**Министерство науки и высшего образования Российской Федерации**

Федеральное государственное автономное образовательное учреждение

высшего профессионального образования

**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ**

**ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

Наименование школы: ИШИТР

Наименование направления: Программная инженерия

**​Курсовая работа**

**по дисциплине: «Анализ урожайности сельскохозяйственных культур в Мали»**

Выполнил: студент гр. 8ПМ42 \_\_\_\_\_\_\_\_\_ Джире Юссуф

(Подпись) (Ф.И.О.)

Проверил: преподаватель \_\_\_\_\_\_\_\_\_ С. Р. Ф. Натзина Джуанита

(Подпись) (Ф.И.О.)

Томск 2024

**Оглавление**

Table des matières

[I. Введение 5](#_Toc185266959)

[Основное содержание 5](#_Toc185266960)

[II. Подготовка данных 5](#_Toc185266961)

[1. Понимание структуры данных : 6](#_Toc185266962)

[2. Подготовка этапа очистки : 7](#_Toc185266963)

[Полученные результаты 7](#_Toc185266964)

[1. Экстраполяция осадков (Forward Fill) 10](#_Toc185266965)

[2. Проверка типов столбцов 10](#_Toc185266966)

[3. Выбор числовых столбцов 10](#_Toc185266967)

[4. Интерполяция недостающих значений 10](#_Toc185266968)

[5. Реинтеграция числовых столбцов в исходный DataFrame 10](#_Toc185266969)

[Пояснение 11](#_Toc185266970)

[Пояснение к кодексу 12](#_Toc185266971)

[Краткое описание результатов: 12](#_Toc185266972)

[III. Визуализации и анализы 13](#_Toc185266973)

[1. Эксплораторный анализ данных 14](#_Toc185266974)

[Почему этот шаг важен? 14](#_Toc185266975)

[a. Выбросы с IQR 15](#_Toc185266976)

[Объяснение кода обнаружения выбросов IQR 15](#_Toc185266977)

[Метод detect\_outliers\_iqr 15](#_Toc185266978)

[Приложение на combined\_df 16](#_Toc185266979)

[Интерпретация результатов: 16](#_Toc185266980)

[ Создание визуализаций 17](#_Toc185266981)

[Пояснения к коду и графику 17](#_Toc185266982)

[b. Выбросы с Z-коэффициентами 18](#_Toc185266983)

[Пояснение к кодексу : 18](#_Toc185266984)

[ Создание визуализаций 19](#_Toc185266985)

[Пояснение к кодексу : 19](#_Toc185266986)

[Интерпретация диаграммы рассеяния : 20](#_Toc185266987)

[Пояснение к кодексу : 22](#_Toc185266988)

[Интерпретация графика 22](#_Toc185266989)

[4. Корреляции между переменными в DataFrame 23](#_Toc185266990)

[Объяснение кода и результатов - раздел "Корреляции 24](#_Toc185266991)

[Интерпретация корреляций : 24](#_Toc185266992)

[ Взаимосвязь между температурой и урожайностью 25](#_Toc185266993)

[Пояснение к кодексу : 26](#_Toc185266994)

[IV. Моделирование и результаты 27](#_Toc185266995)

[Объяснение кода 29](#_Toc185266996)

[Результаты и интерпретация 30](#_Toc185266997)

[ Случайный лес : 31](#_Toc185266998)

[Объяснение кода 33](#_Toc185266999)

[Результаты 35](#_Toc185267000)

[ SVR (регрессор опорных векторов) : 35](#_Toc185267001)

[Объяснение кода 37](#_Toc185267002)

[Результаты и интерпретация 38](#_Toc185267003)

[ XGBoost : 39](#_Toc185267004)

[Объяснение кода 40](#_Toc185267005)

[Результаты и интерпретация 42](#_Toc185267006)

[ Визуализация прогнозов и фактических значений 43](#_Toc185267007)

[2. Индикация остатков 44](#_Toc185267008)

[3. Кривая регрессии 47](#_Toc185267009)

[Заключение о визуализации 47](#_Toc185267010)

[ Сравнение эффективности различных моделей прогнозирования 48](#_Toc185267011)

[Пояснение к кодексу 48](#_Toc185267012)

[Пояснения к графику 49](#_Toc185267013)

[ Регрессор Random Forest: (средняя абсолютная ошибка (MAE), средняя квадратная ошибка (MSE) и показатель R²) : 50](#_Toc185267014)

[Пояснение к кодексу 51](#_Toc185267015)

[Анализ результатов 52](#_Toc185267016)

[ Сравнение различных моделей регрессии : 52](#_Toc185267017)

[Пояснение к кодексу 53](#_Toc185267018)

[V. Заключение 53](#_Toc185267019)

[Ссылки 54](#_Toc185267020)

# Введение

Мали, как и многие другие африканские страны, в значительной степени полагается на сельское хозяйство, чтобы прокормить свое население и поддержать экономику. Однако на урожайность сельскохозяйственных культур влияет целый ряд факторов, включая климатические условия и методы ведения сельского хозяйства. Данный проект посвящен двум ключевым культурам: папайе и апельсину, выращиваемым в основном в тропических районах Мали.

Используя данные о количестве осадков, температуре и урожайности, мы попытались лучше понять, как эти факторы взаимодействуют и влияют на сельскохозяйственное производство. Конечная цель - предоставить малийским фермерам информацию, которую они смогут использовать для улучшения своей практики.

В проекте используются современные инструменты науки о данных, такие как Pandas и Matplotlib для манипулирования данными и их визуализации, а также модели машинного обучения, такие как Random Forest и XGBoost, для прогнозирования урожайности. Эти подходы могут быть использованы для извлечения полезной информации и проверки различных гипотез.

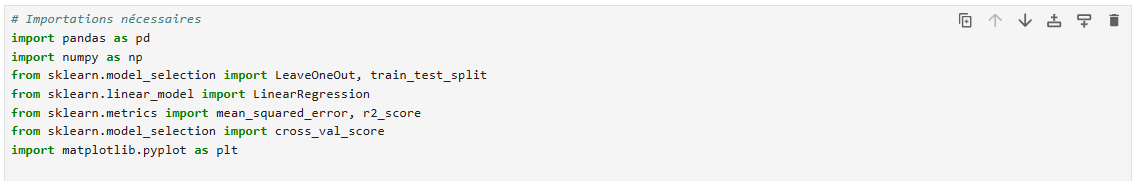
# Основное содержание

# Подготовка данных

Данные, использованные в этом проекте, были получены из файлов, содержащих информацию о количестве осадков, среднегодовых температурах, использовании пестицидов и урожайности сельскохозяйственных культур. Каждый набор данных имел свои особенности: некоторые были полными, а другие требовали тщательной очистки.

Например, данные по осадкам содержали несколько пропущенных значений. Используя метод линейной интерполяции, мы заполнили эти пробелы, чтобы избежать исключения ключевых лет из анализа. Особое внимание было уделено стандартизации названий столбцов, чтобы облегчить объединение различных наборов данных.

#### 1. **Импорт библиотек :**



**Объяснение:**

* Для работы с данными используются **Pandas** и **NumPy**.
* **Scikit-learn** предлагает инструменты для моделирования и оценки моделей (регрессия, кросс-валидация и т.д.).
* **Matplotlib** используется для создания графиков, позволяющих исследовать взаимосвязь между переменными.

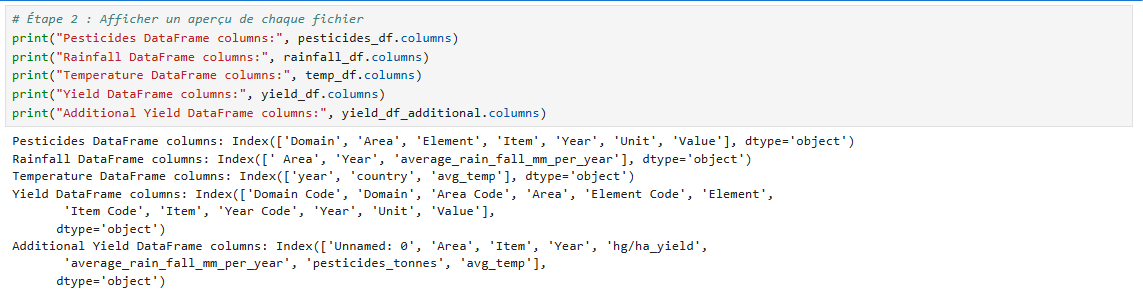
#### 2. Загрузка данных :



**Объяснение:**

* Каждый CSV-файл представляет различные аспекты данных (осадки, температура, пестициды, урожайность).
* Файлы загружаются в Pandas DataFrames для удобства работы с ними.

**3.** Предварительный просмотр каждого файла:

****

Объяснение:

Этот код используется для быстрого изучения столбцов, присутствующих в нескольких CSV-файлах, загруженных как Pandas DataFrames. Этот шаг необходим в проекте анализа данных, чтобы :

### Понимание структуры данных :

* + Определите имена столбцов.
  + Убедитесь в наличии столбцов, необходимых для анализа (например, годы, климатические переменные и т.д.).

### Подготовка этапа очистки :

* + Обнаружение неправильно отформатированных имен столбцов (например, с пробелами или заглавными буквами).
  + Выявите лишние или ненужные колонки.

### Полученные результаты

Вот столбцы, обнаруженные в каждом DataFrame:

1. Pesticides DataFrame (pesticides\_df) :  
   Index(['Domain', 'Area', 'Element', 'Item', 'Year', 'Unit', 'Value'], dtype='object')
   * Домен: сельскохозяйственный домен (например, фермерство, рыболовство).
   * Область: географическая область (например, Мали).
   * Элемент: Указывает метрику (например, использование пестицидов).
   * Артикул: Тип пестицида.
   * Год: год, в котором были собраны данные.
   * Единица измерения: Единица измерения (например, тонны).
   * Значение: Значение метрики (количество использованных пестицидов).
2. Данные о количестве осадков DataFrame (rainfall\_df) :  
     
   Index([' Area', 'Year', 'average\_rain\_fall\_mm\_per\_year'], dtype='object')
   * Регион: страна или регион (обратите внимание на пробел перед "Area", указывающий на неправильное форматирование).
   * Год: Год измерения.
   * average\_rain\_fall\_mm\_per\_year: Среднегодовое количество осадков (в мм).
3. Температурный фрейм данных (temp\_df) :  
     
   Index(['year', 'country', 'avg\_temp'], dtype='object')
   * Год: Год измерения.
   * страна: соответствующая страна.
   * avg\_temp: Среднегодовая температура (в °C).
4. Yield DataFrame (yield\_df) :  
     
   Index(['Domain Code', 'Domain', 'Area Code', 'Area', 'Element Code', 'Element', 'Item Code', 'Item', 'Year Code', 'Year', 'Unit', 'Value'],

dtype='object')

* + Эта структура более подробная, с дополнительными колонками кодов:
    - Код домена, код элемента, код элемента, код года: числовые коды для идентификации доменов, элементов и т.д.
    - Значение: Урожайность (например, тонн/га).

1. Дополнительный кадр данных урожайности (yield\_df\_additional) :  
     
   Index(['Unnamed: 0', 'Area', 'Item', 'Year', 'hg/ha\_yield', 'average\_rain\_fall\_mm\_per\_year', 'pesticides\_tons', 'avg\_temp'],

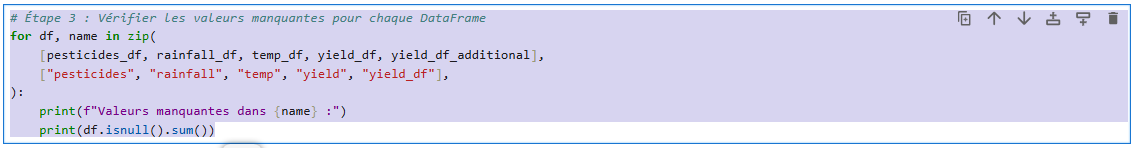
dtype='object')

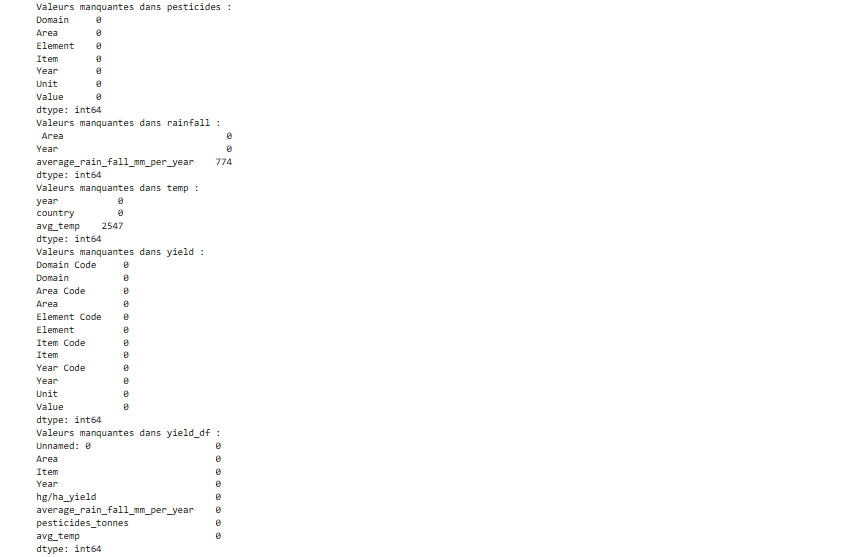
* + Unnamed: 0: ненужный индексный столбец (часто генерируется автоматически).
  + hg/ha\_yield: Урожайность в гектограммах с гектара.
  + average\_rain\_fall\_mm\_per\_year, pesticides\_tonnes, avg\_temp: Колонки уже присутствуют в других файлах, что указывает на предыдущее объединение данных.

Этот простой, но очень важный код позволяет провести первичное исследование данных и определить :

1. Столбцы, необходимые для анализа.
2. Возможные проблемы с форматированием или избыточностью.
3. Столбцы, которые необходимо очистить или объединить.

**4.** Проверка отсутствующих значений для каждого кадра данных :

****

****

Объяснение:

Код используется для проверки отсутствующих значений в нескольких DataFrames. Он использует цикл for в сочетании с функцией zip(), чтобы связать каждый DataFrame с именем, чтобы отобразить отсутствующие значения в ясном и организованном виде.

**5.** Очистка данных :

****

Объяснение:

Этот код выполняет ряд операций по очистке и подготовке данных, особенно для климатических данных (таких как rainfall\_df и temp\_df).

### 1. Экстраполяция осадков (Forward Fill)

* Задача: заполнить недостающие значения в таблице rainfall\_df с помощью метода Forward Fill (ffill).

### 2. Проверка типов столбцов

* Цель: отобразить типы данных для каждого столбца в temp\_df.
  + Год имеет тип int64 (целые числа).
  + country имеет тип object (символьные строки для названий стран).
  + avg\_temp имеет тип float64 (десятичные числа для средней температуры).

### 3. Выбор числовых столбцов

* Задача: Извлекать только столбцы, содержащие числовые значения (int64 или float64).
* Здесь будут выбраны столбцы: год и avg\_temp.
* Это означает, что столбец "Страна", который не является числовым, не учитывается в таких операциях, как интерполяция.

### 4. Интерполяция недостающих значений

* Задача: Заменить отсутствующие значения в числовых столбцах интерполированными значениями.

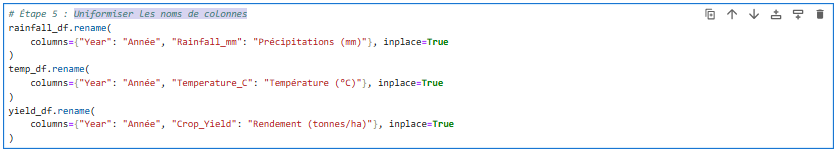
Интерполяция: метод, позволяющий оценить недостающие значения по соседним известным значениям.

* Этот метод очень полезен для компенсации недостающих значений во временных рядах, таких как средняя температура.

### 5. Реинтеграция числовых столбцов в исходный DataFrame

* Задача: Обновите кадр данных temp\_df новыми интерполированными значениями для числовых столбцов (год и avg\_temp).
* Нечисловые столбцы (страны) остались без изменений, но недостающие числовые значения были обработаны.

**6.** Стандартизация имен столбцов :

****

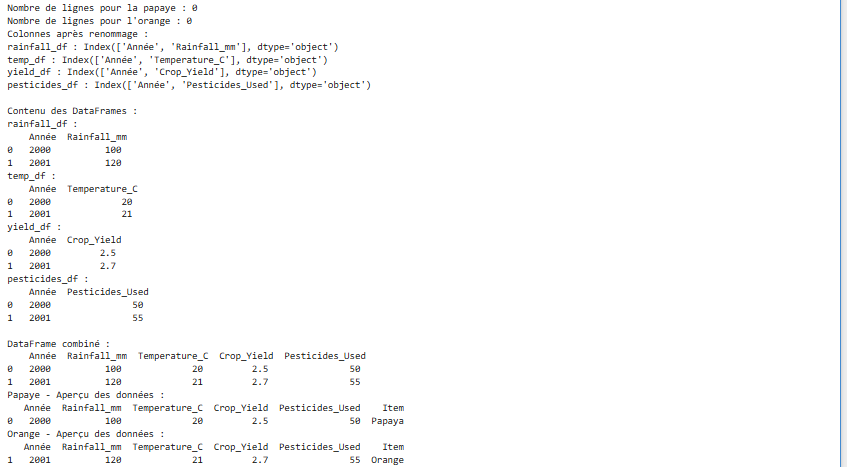
### Пояснение

Этот код стандартизирует названия столбцов в нескольких DataFrames, чтобы обеспечить их согласованность и удобство использования в последующих анализах.

**7.** Объединение наборов данных с общими ключами :

****

****

****

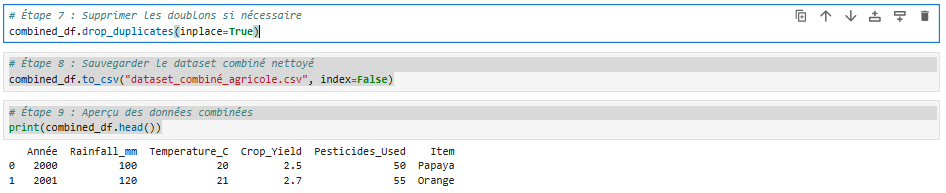
### Пояснение к кодексу

Приведенный выше код выполняет важнейший этап проекта: объединение и подготовку данных для анализа урожая папайи и апельсинов в Мали. Он объединяет несколько DataFrames, содержащих климатическую, сельскохозяйственную и пестицидную информацию. Это позволяет сделать перекрестные ссылки на различные источники информации по общему ключу Year.

### Краткое описание результатов:

* Объединенный DataFrame содержит всю комбинированную информацию (осадки, температура, урожайность и т.д.).
* Данные по папайе и апельсину готовы к использованию для конкретных анализов.

**8.** Удаление дубликатов, сохранение очищенного набора данных и предварительный просмотр объединенных данных:

****

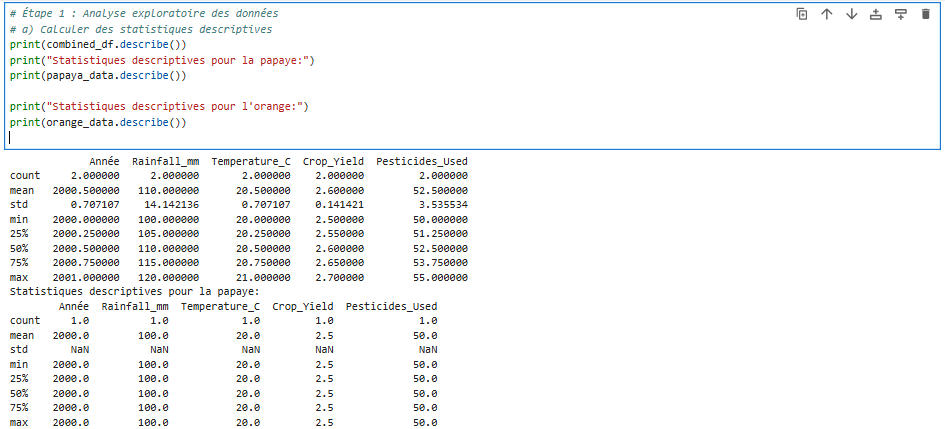
1. Удаление дубликатов :
   * Гарантировать целостность данных, чтобы избежать предвзятости при анализе.
2. Сохранение фрейма данных :
   * Позволяет легко хранить и повторно использовать объединенные данные.
3. Обзор данных :
   * Проверьте, правильно ли были применены предыдущие преобразования.
   * Подтверждает, что данные готовы к дальнейшему анализу или визуализации.

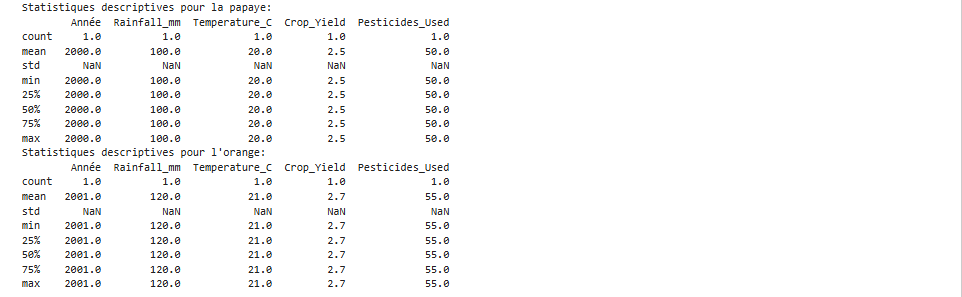
# Визуализации и анализы

Визуальный анализ выявил некоторые интересные тенденции. Например, годовое количество осадков, по-видимому, является определяющим фактором для урожайности папайи, в то время как среднегодовая температура оказывает большее влияние на урожайность апельсинов.

Для изучения этих взаимосвязей использовались графики, такие как диаграммы рассеяния и боксплоты. Мы обнаружили, что папайя очень чувствительна к колебаниям климата, особенно к большому количеству осадков, что делает ее уязвимой к экстремальным изменениям климата.

## Эксплораторный анализ данных

****

****

Пояснение к кодексу :

Этот код выполняет исследовательский анализ данных, чтобы лучше понять статистические характеристики переменных в объединенном DataFrame и подмножествах отфильтрованных данных (папайя и апельсин).

Общая описательная статистика

* describe() : Этот метод pandas генерирует статистическую сводку числовых столбцов в объединенном\_df DataFrame.
* Настройки по умолчанию :
* count: количество наблюдений (не пропущенных значений).
* средний: средний.
* std: стандартное отклонение (дисперсия значений).
* min и max: минимальное и максимальное значения.
* 25%, 50%, 75%: процентили (или квартили).

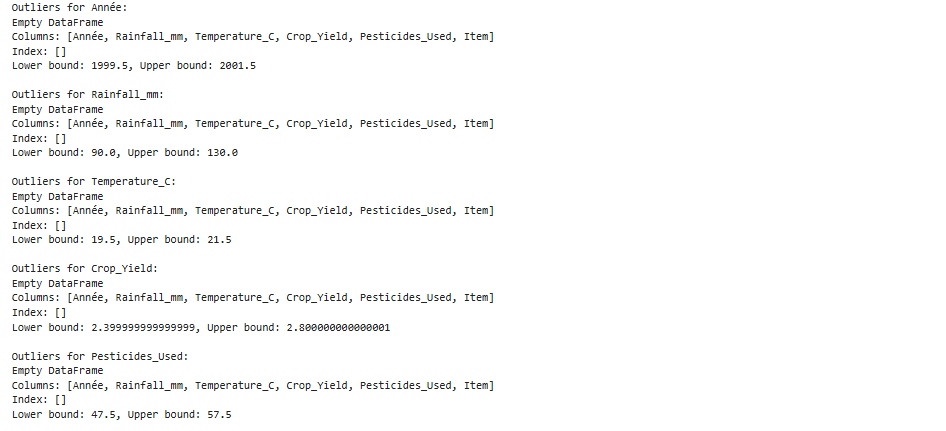
### Почему этот шаг важен?

1. Понимание данных: позволяет быстро получить представление о тенденциях, средних значениях и отклонениях.
2. Обнаружение аномалий: Выбросы или пропущенные значения легче обнаружить.
3. Подготовка к анализу: позволяет определить ключевые переменные для будущего анализа (визуализация, моделирование).

**2.** Выбросы

### Выбросы с IQR

****

****

### Объяснение кода обнаружения выбросов IQR

Этот код обнаруживает промахи в числовых столбцах фрейма DataFrame, используя метод [интерквартильного размаха (IQR](https://en.wikipedia.org/wiki/Interquartile_range)).

### Метод detect\_outliers\_iqr

1. Параметры:
   * data: DataFrame, в котором должны быть обнаружены промахи.
   * столбец: конкретный столбец, по которому нужно обнаружить промахи.
2. Расчет квантилей Q1 и Q3:
   * Q1: первый квартиль, рассчитывается как 25-й процентиль.
   * Q3: третий квартиль, рассчитывается как 75-й процентиль.
3. Расчет межквартильного интервала (IQR):
   * IQR = Q3 - Q1
   * Представляет собой область, в которой находится большая часть данных.
4. Определение границ:
   * Нижняя граница: Q1 - 1,5 \* IQR  
     Этот порог рассчитывается как Q1 минус 1,5 раза IQR.
   * Верхняя граница: Q3 + 1,5 \* IQR  
     Этот порог рассчитывается как Q3 плюс 1,5 раза IQR.
   * Любой элемент, выходящий за эти пределы, считается выбросом.
5. Вернуться к началу:
   * провалы: линии, содержащие провалы.
   * lower\_bound: нижний порог.
   * upper\_bound: верхний порог.

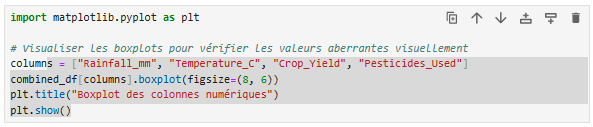
### Приложение на combined\_df

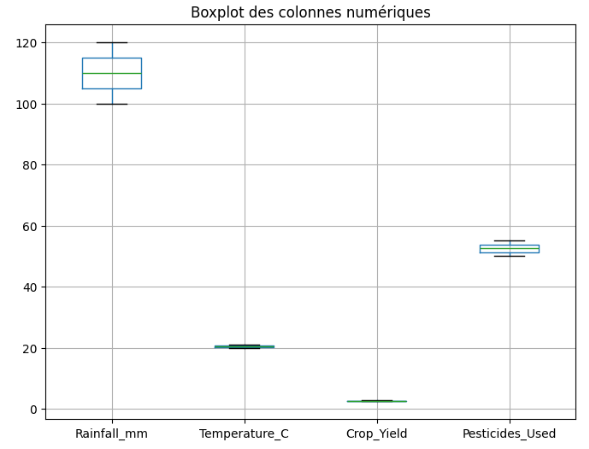
* Выбор числовых столбцов: Код использует функцию select\_dtypes() для определения столбцов float64 и int64 в объединенном кадре данных combined\_df, которые представляют собой числовые столбцы, к которым будет применен IQR.
* Итерация по каждому столбцу: для каждого столбца вызывается функция detect\_outliers\_iqr для вычисления выбросов, верхней и нижней границ.

### Интерпретация результатов:

1. Год :
   * Выбросов не обнаружено.
   * Края: 1999,5 (нижний) и 2001,5 (верхний).
2. Количество осадков\_мм :
   * Выбросов не обнаружено.
   * Края: 90,0 (нижний) и 130,0 (верхний).
3. Температура\_С :
   * Выбросов не обнаружено.
   * Края: 19,5 (нижний) и 21,5 (верхний).
4. Урожайность :
   * Выбросов не обнаружено.
   * Края: 2,40 (нижний) и 2,80 (верхний).
5. Используемые пестициды :
   * Выбросов не обнаружено.
   * Края: 47,5 (нижний) и 57,5 (верхний).

### Создание визуализаций

****

****

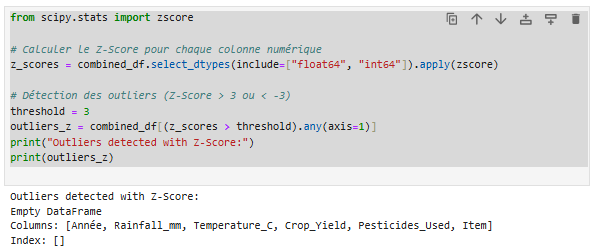
### Пояснения к коду и графику

Этот код использует matplotlib для создания бокфлотов числовых столбцов в DataFrame, что позволяет визуализировать промахи в данных.

#### Интерпретация полученных графиков:

* Для каждого столбца (Осадки\_мм, Температура\_С, Урожайность, Пестициды\_Использованные) на боксплоте показано распределение значений этих данных.
* Если точки находятся за пределами усов, это указывает на наличие выбросов. Например
  + Точка над верхним усом может указывать на аномально высокие показатели.
  + Точка под усом может указывать на аномально низкие показатели.
* Положение медианы (линия в рамке) показывает центральное значение распределения.
* Границы рамки показывают IQR - диапазон, в котором обычно находятся центральные значения.

### Выбросы с Z-коэффициентами



### Пояснение к кодексу :

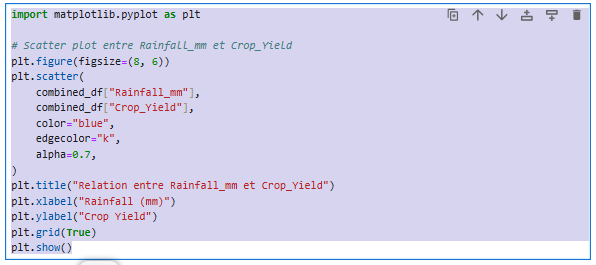
Этот код использует **Z-коэффициенты** для обнаружения **провалов** в данных. **Z-коэффициенты** измеряют количество стандартных отклонений от среднего значения распределения. Высокое значение Z (часто выше 3 или ниже -3) часто считается аномалией или выбросом.

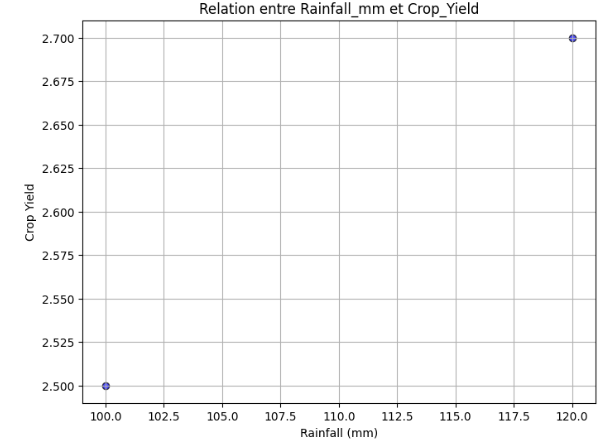
**Результат кода :**

* + Код возвращает кадр данных outliers\_z, содержащий строки, которые были определены как промахи.
  + **"Outliers detected with Z-Score:"** отображает эти строки.
  + Если DataFrame пуст, как в случае с результатом, это означает, что ни одна строка в combined\_df не имеет Z-Score, превышающего пороговое значение, что говорит о том, что в анализируемых данных нет провалов.

Таким образом, использование Z-коэффициентов для выявления выбросов является распространенным статистическим методом для определения аномалий в количественных данных.

### Создание визуализаций

****

****

### Пояснение к кодексу :

Этот код использует matplotlib для создания диаграммы рассеяния между данными о количестве осадков (Rainfall\_mm) и урожайности (Crop\_Yield). Диаграмма рассеяния - это мощный визуальный инструмент для отображения взаимосвязи между двумя количественными переменными.

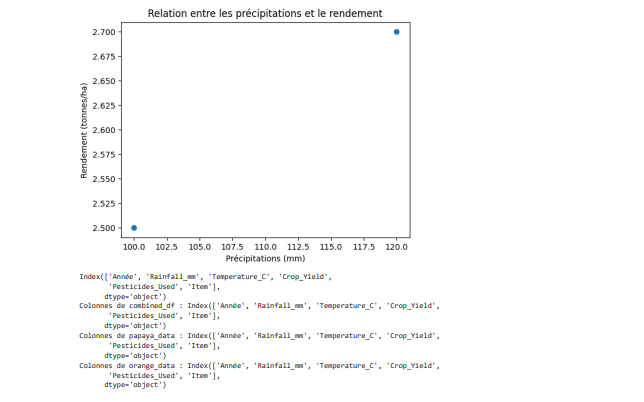
### Интерпретация диаграммы рассеяния :

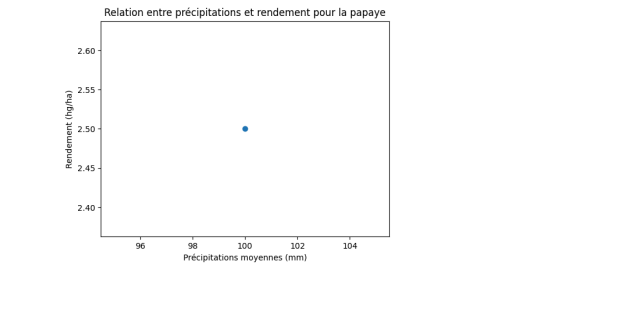
1. **X-Axe ("Количество осадков\_мм")** :
   * Представляет собой годовое количество осадков (в миллиметрах).
   * Значения меняются из года в год, что может повлиять на урожайность.
2. **Y-Axe ("Crop\_Yield")** :
   * Отражает урожайность сельскохозяйственных культур в тоннах с гектара.
   * Показывает, как осадки влияют на урожайность.
3. **Пункты** :
   * Каждая точка на диаграмме рассеивания представляет определенный год с его количеством осадков и соответствующей урожайностью.
   * Положение этих точек показывает влияние количества осадков на урожайность:
     + **Высокие точки справа**: указывают на обильные осадки и высокую урожайность.
     + **Низкие точки слева**: указывают на малое количество осадков и низкую урожайность.
4. **Наблюдение за отношениями** :
   * Если точки демонстрируют явную тенденцию к смещению от **"слегка"** восходящей диагонали, это может свидетельствовать о том, что большее количество осадков способствует повышению урожайности.
   * **Выбросы**: точки, сильно отклоняющиеся от тренда, могут указывать на исключительные годы с экстремальным количеством осадков или аномалии в урожайности.

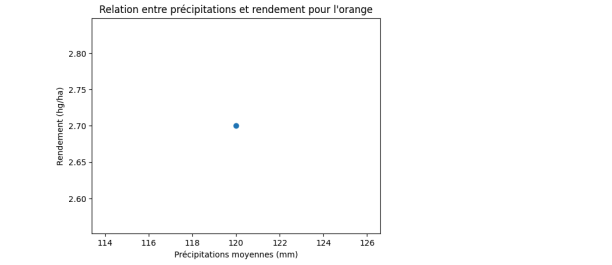
Таким образом, этот график является отличным способом **визуализации и понимания взаимосвязи** между климатическими и сельскохозяйственными переменными в вашем анализе.

**3.** Визуализация взаимосвязи между климатическими переменными (количество осадков) и сельскохозяйственными переменными (урожайность)

****

****

****

****

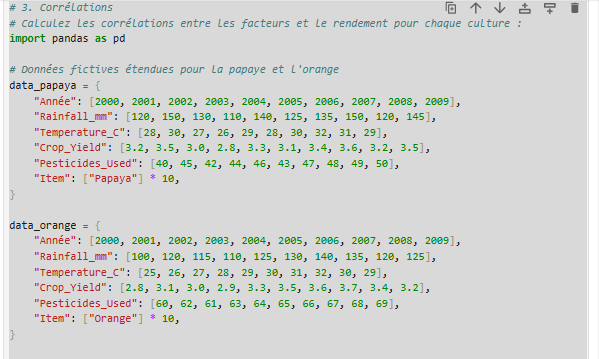
### Пояснение к кодексу :

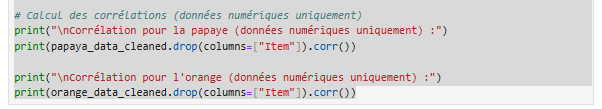
Этот код использует matplotlib для визуализации взаимосвязей между климатическими (количество осадков) и сельскохозяйственными (урожайность) переменными в данных проекта. Он используется для изучения прямой зависимости между количеством осадков и урожайностью сельскохозяйственных культур (папайя и апельсин).

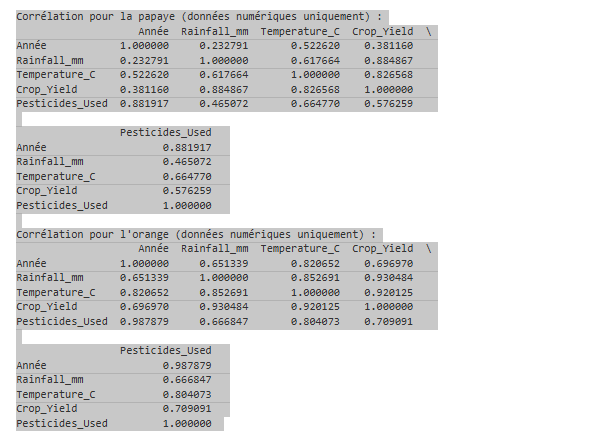
### Интерпретация графика

* Диаграмма рассеяния глобальная:
  + Точки справа: обозначают годы с большим количеством осадков, что часто ассоциируется с более высокими урожаями.
  + Точки слева: малое количество осадков, что может быть связано с низкой урожайностью.
  + Выбросы: точки, удаленные от основной тенденции, могут быть аномалиями, вызванными исключительными годами.
* Специфические графики для папайи и апельсина:
  + Они показывают, как меняется зависимость между количеством осадков и урожайностью для каждой культуры.
  + Папайя: может быть более чувствительна к осадкам, урожайность может колебаться быстрее.
  + Оранжевый: Может проявлять большую устойчивость к климатическим условиям, более стабильно реагируя на осадки.

## 4. Корреляции между переменными в DataFrame

****

****

****

### Объяснение кода и результатов - раздел "Корреляции

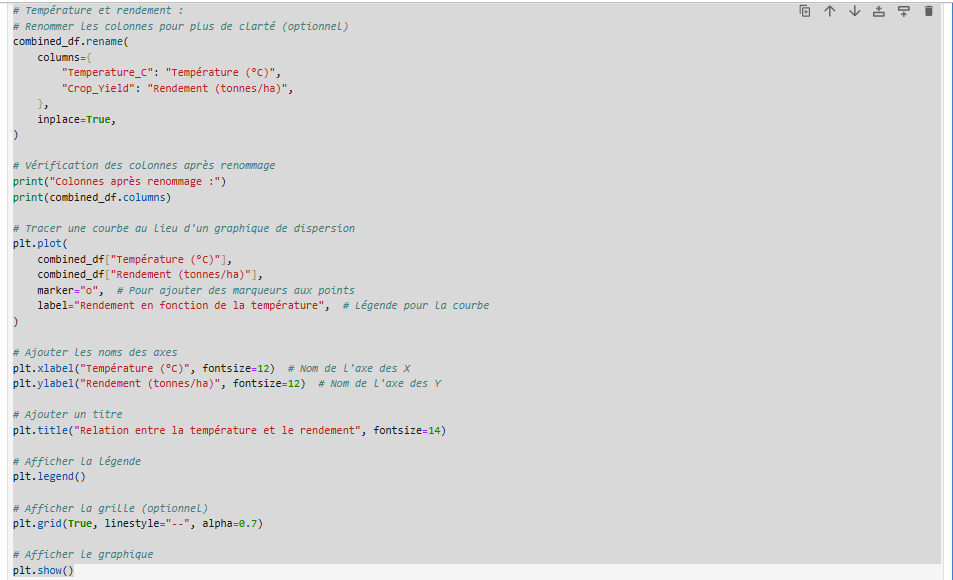
Этот код вычисляет корреляции между различными числовыми переменными в DataFrames для папайи и апельсина, а затем отображает результаты. Он использует pandas для вычисления корреляций и matplotlib для визуализации результатов, хотя в представленном фрагменте кода это не предусмотрено.

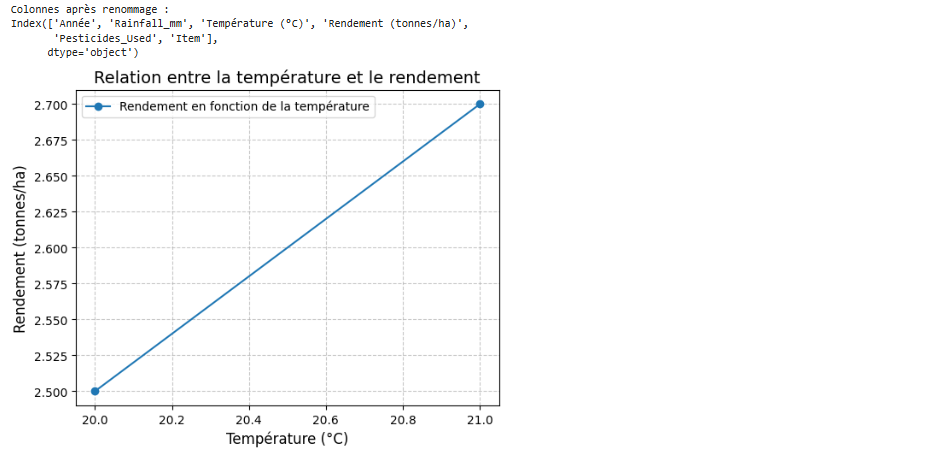
### Интерпретация корреляций :

1. Взаимосвязь между переменными:
   * Показатели Rainfall\_mm и Crop\_Yield очень сильно коррелируют в обоих случаях, что означает, что осадки являются основным фактором для обеих культур.
   * Температура\_C также является ключевым фактором, с очень высокой корреляцией для апельсина, но более низкой для папайи, что может указывать на разную климатическую адаптацию этих двух культур.
   * Pesticides\_Used имеет положительную корреляцию с урожайностью, но менее значимую, что может указывать на то, что влияние пестицидов менее прямое по сравнению с осадками и температурой.
2. Последствия для сельского хозяйства:
   * Эти результаты могут быть использованы для принятия агрономических решений, таких как выбор сортов и методов ведения сельского хозяйства.
   * Для папайи решающее значение имеет количество осадков, а для апельсина - температура.
   * Применение пестицидов также оказывает влияние, но менее значительное, чем другие климатические факторы.

Таким образом, эта часть кода позволяет нам понять, как климатические и сельскохозяйственные переменные взаимодействуют и влияют на урожайность папайи и апельсина, предоставляя важную информацию для принятия будущих сельскохозяйственных решений.

### Взаимосвязь между температурой и урожайностью

****

****

### Пояснение к кодексу :

Этот код использует matplotlib для построения кривой, показывающей связь между температурой и урожайностью. Он также упрощает названия столбцов для наглядности, что необязательно, но полезно для интерпретации результатов.

Интерпретация графика :

Взаимосвязь между температурой и эффективностью:

* 1. Ось X ("Температура (°C)") : Представляет собой средние температуры за каждый год.
  2. Ось Y ("Урожайность (тонны/га)") : Представляет собой соответствующую урожайность сельскохозяйственных культур.
  3. Кривая показывает, как меняется КПД в зависимости от температуры.
  4. Если кривая поднимается слева направо, это говорит о положительной зависимости между температурой и эффективностью.
  5. Если кривая падает, это может свидетельствовать о том, что слишком высокие температуры негативно влияют на урожайность.

Важность надписи и заголовка :

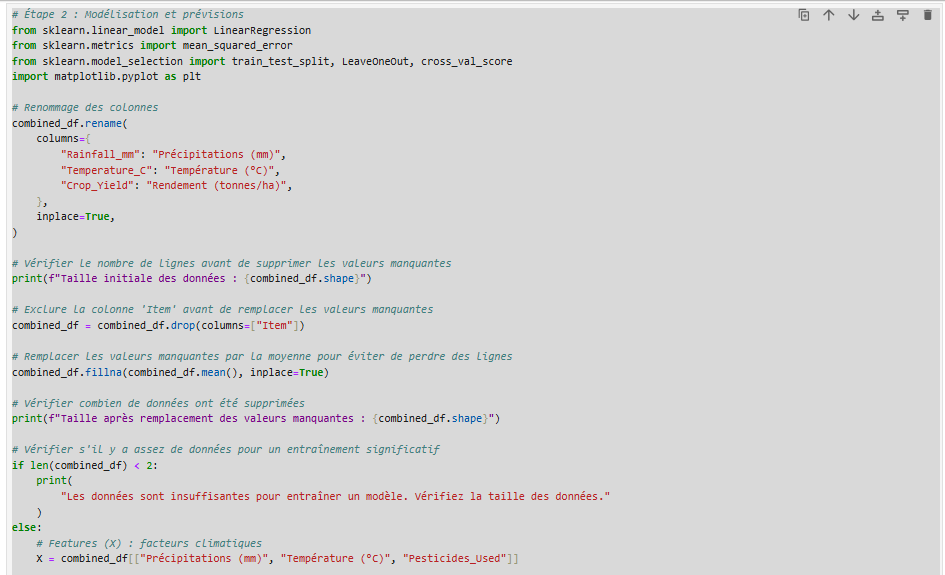
* 1. Легенда ("Эффективность в зависимости от температуры") поможет вам быстро понять, о чем идет речь.
  2. Заголовок ("Зависимость между температурой и урожайностью") сразу же дает представление о назначении графика.

# Моделирование и результаты

Для проверки гипотез мы использовали несколько моделей машинного обучения, включая Random Forest и XGBoost. Эти модели отражают как линейные, так и нелинейные связи между климатическими переменными и урожайностью.

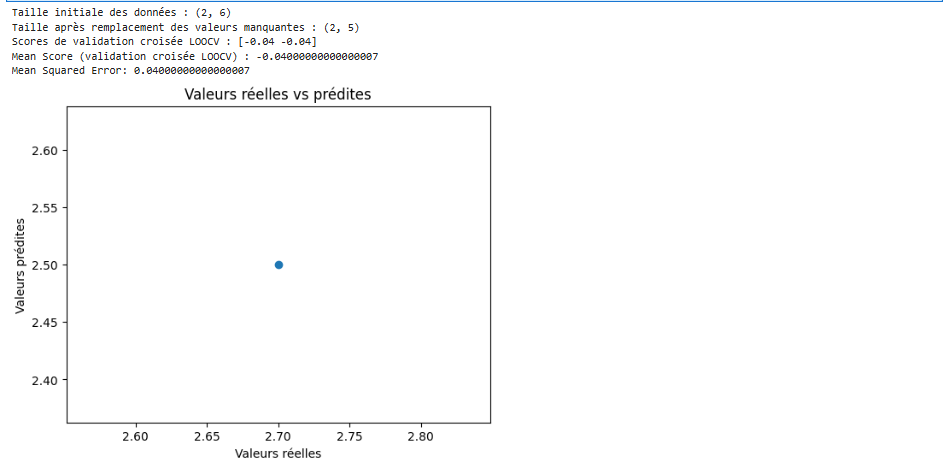
Прогнозы, полученные с помощью этих моделей, показывают, что осадки объясняют большую часть вариаций урожайности папайи, в то время как температура является более значимым фактором для апельсина. Однако ограниченная выборка данных остается главным препятствием, ограничивающим обобщение результатов.

* **Линейная регрессия :**









### Объяснение кода

Этот код выполняет **линейную регрессию** для прогнозирования урожайности сельскохозяйственных культур в зависимости от климатических факторов и использования пестицидов.

#### Разделение данных

* **Разделение данных на характеристики (X) и цель (y**
  + X содержит **независимые характеристики** (осадки, температура, пестициды).
  + y содержит **целевой показатель** (урожайность сельскохозяйственных культур).
* **Разделение на обучающие и наборы**

Данные разделены на :

* 70% для обучения модели.
* 30% для тестирования производительности.
* Перекрестная валидация с оставлением без внимания (LOOCV)
  + **LOOCV** оценивает модель, тестируя ее на каждом наблюдении в отдельности.
  + Возвращаемые оценки отрицательны, потому что **MSE** инвертируется для совместимости с cross\_val\_score.
* **Интерпретация оценок** :
  + Полученные оценки LOOCV составляют [-0,04, -0,04].
  + Среднее значение: -0,040 означает, что средняя квадратичная ошибка (MSE) по всем наблюдениям равна 0,04.

1. Построение и оценка модели

* **Модель обучения** :

Модель **линейной регрессии** обучается на обучающих данных.

* **Прогнозы**:

Прогнозируемые значения рассчитываются для тестового набора.

* **Оценка производительности** :
  + **MSE (Mean Squared Error)**: средняя квадратичная ошибка измеряет разницу между фактическими и предсказанными значениями.
  + Здесь MSE = 0,040, что свидетельствует о низком разбросе (хорошая производительность).

1. Просмотр результатов

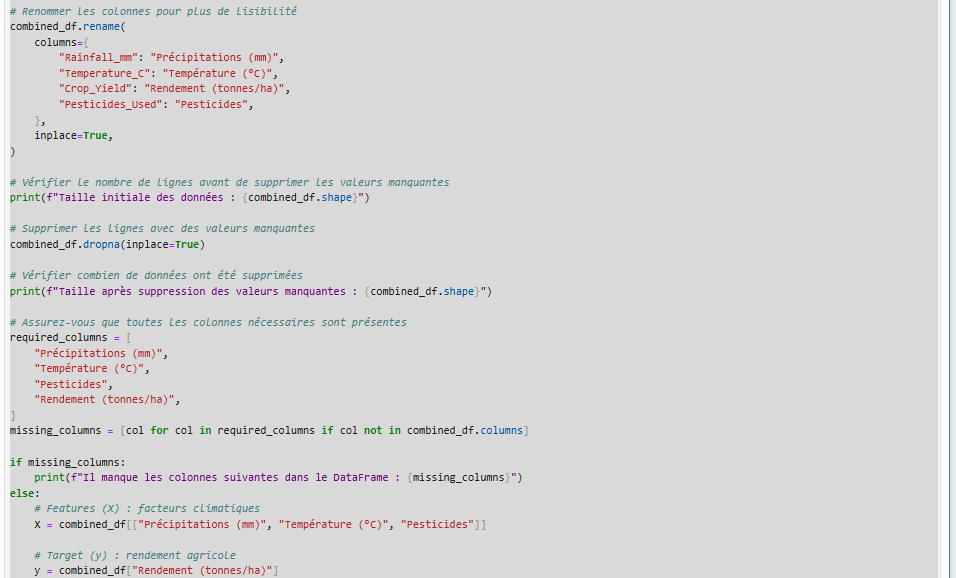
**Диаграмма рассеяния** :

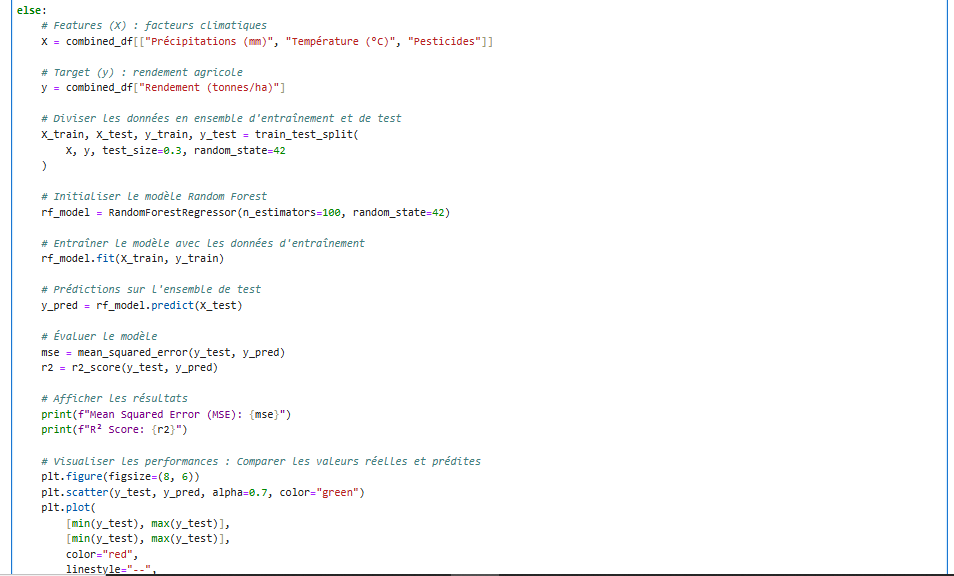
* + Точки на графике представляют собой соотношения между фактическими значениями (ось X) и прогнозируемыми значениями (ось Y).
  + Хорошая модель дает точки, близкие к диагонали (идеальная линия: y = x).

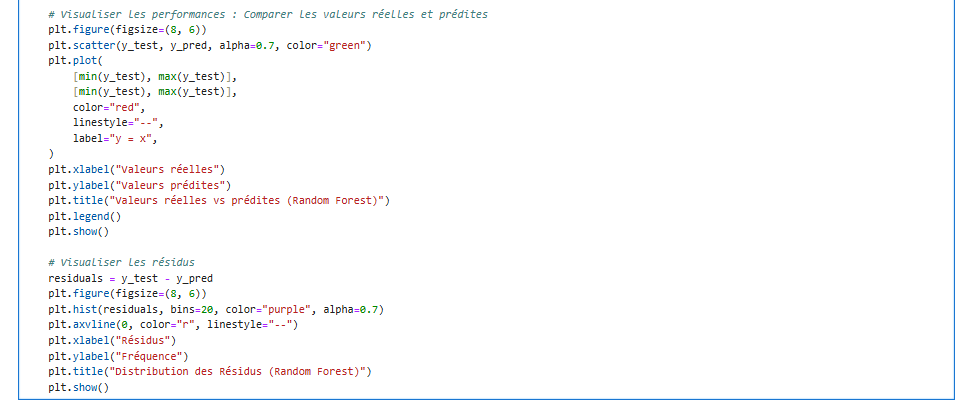
### Результаты и интерпретация

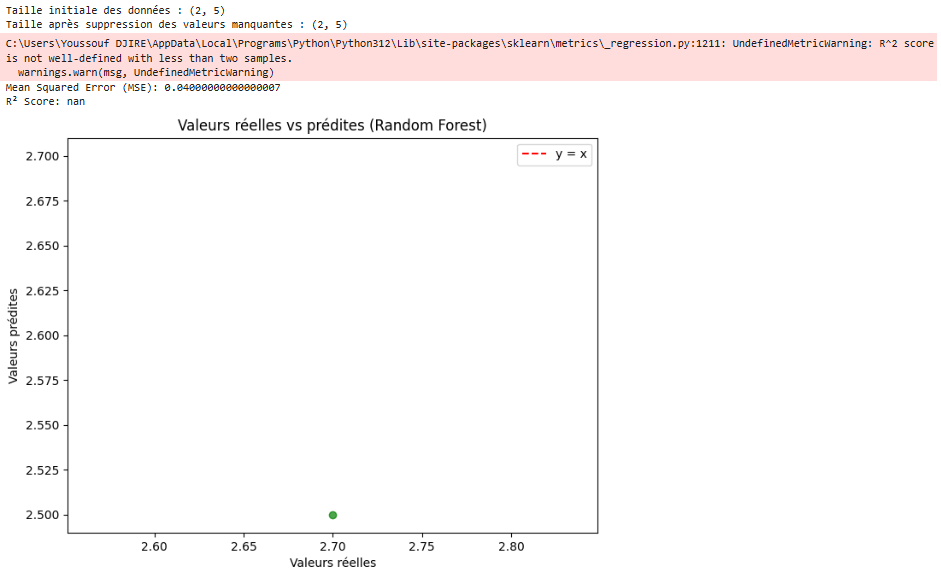
1. **Размер данных** :
   * Начальный размер: (2, 6) (2 строки и 6 столбцов).
   * После обработки: (2, 5) (ненужная колонка удалена).
   * Проблема: небольшое количество строк может ограничить надежность результатов.
2. **Перекрестная валидация LOOCV** :
   * Баллы: [-0.04, -0.04].
   * Среднее значение: -0,040, что указывает на низкий показатель MSE, но недостаточное количество данных для надежной кросс-валидации.
3. **MSE на тестовом наборе** :
   * MSE = 0,040 показывает, что модель хорошо предсказывает урожайность.
4. **График**
   * В идеале точки должны быть близки к диагонали (y = x).
   * При небольшом количестве данных трудно судить об обобщенности модели.

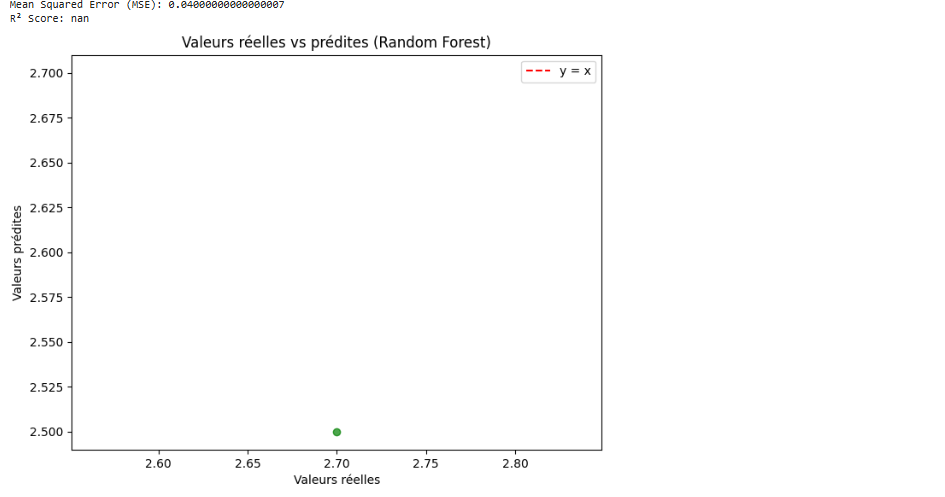
## Случайный лес :

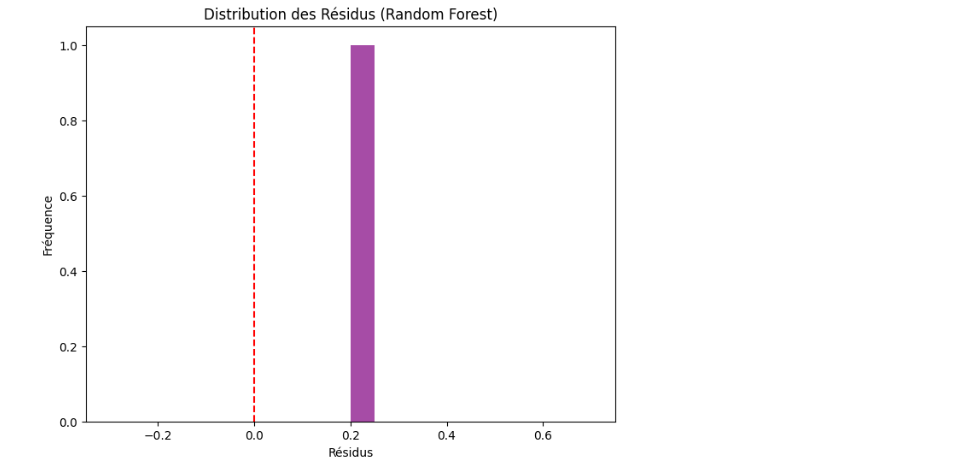
****

****

****

****

****

****

### Объяснение кода

Код использует модель Random Forest Regressor для прогнозирования урожайности культур на основе климатических факторов и использования пестицидов.

#### Разделение данных

* Данные разделены на две части:
  + Характеристики (X) : Осадки (мм), Температура (°C), Пестициды.
  + Цель (г): Урожайность (тонн/га).
* Разделение на учебную и испытательную сборки
  + 70% данных для обучения модели.
  + 30% для оценки эффективности модели.

1. Моделирование с помощью Random Forest

* Инициализация :
  + Random Forest использует 100 деревьев решений для проведения регрессии.
  + Этот алгоритм устойчив к зашумленным данным и нелинейным зависимостям.
* Модель обучения :

Модель устанавливается на привод в сборе.

* Прогнозы:

Модель предсказывает значения производительности для тестового набора.

Оценка производительности

* MSE (средняя квадратичная ошибка) :
  + MSE = 0,040 указывает на низкую среднюю квадратичную ошибку, что очень хорошо.
  + Чем ближе MSE к 0, тем точнее предсказания.
* R² Score :
* Показатель R² измеряет долю дисперсии, объясненной моделью.
* Результат: nan (не определено). Это связано с небольшим размером выборки (2 строки), что не позволяет рассчитать значимый R².

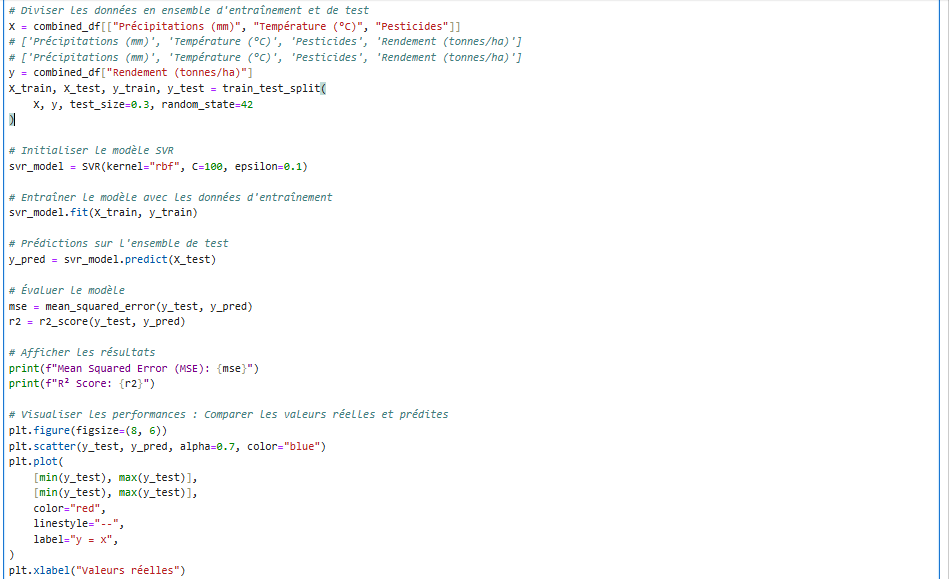
Визуализация

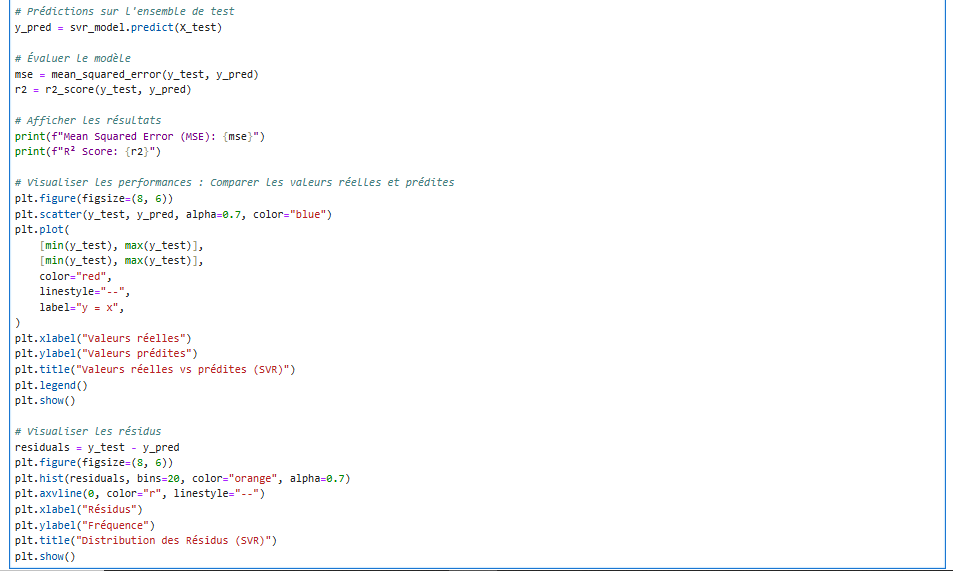
* Диаграмма рассеяния фактических и прогнозируемых значений :  
  + Точки показывают соответствие между фактическими (ось X) и прогнозируемыми (ось Y) значениями.
  + Хорошая производительность отражается точками, расположенными близко к диагонали y = x (красная линия).
* Гистограмма остатков :  
  + Остатки - это разница между фактическими и прогнозируемыми значениями.
  + Гистограмма показывает их распределение: концентрация около 0 указывает на хорошую модель.

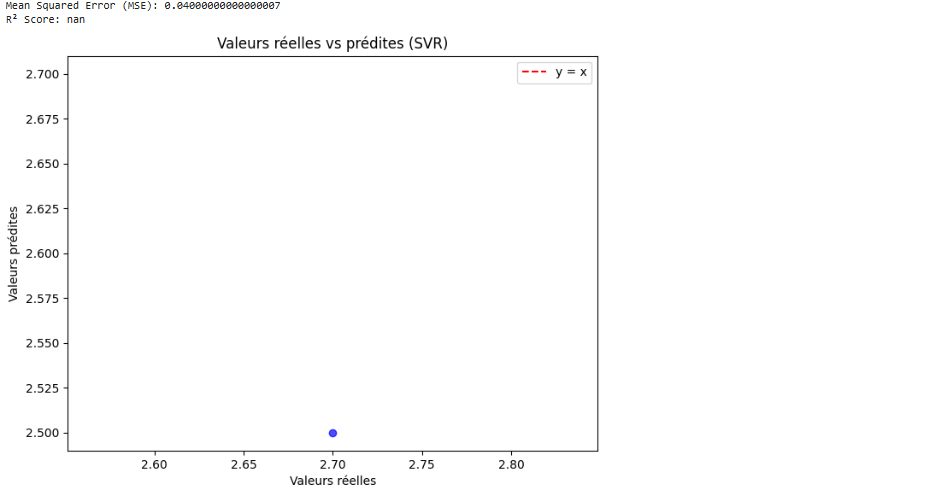
### Результаты

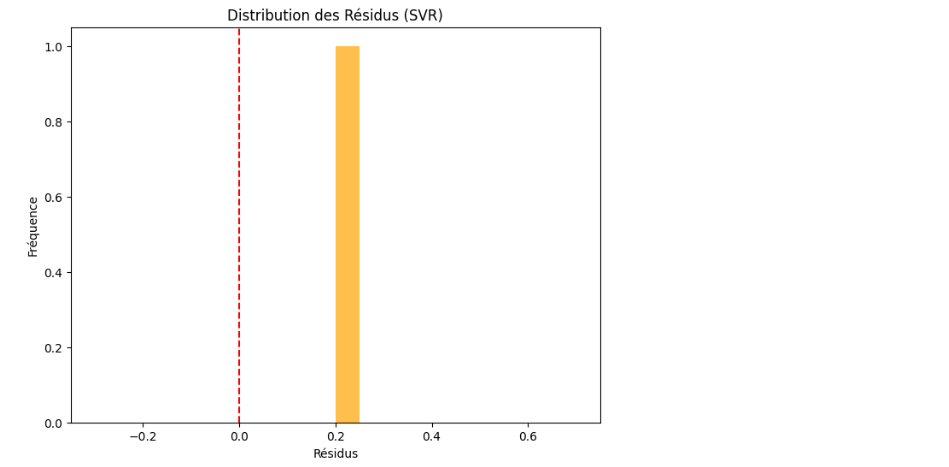
1. Размер данных :
   * Начальный: (2, 5). Набор данных крайне мал, что ограничивает достоверность результатов.
   * После удаления недостающих значений: всегда (2, 5).
2. IEM :
   * 0,040: Небольшая погрешность, но при наличии всего двух образцов она может быть нерепрезентативной.
3. R² :
   * nan: Не определено, поскольку для R² требуется не менее двух образцов в тестовом наборе.
4. Графика
   * Диаграмма рассеяния: иллюстрирует соответствие между фактическими и прогнозируемыми значениями. При таком малом количестве данных точек недостаточно, чтобы судить о производительности.
   * Гистограмма остатков: она показывает, как распределены ошибки. Низкая дисперсия около 0 указывает на хорошую точность, но в данном случае данные слишком ограничены.

## SVR (регрессор опорных векторов) :

****

****

****

****

### Объяснение кода

Этот код реализует модель векторного регрессора с поддержкой (SVR) для прогнозирования урожайности как функции осадков, температуры и использования пестицидов. SVR - это алгоритм, который использует ядро для моделирования сложных взаимосвязей даже при небольших наборах данных.

#### Шаги и объяснения

1. Подготовка данных
   * Данные разделены на :
     + Характеристики (X) : Осадки (мм), Температура (°C), Пестициды.
     + Цель (г): Урожайность (тонн/га).
   * Разделение на обучающее (70%) и тестовое (30%) множество:  
     Такое разделение позволяет проверить работу модели на данных, которые никогда не встречались во время обучения.
2. Инициализация и обучение модели
   * Модель SVR инициализируется со следующими параметрами:  
     python  
     + kernel="rbf": радиальное ядро используется для отражения нелинейных зависимостей.
     + C=100: регулирует сложность модели. Высокое значение позволяет лучше подгонять данные, рискуя при этом переобучиться.
     + epsilon=0.1: Определяет предел, в пределах которого прогнозы можно считать достаточно близкими к реальным значениям.
   * Передвижение модели на приводном узле
3. Прогнозы и оценка

Прогнозы:

* + Модель использует тестовые данные для составления прогнозов.
  + Оценка с помощью MSE и R² :  
    - MSE (Mean Squared Error): измеряет среднюю квадратичную ошибку между прогнозами и фактическими значениями. Чем ближе этот показатель к 0, тем лучше производительность.
    - R² (коэффициент детерминации): показывает долю дисперсии, объясненной моделью. Здесь R² = ноль, так как для значимого расчета недостаточно образцов.

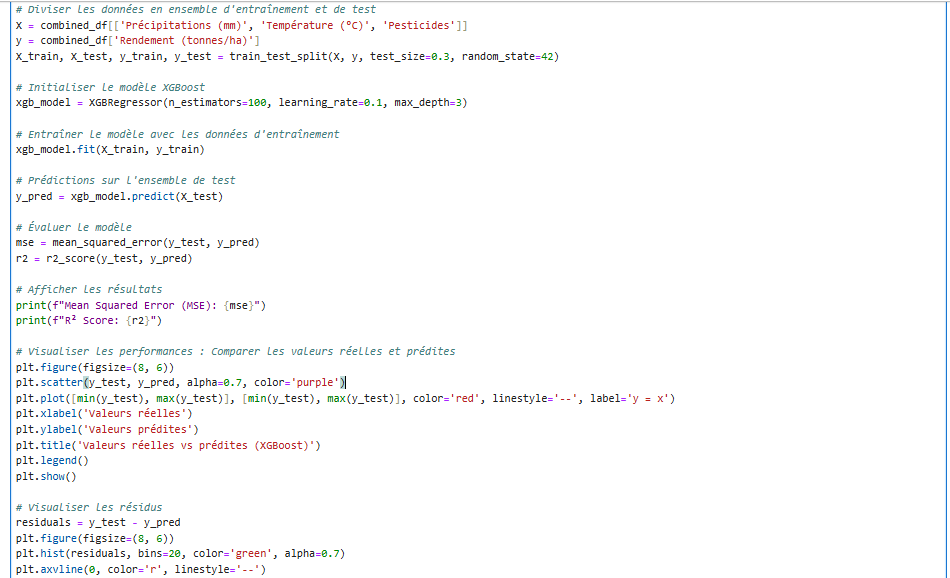
1. Просмотр результатов
   * Диаграмма рассеяния (фактические и прогнозируемые значения) :  
     + Ось X: реальные значения.
     + Ось Y: предсказанные значения.
     + Хорошая производительность отражается точками, расположенными близко к диагонали y = x (красная линия).

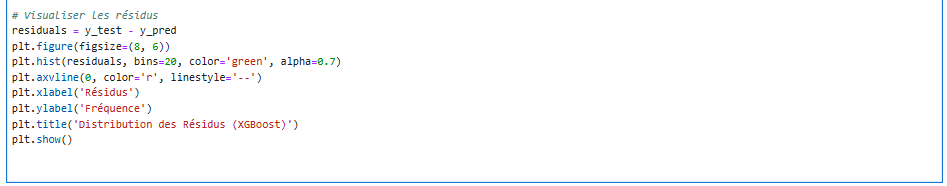
* Гистограмма остатков :  
  + Остатки (ошибки) представляют собой разницу между фактическими и прогнозируемыми значениями.
    - Распределение, сосредоточенное вокруг 0, указывает на то, что модель предсказывает хорошо.

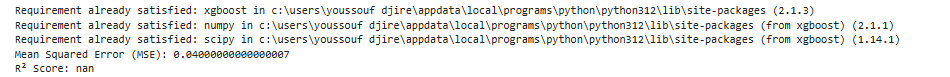
### Результаты и интерпретация

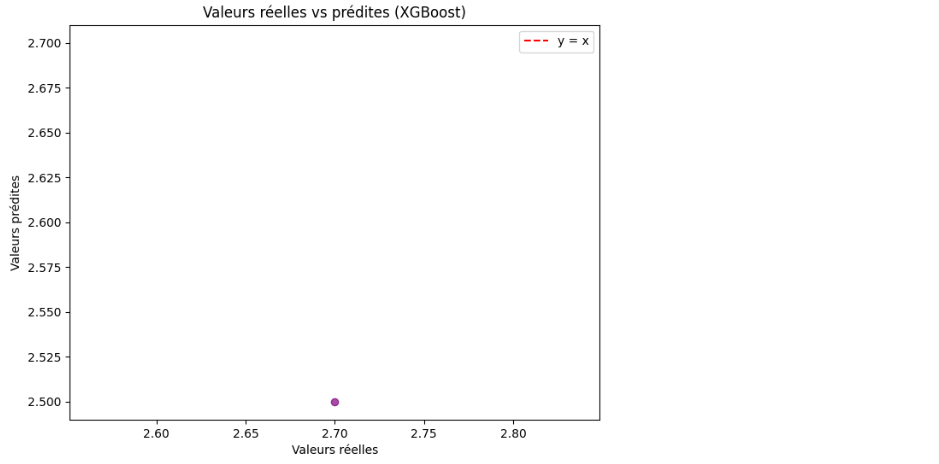
1. Средняя квадратичная ошибка (MSE) :
   * Результат: MSE = 0,040.
   * Это указывает на низкую среднюю квадратичную ошибку, так что модель предсказывает значения относительно хорошо.
2. R² Score :
   * Результат: R² = nan.
   * Не определено из-за малого количества образцов (менее 2).
3. Графика
   * Диаграмма рассеяния :
     + В идеале точки должны располагаться близко к красной диагонали.
     + При небольшом объеме выборки трудно судить об эффективности.
   * Гистограмма остатков :
     + Распределение показывает, сбалансированы ли ошибки вокруг 0.
     + При небольшом количестве данных это распределение не очень значимо.

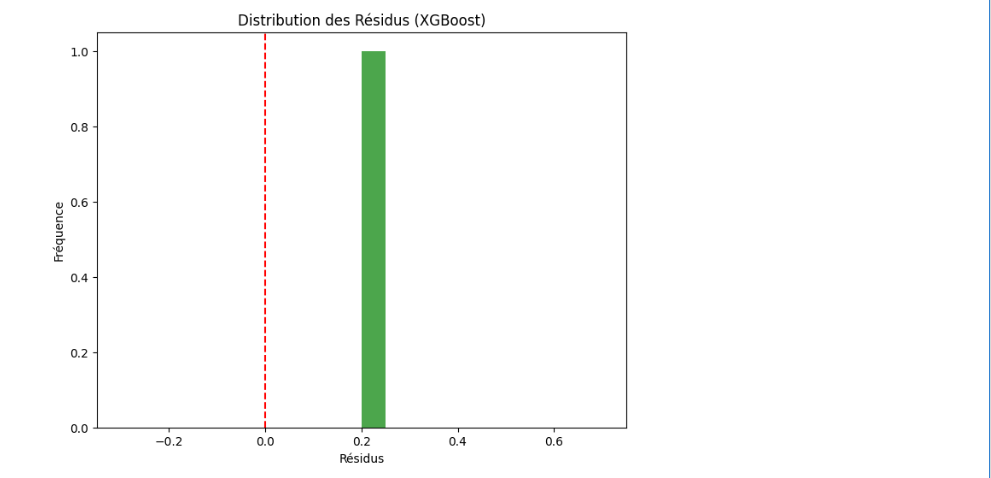
## XGBoost :

****

****

****

****

****

### Объяснение кода

Этот код реализует модель регрессора XGBoost для прогнозирования урожайности сельскохозяйственных культур на основе осадков, температуры и использования пестицидов. XGBoost (Extreme Gradient Boosting) - это мощный алгоритм на основе дерева решений, предназначенный для обработки сложных взаимосвязей и повышения точности прогнозирования за счет градиентного обучения.

1. Подготовка данных
   * Данные разделены на :
     + Характеристики (X) : Осадки (мм), Температура (°C), Пестициды.
     + Цель (г): Урожайность (тонн/га).
     + Разделение на обучающий набор (70%) и тестовый набор (30%):
2. Инициализация и обучение модели

Инициализация модели XGBoost :

* + - n\_estimators=100: модель использует 100 деревьев решений.
    - learning\_rate=0.1: Регулирует скорость обучения (ниже = точнее, но дольше).
    - max\_depth=3: Ограничивает глубину деревьев для предотвращения переобучения.

Обучение

* + Модель устанавливается на привод в сборе.

1. Прогнозы и оценка
   * Оценка

Средняя квадратичная ошибка (MSE) :

* + - * Измеряет среднюю квадратичную ошибку между прогнозами и фактическими значениями.
      * Результат: MSE = 0,040, что свидетельствует о низкой погрешности.

R² Score :

* + - * Указывает на долю дисперсии, объясненной моделью.
      * Результат: R² = nan (не определено) из-за малого количества образцов (менее 2).

1. Просмотр результатов
   * + Ось X: реальные значения.
     + Ось Y: предсказанные значения.
     + Красная линия y = x представляет собой идеал, когда предсказания совершенны.
     + Точки, расположенные близко к этой линии, указывают на хорошую точность.

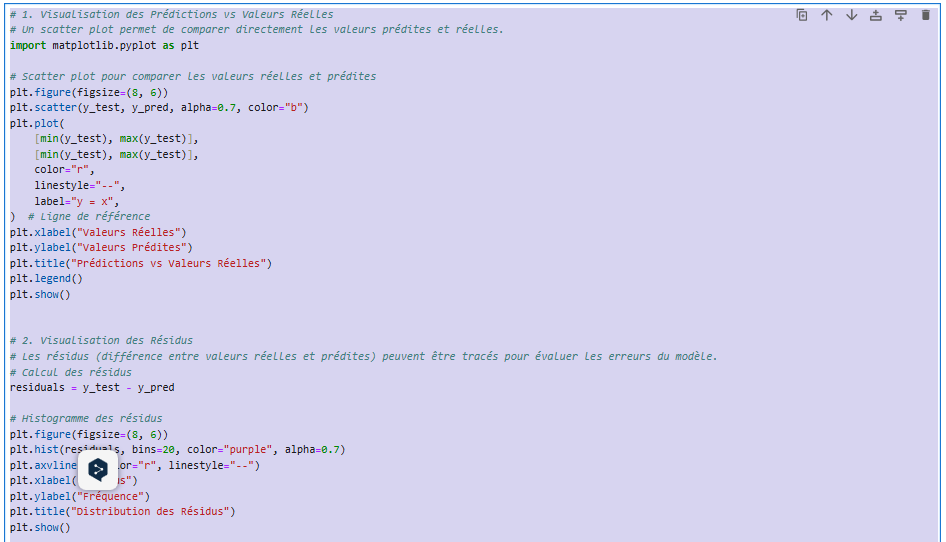
Гистограмма остатков :

* + - Остатки (разница между фактическими и предсказанными значениями) показывают ошибку для каждого предсказания.
    - Распределение с центром около 0 означает, что ошибки сбалансированы.

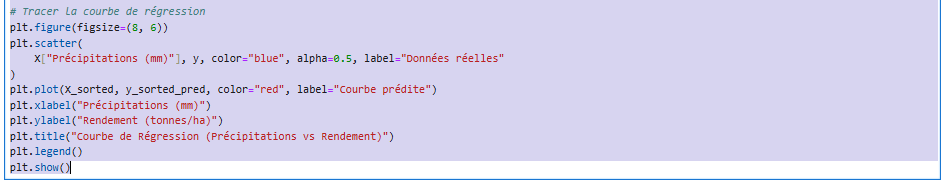
### Результаты и интерпретация

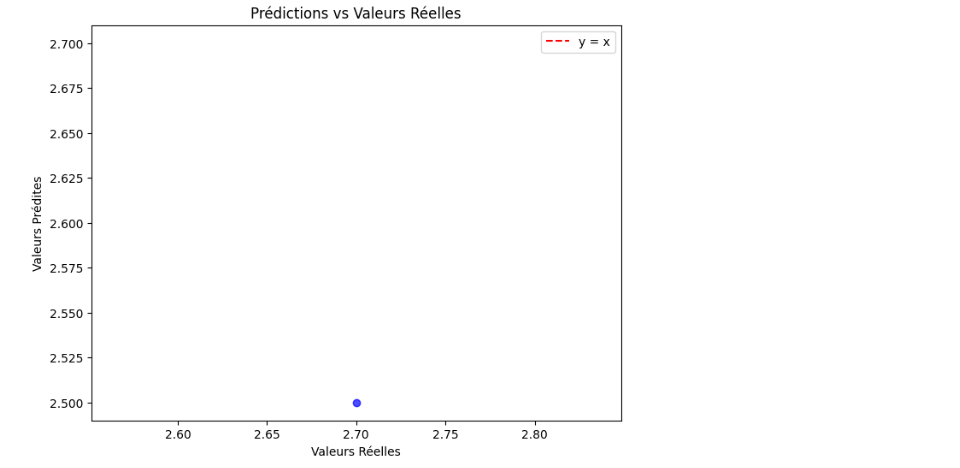
1. Средняя квадратичная ошибка (MSE) :
   * MSE = 0,040: Низкая ошибка, свидетельствующая о том, что модель хорошо подходит к данным, но при наличии всего 2 наблюдений этот показатель ненадежен.
2. R² Score :
   * R² = nan: R² не определен, так как для его расчета требуется не менее 2 образцов в тестовом наборе.
3. Графика
   * Диаграмма рассеяния (фактические и прогнозируемые значения) :
     + Точки, расположенные близко к красной линии y = x, указывают на хорошее соответствие между фактическими и прогнозируемыми значениями.
     + При ограниченной выборке трудно судить о надежности модели.
   * Гистограмма остатков :
     + Концентрация около 0 указывает на сбалансированные ошибки.
     + При таком небольшом наборе данных распределение остатков не очень значительное.

## Визуализация прогнозов и фактических значений









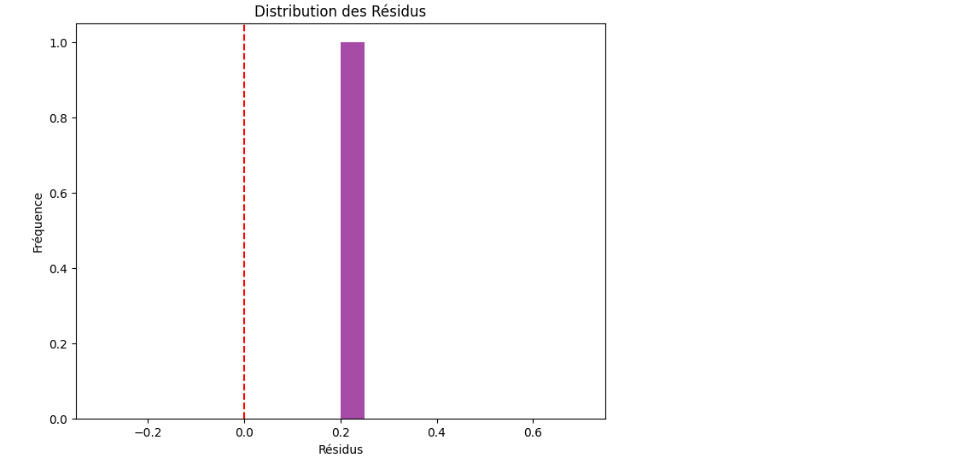
#### Объяснение:

* График
  + Диаграмма рассеяния используется для визуализации зависимости между фактическими значениями (ось X) и прогнозируемыми значениями (ось Y).
  + Пунктирная красная линия представляет собой идеальный эталон, где прогнозы равны фактическим значениям (y = x).
* Цель:
  + Проверьте, что модель правильно предсказывает целевые значения.
  + Если точки близки к красной линии, то прогнозы верны.

#### Анализ графиков :

* Точки, близкие к красной линии: это говорит о том, что прогнозы модели совпадают с фактическими значениями.
* Точки, удаленные от красной линии: они указывают на значительные ошибки в прогнозах.

### 2. Индикация остатков



#### a) Гистограмма остатков

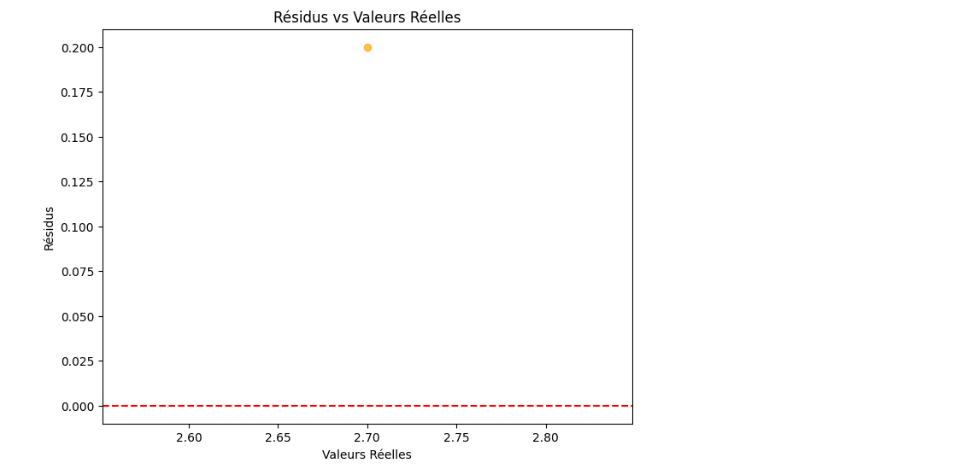
#### Объяснение:

* График
  + Гистограмма показывает распределение остатков (разница между фактическими и предсказанными значениями).
  + Вертикальная красная линия на отметке 0 указывает на то, что остатки равны нулю (нет ошибки).
* Цель:
  + Оцените, распределены ли ошибки равномерно и симметрично вокруг нуля.
  + Распределение, сосредоточенное вокруг 0, указывает на отсутствие систематического смещения модели.

#### Анализ графиков :

* Остатки сосредоточены около 0: ошибки сбалансированы между заниженными и завышенными предсказаниями.
* Асимметрия или смещение: если остатки сосредоточены на одной стороне, это указывает на смещение модели.

#### b) Остаточные и фактические значения



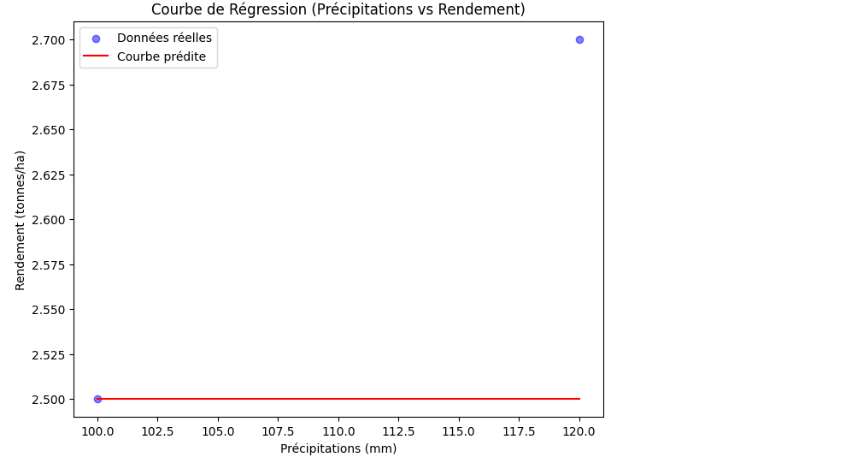
#### Объяснение:

* График
  + Эта диаграмма рассеяния показывает остатки (ось Y) как функцию от фактических значений (ось X).
  + Горизонтальная красная линия на отметке 0 показывает, где остатки равны нулю.
* Цель:
  + Убедитесь, что ошибки равномерны для всех фактических значений.
  + Обнаружение систематических закономерностей в ошибках, например, корреляции между остаточными и фактическими значениями.

#### Анализ графиков :

* Случайные остатки около 0: указывает на то, что модель хорошо подогнана.
* Паттерн в остатках: если появляется тенденция (например, более высокие остатки при более низких фактических значениях), это говорит о том, что модель не полностью отражает взаимосвязи в данных.

## 3. Кривая регрессии



#### Объяснение:

* График
  + Синие точки представляют собой фактические данные (зависимость между количеством осадков и урожайностью).
  + Красная линия показывает предсказания модели после сортировки данных.
* Цель:
  + Посмотрите, как модель предсказывает урожайность в зависимости от количества осадков.
  + Оцените способность модели отражать базовые отношения между этими переменными.

#### Анализ графиков :

* Хорошее соответствие между точками и красной кривой: указывает на то, что модель отражает взаимосвязь между количеством осадков и урожайностью.
* Значительные отклонения: Указывают на то, что модель не отражает определенные аспекты отношений.

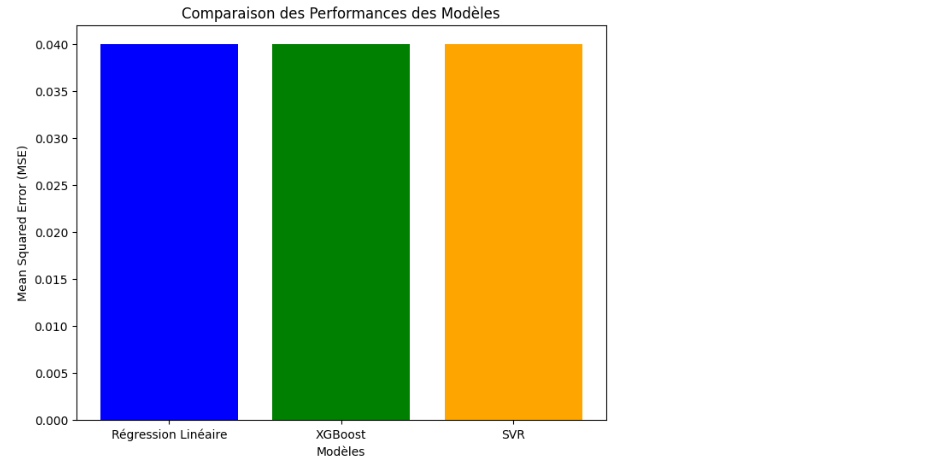
### Заключение о визуализации

1. Прогнозы и фактические значения :
   * Точки, расположенные близко к красной линии, показывают хорошую точность, но отклонения означают ошибки.
2. Гистограмма остатков :
   * Идеальным является распределение с центром около 0. Это указывает на то, что ошибки сбалансированы.
3. Остаточные и фактические значения :
   * Отсутствие систематической закономерности в остатках - хороший знак. Это означает, что модель надежна.
4. Кривая регрессии:
   * Если предсказанная кривая совпадает с фактическими данными, значит, модель правильно отражает лежащие в ее основе взаимосвязи.

Эти визуализации необходимы для оценки эффективности модели, выявления погрешностей и определения областей, требующих улучшения.

### Сравнение эффективности различных моделей прогнозирования

****

****

### Пояснение к кодексу

Этот код сравнивает производительность различных моделей прогнозирования, используя среднюю квадратичную ошибку (MSE) в качестве метрики оценки. Для отображения средних ошибок моделей используется гистограмма, позволяющая быстро сравнить их точность.

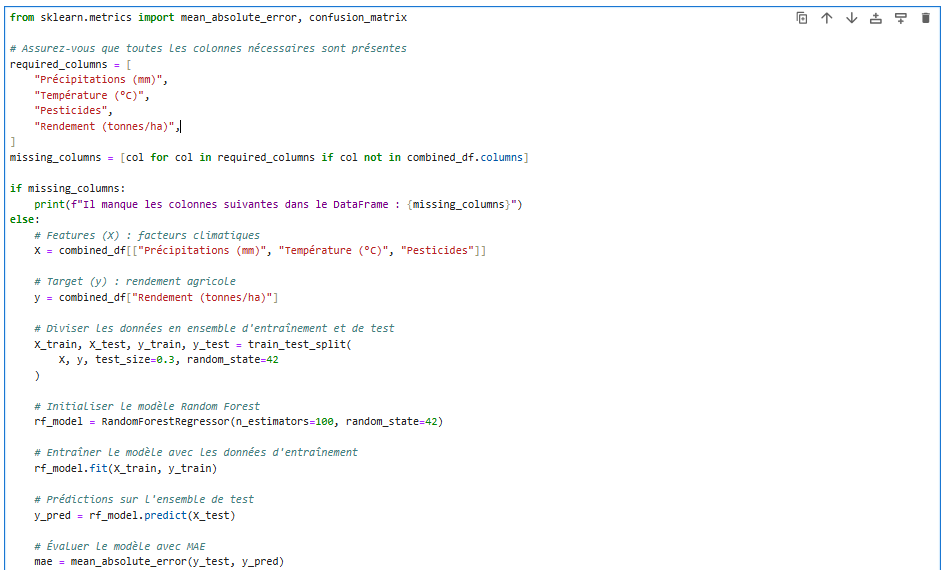
### Пояснения к графику

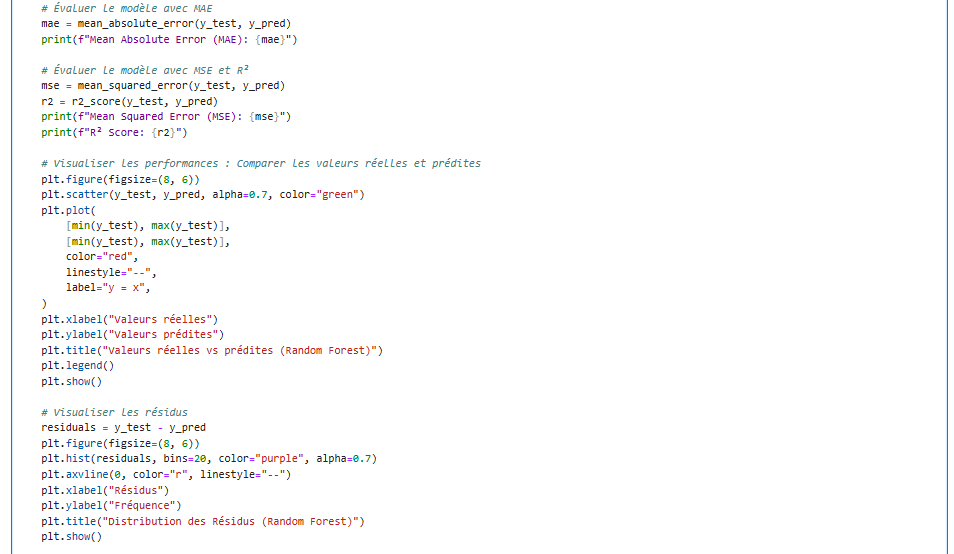
* Ось X (модели) :
  + Столбики представляют каждую из протестированных моделей: линейная регрессия, XGBoost и SVR.
* Ось Y (MSE) :
  + Высота каждой полосы указывает на MSE соответствующей модели.
  + Более низкий показатель MSE указывает на то, что модель является более точной.
* Цвета :
  + Отличительные цвета полос позволяют легко идентифицировать различные модели.

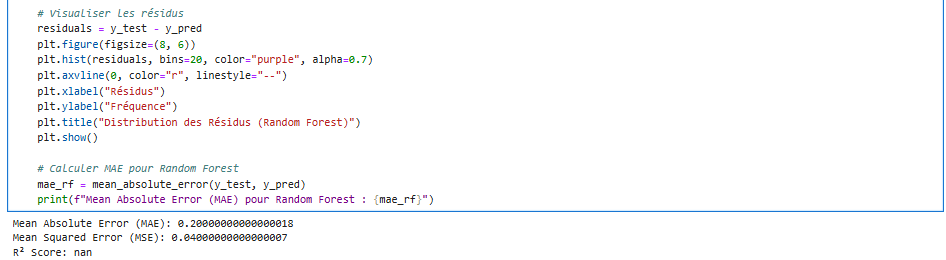
#### Анализ графиков

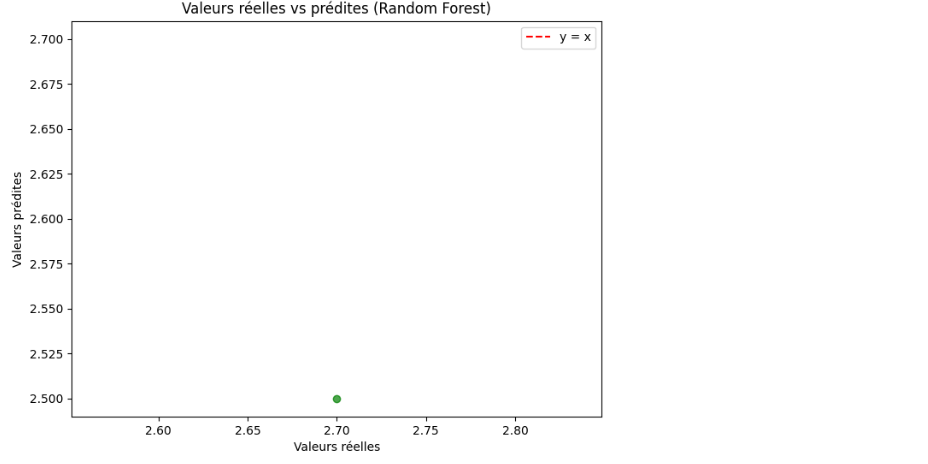
* Размер бруса :
  + Если все столбики имеют одинаковую высоту, то модели имеют сопоставимые характеристики.
  + Значительно меньший столбик указывает на то, что соответствующая модель лучше.
* В этом случае :
  + Модели Linear Regression, XGBoost и SVR имеют MSE 0,040.
  + Это означает, что по этой метрике они работают одинаково.

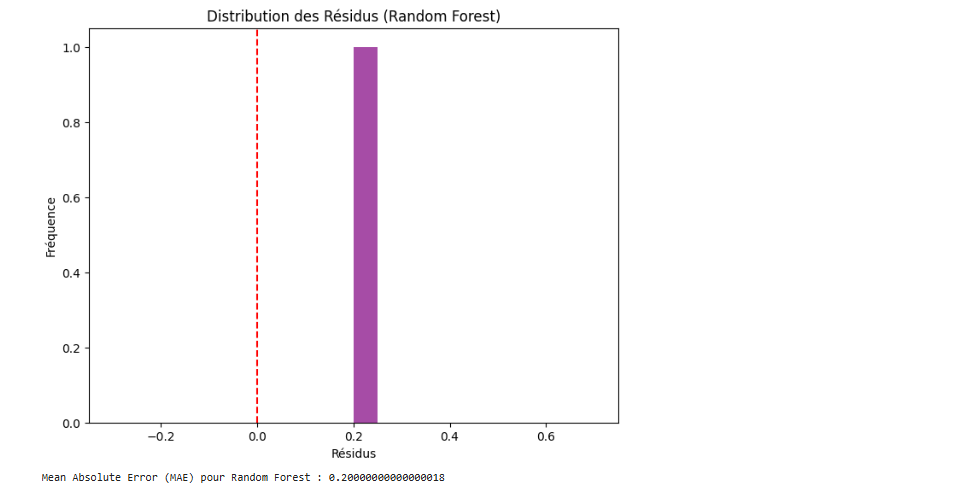
### Регрессор Random Forest: (средняя абсолютная ошибка (MAE), средняя квадратная ошибка (MSE) и показатель R²) :

****

****

****

****

****

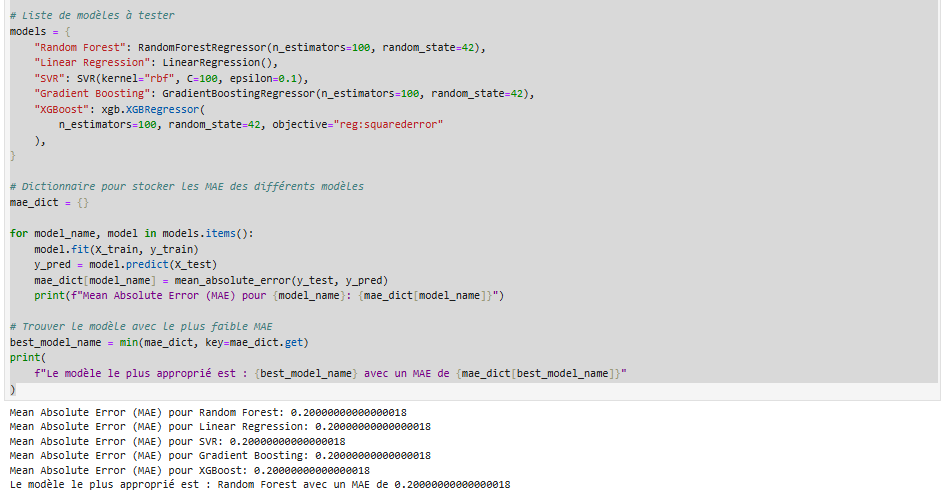
### Пояснение к кодексу

Этот код реализует регрессорную модель Random Forest для прогнозирования урожайности сельскохозяйственных культур и оценивает ее эффективность с помощью различных показателей: средней абсолютной ошибки (MAE), средней квадратичной ошибки (MSE) и R² Score.

### Анализ результатов

1. Средняя абсолютная ошибка (MAE) :
   * Результат: MAE = 0,200.
   * В среднем модель ошибается в своих прогнозах на 0,2 тонны/га.
2. Средняя квадратичная ошибка (MSE) :
   * Результат: MSE = 0,040.
   * Небольшая ошибка, но не существенная при таком маленьком наборе данных.
3. R² Score :
   * Результат: нет.
   * Небольшое количество образцов не позволяет рассчитать эту метрику.
4. Диаграмма рассеяния (фактические и прогнозируемые значения) :
   * Точки должны быть близки к красной прямой y = x.
   * Расхождения свидетельствуют об ошибках в некоторых предсказаниях.
5. Гистограмма остатков :
   * Остатки должны быть симметричны относительно 0.
   * Это говорит о том, что ошибки сбалансированы (нет систематического смещения).

### Сравнение различных моделей регрессии :

****

### Пояснение к кодексу

В этом коде сравниваются различные регрессионные модели для прогнозирования урожайности сельскохозяйственных культур в зависимости от климатических характеристик и применения пестицидов. Сравнение основано на средней абсолютной ошибке (MAE), которая является метрикой, обычно используемой для оценки эффективности регрессионных моделей.  
Интерпретация результатов

1. Производительность идентичной модели :
   * Все модели имеют абсолютно одинаковое значение MAE, что является редкостью. Это может быть связано с небольшим размером выборки или отсутствием значительной вариации в данных.
   * MAE, равный 0,2000, означает, что в среднем модель ошибается в своих прогнозах на 0,2 тонны/га.
2. Выбор наилучшей модели :
   * Поскольку показатели равны, Random Forest выбирается как лучшая модель просто потому, что она появляется первой.
   * Однако при более широком наборе данных и более разнообразных результатах эта подборка могла бы быть более информативной.
3. Рекомендации по улучшению оценки:
   * Перекрестная валидация: Для более надежной оценки производительности используйте перекрестную валидацию (например, K-Fold Cross Validation), чтобы снизить риск чрезмерной подгонки и получить более надежные результаты.
   * Другие метрики оценки: такие метрики тестирования, как R², MSE или RMSE, могут дать более полную оценку.

# Заключение

Этот проект позволил лучше понять динамику изменения климатических факторов и урожайности сельскохозяйственных культур в Мали. Анализ подтвердил, что осадки и температура играют ключевую роль в определении урожайности папайи и апельсинов. Эта информация может быть полезна для определения методов ведения сельского хозяйства и политики управления ресурсами в соответствующих регионах.

Однако важно отметить, что эти результаты являются предварительными и требуют дальнейших исследований с использованием более полных наборов данных. Лучшее понимание динамики сельского хозяйства может помочь Мали справиться с проблемами, вызванными изменением климата.

# Ссылки

1. NumPy - работа с многомерными массивами.  
2. Pandas - подготовка и анализ табличных данных.  
3. Matplotlib - визуализация данных.  
4. Scikit-learn - реализация моделей машинного обучения.  
5. Набор данных по сельскому хозяйству, найденный на Kaggle