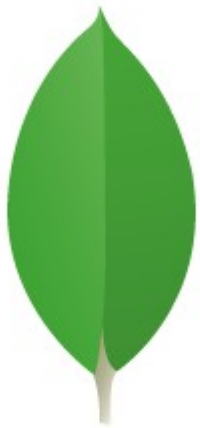
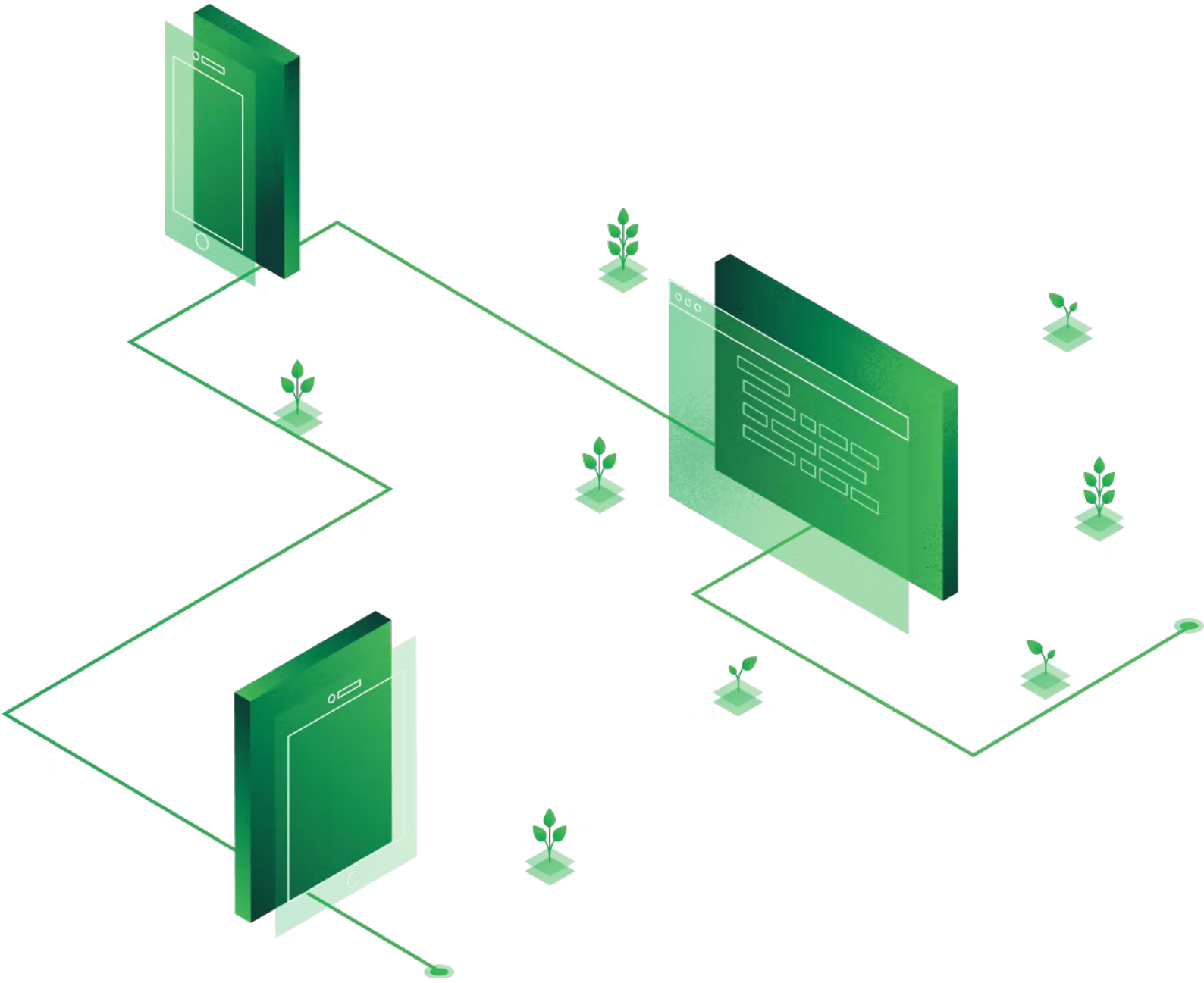


Mongo DB

Joan Paz - 1598851
Carla Martínez - 1606635
Dídac Alvarez - 1601114
Beatriz Salgado - 1605534



INTRODUCCIÓ

Aquest informe tracta del projecte de l'assignatura de Bases de Dades, amb Mongo DB. Mongo DB és un programari de codi obert, per a la creació i gestió de base de dades orientada a documents. Durant els primers mesos de semestre, hem estat estudiant el sistema i ara ens toca posar-lo en pràctica amb aquest treball, que tracta sobre la creació d'una base de dades mèdica, amb pacients i diagnòstics.

REPARTIMENT DEL TREBALL – ORGANITZACIÓ

¿Com ens hem repartit el treball? En començar el projecte vam quedar en què la repartició individual del treball no era la nostra primera opció, ens agradava la idea de fer les coses conjuntament, sabent què feia cada persona i entenent el projecte des d'una visió global i no "a parts". Per tant, la idea va ser aprofitar les sales d'estudi de les biblioteques i anar tots en conjunt a fer els diferents exercicis del projecte. Quan teníem els exercicis fets, decidíem qui els pujava per tal que al github quedés també de la forma més igualada possible.

A la part de les consultes, sí que va haver-hi una repartició del treball, de forma que tots els integrants del grup fessin la mateixa quantitat de consultes aproximadament. Posteriorment, les vam ajuntar totes en un mateix script i un dels integrants va fer el lliurament la github.

EXERCICI 1

En la col·lecció cases hem relacionat els pacients amb els seus nòduls i amb els escàners utilitzats. Hem fet servir **embedded** pels nòduls perquè és una relació 1-n amb pocs n. Per tant, la millor solució és encastar.
Per altra banda, per relacionar tot això amb els escàners, hem fet servir **referència** amb l'id de cadascun dels escàners. Hem decidit fer-ho així perquè fos més fàcil a l'hora de fer les consultes i evitar duplicació d'informació.

En la col·lecció method_output, hem relacionat les entitats method i experiment amb **referència extesa**. Com és una relació 1-n, ens hem decantat per aquesta opció perquè, com accedirem molts cops, és recomanable utilitzar referència per evitar redundàncies. Hem cregut convenient identificar els experiments de cada mètode amb el seu id i al mètode al qual pertanyen per poder diferenciar-los. D'aquesta manera podrem accedir a ells per fer les consultes. La referència extesa, l'hem aplicat a la col·lecció method_output on, per cada mètode, identifiquem els experiments que té per referència extesa, posant l'id del mètode i l'id de l'experiment.

A més, en la col·lecció experiments, hem creat un array que guarda per **referència** els id dels pacients que utilitzen aquell mètode. Això ho fem per tenir més facilitat de fer joins quan fem les consultes.

Per últim, en la col·lecció training_patient, fem **referència** del nòduls per poder realitzar els joins en les consultes.

EXERCICI 2

COL·LECCIONS

1. Col·lecció cases

Hem creat la col·lecció cases i hem decidit que cada document d'aquesta col·lecció es guardaria amb els següents atributs:

- **_id** → ID del pacient, ex: LIDC-IDRI-0071
- **Age** → L'edat del pacient, ex : 21
- **Gender** → Man o Women
- **Diagnosis_Patient** → Bening o Maling
- **CTID** → ID del escaner , ex: 5

Cada pacient de la nostra col·lecció té aquests atributs disponibles. A més afegim un atribut de tipus array anomenat 'Noduls' que és una llista amb els nòduls que té el pacient. Cada nòdul és un document i fem servir el patró **embedded** per desar tota l'informació de cada nòdul.
Els documents dels nòduls tenen els següents atributs:

- **NoduleID:** El número del nòdul (1,2,..)
- **Diagnosis_nodul:** Benign o Malign
- **Position:** {x: la posició x, y: la posició y, z: la posició z} més fàcil per accedir
- **Diàmetre**: El diàmetre en mm

L'estructura queda de la següent manera:

{PatiendID: LIDC-IDRI-0071,

Age: 80,

Gender: Man,

DiagnosisPatient: Benign,

Nodules: [{

NoduleID: 1,

DiagnosisNodule: Benign,

Position: {'x': 376, y: 142 , 'z': 165}

Diameter: 5,5

CTID: 1

}]}

2. Col·lecció experiments

Hem creat la col·lecció experiments i hem decidit que cada document d'aquesta col·lecció es guardaria amb els següents atributs:

- **method_id** → ID del mètode, ex : 19
- **Repetition** → Nombre de repeticions del experiment, ex : 1
- **Train_percentage** → Percentatge del experiment, ex :
- **BenignPrec** → ex : 33.33
- **BenignRec** → ex : 2.44
- **MalignPrec** → ex : 64.91
- **MalignRec** → ex : 97.91
- **pacients_experiments** → Array amb els pacients que han participat en l'experiment.

Cada experiment realitzat de la nostra col·lecció té aquests atributs. Com en el cas anterior, també aquesta col·lecció té un atribut de tipus array anomenat *'patients_experiments'* que és una llista amb els pacients sobre els que s'han realitzat l'experiment. En aquest cas, fem servir el patró de **Referència** per guardar el pacients ja que ja existeix una col·lecció amb la informació d'aquests.

L'estructura queda de la següent manera:

{{'method_id': 19,

'Repetition': 1,

'Train_percentage': 70,

'BenignPrec': 33.33,

'BenignRec': 2.44,

'MalignPrec': 64.91,

'MalignRec': 97.37, 'pacients_experiments' :

['pacienrt_id_1', 'pacient_id_2'] }}

3. Col·lecció method_output

Hem creat la col·lecció method_output i hem decidit que cada document tindrà els següents atributs:

- **_id** → ID del mètode, ex: Method19
- **FeatDescriptor** ex : LBP_HoG_PyFirstOrderShape
- **FeatSelection** ex : mRMR
- **Classifier** ex : Logit
- **Experiments** → Array amb els mètode utilitzat amb les seves repeticions.

En el cas del atribut *"experiments"* utilitzem referència estesa per guardar per cada output d'un mètode el mètode utilitzat amb les seves repeticions, no cal crear una col·lecció nova ja que els experiments ja están guardats en la seva col·lecció propia, així que només cal referenciar-los.

L'estructura queda de la següent manera:

{'_id': 'Method19',
'FeatDescriptor': 'LBP_HoG_PyFirstOrderShape',
'FeatSelection': 'mRMR',
'Classifier': 'Logit',
'Experiments':
[{'Method_id': 19, 'Rep': 1}, {'Method_id':19, 'Rep': 2}] }

4. Col·lecció training_patient

Hem creat la col·lecció training_patient i hem decidit que cada document tindrà els següents atributs:

- **pacient_id** → ID del pacient, ex: LIDC-IDRI-1011
- **Nodules** → Array que conté els nóduls, identificats per l'ID, amb el seu diagnòstic i el train utilitzat.

L'estructura queda de la següent manera:

{'Pacient_id' : 'LIDC-IDRI-1011',
'Nodules':
[{'NoduleID': 1, 'Radiomics_diagnosis': 'Benign', 'train' : 1},
{'NoduleID': 2, 'Radiomics_diagnosis': 'Benign', 'train' : 0}] }

5. Col·lecció scanners

Hem creat la col·lecció scanners on cada document tindrà els següents atributs:

- **_id** → ID del scàner ex: 1
- **Device** → ID del scàner ex: Siemens sensation
- **dataCT** → La data en la que s'ha fet la prova el pacient, ex : 7/15/2018
- **Resolution** → Array que conté les resolucions de les proves realitzades a un pacient.

L'estructura queda de la següent manera:

{'_id' : 1,
'Device': 'Siemens Sensation',
'dataCT': '7/15/2018',
'Resolution':
[{'T' : 0.625, 'TV' : 120, 'TC' : 206}] }

EXERCICI 3

Consulta 1

Hem agrupat els scanners pel seu nom **(Device).** Quan es fa un *"group"* per un id només, et mostra únicament el resultat pel qual agrupes. Per tant, és innecessari fer un *"project".*

Consulta 2

Primer de tot, hem buscat aquells experiments que fossin del mètode 2 i de l'experiment 1. No hem mirat que fos de l'entrenament (train = 1) perquè hem importat les dades de tal manera que només se'ns guardessin els experiments que realment s'utilitzarien. Un cop seleccionat el que volem, veiem que un dels atributs és una llista d'ID que referencien a pacients. Per trobar la informació dels nòduls, és necessari accedir a les dades dels pacients. Per aquesta raó, hem cregut convenient fer un join per poder relacionar els ID de la col·lecció *"experiments"* amb els ID de la col·lecció *"training_patient".*
Abans de fer això, hem fet un *"unwind"* per dividir la llista, ja que no es pot fer un *"lookup"* amb un array. Un cop hem aconseguit això, ja és possible fer el join entre col·leccions. Guardem la informació del pacient relacionada amb l'experiment i mètodes determinats en un array anomenat *"noduls_pacients".*
D'aquest nou array, ens interessen els nòduls. Com no podem fer un recompte idoni, hem fet dos *"unwind"* per poder tenir els diferents nòduls (1 per tupla del resultat). El primer per eliminar la posició on es trobava la informació del pacient (sortia per defecte) i el segon per separar cadascun dels nòduls que contenia un pacient.
Finalment, hem fet el count de tot aquest pipeline.

Consulta 3

Per una banda, tenim el classificador que està a la col·lecció *"methodOutPut"* i, per una altra, tenim la informació dels Benign i Malign a experiments. El que voldrem és relacionar l'atribut de l'ID del mètode, que és comú en ambdues col·leccions. Com tenim els experiments per referència dins d'un llistat, és necessari fer un *"unwind"* d'aquest. A continuació, fem el join, guardant la informació dins d'un atribut class. D'aquest atribut, com és un array que té diferents posicions, hem de fer un *"unwind".* Finalment, agrupem pel classificador que ens interessa i afegim el màxim, el mínim i la mitjana.

Consulta 4

Simplement agrupem per l'atribut gènere de la col·lecció *"cases"*. Això fa que ens separi la informació en homes i dones. Per acabar, fem un "count" de cadascun dels gèneres.

Consulta 5

Busquem aquells pacients els quals la seva llista Nodules sigui més gran que 2 (solució alternativa: *$size:{$gt:2}*). Fem la projecció dels atributs que ens demanen amb el *"project"*.

Consulta 6

El que fem és dir sobre quin element volem ordenar amb el *"sort"* en ordre decreixent. Després, per agafar els 4 mètodes amb més repeticions, agafem del pipeline anterior els 4 primers amb el *"limit"*. Per últim, fem la projecció del que ens demana la consulta.

Consulta 7

Per començar, hem de fer un join dels CTID (id dels escaners) de la col·lecció *“scanners”* i *“cases”*. Guardarem la informació de l'escàner dins de l’atribut *“scanners_info”*. Com té un element 0 en l’array que ens impedeix accedir als seus atributs, hem de fer un *“unwind”*. D’aquesta manera, ja podem agrupar pel pacient i dir amb quin “device” li fan les proves i quina és la data en que es va fer l’escàner.

Consulta 8

El que hem entès en aquesta consulta ha estat que, si tots els nòduls d’un pacient són benignes, aleshores el diagnòstic del pacient també ho serà. Només que un nòdul sigui maligne, el seu diagnòstic també ho serà.
Per tant, simplement busquem aquells pacients que tinguin l’atribut *"Diagnosis_Patient"* a benigne. No fa falta fer un recompte perquè en el resultat de la consulta ja es pot observar.

Consulta 9

Fem un updateMany d’aquells escàners que compleixin la condició que ha estat donada. El que canviarem serà el valor de l’atribut *“ResolutionTV”,* multiiplicant-lo per 1.2 .