



IA per Turismo

Informe final



CIDAI-PAI 03-2021-DO4

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

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

En qualsevol ús d'aquest treball, no s'ha de suggerir que el CIDAI doni suport a cap organització, producte o servei específic. No es permet l'ús del logotip CIDAI. Si adapteu l'obra, heu de llicenciar-la amb la mateixa llicència Creative Commons o equivalent.

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

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

Cita suggerida. CIDAI-PAI 03-2021-DO4 // IA per Turisme – Informe Final per Aina Pedret - Ajuntament de Barcelona; Alberto Pinedo – Microsoft; Esteban Galli, Joan Borràs, Mohammed Mazroua, Salvador Anton Clavé – Eurecat; Javier Ibáñez - CaixaBank Research; Karina Gibert, Mónica Sánchez – IDEAI-UPC, Shutao Zheng, Vicente Gardulski – SDG Group, Verónica Enríquez, Mercè Escrichs – Generalitat de Catalunya, CIDAI 2021.

Llicència: [CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/)

Entitats participants

El projecte s'ha dut a terme amb la participació de les següents entitats:

 Centre Tecnològic de Catalunya	Coordinació del projecte i tractament de les dades de transaccions bancàries i procés de clustering de visitants
	Tractament de les dades d'opinions de TripAdvisor
	Servei API de tractament de text
 UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH Intelligent Data Science and Artificial Intelligence Research Center	 Conceptualització via enginyeria del coneixement dels conceptes de viatge i destinació per als clusterings de visitants i destinacions, especificacions d'ètica en IA i procés de clustering de destinacions
 Ajuntament de Barcelona	 Definició de requisits i validació de resultats
 Generalitat de Catalunya Departament de la Vicepresidència i de Polítiques Digitals i Territori	Definició de requisits i validació de resultats
	Cessió de dades, definició de requisits i validació de resultats i validació de resultats
	Definició de requisits i validació de resultats

Autors

El projecte s'ha dut a terme amb la participació dels següents experts:

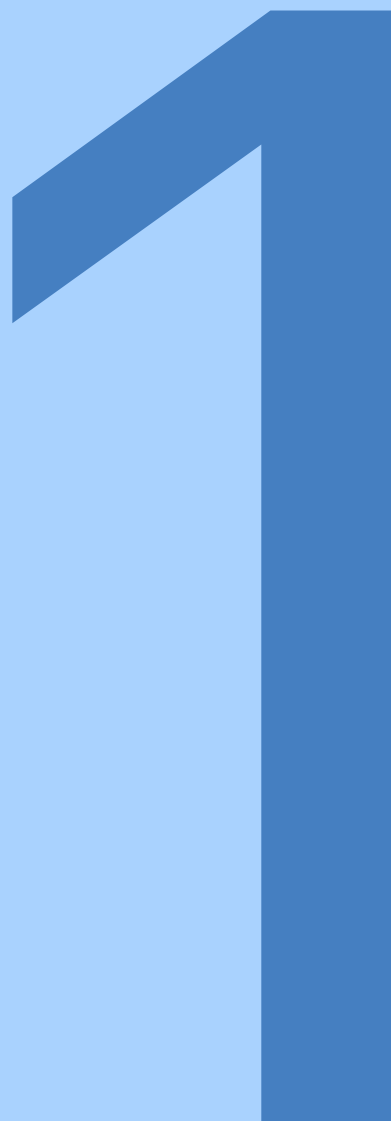
Nom	Entitat	Participació en el projecte
Aina Pedret	Ajuntament de Barcelona (Direcció de Turisme, Esdeveniments i Indústries Creatives)	Experta en coneixement del sector turístic.
Alberto Pinedo	Microsoft	Suport en l'ús de l'API de tractament de text
Esteban Galli	Eurecat (Data Science and Big Data Analytics)	Execució d'anàlisis de dades de transaccions bancàries i clusterització de visitants.
Javier Ibáñez	CaixaBank Research	Extracció i preparació de les dades originals, i validació de resultats. Expert en el coneixement de despesa turística.
Joan Borràs	Eurecat (Data Science and Big Data Analytics)	Coordinador de projecte i participació en el disseny i execució d'anàlisis de dades de transaccions bancàries i clusterització de visitants.
Karina Gibert	IDEAI-UPC (Directora i Catedràtica de Intel·ligència Artificial i Ciència de les Dades)	Enginyeria del coneixement i ciència de les dades i Intel·ligència Artificial eines de suport a la interpretació automàtica de perfils
Mónica Sánchez	IDEAI-UPC (Professora titular d'universitat)	Matemàtica i experta en Intel·ligència Artificial i eines multicriteri i d'aprenentatge automàtic
Mercè Escrichs	Generalitat (Departament d'Empresa i Coneixement)	Experta en coneixement del sector turístic.
Mohammed Mazroua	Eurecat (Consultoria Tecnològica)	Expert en coneixement del sector turístic.
Salvador Anton Clavé	Eurecat (Director d'Innovació Turística) / URV (Catedràtic d'Anàlisi Geogràfica Regional)	Expert en coneixement del sector turístic.
Shutao Zheng	SDG Group	Execució del tractament de dades d'opinions.
Verònica Enríquez	Generalitat (Departament d'Empresa i Coneixement)	Experta en coneixement del sector turístic.
Vicente Gardulski	SDG Group	Execució del tractament de dades d'opinions.

1. Resum Executiu	6
2. Introducció al problema	8
3. Objectius i abast del projecte	10
4. Treball realitzat	12
5. Resultats principals	15
6. Lliurables	27
7. Conclusions	28
8. Lliçons apreses	29
9. Recomanacions futures	30



Resum executiu

Índex



El principal objectiu del projecte ha estat descobrir els patrons de comportament dels visitants a Catalunya i estudiar-ne els canvis en el temps a partir del model de representació del viatge que es pot derivar de l'anàlisi de transaccions bancàries de CaixaBank. Aquesta anàlisi inclou indicadors de consum i comportament dels visitants, com ara la repetició de visites, la despesa per perfil, els fluxos entre municipis, entre altres. Els indicadors s'han pogut segmentar a partir tenint en compte tant el comportament turístic com la mobilitat, i s'han pogut analitzar en moments pre-pandèmia i post-pandèmia, per les diferents marques turístiques de Catalunya.

A més, s'han aplicat tècniques de clusterització en els viatges dels visitants de Catalunya, a partir de les transaccions bancàries. Amb aquest mètode s'han pogut identificar els perfils principals de Catalunya en base als seus patrons de comportament. En una primera fase inicial, s'han obtingut els següents 6 perfils: a) amb presència a Barcelona; b) amb despesa en allotjaments i caixers; c) excursionistes; d) pernoctants; e) de llarga distància; i e) de despesa molt alta. De tots els perfils, se n'han analitzat les seves característiques principals. En una segona fase, s'ha profunditzat en desglossar encara més els 6 perfils principals, obtenint un total de 17 perfils, alguns d'ells d'interès com el d'estada llarga, de despesa en parcs temàtics, repetidors o de primera visita, entre d'altres.

A banda de les transaccions bancàries també s'han analitzat les opinions realitzades als restaurants, allotjaments i atraccions de Catalunya de la plataforma TripAdvisor. En les opinions, s'ha aplicat tècniques d'anàlisi de sentiment, que ha permès conèixer la percepció dels visitants en les diferents marques turístiques. S'ha pogut observar també quines eren les queixes negatives o fets positius que els usuaris tenien en consideració abans de la pandèmia o durant la pandèmia.

Finalment, s'han creuat les dades de transaccions bancàries amb les dades d'opinions de TripAdvisor amb granularitat geogràfica de codi postal. Amb aquesta integració s'ha procedit a clusteritzar les diferents destinacions turístiques de Catalunya amb les dues fonts de dades. Els resultats obtinguts ha permès agrupar aquells municipis turístics que tenen comportaments similars, en base al tipus

de despesa que s'hi efectua i les opinions que s'hi publiquen. En total han sorgit 13 perfils de codis postals, dels quals 4 especialitzen diferents zones de Barcelona i la resta són de territori.

Un dels objectius del projecte també era poder validar les tecnologies aplicades, com ara l'enginyeria del coneixement, o les tecnologies d'Intel·ligència Artificial (anàlisi de sentiment i procediments de clusterització) per a extreure resultats de gran valor per a una millor gestió en la presa de decisions. Aquesta fet ha sigut validat amb experts del sector turístic, i s'ha fet èmfasi en la importància del coneixement turístic per tal de definir la metodologia del tractament de les dades de forma correcta, així com la interpretació dels resultats, obtinguts la definició dels objectius i la correcció dels possibles biaixos derivats de les pròpies fonts de dades.

Introducció al problema

Index



El sector que anomenem de les experiències comprèn el que coneixem com a sector turístic i cultural. La tecnologia aplicada a aquestes dues indústries permet millorar i augmentar la productivitat i la promoció dels seus productes i serveis, conèixer l'oferta de la competència, anticipar el comportament dels potencials clients i millorar la planificació i gestió de les accions dels gestors de la destinació.

Cal destacar que les dinàmiques de l'oferta i els comportaments dels clients són canviants per causa de l'aparició de noves tendències i productes i, sobretot, a conseqüència dels canvis de necessitats i restriccions que han aparegut arran de la pandèmia de la covid-19.

En aquest sentit, les tècniques d'intel·ligència artificial tenen un gran potencial per a anar més enllà dels estudis tradicionals podent extreure coneixement a partir de patrons ocults en el gran volum de dades que es genera a partir de la petjada digital. Aquest coneixement ha de permetre millorar la planificació de les accions de màrqueting, la creació de nous productes, la personalització de les experiències la fidelització dels clients, i, especialment, la gestió i planificació estratègica de destinacions, atractius i equipaments culturals i d'entreteniment, entre d'altres.

En aquest projecte s'ha convidat a participar al Centre d'Excel·lència en Innovació Turística d'Eurecat (CoE (<https://coeintourisminnovation.org/>)). El CoE en Innovació Turística és una iniciativa públicoprivada que té com a objectiu promoure la sostenibilitat, la innovació i la competitivitat del sector turístic. El CoE està promogut per Eurecat amb la col·laboració inicial de la Direcció General de Turisme de la Generalitat de Catalunya, les quatre diputacions catalanes i els ajuntaments de Barcelona i Vila-seca, i també d'empreses líders vinculades al sector (<https://coeintourisminnovation.org/promoters/>).

Per a la realització del projecte ha estat fonamental la participació de CaixaBank, que ha permès l'accés a dades de transaccions bancàries. Aquestes dades permetran conèixer la traçabilitat de despesa dels visitants a Catalunya durant el període dels últims anys, incloent el tipus de despesa realitzada, el valor monetari, la localització per codi postal i el país d'origen del client. L'aplicació d'algoritmes

d'intel·ligència artificial ha de permetre obtenir informació rellevant dels perfils dels visitants basats en els seus patrons de comportament de compra.

Així, per exemple, es pot donar resposta a preguntes, com ara: Què consumeixen els que repeteixen diverses vegades la visita? I els que visiten un sol cop? I els que dormen un determinat nombre de dies? I els que fan molts viatges durant l'any respecte als que en fan pocs? Quins són els principals perfils de viatge de Catalunya? Què consumeixen aquests perfils? Quines altres destinacions fan excursions els visitants que pernocten a una destinació concreta?

A banda, també es disposa de dades de comentaris realitzats pels usuaris de Tripadvisor a Catalunya. La informació inclou les valoracions i comentaris a atraccions turístiques i el perfil dels usuaris que realitzen les valoracions. L'aplicació d'algoritmes d'intel·ligència artificial a les dades de text d'opinions ha de permetre respondre a preguntes com ara: Quines són les principals queixes negatives dels diferents perfils o zones? I els aspectes positius? I segons la tipologia de productes o la marca turística?

A més, no es tracten les dades de forma aïllada, si no que es realitzen associacions entre les transaccions bancàries i les opinions a nivell de codi postal. Així doncs, es poden respondre preguntes com ara: En les zones on es reben millors valoracions, la despesa és més alta? I en les que reben més queixes negatives?

Objectius i abast del projecte

Index



3.1. Objectius generals

Els objectius generals del projecte són:

- OG-01: Detectar els canvis de patrons de consum i visita dels visitants.
- OG-02: Generar un conjunt de resultats que permeti als agents una millor gestió en la presa de decisions.
- OG-03: Aplicar algoritmes d'IA a les dades per extreure coneixement de valor a partir de patrons de comportament dels visitants.
- OG-04: Identificar les limitacions de la tecnologia i de les dades per a poder millorar els models de detecció de patrons.
- OG-05: Detectar punts de millora a futur.

3.2. Objectius tècnics

Els objectius tècnics del projecte són:

- OT-01: Desenvolupar processos d'integració i processament de les dades disponibles.
- OT-02: Desenvolupar models de classificació automàtica (clustering) de visitants i de territoris.
- OT-03: Caracteritzar els perfils de visitants i analitzar si hi ha canvi de patrons entre 2019 i 2021, posant especial èmfasi en el seva despesa econòmica i en l'ús del territori.
- OT-04: Caracteritzar els perfils de territoris combinant els perfils de despesa i reputació online.
- OT-05: Aplicar algoritmes de Natural Language Processing, per extreure opinions, temàtiques i sentiments, a partir de les valoracions dels usuaris.

3.3. Abast

L'abast del projecte és la realització d'un estudi pilot que combini les fonts de dades de transaccions bancàries i les opinions de TripAdvisor, per tal de complir amb els objectius previstos anteriorment.

Els resultats obtinguts són de caire exploratori i en format pilot, per tal de veure quina informació de valor es poden extreure de les dades.

L'abast temporal del projecte són dels anys 2018 fins a 2021 per les dades de transaccions bancàries, i de 2019 i 2020 per les dades de TripAdvisor.

L'abast geogràfic dels resultats és a nivell de marques turístiques de Catalunya. En els casos possibles, també es mostren resultats a nivell de codi postal i municipis turístics.

L'exploració de dades de transaccions bancàries incorporen els següents indicadors: despesa, destinacions visitades, fluxos entre destinacions; i les següents segmentacions: pernoctant/excursionista, categoria de negoci consumit, nits d'estada, repetició de visita, origen del consumidor (província dels nacionals i país dels estrangers) i destinació turística visitada.

S'apliquen algoritmes de classificació per tal de perfilar els visitants de Catalunya basat en el comportament detectat a través de les dades de transaccions.

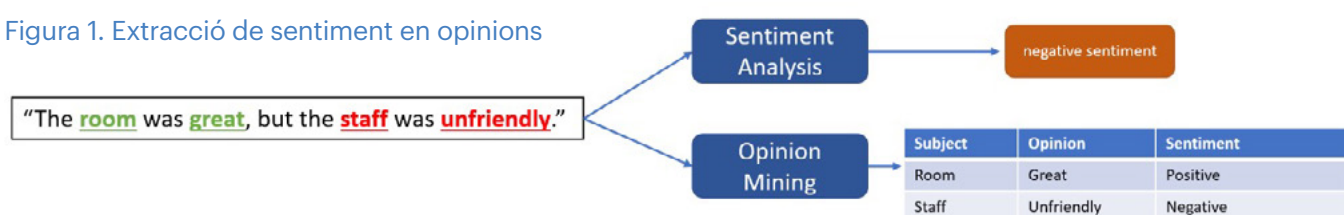
L'exploració de dades de TripAdvisor incorporen els següents indicadors: valoracions, sentiment, freqüència de frases clau, freqüència d'entitats detectades i opinió per concepte (Figura 1); i les següents segmentacions: categories de productes (restaurants, hotels, museus, monuments, etc.), origen de l'usuari, tipus d'usuari (sol, parella, família, etc.) i destinació turística.

S'apliquen algoritmes de classificació per tal d'agrupar destinacions turístiques basat en les opinions que els usuaris han deixat a la destinació, i les transaccions realitzades dels seus visitants.

Els resultats obtinguts permeten comparatives entre els diferents anys i segmentacions rellevants.

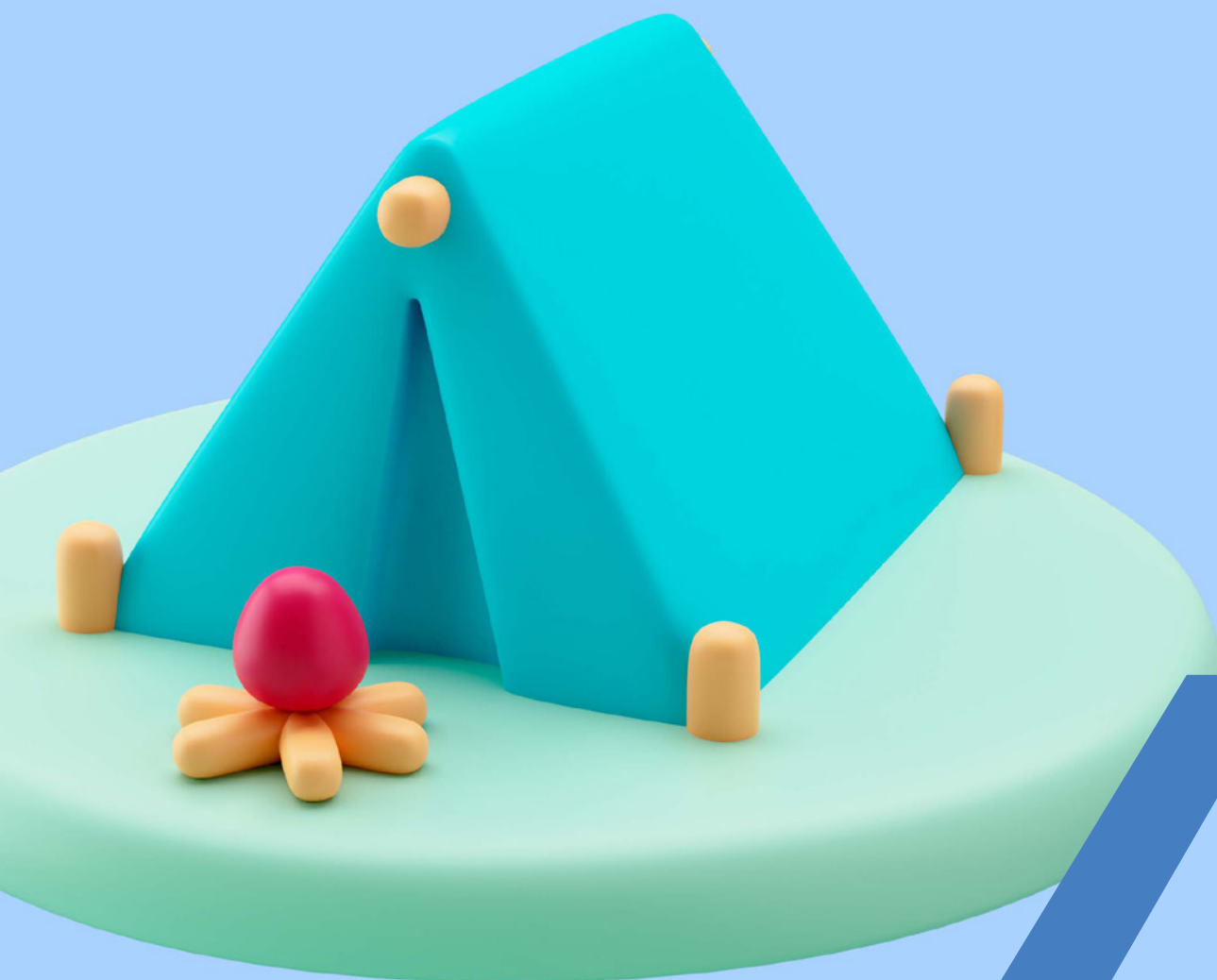
Els resultats segueixen les recomanacions del Parlament Europeu sobre la regulació ètica en la IA.

Figura 1. Extracció de sentiment en opinions



Treball realitzat

Índex



El projecte s'estructura en els 6 paquets de treball (PT) següents:

- PT0: Gestió de projecte
- PT1: Definició de requeriments
- PT2: Tractament de dades transaccions bancàries
- PT3: Tractament de dades d'opinions
- PT4: Integració de dades d'opinions i transaccions bancàries
- PT5: Generació de resultats i conclusions

A continuació es detallen els objectius de cada paquet de treball, les seves tasques i les entitats participants.

PT 0 - Gestió de projecte

El paquet de treball de gestió de projecte tracta de coordinar i realitzar el seguiment de tots els paquets de treball del projecte així com també de les sessions de treball pertinents amb les entitats involucrades.

Tasques	Participants
T0.1. Coordinació i seguiment de les tasques del projecte	Eurecat
T0.2. Coordinació de sessions de treball sobre la metodologia del tractament de dades	Eurecat

PT 1 - Definició de requeriments

L'objectiu principal d'aquest paquet de treball és definir concretament quines hipòtesis es volen resoldre a partir de l'aplicació de la intel·ligència artificial sobre les dades disponibles. Les tasques desenvolupades i els participants, han sigut:

Tasca	Participants
T1.1. Comprensió detallada de negoci i de les dades	Tots
T1.2 Definició de requeriments	Tots

PT 2 - Tractament de dades transaccions bancàries

Aquest paquet de treball es centra en la generació de coneixement sobre els visitants i els viatges, a partir de l'anàlisi de dades de transaccions bancàries de CaixaBank. Els objectius concrets són:

- Auditar i preparar les dades, així com validar la qualitat d'aquestes.
- Generar indicadors sobre les característiques dels visitants i els seus viatges.
- Perfilar viatges en base el seu comportament.

Les tasques desenvolupades i els participants, han sigut:

Tasques	Participants
T2.1. Disseny de la caracterització de viatges	IDEAI-UPC, Eurecat
T2.2. Extracció de dades i auditoria	CaixaBank, Eurecat
T2.3. Processament de dades	Eurecat
T2.4. Anàlisis exploratori de les dades	Eurecat
T2.5. Desenvolupament de models de classificació de viatges	Eurecat

PT 3 - Tractament de dades d'opinions

Aquest paquet de treball es centra en la generació de coneixement a partir de l'anàlisi de dades de d'opinions de TripAdvisor. Els objectius concrets són:

- Auditar i preparar les dades, així com validar la qualitat d'aquestes.
- Generar indicadors sobre les opinions dels usuaris.

Tasques	Participants
T3.1. Auditoria de dades	SDG
T3.2. Preparació de dades	SDG Group, Microsoft
T3.3. Anàlisis exploratori de les dades	SDG Group, Microsoft

PT 4 - Integració de dades d'opinions i transaccions bancàries

Aquest paquet de treball es centra en la generació de coneixement a partir del creuament de dades de transaccions bancàries i d'opinions per zones geogràfiques. Els objectius concrets són:

- Integrar i associar per zones els indicadors principals de transaccions bancàries i opinions.
- Perfilar destinacions en base les opinions que aquestes reben dels usuaris de TripAdvisor i de les característiques dels viatges detectats a través de les transaccions bancàries.

Les tasques desenvolupades i els participants, han sigut:

Tasques	Participants
T4.1. Disseny de la caracterització de perfil turístic de les destinacions	IDEAI-UPC
T4.2. Integració de dades de destinacions	IDEAI-UPC, Eurecat, SDG Group
T4.3. Desenvolupament de models de classificació de les destinacions i conceptualització dels perfils amb eines de suport intel·ligent a la interpretació	IDEAI-UPC

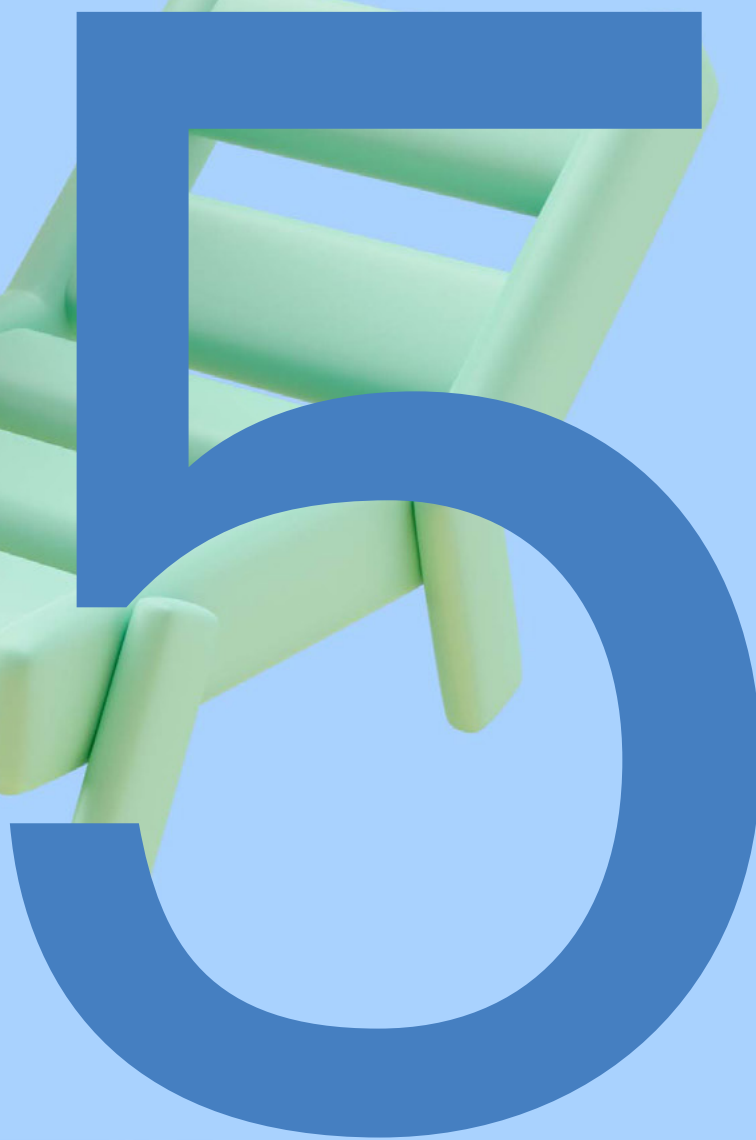
PT 5 - Generació de resultats i conclusions

En aquest últim paquet de treball, els objectius es centren en analitzar, interpretar i validar els resultats obtinguts. Així com també la documentació dels resultats.

Tasca	Participants
T5.1. Anàlisi i validació dels resultats	Tots
T5.2. Generació de l'informe de resultats	Eurecat, SDG, IDEAI-UPC

Resultats principals

Index



5.1. Representació del coneixement per a la definició de l'estructura del viatge

En la tasca "T2.1. Disseny de la caracterització de viatges" s'han utilitzat tècniques d'enginyeria del coneixement per a identificar com representar un viatge a partir de les dades de transaccions bancàries.

La metodologia permet convertir el registre de transaccions, on les transaccions d'una mateixa targeta estan desestructurades i representades de forma independent, en una nova base de dades on la fila representa un viatge amb tota l'activitat de la targeta bancària durant el viatge. Partint d'un model de viatge constituït per seqüències cronològicament ordenades de transaccions, és possible establir els criteris d'agregació de les dades sobre les transaccions que fan emergir la informació rellevant del viatge, com pot ser la durada, la destinació principal, si s'han fet excursions o no, on, en quina categoria d'establiment han realitzat despesa, etc.

En la conceptualització del viatge, a més de variables agregades a nivell de targeta bancària, es proposa també la construcció d'indicadors nous que es poden derivar dels existents tals que, són de rellevància en la caracterització del viatge i són oportuns per a la posterior realització del clustering de viatges.

El resultat d'aquest procés és un nou constructe que representa el viatge en les següents 5 dimensions:

1. Id: identificació del viatger (de la seva tarja)
2. Quan s'ha viatjat (Tini, Tfi, durada, període anual, vacances, setmana santa, franja horària, etc)
3. On? (destinació principal del viatge, tipus de viatge i freqüència de visites al destí)
4. Perfil de despesa? En què s'han gastat els diners? (volum global de despesa, nombre de transaccions, en què s'han gastat més i menys, distribució de la despesa per tipus d'establiment, %, volums, etc).
5. Nivell de mobilitat del viatge: Conté excursions? Quantes? Quants municipis s'han visitat? Distància a la destinació?

5.2. Representació del coneixement per a la definició del perfil turístic de les destinacions

En la tasca "T4.1. Disseny de la caracterització de perfil turístic de les destinacions" s'han utilitzat tècniques d'enginyeria del coneixement per a identificar com representar una destinació a partir del que opinen els visitants dels seus establiments.

La metodologia permet convertir una base de dades d'opinions en text obert, en una nova base de dades on la fila representa una destinació turística. En aquest cas s'ha decidit treballar a nivell de codi postal. Així, s'ha partit d'un model conceptual de destinació turística on la destinació està lligada a llistes d'establiments cadascun amb les seves opinions i valoracions per part dels usuaris. En aquest cas es combina una base de dades textual amb informació estructurada addicional sobre els establiments. Sobre el model conceptual de codi postal és possible establir els criteris d'agregació de les dades sobre el resultat d'haver tractat amb tècniques de llenguatge natural i tenir el sentiment i les paraules clau de les opinions, que fan possible explicitar la informació rellevant de quins tipus d'establiments hi ha a la destinació i com els perceben els visitants.

En la conceptualització de la destinació, a més de variables agregades a nivell de codi postal, es proposa també la construcció d'indicadors nous que es poden derivar dels existents tals que, són de rellevància en la caracterització del CP i són oportuns per a la posterior realització del clustering de CPs.

El resultat d'aquest procés és un nou constructe que representa el viatge en les següents 7 dimensions:

1. característiques de la destinació (id, nom i marca turística del CP)
2. perfil d'establiments informats en la destinació (nombre de comentaris, % de comentaris per tipus d'establiment)
3. perfil dels visitants (distribució territorial dels orígens dels visitants)
4. Reputació dels establiments (nombre de valoracions, puntuacions mitjanes i índexos de popularitat)
5. opinió dels visitants (%comentaris per tipus d'establiment amb sentiment positiu i per paraula clau)

6. Perfil temporal de les visites (%comentaris per període de l'any)
7. Perfil de despesa realitzada en la destinació (tipus d'establiment amb màxima despesa, % despesa per tipus, per períodes)

- Nombre de visitants diaris per país d'origen
- Nivell d'interacció entres municipis/barris
- Distribució territorial del nombre de visites amb pernoctació
- Distribució territorial del nombre d'excursions
- Distribució territorial del nombre d'excursions segons lloc de pernoctació
- Municipi de pernoctació dels visitants de cada municipi

5.3. Exploració de dades de transaccions bancàries

En la tasca "T2.4. Anàlisi exploratori de les dades" s'ha generat tot un seguit d'indicadors a partir de les transaccions bancàries dels visitants de Catalunya, des del gener de 2018 fins a setembre de 2021. Els indicadors que s'han generat són els següents:

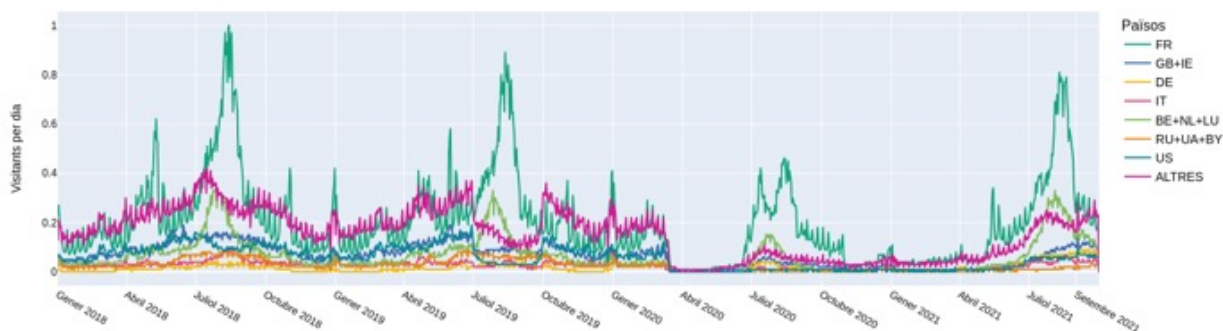
- Distribució anual de la despesa per origen
- Evolució de la despesa mitjana de les transaccions per origen
- Distribució anual de la despesa per sector
- Distribució de la despesa per sector i lloc d'origen
- Distribució de la despesa per sector i nombre de la visita
- Distribució de la despesa per sector i durada en nits de la visita

Tots els indicadors s'han calculat a través de les tècniques de representació del coneixement definides en les tasques anteriors. Aquests s'han generat per a totes les marques turístiques de Catalunya, inclosa la marca Catalunya de forma agregada. Els resultats de totes les marques es poden consultar a l'informe CIDAI-PAI 03-2021-DO1 // "Exploració de dades de transaccions bancàries", Joan Borràs i Esteban Galli, CIDAI, Eurecat 2021.

El nombre de visitants diaris es comptabilitza a partir del nombre de targetes úniques detectades com a visitants en el dia determinat. El nombre es normalitza entre 0 i 1.

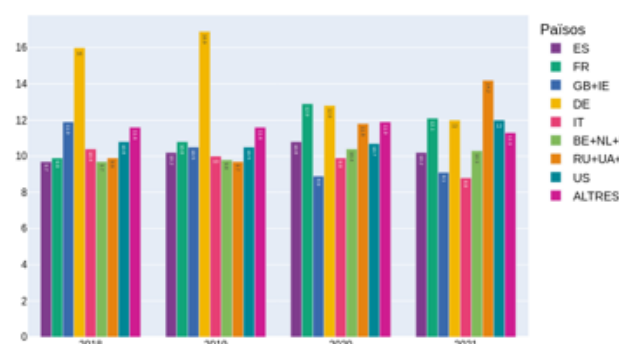
A continuació es mostra un exemple de resultats per la marca Catalunya. Podem veure clarament la caiguda de visitants degut a les restriccions de mobilitat de la pandèmia. També s'observa l'alta estacionalitat d'alguns mercats com el francès, amb pics molt alta durant les èpoques de temporada alta.

Nombre de visitants (normalitzat) per dia, segons país d'origen (països estrangers)



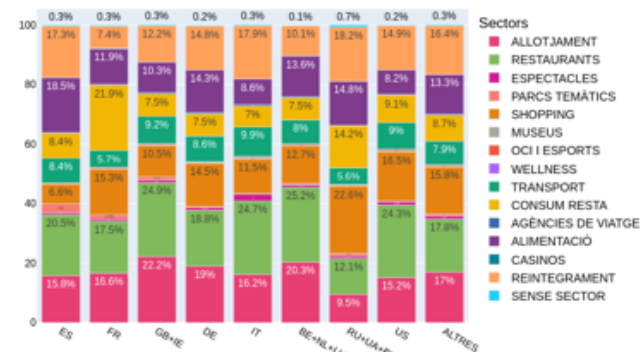
La despesa mitjana de les transaccions per país d'origen permet conèixer el valor mig de les transaccions efectuades dels visitants a Catalunya, segons el país d'emissió de la targeta. Cal tenir en compte els possibles biaixos de comportaments de l'ús de la targeta entre les diferents països o inclús entre diferents períodes. Hi pot haver mercats on l'ús de la targeta és menor, o s'utilitza només per a compres de valors més alts. Després de la pandèmia, també hi ha hagut una tendència a realitzar més compres cost més baix amb targeta bancària, enlloc d'efectiu.

Evolució de la despesa mitjana de les transaccions per origen



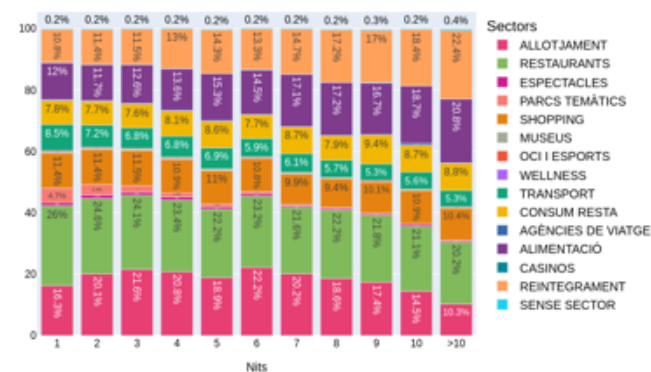
La distribució de la despesa per sector segons país d'origen permet conèixer quina tipologia de producte tendeixen a consumir determinats mercats. Es pot observar, per exemple, que RU+UA+BY tendeixen a consumir més en "Shopping" que altres mercats, o que els mercats que consumeixen més en allotjaments són GB+IE i BE+NL+LU.

Distribució de la despesa per sector i país d'origen (any 2021)



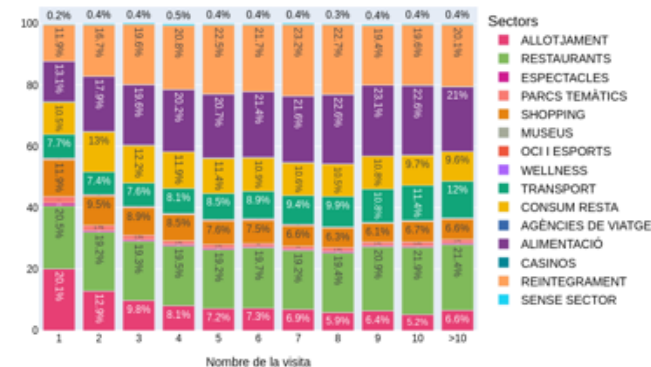
La distribució de la despesa per sector i durada en nits de la visita permet veure com evoluciona la tipologia de la despesa segons si els viatges són de més o menys durada. Veient per exemple que a partir d'estades de més de 6 nits, el pes de la despesa en allotjament disminueix. També observem com el pes de la despesa en alimentació augmenta a mesura que la durada del viatge també augmenta.

Distribució de la despesa per sector i durada en nits de la visita (any 2021)



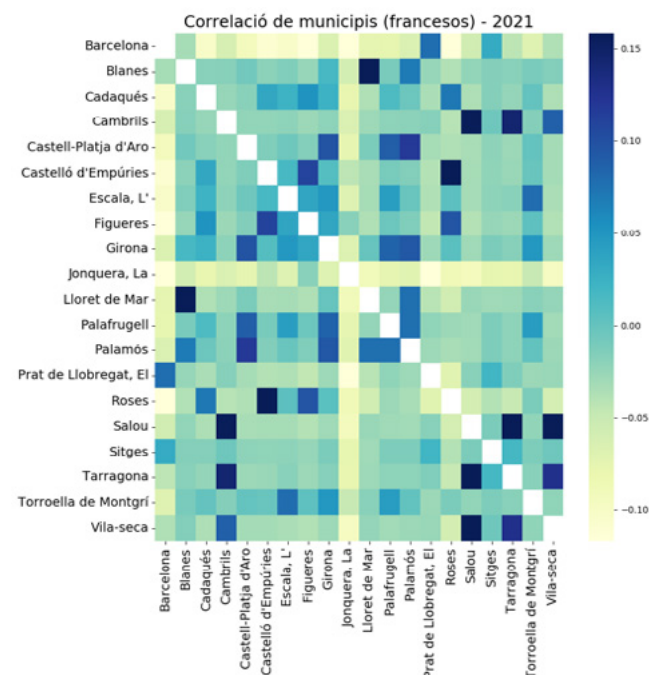
El distribució de la despesa per sector i nombre de la visita ens permet veure com evoluciona la tipologia de despesa segons el nombre de vegades que s'ha visitat la destinació. Un mateix visitant pot tenir visites en diferents nombres segons la visita efectuada. S'observa clarament com aquells visitants que repeteixen la visita més vegades, el pes de la despesa en allotjaments disminueix considerablement. El mateix passa amb la despesa en shopping. Tot el contrari passa amb la despesa en alimentació i transport, que augmenten a mesura que la repetició de la visita també augmenta.

Distribució de la despesa per sector i nombre de la visita (any 2021)



El nivell d'interacció de visites entre municipis permet conèixer la relació que hi ha entre els municipis tenint en compte que un mateix visitant realitzi visites als dos municipis en qualsevol moment de l'any.

Nivell d'interacció entres municipis (origen dels visitants: França; any: 2021)



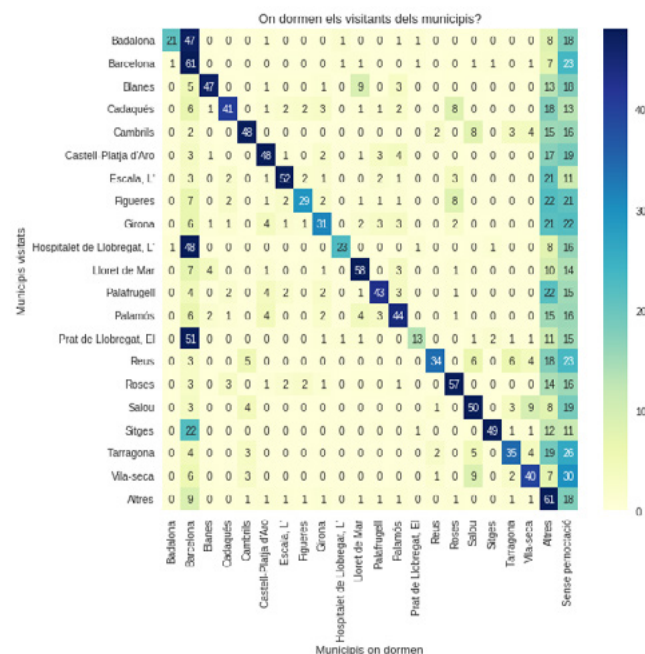
La següent gràfica mostra per a cada municipi visitat (municipis eix vertical) el percentatge de visitants que han pernoctat a cadascun dels municipis de pernoctació (municipis eix horitzontal). Es seleccionen els municipis visitats amb major freqüència de visita.

Així, per exemple, podem observar que el 22% de visitants de Sitges pernocta a Barcelona.

El 61% de visitants de Barcelona pernocta a la mateixa ciutat.

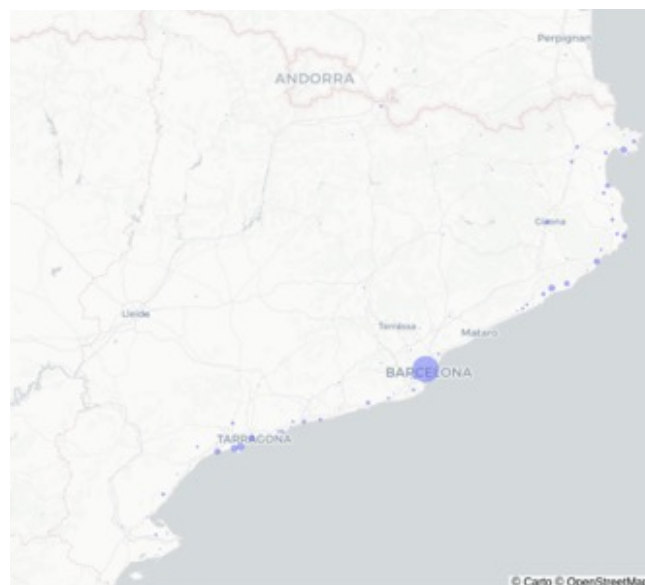
Vila-seca és el municipi amb més percentatge de visitants sense pernoctació, és a dir, que no pernocten a Catalunya o venen de visita d'un dia, directament des de el seu lloc de residència.

Municipi de pernoctació dels visitants de cada municipi (any: 2021)

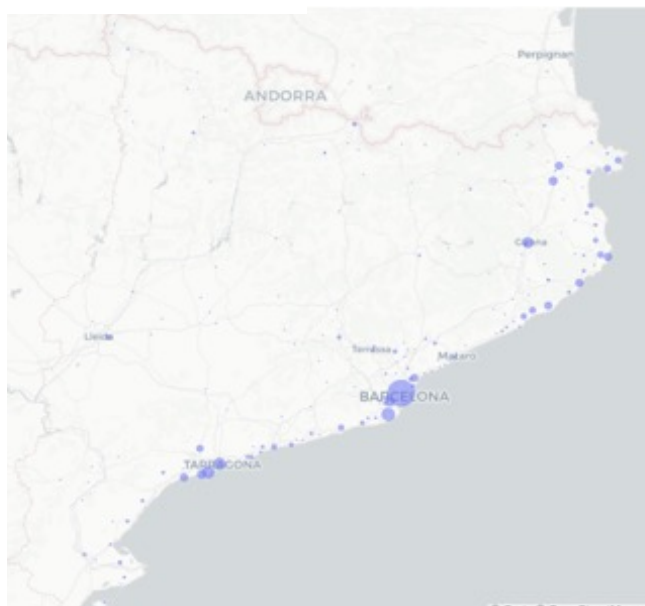


La detecció de viatges ens permet visualitzar en quins municipis es realitzen més visites amb pernoctació i en quins es realitzen més visites sense pernoctació (excursions). Els següents mapes permeten veure, amb la grandària dels cercles, quins són els municipis de Catalunya amb major nombre (normalitzat de 0 a 1) de pernoctacions i excursions.

Nombre de pernoctacions

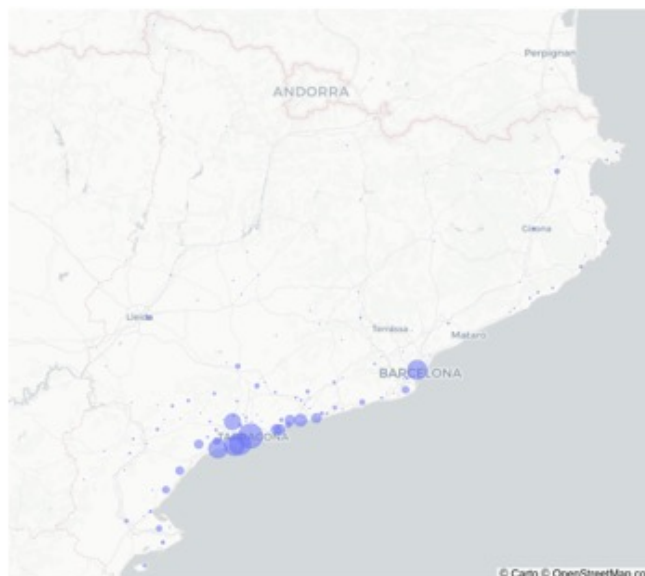


Nombre d'excursions

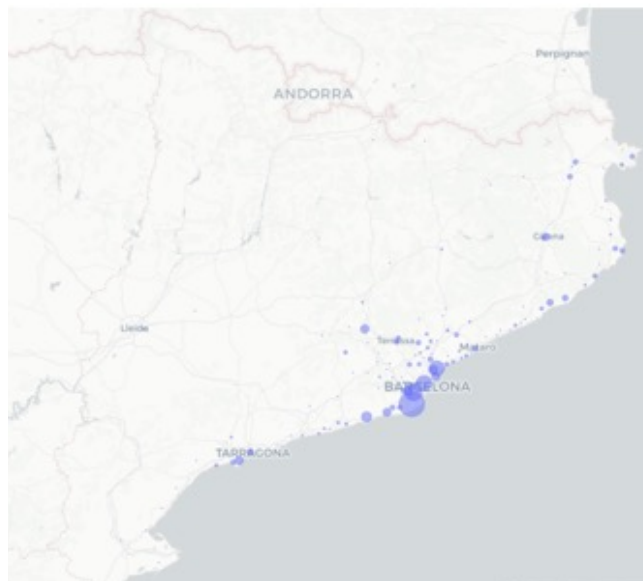


Els gràfics anteriors també es poden segmentar segons el lloc de pernoctació dels visitants. Així per exemple, podem veure les excursions d'un dia que realitzen els visitants que pernocten a la marca Costa Daurada o a la marca Barcelona.

Nombre d'excursions dels visitants que pernocten a la Costa Daurada



Nombre d'excursions dels visitants que pernocten a la marca Barcelona



5.4. Clusterització dels visitants

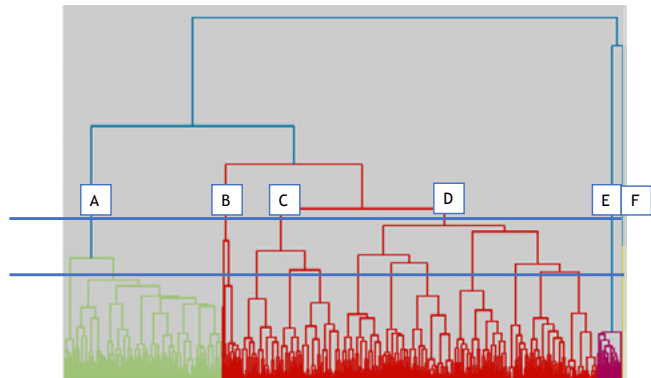
En la tasca "T2.5. Desenvolupament de models de classificació de viatges" s'han desenvolupat els models de classificació per tal de perfilar els viatges. En aquesta tasca, primera s'ha implementat el procés de tractament de dades per tal de calcular les variables. Les variables calculades per al model, són per exemple, la durada del viatge, el nombre de municipis visitats, si és la primera visita al municipi, la despesa mitjana, entre d'altres.

El procés de clusterització dels viatges es realitza amb una classificació jeràrquica amb mètode de Ward i distància Euclídia normalitzada. Cada fila que s'incorpora al model és un viatge, podent haver varis viatges d'una mateixa targeta. Per a la realització del procés es selecciona una mostra aleatòria de 100,000 viatges. El procés es realitza per aquells viatges realitzats durant el període gener i setembre de 2021.

Tot el procés de clusterització i el detall dels resultats es poden trobar a l'informe CIDAI-PAI 03-2021-D02 // "Model de clustering de visitants", Joan Borràs i Esteban Galli, CIDAI, Eurecat 2021.

A continuació es mostra el dendrograma resultant de l'aplicació del procés.

Dendrograma resultant del procés de clustering



Les línies verticals indiquen la diferència entre els dos clústers que uneix (com més altes, més diferència)

A partir del dendrograma resultant, es realitza un primer tall amb 6 classes ($k = 6$), dels quals les seves característiques distintives són les següents:

ID	Clúster	Tasca	%
A	Amb presència a Barcelona	- La majoria (87%) visiten Barcelona - Orígens principals: 23% FR, 22% RESTA-ES	28.5%
B	Despesa en allotjaments i caixers	- Amb despesa aproximada de 250€ per transacció - Despesa principal en allotjaments i caixers 22% visiten Barcelona - Orígens principals: 32% CAT, 22% RESTA-ES i 20% FR	3.3%
C	Excursionistes	- La majoria (88%) són excursionistes 21% visiten Barcelona - Orígens principals: 53% CAT, 18% RESTA-ES i 15% FR	18.0%

D	Pernoc-tants	- La majoria (84%) són visitants amb pernoctació 3% visita Barcelona - Orígens principals: 50% CAT, 17% RESTA-ES i 19% FR	45.6%
E	Llarga distància	- Visitants de llarga distància (>8,000KM) 83% visiten Barcelona - Orígens principals: 49% USA i 40% ALTRES	4.6%
F	Despesa molt alta	- Despesa molt alta (>6,000€ per transacció) 100% visita Barcelona - Orígens principals: 60% RU+UA+BY i 40% ALTRES - El 71% de la despesa en shopping i el 19% en allotjaments	<0,1%

Posteriorment s'aprofundeix amb el segon millor tall de l'arbre que resulta en 17 classes ($k = 17$).

A continuació es detallen els noms dels clústers resultants a partir del tall de l'arbre especificat en 17 clústers:

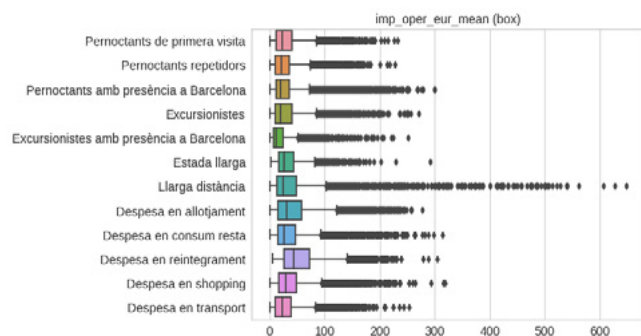
ID	Cluster	%
A1	Estada llarga	2.1
A2	Pernoc-tants amb presència a Barcelona	26.4
B1	Despesa en allotjaments i caixers	3.1
B2	Despesa alta en allotjaments i caixers	0.1
C1	Excursionistes	7.2
C2	Excursionistes amb presència a Barcelona	3.9
C3	Despesa en transport	7.0
D1	Despesa en allotjament	6.4
D2	Despesa en consum resta	7.3
D3	Despesa en reintegrament	4.8
D4	Despesa en shopping	9.7

D5	Despesa en parcs temàtics	3.8
D6	Pernoctants de primera visita	5.6
D7	Pernoctants repetidors	8.1
E	Llarga distancia	4.6
F1	Despesa molt alta	<0.1
F2	Despesa extremadament alta	<0.1

Nota: La lletra de l'identificador (ID) de la taula indica la classe de la qual s'ha subdividit de les classes del tall anterior (k=6).

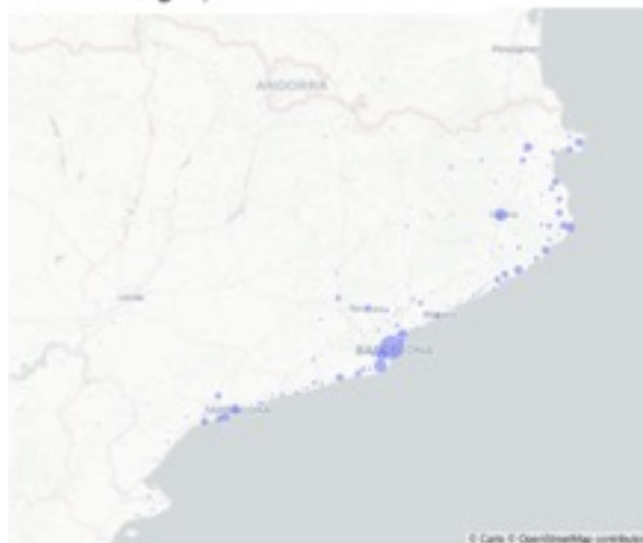
Per analitzar les característiques de cada clúster, es generen gràfics de mitjana i distribució de cada una de les variables incloses en el procés de tots els viatges que formen part en cada clúster. Així per exemple, tal i com mostra la següent figura, podem veure quina és la distribució de la despesa de les transaccions de cada clúster.

Distribució de la despesa per transacció

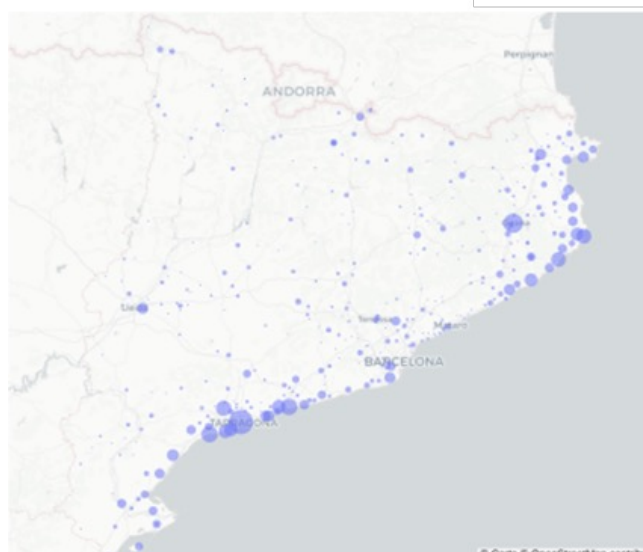


Per altra banda, també es visualitza la distribució geogràfica de les pernoctacions i excursions dels diferents clústers. Així per exemple, en les següents figures podem veure com es distribueixen en el territori català les pernoctacions dels clúster anomenats "Estada Llarga" i "Pernoctants de primera visita".

Nombre d'excursions (clúster "Estada llarga")



Nombre d'excursions (clúster "Pernoctants repetidors")



Els resultats detallats de les característiques de tots els perfils es detallen en l'informe CIDAI-PAI 03-2021-D02 // "Model de clustering de visitants", Joan Borràs i Esteban Galli, CIDAI, Eurecat 2021.

5.5. Exploració de dades d'opinions

En la tasca "T3.3. Anàlisi exploratori de les dades", del paquet de treball "PT 3 - Tractament de dades d'opinions" s'ha realitzat l'exploració de les dades obtingudes amb l'acord de col·laboració amb la plataforma TripAdvisor. La font de dades disponible inclou totes les opinions realitzades a la plataforma durant els anys 2019 i 2020 a Catalunya. Les dades inclouen el text de la opinió, la valoració d'aquesta (valor entre 1 i 5), i la informació del punt que s'ha realitzat l'opinió (nom, categoria i localització).

En aquesta tasca s'hi ha aplicat la API de Text Analytics de Microsoft a les opinions, que ha permès extreure els conceptes positius i negatius que mencionen els usuaris.

Els indicadors que s'han generat són els següents:

- Nombre d'opinions per idioma.
- Mitjana de la valoració per idioma.
- Nombre d'opinions per codi postal / marca turística / perfil de viatger / nacionalitat.
- Sentiment de les opinions per codi postal / marca turística / perfil de viatger / nacionalitat.
- Frases freqüents positives / negatives.

En la següent figura podem observar la distribució dels sentiments de les opinions segons nacionalitat. Els sentiments poden ser: positius, negatius, mixes o neutrals, segons el model de classificació de la API Text Analytics.

Distribució dels sentiments en el opinions segons la nacionalitat dels usuaris



La següent taula mostra quines són els conceptes positius i negatius més mencionats durant l'any 2019 en els productes de categoria atraccions de TripAdvisor. Totes les opinions han sigut traduïdes a l'anglès, per tal de poder agregar tots els conceptes, independentment de l'idioma.

Conceptes positius

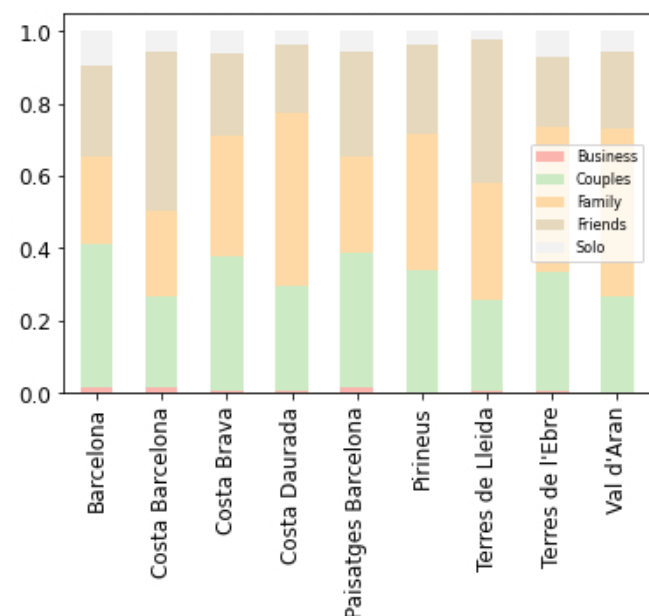
Concepte	Evaluació	Freqüència
experience	great	1145
time	great	1131
tour	recommend	1062
staff	friendly	1060
place	beautiful	972
place	nice	933
service	excellent	821
place	great	818
atmosphere	good	611
tour	great	585

Conceptes negatius

Concepte	Evaluació	Freqüència
price	high	157
food	expensive	112
entrance	expensive	112
place	small	97
prices	high	95
flight	delayed	94
price	expensive	87
place	crowded	84
park	small	81
experience	bad	78

Gràcies a la informació del perfil dels usuaris, podem analitzar quin percentatge d'usuaris segons perfil (Sol, Amics, Família, Parelles o Negoci) han publicat opinions en cada marca turística.

Distribució dels perfils d'usuaris segons marca turística



El document de resultats complet es pot consultar en l'informe CIDAI-PAI 03-2021-D03 // "Exploració de dades d'opinions", Vicente Gardulski i Shutao Zheng, CIDAI, SDG group 2021.

5.6. Clusterització de destinacions

En la tasca "T4.3. Desenvolupament de models de classificació de zones" s'identifiquen els perfils de destinacions turístiques a través de la seva activitat normal en la destinació.

Per a fer-ho es parteix de la base de dades de codis postals, on s'ha ja preprocessat, incloent les tècniques de lingüística computacional. Aprofitant que aquesta base de dades està definint un constructe que permet representar els codis postals segons diferents dimensions, s'opta per fer un ús dels mètodes multivariants de Multiview clustering i fragmentar l'anàlisi.

Les dimensions del constructe han portat a fer 6 vistes i a integrar la darrera dimensió, corresponent al perfil de la despesa, ja un agregat en sí mateix, en projecció ortogonal sobre l'espai de les classes. Cada vista s'ha classificat utilitzant el mètode de Ward i la distància mixta de Gibert, que per aquelles vistes formades només per indicadors numèrics és equivalent a la distància euclídia normalitzada, l'habitual en aquests casos. Cada vista s'ha analitzat amb tècniques d'interpretació automàtica de les classes amb tècniques intel·ligents de profiling (Panell de Classes, quadres Semàfor, termòmetres, valors-test) per poder – ne oferir la interpretació final.

La metodologia emprada per analitzar les dades i interpretar les classes és de desenvolupament propi de IDEAI-UPC i ha permès caracteritzar els perfils de codi postal i alhora comunicar-los.

El resultat final, de la fase de fusió de les vistes, més el creuament amb les dades de transaccions bancàries, han permès identificar 13 perfils, dels quals 4 són la identificació d'algunes característiques conegudes d'entrada.

A Barcelona majoritàriament

1. C908 són CP amb restaurants i bastants allotjaments molt populars i entreteniment i oferta d'oci cultural, i aire lliure també. Un 47% de la despesa s'ha fet en període no vacacional. Ex: BCN Rambla, Passeig de Gràcia Diagonal
2. C9 Restaurants ben valorats, pocs allotjaments, alguns establiments de compres i més varietat. Els restaurants són molt populars. Hi ha oci nocturn wellness, espectacles, museus i shopping molt populars també. EL 46% de la despesa s'ha fet en període no vacacional. Ex: BCN Montjuïc, Paral·lel o Raval

3. C831 Són CPs amb entreteniment i restaurants molt populars i allotjament bastant popular (Ex: dreta eixample, Diagonal mar...)
4. C887 CP amb restaurants i allotjaments mitjanament comentats; entreteniment, oci nocturn, wellness, esports i casino molt populars; restaurants, allotjaments, oci cultural, museu, aire lliure i parcs temàtics bastant populars. No hi ha excursionistes (0.02%), El 47% fan despesa en restaurants Reuneix només dos CPs: Port Vell i Port Olímpic

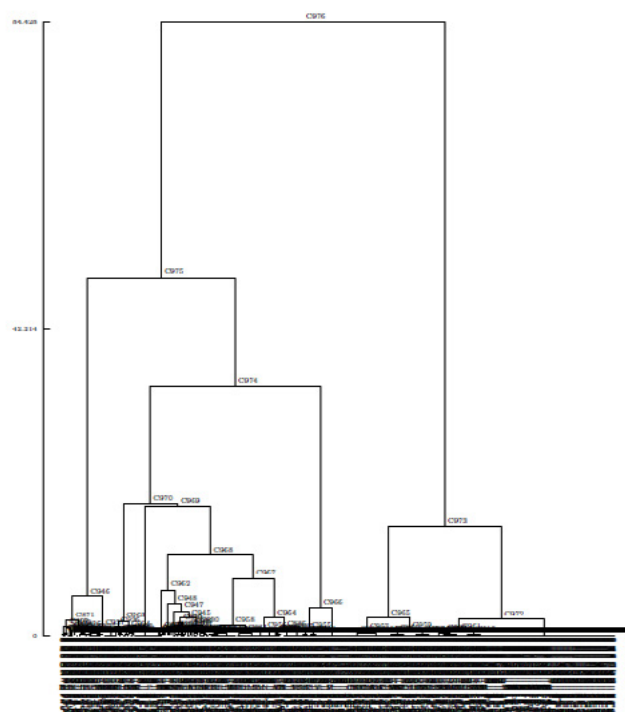
En la tasca "T4.3. Desenvolupament de models de classificació de zones" (continua)

Per la resta de territori han aparegut els següents perfils de destinacions:

5. C904 conté molta dispersió. Restaurants bastant populars i un 88% d'excursionistes. Un 39% fan despesa en transport (Ex. Masnou--El, Igualada, Palamós)
6. C940 Ratings bons de restaurants, pocs allotjament, zona de cultura i museus i una mica de lleure i de benestar. Sobretot excursionistes (86%) (Ex: Caldes-de-Montbui, o8172:Sant-Cugat-del-Valles, o8360:Canet-de-Mar, o8490:Tordera, o8696:Castellar-de-n-Hug--Pobla-de-Lllet—La)
7. C951 No hi ha gaire opinions sobre restaurants però moltes sobre allotjaments. Els ratings indiquen molt bons hotels i molt bons restaurants però la gent no opina sobre cap altre tipus d'activitat (Ex: o8529:Muntanyola, o17744:Navata, o25612:Avellanes-et-al)
8. C966 CP amb menys popularitat que la resta: (Ex o8110:Montcada-i-Reixac, Santa-Perpetua-de-Mogoda, Parets-del-Valles, Cervello)
9. C968 Moltes opinions sobre restaurants (la que més en té) però no n'hi ha sobre allotjaments (Ex: Martorelles, Montmeló, Monistrol-de-Montserrat)
10. C944 CPs que es visiten en cap de setmana, fora de vacances, primavera i hivern (Ex: Montmajor, Sils, Agramunt)
11. C947 CPs No hi ha gaire opinions sobre restaurants però moltes sobre allotjaments emeses el cap de setmana, fora de vacances i a la tardor. Sembla que s'opina només sobre una de les dues coses: (Ex. Pontós, Tarrega, Asco)
12. C957 Sense restaurants ni altres activitats, ni opinions enregistrades: (Ex. Vilobi-d-Onyar, Alcano, Sant-Miquel-de-Fluvia--Ventalló)

13. C698 CP amb un nombre mitjà d'opinions sobre restaurants i allotjaments. Molt bons rankings d'hotels i restaurants però la gent no hi va a fer cap altre tipus d'activitat. Activitat durant tot l'any, però no hi ha opinions després de Setmana Santa i a l'hivern. Tot son destins de Vall d'Aran.

Per exemple, els resultats de la vista per tipus d'establiment parteixen de les variables: Nro d'opinions formulades en CP, % d'opinions per tipus d'establiment (Allotjament, Aire Lliure, Oci cultural, Oci i entreteniment, Oci i Esports, Oci Nocturn, Oficina_Turisme, Shopping, Casinos, Espectacles, Museus, Parc_Temàtic, Wellness). El dendrograma resultant es mostra a la següent figura.



El Panells de classe mostra que només hi ha 4 de totes les variables actives que permeten discriminar les classes resultants: Les variables discriminants són el % de despesa en Restaurants, Allotjament, Museus i Parcs temàtics.

S'utilitza el termòmetre que marca en verd % d'operacions superiors al 0.8 i en vermell les inferiors al 0.2 i s'injecta al semàfor per visualitzar el contingut i particularitat de les classes.



Podem observar aquí el semàfor original i el que se superposa al panell de classes

Classe	n _c	X.Restaurant-2019	X.Allotjament-2019	X.Museu-2019	X.Parc-Temàtic-2019
C965	158				
C972	313				
C963	35				
C968	256				
C946	96				
C966	80				
C960	40				

Classe	X.Restaurant-2019	X.Allotjament-2019	X.Museu-2019	X.Parc-Temàtic-2019
C965				
C972				
C963				
C968				
C946				
C966				
C960				

Les classes ordenades i les seves descripcions són les següents:

- C965 CP amb moltes opinions sobre restaurants i forces sobre allotjaments
- C972 CP amb moltes opinions sobre restaurants (la que més restaurants té) però no hi ha opinions sobre allotjaments
- C963 CP amb bastantes opinions sobre restaurants i allotjaments i amb opinions sobre zones de parcs temàtics
- C968 CP amb un nombre mitjà d'opinions sobre restaurants i allotjaments
- C946 CP amb algunes opinions sobre restaurants i museus
- C966 CP no hi ha gaire opinions sobre restaurants però moltes sobre allotjaments
- C960 no té opinions sobre cap establiment.

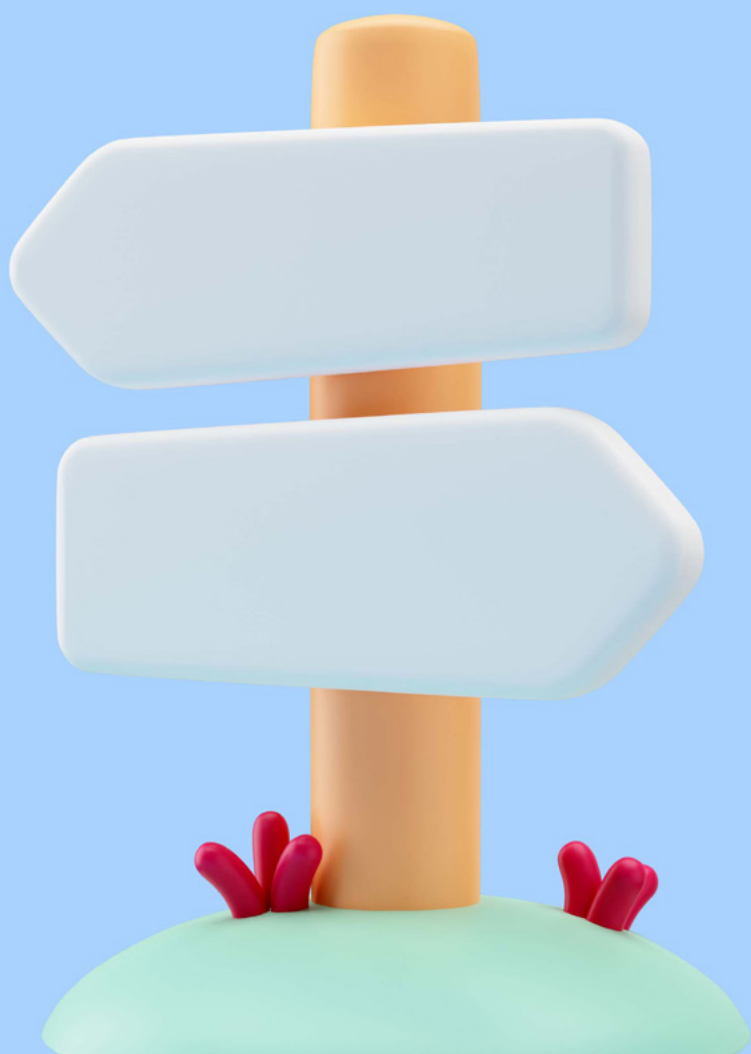
Un procés similar es repeteix per cada vista i al final es disposa d'una matriu de variables de classe per cadascuna de les vistes, que es classifica amb un mètode jeràrquic, l'algorisme de Ward i la mètrica de Xhi2 per tenir ara totes les variables qualitatives Això genera les classes finals que s'han descrit a l'inici d'aquest apartat 5. I les variables il.lustratives són les de perfil de despesa que han ajudat a enriquir la informació dels perfils.

Lliurables

[Índex](#)

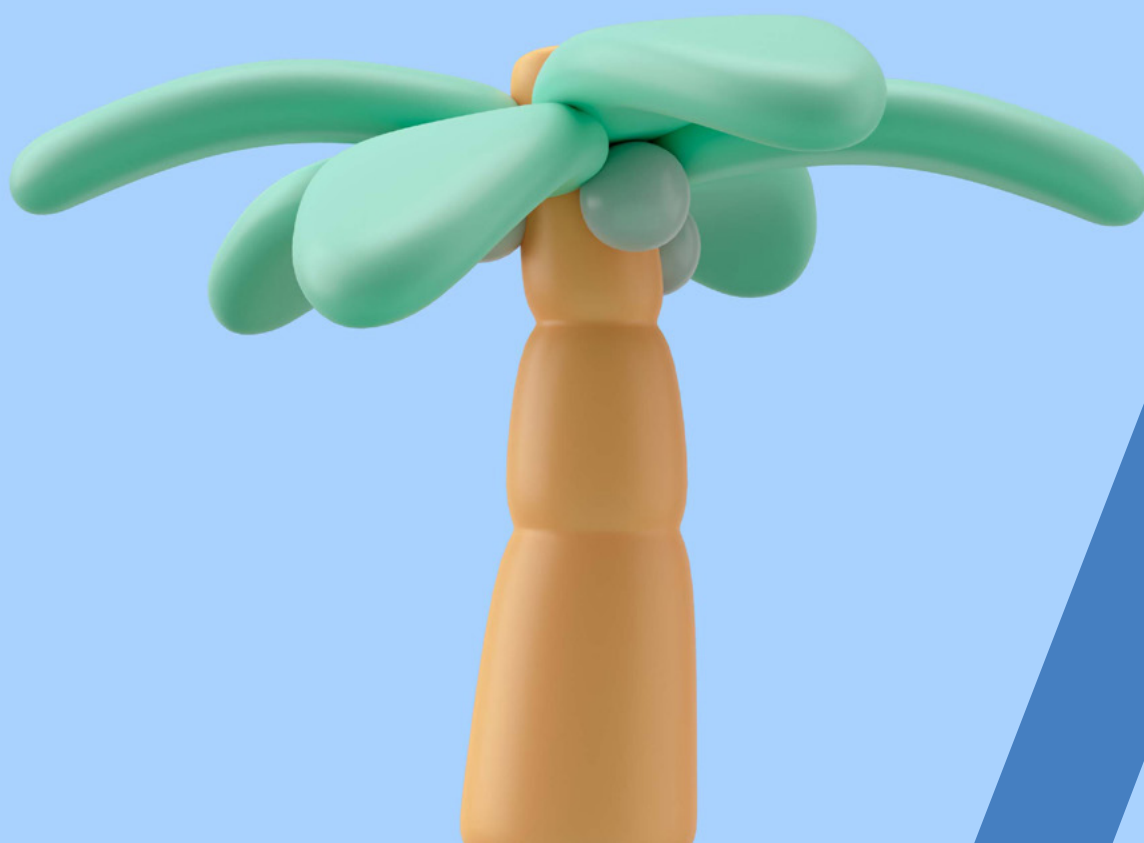
La següent taula llista els documents i fitxers lliurats en el projecte amb els seus enllaços corresponents.

Nom del document	Responsable
CIDAI-PAI 03-2021-D01 // "Exploració de dades de transaccions bancàries", Joan Borràs i Esteban Galli, CIDAI, Eurecat 2021.	Eurecat
CIDAI-PAI 03-2021-D02 // "Model de clustering de visitants", Joan Borràs i Esteban Galli, CIDAI, Eurecat 2021.	Eurecat
CIDAI-PAI 03-2021-D03 // "Exploració de dades d'opinions", Vicente Gardulski i Shutao Zheng, CIDAI, SDG group 2021.	SDG group



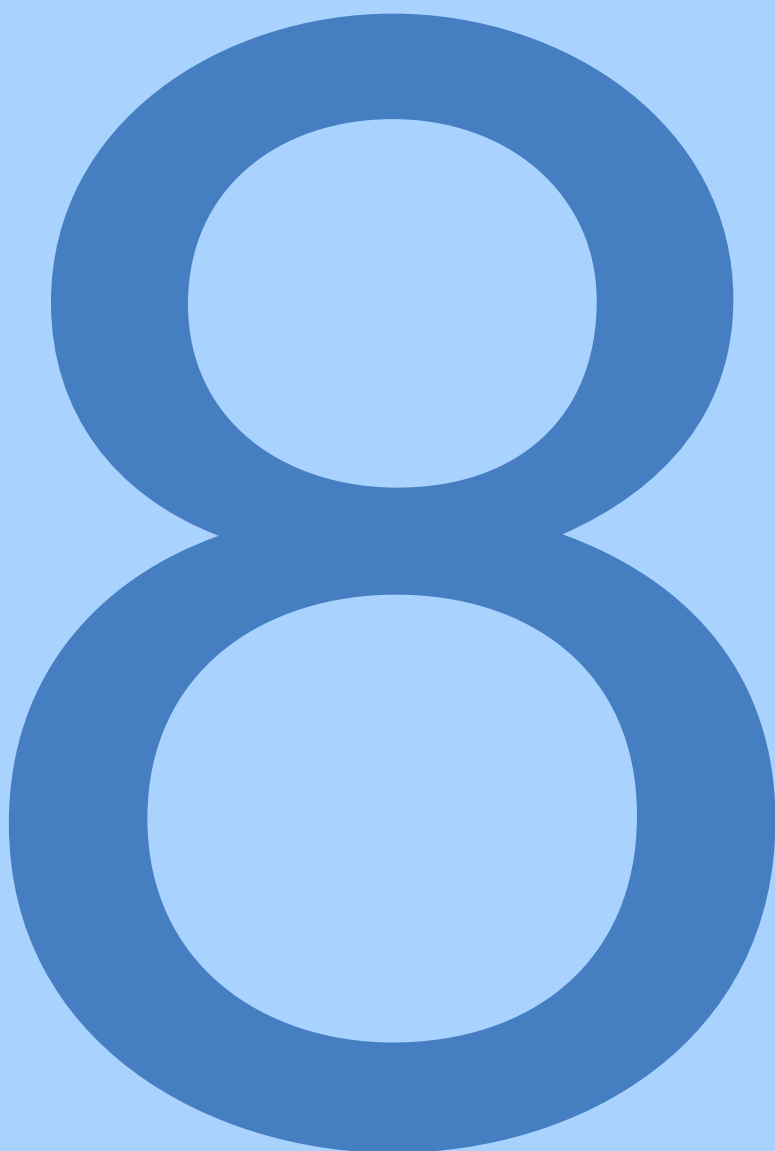
Conclusions

- Les dades de transaccions bancàries permeten obtenir una descripció exhaustiva del tipus de despesa turística d'interès amb segmentacions segons el seu comportament (turístic i de mobilitat).
- A través del seguiment de transaccions bancàries podem modelar comportaments de visita (com ara la repetició de visites, la despesa per perfil, els fluxos entre municipis, etc.).
- A través de tècniques de clusterització podem caracteritzar els principals perfils de visitants de Catalunya.
- L'aplicació de tècniques de sentiment a les opinions online són una eina molt útil per conèixer la percepció dels visitants, tenint present que els usuaris de realitzen les opinions online són un perfil concret de visitants.



Lliçons apreses

- Cal tenir presents els possibles biaixos existents segons la font de dades.
- L'expertesa en turisme és clau per tal de definir la metodologia del tractament de les dades de forma correcta, així com la interpretació de resultats i la definició dels objectius.
- El procediment de clusterització de visitants es limita als comportaments dels visitants en base a les transaccions bancàries, tenint molt de pes en alguns perfils concrets marcats per la tipologia de la despesa.
- El projecte executat de forma col·laborativa entre diferents entitats i experts de diferents dominis enriqueix el coneixement dels resultats.



Recomanacions futures

- Creuar els perfils de destinacions amb altres dades del codi postal (població, PIB, perfil d'activitat...)
- Analitzar els perfils de visitants en els diferents períodes temporals, especialment en època prèvia a la pandèmia i post-pandèmia, i investigar sobre models d'Intel·ligència Artificial per a realitzar previsions per perfils.
- Aplicar models de tractament de text que estiguin entrenats per al domini del turisme i no de forma general.
- Establir indicadors que permetin avaluar decisions de polítiques públiques i veure els canvis de tendència.
- Realitzar el mateix treball per a dades dels propers dos anys

