

No obstant això, no tots els models contemplen totes les variables d'entrada, i.e., no totes les variables *d'entrada* tenen un pes diferent de zero. Per exemple, l'algoritme de predicció de risc de *Violència Intrainstitucional* de l'escala *Screening* contempla les següents variables: situació i els factors de risc 2, 3, 5 i 8. Cada variable té un pes assignat de la següent

manera: per cada factor, si el factor pren el valor de la taula, se suma el valor corresponent al pes d'aquell factor, si no, no contribueix a la suma final.

Variables del model	Valor	Pesos
Situació	Penat	-3
FAC_2: Història de violència	Sí	3
FAC_3: Problemes de conducta penitenciària	Sí	3
FAC_5: Problemes amb el consum de drogues o alcohol	Sí	3
FAC_8: Manca de recursos econòmics	Sí	2

Pesos del model de Suma Pondera per Violència Intrainstitucional Screening

Així doncs, suposem que avaluem a un intern A, amb situació de penat, amb una història de violència i amb una manca de recursos econòmics. Per calcular la seva puntuació, sumem els pesos de les variables corresponents, com es mostra a la següent taula:

Variables del model	Valor	Pesos	Exemple Intern A	Score
Situació	Penat	-3	**Penat**	**-3**
FAC 2: Història de violència	Sí	3	**Sí**	**3**
FAC_3: Problemes de conducta penitenciària	Sí	3	No	0
FAC_5: Problemes amb el consum de drogues o alcohol	Sí	3	No	0
FAC_8: Manca de recursos econòmics	Sí	2	**Sí**	**2**
Puntuació Screening (suma dels pesos dels factors)				**2**
Valoració final (té en compte el punt de tall per l'etiqueta "Risc Alt": 5 punts aquí)				**BAIX**

Exemple de càlcul de nivell de risc amb el model de Suma Pondera per Violència Intrainstitucional Screening

Finalment, determinem que l'intern *A* té un score RisCanvi de 2 punts. Per assignar el nivell de risc, s'aplica el punt de tall definit per experts en aquesta versió de l'algoritme (5 punts). Com que el seu score és inferior al punt de tall, l'intern té un nivell de risc Baix. Aquest procediment s'aplica anàlogament per la resta d'escales i reincidències.

Com s'ha mencionat anteriorment, el 2011 es van actualitzar els punts de tall per separar els nivells de risc, però l'algoritme original de suma ponderada (amb els seus pesos corresponents) continua estant vigent avui en dia per avaluar la *Reincidència Violenta, Violència Intrainstitucional,* i *Violència Autodirigida.*

Regressió Logística

El 2017 es va fer una revisió de l'algoritme de Trencament de Condemna i es va incorporar a dins del sistema la predicció del risc de Reincidència General. Assumim que és un model de Regressió Logística i, per tant, un algoritme d'intel·ligència artificial. Va ser desplegat amb la versió 3 del sistema RisCanvi el juny del 2019.

Igual que en el model de suma ponderada, cada factor de risc té un pes assignat. En aquest cas, però, els pesos de cada factor de risc s'obtenen mitjançant un algoritme que minimitza els errors de predicció de risc donat un històric de dades. Per això cal entrenar el model amb dades de reincidència del passat, que anomenem conjunt d'entrenament. Un cop tenim els pesos definits, de la mateixa manera que en el model de suma ponderada, l'algoritme calcula una puntuació de risc d'un intern com a la suma dels pesos corresponents. Al resultat d'aquesta puntuació s'aplica una funció matemàtica, coneguda com a funció logística, per acabar amb un valor entre 0 i 1. A més a més, en aquesta versió de l'algoritme, veiem que s'aplica una transformació del score detallat al document [2]. Una vegada que arribem a la puntuació final, apliquem els punts de tall i aconseguim el nivell de risc (Alt-Baix per a Screening o Alt-Mig-Baix per a Completa).

Tant els pesos assolits amb l'entrenament del model el 2017, com els punts de tall no han estat actualitzats des del maig del 2017.

Febleses dels models actuals

Per una banda, la principal feblesa del model de Suma Ponderada és el seu alt cost d'actualització. No és possible automatitzar l'actualització dels pesos seguint aquest model, ja que requeriria desenvolupar estudis com el fet l'any 2010 pel Dr. Pueyo de manera periòdica. Com a conseqüència, és un model desactualitzat, que no s'ha revisat completament des de fa dotze anys.

Així doncs, el model de suma ponderada actual està basat en el coneixement expert sobre la reincidència l'any 2010, que pot haver evolucionat amb el pas dels anys. A més, ara com ara, comptem amb molta més informació i dades de les variables relacionades amb reincidència gràcies a la implantació del protocol RisCanvi i a la informatització del sistema penitenciari.

Els pesos han estat seleccionats per experts en la matèria amb molt de coneixement sobre les variables. Tot i això, diferents experts poden assignar diferents pesos a cada variable. És per això que és important comptar amb eines matemàtiques que reflecteixin els objectius de la direcció general i que estiguin basades en les dades de reincidència actualitzades, tant per assignar els pesos com els punts de tall, sempre amb el suport de l'equip d'experts.

Per una altra banda, la principal feblesa del model de Regressió Logística és que per poder-lo actualitzar cal tenir les dades de reincidència real disponibles. No es tracta d'una feblesa del model de Regressió Logística en si mateix, si no de la infraestructura informàtica. Per a cada avaluació de RisCanvi hauríem de tenir gravada, a posteriori, una etiqueta assenyalant si hi ha hagut reincidència durant un període de temps fixat. Aquest requisit és un denominador comú per tots els algoritmes d'aprenentatge automàtic supervisat que es puguin aplicar. A més, també permet fer una correcta avaluació del funcionament de l'algoritme a posteriori.

Una altra feblesa de l'algoritme actual de Regressió Logística és que no tenim constància de l'existència de documentació relacionada amb l'entrenament de l'algoritme implementat (quines dades s'han fet servir, quins processos de neteja de dades s'han aplicat, etc.). Per aquest motiu, no és possible recrear l'entrenament que s'ha fet per arribar als pesos donats a través de la documentació.

Finalment, existeixen altres models de Machine Learning que podrien obtenir millors resultats, com mostrarem més endavant en aquest informe, i que es podrien implementar en futures versions de RisCanvi.

Recreació dels algoritmes RisCanvi

Per a poder avaluar correctament els algoritmes de Riscanvi, s'ha de tenir accés al codi implementat. Ara mateix, l'accés a aquest codi no és públic però durant la segona etapa del projecte, Dribia ha tingut accés al codi font a través d'arxius de text pla. Basant-nos en la documentació aportada ([1], [2], [3]) i discussions amb l'equip d'Afers penitenciaris, hem recreat els algoritmes per als 5 tipus de risc i escales Screening i Completa (en total 10 algoritmes).

Per als casos de Violència autodirigida, Violència intrainstitucional i Reincidència violenta, s'han utilitzat els pesos de l'informe de 2010 [1] i els punts de tall de l'informe de 2011 [2]. En aquests casos, l'algoritme és de suma ponderada. Pel Trencament de condemna i la Reincidència general, s'han emprat els pesos i punts de tall de l'informe de 2017 [3]. En aquests casos, l'algoritme és de regressió logística.

Després d'analitzar el codi implementat al servidor, hem trobat un parell de diferències entre la documentació aportada i la implementació en `psql`. La primera és que el punt de tall al sistema de Trencament de condemna escala Screening és 0,97, en contraposició al 3,20 descrit al document [3]. La segona diferència està relacionada en l'arrodoniment dels decimals a dos dígits, quan a la documentació trobem tres o quatre xifres. En la nostra

versió del codi, hem aplicat el punt de tall 0,97, d'acord amb el codi implementat als sistemes del SIPC. Això no obstant, hem mantingut totes les xifres decimals com es detalla a la documentació.

Amb aquests models recreats a partir de la documentació proporcionada, obtenim resultats pràcticament idèntics als outputs reals de RisCanvi excepte per algun cas concret. Per mesurar les discrepàncies entre el model que hem recreat i el model en producció, hem generat el nivell de risc amb la nostra versió dels algorismes a partir de totes les dades d'avaluacions Screening i Completa de l'extracció d'AP. Per cada algorisme, hem agafat les avaluacions corresponents a les dates on cada algorisme està en funcionament (per exemple, pels models de regressió logística només hem agafat avaluacions corresponents a la versió 3 de RisCanvi). Després, hem comptat el número de vegades que la nostra recreació dona un nivell de risc diferent de la que ve recollida a les dades (e.g. el nostre algorisme diu Mig mentre que a les dades hi ha recollit risc Alt). Finalment, amb el nombre de discrepàncies per cada algorisme hem calculat el percentatge que representa del conjunt d'avaluació. A la següent taula presentem aquest percentatge de discrepàncies entre l'output de la nostra recreació i l'output del model actual.

RISC	MODEL	ESCALA	PERCENTATGE DE DISCREPÀNCIES
Trencament condemna	Suma ponderada	SCREENING	0.32%
		COMPLETA	2.53%
Violència autodirigida	Suma ponderada	SCREENING	3.55%
		COMPLETA	7.99%
Reincidència general	Suma ponderada	SCREENING	0.71%
		COMPLETA	0.81%
Reincidència violenta	Regressió logística	SCREENING	0.36%
		COMPLETA	1.23%
Violència intrainstitucional	Regressió logística	SCREENING	0.47%
		COMPLETA	0.94%

Percentatge de discrepàncies entre l'output de la nostra recreació i l'output de l'algorisme en producció present a les dades d'avaluació (extracció AP)

La recreació sembla molt acurada en la majoria d'algorismes, amb un error per sota de l'1 %, excepte en el cas de *Violència Autodirigida*, i pel cas de completa de *Trencament de condema* i *Reincidència violenta*. Per poder reduir a 0 % aquestes diferències, seria convenient tenir accés a un entorn de simulació on es pogués analitzar pas per pas el procés de càlcul per aquests casos concrets i veure si és un tema de la implementació concreta de l'algorisme o de les dades utilitzades.

Avaluació

En aquest apartat, mesurem la qualitat de les prediccions de risc de reincidència de Riscanvi. Per dur a terme aquesta tasca, és imprescindible comptar amb les dades reals de reincidència, ja que per determinar si la predicció de l'algoritme és errònia o acurada necessitem saber, per cada avaluació, si hi ha hagut reincidència/violència/trencament posterior. Per això, avaluarem les violències Autodirigida, Instrainstitucional i el risc de Trencament de Condemna, amb les dades de l'extracció d'AP, i les Reincidències General i Violenta amb les dades del CEFJE.

Per a les dades d'AP, s'han analitzat només les avaluacions fetes amb l'última versió de l'algoritme. Pel que fa a les dades del CEFJE, cal tenir en compte que aquest conjunt de dades és molt petit en comparació a l'extracció d'AP i, per tant, la seva avaluació tindrà una fiabilitat més limitada: en aquest cas només podem utilitzar l'última avaluació RisCanvi per excarcerat, és a dir una per persona, a diferència d'emprar totes les avaluacions durant l'estada al centre que es pot fer per les altres escales. A més a més, mentre que els algoritmes amb dades d'AP els podem avaluar durant les dates on han estat en funcionament, i.e. a partir de la implementació de l'última versió, l'avaluació pels casos de reincidència general i violenta l'estem fent "a posteriori". És a dir, estem avaluant la qualitat de l'algoritme actual si hagués estat aplicat abans del 2015. En el cas de reincidència general, aquest algoritme es comença a fer servir al 2017 i les dades d'avaluacions RisCanvi són, com a màxim, del 2015. D'aquesta manera, podem recrear quins nivells de risc hauria obtingut l'intern i així avaluar la qualitat de les prediccions. D'una manera semblant, pel cas de reincidència violenta, estem fent servir l'algorisme actual (RisCanvi 2.0) sobre tot el conjunt de dades del CEFJE i així evitem reduir encara més el conjunt d'avaluació.

Per avaluar els algoritmes RisCanvi, farem servir diferents mètriques que ens donaran una visió de les fortaleses i les febleses del model. En concret, centrarem l'avaluació en els resultats en termes de la corba ROC (*Receiver Operating Characteristic)*, l'AUC (*Area Under the Curve)* i l'exactitud (*accuracy, FPR, FNR, TPR, TNR).* També fem una avaluació dels possibles biaixos potencialment discriminatoris de cara a grups històricament discriminats i analitzem altres possibles biaixos que pugui tenir l'algoritme.

Anàlisi AUC-ROC

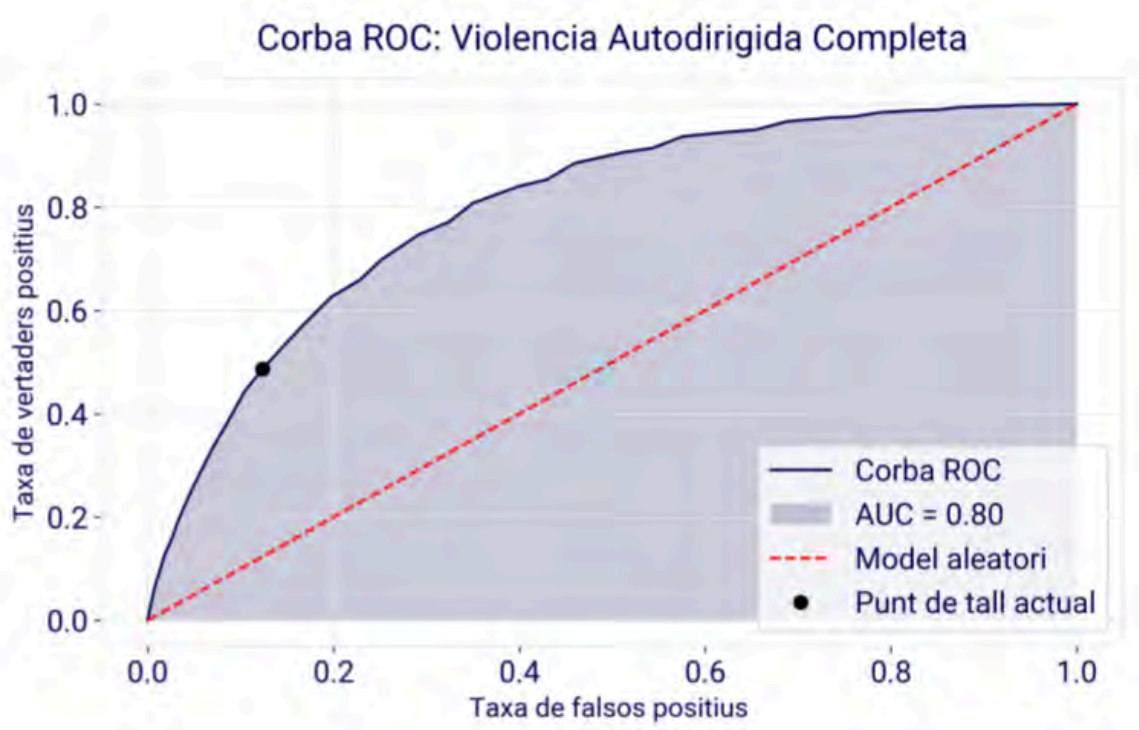
La corba característica de funcionament del receptor, coneguda com a corba ROC, és una gràfica que ens permet visualitzar la capacitat predictiva d'un algoritme classificador binari. Per crear aquestes gràfiques, a partir de la puntuació final del model variem els possibles punts de tall i mesurem la taxa de falsos positius (l'eix *x)* i la taxa de veritables positius (l'eix *y)*. En aquest cas, els veritables positius són les avaluacions etiquetades amb risc alt i que n'hem detectat reincidència, i els falsos positius són les avaluacions etiquetades amb risc alt, però de les que no hi ha reincidència detectada.

L'àrea sota la corba ROC, *Area Under the Curve (AUC)*, és una mètrica d'avaluació que ens dona una indicació de com de bé l'algoritme distingeix entre la classe positiva (incidents) i la

classe negativa (no incidents). Específicament, és la probabilitat de què, donada una mostra positiva seleccionada aleatòriament i una mostra negativa seleccionada aleatòriament, l'algoritme sàpiga predir correctament la classe de cadascuna. Així doncs, un AUC = 1 seria un algoritme perfecte i un AUC = 0,5 seria aleatori.

Per avaluar un algoritme en termes de l'AUC-ROC és important disposar de la puntuació final abans d'aplicar-hi el punt de tall. A les dades d'*output* proporcionades només hi consten els nivells del risc (alt-mig-baix). Per aquest motiu, les anàlisis en aquesta secció fan referència a les recreacions que hem fet, no al model en producció.

A la següent figura, veiem la corba ROC per a l'algoritme de Violència Autodirigida en escala Completa. En vermell, pintem la corba ROC d'un algoritme aleatori (amb AUC=0.5). Observem que el model actual te una AUC=0.8 i, per tant, una capacitat d'encert bastant alta.



Corba ROC per la recreació de l'algoritme RisCanvi per Violència Autodirigida escala Completa

Com veiem a la següent taula, per la resta d'algoritmes s'obtenen bons resultats, amb AUCs per sobre de 0,65, excepte per Tr. C. Pels tres casos on podem fer la comparativa amb l'anàlisi feta el 2017 [4], veiem que els resultats són lleugerament millors o molt semblants. Per la comparativa amb l'estudi original del 2010, veiem que els valors són semblants tot i que lleugerament inferiors, excepte la millora de Trencament de condemna Screening que l'atribuïm al canvi d'algoritme per a aquest risc. Destaca sobretot la baixada de AUC dels algoritmes de reincidència violenta, ja obtinguda el 2017 [4]. Sospitem que aquestes diferències es poden justificar per diferents motius: l'avaluació del propi model del 2010

sobre les dades que han creat al model (risc d'*overfitting* i de no obtenció del rendiment real del model en dades noves), el poc volum de dades amb el que es comptava en 2010 que dona lloc a avaluacions menys robustes, o les diferències tècniques d'implementació. El decrement de l'AUC també podria tenir origen en un possible canvi en els patrons de reincidència des de 2010 a 2022, que hagin ocasionat una pèrdua de capacitat predictiva de l'algorisme.

RISC	ESCALA	AUC-ROC 2022	AUC-ROC 2017	AUC-ROC 2010
Trencament condemna	SCREENING	0.57	*	0.46**
	COMPLETA	0.64	*	0.84**
Violència autodirigida	SCREENING	0.85	*	0.87
	COMPLETA	0.8	0.763	0.83
Reincidència general	SCREENING	0.68	-	-
	COMPLETA	0.65	-	-
Reincidència violenta	SCREENING	0.68	*	0.79
	COMPLETA	0.68	0.69	0.8
Violència intrainstitucional	SCREENING	0.76	*	0.82
	COMPLETA	0.75	0.726	0.83

** No analitzat a l'estudi del 2017*
*** Des del 2010, el model de predicció ha canviat (de suma ponderada a regressió logística).*

Finalment, els nivells d'AUC-ROC són comparables a altres algoritmes en el panorama internacional (COMPAS 0,7 [9], OASys ~0,75 [10]).

Anàlisi d'exactitud

L'exactitud, o *accuracy*, no és més que la probabilitat d'encert de l'algoritme i es calcula de la següent manera

$$accuracy = \frac{nombre\ d'encerts}{nombre\ d'avaluacions} = \frac{VP + VN}{VP+VN+FP+FN}$$

on VP són els veritables positius, VN els veritables negatius, FP els falsos positius i FN els falsos negatius. En aquest estudi, hem considerat les següents definicions:

- Veritables positius (VP): avaluacions amb predicció de risc alt i posterior incidència/violencia
- Veritables negatius (VN): avaluacions amb predicció de risc baix-mitjà i no consta posterior incidència/violencia
- Veritables positiu (FP): avaluacions amb predicció de risc alt i no consta posterior incidència/violencia
- Veritables positiu (FN): avaluacions amb predicció de risc baix-mitjà i posterior incidència/violencia

És a dir, a l'escala completa, agrupem els riscos mitjans amb els riscos baixos. Un enfocament diferent agruparia els riscos mitjans amb els riscos alts. A l'annex, hem inclòs també els resultats d'exactitud amb aquesta segona opció d'agrupament.

A la següent taula mostrem els resultats d'exactitud per a l'output del model en producció, on agafem com a predicció positiva els interns amb risc alt (tant a l'escala Screening com a la Completa):

RISC	ESCALA	EXACTITUD
Trencament condemna	SCREENING	96.32%
	COMPLETA	76.25%
Violència autodirigida	SCREENING	97.85%
	COMPLETA	86.61%
Reincidència general	SCREENING	66.33%
	COMPLETA	69.26%
Reincidència violenta	SCREENING	89.86%
	COMPLETA	85.47%
Violència intrainstitucional	SCREENING	85.96%
	COMPLETA	73.97%

Exactitud dels models RisCanvi (accuracy)

Els valors d'exactitud són molt alts per alguns casos, però poc informatius, degut a la baixa taxa de reincidència per als 5 riscos. Per exemple, només un 0,5 % de les avaluacions RisCanvi tenen una incidència associada amb trencament de condemna en els 6 mesos posteriors. Així doncs, un algoritme que donés risc baix a totes les avaluacions, tindria una exactitud del 99,5 %, ja que només s'equivocaria en el 0,5 % dels casos. L'algoritme de Reincidència General és el que té una exactitud més baixa i, per tant, el que té més marge de millora. Tot i així, s'ha de tenir en compte que el conjunt de dades per R. general i violenta és més baix que per als altres riscos.

Tot i que a l'hora d'avaluar qualsevol algoritme és important tenir present el nivell d'exactitud, hi ha altres mètriques que ens poden donar una visió més informativa de la capacitat d'encert i la mesura de l'error dels algoritmes RisCanvi.

Per començar, quantifiquem com de bé l'algoritme detecta cadascuna de les conductes amb les següents mètriques:

- Taxa de vertaders positius (TPR): el percentatge de riscos alts entre les avaluacions amb una incidència en els 6 mesos posteriors o el percentatge d'excarcerats amb risc alt amb una reincidència en societat, en cada cas corresponent.
- Taxa de falsos negatius (FNR): el percentatge de riscos mitjans i baixos entre les avaluacions amb una incidència en els 6 mesos posteriors o el percentatge

d'excarcerats amb risc mig-baix amb una reincidència en societat, en cada cas corresponent.

També quantifiquem com de bé l'algoritme detecta l'absència de conductes de risc amb les següents mètriques:

- Taxa de vertaders negatius (TNR): el percentatge de riscos mitjans i baixos entre les avaluacions sense una incidència en els 6 mesos posteriors o el percentatge d'excarcerats sense risc mig-baix sense una reincidència en societat, en cada cas corresponent.
- Taxa de falsos positius (FPR): el percentatge de riscos alts entre les avaluacions sense una incidència en els 6 mesos posteriors o el percentatge d'excarcerats amb risc alt sense una reincidència en societat, en cada cas corresponent.

Un algoritme perfecte tindria unes mesures d'error (la FNR i la FPR) properes al 0 % i unes mesures d'encert (la TPR i la TNR) properes al 100 %. A la següent taula, mostrem aquestes dades per a tots els riscos i escales:

RISC	ESCALA	FPR	FNR	TPR	TNR
Trencament condemna	SCREENING	3.63%	100.00%	0.00%	96.37%
	COMPLETA	23.70%	56.00%	44.00%	76.30%
Violència autodirigida	SCREENING	1.17%	88.87%	11.13%	98.83%
	COMPLETA	12.34%	51.36%	48.64%	87.66%
Reincidència general	SCREENING	31.12%	39.97%	60.03%	68.88%
	COMPLETA	6.39%	81.28%	18.72%	93.61%
Reincidència violenta	SCREENING	4.77%	84.25%	15.75%	95.23%
	COMPLETA	9.93%	72.29%	27.71%	90.07%
Violència intrainstitucional	SCREENING	6.95%	71.71%	28.29%	93.05%
	COMPLETA	21.48%	43.87%	56.13%	78.52%

Taxes de falsos positius i negatius i de vertaders positius i negatius

A partir d'aquestes mètriques podem extreure les següents conclusions:

- Les taxes de falsos positius són baixes (~10%) i, per tant, les taxes de vertaders negatius són altes (~90%), sobretot per l'escala Screening. És a dir, l'algoritme detecta majoritàriament els casos les avaluacions de baix risc correctament.
- Les taxes dels falsos negatius són força altes i, en conseqüència, les taxes de vertaders positius són baixes, especialment a l'escala Screening. És a dir, l'algoritme classifica amb risc baix amb més proporció que amb risc alt a les avaluacions amb incidències posteriors. Destaca el cas de trencament de condemna screening, on la taxa de TPR es 0%, i.e., l'algoritme en els casos que passen per screening no etiqueta com a risc alt a cap dels interns que cometran una incidència posterior d'aquest tipus.

Per tant, l'algoritme té una bona capacitat d'encert de baix risc però una baixa capacitat d'encert en alt risc. Aquestes mètriques només quantifiquem uns resultats esperats, donada la naturalesa del problema. Hem de tenir present que el percentatge de prevalença de les conductes de risc és molt baixa (vegeu l'apartat Anàlisi de les dades de reincidència).

A l'annex, incloem les mètriques d'avaluació per l'escenari on les avaluacions amb risc mitjà s'agrupen amb les de risc alt. Amb aquest plantejament, les prediccions negatives serien només les avaluacions amb risc baix i les prediccions positives serien les avaluacions amb risc mitjà i alt.

Biaixos potencialment discriminatius

En aquesta secció analitzem els possibles biaixos de l'algoritme respecte a grups de població històricament discriminats. En concret, analitzarem els potencials biaixos amb relació al sexe, l'origen i l'edat de l'intern/a. Aquestes variables reben el nom *d'atributs protegits*. Els grups històricament discriminats reben el nom de *grups protegits* i els grups històricament privilegiats com a *grups de referència*.

En aquest estudi, considerem les següents particions de la població:

Atribut protegit	Grup protegit	Grup de referència
SEXE	Dones	Homes
ORIGEN	Estrangers	Espanyols
EDAT	Menors de 30 anys	Majors de 30 anys

Hem assumit que totes les variables són dicotòmiques. Per l'atribut *EDAT*, hem agafat el punt de tall de 30 anys seguint la literatura prèvia ([1], [6]). El percentatge de nombre d'avaluacions per cada atribut protegit es pot trobar a l'annex.

Per avaluar l'existència de biaixos de l'algoritme fem servir la mesura d'equitat coneguda com a *error-rate balance* [8]. Considerem que un algoritme té biaixos discriminatoris si la taxa d'error del grup protegit és significativament major que la del grup de referència. Per això, analitzem si la taxa de falsos positius (FPR) i la taxa de falsos negatius (FNR) és similar entre grups. Aquestes mètriques reben el nom de *predictive equality* i *equal opportunity*, respectivament.

Centrem l'anàlisi en les reincidències assignades a incidents pels motius explicats a l'apartat *Dades*. Per RisCanvi Screening, ens centrarem en l'anàlisi dels falsos negatius (*equal opportunity)*, és a dir, les avaluacions que han estat classificades amb risc baix, però que tenen una reincidència associada en els 6 mesos posteriors. Hem seguit aquesta lògica perquè un intern classificat erròniament amb risc baix a totes les escales, no passarà per l'escala completa i, per tant, no serà possible detectar el risc en una reavaluació. Per

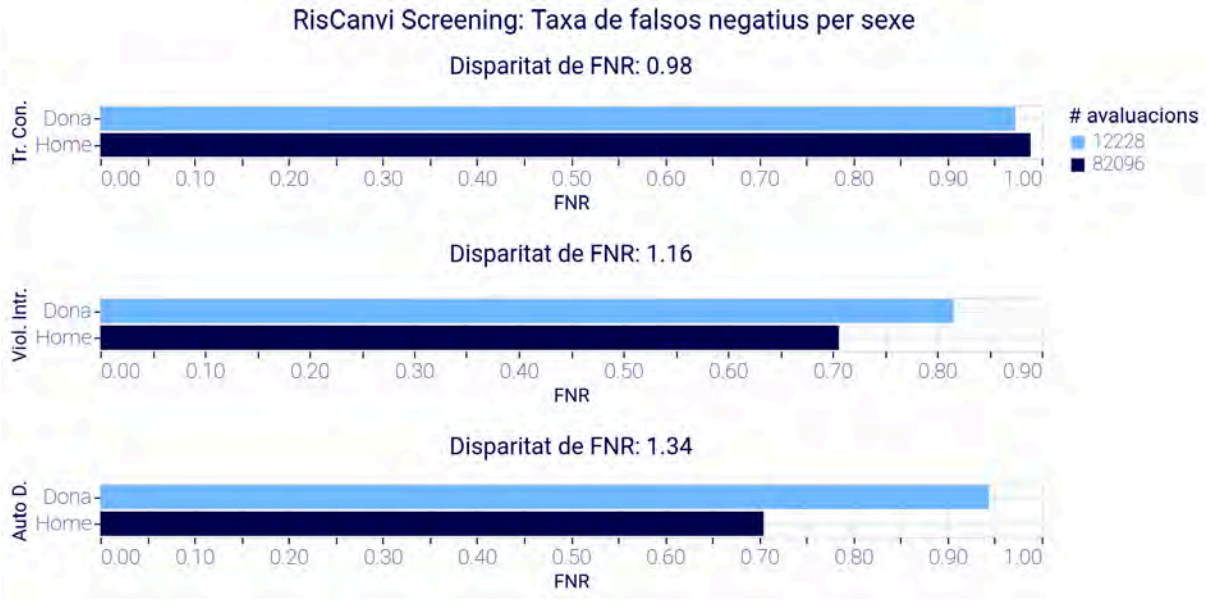
RisCanvi Completa ens centrarem en l'anàlisi dels falsos positius (*predictive equality),* les avaluacions que han estat classificades amb risc alt, però que no tenen una reincidència associada en els 6 mesos posteriors. Així doncs, centrem l'atenció en aquells interns amb risc alt, als que potencialment se l'apliquin mesures de prevenció del risc, però que no cometran cap incidència en els 6 mesos posterior. Seguidament, detallem l'anàlisi per a un dels factors i incloem la resta de resultats a l'apèndix.

Si els valors de FPR i FNR són similars per a tots els grups (les barres de les figures a continuació tenen la mateixa llargada), no hi ha biaix. El valor que ens indica si existeix biaix és la disparitat de FNR o FPR (la proporció entre els valors d'un grup i un altre). Si el valor és 1, no hi ha biaixos; si és >1, el grup protegit té més possibilitat de ser erròniament classificat que el grup de referència; si és <1, el grup de referència és classificat erròniament més sovint. Per determinar si hi ha biaix o no posem els punts de tall en la taxa de disparitat en 0,2 [11]. Les mètriques fora de l'interval [0,8, 1,2] indicarien la presència de biaix.

Anàlisi de biaixos per sexe

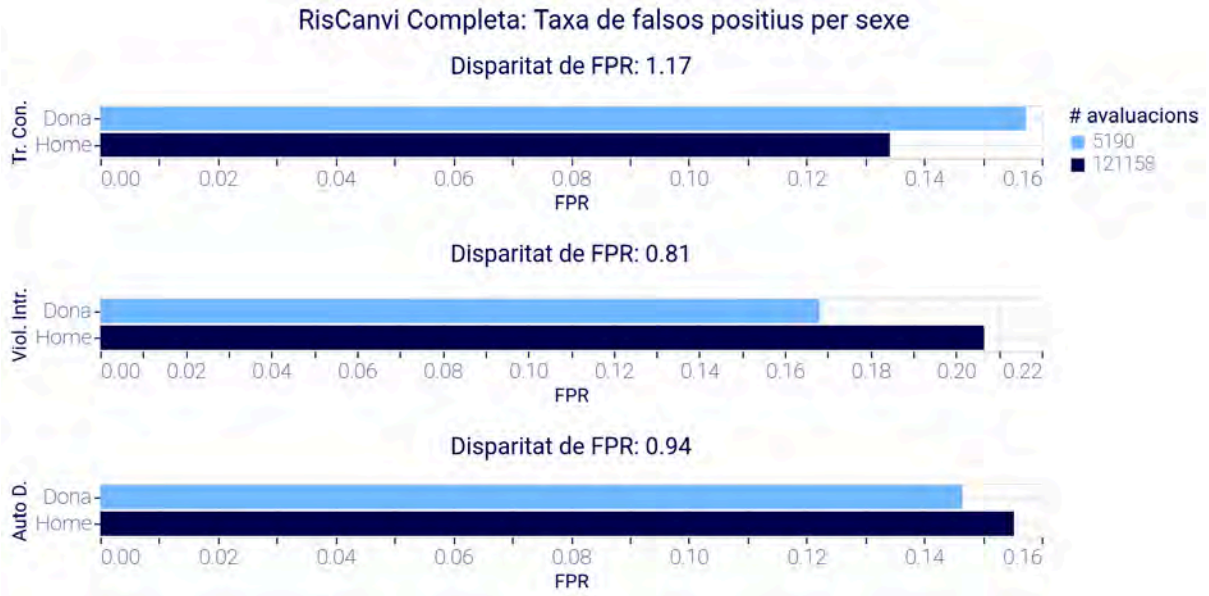
En la següent imatge presentem la taxa de falsos negatius (RisCanvi S) i falsos positius (RisCanvi C) per sexe amb la seva corresponent taxa de disparitat per a les tres reincidències amb major volum de dades. Els casos de Reincidència General i Violenta es troben a l'annex.

Per al cas de Violència Autodirigida a l'escala Screening, veiem que la disparitat és més gran que 1. Això vol dir que l'algoritme té una taxa de falsos negatius més alt per dones i, per tant, hi ha més dones classificades amb risc baix que després reincideixen en violència autodirigida.



Exemple d'anàlisi de biaixos per sexe per a RisCanvi S per a: violència Autodirigida (barres superiors), trencament de condemna (central) i violència intrainstitucional (inferior).

Per a l'escala Completa, veiem que la ràtio de falsos positius és semblant entre grups i en conseqüència la disparitat és molt propera a 1. Així doncs, podem concloure que els algoritmes no tenen biaixos respecte al sexe excepte en el cas de Violència Autodirigida escala Screening.



Exemple d'anàlisi de biaixos per sexe per a RisCanvi C per a: violència Autodirigida (barres superiors), trencament de condemna (central) i violència intrainstitucional (inferior).

En la següent taula presentem el resum de l'anàlisi dels potencials biaixos discriminatoris i a l'annex adjuntem els gràfics d'error-*rate balance* per a la resta d'atributs protegits. No incloem en aquesta avaluació les anàlisis on el nombre de reincidents del qualsevol dels grups sigui baix (<30 casos).

Atribut	**Grup**		**Discriminació per algun risc**	
	protegit	**de referència**	**RisCanvi Screening**	**RisCanvi Completa**
SEXE	Dones	Homes	Per **AD**, s'escapen més casos de dones (d=1,34) i per **RG,** més d'homes (d=0,62). No significativa per les altres.	No significativa.
ORIGEN	Estrangers	Espanyols	Estrangers lleugerament menys etiquetats per **AD** (d=1,12), i notablement menys etiquetats per **RG** (d=1,90) i **RV** (d=1,25).	Per als espanyols més error per **TrC** (disparitat=0,72) i molt més error per **AD** (d=0,5) i **RG** (d=0,56).
EDAT	<30 anys	≥30 anys	Per **AD i RG**, s'escapen més casos de <30 anys (d=1,27, d=1,13).	Més errors pels <30 anys per **Intra** (d=1,16), **RG** (d=1,20) i **RV** (d=1,12). Més errors pels >30 anys per **AD** (d=0,81), i **TrC** (d=0,83).

En general, arribem a les següents conclusions:

- S'observa falta de detecció de risc de Violència Autodirigida per als grups protegits, especialment les dones i els menors de 30 anys, en l'algoritme de Screening. En una futura anàlisi s'hauria de determinar si és necessari definir noves variables per RisCanvi Screening o bé triar uns altres pesos, punts de tall o algoritme (com p. ex. Catboost). Fins i tot, podria ser recomanable fer servir diferents algoritmes per cada grup.
- Pel cas de Reincidència General Screening, hi ha més homes i més estrangers erròniament etiquetats com a no reincidents.
- Pel Riscanvi Complet, l'algoritme s'equivoca més per al grup de referència (espanyols i >30 anys) per Trencament de Condemna, Violència Autodirigida i Reincidència General.
- Pel Riscanvi Complet, en el cas d'edat i Violència Intrainstitucional i Reincidència General i Violenta, sí que hi ha més error amb el grup protegit (<30 anys), tot i que dins el marge de disparitat acceptat (menys d'1,2).

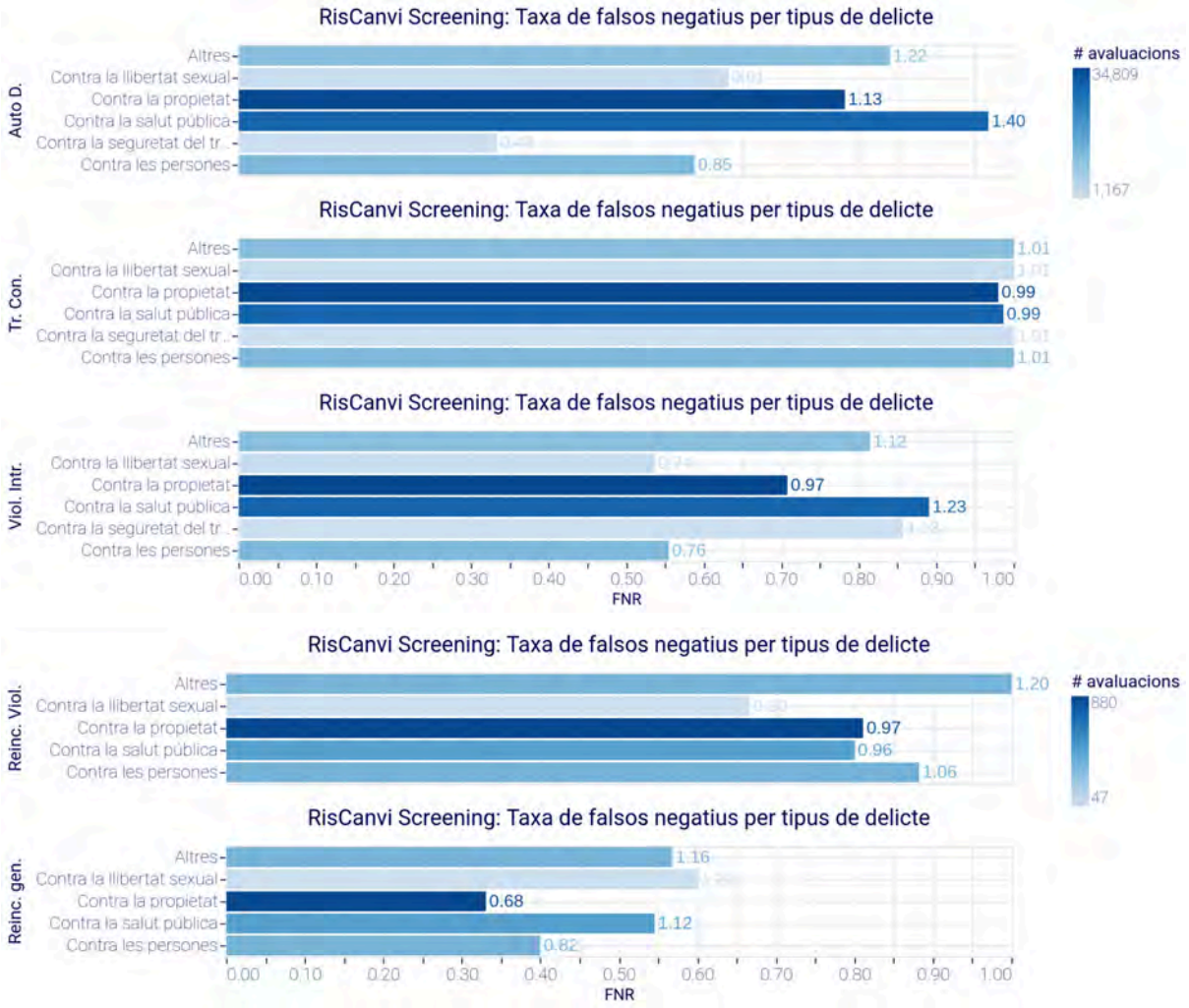
Altres biaixos

Apliquem el mateix tipus d'anàlisi de l'apartat anterior de Biaixos potencialment discriminatoris per a avaluar l'existència de biaixos o comportaments diferenciats de l'algoritme respecte variables no discriminatòries en altres grups de població. S'han analitzat els biaixos per tipus de delicte base i situació (a l'annex) de l'intern revisant la mateixa mètrica d'*error-rate balance* de FNR (Screening) i FPR (Completa).

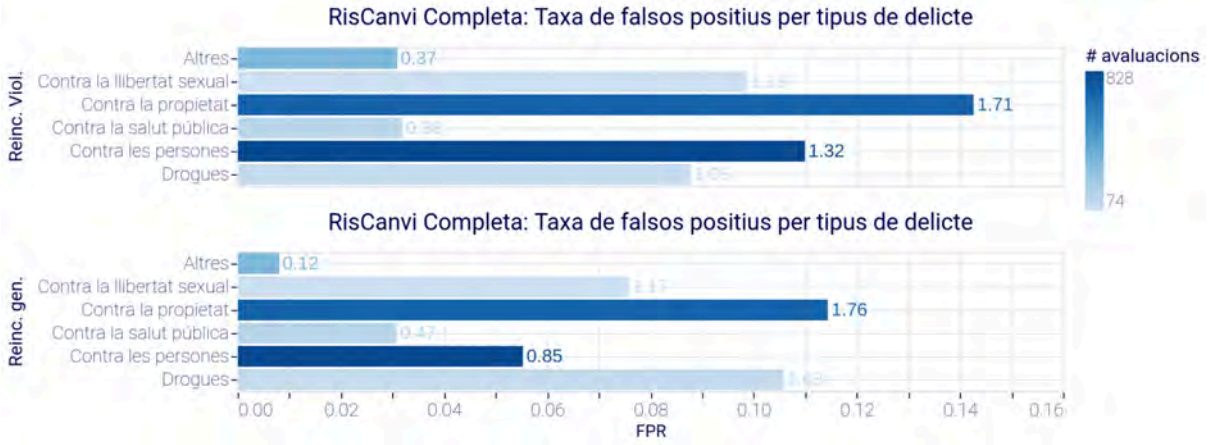
En aquest cas, com que les variables analitzades no són dicotòmiques, la disparitat es mesura respecte a la mitjana de FNR o FPR dels grups considerats. Si el valor és 1, no hi ha biaixos entre les opcions; si és >1, el grup en qüestió té més possibilitat de ser erròniament classificat que la mitjana; si és <1, el grup en qüestió és classificat erròniament menys sovint que la mitjana. Als gràfics mostrem la disparitat al costat de les barres.

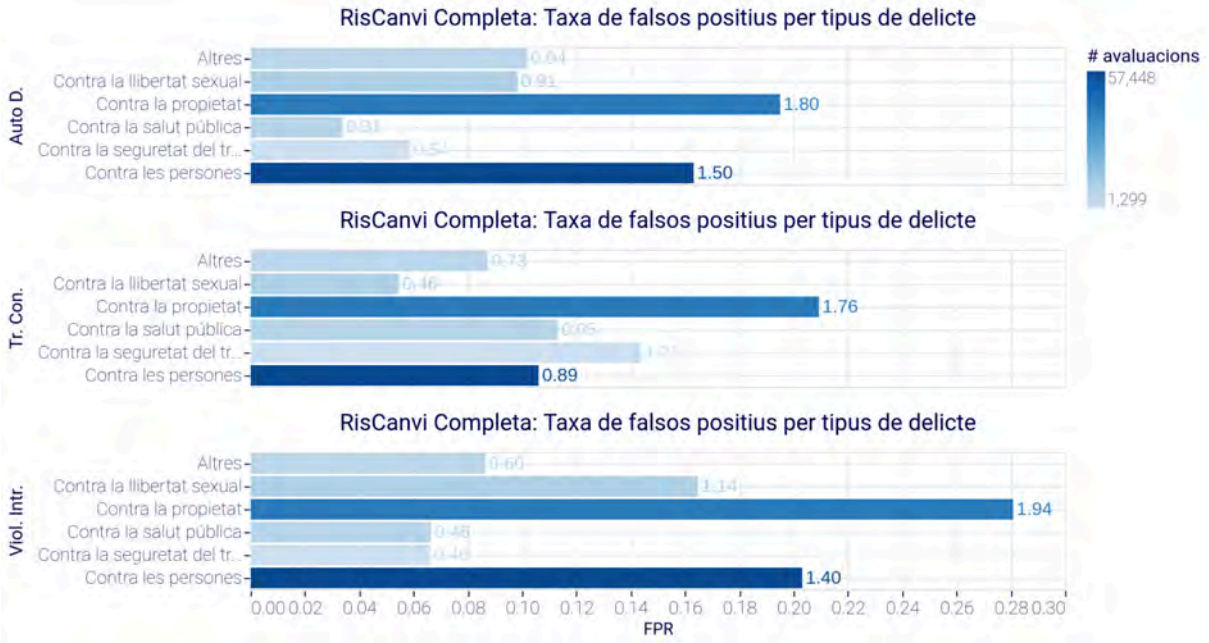
Anàlisi de biaixos per tipus de delicte

En la següent imatge presentem la taxa de falsos negatius (RisCanvi S) i falsos positius (RisCanvi C) per tipus de delicte base amb la seva corresponent taxa de disparitat.



Exemple d'anàlisi de biaixos per tipus de delicte per a RisCanvi S per a les cinc variables.





Exemple d'anàlisi de biaixos per tipus de delicte per a RisCanvi C per a les cinc variables.

En l'escala Screening, trobem paritat d'errors per tots els tipus de delictes per les prediccions de Violència Intrainstitucional i Trencament de Condemna. Per al cas de Violència Autodirigida i pel cas de Violència Intrainstitucional, els interns amb tipus de delicte Contra la salut pública tendeixen estar pitjor classificats (amb més taxa de falsos negatius) que la resta de delictes. Per als casos de Reincidència General, trobem que els delictes contra la llibertat sexual estan lleugerament pitjor classificats respecte a la mitjana.

En l'escala Completa, els interns amb tipus de delicte Contra la propietat tendeixen a estar més mal classificats pels cinc riscos. En el cas de delictes Contra les persones, solen tenir més mala classificació per Autodirigida, Intrainstitucional i Violenta. Els interns amb delictes Contra la llibertat sexual tenen lleugera mala classificació per Intrainstitucional i tenen molts menys falsos positius per Trencament de Condemna que els altres tipus. Els interns amb delictes per drogues tenen una taxa de falsos positius en Reincidència general més alta en comparació amb altres delictes.

A l'annex adjuntem els mateixos gràfics per la variable situació. Veiem que a l'escala Screening tenim una taxa de falsos negatius lleugerament superior pels penats en les Violències autodirigides i Intrainstitucional, però hem de tenir present que també tenim moltes més avaluacions i prediccions precisament pels interns en aquesta situació. No s'ha inclòs aquesta anàlisi per les reincidències General i Violenta, perquè les dades pertanyen a excarcerats i, per tant, un estudi de biaixos per situació de l'intern perd sentit. Massa poca estadística per avaluar correctament.

Conclusions algoritme actual

Els resultats principals són:

- Els algoritmes actuals presenten uns bons nivells en termes d'AUC (~0,8 AUC) i d'exactitud (~90%), excepte per al cas de Trencament de condemna.
- Les exactituds són altes, essent la Reincidència General la que té més marge de millora. Tot i que remarquem que aquesta mètrica no és la més adecuada pel petit percentatge de positius que trobem a les dades.
- Els punts de tall escollits en totes les violències prioritzen tenir una taxa de falsos positius baixa (FPR entre 1 i 31 %) per reduir els casos de predicció de risc alt erronis. Això vol dir que Riscanvi té més tendència a donar un risc baix o mig.
- En conseqüència, el nombre de falsos negatius és alt (FNR entre 40 i 100 %). Això, sumat a l'extrema baixa reincidència (~5 %), fa que el model no identifiqui com a risc alt correctament els interns que sí que acaben reincidint (TPR), sobretot en el Trencament de condemna (on la prevalència és encara inferior, 0,5 %).
- A causa també a la baixa reincidència, els interns amb risc baix o mig no acaben reincidint (TNR > 88 % de mitjana) en totes les violències.
- No s'han detectat biaixos discriminatoris en el RisCanvi Complet cap a grups protegits en termes de sexe, edat ni nacionalitat, excepte pel cas de violència intrainstitucional en el cas de menors de 30 anys (però dins del marge de disparitat acceptat).
- Amb l'algoritme Screening i per Violència Autodirigida, hi ha més proporció de falsos negatius en els casos protegits (dones, estrangers i menors de 30 anys).
- En la versió actual de punts de tall, l'algoritme es comporta diferent quan s'analitza per tipus de delicte. Per situació de l'intern hi ha massa poca estadística per avaluar.

El cas que s'hauria d'atacar abans seria el de l'algoritme de Screening per a Violència Autodirigida. Essent que en el cas de RisCanvi Completa no s'observa discriminació, s'hauria de valorar si considerar més factors de risc en l'Screening per Autodirigida (utilitzar-ne més dels actuals 8, tot i que ens trobem bastant al límit) o incloure nous factors de risc a Screening que capturin millor els casos de Violència Autodirigida.

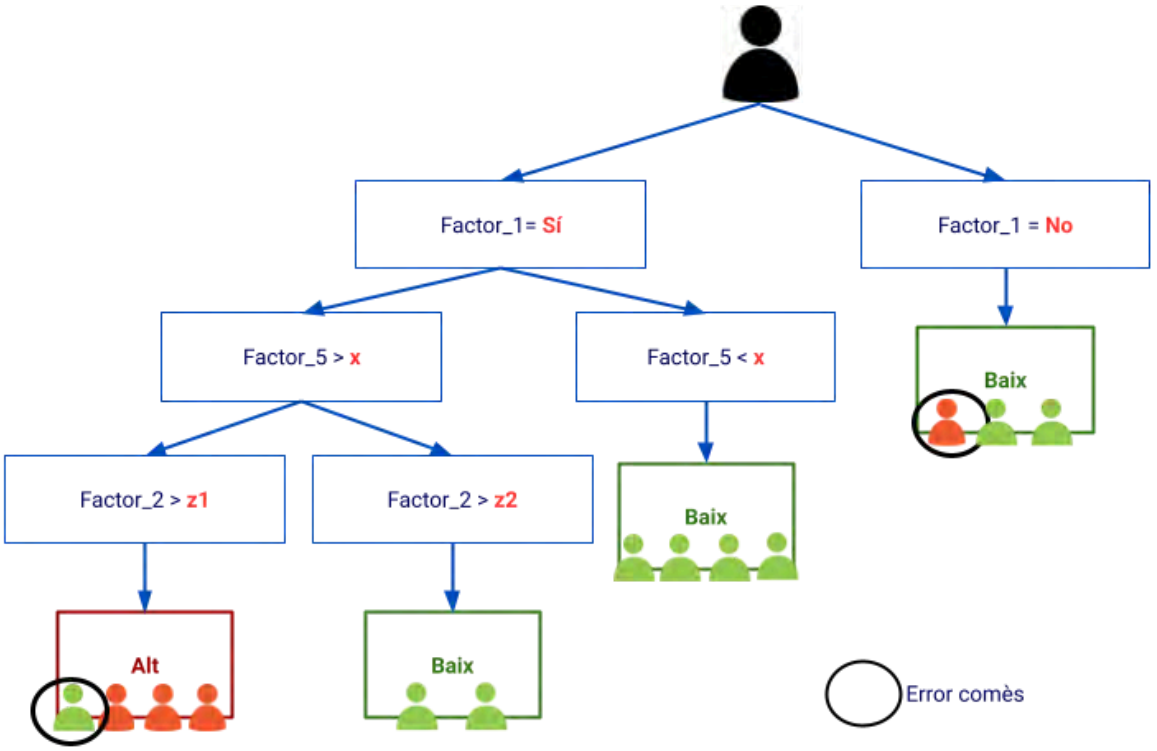
Nous algoritmes

A part d'analitzar el Riscanvi actual, un altre objectiu d'aquesta auditoria algorítmica és explorar algoritmes d'intel·ligència artificial més moderns per veure si actualitzant els models es poden obtenir millors resultats. Després de fer diverses proves, ens hem centrat en el model Catboost, un algoritme especialment dissenyat pel cas en què hi ha diverses variables d'entrada categòriques, com en les avaluacions RisCanvi, i amb el que hem obtingut millors resultats a les proves prèvies. A més, a través del mètode SHAP, afegim una capa d'explicabilitat a les prediccions de Catboost, que ens dona una mesura de la importància de cadascun dels factors de risc i el seu impacte en la predicció. Es tracta d'un primer pas per validar la utilitat d'algorismes d'aquest tipus, que permetrien generar prediccions més acurades, fiables i transparents.

Funcionament

El model Catboost, com la regressió logística, i a diferència de la suma ponderada, és un algoritme supervisat. És a dir, un algoritme que optimitza de manera autònoma els seus paràmetres mitjançant la minimització de l'error de les seves prediccions en un subconjunt de dades d'entrenament en el qual la variable a predir és coneguda. En el cas de RisCanvi, les variables d'entrada són les avaluacions i la variable a predir són les dades de reincidència observada. El nostre conjunt d'entrenament és, per tant, el nostre històric d'avaluacions i casos de reincidència.

Catboost pertany a la família d'algorismes de *Gradient Boosting*, i per generar la predicció fa servir arbres de decisió. D'aquesta manera, l'algoritme aprèn una sèrie de regles basades en els factors d'entrada per arribar a la predicció final. A la següent figura veiem un exemple de com funcionaria un algorisme de la família dels arbres de decisió.



Exemple de funcionament d'un arbre de decisió. Element utilitzat per Catboost i altres algoritmes de la família de Gradient Boosting Trees.

L'algoritme Catboost és de codi font obert (*open source*, codi verificable i reproduïble) i és fàcil d'utilitzar i d'implementar. Presenta bon funcionament amb dades heterogènies i permet obtenir explicabilitat del valor predit. Per aquests motius, pensem que és un bon candidat per reemplaçar els algoritmes actuals.

Per extreure el màxim potencial a l'aprenentatge automàtic, s'ha alimentat el model amb els 43 factors disponibles per a cada risc. D'aquesta manera, el model escull les variables que més l'ajuden a determinar la probabilitat de reincidència. L'entrenament s'ha fet amb l'històric de dades disponibles i s'ha reservat un 20 % de les dades per fer la validació dels resultats.

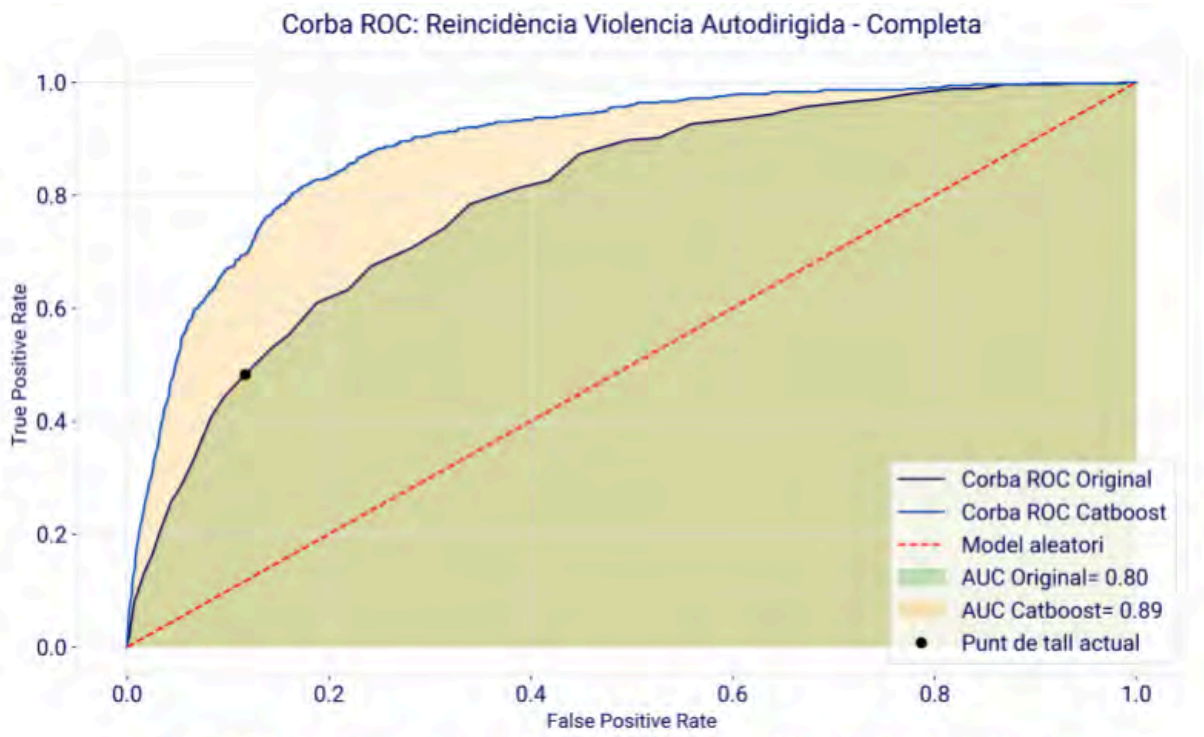
Resultats de l'algoritme Catboost

Les prediccions del model s'han avaluat sobre un conjunt de dades de test (dades mai vistes pel model durant la fase d'entrenament). Sobre aquest mateix grup d'avaluació, s'ha calculat la mètrica AUC-ROC per comparar amb el model actual per RisCanvi Complet. Els resultats en termes d'AUC-ROC són entre un 3 i un 30 % millors que en l'algoritme actual, on el cas de Trencament de condemna destaca per la gran millora. Les millores més petites es donen els els casos on hi ha menys dades disponibles (p. ex., Reincidència violenta).

RISC	ESCALA	AUC-ROC	
		Nou model: Catboost	Model actual
Violència autodirigida	completa	0,89	0,80
Violència intrainstitucional	completa	0,80	0,75
Trencament condemna	completa	0,74	0,57
Reincidència general	completa	0,69	0,62
Reincidència violenta	completa	0,73	0,71

Taula de resultats AUC per ROC per risc i escala
Un AUC = 1 seria un algoritme perfecte. Un AUC = 0,5 seria aleatori.

Per exemple, podem visualitzar els gràfics ROC-AUC pel cas de Violència Autodirigida on, clarament, el model Catboost obté resultats molt millors.



Corba ROC per la recreació de l'algoritme RisCanvi i pel model Catboost per Violència Autodirigida escala Completa

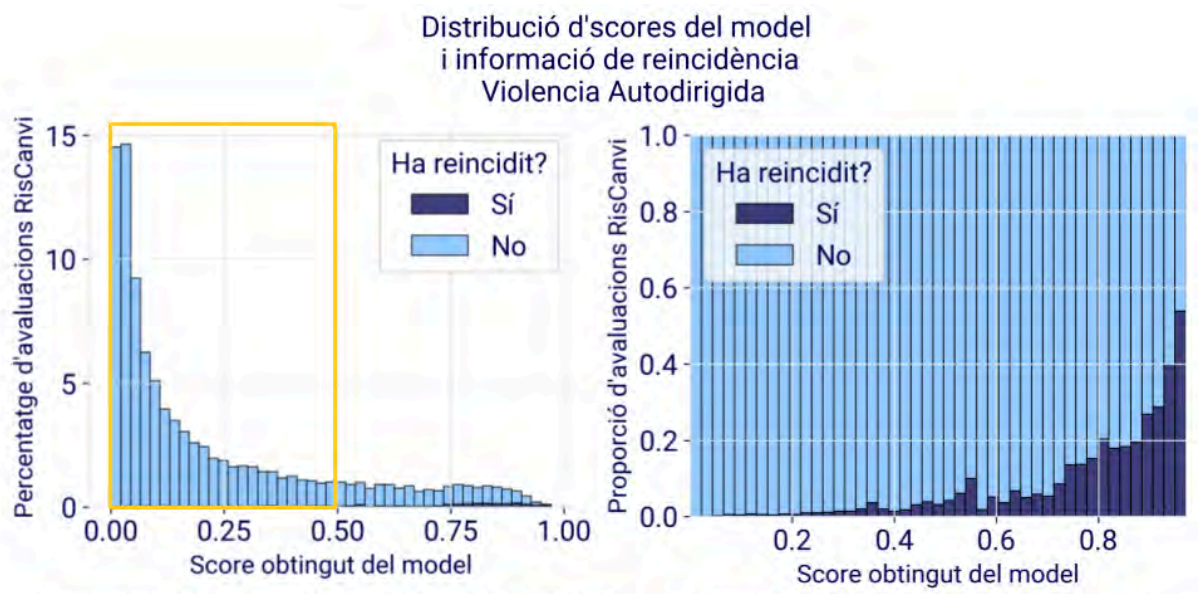
En aquesta gràfica, es pot observar que, per exemple, si fixem la freqüència de falsos positius del model i punts de tall actuals (~13 %), amb el nou model aconseguiríem aproximadament 20 punts percentuals més de freqüència de positiu veritable (TPR d'un 50 % a un ~70 %). És a dir, assumint la mateixa taxa de falsos positius pels 2 models, el nou model etiquetaria correctament 7 de cada 10 interns reincidents, en lloc de 4 de cada 10 del model actual. O al revés, en un punt de tall on tots dos models tinguessin el mateix grau d'encert assignant risc alt als interns reincidents, la taxa de positius veritables (TPR~ 50 %),

el model actual tindria una taxa d'error de falsos positius més del doble que el nou model que proposem. A l'annex podreu comprovar com les millores en les altres 4 violències analitzades és similar. La millora en precisió del model Catboost que proposem és molt remarcable.

A continuació, analitzem en més detall en els resultats del model cas per aquest mateix cas de Violència Autodirigida.

Predicció de risc de reincidència en Violència Autodirigida amb Catboost

Per començar, hem de tenir present que hi ha pocs casos de Violència Autodirigida i, per això, són tan difícils de capturar. Per analitzar el comportament de l'algoritme, hem comptat les vegades que l'algoritme assigna una probabilitat de violència autodirigida i, d'aquestes, en quins casos realment hi ha hagut un incident d'aquest tipus. En relació amb la sortida del model, puntuacions (*scores*) més altes donen lloc a prediccions de risc més altes. A la següent figura mostrem els resultats.



Histograma dels scores obtinguts pel model per Violència Autodirigida. El requadre groc inclou ~83 % de les avaluacions. Hi afegim la informació de si ha reincidit o no.

Proporció d'avaluacions on l'intern ha reincidit o no en funció de l'score assignat pel model (Info complementària a l'histograma)

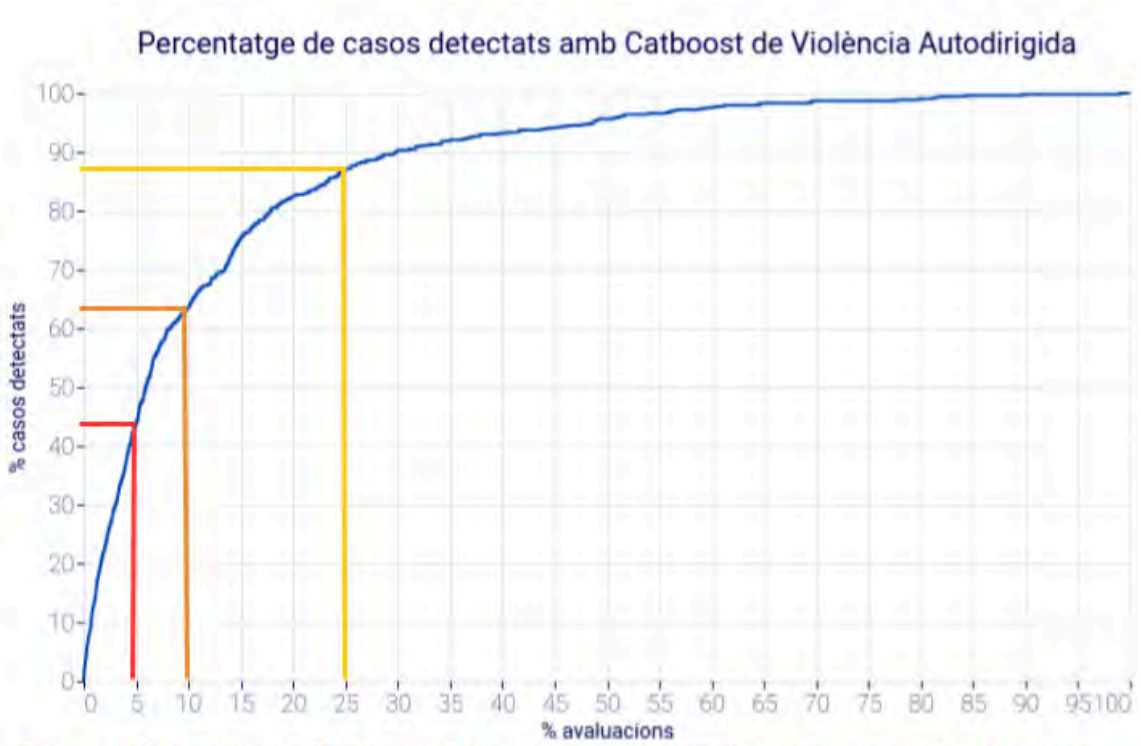
Al gràfic de l'esquerra, veiem que la immensa majoria de les avaluacions (més del 83 %) tenen una probabilitat assignada menor de 0,5. Els casos de violència es concentren a probabilitats altes del model i, per tant, concloem que el model té informació per determinar la probabilitat de reincidència.

Així i tot, al gràfic de la dreta observem que fins i tot en el 47 % de les avaluacions amb l'score per sobre de 0,95 acaben sense cometre Violència Autodirigida. No obstant això, s'ha de tenir en compte que aquest fenomen pot veure's afectat del fet d'haver aplicat mesures preventives i, així doncs, aquesta persona amb risc alt ha acabat sense cometre violència contra ella mateixa.

Per una altra banda, podem mirar els resultats en termes de la capacitat d'actuació dels equips. Algunes de les preguntes que volem contestar són:

- Sobre quants interns puc actuar?
- Si actuo sobre un 10 % de les avaluacions, quants casos de reincidència estaré capturant?

A la següent gràfica mostrem el percentatge de casos detectats versus el percentatge d'avaluacions, ordenades per probabilitat de reincidència calculada amb Catboost, i per score de l'algoritme RisCanvi actual. També, mostrem en vermell, taronja i groc diferents punts de tall que podríem aplicar per implementar mesures correctives.



Proporció de casos reals de reincidència detectats en funció del punt de tall escollit, que està directament relacionat amb la quantitat d'avaluacions que es consideren com a Alt risc.

Així doncs, si considerem només el 5 % de les avaluacions amb més probabilitat de reincidència, ja estarem atacant gairebé el 50 % dels casos que patiran violència Autodirigida (punt de tall vermell). Si agafem el 10 % de les avaluacions amb més probabilitat de reincidir, capturem pràcticament el 70 % dels casos de reincidència. Finalment, si podem actuar sobre un 25 % de les avaluacions, capturarem el 90 % dels incidents de Violència Autodirigida.

Els gràfics d'avaluació per la resta de reincidències es troben a l'annex.

Explicabilitat

L'ús d'aquest tipus d'algoritmes permet també entendre com s'ha obtingut el valor final: quins factors han influït a assignar un nivell de risc i amb quin grau. L'algoritme determina automàticament quins són els factors importants per calcular la probabilitat de reincidència. Incloent aquesta informació en el sistema RisCanvi permetria donar una explicació del risc predit amb l'eina a l'equip avaluador.

Els algoritmes, igual que les persones, es poden equivocar. Per això, és crucial que el personal que pren decisions segons les prediccions dels algoritmes RisCanvi tingui una intuïció de com s'ha generat la predicció. Acompanyar el nivell de risc amb una explicació dels factors més rellevants fomentaria l'esperit crític dels professionals, que podrien decidir de manera fonamentada si estan d'acord amb el nivell predit per l'algoritme o no. D'aquesta manera, l'algoritme deixaria de ser una *caixa negra* en la qual confiar cegament.

Tot i que l'article 22 del GDPR no s'aplica en aquest cas d'ús, ja que sempre hi haurà una validació manual per un humà, és recomanable comptar amb eines per justificar la presa de decisions, no només per al personal avaluador sinó també a interns o jutges que ho poguessin demanar.

Variables més rellevants per a l'etiquetatge de risc

L'algorisme de Catboost, ens dona indicacions de quins criteris són més o menys importants per predir una probabilitat alta de reincidència. En la següent taula, per cada reincidència o violència a predir, llistem els 10 factors de risc més rellevants, i.e., que tenen un pes més gran quan l'algorisme construeix l'output. Hem pintat de color blau, els factors de risc que l'algoritme actual té en compte per construir l'output.

RISC	TOP1	TOP2	TOP3	TOP4	TO5	TOP6	TOP7	TOP8	TOP9	TOP10
Viol. autod.	F37	F2	F10	F30	F21	F36	F26	F5	F19	F41
T. condemna	F38	F34	F33	F2	F9	F18	F5	F14	F30	F40
Violència Intra.	F10	F12	F2	F30	F43	F38	F37	F5	F6	F19
Reinc. General.	F8	F30	F12	F26	F19	F14	F38	F9	F29	F2
Reinc. Violenta	F8	F7	F12	F30	F41	F10	F15	F26	F21	F9

Factors de risc (FR) més importants per la predicció de la incidència segons l'algoritme Catboost. En blau, els FR considerats en l'algoritme actual de RisCanvi Complet.

Veiem que per als casos de Violència Intrainstitucional i Reincidència General, l'algoritme Catboost i els algoritmes actuals estan d'acord en la majoria dels factors. Tot i això, els factors amb més pes per al cas de Reincidència General amb el model actual, i.e., temps ininterromput a presó (F6) i nivell educatiu (F18), no són detectats com a importants per Catboost. Per als casos de Violència Autodirigida i Reincidència Violenta, veiem que només la meitat dels factors rellevants són considerats als models actuals i que, per al cas de Trencament de condemna, només un factor important es té en compte actualment. És per aquest motiu que l'algoritme Catboost pot captar millor el comportament dels interns i presenta una gran millora de resultats en termes d'AUC-ROC.

També és interessant veure quins d'aquests factors importants es tenen en compte en l'escala Screening. A la següent taula veiem que per Trencament de condemna cap factor important per predir la reincidència segons Catboost s'està tenint en compte en la versió actual, i per la resta de reincidències i violències només trobem entre 2 i 4 coincidències. És a dir, podem dir que hi ha informació rellevant per la predicció que no està sent contemplada en la fase de cribratge (escala Screening)[3].

RISC	TOP1	TOP2	TOP3	TOP4	TOP5	TOP6	TOP7	TOP8	TOP9	TOP10
Viol. autod.	F37	F2	F10	F30	F21	F36	F26	F5	F19	F41
T. condemna	F38	F34	F33	F2	F9	F18	F5	F14	F30	F40
Violència Intra.	F10	F12	F2	F30	F43	F38	F37	F5	F6	F19
Reinc. General.	F8	F30	F12	F26	F19	F14	F38	F9	F29	F2
Reinc. Violenta	F8	F7	F12	F30	F41	F10	F15	F26	F21	F9

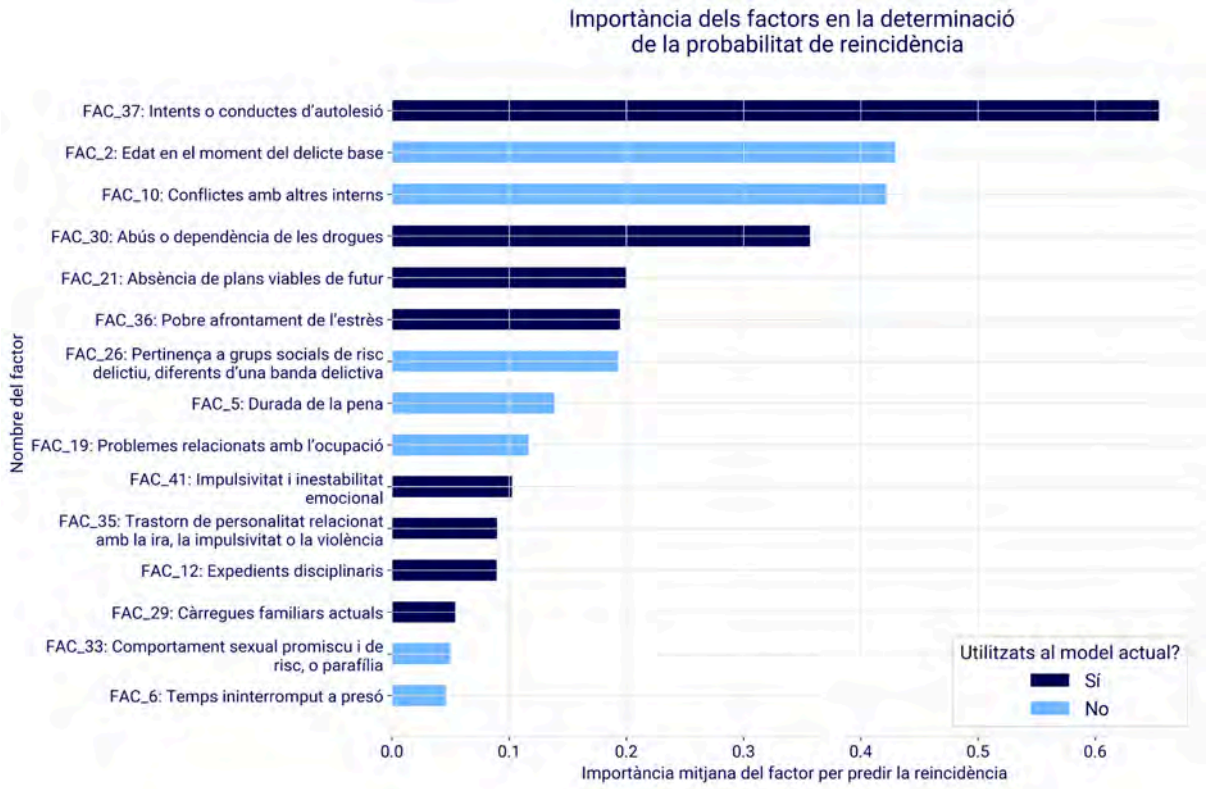
Factors de risc (FR) més importants per la predicció de la incidència segons l'algoritme Catboost. En taronja, els FR considerats en l'algoritme actual de RisCanvi Screening.

A continuació analitzem amb més detall l'exemple concret per al cas de Violència Autodirigida.

Explicabilitat per al cas de Violència Autodirigida amb Catboost

En la següent gràfica mostrem la importància dels factors de risc per determinar la probabilitat de reincidència de violència autodirigida amb Catboost. A més, en blau fosc, mostrem les variables que ja es contemplen en el model actual.

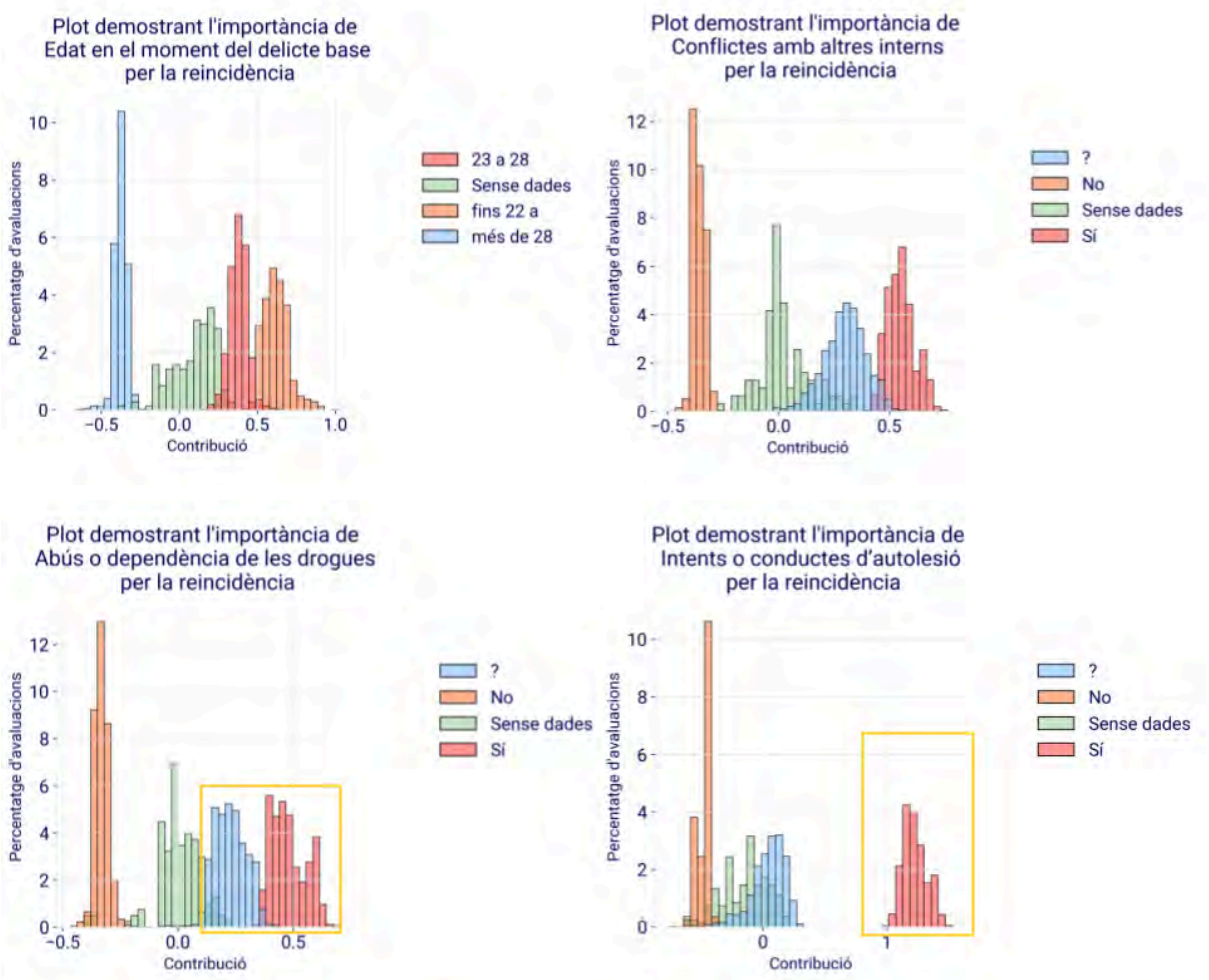
[3] Noteu que no hi ha una correspondència 1 a 1 entre els factors de l'escala Completa i Screening. Per exemple, el F5 de l'escala Screening (*Problemes amb el consum de drogues o alcohol*) correspondria al conjut de F30 (*Abús o dependència de les drogues*) i F31 (*Abús o dependència a l'alcohol*) de l'escala Completa.



Factors més importants per a l'algoritme de risc de Violència autodirigida i quins coincideixen amb els utilitzats pel model actual (proposats per l'equip d'experts)

Catboost ha determinat de manera automàtica que el factor més rellevant per assignar la reincidència en Violència Autodirigida és el número 37, que correspon a intents o conductes d'autolesió. Aquest fet concorda perfectament amb la intuïció humana que podríem tenir. No obstant això, les següents dues variables més rellevants pel model són l'edat en el moment del delicte base i l'existència de conflictes amb altres interns. Aquests factors no semblen tan obvis a l'hora de predir violència autodirigida i, de fet, no es tenen en compte actualment en l'algoritme dissenyat per l'equip del Dr. Pueyo. En total, dels 15 factors més importants per predir el risc de Violència Autodirigida amb l'algorisme Catboost, 8 coincideixen amb els proposats per l'equip d'experts.

És possible aprofundir més en cadascun dels factors i observar el pes que té cada categoria per cada factor. Per exemple, a la següent imatge mostrem la distribució de la importància per cada resposta pels factors 2, 10, 30 i 21.

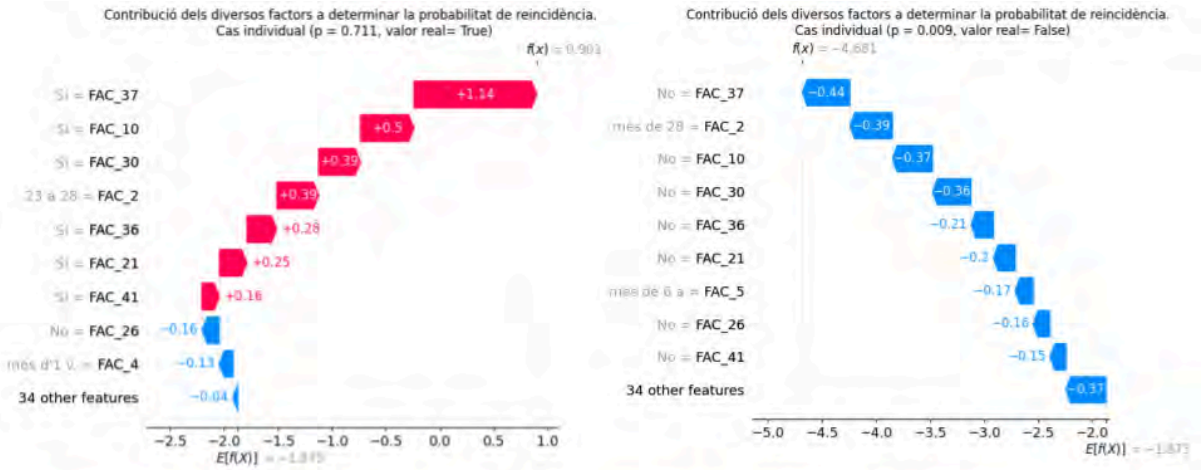


Importància de cada categoria per als factors de risc 37 (intents o conductes d'autolesió), 2 (edat en el moment del delicte base), 10 (conflictes amb altres interns) i 30 (abús o dependència de les drogues)

Així doncs, pel factor 2 (edat en el moment del delicte base) veiem que l'intern tindrà una puntuació més alta de reincidència en violència autodirigida si ha comès el delicte base sent menor de 22 anys que un intern que l'hagi comès amb més de 28 anys. Pel cas del factor 10, un intern amb conflictes amb altres interns també tindrà més probabilitat de cometre reincidència en violència autodirigida que un intern que no tingui conflictes dins de la presó. Aquestes variables no es tenen en compte en el model actual i podrien ser claus a l'hora d'obtenir prediccions més acurades.

Per als factors 30 (abús o dependència de drogues) i 37 (intents o conductes d'autolesió) podem fer la comparativa dels pesos del model Catboost amb el model dissenyat per l'equip del Dr. A. Andrés Pueyo. En el cas del factor 30, el model de suma ponderada posa el màxim pes (3 punts) a les respostes *Sí* i *?*. De la mateixa manera, veiem que l'algoritme Catboost dona més pes a aquestes categories. En el cas del factor 37, l'algoritme de suma ponderada puntua amb 3 punts a la resposta *Sí* i amb 0 punts a la resta de respostes. Veiem que el nou algoritme segueix un patró semblant, donant màxima importància a la resposta *Sí* i importàncies semblants i més baixes a la resta de respostes. Així doncs, observem que l'algoritme tot sol aprèn a extreure informació rellevant de les dades reals de reincidència.

Les anàlisis mostrades prèviament formen part dels mètodes d'explicabilitat *global,* és a dir, *a nivell poblacional*. De la mateixa manera, podríem estar interessats a obtenir una explicació *local*, és a dir, obtenir una explicació per una avaluació en concret. A continuació, mostrem la importància de cadascun dels factors per a una avaluació amb alta probabilitat de reincidència en violència autodirigida i que finalment l'intern comet aquesta violència (esquerra), i per a una avaluació que té baixa probabilitat i que finalment l'intern no comet violència autodirigida (dreta).



Contribució de cada un dels factors per determinar la probabilitat de reincidència en Violència Autodirigida per a dos RisCanvis concrets (intern i data fixa)

Gràcies a aquestes visualitzacions, podem veure en el mateix gràfic les respostes per cada factor de l'avaluació (eix *y)* i com ha contribuït cadascuna en la predicció final. El color rosa indica que el valor de la variable d'aquest intern concret respecte la resta augmenta el seu risc de cometre violència autodirigida, mentre que el color blau indica que el valor de la variable d'aquest intern concret respecte la resta indica una menor probabilitat de cometre violència autodirigida. Així doncs, veiem que per l'avaluació de l'esquerra, la majoria dels factors contribueixen positivament, incrementant la probabilitat de reincidència en violència autodirigida, mentre que en el gràfic de la dreta, totes les respostes disminueixen la probabilitat final. Cal notar també que el punt d'inici és el mateix, -1,873, (la mitjana dels log-odds), i a partir d'aquí els factors de risc disminueixen o augmenten la puntuació final.

En resum, l'algoritme Catboost i altres algoritmes d'aprenentatge automàtic semblants, ens dona eines per explicar els resultats obtinguts, tant en l'àmbit global del funcionament del model com en l'àmbit local per cada avaluació, que pot ser molt útil.

Implementació

El protocol RisCanvi té diversos nivells d'implementació. En aquest document ens centrem en l'algoritme que hi ha darrere de les estimacions del risc de reincidència i, per tant, comentarem alguns detalls de la implementació de l'algoritme. No entrarem en el funcionament del protocol a altres nivells perquè s'escapa de la nostra àrea de coneixement. Hem dividit la implementació de l'algoritme en dos nivells:

- **Tècnic**: com s'ha traduït la feina feta per l'equip d'experts de definició de l'algoritme i punts de tall al codi que s'executa als sistemes del departament.
- **Experiència d'usuari**: com interactua l'usuari final amb l'eina i la facilitat d'ús de l'aplicació.

Tècnica

Degut a deficiències en l'accessibilitat que comentem més endavant, no s'ha tingut accés al codi que corre als sistemes del departament fins a la segona fase del projecte. L'accés al codi ha estat fonamental per revisar la implementació algorítmica i poder dur a terme aquesta auditoria de manera més realista. Ens ha permès entendre com es codificaven les variables, comprovar els punts de tall i pesos, com es tracten els casos Sense dades i, sobretot, validar que els algoritmes descrits a la documentació corresponen amb els implementats al sistema.

Durant la fase final del projecte, l'equip de Dribia vam dur a terme una reunió amb un representant del proveïdor de serveis informàtics, amb l'objectiu d'entendre quin tipus de codi s'havia implementat, com estava organitzat i el seu nivell d'accessibilitat. En aquesta reunió, vam veure que:

- El codi que implementava els algoritmes RisCanvi està introduït com a funcions a dins de la mateixa base de dades, en llenguatge del tipus SQL (en concret PSQL).
- Els pesos i els punts de tall estan directament afegits al codi.
- Per realitzar qualsevol mena de canvi, s'ha de realitzar una petició al proveïdor de la base de dades.
- Des de la introducció de la versió 3, no s'ha fet cap modificació del codi.

Aquest tipus d'implementació presenta diversos problemes:

- La integració de les funcions de càlcul de risc a dins de la base de dades limita la possibilitat d'obrir el codi, ja que aquestes funcions estan connectades directament a les taules i per poder fer proves de càlcul dels algoritmes RisCanvi seria necessari publicar tota la base de dades. Per temes de privacitat, això no seria possible. Fins i tot, per poder fer una auditoria algorítmica basada en el funcionament real de RisCanvi o per fer estudis acadèmics per part d'universitats o altres institucions, seria necessari tenir accés a les bases de dades del SIPC. Tot això fa que, a dia d'ara, el nivell d'accés al codi sigui molt restringit i amb intermediaris.

- No segueix els estàndars de codi que demana tenir el codi que sigui fàcil de mantenir, transparent i llegible. Aquí, les fórmules de RisCanvi estan escrites directament al codi, que fa difícil implementar canvis, actualitzar o revisar el model sense impactar en el funcionament de l'eina. A més, seguint aquest paradigma és fàcil cometre errors d'implementació que són difícilment detectables (s'haurien d'implementar testos unitaris, més difícil en aquest entorn).
- Els llenguatges de SQL estan dissenyats i optimitzats per escriure i llegir en una base de dades, però no per a l'entrenament d'algoritmes d'intel·ligència artificial. Per tant, tot i que el llenguatge actual permet implementar un algoritme de regressió logística, en cap cas permet l'entrenament i ús d'un algoritme modern com el que proposem en aquesta auditoria.

Una altra feblesa que hem trobat és la manca d'una documentació tècnica on s'expliqui el funcionament del codi. Això fa que la implementació actual sigui altament dependent del personal que treballa actualment al proveïdor extern. Un canvi de la plantilla que ha escrit el codi suposaria una pèrdua de coneixement tècnica molt alta.

Per aquests motius, de cara a versions futures de RisCanvi, recomanem substituir la crida d'aquestes funcions SQL per un microservei extern a la base de dades. Aquest microservei calcularia la puntuació i el nivell de risc i es podria cridar a través d'una API connectada al frontend. Això faria possible la separació de l'eina de càlcul i la base de dades del SIPC i, per tant, facilitaria les futures auditories i les actualitzacions dels pesos, punts de tall o canvi de l'algoritme sencer. A més, seria possible fer aquesta nova implementació en altres llenguatges com python, que compten amb eines per desenvolupar algoritmes d'intel·ligència artificial.

També recomanem obrir el codi (fer-lo *open source*) ja que permet:

- Fer auditories basades en el funcionament real del model, no en interpretacions de la documentació.
- Detectar errors i detalls tècnics d'implementació. Per tant, es poden tenir algoritmes més robustos i fiables.
- Incrementar la freqüència de les auditories i de l'actualització del model. Així s'aconseguiran algoritmes més precisos i ajustats a les necessitats del sistema penitenciari i a la seva evolució al llarg del temps.
- Donar tant als professionals de l'àmbit, com als interns que ho demanin, una explicació de la lògica que segueix l'algoritme per assignar els nivells de risc, assolint uns algoritmes més transparents.

Experiència d'usuari

Hem tingut accés a la documentació de la plataforma utilitzada pels usuaris de l'eina en format documentació i ens han fet exemple d'ús. No hi ha un entorn de proves on es pugui treballar amb dades no reals, per tant, no hem pogut usar l'eina directament i els comentaris a continuació reflecteixen el resultat de les entrevistes i de la sessió de demostració d'ús de l'eina.

La plataforma utilitzada no s'integra totalment amb totes les fonts de dades disponibles al SIPC. En conseqüència, no es poden validar les evidències introduïdes de forma automàtica o assistida mitjançant un algoritme de validació amb suport. Creiem que seria una millora important poder comptar amb aquest sistema perquè no s'escapessin evidències ja documentades al sistema. En aquesta línia tampoc hi ha accés a les intervencions per evitar reincidències o tipus de violències amb motiu d'una avaluació de risc alt. Aquestes intervencions s'haurien de registrar a un sistema on Riscanvi hi tingués accés, ja que afecten el seu entrenament i a l'avaluació de la seva precisió.

No hi ha una visualització clara dels motius pels quals l'algoritme ha escollit una classificació o una altra. Per a facilitar la feina als equips que treballen amb els interns, creiem que seria molt útil tenir una descripció de quins factors han influït més en la determinació d'un cert nivell de risc de l'algoritme.

Recomanacions de millora

Durant el treball d'auditoria s'han detectat possibles millores a implementar, recollides a continuació:

- **Dades i algoritme**
  - Sistematitzar la recollida de dades d'incidència / reincidència o de no incidència / reincidència d'un intern: recopilar si l'intern ha comès un incident associat a reincidència durant 6 mesos (per trencament de condemna, violència autodirigida, violència intrainstitucional) o 5 anys (reincidència general i violenta) després d'una avaluació de RisCanvi. Per millorar els algoritmes i la seva avaluació
  - Implementar un nou model predictiu basat en models d'intel·ligència artificial més moderns, com el Catboost analitzat en aquest informe, amb nous factors de risc i punts de tall. Aquest model s'ha de desenvolupar en col·laboració d'experts en reincidència amb experts en algoritmes d'intel·ligència artificial. Si això no fos possible, com a mínim s'haurien de reavaluar els punts de tall dels models actuals.
  - Desplegar el nou algoritme amb programari més modern, mitjançant un microservei connectat via API amb la base de dades i interfície gràfica actual, que permeti la implementació del nou algoritme.
  - Creació d'un repositori del codi de l'algoritme així com d'una documentació viva que permeti l'accés ràpid a qualsevol usuari que vulgui consultar o modificar l'algoritme actual.
  - Obrir les dades del model per a anàlisi científica i potencialment al públic en general (anonimitzades adequadament).
  - Fer públic el codi i la documentació de l'algoritme per transparència i a revisió per part de tercers i possibles propostes de millora.
- **Gestió de RisCanvi**
  - Implementar un equip de manteniment i millora contínua de l'algoritme. El control de l'algoritme per part de la direcció general és clau per poder revisar el seu funcionament, mantenir-lo actualitzat i millorar-lo amb les noves dades i coneixement disponibles.
  - S'hauria de posar en marxa un sistema de generació automàtica d'un informe de precisió i control de biaixos, que s'executés periòdicament.
  - Per poder aprofitar el valor de l'informe i entendre i actuar en conseqüència amb els resultats més recents, s'hauria de fer una formació en el funcionament i interpretació dels resultats de l'algoritme RisCanvi per als equips involucrats (a nivell direcció general i als centres).

Bibliografia

Bibliografia referenciada en aquest resum de l'informe final:

[1] *INFORME RISCANVI (VOLUMEN PRIMERO)*. Dr. D. Antonio Andrés Pueyo, Dra. Dña. Karin Arbach Lucioni, Dr. D. Santiago Redondo Illescas. 2010

[2] *RisCanvi: modificación puntos de corte*. Grup d'Estudis Avançats en Violència - UB. 2011

[3] Conjunt de documents de "Modificació algoritme RC trencament condemna", Modificació algoritme RS trencament condemna", "Reincidència delictiva general RC algoritme", "Reincidència delictiva general RS algoritme". 2017

[4] Riscanvi 2.0 Informe final. Kernel Analytics 2017.

[5] *La reincidència en les excarceracions d'alt risc*. Àrea d'Investigació i Formació en Execució Penal. 2022.

[6] *Technical report for the HUMAINT Project at JRC (Science for policy report 2022 - Milestone 2 (2022))*. Dr. Carlos Castillo. 2022.

[7] *Manual d'aplicació del protocol de valoració RisCanvi*. Subdirecció de Programes de Rehabilitació i Sanitat. 2019.

[8] *Fair prediction with disparate impact: A study of bias in recidivism prediction instruments*. A. Chouldechova. 2017.

[9] *Fairness and Algorithmic Decision Making*. Lecture Notes for UCSD course DSC 167, A. Fraenkel. 2020.

[10] *A compendium of research and analysis on the Offender Assessment System (OASys)* Robin Moore (editor). National Offender Management Service 2009–2013.

[11] *Why Machine Learning May Lead to Unfairness: Evidence from Risk Assessment for Juvenile Justice in Catalonia*. Songül Tolan et al. 2019.

[12] *Taxa de reincidència penitenciària 2020*. Àrea d'Investigació i Formació en Execució Penal. 2023.

Annexos

Fitxa tècnica

Participants, període i tasques d'execució de l'auditoria.

Equip

- Afers penitenciaris: Jordi Camps, Marian Martinez, Xavier Buscà
- Dribia: Claudia Herron, Xavier Clotet, Pol Colomer

Dedicació

- Primera fase
  - Inici fase: 15/10/2022
  - Final fase: 23/12/2022
- Segons fase
  - Inici fase: 27/06/2023
  - Final fase: 17/10/2023

Execució

- Recerca prèvia: revisió documentació.
- Entrevistes: equip d'Afers penitenciaris, Professor A. Andrés Pueyo, Sr. Manel Capdevila, Dr. Carlos Castillo[4].
- Revisió estat actual de l'agoritme.
- Proves algorítmiques amb altres models.
- Redacció de l'informe.
  - v1.0. 6 de febrer de 2023: tres tipus de reincidència.
  - v2.0. 17 d'octubre de 2023: inclusió de reincidència general i violenta.
  - v2.1. 9 de gener de 2024: correcció d'errates.

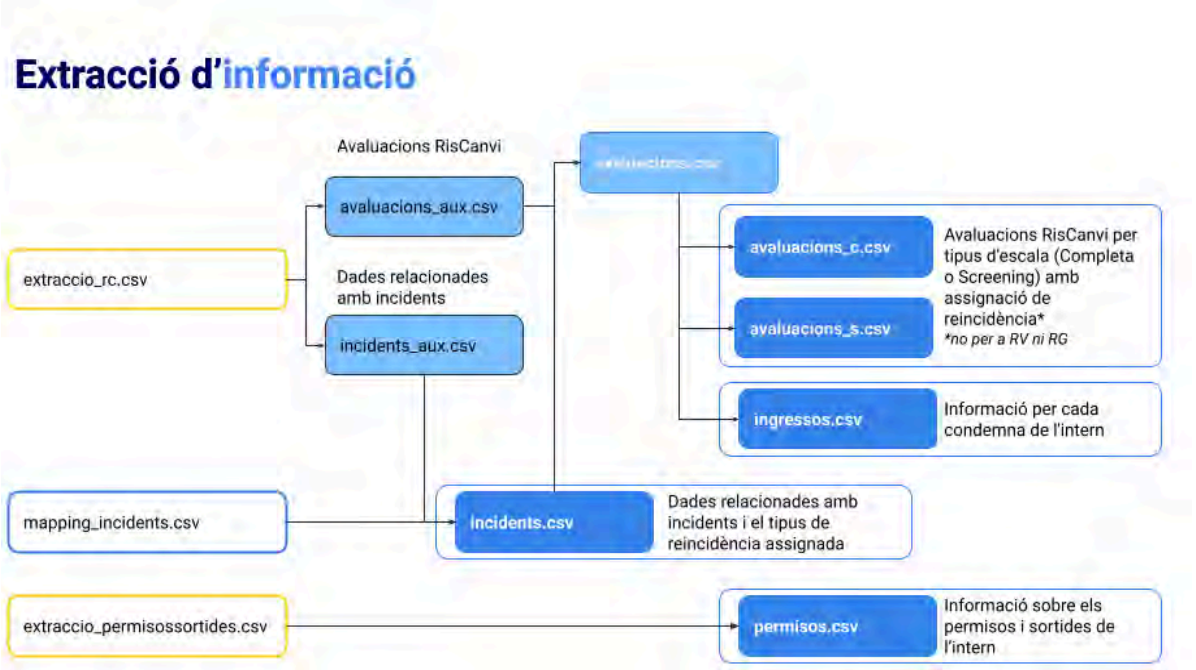
Autoria de l'informe

- Claudia Herron Mulet
- Xavier Clotet Fons
- Pol Colomer de Simon

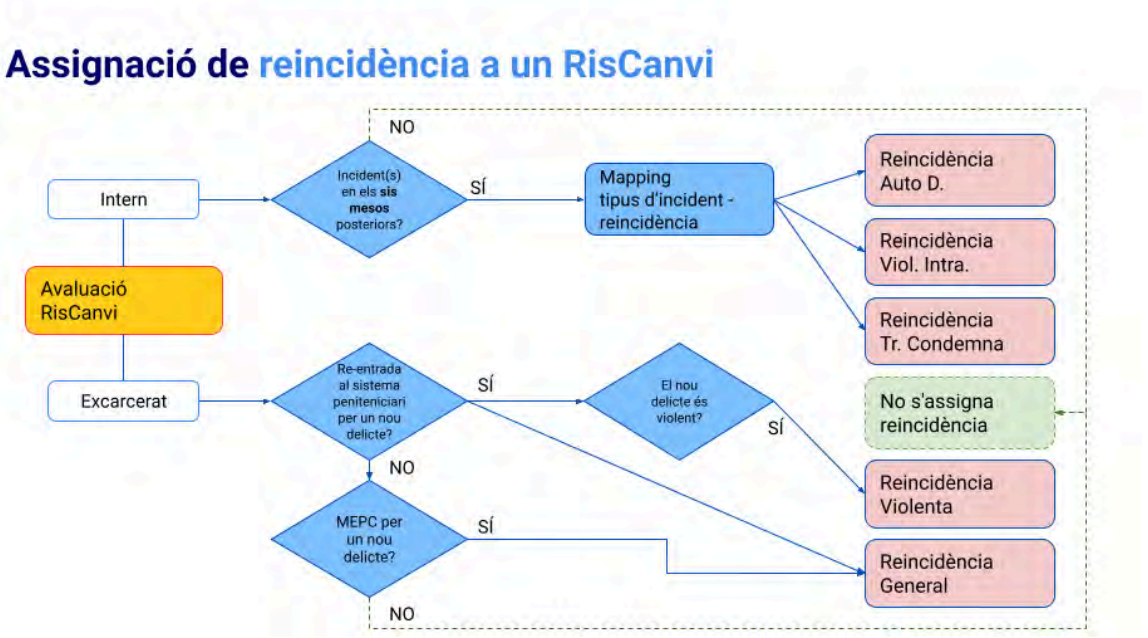
[4] L'equip de Dribia agraeix la disponibilitat i col·laboració de tots els participants en les entrevistes de recerca.

Dades

Organització de les dades en fitxers després de la neteja i transformacions

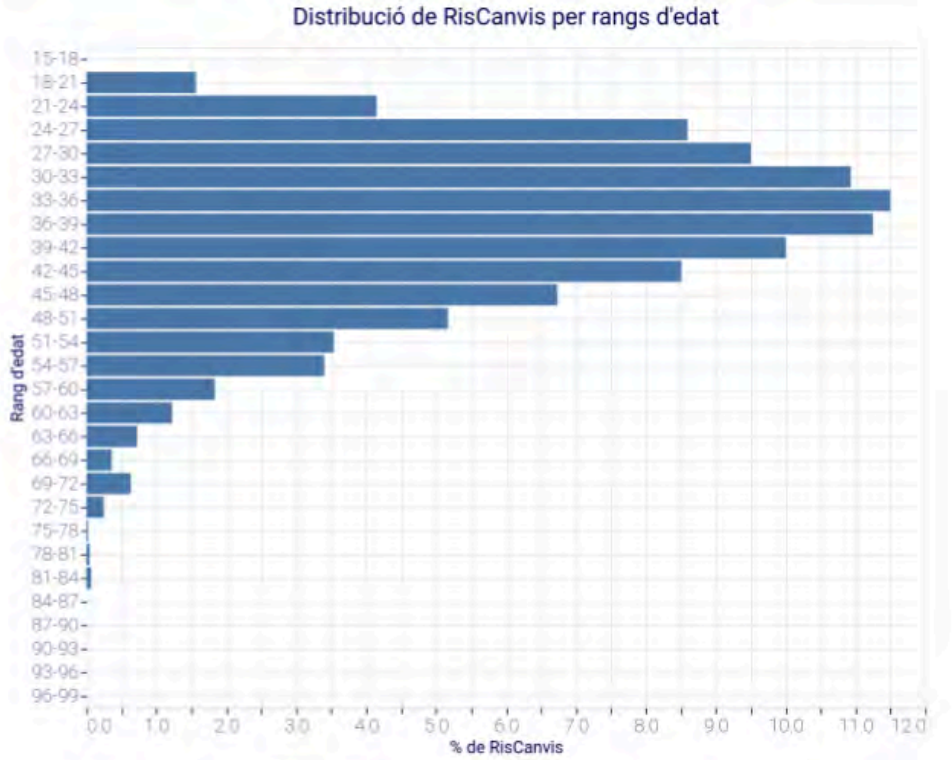


Assignació de reincidència posterior per cada avaluació RisCanvi



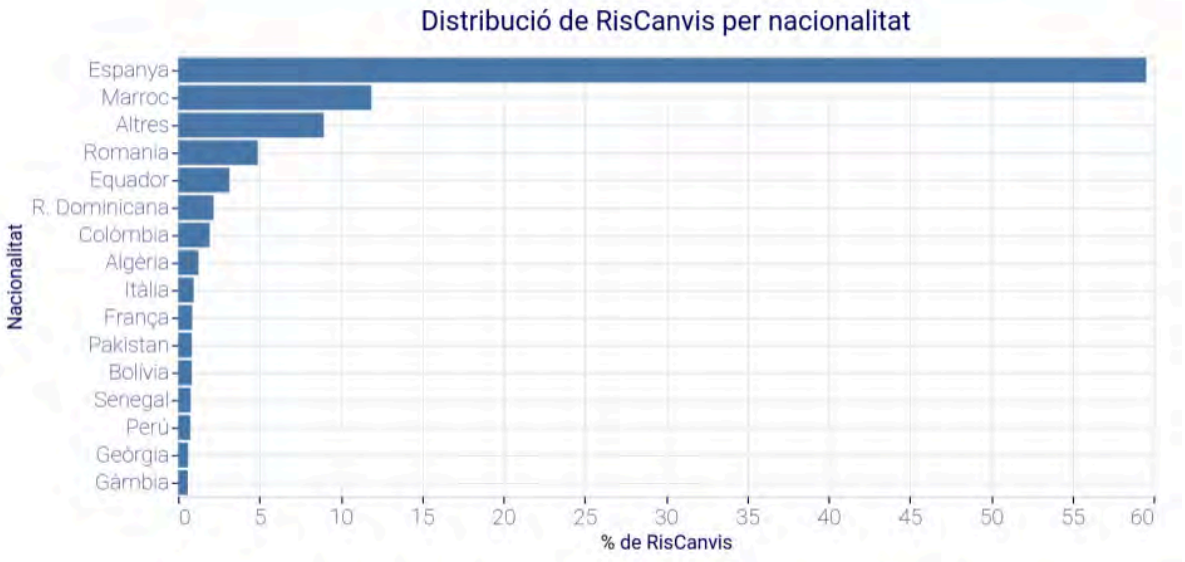
Anàlisi de dades

Distribució d'edats per dades CEFJE



Distribució d'avaluacions RisCanvi per rangs d'edat (dades CEFJE)

Distribució de nacionalitats per dades CEFJE

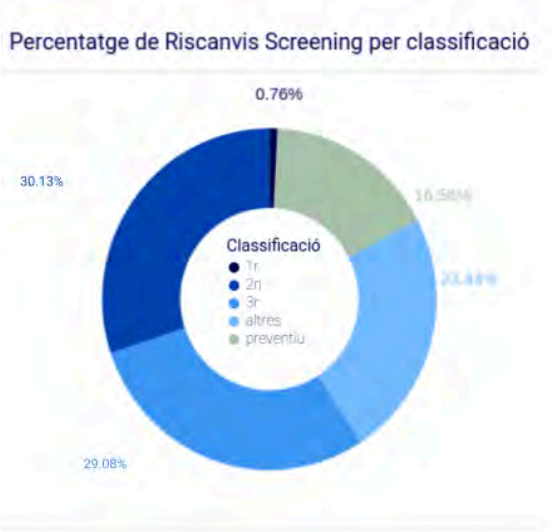


Percentatge d'avaluacions RisCanvi per nacionalitats (dades CEFJE)

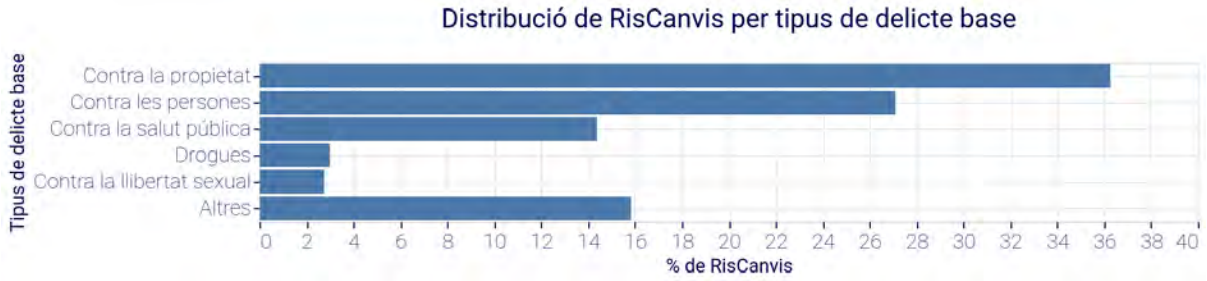
Distribució de sexe per dades CEFJE

Percentatge d'avaluacions RisCanvi per sexe (dades CEFJE)

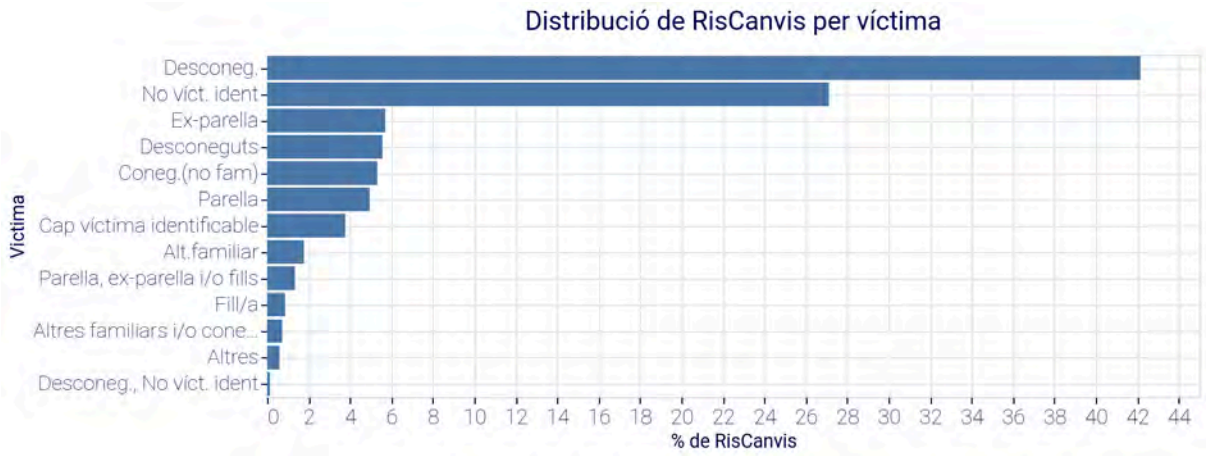
Percentatge de RisCanvis per escala i grau de classificació (dades AP)



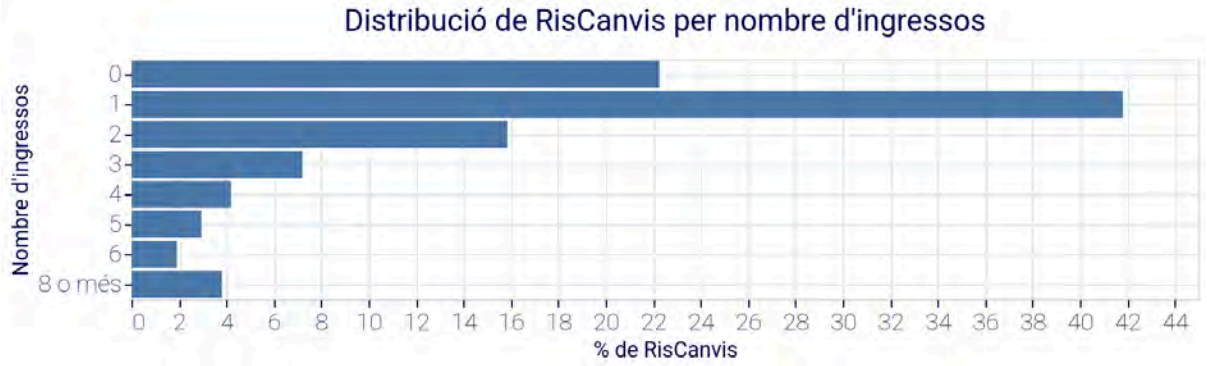
Percentatge de RisCanvis per tipus de delicte base (dades CEFJE)



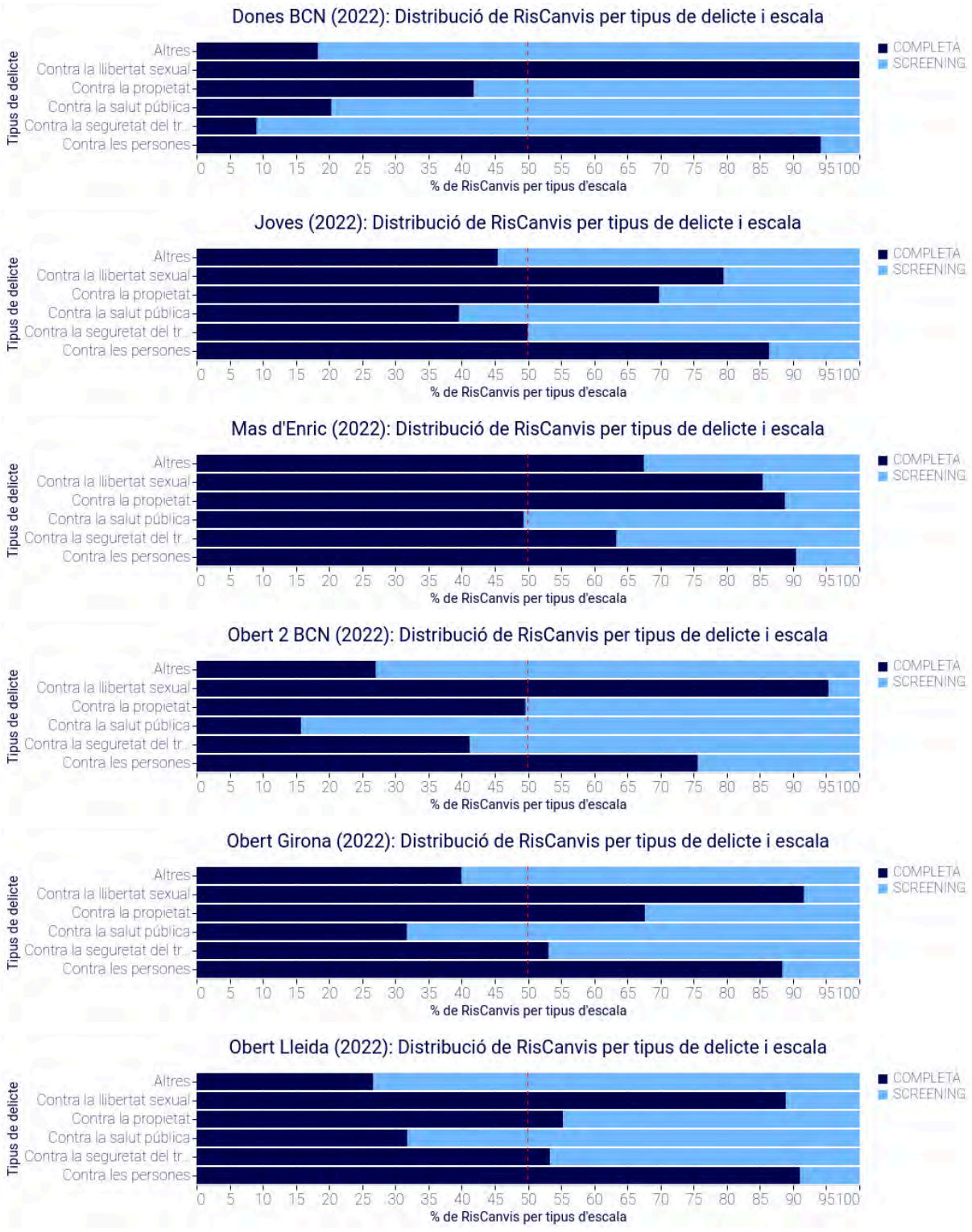
Percentatge de RisCanvis per victima (dades CEFJE)



Percentatge de RisCanvis per nombre d'ingressos (dades CEFJE)

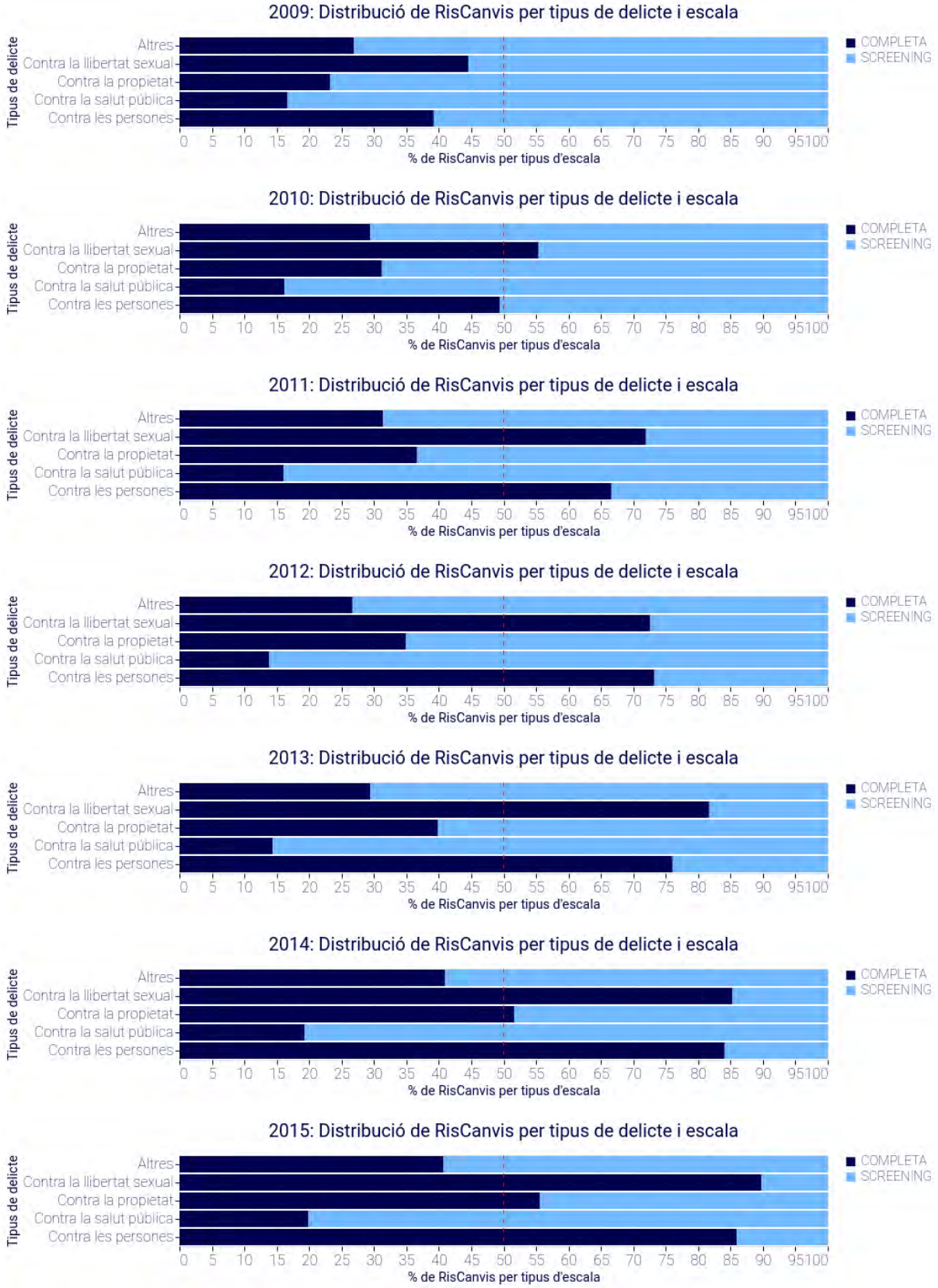


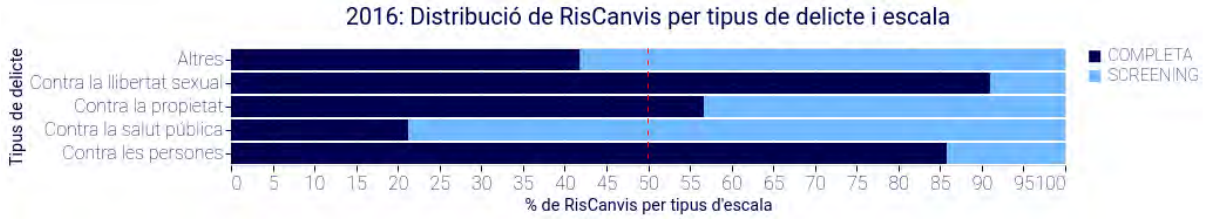
Percentatge d'avaluacions RisCanvi en 2022 per tipus de delicte i escala per centre (dades AP)

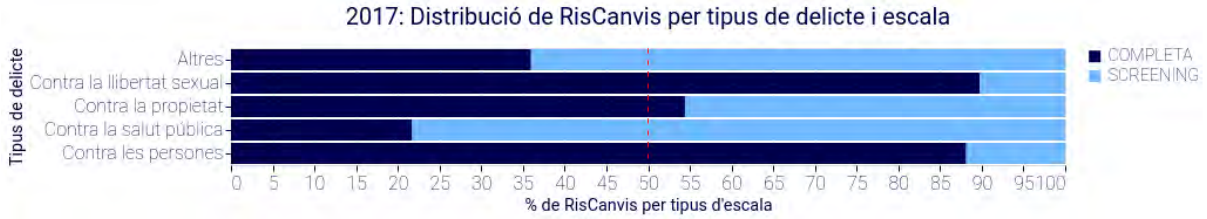


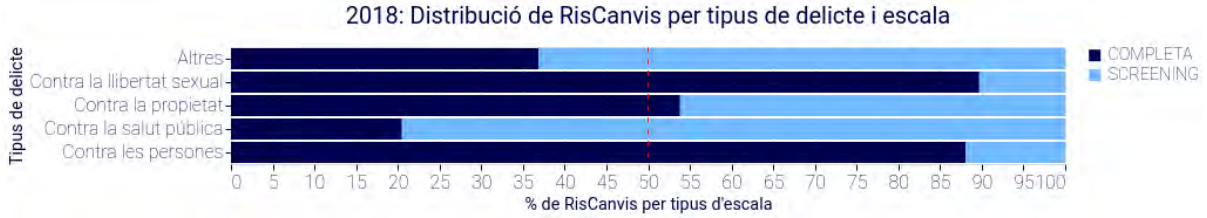

Ponent (2022): Distribució de RisCanvis per tipus de delicte i escala
Puig Basses (2022): Distribució de RisCanvis per tipus de delicte i escala
Q. Camins (2022): Distribució de RisCanvis per tipus de delicte i escala
SSAEP GIR (2022): Distribució de RisCanvis per tipus de delicte i escala
SSAEP Lleida (2022): Distribució de RisCanvis per tipus de delicte i escala
SSAEP TRG (2022): Distribució de RisCanvis per tipus de delicte i escala
Tipus de delicte
Altres
Contra la llibertat sexual
Contra la propietat
Contra la salut pública
Contra la seguretat del tr...
Contra les persones
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95100
% de RisCanvis per tipus d'escala
COMPLETA
SCREENING

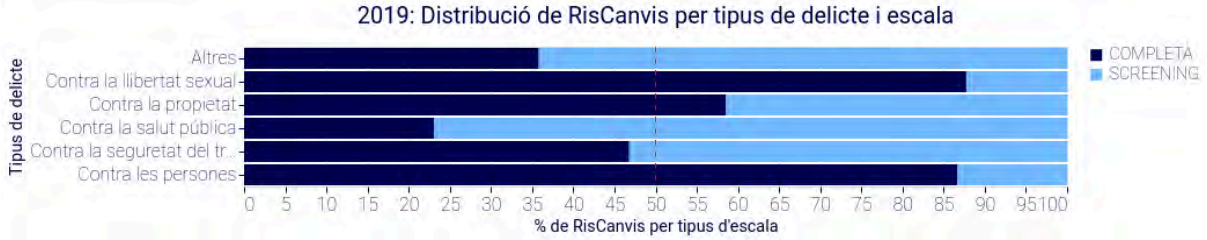
Percentatge d'avaluacions RisCanvi per tipus de delicte i escala per any.

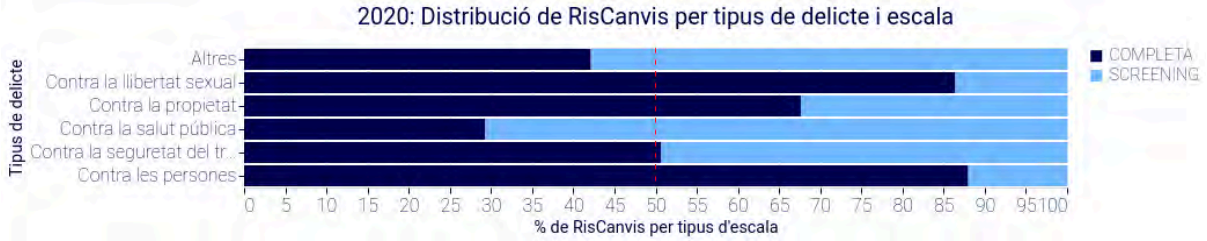


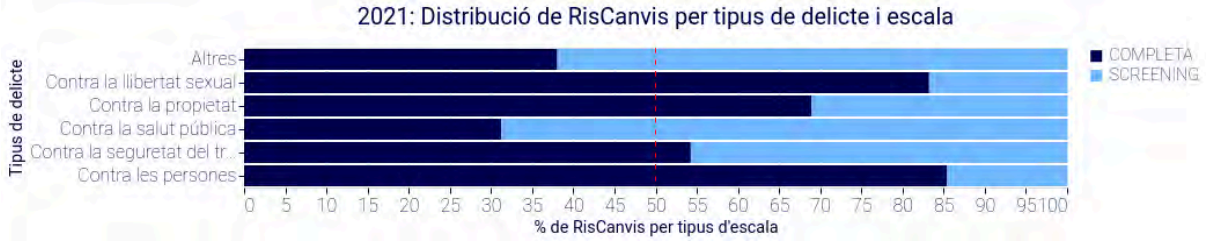

2016: Distribució de RisCanvis per tipus de delicte i escala
Tipus de delicte
Altres
Contra la llibertat sexual
Contra la propietat
Contra la salut pública
Contra les persones
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95100
% de RisCanvis per tipus d'escala
COMPLETA
SCREENING

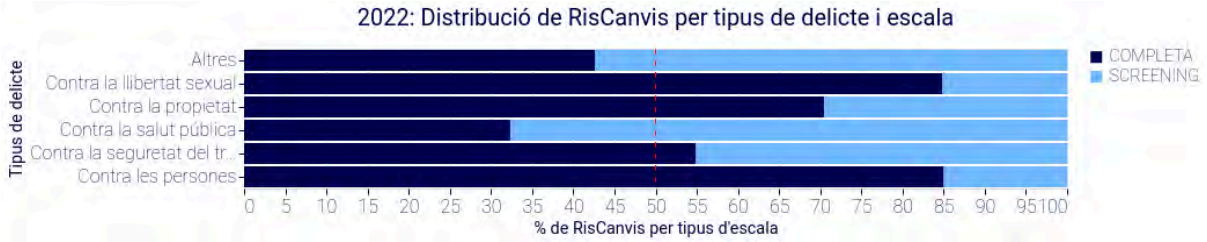

2017: Distribució de RisCanvis per tipus de delicte i escala
Tipus de delicte
Altres
Contra la llibertat sexual
Contra la propietat
Contra la salut pública
Contra les persones
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95100
% de RisCanvis per tipus d'escala
COMPLETA
SCREENING


2018: Distribució de RisCanvis per tipus de delicte i escala
Tipus de delicte
Altres
Contra la llibertat sexual
Contra la propietat
Contra la salut pública
Contra les persones
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95100
% de RisCanvis per tipus d'escala
COMPLETA
SCREENING

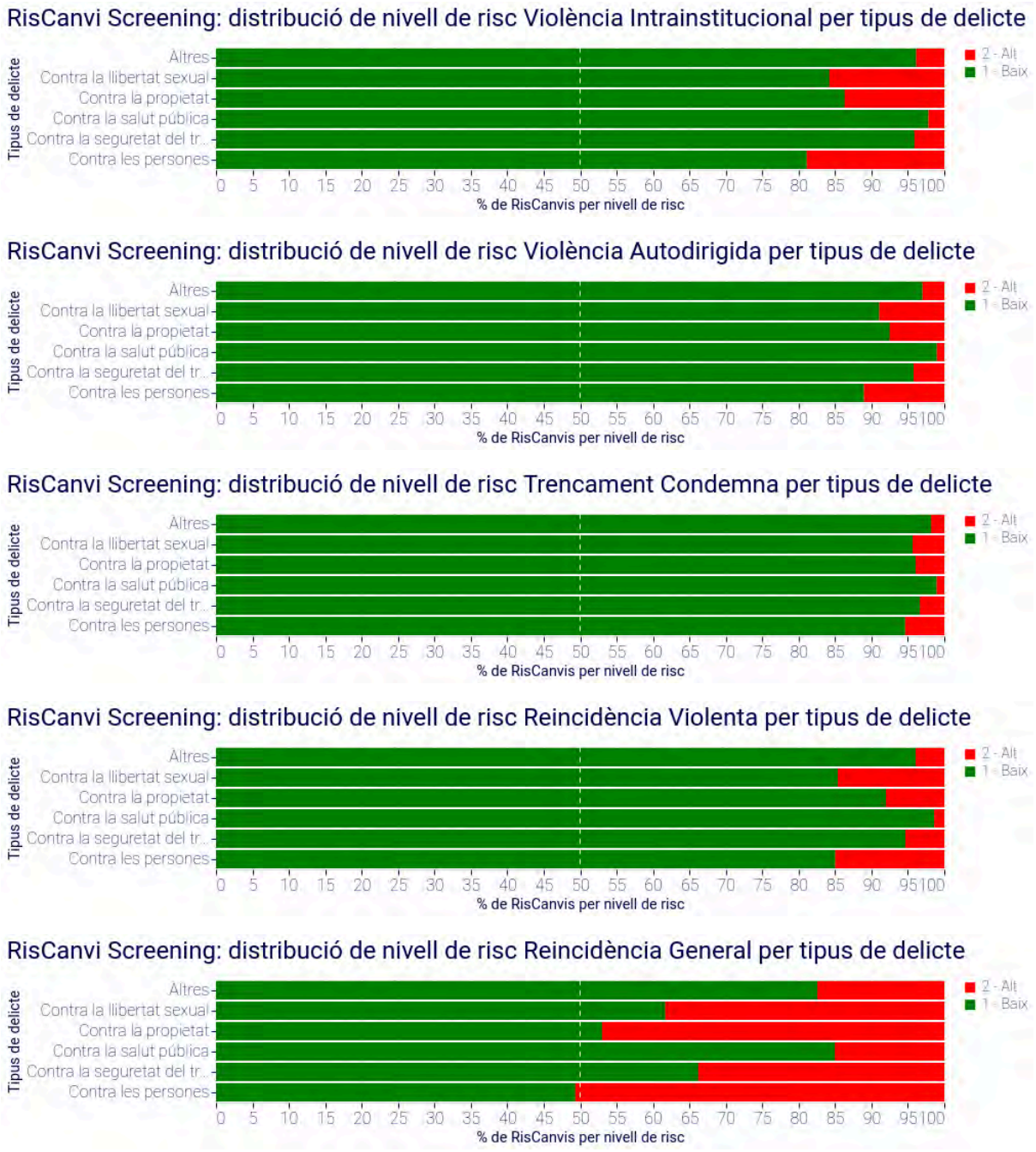

2019: Distribució de RisCanvis per tipus de delicte i escala
Tipus de delicte
Altres
Contra la llibertat sexual
Contra la propietat
Contra la salut pública
Contra la seguretat del tr...
Contra les persones
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95100
% de RisCanvis per tipus d'escala
COMPLETA
SCREENING


2020: Distribució de RisCanvis per tipus de delicte i escala
Tipus de delicte
Altres
Contra la llibertat sexual
Contra la propietat
Contra la salut pública
Contra la seguretat del tr...
Contra les persones
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95100
% de RisCanvis per tipus d'escala
COMPLETA
SCREENING


2021: Distribució de RisCanvis per tipus de delicte i escala
Tipus de delicte
Altres
Contra la llibertat sexual
Contra la propietat
Contra la salut pública
Contra la seguretat del tr...
Contra les persones
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95100
% de RisCanvis per tipus d'escala
COMPLETA
SCREENING


2022: Distribució de RisCanvis per tipus de delicte i escala
Tipus de delicte
Altres
Contra la llibertat sexual
Contra la propietat
Contra la salut pública
Contra la seguretat del tr...
Contra les persones
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95100
% de RisCanvis per tipus d'escala
COMPLETA
SCREENING

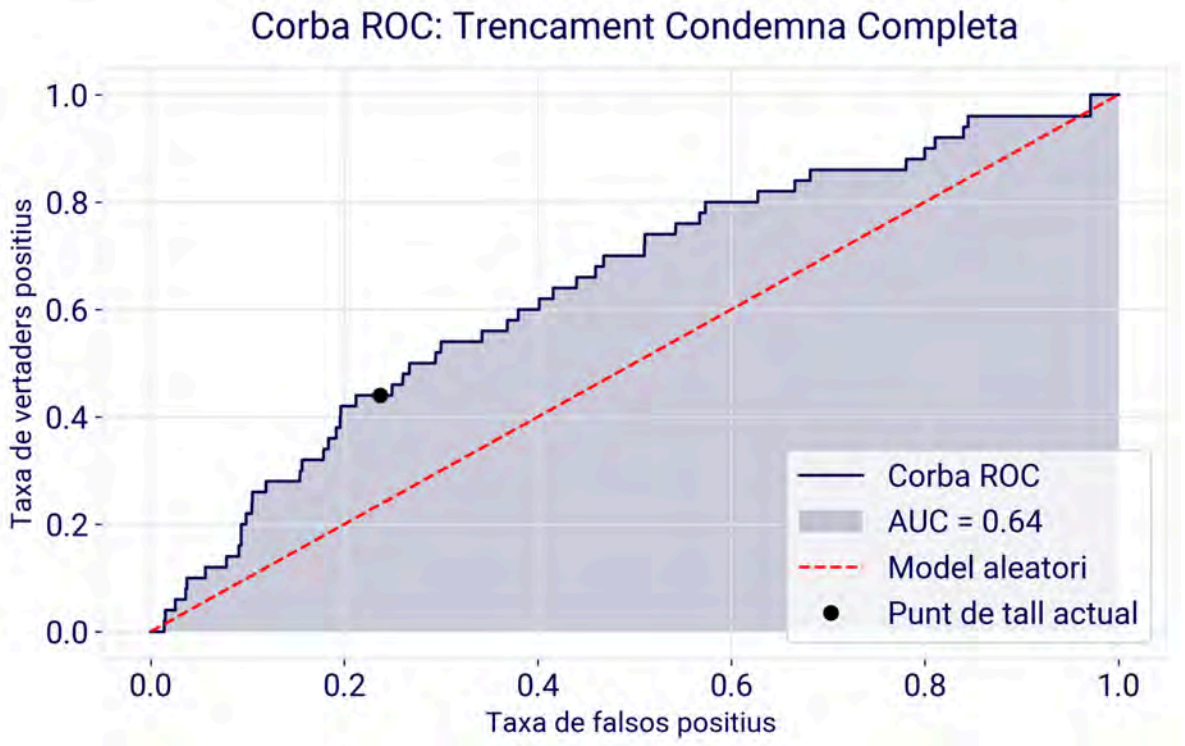
Nivells de risc per tipologia del delicte base per l'escala Screening (dades AP)

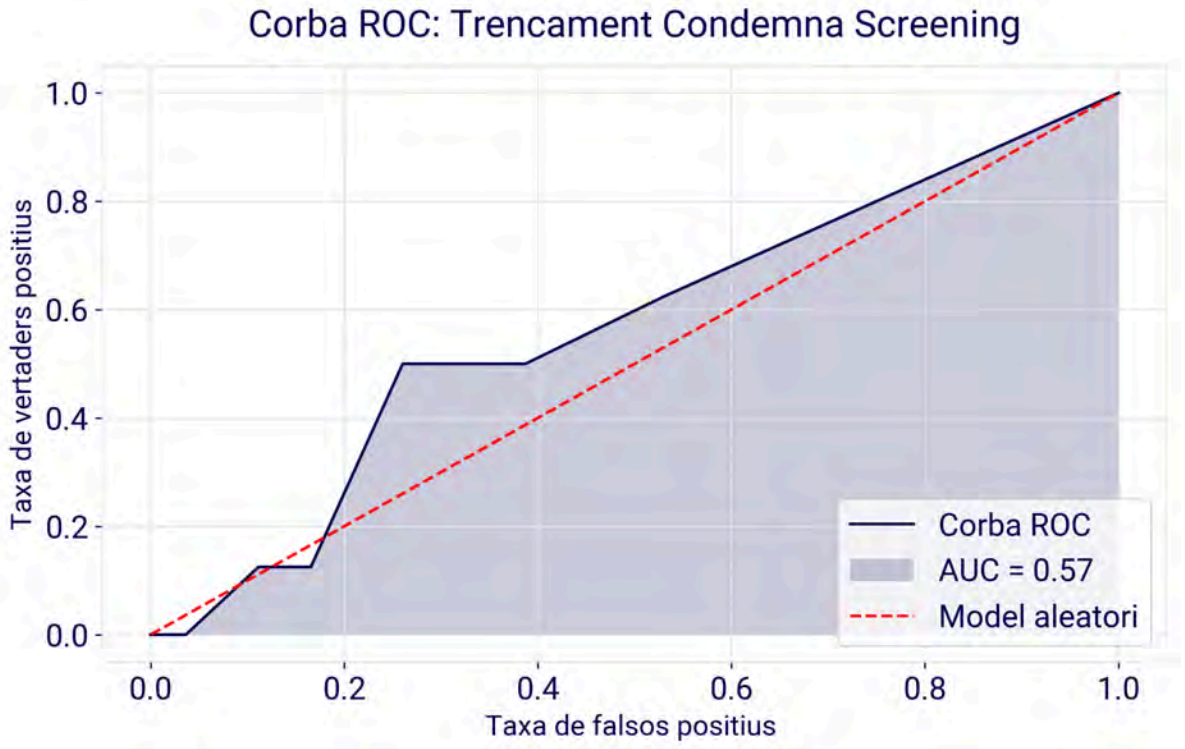


Algoritme actual

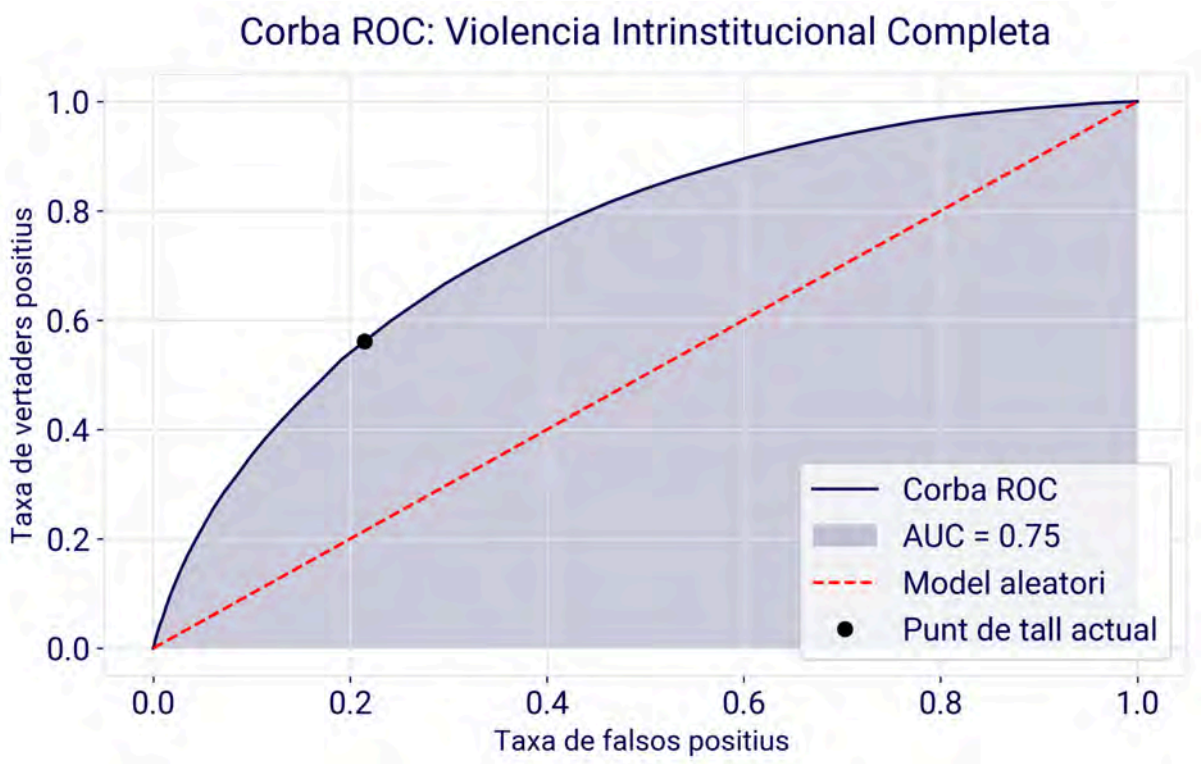
Anàlisi AUC-ROC

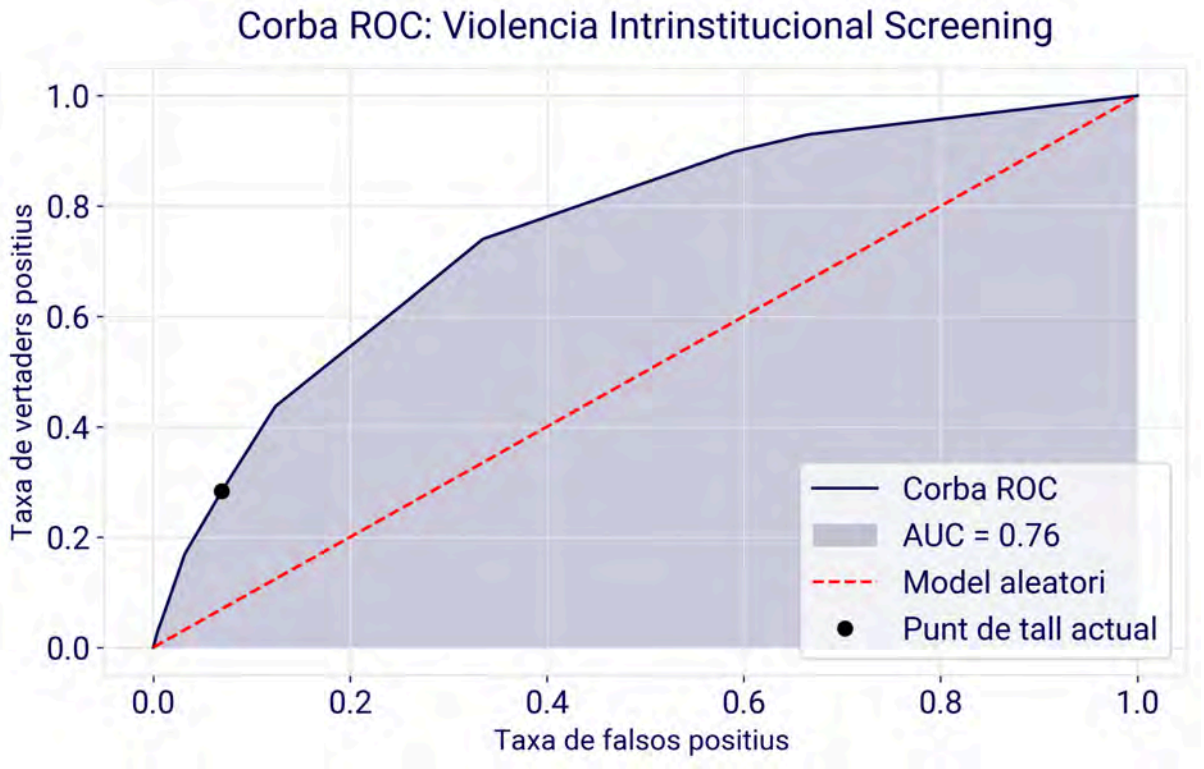
Corbes ROC per al risc de Trencament de Condemna en escala Completa i Screening



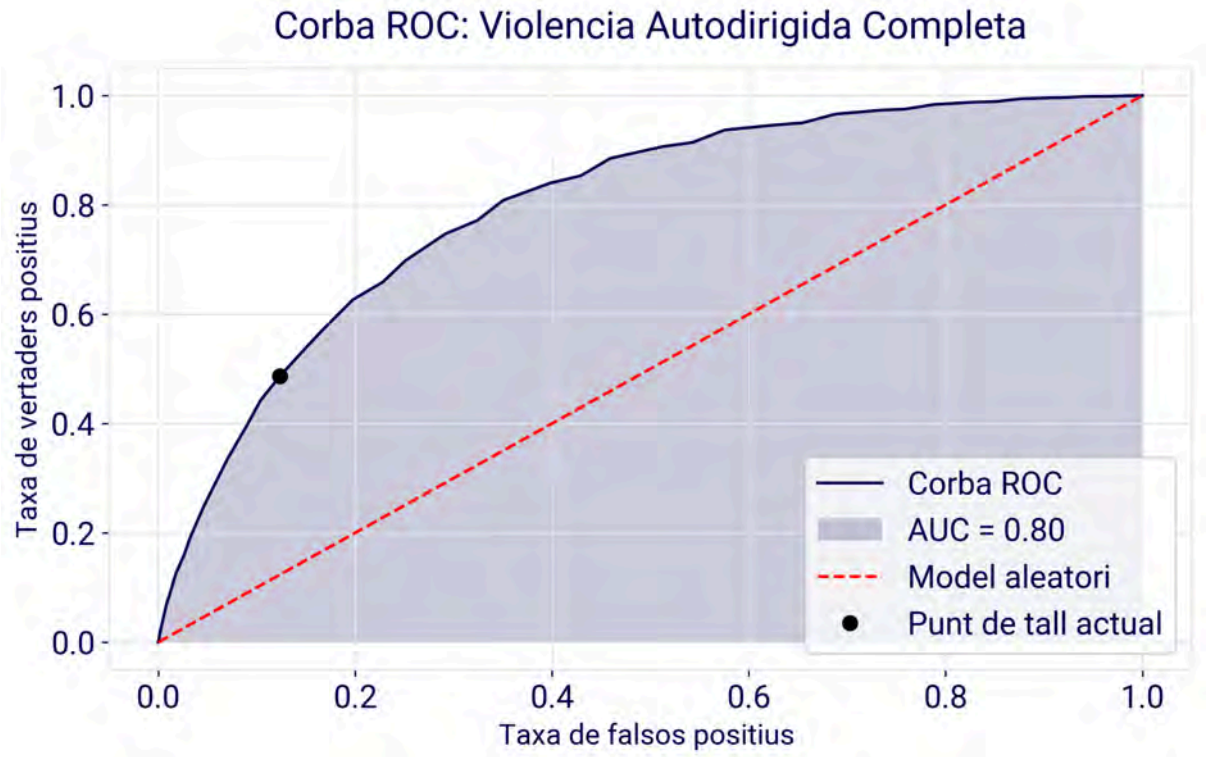


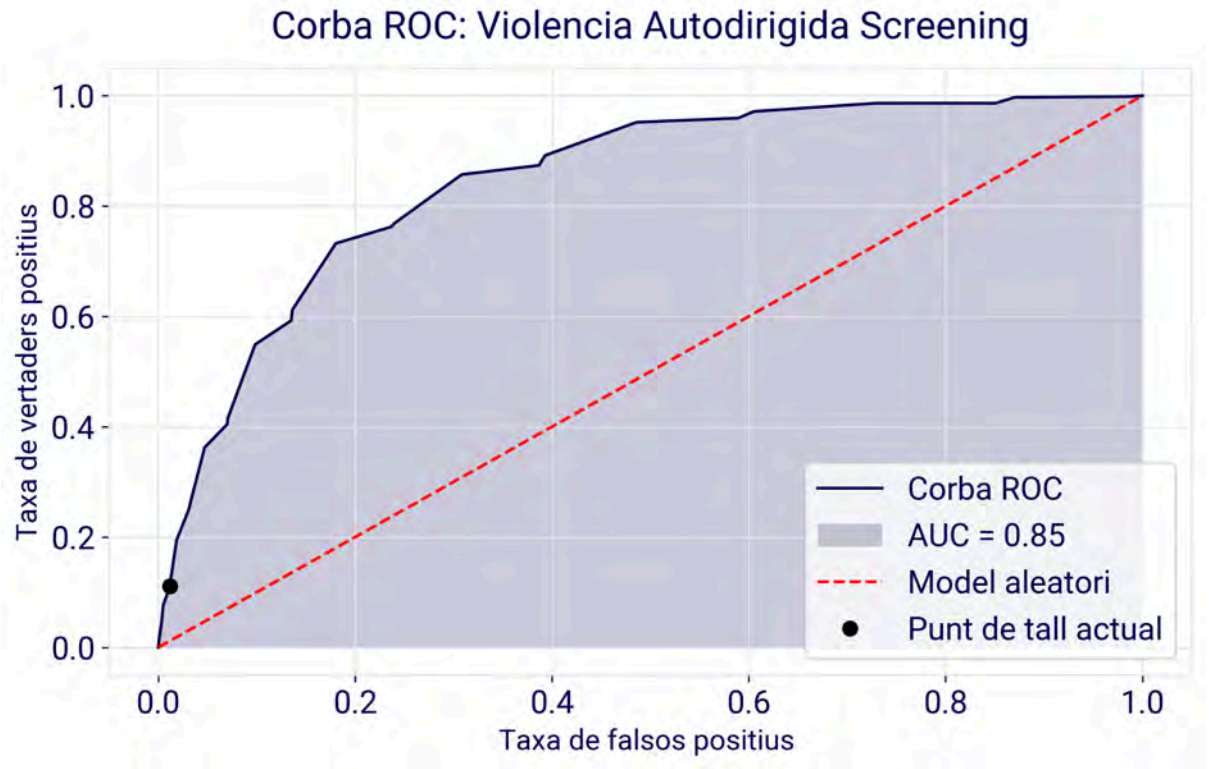
Corbes ROC per al risc de Violència Intrainstitucional en escala Completa i Screening



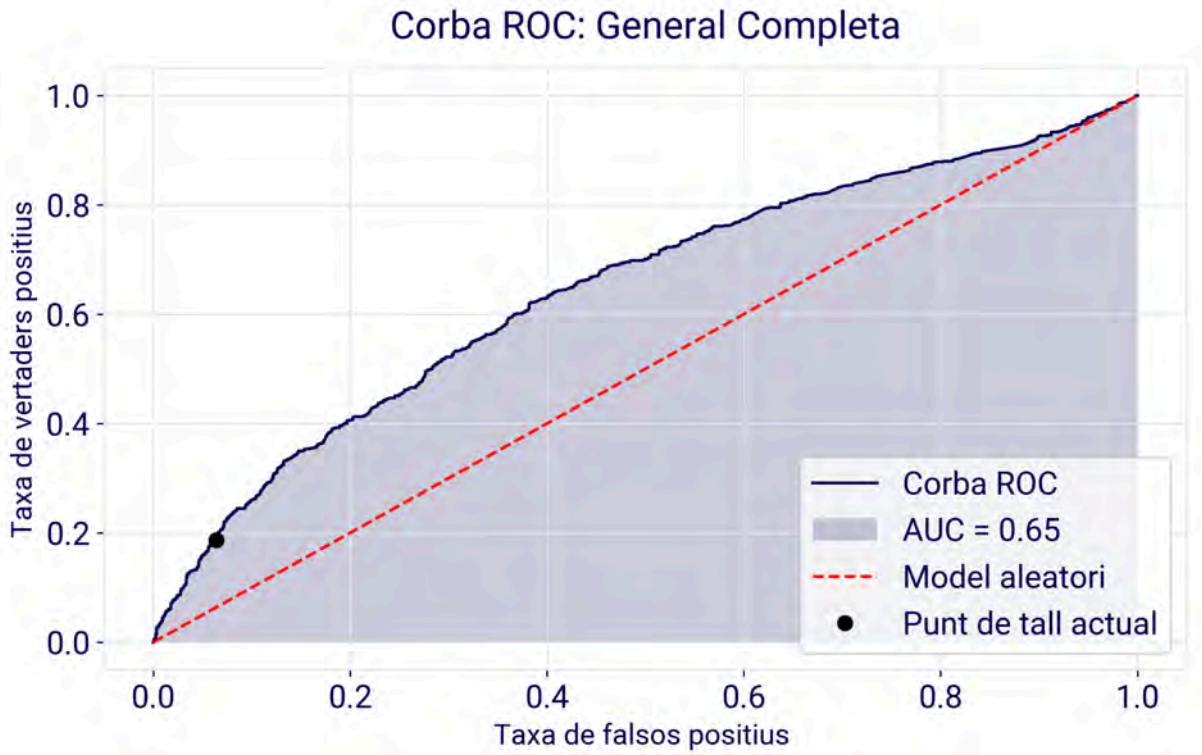


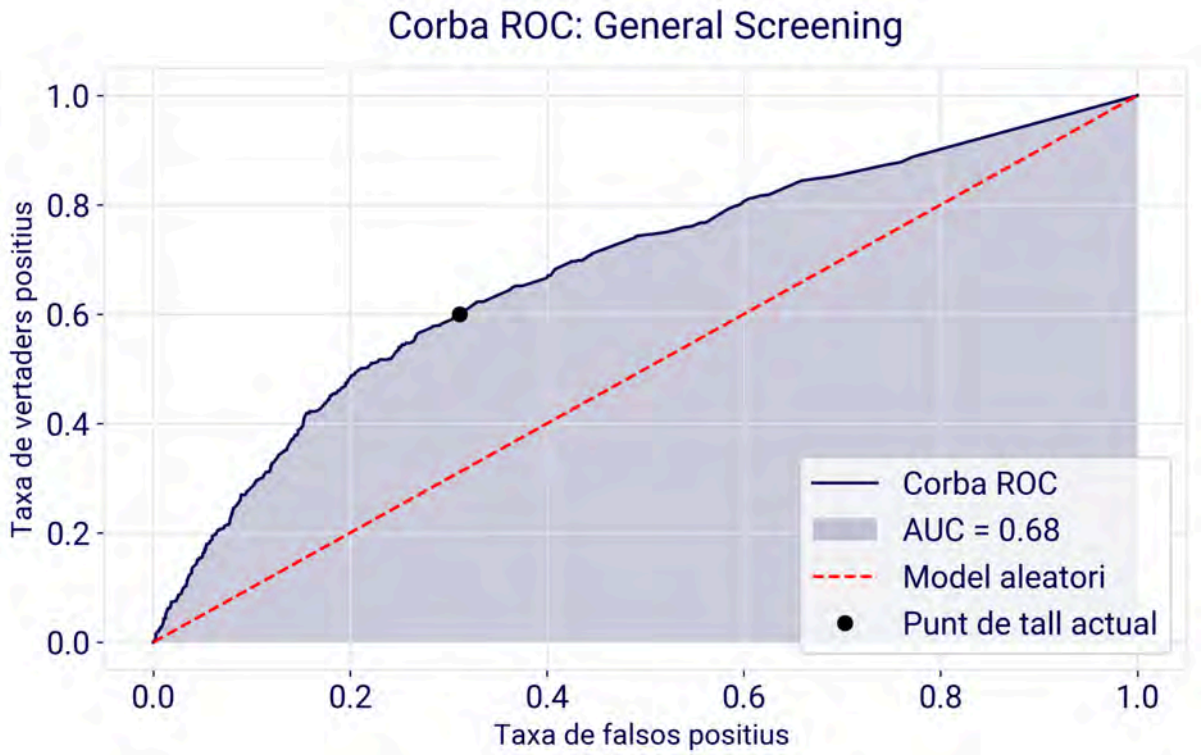
Corba ROC per al risc de Violència Autodirigida en escala Completa i Screening



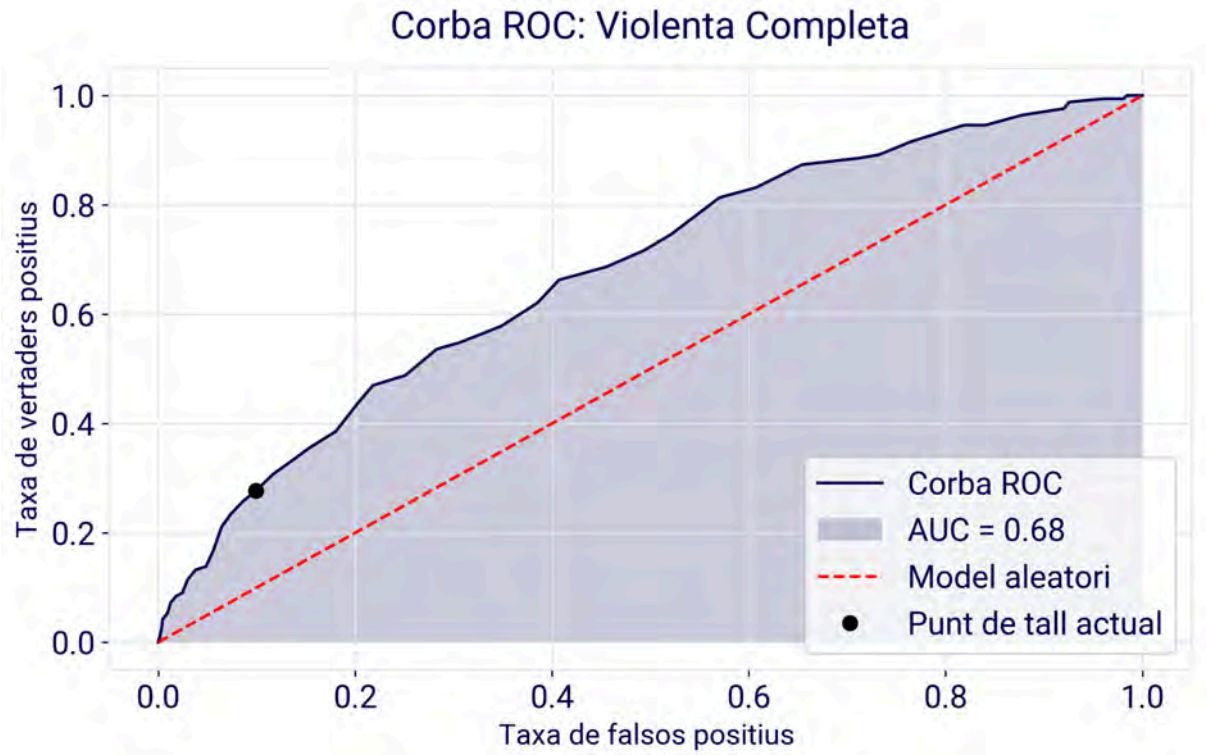


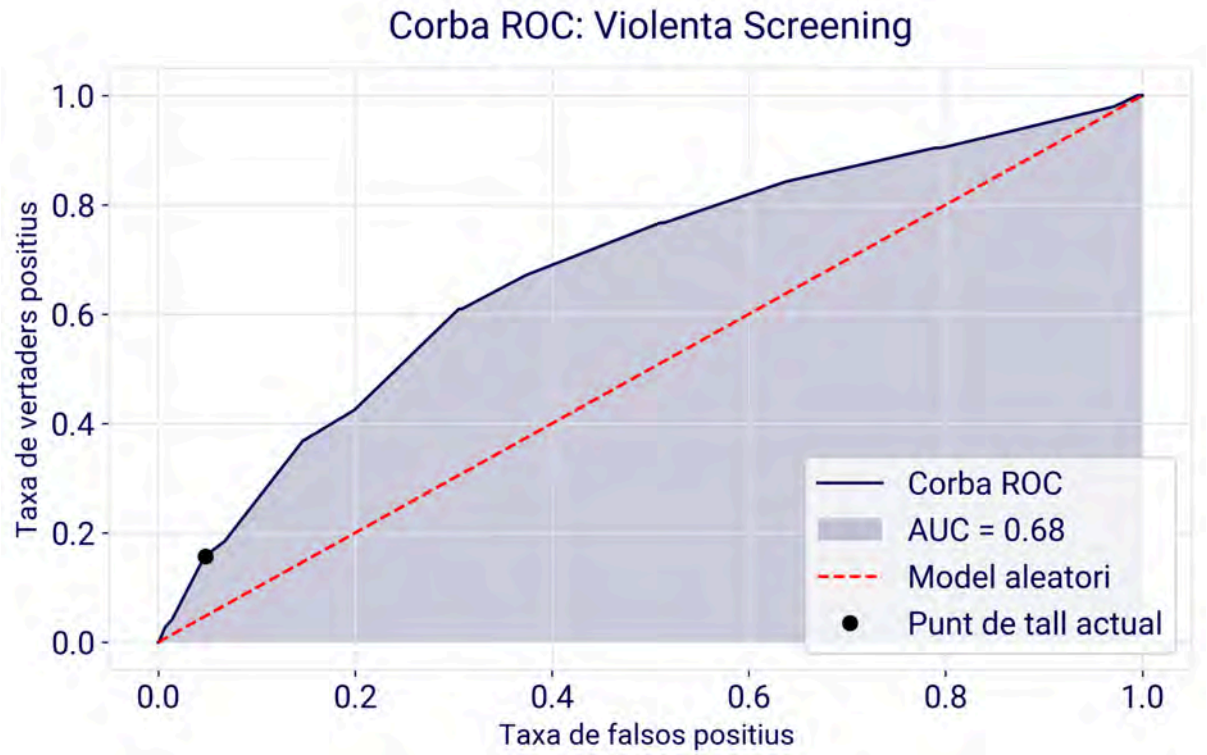
Corba ROC per al risc de Reincidència General en escala Completa i Screening





Corba ROC per al risc de Reincidència Violenta en escala Completa i Screening





Anàlisi d'exactitud

Anàlisi del RisCanvi COMPLETA considerant riscos Alt i Mitjà com a Positius

RISC	ESCALA	EXACTITUD
Trencament condemna	COMPLETA	39.21%
Violència autodirigida	COMPLETA	60.81%
Reincidència general	COMPLETA	67.75%
Reincidència violenta	COMPLETA	68.27%
Violència intrainstitucional	COMPLETA	59.36%

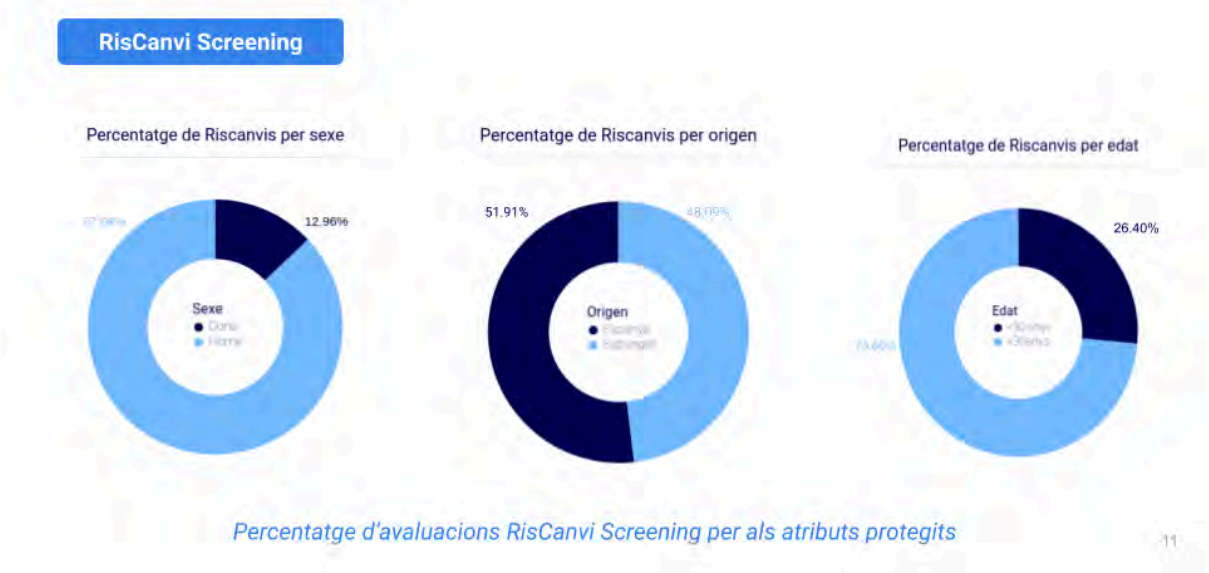
Exactitud dels models RisCanvi (accuracy), prenent al risc mitjà com a positiu

RISC	ESCALA	FPR	FNR	TPR	TNR
Trencament condemna	COMPLETA	60.84%	20.00%	80.00%	39.16%
Violència autodirigida	COMPLETA	39.82%	16.02%	83.98%	60.18%
Reincidència general	COMPLETA	17.31%	63.25%	36.75%	82.69%
Reincidència violenta	COMPLETA	30.69%	44.79%	55.21%	69.31%
Violència intrainstitucional	COMPLETA	46.31%	18.46%	81.54%	53.69%

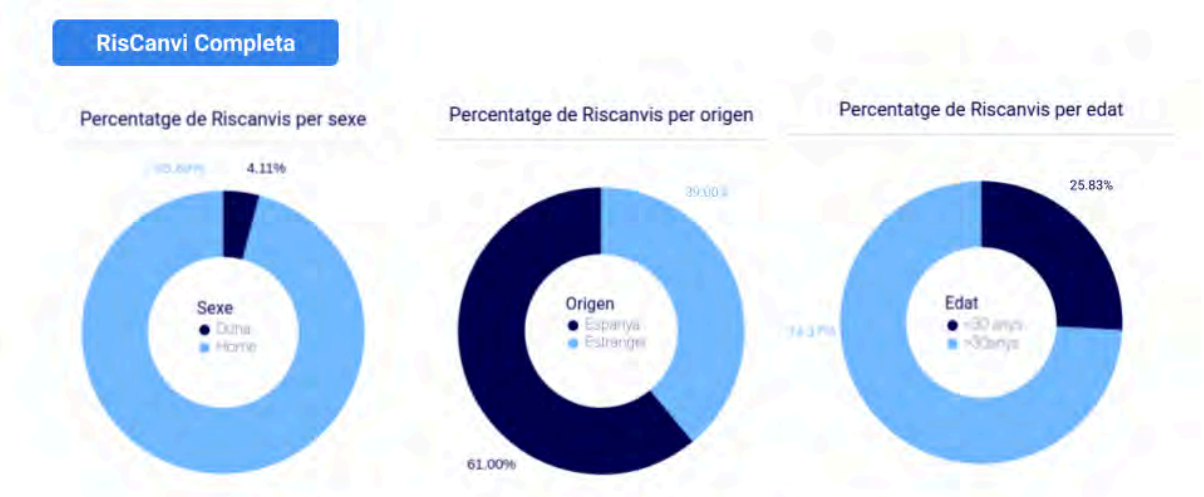
Taxes de falsos positius i negatius i de vertaders positius i negatius, prenent al risc mitjà com a positiu

Biaixos potencialment discriminatius

Percentatge d'avaluacions RisCanvi per cada atribut protegit i escala



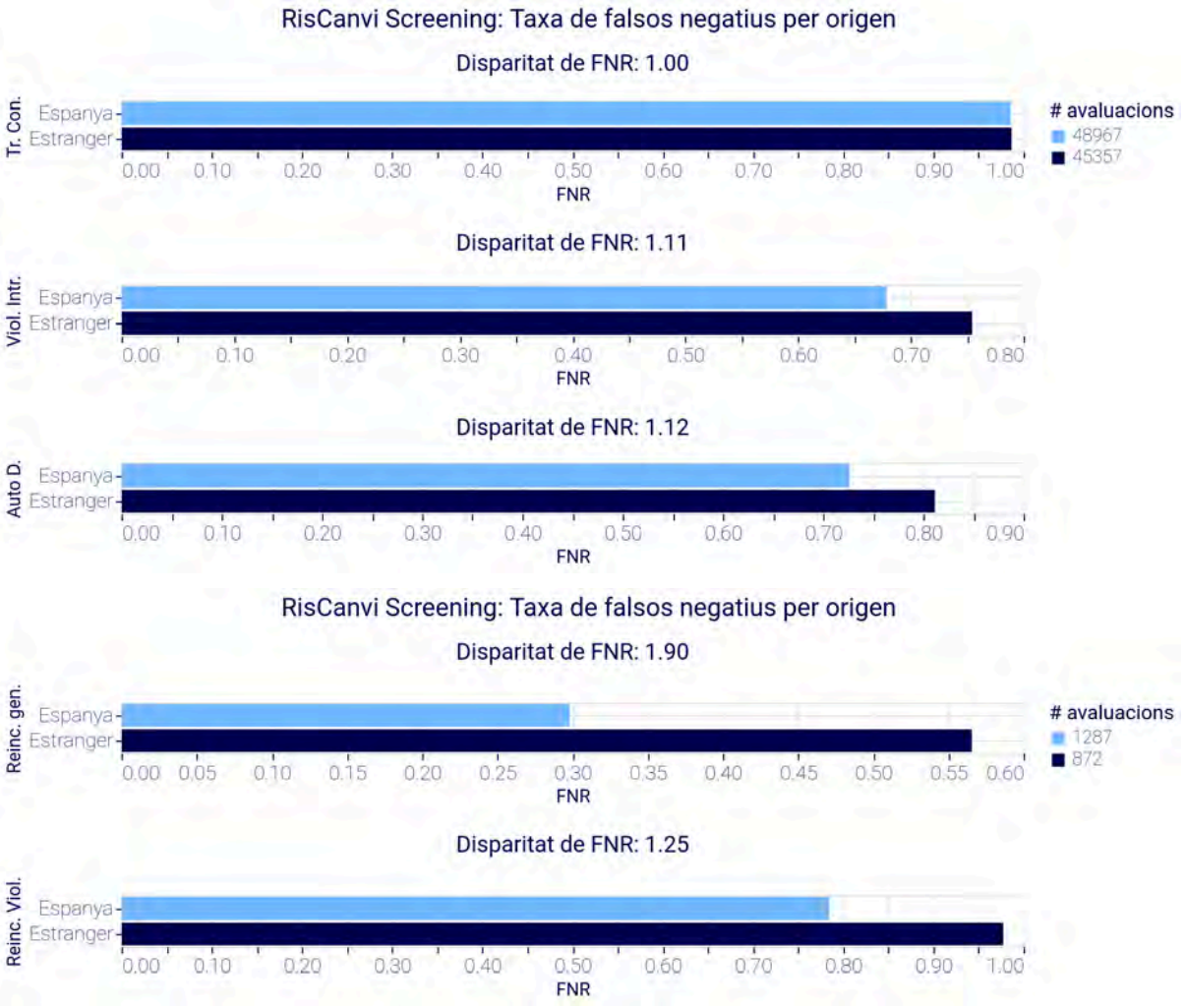
Percentatge d'avaluacions RisCanvi Screening per als atributs protegits



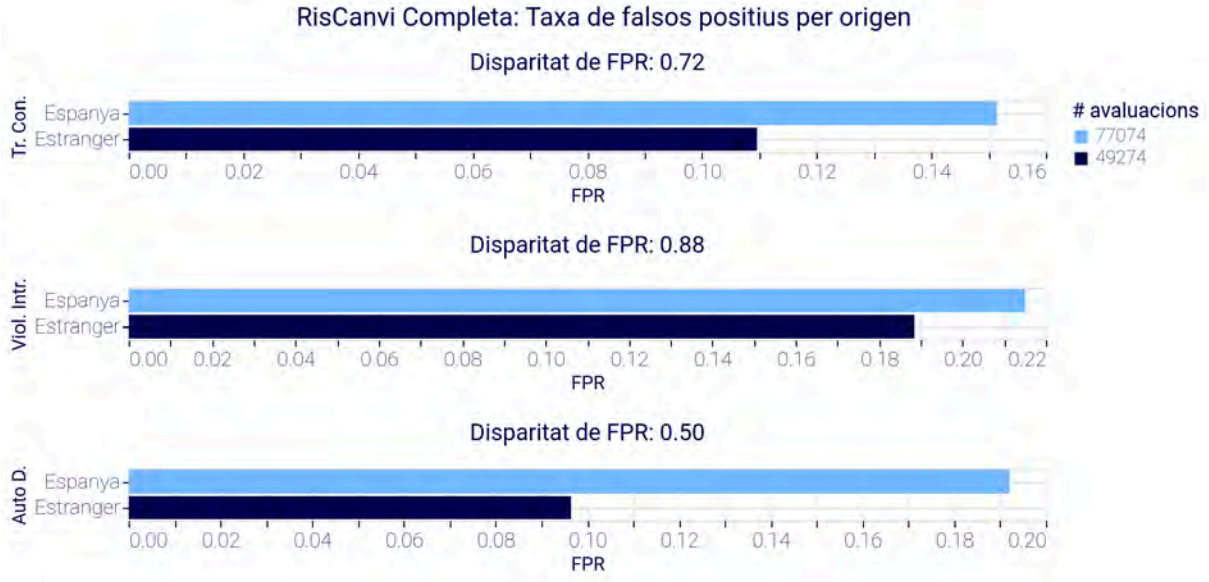
Percentatge d'avaluacions RisCanvi Completa per als atributs protegits

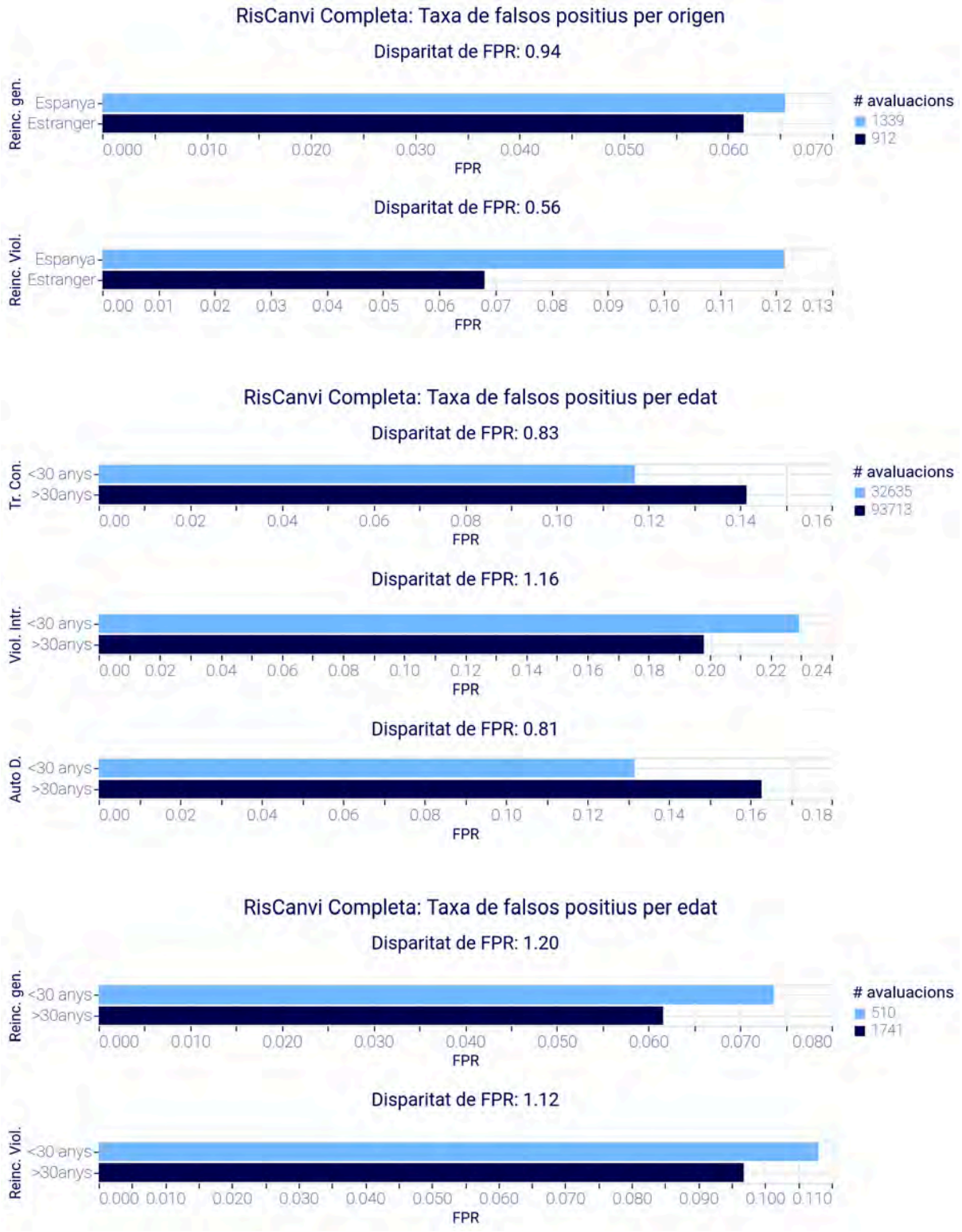
12

Ratio de falsos negatius per origen i per edat per cada risc (RisCanvi Screening)

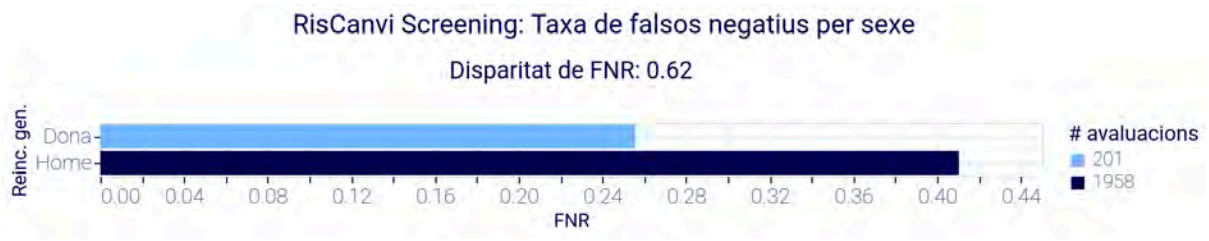


Ratio de falsos positius per origen i per edat per cada risc (RisCanvi Completa)



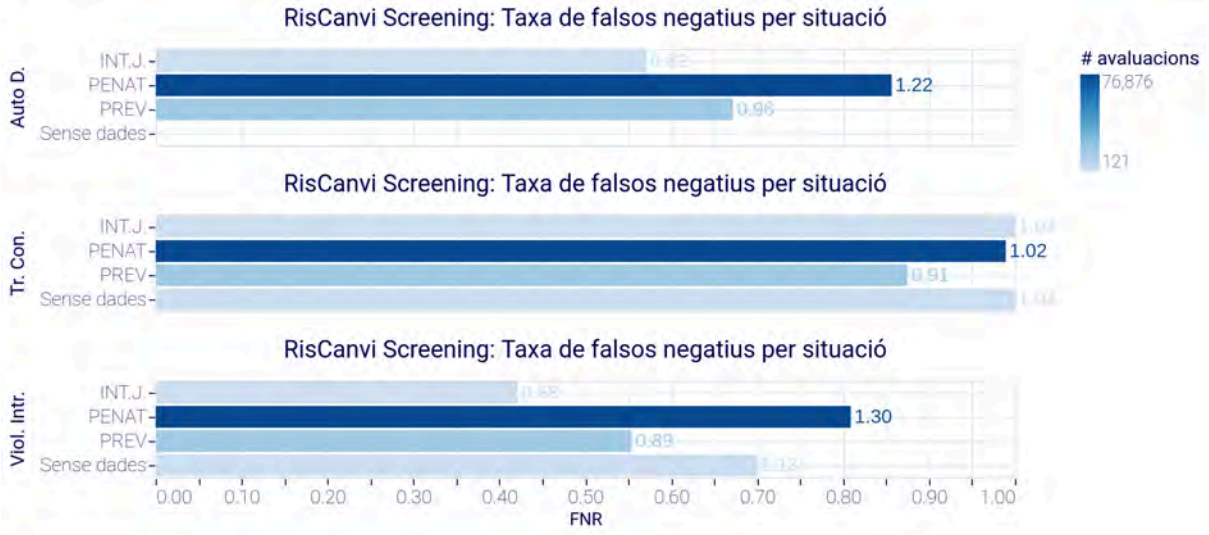

RisCanvi Completa: Taxa de falsos positius per origen
Disparitat de FPR: 0.94
Reinc. gen.
Espanya
Estranger
0.000 0.010 0.020 0.030 0.040 0.050 0.060 0.070
FPR
avaluacions
1339
912
Disparitat de FPR: 0.56
Reinc. Viol.
Espanya
Estranger
0.00 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.10 0.11 0.12 0.13
FPR
RisCanvi Completa: Taxa de falsos positius per edat
Disparitat de FPR: 0.83
Tr. Con.
<30 anys
>30anys
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16
FPR
avaluacions
32635
93713
Disparitat de FPR: 1.16
Viol. Intr.
<30 anys
>30anys
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16 0.18 0.20 0.22 0.24
FPR
Disparitat de FPR: 0.81
Auto D.
<30 anys
>30anys
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16 0.18
FPR
RisCanvi Completa: Taxa de falsos positius per edat
Disparitat de FPR: 1.20
Reinc. gen.
<30 anys
>30anys
0.000 0.010 0.020 0.030 0.040 0.050 0.060 0.070 0.080
FPR
avaluacions
510
1741
Disparitat de FPR: 1.12
Reinc. Viol.
<30 anys
>30anys
0.000 0.010 0.020 0.030 0.040 0.050 0.060 0.070 0.080 0.090 0.100 0.110
FPR

Ratio de falsos negatius per sexede reincidència general i violenta (RisCanvi Screening)

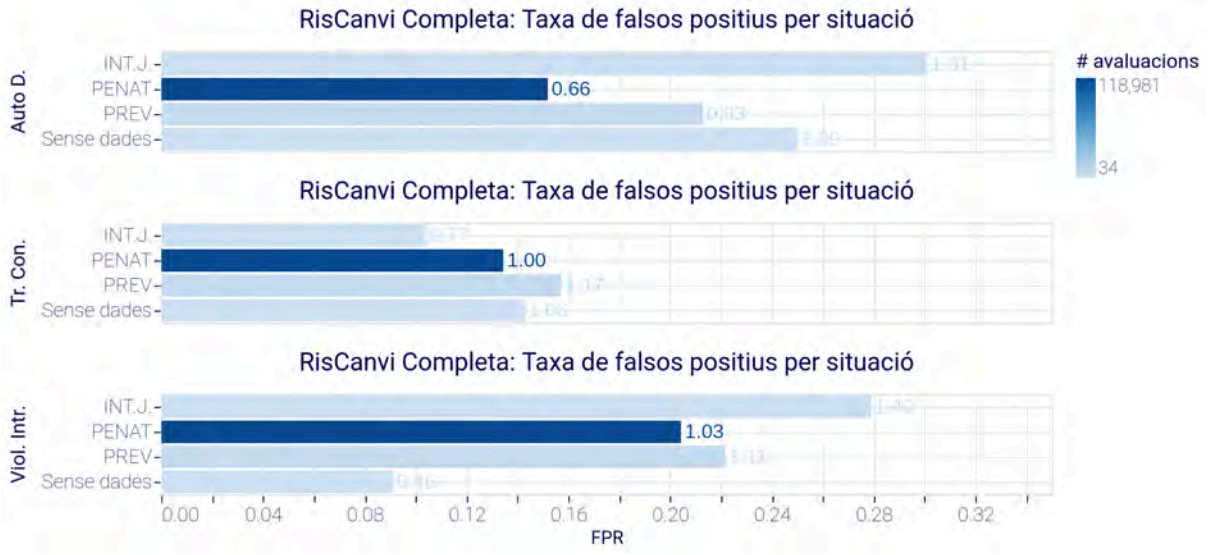


Altres biaixos

Ratio de falsos negatius per situació per cada risc (RisCanvi Screening)



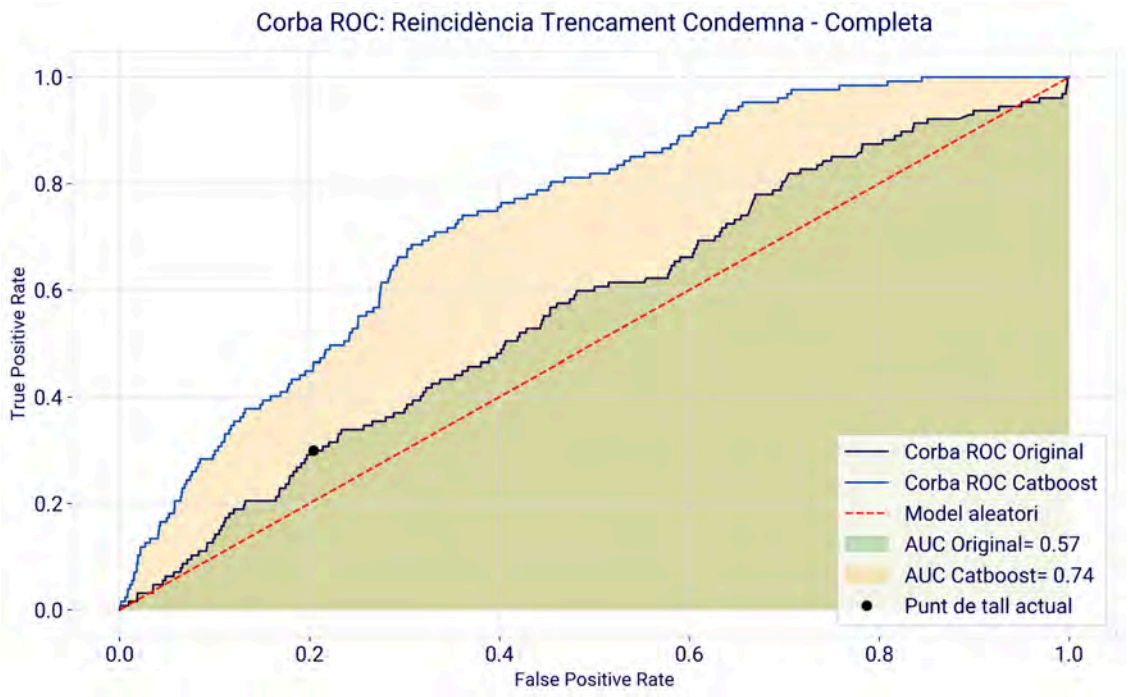
Ratio de falsos positius per situació per cada risc (RisCanvi Completa)

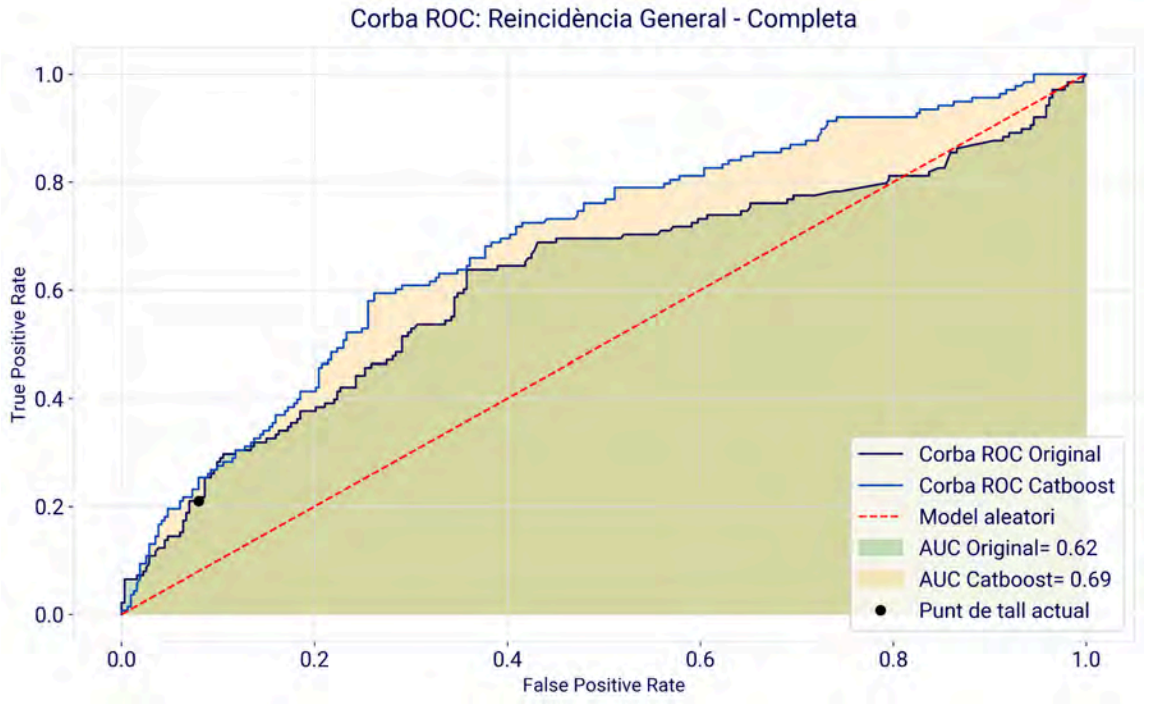


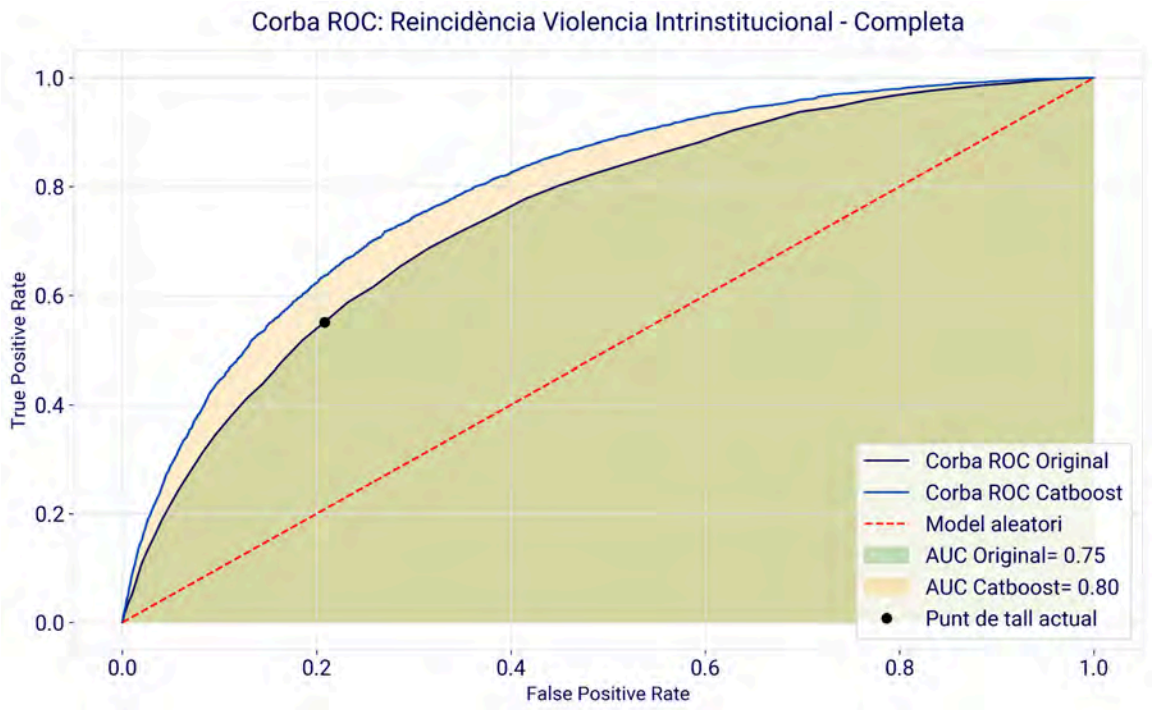
Nous algoritmes

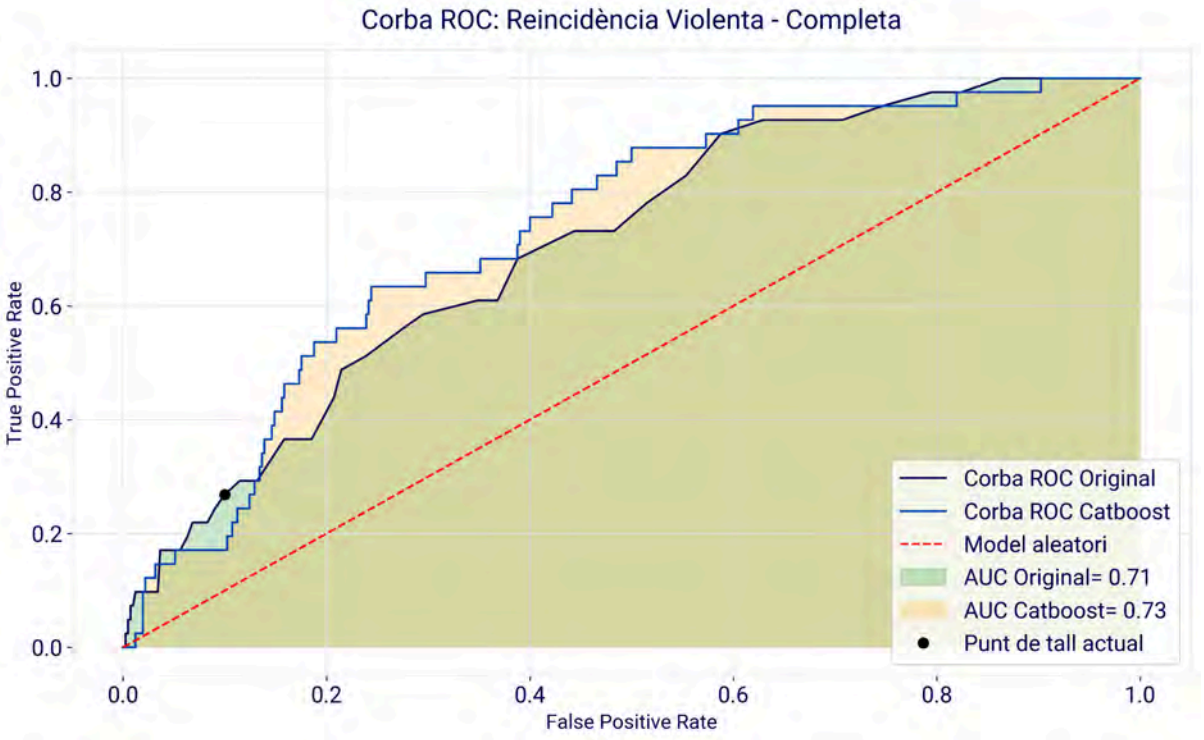
Resultats de l'algoritme catboost

Corbes ROC comparatives entre els algoritmes actuals i l'algoritme Catboost sobre el mateix conjunt de test per a Violència Intrainstitucional, Trencament de Condemna, Reincidència General i Reincidència Violenta

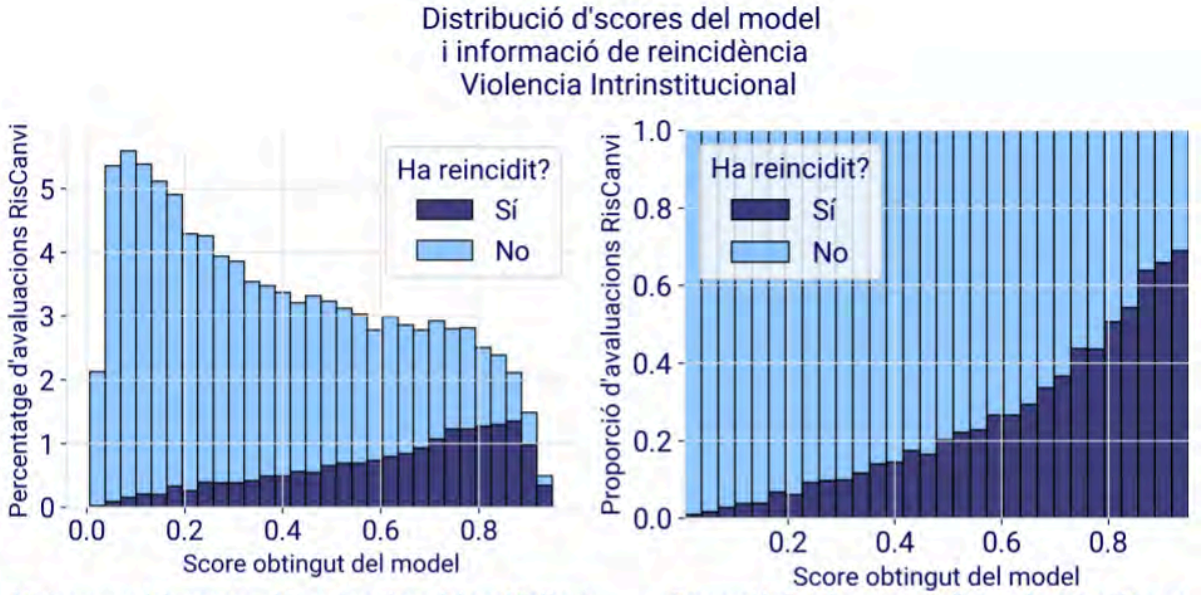



Corba ROC: Reincidència General - Completa
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Corba ROC Original
Corba ROC Catboost
Model aleatori
AUC Original= 0.62
AUC Catboost= 0.69
Punt de tall actual


Corba ROC: Reincidència Violencia Intrinstitucional - Completa
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Corba ROC Original
Corba ROC Catboost
Model aleatori
AUC Original= 0.75
AUC Catboost= 0.80
Punt de tall actual

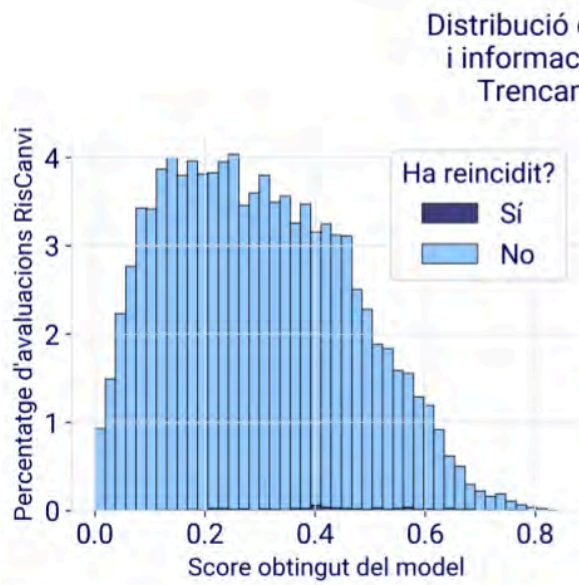


Distribució dels scores del model Catboost per a Violència Intrainstitucional, Trencament de Condemna, Reincidència General i Reincidència Violenta

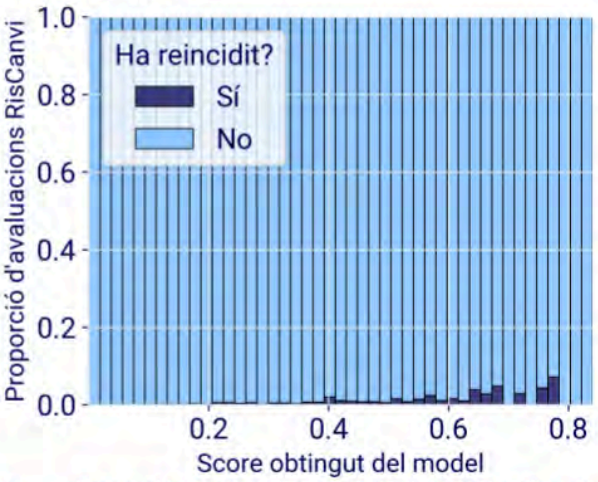


Histograma dels scores obtinguts pel model per Violència Intrainstitucional.
Hi afegim la informació de si ha reincidit o no.

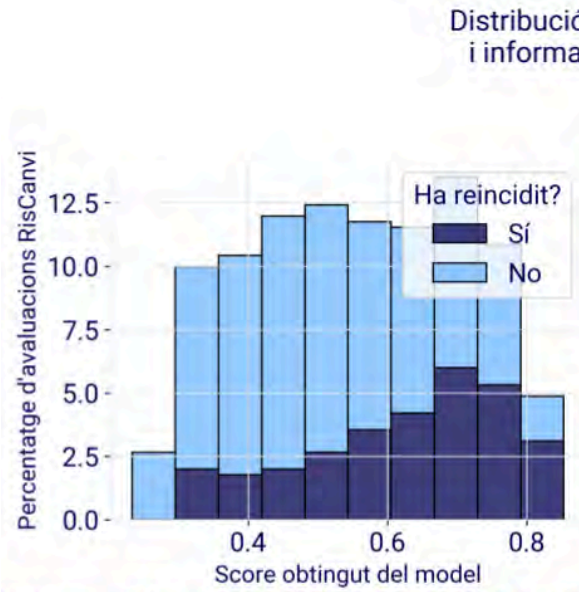
Proporció d'avaluacions on l'intern ha reincidit o no en funció de l'score assignat pel model
(Info complementària a l'histograma)



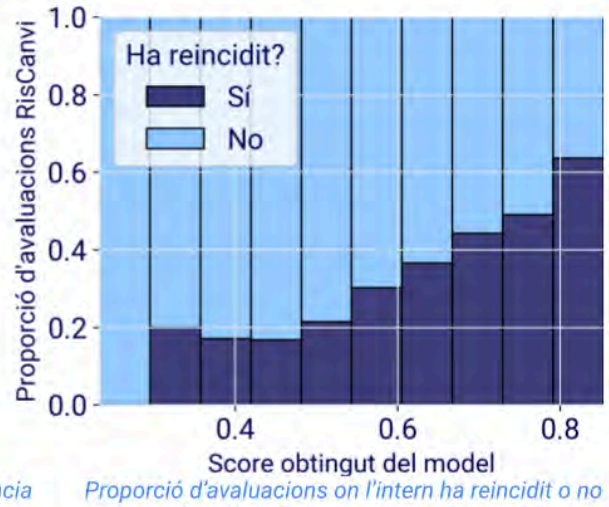
Histograma dels scores obtinguts pel model per Trencament de Condemna
Hi afegim la informació de si ha reincidit o no.



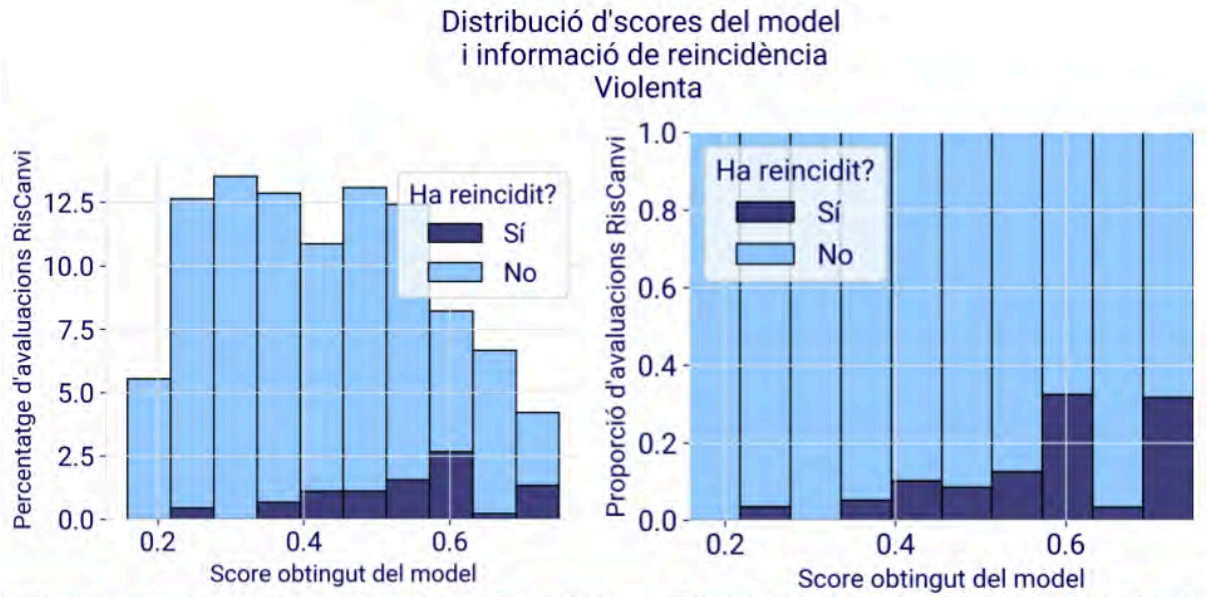
Proporció d'avaluacions on l'intern ha reincidit o no en funció de l'score assignat pel model (Info complementària a l'histograma)



Histograma dels scores obtinguts pel model per Reincidència General.
Hi afegim la informació de si ha reincidit o no.



Proporció d'avaluacions on l'intern ha reincidit o no en funció de l'score assignat pel model (Info complementària a l'histograma)



Histograma dels scores obtinguts pel model per Reincidència Violenta.
Hi afegim la informació de si ha reincidit o no.

Proporció d'avaluacions on l'intern ha reincidit o no en funció de l'score assignat pel model (Info complementària a l'histograma)

Explicabilitat

Importància dels factors de risc per determinar la probabilitat de reincidència amb Catboost per Trencament de Condemna, Violència Intrainstitucional, Reincidència General i Reincidència Violenta

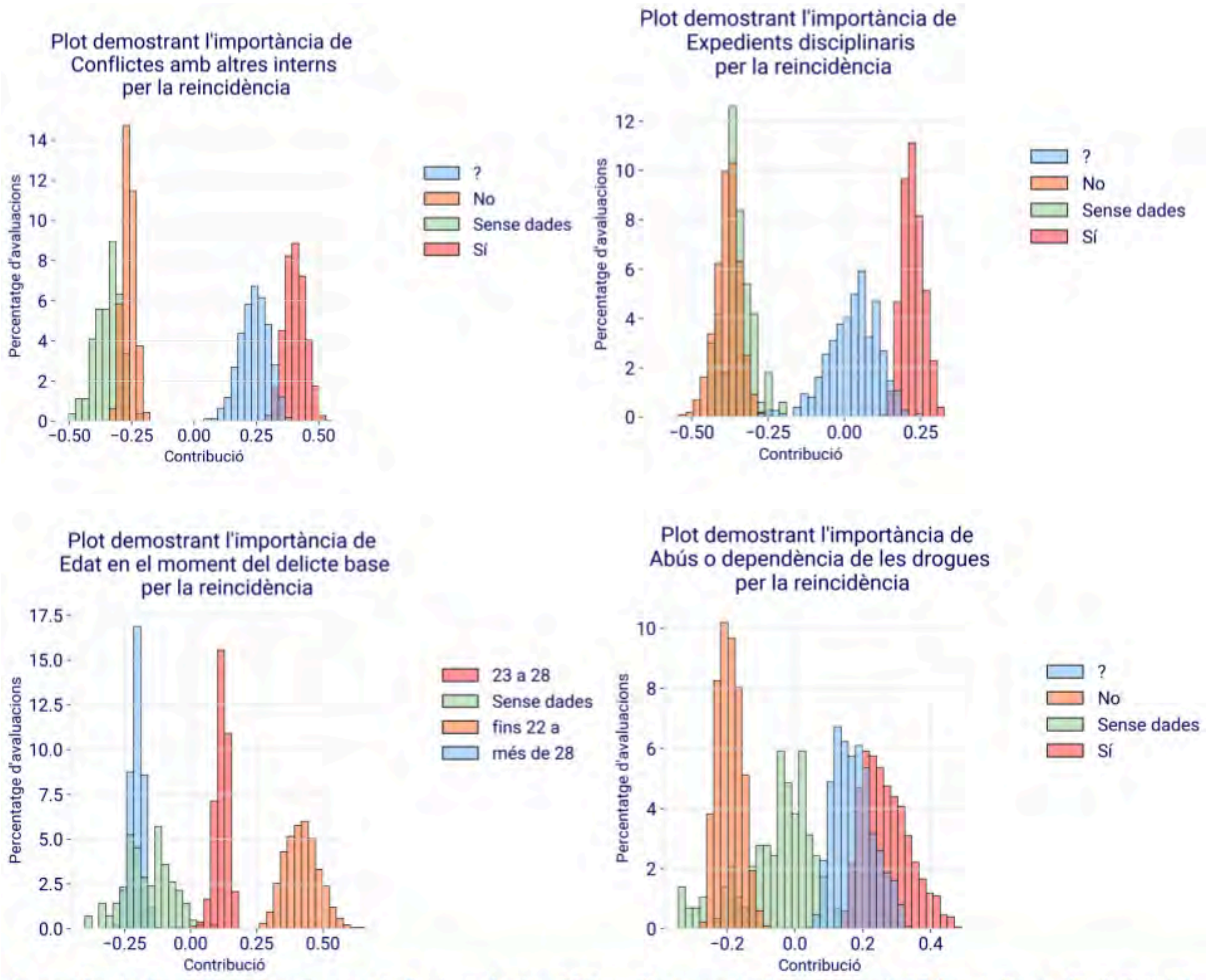
Factors més importants per a l'algoritme de risc de Violència Intrainstitucional i quins coincideixen amb els utilitzats pel model actual (proposats per l'equip d'experts)

Factors més importants per a l'algoritme de risc de Trencament de condemna i quins coincideixen amb els utilitzats pel model actual (proposats per l'equip d'experts)

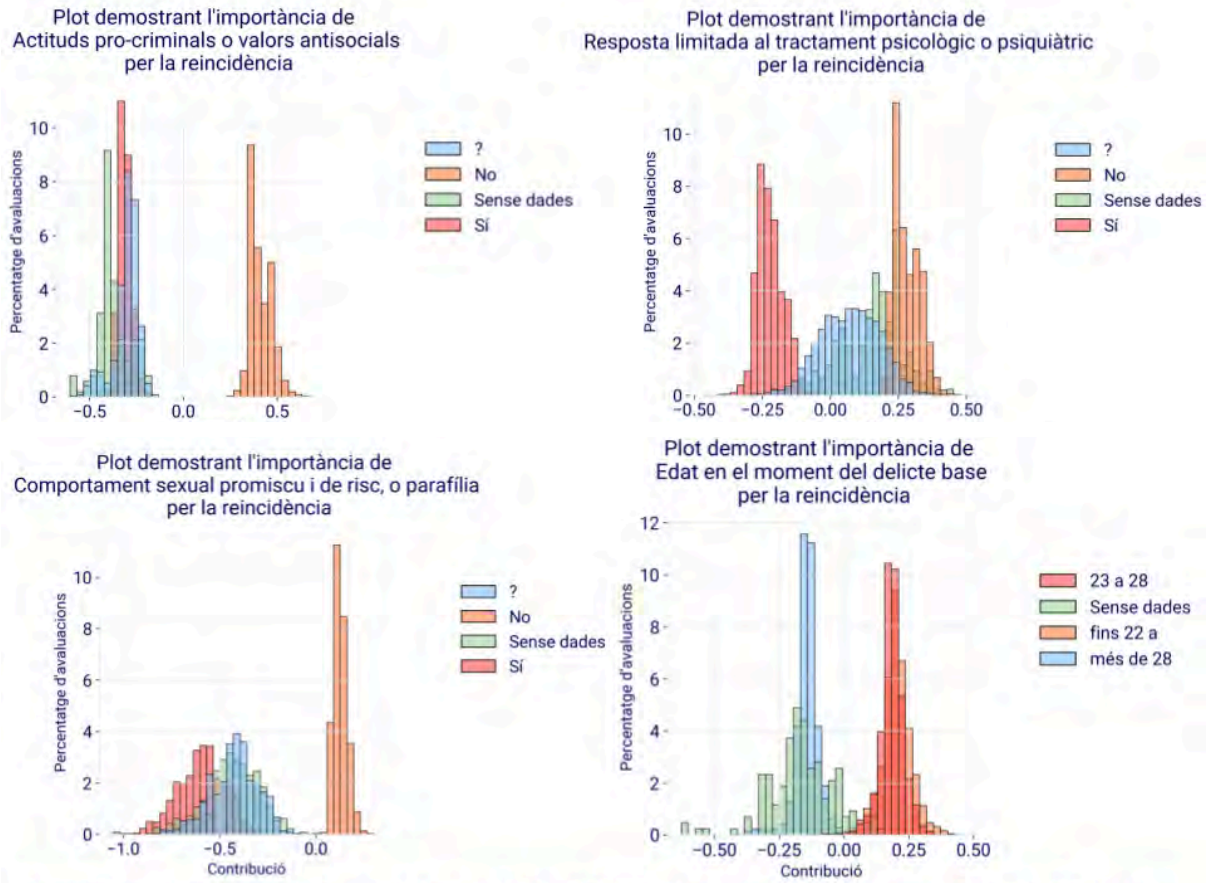
Factors més importants per a l'algoritme de risc de Reincidència General i quins coincideixen amb els utilitzats pel model actual (proposats per l'equip d'experts)

Factors més importants per a l'algoritme de risc de Reincidència Violenta i quins coincideixen amb els utilitzats pel model actual (proposats per l'equip d'experts)

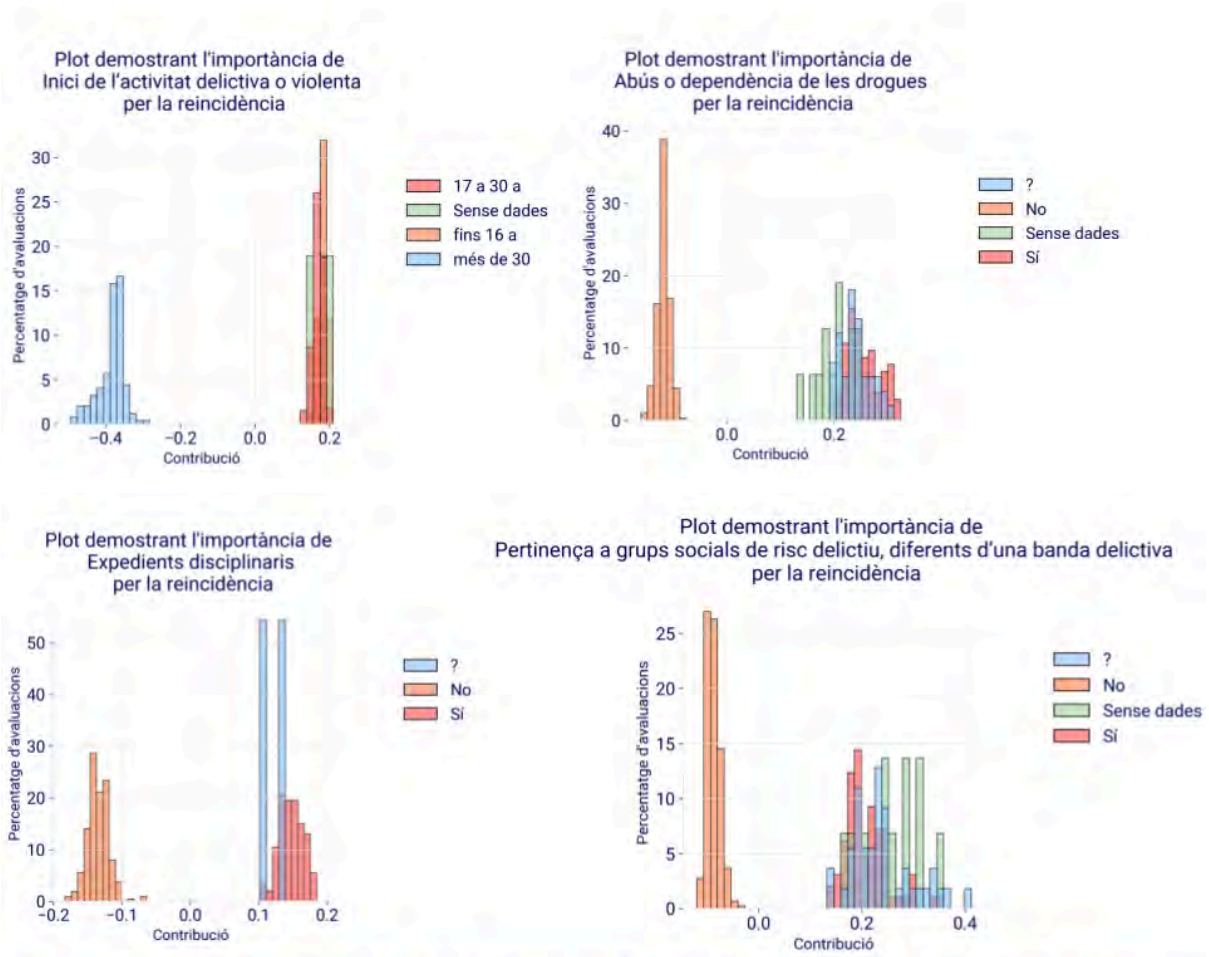
Contribució de cada resposta per la predicció de risc de Violència Intrainstitucional, Trencament de Condemna, Reincidència General i Reincidència Violenta



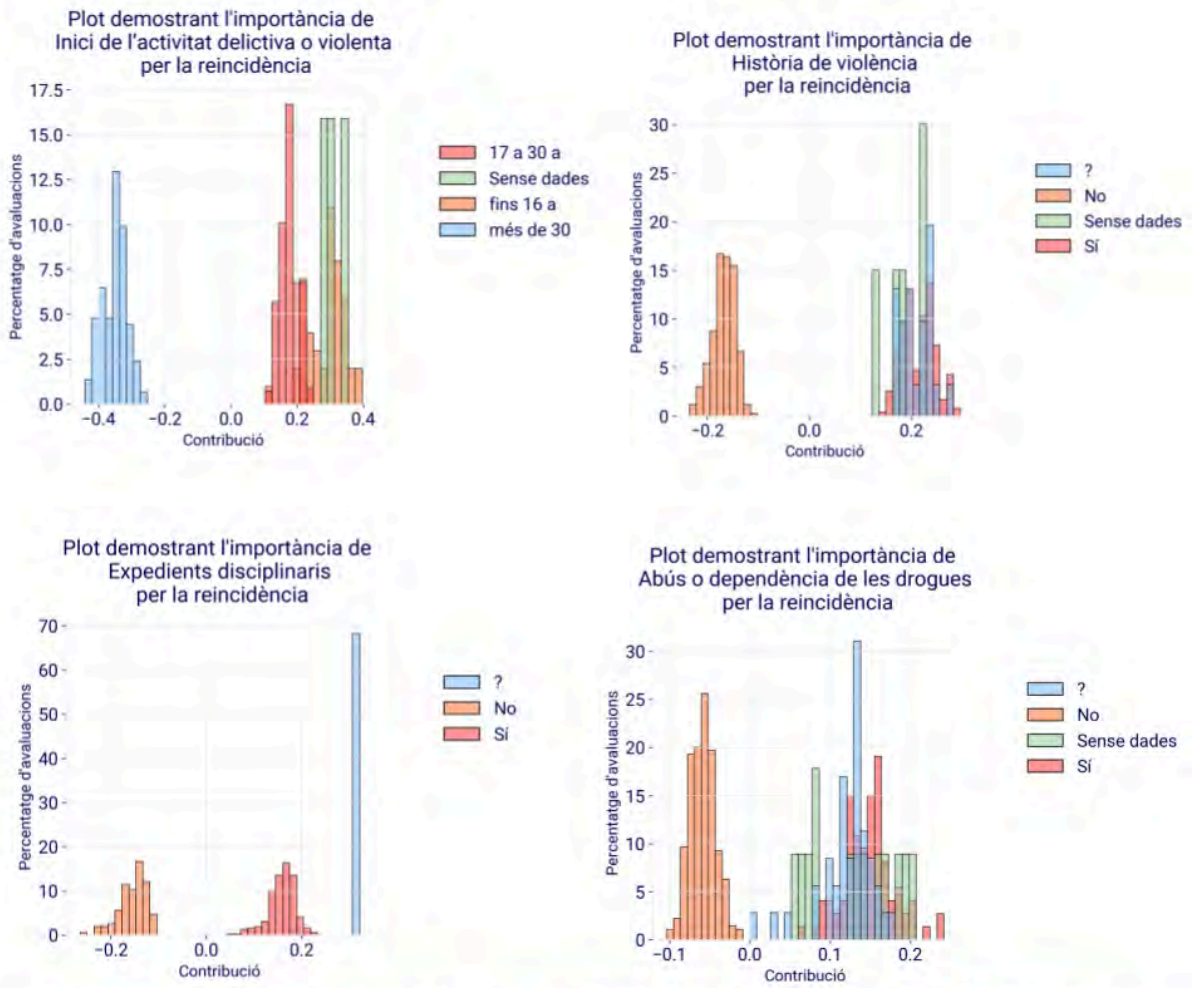
Importància de cada categoria per als factors de risc 10 (conflictes amb altres interns), 12 (expedients disciplinaris), 2 (edat en el moment del delicte base) i 30 (abús o dependència de les drogues) per Violència Intra.



Importància de cada categoria per als factors de risc 38 (actituds pro-criminals o valors antisocials), 34 (resposta limitada a tractament psicològic o psiquiàtric) , 33 (comportament sexual promiscu i de risc, o parafilia) i 2 (edat en el moment del delicte base) per Trencament de Condemna



Importància de cada categoria per als factors de risc 8 (inici de l'activitat delictiva o violenta), 30 (abús o dependència de drogues) , 12 (expedients disciplinaris) i 26 (pertinença a grups socials de risc delictiu, diferents d'una banda delictiva) per Reincidència General



Importància de cada categoria per als factors de risc 8 (inici de l'activitat delictiva o violenta), 7 (història de violència) , 12 (expedients disciplinaris) i 30 (abús o dependència de drogues) per Reincidència Violenta