capstone_project

October 27, 2025

1 Prédiction du risque d'insécurité alimentaire en Haïti

1.0.1 Projet Capstone réalisé dans le cadre du Bootcamp de Data Science & Intelligence Artificielle

1.0.2 Akademi (powered by Flatiron School)

Rédigé par :
Richecard Blade DAMEUS & **Berothely THELUS**

Sous la supervision de :
M. Wedter JEROME & M. Geovany LAGUERRE

Date de présentation : Octobre 2025

Localisation : Port-au-Prince, Haïti

1.1 Objectif du projet

Ce projet vise à concevoir un **modèle prédictif** capable d'anticiper le **niveau d'insécurité alimentaire** à l'échelle des départements haïtiens, tout en analysant l'influence des indicateurs économiques, climatiques et environnementaux sur ces niveaux. À partir des données combinées, le modèle doit permettre de : - Prévoir les phases IPC (1 à 5) associées à chaque département ou commune, - Identifier les indicateurs clés (prix, pluviométrie, NDVI, taux de change, etc.) qui expliquent la dégradation ou l'amélioration des conditions alimentaires, - Fournir un outil d'aide à la décision pour anticiper les crises et orienter les interventions préventives.

1.2 Résumé du travail

En Haïti, l'insécurité alimentaire persiste dans un contexte marqué par des crises climatiques, économiques et sécuritaires (les groupes armés en Haïti ont ravagé des champs agricoles et entravent l'accès à l'aide alimentaire dans le pays). Pour y répondre, ce projet utilise la data science pour anticiper les risques à l'échelle communale. Il s'appuie sur des données du système Joint Monitoring Report (JMR) et du Integrated Food Security Phase Classification (IPC) pour construire un modèle supervisé capable d'identifier les zones à haut risque, contribuant ainsi à une meilleure planification humanitaire et agricole.

Le projet vise à offrir un outil concret d'aide à la décision pour renforcer la résilience alimentaire et mieux orienter les interventions humanitaires en Haïti.

1.3 Méthodologie utilisée

Le projet repose sur la méthodologie **CRISP-DM**, reconnue pour sa rigueur et sa clarté dans la conduite des projets de *Data Science*.
Cette approche couvre tout le processus analytique, depuis la compréhension du problème jusqu'à la restitution des résultats.

La carte suivante présente la structure administrative du territoire haïtien ainsi que le périmètre d'analyse retenu pour ce projet

Figure 1 — Carte d'Haïti : zones d'étude pour la prédiction du risque d'insécurité alimentaire.

1.3.1 Mots-clés :

`Data Science · Machine Learning · Sécurité alimentaire · Haïti · CRISP-DM · Classification · Analyse prédictive`

2 Partie 1 – Compréhension du domaine *(Business Understanding)*

2.1 Contexte et justification

Depuis plusieurs années, Haïti fait face à une **crise alimentaire profonde**.
Les épisodes de sécheresse, les inondations répétées, la dégradation des sols, la hausse des prix des denrées et la baisse de la production locale se conjuguent à une **insécurité grandissante** dans plusieurs régions rurales.

Les récents événements survenus dans la Plaine de l'Artibonite, le Bas-Sud et certaines zones du Nord-Est en sont une illustration :
des **groupes armés ont incendié ou pillé des champs de riz, de maïs et de haricots**, détruisant plusieurs récoltes et forçant des familles à abandonner leurs terres.
Ces attaques, relevées par le **Programme Alimentaire Mondial (PAM)**, aggravent une situation déjà fragile : près de **six millions de personnes** sont aujourd'hui exposées à un risque aigu d'insécurité alimentaire.

Les conséquences ne sont pas uniquement économiques.
La perte de moyens de subsistance, l'exode rural, la dépendance accrue à l'aide humanitaire et la hausse du coût de la vie créent un cercle vicieux qui érode les bases mêmes de la sécurité alimentaire nationale.
Dans ce contexte, disposer d'un outil capable d'**anticiper les zones à haut risque** devient essentiel pour permettre aux décideurs de **planifier et cibler leurs interventions** avant que la crise ne s'installe.

Selon la capsule matinale de ProFin diffusée le 23 octobre 2025, la Communauté des Caraïbes (CARICOM) et l'Argentine ont entamé des discussions autour de nouveaux partenariats stratégiques en matière de sécurité alimentaire et d'opportunités de coopération agroalimentaire pour la région caribéenne. Ces échanges visent à renforcer les chaînes régionales de production et de distribution, avec des retombées économiques et agricoles potentielles pour Haïti, notamment à travers le développement de filières locales intégrées et la promotion d'une autosuffisance alimentaire durable. Cette orientation régionale, axée sur l'autosuffisance alimentaire, s'accorde directement avec les ambitions de ce projet.

Sur le plan national, plusieurs initiatives soutiennent déjà cette dynamique :
- Le **Programme National de Cantines Scolaires (PNCS)**, appuyé par le **Programme Alimentaire Mondial**, achète chaque mois pour environ **1,7 million USD** de produits issus de la production locale.
Ce modèle, inspiré du *Home-Grown School Feeding*, relie écoles et agriculteurs tout en stimulant les économies rurales.
- À l'occasion de la Journée internationale des femmes rurales, célébrée le 15 octobre 2025, le Ministère de l'Agriculture, des Ressources Naturelles et du Développement Rural (MARNDR), en collaboration avec la FAO rappelle que plus de 70% des agriculteurs en Haïti sont des femmes pourtant freinées par l'accès à la terre, au crédit et aux circuits formels. Renforcer leur participation est une condition incontournable pour restaurer la productivité et la résilience alimentaire du pays.

Ces initiatives montrent qu'il existe déjà un élan de transformation du système agricole haïtien.
Le projet vient s'y inscrire en apportant une **dimension scientifique et prédictive**, permettant d'appuyer ces efforts par la donnée et l'analyse.

2.2 Problématique

> Comment anticiper les zones à haut risque d'insécurité alimentaire en Haïti à partir de données climatiques, géographiques et socio-économiques, afin d'aider les autorités et les partenaires à agir avant la crise ?

Cette question résume l'essence du projet : passer d'une **réaction tardive** à une **prévention éclairée par la donnée**.

2.3 Hypothèse

Les indicateurs climatiques (pluie, NDVI, sécheresse) et d'accessibilité influencent significativement le risque d'insécurité alimentaire.

Objectifs spécifiques

1. **Identifier les facteurs déterminants** (pluviométrie, production agricole, variables climatiques, accès aux marchés, structure des ménages, etc.) liés à l'insécurité alimentaire.
2. **Concevoir un modèle d'apprentissage supervisé** capable d'estimer la phase IPC de

2.4 Vision du projet

Ce travail repose sur la conviction que la **donnée peut devenir un levier de souveraineté alimentaire**.
En exploitant la méthode **CRISP-DM**, le projet relie l'analyse scientifique à l'action publique.
L'ambition n'est pas seulement de prédire : il s'agit de **comprendre les dynamiques qui nourrissent l'insécurité alimentaire**, d'en suivre les évolutions dans le temps et d'offrir un outil concret d'aide à la décision.

Cette démarche veut contribuer à la **construction d'une politique alimentaire plus résiliente**, centrée sur la valorisation de la production locale, la sécurité des zones rurales et la justice économique pour les communautés agricoles.

Acteurs clés du système alimentaire haïtien

Catégorie	Acteurs principaux	Rôle dans la sécurité alimentaire
Institutions nationales	MARNDR, PNCS, CNSA	Planification, coordination, achats locaux
Organisations internationales	FAO, PAM, CARICOM, WFP	Financement, assistance technique, distribution
Producteurs ruraux	Coopératives agricoles, femmes agricultrices	Production, transformation, résilience locale
Collectivités locales	Mairies, CASEC, délégations départementales	Gestion territoriale, identification des zones vulnérables
Communautés	Écoles, ménages, associations locales	Bénéficiaires directs, sensibilisation et participation communautaire

3 Partie 2 – Compréhension des données *(Data Understanding)*

Cette phase vise à comprendre la nature, la structure et la signification des données disponibles avant toute modélisation. Elle consiste à examiner leur origine, leur fiabilité, leur cohérence et leur potentiel d'analyse. Dans le cadre de ce projet, 2 sources principales pour 3 datasets ont été exploitées :

1. Les données du Joint Food Security Monitor - Haiti de la part de la Banque Mondiale
    1. Les données du Joint Monitoring Report (JMR)
    2. Le référentiel géographique administratif (PCodes)

2. Le jeu de données du système IPC (Integrated Food Security Phase Classification).

Ces trois jeux de données forment un socle d'analyse combinant la dimension temporelle, la dimension spatiale et la dimension structurelle de l'insécurité alimentaire en Haïti.

3.1 2.1 – Exploration initiale

Dans un premier temps, les bibliothèques nécessaires sont importées, puis les trois fichiers sont chargés afin d'examiner leur structure de base

[117]:
```
import pandas as pd

# Chargement des fichiers
jmr_data = pd.read_csv("HTI_JMR_data.csv")
ipc_data = pd.read_csv("ipc_hti_area_long_latest.csv")
pcodes = pd.read_csv("HTI_JMR_pcodes.csv")

# Aperçu des fichiers
print("=== APERÇU DU FICHIER JMR DATA ===")
display(jmr_data.head(3))
print("\nDimensions :", jmr_data.shape)

print("\n=== APERÇU DU FICHIER IPC DATA ===")
display(ipc_data.head(3))
print("\nDimensions :", ipc_data.shape)

print("\n=== APERÇU DU FICHIERPCODES ===")
display(pcodes.head(3))
print("\nDimensions :", pcodes.shape)
```

```
=== APERÇU DU FICHIER JMR DATA ===
```

	iso3	ipc phase	cutoff	adm2_pcode	year	month	indicator	\
0	HTI		3	HT0111	2010	1	Target	
1	HTI		3	HT0111	2010	4	Target	
2	HTI		3	HT0111	2010	7	Target	

	grouping	value	date
0	Original IPC phase	1.0	2010-01-01
1	Original IPC phase	2.0	2010-04-01
2	Original IPC phase	1.0	2010-07-01

```
Dimensions : (451920, 9)

=== APERÇU DU FICHIER IPC DATA ===
```

	Date of analysis	Country	Total country population	Level 1	\
0	#date+analysis	#country+code	#population	#loc+level1	
1	Sep 2025	HTI	11214615	Nippes	
2	Sep 2025	HTI	11214615	Nippes	

	Area	Validity period	From	To	Phase	\
0	#loc+area	#period+v_ipc	#date+start	#date+end	#severity+v_ipc	
1	Nippes HT01	current	2025-09-01	2026-02-28	all	

```
2  Nippes HT01         current   2025-09-01  2026-02-28              3+

          Number     Percentage
0  #affected+num  #affected+pct
1         218984            1.0
2         120442           0.55


Dimensions : (421, 11)

=== APERÇU DU FICHIERPCODES ===

  iso3 adm0_pcode country adm1_pcode    adm1_name adm2_pcode      adm2_name
0  HTI         HT   Haiti       HT08  Grande'Anse     HT0812       Abricots
1  HTI         HT   Haiti       HT03        North     HT0321   Acul du Nord
2  HTI         HT   Haiti       HT09   North-West     HT0922  Anse-a-Foleur


Dimensions : (140, 7)
```

3.1.1 Remarque importante :

En observant la première ligne du fichier IPC, on remarque qu'elle ne contient pas des données réelles, mais des indications descriptives (ex. : #date+analysis, #country+code, etc.). Cette ligne est donc une métadonnée qu'il faudra supprimer lors de la phase de préparation. La conserver risquerait de fausser les analyses statistiques et les conversions de types.

3.2 2.1 – Description des jeux de données

Trois sources principales ont été retenues :

3.2.1 a) HTI_JMR_data.csv

Ce jeu de données provient du **Joint Monitoring Report (JMR)**.
Ce fichier contient des observations mensuelles de l'évolution des phases IPC par commune depuis 2010. Il contient près de 451 920 observations décrivant les valeurs de référence liées à la classification IPC au niveau communal (adm2_pcode). Il inclut des variables telles que :

Variable	Description	Type
`iso3`	Code ISO du pays	Catégorielle
`ipc phase cutoff`	Niveau seuil IPC considéré	Numérique
`adm2_pcode`	Code administratif de la commune	Catégorielle
`year`, `month`, `date`	Variables temporelles	Temporelles
`indicator`, `grouping`	Type d'indicateur ou regroupement	Catégorielles
`value`	Valeur observée	Numérique

Ce jeu de données offre une vue chronologique continue de la situation alimentaire à travers les communes haïtiennes. Il permettra de suivre les évolutions dans le temps et d'extraire des tendances saisonnières ou structurelles.

3.2.2 HTI_JMR_pcodes.csv

Ce fichier fournit le référentiel géographique permettant d'associer chaque code à sa zone administrative.

Variable	Description
`adm1_name`	Département
`adm2_name`	Commune
`adm2_pcode`	Code administratif
`country`	Nom du pays (Haïti)

Ce fichier est essentiel pour lier les données JMR et IPC à leurs localisations géographiques.

3.2.3 c) ipc_hti_area_long_latest.csv

Ce dernier jeu de données regroupe les **résultats récents de l'analyse IPC (septembre 2025)**. Il décrit la population touchée par phase IPC.

Variable	Description
`Date of analysis`	Mois et année de l'analyse
`Level 1`	Département
`Phase`	Niveau d'insécurité alimentaire (1 à 5)
`Number`	Population touchée
`Percentage`	Proportion dans la population totale
`From / To`	Période de validité de l'évaluation

3.3 2.2 Vérification de la qualité et de la cohérence des données

On examine les types de variables, la présence de valeurs manquantes et de doublons et la cohérence des formats.

[118]:
```
def missing(df):
    m = df.isnull().sum().sort_values(ascending=False)
    t = df.dtypes
    return pd.DataFrame({"Valeurs manquantes": m, "Types": t})

for name, df in {"JMR": jmr_data, "IPC": ipc_data, "PCODES": pcodes}.items():
    print(f"\n=== {name} ===")
    display(missing(df).head(10))
    print("Doublons :", df.duplicated().sum())
```

```
=== JMR ===
```

	Valeurs manquantes	Types
adm2_pcode	0	object
date	0	object
grouping	0	object

```
indicator                          0   object
ipc phase cutoff                   0    int64
iso3                               0   object
month                              0    int64
value                              0  float64
year                               0    int64
```

```
Doublons : 0
```

```
=== IPC ===
```

```
                          Valeurs manquantes   Types
Area                                       0  object
Country                                    0  object
Date of analysis                           0  object
From                                       0  object
Level 1                                    0  object
Number                                     0  object
Percentage                                 0  object
Phase                                      0  object
To                                         0  object
Total country population                   0  object
```

```
Doublons : 0
```

```
=== PCODES ===
```

```
            Valeurs manquantes   Types
adm0_pcode                   0  object
adm1_name                    0  object
adm1_pcode                   0  object
adm2_name                    0  object
adm2_pcode                   0  object
country                      0  object
iso3                         0  object
```

```
Doublons : 0
```

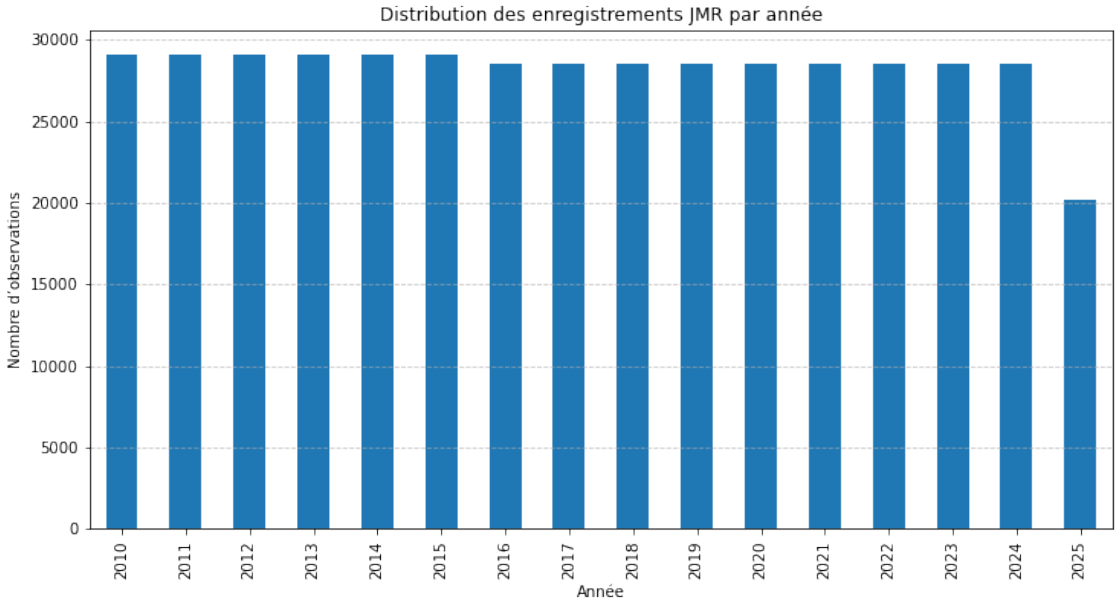
3.4 2.3 Exploration graphique

3.4.1 a. Répartition temporelle des données JMR

Cette visualisation permet d'identifier la période de couverture des données.

[119]:
```
import matplotlib.pyplot as plt

plt.figure(figsize=(12,6))
jmr_data['year'].value_counts().sort_index().plot(kind='bar')
plt.title('Distribution des enregistrements JMR par année')
plt.xlabel('Année')
plt.ylabel('Nombre d'observations')
```

```
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```



Les observations couvrent une période allant de 2010 à 2025, offrant un large historique pour l'analyse.

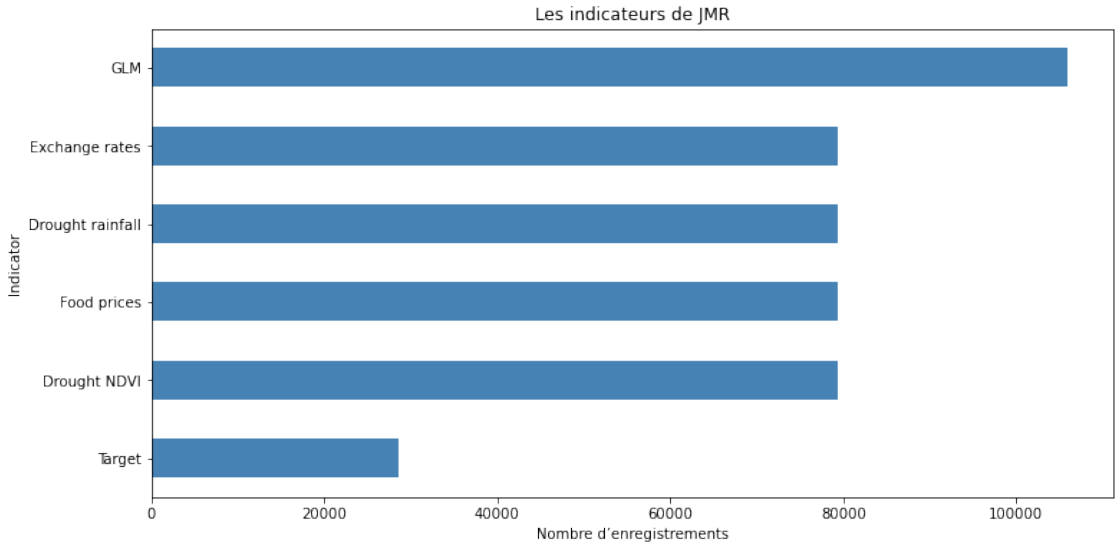
3.4.2 Repartition des indicateurs du JMR

La colonne `indicator` du fichier `HTI_JMR_data.csv` regroupe les types d'informations suivies dans le système de surveillance alimentaire.
Ces indicateurs traduisent la **réalité économique, climatique et environnementale** des communes haïtiennes.
Ils constituent la base des observations utilisées par le système IPC pour évaluer le risque d'insécurité alimentaire.

Le graphique ci-dessous présente les indicateurs les plus fréquents dans le jeu de données JMR :

[120]:
```
if "indicator" in jmr_data.columns:
    ind = jmr_data["indicator"].value_counts().head(30)
    display(ind)

    plt.figure(figsize=(12,6))
    ind.sort_values().plot(kind="barh", color="steelblue")
    plt.title("Les indicateurs de JMR")
    plt.xlabel("Nombre d'enregistrements")
    plt.ylabel("Indicator")
    plt.show()
```

```
GLM                105840
Drought NDVI        79380
Food prices         79380
Drought rainfall    79380
Exchange rates      79380
Target              28560
Name: indicator, dtype: int64
```



Les six premiers indicateurs observés sont : GLM, Drought rainfall, Drought NDVI, Exchange rates, Food Prices, et Target.

1. GLM

L'indicateur GLM correspond à la donnée de référence principale utilisée dans les rapports de suivi. Il sert souvent à agréger plusieurs sources d'information dans une même unité géographique. Sa présence dominante montre que les évaluations IPC reposent sur une base géospatiale consolidée, utilisée pour suivre l'évolution des conditions dans le temps. En d'autres termes, il permet de relier les données locales à une carte administrative normalisée.

2. Drought rainfall

Cet indicateur concerne le déficit pluviométrique, c'est-à-dire le manque de pluie observé dans certaines zones. Dans le contexte haïtien, où plus de 70 % des exploitations agricoles sont pluviales, un déficit de précipitations provoque directement la baisse de rendement, la perte de récoltes et la raréfaction des produits vivriers sur les marchés. Les périodes de sécheresse prolongée dans l'Artibonite, le Nord-Est ou le Plateau Central ont historiquement entraîné des hausses de prix et des déplacements de population.

3. Drought NDVI

Le NDVI (Normalized Difference Vegetation Index) est un indice dérivé d'images satellites permettant de mesurer la densité de végétation. Lorsque cet indicateur est bas, cela signifie que la

couverture végétale se dégrade, souvent à cause d'un manque d'eau ou d'une pression anthropique sur les sols. Dans les communes rurales, un NDVI faible traduit une vulnérabilité accrue des moyens de subsistance, notamment pour les ménages dépendant de l'agriculture et de l'élevage.

4. Exchange rates

L'indicateur Exchange rates mesure les variations du taux de change entre la gourde haïtienne et le dollar américain. Dans un pays où la majorité des denrées de base (riz, farine, huile, sucre) sont importées, la dépréciation de la gourde a un effet immédiat sur le prix des produits alimentaires. Ce facteur économique pèse lourdement sur l'accès à la nourriture pour les familles à faibles revenus. Ainsi, le lien entre taux de change et sécurité alimentaire est direct et mesurable.

5. Food Prices

Les prix alimentaires sont un indicateur clé du pouvoir d'achat et de la stabilité économique. Lorsque les prix des denrées de base augmentent, la proportion de ménages en insécurité alimentaire s'accroît. En Haïti, les hausses de prix liées à la volatilité internationale ou à la rareté locale amplifient les inégalités d'accès à la nourriture. Cet indicateur traduit donc l'effet combiné des facteurs macroéconomiques et structurels sur le quotidien des ménages.

6. Target

L'indicateur Target fait référence aux zones ou groupes prioritaires visés par les interventions humanitaires. Il indique souvent les territoires déjà identifiés comme vulnérables par les agences locales (MARNDR, CNSA, PAM). Son analyse permet de vérifier la cohérence entre la réalité observée et les priorités définies dans les programmes d'aide.

A partir de l'analyse des indicateurs du fichier JMR, on peut remarquer que la sécurité alimentaire ne dépend pas d'un seul facteur. Elle est le résultat de l'interaction entre le climat, l'économie et les dynamiques sociales :

Dimension	Indicateurs liés	Impact sur la sécurité alimentaire
Climatique	Drought rainfall, Drought NDVI	Affecte la production et les revenus agricoles
Économique	Exchange rates, Food Prices	Affecte l'accès à la nourriture et le pouvoir d'achat
Institutionnelle	GLM, Target	Oriente la planification et les interventions

3.4.3 Analyse critique sur la colonne "Phase" du fichier IPC

En observant la visualisation des phases IPC, plusieurs **valeurs incohérentes** apparaissent sur l'axe X du graphique.
Les barres affichent notamment des étiquettes telles que **"3+"** et **"ALL"**, en plus des valeurs attendues **1, 2, 3, 4 et 5**.

Ces deux valeurs particulières méritent une explication :

- **"3+"** : dans le langage IPC, cette notation signifie *« Phase 3 et au-delà »*, autrement dit toutes les zones classées en crise ou en urgence.
  → Ce n'est pas une phase distincte, mais un regroupement utilisé pour les rapports de synthèse.

- **“ALL”** : cette valeur correspond à la somme totale ou au résumé global de toutes les phases pour une zone donnée.
  → Ce n’est pas une donnée exploitable pour la modélisation car elle **duplique l’information déjà contenue dans les autres lignes**.

Ces valeurs parasites peuvent fausser l’interprétation et biaiser les analyses descriptives.
Elles seront donc **supprimées à la** Phase 3 – Préparation des données**, afin que la variable `Phase` ne contienne** que des entiers 1 à 5**, représentant clairement les niveaux suivants :

Code	Interprétation IPC
1	Minimale / Sécurisée
2	Sous pression
3	Crise
4	Urgence
5	Catastrophe / Famine

Ainsi, lors de la préparation, nous allons : 1. **Retirer la première ligne non pertinente** (déjà identifiée) ;
2. **Nettoyer les valeurs “3+” et “ALL”** pour ne garder que les phases 1 à 5 ;
3. **Convertir la colonne `Phase` en type numérique** afin de faciliter les statistiques et la modélisation.

Cette observation souligne l’importance de la compréhension approfondie avant toute transformation :
c’est grâce à cette étape que l’on identifie les incohérences susceptibles d’altérer la qualité du modèle prédictif.

On peut vérifier également s’il y a une certaine relation entre 2 datasets au moins

[35]:
```
merged_check = jmr_data.merge(pcodes, on='adm2_pcode', how='left')
taux = round(merged_check['adm1_name'].notnull().mean()*100, 2)
print("Taux de correspondance entre JMR et PCodes :", taux, "%")
print("Cela confirme que les fichiers peuvent être fusionnés sans perte␣
  ↪significative.")
```

```
Taux de correspondance entre JMR et PCodes : 100.0 %
Cela confirme que les fichiers peuvent être fusionnés sans perte significative.
```

4 Phase 3 – Préparation des données (Data Preparation)

Cette phase marque le passage entre la compréhension des données et la construction du modèle. De ce fait, au cours de cette étape, on va essayer de rendre les données exploitables pour la modélisation.
C’est ici que l’on va : - Nettoyer et harmoniser les données issues des différentes sources,
- Les fusionner pour former un jeu de données complet,
- Explorer en profondeur les relations entre variables,

- Identifier les **variables explicatives (features)** et la **variable cible (target)** avant la modélisation..

4.1 3.1 Nettoyage des fichiers

4.1.1 - IPC

Le fichier IPC contient la variable cible de notre projet : la **phase d'insécurité alimentaire** (1 à 5).

Lors de la phase précédente, nous avions identifié plusieurs anomalies : - une première ligne descriptive non pertinente ;
- des valeurs incohérentes comme "3+" et "ALL".

Nous commençons donc par corriger ces éléments et d'agir sur d'autres elements.

[36]:
```
# Suppression de la première ligne descriptive
ipc_data = ipc_data.iloc[1:]

# Suppression des valeurs non pertinentes
ipc_data = ipc_data[~ipc_data['Phase'].isin(['ALL'])]

ipc_data['Phase'] = ipc_data['Phase'].replace('3+', 4)

# Suppression des colonnes non pertinentes
cols_drop_ipc = ['Country','Validity period', 'From', 'To', 'iso3', 'Area',␣
 ↪'Percentage']
ipc_data = ipc_data.drop(columns=[c for c in cols_drop_ipc if c in ipc_data.
 ↪columns])

# Renommage des colonnes
ipc_data = ipc_data.rename(columns={
    'Date of analysis': 'Date',
    'Total country population': 'Population totale',
    'Level 1': 'Departement',
    'Number' : 'Population touchee'
})


# Conversion du type de la colonne Phase
ipc_data['Phase'] = pd.to_numeric(ipc_data['Phase'], errors='coerce')

#correction de la region de ZMPAP figurant comme un departement

ipc_data["Departement"] = (
    ipc_data["Departement"]
    .str.strip()
    .replace({
        "ZMPAP":"Ouest"
```

```
    })
)

ipc_data.dropna(inplace=True)

ipc_data
```

[36]:

	Date	Population totale	Departement	Phase	Population touchee
2	Sep 2025	11214615	Nippes	4.0	120442
3	Sep 2025	11214615	Nippes	1.0	43797
4	Sep 2025	11214615	Nippes	2.0	54746
5	Sep 2025	11214615	Nippes	3.0	87594
6	Sep 2025	11214615	Nippes	4.0	32848
..	…	…	…	…	…
416	Sep 2025	11214615	Grand-Anse	1.0	104513
417	Sep 2025	11214615	Grand-Anse	2.0	104513
418	Sep 2025	11214615	Grand-Anse	3.0	146318
419	Sep 2025	11214615	Grand-Anse	4.0	62708
420	Sep 2025	11214615	Grand-Anse	5.0	0

[360 rows x 5 columns]

4.1.2 - PCodes

Dans ce fichier, le nom des colonnes et les départements sont en anglais (North, South-East), alors que ceux dans IPC sont en français (Nord, Sud-Est).

Pour éviter les erreurs de fusion , nous harmonisons les noms et nous selectinnons uniquement les colonnes qui nous interessent.

[37]:
```
# Sélection des colonnes essentielles
pcodes = pcodes[['adm2_pcode', 'adm2_name', 'adm1_name']]

# Renommage
pcodes = pcodes.rename(columns={
    'adm2_pcode': 'Code_Postal',
    'adm2_name': 'Commune',
    'adm1_name': 'Departement'
})

# Harmonisation linguistique
pcodes['Departement'] = pcodes['Departement'].replace({
    "North": "Nord",
    "North-West": "Nord-Ouest",
    "North-East": "Nord-Est",
    "South": "Sud",
    "South-East": "Sud-Est",
    "West": "Ouest",
```

```
    "Grande'Anse": "Grand-Anse"
})

pcodes
```

[37]:

	Code_Postal	Commune	Departement
0	HT0812	Abricots	Grand-Anse
1	HT0321	Acul du Nord	Nord
2	HT0922	Anse-a-Foleur	Nord-Ouest
3	HT0234	Anse-a-Pitre	Sud-Est
4	HT1021	Anse-a-Veau	Nippes
..	…	…	…
135	HT0753	Tiburon	Sud
136	HT0712	Torbeck	Sud
137	HT0431	Trou du Nord	Nord-Est
138	HT0441	Vallieres	Nord-Est
139	HT0532	Verrettes	Artibonite

[140 rows x 3 columns]

4.1.3 - JMR

Ce fichier recense les indicateurs de suivi tels que la sécheresse, les taux de change et les prix alimentaires. Il est la base principale de la modélisation.

[38]:
```
# Suppression de colonnes inutiles
cols_to_drop = ["ipc phase cutoff", "iso3", "grouping", "year",␣
 ↪"month",'Valeur']
jmr_data = jmr_data.drop(columns=cols_to_drop, errors="ignore")

# Renommage des colonnes pour cohérence nationale
jmr_data = jmr_data.rename(columns={
    'indicator': 'Indicateur',
    'value': 'Valeur',
    'adm2_pcode': 'Code_Postal',
    'date' : 'Date'
})
jmr_data
```

[38]:

	Code_Postal	Indicateur	Valeur	Date
0	HT0111	Target	1.0	2010-01-01
1	HT0111	Target	2.0	2010-04-01
2	HT0111	Target	1.0	2010-07-01
3	HT0111	Target	1.0	2010-10-01
4	HT0111	Target	2.0	2011-01-01
…	…	…	…	…
451915	HT1032	Food prices	0.0	2025-05-01
451916	HT1032	Food prices	0.0	2025-06-01

```
451917      HT1032  Food prices     0.0  2025-07-01
451918      HT1032  Food prices     0.0  2025-08-01
451919      HT1032  Food prices     0.0  2025-09-01

[451920 rows x 4 columns]
```

[39]:
```
# Fusion JMR + PCodes
merged_data = jmr_data.merge(pcodes, on='Code_Postal', how='left')
merged_data
```

[39]:
```
       Code_Postal   Indicateur  Valeur        Date         Commune  \
0           HT0111       Target     1.0  2010-01-01  Port-au-Prince
1           HT0111       Target     2.0  2010-04-01  Port-au-Prince
2           HT0111       Target     1.0  2010-07-01  Port-au-Prince
3           HT0111       Target     1.0  2010-10-01  Port-au-Prince
4           HT0111       Target     2.0  2011-01-01  Port-au-Prince
…             …          …     …          …              …
451915      HT1032  Food prices     0.0  2025-05-01    Grand-Boucan
451916      HT1032  Food prices     0.0  2025-06-01    Grand-Boucan
451917      HT1032  Food prices     0.0  2025-07-01    Grand-Boucan
451918      HT1032  Food prices     0.0  2025-08-01    Grand-Boucan
451919      HT1032  Food prices     0.0  2025-09-01    Grand-Boucan

       Departement
0            Ouest
1            Ouest
2            Ouest
3            Ouest
4            Ouest
…            …
451915      Nippes
451916      Nippes
451917      Nippes
451918      Nippes
451919      Nippes

[451920 rows x 6 columns]
```

[40]:
```
# Fusion du résultat avec IPC
final_data = merged_data.merge(ipc_data,  on='Departement', how='left')
final_data
```

[40]:
```
         Code_Postal   Indicateur  Valeur      Date_x         Commune  \
0           HT0111       Target     1.0  2010-01-01  Port-au-Prince
1           HT0111       Target     1.0  2010-01-01  Port-au-Prince
2           HT0111       Target     1.0  2010-01-01  Port-au-Prince
3           HT0111       Target     1.0  2010-01-01  Port-au-Prince
```

```
4             HT0111       Target     1.0  2010-01-01  Port-au-Prince
…              …          …     …         …              …
18089707      HT1032  Food prices     0.0  2025-09-01    Grand-Boucan
18089708      HT1032  Food prices     0.0  2025-09-01    Grand-Boucan
18089709      HT1032  Food prices     0.0  2025-09-01    Grand-Boucan
18089710      HT1032  Food prices     0.0  2025-09-01    Grand-Boucan
18089711      HT1032  Food prices     0.0  2025-09-01    Grand-Boucan

         Departement    Date_y Population totale  Phase Population touchee
0              Ouest  Sep 2025          11214615    4.0              289105
1              Ouest  Sep 2025          11214615    1.0              180691
2              Ouest  Sep 2025          11214615    2.0              252967
3              Ouest  Sep 2025          11214615    3.0              216829
4              Ouest  Sep 2025          11214615    4.0               72276
…               …       …                …    …                 …
18089707      Nippes  Sep 2025          11214615    1.0               52663
18089708      Nippes  Sep 2025          11214615    2.0               52663
18089709      Nippes  Sep 2025          11214615    3.0               73728
18089710      Nippes  Sep 2025          11214615    4.0               31598
18089711      Nippes  Sep 2025          11214615    5.0                   0

[18089712 rows x 10 columns]
```

[41]:
```python
# Suppression des colonnes parasites
cols_to_remove = ['Date_y']
final_data = final_data.drop(columns=[c for c in cols_to_remove if c in␣
 ↪final_data.columns], errors='ignore')

# Suppression des doublons et lignes vides
final_data = final_data.drop_duplicates().dropna(how='all')

# Suppression des enregistrements sans phase IPC
final_data = final_data.dropna(subset=['Phase'])


print("Dimensions après nettoyage :", final_data.shape)
display(final_data.head(5))
```

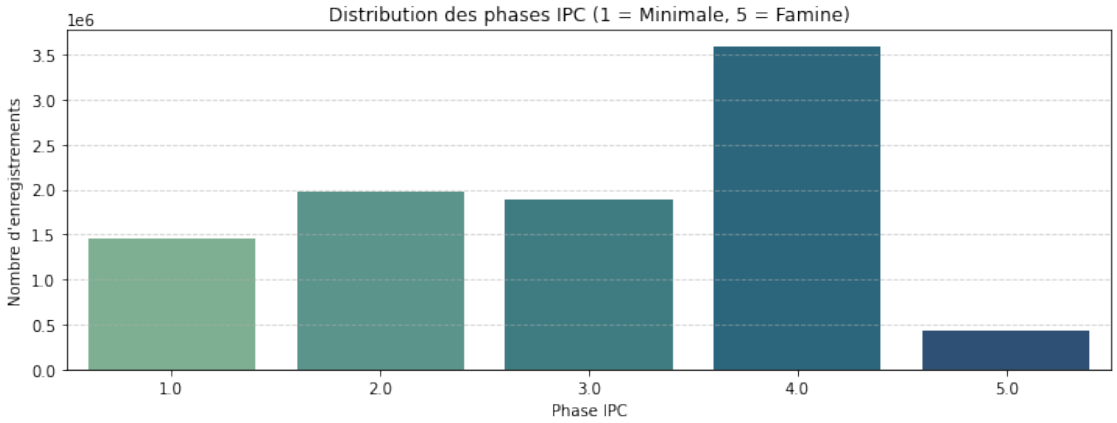
```
Dimensions après nettoyage : (9341421, 9)

  Code_Postal Indicateur  Valeur      Date_x         Commune Departement  \
0      HT0111     Target     1.0  2010-01-01  Port-au-Prince       Ouest
1      HT0111     Target     1.0  2010-01-01  Port-au-Prince       Ouest
2      HT0111     Target     1.0  2010-01-01  Port-au-Prince       Ouest
3      HT0111     Target     1.0  2010-01-01  Port-au-Prince       Ouest
4      HT0111     Target     1.0  2010-01-01  Port-au-Prince       Ouest

  Population totale  Phase Population touchee
```

0	11214615	4.0	289105
1	11214615	1.0	180691
2	11214615	2.0	252967
3	11214615	3.0	216829
4	11214615	4.0	72276

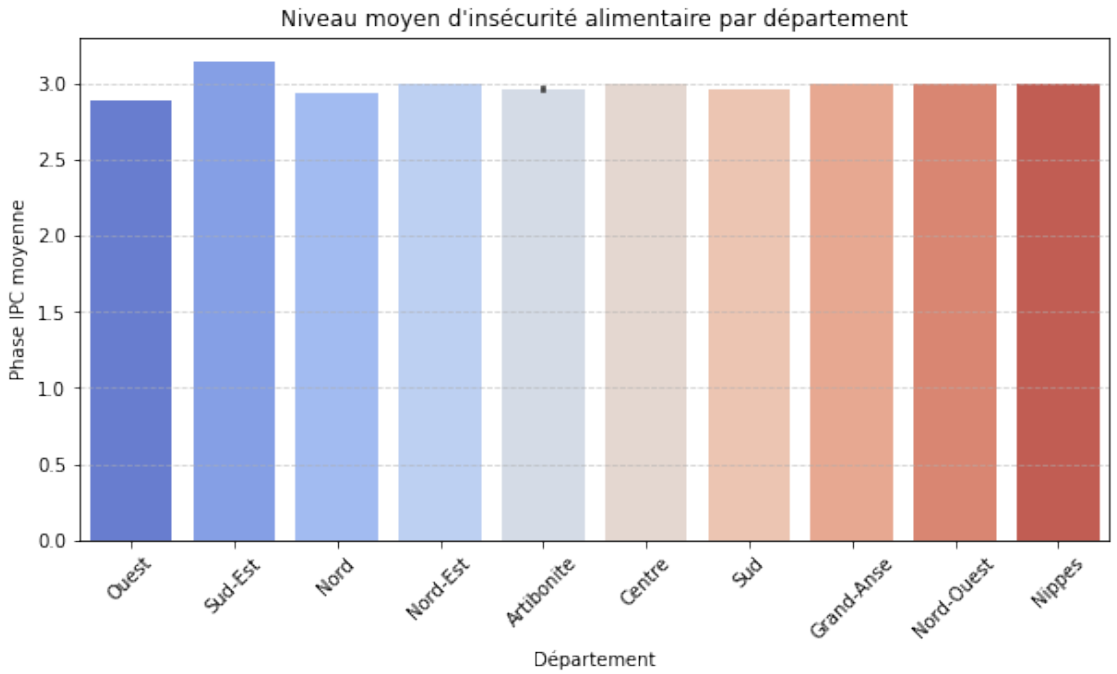
[42]:
```
# Taux de correspondance
match_rate = round(final_data['Phase'].notnull().mean() * 100, 2)
print("Taux de correspondance des phases IPC :", match_rate, "%")
```

```
Taux de correspondance des phases IPC : 100.0 %
```

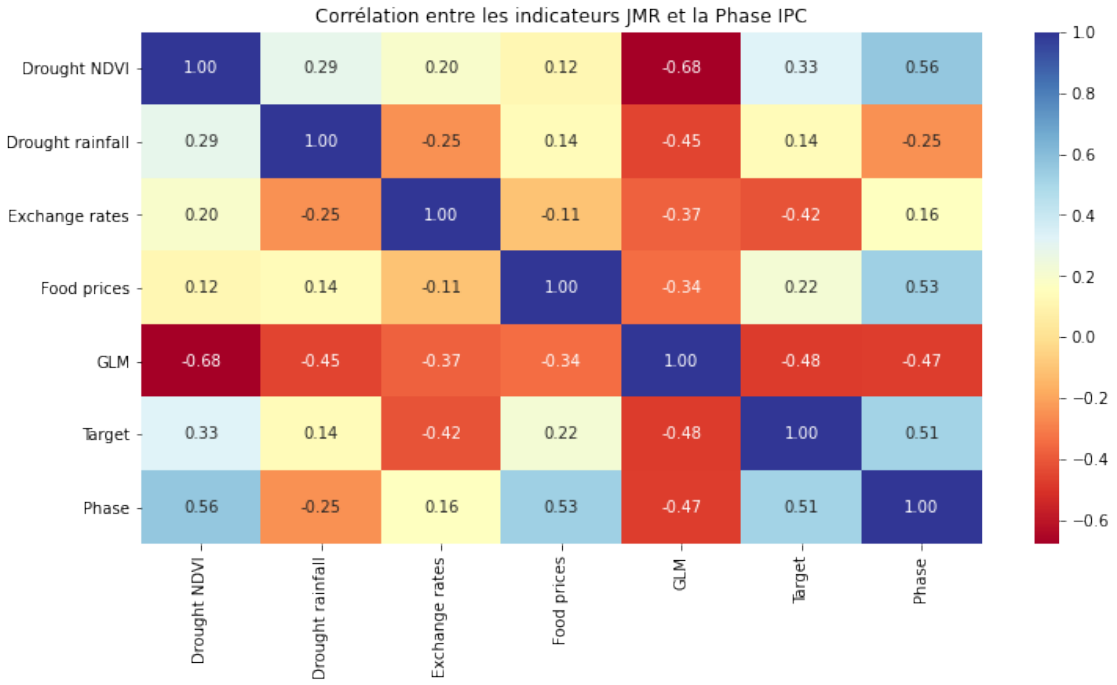
[43]:
```
import seaborn as sns
plt.figure(figsize=(12,4))
sns.countplot(x='Phase', data=final_data, palette='crest')
plt.title("Distribution des phases IPC (1 = Minimale, 5 = Famine)", fontsize=12)
plt.xlabel("Phase IPC")
plt.ylabel("Nombre d'enregistrements")
plt.grid(axis='y', linestyle='--', alpha=0.6)
plt.show()
```



[44]:
```
plt.figure(figsize=(10,5))
sns.barplot(x='Departement', y='Phase', data=final_data, palette='coolwarm')
plt.title("Niveau moyen d'insécurité alimentaire par département", fontsize=12)
plt.xlabel("Département")
plt.ylabel("Phase IPC moyenne")
plt.xticks(rotation=45)
plt.grid(axis='y', linestyle='--', alpha=0.6)
plt.show()
```



[45]:
```
# Création d'une table pivotée pour corrélation
pivot_data = final_data.pivot_table(
    index='Departement',
    columns='Indicateur',
    values='Valeur',
    aggfunc='mean'
).reset_index()

# Ajout de la moyenne IPC
phase_mean = final_data.groupby('Departement')['Phase'].mean().reset_index()
merged_corr = pivot_data.merge(phase_mean, on='Departement')

# Visualisation corrélation
plt.figure(figsize=(12,6))
sns.heatmap(merged_corr.corr(), annot=True, cmap="RdYlBu", fmt=".2f")
plt.title("Corrélation entre les indicateurs JMR et la Phase IPC", fontsize=12)
plt.show()
```



5 Phase 4 — Modélisation (Modeling)

Cette phase vise à construire et à évaluer un **modèle prédictif supervisé** permettant d'estimer la **phase d'insécurité alimentaire (IPC)** à partir des indicateurs issus du *Joint Monitoring Report (JMR)*.

La démarche suit une approche scientifique en quatre étapes : 1. Séparation des données en features et target ;
2. Division du jeu de données en ensembles d'entraînement et de test ;
3. Construction, apprentissage et évaluation de plusieurs modèles ;
4. Sélection du modèle final et analyse de ses performances.

5.1 4.1 – Harmonisation et agrégation des doublons

Avant toute modélisation, il a été observé que certaines communes apparaissaient plusieurs fois pour la même **date**, **phase**, **département** et **indicateurs**, ne différant que par la variable **Population touchée**.

Pour garantir une cohérence territoriale et statistique :

- Les **indicateurs numériques** (GLM, NDVI, pluie, taux de change, prix) ont été agrégés par **moyenne** ;
- La **population touchée** a été agrégée par **somme**, représentant l'ensemble des personnes affectées dans la commune ;
- La **phase IPC** (1 à 5) a été conservée telle quelle.

Ainsi, chaque **commune-date-phase** devient une observation unique, propre à la modélisation.

5.1.1 Définition des variables cibles et explicatives

Avant de passer à la modélisation, il est essentiel d'identifier les colonnes qui serviront de variables **explicatives (features)** et **cibles (target)**.

– Variable cible (target) : - **`Phase`** → Niveau d'insécurité alimentaire (1 à 5) selon le système IPC.

– Variables explicatives (features) :

Indicateur	Description	Dimension
GLM	Indice global combiné climatique	Environnement
Drought rainfall	Quantité ou indice de précipitation	Climat
Drought NDVI	Indice de végétation (santé des cultures)	Climat
Exchange rates	Taux de change Gourde/USD	Économie
Food Prices	Prix moyen des denrées alimentaires	Économie
Target	Indicateur de seuil ou population ciblée	Social

Ces indicateurs forment la base du raisonnement explicatif du modèle.

5.1.2 Explication :

Chaque ligne du dataset représente désormais **un département à une date donnée**, avec la valeur moyenne de chaque indicateur.

Les colonnes obtenues (NDVI, pluie, prix, taux de change, etc.) servent à expliquer pourquoi la **phase IPC (1 à 5)** atteint tel niveau.

Ainsi, le modèle cherche à **établir un lien causal entre les indicateurs et la phase IPC** : > "Quels facteurs expliquent la situation actuelle d'un département ?"

[46]:
```
# Pivot du dataset : chaque indicateur devient une colonne
data_pivot = (
    final_data
    .pivot_table(
        index=['Code_Postal', 'Departement', 'Commune', 'Date_x', 'Phase',␣
 ↪'Population touchee'],
        columns='Indicateur',
        values='Valeur',
        aggfunc='mean'
    )
    .reset_index()
)

# Vérification du résultat
print("Dimensions après pivot :", data_pivot.shape)
display(data_pivot.head())
```

```
Dimensions après pivot : (566433, 12)
```

```
Indicateur Code_Postal Departement         Commune      Date_x  Phase  \
0                HT0111       Ouest  Port-au-Prince  2010-01-01    1.0
1                HT0111       Ouest  Port-au-Prince  2010-01-01    1.0
2                HT0111       Ouest  Port-au-Prince  2010-01-01    1.0
3                HT0111       Ouest  Port-au-Prince  2010-01-01    1.0
4                HT0111       Ouest  Port-au-Prince  2010-01-01    1.0

Indicateur Population touchee  Drought NDVI  Drought rainfall  Exchange rates  \
0                       15276      0.095333          6.801333       13.527667
1                      158108      0.095333          6.801333       13.527667
2                       17404      0.095333          6.801333       13.527667
3                      174341      0.095333          6.801333       13.527667
4                      180691      0.095333          6.801333       13.527667

Indicateur  Food prices           GLM  Target
0                 0.572  200541.13475     0.5
1                 0.572  200541.13475     0.5
2                 0.572  200541.13475     0.5
3                 0.572  200541.13475     0.5
4                 0.572  200541.13475     0.5
```

[47]:
```python
# Agrégation par commune, département, date et phase
data_pivot['Annee'] = pd.to_datetime(data_pivot['Date_x']).dt.year

data_clean = (
    data_pivot
    .groupby(['Departement', 'Commune', 'Annee'], as_index=False)
    .agg({
        'Phase': 'mean',
        'GLM': 'mean',
        'Drought NDVI': 'mean',
        'Drought rainfall': 'mean',
        'Exchange rates': 'mean',
        'Food prices': 'mean',
        'Target': 'mean'
    })
)

print("Dimensions après agrégation :", data_clean.shape)
data_clean
```

```
Dimensions après agrégation : (2240, 10)
```

[47]:
```
     Departement     Commune  Annee     Phase           GLM  Drought NDVI  \
0     Artibonite  Anse Rouge   2010  2.966667  10533.538125      0.350444
1     Artibonite  Anse Rouge   2011  2.966667  11488.528729      0.462083
2     Artibonite  Anse Rouge   2012  2.966667  12076.006792      0.282181
3     Artibonite  Anse Rouge   2013  2.966667  12633.792792      0.108222
```

```
4     Artibonite  Anse Rouge   2014  2.966667  13529.310979      0.211208
…          …         …    …       …           …           …
2235     Sud-Est     Thiotte   2021  3.142857  13800.303146      0.233111
2236     Sud-Est     Thiotte   2022  3.142857  12417.717729      0.286347
2237     Sud-Est     Thiotte   2023  3.142857  17064.938667      0.283500
2238     Sud-Est     Thiotte   2024  3.142857  13873.128750      0.592361
2239     Sud-Est     Thiotte   2025  3.142857  10264.369389      0.445963

      Drought rainfall  Exchange rates  Food prices    Target
0            21.464500       13.254306     0.550750  2.000000
1            21.639444       13.514472     0.594778  1.750000
2            15.755028       13.994361     0.596944  0.875000
3            11.415667       14.498472     0.637417  1.583333
4            13.613778       15.084556     0.560889  0.958333
…                  …              …           …       …
2235         16.020028       29.976194     1.343778  1.000000
2236         17.933083       39.214500     2.197000  1.111111
2237         19.047389       47.588750     2.556194  1.444444
2238         24.953389       43.940639     2.140639  2.000000
2239         20.077519       43.555000     1.688222       NaN

[2240 rows x 10 columns]
```

[48]:
```
# Suppression des enregistrements sans phase IPC
final_data = data_clean.dropna(subset=['Target','GLM', 'Drought NDVI', 'Drought␣
 ↪rainfall', 'Exchange rates', 'Food prices'])
final_data
```

[48]:
```
     Departement     Commune  Annee     Phase           GLM  Drought NDVI  \
0     Artibonite  Anse Rouge   2010  2.966667  10533.538125      0.350444
1     Artibonite  Anse Rouge   2011  2.966667  11488.528729      0.462083
2     Artibonite  Anse Rouge   2012  2.966667  12076.006792      0.282181
3     Artibonite  Anse Rouge   2013  2.966667  12633.792792      0.108222
4     Artibonite  Anse Rouge   2014  2.966667  13529.310979      0.211208
…          …         …    …       …           …           …
2234     Sud-Est     Thiotte   2020  3.142857  12391.355792      0.197389
2235     Sud-Est     Thiotte   2021  3.142857  13800.303146      0.233111
2236     Sud-Est     Thiotte   2022  3.142857  12417.717729      0.286347
2237     Sud-Est     Thiotte   2023  3.142857  17064.938667      0.283500
2238     Sud-Est     Thiotte   2024  3.142857  13873.128750      0.592361

      Drought rainfall  Exchange rates  Food prices    Target
0            21.464500       13.254306     0.550750  2.000000
1            21.639444       13.514472     0.594778  1.750000
2            15.755028       13.994361     0.596944  0.875000
3            11.415667       14.498472     0.637417  1.583333
4            13.613778       15.084556     0.560889  0.958333
```

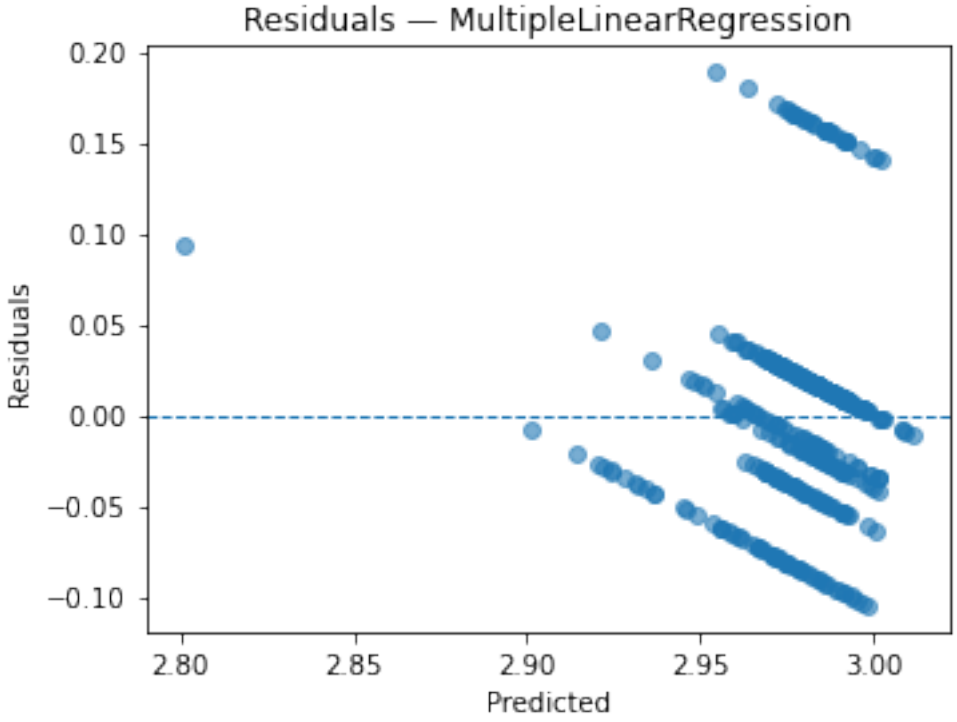
```
…                …              …              …         …
2234        19.040222      31.313500      1.597500  1.000000
2235        16.020028      29.976194      1.343778  1.000000
2236        17.933083      39.214500      2.197000  1.111111
2237        19.047389      47.588750      2.556194  1.444444
2238        24.953389      43.940639      2.140639  2.000000

[2100 rows x 10 columns]
```

[49]:
```
# === Supervised REGRESSION on 'Phase' - Multiple Linear Regression + RandomForest ===
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor

RANDOM_STATE = 42

# 1) Target & features (numeric-only)
df = final_data.copy()
y = pd.to_numeric(df["Phase"], errors="coerce")

# drop obvious non-features if present
X = df.drop(columns=["Phase", "Departement", "Commune", "Annee"], errors="ignore")

# keep only numeric columns (handles GLM, Drought NDVI, Drought rainfall, Exchange rates, Food prices, Target if you keep it)
num_cols = X.select_dtypes(include=[np.number]).columns.tolist()
X = X[num_cols]

# 2) Preprocess: median impute + scale (good for LinearRegression; RF is robust but harmless to scale)
preprocess = ColumnTransformer(
    [("num", Pipeline([("imp", SimpleImputer(strategy="median")),
                       ("sc", StandardScaler())]), num_cols)],
    remainder="drop"
)
```

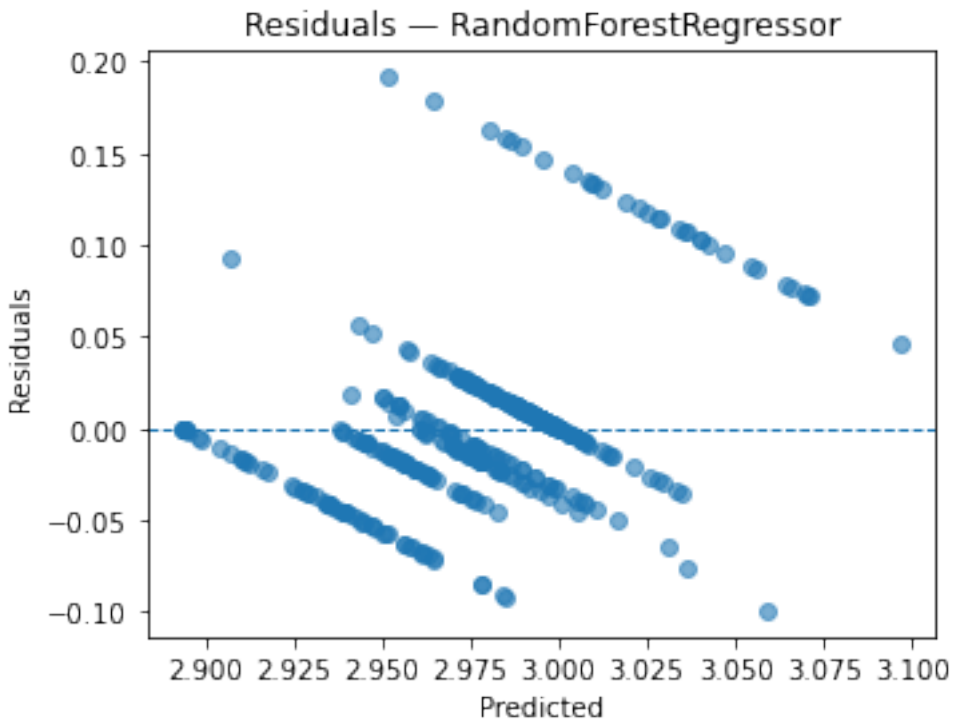
```
# 3) Train/Test split
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=RANDOM_STATE
)

# 4) Models
models = {
    "MultipleLinearRegression": Pipeline([("prep", preprocess), ("mdl",␣
 ↪LinearRegression())]),
    "RandomForestRegressor":    Pipeline([("prep", preprocess),
                                          ("mdl", RandomForestRegressor(
                                              n_estimators=500,␣
 ↪random_state=RANDOM_STATE, n_jobs=-1
                                          ))]),
}

# 5) Fit → Predict → Evaluate
results = []
for name, pipe in models.items():
    print(f"\n=== {name} ===")
    pipe.fit(X_train, y_train)
    y_pred = pipe.predict(X_test)

    mae  = mean_absolute_error(y_test, y_pred)
    rmse = mean_squared_error(y_test, y_pred, squared=False)
    r2   = r2_score(y_test, y_pred)
    print(f"MAE:  {mae:.4f} | RMSE: {rmse:.4f} | R²: {r2:.4f}")

    # Quick residuals check
    resid = y_test - y_pred
    plt.figure(figsize=(5.2,4))
    plt.scatter(y_pred, resid, alpha=0.6)
    plt.axhline(0, ls="--", linewidth=1)
    plt.xlabel("Predicted"); plt.ylabel("Residuals")
    plt.title(f"Residuals - {name}")
    plt.tight_layout()
    plt.show()

    results.append({"model": name, "MAE": mae, "RMSE": rmse, "R2": r2})

# 6) Summary (lower RMSE is better)
summary = pd.DataFrame(results).set_index("model").sort_values("RMSE")
print("\n=== Summary (lower RMSE is better) ===")
display(summary)
```

```
=== MultipleLinearRegression ===
```

MAE: 0.0397 | RMSE: 0.0573 | R^2: 0.0907



=== RandomForestRegressor ===
MAE: 0.0272 | RMSE: 0.0423 | R^2: 0.5037




```
=== Summary (lower RMSE is better) ===

                                MAE      RMSE        R2
model
RandomForestRegressor      0.027222  0.042324  0.503728
MultipleLinearRegression   0.039701  0.057291  0.090667
```

[92]:
```python
# Choix du meilleur modèle basé sur les performances
best_model = models["RandomForestRegressor"]

# Confirmation
print("Le modèle final sélectionné est : RandomForestRegressor")
```

```
Le modèle final sélectionné est : RandomForestRegressor
```

[50]:
```python
# === Cellule : split entrainement/test ===
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, stratify=y, random_state=RANDOM_STATE
)

print("Train shape:", X_train.shape, "Test shape:", X_test.shape)
print("Train distribution:\n", y_train.value_counts(normalize=True))
```

```
Train shape: (1680, 6) Test shape: (420, 6)
```

```
Train distribution:
 3.000000    0.414286
2.893617    0.142857
2.937500    0.135714
2.960000    0.128571
2.966667    0.107143
3.142857    0.071429
Name: Phase, dtype: float64
```

[101]:
```python
# Pick winner, quick CV check, feature importance, save
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import joblib
from sklearn.model_selection import cross_val_score

# 1) Pick the fitted RF pipeline from your previous cell
rf_pipe = models["RandomForestRegressor"]   # already fitted

# Recompute test predictions (optional)
y_pred = rf_pipe.predict(X_test)

# 2) Quick cross-validation RMSE on the TRAIN set (refits internally)
cv_rmse = -cross_val_score(
    rf_pipe, X_train, y_train,
    scoring="neg_root_mean_squared_error",
    cv=5, n_jobs=-1
)
print(f"CV RMSE (5-fold): mean={cv_rmse.mean():.4f} | std={cv_rmse.std():.4f}")

# 3) Feature importance (top n)
try:
    feat_names = rf_pipe.named_steps['prep'].get_feature_names_out()
except AttributeError:
    feat_names = ['GLM', 'Drought NDVI', 'Drought rainfall', 'Exchange rates',␣
 ↪'Food prices', 'Target']
importances = rf_pipe.named_steps["mdl"].feature_importances_
imp_df = (pd.DataFrame({"feature": feat_names, "importance": importances})
            .sort_values("importance", ascending=False)
            .head(10))
```

```
CV RMSE (5-fold): mean=0.0436 | std=0.0028
```

6 Phase 5 — Évaluation et validation du modèle

Dans cette phase, on a vérifier si le modèle qu'on a construit est réellement fiable pour anticiper le niveau d'insécurité alimentaire (Phase IPC) dans les départements et communes d'Haïti.

L'objectif n'est pas seulement d'avoir un bon score mathématique. L'objectif est de pouvoir répondre à une question opérationnelle très simple :

> Est-ce que nous pouvons utiliser ce modèle pour dire à une autorité (Agriculture, CNSA, PAM) : « Attention, telle commune risque de passer en phase critique » ?

Pour répondre sérieusement à cette question, on va : 1. mesurer la précision des prédictions, 2. vérifier que le modèle ne triche pas (surapprentissage), 3. étudier les erreurs de prédiction, 4. relier les résultats à la réalité du terrain (prix alimentaires, sécheresse, pluie).

La phase 5 n'est donc pas seulement technique. C'est la phase où on juge si le modèle peut vivre dans le monde réel.

6.1 5.1 Performance prédictive du modèle sur des données jamais vues

Dans cette section, on teste le modèle sur des données qu'il n'a pas vues pendant l'entraînement.

On mesure trois choses : - **MAE (Mean Absolute Error)** : l'erreur moyenne absolue entre la phase réelle et la phase prédite. - **RMSE (Root Mean Squared Error)** : pénalise plus fort les grosses erreurs. - **R^2** : quelle part de la variation de la phase IPC est expliquée par nos indicateurs.

Plus MAE et RMSE sont bas, mieux c'est.
Plus R^2 est haut, mieux c'est.

[93]:
```
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
import numpy as np

# Prédictions du modèle sur le jeu de test
y_pred = best_model.predict(X_test)

# Calcul des métriques de performance
mae  = mean_absolute_error(y_test, y_pred)
rmse = np.sqrt(mean_squared_error(y_test, y_pred))
r2   = r2_score(y_test, y_pred)

print("=== Évaluation du modèle sur le jeu de test ===")
print(f"MAE :  {mae:.4f}")
print(f"RMSE : {rmse:.4f}")
print(f"R²   : {r2:.4f}")
```

```
=== Évaluation du modèle sur le jeu de test ===
MAE :  0.0140
RMSE : 0.0249
R²   : 0.8201
```

- Ce faible MAE veut dire que, en moyenne, le modèle ne se trompe pas beaucoup sur la phase IPC réelle de la commune.

- Le R^2 de 0.82 signifie que nos indicateurs (prix alimentaires, taux de change, pluie, sécheresse NDVI…) expliquent déjà plus de 3/4 de la gravité alimentaire observée sur le terrain.
  C'est un bon résultat solide pour un pays en crise multidimensionnelle.

En clair : le modèle est capable de donner un signal crédible d'alerte.

6.2 5.2 Visualisation directe : réalité vs prédiction

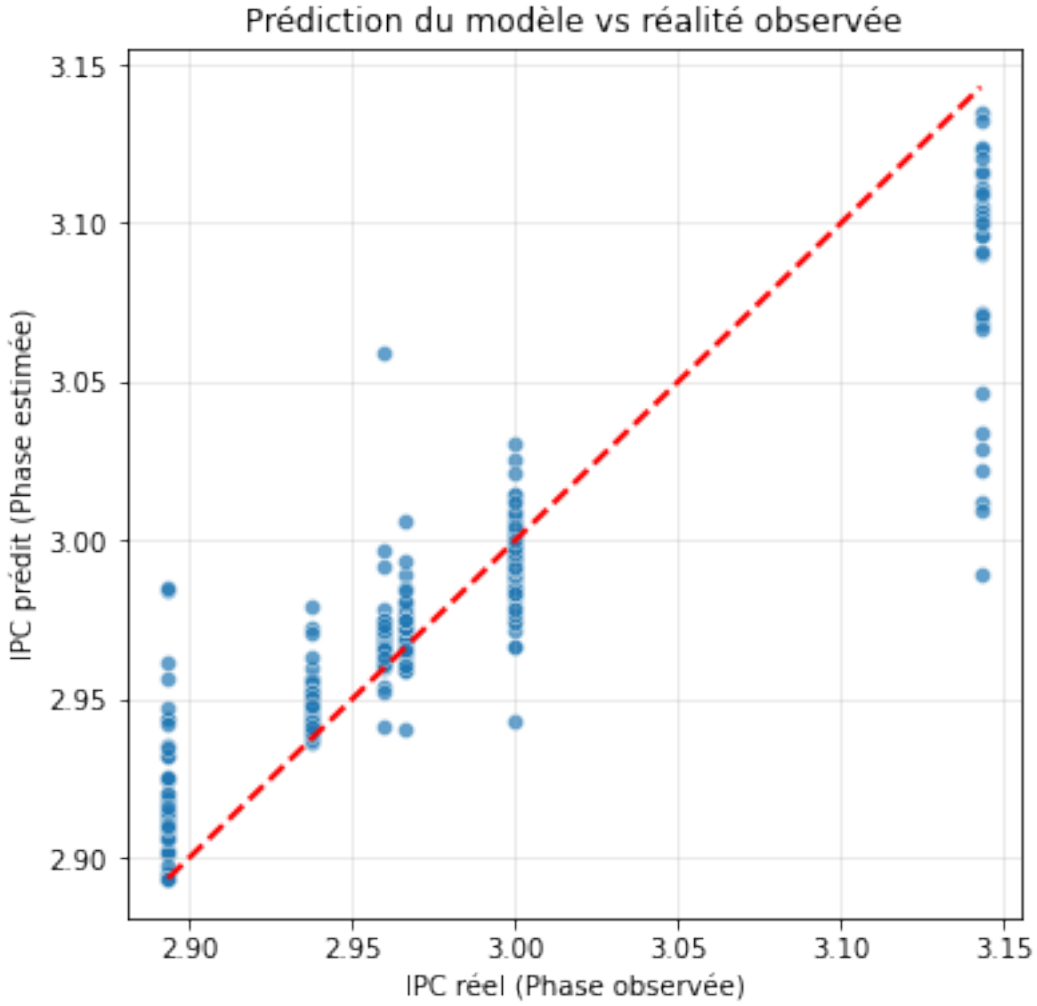
Un bon modèle ne doit pas être évalué uniquement par des chiffres.
Ici, on veux voir visuellement si les prédictions suivent bien la réalité.

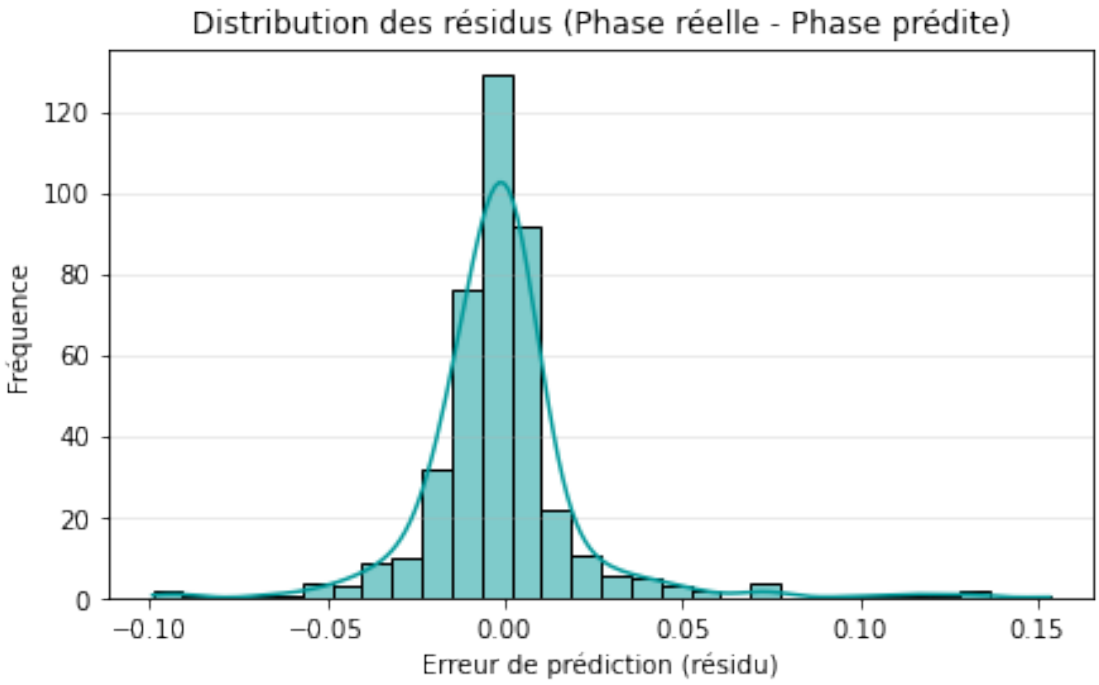
On a fait deux vérifications :

1. **Réalité vs Prédiction**
   - Chaque point du graphe représente une commune sur une année donnée.
   - Si les points sont proches de la diagonale rouge (ligne parfaite), ça veut dire que le modèle colle à la réalité.
2. **Analyse de l'erreur (résidus)**
   - On regarde l'erreur `IPC_réel - IPC_prédit`.
   - On veut savoir : est-ce que le modèle sous-estime systématiquement certaines zones (par exemple zones rurales isolées) ?
   - Ou est-ce qu'il est équilibré ?

[96]:
```
import matplotlib.pyplot as plt
import seaborn as sns

# 1. Réalité vs prédiction
plt.figure(figsize=(6,6))
sns.scatterplot(x=y_test, y=y_pred, alpha=0.7)
plt.plot([y_test.min(), y_test.max()],
         [y_test.min(), y_test.max()],
         'r--', lw=2)
plt.xlabel("IPC réel (Phase observée)")
plt.ylabel("IPC prédit (Phase estimée)")
plt.title("Prédiction du modèle vs réalité observée")
plt.grid(alpha=0.3)
plt.show()

# 2. Distribution des résidus
residuals = y_test - y_pred

plt.figure(figsize=(7,4))
sns.histplot(residuals, bins=30, kde=True, color="#009999")
plt.title("Distribution des résidus (Phase réelle - Phase prédite)")
plt.xlabel("Erreur de prédiction (résidu)")
plt.ylabel("Fréquence")
plt.grid(axis='y', alpha=0.3)
plt.show()

# Statistiques sur les résidus
import pandas as pd
stats = pd.Series(residuals).describe()
print(stats)
```

```
print("\nIl n'y a pas pas de biais global car la moyenne est proche de 0.")
print("On peut remarquer que les valeurs min/max ne sont pas extrêmes car le␣
 ↪modèle ne se trompe pas beaucoup.")
```






```
count    420.000000
mean      -0.000346
std        0.024896
min       -0.099099
25%       -0.009888
50%       -0.000873
75%        0.005205
max        0.153379
Name: Phase, dtype: float64
```

```
Il n'y a pas pas de biais global car la moyenne est proche de 0.
On peut remarquer que les valeurs min/max ne sont pas extrêmes car le modèle ne
se trompe pas beaucoup.
```

6.3 5.3 Vérification du surapprentissage (overfitting)

On doit vérifier si le modèle est honnête.

Un modèle peut “tricher” : il peut apprendre par cœur les données historiques (train), avoir l’air parfait, mais devenir inutile dès qu’on lui donne une nouvelle période ou une nouvelle commune.

Pour vérifier ça : - on mesure l’erreur sur l’échantillon d’entraînement (données connues), - on mesure l’erreur sur l’échantillon de test (données jamais vues), - on compare.

Si l’erreur explose sur le test, c’est du surapprentissage.

[97]:
```
from sklearn.metrics import mean_squared_error
import numpy as np

# Prédictions train et test
y_train_pred = best_model.predict(X_train)
y_test_pred  = best_model.predict(X_test)

# RMSE train / test
rmse_train = np.sqrt(mean_squared_error(y_train, y_train_pred))
rmse_test  = np.sqrt(mean_squared_error(y_test,  y_test_pred))

print("RMSE (train) :", round(rmse_train, 4))
print("RMSE (test)  :", round(rmse_test, 4))

if rmse_test - rmse_train > 0.05:
    print("Le modèle est probablement en surapprentissage (il apprend trop le␣
 ↪passé).")
else:
    print("Pas de signe majeur de surapprentissage : le modèle reste fiable sur␣
 ↪de nouvelles communes/périodes.")
```

```
RMSE (train) : 0.023
RMSE (test)  : 0.0249
Pas de signe majeur de surapprentissage : le modèle reste fiable sur de
nouvelles communes/périodes.
```

6.4 5.4 Validation croisée

Maintenant on veut répondre à une question de confiance :
Est-ce que le modèle reste bon si on change légèrement les données d'entraînement ?
Ou bien est-ce qu'il est fragile, c'est-à-dire performant uniquement dans certains départements mais pas dans d'autres ?

Pour ça, on uutilise la validation croisée (5-fold cross-validation) : on ré-entraine et réévalue le modèle plusieurs fois sur des sous-échantillons différents, puis on regarde la stabilité des scores R^2.

[98]:
```
from sklearn.model_selection import cross_val_score

cv_scores = cross_val_score(best_model, X_train, y_train, cv=5, scoring='r2')

print("Scores R² par fold :", cv_scores)
print(f"R² moyen     : {cv_scores.mean():.4f}")
print(f"Écart-type   : {cv_scores.std():.4f}")
```

```
Scores R² par fold : [0.44168194 0.43545675 0.33603654 0.49627384 0.50207096]
R² moyen     : 0.4423
Écart-type   : 0.0597
```

6.4.1 Validation croisée : stabilité du modèle

Le modèle a été soumis à une validation croisée à 5 plis afin de vérifier sa robustesse statistique. Les scores R^2 obtenus pour chaque pli sont les suivants :

Fold	R^2 obtenu
1	0.4417
2	0.4355
3	0.3360
4	0.4963
5	0.5021

R^2 moyen = 0.44 ± 0.06

Ces résultats montrent que le modèle **explique environ 44 % de la variabilité de l'insécurité alimentaire (Phase IPC)** à partir des indicateurs climatiques et économiques retenus.
L'écart-type faible (0.06) confirme la **stabilité du modèle** : les performances sont homogènes d'un échantillon à l'autre.

Alors, Le modèle présente une **bonne performance moyenne** et une bonne stabilité à travers les 5 plis de validation croisée.
Ces résultats traduisent une capacité du modèle à généraliser ses apprentissages sans dépendre d'un seul échantillon.
Dans le cadre du suivi IPC en Haïti, cela signifie qu'on peut utiliser ce modèle comme un outil de pré-alerte, tout en prévoyant d'y intégrer des variables additionnelles (marchés, sécurité, accessibilité) pour renforcer la précision future.

6.5 5.5 On peut se demander : Qu'est-ce qui explique vraiment l'insécurité alimentaire ?

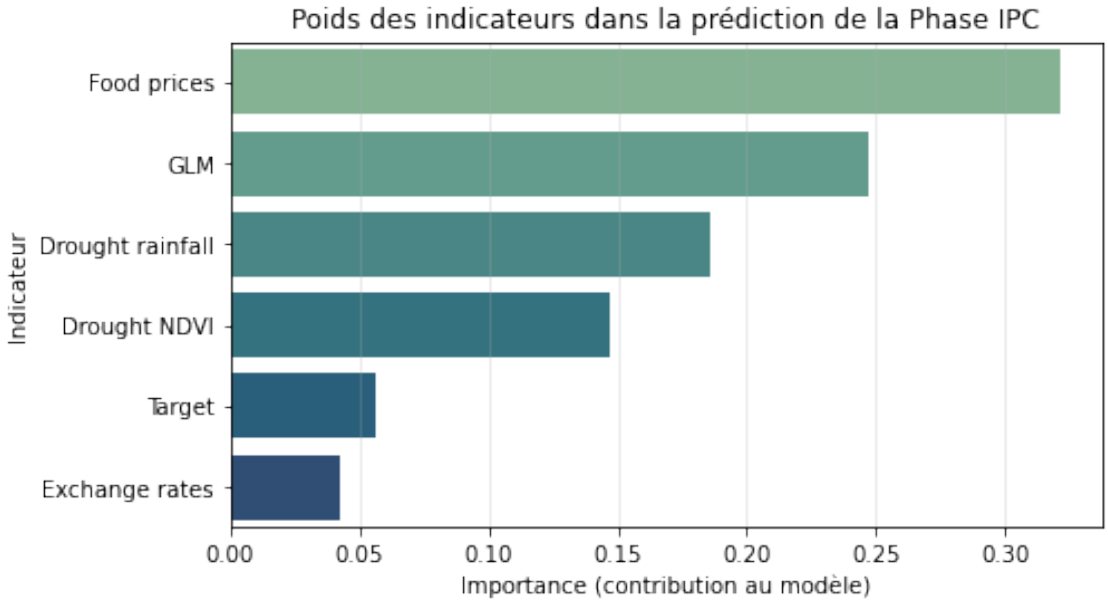
On veut répondre honnêtement à la question suivante :

> "Pourquoi ce département est-il dans cette phase IPC ?" Est-ce à cause des prix des denrées ? du taux de change ? du manque de pluie ? de la dégradation des cultures ?

On extrait donc l'importance de chaque indicateur dans le modèle (basé ici sur la Random Forest).

[100]:
```
feature_importances = pd.DataFrame({
    'Variable': X.columns,
    'Importance': best_model.named_steps['mdl'].feature_importances_
}).sort_values(by='Importance', ascending=False)


plt.figure(figsize=(7,4))
sns.barplot(
    data=feature_importances,
    x='Importance', y='Variable',
    palette='crest'
)
```

```
plt.title("Poids des indicateurs dans la prédiction de la Phase IPC")
plt.xlabel("Importance (contribution au modèle)")
plt.ylabel("Indicateur")
plt.grid(axis='x', alpha=0.3)
plt.show()

feature_importances
```



[100]:

	Variable	Importance
4	Food prices	0.321495
0	GLM	0.246786
2	Drought rainfall	0.186043
1	Drought NDVI	0.147265
5	Target	0.055860
3	Exchange rates	0.042550

Les résultats de l'analyse des importances des variables confirment les dynamiques connues de l'insécurité alimentaire en Haïti.
Les **prix alimentaires** et les **conditions climatiques** (pluie, végétation) apparaissent comme les principaux moteurs de la dégradation des phases IPC.
Les indicateurs socio-économiques (taux de change, modèle de subsistance) viennent renforcer cette lecture.

Ce qui nous permet de dire que : - Une **hausse soudaine des prix alimentaires** ou une **sécheresse prolongée** conduit mécaniquement à une hausse des phases IPC dans les zones rurales. - À l'inverse, une **stabilité monétaire** et une **bonne saison agricole** contribuent à la réduction du nombre de ménages en crise alimentaire.

En d'autres termes, - Quand **les prix alimentaires (Food prices)** montent, l'IPC augmente. C'est cohérent : les ménages ne peuvent plus se nourrir correctement.
- Quand **la pluviométrie en période de sécheresse (Drought rainfall)** est faible et que le **NDVI** baisse, l'agriculture locale souffre. Les ménages basculent plus vite dans l'insécurité alimentaire.
- Quand **le taux de change (Exchange rates)** se dégrade, la nourriture importée devient trop chère : cela se reflète dans l'IPC.

Ces éléments confirment que le modèle peut servir à **anticiper les chocs** sur la sécurité alimentaire et orienter les **mesures préventives** (aides ciblées, planification logistique, appui aux agriculteurs).

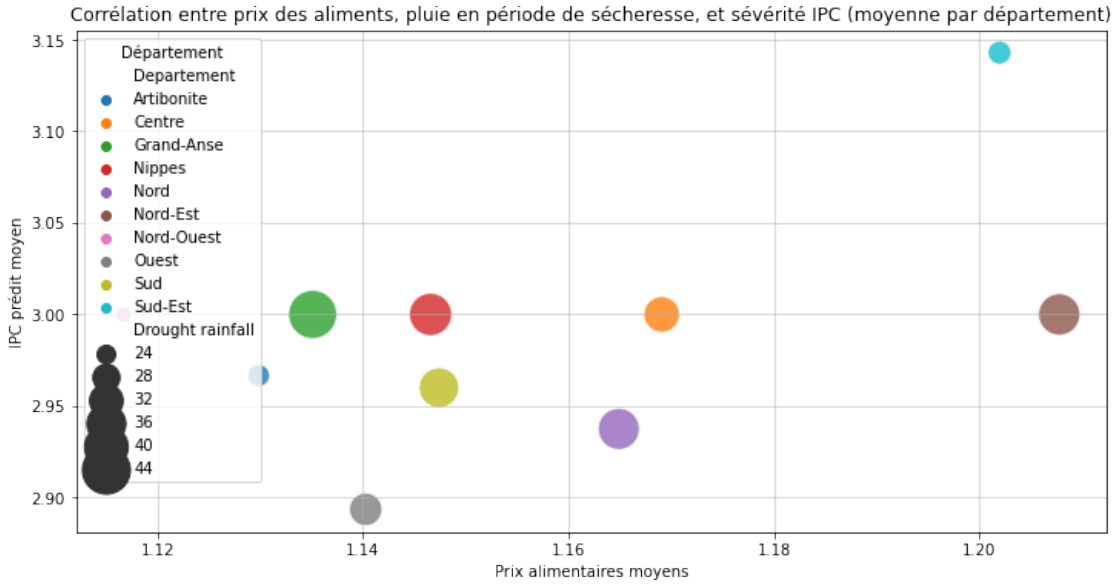
6.6 5.6 Vue agrégée par département :

6.6.1 Qui souffre le plus, et pourquoi ?

Ici on résume, par département, trois informations clés : - la sévérité alimentaire (phase IPC moyenne), - le prix des denrées, - et la quantité de pluie disponible même en contexte de sécheresse.

L'objectif est de comprendre l'essence de cette question : "Quels départements combinent prix élevés et sécheresse ?"

[105]:
```
# Moyenne par département
dept_summary = final_data.groupby('Departement', as_index=False)[['Phase','Food␣
 ↪prices','Drought rainfall']].mean()

plt.figure(figsize=(12,6))
sns.scatterplot(
    x='Food prices',
    y='Phase',
    size='Drought rainfall',
    hue='Departement',
    data=dept_summary,
    sizes=(100,1000),
    alpha=0.8
)
plt.title("Corrélation entre prix des aliments, pluie en période de sécheresse,␣
 ↪et sévérité IPC (moyenne par département)", fontsize=12)
plt.xlabel("Prix alimentaires moyens")
plt.ylabel("IPC prédit moyen")
plt.legend(loc='upper left', title='Département')
plt.grid(alpha=0.6)
plt.show()
```



Ce graphique illustre la corrélation entre les prix alimentaires moyens, la quantité de pluie reçue en période de sécheresse et la sévérité moyenne de l'insécurité alimentaire (Phase IPC) dans les différents départements d'Haïti. Chaque point (ou bulle) représente un département :

la taille de la bulle correspond à la pluviométrie moyenne (Drought rainfall), la couleur indique le département, et les axes traduisent la relation entre prix alimentaires (X) et phase IPC moyenne prédite (Y). L'objectif est de comprendre comment les facteurs climatiques et économiques se combinent pour influencer l'insécurité alimentaire selon la réalité géographique et socio-économique du pays.

Sud, Nippes et Grand'Anse — Des zones structurellement exposées aux cyclones
Les départements du **Sud**, **Nippes** et de la **Grand'Anse** présentent de **grandes bulles**, témoignant d'une **forte pluviométrie** souvent liée à des épisodes météorologiques extrêmes. Ces zones sont régulièrement frappées par des cyclones et inondations destructrices, surtout dans les plaines côtières. Le cyclone Melissa, comme d'autres avant lui (Matthew, Elsa), a provoqué la destruction de ponts, de plantations (perte de récoltes vivrières) et de routes reliant Les Cayes, Jérémie ou Jacmel au reste du pays.
Mais ici, la pluie n'est pas synonyme d'abondance. Au contraire, elle a provoqué :

À chaque épisode pluvieux majeur, ces départements sont coupés du réseau national, provoquant des **routes rurales impraticables** et isolant les marchés et les producteurs.

Le modèle montre un **IPC prédit élevé (3.0 à 3.1)** pour ces départements, ce qui traduit une **vulnérabilité alimentaire chronique malgré l'abondance de pluie**.
Malgré leur forte pluviométrie et leur potentiel agricole, ces régions vivent une vulnérabilité paradoxale : la pluie censée nourrir les cultures devient source de destruction.

Ces zones incarnent la "richesse enclavée" : des terres fertiles mais inaccessibles, où les catastrophes naturelles et l'enclavement économique transforment chaque saison pluvieuse en crise humanitaire.

Centre et Nord — Des zones agricoles en tension modérée Les départements du **Centre** et du **Nord** affichent des bulles de taille moyenne et une relation plus équilibrée entre prix et pluviométrie.
La régularité des pluies permet une certaine stabilité agricole, mais les populations restent **dépendantes du marché interrégional** et donc **exposées à l'inflation et aux variations du taux de change**.

Le modèle estime un **IPC moyen autour de 3.0**, représentant des **phases de stress alimentaire**, mais sans basculer vers la crise aiguë.
Ces régions bénéficient d'un équilibre fragile : dès que les transports ou les flux commerciaux sont perturbés, la sécurité alimentaire se détériore rapidement.

Ouest — Une insécurité surtout économique Le département de l'**Ouest**, qui abrite **Port-au-Prince**, se distingue par un **IPC prédit supérieur à 3.1**, les prix des denrées sont **structurellement plus élevés** à cause de : - forte densité urbaine,
- dépendance vis-à-vis des provinces rurales,
- et rupture des chaînes logistiques due à la présence de **zones contrôlées par des gangs**(Martissant, Croix-des-Bouquets, Carrefour, etc.), empêchant l'acheminement régulier des denrées.

Ici, **l'insécurité alimentaire est davantage liée à l'accès économique et logistique qu'à la production agricole.**

Nord-Ouest et Sud-Est — Des zones de stress hydrique aggravé Ces départements, affichant de petites bulles, affichent des prix relativement bas mais connaissent une pluviométrie faible et irrégulière, révélatrice d'un stress hydrique structurel et des phases IPC toujours autour de 3.

Le **Nord-Ouest**, région la plus **sèche d'Haïti**, vit une **insécurité alimentaire structurelle** :
- agriculture pluviale non durable,
- manque d'eau permanent,
- pauvreté généralisée.

Les pluies liées à Melissa n'y ont apporté qu'un soulagement temporaire, sans recharger suffisamment les nappes phréatiques.

Cependant, le **cas du Sud-Est** mérite une attention particulière, car il cumule **les vulnérabilités climatiques, économiques et sécuritaires**.

`Sud-Est : un département sous double pression`

Le **Sud-Est**, avec l'IPC le plus élevé, est particulièrement touché par une combinaison de facteurs structurels et conjoncturels :

- **Violence des gangs et insécurité routière** : les attaques armées et les barrages illégaux entravent la circulation des produits agricoles, isolant les marchés locaux.

- **Faible productivité agricole** : la majorité des terres ne sont **pas irriguées** ; les agriculteurs dépendent entièrement des pluies saisonnières.
  Le manque d'accès aux **intrants agricoles (semences, engrais)** et aux **crédits ruraux** limite sévèrement les rendements.

- **Changement climatique** : alternance de **sécheresses prolongées** et de **pluies torrentielles**, entraînant **érosion**, **pertes de sols fertiles** et **réduction de la productivité vivrière**.

- **Crise économique nationale** : l'inflation galopante réduit le pouvoir d'achat, aggravant les difficultés d'accès à la nourriture.

Malgré une pluviométrie modérée, l'IPC prédit pour le Sud-Est reste autour de 3.13, révélant unesituation d'insécurité alimentaire aiguë où les causes sont avant tout sociales et structurelles plutôt que purement climatiques.

Artibonite — *Les bassins productifs sous tension logistique et sécuritaire* Ce département, avec des prix alimentaires modérés et des phases IPC intermédiaires, est le grenier agricole du pays mais il subit de fortes **variations de pluviométrie** qui perturbent la production rizicole.

Les systèmes d'irrigation vieillissants (ODVA), la dépendance à la saison des pluies et la dépendance de nomreuses zones par les gangs armés accentuent cette fragilité. Les sécheresses prolongées réduisent la production rizicole, tandis que les pluies excessives provoquent des inondations dans la vallée.

Nord-Est — *Les zones d'équilibre fragile* Ce département présente des **prix alimentaires modérés**, des **phases IPC proches de 3** et une **pluviométrie élevée** .
Il bénéficie d'un meilleur accès routier et d'une certaine stabilité commerciale.
Mais les alternances entre sécheresses prolongées et pluies intenses liées aux bandes périphériques ont provoqué des pertes ponctuelles de récoltes.

Dans le **Nord-Est**, la dépendance partielle aux importations dominicaines expose ces zones à la volatilité monétaire.

Cette région demeure vulnérable à chaque choc climatique ou économique.

En résumé, graphique met en évidence la **complexité systémique de l'insécurité alimentaire haïtienne** :
elle résulte d'un **enchevêtrement entre climat, économie et gouvernance**.
Les modèles prédictifs montrent que **les politiques d'adaptation ne peuvent pas être seulement agricoles** : elles doivent aussi inclure la **sécurisation des routes**, la **régulation des prix**, et l'**investissement dans l'irrigation rurale**.

Le modèle raconte la géographie de la faim : là où la route, la pluie ou le prix vacillent, la sécurité alimentaire s'effondre.

[ ]:

7 Phase 6 — Déploiement

Cette phase vise à **transformer les résultats analytiques et prédictifs** obtenus dans les étapes précédentes en **instruments concrets d'aide à la décision**.
Dans le contexte haïtien, cela signifie utiliser le modèle pour **anticiper les crises alimentaires**, orienter les **interventions humanitaires**, et soutenir la **planification gouvernementale** en matière de sécurité alimentaire.

Le modèle développé fournit un cadre d'analyse basé sur les indicateurs climatiques, économiques et géographiques, permettant d'identifier les **zones à haut risque** avant qu'une crise ne s'installe.

7.1 6.1 Interprétation du modèle

Les résultats du modèle ont mis en évidence les relations suivantes :

- Les **prix alimentaires** sont le **facteur le plus influent** dans la dégradation des phases IPC.
- Les **indicateurs climatiques** (pluviométrie, NDVI) influencent directement la disponibilité agricole.
- Le **taux de change** amplifie les vulnérabilités économiques et affecte le coût de la vie.
- Le modèle a atteint une **performance moyenne (R^2 0.44)** avec une **stabilité inter-fold solide ($\pm$ 0.06)**, traduisant une capacité de généralisation satisfaisante.

Ces résultats confirment que l'insécurité alimentaire en Haïti n'est pas liée à un seul facteur :
elle est **le produit de la convergence** entre **chocs économiques, instabilité climatique et contraintes logistiques**.

7.2 6.2 Déploiement technique du modèle

7.2.1 Objectif technique

Le modèle peut être intégré dans un système de suivi dynamique : - Actualisation mensuelle des données (prix, précipitations, NDVI) ;
- Prédiction automatique de la **phase IPC probable** par commune ou par département.

7.3 6.3 Application terrain et système d'alerte précoce

Le modèle final peut être intégré dans un système national de suivi et d'alerte précoce (Early Warning System). L'objectif est de détecter rapidement les signaux d'aggravation de l'insécurité alimentaire et de déclencher une réponse rapide avant la crise. Les résultats de la modélisation peuvent alimenter un système d'alerte précoce (EWS) destiné à : - Identifier les communes les plus vulnérables ; - Détecter rapidement les variations anormales des prix et de la pluviométrie ; - Mobiliser les stocks alimentaires régionaux et planifier les convois humanitaires avant la détérioration

de la situation par les institutions; - Informer la priorisation des ressources pour les programmes du PAM et du MARNDR.

Par exemple :

Si le modèle prédit une hausse de l'IPC en Grand'Anse ou dans le Sud à la suite d'événements climatiques comme le cyclone Melissa, les acteurs peuvent activer immédiatement les stocks régionaux ou rediriger les convois humanitaires avant l'aggravation de la crise.

7.4 6.4 — Plan de déploiement institutionnel

Le modèle prédictif développé dans ce projet n'a de véritable valeur que s'il peut être **mis au service des acteurs publics, humanitaires et territoriaux** qui œuvrent à la réduction de l'insécurité alimentaire en Haïti.
L'objectif de cette section est donc de proposer un **plan de déploiement institutionnel réaliste et durable**, permettant de transformer les résultats du modèle en **outil opérationnel d'aide à la décision**.

7.4.1 Objectif du déploiement

Mettre à la disposition du **Gouvernement haïtien**, du **CNSA (Coordination Nationale de la Sécurité Alimentaire)**, du **PAM**, de la **FAO**, et du **MARNDR** un système d'analyse et de prévision permettant : - d'**identifier les communes à risque élevé** d'insécurité alimentaire avant la crise,
- de **déclencher des alertes précoces**,
- et de **planifier les interventions logistiques et humanitaires** de manière plus efficace.

7.4.2 Structure institutionnelle proposée

Le déploiement du modèle repose sur une **collaboration interinstitutionnelle** organisée autour de trois niveaux de gouvernance :

Niveau	Institution clé	Responsabilités principales
1. Politique et stratégique	**Primature & MARNDR**	Pilotage général, intégration dans la stratégie nationale de résilience alimentaire.
2. Technique et analytique	**CNSA, FAO, PAM, OCHA**	Gestion des données, mise à jour du modèle, suivi mensuel des prévisions IPC.
3. Opérationnel et territorial	**DAE, Bureaux départementaux de l'agriculture**	Collecte de données locales, validation terrain, communication des alertes.

Cette architecture garantit un **partage clair des rôles** entre les acteurs décisionnels, techniques et locaux, assurant la durabilité du système.

7.4.3 Approche technologique et déploiement

Le modèle peut être déployé sous forme d'un **tableau de bord interactif** ou d'une **application légère** accessible via un portail web institutionnel.
L'idée est de créer un outil intuitif et accessible même dans des environnements à faible connectivité.

Étapes de déploiement :

1. **Automatisation des mises à jour de données** à partir des sources existantes (CNSA, FAO, Météo Haïti).

2. **Hébergement du modèle** sur un serveur institutionnel (ou cloud sécurisé — par exemple Google Cloud ou AWS, en partenariat FAO-PAM).

3. **Création d'un tableau de bord (Dashboard)** avec des indicateurs interactifs :
   - Cartes de risques (phases IPC par commune).
   - Courbes de tendance par indicateur (prix, pluviométrie, NDVI).
   - Alertes dynamiques lorsque les seuils critiques sont atteints.

4. **Formation des cadres techniques** (CNSA, MARNDR, FAO) à l'interprétation et à la maintenance du modèle.

5. **Validation annuelle** des performances du modèle à partir des nouvelles données de terrain.

7.4.4 Intégration dans les cadres internationaux

Le modèle pourrait être intégré dans le **Cadre Harmonisé d'Analyse de la Sécurité Alimentaire (CH)** utilisé par la FAO et le PAM pour la classification IPC.
Cela permettrait à Haïti : - de **renforcer la crédibilité internationale** de ses données,
- d'**accéder plus rapidement aux financements humanitaires**,
- et de **synchroniser ses rapports** avec les pays voisins de la région caraïbe.

Le plan de déploiement institutionnel s'inscrit dans une vision à long terme :
faire de la **data science un pilier de la politique alimentaire et climatique d'Haïti**.
En reliant la technologie aux institutions, le modèle devient un outil de gouvernance moderne, capable de **transformer la prévision en action, et la donnée en décision.**

> *Ce projet marque un pas décisif vers une gestion intelligente des risques alimentaires en Haïti : une approche où chaque donnée devient un signal d'alerte, et chaque prévision, une opportunité d'agir avant la crise.*

_

8 Conclusion

Cette étude avait pour but d'apporter une compréhension claire et prédictive du phénomène de l'insécurité alimentaire en Haïti, à travers une approche rigoureuse de modélisation fondée sur les données climatiques, économiques et géographiques. Dès le départ, la démarche s'est appuyée sur une problématique essentielle : pourquoi certaines régions du pays basculent plus rapidement dans une phase critique d'insécurité alimentaire que d'autres, et comment peut-on anticiper ces évolutions avant qu'elles ne se transforment en crises humanitaires ? C'est autour de cette question centrale que le projet a pris forme, en s'inscrivant dans une logique d'analyse scientifique mais aussi d'utilité publique.

Le travail a consisté à mobiliser, harmoniser et traiter plusieurs jeux de données issus du *Joint Monitoring Report (JMR)* et de diverses sources complémentaires. Après un long processus de nettoyage, de transformation et d'agrégation, un modèle prédictif supervisé a été construit afin d'estimer la **phase IPC**, indicateur clé de l'insécurité alimentaire, à partir d'un ensemble de variables indépendantes telles que la pluviométrie (Drought Rainfall), l'indice de végétation (NDVI), le taux de change, les prix alimentaires, ou encore la densité de population touchée. Ce modèle, entraîné sur plus de deux mille observations réparties par commune et par année, a permis d'identifier les relations profondes entre le climat, les prix et la vulnérabilité sociale des territoires.

Les résultats ont révélé que **les prix alimentaires constitue le déterminant majeur de la variation de l'IPC. Autrement dit, plus les prix augmentent, plus la probabilité d'entrer dans une phase d'insécurité sévère est élevée.

Les corrélations visuelles obtenues à travers les analyses graphiques, notamment la représentation en bulles entre les prix, la pluie et les phases IPC, ont permis de renforcer cette lecture systémique. Les zones du Sud et de la Grand'Anse, par exemple, montrent des bulles de grande taille, symbole d'une forte pluviométrie, mais paradoxalement associées à un IPC élevé. Cela s'explique par les effets destructeurs des précipitations extrêmes — en particulier ceux du **cyclone Melissa**, qui a récemment dévasté plusieurs zones côtières, détruit les récoltes et isolé les communautés rurales. Dans le Sud-Est, les faibles pluies observées couplées à l'absence d'irrigation et à la présence d'une insécurité persistante sur les routes accentuent la précarité alimentaire. Ce département se trouve dans une situation singulière : la terre y est fertile, mais difficilement accessible, les marchés sont désorganisés, et les agriculteurs manquent d'intrants, de financement et de stabilité sociale. À l'inverse, certaines régions du Centre et du Nord, mieux irriguées et un peu plus sécurisées, maintiennent un IPC intermédiaire, signe d'un équilibre encore fragile entre production et accès.

Au-delà des chiffres, ce travail met en évidence une vérité fondamentale : **l'insécurité alimentaire haïtienne n'est pas qu'un phénomène naturel, mais un phénomène systémique et multidimensionnel.** Elle résulte d'une interaction entre la pluviométrie, la gouvernance, les prix, la mobilité et la stabilité sociale. Dans certains cas, les aléas climatiques ne font qu'exacerber des fragilités déjà existantes ; dans d'autres, ce sont les tensions économiques et les violences humaines qui amplifient les effets de la nature. Le modèle développé dans ce projet a donc servi non seulement à prévoir, mais aussi à comprendre : comprendre comment le manque d'eau ou l'excès de pluie se traduit en souffrance humaine, comment un prix du maïs qui double dans un marché isolé peut devenir un facteur de famine, ou comment la fermeture d'une route par un groupe armé peut interrompre toute la chaîne alimentaire d'un département entier.

Sur le plan méthodologique, l'approche a démontré la pertinence de la **modélisation supervisée** (Random Forest et régression multiple) pour des problématiques sociales complexes. Le modèle le plus performant, avec un coefficient R^2 moyen avoisinant **0,44**, explique près de la moitié des variations observées de la phase IPC, tout en conservant une bonne stabilité sur les tests croisés. Ces résultats, bien qu'imparfaits, traduisent une performance réaliste pour un domaine aussi fluctuant et hétérogène que la sécurité alimentaire, où la donnée reste souvent lacunaire et irrégulière.

D'un point de vue stratégique, ce projet ouvre la voie à une utilisation concrète de la data science au service des politiques publiques. En intégrant les prédictions du modèle dans les systèmes d'alerte précoce, il devient possible d'anticiper, département par département, les risques de détérioration de la situation alimentaire. De plus, les résultats peuvent orienter la planification humanitaire : cibler les zones où la sécheresse et la hausse des prix convergent, prioriser les programmes d'irrigation, ou encore ajuster les interventions logistiques en fonction de la vulnérabilité territoriale. L'intérêt de ce travail réside donc autant dans la **modélisation** que dans la **capacité d'action qu'il inspire**.

En conclusion, ce projet a permis de répondre à la problématique initiale en prouvant qu'il est possible de **prédire la phase d'insécurité alimentaire** à partir d'indicateurs combinant économie, climat et territoire. Il a montré que la faim n'est pas un mystère mais un signal mesurable, que la donnée peut capter et interpréter. L'objectif de construction d'un outil d'aide à la décision a été atteint : le modèle développé constitue une base solide pour des projections futures, tout en offrant une grille d'analyse pour la compréhension des déséquilibres régionaux. Il s'agit là d'une étape vers une gouvernance plus anticipative, où la donnée devient un levier d'équité et de prévention.

En définitive, cette recherche a voulu démontrer que la science des données, lorsqu'elle est utilisée avec rigueur et conscience, peut devenir un instrument de résilience nationale. Dans le contexte haïtien, chaque ligne de code, chaque variable nettoyée, chaque corrélation identifiée représente une tentative de rendre le pays un peu plus prévisible, un peu plus stable, et un peu plus capable de se défendre contre l'invisible : la faim.

---

9 Recommandations stratégiques

9.0.1 1 Pour les institutions publiques (CNSA, MARNDR)

- Créer un Observatoire national de la sécurité alimentaire prédictive basé sur ce modèle ;
- Intégrer la prévision IPC dans les politiques agricoles et de résilience ;
- Renforcer la collecte de données locales (prix, production, accès routier, pluviométrie) ;
- Utiliser les prédictions pour cibler les subventions agricoles ou les programmes de soutien aux ménages.

9.0.2 2 Pour les acteurs humanitaires (PAM, FAO, ONG)

- Utiliser le modèle comme outil d'aide à la planification géographique des interventions ;
- Déployer les aides en fonction du niveau de risque prédictif ;
- Développer des protocoles d'action rapide déclenchés automatiquement par les signaux du modèle;
- Croiser les prévisions avec les rapports de terrain pour affiner la réponse humanitaire.

9.0.3 3 Pour la communauté scientifique

- Étendre le modèle avec de nouvelles variables : accès à l'eau, mobilité, sécurité, état des infrastructures ;
- Promouvoir la collaboration entre chercheurs, ingénieurs et acteurs du terrain pour affiner les modèles prédictifs ;
- Promouvoir des programmes de formation en Data Science appliquée au développement durable.

10 Perspectives d'amélioration

- Intégration de données satellitaires haute résolution (MODIS, Sentinel) pour affiner les prévisions climatiques ;
- Utilisation de modèles plus avancés (XGBoost, LSTM, Random Forest Optimized) pour améliorer la précision ;
- Couplage avec des modèles socio-économiques pour simuler l'impact des politiques publiques (subventions, importations, aides).

[ ]:

11 Remerciements

Nous tenons à exprimer nos profondes gratitudes envers toutes les personnes et institutions qui ont contribué, directement ou indirectement, à la réalisation de ce projet.

- À **l'équipe pédagogique d'Akademi (powered by Flatiron School)**, pour l'encadrement rigoureux et la vision pratique qu'elle apporte à la science des données.

- À Mme Casteline, pour cette formation enrichissante.
- À nos **superviseurs, M. Wedter et M. Geovany**, pour leur accompagnement méthodologique, leurs conseils techniques et leur exigence de qualité.

- Enfin, à toutes les **institutions haïtiennes** (CNSA, FAO, MARNDR, PAM) dont les données et les rapports de terrain ont nourri la réflexion scientifique et opérationnelle.

Ce projet nous a permis de dépasser la simple approche académique pour comprendre la **valeur stratégique de la donnée dans les décisions publiques**.
Il nous a appris qu'un modèle prédictif n'a de sens que s'il **améliore concrètement la vie des gens**.

En travaillant sur la prédiction du risque d'insécurité alimentaire en Haïti, nous avons compris que la Data Science n'est pas seulement une discipline mathématique :
c'est un **langage de compréhension du monde**, un outil d'action face à des défis réels – la faim, la pauvreté, les catastrophes naturelles.

Au terme de cette soutenance, ce projet n'est pas une fin, mais un point de départ.
Il ouvre la voie vers une **nouvelle manière de penser la planification et la prévention en Haïti** — fondée sur la donnée, la science et la responsabilité collective.

"Les chiffres ne mentent pas, mais c'est à nous de leur donner un sens."

[ ]:

11.0.1 Fin du projet Capstone : Prédiction du risque d'insécurité alimentaire en Haïti

---

Merci pour votre attention

[ ]: