

PW2-D3: Definició d'un IDSS per a la Gestió de Pandèmies i Preparació de Dades

Presentat per

**Carla ATIENZA, Mireia BRICHES, Jordi GRANJA,
Aleix IBARS, Alberto JEREZ & Pau PRAT**

Supervisors

Dra. Karina GIBERT & Xavier ANGERRI

Grau d'Intel·ligència Artificial

2024-25



Universitat Politècnica de Catalunya
Facultat d'Informàtica de Barcelona

19 Maig de 2025

CONTENTS

1	Definició de l'IDSS	1
2	Descripció de les Dades	2
2.1	Rellevància de Dades	2
2.2	Matrius de Dades Involucrades	3
2.3	Especificació de les Bases de Dades	4
2.4	Consideracions Ètiques	5
2.5	Incertesa de les Dades	5
3	Fonts d'Informació i Coneixement	7
3.1	Fonts d'Informació	7
3.2	Tipus d'Estratègies de Confinament	8
3.3	Regles de Decisió	9
3.4	Esquema Ontològic en Protégé	11
4	Model de Decisió	12
4.1	Definició Formal	12
4.2	Implicació Operativa	12
5	Arquitectura de l'IDSS	14
5.1	Disseny de Models: Entrenament i Validació	14
5.1.1	Dades d'Input	14
5.1.2	Definició	15
5.1.3	Dades d'Output	19
5.1.4	Ús en Producció	19
5.2	Integració i Estructura	20
5.2.1	Inclusive Design	21
5.3	Polítiques Operatives	21
5.3.1	Monitorització del Sistema	21
5.3.2	Comunicació dels Resultats	22
6	Tipus d'Usuari	26
7	Selecció d'eines de Software	28
8	Preprocessament de Dades	30
8.1	Metodologia	30
8.2	Tècniques Aplicades de Preprocessament	30
8.3	Ús de RapidMiner	34
8.4	Esquema general del preprocessament	37
9	Resultats de la Preparació de Dades	38
9.1	Metadades	38
9.1.1	Estructura del Document de Metadades	38
9.1.2	Utilitat de les Metadades per al Nostre IDDS	39
9.2	Matriu de Dades Preprocessades	41
10	Entrenament de Models de Dades	45

10.1	Clustering dels Estats EUA	45
10.2	Prophet: Model de Sèries Temporals	47
11	Representació del Coneixement	51
11.1	Sistema basat en lògica difusa	51
11.1.1	Estructura principal	51
11.1.2	Rol en el sistema	52
11.1.3	Sortida, resultats i validació	52
11.1.4	Integració a l'arquitectura general	52
11.2	Component de models predictius	53
11.2.1	Estructura principal	53
11.2.2	Fonts de contingut	53
11.2.3	Rol en el sistema	53
11.2.4	Sortida, resultats i validació	53
11.2.5	Integració a l'arquitectura general	53
11.3	Component de representació de dades demogràfiques i veïnatge	54
11.3.1	Estructura principal	54
11.3.2	Fonts de contingut	54
11.3.3	Rol en el sistema	54
11.3.4	Sortida, resultats i validació	54
11.3.5	Integració a l'arquitectura general	55
12	Components de Raonament	56
12.1	Components de Raonament	56
12.1.1	Flux i execució del raonament	56
12.1.2	Interacció amb altres components i arquitectura	56
12.1.3	Validació i casos d'ús	57
13	Interaccions de l'IDSS	58
13.1	Integració	60
14	Validació de l'IDSS	62
14.1	Avaluació del Sistema d'Ajuda a la Decisió Intel·ligent (IDSS)	62
15	Interfície i Prova de Concepte	66
16	Línies futures	71
	Bibliografia	71
A	Appendix A: Matrius de dades inicials	73
B	Appendix B: Material de desenvolupament	78

DEFINICIÓ DE L'IDSS

Un *Intelligent Decision Support System* (IDSS) és un sistema informàtic avançat que combina dades, models analítics i eines d'intel·ligència artificial per ajudar en la presa de decisions complexes. Aquest tipus de sistema s'ha demostrat especialment útil en contextos crítics on la rapidesa i l'eficàcia en la presa de decisions poden marcar la diferència, com ara en situacions de crisi sanitària, com les pandèmies.

En el context de la pandèmia de la COVID-19, la gestió de la crisi va evidenciar diverses mancances a nivell global, i especialment als Estats Units, on les polítiques i les restriccions sovint van ser descoordinades i, en molts casos, insuficients per frenar la propagació del virus de manera efectiva. Aquestes deficiències en la presa de decisions van provocar una escalada innecessària dels casos, l'exhauriment dels recursos sanitaris i un gran impacte socioeconòmic. Un exemple notable va ser la manca de gestió òptima dels llits hospitalaris i la implementació de mesures poblacionals poc eficaces en determinades regions, fet que va agreujar la saturació del sistema sanitari.

El nostre IDSS sorgeix com una resposta a aquestes deficiències, proposant una eina capaç d'ofrir suport intel·ligent en la presa de decisions durant pandèmies. Aquest sistema proporciona recomanacions basades en dades sobre com optimitzar les restriccions poblacionals per frenar la propagació del virus, tenint en compte les particularitats de cada estat. A més, integra una funcionalitat per optimitzar la distribució de pacients als llits hospitalaris disponibles, amb l'objectiu de minimitzar la saturació del sistema sanitari.

La importància d'un IDSS en aquest context radica en la seva capacitat per prendre decisions informades i basades en evidències. Això contrasta amb les respostes improvisades que sovint es van donar durant la pandèmia. Per exemple, un IDSS pot analitzar en temps real factors com la densitat poblacional, les taxes d'infecció i les capacitats hospitalàries, per recomanar mesures específiques com confinaments selectius, tancaments escalonats o l'assignació òptima de recursos mèdics. A més, aquestes decisions informades no només poden reduir la propagació del virus, sinó també minimitzar l'impacte socioeconòmic de les restriccions, oferint un equilibri més sostenible entre salut pública i economia.

En conclusió, el nostre IDSS no només destaca per la seva aplicació pràctica, sinó també per la seva rellevància en la millora del sistema de resposta davant futures pandèmies. Mitjançant la integració de dades i ànalisis avançades, aquest sistema pot convertir-se en una eina essencial per evitar la repetició dels errors observats durant la gestió de la COVID-19. Així, el nostre projecte contribueix a un model de presa de decisions més racional, coordinat i efectiu en moments de crisi sanitària global.

DESCRIPCIÓ DE LES DADES

2.1 RELLEVÀNCIA DE DADES

El primer pas per desenvolupar el sistema d'IDSS anterior, consisteix a identificar quines dades poden aportar informació rellevant. En particular, s'han de definir les variables que tenen un impacte directe sobre la situació epidèmica, així com el seu format. Tot això, tenint en compte la viabilitat de recollida d'aquestes dades, balancejant entre la dificultat d'obtenció i la seva potencial utilitat.

- **Variables Epidemiològiques:** permeten monitoritzar l'evolució del brot i anticipar les mesures: casos confirmats, hospitalitzacions, defuncions i recuperacions.
- **Variables sobre Recursos Sanitaris:** permeten optimitzar recursos i aporten coneixement sobre la saturació del sistema sanitari: disponibilitat de llits hospitalaris, equipament mèdic i personal sanitari.
- **Variables Demogràfiques i Geogràfiques:** ajuden a adaptar les mesures segons la densitat de població i vulnerabilitat de cada regió: distribució d'edat, densitat poblacional, veïnatge entre estats.
- **Variables Contextuals:** permeten detectar canvis en les polítiques i la percepció pública: mesures de confinament i dades de xarxes socials.

En una primera anàlisi, aquestes són les dades més generals que són rellevants pel sistema de decisió per pandèmies. Com es pot observar, la majoria dels requisits es tracten de dades estructurades, l'obtenció de les quals dependrà de la seva disponibilitat *open source*. Tot i això, també s'inclouen dades no estructurades, concretament, dades textuales que poden extraure's de les xarxes socials, en aquest cas, per fer *sentiment analysis* i tenir en compte en les recomanacions també les reaccions de la població.

Ara bé, la incorporació d'altres tipus de dades com àudio, imatges o vídeo, és molt rebuscada per aquest context. Aquests tipus de dades requereixen una gran quantitat de recursos per a la seva recollida, processament i anàlisi en temps real, cosa que pot no ser viable en situacions de crisi on la rapidesa i l'eficiència són fonamentals. A més, la informació més rellevant ja es pot obtenir a partir de les altres dades estructurades, que són més fàcils de gestionar i integrar de manera efectiva.

2.2 MATRIUS DE DADES INVOLUCRADES

Una vegada s'ha delimitat el problema que es vol abordar, es procedeix a una selecció rigorosa de bases de dades, la qual garanteix una cobertura exhaustiva dels factors clau en una pandèmia, els quals s'identifiquen en la Sec. 2.1. De cara a poder millorar la gestió en aquestes situacions excepcionals, s'han de recórrer a dades històriques recents, en aquest cas, recopilades amb la Covid-19.

Concretament, es fan servir 11 bases de dades amb informació variada per cada estat dels EUA, des del 13 de gener de 2020 fins al 7 de març de 2021. Tot i que aquest interval no abasta fins al final de la pandèmia (i.e. 18 de setembre de 2022), es tenen en compte les dates més crítiques i la quantitat de dades és suficient per a un IDSS. En la següent taula, es resumeixen les particularitats de cada *database*, les quals compleixen amb els requisits de la pràctica:

Data Source	Nº Records	Nº Variables	Nº Numerical	Nº Binary	Other Qualitative	% NA
<i>states_daily</i>	20780	56	48	0	8	40
<i>us_daily</i>	420	25	21	0	4	9.62
<i>states_info</i>	56	16	1	1	14	10.4
<i>states_pop</i>	52	5	2	0	3	0.38
<i>states_agegp_pop</i>	936	6	2	0	4	0
<i>counties_pop</i>	3220	6	2	0	4	0.48
<i>insurance_cov</i>	1768	13	2	3	8	6.33
<i>beds_per_capita</i>	51	6	4	0	2	4.24
<i>vaccinations</i>	278760	80	75	0	5	61.72
<i>neighbour_states</i>	51	51	0	51	0	0

Table 2.1: Resum de les característiques dels *datasets* emprats.

Respecte al contingut de cada base de dades: les seves variables, el seu tipus i la quantitat de dades mancants, aquesta informació es pot trobar a l'Apèndix A

2.3 ESPECIFICACIÓ DE LES BASES DE DADES

En primer lloc, cal saber quina informació aporta cada base de dades al sistema de suport a la decisió. Això servirà no només per identificar els ingredients disponibles, sinó també per anar intuint quins models de dades seran els més adequats per resoldre el problema.

- ❑ **states_daily**: dades diàries sobre la COVID-19 als EUA en l'àmbit estatal, amb informació sobre casos (positius, negatius, pendents), hospitalitzacions (actuals i acumulades), ingressos a l'UCI, ventilació mecànica, defuncions, recuperacions i proves diagnòstiques, amb indicadors d'increment diari, timestamps i identificadors únics.
- ❑ **us_daily**: els mateixos registres diaris que **states_daily**, però agregat per tots els EUA.
- ❑ **states_info**: informació detallada per cada estat, com el nom, el codi FIPS i enllaços a fonts oficials, acompanyats de notes específiques i detalls sobre els camps i unitats de mesura utilitzats (metadades).
- ❑ **states_pop**: dades poblacionals per estat, amb identificadors com el nom i codi FIPS, que inclouen el nombre d'habitants i la densitat poblacional.
- ❑ **states_agegp_pop**: la mateixa informació que **states_pop**, però desglossada per grups d'edat en cada estat.
- ❑ **counties_pop**: dades poblacionals a nivell de comtat, incloent el nom del comtat, estat, codi FIPS i densitat poblacional.
- ❑ **insurance_cov**: dades sobre la cobertura d'assegurances mèdiques per estat i grup d'edat, presentant tant nombres absoluts com percentatges, així com variables addicionals sobre l'estat ocupacional i marges d'error.
- ❑ **beds_per_capita**: informació sobre el nombre de llits hospitalaris per 1000 habitants, amb el desglossament segons administracions públiques, organitzacions sense ànim de lucre i empreses privades, acompanyat dels identificadors d'estat i codi FIPS.
- ❑ **vaccinations**: registres diaris de vacunació als EUA a partir del 13 de gener de 2020, indicant el nombre d'habitants d'un estat amb la primera dosi, o bé la pauta completa.
- ❑ **neighbour_states**: matriu de dades quadrada que indica de forma binària els estats el territori dels quals són adjacents.

Les anteriors, són bases de dades obertes que es recopilen *off-line* per entrenar els models de dades estàtics. És a dir, proporcionen una base estable a partir de la qual desenvolupar l'IDSS. En particular, es descarreguen un únic cop a partir de dues fonts d'internet principals: [tap-covid-19](#)¹, pels registres d'evolució de la pandèmia i informació dels estats; i <https://data.cdc.gov>, per les dades de vacunació. En ambdós casos, es tracten de fonts confiables i públiques (*open data*), que provenen directament o indirectament d'institucions

¹ Directori de Github que ofereix dades variades de la Covid-19, però que venen principalment de <https://covidtracking.com/>

governamentals dels EUA.

No obstant això, en el context de producció i per garantir la continuïtat i actualització de les dades, serà necessari establir un mecanisme de recopilació més dinàmic, com ara una API en què els estats diàriament reportin les dades.

2.4 CONSIDERACIONS ÈTIQUES

L'anàlisi de dades mèdiques, de vacunació i de població en el context de la COVID-19 es considera un ús secundari de les dades, ja que aquestes van ser recollides inicialment per a finalitats clíniques i ara es reutilitzen per la planificació sanitària, la investigació epidemiològica i l'avaluació de polítiques. Tot i que les dades obtingudes estan anonimitzades, cal garantir que no es puguin tornar a identificar els individus mitjançant la combinació d'altres variables. A més de les mesures tècniques, cal respectar els marcs legals i ètics aplicables:

- ❑ **Protecció de les dades:** Les dades estan subjectes al Reglament General de Protecció de Dades (RGPD) de la UE, que en ús secundari de les dades exigeix garantir la finalitat limitada i la minimització de dades. Amb l'anonymització això està bastant garantit.
- ❑ **Transparència i consentiment:** Si bé les dades anònimes no necessiten consentiment, és interessant informar el públic sobre el seu ús.
- ❑ **Llicència GPL:** Les dades també estan protegides per la Llicència Pública General de GNU (GPL), que permet la seva distribució i ús sense modificacions.
- ❑ **Ètica i IA Explicable en Medicina:** Els models d'IA dissenyats per a assistir en situacions mèdiques han de ser explicables, assegurant que les seves decisions siguin comprensibles pels professionals i pacients. Això facilita la transparència, la validació científica i l'acceptació responsable de recomanacions clíniques. L'ús de models opacs («caixa negra») en àmbits crítics com la salut pot ser èticament qüestionable, ja que pot comprometre la rendició de comptes i la confiança en el procés mèdic.

2.5 INCERTESA DE LES DADES

Les dades involucrades en l'anàlisi reflecteixen majoritàriament registres agregats en l'àmbit estatal, cosa que les situa més com a dades poblacionals que com a mostres. En vista de l'origen de les dades (i.e. fonts oficials i públiques), la recopilació d'aquesta informació garanteix un cert nivell de qualitat i rigor, disminuint la incertesa externa a les dades.

Tanmateix, això no elimina la incertesa inherent a les dades. En aquest cas, es troben grans percentatges de valors mancats (NA) i diferències en com reporten els casos cada estat. També pot existir una incertesa temporal, deguda a la diferència de temps entre que es produeix una infecció i es detecta, fet que pot afectar l'actualització i la precisió de les recomanacions.

A més, determinats grups poden estar infrarepresentats, com ara les comunitats rurals on la infraestructura de recollida de dades no és suficient. Tenir en compte aquests aspectes és fonamental per conèixer els possibles biaixos i imprecisions del sistema, i així implementar tècniques de gestió de la incertesa (e.g. imputació de dades mancants) que permetin millorar la robustesa del sistema de suport a la decisió.

FONTS D'INFORMACIÓ I CONEIXEMENT

3.1 FONTS D'INFORMACIÓ

Aquest treball es fonamenta en una combinació de fonts de dades estructurades i coneixement expert extret de la literatura científica i sanitària. Aquesta doble font permet una presa de decisions basada en evidència, combinant dades empíriques amb regles inferides de l'experiència en gestió de pandèmies. Pel que fa al coneixement expert, s'han analitzat diverses fonts principals:

- **Framework for Equitable Allocation of COVID-19 Vaccine** [1], on s'estableixen criteris de prioritació per a la distribució de vacunes basats en risc, exposició i vulnerabilitat. D'aquest document s'ha extret la regla de prioritzar les zones amb densitat de població en el quartil superior per vacunes i recursos addicionals.
- **Aggressive containment, suppression, and mitigation of COVID-19: lessons learnt from eight countries** [2], que analitza estratègies de contenció en vuit països i defineix llindars com 28 dies seguits amb 0 casos nous per declarar eliminació de transmissió, i més de 50 casos nous per 100k habitants com a llindar per aplicar quarantena a viatgers.
- **Optimal lockdowns for COVID-19 pandemics: Analyzing the efficiency of sanitary policies in Europe** [3], que presenta models de flux basats en SIRD i defineix llindars clau com ara ocupació UCI >80% i taxa de reproducció efectiva $Rt > 1.5$ per activar confinaments.
- **OMS** [4], que estableix que un $Rt > 1$ indica creixement de l'epidèmia i recomana una taxa de positivitats inferior al 5% per considerar l'epidèmia controlada.
- **ECDC** [5], que recomana una incidència acumulada (IA) inferior a 150 casos per 100.000 habitants en 14 dies per evitar risc alt i estableix que una $Rt > 1.5$ requereix mesures estrictes.

Aquest coneixement s'ha traduït en regles de decisió concretes amb llindars numèrics, adaptats a les variables disponibles als datasets. Aquestes regles s'han sistematitzat i integrat en taules de criteris per al seu ús dins de l'IDSS, amb l'objectiu de permetre accions preventives i una millor gestió de recursos durant futures pandèmies.

3.2 TIPUS D'ESTRATÈGIES DE CONFINAMENT

Durant la pandèmia, els països van adoptar diferents estratègies per contenir la propagació del virus i minimitzar el seu impacte sobre la salut pública i l'economia. Aquestes estratègies es poden classificar en tres enfocaments principals: contingut agressiu, supressió i mitigació. Cada enfocament presenta avantatges i limitacions, i la seva elecció sovint depèn de factors com la capacitat sanitària, la densitat de població, la disponibilitat de vacunes i la resposta social a les mesures de confinament.

Aquesta classificació es fonamenta en l'anàlisi realitzada per Wu et al. (2021) [2], que compara l'eficàcia d'aquestes estratègies a nivell internacional.

1. Contingut Agressiu (*Eliminació*)

L'objectiu d'aquesta estratègia és eliminar completament la transmissió comunitària del virus. Per aconseguir-ho, s'implementen mesures estrictes des del primer moment en què es detecten casos, com ara confinaments totals, rastreig exhaustiu de contactes i quarantena obligatòria per als casos sospitosos i per als viatgers procedents de zones d'alt risc. Aquest enfocament també inclou un control molt estricte de fronteres, amb quarantena de 14 dies per a tots els viatgers entrants.

Un llindar clau associat a aquesta estratègia és la consecució de 28 dies consecutius sense cap cas nou, moment en què es considera que la transmissió ha estat eliminada a nivell comunitari i es poden relaxar les mesures internes. Aquest enfocament ha estat seguit amb èxit per països com Nova Zelanda, Taiwan i Vietnam, que han aconseguit mantenir nivells molt baixos de mortalitat i una recuperació econòmica més ràpida.

Dins d'un sistema IDSS, aquesta estratègia implica activar immediatament mesures de contenció davant qualsevol brot i relaxar-les només quan es compleixi el llindar dels 28 dies sense casos.

2. Supressió

L'estratègia de supressió busca reduir la transmissió del virus a nivells manejables, sense necessàriament eliminar-la completament. Aquesta estratègia es basa en l'aplicació de mesures temporals i adaptatives, com ara confinaments intermitents o restriccions puntuals, que s'activen quan determinats indicadors epidemiològics (com el nombre reproductiu efectiu R_t o l'ocupació de llits d'UCI) superen certs llindars de risc.

Un dels llindars clau en aquesta estratègia és $R_t > 1.5$ o una ocupació d'UCI superior al 80%, que actuen com a senyals d'alerta per implementar restriccions. Aquesta estratègia ha estat aplicada per països com Alemanya i Irlanda, que han adoptat una gestió flexible i basada en dades per contenir les onades epidèmiques sense arribar a confinaments totals permanents.

En el context de l'IDSS, aquest enfocament permet monitoritzar constantment els indicadors clau i activar mesures preventives quan es detecta un risc imminent de saturació del sistema sanitari.

3. Mitigació

L'estratègia de mitigació parteix de la premissa que la transmissió no pot ser eliminada i que s'ha d'acceptar un cert nivell de circulació del virus, especialment entre població de baix risc, mentre es protegeixen els grups més vulnerables. L'objectiu principal és evitar el col·lapse del sistema sanitari, minimitzant la mortalitat i l'impacte en la salut pública, però sense aplicar mesures tan restrictives que puguin perjudicar l'economia o la societat de manera excessiva.

Aquesta estratègia prioritza la continuïtat de l'activitat econòmica i social, i pot incloure una estratègia d'immunitat natural progressiva. Els llindars associats són generalment més alts que en les altres estratègies; per exemple, s'activen mesures només quan l'ocupació hospitalària supera el 90% o quan la taxa de mortalitat supera l'1%.

Aquest enfocament va ser adoptat, en fases inicials, per països com Suècia i el Regne Unit, que van optar per una resposta més moderada, especialment en les primeres onades.

En un sistema IDSS, l'estratègia de mitigació es tradueix en una menor freqüència d'intervencions, centrant-se en la protecció de grups vulnerables i en l'optimització dels recursos hospitalaris quan es detecta un risc real de col·lapse.

3.3 REGLES DE DECISIÓ

Les regles de decisió defineixen les accions a prendre davant valors específics de variables epidemiològiques, sanitàries i socioeconòmiques. Aquestes regles han estat dissenyades a partir del coneixement expert extret de la literatura científica i s'han adaptat a les dades disponibles. Els llindars s'han establert per permetre la detecció precoç de situacions de risc i optimitzar la resposta sanitària.

La taula 3.1 resumeix les principals variables utilitzades, els llindars associats i les decisions que en deriven.

Variable	Llindars	Decisió
Ocupació Hospitalària (θ)	$\theta > 80\%$: Saturació crítica del sistema. $70\% < \theta < 80\%$: Saturació imminent.	Reforçar recursos si $\theta > 70\%$. Activar mesures de contenció immediates (restriccions, augment de recursos) si $\theta > 80\%$.
Taxa de Mortalitat (π)	$\pi > 20$: Situació crítica. $10 < \pi < 20$: Risc alt. $\pi < 10$: Risc moderat.	Confinament immediat si $\pi > 20$. Restriccions locals si $10 < \pi < 20$. Monitoratge si $\pi < 10$.
Taxa de Letalitat (CFR)	CFR > 5: Situació crítica. $3 < \text{CFR} < 5$: Risc alt. $\text{CFR} < 3$: Risc moderat.	Aturar activitat CFR > 5. Restriccions locals si $3 < \text{CFR} < 5$. Monitoratge si CFR < 3.
Incidència acumulada (IA)	IA > 150 casos per 100.000 hab. en 14 dies: Risc alt. $50 < \text{IA} < 150$: Risc moderat. $\text{IA} < 50$: Risc baix.	Confinament estricte si IA > 150. Confinament selectiu si $50 < \text{IA} < 150$. Restriccions locals si IA < 50.
Variables Demogràfiques	>20% de població >65 anys: Risc alt de mortalitat. $10\% < \text{població} > 65 \text{ anys} < 20\%$: Risc moderat. població >65 anys < 10%: Risc baix.	Confinaments selectius per a grups d'alt risc si població >65 anys > 20%. Vacunació prioritària si població >65 anys > 30%.
Densitat de Població	Quartil superior ($\geq 75^{\text{è}} \text{ percentile}$).	Prioritzar la zona per vacunes i recursos addicionals.
Casos Diaris (positiveIncrease)	14 dies consecutius amb 0 casos: Epidèmia controlada.	Aixecar confinament si es compleixen els 14 dies seguits amb 0 casos.
Quarantena Viatgers	>50 casos nous per 100k hab. en la darrera setmana.	Quarantena de 14 dies obligatòria per a viatgers procedents de l'estat d'alt risc. Prohibició de viatges si casos nous > 100.
Redistribució Llits (ϵ)	Diferència >30% entre ocupacions UCI d'estats fronterers.	Moure recursos (llits, personal) entre regions veïnes amb saturació desigual.

Table 3.1: Criteris de decisió durant la pandèmia (sense les variables eliminades)

3.4 ESQUEMA ONTOLOGIC EN PROTÉGÉ

L'esquema representat a la Figura 15.6 mostra una versió simplificada de l'ontologia desenvolupada a Protégé, amb l'objectiu de representar visualment com es pren una decisió sobre el tipus de confinament a aplicar a partir de l'anàlisi de variables epidemiològiques. En aquest gràfic, es reflecteix la relació entre l'estat, la seva situació epidemiològica, les variables crítiques (com el nombre reproductiu efectiu R_t , la taxa de mortalitat o l'ocupació hospitalària pertinents a la taula anterior 3.1), i els diferents tipus de confinament.

Tot i que l'ontologia podria incloure més detalls (com els llindars concrets, les regles de decisió o les estratègies globals com supressió o mitigació), s'ha optat per aquesta representació simplificada per tal de fer explícit el flux bàsic de raonament: des de l'observació d'indicadors fins a l'activació de mesures de contenció. Això permet que tots els membres de l'equip tinguem una base compartida sobre com es poden formalitzar i aplicar les idees de coneixement expert en el context del sistema IDSS.

Aquest esquema constitueix, doncs, un primer pas per estructurar el coneixement expert dins del sistema, sobre el qual es poden afegir posteriors capes de detall o refinament semàntic.

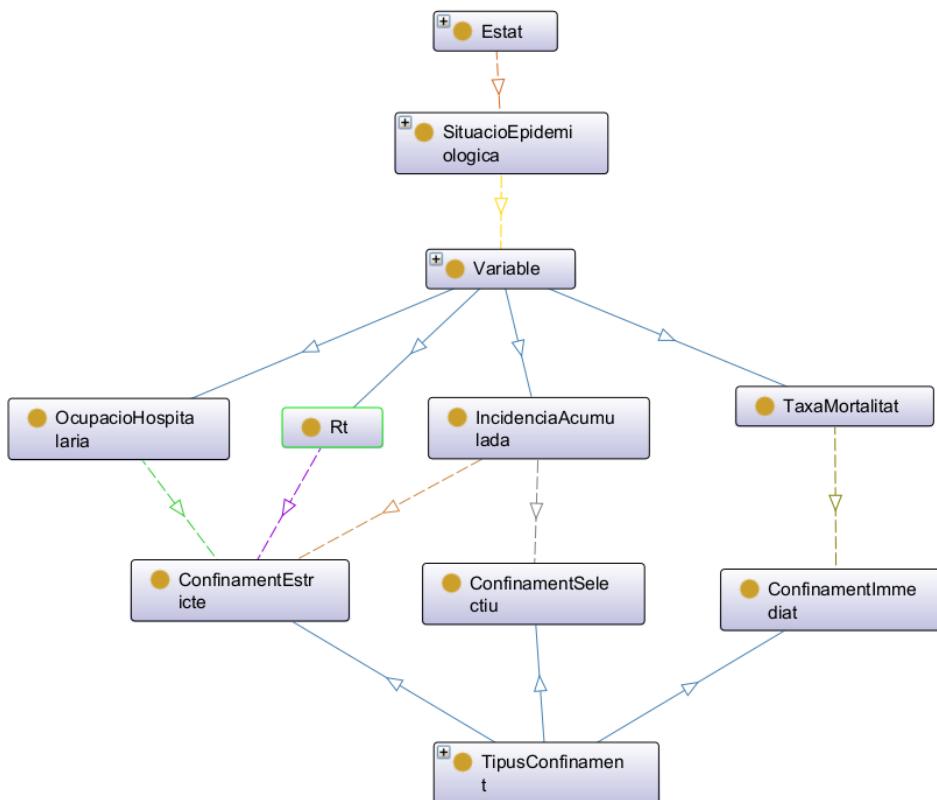


Figure 3.1: Esquema simplificat de l'Ontologia - Protégé

MODEL DE DECISIÓ

4.1 DEFINICIÓ FORMAL

Una vegada s'ha definit el coneixement i les dades disponibles, ja es poden fixar les decisions de suport del sistema.

Decisió	Tipus	Decision-maker	Input Data	Output
Quins estats confinar demà?	Prescripció de mesures de confinament amb prediccio de tendències futures.	Comissió del Centre Nacional d'Emergències d'EEUU	Evolució de la pandèmia: contagis, morts, tests. Vacunes posades. Altres: cens de població per rang d'edat, recursos hospitalaris i assegurances mèdiques.	Puntuació del nivell de risc, junt amb una etiqueta qualitativa de recomanació per cada estat: normalitat, confinament parcial o total.
Com repartir les vacunes disponibles?	Distribució de recursos amb elements predictius.	Comissió del Centre Nacional d'Emergències d'EEUU	Evolució de la pandèmia: contagis, morts, tests. Vacunes posades. Altres: cens de població per rang d'edat, recursos hospitalaris i assegurances mèdiques.	Rànquing de prioritat segons el nivell de risc, acompañat d'un pla de distribució que especifiqui el % o nombre de vacunes per estat.
Com optimitzar la distribució de llits/persones en risc?	Distribució de recursos amb elements predictius	Comissió del Centre Nacional d'Emergències d'EEUU	Les mateixes dades que per la decisió anterior, però afegint informació de la situació en els estats veïns a un estat.	Índex de pressió hospitalària i risc de saturació i, tenint en compte això, recomanació de transferències de pacients entre estats.

Table 4.1: Definició formal de les decisions de suport de l'IDSS.

4.2 IMPLICACIÓ OPERATIVA

El sistema intel·ligent de suport a la decisió proposat proporciona informació estratègica per ajudar les autoritats en la gestió de la pandèmia. Tenint en compte diferents aspectes implicats en l'evolució del brot, com el nivell de vacunació o les poblacions més vulnerables, la decisió final comprendrà millor la situació i, per tant, serà més adequada. D'una banda, la prediccio de tendències aportarà una component preventiva al sistema, molt important en aquest cas sanitari. D'altra banda, considerar les fronteres entre estats i les seves dades, incorpora un enfocament informat localment, cosa que té sentit en un problema de propagació.

Basant-se en l'evidència de les dades, el resultat dels models ajudarà a maximitzar l'eficiència del procés de vacunació, prioritant les regions més vulnerables, minimitzar la càrrega del

sistema hospitalari, o bé controlar els rebrots amb la recomanació de confinaments. Tot això, a partir de coneixement de l'expert i resultats operacionals, com els rankings de prioritat i els índexs de risc, converteixen l'anàlisi predictiva en accions concretes, més preventives i eficients.

ARQUITECTURA DE L'IDSS

5.1 DISSENY DE MODELS: ENTRENAMENT I VALIDACIÓ

5.1.1 Dades d'Input

	Training	Production
Dades	Evolució diària de la pandèmia i vacunació. També, dades estàtiques del cens, recursos hospitalaris i assegurances.	Evolució diària de la pandèmia i vacunació: contagis, morts, hospitalitzacions, etc.
Origen de les Dades	Obertes a les Webs: https://covidtracking.com/ & https://data.cdc.gov	Compartides per l'autoritat sanitària.
Accés a les Dades	off-line CSVs.	Accés amb API a les dades internes diàries de cada estat.
Freqüència de les Dades	Un cop, històric de la Covid-19	Cada dia de la pandèmia.
Transformació de Dades	Creació de variables, e.g diferències diàries de contagi o ràtios.	Creació de les mateixes variables d'entrenament.
Viabilitat de la Transformació	Sí, un preprocessat de les dades adequat ho permet.	Sí, sempre que els estats reportin les dades requerides.

Table 5.1: Descripció dels requeriments de dades d'entrada

Com es pot observar, a diferència de la fase d'entrament en què les dades es recullen *off-line*, en la fase de producció s'utilitzen dades dinàmiques *on-line*, proporcionades per les autoritats sanitàries que facin ús del sistema. També, de cara a perfilar el sistema, s'incorpora el coneixement de l'expert extret de la recerca de les fonts d'informació a la Sec. 3.

5.1.2 Definició

Models de Dades

En vista de la component temporal de les dades, es requereixen models de predicció de sèries temporals per anticipar rebrots o deficiències en el procès de vacunació. En aquest sentit, existeixen diferents opcions, però algunes amb alguna limitació: ARIMA, no pot capturar patrons complexos o no lineals; LSTM, si bé permet un modelatge global, amb un *embedding* fixe per cada estat, la quantitat de dades no es suficient. Aleshores, la proposta es fer servir Prophet de Facebook, un model de regressió additiu que combina tendències flexibles $g(t)$ (lineals o logístiques), estacionalitats $s(t)$ (diàries, setmanals, anuals), efectes de festius $h(t)$ i variables externes X (e.g., dades socioeconòmiques). L'equació bàsica del model és:

$$y(t) = g(t) + s(t) + h(t) + \beta X + \epsilon \quad (5.1)$$

La principal implicació d'aquesta decisió recau sobre el fet que cada estat tindrà un model independent, entrenat amb les seves dades històriques i característiques pròpies. Aquesta separació evita el biaix creuat entre estats amb dinàmiques diferents (e.g., Nova York vs. Wyoming). Això significa que el sistema involucrarà 56 equacions d'aquest tipus, la qual cosa pot ser costós computacionalment.

De cara a la presa de decisions, s'ha considerat necessari poder fer *forecasting* de la quantitat de casos positius, hospitalitzacions i defuncions que tindran lloc en un estat particular. D'aquesta manera, es podran dur a terme mesures preventives que minimitzin els efectes de la pandèmia. La següent qüestió és evident, quin conjunt de variables disponible pot ajudar a predir els requisits anteriors?

Evidentment, els models de Prophet rebran com a *input* les mateixes dades que es volen predir $y(t)$. També, com a regressors externs X amb un potencial impacte sobre els *targets*, s'han identificat les següents variables: taxes de vacunació (`Dose1_Total_pct`, `Complete_Total_pct`), dades demogràfiques (`population`, `pop_density`, `pct_pop_agegroup`) i geogràfiques (`Metro_status`), cobertura d'assegurances (`coverage_type`, `estimate`), recursos hospitalaris (`beds_per_capita`) i indicadors socioeconòmics (`SVI_CTGY`). Aquest conjunt gran de variables s'utilitzarà com a punt de partida per a l'entrenament, i s'avaluarà com impacten sobre les prediccions (i.e. amb els coeficients β). Com que es tracta d'un model de regressió, no hi ha problema a incloure-les totes d'entrada i ajustar segons els resultats.

Després d'entrenar els models amb els conjunts de dades anteriors, el resultat no només serà una predicció de $y(t)$, sinó que també una descomposició dels efectes individuals (tendència, estacionalitat, festius i regressors externs) de la Eq. 5.1. Així, d'una banda, obtindrem prediccions per als següents 7 dies (horitzó), amb per a cada dia el valor puntual i els seus intervals de confiança – que, per defecte, cobreixen un 80% d'incertesa – i, d'altra banda, es podrà analitzar la contribució específica de cada component, cosa que facilita la interpretació dels resultats i la presa de decisions en temps real.

Tot i això, l'enfocament anterior presenta un obstacle evident: no es viable computacionalment entrenar 168 models. (56 estats \times 3 *targets*). A més, com es menciona a l'apartat 2.5, el conjunt de dades disponible presenta algunes incerteses, com valors faltants o dies en què alguns estats no reporten dades. Aleshores, per tal d'augmentar el conjunt d'entrenament dels models Prophet i reduir la càrrega computacional, es troba adient realitzar un *clustering* entre els estats. En el context d'un virus, si les característiques de diferents regions (e.g. la densitat poblacional) són similars, és raonable assumir que la dinàmica temporal seguirà patrons semblants. Amb aquesta estratègia, s'aconsegueix un millor equilibri entre la personalització per territori i l'eficiència computacional i dels models, especialment en aquells casos amb dades escasses.

Però, com es poden agrupar els estats de manera que es mantengui una coherència i consistència, que sigui suficient i no arrisqui la personalització de cada regió?

- **Clustering sobre sèries temporals:** permet capturar la dinàmica epidemiològica de cada estat (casos positius, hospitalitzacions, defuncions). En analitzar les formes i tendències de les sèries (e.g. amb DTW), es poden agrupar els estats segons comportaments temporals similars, independentment de la seva magnitud absoluta.
- **Clustering per característiques dels estats:** com ara la densitat de població, la distribució d'edats, els recursos sanitaris (beds_per_capita), i la cobertura d'assegurances. Aquestes variables són factors estructurals que influeixen sobre la resposta al virus i la capacitat del sistema sanitari. Per tant, un *clustering* en aquest espai permet identificar estats amb condicions estructurals comparables.

La combinació d'aquests dos enfocaments és especialment robusta, perquè integra tant la informació dinàmica com la contextual. Una estratègia viable seria aplicar cada *clustering* per separat i després combinar els resultats, per exemple, amb tècniques de consens. Tot i això, es poden perdre les interaccions entre *clusters*, de manera que s'escull un enfocament híbrid amb fusió de característiques.

En primer lloc, es proposa una reducció dimensional de les sèries temporals a cada *target*, de manera que es pugui incorporar un vector compacte amb aquesta informació en la matriu d'atributs dels estats. Per mantenir la naturalesa del *clustering*, s'utilitza *Dynamic Time Warping* per capturar les similituds entre sèries temporals dels estats. Després, la matriu de distàncies resultant es converteix en un espai de baixa dimensió amb *Multidimensional Scaling*, preservant les relacions originals entre les sèries.

A continuació, els vectors temporals reduïts i les característiques estructurals es fusionen en un únic espai multidimensional, sobre el qual s'aplica un *clustering*, amb una mètrica adaptada que considera el pes relatiu de cada component. En particular, serà la següent distància ponderada:

$$\text{Dist}(A, B) = w_{\text{temp}} \cdot d_{\text{temp}}(A, B) + w_{\text{est}} \cdot d_{\text{est}}(A, B) \quad (5.2)$$

On:

- $w_{\text{temp}} = 0.6, \quad w_{\text{est}} = 0.4$
- d_{temp} : Distància euclidiana entre les components temporals
- d_{est} : Distància euclidiana entre les característiques estructurals.

Finalment, el resultat d'aquest model no supervisat serà un vector d'etiquetes assignat a cada estat (e.g. Cluster1, Cluster2, etc.). Aquest enfocament combinat no només redueix la càrrega computacional, sinó que també permet millorar la robustesa i la generalització dels models de predicció. Això és perquè si alguns estats tenen comportaments molt sorollosos, integrar-los en un grup pot ajudar a suavitzar prediccions i evitar sobreajust.

Models de Coneixement

En el context de la gestió sanitària, la única forma de decidir quan un estat s'ha de confinar, potenciar la vacunació o transferir pacients, és a partir de regles de coneixement expert. Es tracta de decisions crítiques, les quals no poden basar-se només en models de dades, ja que aquests poden ser sensibles a biaixos, limitacions històriques i manca de criteris ètics. Per això, és essencial fonamentar les accions en evidència científica i criteri mèdic per garantir una resposta adaptativa i ètica.

L'ús d'un model de coneixement basat en lògica difusa resulta especialment interessant en aquest context, ja que permet gestionar la incertesa i la imprecisió inherents a la gestió d'una pandèmia. En aquest cas, resulta molt difícil determinar, per exemple, la quantitat de casos exacta a partir de la qual confinar un estat. En lloc de prendre decisions binàries, la lògica difusa ens permet tractar conceptes com "una taxa de vacunació alta" o "una capacitat hospitalària mitjana", aportant una major flexibilitat a l'hora de fer les recomanacions. Per últim, les regles d'aquest tipus faciliten la integració de les múltiples fonts d'informació amb què treballa l'IDSS.

La mineria de regles s'ha realitzat, majoritàriament, basant-se en el coneixement resumit a la Taula 3.1. Tot i això, s'han elaborat algunes de sentit comú i que s'han trobat necessàries, les quals es podrien validar fàcilment amb un expert de l'àmbit. Per exemple, la puntuació de nivell de risc s'obté amb la ràtio reproductiva bàsica (R_t), la indicència acumulada (IA), l'ocupació hospitalària (θ), la taxa de mortalitat (π) i el *Positivity Rate*; seguint les següents regles:

Listing 5.1: Exemple de regles difuses pel nivell de risc

```
if Rt > 1.5 and Incidencia_Acum > 150 and hosp_occ > 80 (%): risc="Critic"
elif Rt > 1.2 and (IA > 100 or idx_mort > 1.5 (%)) and hosp_occ > 70 (%): risc="Alt"
elif IA > 50 and Positivity_Rate > 5 (%): risc="Moderat"
else: risc="Baix"
```

D'aquesta manera, es defineixen un conjunt de regles orientades a mesurar els riscos i les necessitats (e.g. vacunes o transferència de llits). Evidentment, no només es tenen en compte els registres diaris actuals, ja que el sistema perdria la seva raó de ser. Per contra, aprofitant l'horitzó permès pels models de Prophet, es consideraran en les regles les prediccions de casos, defuncions i hospitalitzacions a 7 dies vista (i.e. la sortida dels models).

D'altra banda, també es tindran com a *input* del model de coneixement totes les variables característiques dels estats, ja que aporten informació clau per prendre decisions més precises. Per exemple, es pot augmentar el nivell de risc d'un estat en funció del seu percentatge de població major a 65 anys.

Quant a la sortida d'aquest últim model de coneixement, les regles dispararan una sèrie d'etiquetes qualitatives, com ara ALERTA_Saturació, acompanyades de les seves condicions d'activació: $\theta_{pred_{7d}} > 90\% \text{ and } beds_per_capita < 1.5\%$. D'aquesta manera, es podrà integrar tota aquesta informació en un informe breu, augmentant el nivell d'explicabilitat del sistema.

Representació gràfica dels models

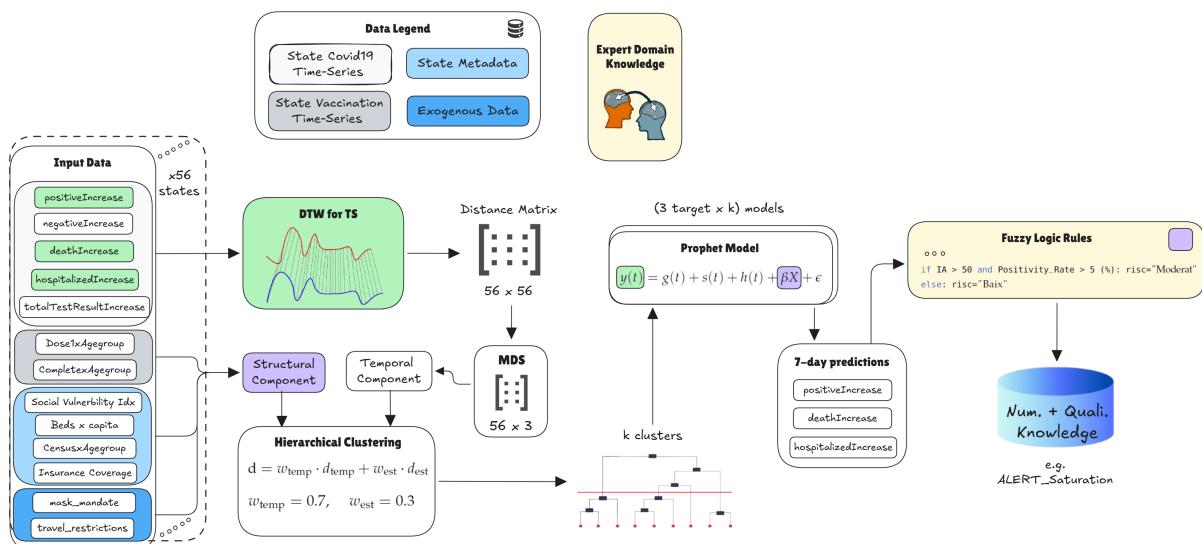


Figure 5.1: Flux de dades pels models del sistema IDSS.

5.1.3 Dades d'Output

Com s'ha mencionat anteriorment, el sistema genera recomanacions a partir de les prediccions temporals del model de dades i regles ben fonamentades.

En particular, la sortida de l'IDSS consisteix en etiquetes que serviran per articular els tres diferents d'outputs dissenyats per donar suport a les decisions que ha d'efectuar el nostre IDDS:

Primerament, unes prediccions numèriques, relacionades amb els valors diaris esperats amb intervals de confiança (casos, hospitalitzacions), que permetran avaluar el risc de confinament de cada estat amb una puntuació juntament diagnòstic qualitatiu extret de les etiquetes que trauran els models definits per la combinació de diverses regles, que indicarà la recomanació a seguir en cada cas (normalitat, confinament parcial o total). (e.g "Risc del 85% a Florida per: (1) Ocupació hospitalària prevista del 80%, (2) Taxa de transmissió ($R_t=1.4$)" "Confinament total"). Per donar suport al pla de distribució de vacunes s'utilitzaran els nivells de risc de cada estat per realitzar un rànquing general de prioritat que especifiqui el nombre de vacunes per estat, utilitzant l'etiqueta del risc de cada estat per ordenar-los. Finalment, per fer l'optimització de la distribució de llits i persones de risc en els centres hospitalaris es farà ús de l'índex de pressió hospitalària i risc de saturació per poder donar una recomanació de transferències de pacient entre estats veïns.

Finalment, s'acompanyaran els resultats amb mapes de calor per visualitzar les regions dels EUA més crítiques en l'evolució de la pandèmia, així com un petit informe explicatiu de quines regles s'han activat per prendre les recomanacions acompanyat de simulacions de l'impacte esperat de cada decisió (ex: "Si Florida rep 10.000 vacunes addicionals, es reduiran les hospitalitzacions un 15% en 14 dies").

5.1.4 Ús en Producció

Una vegada el sistema entri en producció, serà important determinar com s'haurà de fer servir, ja que el procés serà diferent respecte a l'entrenament. En primer lloc, la ingestió de dades ja no podrà ser *off-line*, sinó que els estats hauran d'aportar diàriament les seves dades de la Covid a través d'una API, a la qual es connectarà l'IDSS. Això, obre la porta a fer un reentrenament periòdic dels models, en aquest cas, cada setmana de producció, ja que el cost és petit. En vista que les pandèmies donen lloc a situacions excepcionals, és important fer un *fine-tuning* de les tendències lineals i, sobretot, estacionals, perquè el model inicial pot no respondre de forma efectiva.

En aquest punt, el model per cada estat genera previsions per a cada dia dins d'un horitzó especificat que, inicialment, serà de 7 dies, donat que l'accionament de mesures de la Covid necessita prou previsió. La predicción més immediata i, per tant, amb menys error acumulat, serveix per detectar anomalies, mentre que les altres tenen un impacte en el càlcul del risc de confinament. Seguidament, les prediccions s'envien a un motor de regles adaptatives amb

coneixement de l'expert que calcula el risc de confinament. Per exemple, si les hospitalitzacions previstes superen el 80% de la capacitat ajustada per llits disponibles, el risc s'incrementa.

5.2 INTEGRACIÓ I ESTRUCTURA

El flux de dades del sistema comença amb l'entrada de les dades diàries de la pandèmia per a un estat particular. Aquestes es preprocessen per generar variables rellevants en el domini, com la ràtio de reproducció del virus R_0 , i posteriorment s'incorporen al procés d'anàlisi.

En aquest mòdul d'anàlisi de dades, es combinen diverses estratègies per extreure el màxim potencial predictiu i interpretatiu. D'una banda, s'estudien models estadístics de sèries temporals, com ARIMA, el Filtre de Kalman o Prophet – un model de regressió additiu desenvolupat per *Meta* que permet descompondre la sèrie en tendència, estacionalitat, efectes de festius i regressors externs. D'altra banda, també es tenen en compte xarxes neuronals com LSTM o Transformers, amb arquitectura *encoder-decoder*, capaces de capturar patrons complexos i no lineals mitjançant embeddings compartits entre estats. Tanmateix, la manca de dades i la necessitat d'interpretabilitat van motivar la decisió d'adoptar models estadístics, amb un enfocament particular: entrenar models Prophet per a cada estat o grup d'estats.

Donat que entrenar un model individual per a cada combinació d'estat i *target* (casos, hospitalitzacions, defuncions) implicaria un cost computacional elevat, s'incorpora un procés de *clustering* entre estats amb dinàmiques temporals i característiques estructurals similars. Això permet reduir la càrrega computacional i, alhora, augmentar la robustesa del model mitjançant la transferència d'informació entre territoris comparables. Les etiquetes resultants d'aquest agrupament es fan servir per construir models compartits entre membres del mateix clúster, mantenint un equilibri entre eficiència i especificitat.

Un cop entrenats els models Prophet, aquests proporcionen prediccions quantitatives a 7 dies vista per a cada *target*, així com una descomposició de la influència de cada component i variable externa. Aquesta sortida es complementa amb un model de coneixement basat en regles de lògica difusa, especialment dissenyat per tractar la incertesa i subjectivitat en les decisions de salut pública. Aquest model utilitza tant dades actuals com prediccions futures dels models de sèries temporals per generar alertes qualitatives, com ara ALERTA_Saturació o risc epidemiològic (e.g. "Alt", "Crític"), a partir de condicions que tenen en compte índexs com el R_t , la incidència acumulada, la taxa de mortalitat o l'ocupació hospitalària. A més, s'inclouen factors estructurals dels estats (e.g. percentatge de població gran, cobertura sanitària), per tal d'ajustar les regles a la realitat local.

Així, la integració entre els models de dades i els models de coneixement permet oferir una doble capa d'anàlisi: d'una banda, prediccions numèriques precises amb intervals de confiança; de l'altra, interpretacions qualitatives i accionables. Aquesta combinació es sintetitza en un informe per a cada estat, que es posa a disposició dels responsables de decisió a través d'una API. Aquests poden consultar-ho via interfície gràfica (vegeu la Sec. 9), i aportar el seu *feedback*,

com ara la taxa d'encert percebuda o la validesa de les alertes generades. Aquest retorn tanca el cicle del sistema, permetent revisar i millorar els models (tant de dades com de coneixement) mitjançant processos com el reentrenament selectiu o l'ajust de regles difuses, assegurant una resposta adaptativa i evolutiva al context de la pandèmia.

5.2.1 *Inclusive Design*

Un disseny inclusiu del sistema assegura que tothom pugui accedir i interpretar la informació sense obstacles. Si bé el sistema proposat ja integra decisions inclusives, mitjançant dades segmentades per rangs d'edat que faciliten l'anàlisi específica de l'impacte en diferents poblacions, el disseny de la interfície és crític en aquest aspecte. Per exemple, es podria incorporar un mòdul d'assistència per veu que converteixi automàticament els informes en àudio, facilitant la comprensió per a usuaris amb discapacitats visuals o dificultats de lectura. També, una configuració personalitzable del contingut amb ajustaments de contrast, mida de lletra o altres paràmetres visuals, podria ajudar a usuaris amb problemes de visió.

5.3 POLÍTIQUES OPERATIVES

5.3.1 *Monitorització del Sistema*

La monitorització contínua del sistema mitjançant **indicadors clau de rendiment (KPI)** permet avaluar-ne l'eficàcia i l'impacte en temps real. A continuació, es presenten els KPI proposats per al sistema:

- **Precisió en la predicció:**

- **Descripció:** Mesura la capacitat del model per predir correctament l'evolució de la pandèmia després de la seva implementació.
- **Obtenció** Comparació entre els valors estimats pel model i les dades observades el dia següent.
- **Justificació:** El model predictiu és l'eix central del sistema; prediccions precises afavoreixen recomanacions més efectives i fiables.

- **Taxa de saturació hospitalària evitada**

- **Descripció:** Avalua l'impacte de les mesures aplicades en la reducció de la saturació dels centres hospitalaris.
- **Obtenció:** Comparació entre les prediccions inicials (sense restriccions) i les dades reals després de l'aplicació de les mesures.
- **Justificació:** La saturació hospitalària està estretament relacionada amb la mortalitat i l'evolució general de la pandèmia; reduir-la és clau per a la sostenibilitat del sistema sanitari.

- **Compliment de les recomanacions**

- **Descripció:** Percentatge de recomanacions emeses pel sistema que han estat adoptades per les autoritats.
- **Obtenció:** Càlcul en temps real del nombre de mesures efectivament aplicades en relació amb el total recomanat.
- **Justificació:** Permet identificar si el sistema proposa accions realment viables o, al contrari, poc aplicables en el context real.

- **Usabilitat percebuda**

- **Descripció:** Grau de satisfacció i percepció d'utilitat del sistema per part dels usuaris en situacions de pandèmia.
- **Obtenció:** Mitjançant enquestes de satisfacció dirigides als usuaris finals.
- **Justificació:** El sistema ha de ser intuïtiu, accessible i eficaç, especialment en situacions crítiques; una bona usabilitat afavoreix l'adopció i la confiança en les seves recomanacions.

Una altra part molt important del procés de desenvolupament de l'IDSS és el disseny de l'estratègia de manteniment. En el cas de la gestió de la pandèmia, els models de sèries temporals són susceptibles de *concept drift*, en quant que deixen d'ajustar-se correctament a la realitat. D'altra banda, el coneixement de domini i, sobretot, els *thresholds* per computar les regles de lògica difusa, poden no funcionar en situacions específiques o excepcions, la qual cosa també s'ha de comunicar.

En aquest sentit, la proposta és incloure un *feedback* per part del *decision-maker* o usuari. En particular, s'haurà d'aportar el valor real de la prediccio, per exemple, el nombre de casos realment reportats un dia específic. Així, el sistema podrà detectar quan els models queden degradats (i.e. amb un llindar) i, consegüentment, reentrenar automàticament el model de TS amb les noves dades, una pràctica coneguda com a *batch-learning*. Quant al *knowledge drift*, s'haurà de realitzar una revisió periòdica de les regles de lògica difusa, en què en funció de l'evidència s'ajustin els resultats qualitatius del sistema (e.g. *High Transmission* pot variar entre regions). Aquest mecanisme de *feedback* permetrà no només refinir el sistema, sinó també mesurar el seguiment i l'adequació de les seves recomanacions.

5.3.2 Comunicació dels Resultats

La comunicació efectiva és un pas fonamental per garantir el correcte funcionament del nostre IDDS. Donat que el sistema està dissenyat per interactuar amb professionals de la sanitat i gestors, a presentació de la informació ha de ser clara, intuïtiva i visual, facilitant la interpretació ràpida i la presa de decisions informades.

Els resultats es presentaran a través d'una plataforma interactiva que permetrà als usuaris explorar les dades de forma dinàmica i personalitzada. Aquesta interfície inclourà:

- **Visualitzacions Gràfiques Dinàmiques:**

- Mapes de calor per identificar ràpidament zones amb major risc. Per exemple, el que podem veure a la imatge [5.2](#).
 - Gràfics evolutius per mostrar tendències temporals
 - Rànquings comparatius entre els diferents estats, així com el seu repartiment de vacunes
- **Navegació Jeràrquica i Detallada:**
 - L'usuari podrà fer clic sobre un estat o regió específica i accedir a un informe diari resumit, amb les dades més rellevants i alertes prioritàries. Per exemple, com el que podem veure a la imatge [5.3](#) on es mostra la situació de Colorado.
 - S'inclourà un registre de logs per permetre el seguiment de les decisions preses.
 - **Panell de Control Personalitzable:**
 - Els gestors podran configurar els indicadors clau (KPIs) que vulguin monitorar en temps real.
 - Opció de descarregar els reports en format PDF/ Excel per la seva anàlisi més extensa.
 - **Interacció amb l'usuari:**
 - Alertes intel·ligents per a quan el sistema detecti anomalies o canvis significatius en indicadors clau.
 - Assistència contextual i eines d'ajuda per guiar l'usuari en la interpretació de les dades

Aquest enfocament visual, interactiu i centrat en l'usuari permet garantir que el IDSS no només proporcioni dades, sinó que els transformi en coneixement accionable, millorant d'aquesta manera l'eficiència i la qualitat de les decisions.

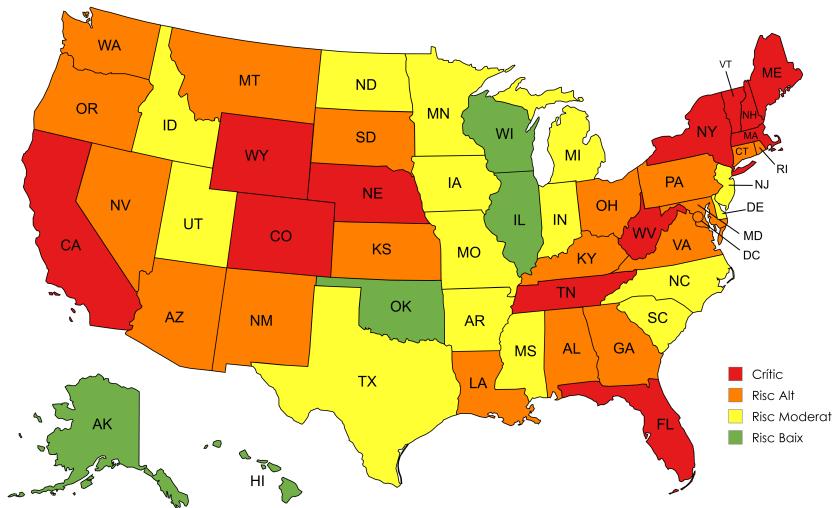


Figure 5.2: Exemple mapa de calor.

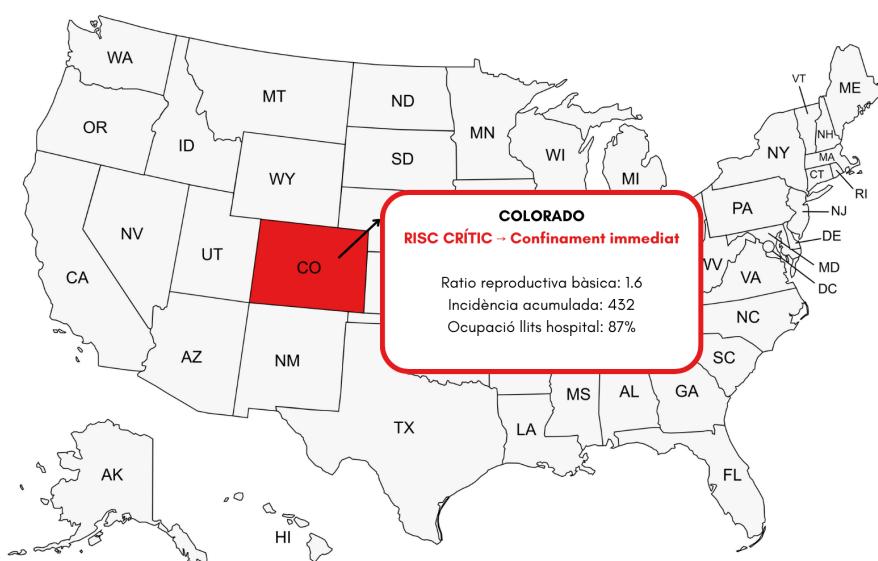


Figure 5.3: Exemple informació quan cliques un estat

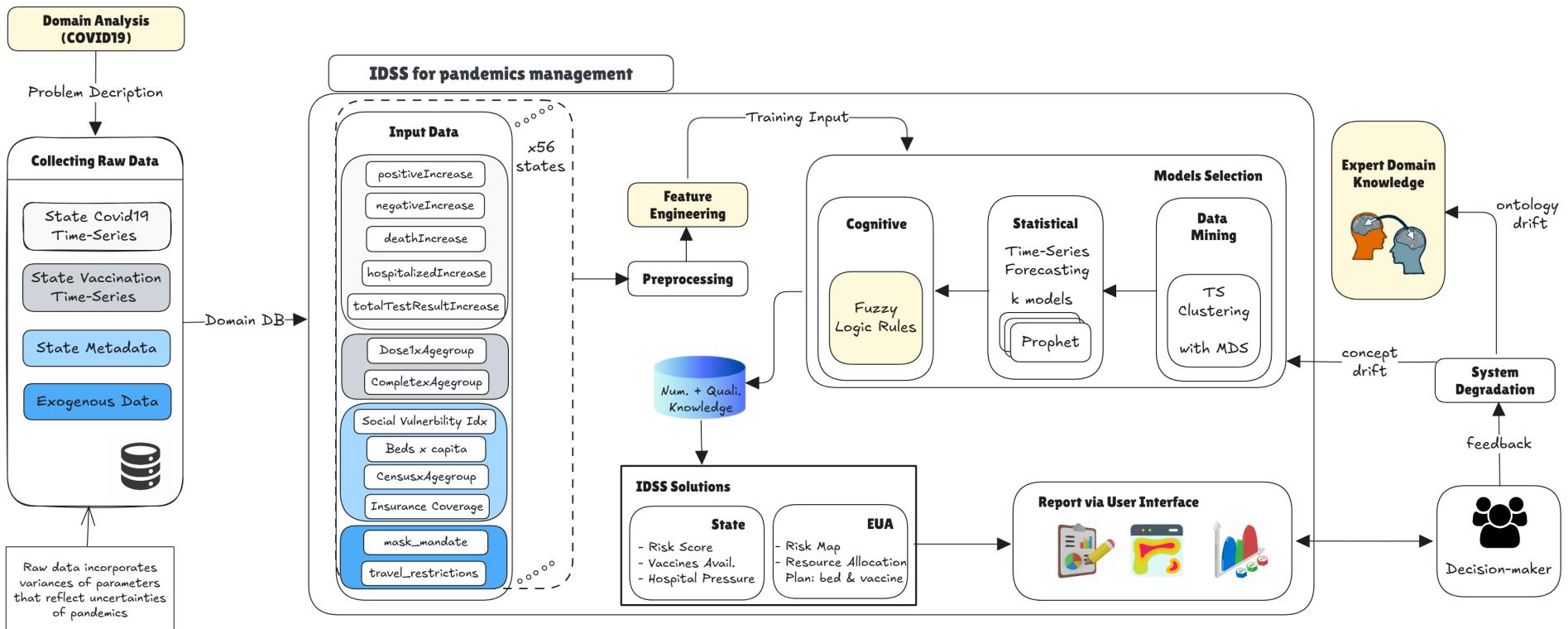


Figure 5.4: Arquitectura del Sistema Completo d'IDSS.

TIPUS D'USUARI

El nostre *Intelligent Decision Support System* (IDSS) està dissenyat per interactuar amb diferents perfils d'usuaris que tenen rols específics en la gestió de la pandèmia. Aquests usuaris inclouen autoritats sanitàries estatals, gestors d'hospitals i responsables de coordinació de recursos. A continuació, es detallen els perfils, les seves necessitats i com el sistema satisfarà aquestes necessitats:

AUTORITATS SANITÀRIES ESTATALS

- **Perfil:** Inclou responsables governamentals i tècnics de salut pública a nivell estatal. Aquestes autoritats són les encarregades de prendre decisions sobre la implementació de polítiques sanitàries i restriccions poblacionals.
- **Necessitats:**
 - Analitzar l'evolució de la pandèmia a l'estat en temps real.
 - Identificar els punts crítics amb major propagació del virus.
 - Obtenir recomanacions òptimes per implementar restriccions efectives, com ara confinaments selectius o limitacions d'aforament.
- **Com el sistema els proporciona suport:**
 - L'IDSS ofereix un panell de control interactiu amb dades actualitzades diàriament que mostren l'evolució dels casos, les taxes de reproducció del virus i altres indicadors clau.
 - Proporciona recomanacions basades en dades sobre les millors accions a emprendre, ajustant les estratègies en funció de la situació de cada estat.
 - Permet simular diferents escenaris per anticipar l'impacte de les mesures abans d'implementar-les.
- **Etiquetes de rol:** *Client, Gestor estratègic, Prescriptor de polítiques.*

GESTORS D'HOSPITALS

- **Perfil:** Inclou directors i coordinadors d'hospitals responsables de gestionar els recursos sanitaris i garantir una resposta efectiva davant l'increment de pacients.
- **Necessitats:**
 - Garantir la distribució òptima dels pacients als llits disponibles.
 - Evitar la saturació d'hospitals concrets mitjançant la derivació coordinada de pacients.
 - Accedir a projeccions sobre l'ús de recursos mèdics, com oxigen o personal sanitari, en funció de l'evolució prevista de la pandèmia.
- **Com el sistema els proporciona suport:**
 - L'IDSS procura optimitzar la distribució més eficient dels pacients en funció de la capacitat de cada hospital.
 - Proporciona avisos preventius quan es detecta una saturació imminent en un hospital, suggerint la reubicació de pacients a centres amb més capacitat.
 - Ofereix eines per planificar millor la disponibilitat de recursos i preparar-se per futurs escenaris crítics.
- **Etiquetes de rol:** *Client, Executor local, Introducció de dades.*

RESPONSABLES DE COORDINACIÓ DE RECURSOS INTER-ESTATALS

- **Perfil:** Són responsables de garantir que hi hagi una coordinació efectiva entre diferents estats, especialment en situacions on els recursos d'un estat són insuficients per atendre la seva població.
- **Necessitats:**
 - Monitoritzar les necessitats i la capacitat de cada estat per prendre decisions de redistribució de recursos.
 - Reduir els desajustos regionals, com la manca de llits hospitalaris en zones altament afectades.
- **Com el sistema els proporciona suport:**
 - L'IDSS ofereix informació consolidada dels diferents estats, permetent visualitzar les necessitats i els recursos disponibles a nivell regional i nacional.
 - Genera alertes per a situacions de desequilibri, com un augment desproporcionat de casos en un estat que supera la seva capacitat hospitalària.
 - Permet establir estratègies de redistribució basades en criteris objectius, millorant la coordinació entre estats.
- **Etiquetes de rol:** *Client, Coordinador interterritorial, Supervisor de recursos*

SELECCIÓ D'EINES DE SOFTWARE

La implementació del nostre Intelligent Decision Support System (IDSS) ha requerit l'ús d'un conjunt d'eines de programari seleccionades estratègicament per garantir una gestió òptima de les dades, la seva anàlisi, la creació de models predictius i la visualització de resultats. Cada eina s'ha escollit en funció de les seves capacitats específiques i la seva adequació a les necessitats del projecte. A continuació es detallen les eines utilitzades i es justifica la seva selecció:

PANDAS (PYTHON)

Descripció: Python és un llenguatge de programació versàtil i àmpliament utilitzat en ciència de dades. Pandas és una llibreria específica per a la manipulació i l'anàlisi de dades estructurades, com ara taules i sèries temporals.

Justificació:

- Ès ideal per al preprocessing de dades, ja que permet netejar, transformar i combinar grans volums d'informació de manera eficient.
- La seva sintaxi intuitiva i la seva integració amb altres llibreries fan que sigui una opció robusta per treballar amb les dades de la pandèmia, que provenen de múltiples fonts i necessiten ser harmonitzades per al seu ús.

RAPIDMINER

Descripció: RapidMiner és una plataforma d'anàlisi de dades que ofereix eines d'extracció, transformació i càrrega (ETL), així com funcionalitats avançades per a l'aprenentatge automàtic.

Justificació:

- En casos específics, s'ha utilitzat per a la seva capacitat de processar dades sense necessitat de programar, accelerant així tasques puntuals de preprocessing o modelatge.
- Ha resultat útil en l'etapa inicial d'exploració de dades i per validar ràpidament hipòtesis amb fluxos de treball visuals.

PROPHET (PYTHON)

Descripció: Prophet és una llibreria desenvolupada per Meta que està dissenyada per a la prediccio de sèries temporals.

Justificació:

- Ès una eina potent i fàcil d'usar per generar prediccions sobre l'evolució dels casos de COVID-19 en els diferents estats.
- La seva capacitat per gestionar patrons estacionals i la seva precisió en dades amb tendències canviants l'han fet indispensable per modelar l'evolució de la pandèmia i proporcionar recomanacions basades en aquestes prediccions.

PROTEGÉ

Descripció: Protegé és una plataforma per al desenvolupament d'ontologies, que permet estructurar i representar el coneixement de manera formal.

Justificació:

- S'ha utilitzat per crear una ontologia que defineix les relacions entre les variables del sistema (com casos, llits hospitalaris disponibles, i restriccions poblacionals).
- La integració de lògica difusa ha permès gestionar la incertesa inherent a la presa de decisions durant la pandèmia, oferint recomanacions adaptades a situacions amb informació incompleta o imprecisa.

RSTUDIO

Descripció: RStudio és un entorn de desenvolupament integrat per al llenguatge R, àmpliament utilitzat per a l'anàlisi estadística i la visualització de dades.

Justificació:

- S'ha utilitzat per desenvolupar scripts auxiliars orientats a l'anàlisi estadística avançada i la creació de gràfics personalitzats.
- La capacitat de R per generar visualitzacions detallades i atractives ha estat fonamental per comunicar els resultats de manera clara i comprensible als usuaris finals.

PREPROCESSAMENT DE DADES

8.1 METODOLOGIA

La qualitat i la consistència de les dades són pilars fonamentals per al correcte funcionament i l'eficàcia del Sistema Intel·ligent de Suport a la Decisió (IDDS). En aquest sentit, el preprocessament de dades emergeix com una etapa crítica, destinada a transformar les dades brutes en un format adequat per a l'anàlisi i la modelització. Aquest procés no és un pas aïllat, sinó una fase integral que es duu a terme tant amb les dades històriques utilitzades per a l'entrenament dels models, com amb les dades en temps real que alimentaran el sistema en producció.

La consistència en l'aplicació del preprocessament entre ambdues fases (entrenament i producció) és de vital importància. Qualsevol divergència en les transformacions aplicades podria generar discrepàncies significatives en el rendiment dels models quan s'enfronten a dades reals, comprometent la fiabilitat de les prediccions i les recomanacions del sistema. Per tant, la metodologia de preprocessament que es detalla a continuació ha estat dissenyada per ser aplicada de manera uniforme en ambdós contextos.

Abans d'implementar les tècniques específiques de preprocessament que es descriuen en les seccions següents, s'ha dut a terme una anàlisi exhaustiva de les característiques de cada conjunt de dades. Aquesta anàlisi ha inclòs la visualització mitjançant gràfics per explorar la distribució, les relacions i la presència de valors atípics de les variables. Paral·lelament, s'ha realitzat un estudi d'estadística descriptiva per quantificar les propietats fonamentals de cada variable, com la tendència central, la dispersió i la presència de valors mancats. Aquesta fase exploratòria ha estat essencial per fonamentar les decisions sobre quines tècniques de preprocessament aplicar a cada conjunt de dades, amb l'objectiu d'optimitzar la seva qualitat i la seva utilitat per al IDDS.

8.2 TÈCNIQUES APLICADES DE PREPROCESSAMENT

Estat de la Covid-19

Un primer conjunt de dades reflecteix l'estat de la COVID-19 (infeccions, hospitalitzacions, casos positius...) amb una granularitat estatal. No obstant això, no tota la informació recopilada és útil per a l'anàlisi, principalment per tres motius:

1. Algunes característiques tenen molt pocs valors únics i no aporten informació rellevant.
2. Altres característiques corresponen a metadades, que no formen part de la matriu de treball.
3. Hi ha característiques presents en alguns estats però absents en altres.

En aquests casos, es proposa un filtratge de columnes per eliminar les dades innecessàries. A continuació, les dades passen per un procés de correcció de format i normalització. En primer lloc, s'estandarditza el format de la data a %Y%M%D i s'ajusten els tipus de les diferents columnes. En el cas de les dades en temps real, les noves instàncies s'afegeixen al final de la matriu de treball, ja que es tracta de sèries temporals.

Un cop normalitzades les dades, es caracteritzen els valors mancants a nivell estatal. Els resultats mostren dos tipus principals de dades absents:

- **MAR (Missing At Random):** Es tracta de dades no reportades en determinats dies pels estats. Tot i que el patró de valors absents sembla aleatori, no hi ha prou evidència per considerar-lo completament aleatori.
- **MNAR (Missing Not At Random):** Aquest tipus de valors absents es troba exclusivament en dades històriques. Es deu a informació que inicialment no es recollia durant la pandèmia, sinó que es va requerir posteriorment.

Per gestionar aquests casos:

- En el cas **MAR**, es proposa una interpolació lineal cap endavant per mantenir la coherència temporal.
- En el cas **MNAR**, es decideix omplir els valors mancants amb zeros. Tot i que aquesta estratègia no és ideal, la manca d'informació impedeix una imputació més precisa. A més, ja que es localitzen només a l'inici, seran fàcils de localitzar. No obstant això, el model de dades accepta valors mancants, per la qual cosa la imputació amb zeros només s'aplica en la fase d'extracció de dades. En les prediccions, els valors es mantindran com a desconeguts.

Les figura 8.1 mostra el preprocessament d'aquesta font de dades en el cas de les dades en streaming o les dades històriques.

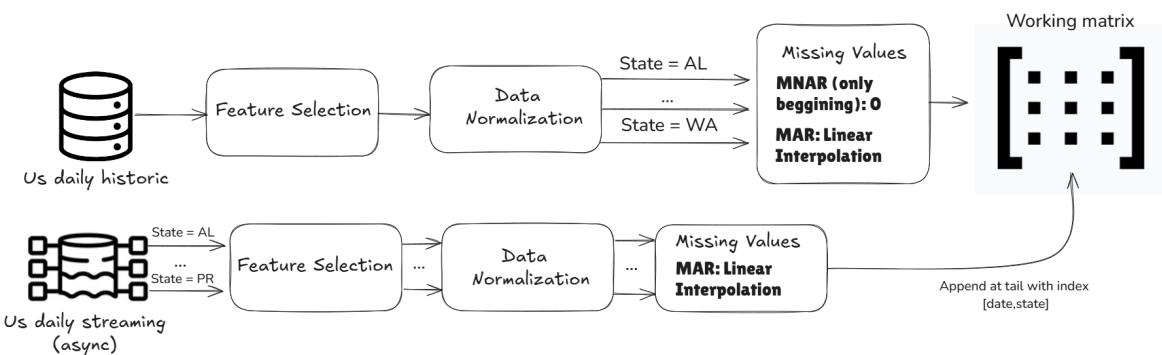


Figure 8.1: Preprocessing proposat per la font de dades us_daily. Adalt es mostra el preprocessament de les dades històriques, abaix les dades que provenen en streaming.

Població i grups d'edat

La incorporació d'estadístiques demogràfiques, com la densitat de població i la distribució per grups d'edat, enriqueix el IDSS amb informació rellevant per a l'anàlisi de l'evolució de la pandèmia i la monitorització de diferents col·lectius. En aquest cas, però, el preprocés és trivial.

Com a pas previ, es filtren les columnes de metadades que no són necessàries per a la matriu de treball. A més, totes les columnes categòriques es transformen en valors numèrics (per exemple, el grup d'edat "Age group 0 to 7" es discretitza com a 0).

Cal destacar que aquestes dades tenen una naturalesa estàtica i no es rebran en format streaming. Un cop processades, s'integren en una de les matrius de treball del sistema.

Comptats, llits i assegurança

Per a tots els conjunts de dades seguim el mateix pas principal, filtrar les columnes sense rellevància.

El conjunt de dades de "counties" conté informació sobre la població i densitat de població dels diferents comtats dels Estats Units. En aquest cas trobem 78 valors nuls, que es corresponen amb comtats de Puerto Rico. Per solucionar-ho, hem recorregut a dades externes del Cens de 2020 per obtenir la densitat de població d'aquests municipis. Hem creat un diccionari amb la densitat corresponent i hem actualitzat els valors nuls al dataframe.

El conjunt de dades de "beds" conté informació sobre el nombre de llits hospitalaris disponibles als diferents estats dels Estats Units, classificats segons la seva titularitat: govern local, organitzacions sense ànim de lucre i entitats amb ànim de lucre. El nombre de llits ja està escalat del 0 al 5. Pels valors nuls de "state_local_government" i "for_profit", hem optat per omplir-los amb 0, assumint que en aquests casos no hi ha llits disponibles sota aquesta categoria. Aquesta decisió es justifica perquè la quantitat de dades faltants és baixa i no afecta el càlcul de la suma total de llits.

El conjunt de dades de "insurance" conté informació sobre la cobertura d'assegurances mèdiques als diferents estats dels Estats Units, incloent-hi estimacions sobre la població coberta per assegurances privades, públiques o sense cobertura. Trobem valors nuls a les columnes "labor_force" i "employed", com que hi ha un 52,94% a "employed" hem decidit eliminar aquesta columna. D'altra banda, "labor_force" presenta un 29,41% de valors nuls, per això hem creat una nova categoria anomenada "Unknown". Això evita esbiaixar les dades en cas d'eliminar les files incompletes. Un cop fet això, utilitzem One-Hot Encoding per convertir les variables categòriques en variables numèriques binàries. Aquest procés crea una nova columna per a cada categoria, assignant 1 si la fila pertany a la categoria i 0 si no. Per fer les variables més comprensibles, reanomenem les columnes resultants del One-Hot Encoding.

Vacunació

Comencem per netejar el dataset eliminant aquelles columnes que estan buides o tenen molts missings. També borrem les files que tenen molts valors perduts o que tenen "Unknown" a la variable County. Procedim modificant el nom de les variables i d'alguns valors per tal de facilitar la interpretació de les dades. A més, ajustem el tipus (type) d'algunes variables.

Les dades d'aquest dataset són temporals i moltes de les variables contenen dades acumulatives. Per tant, per a un mateix estat i comtat, el nombre de persones vacunades ha de ser igual

o major a les persones vacunades del dia anterior. Busquem en el dataset possibles errors que no segueixin aquesta norma i els corregim. Els següents gràfics ens ajuden a trobar a visualitzar els errors:

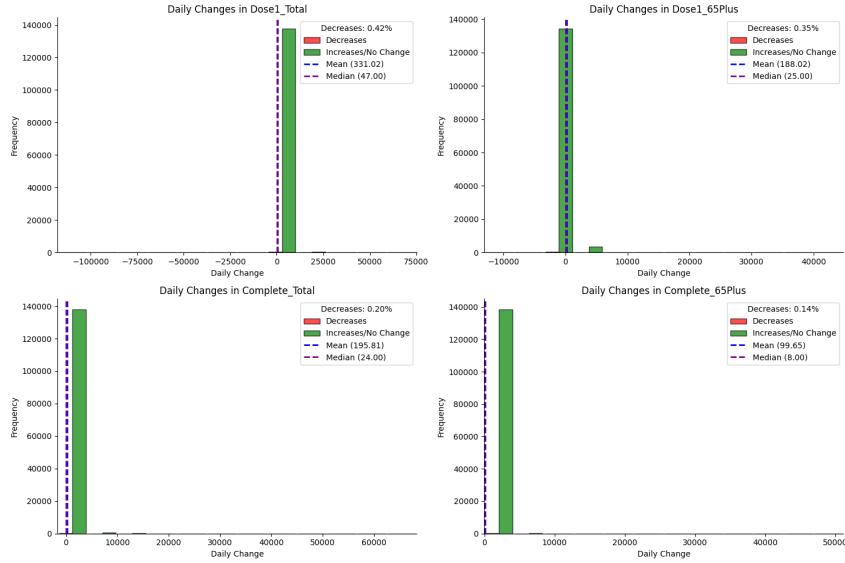


Figure 8.2: Anàlisi variables acumulatives

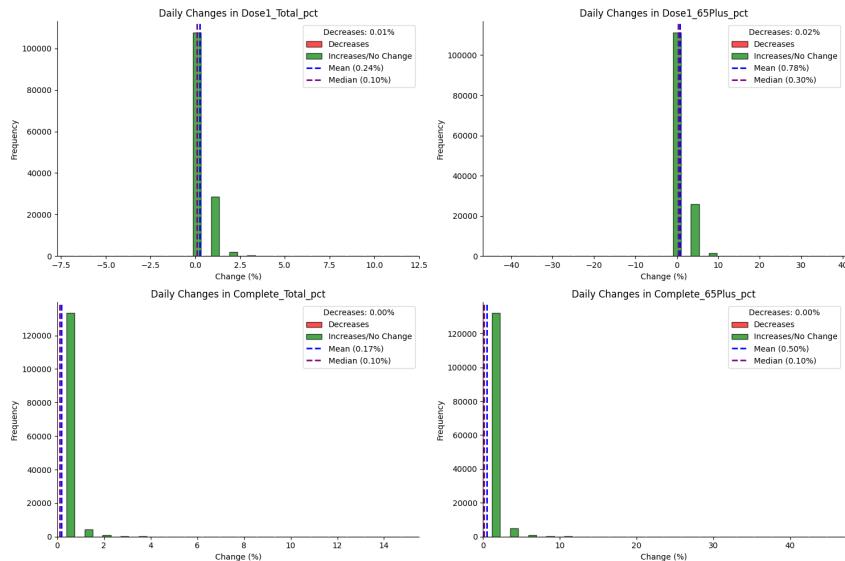


Figure 8.3: Anàlisi variables de percentatge acumulatives

A continuació, mirem els outliers, però arribem a la conclusió que els valors extrems trobats no són outliers, sinó punts rellevants que poden portar a conclusions importants. Per detectar-los hem utilitzat Rolling IQR, un valor es considerarà outlier si és molt diferent que els valors dels 7 dies anteriors i si supera el 2.5% de la població, ja que, sabem que en cap moment es va vacunar tanta gent en un sol dia. Veiem un exemple representat en el gràfic 8.4, on en blau hi ha l'acumulació de vacunes, en taronja les vacunacions diaries, en lila el límit per ser considerat outlier respecte els 7 dies anteriors i en verd el límit del 2.5% que podem veure que mai se supera.

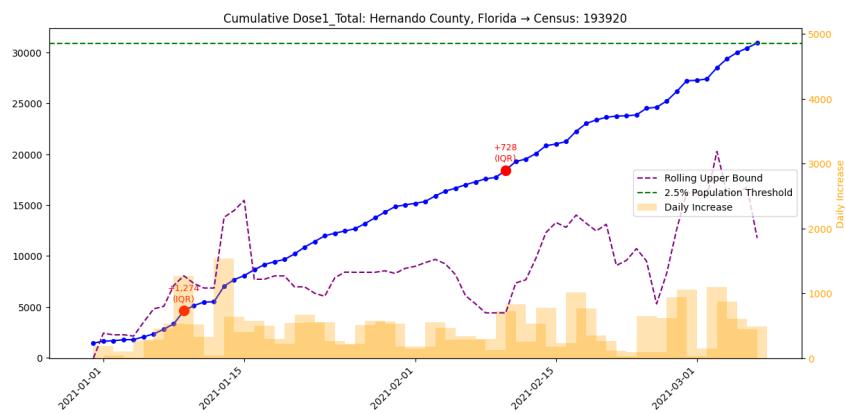


Figure 8.4: Anàlisi outliers Florida

També estudiem els outliers multivariants fent servir la distància de mahalanobis, però observem que corresponen a estats molt més poblats que la resta, motiu pel qual decidim no eliminar-los.

Pel que fa a la imputació missing values, hem utilitzat diferents mètodes segons la variable. Per a les variables que indiquen el nombre de persones vacunades, imputem els valors mancats amb el valor del dia anterior dins del mateix estat i comtat. Si no hi ha dades del dia anterior, assignem un 0. Un cop imputats, recalculem les variables de percentatge corresponents. Els missings de SVI_CTGY es troben tots en un mateix comtat, per la qual cosa imputem la moda d'aquell estat. Finalment, la variable Metro_status també només té missings a un comtat, per tant, busquem el valor per Internet i l'imputem.

Les dades estan per comtats, però a nosaltres ens interessa tenir-les per estats, per tant les agreguem. Per fer-ho, haurem de modificar el format d'algunes variables. La variable "Metro_status" es reconverteix per obtenir, per a cada estat, el percentatge de la població que viu en zones metropolitanes (Metro) i en zones no metropolitanes (Non-metro). El mateix passa amb "SVI_CTGY", que es transforma en quatre columnes indicant el percentatge de població de cada estat en les categories "Low", "Moderate-Low", "Moderate-High" i "High".

La variable "Census2019" s'agrupa per obtenir el cens de cada estat, i s'afegeix una nova variable "Census2019_65Plus", que indica la població major de 65 anys per estat. Per acabar, les variables "Dose1_Total", "Dose1_65Plus", "Complete_Total" i "Complete_65Plus" les sumem ajuntant-les per estat i data i les respectives columnes de percentatge es recalculen utilitzant les variables de "Census".

8.3 ÚS DE RAPIDMINER

Pel que fa al preprocesat de les dades referents al nombre de llits hospitalaris per càpita als Estats Units (dataset beds), també hem volgut implementar un script en *RapidMiner* que automatitza la lectura, imputa els valors perduts, elimina, reanomena i recomputa variables (ja que s'han detectat alguns errors a la variable 'total', que simplement es tracta de la suma

de tres variables diferents). La Figura 8.5 mostra el diagrama visual del procés dissenyat. Tot seguit, es descriu detalladament cadascuna de les etapes:

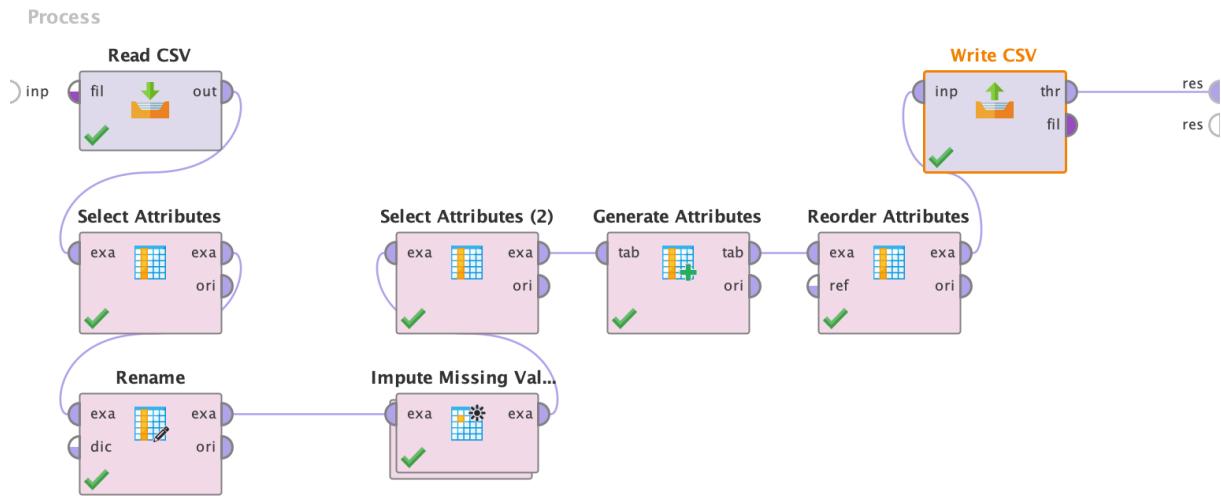


Figure 8.5: Disseny del script en RapidMiner

1. Lectura de dades (Read CSV)

El primer operador llegeix el fitxer `kff_usa_hospital_beds_per_capita_2018.csv`, que com ja hem vist anteriorment conté informació sobre el nombre de llits per 1000 habitants per estat segons el tipus de gestió (govern, organitzacions sense ànim de lucre i sector privat). En aquest primer pas s'especifica el separador de columnes (coma), l'ús de la primera fila com a capçalera, i el format UTF-8. També s'estableix de manera explícita el tipus de cada variable:

- `state, state_name`: variables categòriques (*polynomial*)
- `state_local_government, non_profit,`
- `for_profit, total`: variables numèriques (*real*)

Aquesta configuració assegura que cada atribut es llegeixi amb el tipus correcte per facilitar la manipulació posterior. Cal remarcar que en un principi no especificàvem dins de *Read CSV* de quin tipus era cada variable, sinó que feiem servir operadors que modificaven, per exemple, una variable que estava codificada com a categòrica a numèrica. De la manera en què ho fem ara és molt més senzill i el disseny queda molt més clar.

2. Eliminació de la columna 'State' (Select Attributes)

Un cop llegit el conjunt de dades, s'utilitza l'operador **Select Attributes** per eliminar la columna `state`, ja que es considera redundant. Aquesta variable és un identificador abreujat de l'estat (per exemple, "CA" per Califòrnia), però l'atribut `state_name` conté el nom complet i resulta més informatiu. A més, aquest pas evita problemes de duplicació o confusió en les fases posteriors del procés.

3. Renombrament de la variable 'State_name' a 'State' (Rename)

Per mantenir la coherència amb la resta de datasets del projecte, es renombra la variable `state_name` a `state`. D'aquesta manera, es garanteix l'homogeneïtat de noms de variables claus (com els identificadors d'estat), facilitant la integració posterior d'aquest conjunt amb altres taules (via `join` o `merge`).

4. Imputació de valors perduts (Impute Missing Values)

Un cop detectats els valors perduts, s'aplica una imputació mitjançant un model *k-Nearest Neighbors* (k-NN). Aquesta estratègia permet estimar els valors perduts a partir dels casos més propers segons una distància definida. Els paràmetres clau utilitzats són:

- **`k = 5`**: cada valor es prediu mitjançant la mitjana (ponderada) dels 5 veïns més propers.
- **Weighted vote**: els veïns més propers tenen més pes en la predicció.
- **MixedEuclideanDistance**: distància combinada per atributs numèrics i categòrics.
- **Iterate = true**: s'imputa iterativament fins a completar tots els valors.
- **learn_on_complete_cases = true**: només s'entrena el model amb observacions sense valors perduts.

Aquesta imputació resulta adequada per conjunts amb poques variables i nombres moderats de valors perduts, com és el cas del dataset en qüestió.

5. Eliminació de la columna 'Total' (Select Attributes(2))

El dataset original inclou una variable anomenada `total`, que representa el nombre total de llits hospitalaris per 1000 habitants per estat. No obstant això, s'ha detectat que aquest valor sovint no concorda amb la suma de les tres categories desglossades (`state_local_government`, `non_profit` i `for_profit`), donant lloc a inconsistències. Per aquest motiu, s'elimina aquesta variable, ja que es prefereix reconstruir-la de manera coherent un cop imputats els valors perduts en les tres components.

6. Generar la variable 'Total' (Generate Attributes)

Amb les tres variables base ja netes i sense valors perduts, es genera de nou la variable `total` mitjançant l'operador `Generate Attributes`. La nova definició s'estableix com:

```
total = state_local_government + non_profit + for_profit
```

D'aquesta manera, es garanteix la coherència interna del conjunt de dades i s'eviten inconsistències amb valors computats incorrectament en el fitxer original.

7. Reordenació de les columnes (Reorder Attributes)

Per finalitzar el procés de preprocesat, es reordenen les columnes per tal de millorar la llegibilitat del conjunt de dades exportat. L'ordre establert és:

```
state,state_local_government,non_profit,for_profit,total
```

Aquest ordre situa la variable identificadora en primer lloc, seguida de les components del total i, finalment, la nova variable calculada.

8. Exportació de dades (Write CSV)

Finalment, el conjunt de dades complet (amb valors imputats i la variable total correctament reconstruïda) s'exporta mitjançant l'operador Write CSV. El fitxer resultant es guarda amb el nom kff_usa_hospital_beds_per_capita_2018_rapidminer.csv i conté totes les transformacions aplicades durant el procés de preprocesat.

8.4 ESQUEMA GENERAL DEL PREPROCESSAMENT

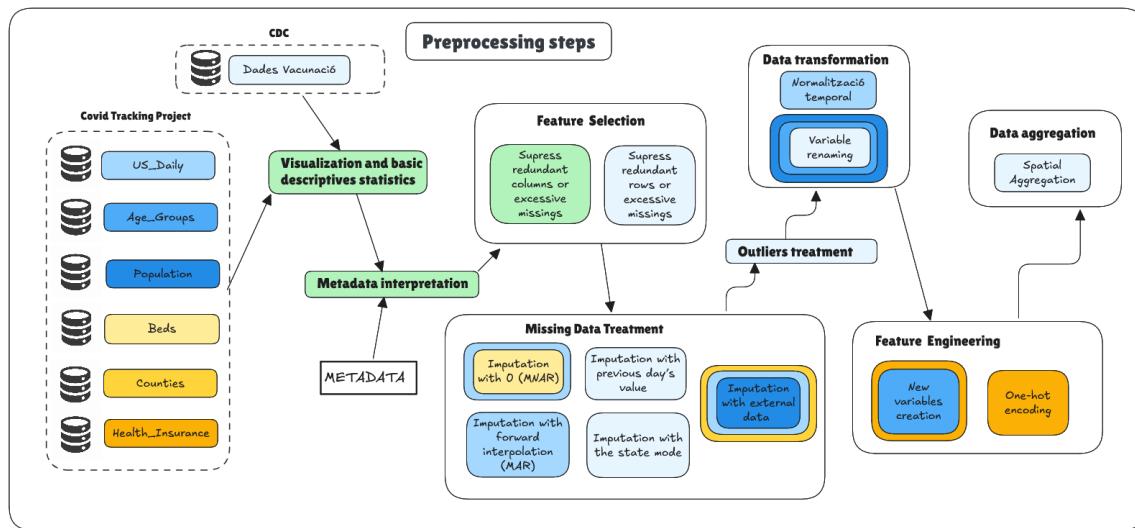


Figure 8.6: Flux de preprocessament dels diferents datasets

RESULTATS DE LA PREPARACIÓ DE DADES

9.1 METADEDES

La gestió exhaustiva de les metadades és fonamental per garantir la comprensió, la traçabilitat i la correcta utilització dels conjunts de dades al llarg de tot el cicle de vida del Sistema Intel·ligent de Suport a la Decisió (IDDS). Aquesta secció presenta una descripció de les característiques principals dels conjunts de dades després d'haver estat sotmesos al procés de preprocessament detallat a la Secció 8.

9.1.1 *Estructura del Document de Metadades*

Per a cada conjunt de dades utilitzat en l'IDDS, es generarà i mantindrà un document de metadades detallat que contindrà les següents seccions principals:

1. Identificació del Dataset

- Nom del dataset
- Font Original de Dades
- Data d'Obtenció
- Versió
- Responsables de la Recollida/Processament Inicial

2. Descripció de les Dades Originals (Abans del Preprocessament)

- Descripció General (basada en la Secció 2.1)
- Nombre Original de Files i Columnes
- Descripció Detallada de Cada Atribut Original

3. Procés de Neteja i Preprocessament Aplicat

- Data del Preprocessament
- Responsable
- Descripció Detallada de Cada Pas

- Tractament de Valors Nuls (tipus, mètode, justificació)
- Transformació de Variables (tipus, justificació, detalls)
- Filtrat de Dades (criteris, justificació)

4. Característiques de les Dades Després del Preprocessament

- Nombre Final de Files i Columnes
- Responsable
- Descripció Detallada de Cada Atribut Final

5.ús de les Metadades al Llarg del Procés

- Explicació detallada de com la informació del document serà utilitzada en cada fase de l'IDDS (Secció 1, Secció 2, Secció 5.1, Secció 5.2, Secció 5.3, manteniment, actualització, reproduïibilitat)

9.1.2 Utilitat de les Metadades per al Nostre IDDS

El document de metadades per a cada conjunt de dades és una peça clau per al correcte funcionament i la sostenibilitat del nostre IDDS. Aporta valor en les següents àrees:

Facilita la comprensió i la interpretació de les dades per a tots els membres de l'equip (científics de dades, desenvolupadors, responsables de la presa de decisions - Secció 6).

Assegura la traçabilitat de les transformacions aplicades, permetent auditjar el procés de preparació de les dades i identificar possibles errors o àrees de millora (relacionat amb la "Monitorització del Sistema" - Secció 5.3.1).

Garanteix la consistència en l'aplicació del preprocessament tant a les dades d'entrenament com a les dades de producció, un requisit essencial per a la fiabilitat dels models (Secció 5.1.4).

Serveix com a guia per a la integració de noves fonts de dades en el futur, proporcionant un marc de referència per a la seva comprensió i el seu preprocessament.

Simplifica les tasques de manteniment i actualització del sistema, oferint una documentació clara de l'estructura i el processament de les dades.

Contribueix a la reproduïibilitat dels resultats, un aspecte fonamental per a la validesa científica i la confiança en l'IDDS.

Per il·lustrar la informació que es detalla en el document de metadades, la següent taula ofereix un resum concís dels principals processos de neteja aplicats a cada conjunt de dades, així com les seves característiques després d'aquest procés:

Nom del Dataset	Transformacions i Neteja	Atributs (Nombre: Nom)	Nombre de Mostres
US DAILY	<ul style="list-style-type: none"> Recodificació de variables Ordenació temporal de vell a nou Imputació de missings <ul style="list-style-type: none"> - MNAR: Imputació amb 0 (excepte per Prophet, on es mantenen els missings). - MAR: Imputació amb interpolació forward. 	12: date, state, positive, totalTestResults, death, positiveIncrease, negativeIncrease, total, totalTestResultsIncrease, posNeg, deathIncrease, hospitalizedIncrease	20780
AGE GROUPS	<ul style="list-style-type: none"> Eliminació de columnes redundants (identificadors) Creació de diccionari per grups d'edat (52 registres per grup) 	4: state, population_agegroup, agegroup, pct_pop_agegroup	918
POPULATION	<ul style="list-style-type: none"> Eliminació de columnes redundants (identificadors) Imputació de dades de Puerto Rico (no presents al dataset) 	3: state, population_state, pop_density_state	51
HEALTH INSUR-ANCE	<ul style="list-style-type: none"> Eliminació de columnes redundants (identificadors i "employed" per exèss de missings) Creació variable "Unknown" a labor_force Aplicació de OneHot encoding 	12: state, acs_variable, estimate, margin_of_error, estimate_type_population, health_insurance_coverage, private_insurance_coverage, public_coverage, age_group_under_19, age_group_overall, not_in_labor_force, labor_force_Unknown	1700
BEDS	<ul style="list-style-type: none"> Eliminació de columnes redundants Imputació de valors faltants amb 0 	5: state, bedsstate_local_government, bedsnon_profit, bedsfor_profit, bedstotal	51
COUNTIES	<ul style="list-style-type: none"> Eliminació de columnes redundants (identificadors) Imputació de dades de Puerto Rico amb informació del Cens de 2020 	4: state, county, population_county, pop_density_county	3142
VACCINATION	<ul style="list-style-type: none"> Eliminació de columnes i files amb molts missings Imputació de dades mancats amb informació externa Agregació per estats, i modificació d'algunes variables per adaptar-les. 	18: State, Date, Dose1_Total, Dose1_Total_pct, Dose1_65Plus, Dose1_65Plus_pct, Complete_Total, Complete_Total_pct, Complete_65Plus, Complete_65Plus_pct, Low_SVI_CTGY, Moderate_Low_SVI_CTGY, Moderate_High_SVI_CTGY, High_SVI_CTGY, Metro, Non-metro, Census2019, Census2019_65Plus	3070

Table 9.1: Taula de metadades

9.2 MATRIU DE DADES PREPROCESSADES

El model de decisió de l'IDDS utilitzarà una matriu de dades preprocessades que integra informació de diverses fonts per a cada estat, tal com s'ha descrit a la Secció 2.2 sobre les "Matrius de Dades Involucrades". Aquesta matriu es construeix a partir de la unió i transformació dels conjunts de dades detallats anteriorment i preprocessats segons la metodologia de la Secció 8. Per tal de capturar tant les característiques inherents de cada estat com la dinàmica temporal de la pandèmia, la matriu final es conceptualitza en dues parts principals: dades estàtiques a nivell estatal i dades diàries (dinàmiques) a nivell estatal.

1. Dades Estàtiques Preprocessades a Nivell Estatal

Aquesta part de la matriu conté característiques que es consideren relativament constants o que canvien a una freqüència menor durant el període d'anàlisi. S'obté a partir de la unió dels conjunts de dades de Health Insurance, Beds, Population i Age Groups.

El procés de construcció d'aquesta part de la matriu va implicar la càrrega dels fitxers de dades i un preprocessament inicial que va incloure el filtratge de dades no rellevants (com Puerto Rico, que no era present a tots els conjunts de dades). Per a una millor comprensió i consistència durant la unió, vam renombrar algunes columnes en els diferents datasets. Finalment, vam unir els datasets mitjançant la variable comuna "state" per crear un únic DataFrame que conté les dades estàtiques a nivell estatal.

Podem veure la forma de la matriu i un exemple de fila a la taula [9.2](#).

Columna	Descripció	Valor d'exemple
state	Nom de l'estat (categorical)	AL
acs_variable	Identificador de l'estat (categorical)	DP03_0096
estimate	Estimació de la variable (numerical)	4307566.0
margin_of_error	Marge d'error de l'estimació (numerical)	8603.0
estimate_type_population	Tipus d'estimació sobre població (binary)	1
health_insurance_coverage	Cobertura d'assegurança mèdica (binary)	1
private_insurance_coverage	Cobertura privada (binary)	0
public_coverage	Cobertura pública (binary)	0
age_group_under_19	Dades del grup d'edat < 19 anys (binary)	0
age_group_overall	Dades generals per grups d'edat (binary)	1
not_in_labor_force	Població fora de la força laboral (binary)	0
labor_force_Unknown	Dades desconegudes sobre força laboral (binary)	1
bedsstate_local_government	Llits hospitalaris públics estatals/locals (numerical)	1.4
bedsnon_profit	Llits en hospitals sense ànim de lucre (numerical)	0.8
bedsfor_profit	Llits en hospitals amb ànim de lucre (numerical)	0.9
bedstotal	Total de llits hospitalaris (numerical)	3.1
population_state	Població de l'estat (numerical)	4887871.0
pop_density_state	Densitat de població de l'estat (numerical)	96.509389
population_agegroup	Població per grup d'edat (numerical)	293203.0
agegroup	Grup d'edat especificat (numerical: representing categorical)	0.0
pct_pop_agegroup	Percentatge de la població per grup d'edat (numerical)	0.059986

Table 9.2: Estructura de la matriu de dades per estat

2. Dades Diàries Preprocessades a Nivell Estatal

Aquesta matriu conté els dades de contagis diaris a nivell d'estat i les dades de vacunació, que també són diàries i a nivell d'estat. La unió s'ha fet per les variables d'estat i dia.

Al fer la unió s'han generat nous missings a les columnes del dataset de vacunació, ja que, les dades de vacunació comencen més tard que les de contagis. Per solucionar-ho, hem imputat amb 0 totes les columnes de dosis. Pel que fa a la resta de columnes, que no depen del dia, sinó de l'estat, hem imputat el valor de l'estat.

Per acabar, com que a l'idss volem tenir en compte les dades de contagis dels estats veïns, hem creat una nova columna. Per fer-ho hem fet servir el dataset de neighbours, que ens diu per cada estat quins son els seus estats veïns, i així hem pogut posar per cada dia i estat quin és l'augment de contagis dels estats veïns.

Podem veure la forma de la matriu i un exemple de fila a la taula [9.3](#).

Columna	Descripció	Exemple
date	Data del registre (date)	2021-02-18
state	Nom de l'estat (categorical)	Alabama
positive	Casos positius acumulats (numerical)	484365.0
totalTestResults	Total de resultats de test acumulats (num)	2243392.0
death	Defuncions acumulades (num)	9424.0
positiveIncrease	Nous casos positius (num)	1198
negativeIncrease	Nous resultats negatius (num)	4081
total	Total de tests realitzats (num)	2347962
totalTestResultsIncrease	Nous resultats de test (num)	5091
posNeg	Suma de positius i negatius (num)	2347962
deathIncrease	Noves defuncions (num)	78
hospitalizedIncrease	Nous hospitalitzats (num)	226
Dose1_Total	Total vacunats amb 1a dosi (num)	462641.0
Dose1_Total_pct	Percentatge vacunats 1a dosi (num)	9.44
Dose1_65Plus	Total vacunats 1a dosi majors 65 (num)	265271.0
Dose1_65Plus_pct	Percentatge vacunats 1a dosi majors 65 (num)	31.05
Complete_Total	Total vacunats pauta completa (num)	159221.0
Complete_Total_pct	Percentatge vacunats pauta completa (num)	3.25
Complete_65Plus	Total majors 65 amb pauta completa (num)	67775.0
Complete_65Plus_pct	Percentatge majors 65 amb pauta completa (num)	7.93
Low_SVI_CTGY	Percentatge població amb SVI baix (num)	8.99
Moderate_Low_SVI_CTGY	Percentatge amb SVI moderadament baix (num)	15.97
Moderate_High_SVI_CTGY	Percentatge població amb SVI moderadament alt (num)	43.99
High_SVI_CTGY	Percentatge població amb SVI alt (num)	31.04
Metro	Percentatge de població en zona metropolitana (num)	76.84
Non-metro	Percentatge de població en zona no metropolitana (num)	23.16
Census2019	Població total segons cens 2019 (num)	4903185.0
Census2019_65Plus	Població >65 segons cens 2019 (num)	854312.0
neighbor_contagions	Nous casos positius a estats veïns (num)	9647

Table 9.3: Estructura de la matriu de dades per estat i data

ENTRENAMENT DE MODELS DE DADES

10.1 CLUSTERING DELS ESTATS EUA

Per identificar patrons comuns en l'evolució de la pandèmia entre els diferents estats dels EUA, s'ha implementat un model de clustering jeràrquic basat en dades temporals de casos diaris de COVID-19. Aquest mòdul del sistema ajuda a agrupar els estats en funció de similituds en la seva evolució pandèmica, per tal de donar suport a decisions més contextualitzades.

El model utilitza l'algoritme de clustering jeràrquic aglomeratiu amb l'estratègia d'enllaç de Ward (method='ward'). Les distàncies entre estats s'han calculat combinant la matriu dinàmica (dades diàries de casos de COVID) i la matriu estàtica (conté variables amb factors com vulnerabilitat social o demografia). Les dues matrius es normalitzen amb MinMaxScaler i es combinen amb pesos $w_{temp} = 0.3$ per a la component temporal i $w_{est} = 0.7$ per a l'eststructural. Es talla el dendograma a una distància de 1 per obtenir els clústers finals (usant fcluster amb criterion='distance').

S'ha fet un clustering diferent per a cada una de les 3 variables target que tenim (hospitalitzacions, casos positius i morts). Aquesta separació ens permet capturar dinàmiques específiques per a cada indicador, ja que la distribució espacial i temporal de morts, hospitalitzacions i positius no sempre segueix el mateix patró. Veiem a continuació els tres clusterings jeràrquics que hem obtingut:

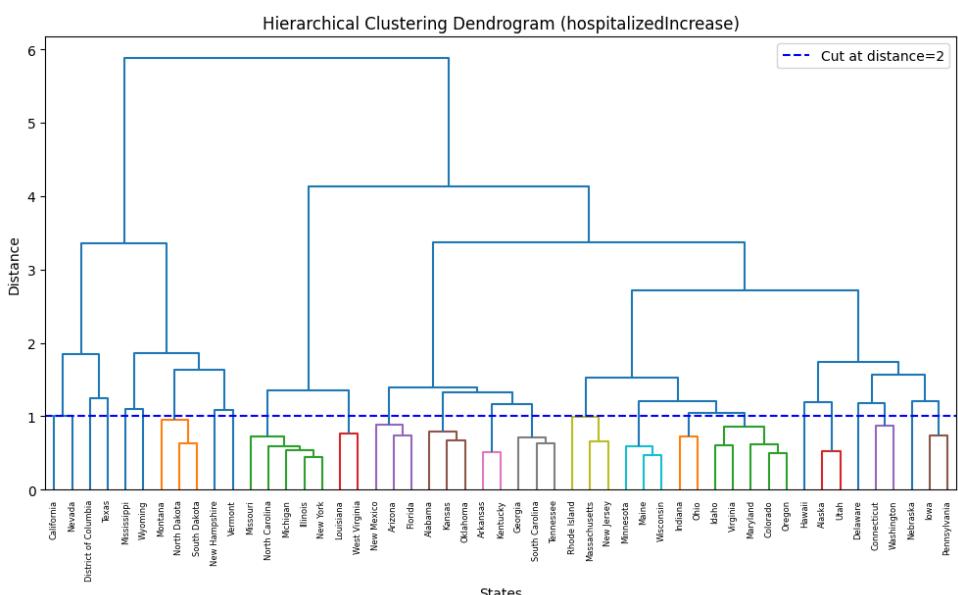


Figure 10.1: Clústering jeràrquic per la variable target de hospitalitzacions

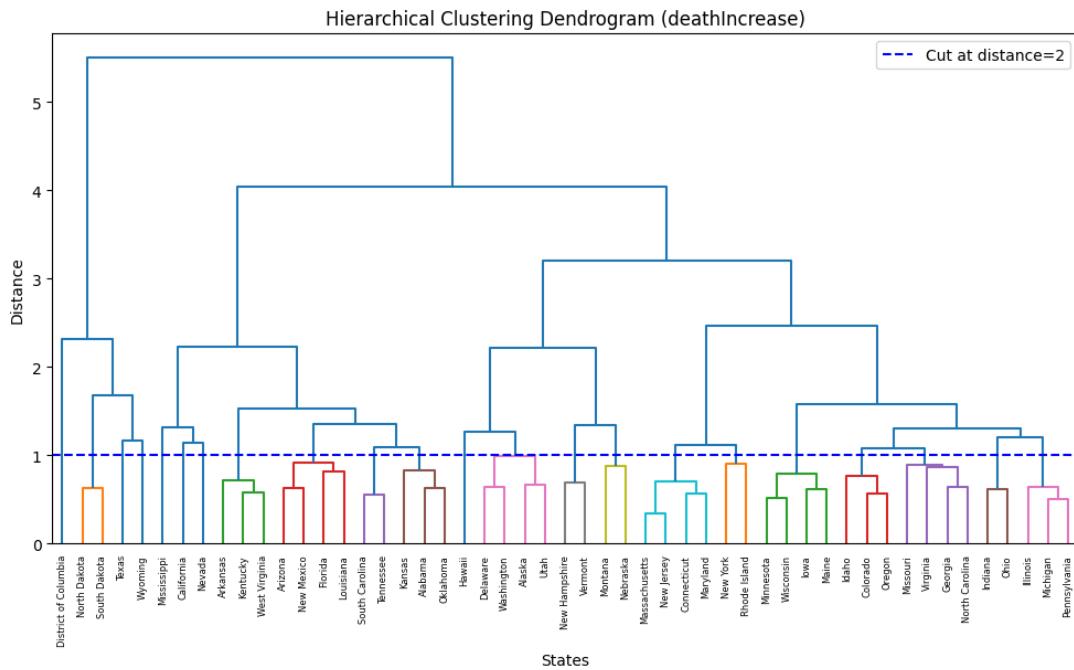


Figure 10.2: Clústering jeràrquic per la variable target de morts

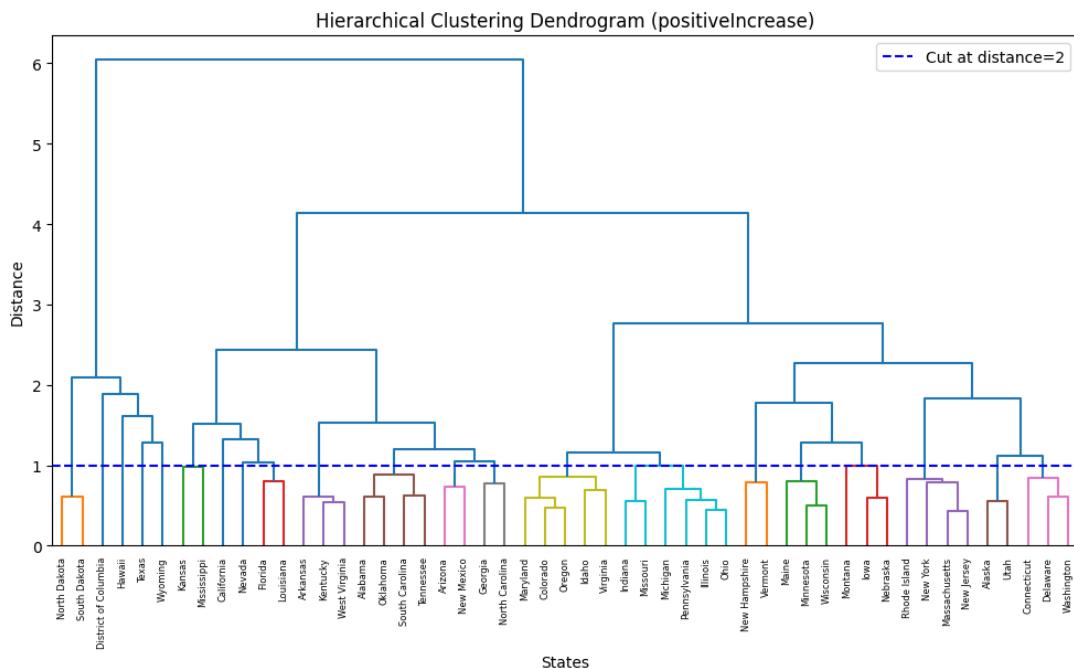


Figure 10.3: Clústering jeràrquic per la variable target de Casos Positius

El model resultant és una classificació dels estats en diversos clústers, cadascun dels quals agrupa territoris amb evolucions pandèmiques similars. Aquests clústers seran utilitzats per entrenar els diferents models predictius de Prophet.

El clustering s'executa com a etapa prèvia dins del pipeline del sistema i els resultats s'emmagatzemen com a metadades associades a cada estat per a la seva consulta en temps real.

La qualitat del clustering s'ha validat mitjançant el coeficient de correlació cophenètica, que quantifica fins a quin punt les distàncies en el dendograma conserven les distàncies originals entre estats. En tots tres casos hem obtingut valors elevats (0,766, 0,766 i 0,75), fet que indica que els clústers formats són consistents i representatius.

10.2 PROPHET: MODEL DE SÈRIES TEMPORALS

El model Prophet ha estat seleccionat per la seva capacitat d'adaptar-se eficaçment a dades temporals amb forta estacionalitat i tendències definides, característiques pròpies de les dades epidemiològiques. Concretament, s'ha utilitzat per predir tres variables clau: nous casos positius diaris de COVID-19, hospitalitzacions diàries i defuncions diàries. Aquestes prediccions constitueixen una entrada fonamental per al càlcul del risc i les decisions que pren l'IDSS.

Configuració dels paràmetres d'entrenament

Per garantir el millor rendiment dels models Prophet, es va dur a terme una optimització dels seus principals hiperparàmetres mitjançant un Grid Search. Aquesta optimització es va centrar en:

- **changepoint_prior_scale (0.01, 0.1, 0.5)**: regula la flexibilitat del model a l'hora d'adaptar-se als canvis sobtats en la tendència. Els valors més baixos imposen suavitats que eviten el sobreajustament, mentre que els més alts permeten capturar disruptions sobtades en l'evolució de la pandèmia.
- **seasonality_prior_scale (0.01, 1, 10)**: controla la força de l'estacionalitat, particularment la setmanal, present en les dades epidemiològiques. S'ajusta aquest paràmetre per evitar sobreinterpretar fluctuacions que poden ser degudes a soroll.

Altres paràmetres es van mantenir constants per estabilitzar el comportament del model:

- **growth="linear"**: s'utilitza una tendència lineal, suficient per a l'horitzó de predicció limitat (7 dies).
- **changepoint_range=0.8**: restringeix la detecció de punts de canvi als primers 80% de les observacions, evitant que el model s'ajusti excessivament a canvis recents o puntuals.
- **weekly_seasonality=True**: activa l'estacionalitat setmanal, molt rellevant en les dades de salut pública.

Més enllà de les variables objectiu, el model es va millorar incorporant regressors externs. Aquests regressors, que representen factors contextials, inclouen variables com la densitat de població i la disponibilitat de llits d'hospital. La seva inclusió permet al model tenir en compte la influència d'aquests factors en les variables predites, millorant potencialment la precisió i proporcionant prediccions més matisades.

Dades i Preprocessament

Les dades utilitzades per entrenar Prophet cobreixen el període entre el 13 de gener de 2020 i l'1 de març de 2021, abastant les fases més crítiques de la pandèmia. La integració de dades va incloure:

- La normalització de variables numèriques per facilitar la convergència i estabilitat dels models.
- La fusió de dades temporals (diàries) amb variables estructurals (població, edat, llits hospitalaris) mitjançant claus geogràfiques i temporals.
- La divisió entrenament/prova, utilitzant dades fins a l'1 de març per entrenar i les posteriors per validar la capacitat de generalització dels models.

Per abordar la variabilitat regional i reduir la càrrega computacional, es va implementar una estratègia d'agrupació d'estats mitjançant tècniques de clustering. En lloc d'entrenar un model per a cada estat, el sistema entrena un model Prophet per a cada variable objectiu dins de cada clúster d'estats amb comportaments similars. Això suposa entrenar tres models per clúster, un per a casos positius, un per a hospitalitzacions i un per a defuncions, optimitzant així el balanç entre precisió i eficiència computacional.

Validació dels Models

Els models es van avaluar en un horitzó de 7 dies, considerat adequat per a la presa de decisions sanitàries immediates. Es van utilitzar les mètriques següents:

- **MAE (Error Absolut Mitjà)**: per avaluar l'error mitjà sense considerar la direcció.
- **RMSE (Arrel de l'Error Quadràtic Mitjà)**: penalitza errors grans, útil per identificar prediccions molt desviades.
- **MAPE (Error Percentual Absolut Mitjà)**: permet comparar errors relatius entre diferents escales.

Per complementar l'anàlisi quantitativa del rendiment dels models Prophet, a continuació es presenten gràficament les mètriques d'error per clúster, diferenciades per variable objectiu: defuncions (**deathIncrease**), hospitalitzacions (**hospitalizedIncrease**) i casos positius (**positiveIncrease**).

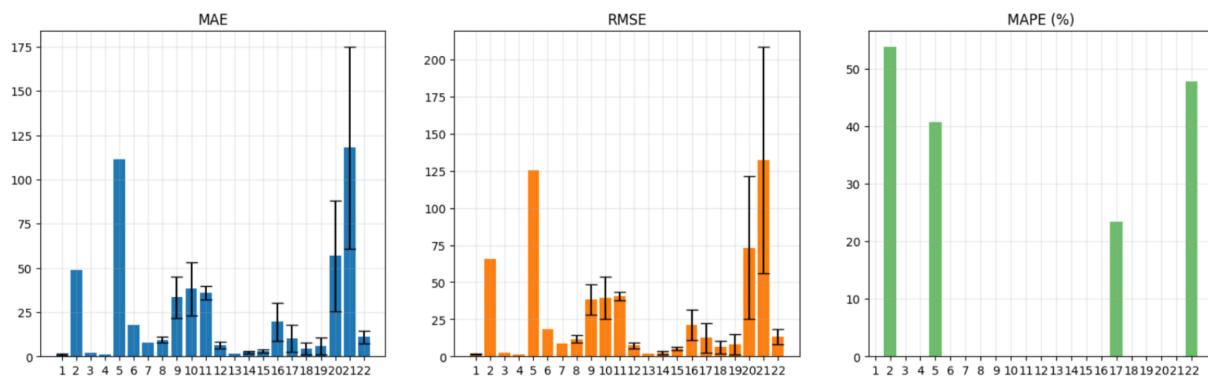


Figure 10.4: Mètriques per clúster - DeathIncrease

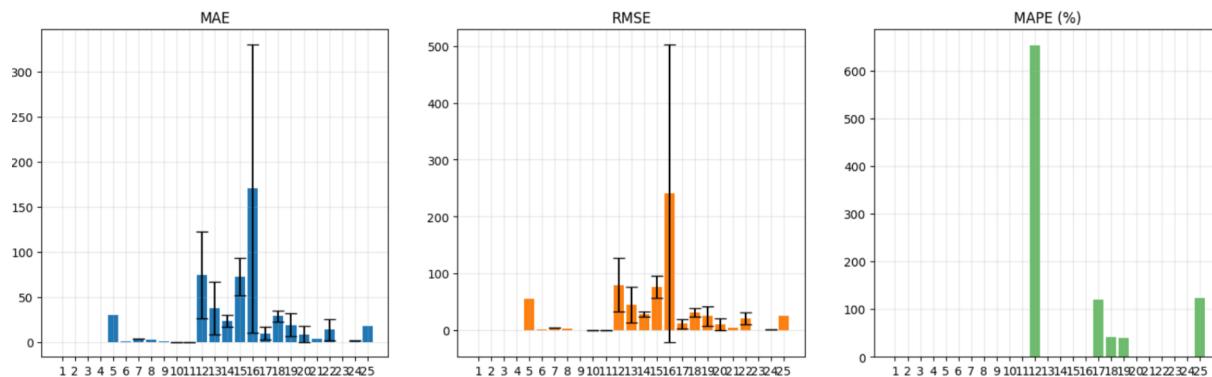


Figure 10.5: Mètriques per clúster - HospitalizedIncrease

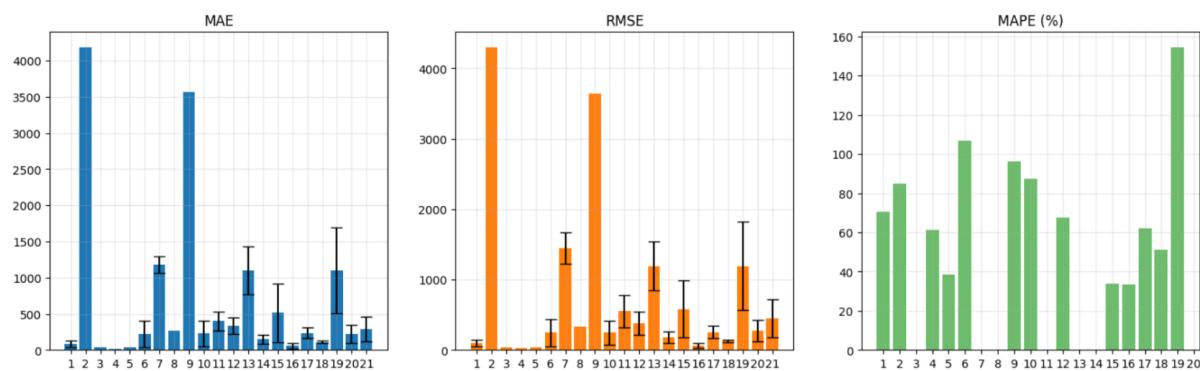


Figure 10.6: Mètriques per clúster - PositiveIncrease

Els errors de predicción per a les defuncions mostren una elevada variabilitat entre clústers, amb valors de MAPE superiors al 50% en alguns casos. Aquesta inestabilitat s'explica per la naturalesa complexa de la variable, que depèn de factors com els retards temporals respecte als contagis, la capacitat hospitalària o les intervencions sanitàries locals, sovint difícils de modelitzar.

Pel que fa a les hospitalitzacions, tot i que l'error mitjà és menor, es detecten clústers amb RMSE elevats, probablement degut a canvis sobtats en les dades o a la heterogeneïtat interna dins del clúster.

En el cas dels casos positius, es registren els majors errors absoluts (MAE i RMSE), degut a la gran diferència de volum entre regions. No obstant això, el MAPE ajuda a identificar clústers amb bon rendiment relatiu, especialment quan se situa per sota del 30%.

Aquests resultats posen de manifest limitacions del model Prophet, com la seva estructura simple o la qualitat i soroll en les dades. Per això, dins l'IDSS, les sortides dels models s'utilitzen com a entrada per al motor de regles difuses, que aporta context i capacitat d'adaptació per millorar la presa de decisions.

Integració al Sistema IDSS

La integració dels models Prophet dins l'IDSS s'ha dissenyat per garantir un funcionament eficient i adaptable en entorns en evolució constant. En primer lloc, els models entrenats es serialitzen mitjançant el format pickle, cosa que permet una càrrega ràpida i la reutilització dels pesos i paràmetres apresos sense necessitat de reentrenament des de zero.

Per mantenir l'actualització constant del sistema davant l'evolució de les dades epidemiològiques, s'implementa un procés de reentrenament automàtic setmanal, que incorpora les dades més recents i assegura que els models reflecteixin les tendències actuals en transmissió, hospitalització i mortalitat.

Els models actualitzats generen prediccions en temps real per a un horitzó de set dies, que serveixen com a entrada per al motor de regles difuses, el qual processa les sortides del model (predicció de casos, hospitalitzacions i defuncions) i les transforma en nivells de risc específics per estat. Així, la combinació entre predicció i coneixement expert proporciona una base sòlida per a la presa de decisions ràpides, contextualitzades i adaptatives durant situacions de crisi sanitària.

REPRESENTACIÓ DEL CONEIXEMENT

En aquesta secció descrivim amb detall el component de coneixement basat en lògica difusa implementat al sistema, que constitueix el nucli del raonament i la presa de decisions en l'IDSS. Aquest component es materialitza en el fitxer `fuzzy_epidemiology.py`, on es formalitza el coneixement expert en forma de regles difuses que permeten inferir un **nivell de risc de confinament** a partir de diverses variables epidemiològiques i demogràfiques.

11.1 SISTEMA BASAT EN LÒGICA DIFUSA

El coneixement es representa mitjançant un sistema de *lògica difusa* (fuzzy logic), un formalisme adequat per modelar sistemes amb incertesa i dades imprecises. Aquest formalisme es basa en la teoria de conjunts difusos i assigna graus de pertinença a conjunts definits per funcions de pertinència.

Concretament, s'utilitza un sistema basat en regles IF-THEN difuses, on cada regla combina etiquetes lingüístiques definides amb funcions de pertinència triangulars aplicades a variables d'entrada (antecedents) i que produeixen una variable de sortida (consequents). La defuzzificació transforma el conjunt difús resultant en un valor numèric que representa el risc.

11.1.1 Estructura principal

La representació del coneixement consta dels següents elements:

- **Variables d'entrada (Antecedents):**
 - IA: Incidència acumulada per 100 000 habitants, amb conjunts des de *Molt Baixa* fins a *Molt Alta*.
 - occupancy: Ocupació hospitalària percentual.
 - mortality: Taxa de mortalitat per 100 000 habitants.
 - lethality: Taxa de letalitat percentual.
 - population65: Percentatge de població major de 65 anys.
 - density: Densitat de població (habitants/km²).
- **Variable de sortida (Consequent):** risk, que representa el risc global de confinament. Està definida amb sis conjunts difusos (*Molt Baix*, *Baix*, *Moderat*, *Alt*, *Molt Alt*, *Extrem*), aportant granularitat per a diferents nivells d'intervenció.
- **Funcions de pertinència:** Totes les variables utilitzen funcions triangulars (`trimf`), amb paràmetres ajustats a la realitat epidemiològica observada durant la COVID-19.

- **Base de regles:** Conjunt de regles IF-THEN difuses que combinen variables d'entrada amb operadors lògics (AND, OR), reflectint combinacions epidemiològiques i demogràfiques crítiques. Per exemple: *Si la incidència és molt alta i l'ocupació hospitalària és molt alta, llavors el risc és extrem.*

Aquest coneixement es basa en les fonts mencionades amb detall en l'apartat 3. Aquestes fonts es combinen per crear un sistema robust que integra dades objectives i coneixement expert.

11.1.2 *Rol en el sistema*

El sistema de coneixement difús actua com a **motor de decisió** dins de l'IDSS, amb les següents funcions:

1. Rebre dades epidemiològiques i demogràfiques actualitzades, incloses prediccions.
2. Aplicar regles difuses per inferir un risc global.
3. Traduir aquest risc en recomanacions sobre mesures de confinament i gestió hospitalària.
4. Fer suport a la planificació estratègica mitjançant recomanacions de trasllats de pacients.
5. Proporcionar una sortida interpretable i explicable per a gestors sanitaris.

11.1.3 *Sortida, resultats i validació*

La sortida principal és un valor numèric entre 0 i 100 indicant el risc de confinament, associat a una etiqueta qualitativa (*No confinar, Selectiu, Estricto, Immediat*). La validació inclou un ajust de regles i funcions de pertinència amb dades reals de la pandèmia, i un posterior ànalisi de casos per garantir recomanacions coherents.

11.1.4 *Integració a l'arquitectura general*

El sistema està integrat dins la classe Python FuzzyEpidemiology, que primer carrega dades demogràfiques i de veïnatge, després calcula mètriques epidemiològiques a partir de prediccions, executa el sistema difús per obtenir el risc, genera recomanacions concretes, i finalment emmagatzema les recomanacions a la base de dades. Aquest mòdul connecta amb models predictius i la capa d'interfície, facilitant manteniment i extensió.

En resum, aquesta representació basada en lògica difusa proporciona un mecanisme transparent, flexible i efectiu per incorporar coneixement expert i dades heterogènies dins de l'IDSS, facilitant una presa de decisions adaptativa en contextos epidemiològics complexos.

11.2 COMPONENT DE MODELS PREDICTIUS

El component de models predictius està basat en tècniques d'anàlisi de sèries temporals i aprenentatge estadístic, que permeten estimar l'evolució futura dels indicadors epidemiològics clau, com ara nous casos, hospitalitzacions i defuncions. En concret, s'utilitzen models Prophet, explicats amb detall en apartats anteriors.

11.2.1 *Estructura principal*

Aquest model pren com a entrada dades històriques temporals de cada estat, incloent casos confirmats, hospitalitzacions i defuncions diàries. La seva sortida consisteix en prediccions puntuals i agregades (per exemple, increment de casos setmanal) que serveixen com a input a la resta del sistema.

11.2.2 *Fonts de contingut*

Aquest component es fonamenta en dades epidemiològiques històriques reals, recopilades de fonts oficials i actualitzades periòdicament, així com en la literatura científica sobre modelització de pandèmies. Es valida contínuament amb dades recents per adaptar-se a canvis en el comportament del virus i les polítiques sanitàries.

11.2.3 *Rol en el sistema*

Actua com a **proveïdor d'informació predictiva**, facilitant una estimació acurada i actualitzada de la situació epidemiològica futura per a cada estat. Aquestes prediccions alimenten el motor de raonament difús i altres mòduls per a la presa de decisions i recomanacions.

11.2.4 *Sortida, resultats i validació*

La sortida principal consisteix en prediccions quantitatives de variables epidemiològiques clau per a horitzons temporals curts (normalment 7 dies). La validació es realitza mitjançant l'avaluació del rendiment predictiu sobre dades històriques i comparacions amb altres models i fonts. Aquesta avaluació assegura la fiabilitat i la utilitat de les prediccions per al sistema.

11.2.5 *Integració a l'arquitectura general*

Aquest component s'integra com un mòdul de backend que actualitza periòdicament les prediccions per estat i data a la base de dades. Les dades generades són consultades i processades pel component de lògica difusa i altres mòduls, facilitant així una presa de decisions informada i oportuna dins de l'IDSS.

11.3 COMPONENT DE REPRESENTACIÓ DE DADES DEMOGRÀFIQUES I VEÏNATGE

Aquest component utilitza una estructura basada en bases de dades relacionals i models de graf per representar la topologia geogràfica dels estats i els seus recursos sanitaris, especialment la capacitat hospitalària. Aquesta representació incorpora coneixement espacial i estructural dins del sistema.

11.3.1 *Estructura principal*

La representació consta de:

- Una taula amb dades demogràfiques estàtiques per estat, que inclou població total, densitat, percentatge de població vulnerable (majors de 65 anys) i nombre de llits hospitalaris per 1000 habitants.
- Una matriu d'adjacència o graf de veïnatge que determina quins estats són veïns directes, facilitant l'anàlisi de transferències de pacients.
- Mecanismes per combinar i consultar aquestes dades amb la informació epidemiològica i predictiva.

11.3.2 *Fonts de contingut*

Les dades provenen de fonts oficials estadístiques i sanitàries, prèviament preprocesades i estructurades per a un accés eficient. Aquesta informació s'actualitza puntualment per garantir la seva validesa i rellevància.

11.3.3 *Rol en el sistema*

Proporciona el **coneixement estructural i contextual** necessari per prendre decisions sobre la gestió de recursos sanitaris i per avaluar la possibilitat de trasllats entre estats. Aquesta representació complementa la informació epidemiològica amb una visió geogràfica i logística essencial.

11.3.4 *Sortida, resultats i validació*

La sortida són consultes i càlculs sobre disponibilitat de recursos i connexions veïnals, que s'utilitzen per generar recomanacions de transferències i gestió de llits hospitalaris. La validació es realitza mitjançant comprovacions de consistència de dades i comparacions amb situacions reals de capacitat hospitalària.

11.3.5 *Integració a l'arquitectura general*

Aquest component està integrat dins la capa d'accés a dades de l'IDSS, facilitant la combinació amb les prediccions i el motor de lògica difusa. Serveix com a font de dades clau per als processos de decisió i recomanació, i es comunica amb la base de dades i la capa d'interfície.

Per tant, veiem que aquests tres components —models predictius, representació demogràfica i veïnatge, i sistema basat en lògica difusa— treballen de manera coordinada dins de l'IDSS per proveir un sistema robust, flexible i capaç d'integrar dades heterogènies i coneixement expert per a la presa de decisions en la gestió de pandèmies.

Aquest disseny modular facilita la mantenibilitat, extensió i adaptabilitat del sistema a futurs canvis i noves fonts de dades o coneixement.

COMPONENTS DE RAONAMENT

12.1 COMPONENTS DE RAONAMENT

En el nostre IDSS, els components de raonament tenen un paper clau en la traducció de dades quantitatives en decisions qualitatives i accions concretes, que poden ser utilitzades per les autoritats sanitàries i gestors hospitalaris. Aquests components de raonament s'encarreguen d'integrar el coneixement expert i la informació epidemiològica obtinguda a partir de models predictius per avaluar el risc i generar recomanacions específiques sobre mesures a adoptar, com ara el tipus de confinament o la necessitat de transferència de pacients entre estats.

12.1.1 *Flux i execució del raonament*

Els components de raonament s'integren en l'arquitectura de l'IDSS en forma d'una classe Python `FuzzyEpidemiology` que encapsula la configuració i execució del motor difús. El seu flux operatiu principal és:

1. Recuperar les dades demogràfiques i de veïnatge des de fitxers i base de dades.
2. Obtenir les prediccions més recents per a cada estat i data.
3. Calcular les mètriques epidemiològiques clau a partir de les dades i prediccions.
4. Executar el sistema difús per inferir el nivell de risc.
5. Traduir el risc numèric en recomanacions concretes sobre confinament, gestió de llits i vacunació.
6. Verificar la disponibilitat de llits a estats veïns per proposar trasllats si cal.
7. Emmagatzemar les recomanacions a la base de dades i posar-les a disposició per a la interfície d'usuari.

12.1.2 *Interacció amb altres components i arquitectura*

Aquest nucli de raonament s'integra dins una arquitectura modular basada en microserveis, comunicant-se amb altres components a través d'una API RESTful. La sortida del motor de raonament (nivell de risc i recomanacions) està disponible en temps real per a dashboards, bots i planificadors de recursos.

12.1.3 Validació i casos d'ús

La validació del sistema de raonament s'ha realitzat mitjançant proves amb dades artificials que simulen escenaris realistes. S'han definit diversos casos d'ús amb paràmetres demogràfics i epidemiològics coneguts per comprovar que el sistema respon correctament a diferents situacions, produint recomanacions coherents i ajustades als nivells de risc.

La taula següent resumeix els resultats obtinguts en aquests casos d'ús, mostrant les mètriques d'entrada i les recomanacions generades pel sistema:

Escenari	1	2	3
<i>Mètriques d'entrada</i>			
IA (per 100k hab.)	23.85	95.38	95.38
Ocupació (θ , %)	0.00	57.23	85.84
Mortalitat (π , per 100k hab.)	0.48	14.31	47.69
Letalitat (%)	2.00	15.00	50.00
Població >65 (habitants)	251618	251618	251618
Densitat (hab/km ²)	17.27	17.27	17.27
<i>Sortida</i>			
Nivell de Risc	6.68	65.00	95.21
Recomanació de Confinament	No	Estricta	Immediat

Table 12.1: Resultats de la validació del sistema IDSS amb dades artificials.

Aquests resultats confirmen que el motor difús respon adequadament als llindars establerts, i evidencien que la formulació de les regles i les funcions de pertinència capturen correctament la relació entre les variables epidemiològiques i demogràfiques i el nivell de risc associat.

Concretament, l'anàlisi dels tres escenaris validats mostra que en el primer cas, amb baixa incidència, ocupació hospitalària nul·la i mortalitat mínima, el sistema assigna un risc molt baix, resultant en una recomanació de no confinar, per tant observem que evita prendre mesures innecessàries.

En el segon cas, amb incidència i mortalitat moderades i una ocupació hospitalària significativa, el risc s'eleva a un nivell alt amb una recomanació de confinament estricta.

Finalment, en l'escenari crític amb alta incidència, mortalitat i ocupació, el motor assigna un risc extrem i recomana confinament immediat. Aquest comportament demostra que les regles IF-THEN i les funcions de pertinència triangulars gestionen efectivament la imprecisió i l'ambigüïtat de les dades d'entrada. Amb això també es confirma que el procés de defuzzificació i el mapeig a accions concretes estan ben implementats, lo qual reforça la confiança en l'ús del motor difús com a eina robusta de suport a la presa de decisions.

INTERACCIONS DE L'IDSS

Una de les tasques més importants en el disseny d'un sistema IDSS és la integració de totes les seves components. En la Sec. 5 es presenten els diferents conjunts de dades, models i coneixement disponible, explicitant les connexions entre aquestes parts. No obstant això, no s'explica com interaccionen exactament, és a dir, què, com i qui intercanvien dades, de qualsevol tipus.

Precisament, això és el que busca aquesta secció, a partir d'una taula que analitza les interaccions entre les components de l'IDSS, tant per la fase d'entrenament com de producció. En particular, s'especifiquen aquelles dependències entre nodes que estan implementades, més que no idees.

Component Origen	Component Destí	Què es transfereix?	Canal	Freqüència	Riscos de Fal·lada
Dades històriques de la Covid-19 (offline)	Mòdul de pre-processament de dades	Fitxers CSV/JSON amb registres diaris i metadades (font, timestamp)	APIs externes (https://covidtracking.com/ , https://data.cdc.gov)	Únic	Pèrdua de connexió; canvis d'esquema; dades incompltes.
Mòdul de pre-processament de dades	Base de Dades SQL	Taules amb dades netejades, imputació i normalització de camps	SQLAlchemy	Únic	Errors d'ETL; valors nuls persistents; incompatibilitat d'esquemes.
Base de Dades SQL	Clustering (DTW + MDS)	Sèries temporals i dades estructurals	SQLAlchemy	Únic	Dades sorolloses; paràmetres DTW/MDS inadaptats.
Clustering (DTW + MDS)	Preparació d'entrenament Prophet	Etiquetes de cluster assignades a cada estat	Directori local / BD interna	Únic	Errors de lectoescritura; assignacions inconsistents.
Preparació d'entrenament Prophet	Entrenament Prophet	Conjunts <i>train/test</i> : sèries temporals i regressors externs	SQLAlchemy	Únic	Dades insuficients; desequilibri de clusters; duplicats.
Entrenament Prophet (per cluster)	Emmagatzematge de models local	Fitxers pickle amb paràmetres (tendència, estacionalitat)	Sistema de fitxers distribuit (NFS)	Únic	Error en serialització; espai insufficient; corrupció de fitxers.

Table 13.1: Interaccions detallades en fase d'entrenament del sistema IDSS

Component Origen	Component Destí	Què es transfereix?	Canal	Freqüència	Riscos de Fallada
APIs externes, Fitxers CSV	Backend / BD SQLite (PandemicData, StaticStateData)	Dades COVID (diàries, demogràfiques, capacitat hosp.).	HTTP/API, Càrrega fitxers	Diària / Segons disponibilitat	Connexió, format, dades incomplades/eròries.
Frontend (HTML)	Backend API (Gestió Usuaris)	Credencials, sol·licituds d'usuari (rols, estats).	API RESTful (HTTPS)	Per sessió / Puntual	Error autenticació, vulnerabilitats, gestió rols.
BD SQLite (PandemicData, StaticStateData)	Servei ProphetTrainer (Preproc.)	Sèries temporals històriques, dades estructurals.	Accés BD (SQLAlchemy)	Periòdica (pre-entrenament)	Dades BD sorolloses/inconsistents, error ETL.
Servei ProphetTrainer	Emmagatzematge Models ML (Fitxers/BD)	Models Prophet serialitzats, mètriques validació.	Escriptura fitxers/BD	Periòdica (actualització)	Error entrenament, overfitting, serialització, emmagatzematge.
Models ML (Prophet) / BD SQLite (Dades recents)	Servei ProphetPredictor	Models carregats, dades recents per predicció.	Càrrega models, Accés BD	Sota demanda / Programada	Model no trobat/corrupte, dades entrada invàlides, error càcul.

Table 13.2: Interaccions (Part 1: Ingesta, Autenticació, Entrenament, Predicció)

La Taula 13.2 il·lustra un sistema en producció dissenyat per a una operació contínua i dinàmica. Això es manifesta en la ingestió de dades provinents d'APIs externes i fitxers CSV amb una freqüència que pot ser diària o segons la disponibilitat de noves dades, i en el reentrenament periòdic dels models Prophet a través d'un servei dedicat (ProphetTrainer) per mantenir la seva rellevància. Aquesta actualització constant és vital, ja que els models Prophet, per la seva naturalesa, tendeixen a perdre precisió si no s'actualitzen regularment amb les dades més recents per capturar noves tendències, estacionalitats o l'efecte d'esdeveniments imprevistos.

Per contra, la fase d'entrenament original (Taula 13.1) es caracteritzava per una càrrega única de dades històriques ("offline") i un procés d'entrenament més puntual i aïllat, enfocat principalment en la creació inicial dels models, de clustering i Prophet. Els components com el "Mòdul de preprocessament de dades" i l'emmagatzematge "local" de models suggereixen una configuració més orientada al desenvolupament inicial, en lloc del manteniment i actualització regulars exigits per un entorn operatiu com el descrit per al nou sistema IDSS-Pandemic Control.

Component Origen	Component Destí	Què es transfereix?	Canal	Freqüència	Riscos de Fallada
Servei ProphetPredictor	BD SQLite (Taula Prediction)	Predictions 7 dies, acumulats, diaris.	Accés BD (SQLAlchemy)	Post-predicció	Error escriptura BD, inconsistència, volum.
BD SQLite (Prediction, PandemicData, StaticStateData)	Servei FuzzyEpidemiology	Predictions, mètriques epidemiològiques, dades demogràfiques.	Accés BD	Sota demanda / Programada	Dades entrada incorrectes, regles difuses errònies, error càcul.
Servei FuzzyEpidemiology	BD SQLite (Taula Recommendation)	Recomanacions, nivells de risc, accions específiques.	Accés BD	Post-càcul Fuzzy	Error escriptura BD, recomanacions inconsistentes.
Backend API (Servint dades SQLite)	Frontend (React) - Dashboard	Dades per a dashboard (mapa, gràfics, indicadors, predictions, recomanacions).	API RESTful (HTTPS, JSON)	Contínua / Actualitzacions	Error API, renderització, dades desactualitzades, sobrecàrrega.
BD SQLite, Models ML, Logs	Sistema Backup / Eines Monitoratge	Backups BD/-models, logs, mètriques rendiment.	Scripts backup, Eines monitoratge	Programada / Contínua	Fallada backup, pèrdua dades, errors no detectats.

Table 13.3: Interaccions (Part 2: Emmagatzematge, Anàlisi, Visualització, Manteniment)

13.1 INTEGRACIÓ

Els components de la Taula 13.3 són essencials perquè transformen els models, prèviament entrenats, d'un estat purament analític a un sistema de suport a decisions (IDSS) plenament operatiu i integrat. Mentre la fase d'entrenament es concentra en la meticulosa creació de models predictius precisos, les interaccions i components descrits posteriorment s'encarreguen de donar vida a aquests models, permetent que el seu potencial es materialitzi en accions i coneixement útil per als usuaris finals en un cicle continu d'actualització i resposta.

Un aspecte fonamental d'aquesta transformació és la decisió estratègica d'emmagatzemar tots els resultats intermedis i finals en una base de dades SQL (SQLite). Aquesta base de dades no és estàtica; s'actualitza contínuament a mesura que arriben noves dades epidemiològiques des de fonts web externes (com APIs públiques) i a través de les aportacions directes que els usuaris o representants dels estats poden pujar al sistema. En lloc d'utilitzar formats més simples com fitxers CSV o JSON, la base de dades SQL ofereix una estructura robusta per gestionar aquest flux constant de dades, facilitant consultes complexes, assegurant la integritat

referencial i servint com a font centralitzada on s'afegeixen les noves sol·licituds, prediccions i recomanacions. Això permet no només la consulta actual, sinó també anàlisis històriques i l'auditoria de decisions. La capacitat d'interpretació de les dades esdevé dinàmica gràcies a aquest flux.

El servei FuzzyEpidemiology no actua només sobre un conjunt fix de dades, sinó que processa les prediccions numèriques que el ProphetPredictor genera a mesura que les noves dades ingressen al sistema i s'actualitza la base de dades. Així, les recomanacions i els nivells de risc es recalculen i s'adapten a la situació més recent, traduint les últimes prediccions en evaluacions qualitatives i consells accionables. Aquesta interpretació contínua és vital per a un sistema que pretén oferir suport rellevant en un context canviant com una pandèmia.

Presentar tota aquesta informació actualitzada de forma entenedora a l'usuari és clau. El dashboard interactiu, alimentat per l'API que consulta la base de dades SQLite, reflecteix aquestes actualitzacions gairebé en temps real. Això significa que quan els usuaris accedeixen al sistema, visualitzen els indicadors, les prediccions i les recomanacions més recents, basades en les últimes dades disponibles, ja siguin obtingudes automàticament de la web o carregades pels propis estats. Això democratitza l'accés a la informació més actualitzada i facilita la presa de decisions informades i oportunes.

Finalment, assegurar la fiabilitat i continuïtat del sistema implica un manteniment actiu que va més enllà dels backups. Crucialment, inclou el reentrenament setmanal dels models Prophet. Aquest cicle de reentrenament programat és fonamental perquè els models s'adaptin a les dinàmiques evolutives de la pandèmia, incorporant les dades més recents ingressades al llarg de la setmana per mantenir i millorar la seva precisió predictiva. El monitoratge continu del sistema complementa aquest procés, assegurant que tant la ingestió de dades, les prediccions basades en els models actualitzats, com la generació de recomanacions funcionin correctament. Totes aquestes funcions són intrínseques a un entorn de producció dinàmic i van molt més enllà de la generació inicial del model.

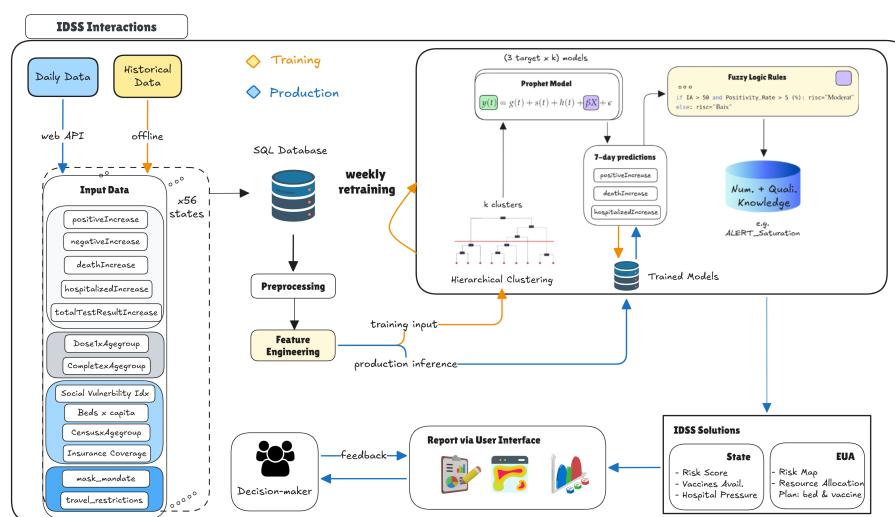


Figure 13.1: Esquema de les Interaccions de l'IDSS en Entrenament i Producció

VALIDACIÓ DE L'IDSS

La validació d'un Sistema Intel·ligent de Suport a la Decisió (IDSS) és un procés crític que assegura la seva fiabilitat, efectivitat i alineació amb les necessitats dels usuaris i els objectius del sistema. Aquest procés engloba l'avaluació individual de cada component, les interaccions entre ells, la gestió de les entrades i sortides, i el rendiment global del sistema en el context de la gestió de pandèmies.

14.1 AVALUACIÓ DEL SISTEMA D'AJUDA A LA DECISIÓ INTEL·LIGENT (IDSS)

La validació del nostre IDSS s'ha enfocat en diversos nivells per garantir una cobertura exhaustiva.

1. Validació de Components Individuals:

- **Models de Dades (Prophet):** El funcionament correcte d'aquests models s'ha validat durant la fase d'entrenament (Secció 10.2) utilitzant mètriques com MAE, RMSE i MAPE en un horitzó de predicció de 7 dies. Els requisits de dades (sèries temporals històriques) i de còmput s'han considerat, i el rendiment actual és acceptable per a l'abast del projecte. Són compatibles amb l'arquitectura, ja que les seves sortides (prediccions) alimenten el motor de regles i la base de dades SQLite.
- **Models de Coneixement (Lògica Difusa):** La validació es va realitzar amb dades artificials que simulen escenaris pandèmics realistes (Taula 3.2, Secció 3.5), demostrant que el motor difús respon adequadament als llindars establerts. La seva arquitectura basada en regles és inherentment compatible i compleix amb els requisits ètics d'explicabilitat (XAI).
- **Mòduls de Preprocessament de Dades:** La correcta transformació i neteja de les dades (Secció 8) s'ha validat assegurant la consistència dels formats de sortida i la capacitat dels components posteriors (models de dades, base de dades) per utilitzar-les sense errors. L'ús de Python (Pandas) i RapidMiner ha estat eficient per a l'actual volum de dades.
- **Backend (API RESTful):** S'ha validat mitjançant proves d'integració, assegurant que les rutes de l'API gestionen correctament les peticions, l'autenticació d'usuaris i les respostes en format JSON. S'han considerat els aspectes de seguretat.
- **Frontend (Interfície d'Usuari):** La validació s'ha centrat en la usabilitat, la correcta visualització de dades, la interacció amb l'API i la robustesa en la gestió d'entrades d'usuari (Secció 15). S'ha dissenyat pensant en els perfils d'usuari definits (Secció 6).
- **Base de Dades (SQLite):** La seva funcionalitat s'ha validat mitjançant la correcta execució d'operacions d'escriptura i lectura, la integritat de les dades i la capacitat de gestionar les consultes necessàries per part de l'API i els serveis de predicció.

2. Validació de les Interaccions entre Components:

- La comunicació entre el frontend, el backend (API), els models de Prophet, el motor de lògica difusa i la base de dades SQLite s'ha dissenyat per ser fluida i amb gestió d'errors. Les interaccions es basen principalment en crides API REST i accés a la base de dades (Secció 13.1, Taules 13.2 i 13.3).
- La latència i el flux de dades són acceptables per a l'escala actual del sistema de prova de concepte. Es consideren aspectes de ciberseguretat bàsics com HTTPS i l'autenticació per a l'accés a l'API.

3. Validació de les Entrades i Sortides (Input/Output):

- La comunicació amb els usuaris es realitza a través de la interfície gràfica (Secció 15), dissenyada per ser intuïtiva i robusta, amb validacions d'entrada per minimitzar errors.
- Les entrades (dades de pandèmia, mètriques de vacunació) i sortides (visualitzacions, nivells de risc, recomanacions) són clares i ben formatades.
- La latència en la presentació de dades després d'una actualització és gestionable, ja que les operacions més costoses (entrenament de models) es fan de forma periòdica i no directament en resposta a una interacció d'usuari.
- Els aspectes ètics es consideren en la presentació de dades agregades, evitant la identificació d'individus (Secció 2.4).

4. Validació del Rendiment Global del Sistema:

Per avaluar el rendiment global, s'han considerat les següents característiques clau i els seus indicadors (KPIs):

- **Precisió en la predicció:**
 - **Descripció:** Mesura la capacitat del sistema per predir correctament l'evolució a curt termini de la pandèmia (e.g., nombre de nous casos, hospitalitzacions) a nivell estatal.
 - **Obtenció:** Diàriament, es compararien les prediccions generades pel sistema per al dia següent (realitzades el dia t per al dia $t+1$) amb les dades reals (ground truth) observades i reportades per cada estat el dia $t+1$. S'utilitzarien mètriques estàndard com l'Error Absolut Mitjà (MAE) o l'Arrel de l'Error Quadràtic Mitjà (RMSE) per quantificar la desviació. (Això complementa la validació dels models Prophet detallada a la Secció 10.2).
 - **Justificació:** Una alta precisió en les prediccions és fonamental, ja que el sistema basa les seves recomanacions i alertes en aquestes estimacions. Aquest KPI indica la fiabilitat de la informació prospectiva que ofereix l'IDSS.
- **Taxa de saturació hospitalària evitada:**
 - **Descripció:** Avalua l'impacte potencial de les mesures recomanades pel sistema (i presumiblement adoptades pels estats) en la prevenció o mitigació de la saturació del sistema hospitalari.

- **Obtenció:** Es compararia la predicción de saturación hospitalaria a 7 días (realitzada el dia t) amb la situació real observada el dia t+7. Per aillar l'efecte de les mesures, idealment es necessitaria una línia base (predicció sense noves intervencions) i comparar-la amb la realitat després de la implementació de les recomanacions. L'anàlisi se centraria en la diferència entre la saturació projectada sense acció i la saturació real amb acció. (Això es relaciona amb les sortides de suport a la decisió de la Secció 5.1.3).
- **Justificació:** Aquest KPI mesura directament un dels objectius crítics de l'IDSS: ajudar a mantenir la capacitat del sistema sanitari. Indica si les intervencions suggerides pel sistema tenen un efecte positiu mesurable en la gestió de recursos crítics.

- **Compliment de les recomanacions:**

- **Descripció:** Percentatge de recomanacions (e.g., tipus de confinament, nivell d'alerta, assignació de recursos) emeses pel sistema que són efectivament adoptades per les autoritats estatals.
- **Obtenció:** A través d'un mecanisme de feedback on els estats, en introduir les seves dades diàries, indicarien quines de les recomanacions del sistema del dia anterior van ser implementades. El KPI seria el ràtio (recomanacions adoptades / total de recomanacions emeses) per estat i agregat. (Això s'alinea amb el mecanisme de feedback esmentat a la Secció 5.3.1).
- **Justificació:** Un alt compliment suggereix que les recomanacions de l'IDSS són percebudes com a rellevants, factibles i útils per les autoritats. Un baix compliment podria assenyalar problemes d'alignació, claredat, confiança o aplicabilitat pràctica de les propostes del sistema.

- **Usabilitat percepuda:**

- **Descripció:** Grau de satisfacció general i percepció d'utilitat del sistema per part dels usuaris finals (e.g., personal de les autoritats sanitàries estatals) en el seu ús quotidià.
- **Obtenció:** Mitjançant un sistema de valoració simple (e.g., una escala d'estrelles d'1 a 5) que els usuaris podrien emplenar diàriament juntament amb la introducció de les dades o la consulta de les recomanacions. Es podria calcular una mitjana de valoració al llarg del temps. (Connecta amb el disseny de la interfície de la Secció 15 i els perfils d'usuari de la Secció 6).
- **Justificació:** La usabilitat és clau per a l'adopció i l'ús continuat i efectiu del sistema, especialment en entorns d'alta pressió com la gestió d'una pandèmia. Una bona percepció d'usabilitat indica que el sistema és intuïtiu, eficient i respon a les necessitats dels usuaris.

5. Proves Finals i Validació Conceptual:

- El sistema ha estat provat utilitzant dades reals de la pandèmia COVID-19 als EUA per a l'entrenament i la validació dels models predictius. S'han utilitzat escenaris

realistes i casos d'ús artificials per validar el motor de lògica difusa i la interacció global del sistema.

- La Prova de Concepte de la interfície (Secció 15) serveix com una validació conceptual amb els usuaris tipus definits, demostrant com es presentaria i s'utilitzaria la informació.

En conclusió, la validació realitzada indica que l'IDSS proposat és capaç de processar dades, generar prediccions, aplicar coneixement expert i presentar resultats de manera coherent i potencialment útil per a la gestió de pandèmies. Tot i que hi ha limitacions inherents a un projecte d'aquesta naturalesa (com l'absència de desplegament real i feedback d'usuaris finals en operació), les proves realitzades sobre els components i el sistema integrat demostren la seva viabilitat conceptual i funcional dins de l'à bast definit.

INTERFÍCIE I PROVA DE CONCEPTE

Una vegada s'ha integrat i avaluat el sistema, com s'especifica a les seccions anteriors, és important presentar els resultats de forma adequada. Això vol dir, conèixer els usuaris del sistema i adaptar tot a les seves capacitats i expectatives. En aquest cas, tal com es defineix en la Sec. 6, qui farà servir el sistema seran gestors d'hospitals, autoritats sanitàries estatals o responsables de coordinar recursos. Aquests perfils no tenen perquè ser hàbils en tecnologia, a més que esperen interpretar els resultats fàcilment, de manera que la aplicació de l'IDSS haurà de tenir-ho en compte.

Per aquesta raó, s'ha considerat imprescindible implementar una interfície gràfica per a la aplicació, a través d'un *front-end* basat en HTML, CSS i JavaScript. D'aquesta manera, s'ha pogut definir cada pantalla de la web detalladament (e.g. l'estil, interacció), connectant les visualitzacions amb les dades del propi IDSS. Això últim, evidentment, a través d'un servei de *back-end* complet, capaç de gestionar totes les dades, prediccions i recomanacions en temps real, tal com s'indica a la Sec. 13.

En aquesta secció, però, s'entra a fons en definir com és la interacció de l'usuari amb l'aplicació, explicant el seu *User journey* a través de la web:

- **Autenticació d'Usuaris** L'inici de sessió es realitza mitjançant nom d'usuari i contrasenya. Es gestionen de forma segura les contrasenyes i les sessions d'usuari, registrant l'hora de l'últim inici de sessió. Un cop autenticat, l'usuari és redirigit al tauler de control principal. Cada usuari està associat a un estat específic, i no es contempla l'acció de registre perquè l'accés d'aquest IDSS hauria de ser limitat.

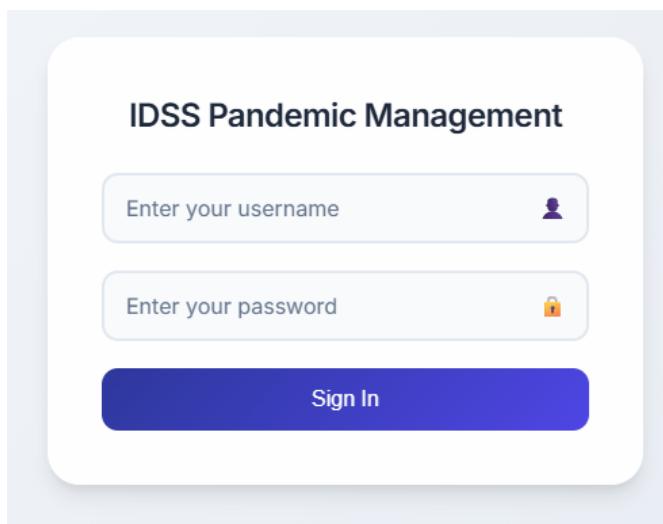


Figure 15.1: Pantalla d'inici de sessió de l'aplicació

- **Interfície del Tauler de Control Principal (Main Dashboard)**

- **Secció de Capçalera:** Inclou un missatge de benvinguda personalitzat amb el nom i la bandera de l'estat de l'usuari. Disposa de botons d'accés ràpid a "Enviar Estadístiques", "Suport a la Decisió" i "Tancar Sessió".
- **Visió General de la Pandèmia:** Presenta un mapa interactiu dels EUA que mostra els nivells de risc pandèmic per estat, codificats per colors: Baix (Verd), Mitjà (Groc), Alt (Taronja), Crític (Vermell). Els estats sota confinament es marquen visualment. En passar el cursor s'ofereixen detalls de l'estat i es pot fer clic per a més informació. Una llegenda explica els nivells de risc.
- **Visualització d'Estadístiques Clau:** Mostra quatre quadres principals amb mètriques nacionals (EUA) i específiques de l'estat de l'usuari per a: Casos Totals, Casos Actius, Casos Recuperats i Morts. Els números estan formatejats per a fàcil lectura i s'actualitzen en temps real.
- **Visualització de Tendències:** Un gràfic interactiu compara els nous casos de l'últim mes (o període configurable) entre la tendència nacional i la de l'estat de l'usuari. Permet obtenir números detallats en passar el cursor, fer zoom, desplaçar-se i seleccionar rangs de dates. El disseny és responsiu.

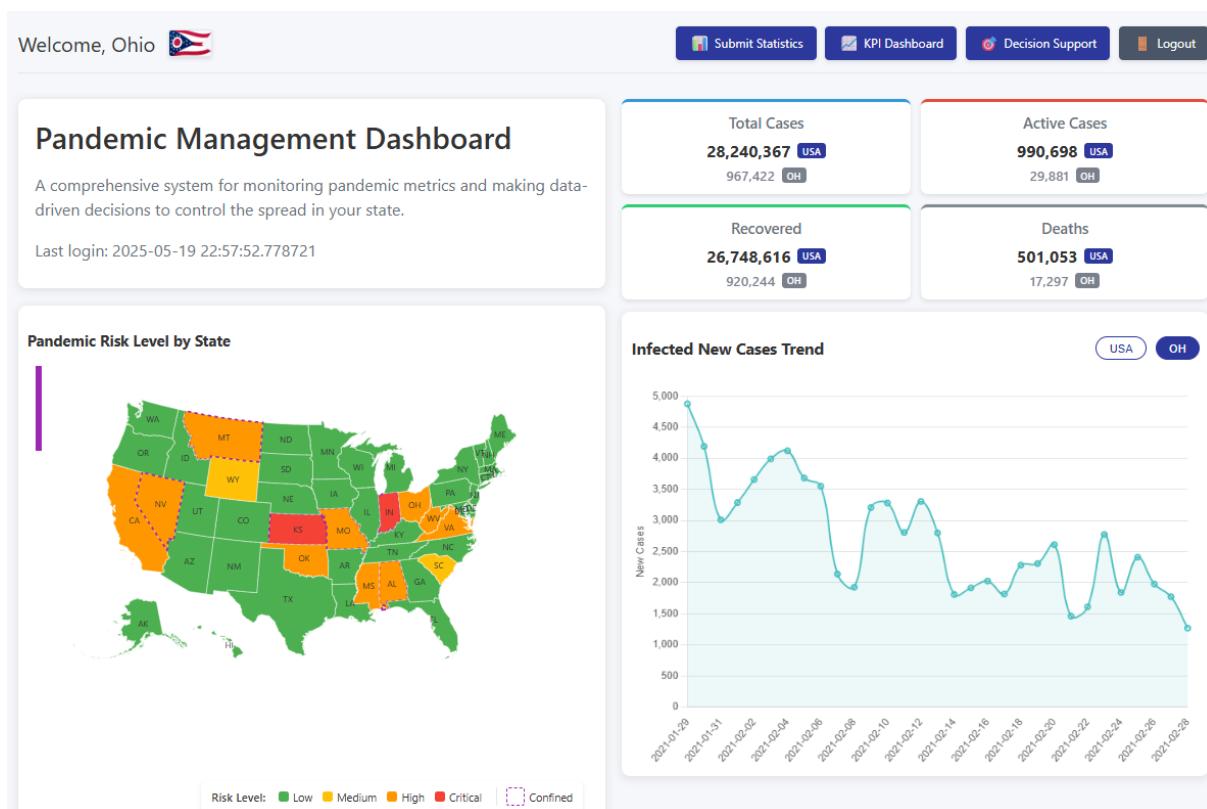


Figure 15.2: Pantalla d'inici de l'aplicació de l'IDSS

- **Enviament d'Estadístiques Diàries**

- **Interfície del Formulari:** L'usuari selecciona una data i introduceix dades per al seu estat a les seccions d'Increments Diaris (nous positius, negatius, proves, morts, hospitalitzacions) i Dades de Vacunació (noves primeres dosis, pautes completades, i el mateix per a majors de 65 anys). Es requereixen valors no negatius i es realitzen càlculs automàtics de percentatges.

Daily Increases

New Positive Cases	New Negative Cases	New Test Results	New Deaths	New Hospitalizations
--------------------	--------------------	------------------	------------	----------------------

Vaccination Data

New First Doses	New Completed Vaccinations	New First Doses (65+)	New Completed Vaccinations (65+)
-----------------	----------------------------	-----------------------	----------------------------------

System Satisfaction

Rate your satisfaction with the system's recommendations and support

★ ★ ★ ★ ★

SUBMIT DAILY STATISTICS

Current Statistics

Active Cases
29,881

First Dose Coverage
10.6%
65+ Coverage: 36.2%

Fully Vaccinated
5.7%
65+ Coverage: 16.8%

Figure 15.3: Pantalla d'enviament de dades diàries de l'IDSS

- **Validació i Procés d'Enviament:** Es valida robustament cada camp, mostrant missatges d'error clars i prevenint enviaments duplicats. Un cop enviat, el sistema actualitza totals, calcula percentatges, mètriques basades en població, confirma l'èxit i pot disparar l'actualització de prediccions i recomanacions.

Daily statistics, satisfaction rating, predictions, and recommendations generated successfully!

Daily Increases

1,000	2,000	3,000	100	200
-------	-------	-------	-----	-----

Vaccination Data

10.6%	5.7%	747,373	346,981	2,062,591
-------	------	---------	---------	-----------

BACK TO DASHBOARD

Figure 15.4: Pantalla de validació de les dades diàries pujades

• Sistema de Suport a la Decisió

- **Visualització de l'Estat Actual:** Un indicador mostra el Nivell de Risc de l'estat (codificat per color, amb indicadors de tendència), la Taxa de Positivitat i una anàlisi de la Tendència dels últims 14 dies.
- **Tauler de Prediccions:** Ofereix una previsió a 7 dies amb desglossament diari per a casos positius, hospitalitzats i morts, mitjançant gràfics interactius amb tooltips i línies de tendència.
- **Secció de Recomanacions:** Proporciona una avaluació del nivell de risc, recomanacions de mesures de confinament, suggeriments per a l'assignació de recursos i objectius de vacunació.

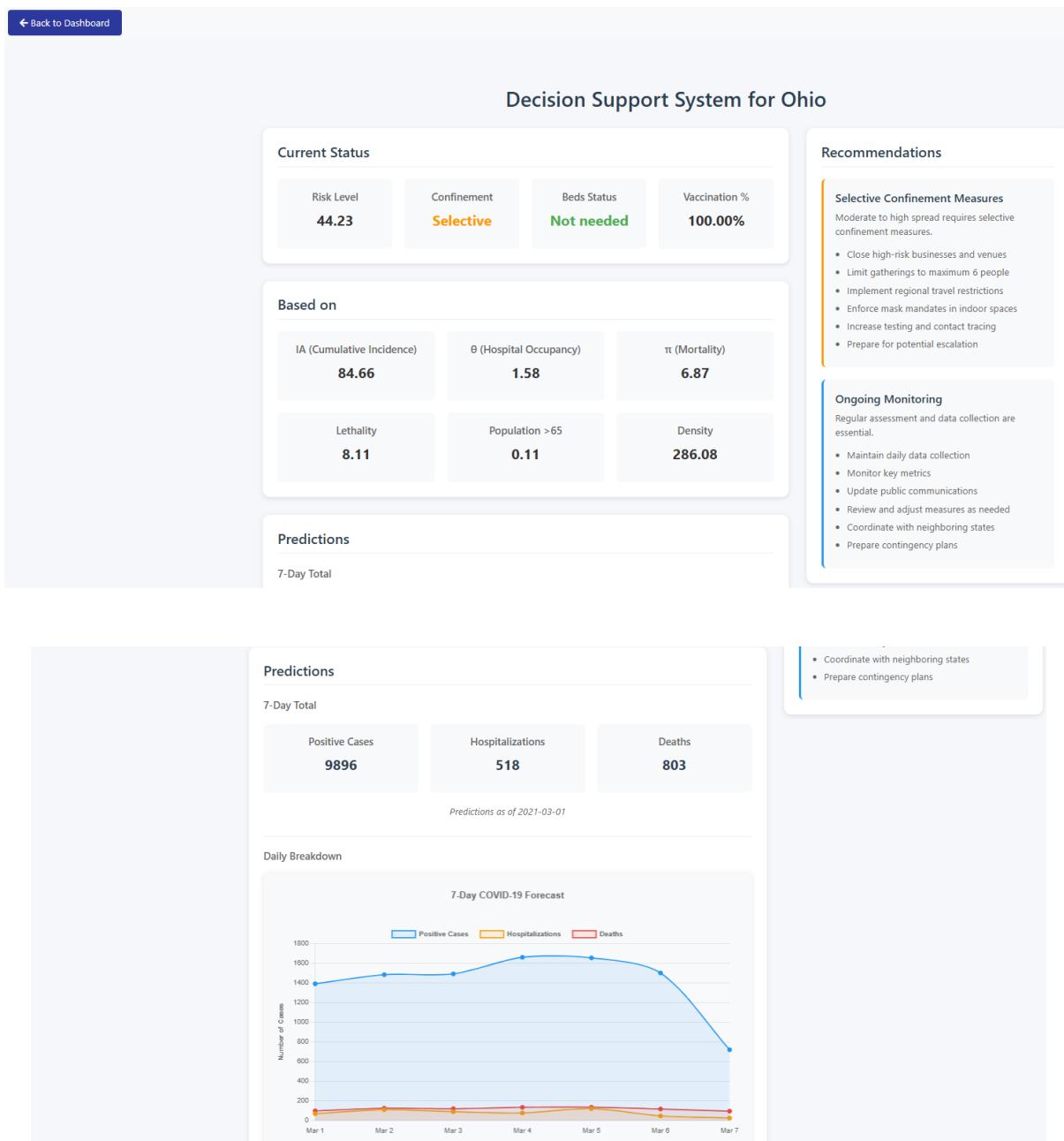


Figure 15.5: Pantalla de suport a la decisió de l'estat d'Ohio.

- Visualització de KPIs

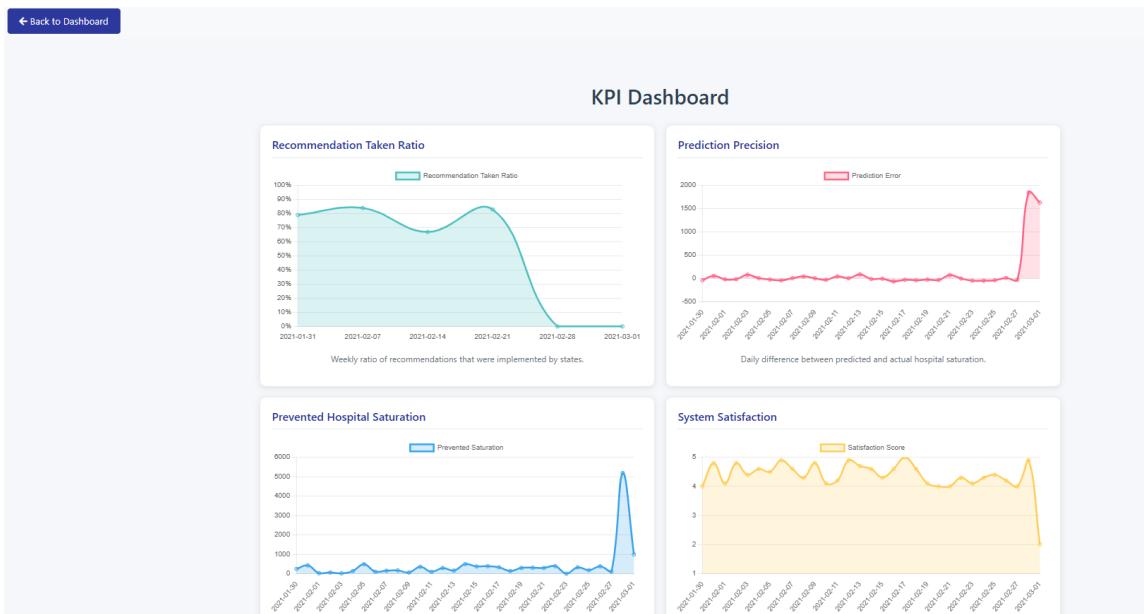


Figure 15.6: Pantalla d'avaluació de l'IDSS amb KPIs.

El sistema està dissenyat per ser intuïtiu i fàcil d'usar, especialment per a professionals de la salut i autoritats sense un perfil tècnic avançat, priorititzant la claredat i la facilitat d'ús mitjançant interaccions directes, retroalimentació visual i tooltips. La gestió de dades es facilita amb una interfície de càrrega de fitxers CSV amb validació automàtica i una pantalla per visualitzar, filtrar, ordenar i exportar dades històriques. El disseny de la interfície és responsiu, modern i accessible, amb actualitzacions dinàmiques en temps real com el refresc de dades i estadístiques en viu.

A nivell de sistema, es garanteix una protecció integral amb gestió segura de contrasenyes, control d'accés basat en rols, validació d'entrades i transmissió segura de dades (HTTPS). Es posa un fort èmfasi en l'accessibilitat, amb opcions d'alt contrast, tipografia llegible, navegació per teclat i suport per a lectors de pantalla. A més, el sistema compta amb una gestió d'errors robusta, oferint missatges clars i accionables, validació, suggeriments de recuperació i registre detallat d'incidències per assegurar la fiabilitat.

LÍNIES FUTURES

El desenvolupament d'aquest IDSS obre la porta a diverses línies de millora i expansió futura, amb l'objectiu d'augmentar la precisió, l'eficiència i l'impacte del sistema en la gestió de pandèmies. Les propostes següents es basen en les limitacions identificades i les oportunitats emergents:

- **Integració de Fonts de Dades Multimodals** Ampliar les fonts de dades amb informació en temps real com dades de mobilitat (e.g., Google Mobility Reports), dades climàtiques (temperatura, humitat) i dades genòmiques del virus per detectar variants emergents. Això enriquiria les prediccions i permetria adaptar les estratègies a factors ambientals o biològics.
- **Models Predictius Avançats** Incorporar arquitectures d'aprenentatge profund (e.g., Transformers, XGBoost) per capturar relacions no lineals i patrons complexos. L'ús d'AutoML (Automated Machine Learning) optimitzaria la selecció d'hiperparàmetres i reduiria el temps de desenvolupament.
- **Personalització de Recomanacions** Incloure variables socioeconòmiques (e.g., índexs de pobresa, accés a serveis bàsics) per ajustar les mesures de confinament i vacunació a les necessitats específiques de cada comunitat, millorant l'equitat en la resposta sanitària.
- **Sistema de Retroalimentació Adaptativa** Habilitar un mecanisme perquè els usuaris (e.g., hospitals, governs) puntuin l'efectivitat de les recomanacions rebudes. Aquesta informació s'utilitzaria per reentrenar els models i ajustar els llindars de risc dinàmicament.
- **Adaptació a Altres Malalties Infeccioses** Reutilitzar l'arquitectura de l'IDSS per a la gestió d'altres brots (e.g., grip aviària, Ebola), adaptant les variables epidemiològiques i les fonts de dades segons el patogen.
- **Optimització Computacional i Escalabilitat** Migrar el sistema a entorns de cloud computing (e.g., AWS, Google Cloud) per gestionar grans volums de dades en temps real i garantir escalabilitat. L'ús de containers (Docker) i orquestració (Kubernetes) simplificaria el desplegament i manteniment.

Aquestes millores no només consolidarien el IDSS com una eina essencial en la lluita contra pandèmies, sinó que també el convertiran en un model versàtil i adaptable per a la gestió de crisis sanitàries futures, sempre amb un enfocament ètic, inclusiu i centrat en l'ésser humà.

BIBLIOGRAPHY

- [1] E. National Academies of Sciences and Medicine, *Framework for Equitable Allocation of COVID-19 Vaccine*. The National Academies Press, 2020. doi: [10.17226/25917](https://doi.org/10.17226/25917) (cit. on p. 7).
- [2] S. Wu *et al.*, ‘Aggressive containment, suppression, and mitigation of covid-19: Lessons learnt from eight countries’, *BMJ*, vol. 375, e067508, 2021. doi: [10.1136/bmj-2021-067508](https://doi.org/10.1136/bmj-2021-067508) (cit. on pp. 7, 8).
- [3] E. Gallic, M. Lubrano and P. Michel, ‘Optimal lockdowns for covid-19 pandemics: Analyzing the efficiency of sanitary policies in europe’, *Journal of Public Economic Theory*, vol. 23, no. 6, pp. 1115–1138, 2021. doi: [10.1111/jpet.12556](https://doi.org/10.1111/jpet.12556) (cit. on p. 7).
- [4] W. H. O. (WHO), *Considerations for implementing and adjusting public health and social measures in the context of covid-19*, 2020 (cit. on p. 7).
- [5] E. C. for Disease Prevention and C. (ECDC), *Covid-19 surveillance guidance*, 2021 (cit. on p. 7).

APPENDIX A: MATRIUS DE DADES INICIALS

Table A.1: Variable descriptions for the *us_daily* dataset.

Variable	Type	Missing %
date	Categorical	0.0 %
states	Numerical	0.0 %
positive	Numerical	0.24 %
negative	Numerical	11.43 %
pending	Numerical	12.14 %
hospitalizedCurrently	Numerical	15.24 %
hospitalizedCumulative	Numerical	12.14 %
inIcuCurrently	Numerical	17.38 %
inIcuCumulative	Numerical	17.14 %
onVentilatorCurrently	Numerical	17.14 %
onVentilatorCumulative	Numerical	18.81 %
death	Numerical	6.67 %
hospitalized	Numerical	12.14 %
totalTestResults	Numerical	0.0 %
recovered	Numerical	100.0 %
total	Numerical	0.0 %
posNeg	Numerical	0.0 %
deathIncrease	Numerical	0.0 %
hospitalizedIncrease	Numerical	0.0 %
negativeIncrease	Numerical	0.0 %
positiveIncrease	Numerical	0.0 %
totalTestResultsIncrease	Numerical	0.0 %
dateChecked	Categorical	0.0 %
lastModified	Categorical	0.0 %
hash	Categorical	0.0 %

Table A.2: Variable descriptions for the *states_daily* dataset.

Variable	Type	Missing %
date	Categorical	0.0 %
state	Categorical	0.0 %
positive	Numerical	0.90 %
probableCases	Numerical	55.38 %
negative	Numerical	36.04 %
pending	Numerical	89.71 %
totalTestResultsSource	Categorical	0.0 %
totalTestResults	Numerical	0.80 %
hospitalizedCurrently	Numerical	16.56 %
hospitalizedCumulative	Numerical	40.41 %
inIcuCurrently	Numerical	44.00 %
inIcuCumulative	Numerical	81.77 %
onVentilatorCurrently	Numerical	56.08 %
onVentilatorCumulative	Numerical	93.79 %
recovered	Numerical	42.24 %
death	Numerical	4.09 %
hospitalized	Numerical	40.41 %
totalTestsViral	Numerical	30.14 %
negativeTestsViral	Numerical	75.82 %
positiveCasesViral	Numerical	31.44 %
deathConfirmed	Numerical	54.43 %
deathProbable	Numerical	63.46 %
totalTestEncountersViral	Numerical	74.83 %
totalTestsPeopleViral	Numerical	55.82 %
totalTestsAntibody	Numerical	76.95 %
totalTestsPeopleAntibody	Numerical	89.41 %
positiveTestsPeopleAntibody	Numerical	94.74 %
negativeTestsPeopleAntibody	Numerical	95.32 %
positiveTestsAntigen	Numerical	89.25 %
positiveIncrease	Numerical	0.0 %
negativeIncrease	Numerical	0.0 %
total	Numerical	0.0 %
totalTestResultsIncrease	Numerical	0.0 %
posNeg	Numerical	0.0 %
deathIncrease	Numerical	0.0 %
hospitalizedIncrease	Numerical	0.0 %

Table A.3: Variable descriptions for the *states_info* dataset.

Variable	Type	Missing %
fips	Numerical	0.0 %
pum	Binary	0.0 %
state	Categorical	0.0 %
notes	Categorical	0.0 %
covid19Site	Categorical	0.0 %
covid19SiteSecondary	Categorical	8.93 %
covid19SiteTertiary	Categorical	26.79 %
covid19SiteQuaternary	Categorical	51.79 %
covid19SiteQuinary	Categorical	76.79 %
twitter	Categorical	1.79 %
covid19SiteOld	Categorical	0.0 %
covidTrackingProjectPreferredTotalTestUnits	Categorical	0.0 %
covidTrackingProjectPreferredTotalTestField	Categorical	0.0 %
totalTestResultsField	Categorical	0.0 %
pui	Categorical	0.0 %
name	Categorical	0.0 %

Table A.4: Variable descriptions for the *states_pop* dataset.

Variable	Type	Missing %
state	Categorical	0.0 %
state_name	Categorical	0.0 %
geo_id	Categorical	0.0 %
population	Numerical	0.0 %
pop_density	Numerical	1.92 %

Table A.5: Variable descriptions for the *states_agegr_pop* dataset.

Variable	Type	Missing %
state	Categorical	0.0 %
state_name	Categorical	0.0 %
geo_id	Categorical	0.0 %
population	Numerical	0.0 %
agegroup	Categorical	0.0 %
pct_pop	Numerical	0.0 %

Table A.6: Variable descriptions for the *counties_pop* dataset.

Variable	Type	Missing %
state	Categorical	0.0 %
county	Categorical	0.0 %
state_name	Categorical	0.0 %
geo_id	Categorical	0.0 %
population	Numerical	0.0 %
pop_density	Numerical	2.42 %

Table A.7: Variable descriptions for the *insurance_cov* dataset.

Variable	Type	Missing %
geo_id	Categorical	0.0 %
state	Categorical	3.84 %
state_name	Categorical	0.0 %
acs_variable	Categorical	0.0 %
estimate	Numerical	0.0 %
margin_of_error	Numerical	0.0 %
label	Categorical	0.0 %
concept	Categorical	0.0 %
estimate_type	Binary	0.0 %
coverage_type	Categorical	0.0 %
age_group	Categorical	0.0 %
labor_force	Binary	29.41 %
employed	Binary	52.94 %

Table A.8: Variable descriptions for the *beds_per_capita* dataset.

Variable	Type	Missing %
state	Categorical	0.0 %
state_name	Categorical	0.0 %
state_local_government	Numerical	15.68 %
non_profit	Numerical	0.0 %
for_profit	Numerical	9.80 %
total	Numerical	0.0 %

Table A.9: Variable descriptions for *vaccinations*.

Variable	Type	Missing %
State	Categorical	0.0 %
County	Categorical	0.0%
Date	Categorical	0.0 %
Dose1_Total	Numerical	0.30 %
Dose1_Total_pct	Numerical	0.24 %
Dose1_65Plus	Numerical	0.32 %
Dose1_65Plus_pct	Numerical	0.24 %
Complete_Total	Numerical	0.24 %
Complete_Total_pct	Numerical	0.24 %
Complete_65Plus	Numerical	0.24 %
Complete_65Plus_pct	Numerical	0.24 %
SVI_CTGY	Categorical	1.85 %
Metro_status	Categorical	1.85 %
Census2019	Numerical	1.79 %
Census2019_65Plus	Numerical	100 %

En el dataset de vacunacions inicialment hi ha 80 variables, però hem decidit posar a la taula A.9 només aquelles que farem servir directa o indirectament en el projecte.

APPENDIX B: MATERIAL DE DESENVOLUPAMENT

WORKSHEET 1: THE MODELS OF MY IDSS

Table B.1: Equivalències Worksheet 1: The models of my IDSS

Question	Report Reference
1. Which models am I using in the training phase? What result does each of these models produce, and what inputs do they work with?	Sec. 5.1.2 (Models de Dades, Models de Coneixement); Sec. 5.1.1 (Dades d'Input); Sec. 5.1.3 (Dades d'Output).
2. What is the result that will be used in the IDSS?	Sec. 5.1.3 (Dades d'Output); Sec. 4.1 (Definició Formal, Tab. 4.1); Sec. 5.3.2 (Comunicació dels Resultats).
3. Which models interact with each other? Draw a diagram showing how the different models interact with each other. (Do they share input variables? Do they feed into each other?...).	Sec. 5.2 (Integració i Estructura); Fig. 5.4 (Arquitectura del Sistema IDSS).
4. Once the flow... is clear, it's time to separate the training phase from the production phase. Describe in detail the format of the result of the training process for each model used (e.g., linear regression generates a linear equation, Apriori model generates a base of association rules, etc.).	Sec. 5.1.2 (Definició); Sec. 5.1 (Disseny de Models: Entrenament i Validació); Tab. 5.1 (Sec. 5.1.1); Sec. 5.1.4 (Ús en Producció).
5. In your design from point 3, replace all trained models with their resulting equivalent components.	Fig. 5.4 (Arquitectura del Sistema IDSS); Sec. 5.1.4 (Ús en Producció).
6. Evaluate how the inputs of each model need to be modified in this new diagram... add... new functions... for the system to work when put into production (do you need to collect data directly from a sensor? From an open database with an API?... do these raw data need to be transformed...?).	Sec. 5.1.4 (Ús en Producció); Sec. 8 (Preprocessament de Dades); Tab. 5.1 (Sec. 5.1.1); Fig. 8.6.
7. What permissions do you need to use the data required in the real application? What permissions do you currently have?	Sec. 2.4 (Consideracions Ètiques).

Continuació a la següent pàgina

Table B.1: Equivalències Worksheet 1: The models of my IDSS (Continuació)

Question	Report Reference
8. What interaction is there with the user? How does the IDSS communicate with the user?	Sec. 5.3.2 (Comunicació dels Resultats); Sec. 6 (Tipus d'Usuari); Fig. 8.6.

WORKSHEET 2: THE DATA SOURCES IN AN INTELLIGENT DECISION SUPPORT SYSTEM

Table B.2: Equivalències Worksheet 2: The data sources in an IDSS

Question	Report Reference
1. Does it support a single decision or multiple decisions?	Sec. 4.1 (Definició Formal, Tab. 4.1).
2. What type of data is relevant? Does it make sense to incorporate other data sources...? Images? Audio? Websites? Video? Etc. Justify your decisions.	Sec. 2.1 (Rellevància de Dades).
3. What are their origins? Where are they located? Do we already have them? Do we still need to collect them? If so, what data collection method...?	Sec. 2.3 (Especificació de les Bases de Dades); Tab. 5.1 (Sec. 5.1.1).
4. Ethical considerations: Are these secondary uses of data? What permissions are required? How can we obtain those permissions? What kind of use can we make of the data?	Sec. 2.4 (Consideracions Ètiques).
5. What uncertainty do the data carry? Is it sample data? Is it population data? How do I manage uncertainty? Do my data represent the entire population? Which part... is not represented...? What biases appear?	Sec. 2.5 (Incertesa de les Dades).
6. Does this system work with dynamic data? Does it incorporate data in real time?	Sec. 5.1.4 (Ús en Producció); Sec. 2.3.
7. What data are used to train the system? How do you train the system? How do I deploy the system into production? Are the data the same in training and testing? What variables are used in each process?	Sec. 5.1 (Disseny de Models: Entrenament i Validació); Tab. 5.1 (Sec. 5.1.1); Sec. 5.1.2 (Definició); Sec. 5.1.4 (Ús en Producció).
8. What preprocessing is applied to the data? Is it the same in training and production? What model do you use to impute missing data...? Do you use the same imputation model...?	Sec. 8 (Preprocessament de Dades) (incl. 8.1, 8.2).
9. Draw the preprocessing flow in training and production, and how it fits with the inputs to the IDSS modules you have already designed.	Fig. 5.4 (Arquitectura del Sistema IDSS); Fig. 8.1; Sec. 8.
10. In addition, the metadata document is very important. What elements will it contain? How will the metadata be used throughout the entire process?	Sec. 9.1 (Metadades); Tab. 9.1.

WORKSHEET 3: DESIGN THE ARCHITECTURE OF AN INTELLIGENT DECISION SUPPORT SYSTEM

Table B.3: Equivalències Worksheet 3: Design the architecture of an IDSS

Question / Instruction	Report Reference
1. Formal definition of the decisions to be supported by the system... (Table format)	Sec. 4.1 (Definició Formal, Tab. 4.1).
2. Distinguish between training and production steps (Referencing Fig 1, Fig 2, Fig 3 from worksheet)	Sec.5.1 (Disseny de Models: Entrenament i Validació); Tab. 5.1; Sec. 5.1.4 (Ús en Producció). Esquema propi: 5.4).
(Instruction) Start the design... by retrieving your Canvases...	Appendix B, Fig. B (Lean canvas).
(Instruction) Describe which data-driven model results will your system provide	Sec. 5.1.3 (Dades d'Output).
(Instruction) Think how you would integrate the results... into a software... production system... For each data-driven model describe:	*(Mirar sub-preguntes posteriors)*
SubQ1: Which are the input data required to train... And which input data are required to run... in production. (Table format)	Sec. 5.1.1 (Dades d'Input), Tab. 5.1.
SubQ2: Nature/form of the data-driven model induced...? Software? Equation? Graph? Other?	Sec. 5.1.2 (Definició).
SubQ3: How is the model used in production? Equation to be computed? Visualization? Set of rules?	Sec. 5.1.4 (Ús en Producció); Sec. 5.3.2 (Comunicació dels Resultats).
SubQ4: Which Will be the result? A numerical prediction? A graph? A qualitative diagnoses? A list of....	Sec. 5.1.3 (Dades d'Output).
SubQ5: How is the result helping to support the target decision?	Sec. 4.2 (Implicació Operativa).
SubQ6: How Will be the result be communicated to the stakeholder? A number is given? Several numbers conveniently sorted? If so, which will be the sorting criterion?	Sec. 5.3.2 (Comunicació dels Resultats).

Continuació a la següent pàgina

Table B.3: Equivalències Worksheet 3: Design the architecture of an IDSS (Continuació)

Question / Instruction	Report Reference
7. Design an schematic architecture for your system	Fig. 5.4 (Arquitectura del Sistema IDSS).
8. While the system works, the use of this components adapts. Which is the policy of training? On-line. Off-line, etc Describe how to implement this return: When should it be done, who or what should be done, which is the use that the IDSS will do from the results, how to measure if the decisions recommended... were followed and... appropriate or not	Sec. 5.1.4 (Ús en Producció); Sec. 5.3.1 (Monitorització del Sistema); Final de Sec. 5.3.2.
9. In the delivery, include answers... Do you detect the need of other pieces...? Please design the components... Include all required (map tools, periodic reports, other data models, knowledge models, optimization models, statistical models, cost models...)	Fig. 5.4 (Arquitectura del Sistema IDSS); Sec. 5.1.2 (Definició).
10. Whit all the pieces properly designed draw how to integrate among them to constitute the final system	Fig. 5.4 (Arquitectura del Sistema IDSS).
11. Make an inclusive design	Sec. 5.2.1 (Inclusive Design).

Dimensió	Comentaris
Tecnologies	R, Python, DBMS, Protegé.
Riscos	Tota decisió vinculable a un risc ha d'estar supervisada per un expert en l'àrea sanitaria. Aquest freqüentment no excusa en cap circumstància la responsabilitat associada a l'aplicació del IDSS
Dades	Les dades sobre l'evolució del covid es troben en 9 dataset diferents, disposen de dades numèriques, categòriques i binàries. Aquest dataset proporciona informació de la pandèmia als Estats Units, així com d'aspects socials i demogràfics. Pel que fa a la vacunació tenim un altre dataset amb informació de la distribució als Estats Units.
Referències	The COVID Tracking Project, GitHub: https://github.com/singer-io/tap-covid-19?tab=readme-ov-file
Viabilitat	La viabilitat del sistema depèn de la fiabilitat de les dades, la seva actualització i fonts .També depèn dels requeriments esmentats al AI Council Canvas.

Figure B.1: Anàlisi de dimensions

Projecte: IDSS Estratègia de Gestió de Pandèmies

Carla Atienza, Mireia Brichs, Jordi Granja, Aleix Ibars, Alberto Jerez i Pau Prat

La tasca i la decisió	Proposta de Valor	Judici (criteri de decisió)	Desplegament
<p>Quina tasca vols analitzar? Quina decisió es pot millorar amb IA? Quin aspecte vols millorar?</p> <p>Analitzar situacions d'emergència sanitària, com la Covid-19, per ajudar i millorar la presa de decisions relacionades amb la gestió de pandèmies:</p> <ul style="list-style-type: none"> • Quins estats confinar/tancar demà? • Es pot aprofitar recursos d'estats propers en cas de saturació? 	<p>Quin valor proporciona una solució basada en IA?</p> <p>Reduir el nombre de contagis que poden acabar en mort.</p>	<p>Quin preu estic disposat a pagar pels errors del sistema?</p> <p>Pagar més pels errors de falsos positius que pels falsos negatius, priorititzant la protecció de la salut pública. Calibrar models per ser més agressius en la detecció de riscos i aplicació de mesures.</p>	<p>Què em falta per posar en marxa aquest projecte?</p> <ul style="list-style-type: none"> • Configuració i establiment de l'entorn de treball • Infraestructura de desplegament • Infraestructura de streaming de dades • Canal de comunicació amb experts • Suport institucional / empresarial.
Riscos	Dades	Resultat	Barreres
<p>Quins són els principals riscos?</p> <p>L'aplicació de decisions errònies en un context sanitari d'emergència. Un sistema completament automatitzat podria prendre decisions sense considerar factors socials o polítics importants.</p>	<p>Tenim accés a les dades necessàries? Es té accés a 9 datasets provinents del Covid Tracking Project i un dataset de vacunació (veure la taula d'abaix). Quant a coneixement, accedim a articles científics que han estudiat la pandèmia.</p>	<p>Com podem mesurar l'èxit del projecte?</p> <p>S'utilitzarà KPI per mesurar l'efectivitat del sistema de predicció de confinaments i de distribució de vacunes en la reducció de la taxa de transmissió</p>	<p>Quins aspectes de l'organització poden posar en perill el projecte?</p> <p>Aspectes legals, transparència i disponibilitat de les dades en temps real, mesura correcta i precisa de les dades sanitàries, infraestructura de suport.</p>
<p>Comentaris: El Canvas està subjecte a modificacions, conforme s'estudiïn les dades i diferents possibilitats. Com no hi ha un client final, a mesura que s'interacció amb el sistema es definiran els objectius més viables, ampliant els riscos, barrers i desplegament.</p>			

Figure B.2: Lean canvas.

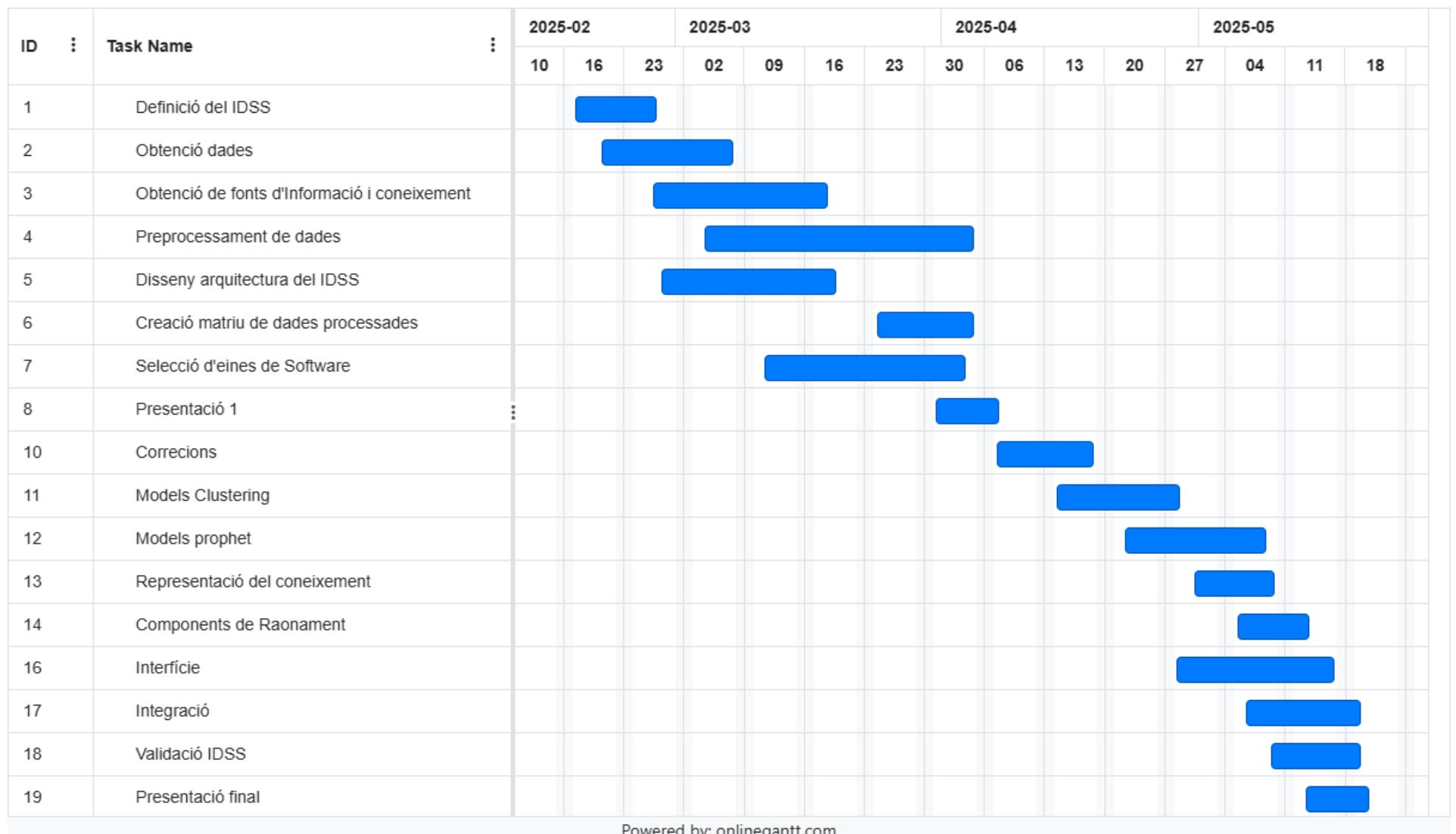


Figure B.3: Gantt.

Repartició de tasques:

Tasca	Alberto	Aleix	Carla	Jordi	Mireia	Pau
Definició del IDSS	X	X	X	X	X	X
Obtenció dades	X		X	X	X	
Obtenció de fonts d'informació concretament		X				X
Preprocessament de dades	X		X	X	X	
Disseny arquitectura del IDSS	X			X		
Creació matriu de dades processades			X		X	
Selecció d'eines de Software		X				X
Presentació 1	X	X	X	X	X	X
Correccions	X	X	X	X	X	X
Models Clustering	X					
Models Prophet			X	X	X	
Representació del coneixement		X				X
Components de Raonament		X				X
Interfície	X	X			X	
Integració				X		
Validació IDSS			X			
Presentació final	X	X	X	X	X	X