Web scrapping price data: dealing with items' classification issues – the experience of AnStat (Côte d'Ivoire)

Presented by :

Mr Franck Arnold Junior MIGONE

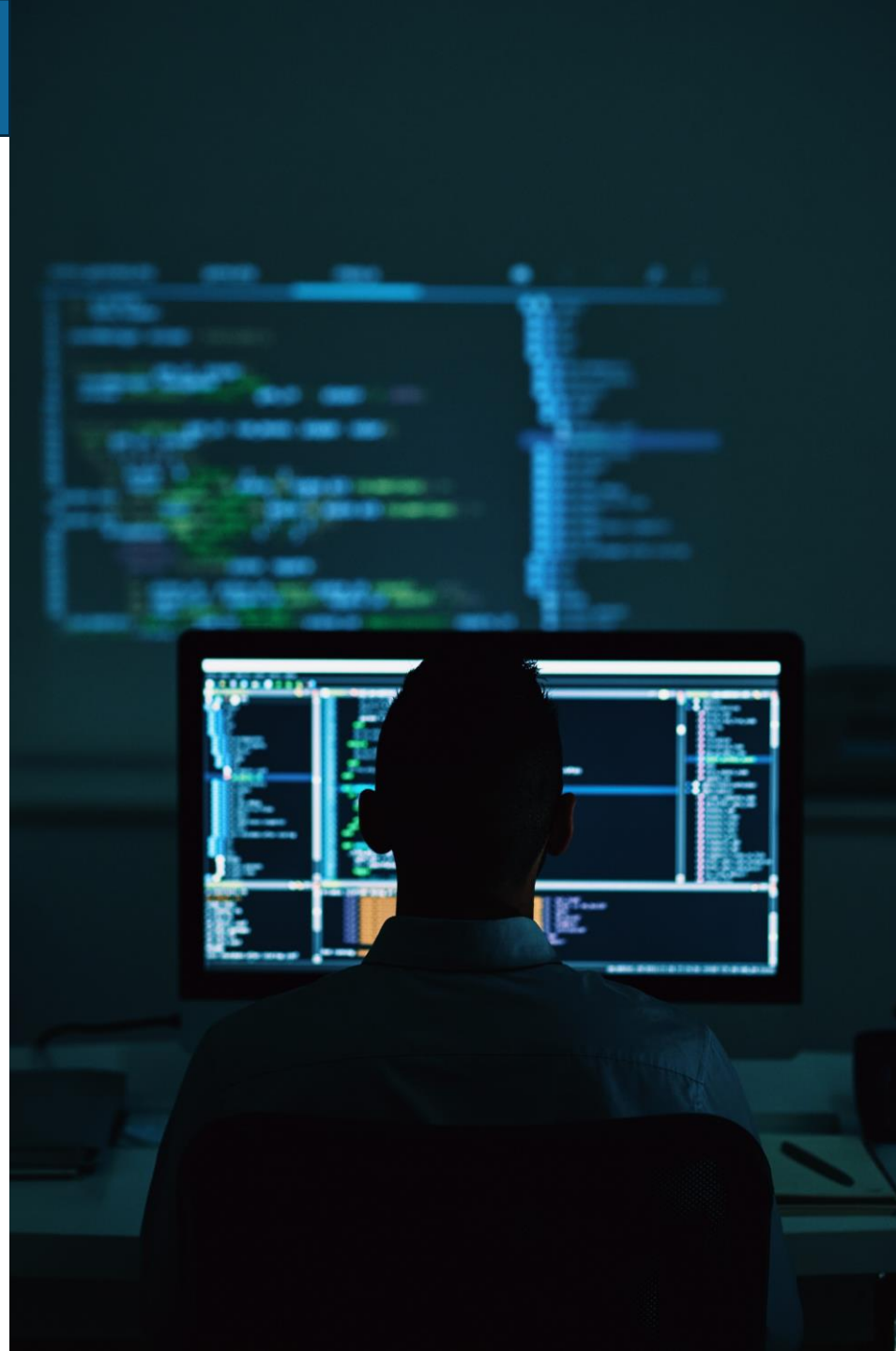
Table of contents

Project Context

With the significant advancement of new technologies, we are witnessing a new consumption behavior on digital platforms. In response, the National Agency of Statistics (**ANStat**) has made it its mission to optimize the calculation of the Consumer Price Index (CPI).

The main idea :

- to collect data from selected platforms, through web scraping methods or by interacting with an API provided by the platform.
- to properly process, and a dedicated processing system will be established to support the business workflow.
- Implement an automated workflow to streamline the entire chain.



COICOP Presentation

Code	Label Function
01	Food and non-alcoholic beverages
02	Alcoholic beverages, tobacco and narcotics
03	Clothing and footwear
04	Housing, water, gas, electricity and other fuels
05	Furniture, household equipment and routine household maintenance
06	Health
07	Transports
08	Communication
09	Recreation and culture
10	Education
11	Restaurants and hotels
12	Miscellaneous goods and services

Categories

- Function
- Group
- Sub group
- Post
- Variety

Data collection

Data Source

Function 1&2

Function 4

. .

Function 11

Function 12

Data Collection

Structure of scraped Data

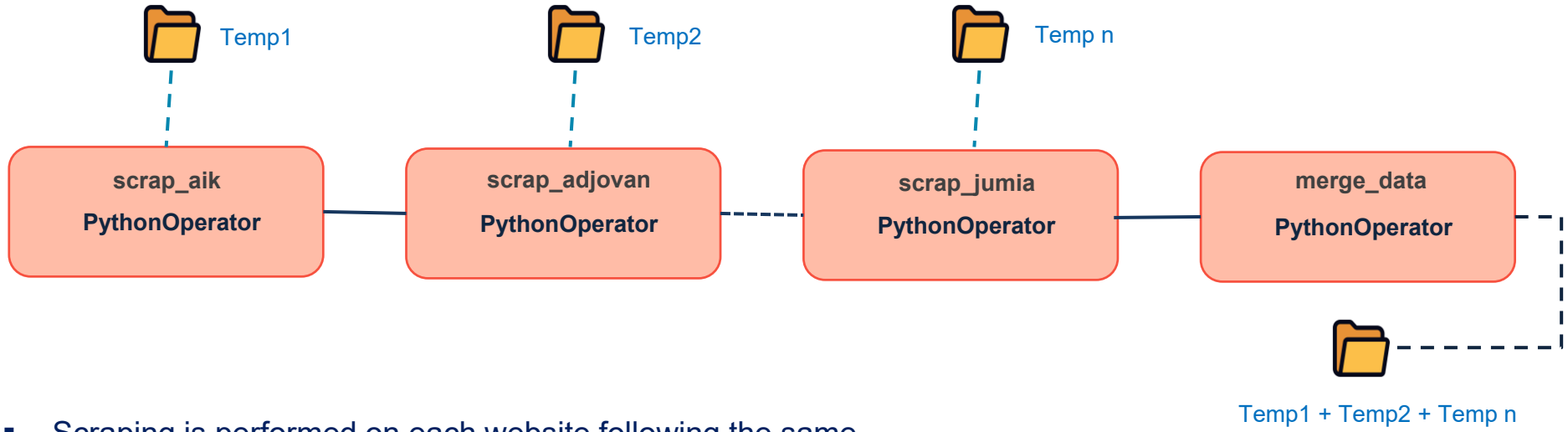
Date_de_Collecte (A)	Code_Site (B)	Libellé_du_produit (C)	Quantite (D)	Prix_du_produit (E)	Caracteristique (F)	Unite (G)	Unite_monétaire (H)

Srapping columns description

- A : Date when data was collected
- B: Scraped site link
- C: Product name extracted during scrapping
- D: Quantity extracted from the site
- E : Product price extracted during
- B: Characteristic extracted from product description
- C: Unit of measurement extracted from the site
- H: Currency extracted (FCFA)

Data Collection

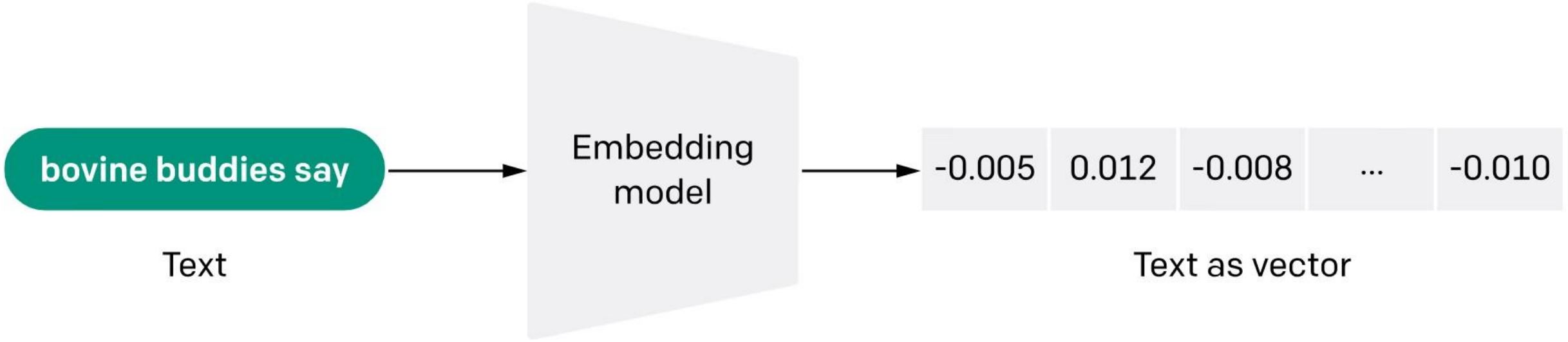
Scraping Mechanism



- Scraping is performed on each website following the same structure in order to enable data merging.
- The scraping scripts are integrated into the Airflow DAG to be executed on a daily basis.

Data Collection

Main Principles : Text Embeddings

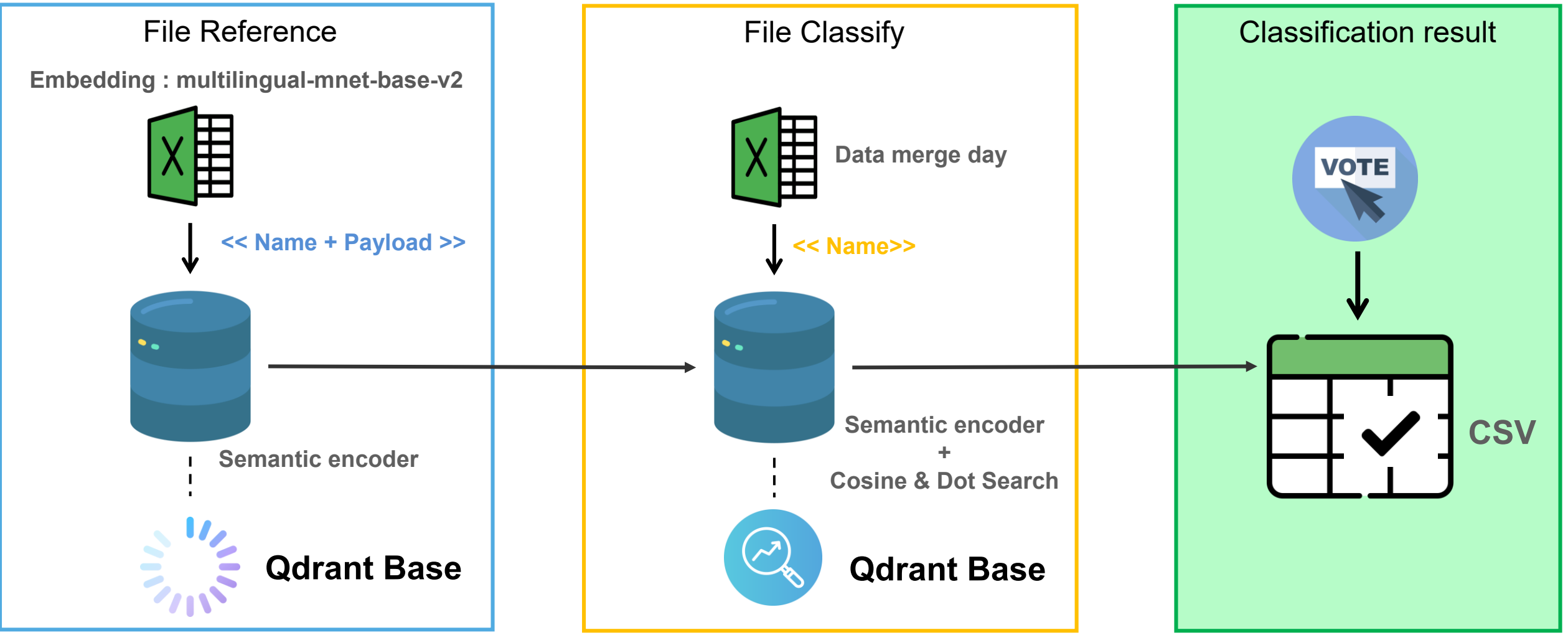


Structure of Input Data

	A	B	C	D	E	F
1	Sous classe (COICOP)	Sous classe (COICOP)	Code_produit	Code_ID	Caracteristique - Description	Libelle_Produit
2	01.1.2.3	01010203	0101020301	01010203010201	Le porc contient plusieurs nutriments essentiels à la santé. Il renf	COTES DE PORC SANS PEAU [500G]
3	01.1.2.3	01010203	0101020301	01010203010201	Tirée de la partie supérieure du cou, l'échine est un morceau sou	ÉCHINE DE PORC [500G]
4	01.1.2.3	01010203	0101020301	01010203010201	Tirée de la partie supérieure du cou, l'échine est un morceau sou	ÉCHINE DE PORC CARTON [10KG]
5	01.1.2.3	01010203	0101020301	01010203010201	Le porc contient plusieurs nutriments essentiels à la santé. Il renf	ENTRECOTE DE PORC [1KG]
6	01.1.2.1	01010201	0101020102	01010201020101	Simplissime à cuisiner, elle se prépare seule ou agrémentée d'une	FILET DE BOEUF [250G]]
7	01.1.2.1	01010201	0101020102	01010201020101	Sa grande valeur nutritive reste toujours la même quelle que soit	FOIE DE BOEUF [1KG]
8	01.1.2.3	01010203	0101020301	01010203010201	Lejarretest lapartiedes pattes soit entre l'épaule et le pied (la patt	JARRET DE PORC [1KG]
9	01.1.2.1	01010201	0101020102	01010201020101	La langue de bœuf est considérée comme un abat. Une langue pè	LANGUE DE BOEUF [ENTIERE]
10	01.1.2.2	01010202	0101020201	01010202010301	Mouton mâle ou femelle de ferme choisi pour le poids.	MOUTON ENTIER [FRAIS ET NETTOYÉ]
11	01.1.2.3	01010203	0101020301	01010203010201	Le porc contient plusieurs nutriments essentiels à la santé. Il renf	OREILLE DE PORC [1KG]

Data Collection

Codification



Data Collection

Codification

payload

```
def predict(input_dict, client, encoder):
    """
    Prediction given the request input
    :param input_dict: [dict], product to be classfied
    :param client: [QdrantClient], Qdrant client
    :param encoder: [SentenceTransformer], SentenceTransformer model for embeddings
    :return: [dict], prediction
    """

    name = input_dict.get("name")


    results_cosine = search_metrics(client, encoder, "cosine", name)
    results_dot = search_metrics(client, encoder, "dot", name)

    results_concat = pd.concat([results_cosine, results_dot])

    result = majority_vote(results_concat)

    result = {
        "ground_truth": input_dict.get("coicop_code"),
        "name": name,
        "classification": result.get("coicop_code"),
        "confidence": result.get("confidence")
    }

    return result
```

Data Collection

Codification

	A	B	C	D	E
1	ground_truth	name	classification	confidence	correct
2	1010703	HARICOT BLANC VRAC [300G]	1010201	1.0	False
3	1010203	CARTON DE COTE DE PORC (STERNUM) [10KG]	1010203	1.0	True
4	1010203	COTES DE PORC SANS PEAU [500G]	1010203	1.0	True
5	1010201	COTES/COTELETTES DE BOEUF FUMÃ‰ES [1/2KG]	1010201	1.0	True
6	1010203	Ã‰CHINE DE PORC [500G]	1010203	1.0	True
7	1010203	Ã‰CHINE DE PORC CARTON [10KG]	1010203	1.0	True
8	1010203	ENTRECOTE DE PORC [1KG]	1010203	1.0	True
9	1010201	FAUX FILET DE BOEUF [500G]	1010201	1.0	True
10	1010201	FILET DE BOEUF [250G]]	1010201	1.0	True
11	1010201	FOIE DE BOEUF [1KG]	1010201	1.0	True
12	1010202	GIGOT Dâ€™AGNEAU [1KG]	1010201	0.5	False
13	1010203	JARRET DE PORC [1KG]	1010203	1.0	True
14	1010201	LANGUE DE BOEUF [ENTIERE]	1010201	1.0	True
15	1010202	MOUTON ENTIER [FRAIS ET NETTOYÃ‰]	1010201	0.5	False
16	1010203	OREILLE DE PORC [1KG]	1010203	1.0	True
17	1010204	PACK 5 POULETS DE CHAIR FRAIS	1010204	1.0	True
18	1010204	PACK FUMÃ‰ [PONDEUSE, BOEUF ET CAPITAINE]	1010204	1.0	True
19	1010204	PACK FUMÃ‰ POULET POISSON VIANDE LIGHT	1010902	1.0	False
20	1010204	PACK MIX POULET 322	12010302	1.0	False

Data Collection

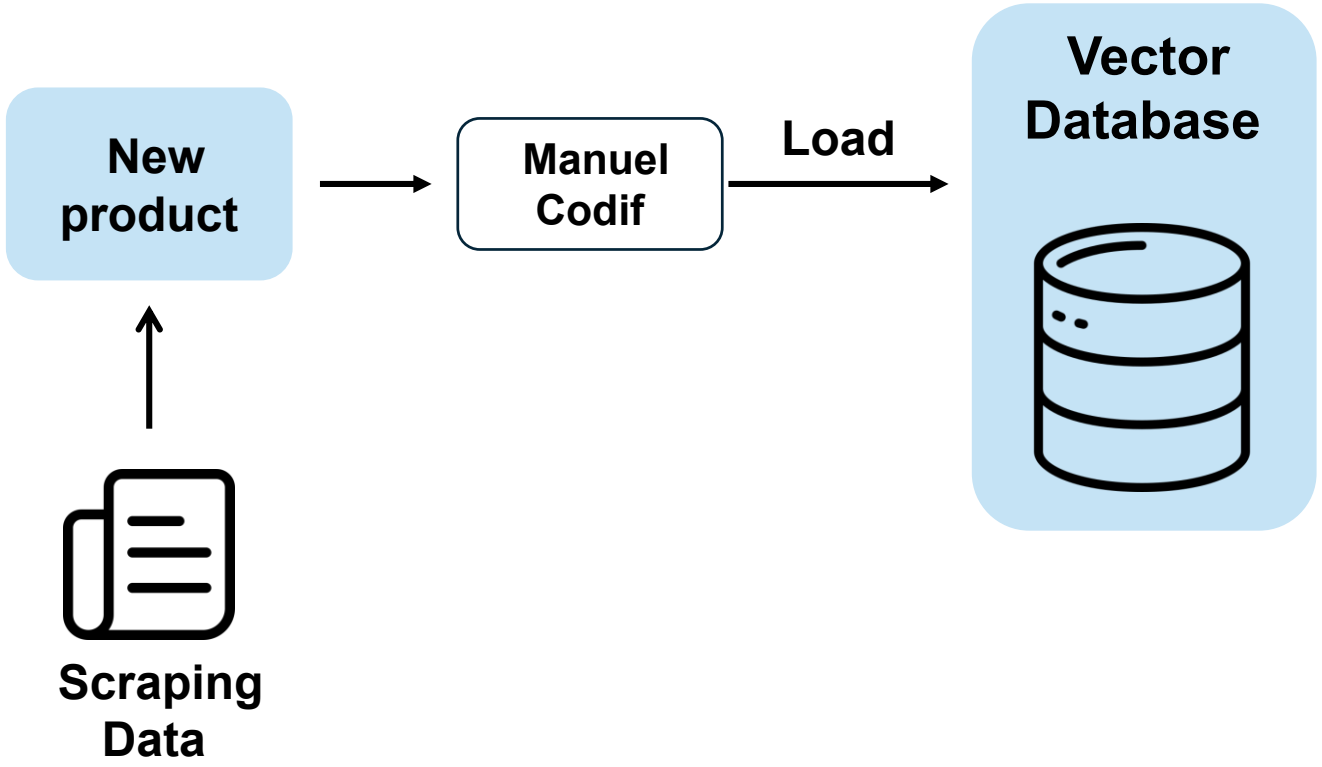
Performance

Accuracy	88,54
Precision	81,20
Recall	75,94

- **EXAMPLE DB :** 788 Product
- **Data Scrapped :** 671 Product for 1 month
- **Product Classified :** 541

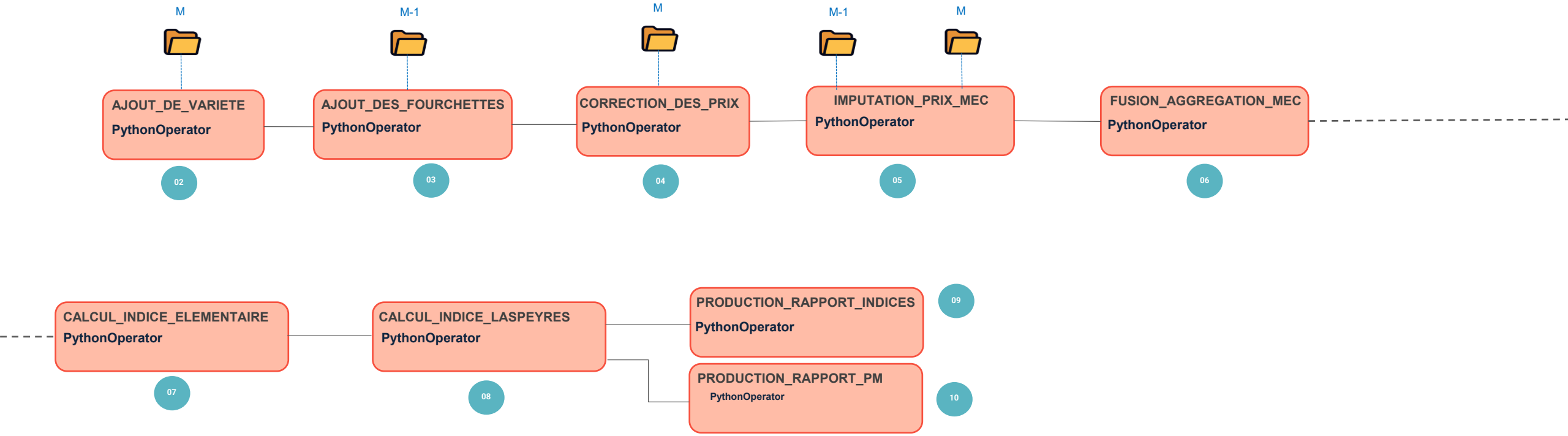
Data Collection

Data Base enrichment



Data Processing Pipeline

Processing Workflow



- Thus, each function generates its report by following the above mechanism.

Data Processing Pipeline

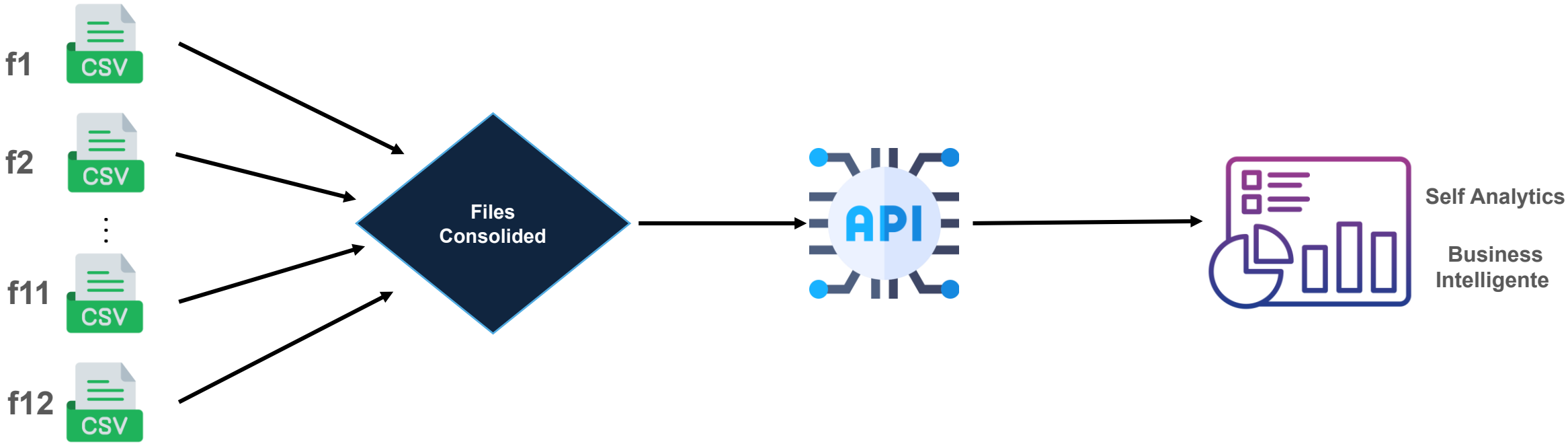
Final Data Structure

Label fonction	First month	. . .	Last month	1 month evolution	3 month evolution
Food products and non-alcoholic beverages					
. . .					
Global					

- **NB** : This structure will be the same for the group, sub-group, and elementary level.

Data Processing Pipeline

Generation of Global report



Languages & Technologies used

AirFlow

Creation, management, and scheduling of workflows

Python

Development of the preprocessing code and Implementation of the API

Minlo

Data Storage

Angula

Implementation of the view

Delta Lake

Analytical DataBase

https://github.com/cae-ins/CPI_Innovation_INS_CAE

https://github.com/cae-ins/codif_rag_ihpc

Perspectives and Improvements

Currently, our codification system is based on a RAG (Retrieval-Augmented Generation) approach, using vector searches with cosine and dot product similarity metrics to suggest the most relevant COICOP codes based on product labels. In the future, a major evolution of the process is planned:

the fine-tuning of a Large Language Model (LLM) dedicated to automatic codification, once a sufficient volume of reference data has been collected.

This approach will enable:

- More refined generalization, even for new or ambiguous formulations
- A fully automated and intelligent classification pipeline

The model will be integrated into our existing workflow, and will operate seamlessly on new products collected via web scraping or API.


THANK YOU FOR YOUR ATTENTION !