Big Data: Mineria de dades amb el framework Apache Spark

Daniel Sànchez Gil

Resum– El tractament i la manipulació de dades és una part inherent de la societat moderna, una disciplina que es posa en pràctica en molts àmbits quotidians. En l'actualitat, però, el volum i la tipologia de les dades (l'anomenat *"Big Data"*) sol desbordar els mètodes de tractament i manipulació tradicionals. Cada cop més freqüentment resulta necessària la utilització d'eines de còmput distribuït i noves tècniques per afrontar l'allau de dades que origina una societat actual al segle XXI. Aquest projecte pretén estudiar la utilització d'un dels *frameworks* de còmput més utilitzats dins l'àmbit de *Big Data*, *Apache Spark*, en un cas d'ús real, com pot ser l'anàlisi del funcionament d'una xarxa de transport en una gran ciutat; en concret, el metro de Londres.

Paraules clau– Dades, *Big Data*, mineria de dades, aprenentatge computacional, còmput distribuit, patrons, xarxa, transport, metro, Londres.

Abstract– Data processing and manipulation is an inherent part of modern society, a field with many everyday applications. In recent times, however, the volume and variety of data (the so-called *"Big Data"*) is becoming a problem that traditional processing methodology is unable to cope with. The use of distributed computing and new data manipulation techniques are becoming increasingly necessary to deal with the information deluge created by 21st Century societies. This project aims to study the use of one of the better known distributed computing networks in the *Big Data* landscape, *Apache Spark*, in a real-world use case: the analysis of operation of a large, metropolitan transport network, such as the London Underground.

Keywords– Data, *Big Data*, data mining, machine-learning, distributed computing, patterns, network, transport, underground, London.

INTRODUCCIÓ

LES dades (i el coneixement que es pot extreure del seu tractament) són una de les mercaderies més valorades que podem trobar en el món del segle XXI, digital i interconnectat. Es tracta d'una mercaderia abundant: es calcula que per l'any 2020 es generaran 600 ZB (*Zettabytes*, 10^{12} *Gigabytes*) de dades, entre persones i màquines, una evolució notable respecte els 145 ZB generats l'any 2015[2]. El problema, però, és la qualitat extraordinàriament variada d'aquesta mercaderia; normalment, no es poden utilitzar les dades tal i com s'obtenen i sovint es requereix d'un procés de refinament.

El terme "*Big Data*" és una de les paraules més de moda en l'actualitat tecnològica, però el seu significat pot considerar-se un tant difós. La concepció popular és que "*Big Data*" implica gran quantitat de dades, tantes que es requereix de grans equips per processar i analitzar-les. Aquesta, però, és només una de les facetes del que realment representa aquest terme; per entendre'l millor, cal pensar en les quatre "V"[1]:

- **Volum:** la concepció popular del terme: grans quantitats de dades que impossibiliten el tractament de les mateixes mitjançant mètodes tradicionals.
- **Varietat:** el conjunt de dades és heterogeni, englobant diferents formats i fonts d'informació, fet que requereix que les dades passin per un procés previ de refinament.
- **Veracitat:** la qualitat i fiabilitat de les dades no és

• E-mail de contacte: daniel.sanchezgil@e-campus.uab.cat
• Menció realitzada: Tecnologies de la Informació
• Treball tutoritzat per: Jordi Casas Roma (Departament d'Enginyeria de la Informació i de les Comunicacions)
• Curs 2016/17

[1] Infografia d'IBM: `http://www.ibmbigdatahub.com/infographic/four-vs-big-data`

constant ni està assegurada, es requereix d'un procés de discriminació previ al processat.

- **Velocitat:** el conjunt de dades no pot ser gestionat a temps pels sistemes tradicionals per resultar d'utilitat (operacions en temps real o quasi-temps real).

Qualsevol conjunt de dades que compleixi un, o més, d'aquests indicadors pot ser considerat com "*Big Data*".

El procés que es realitza per transformar les dades (matèria prima) en coneixement (el producte desitjat) s'anomena Mineria de Dades (*Data Mining*, en anglès). Aquest procés comprèn tots els passos realitzats des que s'obtenen les dades fins a la representació del coneixement obtingut. Les fases i tècniques pròpies d'aquesta activitat formen el nucli d'aquest treball i es discutiran en detall més endavant.

Aquest article consistirà de la següent estructura. En primer lloc, s'exposaran els objectius i les motivacions del treball. Seguidament, es detallarà l'estat de l'art en l'àmbit proposat del treball (*"Big Data"* i Mineria de Dades) i la metodologia i la planificació seguida pel projecte. A continuació, s'explicarà quins conjunts de dades s'han utilitzat per la realització del treball i la seva obtenció, la infraestructura utilitzada per l'anàlisi de les dades i la metodologia d'aquest, seguits d'una exposició dels resultats obtinguts. Finalment, es descriuran les conclusions del treball.

OBJECTIUS I MOTIVACIÓ

Com ja s'ha esmentat en la introducció del document, el tractament i la manipulació de dades és una tasca indispensable en una societat informatitzada com la nostra. La Mineria de Dades és una disciplina amb aplicacions en molts àmbits del dia a dia, un fet que s'està accentuant notablement amb l'arribada de l'Internet de les Coses (*Internet of Things, IoT*) i l'increment de la integració de Tecnologies de la Informació en molts aspectes de les nostres vides. Una conseqüència d'aquesta tendència son les anomenades "***Smart Cities***", ciutats on s'estan implantant aquestes noves tecnologies per controlar i millorar el funcionament d'elements com xarxes de subministrament (aigua, electricitat, gas) i transport (gestió de carreteres, transport públic, etc), i un dels processos indispensables per garantir el funcionament d'aquests nous sistemes és el tractament i interpretació de les dades generades per aquestes noves tecnologies.

Un altre concepte emergent d'aquesta transformació i de les diferents iniciatives de foment de la transparència en l'àmbit públic és l'***"Open Data"***, l'alliberament de dades per part d'empreses i administracions públiques per ser utilitzades pels usuaris i ciutadans[4]. Aquest fenomen té com a conseqüència la incentivació de l'aparició d'aplicacions per satisfer les necessitats de la població, constituint una estratègia atractiva per les administracions o entitats que, per exemple, disposen de grans quantitats de dades però no de la capacitat d'explotar-les adequadament (ajuntaments, operadors de serveis locals).

Aquest treball pretén explotar les possibilitats que obre aquest nou ecosistema, realitzant un anàlisi del funcionament i la utilització d'una infraestructura clau, com és el Metro de Londres. Aquest anàlisi s'ha realitzat amb quatre objectius principals:

- Determinar les zones de més i de menys afluència a la xarxa de metro de Londres.
- Determinar les franges horàries de màxima i mínima utilització per les zones anteriors.
- Determinar quins patrons de moviment existeixen en la xarxa de metro i si varien al llarg del temps.
- Determinar quines parts de la infraestructura resulten més conflictives i determinar-ne les causes principals.

La finalitat de les dades recaptades en aquest treball i el coneixement descobert en el tractament de les mateixes serà proporcionar una visió general del funcionament del sistema de transport de Londres i proposar, si s'escau, canvis per adaptar el sistema a les tendències observades.

ESTAT DE L'ART

En l'àmbit de *"Big Data"* és habitual la utilització de sistemes distribuïts, degut a la necessitat de més prestacions que les que poden oferir els sistemes unitaris. La quantitat d'empreses[2] i productes[3] dedicats a cobrir les necessitats d'aquest ecosistema resulta inabastable; existeixen, però, productes àmpliament utilitzats i que mereixen una especial menció en l'àmbit en que s'orienta aquest treball, com *Apache Hadoop* i *Apache Spark*.

Apache Hadoop[4] és un *framework* de processament distribuït de grans conjunts de dades. Consta de dos grans parts: *Hadoop MapReduce*, un framework de processat paral·lel; i *Hadoop Distributed File System* (*HDFS*, "Sistema de Fitxers Distribuit Hadoop"). En recents versions, s'ha afegit un *framework* de planificació de treballs i gestió de recursos (*Hadoop YARN*). Hadoop és un dels programaris més utilitzats en l'àmbit de sistemes distribuïts, incloent moltes de les grans empreses del sector de Tecnologies de la Informació[5].

Apache Spark[6] és un framework de còmput distribuït, ràpid i de propòsit general. És capaç de ser executat per sobre de diversos sistemes distribuïts (Apache Hadoop, Apache Mesos), en el núvol (Amazon EC2), i en solitari (*standalone*), adaptant-se a pràcticament qualsevol entorn d'ús. A més, disposa de llibreries d'aprenentatge computacional (*MLLib*), operacions SQL (*Spark SQL*), tractament de gràfics (*GraphX*) i tractament de fluxos de dades (Spark Streaming); aquestes llibreries es troben disponibles, en la seva majoria, per Java, Scala i Python. Recentment, s'ha començat a afegir suport per R, un llenguatge àmpliament utilitzat en estadística i anàlisi de dades. A més, al igual que Hadoop, Spark és un programari amb un ampli grau d'utilització en les grans empreses del sector[7].

[2]Llistat d'empreses de l'entorn de *"Big Data"*: `http://dfkoz.com/big-data-landscape/`

[3]Escenari de *"Big Data"*, 2016: `http://mattturck.com/wp-content/uploads/2016/03/Big-Data-Landscape-2016-v18-FINAL.png`

[4]Apache Hadoop: `http://hadoop.apache.org`

[5]Powered by Apache Hadoop: `http://wiki.apache.org/hadoop/PoweredBy`

[6]Apache Spark: `http://spark.apache.org`

[7]Powered by Apache Spark: `http://spark.apache.org/powered-by.html`

Altres tecnologies utilitzades en l'emmagatzemat de grans quantitats de dades són *Apache Hive*[8] (programari magatzem de dades SQL distribuït), *Apache Cassandra*[9] (base de dades distribuïda amb èmfasi en l'escalabilitat i la disponibilitat) i *MongoDB* (base de dades NoSQL, orientada a aplicacions web).

En el camp de la Mineria de Dades, un altre *software* que mereix ser esmentat és *Weka*[10], dissenyat per la Universitat de Waikato. *Weka* és un *framework* de *Data Mining* implementat en Java, amb especial èmfasi en la utilització d'algoritmes d'aprenentatge computacional. En les darreres versions, disposa també de *wrappers* per *Hadoop* i *Spark*.

METODOLOGIA DE TREBALL I PLANIFICACIÓ

El procés de Mineria de Dades és una activitat metòdica, que segueix, generalment, la següent estructura[7]:

1. **Recol·lecció**: s'obtenen dades d'una quantitat i varietat adequada pel propòsit de l'anàlisi.
2. **Neteja**: les dades són tractades per tal d'eliminar "soroll" i inconsistències. El procés de discriminació de dades no fiables té lloc en aquesta fase.
3. **Integració**: es combinen dades de diferents fonts i formats, si s'escau, en un sistema comú.
4. **Selecció**: es fa una tria de les dades pertinents a l'anàlisi a realitzar.
5. **Transformació**: s'apliquen transformacions sobre el conjunt de dades seleccionats, obtenint dades en formes que faciliten operacions d'agregació o resum.
6. **Mineria o processat**: s'apliquen mètodes intel·ligents per extreure patrons i tendències del conjunt de dades (aprenentatge computacional i d'altres tècniques).
7. **Avaluació de patrons**: s'identifiquen els patrons que representen coneixement valuós per l'objectiu de l'anàlisi mitjançant l'aplicació d'heurístiques i procediments estadístics.
8. **Presentació de coneixement**: s'utilitzen tècniques de representació i visualització per presentar el coneixement resultant de la mineria als usuaris finals.

Tot i que es presenta aquest procés com una seqüència lineal, de vegades és necessari, segons el tipus i origen de les dades, o la finalitat del processat, que certes fases (com la recol·lecció i pre-processat de les dades) es realitzin iterativament al llarg del procés. Un exemple d'aquesta situació és el tractament de dades recollides en temps real, com ha estat el cas d'aquest treball.

Per la realització d'aquest treball, el procediment descrit al principi de la secció es va dividir, a priori, en tres parts, corresponent a les fites d'informe de progrés. Així doncs, la primera part del procés, que comprèn la recol·lecció de les dades i el seu tractament previ, corresponia al primer informe de progrés; la transformació i mineria, al segon informe de progrés; i l'avaluació de patrons i presentació de coneixement, a l'informe final. Aquesta planificació, realitzada prèviament a l'inici del treball, va resultar excessivament optimista, subestimant la complexitat de les fases inicials i obligant a una divisió en dues parts (recol·lecció i preprocessat, en la primera part; mineria i avaluació de patrons en la segona), i posposant la realització efectiva de la presentació del coneixement en una etapa posterior a l'elaboració i entrega de l'informe final. Un altre punt no contemplat inicialment en la planificació del treball va ser el muntatge d'un entorn de treball distribuït per la realització del treball, en la forma d'un clúster de tres màquines domèstiques, detallat més endavant en aquest article.

Per avaluar el seguiment de la planificació, es van realitzar reunions periòdiques entre el tutor i l'estudiant, coincidint amb cada fita d'entrega d'informe. Aquestes reunions tenien com objectiu analitzar el progrés realitzat, ajustar la planificació i marcar els objectius a assolir en la següent fita.

DADES

La gran majoria de dades recol·lectades per aquest treball provenen de la iniciativa d'*Open Data* de l'operador del transport metropolità de Londres (*Transport for London*, o *TfL*[11]). *TfL* ofereix una API REST pública per usuaris registrats que proporciona accés a informació tant de les instal·lacions com de les arribades dels vehicles a les diferents estacions d'autobús, tren i metro, entre d'altres. Altres conjunts de dades oferts per *TfL* són conjunts estadístics recopilats regularment, com el registre d'entrades i sortides anuals de les estacions de metro o enquestes sobre els viatges que realitzen els passatgers (*RODS*, *Rolling Origin & Destination Survey*).

Altres dades de rellevància recaptades per l'ús d'aquest treball han estat les localitzacions dels diferents edificis històrics registrats a la ciutat de Londres (anomenats *Listed Buildings*[12]), i informació meteorològica local (obtinguda mitjançant l'API de *Dark Sky*[13]).

Les dades utilitzades en el treball poden classificar-se en tres categories funcionals:

- **Dades d'entorn**: representen "l'escenari" que serà objecte d'anàlisi, és a dir, l'entorn d'estudi. En aquesta categoria incloem l'estructura del Metro de Londres (informació geo-localitzada de les estacions i les línies que conformen la xarxa de transport), així com la informació dels vehicles que hi circulen (els diferents models de la flota de trens del Metro, la seva capacitat i on operen[14]).
- **Dades de mesura**: representen el gruix del conjunt de dades utilitzades en l'anàlisi i proporcionen una visió objectiva de l'estat de la infraestructura en el temps. Comprenen, principalment, la informació d'estat de

[8] Apache Hive: `http://hive.apache.org`
[9] Apache Cassandra: `http://cassandra.apache.org`
[10] Weka: `http://www.cs.waikato.ac.nz/ml/weka`
[11] Transport For London Open Data: `http://tfl.gov.uk/info-for/open-data-users/`
[12] Listed Buildings: `http://historicengland.org.uk/listing/what-is-designation/listed-buildings`
[13] Dark Sky API: `http://darksky.net/dev`
[14] London Underground Rolling Stock: `http://tfl.gov.uk/corporate/about-tfl/what-we-do/london-underground/rolling-stock`

les diferents línies de la xarxa i la localització aproximada dels trens en un instant de temps determinat (informació d'arribada).

- **Dades de context**: són aquelles dades que ens aporten informació addicional sobre el context de l'anàlisi, complementant la informació d'entorn i de mesura. En aquesta categoria podem incloure les dades de posicionament d'edificis d'interès històric, social o econòmic (que aporten valor a les estacions properes), i dades sobre esdeveniments en el temps (calendaris d'esdeveniments esportius o socials).

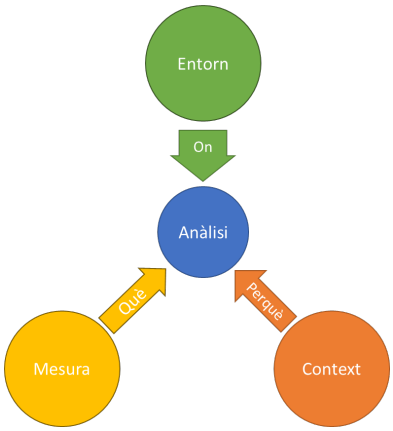

Fig. 1: Tipus de dades

INFRAESTRUCTURA D'ANÀLISI

En entorns d'indústria, els equipaments utilitzats per dur a terme anàlisi i mineria de grans conjunts de dades solen comprendre grans sistemes distribuïts, amb centenars de nuclis de còmput i gran quantitat de memòria, tant en forma física (*in-situ* a l'empresa o institució corresponent) com en la xarxa (*Cloud Computing*). En tots els casos, l'estructura lògica del sistema de mineria sol estar composta per dos elements principals: un sistema d'emmagatzemat massiu comú on es guarden les dades (segons la mida del mateix, i l'origen i la naturalesa de les dades emmagatzemades, pot adoptar diferents noms: *Data Lake*[3], *Data Warehouse*[8], *Data Mart*[15]) i un sistema de còmput i processat que treballa sobre les dades de la part anterior.

Donat que un entorn d'anàlisi de nivell d'indústria no és, evidentment, assolible per la realització d'aquest treball, es va prendre la decisió de construir-ne un a escala reduïda, utilitzant equipament domèstic. Per aquest objectiu es va aprofitar el fet de que moltes tecnologies utilitzades en el tractament de *Big Data* són escalables, tant per grans entorns de computació com per petits clústers de desenvolupament (com l'utilitzat en aquest treball).

El *clúster* utilitzat en aquest treball consta de tres màquines: un node de gestió i control (*master*), i dos nodes de càlcul i emmagatzemat (*slave-1* i *slave-2*). Les especificacions tècniques d'aquestes màquines poden ser consultades en l'apèndix inclòs en el present article.

[15] `http://en.wikipedia.org/wiki/Data_mart`

Una quarta màquina utilitzada en el treball (no inclosa en el *clúster*) és una màquina virtual d'*Amazon EC2*[16] utilitzada per recaptar dades en temps real del funcionament de la xarxa del Metro de Londres. Inicialment, aquesta funció era realitzada per una *Raspberry Pi 2 B*, però la necessitat de mantenir un funcionament constant de manera viable va suposar la migració a EC2. Aquesta màquina va recaptar la informació d'estat i d'arribades de totes les línies del Metro de Londres cada 5 minuts, de manera quasi ininterrompuda, des de la nit del 24 d'octubre de 2016 fins el matí del 19 de gener de 2017.

El programari utilitzat en el *clúster* d'aquest treball ha estat *Apache Hadoop* (utilitzant HDFS com sistema d'emmagatzemat distribuït per tot el sistema) i *Apache Spark* (configurat per utilitzar IPython[10] com intèrpret interactiu).

Inicialment, abans del muntatge del sistema distribuït, es va utilitzar *MongoDB* com repositori provisional de dades i per avaluar inicialment com es podrien organitzar i tractar aquestes. Posteriorment, totes les dades es van traslladar a HDFS. Es va considerar també la necessitat de desplegar una instal·lació dedicada d'*Apache Hive*, per gestionar el magatzem de dades, però Spark ja inclou una implementació bàsica d'aquest sistema i l'emmagatzemat de metadades (*Apache Derby*), amb la qual cosa no va resultar necessària la instal·lació de programari addicional.

ANÀLISI: PROCEDIMENTS I TÈCNIQUES

Anàlisi preliminar de dades

Per determinar la naturalesa dels conjunts de dades i preparar el camí de l'anàlisi posterior, se solen aplicar funcions estadístiques i de classificació sobre la informació recollida durant les fases prèvies d'un treball de mineria de dades. La idea en aquesta fase és adquirir pistes sobre com encarar posteriors anàlisis, més concrets i exhaustius.

Un dels conjunts de dades més importants recollit al llarg del treball ha estat un **registre de les estimacions d'arribades** de trens a cada una de les estacions del Metro de Londres. A diferència d'altres mitjans de transport, en un sistema ferroviari no se sol disposar de les localitzacions exactes dels trens a la línia, que opera mitjançant un sistema de blocs (trams reservats que només poden ser ocupats per un únic vehicle). Aquest conjunt de dades recull l'abstracció que *TfL* realitza d'aquestes posicions aproximades, obtenint un compte enrere estimat de l'arribada del tren (en forma d'un enter indicant els segons fins l'arribada estimada) a l'andana i la posició general del tren (en forma d'una breu cadena de text).

Donat que els identificadors dels vehicles registrats en el conjunt de dades són només únics en una línia específica, resulta difícil deduir la quantitat de trens que circulen pel sistema. El que sí que es pot mesurar objectivament, però, és el nombre d'identificadors independents registrats en cada una de les línies, que proporciona un indicatiu del volum de tràfic (o, més acuradament, del nombre de trens diferents que roten) en les mateixes:

[16] Amazon Elastic Compute Cloud: `http://aws.amazon.com/ec2`

TAULA 1: IDENTIFICADORS REGISTRAT PER LÍNIA

Nom de línia	Identificadors registrats
Circle	70
District	148
Victoria	112
Central	177
Waterloo & City	33
Jubilee	90
Hammersmith & City	150
Bakerloo	60
Northern	198
Piccadilly	267
Metropolitan	120

Un anàlisi de la variació i el nombre d'aparicions d'un mateix identificador en una línia pot servir, a més, per identificar el patró de rotació de cada vehicle (si s'utilitza regularment o, per exemple, és un vehicle de reserva enviat per suplir una manca puntual de trens).

Un altre conjunt de dades important és la informació sobre l'**estat de la xarxa**. Aquest conjunt ens ofereix una descripció categòrica de l'estat de cada línia de la xarxa (un resum, com "Bon Servei" o "Retards menors") i, si s'escau, una explicació més detallada en format de text. Algunes de les descripcions categòriques estan associades al funcionament quotidià del metro ("Servei Tancat", per exemple, correspon a la finalització de l'horari de servei), mentre que d'altres ("Servei Suspès", "Retards Greus") indiquen una situació anòmala a la línia.

De nou, aplicant una funció de recompte, podem veure una indicació de la freqüència observada d'aquests estats (aparicions en el conjunt de dades):

TAULA 2: ESTATS DE LA XARXA PER TIPUS

Tipus d'estat	Nombre d'ocurrències
Servei Suspès	581
Servei Suspès Parcialment	2420
Servei Tancat	27416
Tancament Parcial	10163
Tancament Planejat	4160
Retards Menors	6686
Retards Greus	8971
Bon Servei	144083
Servei Especial	1264

També en la fase preliminar de l'anàlisi, es van associar les posicions dels edificis històrics de la ciutat amb les de les estacions de metro de la xarxa. L'objectiu d'aquest anàlisi era **determinar possibles estacions interessants** per la seva proximitat a localitzacions d'interès turístic i cultural. Aquesta associació va ser possible aplicant l'algoritme *KMeans*[9], un algoritme d'aprenentatge computacional de classificació no supervisada.

Aquest anàlisi va identificar un total de 14 estacions properes a localitzacions d'interès històric:

TAULA 3: EDIFICIS HISTÒRICS PER ESTACIÓ

Nom d'estació	Nombre d'edificis
Aldgate	7
Bank	19
Barbican	10
Blackfriars	18
Cannon Street	11
Chancery Lane	11
Farringdon	2
Liverpool Street	8
Mansion House	9
Monument	16
Moorgate	4
St. Paul's	22
Temple	15
Tower Hill	10

Anàlisi definitiu

Un cop efectuats l'anàlisi preliminar i una revisió dels objectius plantejats, es va procedir a l'anàlisi definitiu dels conjunts de dades del treball.

Una de les primeres decisions preses en aquesta fase va ser la d'elaborar un **graf de connexions** de la xarxa de metro, utilitzant les dades geo-localitzades de les estacions i la informació de seccions i sentits de línia. Apache Spark incorpora una llibreria de computació paral·lela orientada a grafs, *GraphX*[17], però aquesta no té implementació en Python, el llenguatge principal utilitzat en aquest treball, amb la qual cosa es va optar per una llibreria alternativa, *GraphFrames*[18], amb funcionalitat similar.

Poder utilitzar un graf per representar la xarxa de metro va suposar poder avaluar fàcilment la connectivitat entre diferents estacions individuals i obtenir una mesura de la importància de cada estació segons les seves connexions, aplicant l'algoritme **PageRank**[1], desenvolupat originalment per Google per indexar pàgines web en funció de les seves relacions amb altres pàgines web (referents i referits). El resultat d'aquest algoritme és un escalar arbitrari que denota la importància relativa del node avaluat. Donat que PageRank és un algoritme agnòstic en quant a context, és a dir, la seva aplicació és la mateixa per qualsevol tipus de graf, va resultar una eina convenient, juntament amb els graus de referents i referits, per obtenir una visió global sobre la importància de les estacions dins de la xarxa de metro a nivell de connectivitat.

Aquestes mesures, juntament amb l'associació d'estacions amb edificis històrics esmentada a la secció anterior i el *borough* de l'estació, es van combinar en un únic conjunt, del qual es va elaborar una **regressió per arbre de decisió**[6] per avaluar la importància de possibles factors que poguessin influir en el nombre d'usuaris anuals de l'estació. L'arbre resultant d'aquest anàlisi es pot consultar en l'apèndix d'aquest article.

Determinar les zones de més i menys afluència de la xarxa va resultar una tasca més simple, gràcies als registres d'entrada i sortida de passatgers recollits pel sistema de màquines d'accés de la xarxa. Aquests conjunts de dades contenen estadístics processats regularment per l'operador,

[17] GraphX: `http://spark.apache.org/graphx/`
[18] GraphFrames: `http://graphframes.github.io`

mostrant l'agregat anual d'entrades i sortides a cada estació dividit en dies laborables, dissabtes i diumenges, i un recompte anual de passatgers (en milions). Mitjançant consultes contra la base de dades, es pot obtenir un llistat d'estacions, ordenades en funció del nombre de passatgers registrats, de major a menor, o viceversa.

Un altre conjunt de dades (de novembre de 2015) ofereix una resolució més detallada, oferint mitjanes d'entrades i sortides per estació (per separat), segons tipus de dia (laborable, dissabte o diumenge) i en franges de 15 minuts. D'aquesta manera es poden deduir les franges de màxima i mínima afluència per cada estació.

Un altre conjunt de dades analitzat en aquesta fase és el ***RODS*** (*Rolling Origin Destination Survey*). Aquest és un conjunt estadístic detallat sobre la tipologia dels viatges realitzats en la xarxa de metro durant tot l'any (en el moment de l'elaboració d'aquest article, el conjunt més recent disponible corresponia a l'any 2015). Entre la informació continguda al *RODS* podem destacar el desglossament de viatges per motivació (de casa al treball, del treball a casa, d'oci, *commuting*, etc) i estació, per tipus de transport previ (cotxe estacionat, cotxe no estacionat, autobús, bicicleta, etc) , i una matriu d'origen-destí de viatges, també dividida en franges horàries. Gràcies a aquest conjunt de dades s'obté, no només la informació per explicar les franges de més i menys ocupació a la xarxa, sinó que permet deduir patrons de moviment generals a la ciutat.

RESULTATS

Zones de més i menys afluència

Mitjançant una consulta sobre la base de dades del treball, ordenant les estacions segons el nombre de passatgers, podem obtenir la següent taula:

TAULA 4: ESTACIONS MÉS TRANSITADES (2015)

Estació	Recompte anual (M)
Waterloo	95,13837
King's Cross St. Pancras	93,41302
Oxford Circus	92,35562
Victoria	82,88556
Liverpool Street	73,25732
London Bridge	71,96443
Stratford	61,44263
Monument	57,51297
Bank	57,51297
Canary Wharf	54,44142
Paddington	49,63730
Leicester Square	43,74595
Piccadilly Circus	42,79683
Euston Square	42,15933
Holborn	40,53020

Les estacions esmentades en aquesta taula corresponen a importants nodes de transport (Waterloo, King's Cross/St. Pancras, Victoria, Stratford, Paddington, London Bridge, Liverpool Street i Euston Square són importants enllaços ferroviaris) o importants centres econòmics (Canary Wharf, Bank i Monument[19]). Oxford Circus és molt probablement la cruïlla més transitada de tot Londres[5], mentre que Piccadilly Circus és, sens dubte, una de les localitzacions més conegudes de la ciutat. Leicester Square és un destacat centre cultural, agrupant un nombre de cinemes de renom nacional i on se solen realitzar un nombre elevat d'estrenes. Holborn és una estació que serveix a dues línies, prop del nucli històric de la ciutat i localitzacions com el *British Museum* i *Lincoln's Inn Fields* (un dels espais verds més grans de la ciutat).

Aplicant la mateixa consulta, amb l'ordenació inversa, obtenim la següent taula:

TAULA 5: ESTACIONS MENYS TRANSITADES (2015)

Estació	Recompte anual (M)
Tufnell Park	0,00000
Roding Valley	0,26065
Chigwell	0,55906
Grange Hill	0,65766
Theydon Bois	0,85470
Chesham	0,87526
Moor Park	0,88639
North Ealing	0,89259
South Kenton	0,95720
Croxley	1,05941
Ruislip Manor	1,11262
Fairlop	1,13427
Chorleywood	1,13768
Upminster	1,14677
Ickenham	1,19036
Mill Hill East	1,31816

L'estació de Tufnell Park va romandre tancada durant els anys 2015 i 2016 per feines de manteniment, per tant aquella estació no va servir cap viatger en el període que comprèn el conjunt analitzat. La gran majoria d'aquestes estacions pertanyen als anells exteriors de la xarxa del metro, algunes d'elles localitzant-se a la perifèria o, fins i tot, completament fora del comtat de Londres. Aquestes són zones suburbanes, amb poca relativa poca densitat de població i, per tant, d'usuaris.

Franges de màxima i mínima utilització

Combinant la nostra informació sobre les estacions més transitades de la xarxa i el registre d'entrades i sortides de novembre del 2015, podem obtenir les franges de màxima i mínima utilització de les estacions amb una resolució de 15 minuts. Donat que aquestes dades són la mitjana del mes de novembre del 2015, els valors actuals poden variar. Tot i això, la tendència observada hauria de mantenir-se en el temps. A continuació, es mostren les franges de més entrades per un dia laborable a les sis estacions més transitades de la xarxa:

TAULA 6: FRANJA DE MÀXIMES ENTRADES (NOVEMBRE 2015, LABORABLE)

[19] Bank i Monument són dues estacions que serveixen a diferents línies però ocupen aproximadament la mateixa localització general. Les estadístiques les resumeixen en una única estació i, per tant, les dades de recompte de viatgers són les mateixes per les dues estacions.

Estació	Franja de màxima afluència
Waterloo	07:00 - 10:00
King's Cross St. Pancras	07:00 - 10:00
Oxford Circus	16:00 - 20:45
Victoria	07:30 - 10:00
Liverpool Street	16:00 - 20:45
London Bridge	16:00 - 20:45

A la taula anterior es pot apreciar l'existència de dues franges clarament marcades, matí i tarda. Waterloo, King's Cross/St. Pancras i Victoria, reben el seu major nombre d'entrades durant les hores del matí, coincidint amb l'inici de la jornada laboral; en canvi, Oxford Circus, Liverpool Street i London Bridge tenen el seu període de màxima afluència de viatgers entrants durant les hores de la tarda i nit, coincidint amb el final de l'horari laboral. D'aquestes dades es pot inferir un patró d'estacions emissores-receptores regular en el dia a dia.

Podem invertir la situació, observant les xifres de sortides, i observar que, efectivament, els rols s'inverteixen:

TAULA 7: FRANJA DE MÀXIMES SORTIDES (NOVEMBRE 2015, LABORABLE)

Estació	Franja de màxima afluència
Waterloo	16:15 - 19:45
King's Cross St. Pancras	07:30 - 10:00
Oxford Circus	07:30 - 10:00
Victoria	07:30 - 10:00
Liverpool Street	07:30 - 10:00
London Bridge	07:30 - 10:00

Aquest patró canvia completament en arribar el cap de setmana. Vegeu a les següents taules, les entrades i sortides en dissabte:

TAULA 8: FRANJA DE MÀXIMES ENTRADES (NOVEMBRE 2015, DISSABTE)

Estació	Franja de màxima afluència
Waterloo	10:00 - 14:30
King's Cross St. Pancras	09:15 - 19:30
Oxford Circus	13:15 - 21:45
Victoria	10:15 - 20:15
Liverpool Street	10:15 - 19:00
London Bridge	15:15 - 20:45

TAULA 9: FRANJA DE MÀXIMES SORTIDES (NOVEMBRE 2015, DISSABTE)

Estació	Franja de màxima afluència
Waterloo	15:15 - 20:00
King's Cross St. Pancras	16:00 - 20:00
Oxford Circus	09:00 - 20:00
Victoria	16:00 - 19:30
Liverpool Street	16:30 - 20:45
London Bridge	11:00 - 14:45

Com es pot apreciar, durant els dissabtes desapareixen els períodes en els que aquestes estacions esdevenen clarament "emissores" o "receptores", fenomen atribuïble al menor nombre de treballadors anant i tornant de la feina i l'augment d'usuaris que es desplacen per oci. Oxford Circus, per exemple, manté una afluència considerable, tant d'entrada com de sortida, durant gran part del dia, que pot explicar-se amb la seva localització cèntrica, en la zona comercial més important de la ciutat (Oxford Street, a data de 2012, era el carrer comercial més transitat d'Europa, amb mig milió de visitants diaris).

A continuació es mostren les gràfiques base de les que s'han extret aquests resultats:

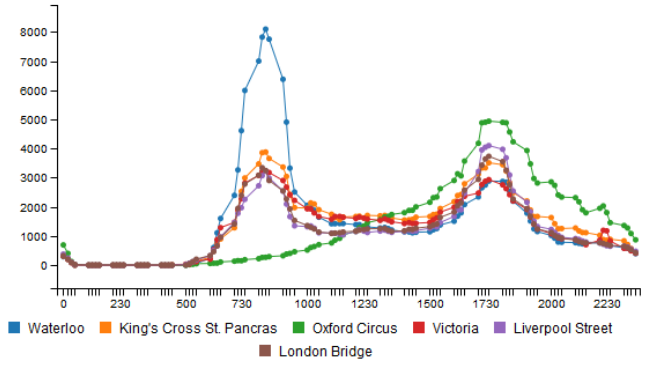

Fig. 2: Usuaris entrants en dia laborable (Novembre 2015)

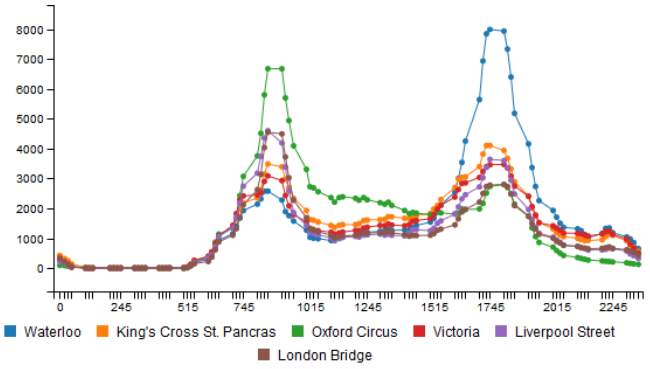

Fig. 3: Usuaris sortint en dia laborable (Novembre 2015)

A la figura 2 es poden apreciar clarament els dos pics d'afluència d'entrades, corresponents a la entrada i sortida de treballadors. És destacable el nivell que assoleix Waterloo al matí, considerablement superior a les altres estacions (més del doble que Oxford Circus). Per contra, la figura 3 mostra nivells d'afluència similars en quant a sortides (menys diferència entre Waterloo i Oxford Circus, les dues estacions amb més sortides de la gràfica).

Aquesta particularitat resulta rellevant, ja que implica que Waterloo rep, proporcionalment, més usuaris que no les altres estacions de la gràfica; és a dir, Waterloo té una utilització destacada com punt d'entrada al sistema que representa la xarxa de metro. Per contra, si bé Waterloo segueix dominant la gràfica de sortides, la importància de la resta d'estacions com punts de sortida de la xarxa és més igualat (relativament parlant, segons la quantitat d'usuaris de l'estació).

Aquesta tendència, com ja s'ha esmentat, no es manté fora de la setmana laboral. A les figures 4 i 5 es pot comprovar com, tant en entrades com en sortides, Oxford Circus

manté una posició predominant a les gràfiques, seguida de King's Cross. Waterloo, per contra, ocupa un modest tercer lloc, molt pròxim a les corbes de la resta d'estacions de la gràfica.

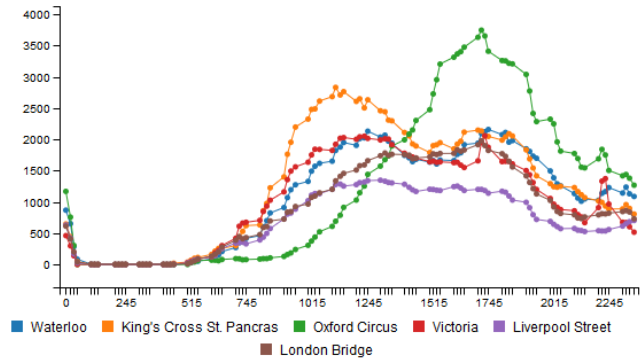

Fig. 4: Usuaris entrant en dissabte (Novembre 2015)

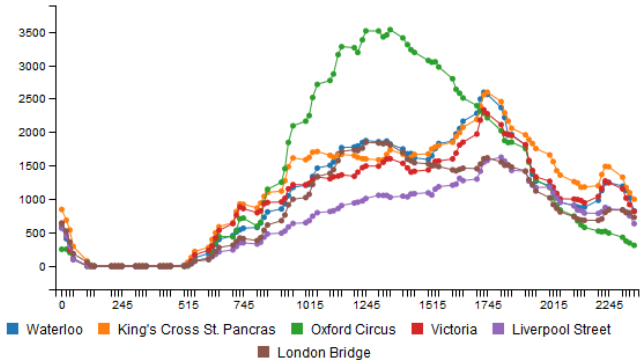

Fig. 5: Usuaris sortint en dissabte (Novembre 2015)

Patrons de moviment en el temps

La idea inicial d'aquest treball era determinar patrons de moviment a nivell diari, relacionant-los amb informació temporal d'esdeveniments i altres factors de possible rellevància (temps atmosfèric, per exemple). Degut a que les dades amb les que s'ha pogut comptar per aquest treball són resums anuals o, puntualment, mensuals, aquest nivell de resolució no ha estat assolible. Tot i això, com ja s'ha esmentat en la secció anterior, les dades tractades permeten determinar patrons generals. Combinant les dades del *RODS* amb la informació ja obtinguda de l'anàlisi temporal i quantitatiu de l'entrada i la sortida d'usuaris a les estacions del metro, es pot aportar una dimensió addicional a l'anàlisi, afegint informació sobre la motivació del viatge i, l'origen i/o destinació del viatger.

Per raons d'espai, aquesta secció es centrarà en els resultats d'una única estació, Bank, situada al costat de l'edifici del Banc d'Anglaterra, a la City de Londres.

En la secció anterior, s'ha esmentat que les estacions de la xarxa, en dies laborables, poden considerar-se "emissores" o "receptores" segons aporten o reben viatgers de la xarxa en un moment determinat. A les dues taules següents es poden apreciar les estacions que més reben i envien viatgers a Bank:

TAULA 10: ESTACIONS RECEPTORES DE BANK

Estació destí	Franja pic	Total
Waterloo	16:00 - 19:00	25862
King's Cross St. Pancras	16:00 - 19:00	4626
Holborn	07:00 - 10:00	4369
Tottenham Court Road	10:00 - 16:00	4264
Oxford Circus	10:00 - 16:00	3755

TAULA 11: ESTACIONS EMISSORES A BANK

Estació origen	Franja pic	Total
Waterloo	07:00 - 10:00	21760
Moorgate	07:00 - 10:00	4256
Tottenham Court Road	10:00 - 16:00	4061
Liverpool Street	19:00 - 22:00	3633
King's Cross St. Pancras	07:00 - 10:00	3481

El principal emissor i receptor de Bank és Waterloo, un important enllaç ferroviari que serveix al sud-oest d'Anglaterra i, fins el 2007, l'estació termini de l'Eurostar. Aquest patró és consistent amb el moviment de treballadors de fora de Londres (*commuters*) cap a i des de les oficines de la City.

RODS ens ofereix informació més detallada sobre el propòsit del viatge de l'usuari. Analitzant de nou els usuaris que surten i entren a Bank segons la tipologia del seu viatge, obtenim les següents taules:

TAULA 12: TIPOLOGIA DE VIATGE (BANK, *Commuters* AL LLOC DE TREBALL)

Franja horària	Total
07:00 - 10:00	48036
10:00 - 14:00	6308
14:00 - 19:00	952
19:00 - 21:00	328
21:00 - 24:00	116

TAULA 13: TIPOLOGIA DE VIATGE (BANK, *Commuters* CAP A CASA)

Franja horària	Total
07:00 - 10:00	842
10:00 - 14:00	4150
14:00 - 19:00	28938
19:00 - 21:00	10809
21:00 - 24:00	2161

Com es pot comprovar, les xifres anteriors corroboren que el principal segment d'usuaris de Bank són *commuters*. Resulta interessant, però, que el nombre de retorns estigui més difós en el dia que el nombre d'arribades, probablement denotant diferents tipus de jornada entre els treballadors.

Parts conflictives de la xarxa

Cap infraestructura de transport és aliena a les incidències. El Metro de Londres, sent com és una xarxa ferroviària, és especialment susceptible a les incidències puntuals i localitzades, que poden provocar retards al llarg d'una línia, propagar-se a d'altres (per acumulació de passatgers) i, en

el pitjor dels casos, provocar suspensions del servei. Per avaluar el grau d'incidències en la xarxa en aquest treball, s'ha utilitzat la informació d'estat de les línies, recaptada a quasi-temps real per un *script* en una màquina virtual al núvol.

Piccadilly, una de les línies més llargues i més transitades de la xarxa, serà l'objecte d'anàlisi d'aquesta secció, començant amb una valoració a alt nivell del funcionament de la línia.

Fixant-se en la figura 6, es pot comprovar que la línia Piccadilly és una part de la xarxa especialment conflictiva, amb gairebé un 47% del temps observat amb retards menors (11,5%) i retards greus (35,4%). Per analitzar les causes de les incidències cal realitzar un anàlisi dels missatges d'explicació de la xarxa.

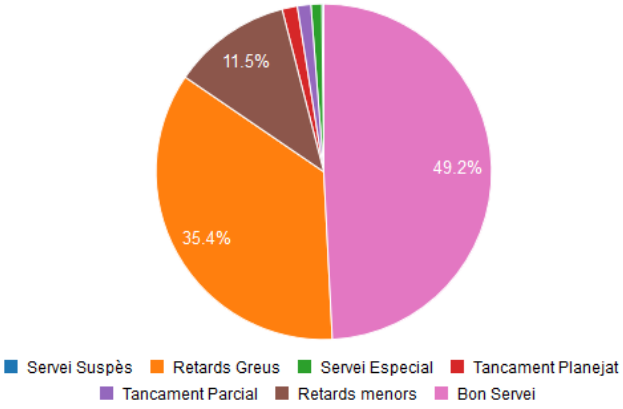

Fig. 6: Estat general de la línia Piccadilly

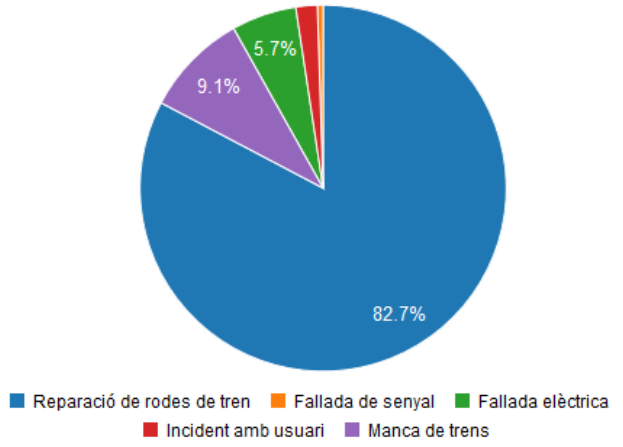

Fig. 7: Raons de retard de la línia Piccadilly

En la figura 7 es poden observar les causes registrades de les incidències. Durant un període d'unes setmanes, la línia Piccadilly va sofrir una manca greu, però puntual, de trens degut a reparacions en la flota del metro, cosa que esbiaixa el resultat. Si es deixa de banda la causa de les reparacions de rodes (figura 8), un percentatge elevat segueix estant relacionat amb una manca general de trens, ja sigui per redistribució de vehicles o per absència de personal. La tercera causa rellevant són incidències amb el subministrament elèctric, seguida d'incidents amb usuaris. La darrera categoria correspon a fallades de senyalització en el recor-

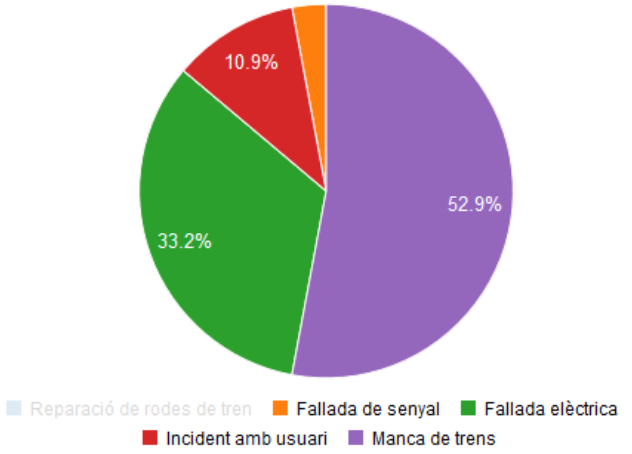

Fig. 8: Raons de retard de la línia Piccadilly (sense reparació de rodes

regut. Aquest tipus de fallades són especialment greus, ja que obliguen a aturar els trens en el segment afectat o suspendre parcialment el servei, donat que no es pot circular amb seguretat.

CONCLUSIONS

La Mineria de Dades és una activitat cada cop més important en el sector de les Tecnologies de la Informació, especialment amb la integració creixent de sistemes informàtics en la nostra vida quotidiana. En aquest treball s'ha intentat posar aquesta activitat en pràctica, en el marc d'un cas d'utilització real, com és l'anàlisi del funcionament d'una infraestructura de transport metropolità, utilitzant eines de primer ordre en el sector.

Gràcies a la informació disponible com a resultat de les polítiques d'*"Open Data"*, en aquest treball s'ha pogut dur a terme una examinació d'alt nivell de la xarxa del Metro de Londres, analitzant l'estat de les línies, incidències en la xarxa, la utilització general del sistema, i els patrons de moviment dels seus usuaris. Tot i això, alguns objectius no s'han pogut assolir tal i com es plantejaven en un principi degut a una insuficient resolució de les dades estadístiques.

Dades d'utilització a nivell diari o mensual de domini públic podrien utilitzar-se per determinar patrons puntuals, que poden passar per alt en anàlisis a nivell anual, i poden aportar valor addicional al coneixement generat. Les possibles aplicacions d'aquest coneixement són variades: des d'ajudar en la planificació de dispositius especials de transport a aportar informació per possibles reestructuracions de la xarxa o canvis en els protocols en cas d'incidències. També serien d'utilitat pel sector comercial, permetent als venedors tenir una idea més acurada de quan poden esperar una afluència major o menor de viatgers.

Els resultats d'aquest treball són un exemple de com l'aplicació de tècniques de Mineria de Dades i Aprenentatge Computacional poden, i ho estan fent ja, aportar coneixement valuós a la societat en el seu conjunt.

REFERÈNCIES

[1] Sergey Brin, Lawrence Page, Rajeev Motwani, i Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.

[2] Cisco. Cisco global cloud index: Forecast and methodology, 2015-2020, 2016.

[3] Josep Curto Díaz. *Fundamentos de Big Data*, chapter 4, page 47. UOC.

[4] Montserrat Garcia Alsina. Open government, open data, big data y transparencia: la información como nexo de unión. *COMeIN*, (39), Dec 2014.

[5] Chris Greenwood. 'Scrambled' pedestrian crossings at signal controlled junctions - a case study. Technical report, Atkins.

[6] Mohammed Guller. *Big Data Analytics with Spark*, chapter 8, pages 160–161. Apress, 2015.

[7] Jiawei Han, Micheline Kamber, i Jian Pei. *Data Mining: Concepts and Techniques*, chapter 1, pages 5–8. Elsevier, 2012.

[8] Jiawei Han, Micheline Kamber, i Jian Pei. *Data Mining: Concepts and Techniques*, chapter 1, pages 10–13. Elsevier, 2012.

[9] Stuart P. Lloyd. Least squares quantization in PCM. 28(2):129–137, Mar 1982.

[10] Fernando Pérez i Brian E. Granger. IPython: a system for interactive scientific computing. *Computing in Science and Engineering*, 9(3):21–29, May 2007.

APÈNDIX

Especificacions tècniques del clúster

Component	*Master*
CPU	Intel Core i5-4200U (2 nuclis Haswell, 1.60 GHz)
Cache L1	128 KB
Cache L2	512 KB
Cache L3	3 MB
RAM	4 GB (DDR3L, 1333 MHz)
Interfície de xarxa	Realtek R8169 (Gigabit Ethernet)
Espai d'emmagatzemat	500 GB

TAULA 14: ESPECIFICACIONS TÈCNIQUES DE LA MÀQUINA *Master*

Component	*Slave-1*
CPU	Intel Core i5-3570K (4 nuclis Ivy Bridge, 3,4 GHz)
Cache L1	256 KB
Cache L2	1 MB
Cache L3	6 MB
RAM	8 GB (DDR3, 1600 MHz)
Interfície de xarxa	Intel 82579V (Gigabit Ethernet)
Espai d'emmagatzemat	500 GB

TAULA 15: ESPECIFICACIONS TÈCNIQUES DE LA MÀQUINA *Slave-1*

Component	*Slave-2*
CPU	Intel Core i3-540 (4 nuclis Clarkdale, 3,07 GHz)
Cache L1	128 KB
Cache L2	512 KB
Cache L3	4 MB
RAM	8 GB (DDR3, 1333 MHz)
Interfície de xarxa	Realtek R8169 (Gigabit Ethernet)
Espai d'emmagatzemat	1 TB

TAULA 16: ESPECIFICACIONS TÈCNIQUES DE LA MÀQUINA *Slave-2*

Arbre de decisió

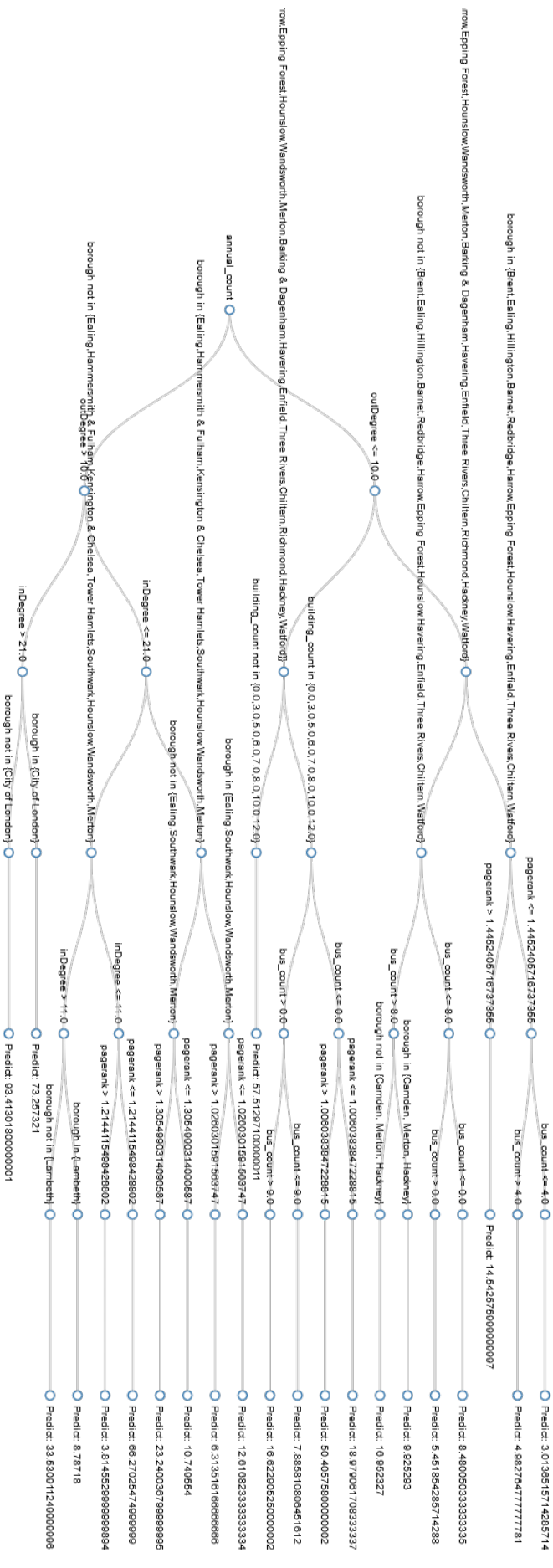

Fig. 9: Arbre de decisió dels factors que influeixen en la quantitat d'usuaris