ESILV LÉONARD DE VINCI | ÉCOLE D'INGÉNIEURS PARIS-LA DÉFENSE

SEISMIC BUMPS

Projet de data analysis
Alexandre BEAUJOUR
Kheir Eddine KOUIDER
Fevrier 2019

PRESENTATION

❖ *Objectif:*

- Créer un script python qui automatise toute la procédure de modélisation de la base de données.

Python 

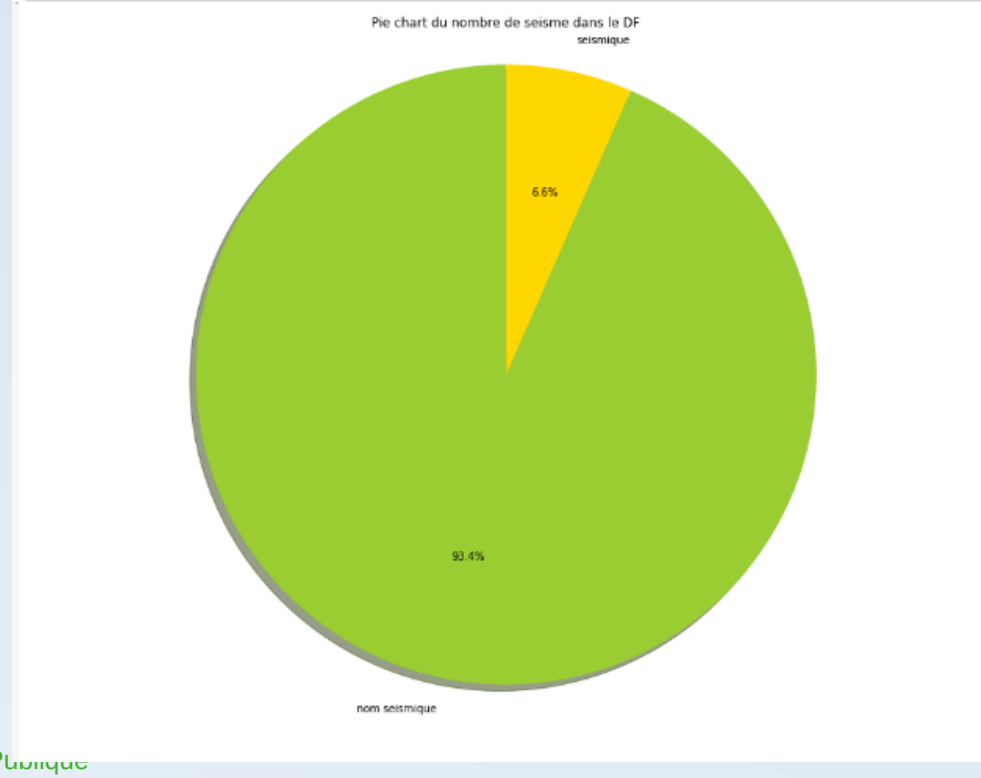
❖ *Etape*

- Charger les données
- Visualiser les données
- Préparer les données
- Construire un model Machine Learning

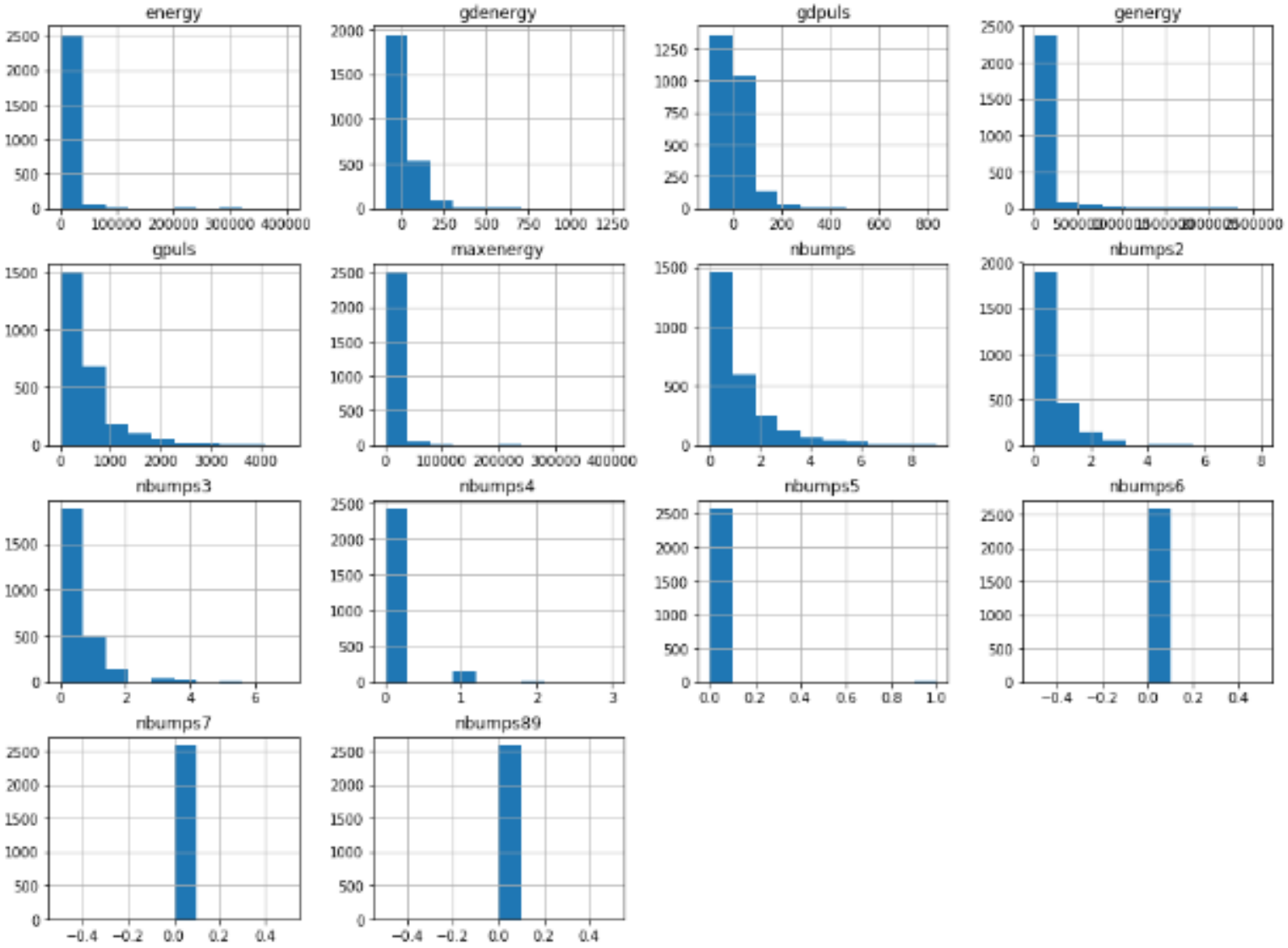
VISUALISATION DES DONNEES

- Pourcentage des données sismiques :
  - 93.4% des données sont non-sismiques
  - 6.6% des données sont sismiques

```
import matplotlib.pyplot as plt

labels = 'nom seismique', 'seismique'
sizes = [df['class'].value_counts()[0], df['class'].value_counts()[1]]
colors = ['yellowgreen', 'gold', 'lightskyblue', 'lightcoral']

plt.pie(sizes, labels=labels, colors=colors,
        autopct='%1.1f%%', shadow=True, startangle=90)

plt.axis('equal')
plt.title('Pie chart du nombre de seisme dans le DF')
plt.savefig('PieChart01.png')
plt.show()
```



HISTOGRAMME DES VARIABLES



PREPARATION DES DONNEES

```
seismic            0
seismoacoustic     0
shift              0
genergy            0
gpuls              0
gdenergy           0
gdpuls             0
ghazard            0
nbumps             0
nbumps2            0
nbumps3            0
nbumps4            0
nbumps5            0
nbumps6            0
nbumps7            0
nbumps89           0
energy             0
maxenergy          0
class              0
dtype: int64
```

```
In [151]: #type des variables
          print(df.dtypes)
```

```
seismic            object
seismoacoustic     object
shift              object
genergy           float64
gpuls             float64
gdenergy          float64
gdpuls            float64
ghazard            object
nbumps            float64
nbumps2           float64
nbumps3           float64
nbumps4           float64
nbumps5           float64
nbumps6           float64
nbumps7           float64
nbumps89          float64
energy            float64
maxenergy         float64
class              object
dtype: object
```

```
#dimension
print(df.shape)
```

```
(2584, 19)
```

- 19 variables
- 2584 lignes
- 0 valeur manquante

- 5 variables catégorielles
- 14 variables numériques

PREPARATION DES DONNES

Python 

- Conversion des variables catégorielles en numérique
- Hot encoding des variables catégorielles
- Séparation des données de test et des données d'entrainement

```
X_entire.shape: (2584, 24) Y_entire.shape: (2584,)
X_train.shape: (1808, 24) Y_train.shape: (1808,)
X_validation.shape: (776, 24) Y_validation.shape: (776,)
Total time for data handling and visualization: 3:32:46.399145
```

MODEL MACHINE LEARNING

- Choix d'un algorithme initialement prévu pour des classifications binaires : SVM

- Les machines à vecteurs de support sont un ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de discriminationnote 1 et de régression. Les SVM sont une généralisation des classifieurs linéaires.

  -> Problématique des seismes

- L'algorithme avec le meilleur AUC : 0.9355670103092784

```
Best: 0.933628 using {'C': 0.25}
```

VALIDATION DU DATASET

	PRECISION	RECALL	F1 SCORE	SUPPORT
0	0.94	1.00	0.97	726
1	0.00	0.00	0.00	50
avg / total	0.88	0.94	0.90	776