Informe Tiresias

Resum executiu

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

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.