Universitat Oberta de Catalunya

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

Fidel Bonet Vilela

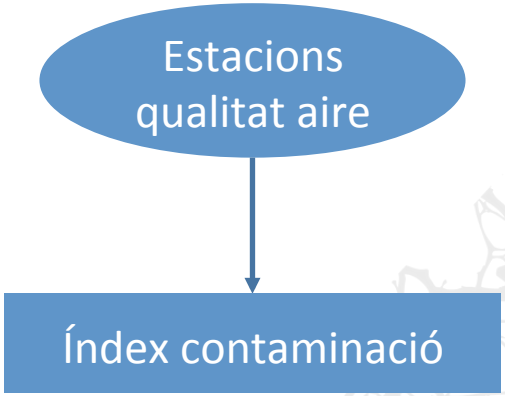
Treball Final de Grau **ENGINYERIA INFORMÀTICA**
Àrea **INTEL·LIGÈNCIA ARTIFICIAL**

Consultor **DR. DAVID ISERN ALARCÓN**
Professor **DR. CARLES VENTURA ROYO**

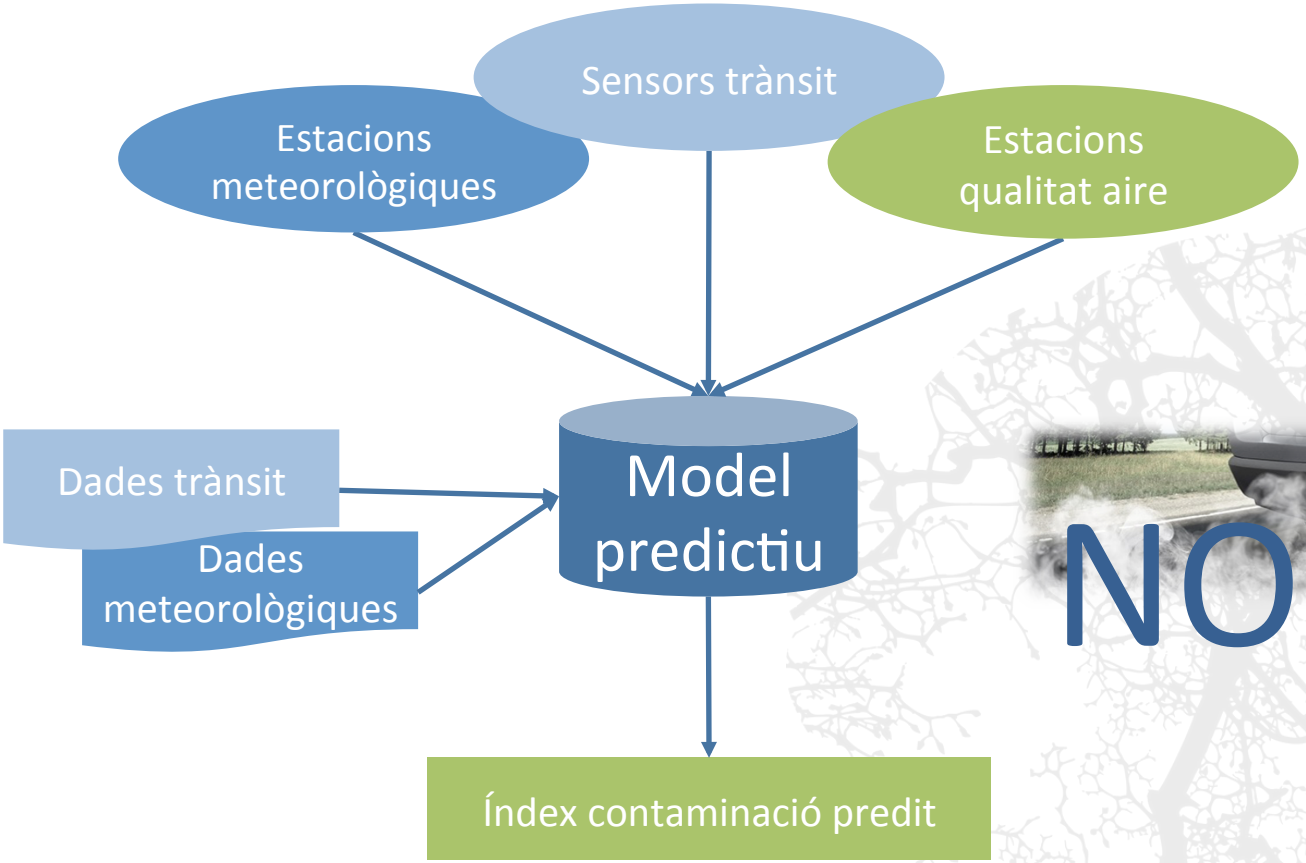
31 DE MAIG DE 2017

ÍNDEX

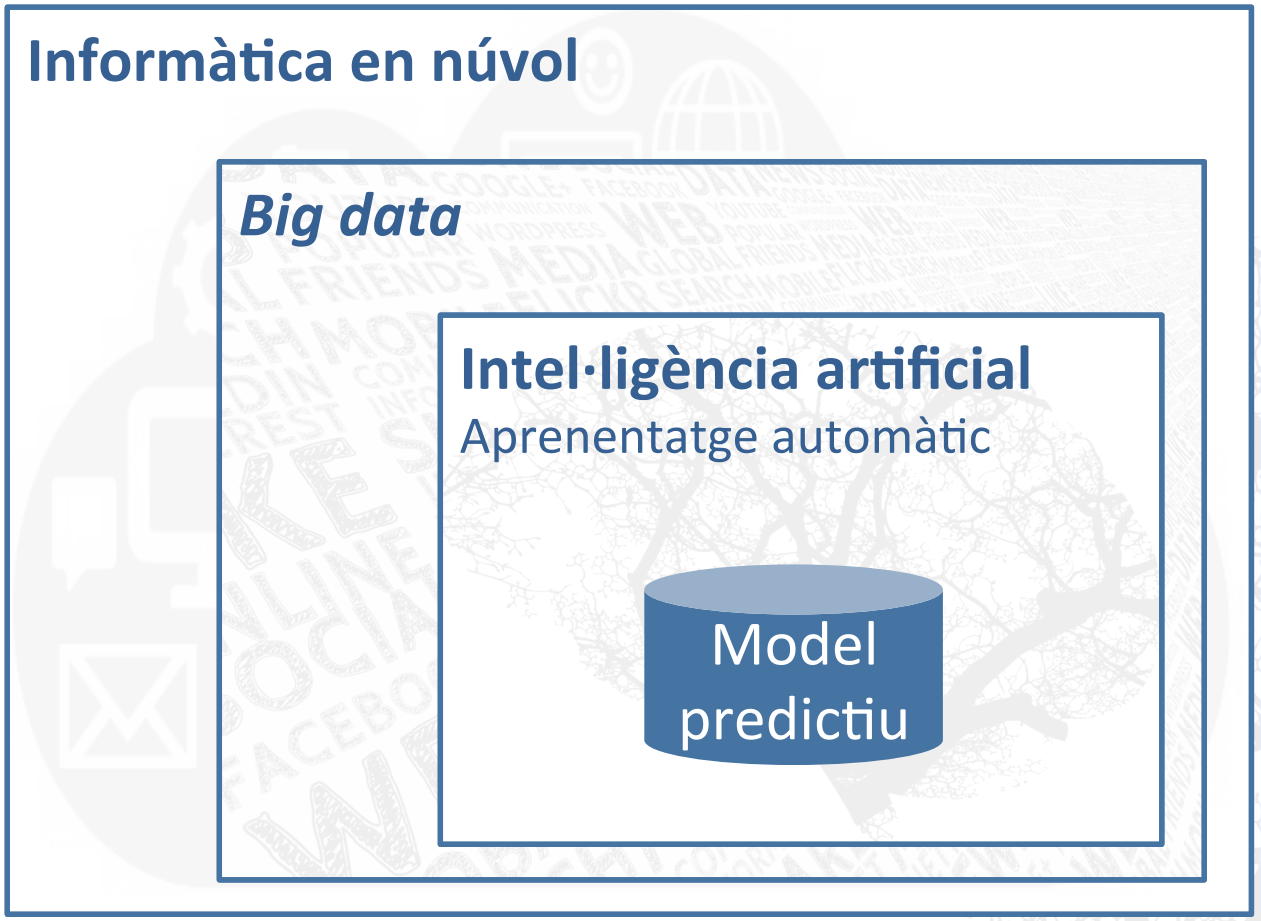
1. Introducció



1. Introducció



1. Introducció

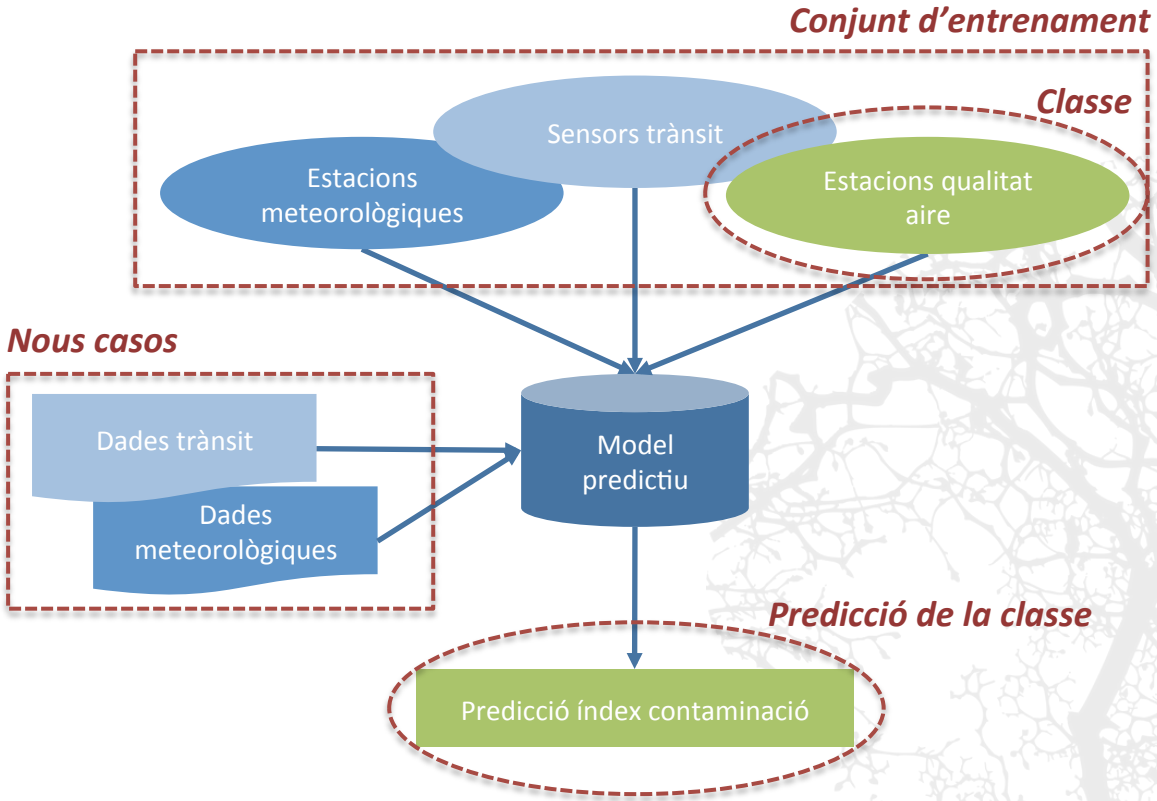


2. Aprenentatge automàtic (*machine learning*)

Tècniques per a proveir a les màquines de la capacitat d'aprendre i millorar a partir de l'experiència.

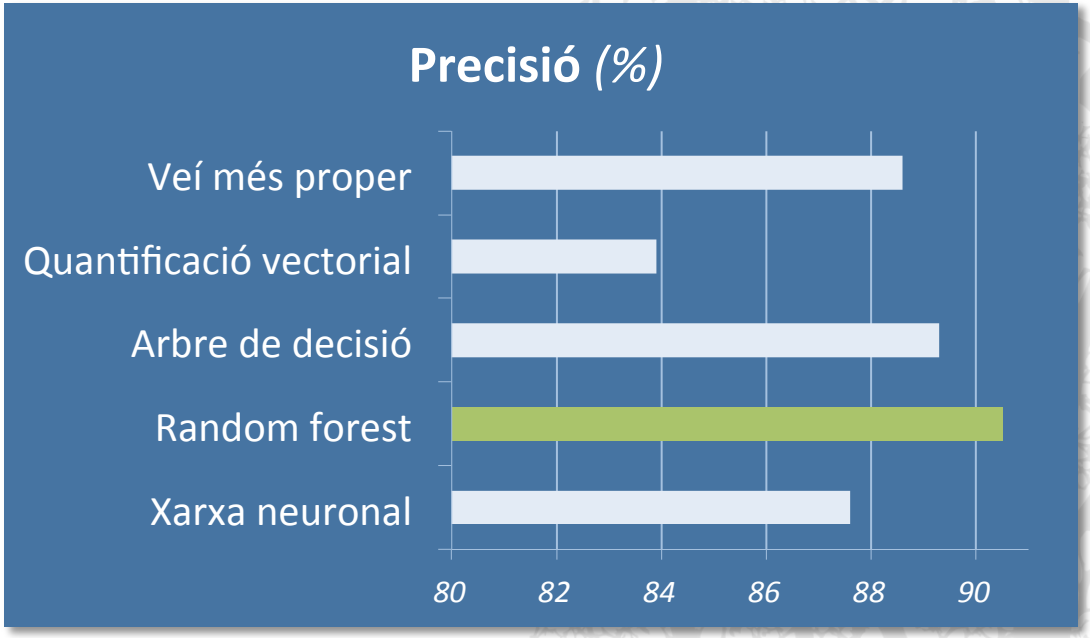
- No supervisat
- Per reforç
- Supervisat ➡ *Classificació*

2. Aprenentatge automàtic (*machine learning*)



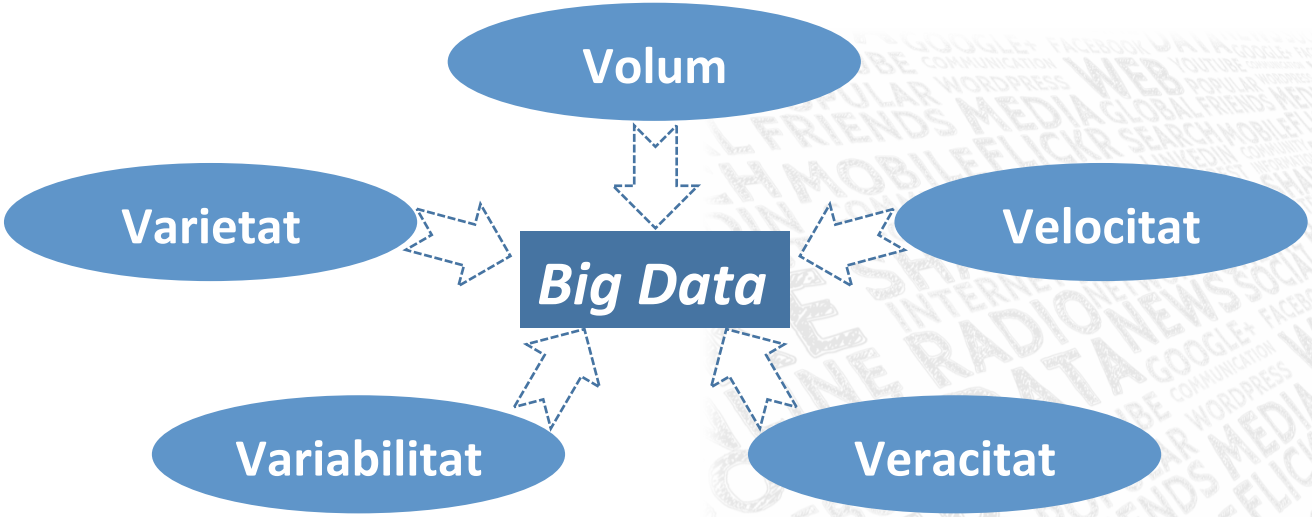
Classificació

- Veí més proper
- Classificació basada en quantificació vectorial
- K-mitjanes (*K-means*)
- Màquina de vectors de suport (SVM)
- Arbres de decisió
- *Random forest*
- Xarxes neuronals
- Classificadors lineals



3. *Big data*

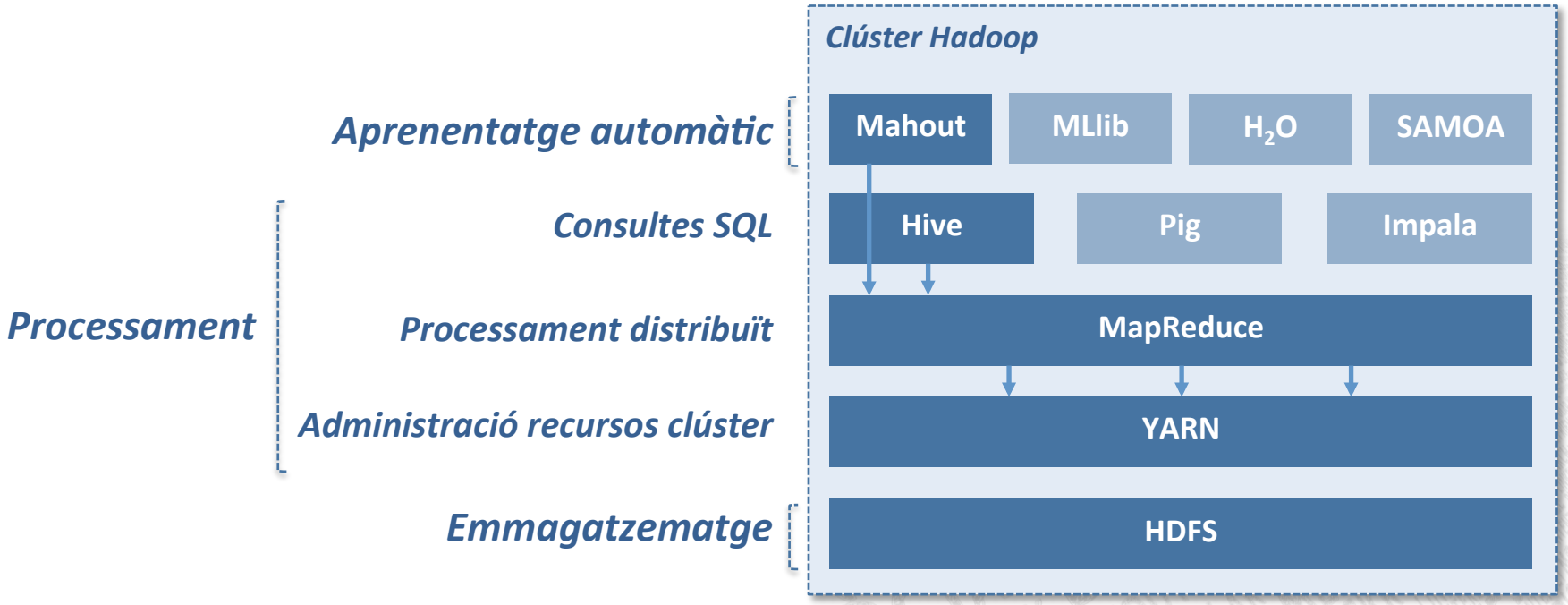
*"Conjunt de tècniques i tecnologies per al tractament de dades en entorns de gran **volum**, **varietat** d'orígens i en els que la **velocitat** de resposta és crítica" (Laney, 2001).*



Principals entorns de treball *big data*

Processament per lots	Híbrids	Processament de fluxos
Hadoop	Spark	Storm
	Flink	Samza

Apache Hadoop



4. Informàtica en núvol (*cloud computing*)

"Model que permet l'accés en xarxa de forma ubiqua, adequada i sota demanda a un conjunt compartit de recursos informàtics configurables" (Mell, 2011).

Models

- *Infraestructura com a servei (IaaS)*
- *Plataforma com a servei (PaaS)*
- *Programari com a servei (SaaS)*

Plataformes

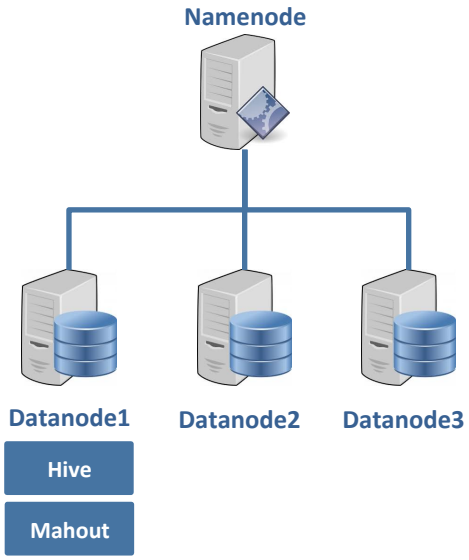
Microsoft Azure

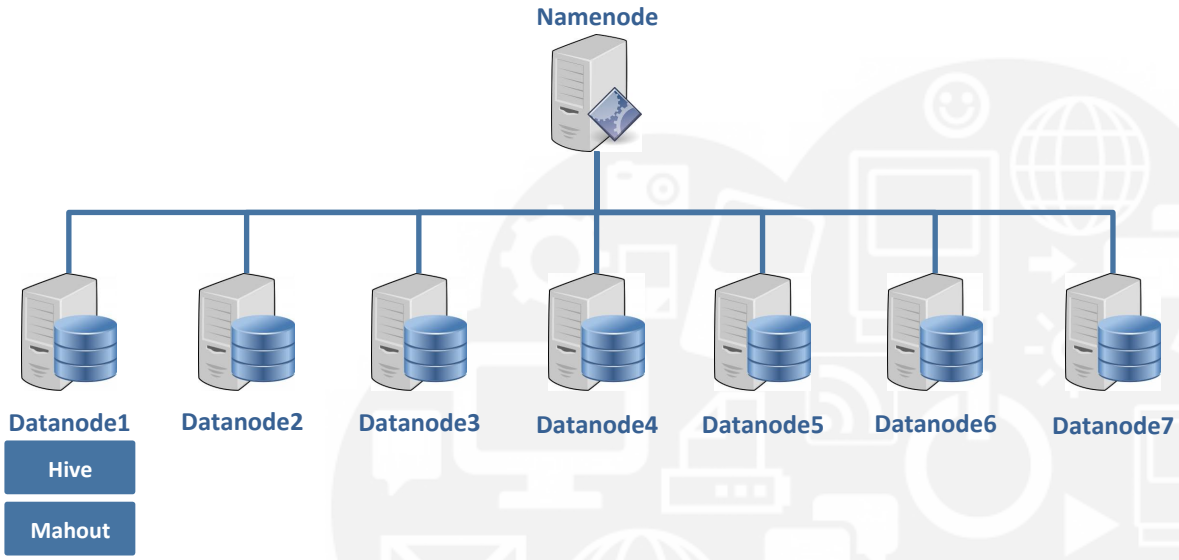
Google Cloud Platform

amazon web services™

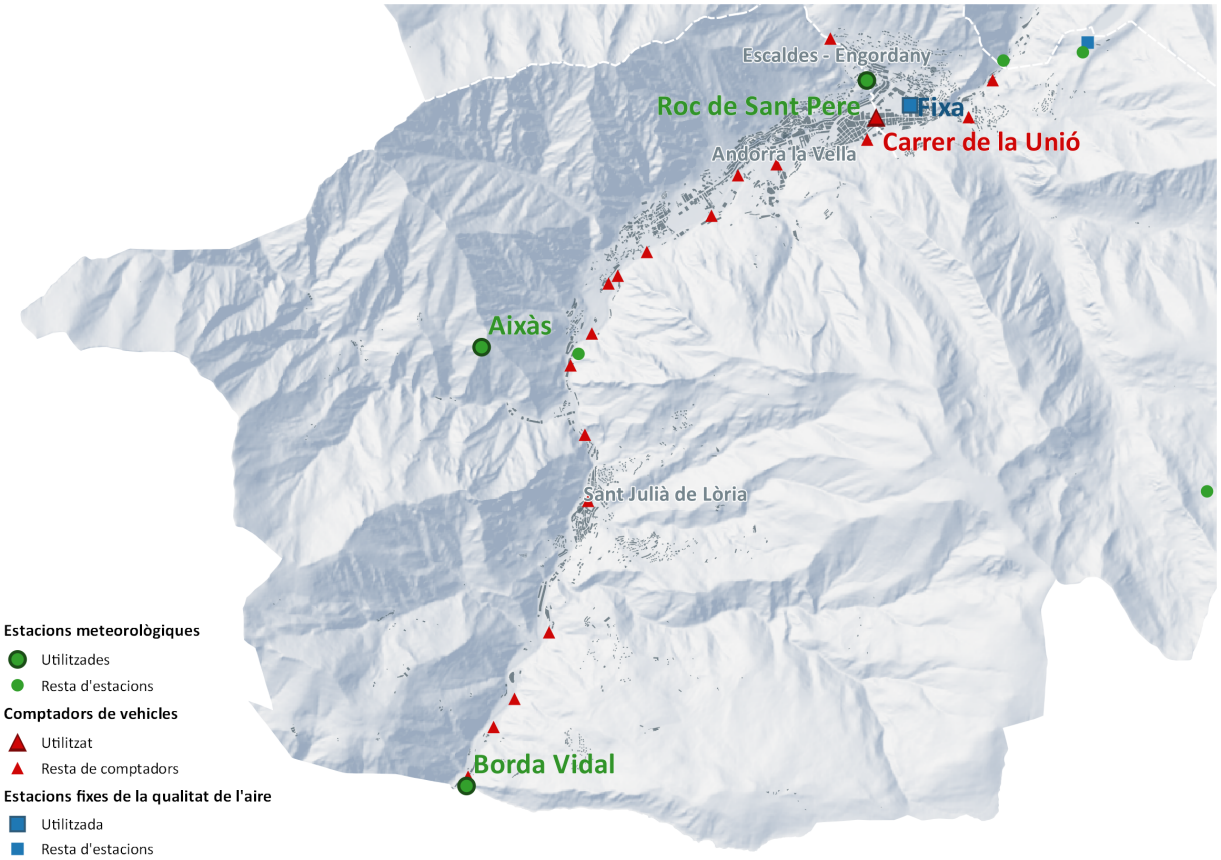
Clústers Hadoop

- *Pseudodistribuït: 2 nodes*
- *Distribuïts: 4 i 8 nodes*





5. Obtenció del model predictiu



Dades

Meteorològiques

- *Hora*
- *Data*
- *Temperatura*
- *Humitat relativa*
- *Direcció del vent*
- *Velocitat del vent*
- *Irradiació solar*
- *Insolació*
- *Precipitació*
- *Inversió tèrmica*

Trànsit

- *Intensitat*

Contaminació

- *Diòxid de nitrogen (NO_2)*

5. Obtenció del model predictiu

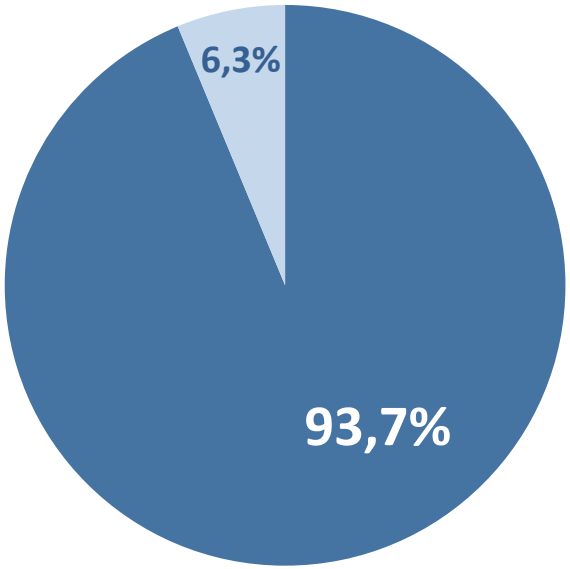
Random forest

- Variant dels arbres de decisió
- Selecció aleatòria d'atributs per obtenir conjunt d'arbres de decisió
- Agregació de bootstrap
- Millor estabilitat i precisió
- Reducció de la variància
- Evita el sobreajustament

Paràmetre	Valor
Arbres	100
Implementació parcial	Sí
Atributs	5
Mida particions *(bytes)*	2.336.740

Anàlisi resultats obtinguts

Precisió del classificador

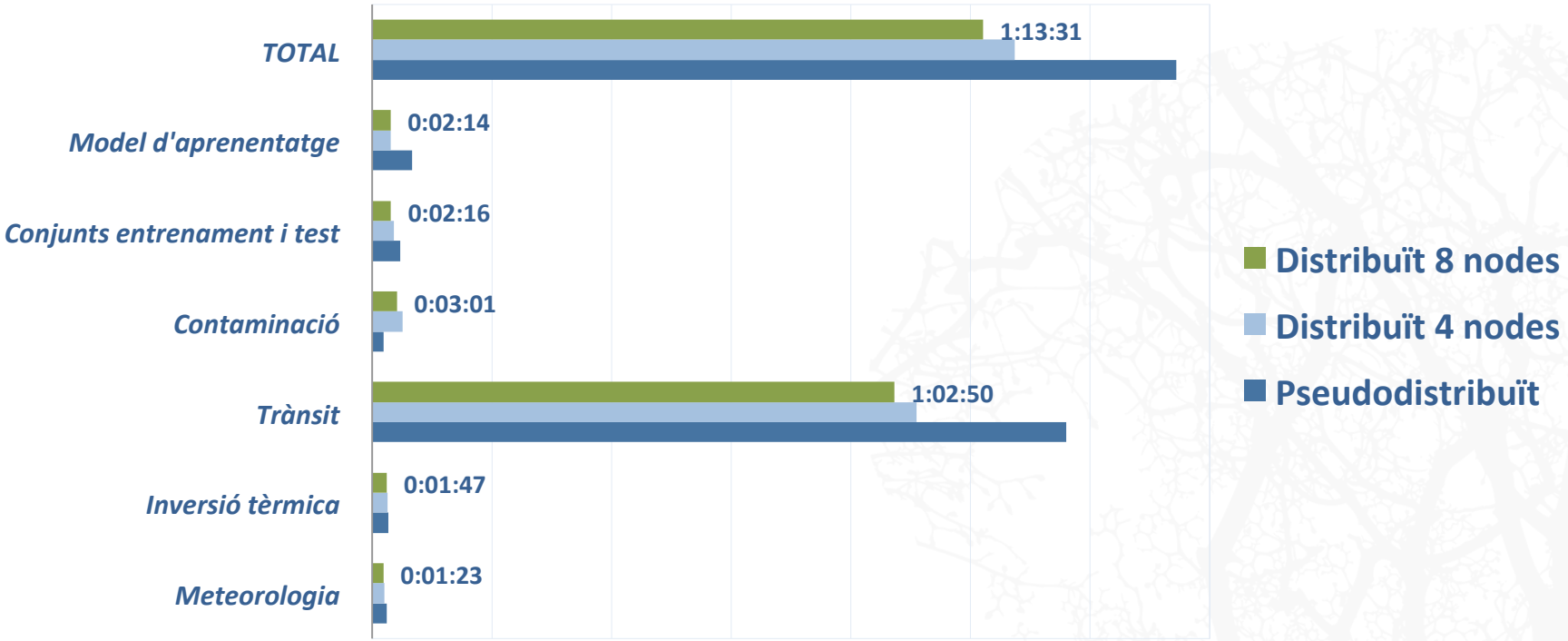


Pes dels atributs en el model

1. **Intensitat de circulació**
2. **Inversió tèrmica**
3. **Temperatura**
4. Humitat
5. Irradiació solar
6. Dia de la setmana
7. Direcció del vent
8. Velocitat del vent
9. Insolació
10. Precipitació

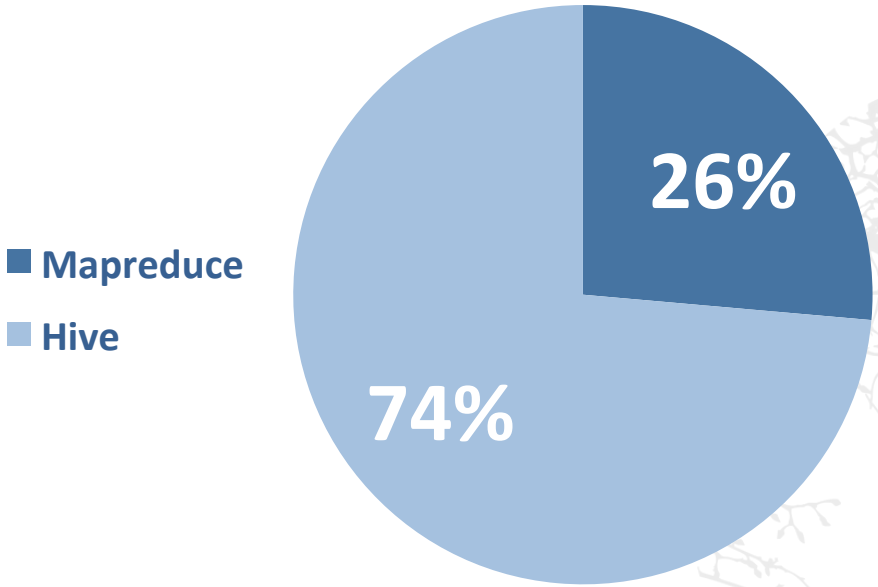
Anàlisi de rendiment

Temps de processament



Anàlisi de rendiment

Temps de processament dels processos MapReduce



6. Conclusions i futures línies de treball

- Hipòtesi inicial confirmada.
- Ús del model d'aprenentatge automàtic per evitar episodis de contaminació:
  - Gestors públics: regulació del trànsit.
  - Ciutadans: ús del transport públic.
  - El model es pot utilitzat en anys posteriors.
- Tècniques d'intel·ligència artificial vàlides.
- Algorisme amb millors resultats: *Random forest*.
- Hadoop:
  - Dificultat en instal·lar directament algunes eines.
  - Bons resultats. Nombroses eines *big data*.
- Informàtica en núvol: implementació solucions *big data* sense adquirir maquinari i a un cost molt reduït.
- Futures línies de treball:
  - Ús del model d'aprenentatge amb dades a temps real en un entorn de processament de fluxos.
  - Anàlisi espacial del comportament del model.
  - Obtenció de nous models per a altres contaminants o altres àmbits.

Universitat Oberta de Catalunya

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

Fidel Bonet Vilela

Gràcies per al seva atenció

git

https://github.com/fbgis/TFG