Universitat de les Illes Balears

TREBALL DE FI DE MÀSTER

APROXIMACIÓ A LA IMPLEMENTACIÓ DE MÈTODES INTEL·LIGENTS PER ANALITZAR LA MOBILITAT A PARTIR DE DADES DE LA SMARTWIFI DE PALMA

Pau Salas Cerda

Màster Universitari en Sistemes Intel·ligents (MUSI)

Especialitat: IoT i Ciències de dades

Centre d'Estudis de Postgrau

Any Acadèmic 2020-21

APROXIMACIÓ A LA IMPLEMENTACIÓ DE MÈTODES INTEL·LIGENTS PER ANALITZAR LA MOBILITAT A PARTIR DE DADES DE LA SMARTWIFI DE PALMA

Pau Salas Cerda

Treball de Fi de Màster

Centre d'Estudis de Postgrau

Universitat de les Illes Balears

Any Acadèmic 2020-21

Paraules clau del treball:

Smart city, mobilitat, Suavització de sèries temporals, Machine learning, Regles d'associació, Wifi

Nom Tutor/Tutora del Treball: Bartomeu Alorda

Aproximació a la implementació de mètodes intel·ligents per analitzar la mobilitat urbana a partir de dades de la SmartWifi de Palma

Pau Salas Cerda
Tutor: Bartomeu Alorda
Treball de fin de Màster Universitari en Sistemes Intel·ligents (MUSI)
Universitat de les Illes Balears
07122 Palma, Illes Balears, Espanya
pausalas85@gmail.com

***Resum*—Estudi que analitza el camp de la mobilitat dins la ciutat de Palma de Mallorca a partir de dades de geolocalització proporcionades per la xarxa SmartWifi de Palma, mitjançant diferents estratègies per obtenir els resultats en temps real. Aquesta xarxa funciona com a una font dades IoT on els sensors són els terminals mòbils. Aquesta infraestructura obté dades en brut dels dispositius vists indicant la posició i el temps, amb les que realitzaran tres classificacions diferents: la primera es detecta si els dispositius són habituals o esporàdics i també si són mòbils o estàtics. La segona classificació vol conèixer quin mitjà de transport utilitzen els vianants, mitjançant una suavització de dades on l'algoritme elegit ha estat el simple exponential smoothing i aplicant multinomial logistic regression un algoritme de machine learning. Per últim cercar les rutes, patrons, més habituals a dins la ciutat a través de regles d'associació per conèixer com es mouen els vianants. Els resultats d'aquests mètodes s'han aplicat en dades reals de mobilitat de 2019.**

Study that analyse the mobility in the city of Palma de Mallorca using geolocation data gived by SmartWifi of Palma network, through different strategies to get results in real time. This network works as a IoT database where the sensors are mobile terminals. This infraestructure get raw data from seen devices with their positions and their times, that the study will do three classifications: the first one, detects if the devices are sporadic or usual and if them are statics or mobiles. The second one, wants to know what modes of transport uses people to move in to the city, through a simple exponential smoothing and using multinomial logistical regression, a machine learning algorithm. The third one, find the movement patrons in the city with association rules, where the items are the touristic places of the city. The results from this methods applied on real mobility data from 2019.

***Index Terms*—Smart city, mobilitat, Suavització, Machine learning, Regles d'associació, Wifi**

I. Introducció

Cada dia que passa les ciutats actuals han d'afrontar nous reptes. La població creix de manera exponencial i s'espera que quan arribem a l'any 2050, la població mundial creixerà un 72 % i la població a dins les ciutats passarà a ser del 67 % [1]. Davant aquest creixement, les ciutats amb els recursos que disposen i el seu actual disseny no podran fer front a la futura densitat de població que arribarà. Per resoldre aquesta situació s'ha de cercar una nova aproximació. Així es presenta la ciutat intel·ligent, Smart City, com una solució que aplica la tecnologia per poder afrontar la futura densitat de població i els problemes associats a aquests.

Una Smart City es pot definir de diferents formes. Però, una definició general seria: «una ciutat per millorar la qualitat de vida dels ciutadans en un entorn sostenible utilitzant les tecnologies de la informació» [2]. La Smart City s'organitza en verticals diversos. Que contemplen les parts de la ciutat: la gestió energètica, el cicle de l'aigua, la seguretat a dins la salut... A més d'afrontar els reptes i problemes d'aquesta com la contaminació de les ciutats, canvi climàtic, pandèmies, entre altres.

La Smart City conté una varietat de característiques que es poden associar amb diferents objectius. Un estudi proposa sis vèrtebres que ha de consolidar una ciutat intel·ligent [3]:

- **Smart Governance**: Fusionar la ciutat i els voltants, incloent-hi els serveis i interaccions, A més dels serveis públics, privats, civils i de la Comunitat europea, tot això perquè la ciutat pugui funcionar eficientment com un sol organisme.
- **Smart Economy**: En el camp de la e-business i e-commerce. Augmentar la productivitat, mitjançant les ICT (tecnologies de la informació), així com crear nous productes, serveis i comerços.
- **Smart Mobility**: Utilitzar les ICT per integrar el transport i els seus sistemes lògistics. Un exemple seria: utilitzar els sistemes de transport englobant-los com és el tramvia, autobusos, trens, metro, cotxe, bicicleta, caminar... En diferents situacions utilitzant un o més modes de transport.
- **Smart Environment**: A través de les ICT, convertir la ciutat en un lloc amigable amb el mediambient a través d'utilitzar energies renovables, controlar i monitorar la pol·lució de l'aire, controlar el cicle del fem, controlar el cicle de l'aigua, etcètera.
- **Smart People**: Dotar a la ciutat, d'accés a l'educació, recursos humans i capacitat de gestió. A dins una societat inclusiva, que millora la creativitat i acull la innovació.
- **Smart Living**: Viure segur en una ciutat multicultural amb els seus propis edificis culturals, i que incorpora una

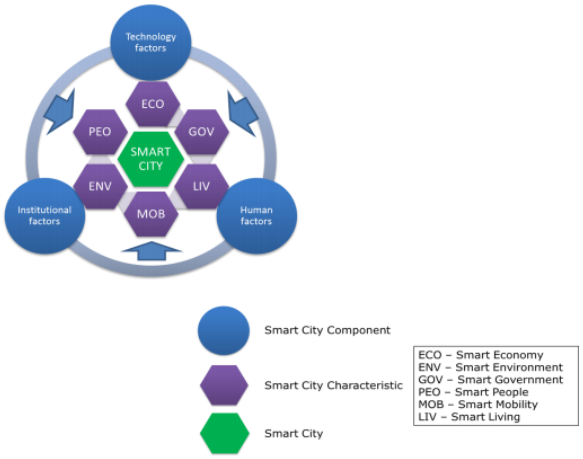

Figura 1. La relació entre els components i les característiques de la Smart City[3]

qualitat mínima de casa i acomodació per a cada ciutadà.

Aquests verticals tenen persones interessades, stakeholders, que els duran a terme mitjançant uns medis, també coneguts com a components. Un exemple de component seria un camp fotovoltaic per donar energia neta a dins la ciutat i reduir la contaminació. Els components es poden conceptualitzar com a eixos bàsics de la Smart City. Aquests es poden categoritzar en tres nuclis diferents:

- **Factors tecnològics**: Infraestructura física, tecnologies Smart, tecnologies mòbils, tecnologies virtuals i xarxes digitals.
- **Factors humans**: Infraestructura humana i capital social.
- **Factors institucionals**: Govern, polítiques, regulacions i directives.

La relació entre els components i les verticals principals donen a lloc a la Smart City, com es mostra a la figura 1.

En comptes de centrar-se en totes les vèrtebres de la Smart City, aquest projecte s'ha dirigit en treballar només en el camp de la mobilitat a dins la ciutat. A causa de la globalització i l'augment de la població la demanda de transport creix, així com la seva respectiva conseqüència: renou, congestions a les carreteres i dependència a la utilització del cotxe propi en contra de l'ús del transport públic [4]. Per aquestes raons està en desenvolupament innovacions pel transport a dins les Ciutats intel·ligents. Un exemple seria la **Mobilitat com a servei (MaaS)** [5], és un paradigma que uneix tots els mitjans de transport, sent amigable amb el medi ambient, baixant el cost de manteniment i millorant el temps per arribar a destí. Un altre exemple, seria l'ús de la xarxa Wifi a dins una ciutat, ja que aporta accés local a les dades en brut sobre la geolocalització precisa en punts d'interès municipal i en quin moment. Aquestes dades donen oportunitats per la innovació del transport amb una col·laboració pública-privada.

L'any 2015 l'Ajuntament de Palma amb l'assessorament de la UIB (Universitat de les Illes Balears) [6] donava inici a un projecte de gran magnitud. Aquest projecte consistia a dotar als turistes i ciutadans d'internet a través d'una xarxa

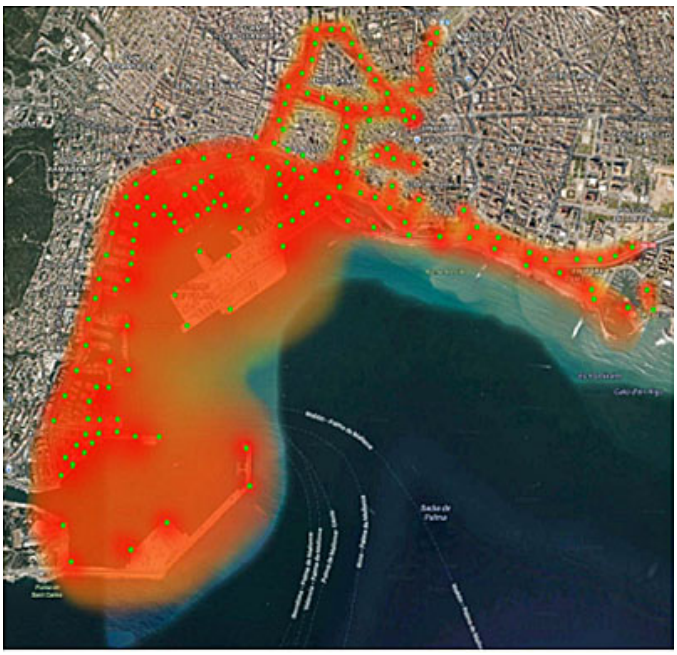

Figura 2. Distribució dels APs i la seva cobertura a la ciutat de Palma. [6]

Wifi extensa. L'execució d'aquest projecte el va dur a terme l'empresa MallorcaWifi, on primerament va realitzar una prova pilot a la platja de Palma en el 2014 a partir del disseny creat pel grup de treball SmartDestination de la UIB. Així donava inici el projecte on la policia Local de Palma utilitzaria el 50 % de la xarxa per temes de seguretat i serveis públics i l'altre 50 % de la xarxa donaria servei d'internet als vianants i turistes de la ciutat. Els APs (punts d'accés, Access Points) es van instal·lar al llarg de les zones turístiques de major afluència de la ciutat. Es pot observar a la figura 2, els diferents APs instal·lats a la ciutat de Palma (els punts de color verd) i la seva àrea de cobertura dins la ciutat (àrea ressaltada).

A. Motivació i objectius

Els APs formen la xarxa de monitoratge que anirà recopilant dades sobre quins dispositius han vist, l'hora i la ubicació estimada d'aquest. A partir d'aquestes dades en brut. S'intentarà, a través de diferents mètodes, obtenir informació i coneixement sobre l'estat actual de la mobilitat dins la ciutat. Aquest coneixement generat permetrà millorar la infraestructura de la mobilitat de la mateixa ciutat atenent a les necessitats detectades.

En aquest estudi es treballarà a partir d'un subconjunt de les dades en brut de la xarxa Wifi (Estius de 2018 i 2019). S'apliquen tres models diferents, que s'implementaran, per obtenir informació sobre la mobilitat a dins la ciutat, aquests són:

- Classificació del tipus de dispositiu. Classificar dispositius segons la seva aparició a dins la xarxa: si són habituals, s'observen diverses vegades al llarg del dia, o si són esporàdics, solament s'observen una o dues vegades.

També es classificarà el seu moviment: si són de caràcter estàtic, no es mouen de la posició al llarg del dia, o de caràcter mòbil, canvien de posició al llarg del temps.
- La classificació del tipus de transport. Determinar el mode de transport dels usuaris mòbils, un exemple seria determinar si un usuari en un trajecte va en cotxe o caminant.
- Aconseguir les rutes més habituals. Conèixer quines rutes són les més utilitzades i quines zones tenen més trànsit de persones.

A la següent secció, es mostrarà el treball relacionat realitzat per altres investigadors que han elaborat una anàlisi que tenen relació amb la mobilitat. En la secció III podem trobar l'estructura de la font de dades. En les seccions IV, V i VI es troben les diferents tècniques explicades, i per últim, en la secció VII, la discussió global i la conclusió del treball on s'analitzarà els resultats obtinguts.

II. Treball relacionat

En aquesta secció, es presentarà el treball relacionat en tres àmbits diferents: tipus de classificació de dispositius, classificació de transports a partir de dades Wifi i implementació de regles relacionals en Smart Cities.

A. *Classificació de dispositius*

Per classificar els dispositius no solament es realitza a través de la seva aparició i el seu moviment. Hi ha altres maneres de classificar un dispositiu, un exemple seria classificar segons el seu tràfic de dades, si consumeixen un gran volum de dades per exercir la seva activitat o necessiten un petit volum per funcionar. A continuació es parlarà de dos articles diferents on es mostra com es poden classificar els dispositius.

El primer estudi classifica dispositius IoT (Internet of Things, Internet de les Coses) connectats a dins una mateixa xarxa [7]. L'escenari de l'estudi és: a dins d'un edifici, on hi ha diferents aparells electrònics com pugui ser un Amazon Echo, una impressora, una càmera... Identificar-los a través del tràfic de xarxa. Una consideració que té en compte, és que alguns aparells són similars, com seria entre els dispositius *Insteon Camera* i la *Samsung SmartCam* que són càmeres. Llavors no es determinarà el dispositiu, sinó a la categoria a la qual pertany. Aquest estudi divideix els dispositius IoT en quatre categories diferents:

1) **Hubs**: Són els controladors domòtics, que serveixen per controlar els llums, la calefacció...
2) **Electrònica**: Aquesta categoria inclou els altaveus i impressores de l'edifici.
3) **Càmeres**: Aparells per capturar fotogrames quiets o en moviment.
4) ***SwitchesTriggers***: Sensors i actuadors de l'edifici. Està format per endolls intel·ligents, repetidors de Wifi i sensors de moviment.

La següent passa ha estat obtenir indicadors que ajudin a poder categoritzar els dispositius a través del seu tràfic de xarxa. Com es pot observar a la figura 3. A través d'una eina de

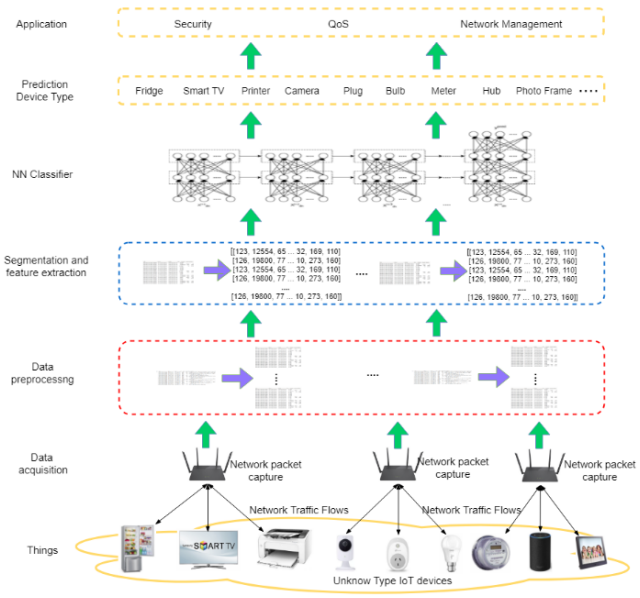

Figura 3. Proposta per la classificació de dispositius [7]

lectura de paquets, com podria ser wireshark[1]. Els indicadors per categoritzar els paquets són:

- Protocols utilitzats per enviar informació. Aquestes estarien dividides en dues categories: paquets d'usuari i control. Els d'usuari contendrien els protocols TCP, UDP, HTTP i altres protocols de la capa superior. Per altra part, estan els paquets de control que usen els protocols ICMP, ARP, DNS i NTP [2].
- La quantitat de paquets generats durant un període de temps determinat. No hi ha el mateix flux de missatges en el cos d'una estació de meteorologia o d'una càmera.

Una vegada obtinguts els indicadors necessaris. La classificació es realitzaria a través d'un classificador end-to-end basat en algoritmes d'aprenentatge profund. Es crea un model que anirà entrenant mitjançant una base de dades d'entrenament per obtenir el resultat final.

Un segon estudi analitzat, ha estat la classificació dels usuaris segons l'ús dels seus aparells electrònics, per així els operadors de telefonia tenir informació per aplicar mesures de control i reduir els costs de manera significativa [8]. L'objectiu d'aquest estudi és analitzar l'ús dels dispositius mòbils, categoritzar-los i proporcionar als clients promocions personalitzades i augmentar els beneficis de l'operadora. Per exemple posar advertiments de descompte o promocions a un tipus d'usuaris en un període de temps.

L'estudi proposa dues classes:

- **Classe 1**: Usuaris amb una preferència específica, són constants en la seva actitud davant l'aparell electrònic.
- **Classe 2**: Usuaris que no tenen una preferència específica, la seva actitud no està clara, ja que canvia amb el temps.

[1]https://www.wireshark.org/
[2]Podeu trobar informació dels protocols a https://www.redeszone.net/tutoriales/internet/protocolos-basicos-redes/

Per poder categoritzar els dispositius s'utilitzen 33 característiques diferents que es poden agrupar en 3:

- **Electronic devices attribute features**: Format per 1 característica. Correspon a l'avaluació de l'aparell.
- **User feedback related features**: Format per 5 característiques. Treu informació dels usuaris, si en cada vegada que es connecten a internet cerquen una informació en concret o no.
- **Electronic devices usage related features**: Conformat per 27 característiques. Descriu el total d'informació utilitzada al llarg del temps.

En tenir ja els camps descriptius necessaris, s'ha de seleccionar un algoritme per fer la classificació. El problema està, que hi ha diversos algoritmes que fan el mateix, però de distinta forma. Per la selecció s'ha d'escollir l'algoritme que obtengui el percentatge més gros d'encerts amb aquestes característiques. El que es fa a l'estudi és a partir un dataset d'entrenament que conté les respostes, es proven diversos algoritmes per elegir el millor. A l'estudi s'usen: logistic regression, random forest i support vector machine (SVM).

Els investigadors per avaluar varen fer dues anàlisis: un amb els 33 trets i una altra reduint el nombre a solament 8. Els resultats obtinguts a través d'aquest estudi han donat que els tres algoritmes han donat un millor resultat utilitzant els 33 trets en comptes de solament 8. I que l'algoritme seleccionat ha estat el del random forest que ha obtingut una exactitud del 79.29 %.

En aquests dos articles s'han classificat dispositius de dues maneres diferents, ja que es On cada cas tenia un conjunt d'entrada diferent de dades. A partir d'aquests dos articles es pot definir una idea general de les passes a seguir per classificar els dispositius.

B. Classificació de transports

En el camp de la mobilitat, és important realitzar estudis per classificar diferents transports que permetran conèixer l'estat actual de les ciutats en aquest tema. I una tècnica utilitza la xarxa Wifi en una ciutat [9]. L'objectiu de l'estudi és distingir quin tipus de transport utilitzen les persones. Les opcions presents són: anar caminant, anar amb bicicleta o anar conduint. Aquesta distinció s'analitza mitjançant els beacons (paquets per descobrir les xarxes Wifi) enviats entre els mòbils i els APs de la xarxa Wifi.

Abans de l'estudi es va elaborar una anàlisi de possibles tecnologies que es podrien utilitzar. Aquestes van ser: GPS (Sistema de posicionament global), GSM (Sistema global per a comunicacions mòbils) o WiFi.

L'ús del GPS permet conèixer la ubicació de l'individu en temps real, aquesta tecnologia va set utilitzada en anteriors estudis i van donar bons resultats. Però, alhora de la implementació, aquesta tecnologia requereix intervenció de l'usuari, ha de permetre que el puguin rastrejar constantment, un alt consum de la bateria del dispositiu i dóna problemes

Figura 4. Àrea on és realitza l'experiment [9]

en espais molt grans. Com a alternativa el GSM. Amb una precisió menor que el GPS, es basa a conèixer la posició dels dispositius a través dels Based Transceiver Stations (BTS). El principal problema d'aquesta opció és la violació sobre la privatesa de les persones a l'hora d'obtenir les dades, però a més, hi ha possibles àrees amb una densitat de BTS baixa.

Finalment, el Wifi. Permet el descobriment de dispositius a dins la seva àrea de cobertura (un conjunt APs repartits en una zona) a través de beacons que emeten els usuaris, sense que hi hagi una violació a les dades personals. Per aquestes raons l'estudi [9] escull aquesta tecnologia.

Una vegada ja seleccionada la tecnologia, l'escenari per realitzar l'estudi ha estat: un circuit tancat que conté 4 APs repartits un en cada recta, aquests dispositius a través dels beacons registraran el dispositiu observat amb el corresponent. Hi ha un conjunt de voluntaris que fan fent voltes al circuit, en el sentit horari, usant un mode de transport assignat que varia entre anar caminant, anar en bicicleta o anar conduint; tot això a dins un període de temps. Es pot observar en la figura 4 el circuit realitzat i la ubicació dels APs.

Per identificar els usuaris s'utilitza la seva direcció MAC (medium access control) dels dispositius. Cada registre a part de la MAC conté 15 trets diferents utilitzats per la classificació dels transports. Aquestes es poden dividir en tres classes diferents:

- **El temps**: On es calcula la velocitat i el temps d'origen i destí dels usuaris.
- **El nombre de connexions**: Consta del nombre de connexions origen, destí i la mitja.
- **La senyal**: La força de la senyal rebuda en els diferents punts.

La metodologia utilitzada és la d'emprar diferents

algoritmes de machine learning, comprovar quin és el millor i utilitzar-lo per predir tots els valors. Els algoritmes que s'han valorat han estat: ReliefF, Decision Tree, Bagged decision Tree, Random Forest i Multi-layer Percepton. Finalment amb la calibració i l'entrenament, el millor algoritme ha estat el de Multilayer Perceptron algorithm amb una precisió del 86,52 %.

Aquest article ens ha demostrat com identificar modes de transport a través de la xarxa Wifi, en un entorn controlat i tenint informació sobre quins transports han anat els voluntaris.

C. Regles d'associació en les Smart Cities

En el camp de la mobilitat en les Smart Cities s'analitza un article per obtenir patrons de rutes de la mobilitat dels ciutadans a dins una ciutat per instal·lar i millorar una xarxa DTNs (Delay Tolerant Networks) en entorn urbà [10].

L'objectiu del protocol DTNs és encaminar dades (un exemple de dades seria informació del clima) d'origen a destí a partir de missatges oportunistes i que toleren retards que es basen típicament en comunicacions de baix rang, com podria ser Bluetooth.

A les xarxes DTNs, les dades no són distribuïdes a tots els dispositius, sinó a un subset d'aquests que reenviarà aquesta informació a altres dispositius successivament fins que no hi hagi altres dispositius per rebre-ho. El problema principal en DTNs és elegir el primer dispositiu perquè la informació que envii arribi al màxim de nombre de dispositius.

L'article identifica els patrons de la mobilitat en la regió funcional de tres ciutats: Roma, San Francisco i Beijing. Una regió funcional és una àrea que conté atraccions turístiques, centres comercials, zones d'educació, llocs de feina i residencies. L'estudi categoritza les regions de la ciutat en 4 tipus:

- **Llocs de feina**: Llocs on els ciutadans fan els treballs
- **Llocs d'entreteniment**: Llocs on els ciutadans gasten el seu temps d'oci.
- **Llocs de residència**: Llocs on van els ciutadans per descansar i dormir.
- **Altres zones**: És la resta de zones que no es poden classificar.

Per l'obtenció de dades, es varen utilitzar la base de dades del GPS dels taxis a dins la ciutat. Solament es van tenir en compte els trajectes entre dos punts: on recullen un passatger fins que el deixen.

Una vegada ja obtingudes les rutes, el següent pas va ser dividir l'àrea de l'estudi en regions a través de les regles d'associació. Primer es divideix la ciutat en regions, si una regió supera l'1 % de visitants, llavors, la regió es dividirà en més petites, d'aquesta manera fins a tenir tota una ciutat dividida en regions.
Com que els patrons dels ciutadans canvia durant el dia com és el cas d'un ciutadà que el matí va fer feina i el vespre dorm a ca seva. El següent pas és categoritzar les regions segons el trànsit durant un període de temps, la classificació queda de la següent manera:

- **Llocs de feina** Aquests es consideren en llocs que es visiten en hores de feina, aquestes són:
  - Dilluns a divendres de 8:00 - 17:00
- **Llocs d'entreteniment** Zones que van els ciutadans per passar el seu temps d'oci, format per:
  - Dilluns a divendres de 17:00 - 23:00
  - Dissabte a diumenge de 8:00 - 20:00
- **Hores de descans** Zones que van els ciutadans per passar el seu temps d'oci, format per:
  - Dilluns a divendres de 23:00 - 8:00
  - Dissabte a diumenge de 20:00 - 8:00

A partir de les regions identificades i classificades, es provaren tres algoritmes diferents per enviar missatges Oracle-Based, History-based i random. Per comprovar l'eficàcia dels algoritmes es miraran els missatges rebuts pels dispositius entre els missatges enviats.

L'algoritme random envia missatges aleatòriament a dins la ciutat. Sense tenir en compte la divisió de les regions ni horaris.
L'algoritme Oracle-Based, tracta d'enviar els missatges més importants a les àrees calentes (àrees on s'agrupa molta gent), un exemple seria, durant les hores de feina s'enviaran a les zones de feina, i de la mateixa forma en les diferents zones.
Per últim, l'algoritme History-based, aquest és similar a l'algoritme Oracle-Based, però amb l'excepció que un taxi visiti les àrees calentes és obtinguda a partir d'un historial i no les dades d'un sol dia. Un exemple si una persona sempre va cada dia a les 8 del matí a fer feina, té major probabilitat que el seu dispositiu sigui seleccionat el primer per enviar missatges DTNs.

A partir dels resultats obtinguts, si es compara l'algoritme random amb el del History-Based. El darrer ha demostrat una eficiència major del 183 % comparant-lo a l'algoritme random.

Aquest estudi utilitza una aplicació de la implementació de les regles d'associació per resoldre el problema a les xarxes DTNs. Aquest estudi demostra una possible aplicació d'utilitzar les regles d'associació a dins una ciutat.

III. Descripció de la font de dades

Aquesta secció tractarà sobre el procés de captura de les dades i l'anàlisi de les fonts de dades que s'utilitzaran per realitzar els algoritmes.

A. Procés de captura

El primer pas és la recopilació de les dades en brut dels dispositius que han vist els diferents APs. Aquest procés de descàrrega de dades la realitzarà una API « Meraki Location Scanning API » [11]. Aquesta permet enviar dades en temps real dels dispositius que observa, on s'utilitza per detectar dispositius dins la xarxa Wifi o Bluetooth Low Energy (BLE). Els elements són exportats via HTTP POST o JSON

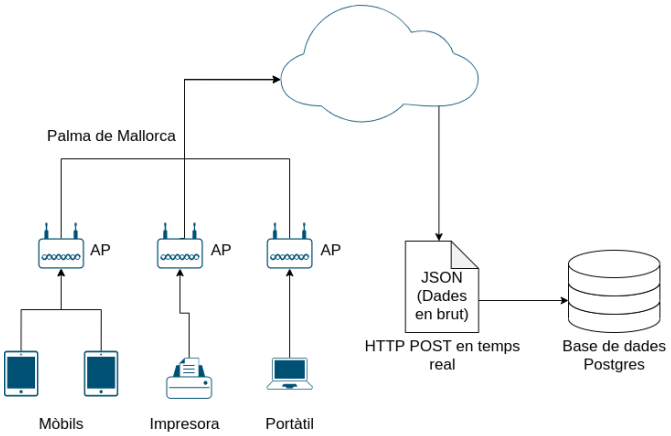

Figura 5. Procés de la captura de dades del Smart Wifi de Palma.

(JavaScript Object Notation) a un servidor específic.
En el nostre cas les dades són de tots els APs de la ciutat de Palma i aquests són enviats directament des del núvol fins a una base de dades POSTGRES. Els POSTS dels paquets ocorren amb una freqüència d'un missatge per minut de cada AP. Aquest procés es pot observar a la figura 5.

B. Organització de les dades

Una vegada ja obtingudes les dades, aquestes es divideixen en diferents taules amb la mateixa estructura, on cada taula conté tots els registres d'un dia i està formada per 9 columnes diferents:

1) **idreg**: Són els identificadors per les entrades de la base de dades.
2) **clientmac**: Identifica la direcció MAC (Media access control) és una direcció única per cada dispositiu que l'identifica. Conformat per 48 bits (24 inicials que identifica el fabricant i els 24 identifiquen al dispositiu).
3) **lat**: Primer indicador de la posició dels dispositius, la latitud.
4) **lng**: Segon indicador de la posició dels dispositius, la longitud.
5) **seendate**: Indica el dia, mes i any de quan ha estat a la posició, en format: dd/mm/yyyy.
6) **seentime**: Temps en què es va recollir la mostra. Aquest està en format Epoch. Són els segons que han passat des d'un instant de temps (1 de gener de 1970 00:00:00).
7) **apmac**: La direcció MAC de l'AP (access point) on ha establert la connexió.
8) **model**: Fabricant del dispositiu recollit a partir de la mac dels dispositius.
9) **ssid**: Columna buida.

C. Processat previ de les dades

La recaptació de les dades, de qualsevol sistema, no és acurat al 100 % perquè poden presentar possibles reptes, com són:

- La recollida de dades no és en temps real, és a dir, la recaptació de les dades tarda més que l'establert pel sistema
- La semàntica de dades és dèbil, les columnes no estan ben definides o podrien faltar camps.
- Possible inexactitud de les dades, els valors poden contenir errors.

Aquests fets ocorren, ja que han de passar a través de diferents processos i aspectes com factors espacials i temporals, és a dir, qualsevol sistema utilitzat per recaptar dades no és perfecte a més de no ser precís [12]. A causa de la falta de precisió, abans d'utilitzar les dades en brut, aquestes hauran de passar a través d'una fase de transformació i plantejar-se alguns escenaris que necessitin o no processar abans les dades pels models. A la revisió de les dades s'han plantejat aquests fets:

- **Problemes de lag**: Signifiquen que les entrades no estan ordenades cronològicament i per tant en alguns casos s'ha de realitzar una ordenació d'aquestes.
- **Errors d'ubicació**: Registres on les posicions estan fora de l'àrea de cobertura del Wifi de Palma. Per eliminar aquests casos s'ha aplicat un filtre on si el registre supera els límits, aquest s'esborra.
- **Problemes espacials-temporals**: En el cas de mirar els registres d'un usuari poden aparèixer registres en ubicacions diferents en el mateix instant de temps o diferents ubicacions, llunyanes, amb poca diferència de temps.
- **Possible violació a la privadesa**: En ser Espanya part de la unió europea, aquest projecte està sotmès a les regles de privadesa de la mateixa unió. En el nostre cas en utilitzar la direcció MAC dels dispositius per classificarlos no viola cap regla, però en altre cas si s'utilitzàs amb altra informació sobre com per exemple, l'adreça IP, llavors s'estaria incomplint la llei de protecció de dades [13].
- **Possibles valors nulls**: Està la possibilitat que a causa de la imperfecció del procés de monitoratge, apareguin entrades nulls a la base de dades. En aplicar els diferents algoritmes a les dades no han sorgit cap error a causa d'aquest fet.
- **Direccions MAC vàlides**: S'ha de revisar si els tipus de macs són: unicast, multicast o broadcast a través d'una revisió d'un dia; per una altra part, també es mirarà si les macs són gestionades localment pel dispositiu o si són globalment úniques (reforçades per la OUI organizationally unique identifier). Els resultats obtinguts són que en més de 100.000 dispositius sortia que el 100 % eren globalment úniques i que el 99.99916 % eren del tipus unicast (les altres eren direccions multicast). Amb aquestes podem concloure que no importa filtrar les MACs dels dispositius.

Aquestes dades ja han passat un primer filtratge. On s'han eliminat els valors nulls de la taula de dades i punts fora de la zona de cobertura.

D. Descripció de la base de dades dels AP

A part de la taula dels registres dels usuaris, està la taula dels APs. Aquesta taula ens aporta informació sobre la ubicació,

el seu identificador i la zona on pertany. Aquesta taula inclou els camps:

- **mac**: Indica la direcció MAC del AP.
- **name**: Nom del AP, que va acord per la seva ubicació.
- **lat**: Latitud del dispositiu per geolocalitzar-lo.
- **lng**: Longitud del dispositiu per geolocalitzar-lo.
- **zona**: Identificador numèric de la zona on pertany el dispositiu.

Aquesta taula ha presentat dos problemes: el primer és que no tots els APs que estan presents als registres dels usuaris estan a dins aquesta base de dades, per fer-se una idea solament estan representats 114 dispositius de 198, solament un 58 %. El segon problema que s'ha trobat és la classificació dels APs per zones. Això és deu a què alguns APs no pertany a la zona que estan classificats. També està el cas que una zona és tan àmplia, el que significa és que conté un gran nombre d'APs i que podria ser dividida en diferents zones fàcilment.

Per resoldre els problemes primer s'ha escollit resoldre el problema de les zones. Primerament s'han recollit els 114 dispositius i s'han plasmat a un mapa amb les seves respectives zones. Llavors manualment s'han reagrupat els APs i creat noves zones. En actualitzar la taula s'ha afegit una columna nova **zona name** que és el nom de la zona, ja que l'identificador numèric dificulta la lectura de la distribució de zones. Finalment s'ha creat una nova taula amb informació de les zones, que contenen: nom, identificador i la seva àrea de cobertura (longitud-latitud màximes i mínimes per delimitar una zona en forma de rectangle).

A continuació per la resolució del problema de la falta d'APs. La primera passa és obtenir totes les entrades dels APs que no estaven presents inicialment, a partir de les mostres d'un dia. El següent pas ha sigut obtenir la ubicació dels APs que faltaven, aquesta tasca s'ha fet a través d'obtenir la mitja de la latitud i la longitud de tots els registres dels dispositius vists. Els resultats permeten valorar si aquests APs pertanyen a una zona ja creada. Pel cas que no pertanyin a cap zona creada, els no classificats es visualitzaran a un mapa i manualment s'assignarà un nom i una zona (es pot incloure a una zona ja creada o per altra part es crea una nova zona). Per últim, en haver finalitzat aquesta tasca, s'actualitzarà la taula de les zones APs amb les noves àrees creades i modificades (per la incorporació de nous APs). L'algoritme utilitzat per classificar

Figura 6. Tots els APs distribuïts en zones

els altres APs es pot veure a l'algoritme 1.

Algorithm 1: Classificació dels APs sense registre

```
for Per cada AP do
    Obtenir llistat de registres de que contengui AP;
    lat = latitud mitja de tots els registres ;
    lng = longitud mitja de tots els registres ;
    posicioAP = lat, lng ;
    if Si la posicioAP pertany a una àrea ja creada
      then
        nomAP = esciure nou nom ;
    else
        nomAP = esciure nou nom ;
        àrea = crear una àrea nova ;
    end
end
Actualitzar taula àrees ;
```

En un principi hi havia 9 zones declarades, però en realitzar la revisió feta ha permès crear més de 50 zones diferents. On algunes zones abasten una gran zona com pot ser Son Sardina i d'altres ocupen menys com Plaça d'Espanya a causa del nombre d'APs. Es pot observar a la figura 6 tots els APs que estan pintats d'un color segons la zona que pertanyen.

IV. Classificació de dispositius

El primer objectiu d'aquest projecte és proposar un mètode de classificació dels dispositius a partir d'informació temporal i espacial per realitzar dues classificacions:

1) Per la seva aparició dins la xarxa, el dispositiu es pot classificar com a esporàdic o habitual.
   - **Esporàdic**: Dispositiu que apareix poc a la xarxa Wifi de Palma, a raó entre una, dues o tres vegades en un dia.
   - **Habitual**: Dispositiu que apareix un nombre considerable de vegades, un exemple seria més de 5 entrades, al llarg del dia a la xarxa Wifi de Palma.
2) Per el seu tipus de moviment es poden classificar com a estàtics o mòbils.

- **Estàtic**: Dispositiu que no canvia la seva geocalització al llarg del dia.
- **Mòbil**: Dispositiu que canvia de posició al llarg del dia.

Abans d'executar el codi, s'ha realitzat una ordenació cronològica a partir del camp seentime. Això és a causa dels problemes de lag citats anteriorment en el processat previ de les dades.

A. *Descripció de la implementació*

La implementació està dividit en dues parts diferents:

1) Distinció entre habitual o esporàdic.
2) Distinció entre mòbil o estàtic.

Cada part s'explicarà per separat per facilitar la seva descripció.

1) Distinció de dispositius habituals o esporàdics: Aquesta classificació s'obtendrà a partir del nombre de registres de cada aparell. Si les entrades d'un dispositiu superen un nombre mínim, aleshores el dispositiu quedarà seleccionat com a habitual, en altre cas, serà classificat com a esporàdic.

Una consideració a tenir en compte, és el cas d'un dispositiu que s'ha vist moltes vegades en un període de temps molt petit (un exemple, seria un telèfon mòbil que conté solament 20 visualitzacions en un període de 5 segons al llarg d'un dia), el resultat de la classificació seria habitual, però hauria de ser esporàdic perquè només es veu en aquest període de 5 segons. La solució en aquest escenari és: alhora de comptabilitzar les entrades es posarà un limit de temps perquè les mostres que estan juntes en el marc temporal solament contin una sola vegada, aquest temps s'anomenarà el **temps mínim entre mostres**, aquest temps dependrà del temps que hi ha entre registres dels usuaris. Perquè la definició d'habitual es que s'ha de veure al llarg del dia i no solament en un període de temps concret, com a conseqüència directa el dispositiu haurà de visualitzar-se al llarg del dia un parell de vegades.

Per diferenciar entre habitual i esporàdic es necessari delimitar un límit de mostres on si un dispositiu el supera es considerarà habitual i en altre cas esporàdic, i també per indicar les mostres vàlides, per fer la diferència entre habitual i esporàdic, també es necessari cercar el temps mínim entre mostres. És realitzarà un estudi d'una mostra dels registres d'un dia. On és mirarà el nombre de connexions diàries i, per altre part, és recollirà un usuari amb moltes visualitzacions per observar la diferencia de temps entre registres.

Després d'haver realitzat l'estudi, en el nombre de connexions, han sorgit uns 21365 dispositius on la mitja de connexions és de 1.794 i la mitjana és 1, es poden mirar els resultats a la figura 7. Com que la gran majoria de dispositius solament té 1 mostra, resultat impossible determinar el nombre mínim de mostres. A arrel d'aquest fet, és farà el mateix estudi però amb

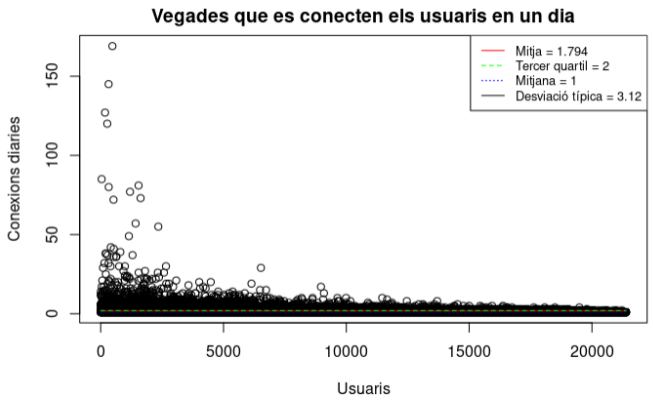

Figura 7. Connexions diàries dels dispositius en un dia.

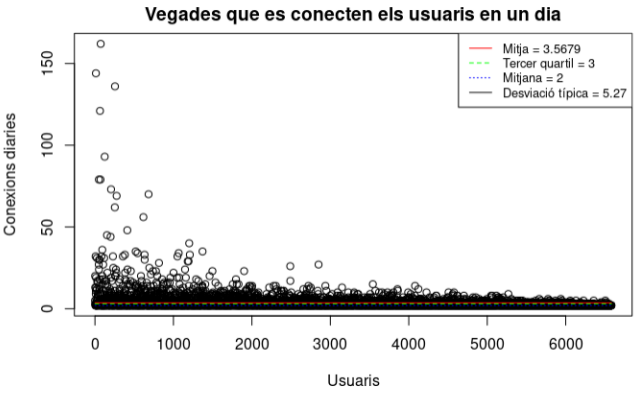

Figura 8. Connexions diàries dels dispositius que tenen 2 o més registres.

aquells dispositius que tinguin dues o més mostres. Aquesta vegada, els resultats han donat una reducció en el nombre de dispositius on passa de 21365 a 6000 dispositius, una reducció de més del 50 %, on la mitja de connexions queda en 3.57, es pot observar els nous resultats a la figura 8. Aquest valor, al arrodonir-lo, quedarà a 4 connexions diàries mínimes per a que un dispositiu sigui considerat com a habitual.

Per altre part està el temps mínim entre mostres. Al final, la selecció ha estat un usuari que té més de 350 registres. En visualitzar els resultats, com s'observa a la figura 9, s'obté un escenari, del qual s'ha comentat abans. És l'existència de mostres (registres) que casi no hi ha diferència ja que la diferència es casi zero. A través d'aquesta visualització de les mostres en el marc temporal, la selecció del temps mínim entre mostres ha estat la meitat del primer quartil $\frac{243.15}{2} \simeq 120s$, que en arrodonir-lo queden 2 minuts.

Una vegada ja obtinguts els valors necessaris, donarà començament l'algoritme de classificació. El primer pas és obtenir una llista de tots els usuaris i anar recorrent-los un a un recollint els seus registres. A partir de l'últim registre vist és recollirà com a referència i s'anirà calculant la diferència de temps, si és supera el mínim de temps entre mostres és comptabilitza per realitzar la diferència entre habitual i esporàdic i s'escull la nova mostra per calcular la nova diferencia

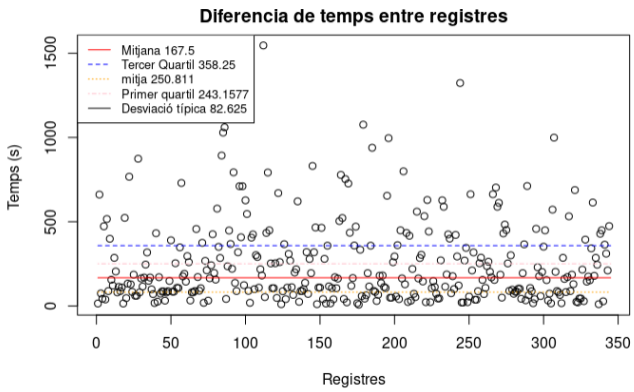

Figura 9. Distribució del temps entre mostres d'un usuari.

amb el següent. Si la diferència fos menor que l'establerta, no es comptabilitza però és manté el registre per calcular la nova diferència, així successivament fins que és comptabilitzés la següent. Finalment sí el comptador del dispositiu supera el llindar és classifica com a habitual, en altre cas, com esporàdic. És pot observar l'algoritme a 2.

Algorithm 2: Classificació de dispositius: habituals o esporàdics

```
PeriodeMinim = 120 ;
ContadorMin = 4;
ContadorUsuari ;
for Per cada usuari do
    Contador usuari = 0;
    Obtenir llistat de registres de l'usuari;
    for Registres de l'usuari do
        Diferencia de temps = Registre següent -
          Registre actual ;
        if Si no supera el PeriodeMinim then
            És manté el registre actual per realitzar la
              següent diferencia ;
        else
            ContadorUsuari ++ ;
        end
    end
    if Si el contadorUsuari supera el contadorMin
      then
        Usuari Habitual ;
    else
        Usuari esporàdic ;
    end
end
```

2) Distinció de dispositius entre mòbils o estàtics: En aquesta secció es classificarà la mobilitat dels dispositius si són de caràcter estàtic o mòbil.

El primer punt a considerar, a partir dels resultats anteriors, és que els dispositius esporàdics no es poden classificar per la falta de registres, no hi ha suficient informació per conèixer l'estat de l'aparell. Per aquesta raó, aquesta nova distinció es realitzarà a continuació de la classificació anterior i no tenir un registre buit, en els casos d'esporàdics, quedaran marcats com a inclassificables, es desconeix la seva actitud en el tema de la mobilitat.

Per classificar els dispositius es comença com en el cas anterior, on primer s'obtendran els registres dels dispositius habituals. I si en tots els registres no canvia l'AP que el visualitza al llarg del dia, llavors aquest es classificarà com a estàtic, en canvi, si l'AP canvia amb el temps es classificarà com a mòbil. Es pot observar a l'algoritme 3.

Algorithm 3: Classificació de dispositius: mòbils o estàtics

```
for Per cada usuari habitual do
    Obtenir llistat d'AP que s'han connectat;
    Eliminar APs repetits ;
    AP contador = APs usuari ;
    if AP contador == 1 then
        Estàtic ;
    else
        Mòbil ;
    end
end
```

Un possible canvi a l'algoritme seria utilitzar la geolocalització de les mostres en comptes dels APs que els visualitzen. On si la longitud i la latitud del dispositiu no varia al llarg de tots els registres(donant un cert marge d'error del 5 %) es classificaria com a estàtic, sinó, es consideraria mòbil.

V. Classificació de la mobilitat

En aquesta secció es proposa una metodologia per analitzar quins modes de transport que utilitzen els vianants a la ciutat.

A. Descripció del problema

A dins la ciutat Palma cada dia circulen una gran varietat de modes de transports (mobilets, patins elèctrics, autobusos, autocars, caminar, córrer, trotar, cotxes...) cada un d'aquestes modalitats tenen les seves característiques com la velocitat mitja, les rutes que poden usar, la velocitat màxima que poden arribar, la velocitat mínima, etcètera. En aquesta secció intentarà obtenir quines són les modalitats de transport que s'utilitza per moure's entre les barriades i llocs emblemàtics de Palma.

El problema principal és la falta d'una referència sobre la relació directa dels modes de transport dels dispositius amb les dades en brut que hi ha sobre la xarxa Wifi. Si per una casualitat hi hagués aquestes referències, resulta que hi ha un altre problema, com es pot diferenciar entre transports que solen fer la mateixa ruta i que les velocitats són similars. Un escenari seria diferenciar una bicicleta i un corredor, solen anar en velocitats similars i solen compartir rutes.

D'altra banda està el problema de les dades aberrants i renou en la captura de mostres. En analitzar les velocitats poden sorgir punts aberrants que superen els 100 km/h, fet impossible a dins la ciutat. També poden aparèixer

dos registres d'un dispositiu en localitzacions diferents en el mateix instant de temps, fet que provoca una velocitat infinita. Aquestes situacions s'han de resoldre abans d'analitzar models de transport.

Per últim s'ha de tenir en compte un fet important: no totes les mostres d'un usuari en un dia corresponen al mateix mètode de transport, poden correspondre a un variat. Un escenari pràctic: *Un usuari al matí pot estar assegut en un restaurant i l'horabaixa se'n va a caminar pel passeig marítim*. Significa que s'hauran d'identificar els diversos mitjans que utilitza cada dispositiu i, a més, existeix la possibilitat que l'usuari estigui un temps en repòs encara que s'hagi classificat com a mòbil.

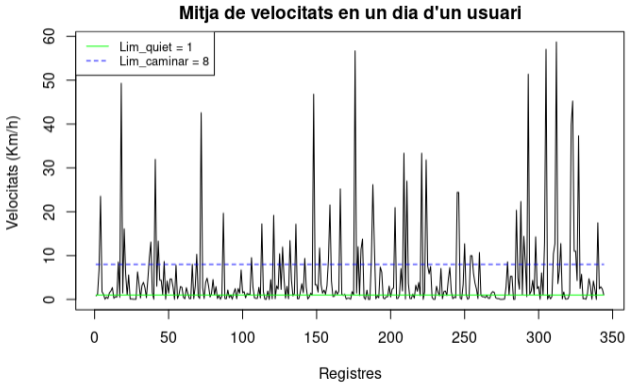

Figura 10. Conjunt de velocitats sense suavitzar

B. Descripció de la implementació

El primer pas és realitzar un filtratge dels dispositius, solament es tendran en compte els dispositius classificats com a mòbils per l'anterior algoritme, ja que els dispositius estàtics no van en cap transport i els esporàdics no es poden classificar. Aquesta acció proporciona una millora a l'eficiència del programa en reduir el nombre de dispositius per classificar.

Respecte a les dades aberrants abans de posar en marxa l'algoritme es tractaran amb una transformació, que tractarà dos fets: primer, casos on les velocitats superen un llindar i, segon, les que donen infinit (diferents localitzacions en el mateix temps).

Cada transformació es realitzarà en ordre. Hi ha dues iteracions:

1) **Dos punts recollits al mateix instant de temps**: En aquest cas es collirà la ubicació mitjana entre els punts, se sobreescriurà al primer i s'esborrarà el segon.
2) **Velocitats que sobrepassen un llindar**: En aquests casos, s'obtindrà en punt entremig i un temps entremig entre l'anterior registre i el següent, sobreescrivint l'actual. Tots els punts que superen els 60 km/h es consideraran aberrants, per les mateixes limitacions a dins la ciutat (50 km/h)[3] donant un marge pels serveis d'emergència (que per motius de seguretat poden superar la velocitat límit).

Aquestes dues iteracions no es realitzaran de manera simultània, sinó que s'executaran en ordre: primer els de punts diferents en el mateix instant de temps i segon, les velocitats que sobrepassen un llindar, ja que en executar la primera execució pot donar com a resultat que s'hagi d'executar la segona (Un dispositiu té dues mostres en el mateix temps, s'executa la primera iteració per arreglar-ho, però dona que la nova velocitat és de 90 km/h que supera el límit establert, aleshores també s'executarà la segona).

[3]En aquest estudi no es té constància de la reducció de la velocitat màxima en alguns carrers

El següent pas és la distinció si una persona sol utilitzar un mitjà de transport o varis. La solució: dividir el grup de velocitats d'un usuari en subgrups on la diferència de temps entre l'última mostra del subgrup i la primera del següent subgrup sigui suficient perquè es considerin independents i es puguin treballar per separat (podria estar el cas que un usuari tingui tres subgrups i que tots fossin el mateix transport). Possiblement en dividir les dades en subgrups, aquests poden tenir poques mostres, en aquests casos no es classificaran per la falta de dades i passarà com el subgrup no hagués existit (si tots els subgrups d'un usuari falten dades, s'indicarà que falten dades per classificar el transport). Per trobar el límit de temps es revisarà la gràfica de les mostres d'un usuari. Com es pot observar a la figura 9, ensenyada anteriorment. Es pot observar un registre, entre altres, on la diferència entre l'anterior i aquest és de 1500 segons (25 minuts). Significa que possiblement el dispositiu ha sortit de la zona de cobertura i ha tornat a entrar, per tant aquest tram no es pot classificar. Aquest no és l'únic cas, es poden observar registres sobrepassant el tercer quartil, on la superen amb una gran diferència. En conclusió, se situarà el límit de temps a 358.2 segons per diferenciar els subgrups d'un aparell electrònic.

Per calcular l'espai en metres entre els registres s'ha utilitzat la fórmula de Harvesine [14] la qual a partir de dues latituds i longituds ens torna la distància entre aquestes en metres o kilòmetres. La qual s'utilitzarà per calcular les velocitats a partir de la fórmula $\frac{espai}{temps}$ en cada instant de temps. Per visualitzar-les s'escollirà l'usuari anterior (el qual s'elabora l'anàlisi d'aquest algoritme). Es pot veure a la figura 10 les velocitats de l'individu sense dividir-les en subgrups. Aquesta imatge mostra el que s'ha comentat prèviament, hi ha conjunts de velocitats a zero i pujades importants, que mostra indicis que no totes les mostres de l'individu pertanyen al mateix transport. També s'observa pics importants en les velocitats que passen de 5 a 50 km/h en un instant de temps i possible renou de les dades. Pel que s'executarà una suavització.

Les velocitats de l'usuari es poden considerar una sèrie

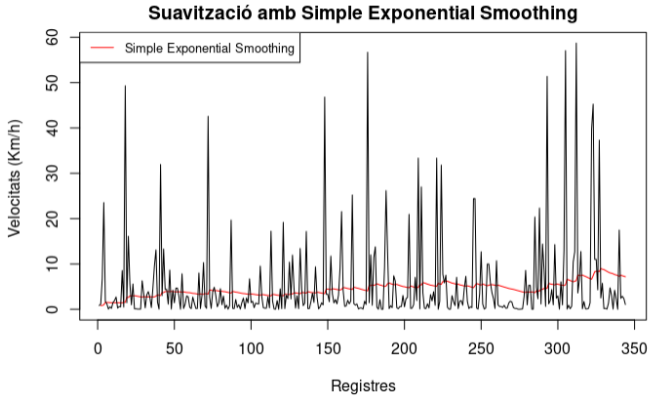

Figura 11. Conjunt de velocitats suavitzades a través del mètode Simple Exponential Smoothing

temporal, amb la qual es pot aplicar una suavització de les dades [15] per solucionar pics importants i renou a les dades. La suavització d'una sèrie de dades, utilitzada en sèries temporals, permet atenuar o eliminar els possibles errors que s'han detectat a les mostres. Per realitzar-la es poden aplicar diferents mètodes, entre altres:

1) **Moving average k=1,3,5...**: Es fa la mitja amb els valors propers dels dos costats, amb un valor k petit dóna més importància als valors recents.
2) **Weighted moving average k=1,3,5...**: És el moving average d'un costat, valors anteriors, en el que s'assigna cada valor de la sèrie un pes diferent per fer la mitja.
3) **Simple exponential smoothing**: Un cas específic del weighted moving average. Que inclou totes les dades passades de la sèrie. On el pes de les mostres depèn de la constant α, on $0 < \alpha < 1$.

A l'hora d'elegir un mètode de suavització no existeix un mecanisme directe que ens indiqui el millor suavitzat possible. Per aquesta raó l'elecció final és el del mètode Simple exponential smoothing, ja que aquest mètode té en compte totes les velocitats restant importància a les més antigues i aconsegueix eliminar més bé els pics de velocitat. Com els pesos depenen de α, el propi suavitzat permet obtenir una α optimitzada per cada cas. Es pot observar a la figura 11 una suavització de les velocitats mostrades anteriorment amb el mètode seleccionat.

Una vegada ja obtinguts els subgrups amb les velocitats suavitzades. Començarà la classificació segons el mitjà de transport. Els transports s'han dividit en tres blocs que són fàcilment diferenciables entre ells. A cada bloc se li ha assignat un rang de velocitats, on és dificultós confondre els blocs. El rang per cada bloc s'ha recollit a partir d'un estudi realitzat on obtenien informació de diferents transports [16], aquests són:

1) **Quiet o assegut**: Els individus poden realitzar accions on estiguin estàtics o que no es moguin molt, un exemple seria anar a una tenda. Menys de 2 km/h
2) **Caminar**: Individus que van a una velocitat constant i lenta per la ciutat de Palma. Entre els 3 i 7 km/h.
3) **Altres**: Tots els altres mitjans de transport com podria ser el patí elèctric, la bicicleta, el cotxe... Entre els 15 km/h i 50 km/h.

En un principi es va pensar a usar un comptador per cada bloc, on cada vegada que una velocitat entri en el rang d'un bloc augmentés el comptador i al final el bloc que tengués més seria el bloc seleccionat. Si es fes d'aquesta manera, seria una mala implementació, ja que hi ha altres factors com els diferents APs que l'han vist o en quantes zones s'ha mogut, a més està el cas que cap velocitat entrés a dins els blocs i per tant la seva classificació quedi buida.

Llavors per la classificació s'ha optat en seguir les pautes del machine learning [17] simplificant-ho a un únic algoritme, ja que una de les passes és l'avaluació de l'algoritme que contrasta els resultats del model amb els reals i avalua la precisió del model, cosa que no es realitzarà per la falta d'una font de dades que indiqui el dispositiu amb el transport que ha anat aquell dia. L'algoritme seleccionat ha estat el de Multinomial Logistic Regression [18]. Que permet categoritzar una variable dependent (el tipus de transport) basat en múltiples variables independents (com són les probabilitats dels transports o el nombre diferents d'APs que es registren). Les passes a seguir per donar els resultats finals són:

1) Obtenir la font de dades per iniciar la classificació.
2) A partir de la font de dades, aconseguir els indicadors que s'utilitzaran per classificar els transports.
3) Crear una llista que faci l'associació dels indicadors i els transports.
4) A partir de la llista creada crear un model de predicció de transports.
5) Utilitzar el model per classificar tots els usuaris.

Els indicadors que s'han recollit per realitzar el model són:

- *Probabilitat de Quiet*: El nombre de vegades que la velocitat entra en aquest rang dividit el nombre de velocitats.
- *Probabilitat de Caminar*: El nombre de vegades que la velocitat entra en aquest rang dividit el nombre de velocitats.
- *Probabilitat de Altres*: El nombre de vegades que la velocitat entra en aquest rang dividit el nombre de velocitats.
- *Diferents APs*: Percentatge d'APs únics que l'han vist entre tots els APs.
  - $diferentsAP \approx 1$: L'han vist molt d'APs diferents, indica que s'ha mogut dins la ciutat.
  - $diferentsAP \approx 0$: L'han vist pocs APs diferents, indica que no s'ha mogut molt a dins la ciutat.
- *Diferents àrees Wifi que s'han visitat*: Nombre d'àrees úniques que han vist el dispositiu, en la franja del subgrup.

Per crear el model es requereix una font de dades que mostri una relació amb aquestes dades i el transport usat, però aquesta relació no es coneix. Per solucionar-lo es crearà un artificial. On es realitzarà una iteració i els subgrups que tenguin una alta probabilitat de quiet, caminar o altres es classificaran directament per crear la font de dades (casos on superen més 40 %[4] la probabilitat de quiet, caminar o altres).

[4]El valor original era del 50 %, però com que el resultat no era l'esperat s'ha reduït

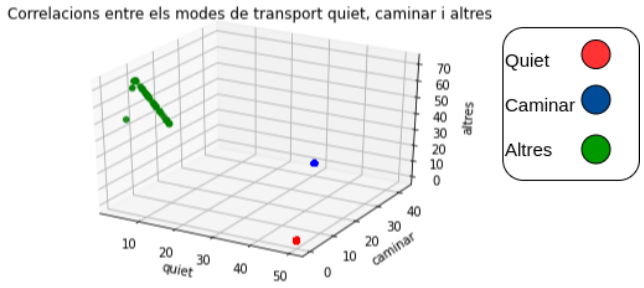

Figura 12. Correlació entre les probabilitats de cada mode segons les seves velocitats

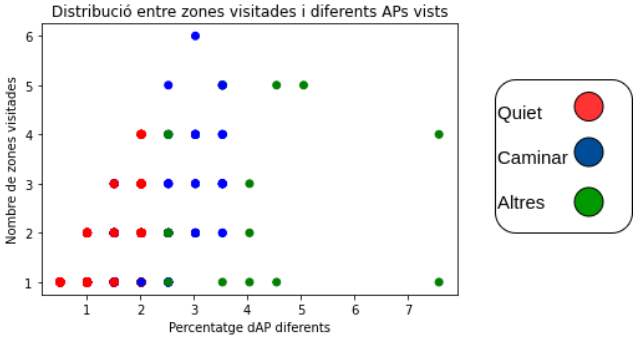

Figura 13. Correlació entre les zones visitades i el percentatge d'APs diferents

Si s'observa la correlació entre els diferents punts a través de les probabilitats d'anar en diferents transports, com es pot veure a la figura 12, mostra que els modes de quiet són fàcils de diferenciar perquè no tenen cap punt en altres blocs, llavors estan els d'altres que no són perfectes, ja que contenen alguna probabilitat de caminar, però en canvi, els que es classifiquen com a caminar són difícils, ja que compten amb una probabilitat considerable que estiguin quiets. Si es mirés la correlació entre les probabilitats de les zones visitades i diferents APs, com s'observa a la figura 13, els que visiten diferents zones són més fàcils de classificar perquè tots són altres i caminar, a més s'observa que la majoria dels quiets tenen molt poca probabilitat que els observin diferents APs i solament estan a una zona Wifi.

Algorithm 4: Classificació dels tipus de transports primera part

```
for Per cada usuari mòbil do
    Corregir errors de velocitats ;
    for Per cada mostra do
        if temps mostra < deadline o no hi ha més
         mostres then
            Afegir mostra al subgrup ;
        end
        else
            Creació del subgrup: ;
            Suavització de les velocitats ;
            Obtenció dels diferents APs connectats ;
            Obtenció de les diferents zones que han
             passar ;
            Obtenir les probabilitats d'anar en els
             diferents modes de transport ;
            Crear un llistat amb les dades ;
        end
    end
    Obtenir llistat transports obvis ;
    Crear el model ;
    Obtenir els resultats ;
end
```

Una vegada ja creat el model, es poden classificar tots els valors i donarà com a resultat els transports, cada transport està format per una tupla de valors:

1) El transport elegit.
2) L'hora de començament del transport
3) L'hora final del transport

En cas de que un usuari mòbil tengui subgrups on cada un d'ells no superen el mínim per classificar el transport quedarà com a *Falten dades de transport*, es pot observar l'algoritme complet a 4.

En aplicar aquest algoritme a l'usuari que s'han mostrat les velocitats per realitzar les probes, dividint el grup de velocitats en subgrups, suavitzant-los i aplicant l'algoritme queda el següent gràfic 14. Es pot observar tres nivells que corresponen als modes de transport i les hores d'inici quan comença el transport. S'observa que hi ha períodes que l'usuari manté el mateix mode de transport i que no hi ha casi ningun canvi brusc entre els transports.

VI. Obtenció de les rutes habituals

L'objectiu de l'últim punt és obtenir els patrons de mobilitat més habituals i importants que es realitzen a la ciutat de Palma. Aquest objectiu s'aconseguirà a través de les regles d'associació. Aquest concepte intenta cercar patrons freqüents, correlacions, associacions o estructures casuals sobre sets d'items o objectes a dins bases de dades transaccionals, relacionals i altres repositoris [19], aquest concepte és molt utilitzat en el camp dels supermercats. On cada regla està format per un o més antecedents i un o més conseqüents a partir d'un conjunt definit d'objectes (En el cas de l'estudi són les diferents zones). Un exemple de regla d'associació a

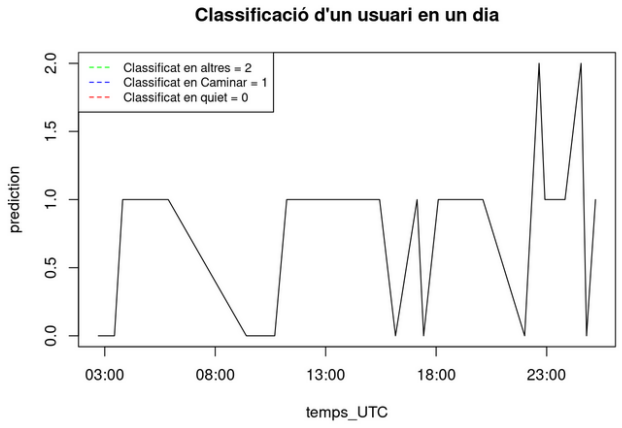

Figura 14. Resultat final al haver classificat el model

dins la ciutat de Palma seria:

$$\{Plaça\ Espanya\} \rightarrow \{Plaça\ Major\}$$

Aquesta regla indica que les persones que van a la Plaça Major passen per la Plaça d'Espanya, aquest antecedent i conseqüent formen una regla. Cada regla ve acompanyada per diversos indicadors que mesuren la seva importància. Les regles d'associació contenen diferents possibles aplicacions reals a dins les ciutats:

- A partir de les rutes més habituals, optimitzar les rutes i la freqüència dels transports públics.
- Realitzar l'estudi de la mobilitat de les persones segons els diferents dies de l'any.
- Optimitzar la implementació de nous serveis, com per exemple, instal·lar-les a les localitzacions més calentes (on hi ha molta gent).

A. *Descripció del problema*

Les regles d'associació són un dels conceptes del machine learning que generarà regles a partir d'un conjunt d'items, on sorgeix el primer problema, quina ha de ser la selecció del conjunt d'objectes, ja que hi ha zones on els APs estan uns sobre els altres. Cada regla necessita indicadors per mesurar la seva importància dels quals, a l'estudi s'han usat els següents de la gran quantitat d'indicadors d'importància [20] existents.

- **Suport**: Indica la freqüència d'un itemset en tots els conjunts d'elements.

$$Support(\{X\} \rightarrow \{Y\}) = \frac{Viatges\ amb\ X\ i\ Y}{Tots\ els\ viatges}$$

- **Confiança**: Juntament amb el suport són dels més importants. Aquesta mesura l'ocurrència del conseqüent donat l'antecedent.

$$Confiança(\{X\} \rightarrow \{Y\}) = \frac{Viatges\ amb\ X\ i\ Y}{Viatges\ amb\ X}$$

- **Lift**: Mesura que compta les vegades que ocorren X i Y junts que les que haurien de sortir si són estadísticament independents. Un valor d'1 índica independència entre X i Y.

$$Lift(\{X\} \rightarrow \{Y\}) = \frac{Confiança(\{X\} \rightarrow \{Y\})}{Support\{Y\}}$$

- **Leverage**: També conegut com la mesura de Piatetsky-Shapiro (PS). Un valor, que mesura la diferència entre que apareguin junts X i Y i l'expectant si X i Y fossin estadísticament dependents. Es tracta de cercar si visiten més (X i Y) que X i Y independents.

$$PS(X \rightarrow Y) = supp(X \rightarrow Y) - supp(X) * supp(Y)$$

- **Conviction**: Valor de mesura sorgit en contrapartida de la confiança. Compara la probabilitat que aparegui X sense que aparegui Y si són dependents i l'actual freqüència que aparegui X sense Y.

$$conviction(X \Rightarrow Y) = \frac{1 - supp(Y)}{1 - conf(X \Rightarrow Y)}$$

Les regles d'associació, comparant-les als altres algoritmes, tenen un gran cost computacional. En el nostre cas, on les dades d'entrada superen el mig giga de bytes cada dia. Aquest fet significa un retard important en l'execució del codi, amb la possibilitat que s'interrompi per falta de recursos de la mateixa màquina on s'executa. Per aquesta raó el codi s'hauria d'optimitzar el màxim possible.

Les passes a seguir per obtenir les regles d'associació són: primer obtenir el conjunt d'itemsets dels usuaris, segon obtenir el conjunt d'itemsets freqüents indicant un suport mínim i finalment crear les regles indicant una confiança o lift mínima. El problema sorgeix en les dues últimes fases, on és necessari indicar un paràmetre de suport mínim per obtenir l'itemset freqüent i dos paràmetres per crear les regles: la mètrica (lift o la confiança) i el seu valor mínim. Aquests valors no són sempre els mateixos, sinó que els paràmetres d'entrada no estan fixats depenen de la persona que les està cerca en aquell moment. Pot ocórrer que en introduir els paràmetres d'entrada les regles obtingudes siguin un conjunt buit, ja que no hi ha regles que compleixen els requisits donats per l'usuari que les cerca, com a conseqüència directa haurà de tornar a executar el codi per ficar-li uns altres paràmetres afegint el retard corresponent de tot el procés.

B. *Descripció de la implementació*

El primer pas és seleccionar el conjunt d'items per poder crear les regles. En primera instància estava planejat utilitzar el llistat d'APs, però quedà descartada, ja que estan aferrats uns als altres el qual donaria regles entre veïnats. També estava l'opció de recollir un subconjunt d'aquests APs, però amb aquest procés s'elimina informació i no agrada. Finalment s'ha decidit per utilitzar les zones dels APs, on cada zona està format per un subconjunt dels APs. Aquest fet permetrà obtenir regles més acurades, ja que les zones estan separades i tenint un nombre reduït d'items provoca una baixada del cost computacional.

Com passava a l'anterior algoritme, no tots els dispositius ens aporten informació, els esporàdics i estàtics no aporten res sobre els patrons de moviment. Per aquesta raó solament es tendran en compte els dispositius mòbils, fet que reduirà el nombre de dispositius, d'aquesta manera baixarà el cost computacional.

En un principi per realitzar l'estudi sencer apareixia un factor que no es tenia en compte, era el factor del temps. Fet que pot ocasionar canvis en els resultats. Aquest factor s'ha de tenir present, ja que les persones canvien d'actitud al llarg del dia, l'exemple directe és un individu on el matí se'n va a fer feina a l'oficina, l'horabaixa va a passejar i el vespre va a ca seva a dormir. Per aquesta raó es dividiran els dies en franges horàries segons en quin dia de la setmana cau, la idea està treta de l'article comentat prèviament [10]. La divisió serà la següent:

1) **Feina**: Període que consta entre dilluns i divendres entre les 8:00 i les 17:00.
2) **Entreteniment**: Període que consta de dilluns i divendres entre les 17:00 i les 00:00, a més dels dissabtes i diumenges de les 10:00 i les 00:00.
3) **Descans**: Període que consta de dilluns i divendres entre les 00:00 i les 8:00, a més dels dissabtes i diumenges de les 00:00 i les 10:00.

A l'inici de l'algoritme l'usuari seleccionarà la franja horària com entrada a l'algoritme de classificació (si vol obtenir els resultats de dues franges horàries diferents, haurà d'executar el codi dues vegades). El següent pas és obtenir una llista on cada element conté les zones visitades per cada usuari mòbil en el període de la franja seleccionada. Aquest llistat conté etiquetes amb els noms, però suposa un problema pels algoritmes d'aprenentatge automàtic, ja que la majoria d'aquests necessiten taules del tipus *One-hot encoding* [21]. Aquesta taula elimina les etiquetes dels noms i afegeix una nova variable binària que indica si les zones estan presents a les files. Finalitzant els passos, s'aniran eliminant variables pesades per alliberar espai de memòria.

Per crear els itemsets freqüents és necessari indicar un suport mínim, aquest valor no és fix i depèn de la situació. A partir d'aquest requeriment i que cada usuari és diferent s'implantaran entrades al llarg del codi, les entrades serveixen perquè l'usuari indiqui el valor del paràmetre via consola. Aquests valors s'utilitzaran per obtenir l'itemset freqüent posant el suport mínim i també per crear les regles on elegiran per realitzar-lo entre el lift o la confiança i el seu valor mínim. Pel fet que l'usuari interaccioni amb el codi s'ha de preveure fer un tractament d'errors per si l'input donat pugui ser erroni.

El resultat final és una taula de regles, on cada fila és un marc de dades format per un o més antecedents, un conseqüent (es limita a ser de longitud 1, per eliminar redundància), els suports de l'antecedent, conseqüent i de la regla i per últim els altres indicadors d'importància.

Algorithm 5: Creació de les regles d'associació

```
Elegir zona horària ;
for Per cada usuari mòbil do
    Obtenir llistat de zones que s'han connectat;
    Eliminar zones repetides ;
end
Transformar les dades al tipus data 1-hot encoded ;
Eliminar variables pesades que ja no s'utilitzen ;
case Opció crear itemsets freqüents do
    passa això
end
case Opció crear regles do
    Indicar el suport mínim Crear el subconjunt
     d'itemsets freqüents ;
end
case Visualitzar regles do
    Crear el conjunt de regles a partir del lift o
     confiança ;
end
case Visualitzador de regles do
    Visualitza les regles a partir d'uns paràmetres. ;
end
case Guardar una còpia en local do
    Guarda una copia en local en format .csv ;
end
case Guardar una còpia a la base de dades do
    Guarda una copia de les regles a la base de dades ;
end
case exit do
    Sortir del programa ;
end
```

Un resultat possible de les regles o els itemsets freqüents és un conjunt buit o amb poques entrades. L'usuari en observar aquest fet hauria de reiniciar el codi i hauria de tornar a esperar el temps perquè l'algoritme crei els itemsets. Per solucionar aquest escenari es crea un menú amb opcions diferents per si un resultat no és el desitjat per l'usuari (conjunt buit o que contengui menys de 10 regles o itemsets), quan passés l'usuari tornaria a repetir la mateixa opció sense haver de sortir i tornar a executar el programa. Les opcions disponibles són:

- **Creació itemsets freqüents**: A partir de l'itemset creat, l'usuari posa un suport mínim per obtenir els més freqüents.
- **Creació de les regles**: L'usuari insereix la mètrica per crear-les (lift o confiança) i també el valor mínim.
- **Visualitzador de les regles**: L'usuari observa el conjunt de regles a partir del millor resultat a una avaluació o regles que contenguin un item en concret als antecedents o al conseqüent.
- **Guardar còpia en local**: Guarda una còpia en local del resultat obtingut en format .csv, on l'usuari incert el nom.
- **Guardar el resultat a la base de dades**: Guarda el resultat a la base de dades (l'explicació vendrà a la següent secció).
- **Exit**: Sortir del programa.

Finalment l'algoritme es pot observar a 5, on consta una

primera fase de la creació dels itemsets i una segona, el menú, per crear les regles i guardar els resultats.

VII. Implementació i optimització

A. Optimització i concurrència

En executar les primeres versions dels diferents codis el temps d'execució ha resultat ser molt llarg, més de 8 hores per algoritme. La principal causa ha estat el gran volum de dades d'entrada, que una vegada tractada segueix essent molt gran. Per tant, els tres algoritmes han requerit una optimització en els seus codis per augmentar l'eficiència. En el cas del primer algoritme s'han fusionat els dos classificadors en un, en el cas del tercer algoritme eliminar variables pesades. Però no ha estat suficient, ja que la reducció del temps era relativament baixa, ara bé, una altra opció que es va plantejar pel retard eren eliminar les operacions de lectura i escriptura a memòria, però en observar que tardaven menys d'un minut per algoritme es descartaren. Finalment el gran canvi que va permetre millorar en gran manera l'eficiència dels codis ha estat implementar multiprocessing [22].

Per entrar en context, en un principi solament es feia ús d'un únic core de la MV on s'executaven tots els codis. Però la mateixa computadora és multi-core (conté més d'un processador, en el nostre cas contenia 4). Així que per millorar l'eficiència dels temps d'execució ha estat utilitzar tots els cores de la màquina com fils de treball per cada algoritme, on cada core treballarà en paral·lel a través del mòdul multiprocessing de python.

Aquest mòdul segueix aquests passos per implementar-se:

1) Crear un fil principal, on s'inicia l'objecte i es divideix les dades per els diferents cores.
2) Cada procés segueix les mateixes instruccions amb inputs de dades diferents i envien els resultats.
3) Rebre les dades dels processos, no avançarà més en l'execució fins que hagi rebut els resultats de tots els fils, és un punt per sincronitzar les dades.
4) Finalment amb les dades es processa l'última part.

El diagrama de flux del funcionament es pot observar a la figura 15, que correspon a un quatre-core.

Per comprovar la millora s'han executat dos codis (una nova versió del primer algoritme aplicant concurrència i una versió antiga sense implementar-lo) amb una mostra de 100 000 registres. Els resultats ha donat que la versió antiga ha tardat 395 segons i la nova, aplicant multiprocessing, ha tardat 207 segons. S'observa una millora del 52 % en el temps d'execució amb una petita mostra. Que escalant-lo a les dades d'un dia provoca una reducció important en el temps d'execució del codi.

B. Base de dades

En l'execució dels algoritmes. Els resultats obtinguts es guardaran de dues formes:

- La possibilitat de guardar els resultats en format .csv a la màquina virtual, depèn si l'usuari vol o no, no és recomanable guardar tots els resultats a la memòria de la VM, per la seva memòria limitada.
- Tots els resultats es guardaran a la base de dades Postgres, ubicació on estan els registres dels diferents dies.

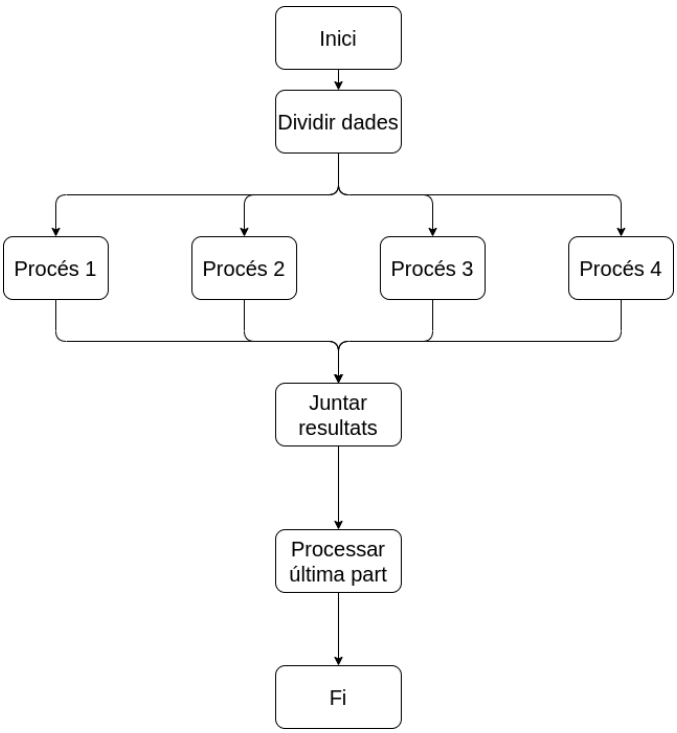

Figura 15. diagrama de flux del multiprocessing.

Per motius de què aquestes taules estan creades per un tercer, l'empresa MallorcaWifi. No es realitzarà cap referència a aquestes, però es mostrarà com s'hauria de realitzar. Finalment l'estructura de la base de dades és la següent:

- **Dispositiu**: Taula que conté la informació de la classificació de dispositius.
  - clientmac: Direcció mac del dispositiu PK (Primary key).
  - dia: Dia en la que s'ha realitzat la classificació PK.
  - class_hab_esp: Classificació de si el dispositiu és habitual o esporàdic.
  - class_mob_est: Classificació de si el dispositiu és mòbil o estàtic.
- **Transport**: Taula que conté informació dels transports utilitzats pels usuaris.
  - idtransport: Índex de la fila per diferenciar les entrades, PK.
  - clientmac: Direcció mac del dispositiu, fa referència a la taula dispositiu.
  - dia: Dia que s'ha realitzat la classificació, fa referència a la taula dispositiu.
  - temps inici: Inici en UTC del transport.
  - temps final: Hora final en UTC del transport.
  - transport: Tipus de transport.
- **AP**: Taula que conté tota la informació dels APs a dins la ciutat de Palma.

  - apmac: Direcció del AP, PK.
  - name: Nom del AP.
  - lat: Indica la latitud del AP.
  - lng: Indica la longitud del AP.
  - zona: Indica a quina zona pertany, fa referència a la taula zona Palma.
- **zona Palma**: Informació sobre les zones a dins la ciutat de Palma.
  - zona name: Nom de la zona, PK.
  - zona nombre: Indicador numèric que indica la zona.
  - lat max: Latitud màxima per determinar la zona.
  - lat min: Latitud mínim de l'àrea que la conforma.
  - lng max: Longitud màxima de l'àrea que la conforma.
  - lng min: Longitud mínima de l'àrea que la conforma.
- **Ruta**: Informació sobre les regles d'associació més importants segons la franja seleccionada.
  - id ruta: Id de la ruta, PK.
  - dia: Dia en el qual es realitza l'estudi.
  - franja: Franja horària en el que s'han creat les regles.
  - antecedents: Antecedents, un conjunt.
  - conseqüents: Els conseqüents, solament hi ha un item.
  - antecedent suport: El suport del/s antecedents.
  - conseqüent suport: El suport del conseqüent.
  - confidence: Valor de mesura d'importància.
  - lift: Valor de mesura d'importància.
  - leverage: Valor de mesura d'importància.
  - conviction: Valor de mesura d'importància.

La figura 16 representa el model de la base de dades.

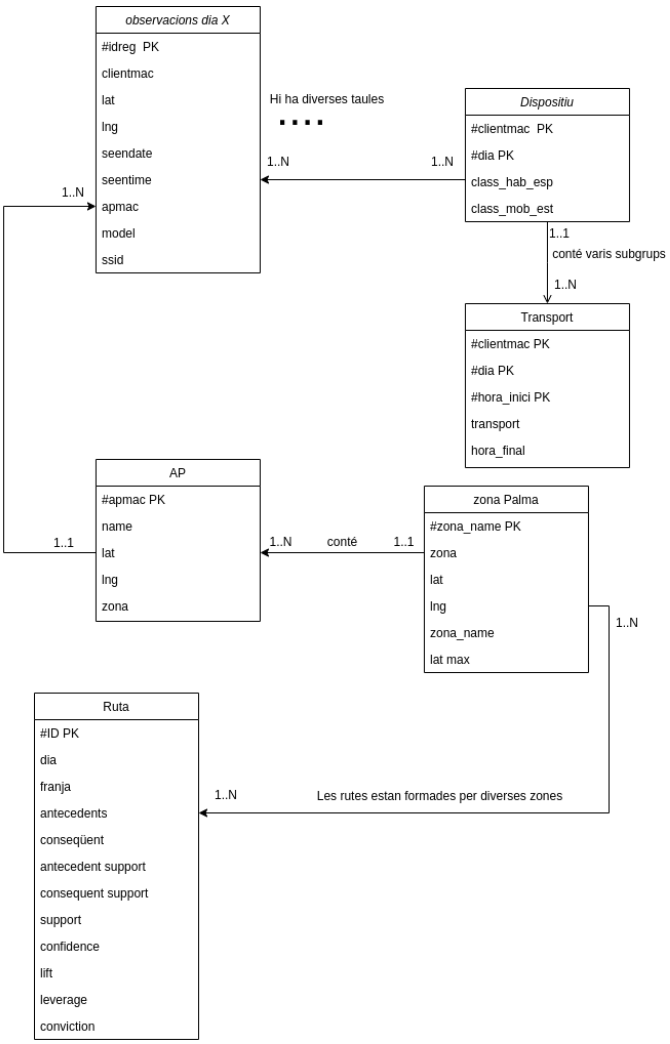

Figura 16. Estructura de la base de dades creada.

C. Codi

Tot el programari es troba localitzat a una pàgina de GitHub [23], format per les següents seccions:

1) **APinformacio**: Conté els algoritmes sobre l'ampliació i reforma dels APs.
2) **AlgoritmesPython**: Són els algoritmes que s'executen a la màquina virtual.
3) **Rscripts**: Conté els processos per crear els algoritmes, està en notebook R i html.
4) **SQLscripts**: Són els SQL scripts utilitzats per crear la base de dades i ficar les dades sobre els APs i les zones.
5) **README**: Breu introducció al Github del TFM.

VIII. Resultats i Discussió

A. Resultats en un dia

Per crear les primeres versions s'ha escollit el 23 de juliol de 2018, les raons han estat que és un dels dies, disponibles a la base de dades, que tenia una gran quantitat de dades i la segona raó ha estat l'aparició de tots els APs que ha permès completar la taula d'APs que estava a mig completar.
Una vegada ja obtinguts els algoritmes creats amb les últimes versions, s'ha escollit un altre dia diferent per provar el seu funcionament i obtenir els resultats, el dia seleccionat ha estat el 6 de juliol de 2019, que cau en dissabte. A diferència del dia seleccionat per les proves té menor longitud i no apareixen tots els APs de la llista, però aquesta ja bé amb un primer tractament on elimina els errors d'ubicació i valors nuls, comentats en la secció de l'organització de les dades.

1) Classificació dels dispositius: En executar la primera classificació han sorgit 124210 dispositius, i s'han categoritzat en els següents resultats:

- Percentatge de dispositius habituals: 19 %
- Percentatge de dispositius esporàdics: 81 %
- Percentatge de dispositius mòbils, en el marc dels habituals: 96 %
- Percentatge de dispositius estàtics, en el marc dels habituals: 4 %

Els resultats mostren que la majoria dels dispositius són esporàdics, aquest fet es pot explicar que una gran majoria de dispositius solament s'han vist una vegada, com es mostra a l'anterior gràfica 7 mostrada per realitzar l'algoritme. Per altra part, en el camp dels habituals la gran majoria d'aquests són de caràcter mòbil, solament hi ha un petit percentatge que són estàtics.

2) Classificació dels transports: Els transports classificats durant el dia han estat de 53710 moviments, que s'han dividit

en:

- Percentatge de moviments categoritzats com a quiet: 21 %.
- Percentatge de moviments categoritzats com a caminar: 45 %.
- Percentatge de moviments categoritzats com a altres: 34 %.

La solució dona que solament el 34 % dels moviments han estat categoritzats com a altres i que la gran majoria s'han categoritzat com a caminar. La durada mitja d'aquests moviments ha estat d'uns 20 minuts per transport, i cada usuari mòbil té una mitja de 2.37 mesures durant el dia.

3) Classificació de les rutes: Com que el dia seleccionat cau en dissabte solament hi ha dues franges horàries, que són la de descans i entreteniment. Per obtenir l'itemset freqüent s'ha posat com a suport mínim el 0.01 % i per crear les regles s'ha post com a mínim lift 0.1.

A la franja de descans, es troba que la zona del passeig del born apareix sovint amb un suport del 0.74%. La regla amb més suport ha estat la de:

$$\{Passeig\ del\ Born\} \rightarrow \{Catedral\}$$

En canvi per mirar les regles amb més lift i confiança sorgien regles amb molt poc suport, per aquesta raó s'han filtrat les regles perquè superin el 10 % del suport. Finalment s'ha identificat que la regla amb millor avaluació del lift amb un valor de 2.34 és la de:

$$\{Plaça\ Major\} \rightarrow \{Plaça\ de\ Cort\}$$

Amb la mateixa limitació del suport si miram la que té més confiança té un 99 % i un suport del 17 %, aquesta és:

$$\{Catedral,\ Plaça\ del\ Mercat\} \rightarrow \{Passeig\ del\ Born\}$$

A la franja d'entreteniment han sortit els següents resultats, la regla que conté el suport més gran, 52 % ha estat la de:

$$\{Catedral\} \rightarrow \{Passeig\ del\ Born\}$$

Seguint el mateix procediment d'augmentar el suport mínim, ha donat que la regla més important segons el lift amb un valor de 3.11 i un suport de l'11 % ha estat:

$$\{Plaça\ de\ Cort\} \rightarrow \{Plaça\ de\ Sant\ Eullia\}$$

I la que té més confiança amb un 99.4 % ha estat la de:

$$\{Llotja\ de\ Palma,\ Catedral,\ Plaça\ del\ Mercat\} \rightarrow$$

$$\{Passeig\ del\ Born\}$$

B. Resultats en una setmana

Pels resultats d'una setmana s'han recollit les dades de la primera setmana de juliol de 2019, de dia 1 a dia 7, de dilluns a diumenge.

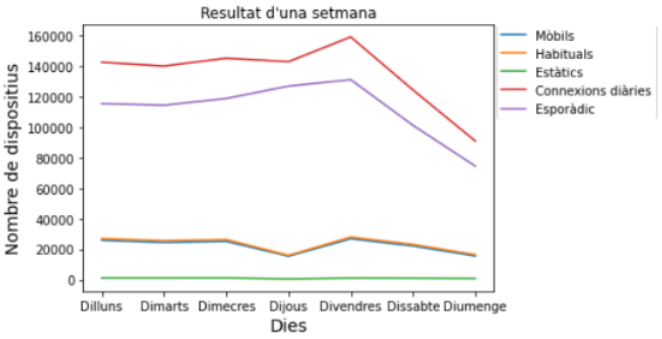

Figura 17. Evolució de la classificació dels dispositius al llarg de la setmana.

1) Classificació dels dispositius: Per obtenir la base de dades d'aquesta classificació s'ha creat una base de dades amb una longitud de 943839 registres verificant que no hi ha cap entrada errònia.

A la figura 17 es pot observar l'evolució de la classificació dels dispositius. El nombre de dispositius a la setmana ronda els 140000, arribant el seu màxim en divendres amb uns 160000 i una baixada el cap de setmana que arriba als 90000. El nombre de dispositius esporàdics manté la mateixa similitud que la corba de dispositius amb una mitja de 120000 dispositius. Per altra part, la corba dels dispositius mòbils i habituals tenen quasi els mateixos valors, amb una baixada el dijous. També estan els estàtics que n'hi ha pocs, però són constants a durant la setmana, ja que la corba es pot considerar una recta, amb una mitja de 1000 dispositius diaris.

També s'ha vist que el 48 % dels dispositius s'observen una sola vegada a la setmana i que la resta, el 52 %, es connecten diverses vegades. També dir que el 90 % dels que s'han connectat una sola vegada a la setmana han estat classificats com esporàdics.

2) Classificació dels transports: La corba dels moviments classificats segueix una similitud amb el nombre de dispositius, amb un màxim el divendres i una baixada el cap de setmana. Els transports com a classificats de caminar són els que hi ha més amb una mitja de 30000. El següent moviment són el d'altres amb una mitja de 20000 i per últim el de quiets amb una mitja de 10000 transports. Es pot observar a la figura següent 18 els resultats de la setmana.

3) Classificació de les rutes: A causa del retard en l'execució dels codis a la màquina virtual s'ha escollit una mostra reduïda per l'estudi, es pot observar a la taula 19 les franges escollides, però es hauria de recollir tota la setmana. Per crear les regles s'ha posat un suport mínim del 10 % i una avaluació del lift del 0.1. En aquest estudi es mostraran taules amb les regles més importants segons els tres paràmetres principals: suport, confiança i lift.

La primera franja estudiada la de descans, la taula es pot observar a la figura 20. Trobam que la primera regla entre setmana no supera el 40 % de la confiança, però en dissabte aquesta augmenta fins al 46,8 %. Per altra part la regla en la

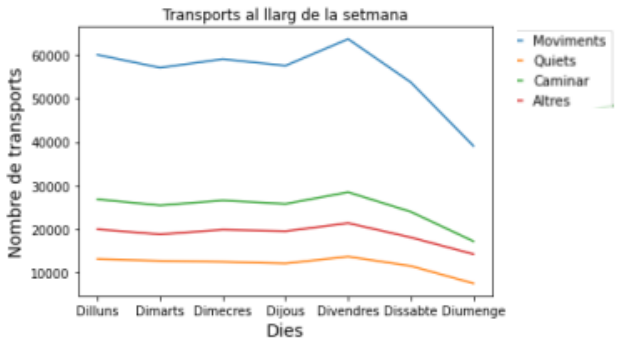

Figura 18. Evolució de la classificació dels moviments dels dispositius al llarg de la setmana.

Dies	Descans	Feina	Entreteniment
Dilluns 1		X	
Dimarts 2		X	X
Dimecres 3	X		
Dijous 4	X		
Divendres 5			X
Dissabte 6	X	////////////////////	X
Diumenge 7		////////////////////	X

Figura 19. Franges de temps seleccionats per realitzar l'estudi de la setmana

qual tenen més confiança són pels tres casos iguals on cada un supera el 95 %, però el suport baixa molt. Per últim s'observa que a través del lift els tres dies són diferents i destacar que el dissabte arriba fins a un valor de 2,34. També indicar que hi ha regles que no apareixen cada dia.

La segona franja estudiada ha estat la de feina, la taula es pot observar a la figura 21. S'observa que les regles més suportades dels dos dies varien, però tenen una diferència molt baixa entre elles. La regla que té més confiança són iguals per a tots dos amb quasi els mateixos valors en tots els camps, i per últim el lift també passa com el cas anterior la similitud dels valors, cal destacar que el lift a un cas arriba a 2,87 i el següent arriba a 3,09 valors alts.

La tercera i última franja estudiada ha estat la d'entreteniment, la taula es pot observar a la figura 22. En un primer moment es pot pensar que la primera regla i la tercera són iguals, però un valor que canvia és la confiança, també queda que la regla amb més suport la té diumenge. El tema de la confiança queda pels quatre dies la mateixa on cada una supera el 99 %. Per últim en el lift mostra que els dies feiners són contraris als del cap de setmana, no apareixen directament, i les regles del cap de setmana superen 3 de lift cada dia.

C. *Conclusió*

A partir dels resultats obtinguts a través d'una setmana de juliol es poden obtenir algunes observacions. La primera, en la classificació dels dispositius, la majoria d'aquests s'han classificat com a esporàdics, la qual cosa afecta directament en una reducció de dades útils d'entrada pels algoritmes i altres futurs, però necessari, ja que no aportaran informació i solament retardaran l'execució dels algoritmes. Per altra part, en tenir la gran majoria dels habituals com a mòbils ha permès que millorin els resultats. La segona, en la classificació dels modes de transport, la classificació obtinguda em satisfà, perquè a causa de no tenir un dataset d'entrenament, la classificació ha quedat variada (que abraça diferents transports) on cap supera el 50 % en cap cas. He d'indicar que en aquest algoritme en poder tenir un dataset d'entrenament, es podria haver ampliat els transports, ampliar les possibilitats de plantejar-se altres algoritmes per avaluar-los. En el tercer algoritme s'han cercat les rutes més habituals dels vianants a partir d'una mostra, com es pot indicar està incomplet, a causa dels recursos de màquina disponible que tardava molt de temps en executar el codi per franja, per això solament s'ha escollit una mostra i no cada dia com s'hauria de realitzar. S'ha tret que el Passeig del Born és una zona important, ja que està present en totes les taules de regles. El període elegit ha estat a principis de juliol on la ciutat de Palma rep a milers de turistes i molta gent té vacacions, també està present que molta gent no té feina i hi ha vacacions a les escoles, cosa que provoca resultats diferents que un altre mes on la temporada turística és baixa i no hi ha vacacions. Per ampliar l'estudi es podria incorporar les diferents èpoques de l'any i observar com evolucionen algunes regles. Un possible estudi seria comparar aquestes dades amb la mateixa franja de la pandèmia i es podria observar com ha evolucionat i que ha canviat en aquests dies.

Un punt negatiu d'aquest treball han estat els recursos hardware on s'han implantat els algoritmes. Ja que la idea inicial era obtenir els resultats en temps real dels diferents algoritmes. Però l'execució dels codis permet observar que tarden molt a executar-se on també cal destacar que en un algoritme s'aturava l'execució d'aquest a causa d'un error de falta de memòria, cal destacar que una vegada optimitzats i arreglats, el temps d'execució eren molt llargs.

Treballar amb una gran base de dades per intentar aconseguir que cap fet passi per alt abans de tractar amb les dades és difícil i s'ha passat un fet per alt el que jo anomenaria **Formes de cada empresa**. En principi s'ha escollit el dia 23 de juliol de 2018 per crear els algoritmes perquè aquesta contenia moltes dades i estaven representats tots els APs. Aquest dia provenia directament de l'API de recaptació de dades on tenia una estructura de noms i posicions de les columnes. En canvi, per l'estudi s'ha escollit la setmana de juliol de 2019 que havia estat ja tractada per l'empresa MallorcaWifi i revisada per la UIB. El problema

Regla	Suport			Confiança			Lift		
	Dimecres	Dijous	Dissabte	Dimecres	Dijous	Dissabte	Dimecres	Dijous	Dissabte
{Passeig del Born}→{Catedral}	**36,60 %**	**39,00 %**	**46,80 %**	52,00 %	55,00 %	63,20 %	1.12	1.14	1.14
{Plaça del Mercat}→{Passeig del Born}	36,20 %	34,60 %	33,98 %	88,90 %	89,70 %	94,00 %	1.26	1.26	1.27
{Catedral, Plaça del Mercat}→{Passeig del Born}	13,70 %	14,00 %	17,00 %	**97,00 %**	**96,70 %**	**98,90 %**	**1.38**	1.36	1.33
{Plaça Major}→{Plaça de Cort}	NA	NA	12,20 %	NA	NA	51,10 %	NA	NA	**2.34**
{Plaça d'Espanya, Passeig del Born}→{Plaça del Mercat}	NA	10,60 %	10,60 %	NA	60,70 %	56,00 %	NA	**1.57**	1.55

Figura 20. Resultats de la setmana en el període de descans

Regla	Suport		Confiança		Lift	
	Dilluns	Dimarts	Dilluns	Dimarts	Dilluns	Dimarts
{Catedral}→{Passeig del Born}	**50,10 %**	45,70 %	89,50 %	87,90 %	1.13	2.14
{Plaça del Mercat}→{Passeig del Born}	45,60 %	**48,30 %**	93,90 %	92,60 %	1.19	1.2
{Llotja de Palma, Catedral, Plaça del Mercat}→{Passeig del Born}	12,20 %	11,90 %	**99,60 %**	**99,60 %**	1.27	1.29
{Catedral, Plaça de Sant Eullia}→{Plaça de Cort}	10,60 %	10,80 %	84,20 %	86,70 %	**2,87**	**3.09**

Figura 21. Resultats de la setmana en el període de feina

Regla	Suport				Confiança				Lift			
	Dimarts	Divendres	Dissabte	Diumenge	Dimarts	Divendres	Dissabte	Diumenge	Dimarts	Divendres	Dissabte	Diumenge
{Passeig del Born}→{Catedral}	43,50 %	**48,70 %**	**52,67 %**	**56,00 %**	56,1%%	61,20 %	66,20 %	67,40 %	1.12	1.09	1.13	1.09
{Passeig del Born}→{Plaça del Mercat}	**47,80 %**	44,70 %	44,70 %	49,40 %	61,5%%	56,30 %	56,30 %	59,50 %	1.2	1.19	1.19	1.15
{Catedral}→{Passeig del Born}	43,50 %	**48,70 %**	**52,67 %**	**56,00 %**	86,70 %	87,20 %	90,00 %	90,00 %	1.12	1.09	1.13	1.09
{Llotja de Palma, Catedral, Plaça del Mercat}→{Passeig del Born}	11,90 %	12,76 %	14,65 %	17,70 %	**99,20 %**	**99,67 %**	**99,39 %**	**99,60 %**	1.27	1.25	1.25	1.2
Plaça Major}→{Plaça de Cort}	11,60 %	11,80 %	16,70 %	16,50 %	40,90 %	45,24 %	48,47 %	55,30 %	**2.07**	**2.22**	1.84	2.15
{Plaça de Sant Eullia}→{Plaça de Cort}	NA	NA	11,30 %	10,70 %	NA	NA	81,50 %	83,66 %	NA	NA	**3.11**	**3.25**
{Plaça de Cort}→{Plaça de Sant Eullia}	NA	NA	11,30 %	10,70 %	NA	NA	43,12 %	42,05 %	NA	NA	**3.11**	**3.25**

Figura 22. Resultats de la setmana en el període d'entreteniment

ha estat que l'empresa ha modificat els noms i les ubicacions de les columnes. Com a resultat, les primeres execucions donaven pocs dispositius, més o menys 50 o 70. Finalment s'ha hagut de revisar l'algoritme inicial que recapta les dades per fer front a la nova estructura. Per possibles treballs futurs s'ha de revisar com estan la base de dades de cada institució i tractar-les per separat.

Finalment, en aquest estudi s'ha passat de dades en brut a coneixement a partir de tres algoritmes on el primer es tracta de classificar dispositius segons les connexions i les ubicacions, la segona a partir d'una suavització a les dades i aprenentatge automàtic conèixer quins són els modes de transports que la gent utilitza dins la ciutat i per últim a través de les regles d'associacions trobar quines han estat les rutes més habituals i les zones més visitades. Aquest coneixement obtingut dóna noves portes per obtenir la saviesa necessària en el camp de la mobilitat per oferir millors serveis i instal·lacions ciutadanes a dins la mateixa ciutat. Per així donar pas a què la ciutat de Palma es transformi en una Smart City exemplar per Espanya i la comunitat europea.

Referències

[1] IEC. Orchestrating infrastructure for sustainable Smart Cities ® White Paper. page 62, 2014.

[2] Junfeng Xie, Helen Tang, Tao Huang, F. Richard Yu, Renchao Xie, Jiang Liu, and Yunjie Liu. A Survey of Blockchain Technology Applied to Smart Cities: Research Issues and Challenges. *IEEE Communications Surveys and Tutorials*, 21(3):2794–2830, 2019.

[3] Kristin Archick. The European parliament. *Democratic Credentials of the European Union: Background and Analysis*, pages 23–49, 2014.

[4] Eric Hannon, Stefan Knupfer, Sebastian Stern, Ben Sumers, and Jan Tijs Nijssen. An integrated perspective on the future of mobility, Part 3: Setting the direction toward seamless mobility. *McKinsey Quarterly*, 2019(1), 2019.

[5] Luis Barreto, Antonio Amaral, and Sara Baltazar. Urban Mobility Digitalization: Towards Mobility as a Service (MaaS). *9th International Conference on Intelligent Systems 2018: Theory, Research and Innovation in Applications, IS 2018 - Proceedings*, pages 850–855, 2018.

[6] Tomeu Crespí Seguí. Palma de mallorca destino inteligente – free wifi 365 días. smart wifi palma. In *esmartcity.es*, November 2016.

[7] Lei Bai, Lina Yao, Salil S. Kanhere, Xianzhi Wang, and Zheng Yang. Automatic device classification from network traffic streams of internet of things. *arXiv*, pages 597–605, 2018.

[8] Xinglu Liu, Wan Wang, Wai Kin Victor Chan, Chiung Ying Kuan, and Junyoung Lee. User Classification in Electronic Devices Using Machine Learning Methods. *IEEE International Conference on Industrial Engineering and Engineering Management*, pages 1553–1556, 2019.

[9] Arash Kalatian and Bilal Farooq. Mobility mode detection using WiFi signals. *arXiv*, 2018.

[10] Kai Zhao, Mohan Prasath Chinnasamy, and Sasu Tarkoma. Automatic City Region Analysis for Urban Routing. *Proceedings - 15th IEEE International Conference on Data Mining Workshop, ICDMW 2015*, pages 1136–1142, 2016.

[11] Meraki. Location scanning api. *CISCO*. https://developer.cisco.com/meraki/scanning-api/!introduction/scanning-api.

[12] Shree Krishna Sharma and Xianbin Wang. Live Data Analytics with Collaborative Edge and Cloud Processing in Wireless IoT Networks. *IEEE Access*, 5:4621–4635, 2017.

[13] Article 29 Data Protection Working Party. Opinion 02/2013 on apps on smart devices. *October*, (February):1–11, 2003.

[14] anonimus. Harvesine formula. *wikipedia*, September 2020. Revisat el 30 de novembre de 2020.

[15] Jacquie Nesbitt. Suavización 1: Métodos distintos de suavizado. *numxl*, 2016. https://support.numxl.com/hc/es/articles/115000144963-Suavización-1-Métodos-Distintos-de-Suavizado.

[16] Asad Lesani and Luis Miranda-Moreno. Development and Testing of a Real-Time WiFi-Bluetooth System for Pedestrian Network Monitoring, Classification, and Data Extrapolation. *IEEE Transactions on Intelligent Transportation Systems*, 20(4):1484–1496, 2019.

[17] Na8. 7 pasos del machine learning para construir tu máquina. *aprendre machine learning*, 2017. https://www.aprendemachinelearning.com/7-pasos-machine-learning-construir-maquina/.

[18] Jon Starkweather and Amanda Kay Moske. Multinomial logistic regression. *Consulted page at September 10th: http://www. unt. edu/rss/class/Jon/Benchmarks/MLR_JDS_Aug2011. pdf*, 29:2825–2830, 2011.

[19] Ben Lutkevich. association rules. *Tech Target*, 2020. https://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining.
[20] Michael Hahsler. A probabilistic comparison of commonly used interest measures for association rules. 2015.
[21] Jason Brownlee. Why one-hot encode data in machine learning? *Machine Learning Mastery*, 2020. Revisat en decembre de 2020.
[22] The Python Standard Library. multiprocessing process-based parallelism. https://docs.python.org/3/library/multiprocessing.html.
[23] Pau Salas Cerda. Programari complet del treball. https://github.com/retolador/TFM, 2021.