
# Ús d'algorismes d'aprenentatge automàtic en entorns *big data* per a l'obtenció de models predictius de contaminació.

**Fidel Bonet Vilela**

**Grau d'Enginyeria Informàtica**
Àrea d'Intel·ligència Artificial
Consultor: Dr. David Isern Alarcón
Professor: Dr. Carles Ventura Royo

31 de maig de 2017

# FITXA DEL TREBALL FINAL

| | |
|---|---|
| **Títol del treball:** | *Ús d'algorismes d'aprenentatge automàtic en entorns* big data *per a l'obtenció de models predictius de contaminació.* |
| **Nom de l'autor:** | *Fidel Bonet Vilela* |
| **Nom del consultor/a:** | *Dr. David Isern Alarcón* |
| **Nom del PRA:** | *Dr. Carles Ventura Royo* |
| **Data de lliurament (mm/aaaa):** | *05/2017* |
| **Titulació o programa:** | *Grau d'Enginyeria Informàtica* |
| **Àrea del Treball Final:** | *Intel·ligència artificial* |
| **Idioma del treball:** | *Català* |
| **Paraules clau** | *Aprenentatge automàtic,* big data*, Apache Hadoop* |

**Resum del Treball:**

*L'objectiu d'aquest treball final de grau és la utilització d'algorismes d'aprenentatge automàtic en entorns* big data *per a l'obtenció de models predictius de contaminació atmosfèrica.*

*A partir de conjunts històrics de dades meteorològiques, de trànsit i de contaminació atmosfèrica provinents de sensors distribuïts en el territori s'han obtingut diversos models d'aprenentatge automàtic. Aquests models s'han generat en un entorn* big data *ja que, avui en dia, el volum de dades recollides pels sensors és molt elevat.*

*Per a dur-ho a terme, en primer lloc s'han implementat clústers Apache Hadoop en dues arquitectures: una de pseudodistribuïda en una màquina virtual i una altra de distribuïda en la plataforma Amazon Web Services. A continuació, s'ha emprat Apache Hive per a carregar les dades en el sistema de fitxers distribuït HDFS i per al tractament previ a la generació del model. Finalment, s'ha utilitzat Apache Mahout com a biblioteca d'aprenentatge automàtic.*

*Els models obtinguts permeten afirmar que la meteorologia i el trànsit tenen una afectació directa en l'augment de la concentració de diòxid de nitrogen. En concret, la inversió tèrmica, la intensitat de vehicles i la temperatura són les variables amb un pes més important alhora de modelar el comportament d'aquest contaminant.*

*Es pot concloure que els models obtinguts confirmen la hipòtesi inicial ja que permeten predir episodis de contaminació mitjançant dades meteorològiques i de trànsit.*

**Abstract:**

*The goal of this project is the use of machine learning algorithms in big data environments for obtaining predictive models of air pollution.*

*Based on historical weather, traffic and air pollution datasets from sensors distributed throughout the territory, several machine learning models have been obtained. These models have been created in a big data environment because, nowadays, the amount of data collected by sensors is very large.*

*In order to accomplish this, firstly, Apache Hadoop clusters have been implemented in two architectures: a pseudo-distributed one, using a virtual machine, and a distributed one in the Amazon Web Services platform. Afterwards, Apache Hive has been used to load the data into an HDFS distributed file system and preprocess it. Finally, Apache Mahout has been used as a machine learning library.*

*The models obtained allow confirming that the weather and traffic have a direct affectation in increasing the concentration of nitrogen dioxide. Specifically, inversion, traffic intensity and temperature are the most important variables in modelling the behaviour of this pollutant.*

*In conclusion, the model obtained confirms the initial hypothesis since it allows predicting pollution episodes from traffic and weather data.*

# ÍNDEX

# LLISTAT DE TAULES

# LLISTAT D'IMATGES

# 1. Introducció

En aquest capítol introductori s'analitza el context del problema que es vol resoldre en el projecte, els objectius d'aquest, el mètode seguit i la planificació del treball. Es finalitza amb un breu sumari dels productes obtinguts i la descripció dels següents capítols d'aquesta memòria.

## 1.1. Context i justificació del treball

En aquest treball final de grau es pretén obtenir models predictius de contaminació atmosfèrica a partir de dades meteorològiques, de trànsit i de pol·lució utilitzant tècniques d'intel·ligència artificial, concretament l'aprenentatge automàtic. Es parteix de la idea que les dades meteorològiques i de trànsit tenen una afectació directa en l'augment de la concentració de contaminants a l'atmosfera i que, per tant, l'anàlisi d'aquestes dades permetrien obtenir models predictius de pol·lució. Es pretén, doncs, demostrar aquesta hipòtesi.

Dels diversos contaminats atmosfèrics existents s'analitzarà el diòxid de nitrogen, ja que és el que té una relació més directa amb el trànsit. Cal tenir en compte que l'exposició prolongada a aquest contaminant pot provocar problemes en la salut de les persones i, per tant, és important conèixer el seu comportament i poder predir-lo.

Actualment, es coneixen els episodis d'augment de la contaminació directament a partir de les dades recollides per sensors distribuïts pel territori. En el cas que els índexs superin uns determinats llindars, es duen a tempre les mesures establertes com, per exemple, l'alerta a la població en els cas dels rangs més alts de concentració. La solució proposada permetria conèixer millor el comportament d'aquests fenòmens, predir-los i, per tant, ajudar tant als gestors com als ciutadans a emprendre mesures per a corregir la situació que ha desencadenat la concentració de pol·lució en aquells factors on es pot incidir, com ara: el control del trànsit en el cas dels gestors o bé l'ús del transport públic en el cas dels ciutadans. Aquestes accions haurien de permetre millorar la qualitat de l'aire i evitar els episodis d'augment de la pol·lució.

Per a l'obtenció dels models caldrà realitzar un anàlisi de diversos dels algorismes d'aprenentatge automàtic disponibles, com, per exemple, les xarxes neuronals, els arbres de decisió, la quantificació vectorial, i el mètode *Random forest*, per avaluar-ne la seva adequació a les dades que s'analitzaran.

Per altra banda, el nombre creixent de sensors de tot tipus distribuïts pel territori dificulta el tractament i l'anàlisi de les dades recollides en sistemes informàtics convencionals. És per això que s'utilitzarà un entorn *big data* per obtenir els diversos models predictius. Concretament es planteja l'ús de l'entorn de treball per a computació distribuïda Apache Hadoop[1] ja que permet l'execució d'aplicacions que usen un elevat volum de dades en clústers de grans dimensions.

En el projecte s'aprofundirà en els dos principals components d'Apache Hadoop: el paradigma de computació distribuïda MapReduce i el sistema de fitxers distribuïts HDFS (*Hadoop Distributed File System*), el qual permet crear múltiples rèpliques de les dades i distribuir-les en els diversos nodes del clúster per, d'aquesta forma, aconseguir computació ràpida i fiable. Per altra banda, caldrà analitzar les diverses llibreries d'aprenentatge automàtic disponibles en l'ecosistema Hadoop.

[1] http://hadoop.apache.org/

Finalment, la informàtica en núvol permet l'accés a un conjunt de recursos de computació com ara: recursos de xarxa, emmagatzematge, servidors i aplicacions. Així, aquest paradigma esdevé útil per a l'obtenció dels models d'aprenentatge proposats anteriorment en un entorn distribuït.

En conclusió, el resultat que es vol obtenir d'aquest treball final de grau és un sistema d'informàtica en núvol capaç d'analitzar grans volums de dades meteorològiques, de trànsit i de contaminació en un entorn *big data* a través de tècniques d'aprenentatge automàtic per a obtenir un model predictiu de contaminació atmosfèrica.

## 1.2. Objectius del treball

Tot seguit s'exposen els objectius generals i específics del projecte per a, tot seguit, analitzar la seva priorització.

| Objectiu general | Descripció |
|---|---|
| G1 | Implementació de solucions *big data*. |
| G2 | Ús d'algorismes d'aprenentatge automàtic en entorns distribuïts. |
| G3 | Utilització d'algorismes d'aprenentatge automàtic per a l'obtenció de models predictius de contaminació atmosfèrica. |

***Taula 1**. Objectius generals.*

| Objectiu específic | Descripció |
|---|---|
| E1 | Anàlisi del paradigma *big data*. |
| E2 | Anàlisi del paradigma de la informàtica en núvol. |
| E3 | Anàlisi de l'entorn de treball per a computació distribuïda Apache Hadoop. |
| E4 | Anàlisi de diversos algorismes d'aprenentatge automàtic en entorns distribuïts útils per a l'obtenció de models predictius: reducció de la dimensionalitat, categorització i classificació. |
| E5 | Anàlisi del model de programació MapReduce per a la implementació dels anteriors algorismes. |
| E6 | Anàlisi de la implementació d'algorismes d'aprenentatge automàtic Apache Mahout. |
| E7 | Utilització de les tècniques anteriors per a l'obtenció de models predictius de contaminació atmosfèrica. |
| E8 | Comprovació de la idoneïtat de les dades meteorològiques i de trànsit per a l'obtenció de models predictius de contaminació. |

***Taula 2**. Objectius específics.*

Pel que fa als objectius generals, el més prioritari és l'ús d'algorismes d'aprenentatge automàtic en entorns distribuïts. En segon lloc, es pretén obtenir models predictius de contaminació atmosfèrica a partir de dades de trànsit, meteorològiques i de contaminació i també analitzar la implementació de solucions *big data* per a resoldre problemes com l'anterior.

Pel que fa als objectius específics, en la taula següent se'n mostra la prioritat de cadascun d'ells.

| Objectiu específic | Descripció | Prioritat |
|---|---|---|
| E1 | Anàlisi del paradigma *big data*. | Mitjana |
| E2 | Anàlisi del paradigma de la informàtica en núvol. | Mitjana |
| E3 | Anàlisi de l'entorn de treball per a computació distribuïda Apache Hadoop. | Mitjana |
| E4 | Anàlisi de diversos algorismes d'aprenentatge automàtic en entorns distribuïts útils per a l'obtenció de models predictius: reducció de la dimensionalitat, categorització i classificació. | Alta |
| E5 | Anàlisi del model de programació MapReduce per a la implementació dels anteriors algorismes. | Mitjana |
| E6 | Anàlisi de la implementació d'algorismes d'aprenentatge automàtic Apache Mahout. | Mitjana |
| E7 | Utilització de les tècniques anteriors per a l'obtenció de models predictius de contaminació atmosfèrica. | Alta |
| E8 | Comprovació de la idoneïtat de les dades meteorològiques i de trànsit per a l'obtenció de models predictius de contaminació. | Alta |

***Taula 3**. Prioritització dels objectius específics.*

En conclusió, els objectius centrals del projecte són l'anàlisi de diversos algorismes d'aprenentatge automàtic en entorns distribuïts, l'obtenció de models predictius de contaminació en entorns *big data* mitjançant els anteriors algorismes i la comprovació de la idoneïtat de l'ús de les dades meteorològiques i de trànsit per a predir episodis de contaminació. Per altra banda, en un segon nivell de prioritat s'hi troben la resta d'objectius: l'anàlisi dels paradigmes *big data*, la informàtica en núvol, el model de programació MapReduce i els algorismes que ofereix la implementació Apache Mahout.

## 1.3. Enfocament i mètode seguit

Per a la realització del projecte s'ha utilitzat l'entorn de treball per a computació distribuïda Apache Hadoop així com la implementació d'algorismes d'aprenentatge automàtic per a entorns distribuïts i escalables Apache Mahout. Ambdues solucions es distribueixen sota llicències lliures.

S'ha escollit aquest entorn *big data* ja que ha estat utilitzat en nombrosos projectes amb resultats exitosos. S'ha considerat, doncs, que és una tecnologia prou consolidada. Per altra banda, disposa de les eines necessàries per a implementar algorismes d'aprenentatge automàtic en entorns distribuïts amb un gran volum de dades, que és un dels objectius prioritaris d'aquest treball.

Per altra banda, s'han emprat les distribucions del sistema operatiu GNU/Linux de codi lliure Ubuntu[2] i CentOS[3].

En una primera fase, el projecte s'ha implementat en un únic node utilitzant un model pseudodistribuït que permetés realitzar les proves necessàries. Un cop s'ha verificat la correctesa de la implementació aquesta s'ha exportat a un sistema distribuït real on s'han dut a terme les proves amb diverses configuracions. En concret, s'ha optat per l'ús de la capa gratuïta d'Amazon Web Services[4], un dels principals serveis d'informàtica en núvol.

[2] https://www.ubuntu.com/
[3] https://www.centos.org/
[4] https://aws.amazon.com/big-data/

## 1.4. Planificació del Treball

En aquest apartat es descriuen els recursos utilitzats per a la realització del projecte tant pel que fa al maquinari i el programari com a les dades. A continuació, es detallen les tasques del projecte i, finalment, s'exposa la planificació temporal prevista inicialment així com les diverses revisions d'aquesta.

### 1.4.1. Recursos per a la realització del projecte

Tot seguit es detallen els recursos de maquinari i programari així com les dades necessàries per al desenvolupament del projecte.

#### *Maquinari i programari*

Les següents taules enumeren els recursos de maquinari i programari utilitzats per a la realització de les proves inicials en mode pseudodistribuït i per a la implementació en un sistema distribuït real.

| | | |
|---|---|---|
| **Ordinador** | **Processador** | *3,06 GHZ Intel Core 2 Duo* |
| | **Memòria RAM** | *4 GiB* |
| | **Sistema operatiu amfitrió** | *macOS Sierra* |
| **Hipervisor** | | *Oracle VM VirtualBox*[5] |
| **Sistema operatiu convidat** | | *Ubuntu 16.04* |
| **Java** | | *Java SE Development Kit 8* |
| **Entorn de treball de computació distribuïda** | | *Apache Hadoop 2.8.0* |
| **Plataforma per a consultes SQL distribuïdes** | | *Apache Hive 2.1.1* |
| **Biblioteca d'algorismes d'aprenentatge automàtic** | | *Apache Mahout 0.11.0* |

***Taula 4**. Recursos necessaris en mode pseudodistribuït.*

| | | |
|---|---|---|
| **Ordinador** | **Processador** | *3,06 GHZ Intel Core 2 Duo* |
| | **Memòria RAM** | *4 GiB* |
| | **Sistema operatiu** | *macOS Sierra* |
| **Clúster Amazon Web Services** | **Nombre de nodes** | *Dues configuracions amb: 4 i 8 nodes* |
| | **Processador** | *1 vCPU de 2,5 GHz* |
| | **Memòria RAM** | *1 GiB* |
| | **Sistema operatiu** | *Ubuntu Server 16.04 LTS* |
| | **Unitat d'emmagatzematge** | *SSD, 8 GiB* |
| **Java** | | *Java SE Development Kit 8* |
| **Entorn de treball de computació distribuïda** | | *Apache Hadoop 2.8.0* |
| **Plataforma per a consultes SQL distribuïdes** | | *Apache Hive 2.1.1* |
| **Biblioteca d'algorismes d'aprenentatge automàtic** | | *Apache Mahout 0.11.0* |

***Taula 5**. Recursos necessaris en mode distribuït.*

[5] https://www.virtualbox.org/

### *Dades*

Les dades emprades per a l'obtenció del model de contaminació s'han obtingut de les següents fonts:

- ***Dades meteorològiques***: Oficina de l'Energia i del Canvi Climàtic del Govern d'Andorra[6].

- ***Dades de trànsit***: Àrea de Mobilitat del Govern d'Andorra[7].

- ***Dades de contaminació atmosfèrica***: Unitat de Medi Atmosfèric del Departament de Medi Ambient i Sostenibilitat del Govern d'Andorra[8].

Cap de les dades que s'han utilitzat per a la realització del projecte estan protegides per les lleis de protecció de dades de caràcter personal.

## 1.4.2. Descripció de les tasques.

A partir dels objectius exposats anteriorment s'han extret les tasques principals del treball. La següent taula en mostra les descripcions així com la relació de cada tasca amb l'objectiu corresponent.

| Objectiu | Tasca | *Descripció* |
|---|---|---|
| **Anàlisi del paradigma *big data*** | Analitzar el paradigma *big data* | *Cerca d'informació sobre aquest paradigma i descripció del seu funcionament.* |
| **Anàlisi del paradigma de la informàtica en núvol** | Analitzar el paradigma de la informàtica en núvol | *Cerca d'informació sobre el paradigma de la informàtica en núvol i descripció del seu funcionament.* |
| **Estudi de l'entorn de treball per a computació distribuïda Apache Hadoop** | Anàlisi d'Apache Hadoop | *Cerca de documentació sobre l'entorna de treball per a computació distribuïda Apache Hadoop.* |
| | Instal·lar i configurar d'Apache Hadoop en un entorn pseudodistribuït | *Instal·lació i configuració en l'entorn pseudodistribuït on s'implementarà el model d'aprenentatge en una primera fase.* |
| | Instal·lar i configurar d'Apache Hadoop en un entorn distribuït | *Instal·lació i es configuració en l'entorn distribuït on es durà a terme la implementació final del model d'aprenentatge.* |
| **Anàlisi de diversos algorismes d'aprenentatge automàtic** | Analitzar els algorismes d'aprenentatge automàtic útils per a l'obtenció del model de contaminació | *Anàlisi de diversos algorismes d'aprenentatge automàtic que s'ajustin a les necessitats concretes del model d'aprenentatge proposat.* |
| **Anàlisi del model de programació MapReduce per a la implementació dels anteriors algorismes.** | Analitzar MapReduce | *Cerca de documentació sobre el model de programació MapReduce i les implementacions que existeixen per als algorismes d'aprenentatge automàtic obtinguts en la tasca anterior.* |
| **Estudi de la implementació d'algorismes d'aprenentatge automàtic Apache Mahout** | Analitzar els algorismes que ofereix Apache Mahout | *Anàlisi els algorismes que ofereix Apache Mahout per a la resolució del problema proposat.* |
| | Instal·lar i configurar Apache Mahout en un entorn pseudodistribuït | *Instal·lació i configuració d'Apache Mahout en l'entorn pseudodistribuït on es durà a terme la implementació del model d'aprenentatge en una primera fase.* |
| | Instal·lar i configurar Apache Mahout en un entorn distribuït | *Instal·lació i configuració d'Apache Mahout en l'entorn distribuït on es durà a terme la implementació final del model.* |

[6] http://www.meteo.ad/climatologia
[7] http://www.mobilitat.ad/
[8] http://www.aire.ad/data_and_statistics_home.php

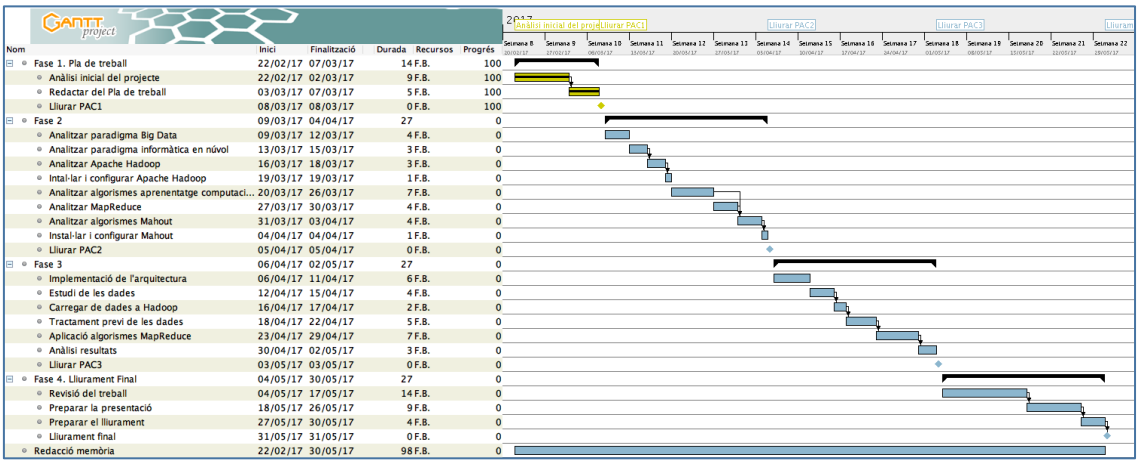
| | | |
|---|---|---|
| **Utilització de les tècniques anteriors per a l'obtenció de models predictius de contaminació atmosfèrica** | Estudiar les dades | *Anàlisi de les dades disponibles en l'entorn R per a comprendre'n l'estructura* |
| | Implementar l'arquitectura d'informàtica en núvol en un entorn pseudodistribuït | *Implementació de l'arquitectura d'informàtica en núvol en un entorn pseudodistribuït* |
| | Càrrega de dades a Hadoop | *Càrrega de les dades a l'entorn Hadoop mitjançant consultes HiveQL en la plataforma Hive.* |
| | Tractament previ de daddes | *Tractament previ de les dades mitjançant consultes HiveQL en la plataforma Hive.* |
| | Aplicar els diversos algorismes MapReduce | *Aplicació d'algorismes MapReduce d'aprenentatge automàtic.* |
| | Implementar la solució obtinguda en un entorn distribuït | *Implementació de la solució obtinguda en un entorn distribuït.* |
| | Anàlisi dels resultats | *Un cop obtingut el model se n'analitzarà la qualitat. També s'analitzaran els resultats obtinguts així com el rendiment del sistema.* |
| **Preparació del lliurament** | Revisar el treball | *Revisió de les diverses fases del treball i redacció final de la memòria.* |
| | Preparar la presentació | *Preparació de la presentació.* |
| | Preparar el lliurament | *Preparació del lliurament: memòria, codi i presentació.* |
| **Redacció de la memòria** | Redactar la memòria | *Redacció del contingut de la memòria durant totes les fases del treball.* |

***Taula 6**. Tasques.*

Per a ordenar les diverses tasques en que s'ha dividit el projecte, s'ha tingut en compte la priorització detallada en l'apartat 1.2 per a dur a terme abans aquelles que són prioritàries sempre que no hi hagués una dependència entre tasques que ho impedís.

### 1.4.3. Planificació temporal.

Un cop definides les tasques que es duran a terme s'ha analitzat la planificació temporal del projecte. El següent diagrama de Gantt mostra aquesta planificació on cada lliurament s'indica com una fita. També s'hi assenyalen les dependències entre les tasques.




***Imatge 1**. Planificació inicial.*

Per altra banda, s'assenyala en color groc aquelles tasques que ja han estat completades. També s'hi mostra la durada de cadascuna així com els recursos assignats a aquestes.

Com es pot observar, s'ha assignat una part important del temps destinat al projecte a l'anàlisi previ dels paradigmes amb els quals es treballarà així com a la tecnologia proposada.

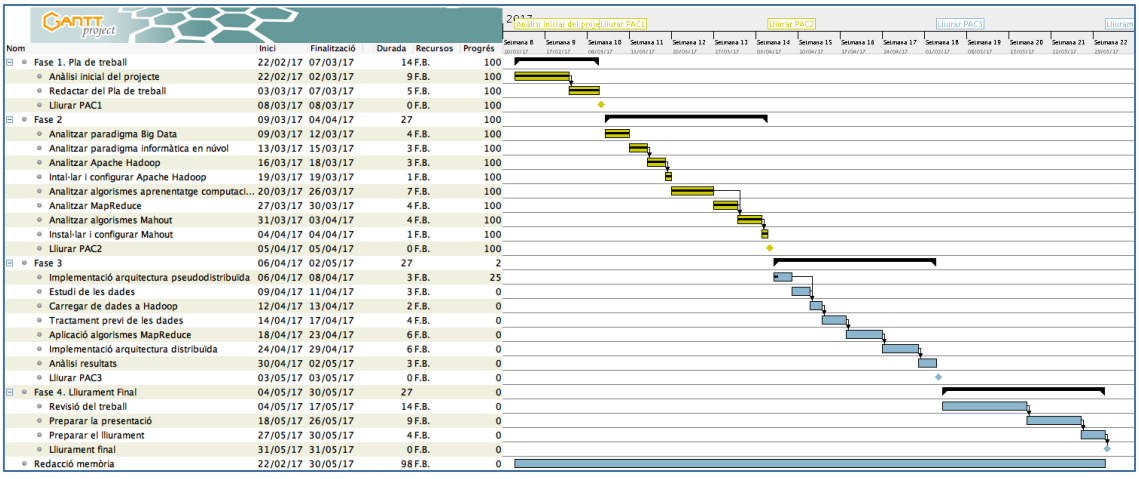
En la planificació no s'han eliminat els caps de setmana ja que, a diferència d'un projecte desenvolupat en un entorn laboral, també s'empraran per a la realització d'aquest. Tampoc es contempla cap període de vacances tot i que s'han tingut en compte cinc jornades festives.

### 1.4.4. Seguiment de la planificació.

Durant la realització de la fase 1 del treball s'ha decidit modificar la planificació de la fase 2 ja que s'han identificat problemes alhora d'implementar la solució en un sistema distribuït. És per això que s'ha afegit una nova tasca en la segona fase del projecte encaminada a implementar l'arquitectura distribuïda. Així, la tasca *Implementació de l'arquitectura* ha estat modificada per reservar la meitat del temps assignat inicialment a aquesta a la implementació de l'arquitectura pseudodistribuïda. Per altra banda, s'han reduït les tasques *Estudi de les dades*, *Tractament previ de les dades* i *Aplicació algorismes MapReduce* en un dia cadascuna. Aquestes reduccions s'han utilitzat per assignar sis dies a la nova tasca *Implementació arquitectura distribuïda* que s'ha previst realitzar un cop finalitzades les tasques anteriors en una arquitectura pseudodistribuïda.

S'ha considerat que la desviació detectada es pot solucionar amb els canvis en la planificació descrits sense que calgui un augment significatiu de la dedicació ni la modificació dels objectius inicials del projecte. Cal tenir en compte que en la planificació inicial s'havia previst la possibilitat d'haver de canviar l'arquitectura de la solució si el maquinari de què es disposava no era suficient. És per això que s'havia assignat suficient temps a la implementació tot preveient possibles desviacions.

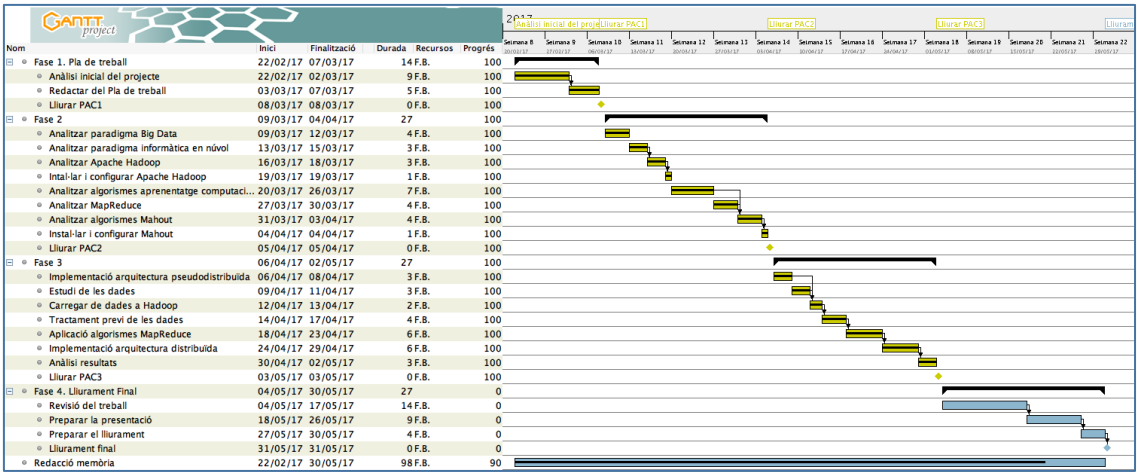
El següent diagrama de Gantt mostra la planificació un cop realitzats els canvis anteriors així com l'estat de la consecució dels treballs en la data de finalització de la fase 2. Com es pot observar, no s'han produït endarreriments.



| Nom | Inici | Finalització | Durada | Recursos | Progrés |
|---|---|---|---|---|---|
| Fase 1. Pla de treball | 22/02/17 | 07/03/17 | 14 F.B. | | 100 |
| Anàlisi inicial del projecte | 22/02/17 | 02/03/17 | 9 F.B. | | 100 |
| Redactar del Pla de treball | 03/03/17 | 07/03/17 | 5 F.B. | | 100 |
| Lliurar PAC1 | 08/03/17 | 08/03/17 | 0 F.B. | | 100 |
| Fase 2 | 09/03/17 | 04/04/17 | 27 | | 100 |
| Analitzar paradigma Big Data | 09/03/17 | 12/03/17 | 4 F.B. | | 100 |
| Analitzar paradigma informàtica en núvol | 13/03/17 | 15/03/17 | 3 F.B. | | 100 |
| Analitzar Apache Hadoop | 16/03/17 | 18/03/17 | 3 F.B. | | 100 |
| Intal·lar i configurar Apache Hadoop | 19/03/17 | 19/03/17 | 1 F.B. | | 100 |
| Analitzar algorismes aprenentatge computaci... | 20/03/17 | 26/03/17 | 7 F.B. | | 100 |
| Analitzar MapReduce | 27/03/17 | 30/03/17 | 4 F.B. | | 100 |
| Analitzar algorismes Mahout | 31/03/17 | 03/04/17 | 4 F.B. | | 100 |
| Instal·lar i configurar Mahout | 04/04/17 | 04/04/17 | 1 F.B. | | 100 |
| Lliurar PAC2 | 05/04/17 | 05/04/17 | 0 F.B. | | 100 |
| Fase 3 | 06/04/17 | 02/05/17 | 27 | | 2 |
| Implementació arquitectura pseudodistribuïda | 06/04/17 | 08/04/17 | 3 F.B. | | 25 |
| Estudi de les dades | 09/04/17 | 11/04/17 | 3 F.B. | | 0 |
| Carregar de dades a Hadoop | 12/04/17 | 13/04/17 | 2 F.B. | | 0 |
| Tractament previ de les dades | 14/04/17 | 17/04/17 | 4 F.B. | | 0 |
| Aplicació algorismes MapReduce | 18/04/17 | 23/04/17 | 6 F.B. | | 0 |
| Implementació arquitectura distribuïda | 24/04/17 | 29/04/17 | 6 F.B. | | 0 |
| Anàlisi resultats | 30/04/17 | 02/05/17 | 3 F.B. | | 0 |
| Lliurar PAC3 | 03/05/17 | 03/05/17 | 0 F.B. | | 0 |
| Fase 4. Lliurament Final | 04/05/17 | 30/05/17 | 27 | | 0 |
| Revisió del treball | 04/05/17 | 17/05/17 | 14 F.B. | | 0 |
| Preparar la presentació | 18/05/17 | 26/05/17 | 9 F.B. | | 0 |
| Preparar el lliurament | 27/05/17 | 30/05/17 | 4 F.B. | | 0 |
| Lliurament final | 31/05/17 | 31/05/17 | 0 F.B. | | 0 |
| Redacció memòria | 22/02/17 | 30/05/17 | 98 F.B. | | 0 |

***Imatge 2**. Replanificació i situació de l'avanç del projecte en el lliurament de la fase 2.*

Per altra banda, el següent diagrama mostra l'estat dels treballs en la finalització de la fase 3 on s'han complert les previsions després de les modificacions de la planificació dutes a terme durant la fase 2.



| Nom | Inici | Finalització | Durada | Recursos | Progrés |
|---|---|---|---|---|---|
| Fase 1. Pla de treball | 22/02/17 | 07/03/17 | 14 F.B. | | 100 |
| Anàlisi inicial del projecte | 22/02/17 | 02/03/17 | 9 F.B. | | 100 |
| Redactar del Pla de treball | 03/03/17 | 07/03/17 | 5 F.B. | | 100 |
| Lliurar PAC1 | 08/03/17 | 08/03/17 | 0 F.B. | | 100 |
| Fase 2 | 09/03/17 | 04/04/17 | 27 | | 100 |
| Analitzar paradigma Big Data | 09/03/17 | 12/03/17 | 4 F.B. | | 100 |
| Analitzar paradigma informàtica en núvol | 13/03/17 | 15/03/17 | 3 F.B. | | 100 |
| Analitzar Apache Hadoop | 16/03/17 | 18/03/17 | 3 F.B. | | 100 |
| Instal·lar i configurar Apache Hadoop | 19/03/17 | 19/03/17 | 1 F.B. | | 100 |
| Analitzar algorismes aprenentatge computaci... | 20/03/17 | 26/03/17 | 7 F.B. | | 100 |
| Analitzar MapReduce | 27/03/17 | 30/03/17 | 4 F.B. | | 100 |
| Analitzar algorismes Mahout | 31/03/17 | 03/04/17 | 4 F.B. | | 100 |
| Instal·lar i configurar Mahout | 04/04/17 | 04/04/17 | 1 F.B. | | 100 |
| Lliurar PAC2 | 05/04/17 | 05/04/17 | 0 F.B. | | 100 |
| Fase 3 | 06/04/17 | 02/05/17 | 27 | | 100 |
| Implementació arquitectura pseudodistribuida | 06/04/17 | 08/04/17 | 3 F.B. | | 100 |
| Estudi de les dades | 09/04/17 | 11/04/17 | 3 F.B. | | 100 |
| Carregar de dades a Hadoop | 12/04/17 | 13/04/17 | 2 F.B. | | 100 |
| Tractament previ de les dades | 14/04/17 | 17/04/17 | 4 F.B. | | 100 |
| Aplicació algorismes MapReduce | 18/04/17 | 23/04/17 | 6 F.B. | | 100 |
| Implementació arquitectura distribuida | 24/04/17 | 29/04/17 | 6 F.B. | | 100 |
| Anàlisi resultats | 30/04/17 | 02/05/17 | 3 F.B. | | 100 |
| Lliurar PAC3 | 03/05/17 | 03/05/17 | 0 F.B. | | 100 |
| Fase 4. Lliurament Final | 04/05/17 | 30/05/17 | 27 | | 0 |
| Revisió del treball | 04/05/17 | 17/05/17 | 14 F.B. | | 0 |
| Preparar la presentació | 18/05/17 | 26/05/17 | 9 F.B. | | 0 |
| Preparar el lliurament | 27/05/17 | 30/05/17 | 4 F.B. | | 0 |
| Lliurament final | 31/05/17 | 31/05/17 | 0 F.B. | | 0 |
| Redacció memòria | 22/02/17 | 30/05/17 | 98 F.B. | | 90 |

***Imatge 3**. Situació de l'avanç del projecte en el lliurament de la fases 3.*

Finalment, la fase 4 del projecte s'ha completat segons les previsions inicials amb el tancament del projecte havent realitzat totes les tasques previstes i segons la planificació temporal establerta a l'inici.

## 1.5. Breu sumari de productes obtinguts

Els productes obtinguts en aquest treball han estat:

- Model predictiu de contaminació així com els Scripts i el codi HiveQL per al tractament d'informació meteorològica, de trànsit i de contaminació i per a l'obtenció del model d'aprenentatge automàtic en un entorn Apache Hadoop.

- Clúster Apache Hadoop en la plataforma Amazon Web Services.

## 1.6. Breu descripció dels altres capítols de la memòria

El capítol 2 se centra en la descripció del paradigma *big data* on s'enumeren les seves principals característiques i categories així com els principals entorns de treball que l'utilitzen. Es conclou el capítol valorant els diversos entorns analitzats i detallant perquè s'ha escollit Apache Hadoop per a la realització d'aquest projecte.

Per altra banda, el capítol 3 introdueix el concepte d'informàtica en núvol, els diversos models de servei que ofereix i els models de desplegament existents. Finalment, s'analitzen alguns

dels principals serveis d'informàtica en núvol disponibles per poder valorar el més adient per a la implementació del model d'aprenentatge proposat.

El capítol 4 descriu amb més profunditat l'entorn per a processament distribuït Apache Hadoop que ja ha estat introduït en el capítol 2. S'hi descriu MapReduce, el model de programació que empra Hadoop, i HDFS, el seu sistema de fitxers.

Tot seguit, s'introdueixen els principals conceptes de l'aprenentatge automàtic en el capítol 5 donant especial èmfasi als algorismes de classificació i es descriuen breument les eines d'aprenentatge més destacades en l'ecosistema Apache Hadoop. A continuació es duu a terme una valoració d'aquestes i es detalla perquè s'ha escollit la biblioteca Apache Mahout per a la realització del projecte. Finalment, s'enumeren els algorismes d'aprenentatge automàtic d'aquesta darrera biblioteca.

Un cop analitzades les bases teòriques del treball, en el capítol 6 es descriu el procés seguit per a la implementat del model predictiu de contaminació: l'anàlisi de les dades disponibles, l'elecció dels algorismes d'aprenentatge, la preparació de l'arquitectura *big data*, la càrrega i el tractament previ de les dades, l'obtenció del model d'aprenentatge i l'anàlisi dels resultats obtinguts.

Per altra banda, en el capítol 7 es detallen les conclusions obtingudes en el projecte, en el capítol 8 s'enumeren els termes i acrònims més important utilitzats en aquesta memòria i en el capítol 9 es mostra la bibliografia emprada durant la realització del treball.

Finalment, el capítol 10 conté els annexos amb els manuals per a la instal·lació de les eines utilitzades tant en mode pseudodistribuït com distribuït, les instruccions per a l'execució del cas pràctic i el codi implementat.

# 2. *Big data*

En aquest capítol s'analitza el paradigma *big data*, una de les tecnologies que s'utilitzarà en aquest treball. També s'hi descriuen els principals entorns de treball *big data* existents i s'analitza quin s'adapta millor a les necessitats del projecte.

## 2.1. Definició

Segons Turner (2014) la mida de les dades digitals creixerà a un ritme del 40% per any durant l'actual dècada, passant dels 4,4 zettabytes l'any 2013 a una estimació de 44 zettabytes per a l'any 2020, és a dir, 44.000 milions de terabytes. El motiu d'aquest augment és el creixent ús d'Internet per a realitzar tot tipus de tasques però també les dades generades per dispositius intel·ligents i la Internet de les coses (en anglès, *Internet of things*, IoT). Aquest gran volum de dades planteja nous reptes pel que fa al seu emmagatzematge i processament.

Les dades massives (en anglès, *big data*) fan referència, precisament, a conjunts de dades que o bé per la seva grandària o bé per la seva complexitat no poden ser processades pels sistemes tradicionals.

Una de les definicions més esteses d'aquest paradigma és la proposada per Laney (2001) segons la qual el *big data* és "el conjunt de tècniques i tecnologies per al tractament de dades en entorns de gran volum, varietat d'orígens i en els que la velocitat de resposta és crítica". Per tant, es presenten tres dimensions, sovint anomenades les "3 *V*" del *big data*:

- *El* ***volum*** fa referència a la mida de les dades. Com s'ha comentat anteriorment, actualment les dades digitals que es generen creixen de forma exponencial, passant de la generació de gigabytes d'informació diària fa uns anys a la generació de terabytes actualment. Aquesta problemàtica es fa evident degut al gran volum de dades que es generen provinents de fonts tan diverses com: Internet, els dispositius mòbils, les imatges satèl·lit, la Internet de les coses, les xarxes socials, els sensors sense fils, les càmeres, la comunicació màquina a màquina (en l'anglès, *machine-to-machine*, M2M), etc.

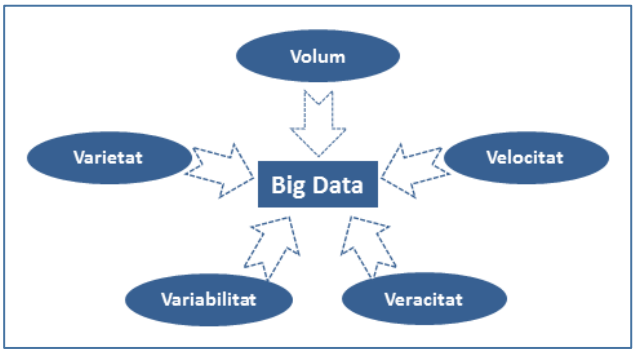
- *La* ***varietat*** es refereix a la diversitat i incompatibilitat dels formats de les dades pel fet que aquestes provenen de diferents fonts. Així, els orígens de dades poden ser: estructurats com en les bases de dades o les fulles de càlcul, semiestructurats com en les pàgines web o els documents XML o bé no estructurats com en els documents de text, les imatges, l'àudio o el vídeo.

- *La* ***velocitat*** indica la rapidesa en què les dades arriben als sistemes, sovint en temps real, i també a la necessitat de processar-les ràpidament. Per altra banda, també fa referència a la taxa de variació d'aquestes. Cal tenir en compte que en molts sistemes el temps de resposta és crític.

Altres autors han estès aquestes tres dimensions afegint-hi la variabilitat i la veracitat:

- La ***variabilitat*** indica el nivell de consistència de les dades, és a dir, la diversitat semàntica que s'hi produeix.

- La ***veracitat*** fa referència a la qualitat de les dades. Aquesta pot variar degut a la diversitat d'orígens i a la seva quantitat. Els sistemes *big data* no només han d'assegurar que els anàlisis són correctes sinó també que les dades que s'utilitzin ho siguin.




***Imatge 4**. Les 5 V del big data.*

Els sistemes tradicionals sovint presenten dificultats alhora de processar aquest tipus de dades i esdevé necessari l'ús de sistemes d'emmagatzematge distribuït així com la computació paral·lela per al seu tractament.

La Comissió Econòmica de les Nacions Unides per a Europa (Vale, 2013) classifica el tipus de sistemes *big data* en tres categories:

- ***Xarxes socials***. Es tracta d'informació d'origen humà digitalitzada i emmagatzemada en xarxes socials i ordinadors personals. Aquestes dades normalment són semiestructurades o no estructurades. Se'n poden destacar els següents entorns: xarxes socials, blogs, documents personals, fotografies, vídeos, cerques a Internet, missatges de text, mapes generats pels usuaris i correus electrònics.

- ***Sistemes de negoci tradicionals***. En aquest cas les dades són estructurades i normalment emmagatzemades en bases de dades relacionals i fan referència a registres de clients, la fabricació de productes, la gestió de les comandes, etc. Aquestes dades poden ser produïdes per les administracions o bé per empreses com és el cas de les transaccions comercials, registres bancaris, comerç electrònic i targes de crèdit.

- ***Internet de les coses***. El creixen nombre de sensors i altres dispositius que mesuren i enregistren elements del món físic són una font de dades estructurades generades per màquines. Se'n poden destacar les dades provinents tant de sensors fixes com mòbils i també les dades provinents de sistemes informàtics com són els diaris (en anglès, *logs*).

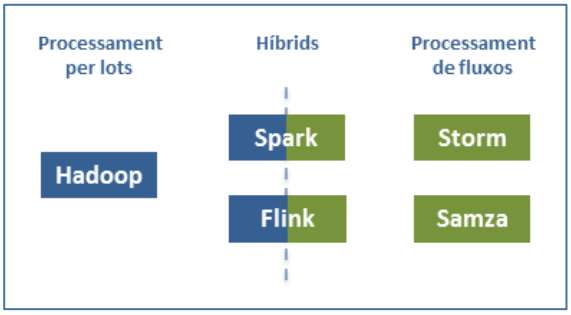
## 2.2. Principals entorns de treball *big data*

Com apunta Ellingwood (2016) els entorns de treball *big data* es poden classificar en tres grups:

- ***Entorns de treball de processament per lots*** (en anglès, *batch processing*). El processament es duu a terme sobre un gran volum de dades que no canvia, sovint dades històriques. Per tant, aquest tipus de processament no és indicat quan el temps és un factor important. Entre aquest tipus d'entorns destaca Apache Hadoop.

- ***Entorns de treball de processament de fluxos de dades*** (en anglès, *stream processing*). En aquests entorns les dades són processades no en lot sinó individualment a mesura que entren al sistema. Per tant, aquest tipus és l'indicat quan es requereix processament proper al temps real. S'hi troben els projectes Apache Storm i Apache Samza.

- ***Entorns de treball híbrids***. Suporten els dos tipus de processaments anteriors. Hi destaquen Apache Spark i Apache Flink.




***Imatge 5***. *Classificació dels principals entorns de treball big data.*

Tot seguit es descriu breument cadascun dels anteriors entorns de treball i s'analitza quin s'adapta millor als objectius del projecte.

### 2.2.1. Apache Hadoop

Apache Hadoop és un entorn de treball de processament només per lots. Està format pel sistema de fitxers distribuït HDFS, el component de coordinació de clústers YARN i el model de programació MapReduce.

Hadoop llegeix les dades del seu propi sistema de fitxers HDFS, les divideix en fragments que tot seguit distribueix pels diversos nodes del clúster, processa cada subconjunt de dades en el respectiu node, combina els resultats de cada node i emmagatzema el resultat altre cop al sistema de fitxers.

Els principals avantatges de Hadoop són la capacitat de processar conjunts de dades molt grans, la possibilitat de treballar amb maquinari convencional, l'alta escalabilitat i el fet de disposa d'un ampli ecosistema d'aplicacions.

Per contra, el seus principals inconvenients són la seva lentitud en comparació amb altres entorns i la corba d'aprenentatge del model de programació MapReduce.

### 2.2.2. Apache Storm

Apache Storm[9] és un entorn de treball de processament de fluxos de grans volums de dades amb una latència molt baixa i, per tant, indicat per al processament proper al temps real.

Storm gestiona el processament de les dades amb un conjunt de *topologies*. Una topologia està formada per:

- Fluxos de dades il·limitats que entren contínuament al sistema.
- Fonts dels fluxos de dades.
- Procés dut a terme sobre els fluxos de dades que genera un flux de dades com a resultat.

Els principals avantatges d'Strom són la baixa latència en el processament de dades, fet que el converteix en la millor solució per al processament proper al temps real, i la possibilitat d'integrar-lo amb Hadoop. Per contra, Storm no garanteix que un missatge no puguin ser duplicats.

### 2.2.3. Apache Samza

Apache Samza[10] és un entorn de treball de processament de fluxos que utilitza el sistema de missatges Apache Kafka per a garantir la tolerància a les fallades, memòria intermèdia i emmagatzematge d'estat. Per altra banda, funciona sobre clústers Hadoop i utilitza el coordinador de clústers YARN per a gestionar-ne els recursos.

Cadascun dels fluxos de dades que entren en el sistema són anomenats *tòpics*. Cada tòpic és dividit en parts i distribuït entre els diversos nodes del clúster. Cada missatge amb la mateixa clau serà processat per un mateix node.

Els principals avantatges de Samza són la baixa latència, la possibilitat que diversos subscriptors accedeixin als missatges resultants de cadascun del passos que es duen a terme en el processament de les dades d'entrada, la no pèrdua de dades en el cas que aquestes arribin al sistema a una velocitat més alta de la que són processades, l'emmagatzematge de l'estat i un nivell d'abstracció superior al d'altres entorns de treball de processament de fluxos com Storm. Per contra, com a principal desavantatge cal destacar que no suporta tants llenguatges de programació com Storm.

### 2.2.4. Apache Spark

Apache Spark[11] és un entorn de treball amb capacitats de processament tant per lots com de fluxos de dades. Malgrat que es basa en el funcionament de MapReduce, Spark utilitza computació totalment en memòria i optimitzacions de processament per a fer més ràpids els processos per lots. D'aquesta forma, Spark només interacciona amb el sistema de fitxers per a obtenir les dades inicials i emmagatzemar els resultats finals. La resta del procés es realitza en la memòria.

[9] http://storm.apache.org/
[10] http://samza.apache.org/
[11] http://spark.apache.org/

Pel que fa al model de processament de fluxos, aquest l'ofereix Spark Streaming gestionant els fluxos de dades com a petits processos per lots que són processats utilitzant el model per lots.

Aquest entorn pot ser utilitzat independentment o conjuntament amb Hadoop per a substituir el motor MapReduce.

En definitiva, el principal avantatge d'Spark sobre MapReduce és la major velocitat del primer. Altres avantatges d'Spark són la seva versatilitat ja que pot ser desplegat com un clúster independent o integrat dins d'un clúster Hadoop i també ser emprat en mode per lots o bé de fluxos de dades. Per altra banda, disposa de biblioteques d'aprenentatge automàtic i consultes interactives i és més fàcil d'implementar que MapReduce.

Per contra, el seu principal desavantatge és l'augment de la latència en sistemes de processament de fluxos quan aquests són molt grans. Això el fa inapropiat quan es requereix una latència baixa. Per altra banda, Spark utilitza més recursos que altres entorns i, per tant, no és apropiat quan aquests han de ser compartits amb altres sistemes. Pel que fa al cost, es requereix més RAM. Malgrat això, el fet de dur a terme els processos en un temps més curt compensa aquests costos ja que pot disminuir el cost total en hores d'ús del clúster.

### 2.2.5. Apache Flink

Apache Flink[12] és un entorn de treball de processament de fluxos de dades amb la possibilitat de treballar en mode per lots com a un cas especial de processament de fluxos. Per altra banda, té la capacitat de guardar estats per a poder-se recuperar en cas de fallades i garanteix l'ordre dels esdeveniments que entren al sistema.

Així, el model de fluxos treballa amb fluxos il·limitats de dades que entren al sistema sempre amb la mateixa estructura. Aquests fluxos són processats amb funcions que generen nous fluxos. El resultat dels processos són enviat a una base de dades o a un altre sistema.

Pel que fa al model per lots, aquest només es diferencia de l'anterior en el fet que les dades són llegides d'un sistema d'emmagatzematge persistents en comptes de ser llegides des d'un flux de dades. Un cop les dades entren al sistema són tractades com en el model anterior.

Els principals avantatges d'aquest entorn són la baixa latència, l'alt rendiment, el fet de no necessitar determinades optimitzacions manuals que sí que requereix Spark, la disponibilitat de diversos mecanismes d'optimització de tasques, un visor web per gestionar les tasques, consultes similars a SQL, biblioteques d'aprenentatge automàtic, computació en memòria i la possibilitat d'executar processos desenvolupats per a altres sistemes com Hadoop o Storm.

El principal desavantatge d'aquest entorn és el fet de ser encara molt recent i menys consolidat que altres sistemes.

### 2.2.6. Elecció de l'entorn de treball *big data*

Un cop analitzats els principals entorns de treball *big data* i tenint en compte que per a la realització del projecte es necessita el processament de dades històriques, que el temps

[12] https://flink.apache.org/

d'anàlisi no és important i que es requereix un entorn amb un cost baix, el sistema més adient és Apache Hadoop ja que compleix aquests requisits. Efectivament, Hadoop permet treballar en el model de processament per lots, funciona bé amb clústers formats per maquinari de baix cost i és compatible amb altres sistemes.

La segona opció és Spark ja que també permet treballar en el model per lots i és més ràpid que Hadoop. Tanmateix, necessita maquinari amb uns requisits superiors i, per tant, més costosos. És per aquest motiu que s'ha utilitzat l'entorn Apache Hadoop per a la realització d'aquest projecte.

# 3. Informàtica en núvol

Com apunta Manyika (2011), una de les principals tecnologies emprades en el camp del *big data* és la informàtica en núvol (en anglès, *cloud computing*).

És per aquest motiu que tot seguit es descriu breument aquesta tecnologia així com els seus principals models de servei i de desplegament i algunes de les plataformes disponibles actualment.

## 3.1. Definició

Una de les definicions més conegudes de la informàtica en núvol és la del National Institute of Standards and Technology: "La informàtica en núvol és un model que permet l'accés en xarxa de forma ubiqua, adequada i sota demanda a un conjunt compartit de recursos informàtics configurables (p. ex., xarxes, servidors, emmagatzematge, aplicacions i serveis) que es poden proveir i publicar ràpidament amb un mínim esforç de gestió i interacció del proveïdor." (Mell, 2011, pàg. 2).

Les principals característiques de la informàtica en núvol són:

- *Autoservei sota demanda*. Els usuaris poden accedir als recursos informàtics que necessitin de forma automàtica sense la necessitat d'intervenció del proveïdor.

- *Major agilitat de les organitzacions*, ja que poden redissenyar els recursos de les seves infraestructures tecnològiques de forma més flexible.

- *Accés als recursos a través de la xarxa*. S'hi accedeix mitjançant estàndards que permeten l'ús de diversos tipus de clients, tant lleugers com pesats. D'aquesta forma, els recursos poden ser utilitzats sense tenir en compte el lloc ni el dispositiu que s'utilitzi. Per altra banda, també es simplifica el manteniment de les aplicacions d'informàtica en núvol ja que no cal que siguin instal·lades en tots els ordinadors.

- *Agrupació de recursos*. Els recursos que s'ofereixen són utilitzats per diversos clients. Aquests són assignats als clients de forma dinàmica en funció de les seves necessitats.

- *Elasticitat i escalabilitat*. Els recursos són subministrats i publicats de forma flexible segons la demanda i, sovint, de forma automàtica.

- *Monitorització dels serveis*.

- *Reducció de costos*. En molts casos suposa un reducció de costos tant pel que fa als recursos de la pròpia infraestructura tecnològica del client com per la menor necessitat de perfils amb coneixements en informàtica en núvol en la pròpia organització. Per altra banda, també pot significar un augment de la productivitat ja que diversos usuaris poden treballar simultàniament amb les mateixes dades.

- *Augment de la fiabilitat,* ja que sovint els recursos són redundants.

- *Seguretat*. Augmenta la seguretat pel fet que es pot focalitzar aquesta ja que les dades estan centralitzades. Per contra, es pot perdre el control de dades sensibles per a la organització.

## 3.2. Models de servei

Existeixen tres models de servei principals:

- **Infraestructura com a servei** (en anglès, *Infrastructure as a Service*, IaaS).

  En aquest model es proveeixen recursos de processament, emmagatzematge i xarxa, entre d'altres. Sovint aquests recursos s'ofereixen mitjançant hipervisors que executen màquines virtuals.

  D'aquesta forma, l'usuari no necessita conèixer els detalls de la infraestructura: recursos físics, localització, seguretat, còpies de seguretat, etc. i es pot centrar en el desplegament del seu programari, des del sistema operatiu fins a les aplicacions.

- **Plataforma com a servei** (en anglès, *Platform as a Service*, PaaS).

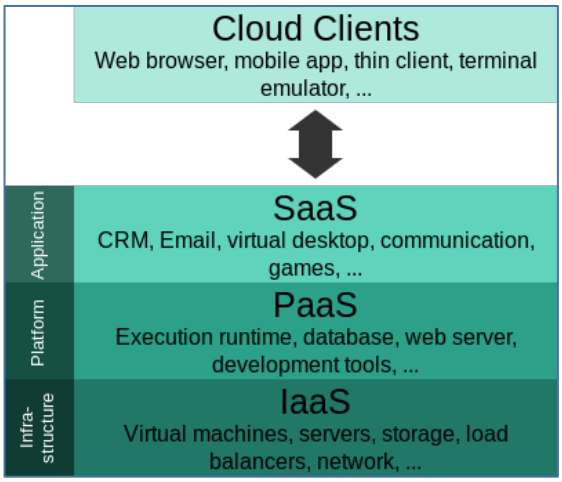
  En aquest model s'ofereixen entorns de desenvolupament on els usuaris poden desplegar aplicacions pròpies o bé de tercers tenint control sobre aquest desplegament i, sovint, sobre la configuració de l'entorn d'allotjament.

  A diferència de l'anterior model, els usuaris no tenen control sobre la infraestructura: xarxa, servidors, sistemes operatius, sistemes d'emmagatzematge, etc.

- **Programari com a servei** (en anglès, *Software as a Service*, SaaS).

  Aquest tercer model ofereix als usuaris aplicacions i bases de dades. A diferència dels dos models anteriors, l'usuari no gestiona ni la infraestructura ni la plataforma sobre les quals s'executen les aplicacions. Es pot accedir a aquests serveis a través de diversos clients lleugers com ara els navegadors web o bé clients pesats.

  El principal avantatge d'aquest model és l'escalabilitat de les aplicacions ja que les tasques es poden repartir entre diverses màquines virtuals. Per altra banda, les aplicacions poden ser actualitzades sense necessitat que l'usuari instal·li nou programari.




***Imatge 6**. Models de servei de la informàtica en núvol (font: Wikipedia).*

## 3.3. Models de desplegament

Existeixen quatre models de desplegament en la informàtica en núvol:

- **Núvol privat**. La infraestructura és utilitzada per una organització i pot estar mantinguda per aquesta mateixa organització o bé per tercers i pot estat allotjada als mateixos locals d'aquesta o fora d'ells.

- **Núvol públic**. La infraestructura és oberta a tot el públic i pot ser tant de pagament com gratuïta. Es troba allotjada en les instal·lacions del proveïdor.

- **Núvol comunitari**. La infraestructura és compartida per diverses organitzacions amb preocupacions similars. Aquesta infraestructura pot ser gestionada per una o més de les organitzacions que en formen part o bé per tercers i pot estar allotjada en les dependències de les organitzacions o fora d'aquestes.

- **Núvol híbrid**. En aquest tipus de desplegament es combinen dos o més núvols de diversos tipus: privat, públic o comunitari normalment de diversos proveïdors.

## 3.4. Principals plataformes d'informàtica en núvol

Tot seguit s'analitzen tres de les principals plataformes d'informàtica en núvol disponibles: Google Cloud Platform[13], Microsoft Azure[14] i Amazon Web Services. Les tres plataformes ofereixen solucions per a *big data* i disposen d'una capa de serveis gratuïts que permeten als desenvolupadors poder obtenir experiència en el seu ús.

### *Google Cloud Platform*

Google Cloud Platform és un servei d'informàtica en núvol que empra la mateixa infraestructura que utilitza Google per a oferir els seus principals serveis. Permet l'allotjament,

[13] https://cloud.google.com/solutions/big-data/
[14] https://azure.microsoft.com/en-us/solutions/big-data/

la computació, l'emmagatzematge i l'ús de diverses interfícies de programació d'aplicacions (en anglès, *application programming interface*, *API*).

Conté els models de servei IaaS com ara màquines virtuals (*Google Compute Engine*), emmagatzematge (*Bigtable*) i bases de dades (*BigQuery*) així com el servei PaaS per a l'allotjament d'aplicacions (*Google App Engine*).

Finalment, disposa de la solució *big data Cloud Dataproc* que permet gestionar serveis Apache Hadoop, Apache Spark, Apache Hive i Apache Pig[15].

### *Microsoft Azure*

Microsoft Azure és la plataforma d'informàtica en núvol de l'empresa Microsoft. Aquesta proporciona els tres models de servei: IaaS, PaaS i SaaS.

Per altra banda, disposa del servei *Microsoft Azure HDInsight* que permet utilitzar clústers Apache Hadoop.

### *Amazon Web Services*

Amazon Web Services són els serveis d'informàtica en núvol de l'empresa Amazon: *Amazon Elastic Compute Cloud* (EC2) i *Amazon Simple Storage Service* (S3) els quals ofereixen serveis de: emmagatzematge, computació, xarxa, bases de dades, aplicacions, desplegament, gestió i desenvolupament, entre d'altres.

Amazon Web Services disposa dels servei *Amazon Elastic MapReduce* (EMR), un servei PaaS que ofereix clústers Apache Hadoop.

Un cop analitzades i testejades les tres plataformes s'ha optat per l'ús de l'Amazon Web Services per a la implementació de la solució *big data* en un entorn distribuït degut al ventall de serveis que ofereix de forma gratuïta i al fet de poder utilitzar aquesta capa gratuïta durant dotze mesos, així com la nombrosa documentació disponible a la plataforma per a la configuració dels diversos serveis.

Tot i que s'ha configurat un clúster Hadoop en la plataforma Microsoft Azure utilitzant el servei HDInsight de forma molt còmoda i ràpida s'ha descartat aquesta opció perquè es cercava un control total de les configuracions del clúster Hadoop així com de les eines utilitzades per a l'obtenció del model d'aprenentatge. Per altra banda, també s'ha tingut en compte la limitació a tan sols trenta dies en l'ús de la capa gratuïta de la plataforma.

Finalment, s'ha descartat Google Cloud ja que Amazon Web Services ha permès més flexibilitat alhora de configurar el clúster Hadoop.

[15] https://pig.apache.org/

# 4. Apache Hadoop

En l'apartat 2.2 s'han analitzat els principals entorns de treball *big data* i s'ha optat per Apache Hadoop en considerar-lo el més adient per a la realització del projecte.

Apache Hadoop és un entorn de treball (en anglès, *framework*) de codi obert que permet el processament distribuït de grans volums de dades a través d'un clúster d'ordinadors possibilitant la resolució de problemes *big data* sense la necessitat d'utilitzar maquinari d'altes prestacions.

Les principals característiques d'Apache Hadoop són la seva escalabilitat, l'alta disponibilitat gràcies a la capacitat de detectar i gestiona fallades en el clúster, l'ús del model de programació MapReduce per al processament distribuït de les dades i el sistema d'arxius propi Hadoop Distributed Filesystem (HDFS).

Més enllà de MapReduce i HDFS, existeix un ampli ecosistema de projectes al voltant de Hadoop, entre els quals podem destacar:

- ***YARN,*** sistema de gestió dels recursos del clúster i monitoratge i planificació de tasques. Permet l'execució en un clúster Hadoop de tot tipus de programes distribuïts més enllà de MapReduce.

- ***HBase***[16], base de dades distribuïda orientada a columnes del tipus clau-valor que funciona sobre el sistema de fitxers propi de Hadoop (HDFS).

- ***Impala***[17] i ***Hive***[18]***,*** sistemes que permeten realitzar consultes distribuïdes SQL en un clúster Hadoop.

- ***Spark***, sistema que permet el processament interactiu que no ofereix MapReduce (vegeu apartat 2.2.4).

- ***Storm***, ***Spark Streaming*** i ***Samza***, són sistemes de processament de fluxos que permeten l'execució de processaments distribuïts de fluxos il·limitats en temps real (vegeu apartat 2.2).

- ***Solr***[19], plataforma d'indexació i cerca dels documents que són emmagatzemats en HDFS.

Tot seguit es descriuen amb més detall els dos components fonamentals d'Apache Hadoop: MapReduce i Hadoop Distributed Filesystem. Finalment, s'analitzen les principals plataformes de tractament de dades distribuïdes de l'ecosistema Hadoop.

[16] https://hbase.apache.org/
[17] https://impala.incubator.apache.org/
[18] https://hive.apache.org/
[19] http://lucene.apache.org/solr/

## 4.1. MapReduce

MapReduce és un model de programació per al processament de dades que utilitza computació paral·lela[20], és a dir, divideix problemes grans en un conjunt de problemes més petits que són processats simultàniament.

Hadoop distribueix el processament dels programes MapReduce als diversos nodes d'un clúster que emmagatzemen les dades que han de ser processades mitjançant el sistema de gestió de recursos YARN.

D'aquesta forma, MapReduce permet realitzar processament per lots sobre grans volums de dades obtenint resultats en un temps raonable. Tot i així, no es pot utilitzar per a resoldre tot tipus de problema, sinó només aquells que són paral·lelitzables. Concretament, no respon bé a les necessitats d'anàlisi interactiu.

MapReduce descompon el processament en dues fase: *map* i *reduce*, les quals utilitzen dades estructurades en forma de clau i valor. Per a dur-ho a terme, s'usen dos tipus de nodes anomenats màster i *workers*. El node màster és l'encarregat d'assignar les diverses tasques *map* o bé *reduce* a la resta de nodes, els nodes *workers*.

La fase *map* és la responsable de filtrar i ordenar les dades. Per a fer-ho, les divideix en un conjunt de particions que són enviades a diverses màquines del clúster les quals processen en paral·lel els elements de l'entrada de dades retornant un llistat de parells clau-valor que posteriorment són agrupats per clau. D'aquesta forma, el processament es pot dividir entre els diversos nodes del clúster.

$$Map(k1, v1) \rightarrow list(k2, v2)$$

Un cop obtingut el llistat de parells clau-valor, el node màster s'encarrega d'assignar als nodes disponibles la corresponent fase *reduce*, que també s'aplica en paral·lel en cadascun dels grups obtinguts en la fase *map* generant un valor de sortida. En primer lloc, ordena les claus per així agrupar les dades que comparteixen clau. Tot seguit, itera sobre cadascuna de les claus i hi executa la funció *reduce*. Per tant, la sortida de la funció *reduce* és un llistat de valors.

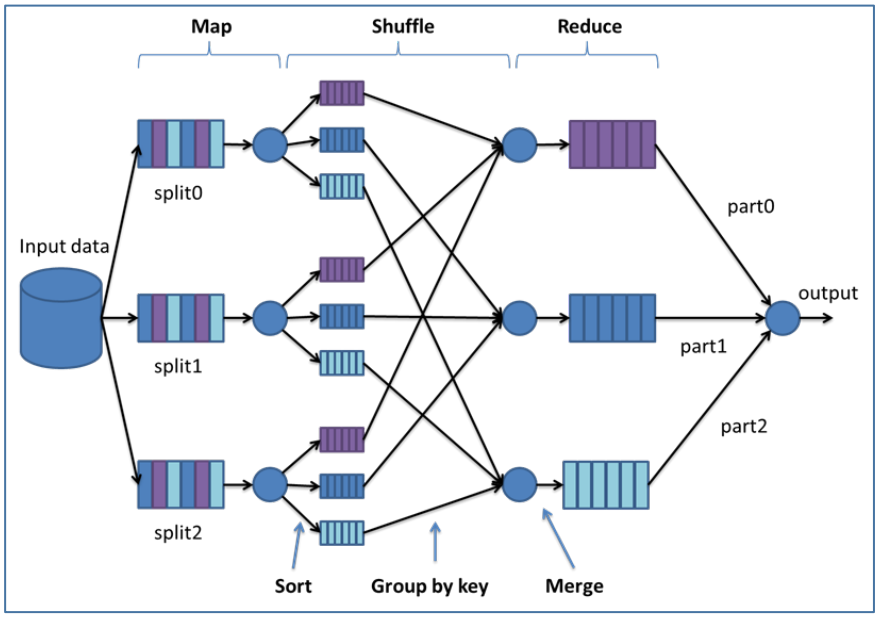
$$Reduce\ \big(k2, list(v2)\big) \rightarrow list(v3)$$

El resultat és emmagatzemat en un fitxer si ja s'ha obtingut el resulta final o bé en tants fitxers com funcions *reduce* executades els quals seran l'entrada a un altre programa MapReduce.

Una de les limitacions de MapReduce es dóna en el transport de grans fitxers entre els diversos nodes del clúster. És per això que, un cop guardat el resultat de la funció *map* a la memòria intermèdia local del node que l'ha executada, s'aplica una funció d'agregació per agregar i ordenar el llistat de valors obtingut i, d'aquesta forma, reduir el transport de grans fitxers. Aquesta funció d'agregació també es pot incloure a la pròpia funció *map*.

La següent imatge mostra el funcionament del model de programació MapReduce.

[20] https://ca.wikipedia.org/wiki/Computació_paral·lela




***Imatge 7**. Funcionament del model de programació MapReduce (font: Universidad de Granada).*

Una de les principals característiques de MapReduce és que és tolerant a fallades. Aquest tret és important ja que MapReduce ha estat dissenyat per executar-se en centenar i, fins i tot, milers de nodes. Això augmenta la possibilitat que algun dels nodes *worker* involucrats en el processament falli durant l'execució.

Per a garantir la tolerància a fallades el node màster utilitza la comanda ping[21] per a verificar l'estat dels diversos nodes *worker*. Si un node no respon i estava executant una tasca *map* o bé *reduce* aquesta és assignada a un altre node *worker*.

Tot i que Apache Hadoop és la implementació més coneguda del model de programació MapReduce n'existeixen d'altres com les bases de dades NoSQL: Couchbase[22], CouchDB[23], Infinispan[24], MongoDB[25] i Riak[26].

## 4.2. Hadoop Distributed Filesystem

Hadoop Distributed Filesystem (HDFS) és el sistema de fitxers distribuït d'Apache Hadoop, el qual s'integra amb altres sistemes d'emmagatzematge com el sistema de fitxers local o Amazon S3.

La principal característica d'HDFS és que aquest és tolerant a la fallada de nodes, la principal problemàtica dels sistemes de fitxers distribuïts.

[21] Protocol de xarxa que permet verificar si un dispositiu d'una xarxa és accessible.
[22] https://www.couchbase.com/
[23] http://couchdb.apache.org/
[24] http://infinispan.org/
[25] https://www.mongodb.com/
[26] http://basho.com/products/#riak

Per altra banda, ha estat dissenyat per emmagatzemar arxius molt grans. De fet, existeixen clústers Hadoop que emmagatzemen petabytes de dades. HDFS es mostra especialment eficient quan el conjunt de dades són escrites un cop i analitzades diversos cops.

Per contra, HDFS no dóna un bon resultat en altres entorn, com quan existeixen grans quantitats d'arxius petits o bé quan es requereix baixa latència en l'accés a les dades.

Per assolir les anteriors característiques HDFS divideix els arxius en diversos fragments de la mida d'un bloc, normalment de 128MB, i els emmagatzema en nodes independents. Això permet que un arxiu pugui ser més gran que l'espai d'emmagatzematge d'un dels nodes del clúster. Tot seguit, i per a garantir la tolerància a les fallades i la disponibilitat de les dades, cada bloc es replica en diverses nodes físicament separats; normalment s'utilitzen tres rèpliques.

El sistema d'arxius HDFS empra dos tipus de nodes: un *namenode*, node màster que gestiona l'espai de noms del sistema d'arxius, gestiona les metadades dels arxius i directoris i coneix en quins nodes es troben els blocs d'un determinat arxiu, i diversos *datanodes*, nodes esclaus que emmagatzemen els blocs.

## 4.3. Principals plataformes de tractament de dades distribuïdes de l'ecosistema Hadoop

L'ecosistema Apache Hadoop disposa de diverses plataformes d'alt nivell que permeten el tractament de grans conjunts de dades distribuïdes sense necessitat d'utilitzar directament llenguatges com MapReduce o Spark. Entre aquests, se'n poden destacar tres: Apache Hive, Apache Pig i Apache Impala.

### *Apache Hive*

Apache Hive és una aplicació de l'ecosistema Hadoop que permet el tractament de grans conjunts de dades emmagatzemades en sistemes de fitxers distribuïts utilitzant consultes HiveQL, llenguatge basat en SQL[27]. Disposa tant d'una eina de línia d'ordres per a realitzar les consultes com d'un controlador JDBC i suporta només dades estructurades. Hive converteix les consultes escrites en el llenguatge declaratiu HiveQL a MapReduce, Apache Tez[28] o bé en tasques Spark.

### *Apache Pig*

Per altra banda, Apache Pig utilitza el llenguatge Pig Latin, un llengutage de consultes similar a SQL tot i que amb majors diferències respecte aquest que HiveQL. Com en el cas de Hive, aquestes consultes són traduïdes a MapReduce, Tez o bé Spark. A diferència de HiveQL, Pig Latin no és un llenguatge declaratiu sinó procedimental. Per altra banda, suporta tant dades estructurades com no estructurades i, per tant, és l'elecció idònia per al processament de fluxos de dades.

[27] https://ca.wikipedia.org/wiki/Structured_Query_Language
[28] https://tez.apache.org/

### *Apache Impala*

Finalment, Apache Impala utilitza la sintaxi HiveQL per a realitzar consultes en al processament de grans conjunts de dades en paral·lel. A diferència de les dues plataformes anteriors que poden funcionar sobre qualsevol sistema operatiu que suporti la màquina virtual Java, Impala només funciona sobre el sistema operatiu Linux. Per altra banda, Impala realitza les consultes en memòria i no les tradueix prèviament a altres llenguatges com MapReduce.

Així, s'ha optat per l'ús d'Apache Hive ja que la sintaxi de consultes que utilitza és molt similar a SQL, està pensat per a processar dades estructurades com són les tractades en aquest projecte i funciona sobre qualsevol sistema operatiu que suporti la màquina virtual Java.

Per tant, l'ús d'aquesta plataforma permetrà dur a terme el tractament inicial de les dades de forma més ràpida sense necessitat d'implementar-lo en MapReduce.

# 5. Aprenentatge automàtic

Un cop descrits els paradigmes *big data* i de la informàtica en núvol i valorades les plataformes més adients per a la realització del projecte, en aquest capítol es descriu el camp de l'aprenentatge automàtic. Tot seguit, es detallen les principals eines d'aprenentatge disponibles en l'entorn de *big data* escollit, Apache Hadoop, i se n'analitza la seva adequació al problema que es vol resoldre en aquest projecte.

## 5.1. Definició

L'aprenentatge automàtic (en anglès, *machine learning*) és un camp de la intel·ligència artificial que estudia tècniques per a proveir a les màquines de la capacitat d'aprendre i millorar a partir de l'experiència.

Es poden classificar aquetes tècniques en els següents tipus:

- ***Aprenentatge supervisat***. Aquest tipus d'aprenentatge es dóna quan es coneix quina és la resposta del sistema. Existeixen tres tipus de problemes d'aprenentatge supervisat:

    - *Problemes de regressió*. En aquest tipus de problemes es disposa d'un atribut solució de tipus numèric i l'objectiu del sistema és reproduir-lo.
    - *Problemes de classificació*. A diferència dels problemes de regressió, en aquest cas l'atribut solució és binari o bé categòric.
    - *Problemes de cerca*. S'utilitzen per a la resolució de problemes aplicant algorismes de cerca.

- ***Aprenentatge per reforç***. A diferència de les tècniques anteriors, no es coneix la sortida del sistema sinó només una gratificació o penalització en funció de la resposta d'aquest.

- ***Aprenentatge no supervisat***. No es disposa de cap tipus d'informació sobre les sortides. Aquests sistemes extreuen el coneixement de la informació d'entrada disponible.

Tot seguit es defineixen breument els algorismes no supervisats de categorització i els algorismes supervisats de classificació, centrant-se en aquests darrers ja que són els que s'utilitzaran per a l'obtenció del model predictiu de contaminació.

### 5.1.1. Algorismes de categorització

Els algorismes de categorització (en anglès, *clustering*) són mètodes d'aprenentatge no supervisat capaços d'obtenir un grup de categories a partir d'un conjunt d'objectes inicials. Per aconseguir-ho es divideix el conjunt inicial en subconjunts d'objectes amb característiques semblants entre si i diferents dels objectes de la resta de categories. Els atributs dels objectes poden ser dels següents tipus: numèrics, booleans, ordinals o bé nominals.

Pel que fa a la mesura de la semblança dels objectes s'utilitzen diversos coeficients de semblança:

- *Basats en distàncies*. Les distàncies poden ser euclidianes o be de Manhattan[29], entre d'altres.
- *Basats en associacions*.
- *Basats en correlacions*.
- *Basats en la semblança probabilística*.

### 5.1.2. Algorismes de classificació

Els algorismes de classificació són mètodes d'aprenentatge supervisat on es disposa d'un conjunt d'objectes dels quals es coneix el valor de sortida. És a dir, es coneix la classe a la qual pertany cada objecte.

D'aquesta forma, a partir d'un conjunt d'entrenament del qual es coneix la classe a la qual pertany cada objecte, l'algorisme de classificació permet classificar nous objectes dels quals no se'n coneix la classe.

Tot seguit es descriuen els principals mètodes de classificació i s'analitza llur adequació al problema que es vol resoldre en aquest projecte.

#### *Veí més proper*

El mètode del veí més proper, i la seva generalització k veïns més propers[30] (en anglès, *k-nearest neighbour*, *k-NN*), és un mètode basat en els algorismes de categorització que classifica un objecte a partir de l'objecte (o objectes) més proper. Cal doncs, definir la semblança entre els objectes del domini.

Aquest algorisme és vàlid per a l'obtenció del model d'aprenentatge que s'està implementant. Altres mètodes similars també basats en mètodes de categorització i que s'ajusten a les necessitats del projecte són: la classificació basada en la quantificació vectorial[31] (en anglès, *vector quantization*) i la classificació basada en K-mitjanes[32] (en anglès, *K-means*).

#### *Màquina de vectors de suport*

El mètode de classificació màquina de vectors de suport[33] (en anglès, *suport vector machines*, *SVM*) permet la classificació d'objectes amb atributs numèrics en dues classes. Aquest mètode s'aplica en problemes de reconeixement de patrons i de sèries temporals, entre d'altres.

Degut a què el número de classes de contaminació que es tindran en compte seran superiors a dos, aquest no és un mètode apropiat per a la implementació del model de predicació de la contaminació atmosfèrica.

[29] https://ca.wikipedia.org/wiki/Geometria_del_taxista
[30] https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm
[31] https://en.wikipedia.org/wiki/Learning_vector_quantization
[32] https://en.wikipedia.org/wiki/K-means_clustering
[33] https://en.wikipedia.org/wiki/Support_vector_machine

### *Arbres de decisió i* Random forest

En el mètode dels arbres de decisió[34] es construeix un arbre on en cada node s'avalua un concepte i en cada fulla s'hi associa una classe. D'aquesta forma, cada nou objecte s'avaluarà seguint els diversos nodes fins a una fulla que el classificarà.

Una variant dels arbres de decisió és el mètode *Random forest*[35]. Aquest algorisme utilitza l'agregació de bootstrap[36] per a millorar l'estabilitat i la precisió de la classificació així com reduir la variància i evitar el sobreajustament. Per altra banda, s'empra la selecció aleatòria d'atributs per a obtenir un conjunt d'arbres de decisió.

Tenint en compte que els arbres de decisió acostumen a estar influïts per soroll, el que pretén el mètode *Random forest* és obtenir la mitjana d'aquests. Per aconseguir-ho, s'utilitza un subconjunt dels casos com a conjunt d'entrenament i la resta de casos com a conjunt de prova per a estimar l'error de l'arbre. Per a cada node, s'empra un subconjunt de les variables d'entrada escollides aleatòriament per determinar la decisió d'aquest. Un cop obtinguts els diversos arbres utilitzant aquesta metodologia es classifica cada cas nou en tots els arbres i s'assigna com a predicció la classe que té una quantitat major d'incidències.

Tant els arbres de decisió com el mètode *Random forest* s'adapten bé a les necessitats del d'aquest projecte.

### *Xarxes neuronals*

Les xarxes neuronals[37] s'inspiren en les xarxes de neurones dels éssers vius. Així, se simulen les propietats dels sistemes neuronals mitjançant models matemàtics. Aquests sistemes tenen la capacitat d'aprendre i són flexibles i tolerants a les fallades.

Per tant, les característiques d'aquests mètodes de classificació els fan útils per al problema que es vol resoldre.

### *Classificadors lineals*

Els classificadors lineals[38] seleccionen la classe d'un determinat objecte a partir del valor de les combinacions lineals de les seves característiques.

Alguns dels classificadors lineals més coneguts són la regressió logística[39] i el classificador bayesià ingenu[40].

La regressió logística és un anàlisi de regressió que permet predir el resultat d'una variable categòrica a partir d'un conjunt de variables predictores, les quals poden ser categòriques o bé numèriques. Aquest algorisme permet modelar la probabilitat de què un determinat esdeveniment tingui lloc en funció d'un conjunt de factors. Per tant, aquest mètode pot ser utilitzat per a la implementació del model d'aprenentatge d'aquest projecte.

[34] https://en.wikipedia.org/wiki/Decision_tree_learning
[35] https://en.wikipedia.org/wiki/Random_forest
[36] https://en.wikipedia.org/wiki/Bootstrap_aggregating
[37] https://ca.wikipedia.org/wiki/Xarxa_neuronal_artificial
[38] https://en.wikipedia.org/wiki/Linear_classifier
[39] https://en.wikipedia.org/wiki/Logistic_regression
[40] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

Per altra banda, el classificador probabilístic bayesià ingenu, basat en el teorema de Bayes, assumeix que la presència o absència d'una determinada característica no està relacionada amb la presència o absència d'una altra característica. El principal avantatge d'aquest mètode és que permet classificar amb poques dades d'entrada. Aquest mètode s'utilitza principalment per a la categorització de text i no és el més indicat per al problema que es vol resoldre on els atributs amb els quals es treballarà són numèrics.

## 5.2. Principals eines d'aprenentatge en l'ecosistema Apache Hadoop

Com descriu Landset (2015) els principals conjunts d'eines d'aprenentatge en l'ecosistema Apache Hadoop són: Mahout, MLlib[41], $H_2O$[42] i SAMOA[43].

Tot seguit se'n descriuen les principals característiques i s'analitza quina és la més adient per a la implementació dels models predictius de contaminació atmosfèrica.

### 5.2.1. Apache Mahout

Apache Mahout[44] és una biblioteca de codi obert d'algorismes d'aprenentatge automàtic escalables escrita en Java que pot ser utilitzada quan el volum de dades que cal processar és molt gran.

És una de les biblioteques més completes i conegudes d'aprenentatge. Des de la seva versió 0.10.0 permet als usuaris desenvolupar els seus propis algorismes distribuïts.

Mahout està format per un extens conjunt d'algorismes per a MapReduce. Molts d'aquests també han estat implementats en Spark, $H_2O$ i Flink. Els principals algorismes es centren en els sistemes de recomanació, la categorització i la classificació. També disposa d'eines complementàries per a la reducció de la dimensionalitat, la vectorització de text i les mesures de similitud, entre d'altres. Per altra banda, disposa d'un entorn d'experimentació amb sintaxi similar a R.

La següent taula mostra els algorismes de sistemes de recomanació, categorització i classificació suportats per Mahout tot indicant si es troben implementats en una sola màquina, en MapReduce o bé en Apache Spark.

[41] http://spark.apache.org/mllib/
[42] https://www.h2o.ai/
[43] https://samoa.incubator.apache.org/
[44] http://mahout.apache.org/

| Tipus | Algorisme | Màquina única | MapReduce | Spark |
|---|---|---|---|---|
| **Sistemes de recomanació** | Filtre col·laboratiu basat en usuaris | ✔ | ✔ | ✔ |
| | Filtre col·laboratiu basat en ítems | ✔ | ✔ | ✔ |
| | Descomposició de matrius amb ALS | ✔ | ✔ | ✖ |
| | Descomposició de matrius amb ALS sobre retroalimentació implícita | ✔ | ✔ | ✖ |
| | Descomposició de matrius ponderada, SVD++ | ✔ | ✖ | ✖ |
| **Categorització** | Canopy | ✔ | ✔ | ✖ |
| | K-mitjanes | ✔ | ✔ | ✖ |
| | K-mitjanes difuses | ✔ | ✔ | ✖ |
| | Streaming k-Means | ✔ | ✔ | ✖ |
| | Espectral | ✖ | ✔ | ✖ |
| **Classificació** | Regressió logística (entrenada per mitjà d'SGD) | ✔ | ✖ | ✖ |
| | Bayesià ingenu / Bayesià ingenu complementari | ✖ | ✔ | ✔ |
| | Model ocult de Markov | ✔ | ✖ | ✖ |
| | Random forest | ✖ | ✔ | ✖ |

***Taula 7**. Principals algorismes d'aprenentatge automàtic de la biblioteca Apache Mahout.*

### 5.2.2. MLlib

MLlib és un entorn de treball que, com en el cas de la biblioteca Mahout, disposa d'algorismes de categorització, classificació i sistemes de recomanació i, a més, de models de regressió. Per altra banda, funciona sobre Spark tant en el mode de processament per lots com en fluxos de dades.

Com a principal inconvenient es pot destacar que, a diferència de Mahout, es tracta d'un projecte poc documentat i encara poc avaluat.

### 5.2.3. $H_2O$

Tot i que $H_2O$ disposa d'una edició empresarial, també és una eina de codi obert amb la característica de disposar d'una interfície gràfica d'usuari. Malgrat que existeixen altres eines d'aprenentatge amb IGU com Weka, RapidMiner i KNIME, aquestes no estan pensades per entorns *big data*.

Per altra banda, $H_2O$ implementa diverses eines per a xarxes neuronals profundes, pot ser programat en Java, Python, R i Scala i disposa del seu propi motor de processament tot i que també es pot integrar amb altres motors com Spark i Storm.

### 5.2.4. SAMOA

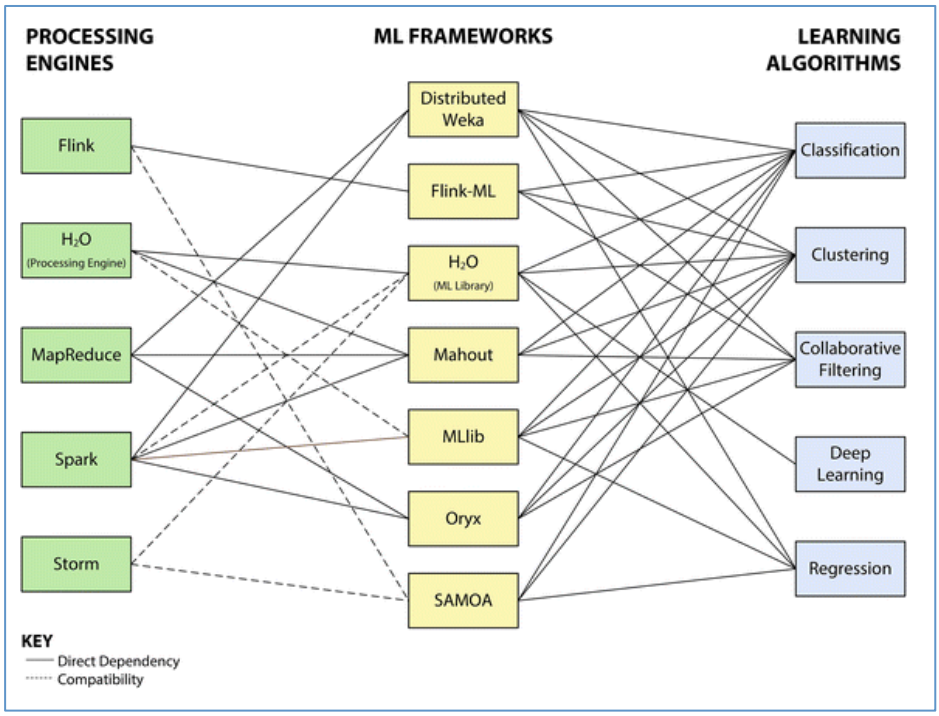
SAMOA és una plataforma dissenyada per a l'aprenentatge sobre dades distribuïdes en temps real. Disposa d'algorismes de classificació, categorització, regressió, mineria de patrons i *boosting*[45] entre d'altres.

Es tracta d'un entorn de treball pensat per a grans volums de dades que s'actualitzen constantment i on cal obtenir els resultats de l'anàlisi a temps real.

[45] https://en.wikipedia.org/wiki/Boosting_(machine_learning)

### 5.2.5. Comparació de les eines d'aprenentatge i elecció de la més adient

La següent imatge mostra la relació entre els diversos motors de processament en entorns *big data*, els principals entorns de treball d'aprenentatge automàtic distribuït i els algorismes d'aprenentatge que aquests implementen.




***Imatge 8**. Relació entre entorns de treball d'aprenentatge, els motors de processament i els algorismes d'aprenentatge (font: Landset).*

Com es pot observar, poden existir diversos motius per escollir l'entorn de treball adient per a un determinat projecte. Tenint en compte que es vol obtenir un model, s'optarà per una eina per lots descartant d'entrada SAMOA ja que és una eina pensada per al processament en temps real. Per altra banda, s'ha escollit MapReduce com a model de programació i, entre les quatre eines analitzades, només Mahout l'ofereix.

Així, tenint en compte l'exposat i que Mahout és una entorn altament escalable i extensible que cobreix un ampli ventall d'algorismes d'aprenentatge amb un nivell d'usabilitat acceptable i que només té com a punt desfavorable la velocitat d'execució i aquest no és un dels requisits del projecte, s'ha optat per aquesta biblioteca per a l'obtenció dels models d'aprenentatge.

# 6. Obtenció de models predictius de contaminació atmosfèrica

Un cop analitzat l'ecosistema Apache Hadoop així com els paradigmes del *big data*, la informàtica en núvol i l'aprenentatge automàtic s'ha iniciat el procés d'obtenció del model predictiu de contaminació atmosfèrica.

En aquest capítol es descriu el problema que es vol resoldre i, a continuació, es detalla el procés seguit: l'anàlisi previ de les dades, l'elecció dels algorismes d'aprenentatge, la preparació de l'arquitectura *big data*, la càrrega i el tractament de les dades en un clúster Hadoop i l'obtenció del model d'aprenentatge. Finalment, s'analitzen els resultats obtinguts així com el rendiment del sistema.

## 6.1. Presentació del problema

La finalitat d'aquest projecte és l'ús d'algorismes d'aprenentatge automàtic en un entorn *big data* per a l'obtenció d'un model que permeti preveure l'índex de contaminació atmosfèrica per diòxid de nitrogen a partir de dades d'intensitat de trànsit i meteorològiques, ja que són dos dels factors principals que poden desencadenar episodis de contaminació del pol·luent analitzat.

A partir de les dades meteorològiques, de trànsit i de contaminació corresponents a l'any 2016 s'ha entrenat un classificador per a obtenir el model de predicció. Tenint en compte que l'evolució dels fenòmens de contaminació és similar tots els anys, el model obtingut per a l'any 2016 podria ser aplicat per a predir episodis de contaminació en altres anys.

El territori en el qual s'ha aplicat el model és el Principat d'Andorra i les dades provenen de varis sensors distribuïts pel territori gestionats per diversos organismes oficials.

## 6.2. Estudi previ de les dades

Amb la finalitat de comprendre les principals característiques de les dades utilitzades per a obtenir el model d'aprenentatge, s'han analitzat els diversos conjunts de dades disponibles: meteorològiques, de trànsit i de contaminació.

### 6.2.1. Dades meteorològiques

L'Oficina de l'Energia i del Canvi Climàtic del Govern d'Andorra disposa de 17 estacions meteorològiques en servei tot l'any i 6 en servei només a l'hivern. D'aquest conjunt d'estacions se n'utilitzen 15 per recollir les dades que permeten obtenir els models meteorològics.

S'han utilitzat les dades de l'estació meteorològica Roc de Sant Pere per a l'obtenció del model d'aprenentatge automàtic ja que és la més propera a l'estació de mesura de la qualitat de l'aire i, per tant, és l'estació que pot ajudar a modelar amb més precisió la influència de les condicions meteorològiques en els episodis de contaminació atmosfèrica. Les dades que recull

aquesta estació són: temperatura, temperatura de rosada, precipitació, humitat relativa, direcció del vent, força del vent, irradiació global i insolació.

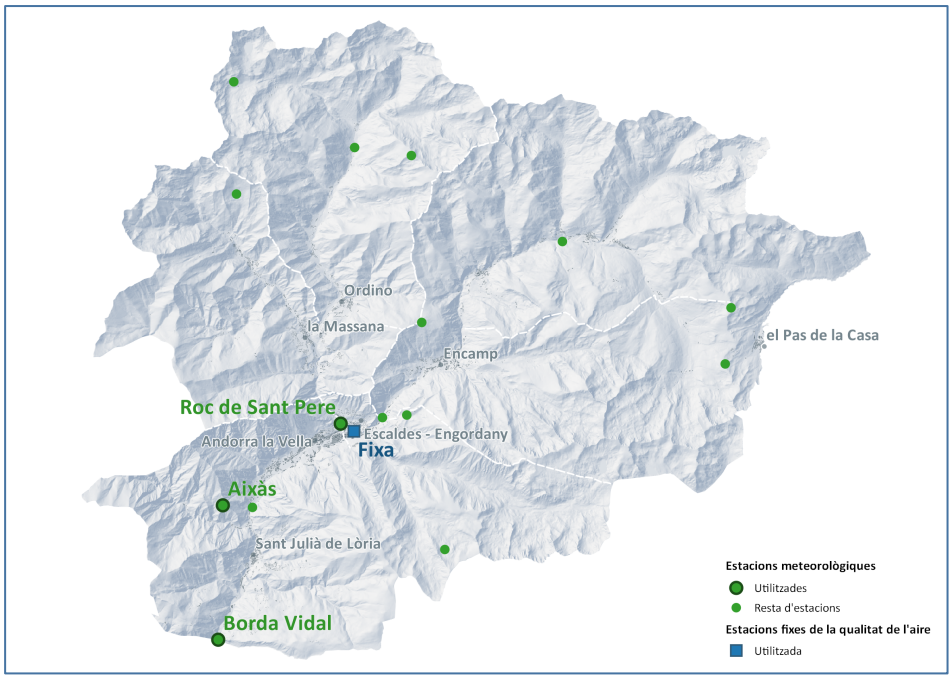
Cal tenir en compte que un dels principals desencadenants de l'augment de la concentració de contaminants és la inversió tèrmica. És per això que s'han utilitzat les dades de temperatura recollides per dues estacions pròximes ubicades, una, al fons d'una vall i, l'altra, a alta muntanya. Concretament s'ha emprat l'estació Borda Vidal ubicada al fons d'una vall i l'estació Aixàs situada a alta muntanya. El desnivell entre ambdues és d'aproximadament 690 metres.

Aquestes localitzacions permeten coneixes el valor d'inversió tèrmica a partir de la següent fórmula:

$$Inversió = T_{Aixàs} - T_{Borda\ Vidal}$$

Per tant, els valors positius indicaran que existeix inversió tèrmica.

El següent mapa mostra la distribució de les estacions meteorològiques així com la ubicació de les que han estat utilitzades en el projecte i la situació de l'estació fixa de mesura de la qualitat de l'aire.




***Imatge 9**. Distribució de les estacions meteorològiques.*

La següent taula mostra les estacions meteorològiques utilitzades així com les seves coordenades geogràfiques i altituds aproximades, la freqüència de lectura i la finalitat per a la qual s'han utilitzat en el projecte.

| Estació | Finalitat | Coordenades | | | Freqüència de mesura |
|---|---|---|---|---|---|
| | | *Longitud* | *Latitud* | *Altitud* | |
| Roc de Sant Pere | Dades meteorològiques | 1°31'56" | 42°30'47" | 1.113 m | Diària, horària i 6-minutal |
| Aixàs | Inversió tèrmica | 1°28'38" | 42°29'01" | 1.564 m | Minutal |
| Borda Vidal | Inversió tèrmica | 1°28'29" | 42°26'23" | 873 m | Horària i 6-minutal |

***Taula 8**. Finalitat, coordenades geogràfiques i freqüència de mesura de les estacions meteorològiques utilitzades.*

Les dades meteorològiques es corresponen a tot l'any 2016. Els registres recullen els següents atributs:

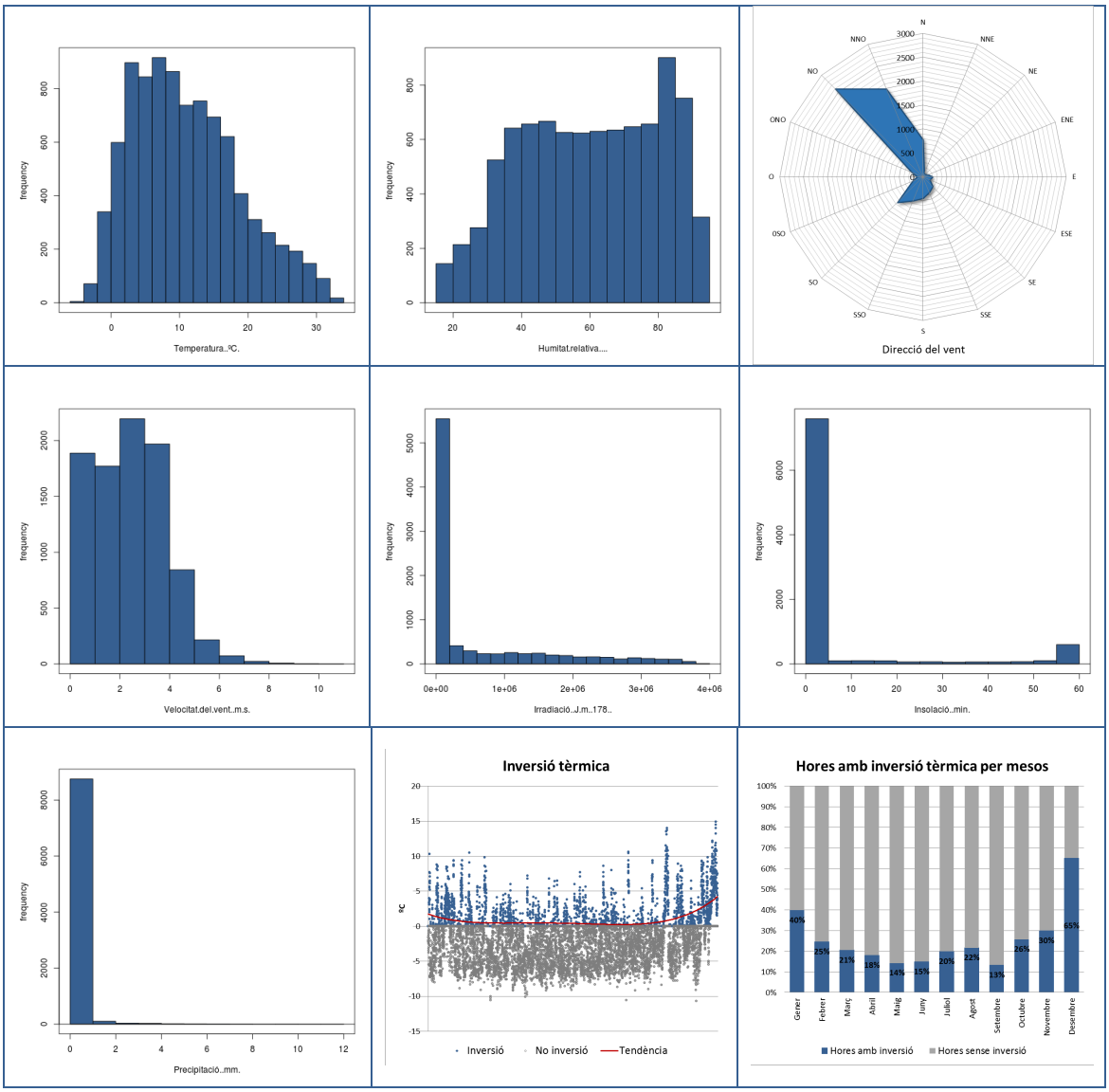
- *Data*.
- *Hora*.
- *Temperatura*. Unitat de mesura: ℃.
- *Humitat relativa*. Unitat de mesura: %.
- *Direcció del vent*. Unitat de mesura: graus.
- *Velocitat del vent*. Unitat de mesura: $m/s$.
- *Irradiació*. Unitat de mesura: $J/m^2$.
- *Insolació*. Unitat de mesura: minuts.
- *Precipitació*. Unitat de mesura: mm.
- *Inversió tèrmica*. Unitat de mesura: ℃.

La majoria de les estacions meteorològiques recullen dades cada minut o cada 6 minuts. Tanmateix, s'han utilitzat dades horàries per a l'obtenció del model d'aprenentatge. La següent taula mostra els principals valors estadístics per a cadascun dels atributs numèrics.

| Estadístic | Temper. (℃) | Humitat relativa (%) | Direcció del vent (graus) | Velocitat del vent ($m/s$) | Irradiació ($J/m^2$) | Insolació (minuts) | Precipit. (mm) | Inversió tèrmica (℃) |
|---|---|---|---|---|---|---|---|---|
| Mínim | -4,20 | 15,0 | 0 | 0,00 | 0 | 0,00 | 0,00 | -10.7 |
| 1r quantil | 4,90 | 44,0 | 200 | 2,00 | 0 | 0,00 | 0,00 | -4.6 |
| Mediana | 10,00 | 61,0 | 310 | 3,00 | 10000 | 0,00 | 0,00 | -2.1 |
| Mitjana | 11,02 | 60,4 | 255 | 2,87 | 618074 | 6,49 | 0,09 | -1.97 |
| 3r quantil | 16,10 | 79,0 | 330 | 4,00 | 1010000 | 0,00 | 0,00 | -0.19 |
| Màxim | 33,70 | 95,0 | 360 | 11,00 | 3960000 | 60 | 11,40 | 14.9 |
| Registres sense valor | 1 | 77 | 52 | 0 | 15 | 10 | 16 | 32 |

***Taula 9**. Principals valors estadístics dels atributs de les dades meteorològiques.*

Els següents diagrames mostren la distribució de cadascun dels anteriors atributs.




***Imatge 10**. Distribució dels valors dels atributs numèrics.*

## 6.2.2. Dades de trànsit

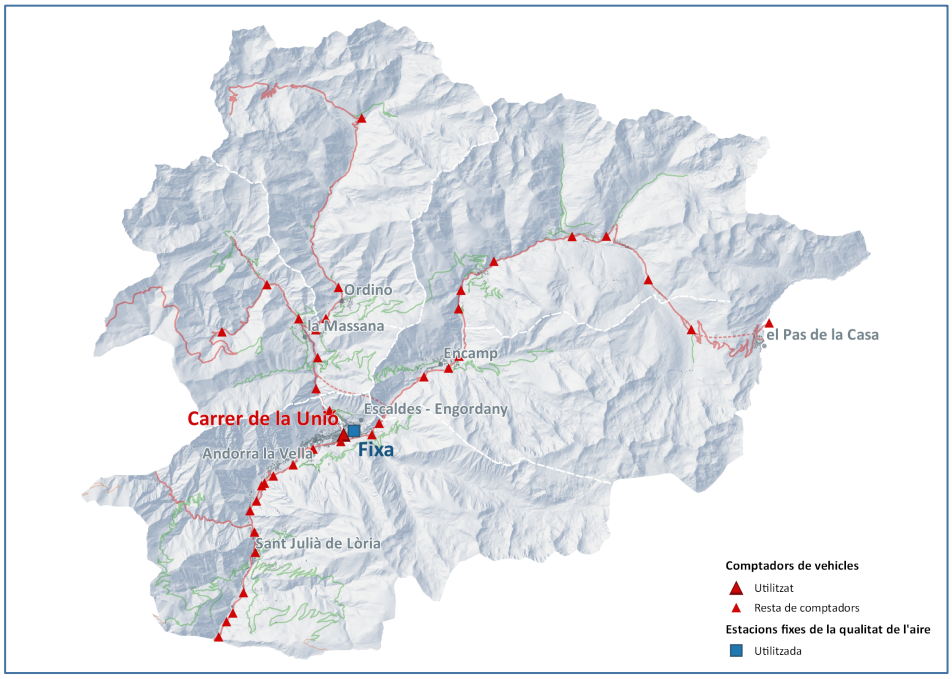
L'Àrea de Mobilitat del Govern d'Andorra disposa de 39 punts de comptatge de vehicles, els quals obtenen el número de vehicles cada minut en els dos sentits de circulació. S'han utilitzat les dades del punt de comptatge més proper a l'estació de recollida de dades de contaminació atmosfèrica. Aquest punt es troba situat al Carrer de la Unió entre les parròquies d'Andorra la Vella i Escaldes-Engordany.

Es disposa de dades per minut  per a l'any 2016 corresponents a tres sensors. Els atributs per a cada lectura són:

- *Data*.
- *Hora*.
- *Velocitat*. Velocitat mitjana de les lectures preses.
- *Ocupació*. Índex de l'estat del trànsit. S'utilitza per a valorar el nivell de retencions, ja que quan més alt és més retencions hi ha.

- *Intensitat*. Nombre de vehicles per minut.
- *No classificats*. Nombre de mesures que no s'han pogut classificar.
- *Vehicles hora*. Càlcul del nombre de vehicles hora a partir dels vehicles mesurats en un minut.
- *Apunts*. Quantitat de lectures en 1 minut.
- *Apunts vàlids*. Mesures vàlides entre les lectures preses.
- *Error*. Indica el percentatge d'errors de les mesures.

El següent mapa mostra la ubicació del punt de comptatge utilitzat així com la situació de l'estació fixa de mesura de la qualitat de l'aire.




***Imatge 11**. Distribució dels punts de comptatge de vehicles.*

S'han eliminat els casos no vàlids del conjunt inicial per evitar que distorsionin el model obtingut. La següent taula mostra els casos vàlids i no vàlids; existeix la mateixa quantitat de lectures no vàlides en els tres sensors.

| Casos | Vàlids | No vàlids | % vàlids | % no vàlids |
|---|---|---|---|---|
| 526.980 | 523.548 | 3.432 | 99,35 | 0,65 |

***Taula 10**. Resum dels casos vàlids i eliminats.*

Com es pot observar, les següents taules mostren els valors estadístics més rellevants de cada atribut numèric de cadascun dels tres sensors utilitzats: Carlemany, Meritxell i Carrer de la Unió.

| Sensor: Carlemany | | | | | |
|---|---|---|---|---|---|
| **Estadístic** | **Velocitat** | **Ocupació** | **Intensitat** | **No classificat** | **Vehicle hora** |
| Mínim | 0,0 | 0 | 0 | 0 | 0 |
| 1r quantil | 0,0 | 0 | 1 | 0 | 60 |
| Mediana | 23,0 | 2 | 4 | 0 | 240 |
| Mitjana | 19,61 | 5,7 | 4,58 | 0,90 | 274,7 |
| 3r quantil | 30,0 | 8 | 7 | 1 | 420 |
| Màxim | 255,0 | 100 | 55 | 55 | 3300 |
| Registres sense valor | 0 | 0 | 0 | 0 | 0 |

***Taula 11**. Principals valors estadístics dels atributs numèrics del sensor Carlemany.*

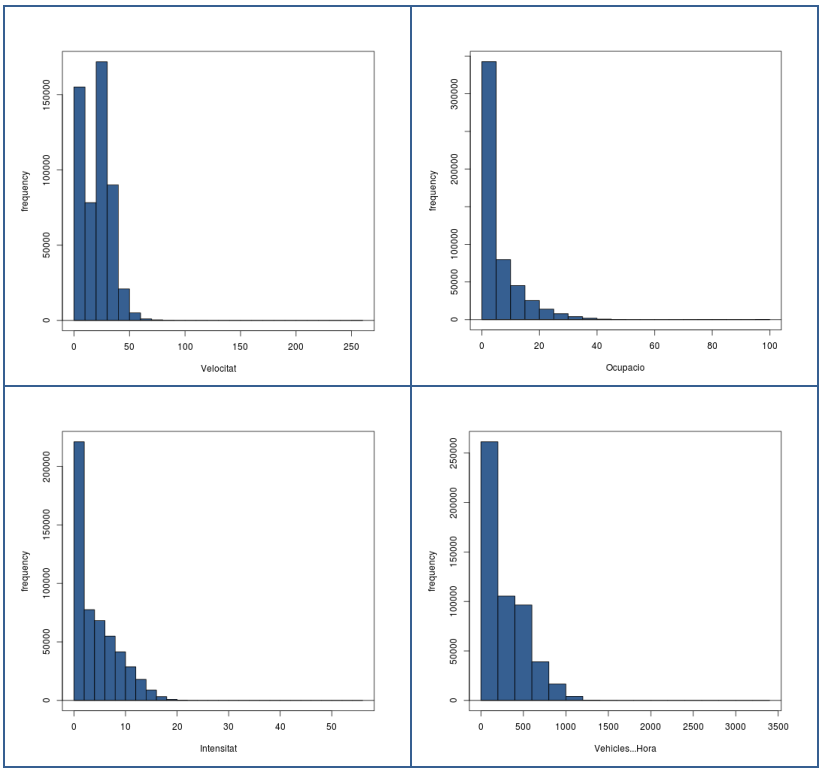
| Sensor: Meritxell | | | | | |
|---|---|---|---|---|---|
| **Estadístic** | **Velocitat** | **Ocupació** | **Intensitat** | **No classificat** | **Vehicle hora** |
| Mínim | 0,0 | 0 | 0 | 0 | 0 |
| 1r quantil | 0,0 | 0 | 0 | 0 | 0 |
| Mediana | 6,0 | 0 | 1 | 0 | 60 |
| Mitjana | 10,42 | 3,12 | 1,87 | 0,30 | 112,2 |
| 3r quantil | 20,0 | 5 | 3 | 0 | 180 |
| Màxim | 255,0 | 50 | 23 | 13 | 1380 |
| Registres sense valor | 0 | 0 | 0 | 0 | 0 |

***Taula 12**. Principals valors estadístics dels atributs numèrics del sensor Meritxell.*

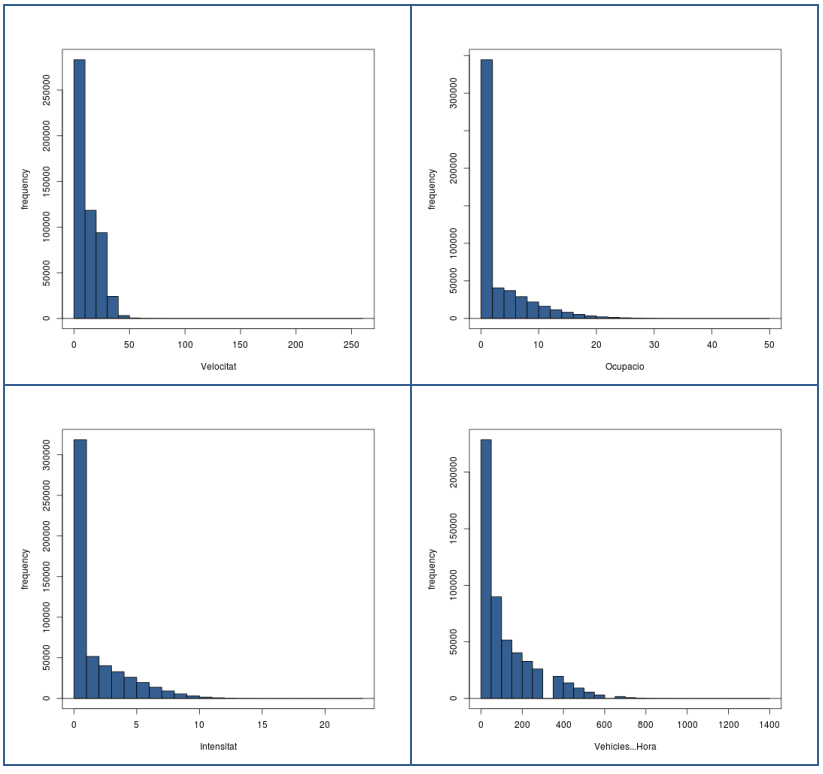
| Sensor: Carrer de la Unió | | | | | |
|---|---|---|---|---|---|
| **Estadístic** | **Velocitat** | **Ocupació** | **Intensitat** | **No classificat** | **Vehicle hora** |
| Mínim | 0,0 | 0 | 0 | 0 | 0 |
| 1r quantil | 12,0 | 0 | 1 | 0 | 60 |
| Mediana | 20,0 | 5 | 4 | 0 | 240 |
| Mitjana | 19,89 | 7,66 | 4,74 | 0,41 | 284,3 |
| 3r quantil | 29,0 | 12 | 8 | 1 | 480 |
| Màxim | 255,0 | 60 | 25 | 15 | 1500 |
| Registres sense valor | 0 | 0 | 0 | 0 | 0 |

***Taula 13**. Principals valors estadístics dels atributs numèrics del sensor Carrer de la Unió.*

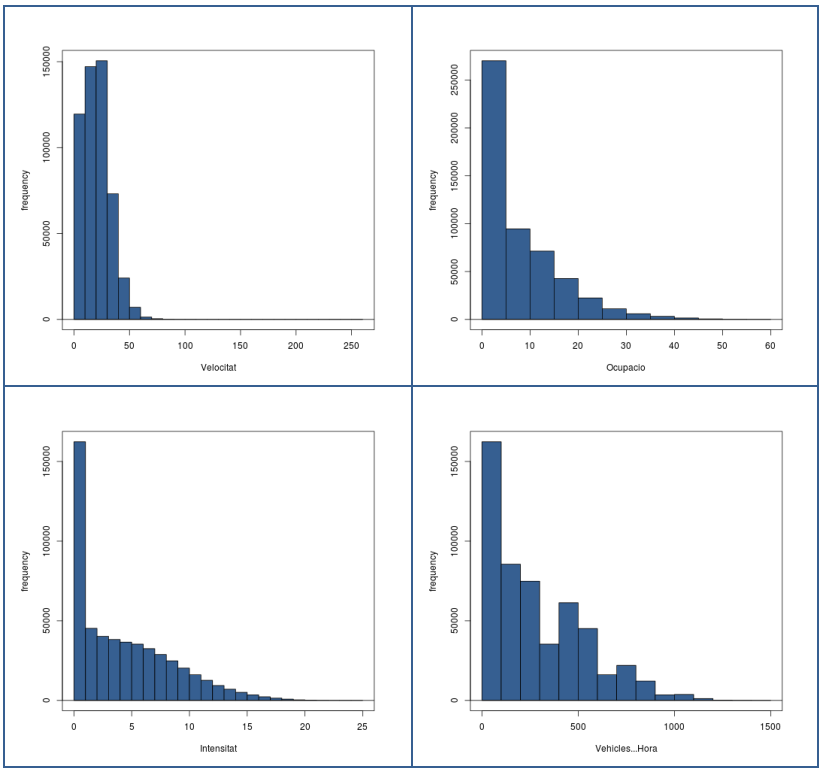
Els següents diagrames mostren la distribució dels atributs: velocitat, ocupació, intensitat i vehicles hora per als tres sensors.




***Imatge 12**. Histogrames dels principals atributs del sensor Carlemany.*




***Imatge 13**. Histogrames dels principals atributs del sensor Meritxell.*



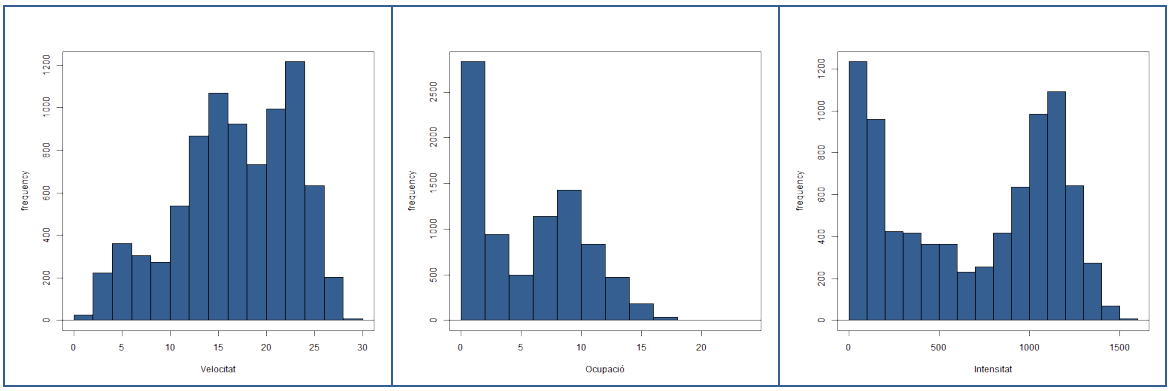

***Imatge 14**. Histogrames dels principals atributs del sensor Carrer de la Unió.*

Tot seguit s'han obtingut les mitjanes horàries dels tres sensors a partir de les lectures per minuts. La següent taula en mostra els principals valors estadístics.

| Estadístic | Velocitat | Ocupació | Intensitat | No classificat |
|---|---|---|---|---|
| Mínim | 0,00 | 0,00 | 0 | 0,00 |
| 1r quantil | 12,95 | 0,86 | 186 | 22,00 |
| Mediana | 17,11 | 5,75 | 782 | 75,00 |
| Mitjana | 16,78 | 5,57 | 680,5 | 97,74 |
| 3r quantil | 21,94 | 9,15 | 1100 | 128,00 |
| Màxim | 28,90 | 22,9 | 1579 | 735,00 |
| Registres sense valor | 0 | 0 | 0 | 0 |

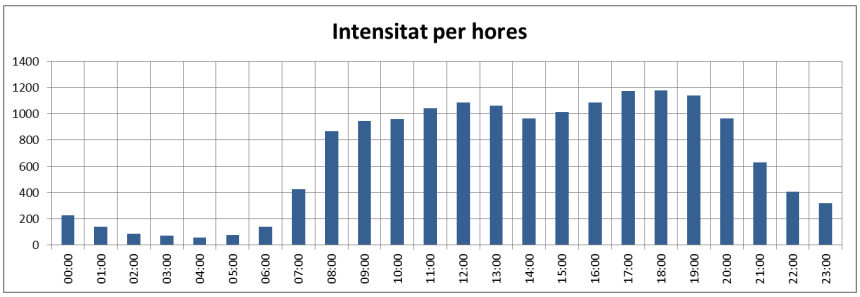
***Taula 14**. Principals valors estadístics dels atributs numèrics horaris.*

Els següents diagrames s'hi indica la distribució de freqüències de les mitjanes horàries dels atributs: velocitat, ocupació i intensitat.

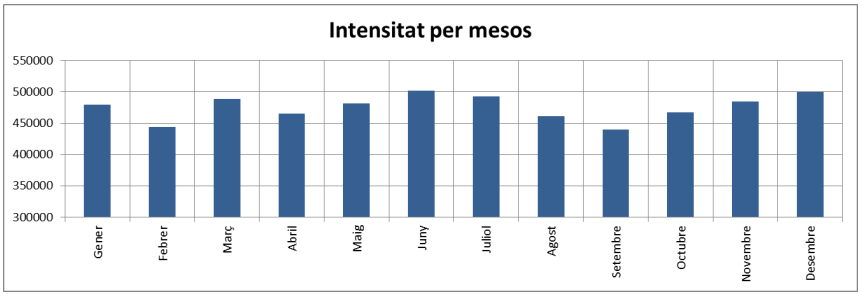



***Imatge 15**. Histogrames dels principals atributs per a les mitjanes horàries.*

Finalment, els següents gràfics mostren la intensitat mitjana de vehicles per hores així com l'acumulada per mesos al llarg l'any 2016. Com es pot observar, el trànsit augmenta entre les 6:00 i les 8:00 del matí, amb un primer màxim al voltant de les 12:00 i un segon màxim entre les 17:00 i les 18:00. Pel que fa a la distribució per mesos, s'observa un mínim el setembre amb 439.272 vehicles i un màxim al juny amb 501.753 vehicles seguit de prop pel més de desembre. Tanmateix, la quantitat total de vehicles es manté força estable durant l'any sempre per sobre dels 430.000 vehicles mensuals.




***Imatge 16**. Distribució de les intensitats mitjanes de vehicles per hores.*




***Imatge 17**. Distribució de les intensitats acumulades de vehicles per mesos.*

### 6.2.3. Dades de contaminació atmosfèrica

La Unitat de Medi Atmosfèric del Departament de Medi Ambient i Sostenibilitat del Govern d'Andorra disposa de tres estacions fixes i dues de mòbils. Les dades que recullen les diverses estacions són les següents:

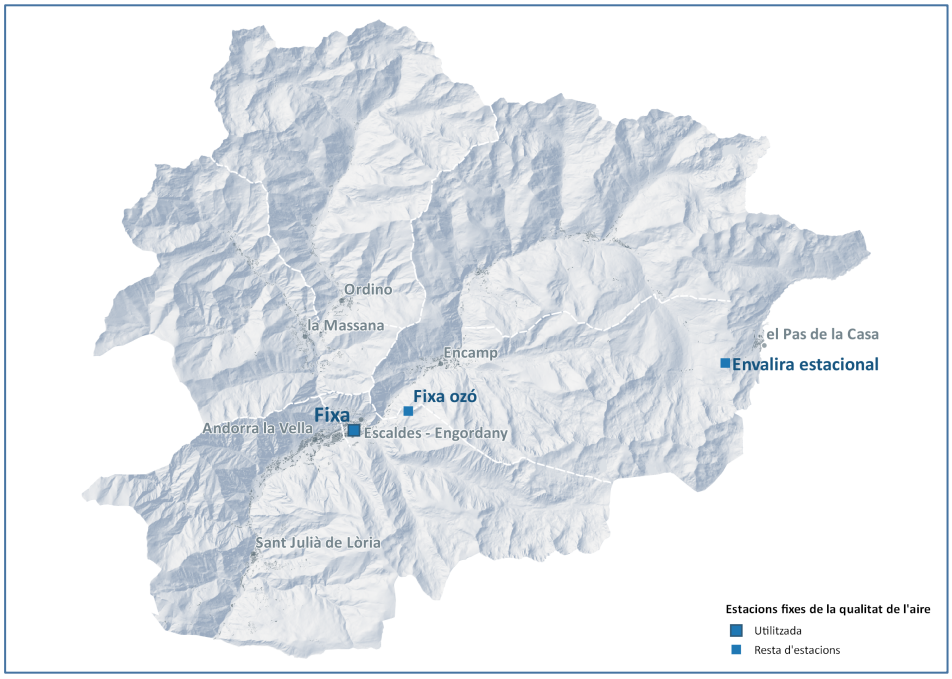
- *Monòxid de carboni.*
- *Monòxid de nitrogen.*
- *Diòxid de nitrogen.*
- *Ozó.*
- *Partícules PM10 i PM2.5.*
- *Diòxid de sofre.*
- *Temperatura ambient i velocitat del vent.*

Dels contaminants que recullen les estacions s'han descartat l'ozó i les partícules en suspensió ja que són produïts per fenòmens que tenen el seu origen fora del territori d'estudi com ara grans centres urbans i industrials en el cas de l'ozó o bé la pols sahariana en el cas de les partícules en suspensió. De la resta de contaminants s'ha tingut en compte tan sols el diòxid de nitrogen ja que és el que té una relació més directa amb la pol·lució produïda pels vehicles.

El diòxid de nitrogen ($NO_2$) s'emeten a l'atmosfera principalment arran dels fenòmens de combustió com la dels motors, les centrals tèrmiques o les estufes. En l'àmbit d'estudi el fenomen de combustió més important és el produït pels vehicles i, per tant, aquest contaminant té una relació directa amb la intensitat de la circulació. Cal tenir en compte que aquest contaminant és tòxic en ser inhalat. De fet, com indica l'Environmental Protection Agency dels Estats Units (2017) l'exposició prolongada a concentracions elevades de $NO_2$ poden augmentar el risc de patir problemes respiratoris, agreujar els problemes respiratoris previs i disminuir la funció pulmonar.

De les estacions disponibles s'han utilitzat les dades de l'estació anomenada "Fixa" situada a Escaldes-Engordany ja que és l'única estació fixa que recull dades del contaminant analitzat.

El següent mapa mostra la distribució d'estacions fixes de mesura de contaminació atmosfèrica en el territori i la ubicació de l'estació escollida per a l'estudi.




***Imatge 18**. Distribució de les estacions de mesura del nivell de la qualitat de l'aire.*

Es disposa de dades de contaminació horàries per al contaminant diòxid de nitrogen corresponents a l'any 2016. En total, el conjunt de dades està format per 8.785 registres amb els següents atributs:

- *Organisme*. Nom de l'organisme responsable.
- *Estació*. Nom de l'estació de mesura.
- *Mesura*. Contaminant mesurat.
- *Component*. Component del contaminant mesurat.
- *Unitat*. Unitat de mesura: $microg/m^3$.
- *Data*. Dia i hora de la mesura.
- *Valor*. Valor de la mesura.

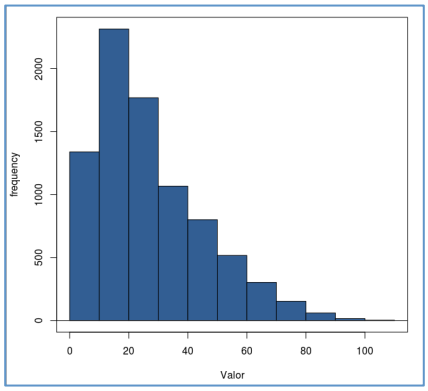
Els valors dels atributs: Organisme, Estació, Mesura, Component i Unitat són els mateixos per a tots els casos ja que es tracta de mesures d'un determinat contaminant per a una mateixa estació. Així, els únics camps útils per a l'obtenció del model són la data i el valor de les mesures.

La següent taula mostra els principals indicadors estadístics que prenen els valors de les mesures:

| Estadístic | Valor $(microg/m^3)$ |
|---|---|
| Mínim | 0,00 |
| 1r quantil | 13,30 |
| Mediana | 22,50 |
| Mitjana | 26,99 |
| 3r quantil | 37,50 |
| Màxim | 105,20 |

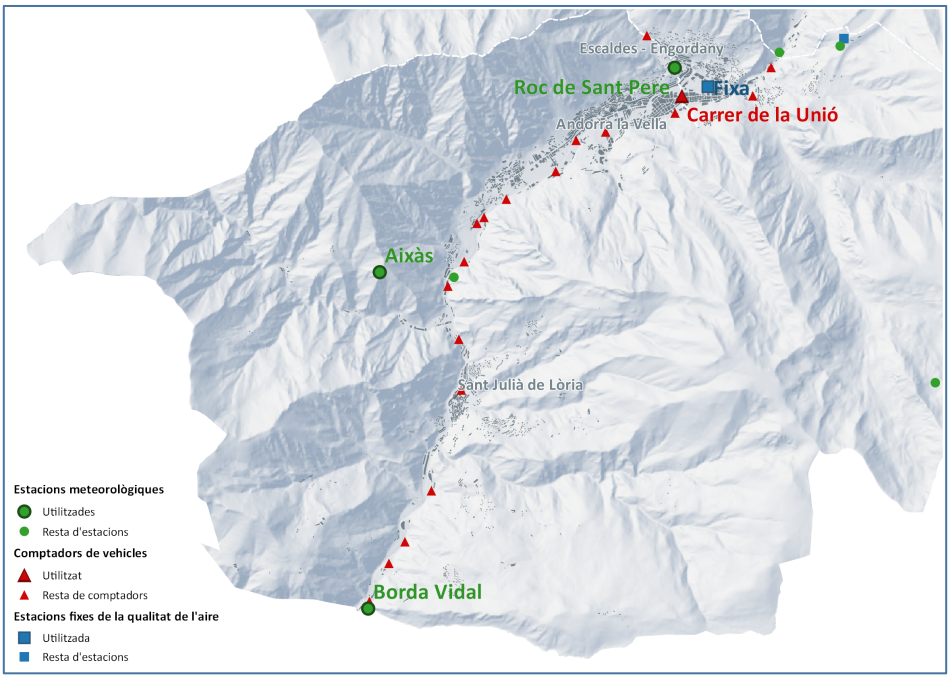
***Taula 15**. Principals valors estadístics de l'atribut Valor.*

Existeixen 442 registres sense cap valor. Com es pot observar en el següent histograma, la majoris dels valors de contaminació són baixo, essent la freqüència amb més casos aquells que prenen valors entre 10 i 20 i la freqüència amb menys caos aquells que es troben per sobre de 100.




***Imatge 19**. Histograma dels valors de contaminació.*

Finalment, el següent mapa mostra la ubicació de les estacions meteorològiques, el punt de comptatge de vehicles i l'estació fixa de mesura de la qualitat de l'aire emprats per a l'estudi.




***Imatge 20**. Ubicació estacions meteorològiques, punt de comptatge de vehicles i estació de mesura de la qualitat de l'aire.*

### 6.2.4. Anàlisi de correlació entre atributs

Un cop analitzades les diverses dades que s'utilitzaran per a l'obtenció del model, s'han avaluat les possibles correlacions lineals entre els atributs utilitzats. L'existència d'una correlació alta entre dos atributs indicarà que ambdós aporten la mateixa informació al model i, per tant, es pot concloure que un dels dos no és necessari.

S'ha utilitzat el coeficient de correlació de Pearson[46] per a obtenir la relació lineal entre les diverses variables quantitatives:

$$R = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$$

on $\sigma_{XY}$ és la covariància dels atributs X i Y i $\sigma_X$ i $\sigma_Y$ són les distribució típica dels atributs X i Y respectivament.

Els valors del coeficient de correlació de Pearson varien entre -1 i 1. Valors propers a 1 indiquen que el creixement d'una variable és proporcional al creixement d'una altra mentre que valors propers a -1 indiquen que aquesta relació és inversa. La independència lineal entre atributs s'obtindrà quan els valor del coeficient corresponent a aquests sigui proper a zero.

La següent taula mostra els coeficient obtinguts.

[46] https://ca.wikipedia.org/wiki/Coeficient_de_correlaci%C3%B3_de_Pearson

| Atribut | Temper. | Humitat | Direcció vent | Velocitat vent | Irradiació | Precipit. | Intensitat vehicles | Inversió tèrmica |
|---|---|---|---|---|---|---|---|---|
| **Temperatura** | 1.00 | -0.46 | -0.15 | -0.04 | 0.57 | 0.15 | -0.02 | 0.35 |
| **Humitat** | -0.46 | 1.00 | 0.14 | -0.18 | -0.60 | -0.36 | 0.19 | -0.44 |
| **Direcció vent** | -0.15 | 0.14 | 1.00 | 0.52 | -0.29 | -0.16 | -0.09 | -0.27 |
| **Velocitat vent** | -0.04 | -0.18 | 0.52 | 1.00 | -0.05 | 0.01 | -0.05 | -0.12 |
| **Irradiació** | 0.57 | -0.60 | -0.29 | -0.05 | 1.00 | 0.57 | -0.07 | 0.46 |
| **Precipitació** | 0.15 | -0.36 | -0.16 | 0.01 | 0.57 | 1.00 | -0.05 | 0.27 |
| **Intensitat vehicles** | -0.02 | 0.19 | -0.09 | -0.05 | -0.07 | -0.05 | 1.00 | 0.03 |
| **Inversió tèrmica** | 0.35 | -0.44 | -0.27 | -0.12 | 0.46 | 0.27 | 0.03 | 1.00 |

***Taula 16**. Coeficients de correlació de Pearson dels atributs numèrics.*

Es poden identificar quatre coeficients amb valors superiors a 0,5 que indiquen quins són els atributs amb major correlació lineal. En primer lloc, existeix correlació entre la direcció del vent i la velocitat d'aquest. En segon lloc, entre la temperatura i la irradiació solar. Tot seguit, s'ha identificat correlació lineal entre la humitat i la irradiació solar. En aquest cas la correlació és inversa, és a dir, quan augmenta la irradiació solar disminueix la humitat. Finalment, també s'ha identificat correlació entre la irradiació solar i la precipitació.

Tot i les correlacions lineals detectades, cap d'aquestes presenten valors alts propers a 1 i, per tant, s'ha optat per mantenir tots els atributs per a la generació del model d'aprenentatge.

### 6.2.5. Tractament previ

El tractament previ de les dades s'ha limitat a la categorització de la variable numèrica dels valors de contaminació per diòxid de nitrogen per a obtenir les categories dels model de classificació.

Per a definir les classes s'han utilitzat els índex europeus definits per Elshout (2006). La següent taula en mostra els rangs així com la distribució de casos per al conjunt d'entrenament.

| Classe | Rang ($\mu g\ m^3$) | Casos |
|---|---|---|
| 1 | 0-50 | 6.544 |
| 2 | 51-100 | 970 |
| 3 | 101-200 | 3 |
| 4 | 201-400 | 0 |
| 5 | >400 | 0 |

***Taula 17**. Rangs de les classes i distribució dels casos.*

No ha estat necessari dur a terme més tractaments de les dades com ara la normalització dels atributs numèrics en el cas dels classificador obtinguts amb els mètodes: xarxes neuronals, arbres de decisió i *Random forest*. En el cas dels arbres de decisió i del mètode *Random forest* l'estructura dels arbres es manté tant amb les dades normalitzades com amb les dades originals.

Per contra, sí que ha estat necessari normalitzar els atributs numèrics per a poder utilitzar els algorismes del veí més proper i la classificació basada en quantificació vectorial.

Tot seguit, s'analitzaran en detall aquests mètodes d'aprenentatge automàtic.

## 6.3. Elecció dels algorismes per a l'obtenció del model de predicció

En aquest projecte es pretén obtenir un model de predicció de la contaminació. Per tant, cal resoldre un problema de classificació on el classificador obtingut sigui capaç de predir el nivell de contaminació d'un determinat cas.

Així, s'han aplicat diversos dels algorismes de classificació descrits a l'apartat 5.1.2 a les dades del projecte.

Aquests classificadors s'han obtingut mitjançant el llenguatge de programació R[47] en un entorn no distribuït. Els models generats no només han servit per a avaluar els diversos algorismes de classificació disponibles sinó que ha permès completar l'anàlisi previ de les dades abans del seu tractament en un entorn distribuït.

Els algorismes d'aprenentatge automàtica utilitzats han estat: veí més proper, classificació basada en la quantificació vectorial, arbre de decisió, *Random forest* i xarxes neuronals.

Un cop eliminades les dades no vàlides s'ha obtingut un conjunt de 7.517 casos. A partir d'aquests s'ha generat un conjunt d'entrenament de 5.262 casos, és a dir del 70% dels exemples inicials. El 30% de casos restant s'ha utilitzat com a conjunt de test per a avaluar la qualitat del model obtingut.

### 6.3.1. Veí més proper

S'ha emprat el paquet *Class*[48] del llenguatge de programació R per a obtenir un classificador mitjançant l'algorisme del veí més proper amb diversos valors de *k*, on *k* és el nombre de veïns.

La següent taula mostra els resultats obtinguts.

| *k* | Precisió (%) | Matriu de confusió |
|---|---|---|
| 1 | 86,92 | Classe \| 1 \| 2 \| 3<br>1 \| 1804 \| 150 \| 1<br>2 \| 143 \| 156 \| 1<br>3 \| 0 \| 0 \| 0 |
| 5 | 88,56 | Classe \| 1 \| 2 \| 3<br>1 \| 1866 \| 175 \| 1<br>2 \| 81 \| 131 \| 1<br>3 \| 0 \| 0 \| 0 |
| 10 | 88,47 | Classe \| 1 \| 2 \| 3<br>1 \| 1884 \| 195 \| 1<br>2 \| 63 \| 111 \| 1<br>3 \| 0 \| 0 \| 0 |
| 20 | 88,65 | Classe \| 1 \| 2 \| 3<br>1 \| 1887 \| 194 \| 1<br>2 \| 60 \| 112 \| 1<br>3 \| 0 \| 0 \| 0 |

***Taula 18**. Resultats de l'aplicació de l'algorisme veí més proper per a diversos valors de* k.

[47] www.r-project.org
[48] https://stat.ethz.ch/R-manual/R-devel/library/class/html/knn.html

Els millors resultats s'han obtingut amb un valor de $k = 20$ amb una precisió del 88,6% en la classificació dels casos del conjunt de test.

### 6.3.2. Classificació basada en la quantificació vectorial

Per a obtenir un model predictiu de contaminació a partir de la classificació basada en la quantificació vectorial s'ha utilitzat, com en el cas anterior, el paquet *Class*[49] del llenguatge de programació R. El model obtingut ha generat deu centroides: nou corresponents a la classe 1 i un corresponent a la classe 2. La precisió obtinguda amb el classificador ha estat del 83,9%.

Tot seguit es mostra la matriu de confusió obtinguda. Com es pot observar, la majoria dels casos de la classe 2 han estat classificats errròniament.

| Classe | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 1819 | 128 | 0 |
| 2 | 232 | 74 | 0 |
| 3 | 1 | 1 | 0 |

***Taula 19**. Matriu de confusió.*

### 6.3.3. Arbre de decisió

L'algorisme utilitzat per a obtenir l'arbre de decisió ha estat C50[50]. La següent taula mostra l'ús que ha fet l'algorisme de cadascun dels atributs.

| Atribut | Percentatge d'ús |
|---|---|
| Inversió tèrmica | 100,0 |
| Temperatura | 84,0 |
| Intensitat | 52,4 |
| Humitat relativa | 9,7 |
| Velocitat del vent | 8,9 |
| Dia de la setmana | 8,2 |
| Irradiació solar | 3,8 |
| Insolació | 0,9 |
| Direcció del vent | 0,2 |

***Taula 20**. Ús dels atributs en l'arbre de decisió.*

Com es pot observar, l'atribut que té un pes més important en el model obtingut és la inversió tèrmica seguit de la temperatura i de la intensitat de vehicles. La resta d'atributs tenen un pes molt menor amb atributs com els minuts d'insolació i la direcció del vent amb un ús residual en el model.

Un cop utilitzat el conjunt de test per avaluar la qualitat del classificador s'ha obtingut un 89,3% de casos ben classificats. La següent taula mostra la matriu de confusió del model de predicció.

[49] https://stat.ethz.ch/R-manual/R-devel/library/class/html/knn.html
[50] https://cran.r-project.org/web/packages/C50/C50.pdf

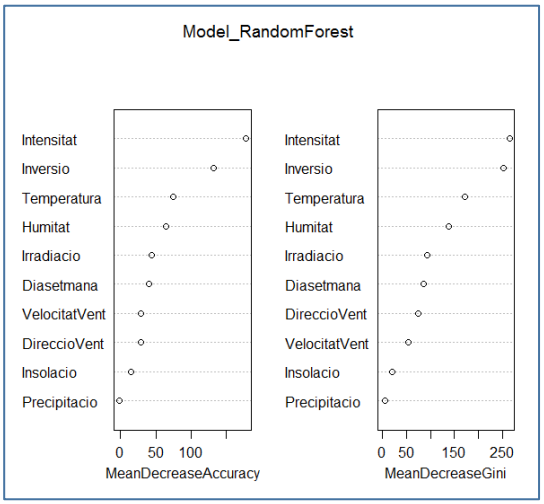
| Classe | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 1857 | 95 | 0 |
| 2 | 145 | 157 | 0 |
| 3 | 0 | 1 | 0 |

***Taula 21**. Matriu de confusió.*

### 6.3.4. *Random forest*

S'ha utilitzat la implementació de Breiman (2015) de l'algorisme *Random forest*. Aquest algorisme millora l'estabilitat i la precisió de la classificació dels arbres de decisió al mateix temps que n'evita el sobreajustament.

S'ha generat un total de 1000 arbres de decisió a partir de subconjunts de tres atributs. El següent gràfic mostra el pes de cada atribut en el model obtingut.




***Imatge 21**. Pes de cada atribut en el model.*

Com en el cas de l'arbre de decisió, els tres atributs amb més pes alhora de classificar han estat: la intensitat de vehicles, la inversió tèrmica i la temperatura, tot i que en aquest darrer model és la intensitat de vehicles el principal atribut alhora de classificar els casos en comptes de la inversió tèrmica com passava en l'anterior model.

Utilitzant el classificador generat en el conjunt de test s'ha obtingut una precisió en la classificació de nous casos del 90,5%, un punt i dues dècimes superior que en el model obtingut a partir de tan sol un arbre de decisió. La següent taula en mostra la matriu de confusió.

| Classe | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 1895 | 52 | 0 |
| 2 | 160 | 146 | 0 |
| 3 | 1 | 1 | 0 |

***Taula 22**. Matriu de confusió.*

### 6.3.5. Xarxes neuronals

S'ha utilitzat el paquet *nnet*[51] del llenguatge R per aplicar el mètode de classificació de xarxes neuronals. La següent taula mostra la matriu de confusió resultant de l'aplicació del classificador generat al conjunt de test.

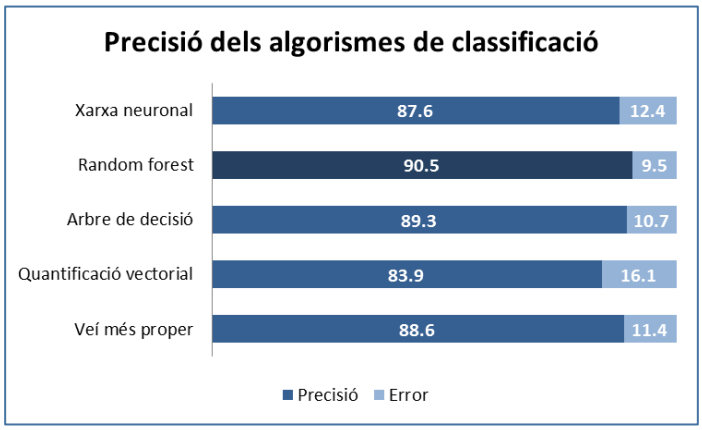
| Classe | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 1909 | 38 | 0 |
| 2 | 204 | 66 | 0 |
| 3 | 2 | 0 | 0 |

***Taula 23**. Matriu de confusió.*

La precisió obtinguda pel classificador ha estat del 87,6%. Tanmateix, s'observa que la xarxa neuronal ha classificat la majoria de casos del conjunt de test en la classe 1 obtenint una classificació pobra per aquells casos que pertanyen a la classe 2.

### 6.3.6. Comparació dels resultats obtinguts i elecció de l'algorisme d'aprenentatge

El següent gràfic resumeix els resultats obtinguts en l'avaluació dels diversos algorismes d'aprenentatge automàtic. La millor precisió és la donada per l'algorisme *Random forest* mentre que la pitjor s'ha obtingut amb la classificació basada en quantificació vectorial.




***Imatge 22**. Precisió obtinguda amb els diferents algorismes de classificació.*

Tots els algorismes utilitzats permeten obtenir classificadors amb una precisió alta. Tanmateix, l'algorisme seleccionat per a l'obtenció del model d'aprenentatge en un entorn *big data* ha estat *Random forest*, ja que és el que ha donat millors resultats en les proves dutes a terme i és també l'únic dels diversos algorismes analitzats implementat en la biblioteca Apache Mahout. Tot i això, l'anterior anàlisi permet concloure que també s'haurien pogut aplicar altres algorismes si aquests fossin disponibles com, per exemple, el veí més proper o els arbres de decisió.

[51] https://cran.r-project.org/web/packages/nnet/nnet.pdf

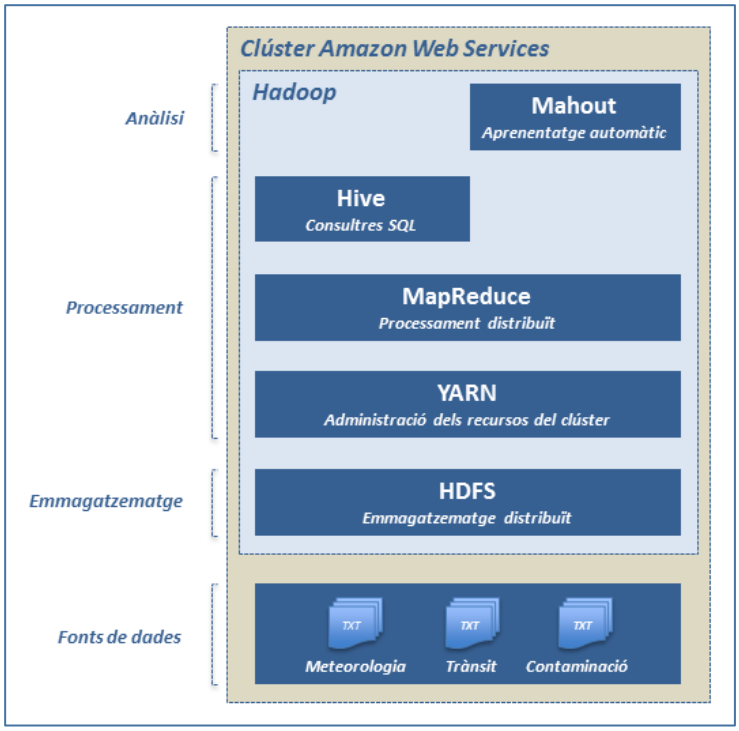
## 6.4. Preparació de l'arquitectura *big data*

S'han utilitzat dues arquitectures per a la implementació de l'entorn *big data*:

- Una primera arquitectura en **mode pseudodistribuït** on s'ha simulat un clúster Hadoop de dos nodes. Aquesta ha estat utilitzada com a entorn de desenvolupament per a dur a terme les proves del programari i validar el flux de treball. Vegeu la taula 4 amb la descripció de les característiques principals d'aquesta arquitectura.

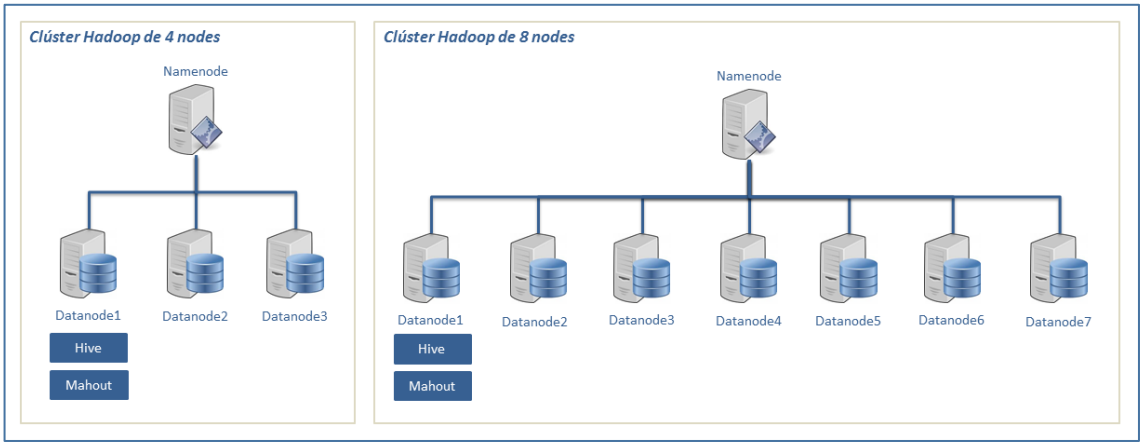
- Una segona arquitectura en **mode distribuït** on s'ha utilitzat computació en núvol per a implementar dos clústers Hadoop en la plataforma Amazon Web Services amb les característiques indicades en la taula 5. El primer clúster està format per un *namenode* i tres *datanodes* metre que el segon consta d'un *namenode* i set *datanodes*.

El següent diagrama mostra un resum dels principals elements de l'arquitectura en mode distribuït.




***Imatge 23**. Arquitectura big data en el clúster distribuït.*

Tot seguit es detalla la configuració dels dos clústers Hadoop implementats en la plataforma Amazon Web Services.




***Imatge 24**. Configuració dels dos clústers Hadoop implementats en la plataforma Amazon Web Services.*

Tots els nodes d'ambdós clústers tenen les mateixes característiques indicades en la següent taula. L'única excepció és el node *Datanode1* on s'ha implementat la plataforma Hive i la biblioteca Mahout. En una configuració inicial aquests dos components s'havien instal·lat en el *namenode* però el rendiment del clúster es veia afectat per una sobrecàrrega de tasques i es va optar per modificar aquesta configuració inicial per, d'aquesta forma, repartir millor la càrrega de treball.

| | |
|---|---|
| **Processador** | *1 vCPU de 2,5 GHz* |
| **Memòria RAM** | *1 GiB* |
| **Sistema operatiu** | *Ubuntu Server 16.04 LTS* |
| **Unitat d'emmagatzematge** | *SSD, 8 GiB* |
| **Java VM** | *Java SE Development Kit 8* |
| **Entorn de treball de computació distribuïda** | *Apache Hadoop 2.8.0* |
| **Plataforma per a consultes SQL distribuïdes** | *Apache Hive 2.1.1*[52] |
| **Biblioteca d'algorismes d'aprenentatge automàtic** | *Apache Mahout 0.11.0*[53] |

***Taula 24**. Principals característiques dels nodes dels clústers distribuïts.*

Tant la mida dels clústers com l'ús que se n'ha fet per a realitzar les proves han excedit els límits de la capa gratuïta de la plataforma Amazon Web Services. Així, les proves realitzades durant dos mesos han tingut un cost de 17,48€ corresponents a: 15,72€ d'excés d'hores de lloguer dels nodes, 0,72€ d'excés transferència de dades i 1,04€ d'excés d'espai d'emmagatzematge.

| Servei | Consum | | Tarifa | Cost | |
|---|---|---|---|---|---|
| | *Gratuït* | *Excés* | *(dòlars)* | *Dòlars* | *Euros* |
| Lloguer d'instàncies | 1250 hores | 1244 hores | 0,014 $/hora | 17,42 | 15,72 |
| Transferència de dades | 11,80 GB | 79,42 GB | 0,010 $/GB | 0,79 | 0,72 |
| Emmagatzematge | 45,31 GB | 9,66 GB | 0,119 $/GB | 1,15 | 1,04 |

***Taula 25**. Costos de la plataforma* big data.

Vegeu l'Annex I on es detalla com implementar l'arquitectura *big data* en ambdós modes.

[52] El programari Apache Hive només s'ha instal·lat en un dels nodes del clúster Hadoop.
[53] El programari Apache Mahout només s'ha instal·lat en un dels nodes del clúster Hadoop.

## 6.5. Càrrega de les dades a Hadoop i tractament previ amb Hive

La càrrega de les dades en el clúster Hadoop i el tractament previ d'aquestes s'han realitzat mitjançant consultes HiveQL en la plataforma Apache Hive (vegeu apartat 4.3). Per agilitzar el procés s'han programat un conjunt d'scripts Bash[54] encarregats, entre d'altres, de les crides a les consultes HiveQL. Aquests scripts són executables en els sistemes operatius basats en UNIX (com GNU/Linux), macOS i Windows 10.

En primer lloc, s'han desenvolupat un conjunt de consultes HiveQL per al tractament de les dades meteorològiques: càrrega de els dades, generació de la clau principal, eliminació dels casos amb valors nuls i eliminació dels duplicats.

Tot seguit, s'han carregat les dades de les dues estacions utilitzades per a obtenir els valors d'inversió tèrmica, se n'han generat les claus principals, s'ha calculat el valor d'inversió tèrmica i s'han eliminat els casos amb valors nuls.

A continuació s'han importat al clúster les dades de trànsit. Les dades dels tres sensors utilitzats es troben repartides en 36 arxius mensuals. Per cadascun dels arxius s'han carregat les dades a Hadoop, s'ha generat la clau principal i s'han convertit les dades per minuts a dades horàries. Un cop carregades totes les dades s'han sumat els valors d'intensitat de trànsit dels tres sensors i s'han eliminat els casos amb valors nuls.

Les darreres dades carregades al clúster han estat les de contaminació. Un cop carregades s'han projectat per a eliminar els camps no utilitzats en el model i finalment s'han eliminat els valors nuls i s'ha generat la classe de cada cas a partir del rangs definits en la taula 17.

Un cop carregades totes les dades, aquestes han estat unides en un sol conjunt de casos. Tot seguit, s'ha obtingut el conjunt d'entrenament a partir d'un subconjunt aleatori del 70% dels casos. La resta de casos s'han utilitzat per generar el conjunt de test que ha servit per avaluar la qualitat del model obtingut.

En l'Annex III es mostra el codi utilitzat per a la preparació de les dades.

## 6.6. Obtenció del model d'aprenentatge automàtic amb Apache Mahout

Un cop les dades han estat carregades a l'entorn *big data* i han estat processades per a obtenir tant el conjunt d'entrenament com el conjunt de test, s'ha iniciat el procés per a generar el model d'aprenentatge final en un entorn distribuït. Per a obtenir-lo, s'ha utilitzat l'algorisme de classificació implementat en la biblioteca Mahout *Random forests*. Aquest algorisme ha estat escollit en l'apartat 6.3.8 ja que ha estat el que ha aportat millors resultats per a les dades del projecte.

Utilitzant l'algorisme per a un primer cas de 100 arbres i conjunts de cinc atributs seleccionats aleatòriament s'han obtingut els resultats que es mostren en la següent taula.

[54] https://en.wikipedia.org/wiki/Bash_(Unix_shell)

| Paràmetre | | Valor |
|---|---|---|
| Temps de processament[55] | *Mode pseudodistribuït* | 4'51'' |
| | *Mode distribuït (4 nodes)* | 2'13'' |
| | *Mode distribuït (8 nodes)* | 2'14'' |
| Número total de nodes | | 24.930 |
| Mitjana de nodes dels arbres de decisió | | 249 |
| Profunditat mitjana dels arbres de decisió | | 16 |

***Taula 26**. Principals paràmetres del model d'aprenentatge.*

Així, s'han generat un total de 24.930 nodes en arbres que tenen de mitjana 249 nodes en un temps de processament en mode distribuït de 2 minuts i 14 segons.

Tot seguit, s'han realitzat un conjunt de proves amb diverses configuracions per a trobar aquelles que obtenen un millor resultat. La taula següent resumeix aquestes proves. Per a simplificar aquesta taula no s'ha tingut en compte la classe 3 que, com s'ha comentat anteriorment, té només tres casos i cap d'ells ha estat assignat aleatòriament al conjunt de test.

Com es pot observar, en les dues primeres proves, 1 i 2, s'ha analitzat l'ús de la implementació parcial. En aquesta implementació cada fase *map* de MapReduce s'encarrega de crear un subconjunt d'arbres a partir de les dades que té disponibles en el node sobre el qual s'executa. Aquesta característica afavoreix el rendiment de l'algorisme ja que cada partició del problema és carregada en memòria per a la seva execució. Efectivament, la implementació parcial s'ha processat en un 16% menys de temps que la prova sense implementació parcial. És per aquest motiu que la resta de proves s'han dut a terme amb aquesta configuració.

Tot seguit, s'han avaluat diverses implementacions amb vàries mides de partició en les proves 3 i 4. Tot i que s'esperava un augment de la precisió en augmentar la mida de les particions s'ha obtingut pràcticament el mateix resultat. Tot i així, el temps d'execució en el cas de poques particions (és a dir, amb mides de particions grans) és pràcticament la meitat que l'obtinguda amb una quantitat elevada de particions. En conseqüència, s'ha optat per aquesta segona configuració amb particions grans per a les següents proves.

Pel que fa al número d'arbres, s'ha observat que diverses configuracions en les proves 4, 5, 6 i 7 amb 100, 10, 1000 i 200 arbres respectivament donen pràcticament la mateixa precisió. Tot i així, un valor al voltant dels 100 arbres és el que permet un temps de processament inferior.

Finalment, s'han realitzat les proves 4, 8, 9, 10 i 11 amb un número diferent d'atributs seleccionats aleatòriament en cada node per a construir els arbres. El valor que ha donat millors resultats ha estat de cinc atributs.

[55] El temps de processament inclou: la preparació de les dades per a l'algorismes, l'obtenció del model i l'avaluació d'aquest.

| Paràmetres | | | | | Resultats | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Prova* | *Núm. arbres* | *Implemen. parcial* | *Núm. atributs* | *Mida particions (bytes)* | *Nodes* | *Profunditat mitjana* | *Casos correct. classificats (%)* | *Casos incorrect. classificats (%)* | *Matriu de confusió* | *Temps[56] (segons)* |
| **1** | 100 | Sí | 5 | 1.874.231 | 25.134 | 16 | 93,67 | 6,33 | Classe 1 2<br>1 1121 36<br>2 48 122 | 72 |
| **2** | 100 | No | 5 | 1.874.231 | 25.696 | 16 | 93,82 | 6,18% | Classe 1 2<br>1 1122 35<br>2 47 123 | 86 |
| **3** | 100 | Sí | 5 | 2.337 | 25.118 | 16 | 93,75 | 6,25 | Classe 1 2<br>1 1121 36<br>2 47 123 | 115 |
| **4** | **100** | **Sí** | **5** | **2.336.740** | **25.098** | **16** | **93,75** | **6,25** | Classe 1 2<br>1 **1122** **35**<br>2 **48** **122** | **61** |
| **5** | 10 | Sí | 5 | 2.336.740 | 2.524 | 16 | 93,67 | 6,33 | Classe 1 2<br>1 1121 36<br>2 48 122 | 86 |
| **6** | 1000 | Sí | 5 | 2.336.740 | 250.396 | 16 | 93,75 | 6,25 | Classe 1 2<br>1 1122 35<br>2 48 122 | 74 |
| **7** | 200 | Sí | 5 | 2.336.740 | 50.350 | 16 | 93,75 | 6,25 | Classe 1 2<br>1 1122 35<br>2 48 122 | 66 |
| **8** | 100 | Sí | 2 | 2.336.740 | 32.418 | 18 | 93,82 | 6,18 | Classe 1 2<br>1 1122 35<br>2 47 123 | 69 |
| **9** | 100 | Sí | 3 | 2.336.740 | 29.294 | 17 | 93,75 | 6,25 | Classe 1 2<br>1 1122 35<br>2 48 122 | 80 |
| **10** | 100 | Sí | 6 | 2.336.740 | 24.032 | 16 | 93,82 | 6,18 | Classe 1 2<br>1 1122 35<br>2 47 123 | 70 |
| **11** | 100 | Sí | 9 | 2.336.740 | 22.498 | 15 | 93,82 | 6,18 | Classe 1 2<br>1 1122 35<br>2 47 123 | 75 |

***Taula 27**. Paràmetres utilitzats en les proves realitzades i resultats obtinguts. Es ressalta la prova amb millors resultats.*

En conclusió, els millor resultats han estat els obtinguts en la 4a prova amb 100 arbres de decisió, implantació parcial, conjunts de 5 atributs i mides de partició grans.

En l'Annex II es detalla el manual d'execució del cas pràctic i en l'Annex III es mostra el codi utilitzat per a l'obtenció i l'avaluació del model d'aprenentatge.

## 6.7. Anàlisi dels resultats obtinguts

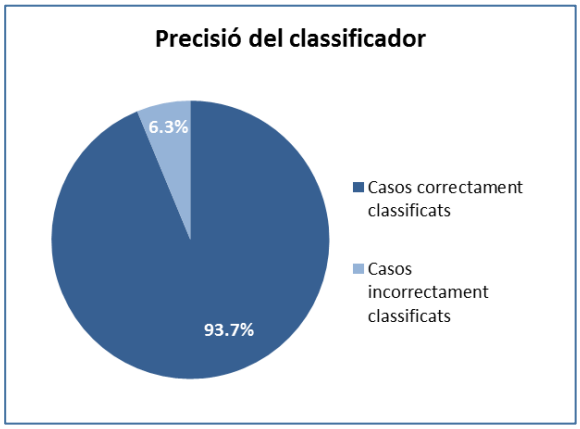
Un cop obtingut el model d'aprenentatge s'ha utilitzat el conjunt de test per avaluar la qualitat d'aquest. La següent taula indica els resultats de la predicció del classificador.

[56] El temps de processament inclou: l'obtenció del model i l'avaluació d'aquest. No inclou la preparació de les dades per a l'algorisme.

| Paràmetre | Valor | % |
|---|---|---|
| Número total de casos del conjunt de test | 1.327 | 100,00 |
| Casos correctament classificats | 1.244 | 93,75 |
| Casos incorrectament classificats | 83 | 6,25 |

***Taula 28**. Resultats de l'avaluació del model.*

Per tant, el conjunt de test permet estimar la precisió del model en un 93,75% com mostra el següent gràfic.




***Imatge 25**. Precisió del classificador.*

Com es pot observar en la següent matriu de confusió, s'han classificat correctament el 97,0% dels casos pertanyents a la classe 1 i el 71,8% dels casos de la classe 2. Es pot explicar aquesta diferència pel fet que la classe 1 és la més nombrosa i, per tant, el model s'hi ajusta millor.

| Classe | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 1122 | 35 | 0 |
| 2 | 48 | 122 | 0 |
| 3 | 0 | 0 | 0 |

***Taula 29**. Matriu de confusió.*

Per a millorar la precisió de la classificació dels casos de la classe 2 es pot plantejar l'ús d'un conjunt de casos major amb dades de diversos anys. Per altra banda, cal tenir en compte que només existeixen 3 casos de la classe 3 i que, en seleccionar de forma aleatòria els casos del conjunt d'entrenament i el de test cap cas de la classe 3 ha esta assignat a aquest darrer conjunt.

Tot i que el classificador obtingut té una precisió alta i, per tant, és capaç de predir els episodis de contaminació a partir de dades meteorològiques i de trànsit de forma satisfactòria, es podrien aplicar algunes millores a aquest que permetessin predir les diverses classes evitant un sobreajustament en una de les classes. És per això que en la següent taula es proposen nous rangs per a les classes que reparteixin millor els casos i, així, permetrien obtenir un millor entrenament del model.

| Classe | Proposta | Casos | % |
|---|---|---|---|
| 1 | 0-15 | 2504 | 30,0 |
| 2 | 15-30 | 2919 | 35,0 |
| 3 | 30-40 | 1067 | 12,8 |
| 4 | >40 | 1853 | 22,2 |

***Taula 30**. Nous rangs per a les classes de les dades de contaminació.*

Dels resultats obtinguts en els diversos models de classificació generats es pot observar que tant les dades meteorològiques com les de trànsit tenen una afectació directa en els episodis de contaminació atmosfèrica per diòxid de nitrogen. Per tant, s'ha confirmat la hipòtesi plantejada inicialment ja que aquestes dades permeten obtenir models predictius vàlids.

Concretament, d'entre les diverses variables meteorològiques emprades per a l'obtenció dels models, la inversió tèrmica és la que ha mostrat tenir un pes més important, essent el segon atribut més utilitzat per l'algorisme *Random forest* i el primer en el cas dels arbres de decisió. Per tant, es pot afirmar que, d'entre les variables meteorològiques utilitzades, la inversió tèrmica és la que més influeix en el desencadenament d'episodis de contaminació atmosfèrica per diòxid de nitrogen. Concretament, l'existència d'inversió tèrmica augmenta la possibilitat de contaminació atmosfèrica.

La segona variable meteorològica que més pes ha tingut en els diversos models obtinguts ha estat la temperatura tant en el cas dels arbres de decisió com en el cas del mètode *Random forest*. Podem afirmar, doncs, que la temperatura també influeix directament en l'aparició dels fenòmens de contaminació atmosfèrica.

La resta d'atributs meteorològics: humitat relativa, velocitat del vent, irradiació solar, insolació i direcció del vent, tot i que han contribuït a la generació del model d'aprenentatge, han tingut un pes molt menor. Podem afirmar, doncs, que aquestes variables meteorològiques tenen una influència secundària en el desencadenament dels episodis de contaminació.

Tot i que inicialment s'havia considerat que el vent podria tenir un pes important en la concentració de contaminants, els models obtinguts han donat un pes secundari tant a la velocitat del vent com a la direcció del vent. Tanmateix, podem afirmar que el model empra més la velocitat del vent que la direcció d'aquest alhora de predir els episodis de contaminació.

Una altre atribut que ha tingut un pes inferior a l'esperat inicialment ha estat el dia de la setmana. Tot i que la concentració de contaminants té un comportament cíclic al llarg de la setmana, aquest atribut té un pes relativament baix en els diversos models obtinguts essent la sisena variable amb un pes més important d'un total de deu en els models obtinguts mitjançant els algorismes *Randorm Forest* i arbres de decisió.

Pel que fa a les dades de trànsit, com s'havia previst, aquestes tenen una afectació directa en l'índex de contaminació atmosfèrica per diòxid de nitrogen. De fet, és la variable amb més pes en el cas de l'algorisme *Random Forest* i la tercera pel que fa als arbres de decisió. Els models obtinguts indiquen que l'augment de la intensitat de circulació té una incidència directa en l'augment de la concentració de $NO_2$.

En conclusió, tant les dades meteorològiques com les de trànsit són necessàries per a l'obtenció de models de predicció de contaminació vàlids ja que el pes d'ambdues fonts de dades és important en els diversos classificadors obtinguts en aquest projecte.

## 6.8. Anàlisi de rendiment

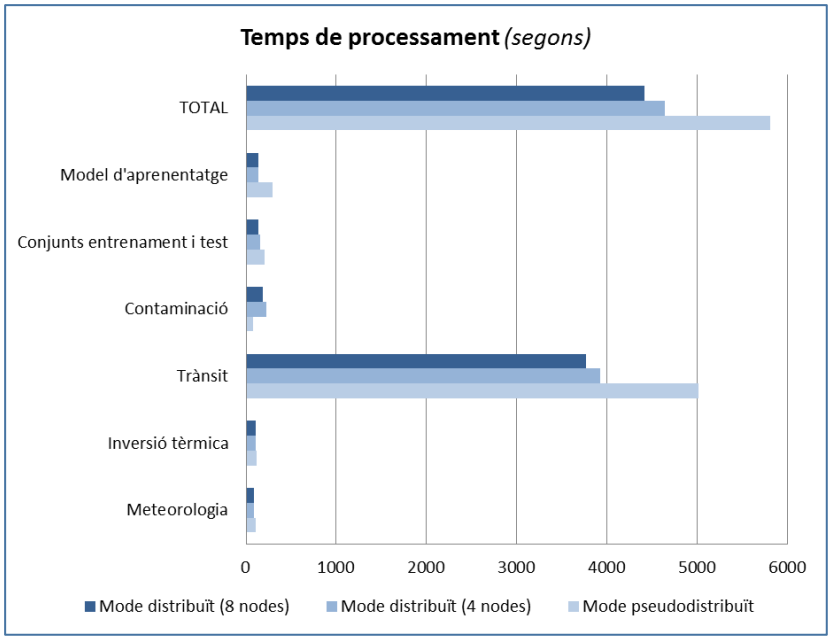
Finalment, s'ha realitat un anàlisi del rendiment del sistema implementat que permet avaluar quines de les fases o bé plataformes utilitzades tenen uns costos de processament més alts.

La següent taula mostra el temps de processament per a la càrrega i el tractament previ de les dades de cadascuna de les fonts: meteorologia, trànsit i contaminació així com el temps utilitzat per a l'obtenció dels conjunts de casos i el model d'aprenentatge. Aquest anàlisi s'ha realitzat en el clúster en mode pseudodistribuït i en els dos clústers implementats en mode distribuït.

| Fase | Mida de les dades | Núm. arxius | Temps de processament | | |
|---|---|---|---|---|---|
| | | | *Mode pseudodistribuït* | *Mode distribuït (4 nodes)* | *Mode distribuït (8 nodes)* |
| Meteorologia | 361 KB | 1 | 1'45" | 1'30" | 1'23" |
| Inversió tèrmica | 363 KB | 2 | 1'56" | 1'50" | 1'47" |
| Trànsit | 62,3 MB | 36 | 1h23'29" | 1h05'29" | 1h02'50" |
| Contaminació | 847 KB | 1 | 1'21" | 3'42" | 3'01" |
| Conjunts entrenament i test | - | - | 3'22" | 2'34" | 2'16' |
| Model d'aprenentatge | - | - | 4'51" | 2'13" | 2'14" |
| **Total** | **63,7 MB** | **40** | **1h36'44"** | **1h17'18"** | **1h13'31"** |

***Taula 31**. Mida de les dades, número d'arxius i temps de processament de les fases d'obtenció del model.*

Com es pot observar en el següent gràfic, s'han obtingut rendiments similars en tots tres clústers. Tanmateix, el clúster en mode pseudodistribuït és el que obté un pitjor rendiment i el clúster distribuït de vuit nodes el que obté els millor resultats amb un temps de processament un 24,0 % inferior al del clúster pseudodistribuït. La diferència en el temps de processament entre els dos clústers distribuïts és inferior a l'esperada essent l'estalvi de temps del clúster de vuit nodes respecte al de quatre nodes de tan sols el 4,9%.




***Imatge 26**. Temps de processament.*

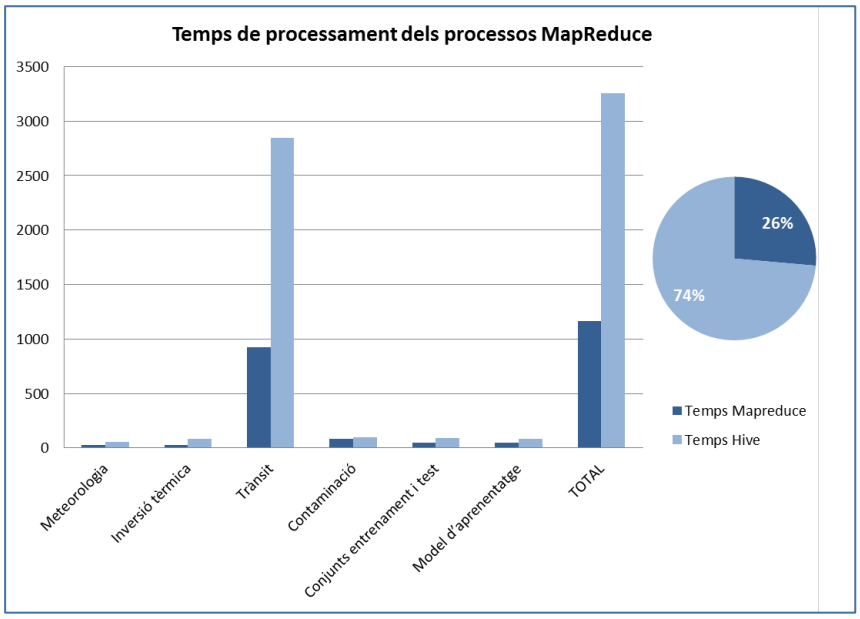
Per altra banda, el conjunt que ha requerit més temps de processament han estat les dades de trànsit degut, principalment al fet d'haver de processar 36 arxius independents amb dades per minuts a diferència de la resta de fons de dades on es disposava d'un o dos arxius de dades horàries.

Tot seguit s'ha analitzat el temps utilitzat pel clúster Hadoop distribuït de vuit nodes per a executar els processos MapReduce i per a realitzar la resta d'accions entre les quals destaquen, principalment, les diverses connexions a Hive necessàries així com la preparació i el llançament de les consultes. Cal tenir en compte que Hive tradueix les consultes HiveQL a diversos processos MapRedude. També s'hi detallen els processos MapReduce i les tasques *map* i *reduce* necessaris per a cadascuna de les fases.

| Fase | Processos MapReduce | Núm. tasques | | Temps *(segons)* | |
|---|---|---|---|---|---|
| | | *map* | *reduce* | *MapReduce* | *Hive* |
| Meteorologia | 3 | 3 | 1 | 30 | 57 |
| Inversió tèrmica | 4 | 4 | 0 | 29 | 81 |
| Trànsit | 90 | 94 | 30 | 925 | 2845 |
| Contaminació | 7 | 10 | 8 | 84 | 100 |
| Conjunts entrenament i test | 4 | 4 | 2 | 49 | 87 |
| Model d'aprenentatge | 7 | 7 | 0 | 50 | 84 |
| **Total** | **115** | **122** | **41** | **1167** | **3254** |

***Taula 32**. Principals paràmetres de rendiment del clúster de vuit nodes.*

El temps mitjà d'execució dels processos MapReduce ha estat de 10,1 segons mentre que el temps mitjà de processament de les tasques ha estat de 7,2 segons. Per altra banda, el temps invertit per Hive representa el 74% del total i, per tant, afecta considerablement al rendiment del sistema. El següent gràfic mostra el temps dedicat als processos MapReduce i el temps dedicat a les consultes de la plataforma Hive.




***Imatge 27**. Comparació del temps de processament dels processos MapReduce i el temps invertit per la plataforma Hive.*

# 7. Conclusions

En primer lloc, es pot concloure que la hipòtesi plantejada a l'inici d'aquest projecte ha quedat confirmada, ja que el model d'aprenentatge obtingut és capaç de predir episodis de contaminació a partir de dades recollides per les estacions meteorològiques i pels sensors de trànsit. Per tant, es pot afirmar que les condicions tant del trànsit com meteorològiques tenen una afectació directa en l'augment de la concentració de diòxid de nitrogen a l'atmosfera. Per altra bada, l'ús conjunt d'ambdues fonts de dades ha donat bons resultats alhora de modelar el comportament de la concentració del contaminant analitzat i, per tant, podem concloure que ambdues fonts són necessàries per a l'obtenció de models d'aprenentatge precisos.

També s'ha observat que els atributs que més pes tenen alhora de modelar aquest comportament per part dels diversos models d'aprenentatge obtinguts han estat la inversió tèrmica, la intensitat de vehicles i la temperatura. En conseqüència, podem concloure que aquests són els factors que més influeixen en la concentració de diòxid de nitrogen d'entre aquells que s'han utilitzat per a obtenir els models predictius.

Per contra, altres atributs dels quals s'havia suposat una afectació important en els episodis de concentració de contaminants a l'atmosfera, finalment no han tingut la importància esperada. És el cas tant del dia de la setmana com de la direcció i intensitat del vent que han tingut un pes relativament baix en els diversos models predictius obtinguts.

Pel que fa als algorismes de classificació analitzats, tots ells han generat models d'aprenentatge amb precisions superiors al 83% per a les dades tractades. Concretament, el model menys precís ha estat l'obtingut mitjançant classificació basat en la quantificació vectorial. Per altra banda, els models obtinguts mitjançant xarxes neuronals, arbres de decisió i el mètode del veí més proper han assolit precisions entre el 87,6% i el 89,3%. Finalment, el mètode amb uns millors resultats ha estat l'algorisme *Random forest* amb una precisió del 93,8%. És per aquest motiu que s'ha considerat aquest darrer mètode l'idoni per a l'obtenció del model final en un entorn *big data*.

Tot i que el model s'ha obtingut a partir de dades de l'any 2016, aquest pot ser utilitzat per a altres anys ja que els episodis de contaminació presenten un comportament similar tots els anys.

En conclusió, l'ús de tècniques d'intel·ligència artificial s'ha mostrat vàlid per a resoldre el problema plantejat en aquest projecte.

Per la seva banda, el paradigma *big data* permet el tractament i l'anàlisi de grans volums de dades de forma eficient i és una bona solució quan es disposa de dades provinent de sensors distribuïts pel territori.

Pel que fa a l'entorn *big data* emprat, tot i que algunes de les eines utilitzades d'aquesta plataforma han presentat dificultats alhora d'instal·lar-les o bé de configurar-les s'han pogut realitzar les tasques que s'havien previst inicialment amb bons resultats.

Malgrat que existeixen distribucions i projectes que simplifiquen les tasques d'instal·lació i configuració de Hadoop com, per exemple, Cloudera[57] que ofereix màquines virtuals amb el programari ja instal·lat o bé Apache Ambari[58], que proporciona des d'instal·ladors fins a eines

[57] https://www.cloudera.com/downloads/quickstart_vms/5-10.html
[58] https://ambari.apache.org/

de gestió i monitoratge d'un clústes Hadoop, aquestes només han estat emprades en una fase inicial amb la finalitat a realitzar les primeres proves. Així, s'ha optat per la instal·lació i configuració de totes les eines usades ja que es cercava un major control d'aquestes. Tot i això, l'ús de projectes com Cloudera o Ambari poden ser una bona solució per a evitar el temps d'instal·lació, configuració i aprenentatge que requereixen les diverses eines de l'ecosistema Hadoop.

Hadoop disposa d'un ampli conjunt d'eines i plataformes desenvolupades sobre aquest que permeten resoldre una gran quantitat de problemes en el camp del *big data*. Tant l'ús del sistema de fitxers distribuït de Hadoop, HDFS, com MapReduce i la plataforma Hive han estat satisfactoris. Malgrat això, cal tenir en compte que les dades que s'han analitzant són històriques i que, per tant, s'ha utilitzat processament per lots. Així, si el rendiment hagués estat un requeriment important del sistema o bé s'hagués necessitat processar dades a temps real, l'arquitectura utilitzada hagués estat inadequada. Per a solucionar-lo es podrien haver usat altres motors de processament com Apache Tez o bé Apache Spark en substitució de MapReduce mantenint el mateix flux de treball així com les mateixes consultes HiveQL sense necessitat d'implementar nou codi.

Per altra banda, la informàtica en núvol ha permès la implementació de la solució *big data* en un clúster sense la necessitat d'adquirir maquinari i amb un cost molt reduït. Aquesta tecnologia ha permès treballar amb diverses arquitectures i realitzar nombroses proves.

Pel que fa als objectius plantejats a l'inici del projecte, tots ells s'han assolit, des de la implementació de solucions *big data* i l'ús d'algorismes d'aprenentatge automàtic en entorns distribuïts fins a la utilització de tècniques d'intel·ligència artificial per a l'obtenció de models predictius de contaminació atmosfèrica. L'únic objectiu que tan sols s'ha assolit de forma parcial ha estat l'anàlisi del model de programació MapReduce ja que, tot i que se n'ha realitzat l'anàlisi, finalment s'ha optat per l'ús de consultes HiveQL per al tractament previ de les dades en substitució de la implementació directa de funcions MapReduce.

Per altra banda, durant la realització d'aquest treball final de grau s'ha seguit la planificació i la metodologia establertes inicialment. Així, es pot afirmar que la metodologia ha estat l'adequada ja que no han sorgit problemes significatius durant la seva aplicació. Els canvis en la planificació temporal han estat poc significatius i en cap moment han posat en perill l'assoliment de cap dels objectius generals i específics plantejats a l'inici.

Una possible línia de treball futur és la modificació de l'arquitectura actual per a permetre el tractament de fluxos de dades a temps real meteorològiques i de trànsit amb l'entorn de treball Spark capaç de dur a terme tant processament per lots com de fluxos o bé un entorn completament de processament de fluxos com Storm o Samza. Això permetria utilitzar el model d'aprenentatge amb dades a temps real i, per tant, activar alertes quan aquest preveiés una episodi de contaminació a partir de les dades meteorològiques i de trànsit recollides pels diversos sensors.

Per altra banda, també es pot continuar el treball iniciat en aquest projecte amb el desenvolupament de nous models predictius d'altres contaminants o, fins i tot, d'altres àmbits emprant la mateixa metodologia.

Finalment, es pot plantejar una segona ampliació de l'estudi dut a terme encaminada a l'ús de dades de tot el territori i l'anàlisi espacial del comportament del model de predicció obtingut. Cal tenir en compte que aquesta ampliació del territori analitzat representaria un augment

significatiu de les dades tractades per a l'obtenció dels classificadors i, per tant, augmentaria els costos en una implementació distribuïda.

# 8. Glossari

**Amazon Web Services** *m* serveis d'informàtica en núvol oferts per Amazon.com.
*sigla* **AWS**.

**Apache Hadoop** *m* entorn de treball de codi obert que fa possible el processament distribuït de grans volums de dades a través d'un clúster d'ordinadors.

**Apache Hive** *m* plataforma d'alt nivell per al tractament de grans conjunts de dades emmagatzemades en sistemes de fitxers distribuïts mitjançant consultes basades en el llenguatge SQL.

**Apache Impala** *m* motor de consultes SQL per a dades emmagatzemades en sistemes de fitxers distribuïts.

**Apache Mahout** *m* biblioteca de codi obert d'algorismes d'aprenentatge automàtic escalables escrita en Java que pot ser utilitzada quan el volum de dades que cal processar és molt gran.

**Apache Pig** *m* plataforma d'alt nivell per al tractament de grans conjunts de dades emmagatzemades en sistemes de fitxers distribuïts mitjançant el llenguatge de consultes Pig Latin.

**AWS** Vegeu **Amazon Web Services**.

**aprenentatge automàtic** *m* camp de la intel·ligència artificial que estudia tècniques per a proveir a les màquines de la capacitat d'aprendre.
*en* **machine learning**.

**aprenentatge no supervisat** *m* tipus d'aprenentatge on no es disposa de cap mena d'informació sobre les sortides.

**aprenentatge supervisat** m tipus d'aprenentatge on es coneix quina és la resposta del sistema.

**arbre de decisió** *m* mètode de classificació.

***batch processing*** Vegeu **processament per lots**.

***big data*** Vegeu **dades massives**.

**categorització** *m* mètodes d'aprenentatge no supervisat capaços d'obtenir un conjunt de categories a partir d'un conjunt d'objectes inicials.
*en* **clustering**.

**CentOS** *m* distribució del sistema operatiu GNU/Linux de codi lliure clon de la distribució Red Hat.

**classificació** *f* mètodes d'aprenentatge supervisat on es disposa d'un conjunt d'objectes dels quals es coneix el valor de sortida.
*en* **classification**.

**classificació basada en la quantificació vectorial** *f* mètode de classificació basat en el mètode de categorització quantificació vectorial.
*en* **vector quantization**.

**classificadors lineals** *m* mètodes de classificació que selecciona la classe d'un objecte a partir del valor de les combinacions lineals de les seves característiques.

**classification** Vegeu **classificació**

***cloud computing*** Vegeu **informàtica en núvol**.

***clustering*** Vegeu **categorització**.

**dades massives** *f* conjunt de tecnologies per al tractament de grans volums de dades de fonts diverses i on la velocitat de processament és important.
*en* **big data**

***datanode*** *m* node esclau que emmagatzema els blocs en un sistema d'arxius HDFS.

**Hadoop** Vegeu **Apache Hadoop**.

**Hadoop Distributed File System** *m* sistema de fitxers distribuït d'Apache Hadoop.
*sigla* **HDFS**.

**HDFS** Vegeu **Hadoop Distributed File System**.

**hipervisor** *m* programari o maquinari que utilitza la virtualització per a facilitar l'ús de diversos sistemes operatius en un mateix ordinador.
*en* **hypervisor**.

**Hive** Vegeu **Apache Hive**.

***hypervisor*** Vegeu **hipervisor**.

**IaaS** m Vegeu **infraestructura com a servei**.

**Impala** Vegeu **Apache Impala**.

**informàtica en núvol** *f* model que permet l'accés en xarxa a recursos informàtics compartits sota demanda i des de qualsevol indret.
*en* **cloud computing**.

**infraestructura com a servei** *f* model d'informàtica en núvol on s'ofereixen recursos de processament, emmagatzematge i xarxa, entre d'altres.
*en* **infrastructure as a service**.
*sigla* **IaaS**.

***infrastructure as a service*** Vegeu **infraestructura com a servei**.

***k-nearest neighbour*** Vegeu **veí més proper**.

**k-NN** Vegeu **veí més proper**.

***machine learning*** Vegeu **aprenentatge automàtic**.

**Mahout** Vegeu **Apache Mahout**.

**MapReduce** *m* model de programació per al processament de grans volums de dades que utilitza computació paral·lela i distribuïda.

**màquina de vectors de suport** *m* mètode de classificació que permet classificar objectes definits per atributs numèrics en dues classes.
*en* **suport vector machines**.
*sigla* **SVM**.

***namenode*** *m* node màster que gestiona l'espai de noms dels sistemes de fitxers HDFS.

**PaaS** Vegeu **plataforma com a servei**.

**Pig** Vegeu **Apache Pig**.

**plataforma com a servei** *f* model d'informàtica en núvol en què s'ofereixen entorns de desenvolupament on els usuaris poden desplegar aplicacions pròpies o bé de tercers tenint control sobre el desplegament i, sovint, sobre la configuració de l'entorn d'allotjament.
*en* **platform as a service**.
*sigla* **PaaS**.

***platform as a service*** Vegeu **plataforma com a servei**.

**processament de fluxos** *m* paradigma de programació que permet a les aplicacions el processament en paral·lel.
*en* **stream processing**.

**processament per lots** *m* processament on les dades rebudes no es processen immediatament ja que el temps de resposta no és rellevant.
*en* **batch processing**.

**programari com a servei** *m* model d'informàtica en núvol on s'ofereix als usuaris aplicacions i bases de dades.
*en* **software as a service**.
*sigla* **SaaS**.

***Random forest*** *m* variant del mètode de classificació arbre de decisió que utilitza l'agregació de bootstrap per a millorar l'estabilitat i la precisió de la classificació tot evitant el sobreajustament.

**SaaS** Vegeu **programari com a servei**.

***software as a service*** Vegeu **programari com a servei**.

***stream processing*** Vegeu **processament de fluxos**.

**suport vector machines** Vegeu **màquina de vectors de suport**.

**SVM** Vegeu **màquina de vectors de suport**.

**Ubuntu** *m* distribució del sistema operatiu GNU/Linux basada en Debian.

**vector quantization** Vegeu **classificació basada en la quantificació vectorial**.

**veí més proper** *m* mètode de classificació basat en els algorismes de categorització que classifica un objecte a partir de l'objecte, o objectes, més proper.
*en* **k-nearest neighbour**.
*sigla* **k-NN**.

**xarxes neuronals** *f* mètode de classificació que simula les propietats dels sistemes neuronals dels éssers vius mitjançant model matemàtics.

**YARN** *m* sistema de gestió dels recursos, monitoratge i planificació de tasques en clústers Apache Hadoop.

# 9. Bibliografia

BREIMAN, Leo (2001). R*andom Forest*. [Data de consulta: 22 de maig de 2017] <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.23.3999&rep=rep1&type=pdf>

BREIMAN, Leo; LÉGER, Karine (2015). *Package 'randomForest'*. The Comprehensive R Archive Network. [Data de consulta: 22 de maig de 2017] <https://cran.r-project.org/web/packages/randomForest/randomForest.pdf>

ELLINGWOOD, Justin (2016). *Hadoop, Storm, Samza, Spark and Flink: Big Data Frameworks Compared*. Digital Ocean. [Data de consulta: 22 de maig de 2017] <https://www.digitalocean.com/community/tutorials/hadoop-storm-samza-spark-and-flink-big-data-frameworks-compared#apache-hadoop>

ELSHOUT, Sef; LÉGER, Karine (2006). *Comparign Urban Air Quality Acroos Borders. A review of existint air quality indices and the proposal of a common alternative*. Schiedam: Environmental Protection Agency Rijnmond. [Data de consulta: 22 de maig de 2017] <http://www.aire.ad/assets/documents/24200707_ComparingUrbanAirQualityAcrossBorders.pdf>

ENVIRONMENT PROTECTION AGENCY (2017). *Nitrogen Dioxide ($NO_2$) Pollution*: US EPA [Data de consulta: 22 de maig de 2017] <https://www.epa.gov/no2-pollution>

ESTEVES, Rui Máximo; RONG, Chinming (2011, desembre). "Using Mahout for clustering Wikiepdia's latest articles" [ponència]. A: *Third IEEE International Conference on Cloud Computing Technology and Science.* Atenes.

Hadoop: http://hadoop.apache.org. [Data de consulta: 22 de maig de 2017].

Hive: https://hive.apache.org/. [Data de consulta: 22 de maig de 2017].

JIAN, Sheng Jen; CHIU, Ruey Kei; WANG, Shin-an (2012, juliol). "A cloud decision suport System for the risk assessment of coronary heart disease" [ponència]. A: *International Conference of Machine Learning and Cybernetics.* Xian.

LANEY, Doublas (2001, febrer). "3D Data management: Controlling Data Volume, Velocity and Variety". *Gartner*.

LANDSET, Sara; KHOSHGOFTAAR, Taghi M.; RICHTER, Aaron N.; HASANIN, Tawfiq (2015). "A survey of open source tools for Machine learning with Big Data in the Hadoop ecosystem". *Journal of Big Data* (núm. 2:24).

Mahout: http://mahout.apache.org. [Data de consulta: 22 de maig de 2017].

MANYIKA, James; CHUI, Michael; BUGHIN, Jaques *et al.* (2011, maig). *Big Data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute. [Data de consulta: 22 de maig de 2017] <http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation>

MapReduce Tutorial: https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html. [Data de consulta: 22 de maig de 2017].

MCCALLUM, Andrew; NIGAM, Kamal; UNGAR, Lyle H. (2000, agost). "3D Data management: Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching". [ponència]. A:

*Sixth ACM SIGKDD International conference of Knowledge discovery and data mining* (pàg 169-178). Boston, Association for Computing Machinery.

MELL, Peter; GRANCE, Timothy (2011). *The NIST Definition of Cloud Computing*. Gaithersbur: National Institute os Standards and Technology. [Data de consulta: 22 de maig de 2017] <http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf>

SARNOVSKÝ, M.; BUTKA, P.; PÓCSOVÁ, J. (2012, juny). "Cloud Computing as a Platform for Distributed Fuzzy FCA Approache in Data Analysis" [ponència]. A: *International Conference of Intelligent Engineering Systems.* Lisbon.

TAN, Pang-Ning; STEINBACK, Michael; KUMAR, Vipin (2005). *Introduction to Data Mining* (1a ed.). Harlow: Pearson.

TURNER, Vernon (2014). *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things*. EMC. [Data de consulta: 22 de maig de 2017] <https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm>

VALE, Steven (2013). *Classification of Types of Big Data*. [article en línia]. UNECE [Data de consulta: 22 de maig de 2017]. <http://www1.unece.org/stat/platform/display/bigdata/Classification+of+Types+of+Big+Data>

WANG, Shin-an; CHIU, Ruey Kei; JIAN, Sheng-Jen (2012, juliol). "The implementation of an intelligent cloud service System for diease risk assessment – chronic kindney diesease as an example" [ponència]. A: *International Conference of Machine Learning and Cybernetics.* Xian.

WHITE, Tom (2015). *Hadoop: The Definitive Guide* (4a ed.). Sebastopol: O'Reilly.

ZHOU, Ping; LEI, Jingsheng; YE, Wenjun (2011, desembre). "Large-Scale Data Sets Clustering Based on MapReduce and Hadoop" [ponència]. A: *Journal of Computational Information Systems* (núm. 7, pàg 5956-5963).

# 10. Annexos

En els següents annexos es detalla el procés d'instal·lació de les eines emprades en el projecte, el procediment per a l'execució del model d'aprenentatge i el codi implementat.

## Annex I. Manual d'instal·lació de les eines

En aquest annex es descriuen els passo que cal seguir per a la instal·lació i la configuració de les eines utilitzades tant en mode pseudodistribuït com en mode distribuït.

### Annex I.1. Configuració de l'entorn en mode pseudodistribuït

Per a la configuració de l'entorn en mode pseudodistribuït cal seguir els següents passos:

*1. Instal·lació de l'hipervisor Oracle VM VirtualBox.*

Accediu a l'URL https://www.virtualbox.org/ i instal·leu la versió de l'hipervisor corresponent al vostre sistema operatiu amfitrió.

*2. Creació d'una nova màquina virtual i instal·lació del sistema operatiu Ubuntu 16.04.*

Seguiu els passos descrits en el següent URL per a configurar una nova màquina virtual: http://download.virtualbox.org/virtualbox/5.1.22/UserManual.pdf. Podeu optar per instal·lar el sistema operatiu des d'una imatge ISO[59] d'aquest o bé obtenir una màquina virtual amb el sistema operatiu ja instal·lat.

*3. Configuració SSH.*

Accediu a la nova màquina virtual i executeu-hi un intèrpret de comandes (terminal) i instal·leu les aplicacions *ssh* i *pdsh* mitjançant les següents instruccions.

```
local$ sudo apt-get install ssh
local$ sudo apt-get install pdsh
```

*3. Instal·lació del JDK de Java.*

Instal·leu la versió 8 del JDK de Java mitjançant la següent instrucció o bé des del web[60].

```
local$ sudo apt-get install openjdk-8-jdk
```

*4. Instal·lació d'Apache Hadoop.*

Instal·leu la versió 2.8.0 d'Apache Hadoop mitjançant la següent comanda o bé descarregueu-vos-la[61].

[59] https://www.ubuntu.com/download
[60] http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html
[61] http://hadoop.apache.org/releases.html

```
local$ wget http://apache.mirrors.tds.net/hadoop/common/hadoop-2.8.0/hadoop-
2.8.0.tar.gz -P ~/Downloads
```

*5. Configuració de Hadoop en mode pseudodistribuït.*

Seguiu les instruccions del següent URL per a la configuració de Hadoop en mode pseudodistribuït:
http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/SingleCluster.html.

*6. Inici del clúster Hadoop.*

Inicieu el clúster Hadoop mitjançant les següents comandes.

```
local$ $HADOOP_HOME/sbin/start-dfs.sh
local$ $HADOOP_HOME/sbin/start-yarn.sh
```

*7. Instal·lació i configuració d'Apache Hive.*

Seguiu les instruccions del següent URL per a instal·lar i configurar Apache Hive: https://cwiki.apache.org/confluence/display/Hive/GettingStarted.

*8. Instal·lació d'Apache Mahout.*

Instal·leu la versió 0.11.0 de la biblioteca Mahout mitjançant la següent comanda o bé directament des del web[62].

```
local$ wget http://www-eu.apache.org/dist/mahout/0.11.0/apache-mahout-distribution-
0.11.0.tar.gz -P ~/Downloads
```

## Annex I.2. Configuració de l'entorn en mode distribuït

Per a la configuració de l'entorn en mode distribuït cal seguir els passos que s'enumeren a continuació.

*1. Registre en la plataforma d'informàtica en núvol Amazon Web Services.*

Accediu a l'URL https://aws.amazon.com i registreu-vos-hi. Els recursos que s'utilitzaran formen part de la capa gratuïta de la plataforma. Tot i així, si aquests s'excedeixen la utilització del clúster pot comportar costos.

*2. Creació d'instàncies EC2.*

Seguiu les instruccions del següent URL per a crear els diversos nodes del clúster: http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/EC2_GetStarted.html#ec2-launch-instance_linux.

[62] http://mahout.apache.org/general/downloads.html

Es pot optar per diverses configuracions d'instàncies. En aquest projecte s'han utilitzat instàncies amb el sistema operatiu Red Hat Enterprise Linux 7.3 i Ubuntu Server 16.04 LTS ja que ambdós formen part de la capa gratuïta de la plataforma. Finalment, els clústers s'han implementat amb instàncies amb el sistema operatiu Ubuntu Server 16.04.

En el procés de creació dels nodes cal generar un parell de claus, pública i privada, que servirà posteriorment per accedir als nodes del clúster. Cal emmagatzemar l'arxiu amb les claus en el propi ordinadors per utilitzar-lo posteriorment.

*3. Configuració SSH.*

Per a poder accedir als nodes des de la unitat local caldrà executar la següent instrucció.

```
local$ ssh -i clau.pem usuari@dns_pública_namenode
```

on *clau.pem* és la clau generada en c*rea*r el clúster, *usuari* és el nom de l'usuari del node (normalment *ubuntu* per a les instàncies EC2 Ubuntu 16.04) i *dns_pública_namenode* és la DNS pública del *namenode*.

Tot seguit cal connectar-se al *namenode* i executar les següents instruccions per a completar la configuració SSH.

```
local$ ssh namenode
namenode$ ssh-keygen -f ~/.ssh/id_rsa -t rsa -P ""
namenode$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
namenode$ cat ~/.ssh/id_rsa.pub | ssh datanode1 'cat >> ~/.ssh/authorized_keys'
```

Cal repetir la darrera instrucció per a cadascun dels *datanodes* del clúster on *datanode1* s'haurà de substituir per cadascun dels noms dels *datanodes*.

*3. Instal·lació del JDK de Java en tots els nodes.*

Instal·leu la versió 8 del JDK de Java en cadascun del nodes mitjançant la següent instrucció o bé des del web[63].

```
nodes$ sudo apt-get install openjdk-8-jdk
```

*4. Instal·lació d'Apache Hadoop en tots els nodes.*

Instal·leu la versió 2.8.0 d'Apache Hadoop en tots els nodes mitjançant la següent comanda o bé descarregueu-vos-la[64].

```
node$ wget http://apache.mirrors.tds.net/hadoop/common/hadoop-2.8.0/hadoop-
2.8.0.tar.gz -P ~/Downloads
```

[63] http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html
[64] http://hadoop.apache.org/releases.html

*5. Configuració de Hadoop en mode distribuït.*

Seguir les instruccions del següent URL per a la configuració en mode distribuït del clúster Hadoop: http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/ClusterSetup.html.

*6. Inici el clúster Hadoop.*

Accediu al node *namenode* i inicieu-hi el clúster Hadoop mitjançant les següents comandes.

```
local$ ssh namenode
namenode$ $HADOOP_HOME/sbin/start-dfs.sh
namenode$ $HADOOP_HOME/sbin/start-yarn.sh
```

*7. Instal·lació i configuració d'Apache Hive.*

Seguiu les instruccions del següent URL per a instal·lar i configurar Apache Hive en un dels *datanodes* del clúster:
https://cwiki.apache.org/confluence/display/Hive/GettingStarted.

*8. Instal·lació i configuració d'Apache Mahout.*

Instal·leu la versió 0.11.0 de la biblioteca Mahout mitjançant la següent comanda o bé directament des del web[65] en un dels *datanodes*.

```
node$ wget http://www-eu.apache.org/dist/mahout/0.11.0/apache-mahout-distribution-0.11.0.tar.gz -P ~/Downloads
```

[65] http://mahout.apache.org/general/downloads.html

# Annex II. Manual d'execució

L'execució del cas pràctic és similar en ambdós modes: pseudodistribuït i distribuït. En mode pseudodistribuït cal copiar les dades en una carpeta anomenada TFG en la màquina virtual i començar l'execució pel pas 3 d'aquest manual. Per contra, en mode distribuït cal començar l'execució pel pas 1.

*1. Còpia de les dades en el clúster.*

Utilitzeu la següent comanda per a copiar les dades en el clúster Hadoop.

```
local$ scp -i clau.pem TFG/dades/*  usuari@dns_pública_datanode:~/TFG
```

on *clau.pem* és la clau generada en crear el clúster, *usuari* és el nom de l'usuari del node (normalment *ubuntu* per a les instàncies EC2 Ubuntu 16.04) i *dns_pública_datanode* és la DNS pública del *datanode* on s'ha instal·lat Apache Hive.

*2. Accés al* datanode *on es troba instal·lat Apache Hive.*

Accediu al *datanode* on s'ha instal·lat Apache Hive mitjançant la següent comanda:

```
local$ ssh datanode1
```

on *datanode1* és el nom del *datanode* on s'ha instal·lat Apache Hive (vegeu l'Annex I.2).

*3. Inici el clúster Hadoop.*

Inicieu el clúster Hadoop si aquest es troba aturat.

```
datanode$ $HADOOP_HOME/sbin/start-dfs.sh
datanode$ $HADOOP_HOME/sbin/start-yarn.sh
```

Recordeu que si treballeu en mode pseudodistribuït haureu d'iniciar el clúster des de la màquina virtual.

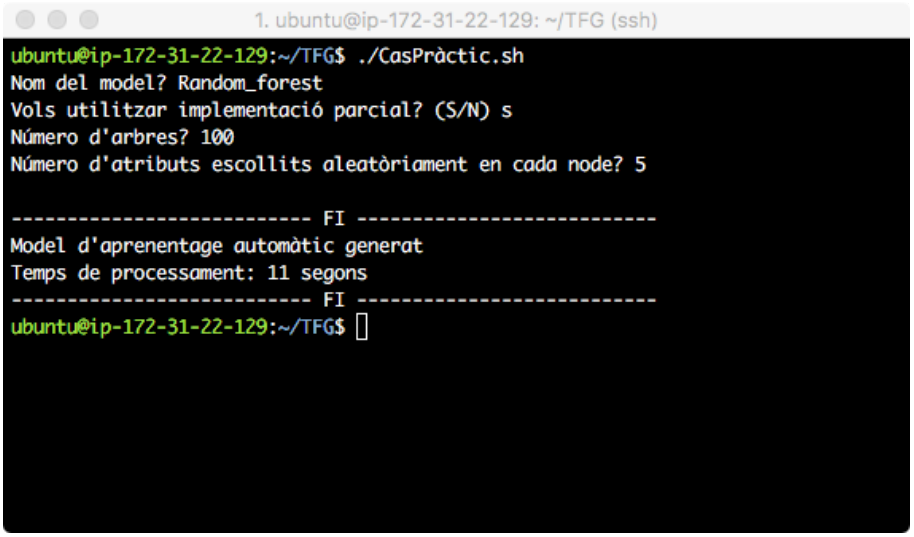
*4. Execució del cas pràctic.*

Executeu el cas pràctic mitjançant les següents comandes.

```
datanode$ cd ~/TFG
datanode$ ./CasPràctic.sh
```

El sistema informarà dels diversos processos MapReduce que s'executen i de la finalització del tractament de cadascun dels conjunts de dades. Un cop tractades les dades i generats els conjunts d'entrenament i de test, el sistema demanarà a l'usuari els paràmetres del nou model d'aprenentatge: el nom del nou model, l'ús de la

implementació parcial, el número d'arbres que es generaran i el número d'atributs escollits aleatòriament en cada node.




***Imatge 28**. Exemple de fragment de l'execució del cas pràctic.*

Finalment, es mostrarà el resultat del model obtingut així com els paràmetres de l'avaluació d'aquest.

# Annex III. Codi

Tot seguit es mostra el codi implementat. Per una banda, s'han desenvolupat un conjunt d'scripts encarregats de realitzat les crides a les consultes HiveQL, d'informar a l'usuari de l'evolució de l'execució, de l'obtenció del model d'aprenentatge i de sol·licitar a l'usuari els paràmetres necessaris. Per altra banda, s'han implementat un seguit de consultes HiveQL per a la càrrega al sistema i el tractament previ de les dades, l'obtenció dels conjunts d'entrenament i de test, la generació del model i l'avaluació d'aquest.

Aquest codi es pot executar des de sistemes operatius basats en UNIX (com GNU/Linux), macOS i Windows 10.

## *Execució del cas pràctic*

El següent script és l'encarregat de l'execució de les diverses fases del cas pràctic: la càrrega a Hadoop i el tractament previ de les dades meteorològiques, d'inversió tèrmica, trànsit i contaminació atmosfèrica, l'obtenció del conjunts d'entrenament i test i la generació del model d'aprenentatge.

```
#! /bin/bash
# Execució del cas pràctic TFG.

inici_s=`date +%s`

#Càrrega a Hadoop i tractament previ de les dades meteorològiques
bash meteorologia/meteorologia.sh

#Càrrega a Hadoop i tractament previ de les dades d'inversió tèrmica
bash inverio_temica/inversio_termica.sh

#Càrrega a Hadoop i tractament previ de les dades de trànsit
bash transit/transit.sh

#Càrrega a Haddop i tractament previ de les dades de contaminació atmosfèrica
bash contaminacio/contaminacio.sh

#Unió de les dades, obtenció del conjunt de casos, del conjunt d'entrenament i del conjunt de test
bash conjunt_casos/conjunt_casos.sh

#Obtenció del model d'aprenentatge i avaluació de la seva qualitat
bash model/model.sh

fi_s=`date +%s`
let total_s=$fi_s-$inici_s

echo
echo "--------------------------- FI ---------------------------"
echo "S'ha finalitzat l'execució del cas pràctic"
echo "Temps total d'execució del cas pràctic: $total_s segons"
echo "--------------------------- FI ---------------------------"
```

***Imatge 29**. Arxiu CasPràctic.sh per a l'execució del cas pràctic.*

## *Càrrega i tractament de les dades meteorològiques*

```
#! /bin/bash

inici_s=`date +%s`
hive -f meteorologia.sql
fi_s=`date +%s`
let total_s=$fi_s-$inici_s

echo "--------------------------- FI ---------------------------"
echo "Dades meteorològiques carregades i tractades correctament."
echo "Temps de processament: $total_s segons"
```

***Imatge 30**. Arxiu meteorologia.sh per a la càrrega i tractament de les dades meteorològiques.*

```
--Càrrega de les dades meteoroògiques
CREATE TABLE meteorologiaTemp (Data String, Hora String, Temperatura Float, HumitatRelativa int,
DireccioVent int, VelocitatVent int, Irradiacio int, Insolacio int, Precipitacio Float)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
LINES TERMINATED BY '\n'
STORED AS TEXTFILE;

LOAD DATA LOCAL INPATH 'meteorologia.txt' INTO TABLE meteorologiaTemp;

--Generació de la clau principal
CREATE TABLE meteorologiaTemp1 AS SELECT concat(Data,'-',Hora) AS id, Temperatura,
HumitatRelativa, DireccioVent, VelocitatVent, Irradiacio, Insolacio, Precipitacio
        FROM meteorologiaTemp;

CREATE TABLE meteorologiaTemp2 (id String, Temperatura Float, HumitatRelativa int, DireccioVent int,
VelocitatVent int, Irradiacio int, Insolacio int, Precipitacio Float)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
LINES TERMINATED BY '\n'
STORED AS TEXTFILE;

--Eliminació dels casos amb valors nuls
INSERT OVERWRITE TABLE meteorologiaTemp2
   SELECT * FROM meteorologiaTemp1
   WHERE Temperatura IS NOT NULL AND
   HumitatRelativa IS NOT NULL AND
   DireccioVent IS NOT NULL AND
   VelocitatVent IS NOT NULL AND
   Irradiacio IS NOT NULL AND
   Insolacio IS NOT NULL AND
   Precipitacio IS NOT NULL;

--Eliminació dels casos duplicats
CREATE TABLE meteorologia AS
SELECT id,
  MAX(Temperatura) AS Temperatura,
  MAX(HumitatRelativa) AS HumitatRelativa,
  MAX(DireccioVent) AS DireccioVent,
  MAX(VelocitatVent) AS VelocitatVent,
  MAX(Irradiacio) AS Irradiacio,
  MAX(Insolacio) AS Insolacio,
  MAX(Precipitacio) AS Precipitacio
FROM meteorologiaTemp2 GROUP BY id;

DROP TABLE meteorologiaTemp;
DROP TABLE meteorologiaTemp1;
DROP TABLE meteorologiaTemp2;
```

***Imatge 31**. Arxiu meteorologia.sql amb les consultes HiveQL per a la càrrega i tractament de les dades meteorològiques.*

### *Càrrega i tractament de les dades d'inversió tèrmica*

```
#! /bin/bash

inici_s=`date +%s`

hive -f inversio_termica.sql

fi_s=`date +%s`
let total_s=$fi_s-$inici_s

echo
echo "--------------------------- FI ---------------------------"
echo "Dades de d'inversió tèrmica carregades i tractades correctament."
echo "Temps de processament: "$total_s" segons"
echo "--------------------------- FI ---------------------------"
```

***Imatge 32**. Arxiu inversio_termica.sh per a la càrrega i tractament de les dades d'inversió tèrmica.*

```
--Càrrega de les dades d'inversió tèrmica
CREATE TABLE estacioAlta (Data String, Hora String, Temperatura Float)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
LINES TERMINATED BY '\n'
STORED AS TEXTFILE;

CREATE TABLE estacioBaixa (Data String, Hora String, Temperatura Float)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
LINES TERMINATED BY '\n'
STORED AS TEXTFILE;

LOAD DATA LOCAL INPATH 'Aixas.txt' INTO TABLE estacioAlta;
LOAD DATA LOCAL INPATH 'BordaVidal.txt' INTO TABLE estacioBaixa;

--Creació de la clau principal
CREATE TABLE estacioAltaID AS SELECT concat(Data,'-',Hora) AS id, Temperatura
        FROM estacioAlta;

CREATE TABLE estacioBaixaID AS SELECT concat(Data,'-',Hora) AS id, Temperatura
        FROM estacioBaixa;

--Càlcul dels valors d'inversió tèrmica
CREATE TABLE inversioTemp AS SELECT a.id, (a.Temperatura-b.Temperatura) AS Inversio
FROM estacioAltaID a JOIN estacioBaixaID b
ON (a.id = b.id);

CREATE TABLE inversio (id String, Inversio Float)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
LINES TERMINATED BY '\n'
STORED AS TEXTFILE;

--Eliminació dels valors nuls
INSERT OVERWRITE TABLE inversio SELECT * FROM inversioTemp  WHERE Inversio IS NOT NULL;

DROP TABLE estacioAlta;
DROP TABLE estacioBaixa;
DROP TABLE estacioAltaID;
DROP TABLE estacioBaixaID;
DROP TABLE inversioTemp;
```

***Imatge 33**. Arxiu inversio_termica.sql amb consultes HiveQL per a la càrrega i tractament de les dades d'inversió tèrmica.*

## *Càrrega i tractament de les dades de trànsit*

```
#! /bin/bash

inici_s=`date +%s`

#Creació de la taula temporal on s'emmagatzemaran les dades prèviament tractades
hive -f 1CrearTaulaTemporal.sql

echo "-------------------------------------------------------------"
echo "Taula Trànsit creada."
echo "-------------------------------------------------------------"

#Obtenció dels directors on es troben les dades de trànsit
ls -d */ > directoris.txt

#Per a cada directori es tracten els arxius amb dades de trànsit
for directori in $(cat directoris.txt)
        do

        direc=${directori%?}

        cd $directori

        ls *.txt > taules.txt

        #Càrrega i tractament de cadascun dels arxius amb dades de trànsit
        for line in $(cat taules.txt)
        do
                #Càrrega i tractament d'un arxiu mensual
          hive -hiveconf flag2=${line%%.*} -f ../2CarregarDadesSensor.sql

          #Càrrega de les dades tractades a la taula temporal
          hive -hiveconf flag1=${direc} -f ../3CarregarDadesTaulaTemporal.sql

          echo "-------------------------------------------------------------"
                echo "Dades del fitxer: " $line " tractades i carregades."
                echo "-------------------------------------------------------------"
        done

        rm taules.txt

        echo "-------------------------------------------------------------"
        echo "S'han tractat les dades del directori: "$direc
        echo "-------------------------------------------------------------"

        cd ..

done

rm directoris.txt

#Suma dels valors dels diversos sensors i elimianció dels casos amb valors nuls
hive -f 4CarregarDadesFinal.sql

fi_s=`date +%s`
let total_s=$fi_s-$inici_s
echo
echo "--------------------------- FI ---------------------------"
echo "Dades de trànsit carregades i tractades."
echo "Temps de processament: $total_s segons"
echo "--------------------------- FI ---------------------------"
```

***Imatge 34**. Arxiu transit.sh per a la càrrega i tractament de les dades de trànsit.*

```
--Creació de la taula temporal
CREATE TABLE TransitTemp (id String, Intensitat int)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
LINES TERMINATED BY '\n'
STORED AS TEXTFILE;
```

***Imatge 35**. Arxiu 1CrearTaulaTemporal.sql amb la consulta HiveQL per a la creació de la taula temporal de trànsit.*

```
CREATE TABLE taula0 (Data String, Hora String, Velocitat int, Ocupacio int, Intensitat int, Noclassificats
int, VehiclesHora int, Apunts int, ApuntsValids int, Error String)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
LINES TERMINATED BY '\n'
STORED AS TEXTFILE;

SET arxiu=${hiveconf:flag2}.txt;

--Càrrega de les dades d'un dels arixus mensuals del sensor
LOAD DATA LOCAL INPATH '${hiveconf:arxiu}' INTO TABLE taula0;

--Generació de la calu principal
CREATE TABLE taula2 AS SELECT concat(Data,'-',concat(substr(Hora, 0, 2),':00')) AS id, Intensitat
        FROM taula0;

--Conversió dels registrs per minuts a registres horaris
CREATE table taula3  AS SELECT id, SUM(Intensitat) AS Intensitat FROM taula2 GROUP BY id;

DROP TABLE taula0;
DROP TABLE taula2;
```

***Imatge 36**. Arxiu 2CarregarDadesSensor.sql amb les consultes HiveQL per a la càrrega de les dades mensuals d'un sensor.*

```
--Càrrega de les dades tractades d'un arxiu mensual a la taula temporal
INSERT INTO TABLE TransitTemp SELECT * FROM taula3;
DROP TABLE taula3;
```

***Imatge 37**. Arxiu 3CarregarDadesTaulaTemporal.sql amb la consulta HiveQL per a la càrrega de les dades mensuals.*

```
--Suma dels valors d'intensitat dels tres sensors
CREATE TABLE taula4  AS SELECT id, SUM(Intensitat) AS Intensitat
FROM TransitTemp
GROUP BY id;

DROP TABLE TransitTemp;

CREATE TABLE transit (id String, Intensitat int)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
LINES TERMINATED BY '\n'
STORED AS TEXTFILE;

--Eliminació dels valors nuls
INSERT OVERWRITE TABLE transit SELECT * FROM taula4 WHERE Intensitat IS NOT NULL;

DROP TABLE taula4;
```

***Imatge 38**. Arxiu 4CarregarDadesFinal.sql amb les consultes HiveQL per a l'obtenció de les dades de trànsit finals.*

## *Càrrega i tractament de les dades de contaminació atmosfèrica*

```
#! /bin/bash
inici_s=`date +%s`

#Càrrega i tractament de les dades de contaminació
hive -f contaminacio.sql

fi_s=`date +%s`
let total_s=$fi_s-$inici_s
echo "--------------------------- FI ---------------------------"
echo "Dades de contaminació carregades i tractades correctament."
echo "Temps de processament: $total_s segons"
echo "--------------------------- FI ---------------------------"
```

***Imatge 39**. Arxiu contaminacio.sh per a la càrrega i tractament de les dades de contaminació.*

```
CREATE TABLE contaminacioTemp (Organisme String, Estacio String, Mesura String, Constituent String,
Unitats String, Data String, Valor Float)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
LINES TERMINATED BY '\n'
STORED AS TEXTFILE;

--Càrrega de les dades de contaminació
LOAD DATA LOCAL INPATH 'contaminacio.txt' INTO TABLE contaminacioTemp;

--Projeccó de la taula inicial per a reduir els camps utilitzats
CREATE TABLE contaminacioTemp1 AS SELECT Data, Valor FROM contaminacioTemp;

CREATE TABLE contaminacioTemp2 (id String, Valor Float)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
LINES TERMINATED BY '\n'
STORED AS TEXTFILE;

--Eliminació dels casos amb valors nuls
INSERT OVERWRITE TABLE contaminacioTemp2 SELECT * FROM contaminacioTemp1
   WHERE Valor IS NOT NULL;

CREATE TABLE contaminacioTemp3 AS
SELECT
   regexp_replace(id, ' ', '-') AS id,
   Valor
FROM contaminacioTemp2;

CREATE TABLE contaminacio(id String ,Valor Float) clustered by (id) into 2 buckets stored as orc
TBLPROPERTIES('transactional'='true');

INSERT INTO TABLE contaminacio SELECT * FROM contaminacioTemp3;

--Afegir la classe a la taula contaminació
ALTER TABLE contaminacio ADD COLUMNS (classe int);

--Càlcul de la classe
UPDATE contaminacio SET classe=1 WHERE Valor BETWEEN 0 AND 50.00;
UPDATE contaminacio SET classe=2 WHERE Valor BETWEEN 50.01 AND 100.00;
UPDATE contaminacio SET classe=3 WHERE Valor BETWEEN 100.01 AND 200.00;
UPDATE contaminacio SET classe=4 WHERE Valor BETWEEN 200.01 AND 400.00;
UPDATE contaminacio SET classe=5 WHERE Valor>400.01;

DROP TABLE contaminacioTemp;DROP TABLE contaminacioTemp1;
DROP TABLE contaminacioTemp2;DROP TABLE contaminacioTemp3;
```

***Imatge 40**. Arxiu contaminacio.sql amb les consultes HiveQL per a la càrrega i tractament de les dades de contaminació.*

### *Obtenció del conjunt de casos, el conjunt d'entrenament i el conjunt de test*

```
#! /bin/bash

inici_s=`date +%s`

#Generació del conjunt de casos, el conjunt d'entrenament i el conjunt de test
hive -f conjunt_casos.sql

fi_s=`date +%s`
let total_s=$fi_s-$inici_s

echo
echo "--------------------------- FI ---------------------------"
echo "Les taules s'han unit."
echo "Temps de processament: $total_s segons"
echo "--------------------------- FI ---------------------------"
```

***Imatge 41**. Arxiu conjunt_casos.sh per a la generació del conjunts de: casos, entrenament i test.*

```
--Unió de les taules amb les dades meteorològiques, d'inversió tèrmica, de trànsit i de contaminació
CREATE TABLE conjuntcasos AS SELECT m.id, m.Temperatura, m.HumitatRelativa, m.DireccioVent,
m.VelocitatVent, m.Irradiacio, m.Insolacio, m.Precipitacio, i.Inversio, t.Intensitat, c.classe
FROM meteorologia m JOIN inversio i ON (m.id=i.id) JOIN transit t ON (i.id=t.id) JOIN contaminacio c ON
(t.id=c.id);

CREATE TABLE entrenament (id String, Temperatura Float, HumitatRelativa int, DireccioVent int,
VelocitatVent int, Irradiacio int, Insolacio int, Precipitacio Float, Inversio Float, Intensitat int, classe int)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
LINES TERMINATED BY '\n'
STORED AS TEXTFILE;

--Creació del conjunt d'entrenament a partir d'un subconjunt aleatòri del 70% dels casos
INSERT OVERWRITE TABLE entrenament SELECT * FROM conjuntcasos SORT BY RAND() LIMIT 4250;

--Creació del conjunt de test a partir dels casos no utilitzats en el conjunt d'entrenament
CREATE TABLE test AS
SELECT *
FROM   conjuntcasos c
WHERE  NOT EXISTS (
  SELECT 1
  FROM   entrenament e
  WHERE  c.id = e.id);
```

***Imatge 42**. Arxiu conjunt_casos.sql amb les consultes HiveQL a la generació del conjunts de: casos, entrenament i test.*

### *Obtenció i avaluació del model d'aprenentatge*

```
#! /bin/bash

inici_s=`date +%s`

hdfs dfs -mkdir /user/mahout/
hdfs dfs -mkdir /user/mahout/tfg

hive -hiveconf flag="entrenament" -f 7CarrerarDadesHDFS.sql
mv dades/000000_0 dades/entrenament.txt
hdfs dfs -put dades/entrenament.txt /user/mahout/tfg
rm dades/entrenament.txt

hive -hiveconf flag="test" -f 7CarrerarDadesHDFS.sql
mv dades/000000_0 dades/test.txt
hdfs dfs -put dades/test.txt /user/mahout/tfg
rm dades/test.txt

mahout org.apache.mahout.classifier.df.tools.Describe -p /user/mahout/tfg/entrenament.txt -f
/user/mahout/tfg/entrenament.info -d I 9 N L

hadoop jar $MAHOUT_HOME/mahout-examples-0.11.0-job.jar
org.apache.mahout.classifier.df.mapreduce.BuildForest -Dmapred.max.split.size=1874231 -d
/user/mahout/tfg/entrenament.txt -ds /user/mahout/tfg/entrenament.info -sl 5 -p -t 100 -o
/user/mahout/tfg/model

hadoop jar $MAHOUT_HOME/mahout-examples-0.11.0-job.jar
org.apache.mahout.classifier.df.mapreduce.TestForest -i /user/mahout/tfg/test.txt -ds
/user/mahout/tfg/entrenament.info -m /user/mahout/tfg/model -a -mr -o
/user/mahout/tfg/prediccions

fi_s=`date +%s`
let total_s=$fi_s-$inici_s

echo
echo "---------------------------- FI ----------------------------"
echo "Dades per a la generació del model preparades."
echo "Temps de processament: $total_s segons"
echo "---------------------------- FI ----------------------------"
```

***Imatge 43**. Arxiu preparacio.sh per a la preparació de les dades en el sistema de fitxers distribuït HDFS.*

```
--Càrrega al sistema de fitxers HDFS del conjunt de casos (entrenament i test)
INSERT OVERWRITE LOCAL DIRECTORY '/home/ubuntu/dades/mapreduce8/model/dades'
ROW FORMAT
DELIMITED FIELDS TERMINATED BY ','
lines terminated by '\n'
STORED AS TEXTFILE
SELECT * FROM ${hiveconf:flag};
```

***Imatge 44**. Arxiu CarrerarDadesHDFS.sql amb la consulta HiveQL per a la càrrega dels conjunts d'entrenament i test al sistema de fitxers distribuït HDFS.*

Finalment, el següent script sol·licita a l'usuari els paràmetres de l'algorisme *Random forest*: nom del model, implementació parcial, número d'arbres i número d'atributs escollits aleatòriament en cada node del clúster.

```
#! /bin/bash

inici_s=`date +%s`

let arg_p
let arg_nom
let arg_arbres
let arg_atributs

while true; do
   read -p "Nom del model? " yn
   case $yn in
            [A-Za-z0-9]* ) arg_nom=$yn; break;;
   esac
done

while true; do
   read -p "Vols utilitzar implementació parcial? (S/N) " yn
   case $yn in
      [Ss]* ) arg_p='-p'; break;;
      [Nn]* ) arg_p=''; break;;
      * ) echo "Sisplau, respon sí (S) o no (N).";;
   esac
done

while true; do
   read -p "Número d'arbres? " yn
   case $yn in
      [0-9]* ) arg_arbres=$yn; break;;
      * ) echo "Sisplau, introdueix el número d'arbres.";;
   esac
done

while true; do
   read -p "Número d'atributs escollits aleatòriament en cada node? " yn
   case $yn in
      [0-9]* ) arg_atributs=$yn; break;;
      * ) echo "Sisplau, introdueix el número d'atributs.";;
   esac
done

hdfs dfs -mkdir /user/mahout/tfg/$arg_nom

hadoop jar $MAHOUT_HOME/mahout-examples-0.11.0-job.jar
org.apache.mahout.classifier.df.mapreduce.BuildForest -Dmapred.max.split.size=2336740 -d
/user/mahout/tfg/entrenament.txt -ds /user/mahout/tfg/entrenament.info -sl $arg_atributs $p -t
$arg_arbres -o /user/mahout/tfg/$arg_nom/model

hadoop jar $MAHOUT_HOME/mahout-examples-0.11.0-job.jar
org.apache.mahout.classifier.df.mapreduce.TestForest -i /user/mahout/tfg/test.txt -ds
/user/mahout/tfg/entrenament.info -m /user/mahout/tfg/model -a -mr -o
/user/mahout/tfg/$arg_nom/prediccions

fi_s=`date +%s`
let total_s=$fi_s-$inici_s

echo
echo "--------------------------- FI ---------------------------"
echo "Model d'aprenentage automàtic generat"
echo "Temps de processament: $total_s segons"
echo "--------------------------- FI ---------------------------"
```

***Imatge 45**. Arxiu model.sh per a la generació i avaluació del model d'aprenentatge.*