Projet Big Data

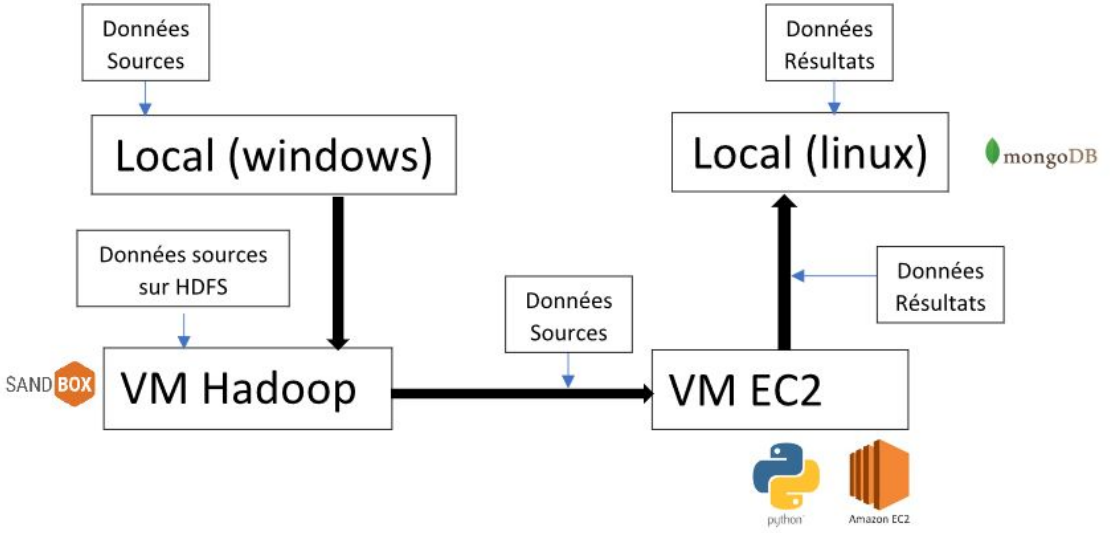
soutenance finale

Chaymae EL ABBADI
Yueming YANG
Xuanlong YU

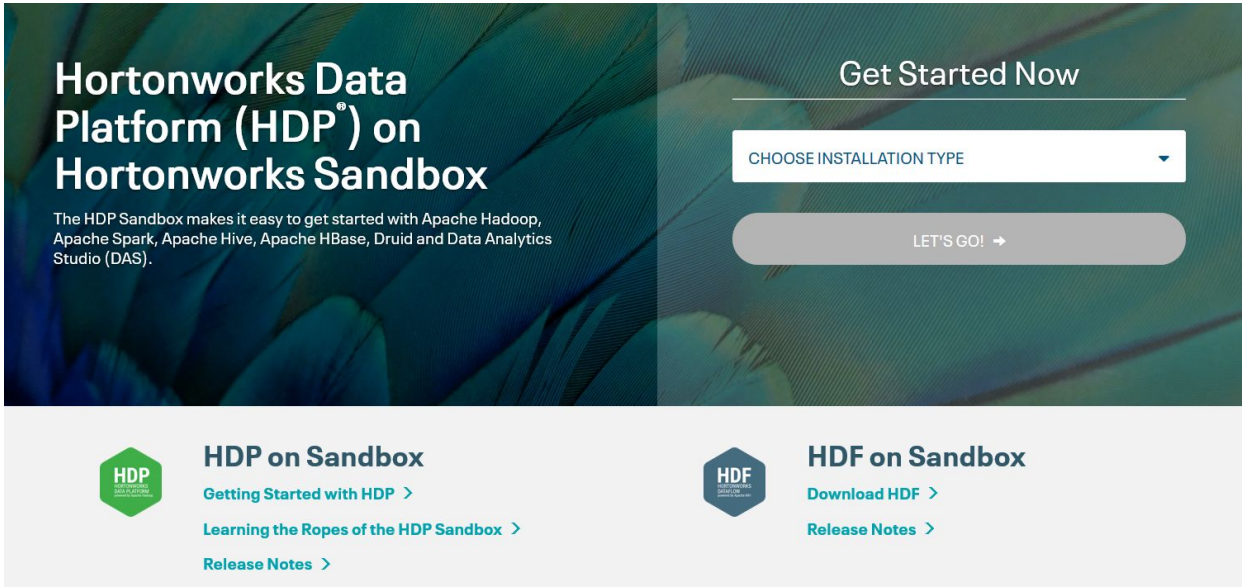
plan

- Introduction
- Etape 0
- Etape 1
- Etape 2
- Etape 3
- Etape 4
- Etape 5
- Conclusion

Introduction



Etape 0 : Installation du VM hadoop



Etape 1 : Machine locale -> HDFS

1.

```
[root@sandbox-hdp ~]# sudo -u hdfs hadoop dfs -mkdir /input
WARNING: Use of this script to execute dfs is deprecated.
WARNING: Attempting to execute replacement "hdfs dfs" instead.

[root@sandbox-hdp ~]# hadoop fs -ls /
Found 14 items
drwxrwxrwt   - yarn   hadoop          0 2018-11-29 17:56 /app-logs
drwxr-xr-x   - hdfs   hdfs            0 2018-11-29 19:01 /apps
drwxr-xr-x   - yarn   hadoop          0 2018-11-29 17:25 /ats
drwxr-xr-x   - hdfs   hdfs            0 2018-11-29 17:26 /atsv2
drwxr-xr-x   - hdfs   hdfs            0 2018-11-29 17:26 /hdp
drwxr-xr-x   - hdfs   hdfs            0 2020-01-30 14:40 /input
drwx------   - livy   hdfs            0 2018-11-29 17:55 /livy2-recovery
drwxr-xr-x   - mapred hdfs            0 2018-11-29 17:26 /mapred
drwxrwxrwx   - mapred hadoop          0 2018-11-29 17:26 /mr-history
drwxr-xr-x   - hdfs   hdfs            0 2018-11-29 18:54 /ranger
drwxrwxrwx   - spark  hadoop          0 2018-11-29 19:21 /spark2-history
drwxrwxrwx   - hdfs   hdfs            0 2018-11-29 19:01 /tmp
drwxr-xr-x   - hdfs   hdfs            0 2018-11-29 19:21 /user
drwxr-xr-x   - hdfs   hdfs            0 2018-11-29 17:51 /warehouse
```

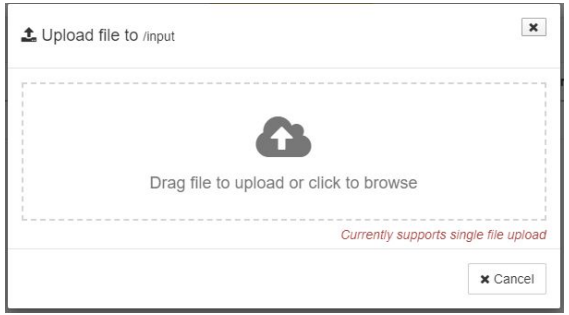
Créer un dossier dans hdfs

2.

```
[root@sandbox-hdp ~]# sudo -u hdfs hadoop fs -chmod 777 /input
```

Modifier l'autorisation de dossier pour le rendre accessible à tous les utilisateurs

3.



Name	Size	Last Modified	Owner	Group	Permission	Erasure Coding	Encrypted
predict.csv	175.8 kB	2020-01-30 11:33	admin	hdfs	-rw-r--r--		No
train.csv	20.1 MB	2020-01-30 11:33	admin	hdfs	-rw-r--r--		No

Etape 2 : Placer des données sur la VM dans le cloud AWS

HDFS -> VM locale -> Amazon EC2

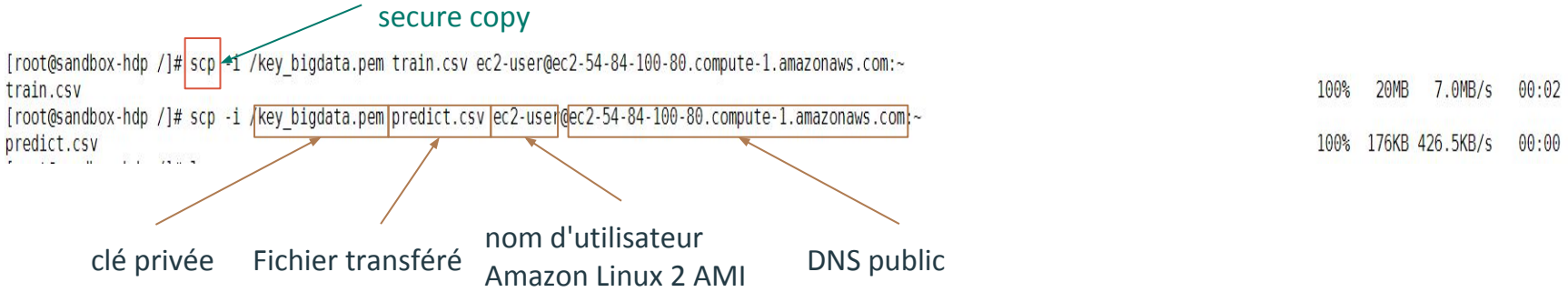
HDFS -> VM locale

```
[root@sandbox-hdp ~]# hadoop fs -get /input/predict.csv /.
[root@sandbox-hdp ~]# hadoop fs -get /input/train.csv /.
[root@sandbox-hdp ~]# hadoop fs -get /input/key_bigdata /.
get: `/input/key_bigdata': No such file or directory
[root@sandbox-hdp ~]# hadoop fs -get /input/key_bigdata.pem /.
[root@sandbox-hdp ~]# ls
anaconda-ks.cfg
[root@sandbox-hdp ~]# cd ..
[root@sandbox-hdp /]# ls
apps  boot         dev  hadoop  kafka-logs      lib    media  mysql-connector-java-5.1.45  packer-files  proc  run      sandbox-flavour  srv  tmp         usr
bin   cgroups_test  etc  home    key_bigdata.pem lib64  mnt    opt                           predict.csv   root  sandbox  sbin             sys  train.csv   var
```

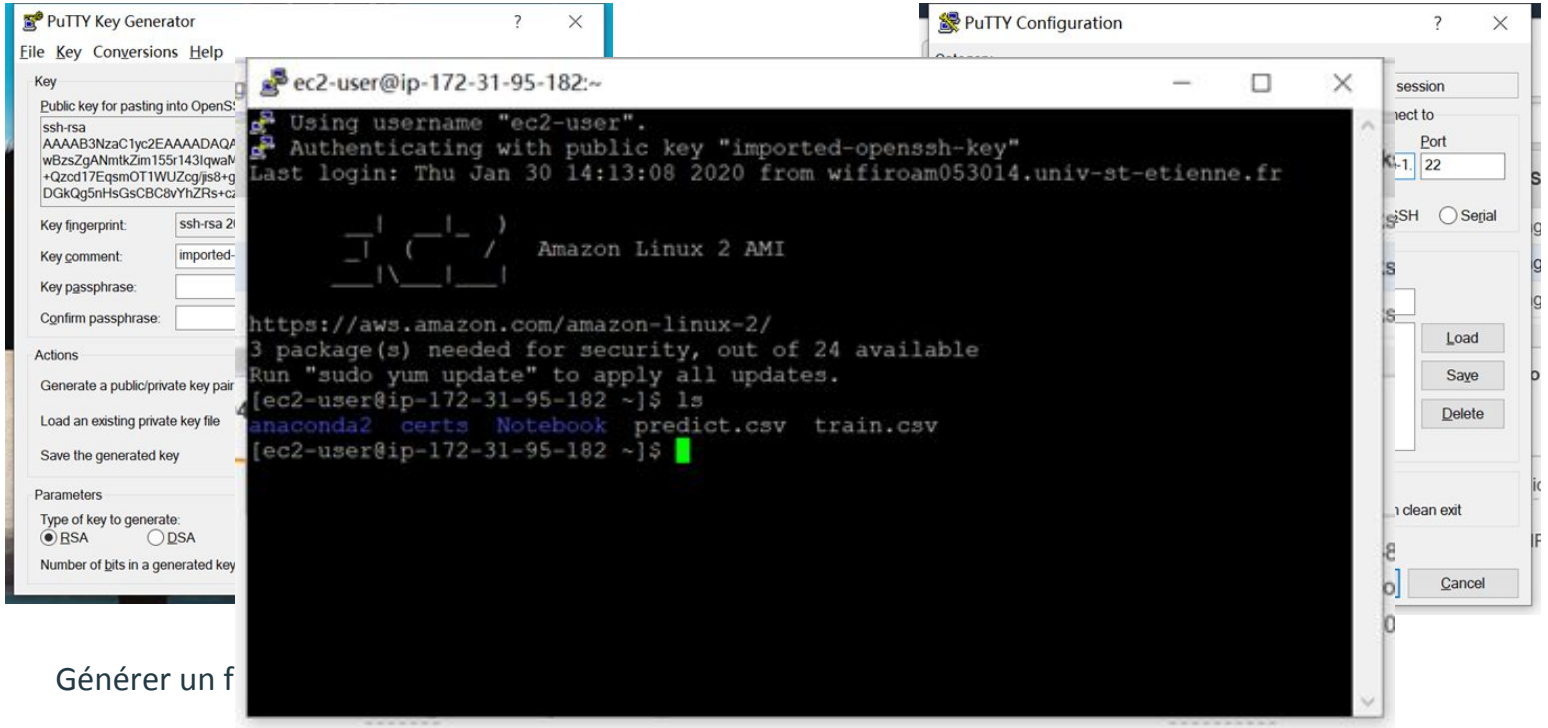
VM locale -> Amazon EC2

sudo chmod 400 key_bigdata.pem

Modifier l'autorisation du fichier pem. Si l'autorisation est trop ouverte, la clé n'a aucun sens.



Accéder à Amazon EC2 instance

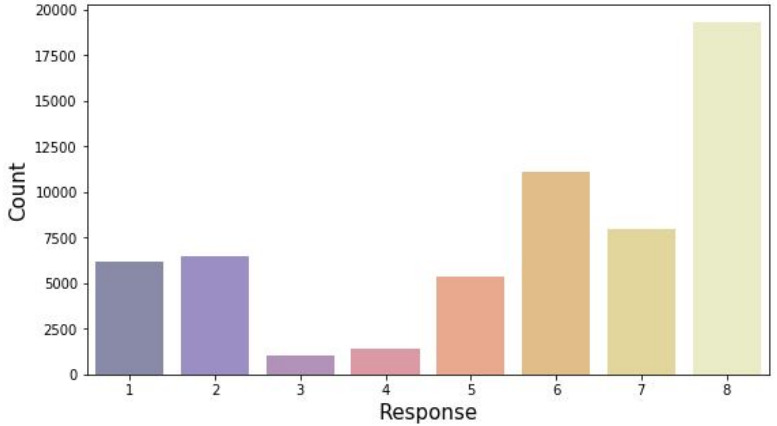


Générer un f

Etape 3: Analyse et traitement des données

- Analyse des données
- Que pouvons-nous savoir? Que pouvons-nous faire?
  - Labels
  - Type de données
  - Distribution de données
  - Informations numériques

- **Labels:**
  - Des données discrètes avec des valeurs 1-8
    - Problème de multi-classification :
      - KNN / SVC / NN / Des modèles basé sur l'arbre de décision*

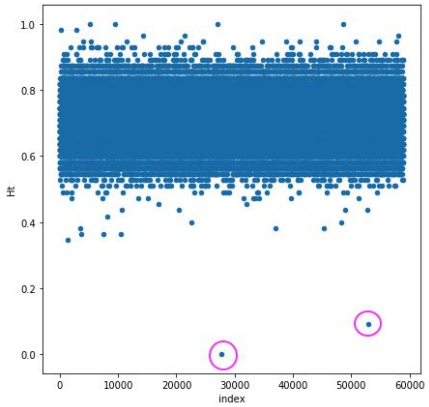


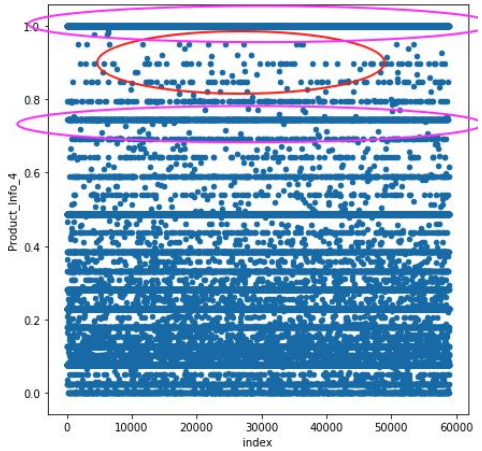
- **Type de données:**
  - Données non numériques :
    - Des données temporelles? - Non
    - Les significations des attributes? - Oui, une partie
      - **InsuredInfo_7 - Sexe (Male, Female)**
      - **InsuredInfo_8 - Race (e.g. Arabic, Asian, etc)**
      - **InsuredInfo_9 - La religion (e.g. Muslim, Jewish)**

```
X_train[X_train.columns[(X_train.dtypes == 'object')==True]].head(5)
```

	Product_Info_2	InsuredInfo_7	InsuredInfo_8	InsuredInfo_9
0	A8	Female	NaN	Muslim
1	D2	Male	NaN	NaN
2	D2	Male	Caucasian	Jewish
3	D4	Female	NaN	NaN
4	E1	Female	NaN	NaN

- **Type de données:**
  - Données continues :
    - Normalisées? - les atrributes de ‘**Medical_History’** ne sont pas Normalisées
    - Des valeurs aberrantes?
    - Les intervalles?



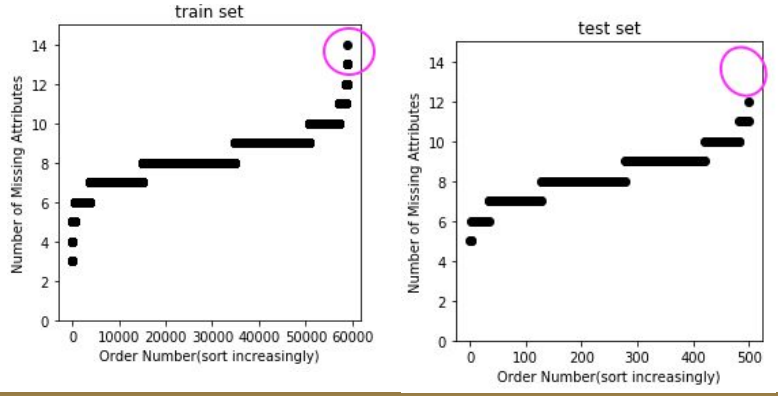


- **Distribution des données:**
  - Les valeurs manquantes
    - Statistiques basées sur les attributes

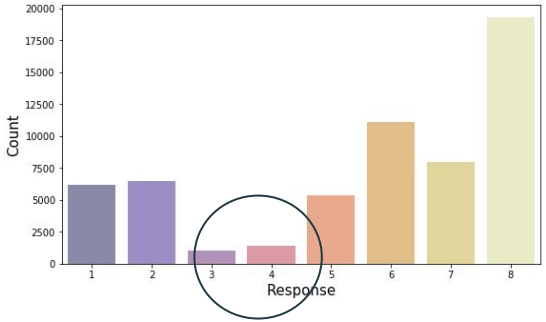
Name - Missing/Nb_row:

	Name	Missing
48	Medical_History_10	0.990642
70	Medical_History_32	0.981318
62	Medical_History_24	0.935939
24	InsuredInfo_8	0.879197
53	Medical_History_15	0.750972
25	InsuredInfo_9	0.749410
38	Family_Hist_5	0.703996
36	Family_Hist_3	0.576621
35	Family_Hist_2	0.482702
30	Insurance_History_5	0.427642
37	Family_Hist_4	0.323177
16	Employment_Info_6	0.182623
39	Medical_History_1	0.149607
14	Employment_Info_4	0.114078
11	Employment_Info_1	0.000323

    - Statistiques basées sur des échantillons entre le test et le training set



- **Distribution des données:**
  - Statistiques sur des labels



- **Informations numériques:**
  - Variance
  - Corrélation

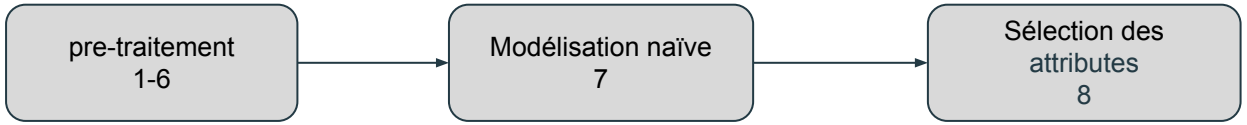
Insurance_History_5	0.000054
Employment_Info_4	0.001083
Medical_History_35	0.004106
Medical_History_38	0.004817
Ht	0.005513
Medical_Keyword_13	0.005892

Variance

	Response
Medical_Keyword_45	0.004952
Product_Info_7	0.004825
Medical_Keyword_8	0.004202
Family_Hist_3	0.003265
Product_Info_5	0.002011
Insurance_History_1	0.001034
History_26	0.000601
Medical_History_36	0.000580
Insurance_History_4	0.000567
Medical_History_25	0.000428

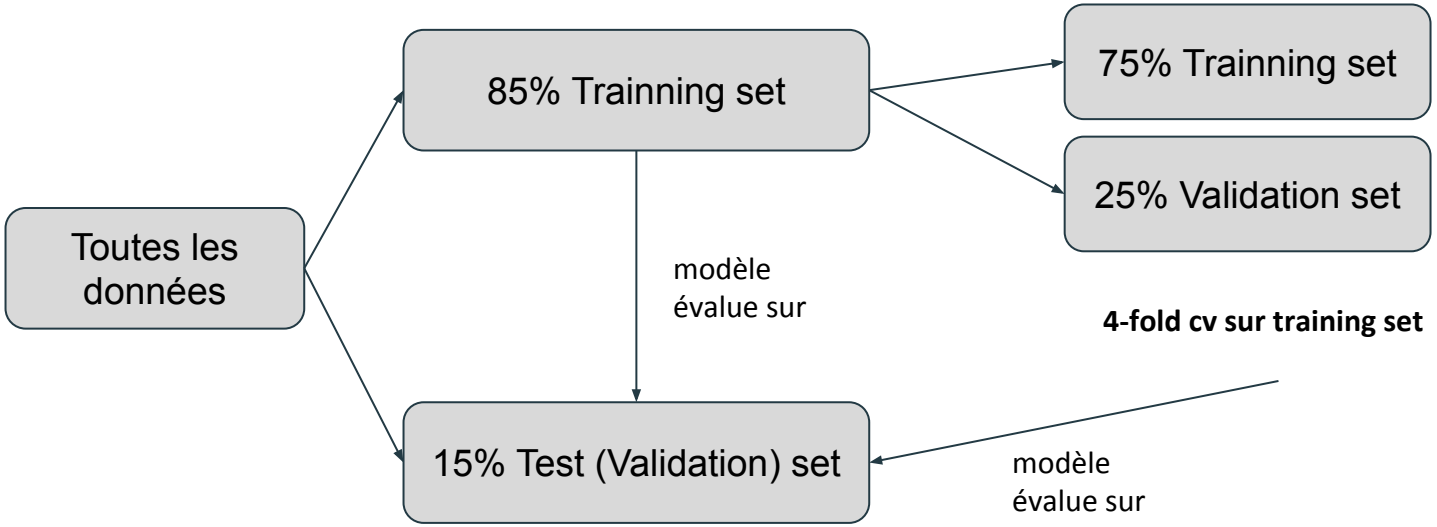
Corrélation

Traitement global des données



1. Supprimer les échantillons avec 14 et 15 valeurs manquantes;
2. Supprimer directement les attributes qui ont plus de 70% de valeurs manquantes;
3. Encoder les données de type objet;
   - Après avoir suivi les étapes ci-dessus, on trouve que on a seulement besoin de coder ‘sexe’
     - One-hot encoding
4. Remplissez les valeurs manquantes en **moyenne/médiane** pour d'autres attributes; ***(* comparison)***
5. Supprimer les top 5 attributes avec la variance la plus faible;
6. Supprimer les top 5 attributes avec la plus faible corrélation avec le label;
7. Former **certains modèles** pour voir le ‘baseline’;***(* comparison)***
8. * Sélectionner **un nombre de** attributes importantes pour tester les performances du modèle;***(* comparison)***
   - Modèle arborescent - random forest + xgboost -> feature_importance_

- **Pour l'ensemble de l'apprentissage, l'ensemble de validation et l'ensemble de test:**



Modélisation naïve
7

- **Comparisons:**
  - moyenne/médiane - modèles différents

In using median, the accuracy result on validation set is:

```
rf  model:  0.5131400090620752
knn:  0.3194381513366561
xgboost0  0.5445174444947893
xgboost1  0.5425917535115542
xgboost2  0.5434979610330766
xgboost3  0.540666062528319
4  xgboost  hard  voting:  0.5423652016311735
4  xgboost+rf  hard  voting:  0.5434979610330766
```

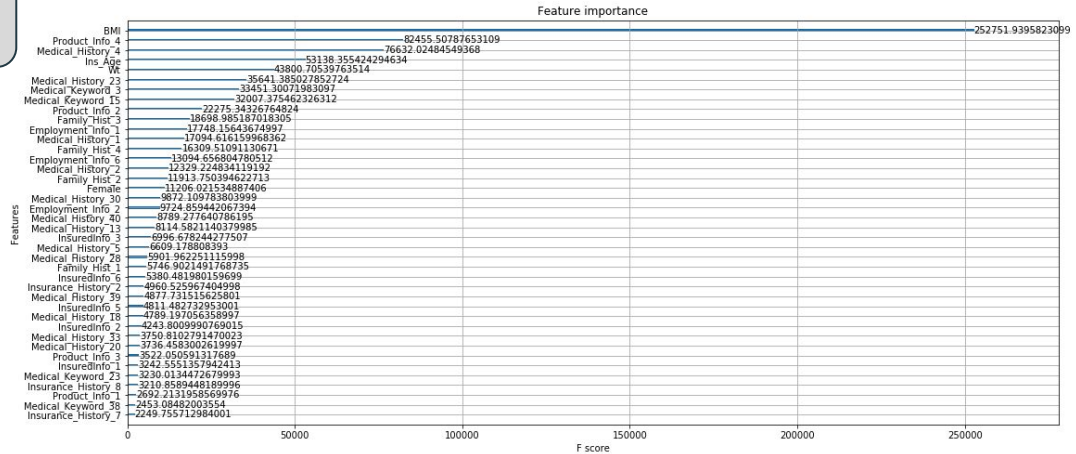
In using mean, the accuracy result on validation set is:

```
rf  model:  0.5131400090620752
knn:  0.3192115994562755
xgboost0  0.544404168554599
xgboost1  0.5425917535115542
xgboost2  0.542931581332125
xgboost3  0.5424784775713638
4  xgboost  hard  voting:  0.5437245129134571
4  xgboost+rf  hard  voting:  0.542931581332125
```

model	mean	median
rf	0.513	0.513
knn	0.319	0.319
xgboost1	0.544	0.544
xgboost2	0.543	0.543
xgboost3	0.543	0.543
xgboost4	0.542	0.541
4xgb_voting	0.544	0.542
4xgb+rf_voting	0.543	0.543

- Hard voting:
  - Pour les résultats de classification de chaque échantillon, le résultat majoritaire est sélectionné comme résultat final.
    [1,1,3,1,5] -> [1]

Sélection des attributes
8

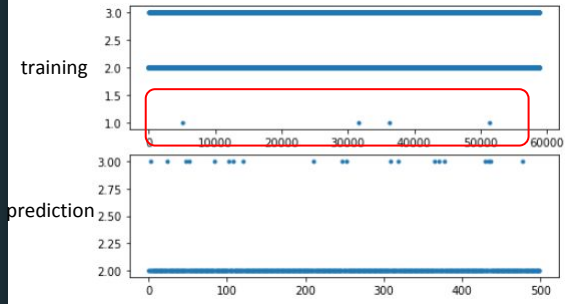


- Sélectionnez les top 50 attributes parmi les cinq modèles (1 RF, 4 Xgb via 4 fold CV), puis prenez leur intersection:

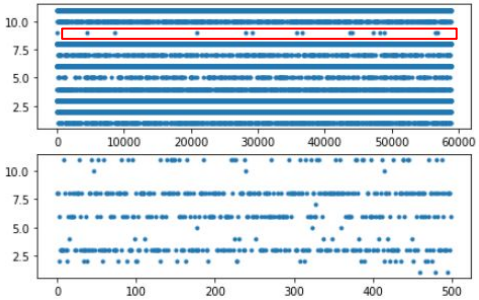
- TOP 35 attributes: `['BMI', 'Product_Info_4', 'Medical_History_4', 'Ins_Age', 'Wt', 'Medical_History_23', 'Medical_Keyword_3', 'Medical_Keyword_15', 'Product_Info_2', 'Employment_Info_1', 'Family_Hist_3', 'Medical_History_1', 'Family_Hist_4', 'Employment_Info_6', 'Medical_History_2', 'Family_Hist_2', 'Employment_Info_2', 'Medical_History_30', 'Female', 'InsuredInfo_3', 'Medical_History_40', 'Medical_History_13', 'Medical_History_39', 'Medical_History_28', 'Insurance_History_2', 'Family_Hist_1', 'InsuredInfo_5', 'InsuredInfo_6', 'InsuredInfo_2', 'Medical_History_18', 'InsuredInfo_1', 'Product_Info_3', 'Insurance_History_8', 'Medical_History_33', 'Medical_Keyword_23'`

- **Traitement détaillé des données**

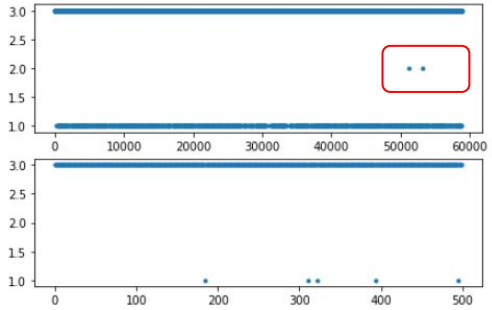
58881 -> 58776



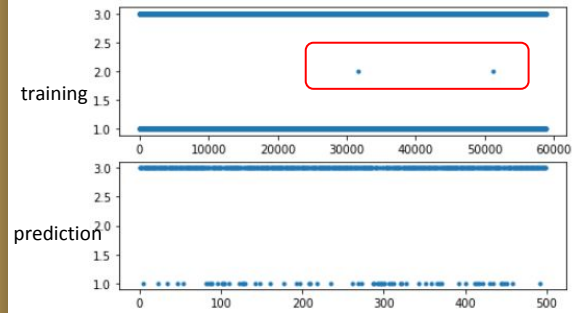
Medical_History_30



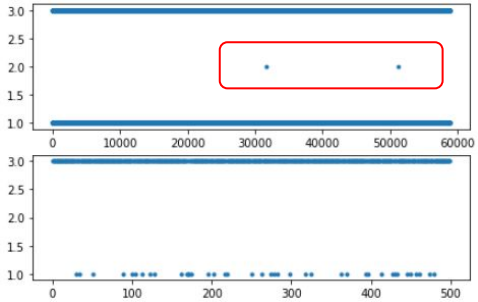
InsuredInfo_3



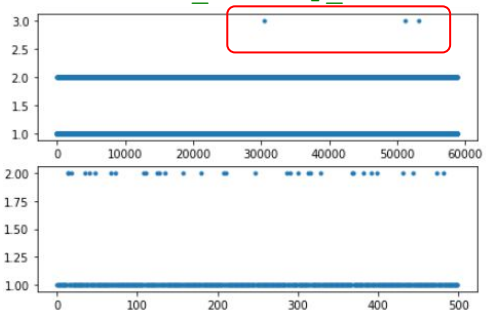
Medical_History_40



Medical_History_13



Medical_History_39



Medical_History_28

- **Comparisons - models:**
  - SVM

```
validation set presicion: 0.32822955653850516
training set precision: 0.328189115074361
```

10 attributes

```
validation set presicion: 0.459906997845072
training set precision: 0.45301146940491205
```

5 attributes

```
validation set presicion: 0.445956674605875
training set precision: 0.43687824015692867
```

3 attributes

  - Random forest

```
(49959, 35)
(8817, 35)
validation set presicion: 0.5353294771464217
```

35 attributes

```
(49959, 15)
(8817, 15)
validation set presicion: 0.5136667800839287
```

15 attributes

```
(49959, 10)
(8817, 10)
validation set presicion: 0.47794034252013157
```

10 attributes

- **Comparisons - models:**
  - La distribution des étiquettes est différente :
    - Over-sampling - Under-sampling - Inchangé sur trainning set
      (Évaluation sur un ensemble de test inchangé)

over-sampling:
SMOTE

```
xgboost0 0.544743109901327
xgboost1 0.5456504479981853
xgboost2 0.5461041170466145
xgboost3 0.5436089372802541
test set precision 4 models hard voting: 0.543949189066576
test set precision 5 models hard voting: 0.5453101962118635
```

under-sampling:
Random

```
xgboost0 0.4411931495973687
xgboost1 0.4417602359079052
xgboost2 0.43574912101621865
xgboost3 0.44414199841215835
test set precision 4 models hard voting: 0.4384711353067937
test set precision 5 models hard voting: 0.44085289781104686
```

Inchangé

```
xgboost0 0.5542701599183396
xgboost1 0.553249404559374
xgboost2 0.5485992968129749
xgboost3 0.5525689009867302
test set precision 4 models hard voting: 0.5557445843257344
test set precision 5 models hard voting: 0.554043325394125
```

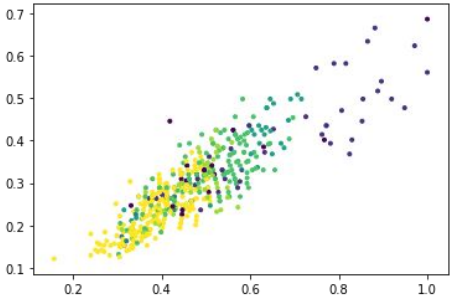
- **Resultat final:**
  - La précision maximale sur l'ensemble de test (15% de données): 56%

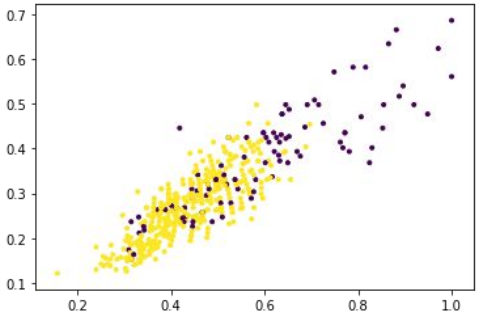
	precision	recall	f1-score	support
1	0.47	0.27	0.35	923
2	0.48	0.25	0.33	972
3	0.36	0.05	0.09	151
4	0.38	0.13	0.19	213
5	0.62	0.56	0.59	807
6	0.43	0.56	0.49	1667
7	0.47	0.37	0.41	1190
8	0.66	0.88	0.76	2894
accuracy			0.56	8817
macro avg	0.48	0.38	0.40	8817
weighted avg	0.54	0.56	0.53	8817

- **Resultat final:**

BMI - Wt

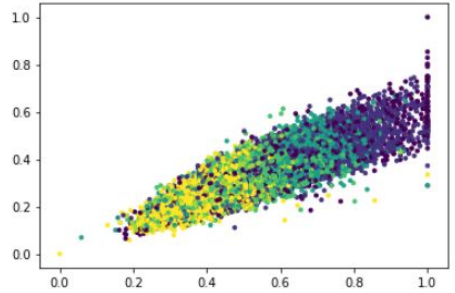
pred





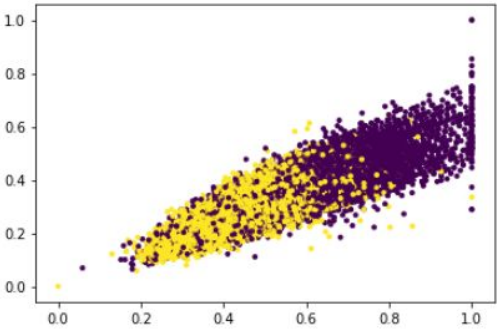
training

```
plt.scatter(data['BMI'][:],data['Wt'][:],c=y_data,ma
<matplotlib.collections.PathCollection at 0x7f9e8421
```



8 labels

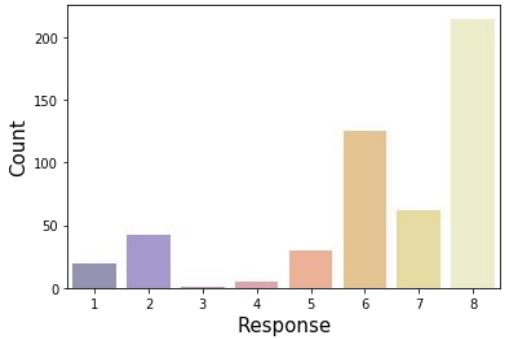
```
plt.scatter(data['BMI'][:],data['Wt'][:],c=y_data2,mark
<matplotlib.collections.PathCollection at 0x7f9e842aa12
```



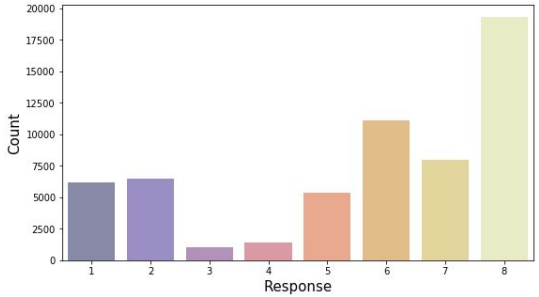
2 labels
1-5: 0
6-8: 1

- **Resultat final:**

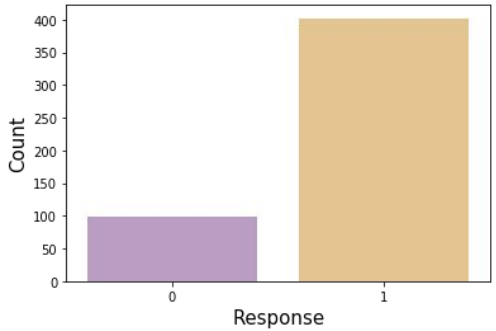
pred

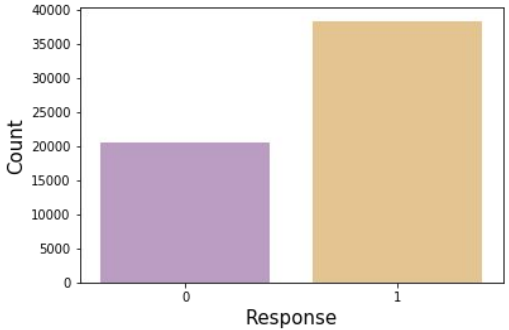


trainning



8 labels





2 labels
1-5: 0
6-8: 1

Exécution d'algorithmes sur le cloud AWS

- Installer du Python 3.7 sur l'instance EC2.
  - Installer du Python 3.7 sur l'instance EC2.

    sudo yum install python3 -y

  - Créer un environnement virtuel dans le répertoire de base ec2-user.

    python3 -m venv my_app/env

  - Activer l'environnement virtuel

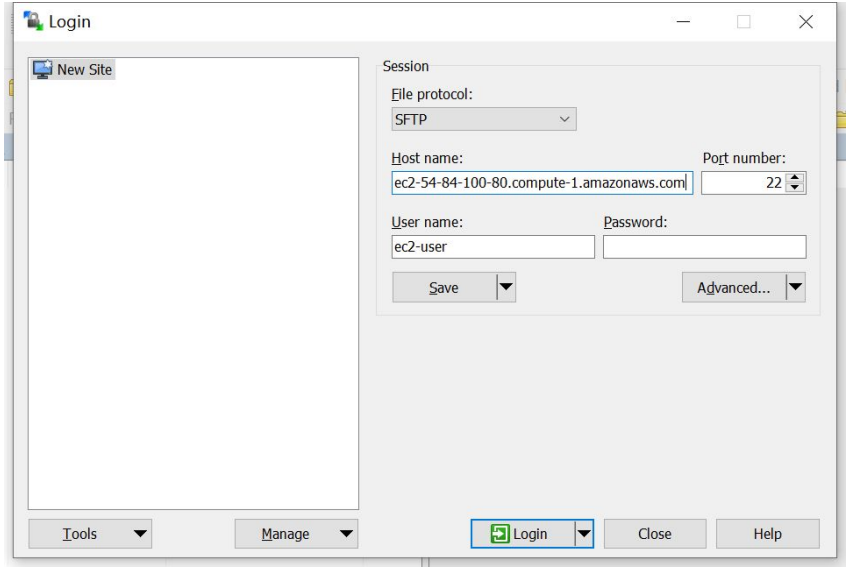
```
(python3) [ec2-user@ip-172-31-95-182 ~]$ python --version
Python 3.7.4
```

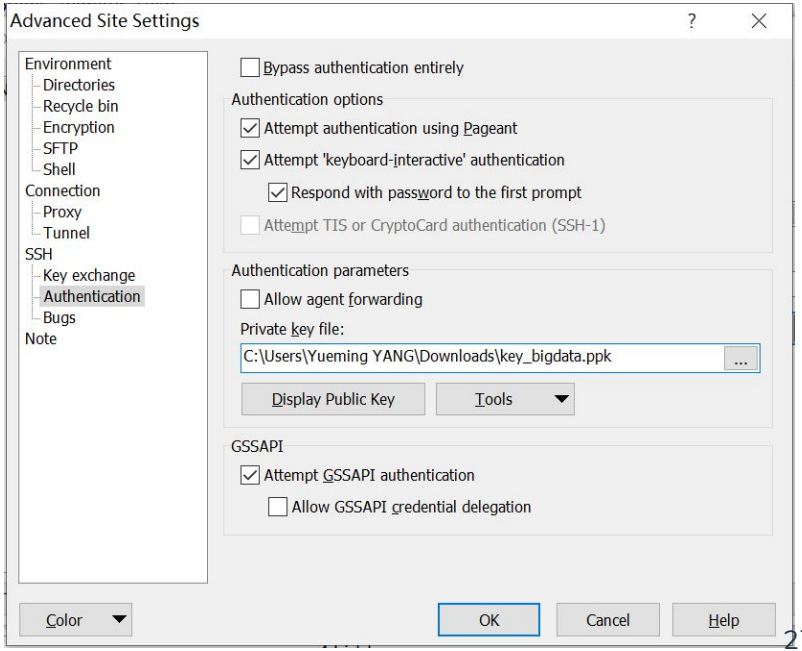
    source ~/my_app/env/bin/activate;

    echo "source ${HOME}/my_app/env/bin/activate" >> ${HOME}/.bashrc;

    source ~/.bashrc

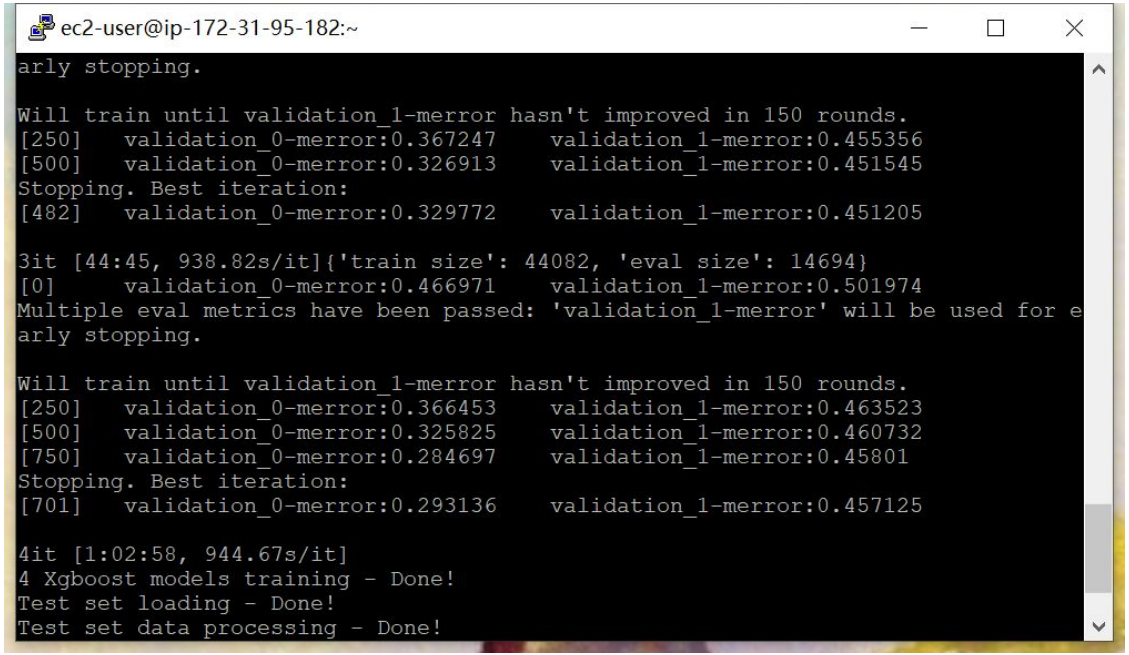
- Le transfért du fichier python de la machine locale vers l'instance EC2 : WinSCP
    - User Name : ec2-user
    - Host Name: le DNS public
    - Private Key File





- Exécution du fichier python par la commande:

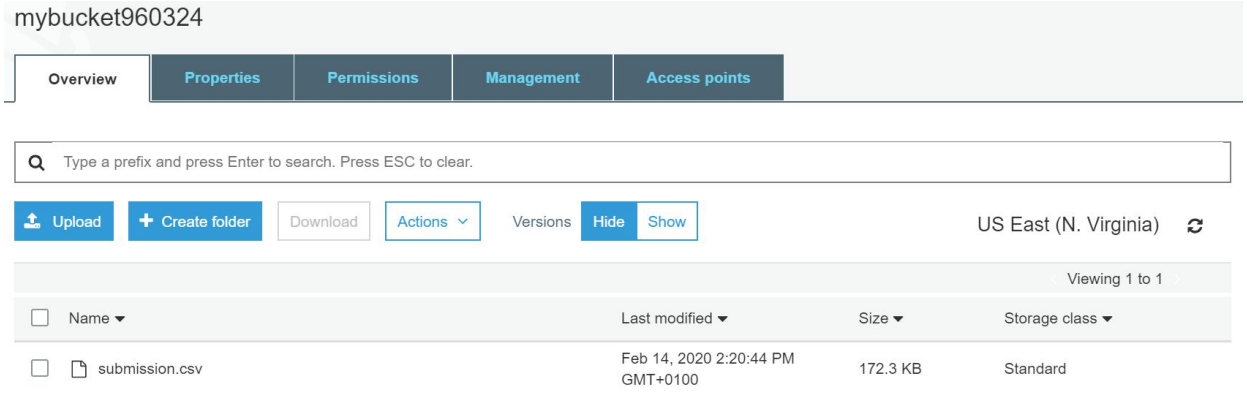
python big_data_project.py



Etape 4:Créer et sauvegarder le fichier resultats

```
X_test = pd.read_csv('predict.csv')
X_test['Response'] = pred
X_test.to_csv("submission.csv",index=False)
```

→ Création d'une bucket → Telechargement du fichier.CSV



Etape 5: Recupération des données dans MongoDB

- Installer : 
- Transférer le fichier python de l'instance EC2 vers VM MongoDB avec la commande suivante:

scp -i /home/yym/Downloads/key_bigdata.pem
ec2-user@ec2-54-84-100-80.compute-1.amazonaws.com:/home/ec2-user/predict.csv /home/yym/Downloads

Créer une base de données et une collection dans MongoDB

- Créer une base de données :

use bigdata →

```
> show dbs
admin    0.000GB
bigdata  0.000GB
config   0.000GB
local    0.000GB
```

- Créer une collection :

db.createCollection(submission) →

```
> show collections
myCollection
predict
submission
```

Importer le fichier csv dans MongoDB

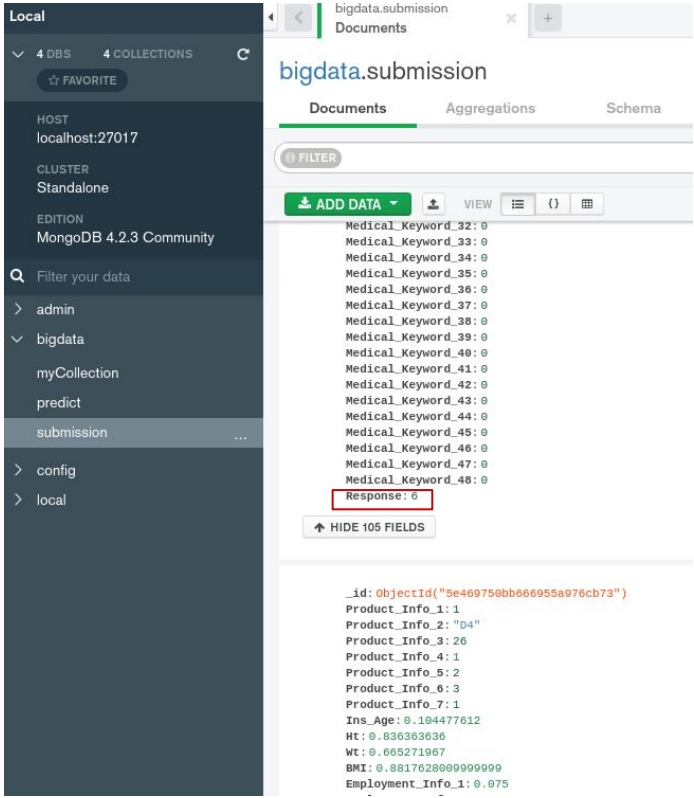
Par la commande:

mongoimport --db bigdata --collection submission --type csv --headerline --file /home/yym/Downloads/submission.csv

```
yym@yym-VirtualBox:~/Downloads$ mongoimport --db bigdata --collection predict --
type csv --headerline --file /home/yym/Downloads/predict.csv
2020-02-13T15:57:59.251+0100    connected to: mongodb://localhost/
2020-02-13T15:57:59.858+0100    500 document(s) imported successfully. 0 documen
t(s) failed to import.
```

Pour ouvrir et voir le fichier dans MongoDB et dans l'outil GUI 'MangoDB Compass'

```
> db.submission.find().pretty()
{
        "_id" : ObjectId("5e469750bb666955a976cb72"),
        "Product_Info_1" : 1,
        "Product_Info_2" : "D4",
        "Product_Info_3" : 26,
        "Product_Info_4" : 0.230769231,
        "Product_Info_5" : 2,
        "Product_Info_6" : 1,
        "Product_Info_7" : 1,
        "Ins_Age" : 0.358208955,
        "Ht" : 0.709090909,
        "Wt" : 0.349372385,
        "BMI" : 0.569404716,
        "Employment_Info_1" : 0.075,
        "Employment_Info_2" : 9,
        "Employment_Info_3" : 1,
        "Employment_Info_4" : 0,
        "Employment_Info_5" : 2,
        "Employment_Info_6" : 0.4,
        "InsuredInfo_1" : 2,
        "InsuredInfo_2" : 2,
        "InsuredInfo_3" : 8,
        "InsuredInfo_4" : 3,
        "InsuredInfo_5" : 1,
        "InsuredInfo_6" : 1,
        "InsuredInfo_7" : "Male",
        "InsuredInfo_8" : "",
        "InsuredInfo_9" : "",
        "Insurance_History_1" : 2,
        "Insurance_History_2" : 1,
        "Insurance_History_3" : 1,
        "Insurance_History_4" : 3,
        "Insurance_History_5" : "",
        "Insurance_History_7" : 3,
        "Insurance_History_8" : 2,
        "Insurance_History_9" : 3,
        "Family_Hist_1" : 2,
        "Family_Hist_2" : 0.47826087,
        "Family_Hist_3" : "",
        "Family_Hist_4" : 0.549295775,
        "Family_Hist_5" : "",
        "Medical_History_1" : 7,
        "Medical_History_2" : 112,
        "Medical_History_3" : 2,
        "Medical_History_4" : 2,
        "Medical_History_5" : 1,
```



Query dans MongoDB

```
> db.submission.find({Response:{$gte:6}})
{ "_id" : ObjectId("5e42b96419b819e1b2479017"), "Response" : 6 }
{ "_id" : ObjectId("5e42b96419b819e1b247901a"), "Response" : 6 }
{ "_id" : ObjectId("5e42b96419b819e1b247901b"), "Response" : 8 }
{ "_id" : ObjectId("5e42b96419b819e1b247901c"), "Response" : 7 }
{ "_id" : ObjectId("5e42b96419b819e1b247901d"), "Response" : 8 }
{ "_id" : ObjectId("5e42b96419b819e1b247901f"), "Response" : 6 }
{ "_id" : ObjectId("5e42b96419b819e1b2479020"), "Response" : 8 }
{ "_id" : ObjectId("5e42b96419b819e1b2479021"), "Response" : 6 }
{ "_id" : ObjectId("5e42b96419b819e1b2479022"), "Response" : 8 }
{ "_id" : ObjectId("5e42b96419b819e1b2479023"), "Response" : 8 }
{ "_id" : ObjectId("5e42b96419b819e1b2479024"), "Response" : 6 }
{ "_id" : ObjectId("5e42b96419b819e1b2479025"), "Response" : 8 }
{ "_id" : ObjectId("5e42b96419b819e1b2479026"), "Response" : 8 }
{ "_id" : ObjectId("5e42b96419b819e1b2479027"), "Response" : 8 }
{ "_id" : ObjectId("5e42b96419b819e1b2479028"), "Response" : 8 }
{ "_id" : ObjectId("5e42b96419b819e1b2479029"), "Response" : 6 }
{ "_id" : ObjectId("5e42b96419b819e1b247902a"), "Response" : 6 }
{ "_id" : ObjectId("5e42b96419b819e1b247902b"), "Response" : 8 }
{ "_id" : ObjectId("5e42b96419b819e1b247902c"), "Response" : 8 }
{ "_id" : ObjectId("5e42b96419b819e1b247902e"), "Response" : 8 }
Type "it" for more
```

Conclusion

+ Différentes disciplines intégrées

- Répartition inégale de la gravité spécifique de chaque partie

Merci pour votre attention