CRÉDIT AGRICOLE
CORPORATE & INVESTMENT BANK

Classification des articles financiers avec Reuters-21578

Plan

1. Objectifs de l’exercise
2. Analyse Exploratoire des Données
3. Classification

Objectifs de l’exercise

Reuters-21578 contient 21 578 articles de presse étiquetés selon différents sujets.

Focus sur 5 catégories :

1. Monnaie/Échange étranger (**MONEY-FX**)
2. Transport maritime (**SHIP**)
3. Taux d'intérêt (**INTEREST**)
4. Fusions/Acquisitions (**ACQ**)
5. Résultats financiers (**EARN**)

Objectifs de l'exercise

- Réaliser une analyse descriptive pour comprendre les caractéristiques des données.
- Construire et évaluer des modèles de classification pour une analyse prédictive

Résultats attendus

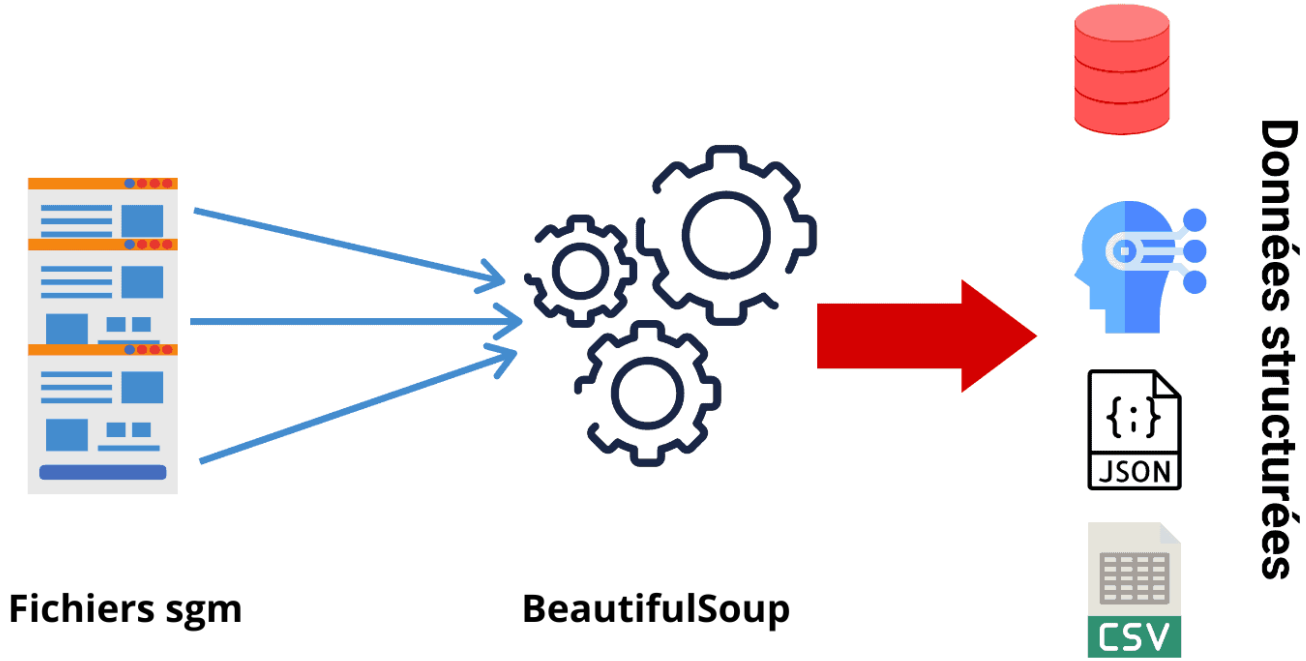
Classifier les articles en fonction de leur catégorie.

Analyse Exploratoire des Données

Extraction des données

Objectif :

L'objectif de l'extraction des données avec **BeautifulSoup** est de parser et extraire efficacement le contenu pertinent des fichiers .sgm, notamment les balises contenant les métadonnées et les articles, afin de structurer les informations sous un format exploitable pour l'analyse.



Automatisation ▶ ***Analyse Exploratoire des Données***

Anciennes

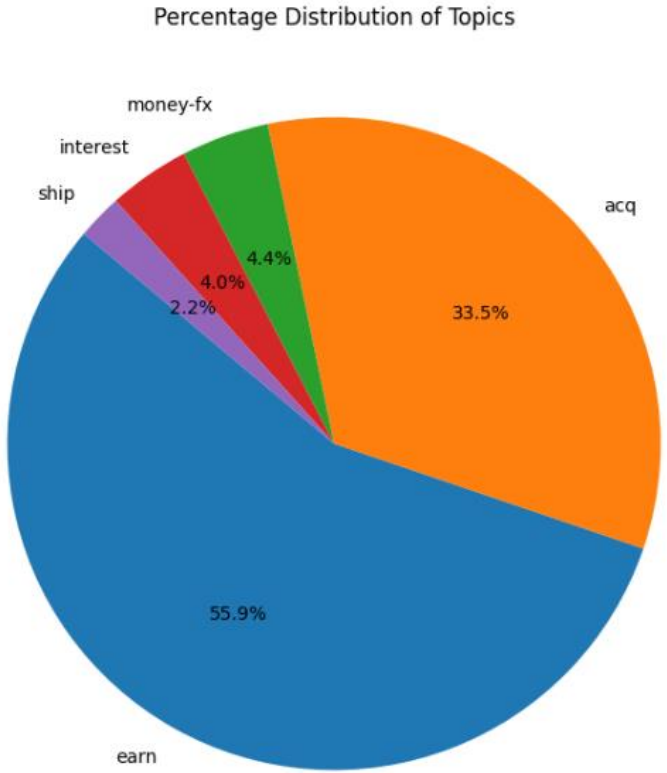
	Nom de la colonne
0	Topics
1	Places
2	People
3	Orgs
4	Exchanges
5	Companies
6	LEWISSPLIT
7	CGISPLIT
8	OLDID
9	NEWID
10	Title
11	Dateline
12	Body

Nouvelles

	Nom de la colonne
0	LEWISSPLIT
1	Text
2	Topics

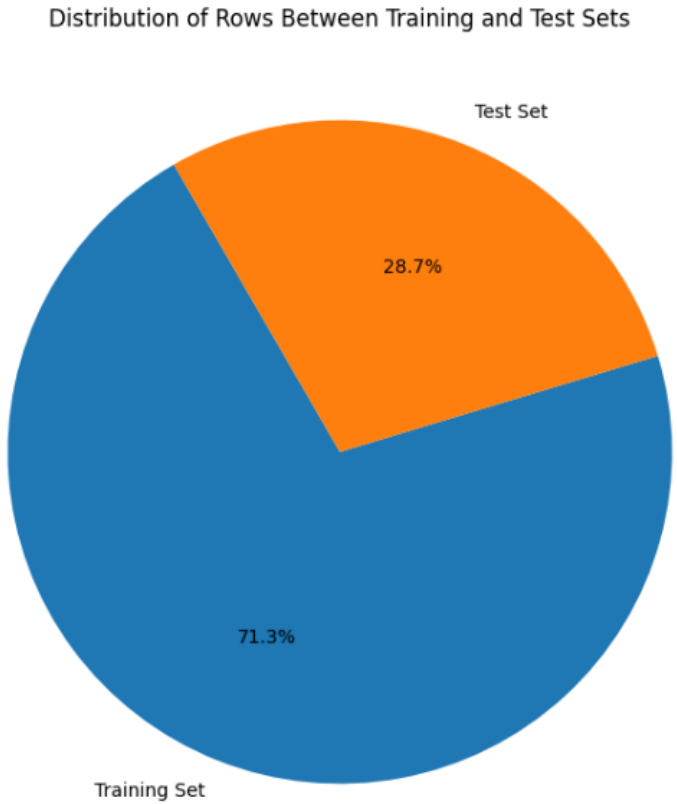
Automatisation ▶ ***Analyse Exploratoire des Données***

Analyse exploratoire des données



Automatisation ▶ ***Analyse Exploratoire des Données***

Analyse exploratoire des données



Automatisation ▶ ***Analyse Exploratoire des Données***

Analyse exploratoire des données

Mots les plus

```
[('vs', 14130),
 ('mln', 13540),
 ('said', 9497),
 ('cts', 7853),
 ('net', 6766),
 ('dlrs', 6714),
 ('\x03', 6486),
 ('reuter', 6437),
 ('loss', 4940),
 ('shr', 4036)]
```

Mots les moins

```
([('inapplicable', 1),
  ('render', 1),
  ('eleventh', 1),
  ('icahn-led', 1),
  ('passengers.', 1),
  ('liberalisation.', 1),
  ('dehesa', 1),
  ('guillermo', 1),
  ('interest."', 1),
  ('enacted', 1)],)
```

Automatisation ▶ ***Analyse Exploratoire des Données***

Analyse exploratoire des données

Bi-gram

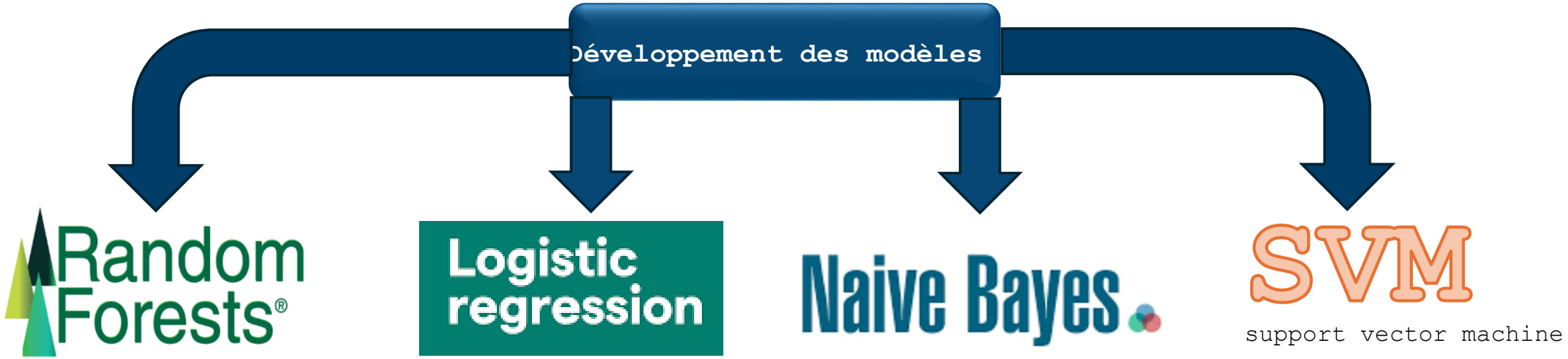
```
the most frequent bi-grams in the dataset:
[(('mln', 'vs'), 3947),
 (('said', 'it'), 3290),
 (('cts', 'vs'), 3285),
 (('of', 'the'), 3204),
 (('in', 'the'), 2634),
 (('cts', 'net'), 2220),
 (('mln', 'dlrs'), 2108),
 (('the', 'company'), 1978),
 (('said', 'the'), 1809),
 (('vs', 'loss'), 1775)]
```

Tri-gram

```
the most frequent tri-grams in the dataset:
[(('the', 'company', 'said'), 742),
 (('cts', 'vs', 'loss'), 659),
 (('mln', 'avg', 'shrs'), 621),
 (('qtr', 'net', 'shr'), 577),
 (('said', 'it', 'has'), 577),
 (('cts', 'net', 'loss'), 521),
 (('inc', 'said', 'it'), 501),
 (('mln', 'dlrs', 'in'), 496),
 (('pct', 'of', 'the'), 434),
 (('nine', 'mths', 'shr'), 414)]
```

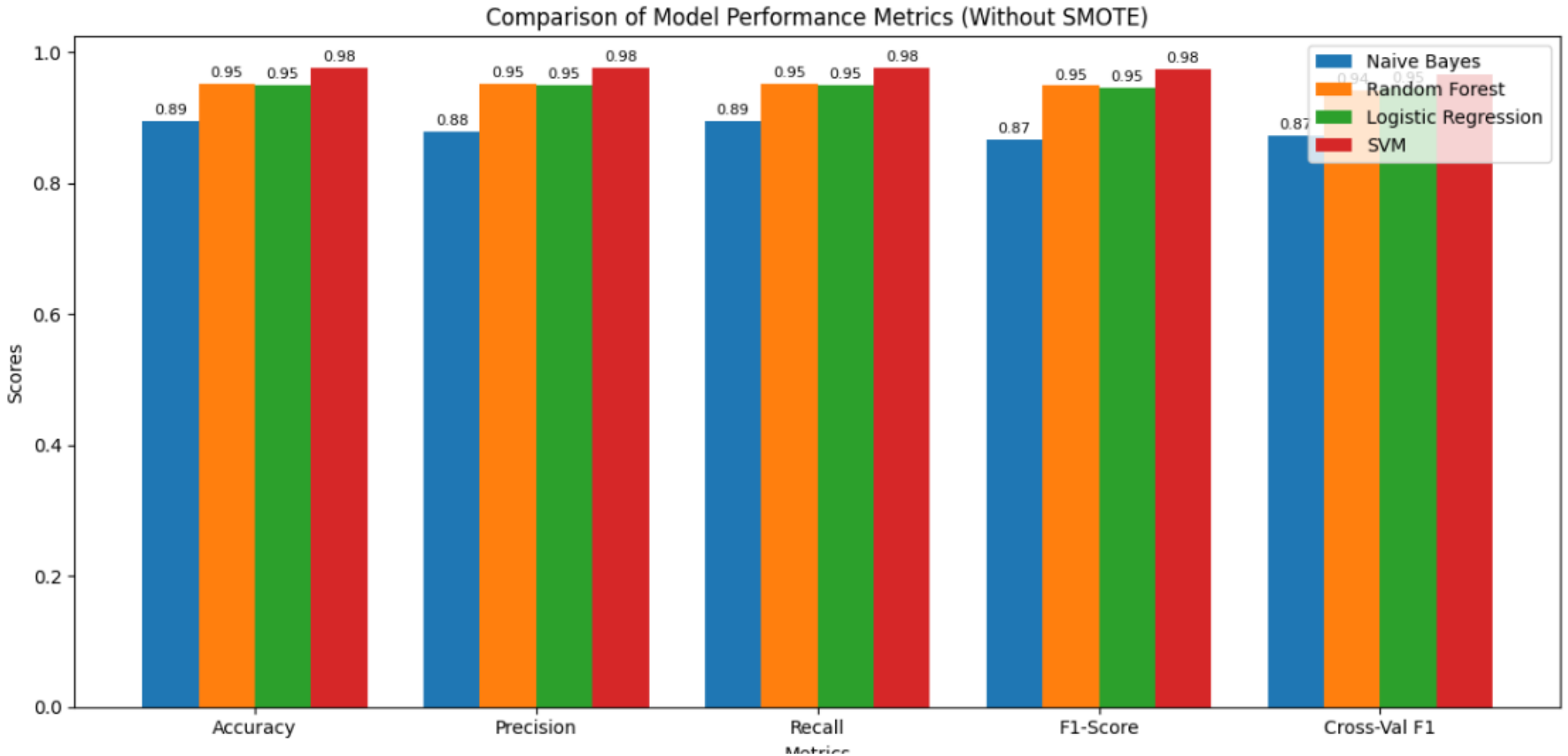
Classification

Ancienne approche *Nouvelle approche*



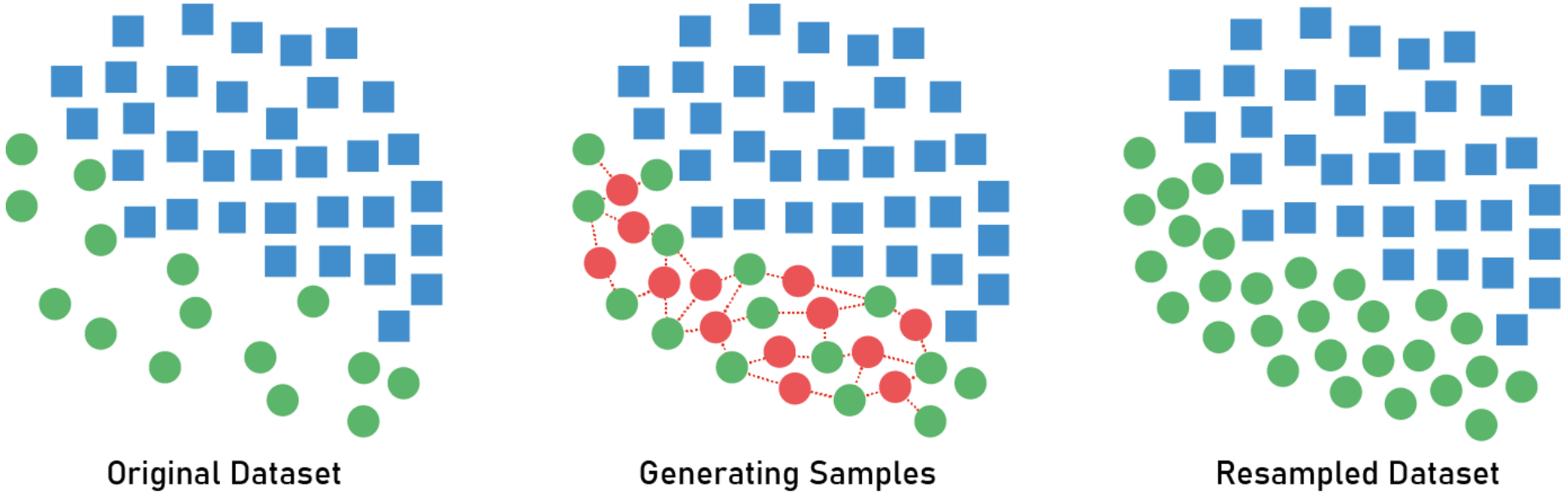
- Chaque algorithme a été entraîné sur l'ensemble d'entraînement.

Ancienne approche *Nouvelle approche*

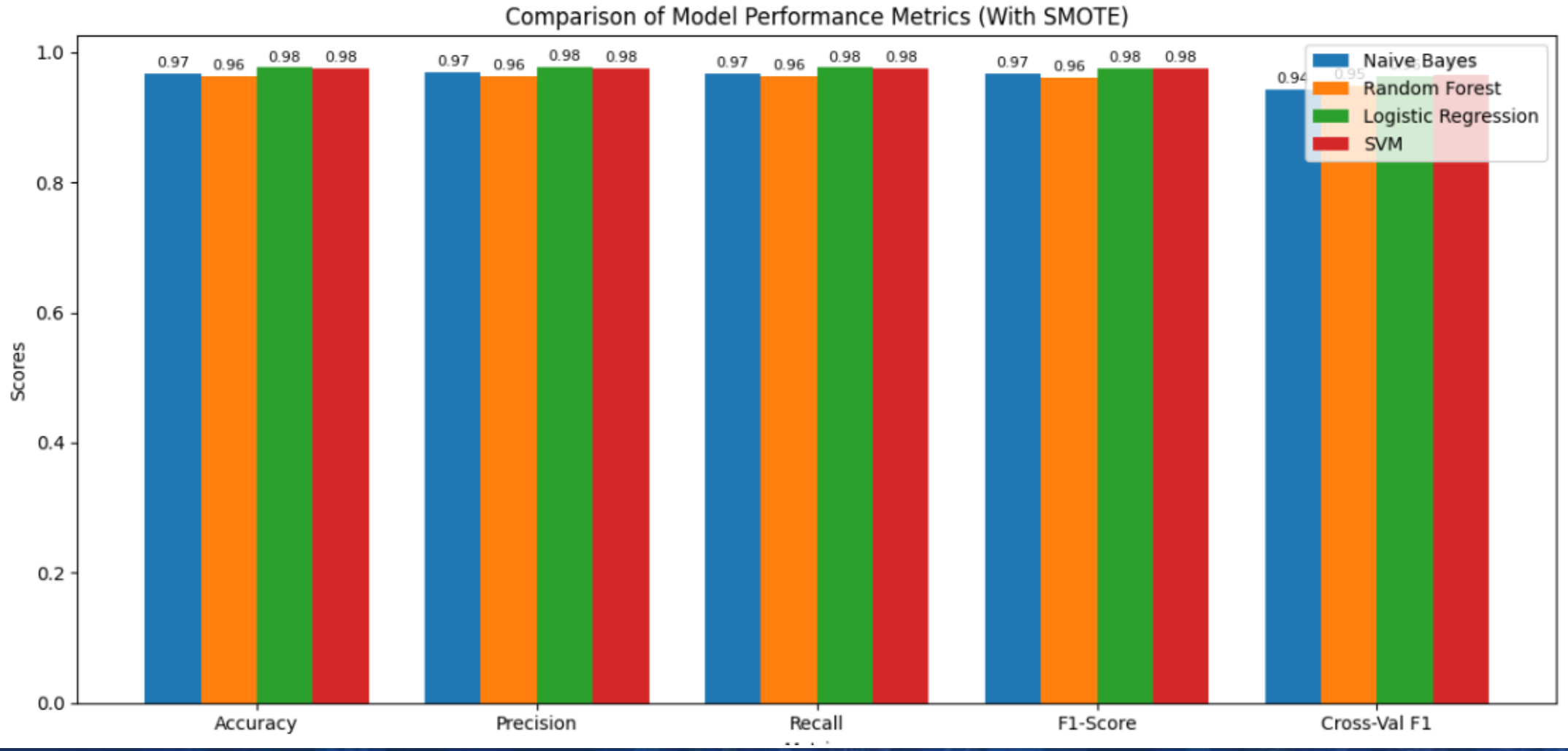


Nouvelle approche propo

Synthetic Minority Oversampling Technique



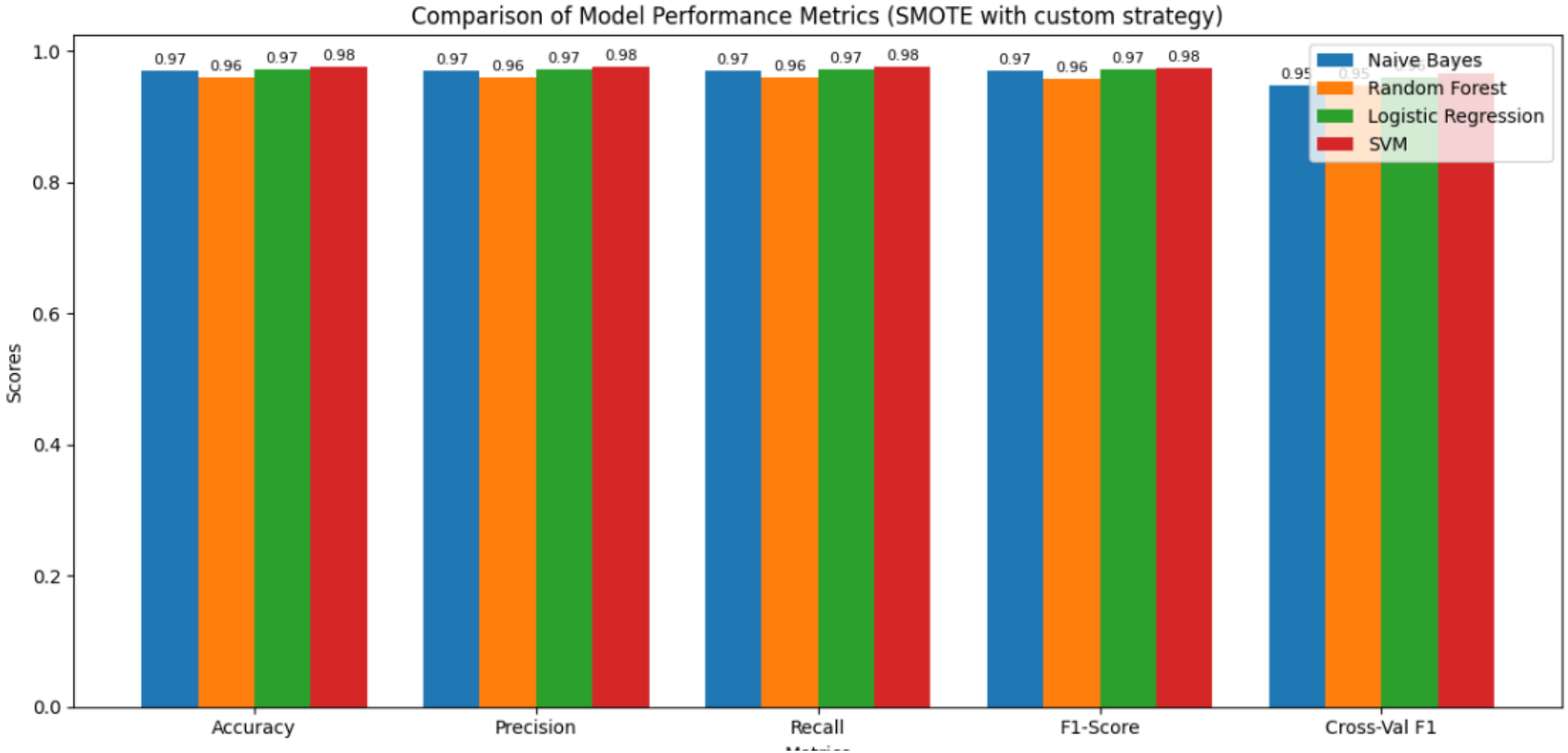
Ancienne approche ▶ ***Nouvelle approche***



Ancienne approche ▶ ***Nouvelle approche***

```
sampling_strategy = {
    'earn': 3945,          # *1
    'acq': 2834,           # *1.2
    'money-fx': 614,       # *2
    'interest': 712,       # *2.5
    'ship': 474            # *3
}
```

Ancienne approche ▶ ***Nouvelle approche***



CRÉDIT AGRICOLE
CORPORATE & INVESTMENT BANK

Classification des articles financiers avec Reuters-21578