

Estimació d'Audiències Per a la Corporació Catalana de Mitjans Audiovisuals



CIDAI-POC-2025-01

Drets reservats. Aquest treball està disponible sota la llicència Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International ([CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/)).

Segons els termes d'aquesta llicència, podeu copiar, redistribuir i adaptar l'obra amb fins no comercials, sempre que l'obra sigui citada adequadament, tal com s'indica a continuació.

En qualsevol ús d'aquest treball, no s'ha de suggerir que el CIDAI doni suport a cap organització, producte o servei específic. No es permet l'ús del logotip CIDAI.

Si adapteu l'obra, heu de llicenciar-la amb la mateixa llicència Creative Commons o equivalent.

Si creeu una traducció d'aquest treball, heu d'afegir la següent exempció de responsabilitat juntament amb la cita suggerida: "Aquesta traducció no la va crear el Centre of Innovation for Data tech and Artificial Intelligence (CIDAI). CIDAI no es fa responsable del contingut ni de l'exactitud d'aquesta traducció. L'edició original en català serà l'edició autèntica i vinculant".

Qualsevol mediació relacionada amb disputes derivades de la llicència es durà a terme d'acord amb les normes de mediació de la [World Intellectual Property Organization](https://www.wipo.int/mediation/).

Cita suggerida. CIDAI-POC-2025-01 // Estimació d'Audiències per a la Corporació Catalana de Mitjans Audiovisuals per Karla Andrea Trejo Ramírez, Fundació i2cat, CIDAI, 2025. Llicència: [CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/)



1.	ABSTRACT	4
2.	IMPLEMENTACIÓ DE LA SOLUCIÓ	6
3.	RESULTATS	10
4.	CONCLUSIONS I POSSIBLES MILLORES	14
5.	AGENTS PARTICIPANTS	15



1. Abstract

1.1. Context i Problemàtica

La Generalitat de Catalunya va llançar, en col·laboració amb i2CAT, un programa per desenvolupar un conjunt de projectes “tractors” orientats cap a facilitar l’adopció de tecnologies basades en intel·ligència artificial (IA) per part de l’administració i estan lligats a l’Estratègia IA de Catalunya.

Aquests projectes tractors tenen dos objectius específics:

1. Codissenyar conjuntament amb la Secretaria de Telecomunicacions i Transformació Digital (STTD) i la Secretaria de Polítiques Digitals (SPD) amb els agents públics interessats possibles solucions als reptes claus de transformació digital.
2. Desenvolupar i executar els projectes tractors *mission driven dual-use* d’Intel·ligència Artificial en coordinació amb empreses i entitats líders del sector, altament especialitzades per a demostrar el valor de l’aplicabilitat d’aquesta tecnologia.

En aquest marc de transformació digital, la Corporació Catalana de Mitjans Audiovisuals (CCMA, ara 3CAT), com a ens públic gestor dels mitjans de comunicació audiovisual de la Generalitat de Catalunya, té un rol clau. Amb el seu ferm compromís amb la qualitat, la independència i la promoció de la cultura i la llengua catalanes, 3CAT també afronta reptes significatius en l’àmbit tecnològic per mantenir-se com a referent en el panorama mediàtic català.

Dins de 3CAT, TV3 destaca com el seu principal canal de televisió de cobertura autonòmica, sent el mitjà de comunicació de referència a Catalunya per a la gran majoria dels seus ciutadans. Tanmateix, la gestió d'aquest lideratge requereix eines i informació actualitzada per a prendre decisions estratègiques.

En aquest sentit, la demora en el lliurament de dades d'audiència televisiva impacta a 3CAT en restringir la seva capacitat per a realitzar anàlisis immediates i fonamentades en informació recent.

1.2. Objectiu

L'objectiu principal serà utilitzar dades històriques d'audiències i dispositius connectats de TDT (televisors) del canal de TV3 per entrenar un model d'intel·ligència artificial que demostrï que es poden fer estimacions d'audiències en temps real.

1.3. Impacte

Poder adaptar els continguts en funció de l'audiència estimada en cada moment per oferir els que siguin més interessants per als espectadors o per 3CAT.

1.4. Solució

Segons els requeriments del problema a resoldre, es defineix una arquitectura senzilla i directa de la solució, tal com es reflecteix en la Figura 1.

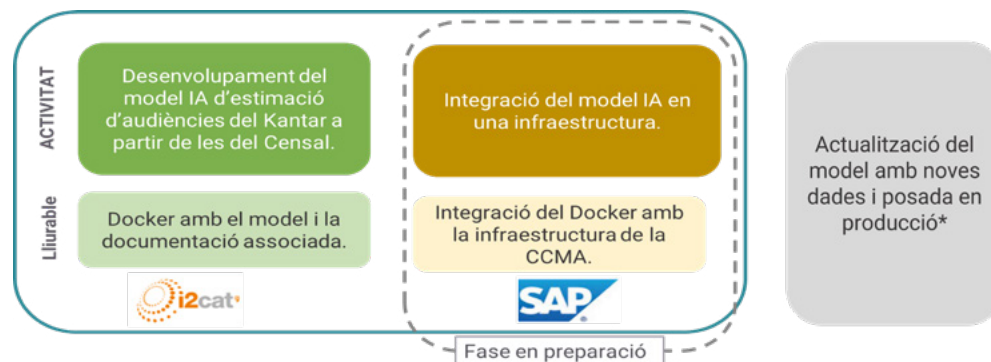


Figura 1: Arquitectura solució

Els elements que componen aquesta arquitectura són els següents:

- Base de dades de dispositius TDT connectats al canal TV3.
- Un model IA de predicció d'audiències basat en aprenentatge automàtic.
- Un contenidor Docker amb API REST per a albergar el model de predicció i interactuar amb ell a través de endpoints.

La integració i la posada en producció del model és a càrrec de l'equip tècnic de 3CAT.

2. Implementació de la solució

2.1. Bases de dades

El dataset de MATR inclou dades en temps real de totes les SmartTV connectades, amb un registre de l'hora exacta i el nombre de dispositius connectats. D'altra banda, el dataset de Kantar ofereix tres tipus de dades diferents: "aud_mitj_milers", "aud_mitj_percent" i "quota". Aquestes variables van ser extretes a partir de dades *raw* en format JSON.

La primera variable reflecteix el nombre total de persones que estan veient la televisió, expressat en milers; la segona indica el percentatge de la població de Catalunya que està sintonitzant TV3; i l'última mostra el percentatge de persones que veuen TV3 en relació amb el total de persones que estan veient la televisió.

Tots dos conjunts de dades abasten el període des de l'1 de juliol de 2023 fins al 8 de febrer de 2024, és a dir, un total de 8 mesos. En total, treballem amb més de 420 000 registres de dades creuades procedents de MATR i Kantar, la qual cosa ens permet una anàlisi detallada i representativa.

La variable principal a predir és l'audiència. No obstant això, durant l'anàlisi exploratòria de les dades (EDA), identifiquem unes certes característiques rellevants que tenen un impacte en la sèrie temporal. En convertir aquestes característiques en metadades de tipus binari, aconseguim incrementar tant la flexibilitat com el poder explicatiu del model.

2.2. Extracció de propietats

En l'anàlisi de les dades, es van identificar diverses variables que tenen una influència significativa en l'audiència i que van ser transformades en predictors booleans per a aportar aquesta informació al model predictiu. Aquests predictors no sols depenen de característiques intrínseques dels programes de televisió, sinó també dels patrons de comportament dels espectadors en relació amb els horaris del dia i els períodes festius, els quals varien al llarg de l'any i són rellevants per a persones de totes les edats.

Un dels predictors clau és **l'estacionalitat**, que diferencia entre els mesos d'estiu (juny-agost) i els mesos d'hivern (desembre-febrer). Durant l'estiu, l'audiència experimenta un augment en el temps lliure degut tant als períodes vacacionals com a la jornada laboral reduïda, la qual cosa impacta en els horaris de consum televisiu. En canvi, durant l'hivern, els hàbits de visualització de televisió poden estar més influenciats per la programació relacionada amb les festivitats, com a especials de Nadal o celebracions de cap d'any, que atreuen a l'audiència en un context més familiar o festiu. En alguns casos, l'audiència pot estar més atreta per contingut relacionat amb les festivitats o esdeveniments especials que per la programació habitual.

Un altre predictor important és el **dia de la setmana**, que distingeix entre els dies laborables i els caps de setmana. Durant la setmana laboral, els horaris de visualització estan més restringits pels compromisos professionals i educatius, mentre que durant el cap de setmana, l'audiència pot disposar de més temps per a veure televisió. Aquest patró és rellevant per a persones de totes les edats, ja que influeix tant en la població laboralment activa com en els estudiants.

En relació amb **l'hora del dia**, es va observar que l'error de predicció es concentra en dues franges horàries específiques: entre les 5-8 hores i les 13-17 hores. Aquestes hores coincideixen amb moments en els quals les persones de diferents edats es troben en transicions entre activitats, com el matí primerenc, quan s'inicia el dia, o la tarda, durant les pauses laborals o escolars. Aquest patró és indicatiu de la disponibilitat de l'audiència en moments clau del dia.

Finalment, el **tipus de programa** també juga un paper fonamental en la predicció de l'audiència. Dins d'aquesta variable, els programes de notícies constitueixen un 44% de la programació, la qual cosa reflecteix una preferència generalitzada per aquesta mena de contingut. Aquesta dada està estretament relacionada amb la rutina diària de les persones, ja que moltes vegades els noticiaris són consumits en horaris específics del dia, com en les primeres hores del matí o a la nit, coincidint amb les pauses diàries o moments de descans. La pregunta associada seria: *"És el programa un noticiari?"*, la qual cosa es vincula amb els horaris clau en els quals les persones, sense importar la seva edat, estan més inclinades a estar enfront del televisor.

2.3. Model d'IA

Abans de determinar el model final utilitzat en aquest projecte, es van considerar diversos enfocaments amb l'objectiu de trobar la millor aproximació per predir l'audiència televisiva. Els models inicialment proposats per a la comparació van ser els següents:

- **Baseline (Multiply by Coefficients):** Aquest model bàsic calcula la mitjana de les dades de Kantar en proporció a les dades de MATR i la multiplica directament per les dades a predir. En aquest cas, s'utilitza la mitjana dels 7 primers dies com a referència.
- **Moving Average Coefficients:** Aquest mètode utilitza el coeficient del dia, setmana o mes anterior entre les dades de Kantar i MATR per calcular les dades de Kantar per al dia actual. Per aconseguir prediccions més precises, s'emptra una mitjana mòbil, que permet captar les tendències a curt termini en l'audiència.
- **Linear Regression:** Aquest model aplica una regressió lineal simple, minimitzant l'error entre les dades del dataset i les prediccions. A diferència dels models anteriors, aquest no utilitza les dades de MATR, sinó exclusivament les de Kantar.

Per avaluar la precisió de cadascun d'aquests models, es van utilitzar les mètriques d'error absolut mitjà (MAE) i l'arrel quadrada de l'error quadràtic mitjà (RMSE). Els resultats preliminars van indicar que el model de regressió lineal proporcionava un millor rendiment en termes d'aquestes mètriques en comparació amb els altres dos models inicials (Figura 2).

Amb aquesta observació, es va decidir expandir el model cap a una aproximació multivariable, incorporant noves característiques derivades de l'extracció de propietats (Secció 2.2). Aquest enfocament ens va portar a implementar un model de **regressió lineal múltiple (MLR)**, que permet aprofitar diverses variables predictives en forma de metadades per millorar la precisió en la predicció de l'audiència televisiva.

Model-id	Model	RMSE	MAE (Audiencia)
1	Baseline (multiply by general average)	45,414.87	28,703.89
2	Previous day coefficients with a filter	39,465.61	26,171.95
2	Previous week coefficients with a filter	34,429.89	23,070.94
2	Previous month coefficients with a filter	34,380.14	23,039.39
3	Linear regression (70%-30%)	34,002.29	22,319.80

Figura 2: Comparativa dels models proposats inicialment

La regressió lineal múltiple (MLR, per les seves sigles en anglès) és una tècnica estadística àmpliament utilitzada per a modelar i analitzar la relació entre una variable dependent, també coneguda com a variable de resposta, i múltiples variables independents o explicatives. Aquest mètode es basa en el cas que existeix una relació lineal entre les variables independents i la variable dependent, la qual cosa permet representar aquesta relació a través d'una equació lineal.

L'objectiu principal de la regressió lineal múltiple és identificar i quantificar com cada variable explicativa contribueix al comportament de la variable de resposta. Això s'aconsegueix mitjançant el càlcul d'un polinomi que descriu aquesta relació (Figura 3). En aquest context, els coeficients associats a cada variable independent representen l'impacte específic de cadascuna sobre la variable dependent, mantenint constants les altres variables del model.

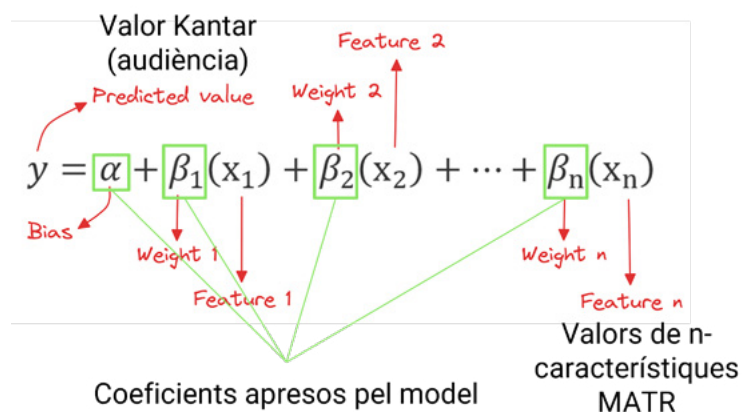


Figura 3: Fórmula i càlcul de la regressió lineal múltiple (MLR)

Un dels principals avantatges de la MLR és la seva capacitat per a avaluar de manera simultània l'efecte de múltiples variables sobre el resultat, la qual cosa la converteix en una eina poderosa per a l'anàlisi predictiva i explicativa. A més, el model permet determinar la significança de cada variable independent, la qual cosa resulta útil per a identificar quines d'elles tenen una major influència en el resultat final.

La regressió lineal múltiple no sols proporciona una predicció numèrica, sinó que també ofereix una interpretació clara de les dades, ja que ajuda a comprendre la interacció entre les variables del model. Això és especialment rellevant en estudis on es busquen patrons en conjunts de dades complexes o quan es desitja realitzar simulacions sobre com canvis en unes certes variables podrien afectar el resultat.

Així, després d'una avaluació exhaustiva i comparativa contra altres algorismes d'aprenentatge automàtic i tècniques estadístiques, es va seleccionar el model MLR a causa de la seva simplicitat, interpretabilitat i eficàcia per a capturar les relacions trobades entre variables. A més, amb MLR es poden incorporar els diversos predictors identificats, com l'estacionalitat, el tipus de programa o l'hora del dia, la qual cosa permet generar prediccions més robustes i fiables, de manera que puguem comprendre d'una millor manera com les diferents variables influeixen en l'audiència televisiva, sense incórrer en models excessivament complexos o difícils d'interpretar.



3. Resultats

A continuació es detallen els resultats obtinguts amb el demostrador.

3.1. Mètriques de rendiment

La Taula 1 mostra el rendiment del model de predicció MLR a través de tres mètriques clau que permeten avaluar l'eficàcia i la precisió del model:

Mean Absolute Error (MAE)

Calcula la diferència absoluta entre els valors observats i els valors predits pel model, proporcionant una mesura directa de l'error mitjà en unitats absolutes. A més, és robusta davant de valors atípics, ja que no amplifica els errors grans.

Root Mean Squared Error (RMSE)

Avalua la distància entre les observacions i la línia de regressió ajustada pel model. Penalitza especialment els errors grans, ja que aquests s'eleva al quadrat, fent-la més sensible a desviacions importants i útil per destacar les discrepàncies significatives.

Coeficient de determinació (R^2 score)

Mesura la proporció de variabilitat en la variable dependent que és explicada pel model. Un valor proper a 1 indica que el model és molt eficaç a l'hora d'explicar la variabilitat en les dades, mentre que un valor més baix suggereix que hi ha factors importants que el model no captura.

Taula 1: Mètriques de rendiment del model MLR

Mètrica	Valor
MAE	21,910.40
RMSE	33,507.00
R² Score	93.30%

Aquest conjunt de mètriques proporciona una visió global i equilibrada del rendiment del model, amb especial èmfasi en el **R² score del 93%**, que evidencia la solidesa i fiabilitat del model en termes predictius.

En la Figura 4, es presenta una gràfica que compara els valors observats amb els valors predits pel model. Aquesta representació visual reforça l'excel·lent resultat obtingut en el R² score, ja que mostra com els valors predits s'alineen de manera molt propera amb els valors reals.

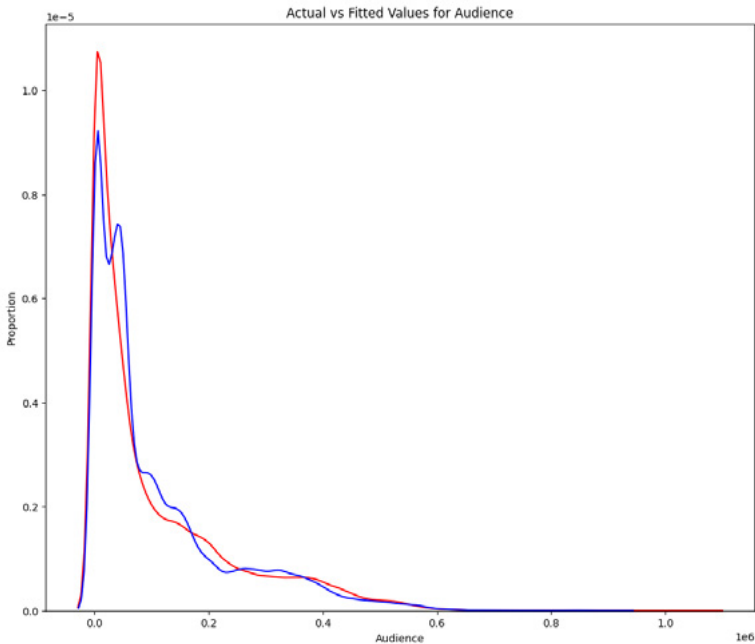


Figura 4: Valors observats vs. Valors predits

3.2. Visualització de l'error

Aquest apartat mostra com la precisió del model ha evolucionat gràcies a la incorporació de les propietats identificades a través dels predictors booleans. Per a il·lustrar aquest progrés, es presenten mapes de calor que permeten contrastar els resultats preliminars del model, abans d'incloure aquestes propietats, amb els resultats després de la seva incorporació.

Els mapes de calor representen de manera granular la precisió en la predicció, desglossada per dia i hora durant una setmana mostra en la Figura 5, i en un mes mostra en la Figura 6. En aquests gràfics, les prediccions amb baixa precisió (0%) apareixen en color vermell, mentre que les zones amb alta precisió (100%) es mostren en color verd, amb una escala de transició entre tots dos extrems.

Aquest tipus de visualització posa de manifest no sols les àrees on el model inicial presentava mancances significatives, sinó també com la introducció de variables com l'estacionalitat, l'hora del dia, o el tipus de programa ha reduït els errors en aquestes àrees. Els resultats mostren una millora clara en la *accuracy*, especialment en franges horàries conflictives, reforçant l'impacte positiu de l'optimització del model.

Figura 5: Precisió en la predicció del model a nivell setmana:
Comparativa abans i després de l'extracció i aplicació de propietats

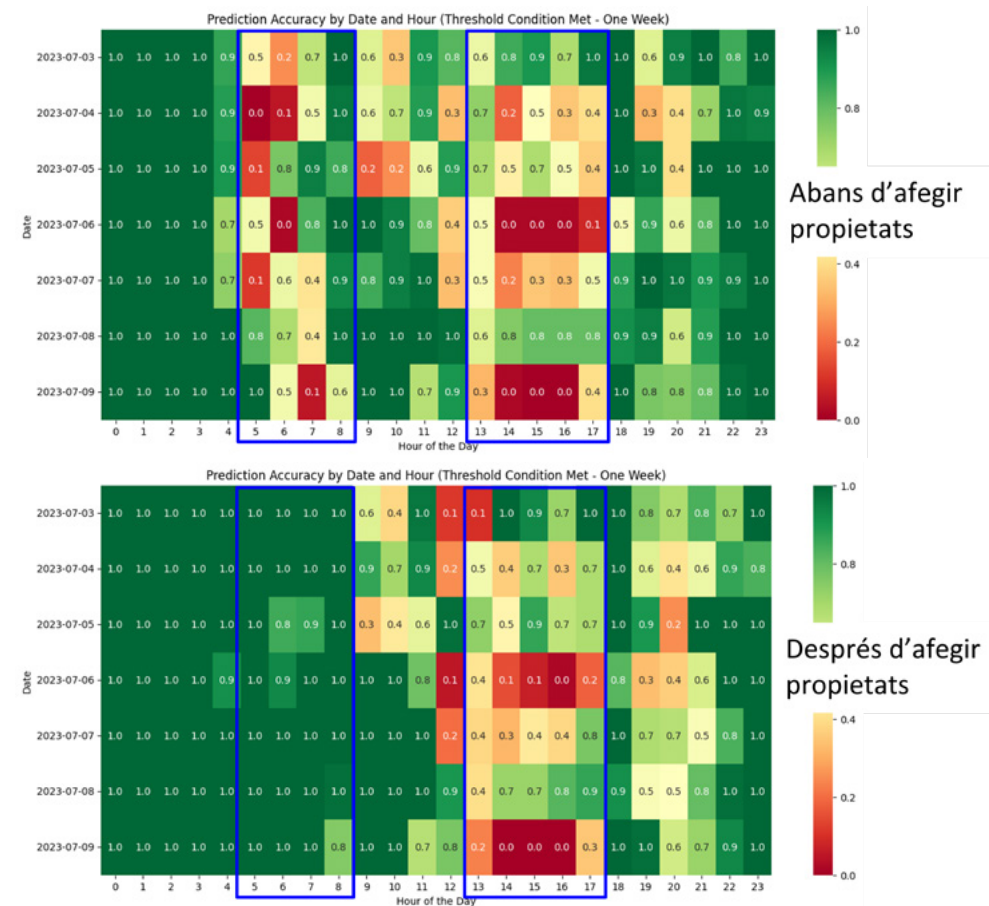
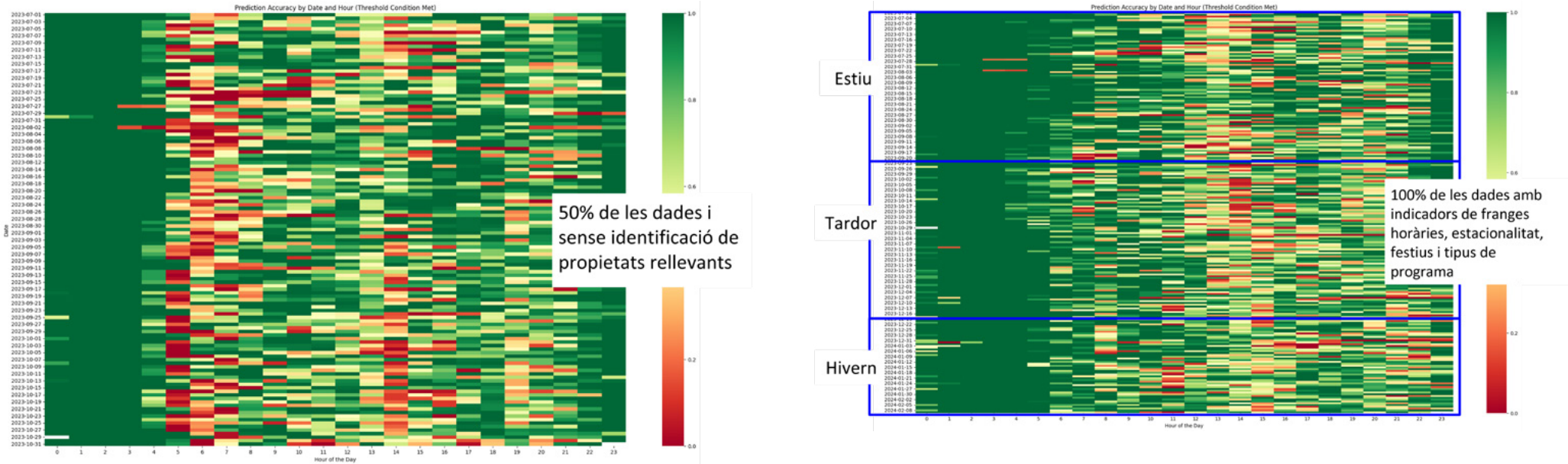


Figura 6: Precisió en la predicció del model a nivell mes:
Comparativa abans i després de l'extracció i aplicació de propietats



4. Conclusions i possibles millores

Les conclusions d'aquest treball són les següents:

- El coeficient R2 és prou bo, superant el 90% tot i que aquestes mesures s'han de prendre amb precaució doncs s'han dut a terme en una finestra de temps concreta i pot haver variabilitat d'un període a un altre. Hem aconseguit reduir l'error de predicció especialment a les zones on hi havia un comportament més diferent.
- S'ha destacat el valor de les característiques extrems durant l'exploració de dades, particularment en relació amb la influència de l'hora del dia i el dia de la setmana en les audiències televisives.
- La visualització mitjançant mapes de calor, que contrasta els valors reals amb les prediccions després de l'entrenament del model, proporciona una gran visibilitat i comprensió de les dades, facilitant així la interpretació i aplicació de les troballes obtingudes.
- Els resultats de les audiències estimades es generen en un format equivalent a l'actual de manera que es pugui aprofitar la informació sense necessitat de realitzar conversions complicades. Això ha permès fer una integració del model de IA en format Docker amb els sistemes de 3CAT per poder visualitzar les estimacions de manera molt directa.
- Els resultats obtinguts en aquest projecte permeten als especialistes enfocar les seves anàlisis i explotar les dades d'acord amb els seus interessos específics, la qual cosa millora significativament la utilitat pràctica de la informació.

- La versatilitat per a l'usuari final s'ha ampliat considerablement des d'aquesta perspectiva, brindant eines més adaptables i útils per a la presa de decisions.

De cara a una evolució futura del sistema, les possibles àrees on dedicar esforços per tal de millorar resultats són les següents:

- Ampliar el *dataset* d'entrenament incorporant un conjunt més ampli de dades per a abastar un marc temporal anual complet, la qual cosa permetria al model aprendre patrons més complets i robustos. Actualment, el dataset cobreix únicament 8 mesos i no inclou els mesos de primavera ni els festius associats a aquestes dates.
- Afinar l'anàlisi de la franja horària del "*prime time*", la qual comprèn generalment les hores entre les 21.00h i les 24:00h, considerant la influència de diversos factors interrelacionats. Això implica ajustar el model per a captar amb major exactitud les variacions i comportaments específics durant aquest període crucial de visualització televisiva.



5. Agents participants

Les següents són les entitats i agents participants en la implementació d'aquesta prova de concepte:

- **CORPORACIÓ CATALANA DE MITJANS AUDIOVISUALS – 3CAT:** Definició del cas d'ús i aportació de dades per a l'entrenament dels models d'IA.
- **FUNDACIÓ i2CAT:** Responsable de la gestió del projecte i de la implementació dels models d'IA.

Socis



Membres

